Faculty of Engineering and Technology

Faculty Research Week
2016 Proceedings

9th – 13th May 2016

Faculty of Engineering and Technology
James Parsons Building, Byrom Street, Liverpool, L3 3AF

MMXVI

Contents

Magomed Muradov

Online monitoring of meat drying process using microwave spectrum.

Magomed Muradov*, Alex Mason, Jeff Cullen, Andrew Shaw and Ahmed I. Al-Shamma'a
Built Environment and Sustainable Technologies (BEST) Research Institute School of Built Environment
Liverpool John Moores University
Liverpool, UK
M.Muradov@2009.ljmu.ac.uk*

Abstract. The aim of this investigation is to monitor the meat drying process and analyse the change in electromagnetic (EM) signature from a bespoke sensor during the process. The sensor has been modelled using High Frequency Structure Simulation Software (HFSS) and then constructed and tested. Experimental work was conducted, involving measurement of meat weight and EM signature (namely the reflected signal in the 1-6GHz frequency range) over a period of one week (it takes approx. 1 week to lose 40% of weight and then the measuring is stopped as the general weight loss of dry-cured meat is 30-35 % in the final product), with measurements recorded once per hour. The change in EM signature and weight loss has been analysed and correlations drawn from the resultant data. The results demonstrate a strong linear relationship (R^2 = 0.98) between the reflection coefficient and weight loss of the meat sample, and it is proposed that this could be used as the basis for future industrial application for measuring the quality of meat products during drying processes, such as those used in curing. The ability of this sensor system to monitor the water content of cured meat could have huge implications for the meat industry, which currently lack appropriate tools for inspection of products in-situ during the processing stages.

Keywords. Dry-curing; electromagnetic wave; non-invasive; sensor; real-time meat analysis.

1. Introduction

Products of certified high quality are increasingly sought by both consumers and manufacturers [1]. Products in the meat industry are no exception to this rule [2, 3] although it is a challenging task since the product exhibits considerable variability as a result of the natural raw material [4]. Meat curing is the process by which food products are preserved and flavoured, typically through the addition of salts, nitrates, nitrites or sugars and perhaps in combination with other cooking or smoking processes. In some countries, such products have lost favour due to widespread use of mechanical refrigeration systems which means that fresh meat produce can be readily purchased in local stores and supermarkets, and indeed very few people can remember a time before refrigeration or when salt was regularly used to preserve meat [5].

However, across much of Europe (e.g. Spain, France and Germany) cured meat products have returned a place in the consumer market, and so the industry is interested to find ways of improving the process. Despite curing being used as early as 1300 BC in China, there are few techniques for monitoring the curing process with producers typically relying upon "rule of thumb" methods or crude measurements such as product weight or solidity. While fresh meat products have benefitted from the advent of sensor systems, with near-infrared and x-ray based techniques being the most prevalent for applications such as meat composition monitoring [6], sorting [7] and foreign object detection (i.e. detection of metal, plastic and bone shards), these technologies have made little impact on monitoring of cured meat production. Thus there is heavy reliance upon the experience of key workers within the industry which leads to problems of consistency if the instincts or opinions of those workers vary [8].

Electromagnetic (EM) sensors, namely those which operate at radio or microwave frequencies, are widely used in a variety of industrial sectors. Examples include, civil engineering materials analysis [9], timber imaging [10], chemical processing [11, 12] and medicine [13, 14]. However, there has yet to be a significant impact of these sensors in the food industry, with the majority of such technology there being centred on cooking or sterilisation. However, the authors have demonstrated the first potentially non-invasive solution for measurement of water holding capacity [15] and water activity [16] in meat products based on EM wave sensors operating at microwave frequencies, and so there is the belief that further benefits could be brought to the cured meat industry; this belief is supported by other authors in the field [17]. The particular benefit of the technology is its non-invasive nature, which, in contrast with current techniques, holds the potential to eliminate fears over instrumentation contributing to food contamination which has disastrous consequences for consumers and suppliers. Thus the purpose of this study is to demonstrate the potential for EM wave sensors in monitoring cured meat products.

2. Planar Electromagnetic Wave Sensor

For the purposes of this work a patch type sensor is utilised; it's simple and well understood design allows for accurate prediction of sensor performance based up on preliminary models created using Ansys HFSS electromagnetic wave simulation software. The physical model is illustrated in Figure 2.

The sensor consists of 4 key parts; (1) active radiating element, (2) ground plane, (3) substrate and (4) the feed. The dimensions of the sensor illustrated in figure 1 allow the sensor to resonate at approximately 2 GHz. The radiating element and ground plane are copper based conductors, while the substrate material is FR4 with dielectric constant of 4.4 and loss tangent of 0.021. The modelled return loss was in good agreement with the physical measurement, as indicated in figure 2.

(a)   (b)
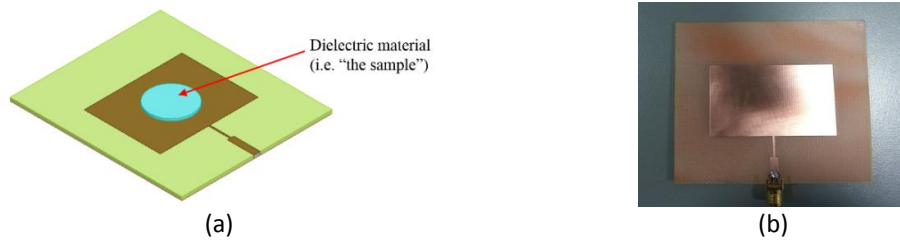
Figure 1. (a) Model of patch type electromagnetic wave sensor, designed to operate at approx. 2 GHz and (b) the constructed sensor.
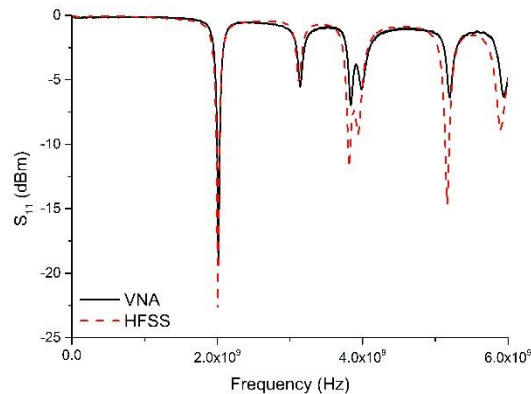
Figure 2. Demonstrating the modelled and measured sensor return loss.

The sensor model, depicted in figure 1(a) and figure 3(a), includes a cylindrical dielectric material which represents a water volume in this work. As this volume was increased, the sensor was parametrically modelled to understand the relationship between the return loss and water volume. As expected, the sensor response changes with the increasing water volume, likely as a result of the changing dielectric constant (ϵr) directly above the radiating element; it is on this basis that the sensor was selected for use in monitoring meat during the curing process.

Generally speaking, sensors of this nature offer a good number of benefits, including being low profile, conformable to planar and non-planar surfaces, simple and cheap to fabricate using modern printed-circuit technology, mechanically robust when mounted on rigid surfaces and very flexible in terms of resonant frequency, polarization, pattern, and impedance [18]. Furthermore, such sensors provide good depth of penetration (> 100 mm) a relatively low power (< 1 mW), which is important given that rapid drying of meat products occurs on surfaces exposed to the atmosphere.

(a)   (b)
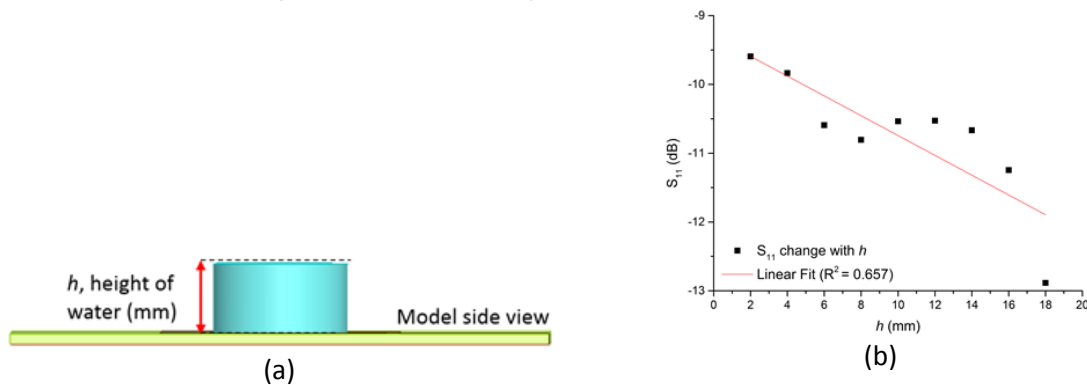
Figure 3. (a) Model of sensor with variable water volume and (b) parametric analysis result with correlation between sensor return loss and water volume (R^2 = 0.657).

3. Methodology

3.1. Sample Preparation

Pork chump ends (24 hours post-mortem) were purchased from a local slaughter house for this investigation. Each meat sample was deboned; fat removed and cut into approximately 100 × 70 × 50 mm pieces. Sliced meat samples are shown in figure 4 (a). The desired weight loss was 40 % of initial weight as, generally, dry-cured meat products have 30-35 % weight loss in the final product [19]. In order to achieve a final salt concentration of approximately 5.5 % in the 30 % weight loss, all pieces were added 3.85 % salt prior to vacuum packing. Then all pieces were stored vacuum packed (figure 4 (b)) for two weeks at 4°C during salting and salt equalization. After salt equalization the pieces were unsealed and placed inside an incubation system at 12-14°C and 72-74 % relative humidity (RH) for microwave measurements until the samples obtain the desired weight loss.

| (a) | (b) |

Figure 4. (a) Sliced meat samples, (b) salted and vacuum sealed sample.

3.2. Experimental Setup

The experimental setup comprises two patch antennas mounted on top two electronic weighing scales. Both sensors are connected to PC via Vector Network Analyser (VNA). The VNA and both weighing scales are connected to a computer for data acquisition via a LabVIEW interface, and are placed inside an incubation system to maintain a consistent temperature (approx. 13°C, ± 1°C).

The sample is placed on a bespoke plastic "water runoff system" (shown in figure 5), designed to prevent water pooling on the weighing scales once lost from the meat

To promote air circulation and promote water loss two fans are fixed inside the incubation system. Both fans are connected to power supply via a relay that is used to switch off the fans while the measurements are taken. The purpose of this is to avoid incorrect weight measurements as the scale is sensitive to small changes in air pressure. The scales are zeroed once the sensors and two "water runoff systems" are fixed to them, before placing the meat samples.

Measurement from the patch sensors is provided by using the S_{11} and S_{22} parameters from the VNA. Data acquisition (i.e. S_{11}, S_{22}, temperature and weight) took place once per hour over a period of 7 days. Weight and microwave measurements were then correlated to determine the relationship between weight loss of the meat and change in EM signature from the sensor.

(a)	(b)
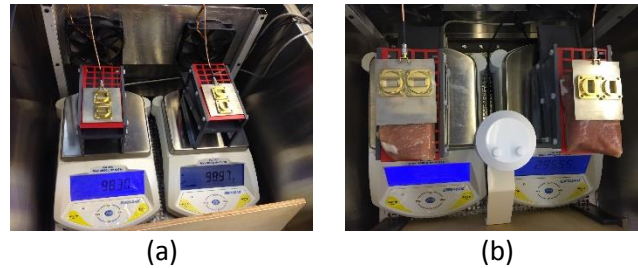
Figure 5. Experimental setup for monitoring of the meat drying process with (a) scales for monitoring weight loss, drip catchers/fans to move excess moisture away from sensor and (b) humidity/temperature sensing to monitor drying condition inside incubation system.

4. Results

Figure 6(a) shows the S_{11} measurements, i.e. S_{11} of sample of pork chump end. The measurements shown in this figure were taken once per hour (i.e. 24 times a day) during one week. Measurements of the weight loss of the sample also were taken at the same time during the week. It can be seen in figure 6(a), that there is a noticeable change in EM signature. The change is thought to be caused by the decreasing amount of water in the meat sample. The sample was not touched or moved during the experimental work, and all other conditions, such as temperature and light remained nominally the same during the test. Figure 6(b) shows that weight loss of the sample in our experimental environment can typically be represented by a linear model; the model used here results in $R^2 > 0.98$ demonstrating a good fit.

(a)	(b)
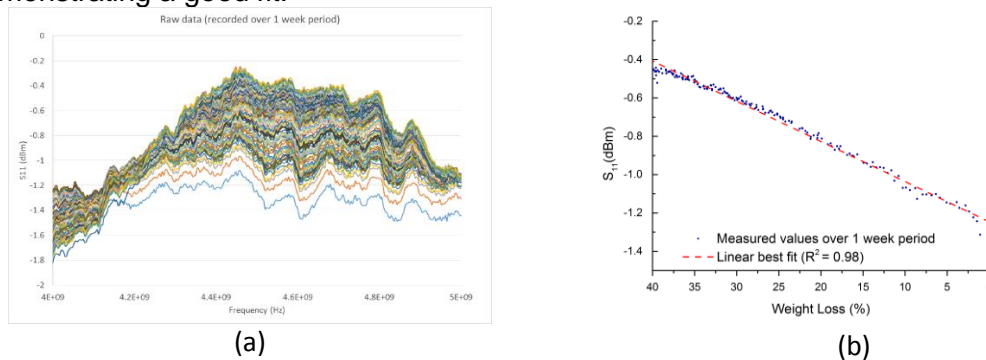
Figure 6. (a) Readings from the electromagnetic wave sensor; measurements were taken once per hour; (b) correlation of weight loss and S_{11} at 4.5 GHz, with $R^2 = 0.98$.

5. Conclusion

The purpose of this work was to demonstrate the capability of electromagnetic wave sensors operating at low GHz frequencies for monitoring the meat curing process. Through a methodology, developed in partnership with experts from the Norwegian food industry, this work has taken a valuable first step in demonstrating the capability of the sensor technology. In particular, the results demonstrate a good correlation ($R^2 = 0.98$) between the weight lost by meat samples, with weight lost being equated to water loss, and the S_{11} output of the sensor utilized.

The ability of this sensor system to monitor the water content of cured meat could have huge implications for the meat industry, which currently lack appropriate tools for inspection of products in-situ during the processing stages. While there have been significant improvements over recent years with the introduction of X-Ray scanners, near-infrared and optical systems, they currently do not provide the opportunity for continuous monitoring of water content, perhaps due to cost, safety or practicality. Alternative testing methods tend to be destructive, either through the use of invasive probes, which pose a contamination risk, or the sampling of meat for short time-based studies. Since the water content of meat has

impact on the saleable value in addition to is quality and shelf-life, the availability of tools to monitor water content, and perhaps more importantly, the loss of water, is certainly desirable.

While this is an issue that is present across the meat industry (i.e. not only the cured meat sector), the reliance on "rule of thumb" or processing techniques passed on through family generations precludes advancement which could improve the sustainability of cured meat production. In particular it is suggested by some in the industry that significant savings in time and energy could be made through knowing at which stage in the curing process a product is. Given that the curing process could well be taking place over days, weeks or even months, knowing whether a product is ready earlier than expected allows it to be shipped for sale and the precious temperature and humidity controlled environments can be utilized effectively for new product. Such information could be derived from knowing the product water content.

There is still considerable work to be conducted in this area, particularly to prove the technology in a real-world scenario. Working with partners in Norway and Spain, the authors plan to do just that and test the technology on a range of meat products, particularly lamb and pork. Given the success so far at relatively low GHz frequencies, the outlook for the future is optimistic as the electronics to generate EM waves at these frequencies is abundant. This means that the technology could be made portable and be presented in a number of different forms for instantaneous and continuous monitoring purposes across the industrial and research domains.

References

[1]   W. Verbeke, "Consumer attitudes and communication challenges for agro-food technologies," *AGRO FOOD INDUSTRY HI-TECH*, vol. 22, no. 5. pp. 34–36, 2011.

[2]   W. Verbeke, L. Van Wezemael, M. D. de Barcellos, J. O. Kügler, J.-F. Hocquette, Ø. Ueland, and K. G. Grunert, "European beef consumers' interest in a beef eating-quality guarantee Insights from a qualitative study in four EU countries.," *Appetite*, vol. 54, no. 2, pp. 289–96, Apr. 2010.

[3]   W. Verbeke, L. Van Wezemael, M. D. de Barcellos, J. O. Kügler, J.-F. Hocquette, Ø. Ueland, and K. G. Grunert, "Future trends and consumer lifestyles with regard to meat consumption.," *Meat Sci.*, vol. 74, no. 1, pp. 149–60, Sep. 2006.

[4]   J. L. Damez and S. Clerjon, "Quantifying and predicting meat and meat products quality attributes using electromagnetic waves: An overview," *Meat Sci.*, vol. 95, no. 4, pp. 879–896, Dec. 2013.

[5]   G. Rentfrow, R. Chaplin, and S. P. Suman, "Technology of dry-cured ham production: Science enhancing art ."

[6]   Prediktor, "Spektron," 2016. [Online]. Available: http://www.prediktor.no/. [Accessed: 01-Apr-2016].

[7]   Tomra, "QVision," 2016. [Online]. Available: https://www.tomra.com/en/. [Accessed: 01-Apr-2016].

[8]   SWISS MEAT, "Production of Meat Specialties," 2014. [Online]. Available: http://www.schweizerfleisch.ch/en/swiss-meat/handling/production-of-meat-specialties/. [Accessed: 18-May-2014].

[9]   M. Adous, P. Quéffélec, and L. Laguerre, "Coaxial/cylindrical transition line for broadband permittivity measurement of civil engineering materials," *Meas. Sci.*

Technol., vol. 17, no. 8, pp. 2241–2246, Aug. 2006.

[10] B. Goy, P. Martin, and J.-M. Leban, "The measurement of wood density by microwave sensor," *Holz als Roh- und Werkst.*, vol. 50, no. 4, pp. 163–166, Apr. 1992.

[11] L. Gradinarsky, H. Brage, B. Lagerholm, I. N. Björn, and S. Folestad, "In situ monitoring and control of moisture content in pharmaceutical powder processes using an open-ended coaxial probe," *Meas. Sci. Technol.*, vol. 17, no. 7, pp. 1847–1853, Jul. 2006.

[12] O. Korostynska, A. Mason, M. Ortoneda-Pedrola, and A. Al-Shamma'a, "Electromagnetic wave sensing of NO3 and COD concentrations for real-time environmental and industrial monitoring," *Sensors Actuators B Chem.*, vol. 198, pp. 49–54, Jul. 2014.

[13] O. Korostynska, A. Mason, and A. Al-Shamma'a, "Microwave sensors for the non-invasive monitoring of industrial and medical applications," *Sens. Rev.*, vol. 34, pp. 182–191, 2014.

[14] J. H. Goh, A. Mason, O. Korostynska, A. I. Al-Shamma'a, P. Browning, and M. Field, "Real-Time Monitoring of Bodily Fluids Using a Novel Electromagnetic Wave Sensor," *J. Public Heal. Front.*, vol. 2, no. 4, pp. 201–206, 2013.

[15] B. M. Abdullah, J. D. Cullen, O. Korostynska, A. Mason, and A. I. Al-Shamma'a, "Assessing Water-Holding Capacity (WHC) of Meat Using Microwave Spectroscopy," *Sens. Technol. Curr. Status Futur. Trends I*, vol. 7, pp. 117–140, 2014.

[16] S. G. Bjarnadottir, K. Lunde, O. Alvseike, A. Mason, and A. I. Al-Shamma'a, "Assessing Quality Parameters in Dry-Cured Ham Using Microwave Spectroscopy," *Meat Sci. In-Press*, 2015.

[17] S. Clerjon and J. L. Damez, "Microwave sensing for an objective evaluation of meat ageing," *J. Food Eng.*, vol. 94, no. 3–4, pp. 379–389, Oct. 2009.

[18] B. A. Constantine, *ANTENNAS THEORY. Analysis and Design*, 3rd ed. New Jersey: John Wiley & Sons, Inc., 2005.

[19] P. J. Fellows, *Water activity*, 2nd ed. Woodhead Publishing Limited, 2000.

The feasibility of using electromagnetic waves in determining moisture content of building fabrics

P. Kot[*], A. Shaw, K. O. Jones, M. Riley, A. Mason and A. I. Al-Shamma'a

Built Environment and Sustainable Technologies (BEST) Research Institute
School of Built Environment
Liverpool John Moores University
Liverpool, UK
P.Kot@ljmu.ac.uk[*]

Abstract. The purpose of this research is to identify the moisture content of building fabrics. Increased amount of moisture content will influence the overall performance of building structures. If building's structural issues are not recognised at an early stage this may result in large refurbishment costs as well as an increased risk of building collapse, which may cause residents' injuries or even death. Currently moisture content is measured by surveyors using commercially available devices however these techniques require taking a sample of material from the building structure to analyse. The aim of this research is to develop a novel non-destructive electromagnetic wave sensor to determine the moisture content of the building fabrics in real time. During experimental work a soaked ceiling plasterboard was monitored by the microwave sensor placed in front of the measured material. The weight of the sample and microwave spectrum was recorded and the data was analysed by correlating two measurements. The results from these experiments provided a strong relationship between reflected power and weight loss of the material. The potential of this new microwave sensor will provide surveyors with a time efficient and non-destructive technique to determine the moisture content of building fabrics.

Keywords. Building materials; electromagnetic waves; moisture content; non-invasive.

1. Introduction

Building constructions represents one of the largest, single investment of national resources. The wellbeing of such economies is therefore heavily dependent upon the satisfactory performance of its constructions [1]. Building materials are the material foundation for all construction engineering. Various buildings and structures are constructed by all kinds of buildings materials on the basis of reasonable design. The varieties, specifications and qualities of building material are directly related to the applicability, artistry and durability of buildings and also the cost of projects [2]. Failures in the built environment occur for a variety of reasons and on a variety of scales. Moisture is the primary factor of deterioration and,

combined with diurnal temperature variation, has the greatest influence upon the overall performance of building materials. Moisture can influence external walling in all its states, i.e. as a solid (ice, snow); liquid (wind-driven rain) and gas (water vapour). Most constructional defects, e.g. movement, cracking, fungal attack, chemical reaction, are initiated and aggravated by the presence of moisture [3]. It is important to improve the overall performance of building structures to avoid large refurbishment costs as well as an increased risk of building collapse, which may cause injury or even death to occupants [4]. Algo Centre Mall, a retail hub for Elliot Lake, Ontario, was constructed in 1980. The two story structure had 190,000 square feet of floor space occupied by commercial units as well as a hotel. A prominent feature of the complex was a roof top parking deck. Throughout the life of the mall there had been reports of water leaking into the occupied units. This water presence and damage had been so severe that drip tarps were put up in plain sight of the customers. This water infiltration lead to a variety of serviceability problems and, eventually, to businesses leaving the mall. The collapse occurred on June 23, 2012. The chemical process of corrosion has several parts. In this case a metallic surface was exposed to an electrolyte. The de-icing salts acted as the electrolytes, and had been dissolved into solution with the presence of water. The solution then migrated through the roof top parking deck and settled on a connection of the frame. After a lengthy exposure time, the chemical process degraded the frame connection and led to collapse [5].

2. Literature Review

2.1 Moisture Content of building fabrics

Most construction materials such as plasters, bricks, concrete, timber and some insulation materials such as Rockwool are porous, with their relative porosity being related to their density. Those materials function well within manufacturer specified levels of moisture content, which depends on their location during the construction process as well as the function they are required to carry out throughout the building lifecycle. It is vital to identify the source of the moisture ingress to terminate it and resolve resultant dampness issues. The source of moisture ingress is responsible for the significant amount of the moisture present within building material and can be identified by the process called elimination or using an appropriate monitoring process [6]. It is significant to identify the source of the moisture in order to terminate the source and solve dampness problems [7]. The source is responsible for the greatest amount of the moisture present which can be identified by an elimination process or by a monitoring process. Sources of moisture in buildings can be various, and some examples are given in Table 1 [6].

Table 1. Sources of moisture in buildings.

Moisture source	Typical examples
Air moisture condensation	Cooking
	Heating
Penetrating dampness	Rainwater, snow
	Leaking external rainwater goods
Internal plumbing leaks	Long-term breakdown of plumbing joints
	Burst pipes
Below-ground moisture	Underground plumbing leaks
	Leaking ponds

2.2 Current methods to determine moisture content

2.2.1 Radiological Measurements

• Neutron Method

Neutrons interact mainly with hydrogen nuclei and give a direct measurement of water content by volume. A measurement of thermal neutron density in the vicinity of a neutron source will be a measure of the concentration of hydrogen nuclei on a volume basis, and thus a measure of the water concentration. In this method the source and detector are placed in a probe. The neutron source is mounted in a lead shield at the bottom of the gas filled detector tube. The unit is constructed to be set over an access hole in the porous material, and the probe lowered through the bottom of the shield into the hole. The resolution of the probe is low however, and it is not possible to measure accurately the water content within 6 in. (150 mm) of the surface [8, 9].

• Gamma Scattering Method

Measuring the attenuation of gamma radiation enables us to determine the density of the material. The essential components of the gamma ray method are a source of gamma rays, a detector, an amplifier, a discriminator and a device to record the pulses from the detector. The gamma ray method offers the greatest accuracy in the determination of moisture content, at a level of ±1.0 per cent. There is no hysteresis and the method has a short measurement time. There is also no effect from temperature change or dissolved conducting materials. Unfortunately, there is a high capital cost and precautions are necessary with radioactive equipment. The gamma scattering method determines the moisture content of concrete by measuring the intensity of the scattered radiation in the limestone concrete. It uses an ingeniously built goniometer and an HPGe spectrometer [10, 11]

2.2.2 Ultrasonic Methods

Non-destructive sonic and ultrasonic testing methods have been used in recent years in the assessment of civil engineering structures and materials. The sonic method refers to the transmission and reflection of mechanical stress waves through a medium at sonic and ultrasonic frequencies. The most commonly used sonic techniques are: Sonic transmission method, Sonic/seismic tomography, Sonic/seismic reflection method, Ultrasonic reflection method and sonic resonance method. Detection of flaws by the sonic transmission method is possible since sonic waves cannot transmit across an air gap which could be caused owing to a crack, a void or delamination at the interface between brick or stone and mortar. Measurement has to be carried out in dry conditions. A propagating wave must find a path around the void, resulting in attenuation and an increase in the transit time of the signal. Data scatter from sonic tomography has the effect of increasing the residual of tomographic velocity reconstruction and may lead to the identification of false anomalies. The accuracy of the velocity reconstruction can be improved by a better understanding of the input signals or by carefully planning the choice of position and number of reading stations. The Sonic reflection method is not one currently recommended since the resolution achievable with its low frequency is poor, and it is often difficult to distinguish reflections from surface waves and refracted arrivals [12, 13].

2.2.3 Electrical Measurements

• Electrical resistance measurement

Electrical resistance measurements are based on the fact that each material possesses a unique electrical resistance and water content has a direct impact on the electrical resistance of the material (that is, more water means lower resistance). Measurement of the electrical resistance is usually carried out using needle-shaped electrodes. Two measuring detectors are placed by driving or drilling into a building element and electrical resistance as a function of the electrical conductivity is measured. This method for determining the moisture content of building elements is simple and fast [9, 14].

- LTCC sensor

Low Temperature Co-Fired Ceramic (LTCC) sensors with design and fabrication of the inductor capacitor (LC) planar sensors have been applied to monitor moisture content of the most frequently used buildings materials [15]. The sensor can be put retrospectively into the building material through a small hole, or it can be buried in plasterwork (for example) during the building process. Variation of water content in the tested specimens is measured wirelessly, with an antenna coil, tracking changes in the sensor resonant frequency. The sensor exhibits 0 to 70 % wide detection range of water absorption with high linearity and fast response. This measuring method is destructive to building materials [14].

- Time Domain Reflectometry (TDR)

The principle of TDR used for the measurement of moisture is to transmit an electromagnetic signal down the parallel electrodes of a waveguide inserted into the dielectric material under investigation. The time taken for the signal to return after reflecting from the end of the waveguide is measured. This gives a direct measure of the permittivity of the surrounding material along the length of the waveguide [9]. Current state of research common applications of TDR technique for monitoring moisture content in all porous materials, regardless of their type, are not possible yet. The method requires further investigations particularly in the data acquisition process. General formulas for determination of moisture content from measured relative permittivity are not yet available [16].

2.2.4 Physical Property Measurements
- Thermal method

The thermal conductivity of a material increases with increasing moisture content. One of the methods used to determine thermal conductivity is to supply a probe in the material with a known heat input, and to measure the temperature rise at a fixed distance from the heat source using thermocouples or thermistors. Thermal conductivity does depend on the environmental temperature and upon the density of the material, necessitating several calibration curves for different densities of the material measured. The thermal conductivity method suffers from the difficulty in obtaining reproducible calibration curves, and each calibration curve is only valid for a specific density of the material [9, 17]. One thermal technique is the 'pad' sensor. The inventors of this sensor demonstrated a non-destructive technique for the measurement of moisture content within a material. It uses thermal diffusion wherein heat is delivered into the surface and the resulting temperature increase is detected at a distance from the injection point. In the 'pad' sensor method the heater and temperature sensor are fixed at the surface of a thermal insulation material block (the 'pad') [18].

- Vapour pressure

Measurement of the equilibrium relative humidity of air in contact with a porous body enables us to deduce directly the water tension of the porous body, which can be correlated with the moisture content. The conversion of water tension values to the moisture content requires an individual calibration for each type of porous material and is not a reliable procedure. Vapour pressure method requires very accurate measurements of the equilibrium relative humidity between the porous body and the surrounding air [9].

3. Methodology

The sensor has been designed to be able to carry out numerous measurements without being concerned with its displacement. Sensor parameters such as: Antenna angle, distance to object and temperature can be monitored in a real time. The sensor frame had two calibrated displacement transducers attached, which enable clarification of the distance between the antennas and the measured material. A third displacement transducer was located between the antennas to calculate their angle. In experiment microwave spectroscopy will be used to identify if microwave technology could be a potential method to

determine moisture levels in the building materials. The preliminary experiment for the measurement of moisture content was conducted with the use of Ceiling plasterboard. The material was soaked in water for two minutes and left to dry off. The sensor was placed 2cm in front of the material. The weight of the sample and microwave spectrum were recorded every 60 seconds by the program developed in LabVIEW for a period of 24 hours. Figure 1 shows the experimental setup used during this experiment.
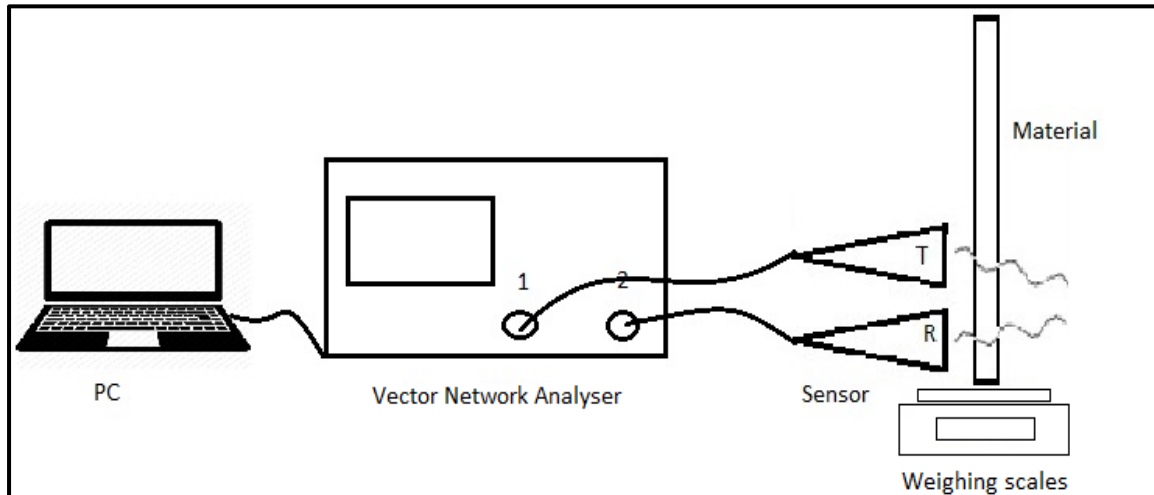
Figure 1. Experiment Setup to monitor ceiling plasterboard drying off process.

4. Results and Analysis

Experimental results for drying process of the ceiling plasterboard over 24 hours are presented in Figure 2. Microwave spectrum and material weight have been captured once an hour. There is significant amplitude shift across 6GHz to 12GHz frequency range owing to the evaporation of the water. Water has high dielectric properties which affects the microwave spectrum. A lower amount of water enables reduction of the power loss of the microwave energy and creates visible changes to the microwave spectroscopy.
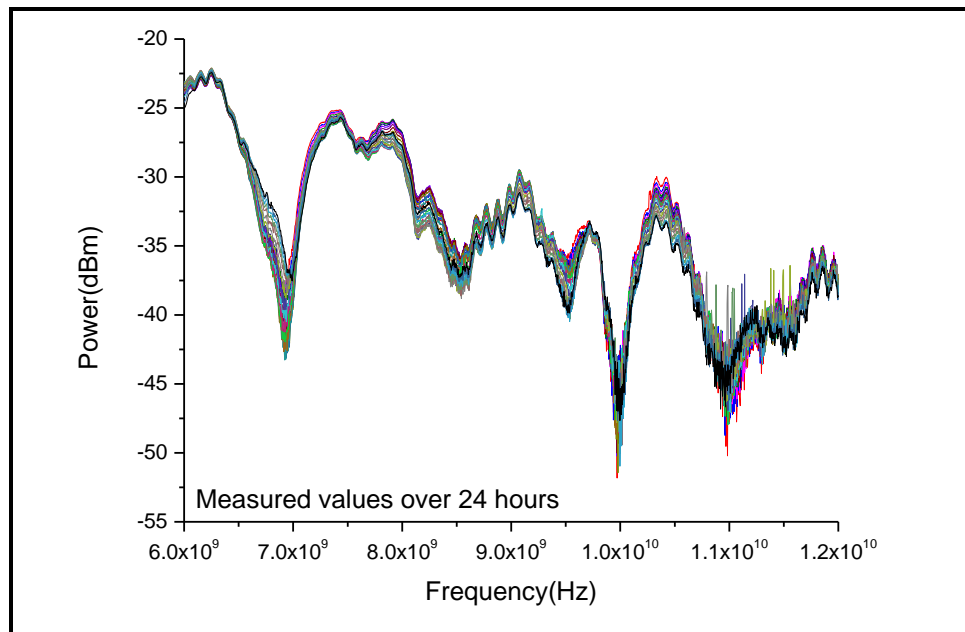
Figure 2. Ceiling plasterboard drying process over 24 hours.

Figure 3 shows the direct comparison between weight loss and attenuation at 10.3GHz. S21 change and material weight loss at the particular frequency decreases linearly (R2 = 0.918) over the period of 24 hours.
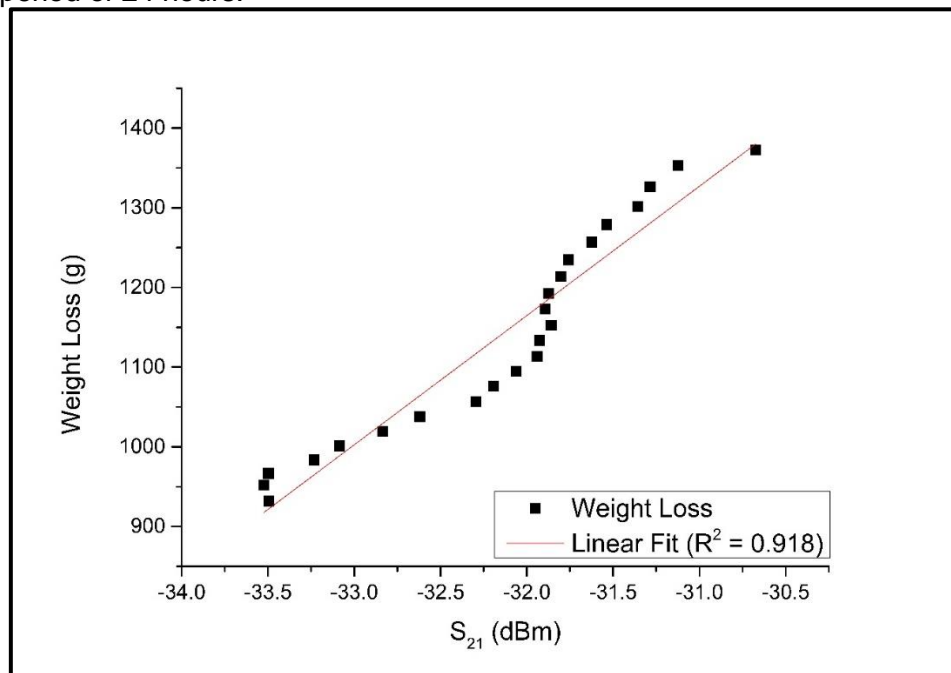
Figure 3. Ceiling plasterboard comparison between weight loss and attenuation at 10.3GHz.

During drying of porous material, a critical level is defined in such a way that for moisture contents above that level, liquid exists mainly as free water and as only bound water below that level. The energy absorption is low when the moisture content is below critical and the dielectric properties have values close to those of a skeleton porous material. Above the critical level the dielectric properties are much higher and have values that tend towards that of the free water. Hence, the dielectric loss will decrease as the material dries and consequently the effectiveness of dielectric heating or drying reduces as the product dries out. There is a high linear correlation between material weight loss and attenuation at 10.3GHz. The prediction model was made by using the Partial Least Squares Regression (PLS). Two models were calibrated for both EM attenuation and moisture content, The EM attenuation model has good accuracy, with R2 = 0.995 and a root mean square error of estimation (RMSEE) of 0.745g for the calibration set and R2 = 0.892, RMSEE = 3.97g for the validation set.

5. Conclusion
This experiment has been undertaken to confirm if microwave spectroscopy can be used to determine the moisture content of the ceiling plasterboard during the drying off process. Results show that there is a significant change in attenuation of EM signal across 6GHz to 12GHz frequency range while the water evaporates from measured material. There is a high linear correlation between material weight loss and attenuation at 10.3GHz. The prediction model using PLS method has good accuracy, with R2 = 0.995 and a root mean square error for estimation (RMSEE) of 0.745g for the calibration set and R2 = 0.892, RMSEE = 3.97g for the validation set.

6. References
[1]   J. F. A. Moore, *Monitoring Building Structures*. Springer Science, 1992.
[2]   H. Zhang, *Building Materials in Civil Engineering*. Woodhead Publishing, 2011.
[3]   C. Stirling, "Building pathology: moisture conditions within external masonry walls," 2011.

[4]   P. Brimblecombe, C. M. Grossi, and I. Harris, "The effect of long-term trends in dampness on historic buildings," *Weather*, vol. 61, no. 10, pp. 278–281, 2006.

[5]   R. McDonald, "Algo Centre Mall Collapse," 2013. [Online]. Available: http://failures.wikispaces.com/x-4.0 Collapse. [Accessed: 09-Mar-2016].

[6]   R. Burkinshaw and M. Parrett, *Diagnosing Damp*. Coventry: RICS Books, 2004.

[7]   C. H. Sanders and M. C. Phillipson, "UK adaptation strategy and technical measures: the impacts of climate change on buildings," *Build. Res. Inf.*, vol. 31, no. 783016864, pp. 210–221, 2003.

[8]   J. Imal, "Measurement of Moisture Fields in the Bridge Structure of Charles Bridge," vol. 42, no. 3, pp. 53–58, 2002.

[9]   R. Wormald and A. L. Britch, "Methods of Measuring Moisture Content Applicable to Building Materials," vol. 3, no. 1, pp. 135–145, 1969.

[10]   D. Bucurescu and I. Bucurescu, "Non-destructive measurement of moisture in building materials by compton scatering of gamma rays," vol. 63, no. 1, pp. 61–75, 2011.

[11]   R. Vijayakumar, L. Rajasekaran, and N. Ramamurhy, "Determing the Moisture content in Limestone Concrete by Gamma Scattering Method: A Feasibility Study," no. 1, pp. 1–10, 2002.

[12]   D. . McCann and M. . Forde, "Review of NDT methods in the assessment of concrete and masonry structures," *NDT E Int.*, vol. 34, no. 2, pp. 71–84, 2001.

[13]   R. Y. Vun, K. Hoover, J. Janowiak, and M. Bhardwaj, "Calibration of non-contact ultrasound as an online sensor for wood characterization: Effects of temperature, moisture, and scanning direction," *Appl. Phys. A*, vol. 90, no. 1, pp. 191–196, Sep. 2007.

[14]   M. Maksimović, G. M. Stojanović, M. Radovanović, M. Malešev, V. Radonjanin, G. Radosavljević, and W. Smetana, "Application of a LTCC sensor for measuring moisture content of building materials," *Constr. Build. Mater.*, vol. 26, no. 1, pp. 327–333, Jan. 2012.

[15]   G. Radosavljevic, "Wireless LTCC sensors for monitoring of pressure, temperature and moisture," *J. Microelectron.*, vol. 42, no. 4, pp. 272–281, 2012.

[16]   R. Černý, "Time-domain reflectometry method and its application for measuring moisture content in porous materials: A review," *Measurement*, vol. 42, no. 3, pp. 329–336, 2009.

[17]   M. A. M. Yunus and S. C. Mukhopadhyay, "Novel Planar Electromagnetic Sensors for Detection of Nitrates and Contamination in," vol. 11, no. 6, pp. 1440–1447, 2011.

[18]   M. Davies and Z. Ye, "A ' pad ' sensor for measuring the moisture content of building materials," vol. 3, no. 30, pp. 263–270, 2009.

Space design: How individual differences influence requirements for the physical environment

Hannah Wilson
School of the Built Environment, Henry Cotton Building
Liverpool John Moores University
H.K.Crawford@2010.ljmu.ac.uk

Abstract. There is a significant amount of individual differences within the students who attend University, from cultural to subject choice (Hassanain & Mudhei, 2006). With this understanding the physical learning environment must be suitable for many differing activities and people, maintaining and developing an appropriate physical space is essential. Ensuring learning is supported by a positive student experience is vital, as interactions with the physical learning environment have been found to have a relationship with satisfaction but more noteworthy with educational outcomes (Chan, 2011; Vidalakis et al., 2013). The aim of this research is to identify the relationship between personality and subject cohort to develop an innovative framework for the design of higher education spaces. Data was collected from students, by surveys and focus groups in Liverpool John Moores University, across four different departments. A survey was distributed to measure students' preferences and the effect individual differences have, in this research personality and subject cohort. Additionally, focus groups allowed further exploration of the students' experience regarding the design of learning environments; to identify factors most preferred by students and why they were considered important. Preliminary findings from the survey found a difference in personalities between subject cohorts. Furthermore, that students from different subjects and with differing personality traits, preferred different features in the physical learning environment. The findings from the focus group show a promising consensus with the survey findings and interesting insights into students' feelings towards design elements. This research has implications for built environment policy in the education sector.

Keywords. Higher education, Physical learning environments, Personality.

1. Introduction

We use buildings day in a day out, in fact, research finds that we spend upwards of 90% of our lives within buildings (Evans & McCoy, 1998). Consequently we take space for granted, and we can fail to notice how space plays a part in our behaviours and intentions. Furthermore user's perceptions of their physical environment significantly influence experiences (Kandiko & Mawer, 2013) and consequently satisfaction (Chan, 2011). Due to this considering how the physical environment needs to be designed to increase the buildings performance is important.

Although research has identified a number of features that affect students learning experiences within physical educational buildings most research still focuses on work places. Riley et al. (2010) review of literature for post occupancy evaluation found that performance

management of buildings attends to commercial and residential buildings. Whereas knowledge on the features that define well performing higher education institutes are lagging behind. To design higher education space it is important to understand what the users, the students, preferences are, (Kasalı & Doğan, 2010) concluded that students are good sources of information in the design and planning of the environments they occupy. Their insight allows academic developers to identify specific features of the space and how they will be used. However, previous research has failed to concentrate on the evidence based needs of the end user (CABE, 2005; Bickford & Wright, 2006). Therefore understanding what is it that students in higher education require out of the physical learning environments is an important area for investigation.

To design appropriate physical learning environments for students, it is important to understand the influence the environment has. Students have been found to make judgements of teacher's capabilities by the physical environment that they are taught in (Weinstein & Woolfolk, 1981). Specifically students made more positive assessments of the capabilities when the room was more orderly. Furthermore, the design of the physical environment has also been found to impact student's health and also the amount of stress they experience (Evans & McCoy, 1998). Therefore not only does the physical environment affect students perceptions but also their overall wellbeing. By improving physical environments maybe therefore consequent in increased student satisfaction and a positive impact on their learning experiences (Vidalakis et al., 2013). The environmental factors that have been found to be importance in perceptions and satisfaction are features such as lighting, ventilation (Winterbottom & Wilkins, 2009) temperature (Douglas & Gifford, 2001; Yang et al., 2013). The colour schemes, comfort of seating (Hawkins & Lilley, 1998) and a new wave of literature on flexibility of space is noted to influence satisfaction (Holm, 2011). However although this research explores individual factors of the environment that should be considered in the design recognising the importance of environmental as a whole is not understood (Rullman & Van den Kieboom, 2012). This redevelopment of the physical learning environment should be undertaken through a review of stakeholder evaluations of the performance of all of the features of the physical learning environment (Amaratunga, 2000). This research will therefore explore student's perceptions and evaluations of the physical learning environment to understand features that can be develop to enhance students learning experiences.

Although the design of the higher education learning environments has been broadly researched (Neary & Saunders, 2011; Yang et al., 2013; Kollar et al., 2014) the literature highlights large gaps in the understanding of student's requirements of the environment and the importance of developing this knowledge. Shouder et al. (2014) noted the importance of the physical environment providing students with the opportunity to be managers of their own space to increase work productivity. Additionally Guney and Al (2012) identified the importance of space design to support individual differences in learning styles. Therefore providing space that students can frequent to suit their own individual needs. Understanding the influence of individual differences in the requirements of the physical learning environment is an important aspect of research to consider, and requires further examination (Pawlowska et al., 2014).

One measure of individual differences, personality, has been found to be a strong predictor of perceptions of the physical learning environment (Ibrahim et al., 2002). Furthermore, Pawlowska et al. (2014) noted a relationship between personality traits and classroom environmental factors such as 'students in class get to know each other really well' and satisfaction. Therefore identifying student's individual requirements would develop understanding of how to design higher education physical learning environments. Allport (1966) ascertained that personality traits do not wait to be aroused by external stimuli, but that an individual actively seek stimulus situations that encourage their traits. Therefore physical environment that students work in needs to suit their personality traits or will not be utilized by them, consequently affecting their behaviour. Another difference in the university environment is the different subjects that on space has to account for. Ibrahim et al. (2002) found that between architectural students and non-architectural students there was a

difference in perceptions of the physical environment, they found this was due to an influence of personality traits. Therefore considering the effect that subject choice has on the requirement of the physical environment may be an important avenue to examine. Perhaps designing environments that suit the individual differences of the students using them will increase student's satisfaction but having a positive impact on learning experiences.

2. Conceptual framework

Considering features of the physical learning environment seems only sensible when understanding the influences of student's satisfaction. Additionally, understanding physical learning environments may need to suit many different requirements at once. A summit on the design of higher education space concluded that although literature does exist on designing physical learning environments, this rarely informs actual planning and design processes (Rullman & Van den Kieboom, 2012); current literature doesn't identify features in the environment as a whole and are therefore convoluted and difficult for academic developer to use. Therefore by identifying a simple model categorising specific factors for academic developer can utilise is absent important. Overall it is established a focus should be placed on forming a functioning model of the design space.

From the overview of the literature it can be seen that these two facets of research, the physical learning environment and personality interlink as they effect students satisfaction and perceptions of the environment. Exploring these factors could enhance the physical learning environment beneficially for students. There is currently a gap in the literature for these elements, although there is a large body of literature focusing on the design of Higher education institutes there is no student focussed preference model for the design of space. Therefore to begin to identify how the design of space could be enhanced for student satisfaction this research will explore the features of the design of physical learning environments. This research will firstly aim to identify if there are differences in personality traits between disciplines. This will then develop to identify if these individual differences has a relationship with preferred factors in the physical learning environment. The aims are outlined to help the research identify factors of the design that should be considered when developing new, and refurbishing old higher education physical learning environments.

3. Methodology

3.1. Design

The aim of this research is to identify features of the physical learning environment that students consider important; and to identify if these features have a relationship with personality traits. Therefore a mixed method approach, utilising self-report surveys and focus groups, was decided upon. This enables the research to gain a better and more comprehensive understanding of the complexity of interactions between personality traits, subject cohort and features of the physical learning environment. Four schools in Liverpool John Moores University were selected but opportunity sampling; School of Engineering, School of the Built Environment, School of Art and Design and the Business School.

3.2. Research tools

3.2.1. Surveys. Learning environment- A 57 item questionnaire was developed through an extensive literature review and from the focus groups findings. The survey was split into three elements building, Functionality and Environment adapted from previous literature on the design of buildings (Gann et al., 2003). Within the build element of the survey factors such as 'lecture halls' and 'specialist teaching rooms' were included. Within the functionality element of the survey factors such as, 'clear signs in building' and 'motivating environment' were identified. And finally for the Environment section of the survey factors such as 'natural lighting' and 'comfortable furniture' were included. The questionnaire was scored on a five point Likert scale scored from Unimportant to Very important.

Personality- To measure personality the Big five measure of personality was used, the Five Factor Model (FFM) (Goldberg et al., 2006). This survey is constructed of questions

measuring the five personality traits identified through empirical research, Openness, Conscientiousness, Agreeableness, Extraversion and Neuroticism (Emotional Stability).

3.2.2. Focus groups. 10 focus groups were conducted with 3-10 students participating in discussion. Students were asked questions such as 'How do you think the learning environments in the universities buildings could be improved? What features or spaces could be added or changed?' and 'What features of spaces do you like best about the learning spaces?'.

4. Findings

4.1. Qualitative findings

The findings from the focus group discussion identified 6 higher level themes; place, environment, design, facilities, cosmetics and operations. Figure 1 displays the high themes and the low level themes these were developed from. In the focus groups students suggested that "if the spaces were designed in a way that facilitated a better teaching method then that would be better". Therefore students consider the design of the space to be an important tool in better teaching. Furthermore students noted that enhancing the physical environment would be better for them as workspaces a student noted "You feel good working in it because it's nice". In identifying these high level themes students discussed features of their physical environments and noted "there are not the resources we need" and "access to technology… you need to research on the internet so that is important too". This discussion allowed for the exploration of students feelings towards features in their physical learning environment to identify specific features that they note as being important. Another interesting factor of the focus group analysis was the discussion of different features by the four different schools, displayed in Figure 1. For example all school discussed easily 'getting around' as an important feature, however how up to date the building was only discussed by two of the schools.
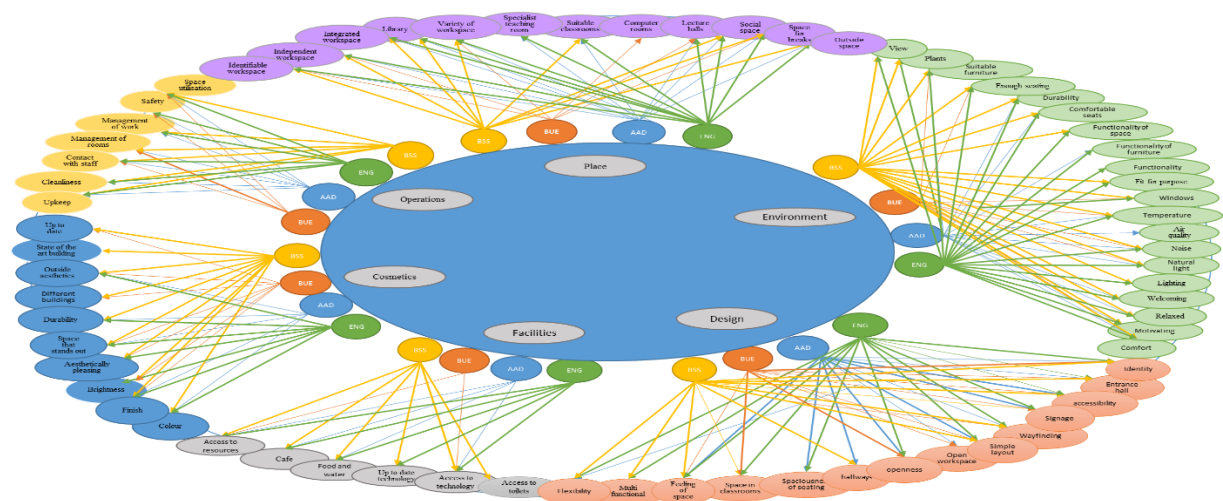
Figure 1. High and low level themes identified from focus group discussion and the discussion by schools.

4.2. Quantitative findings

The results from the survey analysis found through a factor analysis 10 significantly correlated components of the physical learning environment that students consider as important. These are show in Figure 2, wayfinding, aesthetics, rooms, social spaces, environmental (feeling), environmental (traditional), facilities and resources, identity, technology and Decor. As can be seen from Figure 1 these factors are very similar to the themes discussed by the students in the focus group discussion.
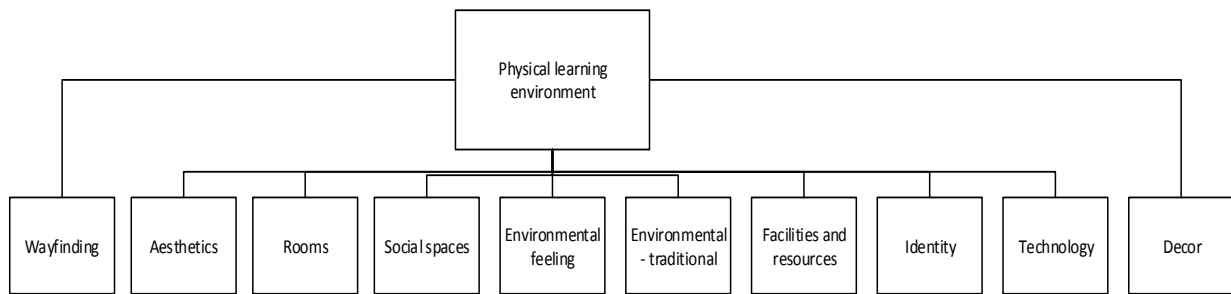
Figure 2. Components from factor analysis for design elements of the Physical learning environment.

To analyse the relationship between personality and features of the physical learning environment structural equation modelling was utilised. The findings identified a relationship between the personality traits and the components of the physical learning environment. The results also showed a difference between personality traits and individual variables. For example there was a relationship between those who scored low on neuroticism, high on emotional stability and a preference for windows. Therefore suggesting that personality traits influence the preferences for features in the physical learning environment. Consequently there is a requirement to design environments for the individual differences in personality traits. This is because as Allport (1966) noted, people seek out situations that encourage their traits, in this case environments that suit their personality.

The current research also found a difference in personality traits between schools, furthermore that there was a relationship between school and preferences for features within the learning environment. For example students from the school of Art and Design rated 'natural lighting' as statistically more important than the Business school. Therefore suggesting that the physical learning environments need to be designed suitably for the school or subject that is going to use the space. Utilising the specific aspects students highlighted in this research, as important for each subject cohort, could enable each building to increase the students learning experiences through the appropriate design of space.

5. Conclusion
This current research approached the design of higher education learning space from a new perspective. By focusing on students requirements in their physical learning environments it will offer the opportunity to use a simple model that practitioners can use to design space to suit the users' needs (Rullman & Van den Kieboom, 2012). This research notes that it is important to consider the users of the physical learning environments when designing the space and facilities. Personality and subject choice appears to influence student's preference for features of the design. When designing physical learning environments this should be considered to enable the space to suit the requirements of the end users. Therefore by considering this in the design process the suitability of the space will be inherent in the structure that is developed. This could consequently lead to a more satisfactory environment for the students but could also help to reduce post- completion adaptations. Ultimately this aim of this research is to develop a model of information that can be utilised to influence sectional policy. Universities across the country could adopt such policy to develop and refurbish their institutional buildings.

6. References
Allport, G. W. (1966). Traits revisited. *American psychologist, 21*(1), 1.
Amaratunga, R. (2000). Building performance evaluation in higher education properties: a facilities management approach. *RICS Research Paper Series, 3*(14).
Bickford, D. J., & Wright, D. J. (2006). Community: The hidden context of learning. In D. Oblinger (Ed.), *Learning Spaces*. Boulder: CO:EDUCAUSE.

CABE, J. (2005). *Design with distinction–the value of good building design in higher education*: CABE Publications London.

Chan, R. Y. (2011). *The effects of student involovement and college environment on students' learning and living experience at world class universities in Shina: A comparitive case study of the University of Hong Kong (HKU) and Shanghai Jiao Tong University (SJTU).* School of education, Shanghai Jiao Tong University.

Douglas, D., & Gifford, R. (2001). Evaluation of the physical classroom by students and professors: A lens model approach. *Educational Research, 43*(3), 295-309.

Evans, G. W., & McCoy, J. M. (1998). When buildings don't work: the role of architecture in human health. *Journal of Environmental psychology, 18*(1), 85-94.

Gann, D., Salter, A., & Whyte, J. (2003). Design quality indicator as a tool for thinking. *Building Research & Information, 31*(5), 318-33.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality, 40*(1), 84-96.

Guney, A., & Al, S. (2012). Effective learning environments in relation to different learning theories. *Procedia-Social and Behavioral Sciences, 46*, 2334-8.

Hassanain, M. A., & Mudhei, A. A. (2006). Post-occupancy evaluation of academic and research library facilities. *Structural Survey, 24*(3), 230-9.

Hawkins, H. L., & Lilley, H. E. (1998). Guide for school facility appraisal *Council of Educational Facility Planners.* Scottsdale, AZ.

Holm, J. (2011). Management: Pros and Cons of Open Plan and Office-based Workplaces.

Ibrahim, A. F., Abu-Obeid, N., & Al-Simadi, F. (2002). The Effect Of Personality Traits On Architectural Aesthetics' Evaluation: Familiar And Non-Familiar Environments As Evaluated By Architectural And Non-Architectural Students. *Architectural Science Review, 45*(3), 197-210.

Kandiko, C. B., & Mawer, M. (2013). Student expectations and perceptions of higher education. 1-82.

Kasalı, A., & Doğan, F. (2010). Fifth-, sixth-, and seventh-grade students' use of non-classroom spaces during recess: The case of three private schools in Izmir, Turkey. *Journal of Environmental Psychology, 30*(4), 518-32.

Kollar, I., Pilz, F., & Fischer, F. (2014). Why it is hard to make use of new learning spaces: a script perspective. *Technology, Pedagogy and Education, 23*(1), 7-18.

Neary, M., & Saunders, G. (2011). Leadership and learning landscapes: The struggle for the idea of the university. *Higher Education Quarterly, 65*(4), 333-52.

Pawlowska, D. K., Westerman, J. W., Bergman, S. M., & Huelsman, T. J. (2014). Student personality, classroom environment, and student outcomes: A person–environment fit analysis. *Learning and Individual Differences, 36*, 180-93.

Riley, M., Kokkarinen, N., & Pitt, M. (2010). Assessing post occupancy evaluation in higher education facilities. *Journal of Facilities Management, 8*(3), 202-13.

Rullman, L. J., & Van den Kieboom, J. (2012). Physical Place on Campus: A Summit on Building Community. *Association of College Unions International: Bloomington.*

Shouder, T., Inglis, G., & Rossini, A. (2014). Listening to Students: Make Learning Spaces Your Own. *Change: The Magazine of Higher Learning, 46*(1), 26-7.

Vidalakis, C., Sun, M., & Papa, A. (2013). The quality and value of higher education facilities: a comparative study. *Facilities, 31*(11/12), 489-504.

Weinstein, C. S., & Woolfolk, A. E. (1981). The classroom setting as a source of expectations about teachers and pupils. *Journal of Environmental Psychology, 1*(2), 117-29.

Winterbottom, M., & Wilkins, A. (2009). Lighting and discomfort in the classroom. *Journal of Environmental Psychology, 29*(1), 63-75.

Yang, Z., Becerik-Gerber, B., & Mino, L. (2013). A study on student perceptions of higher education classrooms: Impact of classroom attributes on student satisfaction and performance. *Building and Environment, 70*, 171-88.

Antoni Dmochowski

Promoting the vitality and viability of town and city centres: Incorporating new regional inner-city retail developments through retail-red regeneration

Antoni Dmochowski
School of the Built Environment, Henry Cotton Building, Liverpool, L3 2ET
E-mail address: A.Dmochowski@2010.ljmu.ac.uk
Director of Study: Dr. Raymond Abdulai
Faculty of Technology and Environment, 15-21 Webster Street, Liverpool, L3 2ET.
E-mail address: R.Abdulai@ljmu.ac.uk

Abstract. The publication of the *National Planning Policy Framework* (NPPF) in 2012 set out the government's objectives for town centres. The key strategies from the paper were to promote their vitality and viability, by planning for their growth and development, as well as promoting and enhancing existing centres. The main focus of the policy however was to promote a 'town centre first' approach in regards to new retail development. Liverpool is one of many UK cities who have recently experienced a new inner city retail development built alongside existing inner city retail areas, in particular, through the building of Liverpool One. Opened in 2008, Liverpool One was built to not only offer a new retail experience in the city, alongside the established inner city retail areas, but to also enhance the existing city centre retail areas through retail-led regeneration. Research in examining the impacts on existing retailers following the opening of Liverpool One was conducted through a survey assessing the level of sales experienced by retailers in Liverpool's existing inner city retail areas as well as interviews with Liverpool's existing inner city shopping centre managers and existing individual retailers. Following the results of the survey and interviews, which suggested that the opening of Liverpool One had impacted on existing inner-city retailers, this paper presents a set of research-based best practice guidelines which could be used as an industry tool in the proposal, development and post development phases of new regional inner-city shopping centres. It is anticipated that these guidelines could aid in promoting the vitality and viability of town and city centres and therefore go some way in minimising the impacts on already established inner-city retailers.

Keywords: (1), Regeneration (2), Town Centres (3), City Centres (4),Retail Developments

1. Background

The late 1990's and early 2000's saw the completion of the first six enclosed regional inner-city shopping centres. These in-town schemes were each over 400,000 sq ft. with a

combination of both retail and leisure space and paved the way for a further twenty new inner-city retail schemes to be built between 2001 and the present day. A significant number of these twenty-five new inner-city retail schemes have been built alongside the concept of retail-led regeneration, which can be argued as being one of the most significant philosophies in shaping the current retail landscape in British town and city centres. As a report by the Retail Strategy Group (2004, p.13) of the Department of Trade and Industry noted "retail invests in people and places, it creates new markets, provides a focus for the implementation of social policies and plays an important role in the regeneration and well-being of towns, cities and urban areas". Retail led urban regeneration also has the potential to act as a catalyst for the wider regeneration of deprived areas and cannot simply be ignored, especially when retailing plays such a vital role in the UK economy (BITC, 2007). Not only is it the third largest service based industry in the UK, the sector is also a major employer, creating new jobs and opportunities and giving communities access to local goods and services (BITC, 2007).

Although the proposed benefits of retail led regeneration are theorised to be profound, it can be argued that it is all too easy to simplify the arguments in favour of retailing as a catalyst for local economic growth. Dixon (2005, p.169) comments that "relying on 'headline figures' for job growth, for example, creates a danger that equally important issues are frequently overlooked". It is without a doubt vital, especially as the UK emerges from the depths of recession, that the Government is seen to be providing opportunities and jobs as well as pumping investment into local economies; however, it is all too easy to say that this does not impact on already existing inner-city retail shops and developments. Dixon (2005, p.169) poses some interesting questions in regard to the possible consequences of retail led regeneration such as "how does the local regeneration 'balance sheet' stack up in terms of jobs, spending, loss of local shops, the property impact in terms of shift of prime rental pitch and overall social inclusion in the local community?" It is therefore naïve to consider that retail led regeneration is without its flaws and the possible negatives from the proposed regeneration schemes need to be highlighted, something that is arguably, following a difficult economic period, being conveniently overlooked.

Inner-city retail development as a tool for regeneration looks set to continue and although this may be the case, only three studies have been conducted into the impacts of new inner-city retail developments on existing centres, of which one was conducted over thirty-five years ago on Eldon Square in Newcastle and the other two (Oracle Centre in Reading and West Quay in Southampton), almost ten and fifteen years ago respectively. Given the fact that over twenty-five inner-city centres have opened their doors since 2000, there have been no recent studies to understand the nature of their impacts in UK town and city centres. There also seems to be a general acceptance and approval of the merits that these developments bring without any concrete evidence.

This paper forms part of a larger study in which the impacts of new retail developments on existing inner-city shopping centres and other retail areas was examined using Liverpool One as a case study. A mixed methods approach was adopted for the study with a questionnaire being developed to investigate existing retailers' sales levels whilst vacancy rates and changes of occupation were measured using Goad maps. Semi-structured interviews were also carried out with existing inner-city shopping centre managers. The findings of the research suggested that the opening of Liverpool One impacted on existing retailers and shopping centres through a loss in annual sales, alterations in footfall patterns, fluctuations in vacancy rates, and changes in occupation of retail space. However, the scale of impact varied between retail areas within the city centre and the economic recession at the time of the study may have also been a contributing factor to these findings. Nevertheless, this research informed the development of a set of best practice guidelines to be used as a possible industry tool when incorporating a new regional inner-city shopping centre into an existing town or city centre.

2. Best Practice Guidelines

Given the lack of detailed research into the impacts of new inner-city regional shopping centres on existing retailers and following analysis of the questionnaire survey results, Goad maps and interviews with the centre managers, the findings suggested that it would be beneficial to propose a set of best practice guidelines that could be used as an industry tool in the proposal, development and post development phases of a new regional inner-city shopping centre. Government guidelines over the past fifteen years have continually promoted the 'town-centre first' approach in regard to new retail development and this looks set to continue with the 'sequential approach' strategy reinforcing an even stronger presumption against out-of-centre developments. Furthermore, given the government's requisite of encouraging sustainable retail development in town centres and that planning authorities should recognise town centres as the heart of their communities, the development of a set of best practice guidelines will go some way in enabling developers and town planners to achieve these goals.

Additionally, at the heart of government policy is to promote city centres' vitality and viability by planning for their growth and development as well as promoting and enhancing existing centres (Office of the Deputy Prime Minister, 2005). Objectives stemming from these key strategies were to enhance consumer choice through shopping, leisure and local services, with the aim of providing a genuine opportunity to meet the needs of local communities (Office of the Deputy Prime Minister, 2005). Further objectives to support these strategies are through making city centres efficient, competitive and innovative, whilst also improving city centre productivity (Office of the Deputy Prime Minister, 2005). The following best practice guidelines therefore aim to promote the vitality and viability of town and city centres when incorporating a new regional inner-city shopping centre and in doing so go some way to minimising the effects of a new shopping centre on the existing retailers. The following will outline each of the best practice guidelines, explain how and why each guideline is appropriate and discuss its relevance in line with government policy, recent literature and the questionnaire survey results, goad map findings and interviews with the existing shopping centre managers which form part of the larger study.

In order to promote and enhance existing centres, retail-led regeneration strategies should be integrated and placed within whole town centre retail strategies for the benefit of all retailers and not just in and around the intended development site.

A key strategy outlined in government policy was to promote city centres' vitality and viability through promoting and enhancing existing centres. A key finding from the Eldon Square study in Newcastle suggested that following its development, those retailers who were based in close proximity to the scheme showed signs of growth whilst a steady decline was observed by retailers who were located further away from the new centre. In regard to this study, results from the retailer survey and comments from participants also showed similar trends. A number of retailers who were located within close proximity to Liverpool One suggested that they were benefiting from the spill over of 'high end' shoppers attracted to the new centre. For example, retailers located in the Cavern Walks shopping centre who are located within close proximity of Liverpool One noted a year on year increase in sales and retailers in Church Street and Lord Street, who are again located close to Liverpool One, predominantly noted increases to their sales. On the other hand, retailers in Clayton Square and Bold Street had felt the largest decreases to their sales of over 10% and these locations, apart from St. Johns, are the furthest distance away from Liverpool One.

Further comments from the retailer survey, in particular from both small and independent retailers, suggested that Liverpool One had moved the primary retail location and they felt that the retail strategy in the town centre was now centred in and around the proximity of the new centre. Additionally, comments from existing retailers in regard to public transport, parking, events and occasions and relocation of retailers were all predominantly of the opinion that the town centre strategy was now focused in and around the proximity of Liverpool One and that the only beneficiaries were either the new centre or surrounding retailers. It is therefore imperative that planners and local councils place new retail

developments within whole town centre retail strategies for the benefit of all retailers, that these strategies are clearly defined for all areas and therefore not allow the new development to dictate the focus of the town centre retail strategy.

Town centres are at the heart of their communities. The individuality and uniqueness of a town centre should be recognised and taken into consideration and retail-led regeneration initiatives should reflect this.

In ensuring the vitality and viability of town centres, the NPPF outlines that town centres should be recognised as the heart of their communities. This is a particularly interesting point, not only because recent Government policy has been structured with an emphasis on community and sustainability but also because comments made, particularly by small and independent existing retailers both in the retailer survey and through brief interviews with participants, in regard to the impacts of a new retail development, have outlined their significance in terms of the effects on their sales, livelihood, well-being, quality of their environment both structurally and socially, ability to trade and to compete and most importantly sustain themselves within the retail hierarchy, particularly when a new retail development has been built alongside them. Each town and city centre has a rich heritage, one that has been built and evolved over time, is individual and unique and retail-led regeneration initiatives should reflect this to both conserve and work alongside existing retailers.

The impact of new retail developments on small, independent and secondary area retailers within city centres should not be overlooked. Small, independent and secondary area retailers should be integrated within the planning process so as to minimise the effects of a new shopping centre being built alongside them and it should be recognised that smaller retailers can significantly enhance and maintain the character and vibrancy of a centre.

Previous studies in both Newcastle and Reading have shown that smaller and secondary area retailers have felt the largest impacts from a new regional inner-city shopping centre being built alongside them through loss of trade, shift in trading patterns and an overall decline in their areas. Results from the retailer survey in this study have also noted similar results with small and independent retailers predominantly reporting decreases in their sales following the opening of Liverpool One. Integrating smaller retailers into the planning process for a new retail development through linking areas to the new centre could significantly reduce the impacts felt through changes in footfall patterns for example. Smaller and independent retailers are usually individual and unique and can add to the character and vibrancy of a town centre. The Goad Map survey in regard to 'types of retailer' showed that Bold Street for example (a secondary retail area) experienced a continual annual growth in occupation by independent retailers. Comments from the retailer survey and brief interviews with participants also suggested that the increase in the high proportion of independents located in Bold Street was down to it offering a different shopping experience to that of Liverpool One.

The above suggests that by possibly integrating smaller and independent retailers through their different retail offering within the planning process for a new retail development, could maintain and even enhance the vitality and viability of the town centre therefore reducing the impacts on smaller retailers. Government policy also highlights the need to support a diverse range of retail, with a strong retail mix of both comparison and convenience retailers and integrating smaller retailers into the planning process could go some way to achieving that.

Studies have shown that the impacts of a new retail development on existing retailers are at their most severe within the first three years of a new development opening. Local councils should recognise this and have policies in place so as to support existing retailers, for example business rate concessions and free parking initiatives.

The study on Eldon Square in Newcastle by Bennison and Davies (1980) was concluded by the authors noting that the effects on established retailers' sales levels can be distinguished between a series of short term effects in the immediate years following the centres opening which then declined over the long term, in this instance three years. In the case of this study, a similar trend was observed with the survey participants' sales decreasing in the first three

years following the opening of Liverpool One in 2008 and then the sales levels beginning to recover. Comments from the retailer survey and brief interviews with participants also suggested that existing retailers felt that they were being neglected and given very little support by the local council with some retailers stating that business rates were too high and increasing annually, or the lack of free or affordable parking initiatives close to the existing retailers. If local councils recognised this trend then business rate cuts and free parking initiatives could go some way to alleviating the impacts felt by the existing retailers in the short term. Local Councils could inform existing retailers that previous studies have shown that the severity of the effects in the immediate years should decline and that local councils, through these initiatives, would be showing support for the existing retailers and it would go some way to maintaining the vitality and viability of the town centre.

Whilst the physical development of a new shopping centre is an important part of the regeneration process, existing retail areas are also important and should not be ignored; preserving, enhancing and investing in established retail areas so as to maintain their vitality and viability is essential so that they can compete alongside new retailers.

Retail-led regeneration focuses on physical developments and is often justified in terms of acting as a stimulus for further investment in the retail sector and other areas such as commercial or residential investment. The physical development will then act as a catalyst for wider economic benefits such as retail spend, attracting new visitors, improving perceptions of place and investment in other areas. Comments from the retailer survey, interviews with centre managers and brief interviews with participants have suggested that this isn't always the case. In some of the secondary retail areas, such as Ranelagh Street and Renshaw Street, the relocation of retailers to Liverpool One has left the street with vacant property, boarded up shop fronts and an area which is in need of investment, investment which as of now has not been stimulated by the development of Liverpool One. Existing retailers also spoke about either the lack of or sub-standard maintenance of their areas, particularly by the local Council, with examples of block paved walkways being refurbished or replaced with concrete, road-works not being completed to schedule as well as litter and pest-control issues. Preserving, enhancing and investing in these areas, particularly when the stimulus from retail-led regeneration theory is not being felt is therefore imperative to both existing retailers and the vitality and viability of the town-centre.

Every effort should be made to link new retail developments with existing retail areas so as to minimise changes in footfall patterns or transport routes.

Previous studies have shown that when a new retail development is built, its impact in regard to pedestrian flow change, footfall, changes to public transport routes and walkways have all had a detrimental effect on existing retailers. Improving transport links so that both the new and existing areas can be genuinely accessed is also seen as an important goal in Government policy for town centre development. Comments from the manager of Clayton Square highlighted that prior to the opening of Liverpool One, the new centre was perceived to be a good thing, as it would bring more people to the town centre, however, what it actually did, was change footfall patterns. Changes in pedestrian flow change and footfall were key issues expressed by participants in the retailer survey and after the opening of Liverpool One and the economy, footfall was selected as one of the main reasons for decreases in annual sales. Existing retailers also mentioned that certain public transport routes and bus stops at which people alight for the town centre were once key to their footfall and these had now also changed. All of the above highlight the need for developers and planners to make every effort to link new retail developments with existing retail areas so as to minimise changes in footfall patters and transport routes.

When planning retail regeneration developments, city councils should incorporate the thoughts, ideas, views and opinions of the existing retailers via consultation exercises to enable a clear vision and strategy for the whole of the town centre. These consultation exercises should continue following the development of a new shopping centre enabling existing retailers to have a line of communication between themselves and the local council.

A recurring theme which became evident after speaking to the retailers who participated in the retailer survey was a lack of communication, particularly between small and independent retailers and the local council. Although this related predominantly to communication following the development of Liverpool One, existing retailers felt frustrated that there were no clear lines of communication between themselves and the local council. Reviewing both the planning and development literature in regards to Liverpool One, there was very little documented evidence to suggest that local retailers had participated in any consultation exercises prior to the centre's development. Goddard's (2013) suggestion that the rise of localism may enable councils to adopt widely different interpretations of policy based on their local circumstances means that consultation exercises would enable existing retailers to have their voices and opinions heard in regard to plans for any new retail developments. Consultations would enable effective and coordinated delivery of plans relating to new retail developments, create partnerships between the council and existing retailers, allow both the public and private sector to share expertise and deliver retail-led regeneration benefits for the whole of the town centre.

Events, celebrations and annual festive occasions supported by the local council should be spread across the whole of the town centre so as to include all retailers and not just the latest retail development.

Following on from existing retailers feeling frustrated that there were no clear lines of communication between themselves and the local council, as well as the opinions expressed in regards to the town centre retail strategy being focused on Liverpool One, many retailers expressed both disappointment and frustration that many of the existing retail areas were neglected by the local council in terms of events, celebrations and annual festive occasions. Town Centre events can attract a wide range of people from local residents to tourists and festive occasions such as Christmas markets are particularly well attended. The revenue that is generated from these events is for the most part extremely advantageous for the town centre economy. According to an article on the BBC website (2007), the Birmingham Christmas markets are estimated to attract well over £200 million in spending in the city over the duration of the event. Including existing retail areas within the boundaries of events or festive occasions can help to generate much needed income for these areas.

The health and vitality of all town centre areas should be monitored by the local council so as to strive for a strong, prosperous and sustainable retail offering and have strategies in place to make improvements where ever necessary. Shop rents, trade and turnover, proportion of vacant street level property, commercial yields, pedestrian flow and accessibility should all be routinely monitored.

A review of literature emphasised that not only is there a lack of detailed research into the impacts of regional inner-city shopping centres but there is also a lack of reliable and essential data. Most of the research has relied on primary data through the use of surveys and gathering historic data has also proved difficult with it being either not available or not detailed enough. To be able to measure and evaluate the health of town centre retail requires it to be statistically monitored with routinely collected data on all of the above performance indicators. Although this will inevitably rely on investment and support to be put in place, monitoring would enable Councils to distinguish which areas of town centres need support and it would be anticipated that this would lead to a sustainable and prosperous retail offering.

It is imperative that recent retail-led regeneration schemes in other UK towns and cities are evaluated with detailed analysis of the impacts on existing retailers over a 10 year period so policy can be directed appropriately to proposed retail-led regeneration projects. Proposals for new retail-led regeneration schemes should be considered alongside independent predictive impact assessments so as to gauge the possible effects on established retailers in existing town centres.

As mentioned previously, there is very little detailed research into the impacts of inner-city retail-led regeneration schemes and there seems to be a general acceptance and approval of the merits that these developments bring without any concrete evidence. Given the fact that over twenty five inner-city centres have opened their doors since 2000, there have been

no recent studies to understand the nature of their impacts in UK towns and city centres. The opportunities are still available to evaluate their impacts longitudinally (10 years or more) with the opening of a number of recent inner-city shopping centres including St. Stephens shopping centre in Hull (2011), the Westfield Centre in Stratford (2011) and most recently the Trinity Centre in Leeds (2013). Evaluating the impacts of these centres would allow for policy to be directed appropriately to proposed retail-led regeneration projects. Furthermore, independent predictive impact assessments, on established retailers in towns and cities designated for future inner-city retail developments, would allow for proposals and initiatives to be put in place both prior, during and after the new developments have opened, with the intention of limiting the impacts on established retailers, particularly in the immediate years of a centre opening.

4. Conclusions

Current Government policy suggests that there is no sign of the 'town-centre first' approach for new retail developments being reconsidered and the concept of promoting, enhancing and maintaining the 'vitality and viability' of town centres with an emphasis on sustainable development and communities will remain a key strategy. Current government policy however lacks guidance on how to adequately implement these policies in practice and the best practice guidelines put forward in this paper offer an opportunity to further develop and build upon the existing 'town-centre first' approach. The best practice guidelines were therefore developed with a focus on the town-centre first approach and their aim is to promote a town centre's vitality and viability by minimising the impacts on existing retailers.

5. References

BENNISON, D. and DAVIES, R. (1980) "The Impact of Town Centre Shopping Schemes in Britain: Their Impact on Traditional Retail Environments". *Progress in Planning*, 14, pp. 1-104.

BITC (2007) *Under-served Markets: Retail and Regeneration* [online]
Available at: www.bitc.org.uk/resources/publications/usm_guide.html
[Accessed: 13th January 2014]

DIXON, T. (2005) "The Role of Retailing in Urban Regeneration". *Local Economy*, 20 (2), pp.168-182.

GODDARD, C. (2013) "Planning, Development and Regeneration Briefing" [online] *The NPPF One Year On: GVA Grimley*, pp.1-10.
Available at: www.gva.co.uk/research/nppf-one-year-on/
[Accessed: 10th November 2014]

OFFICE OF THE DEPTY PRIMEMINISTER (2005) *Planning Policy Statement 6: Planning for Town Centres* [online]
Available at: www.communities.gov.uk/documents/planningandbuilding/pdf/147399.pdf
[Accessed: 7th November 2014]

RETAIL STRATEGY GROUP (2004) *The Retail Strategy Group Report: Driving Change* [online] London: Department of Trade and Industry.
Available at: http://www.bis.gov.uk/files/file10993.pdf
[Accessed: 3rd December 2013]

Basma Abdulaimma

Data Quality Control to WTCCC GWAS Data for Type 2 Diabetes

B. Abdulaimma[1], A. Hussain[1], P. Fergus[1], D. Al-Jumeily[1], C. Aday Curbelo Montañez[1]

[1]Liverpool John Moores University, Applied Computing Research Group, Faculty of Engineering and Technology, Byrom Street, Liverpool, L3 3AF, UK.
B.T.Abdulaimma@2015.ljmu.ac.uk
{A.Hussain, P.Fergus, D.Aljumeily }@ljmu.ac.uk
C.A.Curbelomontanez@2015.ljmu.ac.uk

Abstract. The global growth in incidence of type 2 diabetes (T2D) has been reached a pandemic, and has recently considered as a major international health concern. As such, understanding the etiology of type 2 diabetes is vital. Researchers pointed out that T2D is resulting from the convergence of genetics, environment, diet and lifestyle risk factors, however, genetic susceptibility has been established as a key component of risk. This has led researchers to investigate into genetic variants (single nucleotide polymorphism (SNPs)) associated with increase susceptibility to Type 2 diabetes and related traits using a number of data science approaches. However, in order for the study and the results to be reliable, quality control/filtering steps need to be conducted prior to further association analysis to eliminate bias into the study. Several quality control steps are adopted in this paper includes identification of samples with discordant sex, low genotyping call rate, high/low heterozygosity rate, and markers (SNPs) with low minor allele frequency and also testing for Hardy-Weinberg equilibrium. The results show that a number of sample and Markers have failed the quality control test and thus they had to be removed from the study.
Keywords: Type 2 Diabetes, Quality Control, Data Science, GWAS, Plink

1. Introduction

Genome-wide association studies (GWAS) are usually used for investigating the genetic architecture of human disease. More specifically, to identify common single nucleotide polymorphisms (SNPs), which is a single base-pair change in the genetic code and it is the main cause of human genetic variability [1], that influence human traits [2].

With the increase of genotypic technologies, information related to human genome is growing very fast. Although, studies with very large sample sizes have more power, they are also more likely to be subjected to experiment errors. Such errors include low quality DNA samples, differences in DNA quality that cause differences in the frequency of missing genotype call rate, errors in sample identification (sex identification problem) and poorly performing SNP assays [3]. Consequently, these errors can generate systematic bias into the study, leading to increase the number of false-positive and false-negative associations [4]. Researchers in [5] have demonstrated that loci with low minor allele frequency are more likely to result in false findings. Thus many study related to GWAS have exclude SNPs with MAF<10% [6]. Therefore, to decrease the effect of systematic bias, it is particularly important to perform quality control measures /filters to detect and remove markers and individuals for whom the genotypic quality is problematic [7]. These critical steps are significantly important and necessary before conducting any statistical analysis.

Here in this paper, we use and explore the dataset from Wellcome Trust Case Control Consortium (WTCCC) Type 2 diabetes. The prevalence and the incidence of Type 2 diabetes throughout the world is reaching the peak levels. In 2015, Diabetes UK[1] announced that there are 3.9 million people in UK living with diabetes. Type2 diabetes remains the leading cause of a serious long term health complications. It is responsible for most cases of blindness (Diabetic retinopathy), kidney failure, lower limb amputation and cardiovascular disease [8].

Beyond the human suffering, In the UK, the annual cost of type 2 diabetes to the NHS is approximately 8.8 billion for direct cost and £13 billion for indirect cost. Which [9] [10] . Researchers believed that Type 2 diabetes is resulting from the convergence of genetics, environment, diet and lifestyle risk factors [11]. Although environmental factors play a substantial role in the etiology of T2D, however, genetic susceptibility has been established as a key component of risk [12].

Building on existing works and tools for quality control assessment, this paper performs several steps into genetic case-control type 2 diabetes dataset to identify substandard markers and samples which should be removed prior to further association analysis. This has been achieved using PLINK toolset, a tool for handling SNP data with very large sample size, and R packages.

2. **Background**

Researchers have defined quality control (QC) as steps taken to monitor and control the quality of a GWAS data as it is being produced [13]. The need for careful QC of genotypic data is paramount to produce the subset of reliable markers and sample to serve as a rigorous ground for the subsequent association analysis. Several publications perform and address various aspects and steps into type 2 diabetes case-control studies.

In [14] a systematic meta-analysis was performed on a case-control study to investigate the role of potassium inwardly-rectifying-channel, subfamily-J, member 11 (KCNJ11) variation particularly E23K polymorphism (rs5219) in susceptibility to type 2 diabetes (T2D). In this meta-analysis, 56,349 T2D cases, 81,800 controls, and 483 family trios were collected from 48 published studies. The statistical methods used within the approach included The Standard Q-statistic test, subgroup analysis (ethnicity, sample size, BMI, age and sex) were performed to explore whether the variation in these studies was due to heterogeneity. Furthermore, for family based association studies, the transmission disequilibrium test (TDT) was employed to analyse effect size of the polymorphism. Moreover, the Z-test was used to determine the significance of overall odds ratio (OR). The study also conducted risk allele frequency (RAF) and population attributable risk (PAR) for a comprehensive observation of the effect of the E23K variant on T2D at population level. This study suggested that a modest but statistically affect of the 23K allele of rs5219 polymorphism in susceptibility to T2D, particularly in East Asians and Caucasians. However, the contribution of these genetic variations to T2D in other ethnic populations (e.g. Indian, African, American, Jews, and Arabian) appears to be relatively low.

In [15] researchers performed a case- control study of 400 type 2 diabetes cases and controls of South Indian population to analyse and outline the association of Potassium inwardly rectifying channel, subfamily J, member 11 (KCNJ11) gene on risk of T2D. The study also conducted a systematic review and meta-analysis for KCNJ11 (rs5219) polymorphism in 3,831 cases and 3,543 controls aggregated from 5 published reports from South Asian and East Asian population. In this study, odds ratio (OR) was employed as the measure of association of KCNJ11 polymorphisms (rs5219, rs5215, rs41282930, rs1800467) and T2D with its corresponding 95% confidence interval (CI). Moreover, Cochran's Q, I2 statistics were utilized to assess for heterogeneity within and between the eligible studies. The resulting evidence therefore showed that KCNJ11 rs5215, C-G-C-C haplotype and two loci analysis (rs5219 vs rs1800467) have a significant association with T2D however copy number variations (CNV) analysis did not show significant variation between T2D cases and

---

[1] https://www.diabetes.org.uk/About_us/News/39-million-people-now-living-with- diabetes/

control subjects. Furthermore, the meta-analysis of the study suggested that KCNJ11 (rs5219) polymorphism is associated with risk of T2D in East Asian and Global population however this outcome inapplicable to South Asian population.

3. Methodology

The Wellcome Trust Case Control Consortium (WTCCC) Type 2 diabetes datasets was used in this study. The dataset includes approximately 2,000 individuals for cases and 1,500 individuals for controls. All individuals were genotyping using Affymetrix GeneChip 500 K arrays.

Type 2 diabetes cases were chosen from UK Caucasian subjects who form part of the Diabetes UK Warren 2 repository. The dataset contains individuals aged between 25 and 75 years of age. While, the control individuals consist of 1,500 individuals from the 1958 British Birth Cohort controls (58BC)[2]. Individuals were followed periodically from birth (at ages 7, 11, 16, 23, 33) to age 44-45 years and they recruited based on sex and geographical region. A comprehensive description of the WTCCC dataset can be found in the original paper [3].

The datasets that is used in this paper is in Plink format. Plink is a whole genome data analysis toolset (for handling SNP data). The dataset contains 24 transposed file set (24 files for case and 24 files for control) that include two types of text file (TPED, TFAM).

The TPED file contains SNP and genotype information where one row represents a SNP (single marker). The other file which is TFAM includes individual and family information where each row records details related to one individual. Family ID, Individual ID, Paternal ID, Maternal ID, Sex, and Phenotype are recorded for each individual. For further information about Plink File set the reader can be referred to original website[3]

3.1 Data Pre-Processing

As it has been mentioned above, the dataset that is used in this study is in plink format (transposed file sets) which are tped and tfam files. Generally, as a first step we recoded these two files and convert them to .ped and .map files. While .ped and .map files are very large file (100K +SNP), it is recommended to convert these two files to a binary format files which are .bed ,.fam, and .bim files and this can be performed using PLINK toolset. Transferring to a binary formatted file, results in a considerable reduction in file size and significantly enhancing computational efficiency. Figure1 and Figure 2 is a visual representation corresponding to the standard files and binary files format respectively.

Fam ID	Ind ID	Pat ID	Mat ID	Sex	Pheno	rs1	rs2	rs3	…
FAM_T2D	WTCCC60444	0	0	1	-9	CC	GG	TT	…
FAM_T2D	WTCCC166692	0	0	2	-9	GG	GG	TT	…
FAM_T2D	WTCCC167773	0	0	1	-9	GC	GG	TT	…
FAM_T2D	WTCCC167362	0	0	2	-9	CC	GG	TT	…
FAM_T2D	WTCCC166960	0	0	2	-9	GC	GG	TT	…

Ped File

Chr	SNP	Genetic Distance	Base-Pair
2	SNP_A-1820282	0	24049
2	SNP_A-2056638	0	43652
2	SNP_A-1792446	0	49698
2	SNP_A-2063286	0	64387
2	SNP_A-2260913	0	76644

Map File

Figure 1: Represents the standard format of the files in the Dataset

---

[2] http://www.cls.ioe.ac.uk

[3] http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml

Family ID	Individual ID	Paternal ID	Maternal ID	Sex	Phenotype
FAM_T2D	WTCCC60444	0	0	1	-9
FAM_T2D	WTCCC166692	0	0	2	-9
FAM_T2D	WTCCC167773	0	0	1	-9
FAM_T2D	WTCCC167362	0	0	2	-9
FAM_T2D	WTCCC166960	0	0	2	-9

Fam File

Chr	SNP	Genetic Distance	Base-Pair	Allele1	Allele2
2	SNP_A-1820282	0	24049	G	C
2	SNP_A-2056638	0	43652	G	A
2	SNP_A-1792446	0	49698	T	C
2	SNP_A-2063286	0	64387	G	G
2	SNP_A-2260913	0	76644	C	A

Bim File

01101100	00011011	00000001
00011101	00011100	10010001
11111111	00111110	00011100
11100001	00011000	00101100
11001100	00000001	00110000

Bed File

Figure 2: Represents the binary format of the files in the Dataset

The notable issue with this dataset is that the disease phenotype of the whole study is assigned to -9 which represents missing phenotype. Consequently, it is impossible to distinguish between case and control. Therefore we replaced the disease phenotype value of the sixth column in .fam file to 1 and 2 to point to unaffected (control) and affected (case) phenotype respectively.

The last step is conducted by merging the binary files of both case and control to obtain one binary format dataset. This process accomplished by merging 45 samples from case and control and the final sample includes 3503 individuals in which 1999 cases and 1504 controls, also it includes 500568 markers (SNPs)

3.2 **Data Quality Control and Results**

   In this paper, we conducted quality control processes by using PLINK toolset and standard statistical software R. As a start point we identified individuals with discordant sex information in our dataset. Usually this begins by using genotype information from X-chromosome and check heterozygosity rates. In our dataset 32 samples were reported to have a PROBLEM flag (discordant sex data) and these are removed from the dataset.

Furthermore, we conducted missing genotypes rate for per individual in the dataset. Generally low genotyping call rate refers to low DNA quality or low sample concentration. In addition to that, we performed heterozygosity rate calculation for per individual by using the formula (N(NM)-O(HOM))/N(NM) where N(NM) points to the number of non-missing genotypes per individual, while O(HOM) refers to the observed number of homozygous genotypes. We generated a plot as shown in figure 3, using the proportion of missing SNPs

per individuals that is plotted on the x-axis and the observed heterozygosity rate per individual which is plotted on y-axis.
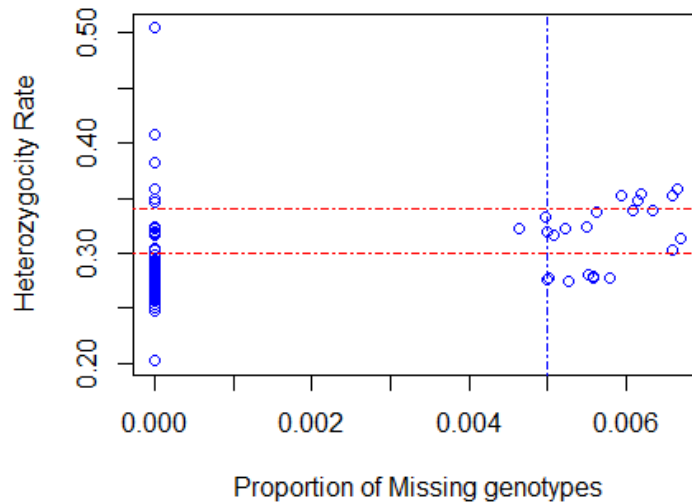
Figure 3: Genotype failure rate vs. Heterozygosity rate. Dashed lines indicates quality control thresholds and the dot represents the observed samples

Based on the plot, we observed that most samples without missing genotype and a few with missing genotype rate, therefore we decided to set thresholds at which to exclude samples. 37 samples with a genotype failure rate >0.005 and heterozygosity rate were omitted from our dataset.

Moreover, we tested the dataset against the identification of high missing data rate for each marker. Based on the results from F_MISS column which represents the proposition of individual missing for this marker, we found that a few SNPs need to be removed while the rest remained for further analysis.We also performed a statistical summary to check SNPs with low minor allele frequency (maf). According to the results from frequency test we chose to remove maf threshold of <0.01. We further conducted the Hardy-Weinberg Equilibrium (HWE) test for all SNPs in the study for the controls. Based on the results we removed those markers with P-value <0.001.

Following those statistical summary test steps, 3471 individuals and 430336 markers remained in the study from the original sample of 3503 and 500568 respectively.

4. Discussion

Using genotypic data for the identification of data quality is enormously fundamental process to the success of a case-control association study. It is very important to begin with quality assessment control before proceed with any statistically testing for association, hence this step will remove poor DNA sample and markers(SNPs) that produce bias to the study. Bias between cases and controls could lead to introduce more false-positive and false-negative associations.

We explore the case-control dataset to identify the main genotypic errors. This was commenced by the identification of samples with discordant sex data and it showed that there were a number of samples that meet this criterion. We additionally considered sample with low genotyping call rate and high/low heterozygosity rate. These findings were considered paramount; however for in-depth identification of DNA sample errors we must consider the identification of related and duplicate individuals. And we further must perform

the detection of samples of divergent ancestry and this step has a number of limitations as this test requires additional dataset with diverse ethnic population groups and further we need to make sure that the SNP genotyping for the two datasets are matched.

For extensive detecting for errors in this datasets we research for markers with the indication of low minor allele frequency and in this dataset we chose the MAF threshold of <0.01. We further perform Hardy-Weinberg equilibrium (HWE) test. And thus, this helped us to identify SNPs that demonstrate a significant deviation from HWE test.

Although we conducted several steps for quality assessment control for our dataset but still there are several summary statistics tests for the detection of per-SNP and per-Individual available in software like PLINK toolset, which will be considered in our future work

5. **Conclusion**

The interpretation of GWAS of case-control association requires the proper use of summary statistic tests for quality assessment and control to reduce bias in the results. This study used an existing data from The Wellcome Trust Case Control Consortium (WTCCC) Type 2 diabetes to perform quality assessment control steps. The dataset is in plink format and the size of the files are very large to be handled, thus the dataset was analysed using rigorous data science techniques using plink toolset and R packages. This first led to store large genome data in more efficient structure and thus improved our methodology to handle this data. Second enhanced our understanding of the data structure and features, and helped us to modify the data in more efficient way. A series of quality control steps were performed to remove data items that introduce bias to the study.

Although these approaches show, the steps that have been conducted for filtering the dataset are promising, a number of explorations into this dataset remain. More specifically, even after considering comprehensive quality control filtering, it is impossible to identify and eliminate all inherent genotyping errors in the dataset. As the genotyping errors of some SNPs and samples are not significantly poor to be marked as a red indicator in the quality control filtering step and thus it will remain in the dataset.

References
[1] D. Altshuler, E. Lander, and L. Ambrogio, "A map of human genome variation from population scale sequencing," *Nature*, vol. 476, no. 7319, pp. 1061–1073, 2010.
[2] W. S. Bush and J. H. Moore, "Chapter 11: Genome-Wide Association Studies," *PLoS Comput. Biol.*, vol. 8, no. 12, 2012.
[3] T. W. T. C. C. Consortium, "Genome-wide association study of 14 000 cases of seven common diseases and 3 000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–678, 2007.
[4] M. E. Tabangin, J. G. Woo, and L. J. Martin, "The effect of minor allele frequency on the likelihood of obtaining false positives.," *BMC Proc.*, vol. 3 Suppl 7, no. Suppl 7, p. S41, 2009.
[5] A. C. Lam, M. Schouten, Y. S. Aulchenko, C. S. Haley, and D.-J. de Koning, "Rapid and robust association mapping of expression quantitative trait loci.," *BMC Proc.*, vol. 1 Suppl 1, p. S144, 2007.
[6] J. C. Florez, A. K. Manning, J. Mcateer, and K. Irenze, "ORIGINAL ARTICLE A 100K Genome-Wide Association Scan for Diabetes and Related Traits in the Framingham Heart Study," *Diabetes*, vol. 56, no. December, pp. 3063–3074, 2007.
[7] C. a Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, P. Morris, and K. T. Zondervan, "Data quality control in genetic case-control association studies," *Nat. Protoc.*, vol. 5, no. 9, pp. 1564–1573, 2011.
[8] S. E. Inzucchi, R. M. Bergenstal, J. B. Buse, M. Diamant, E. Ferrannini, M. Nauck, A. L. Peters, A. Tsapas, R. Wender, D. R. Matthews, American Diabetes Association (ADA), and European Association for the Study of Diabetes (EASD), "Management of hyperglycemia in type 2 diabetes: a patient-centered approach: position statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD).," *Diabetes Care*, vol. 35, no. 6, pp. 1364–1379, 2012.

[9]   N. Hex, C. Bartlett, D. Wright, M. Taylor, and D. Varley, "Estimating the current and future costs of Type1 and Type2 diabetes in the UK, including direct health costs and indirect societal and productivity costs," *Diabet. Med.*, vol. 29, no. 7, pp. 855–862, 2012.

[10]   Diabtes UK, "The cost of Diabetes," *Diabtes UK*, pp. 1–20, 2014.

[11]   J. Gulcher and K. Stefansson, "Clinical risk factors, DNA variants, and the development of type 2 diabetes.," *N. Engl. J. Med.*, vol. 360, no. 13, p. 1360; author reply 1361, 2009.

[12]   R. B. Prasad and L. Groop, "Genetics of type 2 diabetes—pitfalls and possibilities," *Genes (Basel).*, vol. 6, no. 1, pp. 87–123, 2015.

[13]   C. Laurie, K. Doheny, and D. Mirel, "Quality control and quality assurance in genotypic data for genome-wide association studies," *Genet. …*, vol. 34, no. 6, pp. 591–602, 2010.

[14]   L. Qiu, R. Na, R. Xu, S. Wang, H. Sheng, W. Wu, and Y. Qu, "Quantitative assessment of the effect of KCNJ11 gene polymorphism on the risk of type 2 diabetes," *PLoS One*, vol. 9, no. 4, 2014.

[15]   N. M. Phani, V. Guddattu, R. Bellampalli, V. Seenappa, P. Adhikari, S. K. Nagri, S. C. D'Souza, G. P. Mundyat, K. Satyamoorthy, and P. S. Rai, "Population specific impact of genetic variants in KCNJ11 gene to type 2 diabetes: A case-control and meta-analysis study," *PLoS One*, vol. 9, no. 9, 2014.

Sarmad Abdulazeez

Simulation of Massively Multiplayer Online Games Communication using OPNET Custom Application

S A Abdulazeez, A El Rhalibi and D Al-Jumeily

Department of Computer Science, Faculty of Engineering and Technology
Liverpool John Moores University
Liverpool, UK
S.A.Abdulazeez@2013.ljmu.ac.uk, {A.Elrhalibi, D.Aljumeily}@ljmu.ac.uk

Abstract. In recent years, there has been an important growth of online gaming. Nowadays, Massively Multiplayer Online Games (MMOGs) may involve millions of synchronous players scattered across the world and participating with each other within a single shared game. In this paper, we propose a new technique to communicate between players and game server, also among players each other based on hybrid Peer-to-Peer architecture. We propose to use OPNET Modeler 18.0, and in particular the custom application to simulate the new architecture, which required the implementation of the new nodes models and behaviors in the simulator to emulate correctly the new architecture. We use OPNET Modeler 18.0 to simulate the network, applying two transport protocols TCP and UDP, and with different scenarios. The scenarios include both client-server and hybrid P2P system to evaluate the communication of games with (125, 500, and 1000) peers. The results show that the hybrid Peer-to-Peer system produce low delay, low traffic received in game communication compared with client-server system for all scenarios used in the simulation.

Keywords: MMOGs; OPNET Modeler; TCP; UDP; hybrid Peer-to-Peer

1. Introduction

Online games generally involve a lot of actions that require fast response and low delay, and lost packet retransmission is impractical in these circumstances. Massively Multiplayer Online Games (MMOGs) are a type of online interactive game. As a result, traffic generated by online interactive games commonly consists of huge number of UDP packets. The traffic is also highly periodic because this requires consistency and updating between the players and the game server states. One of the key important features in these games is the massive number of players who play concurrently over the Internet. The huge number of players leads to a more interactive, complex, and attractive game environment. The main reason that makes MMOG's popularity increasing so rapidly in recent years is the experience of playing with other human players. With the increasing number of players in MMOGs, It became difficulty to control efficiently the communication between the players and the game server, and between players. There are historically two main architectures that used for MMOGs communication, although new cloud based architectures are being currently developed. The traditional architecture called client-server, in which a centralized server has the authority to perform the majority of the game logic and then send a game state update to the clients for rendering. The responsibility of client is to collect the user commands and send them to the server for processing. In this architecture, the requirement of computational power and bandwidth for a single client is minimal. However, considerable

resources and consecrated support staff are required for a large number of players at the server side. In most case this can lead to a high delay, as well as high traffic of sent and received packets, which may consequently lead to low game performance. The other architecture called Peer-to-Peer architecture. P2P architecture corresponds to a network without a centralized server or control. Peer-to-peer networks allow game designers to transfer a big part of the game processing and load of communication bandwidth to their participant's peers. However, this architecture has drawbacks to manage the system due to there being no centralized control to guarantee each player in the game uses the same software, and does not engage in irregular game activities. In this paper, we use hybrid Peer-to-Peer architecture [1] for game communication between players and game server, which is implemented and deployed in OPNET Modeler 18.0 using custom application. Standard application in OPNET Modeler [2] does not support in game communication, therefore, we use custom application to deal with this aspect of the traffic and carry out evaluation of the different architecture using different scenarios. The rest of the paper is organized as follows. Section 2 presents a review of the state-of-the-art of MMOGs simulation systems. Section 3 describes the MMOGs communication in client-server system, while section 4 introduces the MMOGs communication in P2P system and section 5 MMOGs communication in hybrid P2P system. Section 6 illustrates the custom application in OPNET Modeler. Section 7 and 8 introduce the simulation and results. Section IX presents the conclusion.

2. Literature Review

Game simulation is widely used evaluate new architectures and protocols. The game communication is considered one of the most important aspects in the game simulation. Some researches tried to deal with the simulation of massively multiplayer online game communication. Mohorko et al. present an overview of some suitable simulation tools for research on network technologies and communication infrastructure. The authors also show different possibilities of using the advanced simulation frameworks for simulating the tactical networks. This system is of limited use for OPNET. This limitation leads to difficulties to control and change the communication model or create a new custom application [3]. Shirmohammadi et al. propose a P2P communication architecture for game network to provide reliability for important messages and decreasing network congestion by acknowledging only key messages. The system proposed called (HDSP) Hybrid Distributed Simulation Protocol. HDSP supports both best-effort delivery, for frequently occurring messages, and reliable delivery for important or "key" messages. HDSP simply implements the standard UDP protocol [4]. Denault and Kienzle used simulation in the context of MMOGs. They use Mammoth to simulate MMOGs to evaluate performance measurements such as CPU usage, memory usage. The authors define five simulation setups within the Mammoth MMOG framework, four using a single computer and the others in a distributed setting. The limitations of this system are the use of TCP protocol only, and there is no way to clearly define the game communication in the simulation [5]. Webb et al. develop an application layer Network Game Simulator (NGS) and successfully simulated Client/Server system, Region based, and Neighbor based architectures. The simulator is very flexible and can easily be extended to contain new features and architectures. The disadvantages of using this simulator are that there is no complex routing algorithms, and NGS does not provide support to simulate the network stack. In addition this simulator has been used to just perfum preliminary evaluation [6]. All these researches do not use OPNET Modeler simulation to simulate the game communication or the custom game communication. The used of OPNET custom game communication [7] provide more flexible features to manage the communication and control the traffic flow between network devices in both architecture, and support both TCP and UDP protocols. The next section will introduce the MMOGs communication in client-server system.

3. MMOGs Communication in Client-Server Architecture

The vast majority of the MMOGs have been developed using the classic client-server architecture. This architecture demonstrates some problems such as all the load of handling the virtual world is carried by the server. The players connect to the server through the client application on their own computer and the server is responsible for handling all the rules and the state of the virtual world. One of the first things we have to be aware of is that due to the big number of clients that will connect to the server as well as the massive load that they create it will be impossible for one single server machine to handle everything. A lot of effort is put into developing better and more effective ways to deploy the load over multiple servers. This method can achieve the largest possible number of the participating players and experience with a minimum of needed resources.

One of the most common optimizations used by game communication is area of interest (AoI) management. Individual players usually only interact with a small area of the game world at any one time. Servers only need to update clients about objects in this AoI. This leads to reduce the number of objects transferred from the total set of objects to the number of objects in this area [8].

4. MMOGs Communication in Peer to peer Architecture

The primary requirement for MMOGs based on Peer-to-Peer architecture is to maintain state consistency, and a shared sense of virtual space among great numbers of players without the need for support by the server [9]. The development of a peer to peer architecture for a MMOG is a significant research topic. There is a great deal of research which has been done on the use of peer to peer architectures to design MMOGs, but until now this has not led to it being used in any commercial MMOG. Because the use of peer-to-peer system to design and develop multiplayer games is a new phenomenon and it hasn't reached the mainstream consumer market. P2P architecture allows clients to communicate directly with each other and they are responsible for a small portion of the game state, and therefore do not need for a central game server. Dispensing the need for a central game server, you will decrease the cost of running a game dramatically. However, there are a lot of problems when designing and developing a peer-to-peer architecture for MMOGs. For example, a major problem will be to save the game state, the deployment of game updates will be more troublesome, and the bandwidth used for the clients will be extremely higher when compared with the client-server architecture. Peer-to-peer architectures do not have central server and therefore it becomes most difficult to present single consistent virtual world to all users. In Peer-to-Peer system, distributing the game state over all the peers in the game world, so there will have to be an effective way to set different elements of the game world to different peers [9].

5. MMOGs Communication in Hybrid Peer to Peer Architecture

Massively Multiplayer Online Games are becoming more widespread each year with diverse new technology being released on platforms. These developments make the game more exciting to attract additional numbers of players to the game. The increase in the number of players creates issues to be faced by gaming companies such as adding more servers to deal with this problem. However, this solution is not always possible because of the cost of adding new server is high, and the communication of state dissemination is also more complex when the players are managed from different servers. To better solve this problem, we propose a new technique based on Hybrid Peer-to-Peer system [1].

6. Custom Application in OPNET Modeler

The custom application is a generic model framework that you can use to represent a broad class of applications in OPNET. It can be used when the application needed in simulation does not correspond to any of the standard applications. The standard applications in OPNET Modeler are for example FTP, E-mail, HTTP, Remote Login, Database, Video Conferencing, Print, Peer-to-Peer Sharing, Video Streaming, and Voice. The custom application provides attributes that allow you to configure different aspects of the application

in detail. Custom application can be used to represent complex multi-tier applications. The following components are used within the custom application [10]:

- Task: A basic unit of user activity within the context of an application. For example: reading an e-mail message, obtaining records from a database or send a query to the server. The start and end of a task must be clear defined.
- Phase: An interval of related activity that is contained in a task, for example a processing phase and a data transfer phase are specific phases of a task.

7. Simulation Design

In this research, we have used OPNET Modeler 18.0 to implement and simulate the proposed architecture (MMOGs based on hybrid Peer-to-Peer system) and compare the results with the traditional architecture (Client-Server system). Furthermore, we compared transport protocol TCP versus UDP for each scenario in the project. The network topology consists of three scenarios for each architecture which will be explained in the following sections.

7.1 Network Topology for Client-Server system

This Client-Server architecture consists of the following devices: game server, two Cisco C400 Router, IP Cloud, Ethernet Switch, and Ethernet Workstations as shown in figure 1. The connection mechanism between these devices is described as follow: The game server connects with the IP Cloud through Cisco Router. The IP cloud connects with the Ethernet Workstations (Peers) through both Cisco C400 Router (Client gateway) and Ethernet Switch. The game applications are controlled and managed by the central server. We use task configuration to configure the custom application for this scenario.
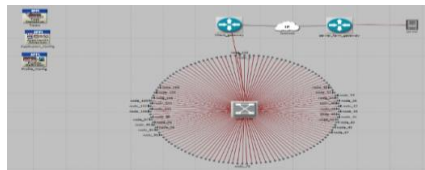
Figure 1. Client/Server scenario
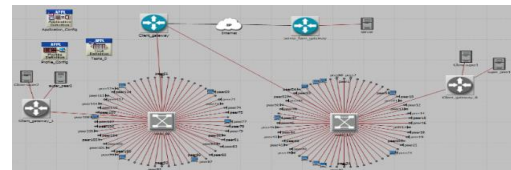
Figure 2. Hybrid P2P scenario

7.2 Network Topology for Hybrid P2P system

The hybrid P2P scenario has the same devices as in scenario 1 but with change in the zone area by adding super-peer and clone-super-peer to the region. Each region has 60 peers connected to both super-peer and clone-super-peer using the super-peer gateway router as shown in figure 2. The game application is controlled and managed by the region super-peer and clone-super-peer. We use task configuration to manually configure the custom game application for hybrid P2P system.

7.3 Custom Application for Game Communication

As the hybrid P2P architecture is not currently supported in OPNET, or any other simulator, we have created a custom application for game communication between client and server in client-server architecture, and between peers, super-peers, clone-super-peers, and server in hybrid P2P architecture. The custom application for client-server architecture consists of two phases. The first phase is to communicate between clients and server; however the second phase is to communicate between server and clients. However, the custom application for hybrid P2P architecture consists of 20 phases for the scenario with 125 peers; however the number of phases will increase whenever the number of peers is increasing till up to 80 phases for the scenario with 1000 peers and 16 regions. The number of regions will increase with the increase the number of peers. Figure 3 shows the custom application for hybrid P2P architecture.
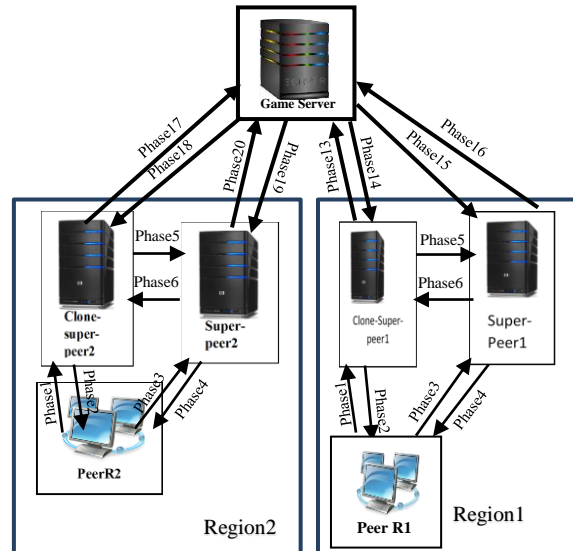
Figure 3. Custom Application phases in Hybrid P2P System

8. Simulation Results

With the scenarios that have been introduced previously, we have get to the following simulation results:

8.1 Overall delay

This parameter is defined as the overall end-to-end delay for all packets received by the station. End- to-end delay for the application used during the simulation is measured from the time from source to destination. End-to-end delay refers to the time it takes to send a packet from source to destination over a network. Figure 4 shows the overall delay for client-server architecture compared to the results of the overall delay for the hybrid P2P architecture using TCP transport protocol. The figure below illustrates a great variation of delay when using client-server system for game communication compared with hybrid P2P system. In addition, figure 8 shows the effect of adding background traffic on base links of the network. However, figure 5 shows the overall delay for the same scenario but using UDP transport protocol instead of TCP. Further, the figure illustrates the impact of adding background traffic on base links of the network. Figure 5 illustrates that using UDP transport protocol is more stable for both architectures, as well as produce low delay.
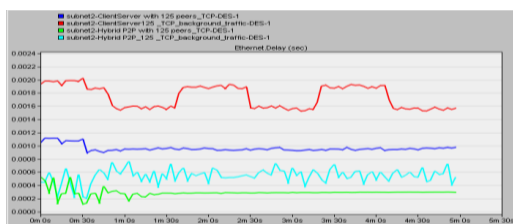
Figure 4. Overall Delay for Client/Server and Hybrid P2P with 125 peers TCP Protocol with Background Traffic
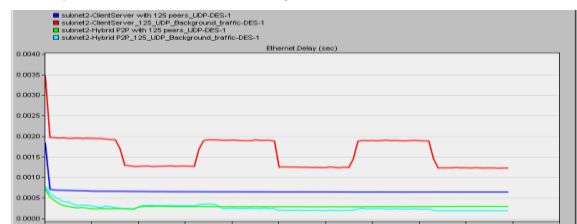
Figure 5. Overall Delay for Client/Server and Hybrid P2P with 125 peers UDP Protocol with Background Traffic

Figure 6 shows the overall delay for both Client/Server and Hybrid P2P architecture with 500 peers using TCP transport protocol and the effect of adding background traffic on base links of the network. The figure below elucidates the using of hybrid P2P system produces less delay compare with client-server system. However, figure 7 shows the overall delay for both Client/Server and Hybrid P2P architecture with 1000 peers using TCP transport protocol without adding background traffic on base links of the network.
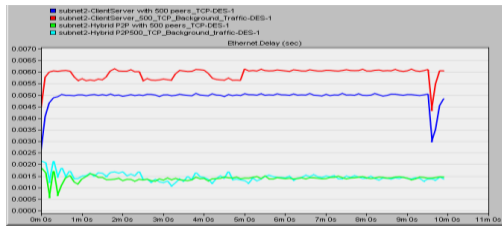
Figure 6. Overall Delay for Client/Server and Hybrid P2P with 500 peers TCP Protocol with Background Traffic
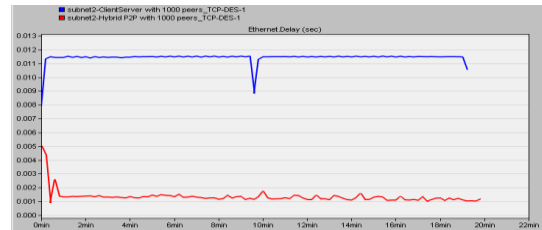
Figure 7. The overall Delay for Client/Server and Hybrid P2P with 1000 peers TCP Protocol without Background Traffic

8.1 Traffic Received

Traffic received is the average number of bytes per second received by all the nodes in the network. In other words, network traffic is the amount of data moving across a network at a certain point of time. Network traffic is one of the main ingredient for measuring network traffic, network traffic control and simulation. The proper regulation for network traffic helps to ensure the quality of service for the network. Figure 8 and 9 shows the custom application traffic received for client server and hybrid P2P system with 125 peers using TCP and UDP transport protocol respectively. As well as the effect of adding background traffic on base links of the network.
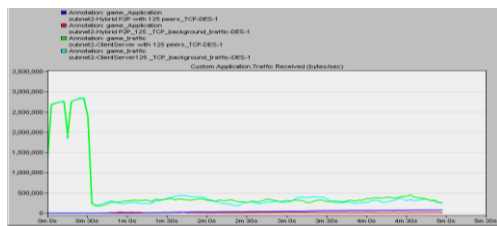
Figure 8. Traffic Received for Client/Server and Hybrid P2P with 125 peers TCP Protocol with Background Traffic
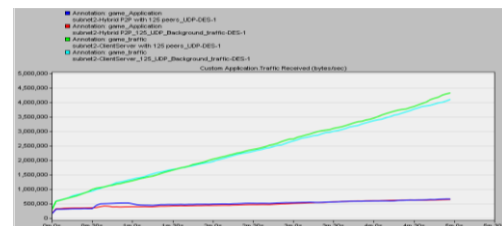
Figure 9. Traffic Received for Client/Server and Hybrid P2P with 125 peers UDP Protocol with Background Traffic
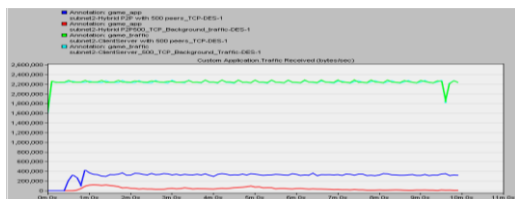
Figure 10. Traffic Received for Client/Server and Hybrid P2P with 500 peers TCP Protocol with Background Traffic
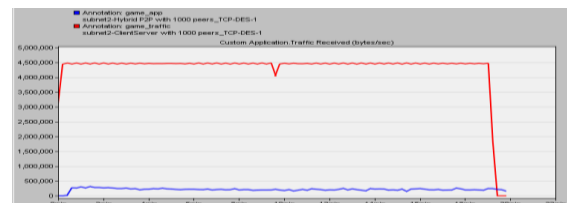
Figure 11. Traffic Received for Client/Server and Hybrid P2P with 1000 peers TCP Protocol without Background Traffic

Figures 10 and 11 illustrate the traffic received for both client-server and hybrid P2P architectures with 500 and 1000 peers using TCP transport protocol. The figures above show that the traffic received by client server architecture is much higher than the traffic received by hybrid P2P architecture for both scenarios. The traffic received for client server is duplicated when the number of clients is duplicated from 500 to 1000 clients; however, the traffic received for hybrid P2P is slightly higher with 1000 peers. Due to the using of hybrid P2P architecture is more organized and more efficient compare with client-server architecture.

9. Conclusions

In this paper, we have reviewed the research related to the MMOGs communication and simulation, as well as explained the advantages and limitation of these researches. We also

highlight the subjects and issues related to our research and explain the main contributions of the paper. We use OPNET Modeler 18.0 to model and simulate the communication networks. We have used OPNET simulation to enable the networks construction, study of communication infrastructure, design of individual devices, and simulation of protocols and applications. The results illustrate that the hybrid Peer-to-Peer system produce low delay and low traffic received in game communication when compared with client-server system for all scenarios used in the simulation.

10. References

[1]   S. A. Abdulazeez, "Survey of Solutions for Peer-to-Peer MMOGs," *ICNC 2015*, no. California, USA.

[2]   S. A. Sethi and V. Y. Huatyshin, *The Practical OPNET® User Guide for Computer Network Simulation*. CRC Press, Taylor & Francis Group, 2013.

[3]   J. Mohorko, F. Matjaž, and K. Saša, "Advanced modelling and simulation methods for communication networks," *Microw. Rev.*, pp. 41–46, 2008.

[4]   S. Shirmohammadi, A. Diabi, and P. Lacombe, "A Peer-to-Peer Communication Architecture for Networked Games," *Communication*, 2005.

[5]   A. Denault and J. Kienzle, "The perils of using simulations to evaluate Massively Multiplayer Online Game performance," *3rd Int. ICST Conf. Simul. Tools Tech. ICST (Institute Comput. Sci. Soc. Telecommun. Eng.*, 2010.

[6]   S. D. Webb, W. Lauel, and S. Soh, "NGS : An Application Layer Network Game Simulator," *Proceeding IE '06 Proc. 3rd Australas. Conf. Interact. Entertain.*, no. December, pp. 15–22, 2006.

[7]   "Applications Model User Guide," in *Guide Standard Models, Modeler Relase 10.0*, pp. 1–36.

[8]   B. Hughes, J. Haggerty, J. Nothman, S. Manickam, and J. R. Curran, "A Distributed Architecture for Interactive Parse Annotation," *Proc. Australas. Lang. Technol. Work.*, pp. 207–214, 2005.

[9]   L. Fan, "Design issues for Peer-to-Peer Massively Multiplayer Online Games," *Int. J. Adv. Media Commun.*, vol. 4, no. 2, pp. 108–125, 2010.

[10]   Z. Lu and H. Yang, *Unlocking the Power of OPNET Modeler*, First publ. Cambridge University Press, The Edinburgh Building, Cambridge CB2 8RU, UK: The United States of America by Cambridge University Press, New York, 2012.

Ruqayah Al-Dahhan

Context-Aware Cloud-Based Access Control for the Internet of Things

Ruqayah Rabeea Al Dahhan and Qi Shi
School of Computing and Mathematical Science, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK
R.R.Aldahhan@2015.ljmu.ac.uk, Q.shi@ljmu.ac.uk

Abstract. Although Cloud computing and the Internet of things (IoT) are very different technologies, in recent years, they are tightly coupled with each other. The combination of them provides a universal environment by enabling data collecting services and powerful processing of the collected data. Whereas IoT produces rich context information, the cloud effectively serves as the brain to improve decision-making based on the information provided. To protect the stored data in clouds, one of the main mechanisms used in the cloud environment is access control with its main responsibility of managing users' access rights. The general aim of this proposed project is to form a spontaneous coalition between the cloud and IoT in a specific way where IoT provides accurate, dynamic context information to the cloud which in turn improves decision-making to access to its resources. This aim will be applied to some applications that have digital contents such as e-books, videos, games and so on. These applications are becoming pervasive in the era of cloud, and need a mechanism to prevent them from being obtained by inappropriate users. This will be achieved by offering a novel model for dynamic access to cloud resources in response to relevant contextual information changes such as locations and identities offered by IoT which in turn allows users to delegate their resources to others providing more flexibility.

Keywords: Cloud computing, Access control, Context awareness, Internet of things.

1. Introduction

Cloud computing in recent years is tightly coupled with the Internet of things (IoT). The two technologies are very different. However, they are both already vital part of our life. Cloud computing is a paradigm for enabling convenient, on-demand network access to a shared pool of computing resources (Botta et al., 2016). These resources are offered easy accessibility through the network anywhere and anytime. Despite several advantages of cloud computing such as unlimited storage and the efficiency of computing processes, there are concerns about the security and privacy in these environments. On the other hand, IoT creates a world where all things around us are connected to the internet and interact with each other almost without human intervention. Though IoT represents one of the most efficient scenarios that enables ubiquitous and persistent computing, the restricted storage and processing capacity, security, and privacy are still its main issues (Gubb, 2013). To provide a universal environment, the combination of cloud computing and IoT can enable data collecting services and powerful processing of the collected data. Exploiting the data collected by IoT with enhanced access control to objects in the cloud environment is crucial

for the wide adoption and deployment of the technologies. Whereas IoT produces rich context information, the cloud effectively serves as the brain to improve decision-making based on the information provided.

Unlimited storage in the cloud environment attracts most of the data owners and the organizations to store their data in it (Bouabana-Tebibel & Kaci, 2015). Much of the stored data is highly sensitive, so security and privacy are very important issues in the cloud computing. To protect the data, one of the main mechanisms used in the cloud environment is access control with its main responsibility of managing users' access rights. It grants access to authorized users and forbids others access to data (Li et al., 2013). Many system models and algorithms for access control have been characterized and described by researchers to provide secure and efficient cloud access control (e.g. Thangavel & Varalakshmi, 2014). Whereas most of these access control schemes are based on static access privileges, the main challenge is how to devise an active/dynamic access control model (Toninelli, 2006), which is aware of the context correlated with ongoing activities in the environment concerned. This enables the context based access control in real time, which is essential for IoT applications. The development of such a model requires an integration of contextual information and access control.

Many applications have been developed in the case of providing active access control. For example, Huang & Su (2015) proposed an access control model for running a rating system in cloud. This system is to prevent specific digital contents from being obtained by inappropriate users. However, this model is limited and only provides Yes or No admission control decision. In terms of active access control, it might be possible to extend this system for access by some content users during a specific time period or under specific conditions.

In this project, developing a flexible, active access control model based on context information gathered by the IoT technology is the proposed work. The rest of this paper is structured as follows. The literature review is presented in section 2 in which we classify some existing access control models and specify the most current problems. Section 3 shows the project aims and objectives. The methodology is presented in section 4. Section 5 draws some conclusion

2. Literature review

Relevant topics will be reviewed further to address the current issues of access control and context awareness in IoT environments.

2.1. Access control

One of critical security mechanisms for data protection in cloud applications is access control. As a result of a dynamic and large user base, the sensitive data stored in the cloud requires fine-grained access control (Ahmadi et al., 2015). To achieve this, attribute-based access control has been introduced. However, to hide the data from the storage server, encrypted data is stored in cloud data servers. Thus data encryption with attribute-based access control is known as an attribute-based encryption (ABE) technique. In general, attributes are classified into two types: 1) non-temporal attributes with discrete attribute values (e.g. age, address, email, etc.), and 2) temporal attributes (e.g. interval, time, etc.).

Each cloud access control mechanism has crucial requirements (Ruj, 2014). One of the most important prerequisites which the system has to support is user revocation. That means that a revoked user, after it has been revoked, will no longer have the previous assigned right for further data access and decryption of related cipher text/data.

Many studies have been carried out on the access control of cloud data using discrete attribute values (Ruj et al., 2011; Yang et al., 2013; Ruj et al., 2014). These researches have a lot of problems which remain unsolved. For example, Ruj et al. (2011) could not handle a revocation problem in their model. Though Yang et al. (2013) provided a solution to the problem efficiently in terms of access computation costs, the work incorporated weak privacy and security considerations. On the other hand, some schemes require intensive computations in return for stronger privacy and security protection,

meaning that they are unsuitable for mobile devices with limited computation power (Khan & Hamlen, 2012; Yassin, 2012; Ruj et al., 2014).

In addition, some work has been carried out using continuous attributes, which is known as temporal access control (Yang et al., 2012; Zhu et al., 2012). This access structure can be of the form of time, e.g. between 8 am to 12 pm. These temporal attributes are very familiar in cloud. For instance, only during a particular period of time, users can access certain data. However, such schemes have their shortcomings. For example, the scheme of Zhu et al. (2012) does not address user revocation, and the scheme of Yang et al. (2012) manages the revocation problem inefficiently by refreshing an update key and sending it to all users at every time slot with a valid set of attributes which represent the revocation.

To achieve effective and efficient access control, we need a model that uses both discrete and temporal attributes/values. For example, new employees in a company have an initial access right to specific data. After that, they are permitted to join a group in the same company and there is a need to access more sensitive data under limited conditions and during a specific period of time. However, little work has been done on cloud access control using both discrete attribute values and temporal attributes (Balani and Ruj, 2014). Even though the schema in (Balani and Ruj, 2014) addresses the user revocation problem, it is not active.

2.2. Context awareness

Due to rapid changes in users' context, the use of the context information is crucial in interactive applications, particularly for ubiquitous computing applications (Kulkarni & Tripathi, 2008). Context can be defined as "any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves" (Abowd et al., 1999). If a system uses the context to supply relevant information to a user who uses it in a specific task, it is considered as context-aware (Perera et al., 2014).

Although different researchers have developed context-aware access control models, the use of context has targeted at detecting devices and network environments that were used to request an access to the cloud data. This information can be useful to detect the computational power of the devices or to check if the requested data will be transmitted through insecure channels (Choi& Kwak, 2013, Gravier et al., 2015).

However, the active access control model requires a much broader scope of context and is centred on the context which consists of all the characterizing information considered relevant to it. This context can be identified as identities, locations, times or activities which are collected and labelled in some meaningful way and represented in terms of attributes (Zhou, 2013).

In our proposed work, we will develop a novel, active access control model which uses context awareness to restrict cloud data access in response to context changes so as to enable a dynamic adaptation scheme. This context is generated by processing the raw data reported by sensors in smart environments.

3. Research objectives
3.1. Objectives
To accomplish the aim of this proposed project, the following objectives are set out to:

1. Examine the current approaches to cloud access control and context computing for IoT by performing a thorough literature review, and identify their benefits and shortcomings.
2. Develop an attribute based encryption technique in a specific way that is able to handle both the temporal and non-temporal attributes with reference to solving the user revocation problem.

3. Construct the proposed framework by aggregating data using IoT technologies and making decisions using cloud based mechanisms, which currently exhibit limitations with respect to context awareness.
4. Implement and evaluate the constructed framework based on case studies.

3.2. Novelties

The novelties of this proposed project include:

1. Developing a new fusion method to collect data from sensors and then extract the context using the IoT technology. The context information produced is passed to the cloud and used to make decisions on active access control of the objects in the IoT environment and data collected and stored in the cloud. The current research mostly grants passive permissions for users based on their attributes which are defined in advance from the given access control policy.
2. Discovering a novel practical way to delegate the access right to someone who will use others' resources under limited conditions according to its current situation. This will provide more flexible delegation than relevant existing models which normally follow access policies.
3. Using both descriptive and temporal attributes, which are computed by the IoT technology, to design an access control framework. Currently the revocation problem has not been addressed efficiently by the existing systems which use just temporal attributes in their access control policies. We will devise an integral model to solve the revocation problem.

4. Methodology

Identifying the key cause of the problems described in section 2 is the first step in this research. It will be followed by creating new solutions to rectify the problems. More specifically, the methodology for this project consists of four distinct phases: (1) literature review, (2) requirement analysis and specification, (3) framework design, and (4) implementation and evaluation. The framework design phase also involves several stages as shown in the sub-sections below.

4.1. Literature review

To fill the knowledge gap and solve the problems identified, we will perform a thorough literature review of cloud oriented access control and current methods for data collection using the IoT technology. In addition, we will analyse their limitations and try to find appropriate solutions to overcome these restrictions.

4.2. Requirements analysis and specification

In this phase, the relevant existing work will be critically analysed to determine the needs or conditions for the newly proposed framework to meet, by taking account of the possibly conflicting requirements of the two different technologies which are cloud and IoT. Since requirements analysis is critical to the success of the proposed framework, the requirements derived must be practical and related to identified system needs. Based on these considerations, the initial requirements will be selected and specified.

4.3. Framework design

Based on the analysis from the previous phase, the essential requirements for the design of our proposed work will be extracted. The design will be divided into two main stages: 1) designing an initial cloud oriented access control model including an algorithm to solve the revocation problem stated earlier, 2) designing a model for the collection of raw data using the IoT technology and extraction of context-awareness information from the data, which is utilized to make access control decisions and activate necessary delegation properties. These two models will be integrated together to produce the complete framework proposed.

4.3.1. Design of the initial access control model

This model will consider temporal and descriptive attributes in its design. It will initially involve granting passive permissions and solving the revocation problem. We intend to adopt one of the most popular variants of attribute based encryption techniques (ABE), Ciphertext-Policy ABE (CP-ABE). This technique requires that data owners encrypt their data and then send it to the cloud for storage, where the secret encryption keys are generated from the given attribute based access policies. Therefore, the authorized users who possess the required attributes in a right combination, i.e. satisfy the access policies, can recover the keys to successfully decrypt the encrypted data.

However, CP-ABE only works fine when the attributes are descriptive, and temporal attributes are not well handled by CP-ABE. Extending the CP-ABE technique in a specific way to handle both temporal and non-temporal attributes is the first step in this stage. The next step will be focused on how to adapt the extended technique to efficiently manage the user revocation. We intend to solve this problem by identifying a user revocation list as well as introducing extra blinding on the secret key components. The final step in this stage will address how to grant passive permissions to authorized users using static data collected from the users during their registration.

4.3.2. Collecting data and designing the active access control model

In this stage, IoT will be integrated with the cloud. Therefore, the user's benefits will be achieved by utilizing services offered by IoT and controlled by the cloud as well as accessing cloud services. As a result, we will explore a way to centralize decision making and manage the devices within the cloud. Two steps will be used to satisfy this vision as described below.

4.3.2.1. Designing system middleware

One of the most important components in the proposed work is middleware. It is a software layer mediating between the application layer and the physical world. It will be designed to: 1) help a customer/ user, who is lack of knowledge about communication and processing, to perform the mediation, 2) simplify the integration of the IoT and cloud technologies, and 3) store necessary facts from users in the designated database for decision making and authority delegation processes. The middleware will be located in the cloud.

4.3.2.2. Designing the active access control model and authority delegation

During this stage, two types of context information are considered. The first one is to show if two authorized actors/users are co-located, meaning they are both at the same place. Then, their locations are determined and stored as a fact in a specific place which is used to build the description of the current situation. The second information which is taken into our consideration is time.

Collecting the required contextual information will be done using sensors deployed in smart environments. Many processes will be applied to the collected data to accurately gain the context which is required by the proposed system. These processes are for modelling and reasoning the data.

After the measures for the collection and determination of the context have been devised, the design of the active access control model will begin. Clearly, the access policy decision will be influenced by the context, which will be used to build the access structure policy, due to its inherently dynamic nature. Thus, designing a flexible access control model/system is a challenging task, which has to manage the dynamic nature of contextual information.

Furthermore, the context information will be used to determine necessary conditions for the delegation property and also to restrict and regulate the privileges delegated to the authorized users. Due to the dynamic nature of the context and IoT, the key challenge of this task is how to design a dynamic authority delegation scheme with proper control over the delegation process.

Therefore, according to the above discussion, some requirements should be considered throughout this stage of design for a fine grained access control model in a

highly dynamic environment. These include the issues of how to design dynamic rules for assigning dynamic access rights to particular users in relation to their attributes, and also how to design a dynamic mechanism to assign these rights to users based on these rules.

4.4. Implementation and evaluation

The final phase of this proposed system will consist of two stages that are the implementation and evaluation of the designed framework. In the implementation stage, we will implement the designed framework including the models and methods mentioned earlier, and test it according to the requirements specified. The IoT and cloud components of the framework will be simulated and the selection of test data will be first by choosing data that is compatible with applications with digital contents. Then, the results will be interpreted.

In the second stage of this project, the implemented framework will be evaluated and compared against the relevant existing work to determine any benefits our framework offers and any problems it may experience, which may lead to improvements of the framework. Finally, using the diverse data of different applications to prevent children from watching inappropriate programmes will be tested as a case study and evaluated to see the effectiveness of the proposed framework and measure its application compatibility over a number of iterations.

5. Conclusion

In this paper, the main issues of the existing models such as user revocation and lack of active access control systems have been addressed. In the proposed system, we are focusing on the integration of IoT and cloud computing to generate an active access control system. This system grants adaptable access rights to users/clients depends on their current situations by utilising dynamic, relevant contextual information which is gathered by the IoT technology. The system will efficiently solve the revocation problem as well as activating the delegation capability.

References:

Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., & Steggles, P. (1999) Towards a better understanding of context and context-awareness. In *Handheld and ubiquitous computing,* pp. 304-307. Springer Berlin Heidelberg.

Ahmadi, M., Chizari, M., Eslami, M., Golkar, M. J., & Vali, M. (2015) Access control and user authentication concerns in cloud computing environments. In *Telematics and Future Generation Networks (TAFGEN), 2015 1st International Conference on,* pp. 39-43.

Balani, N., & Ruj, S. (2014) Temporal access control with user revocation for cloud data. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2014 IEEE 13th International Conference on*, pp. 336-343.

Botta, A., de Donato, W., Persico, V., & Pescapé, A. (2016) Integration of cloud computing and internet of things: A survey. *Future Generation Computer Systems, 56, March 2016*, pp. 684-700.

Bouabana-Tebibel, T., & Kaci, A. (2015) Parallel search over encrypted data under attribute based encryption on the cloud computing. *Computers & Security, 54, October 2015*, pp.77-91.

Choi, S. K., & Kwak, J. (2013) Context-aware information-based access restriction scheme for cloud data. *International Journal of Multimedia and Ubiquitous Engineering, 8*(6), pp.97-104.

Gravier, C., Subercaze, J., Najjar, A., Laforest, F., Serpaggi, X., & Boissier, O. (2015) Context awareness as a service for cloud resource optimization. *Internet Computing, IEEE, 19*(1), pp. 28-34.

Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013) Internet of things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems, 29*(7), 1645-1660.

Huang, W. B., & Su, W. T. (2015) Identity-based access control for digital content based on ciphertext-policy attribute-based encryption. In *Information Networking (ICOIN), 2015 International Conference on,* pp. 87-91.

Khan, S. M., & Hamlen, K. W. (2012) AnonymousCloud: A data ownership privacy provider framework in cloud computing. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on,* pp. 170-176.

Kulkarni, D., & Tripathi, A. (2008) Context-aware role-based access control in pervasive computing systems. In *Proceedings of the 13th ACM symposium on Access control models and technologies*, pp. 113-122, ACM.

Li, X., & Zhao, X. (2013) Survey on access control model in cloud computing environment. In *2013 International Conference on Cloud Computing and Big Data (CloudCom-Asia),* pp. 340-345. IEEE.

Perera, C., Zaslavsky, A., Christen, P., & Georgakopoulos, D. (2014) Context aware computing for the internet of things: A survey. *Communications Surveys & Tutorials, IEEE, 16*(1), pp. 414-454.

Ruj, S. (2014) Attribute based access control in clouds: A survey. In *Signal Processing and Communications (SPCOM), 2014 International Conference on,* pp. 1-6. IEEE.

Ruj, S., Stojmenovic, M., & Nayak, A. (2014) Decentralized access control with anonymous authentication of data stored in clouds. *Parallel and Distributed Systems, IEEE Transactions on, 25*(2), pp. 384-394.

Thangavel, M., & Varalakshmi, P. (2014) A survey on security over data outsourcing. In *Advanced Computing (ICoAC), 2014 Sixth International Conference on*, pp. 341-349.

Toninelli, A., Montanari, R., Kagal, L., & Lassila, O. (2006) A semantic context-aware access control framework for secure collaborations in pervasive computing environments. In *The Semantic Web-ISWC 2006,* pp. 473-486.

Yang, K., Liu, Z., Cao, Z., Jia, X., Wong, D. S., & Ren, K. (2012) TAAC: Temporal attribute-based access control for multi-authority cloud storage systems. *IACR Cryptology ePrint Archive*, pp. 651.

Yang, K., Jia, X., Ren, K., Zhang, B., & Xie, R. (2013) Dac-macs: Effective data access control for multiauthority cloud storage systems. *Information Forensics and Security, IEEE Transactions on, 8*(11), pp.1790-1801.

Yassin, A., Jin, H., Ibrahim, A., Qiang, W., & Zou, D. (2012) A practical privacy-preserving password authentication scheme for cloud computing. In *Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012 IEEE 26th International,* pp. 1210-1217.

Zhou, Z., Wu, L., Hong, Z., Liang, Z., Jun, L., Sheng-Jun, X., & Xilong, Q. (2013) Context-aware access control model for cloud computing. *International Journal of Grid and Distributed Computing, 6*(6), pp. 1-12.

Zhu, Y., Hu, H., Ahn, G. J., Huang, D., & Wang, S. (2012) Towards temporal access control in cloud computing. In *INFOCOM, 2012 Proceedings IEEE,* pp. 2576-2580.

Raul Casana Eslava

PC-algorithm modifications to improve the DAG consistency

R V Casana-Eslava
Department of Applied Mathematics, LJMU. 3 Byrom Street, Liverpool L3 3AF
r.v.casanaeslava@ljmu.ac.uk

Abstract. In Bayesian Networks domain, PC-algorithm is one of the most efficient constrained-based algorithms in Structure Finding problems. Last versions of PC-algorithm [1] provide a consistent skeleton independently of the nodes order. However when Directed Acyclic Graphs (DAG) are obtained from the skeleton (undirected graph), appear multiple solutions that strongly depend on the node order of the data set. This work proposes several algorithm modifications based on node pairwise mutual information to obtain a unique DAG independently of the nodes order. The DAG obtained presents a very good performance (log-likelihood scores) compared with the performance distributions generated by random nodes ordering and for any data sample of the same size.

Keywords. PC-algorithm, DAG, mutual information, skeleton, node order, Bayesian Network

1 Introduction

Probabilistic Graphical Models (PGM) have increasingly been studied, especially from the theoretical point of view. They offer an efficient graphical approach to apply statistical inference in complex systems. They serve as a framework for Bayesian and Markov networks [2], and have two components: a structure or skeleton in the form of a graph, and a set of parameters that can be used to make inferences.

Recently new structure finding algorithms based on PC-algorithm have been developed [1], which can obtain a faithful Bayesian Network (BN) without the need for strong approximations and which execute within a reasonable time. Using these methods complex associative maps can be obtained through Conditional Independence Maps (CI-maps).

One case of interest are directed acyclic graphs (DAGs), that contain directed rather than undirected edges, which restrict in a sense the conditional dependence relations. These graphs can be interpreted by applying the directed Markov property [3]. When ignoring the directions of a DAG, we get the skeleton of a DAG.

Roughly speaking, the PC-algorithm starts from an undirected completed graph and deletes recursively edges based on conditional independence decisions, measuring nodes mutual information (1). This yields an undirected graph which can then be partially directed and further extended to represent the underlying DAG. However, given a partially directed graph (skeleton and v-structures identified) there is a group of DAGs that represent equally the conditional probability distribution of the data. Therefore, there is no a single DAG solution, next step is how to estimate the best DAG from our data.

$$I(i,j|A) = \sum_{x_i, x_j, x_A} p_{ijA}(x_i, x_j, x_A) log_2 \left(\frac{p_{ij|A}(x_i, x_j | x_A)}{p_{i|A}(x_i | x_A) \cdot p_{j|A}(x_j | x_A)} \right) \qquad (1)$$

One of the first problems found is the outcome of the vanilla PC algorithm is influenced by the order in which the conditional independence tests are executed. It means that obtained DAGs strongly depend on the node ordering. In other words, given the same data sample, if the nodes (usually represented as columns) are permuted, the DAG is significantly different, the v-structures (colliders) are in different positions and even sometimes the algorithm yields a different skeleton.

Using certain policies described in [1], PC-algorithm can provide a stable and consistent undirected graph (skeleton) as output, independently of the nodes ordering in the input data set.

This work is about the policies and modifications have to be applied in PC-algorithm and posterior orient graphs algorithms [4] to generate a unique DAG from the partially oriented graph obtained previously, selecting a good DAG candidate between all the compatible DAGs, where the log-likelihood and regularized scores (BIC or AIC) have a good values compared with the distribution of other DAG ones.

2 Methodology

The starting point is the PC-algorithm policy mentioned in [1] to achieve a unique skeleton. This policy is applied after the algorithm first pass, when the pairwise unconditional independence tests are computed. It consists in measuring the strength of each node, defined as the sum of the marginal mutual information (2) between each node and his adjacent nodes (neighbours). The nodes and edges are sorted in ascending strength; this is a non-arbitrary ordering named The Weakest First (TWF) policy. Thus, in the second pass of the algorithm, conditional independence tests are applied firstly to weakest nodes, helping to avoid the error of pruning a strong edge too soon, producing subsequent chain errors.

$$\sigma_i = \sum_{j \in ne(i)} I(i,j) \qquad (2)$$

To construct a DAG from the skeleton there is an algorithm based on four rules mentioned in [5] and [4] that orients the skeleton. Here is where the skeleton diverges in wide range of DAGs if the node order is permuted. Bear in mind that in a network of n nodes exists n! Possible permutations, so it is hard computationally check all the possibilities.

To tackle this problem, in a similar way than before, one may apply TWF policy to the adjacency matrix to reorder the nodes (columns and rows) inside the orient graph algorithm. But this is not enough because the node links and the d-separation list from each node depend on the order of nodes that was when the PC-algorithm was executed. If the PC-algorithm is executed with different nodes order, the node links and d-separation list differ completely from previous runs.

To fix this problem is necessary to reorder the nodes before the skeleton is created; it is before the first pass of unconditional independence test, inside the PC-algorithm. In this way, the node links and d-separation list is consistent in later stages when DAG is created.

There are three possible node sorting, the weakest first (TWF), the strongest first (TSF), and the default ordering. Empirical results, shown in next section, point to the best policy to order the nodes is TSF, before the step where skeleton is created. This make sense, if one thinks the network should start to link the nodes with higher mutual information first. One may imagine this like a tree growing from the strongest branches. Later the pruning will be in opposite direction, cutting the weakest branches first.

In summary, during the first steps of the PC-algorithm the nodes have to been sorted by TSF, then before the second pass, the node edges mainly have to been sorted by TWF, and finally in the orient graph function, the nodes have to been sorted again by TSF.

3 Results
This section will show the DAG performance of applying the TSF policy compared with the DAGs obtained by randomly permuting the nodes order. In addition, the best and worst DAGs just obtained sorting the nodes by chance also will be included in the comparison.
The data set used is the INSURANCE benchmark, obtained from the Bayesian Repository. This a Bayesian network of 27 nodes with 52 edge degree. Initially the sample size has been

fixed to 500 observations, and later in future works, the sample size will be extended to higher ranges.

Regarding the PC-algorithm setup is as follows (a): The test level has been set to α=0.05, False Discovery Rate with basic policy (FDR control is applied a-posteriori), TWF policy is enabled for conditional independence tests and TSF policy for unconditional independence tests, orient graph algorithm also with TSF policy. To assess the DAG performance has been used the BIC score, it is the log-likelihood regularized with the Bayesian Information Criterion (BIC), that penalizes the DAG complexity.

$$score_{BIC} = LL + BIC = M \sum_{i=1}^{n} I_{\hat{p}}(X_i; Pa_{X_i}) - M \sum_{i=1}^{n} H_{\hat{p}}(X_i) - \frac{\log(M)}{2} Dim[G] \quad (3)$$

The first picture depicts the BIC score over the same sample but the node order randomly permuted 500 times. Vertical lines highlight the values for special node order. In red is shown the node order which produced extreme score values in log-likelihood just by chance. Similarly, in magenta is shown the node order which produced extreme score values in BIC just by chance. And the blue line shows the score obtained applying TSF policy, this score is unique and does not depend on node ordering. One may observe it is significantly better than the others.
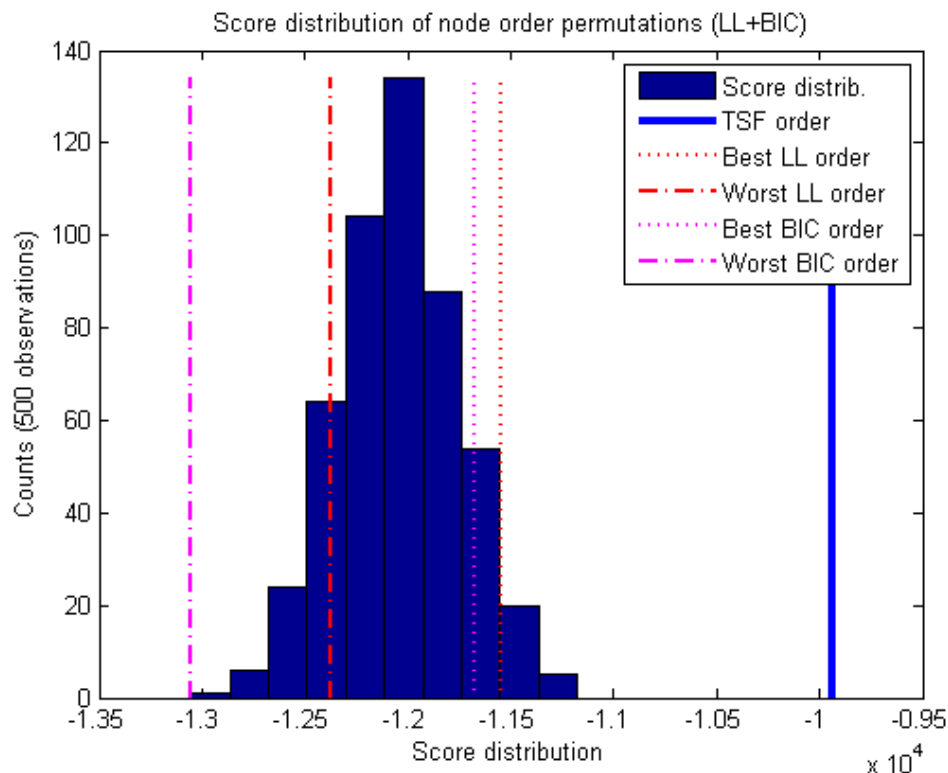
Fig. 1 Score distribution permuting the node order 500 times over the same sample size. Vertical lines highlight extreme values in log-likelihood and BIC. Blue line shows the score with the TSF policy.

The order node with the TSF policy, and subsequently the DAG score, could be obtained also just by chance, but the number of possible combinations taking into account the nodes quantity (27 nodes), and maximizing LL and BIC simultaneously, make it highly improbable. Next picture tries to show that a node order with a good DAG score in one sample, obtained by random permutation, it does not imply that in another sample it will produce also a good score. However using TSF policy, the DAG scores obtained are always better than the average scores using a random node ordering.
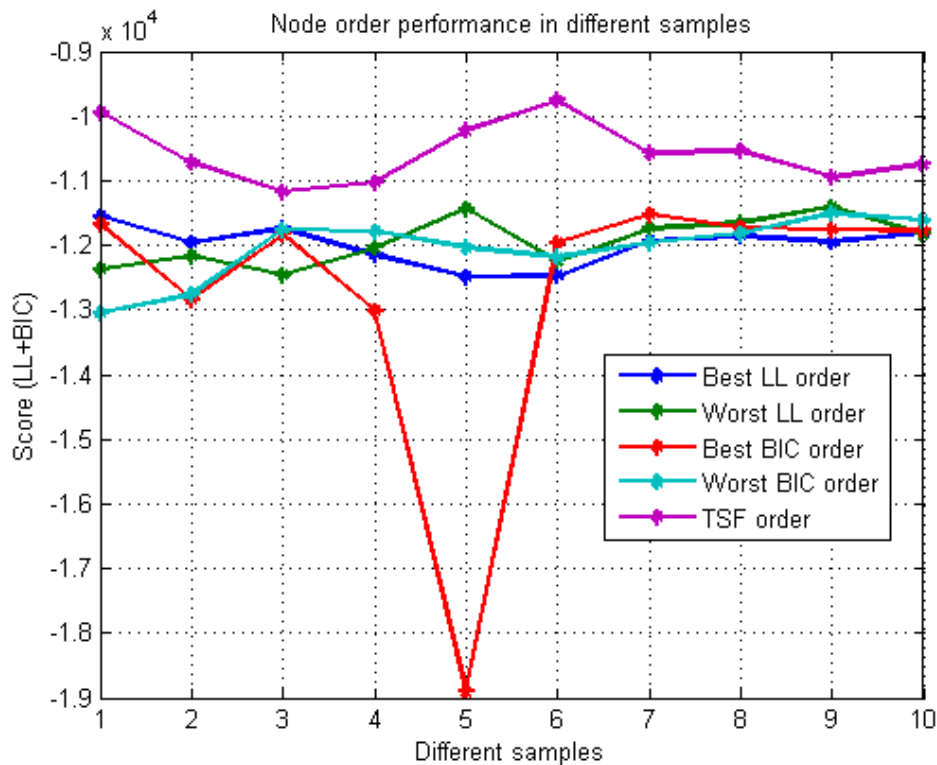
Fig. 2 BIC scores of different samples using node orders which yielded extreme score values in a previous sample. In purple highlights the score with TSF policy, that order the nodes of each sample according the pairwise mutual information.

4 Conclusion

The TSF policy lets apply the PC-algorithm to generate a DAG from the skeleton having guarantees than the BIC score will be better than the average score. This result is very useful because a unique DAG can be generated independently the node order in the input data set, it can serve as a systematic tool to generate DAGs with stability and consistency, overall in unknown data sets.

As future work is pending to implement a stable and improved policy of False Negative Reduction (FNR) to decrease the main skeleton errors (edges erroneously cut). In addition, it is necessary a detailed assessment of this policies when the data size is increased to validate and limit errors against unknown data.

5 References

[1]   D. Bacciu, T. A. Etchells, P. J. G. Lisboa and J. Whittaker, "Efficient identification of independence networks using mutual information," *Computational Statistics,* vol. 28, no. 2, pp. 621-646, 2013.

[2]   D. Bacciu, P. J. G. Lisboa, A. Sperduti and T. Villmann, "Probabilistic Modeling in Machine Learning," in *Springer Handbook of Computational Intelligence*, Springer Berlin Heidelberg, 2015, pp. 545-575.

[3]   S. Lauritzen, Graphical Models, Oxford University Press, 1996.

[4]   J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, 2000.

[5]   C. Meek, "Causal inference and causal explanation with background knowledge," 1995.

[6]   R. Daly, S. Qiang and S. Aitken, "Learning Bayesian networks: approaches and issues," *The Knowledge Engineering Review,* vol. 26:2, pp. 99-157, 2011.

[7]   K. Chalak and H. White, "Causality, Conditional Independence, and Graphical Separation in Settable Systems," *Neural Computation,* vol. 24, no. 7, pp. 1611-1668, 2012.

[8]   D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and techniques, Massachusetts: The MIT Press, 2009.

[9]   M. Jongh, "Algorithms for constraint-based learning of bayesian network structures with large numbers of variables," University of Pittsburgh, Doctoral Dissertation, Pittsburgh, 2014.

[10] N. Friedman, K. Murphy and S. Russell, "Learning the structure of dynamic probabilistic networks," *Fourteenth Conf. on Uncertainty in Artificial Intelligence (UAI),* pp. 139-147, 1998.

[11] T. Duy Le, T. Hoang, J. Li, L. Liu and H. Liu, "A fast PC algorithm for high dimensional causal discovery with multi-core PCs," *arXiv preprint,* p. arXiv:1502.02454, 2015.

[12] J. Whittaker, Graphical Models in Applied Multivariate Statistics, Wiley, 2009.

[13] J. Pearl, Causality: models, reasoning, and inference, 2nd ed., New York: Cambridge University Press, 2009.

[14] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys,* vol. 3, pp. 96-146, 2009.

[15] J. Cheng, R. Greiner, J. Kelly, D. Bell and W. Liu, "Learning Bayesian networks from data: an information-theory based approach," *Artif Intell,* Vols. 1-2, no. 137, pp. 43-90, 2002.

[16] I. Tsamardinos, L. Brown and C. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Mach Learn,* vol. 1, no. 65, pp. 31-78, 2006.

[17] I. Tsamardinos and L. Brown, "Bounding the false discovery rate in local Bayesian network learning," in *AAAI*, Chicago, 2008.

[18] A. Fast, M. Hay and D. Jensen, "Improving accuracy of constraint-based structure learning," Technical report, University of Massachusetts Amherst, Computer Science Department, 2008.

[19] B. Goebel, Z. Dawy, J. Hagenauer and J. Mueller, "An approximation to the distribution of finite sample size mutual information estimates," *IEEE InternationalConference on Communications,* vol. 2, p. 1102–1106, 2005.

[20] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J R Stat Soc Ser B Methodol,* vol. 57, no. 1, p. 289–300, 1995.

[21] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Ann Stat,* vol. 29, no. 4, p. 1165–1188, 2001.

[22] M. J. Ha, W. Sun and J. Xie, "PenPC: A two-step approach to estimate the skeletons of high-dimensional directed acyclic graphs," *Biometrics,* 2015.

[23] M. Kalisch and P. Bühlmann, "Estimating high-dimensional directed acyclic graphs with the PC-algorithm," *The Journal of Machine Learning Research,* vol. 8, pp. 613-636, 2007.

[24] M. Kalisch and P. Bühlmann, "Robustification of the PC-algorithm for Directed Acyclic Graphs," *Journal of Computational and Graphical Statistics,* vol. 17, no. 4, pp. 773-789, 2008.

Darren Conway

Systems biology modelling of the tumour microenvironment

D Conway [1], G Dockray[2], Chris Sanderson[2], A Varro[2] and S Webb[1]

[1] Department of Applied Mathematics, Liverpool John Moore's University, James Parsons Building Byrom Street, Liverpool, L3 3AF, UK

[2] Department of Cellular and Molecular Physiology, University of Liverpool, Crown Street, Liverpool L69 3BX, UK

Abstract

Myofibroblasts are stromal cells that are rare in many normal tissues, but increase dramatically in injury, chronic inflammation and cancer. Also, the gastrointestinal tract is unusual in that myofibroblasts are normally well represented; in the stomach. Using a variety of molecular and cell physiological methods, it has been shown that gut-derived myofibroblasts secrete growth factors which regulate epithelial cell proliferation and migration; in turn, gastric epithelial cells secrete factors that control myofibroblast signalling. The central goal of the proposed research is to develop new integrated *in vitro/in silico* models that will provide a better understanding of how different types of myofibroblasts interact with epithelial cells. Myofibroblasts and gastro adenocarcinoma cancer cell lines were cultured and placed in an Ibidi chamber, to form a monolayer, before being placed in a time-lapse microscope. The Ibidi chamber was removed and images were taken every 15-30 minutes of the cells interactions. Data obtained from time-lapse microscopy show a strong migration and epithelial-mesenchymal transition of AGS cells, and limited but more direct migration of the myofibroblasts. The data indicates an attractive force between the cell lines. The exact mechanism is unclear at this time but the data obtained can help direct *in silico* model development which will in turn aid in the understanding behind it.

Keywords. Myofibroblasts, cell migration, *in silico,* computational, models

1. Introduction

In the gastrointestinal tract, interactions between epithelial cells and stromal cells control cell proliferation, differentiation and apoptosis. The coordination of these cellular interactions is vital in maintaining the organisation and function of tissues. Disruption of these interactions, through inflammation, and subsequent disease progression has driven a growing appreciation of the role of cellular microenvironments, particularly in carcinogenesis. Within the gastro-intestinal tract, the myofibroblast has been highlighted as a key cell involved in this breakdown of homeostasis. Gastro-epithelial cells secrete matrix metalloproteinase-7 which cleaves insulin-like growth factor binding protein-5 (IGFBP-5) secreted by myofibroblasts leading to release of IGF-II which stimulates both epithelial and myofibroblast proliferation (1).

Myofibroblasts are activated fibroblasts; they are dynamic, spindle-like cells sharing the functional characteristics of both fibrocytes and smooth muscle cells (2). Myofibroblasts are not widely distributed in most tissues, however, they are relatively abundant in the gastrointestinal tract where they are localised in a sub-epithelial compartment. The origins of these cells have a variety of different sources, including (a) rapid stimulation of local fibroblasts to differentiate into myofibroblasts (3), (b) transdifferentiation of epithelial cells into myofibroblasts via epithelial mesenchymal transition (EMT) (4,5). In normal tissue environment, these cells play an important role in the maintenance of tissue architecture and wound healing (6,7). Myofibroblasts are a rich source of growth factors, cytokines, chemotaxis, extracellular matrix (ECM) proteins, proteases and their inhibitors (8).

Myofibroblasts that reside in the cancer niche exhibit changes to their phenotype; they are designated "cancer associated myofibroblasts" (CAMs) (9). These cells undergo changes at the epigenetic level and a difference in gene expression (10) has been detected in CAMs from gastric cancer compared with "normal tissue myofibroblasts" (NTMs) (11). This differential expression of genes relates to various cellular functions such as cancer cell proliferation, migration and invasion (12-14). CAMs are phenotypically more migratory and proliferative compared to myofibroblasts from non-cancerous stroma (15,16). The functional role of myofibroblasts in cancer cells is not well understood, however, they influence proliferation and metastasis of cancer through disruption of autocrine and paracrine signalling pathways (8,17). A decreased sensitivity to chemotherapy has been demonstrated in CAMs, partly due to hypermethylation of CpG DNA and reduced expression of caspase and STAT-1 resulting in escape from apoptosis of cancer cells (18). *In* silico models allow us to study dynamics of biological cells which cannot be easily defined using experimental approaches. These computational models of cell migration provide important insights into the dynamics of their motility. They are able to show how important cell-like behaviours such as cell shapes changes, cell adhesion, chemotaxis and proliferation can produce biological patterns and shapes (19).

By using computational approaches to reproduce properties that the biological system displays, a better understanding of how the mechanism of interaction between the myofibroblast cells and the cancer cells can be provided.

2. Materials

Cell culture materials namely 0.25 % w/v trypsin-EDTA, Dulbecco's modified Eagle's medium (DMEM), L-glutamine, non-essential amino acids, antibiotic-antimycotic solution, penicillin-streptomycin solution and phosphate buffered saline (PBS) were obtained from Sigma (Dorset, UK); Recovery^TM cell freezing medium came from Invitrogen (Paisley, UK). Foetal bovine serum (FBS) was purchased from Lonza. Ibidi cell culture inserts for migration assays were purchased from SLS (Nottingham, UK). Gastro adenocarcinoma cancer cell line (AGS) were derived from fragments of a tumour resected from a patient who had received no prior therapy. AGS was obtained from American type culture collection, ATCC, VA, US. Human primary myofibroblasts were obtained from resected oesophageal cancers and adjacent non-cancerous tissue during surgery for removal of tumours at First Department of Surgery, University of Szeged, Szeged, Hungary. Myofibroblasts were prepared in the Department of Medicine, University of Szeged, Hungary as previously described (1) and transported to University of Liverpool in liquid nitrogen and cryopreserved until use. Myofibroblasts were derived from three Barrett's adenocarcinoma, recovered from the oesophagus and oesophagus/cardiac junction. From each patient, CAMs and ATMs were cultured. Also, 6 myofibroblast lines were obtained from the oesophagus of normal

patients who were transplant donors (NTMs: 3 males, 3 females, mean average age 48.7 ± 5.8 yrs and 49.3 ± 2.7 yrs respectively).

2.1 Tissue culture

2.1.1 Human myofibroblasts. Myofibroblasts (This is referred to as "308/1" unless otherwise stated.) were maintained in T-75 flasks in DMEM supplemented with 10% v/v FBS, 1% v/v penicillin-streptomycin, 2% v/v antibiotic-antimycotic and 1% v/v non-essential amino acids. This is referred to as "full medium" (FM) unless otherwise stated. Cells were grown at 37°C in a 5% v/v CO2 atmosphere and the medium was changed every 48 h. For passaging, cells at 80% confluence or above were washed twice with PBS followed by incubation for 5-8 min in 2ml 0.25 % w/v trypsin-EDTA. Trypsinised single cell suspensions were then added to 8ml of FM and cells were re-plated in T-75 flasks. Aliquots of 10µl of cell suspensions were used for cell counting using a haemocytometer.

2.1.2 Human gastro adenocarcinoma cancer cell line. Gastro adenocarcinoma cancer (This is referred to as "AGS" unless otherwise stated) cell lines were cultured in Ham's F-12 Nutrient mixture supplemented with 10% v/v FBS, 1% v/v penicillin-streptomycin, 2% v/v L-glutamine. Cells were maintained at 37oC in a 5 % v/v CO2 atmosphere, and media was changed every 48-72 h. Confluent cells at 80% were washed twice with PBS, trypsinised using 0.25% w/v trypsin-EDTA and cells were added to 8ml FM and re-plated in T-75 flasks.

2.1.3 Cryopreservation of cell lines. Cells were trypsinised at 80% confluence, centrifuged at 800 x g for 7 min at 4oC and supernatants discarded. Pellets were resuspended in 1 ml recovery cell freezing medium and cell suspensions stored in 1.5 ml cryovials. These vials were placed in a plastic holder in a bath containing propane-1, 2,-diol and then transferred to a -80°C freezer overnight and then to liquid nitrogen for long-term storage.

2.1.4 Recovering frozen cell lines. Cells were removed from liquid nitrogen, thawed by hand or in a water bath at 37oC and added to a T-75 culture flask with 19 ml FM and maintained 37oC in a 5 % v/v CO2 atmosphere.

2.2 Migration assay

Cells were grown to confluence in six-well dishes. In co-culture experiments, 100,000 AGS and 50,000 308/1 cells were cultured in each well. Typically, the two cell types were segregated into different parts of the well by applying them to either side of an Ibidi chamber that was subsequently removed. Before removing Ibidi chamber, cells were kept in an incubator for 24 h (in FM). Ibidi chambers were then removed and each well undergoes a double PBS wash before 800µl of 50/50 mixed DMEM and F12-Hams medium with 2% FCS was added to each well. The number of cells crossing a margin of 500 µm into an acellular area was counted. Time-lapse image series were analysed using Scion Image software (Scion, Frederick, MD), which is based on National Institutes of Health Image. Cell tracking and frame-by-frame recording of movements were used to calculate cell speed over 30-min periods. Manual cell tracking was performed with Image J using the *Manual Tracking* plugin. Statistical analysis was performed with excel and MATLAB.

3. Results

A video was made from time-lapse microscopy by taking an image every 30 minutes for 18 hours and stitching them together. These images were then analysed, using Image J software, to produce a graph showing the area of migration covered by the cell types (AGS and 308/1) over time (Figure 1). At t=0, a line was plotted across the leading edge of the

cells, using Image J, and the area was set to zero. At every 3 hour interval until t=18 the area of total cells that crossed the leading edge line was calculated and recorded for each individual experiment (n=3). Collectively, this data was used to produce a graph showing the average cell migration in the two cell types for the experiments (Figure 1). When the Ibidi chamber was removed, a 500 micron acellular gap was present between the two cell colonies. The cells located on the proximal side were labelled as 'AGS' and '308/1' respectively. The cells located on the distal side were used as a control and were labelled appropriately. The red dotted line represents the point of initial contact between the proximal AGS cells and 308/1 cells.

3.1 Increased cell migration in epithelial cancer cells (AGS)

AGS cells exhibited loss of epithelial phenotype from t=0. The cells underwent EMT and began migration into the 500 micron space. The proximity of myofibroblasts had a pronounced effect on the speed of migration of the AGS cells. As the divide between cell types narrows, the AGS cells progression decreases (t=4). Initially, the 308/1 cells maintain cellular phenotype and exhibit low amounts of cell migration. However, as the time course extends and the proximity of the AGS cells increase, there is an increase in the migration of the 308/1 cells.
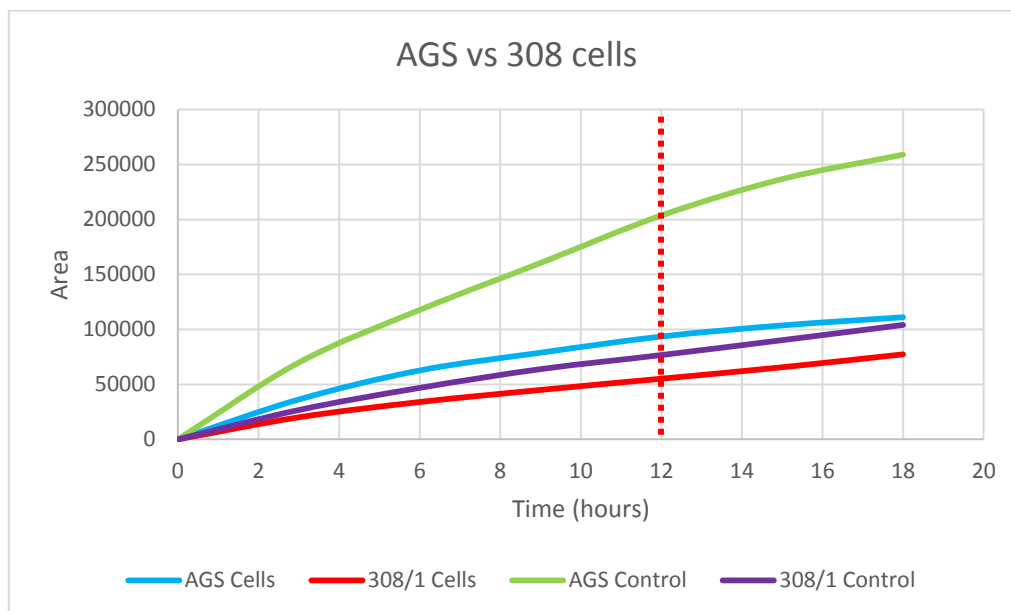
Figure 1: AGS cells exhibit increased migration. Graph showing migration of AGS and 308/1 cells over time. The proximal AGS and 308/1 cells are represented by the blue and red lines respectively. The distal AGS and 308/1 control cells are represented by the green and purple lines respectively. The red dotted line indicates the time point at which the AGS cells and 308/1 cells interact.

At the moment of initial contact between the cell types (red dotted line), there is a decrease in cell migration of AGS cells and inversely, an increase in the cell migration of the 308/1 cells observed in the time-lapse videos. Significantly, a huge increase can be seen in cell migration between the proximal AGS cells and the distal AGS control cells. At t=18, the AGS control cells have migrated into an area 2.5 times larger than that of the proximal AGS cells. Comparatively, there is also a larger increase in the area covered by the 308/1 control cells than that of the proximal 308/1 cells, although on a much smaller scale. Unexpectedly, the AGS control cells exhibit the biggest increase in cell migration.

3.2 Cell migration in cancer cells is directed

Using manual tracking software on Image J individual AGS cells were tracked from t=0 to t=18, using images taken every 15 minutes from the time-lapse microscope (Figure 2). The data obtained from these tracks show a direct and deliberate migration pattern away from the colony of AGS cells towards the 308/1 cells. As stated above, the AGS cells migratory behaviour changes with increased proximity to the 308/1 cells and this behaviour is observed once again with the tracking data, as the cells change direction (Figure 2A). The cells exhibit Brownian motion once that point has been reached. Moreover, the AGS control cells exhibit a similar pattern of direct migration followed by Brownian motion (Figure 2B).
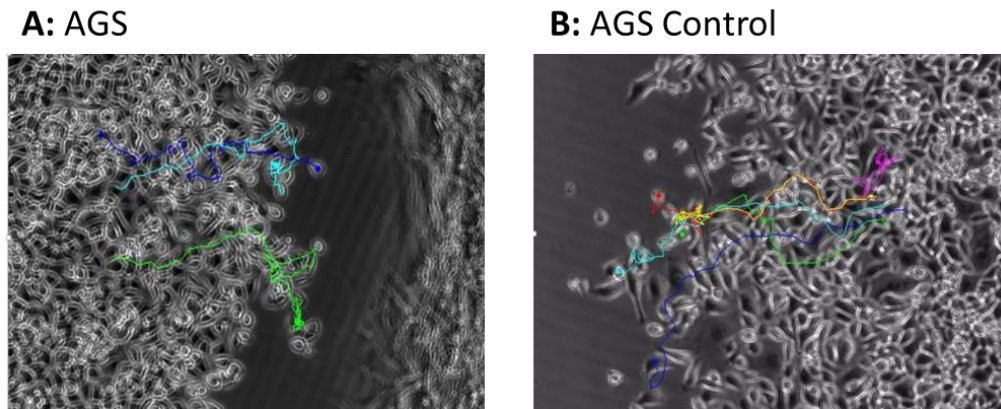
Figure 2: Cell migration tracking data. (A) This image shows the migration track patterns of individual AGS cells on the proximal side. (B) This image shows the migration track pattern of individual AGS control cells on the distal side.

3.3 Quantitative analysis of the cell migration

Fick's first law relates the diffusive flux to the concentration under the assumption of steady state. It postulates that the flux goes from regions of high concentration to regions of low concentration, with a magnitude that is proportional to the concentration gradient (spatial derivative), or in simplistic terms the concept that a solute will move from a region of high concentration to a region of low concentration across a concentration gradient. From the raw data, cell density was calculated at each time interval and plotted against space (Figure 3). The Xmean was calculated and relates to the space that 50% of the cell density has migrated into. Comparatively, the AGS control cells (Figure 3A) migrated into 2.5 times more space than the 308/1 control cells (Figure 3B) over the same time period (~225 vs ~90). The AGS control cells migrate into the space approximately seven times faster than the 308/1 control cells when comparing the diffusion coefficients (1143.25 vs 174.79 μm^2 hr^{-1}).
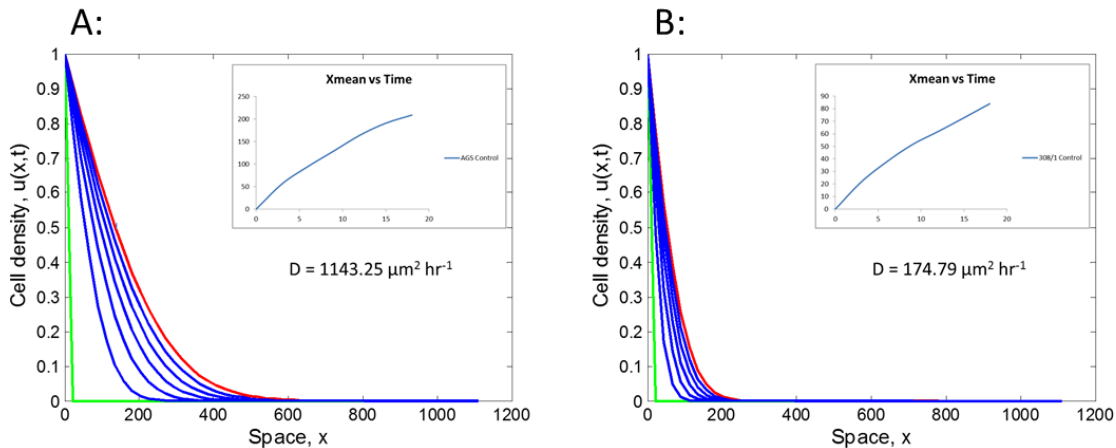
Figure 3: Graph showing the cell density of the migrating cell populations over space.
(A) shows AGS Control cells. (B) shows 308/1 Control cells. Subplot graphs shows Xmean vs Time for each cell type. The diffusion coefficient is stated for each cell type also.

4. Discussion

There is increasing evidence that CAMs contribute to tumour growth (20). The mechanisms of homing and migration remain incompletely understood. In this study, we provide evidence of both positive and negative interactions between the CAMs and the cancer cell line. Previous studies have reported that CAMs evoke a more aggressive phenotype in cancer cells compared with myofibroblasts derived from normal tissue environments (21). This can be observed from our data where, initially, the AGS cells migrate aggressively towards the 308/1 cells. The present data suggest that a chemoattractant is released from the myofibroblasts to promote the recruitment of the AGS cells, Nevertheless, there is also inhibition present between the two cell types as the experiments display a decrease in migration as the AGS cells approach the 308/1 cells. There are several theories as to why this happens: (a) Contact inhibition from the decrease in acellular space for the AGS cells to migrate into (22) (b) the release of a compound from the CAMs that directly inhibits the migration of the AGS cells or (c) the AGS cells slow down due to the energy spend required to remodel the extracellular network that has been laid down by the myofibroblasts (23).

The manual tracking data backs up this attraction and inhibition model as it clearly shows the direct migration of the AGS cells followed by a delayed inhibition. In Figure 2A, the tracked cells migrate in a direct, unidirectional manner and as the acellular space between the AGS and 308/1 cells decreases, so too does the direct migration pattern of the tracked cells. In contrast, the cessation of direct migration by the tracked cells in Figure 2B is less obvious as there is no decrease in acellular space.

Fick's first law describes that the molar flux due to diffusion is proportional to the concentration gradient (24). Qualitatively, the time lapse microscopy data can be used to observe and predict the cellular migration patterns of the cell types. However, quantitatively these experiments offer very little detail. This is where *in silico* modelling approaches prove useful. Here, it is simply shown how quantitative analysis, the diffusion coefficient, can be calculated from the creation of an *in silico* model of the system. The diffusion coefficient represents the driving force which arises from the differential surface energies in the system i.e. the difference in surface energy between homotypic cells, heterotypic cells and between the cells and the external media.

The next step is to further this progression through the introduction of cellular potts modelling (CPM). The CPM is a spatial grid based model where the cell is defined over a region of multiple lattice sites, with a set number of interactions and constraints acting on its boundary and area. Effectively, this means that cell migration is modelled from the overall minimisation of the energy of deformation and stretching of the membrane through stochastic fluctuations, in which the global and local forces upon a cell edge are resolved (25-27). One of the main advantages to using this modelling technique is that it allows for the input of a lot of data while remaining computationally efficient. By tailoring the CPM to exhibit the dynamics and behaviour observed in the experiments described above, a model can be developed and used to test and predict how the two cell types interact with one another *in vivo.*

References

(1) McCaig C, Duval C, Hemers E, Steele I, Pritchard DM, Przemeck S, et al. The role of matrix metalloproteinase-7 in redefining the gastric microenvironment in response to Helicobacter pylori. Gastroenterology 2006 May;130(6):1754-1763.

(2) Hinz B, Phan SH, Thannickal VJ, Galli A, Bochaton-Piallat ML, Gabbiani G. The myofibroblast: one function, multiple origins. Am J Pathol 2007 Jun;170(6):1807-1816.

(3) Serini G, Bochaton-Piallat ML, Ropraz P, Geinoz A, Borsi L, Zardi L, et al. The fibronectin domain ED-A is crucial for myofibroblastic phenotype induction by transforming growth factor-beta1. J Cell Biol 1998 Aug 10;142(3):873-881.

(4) Spaeth EL, Dembinski JL, Sasser AK, Watson K, Klopp A, Hall B, et al. Mesenchymal stem cell transition to tumor-associated fibroblasts contributes to fibrovascular network expansion and tumor progression. PLoS One 2009;4(4):e4992.

(5) Quante M, Tu SP, Tomita H, Gonda T, Wang SS, Takashi S, et al. Bone marrow-derived myofibroblasts contribute to the mesenchymal stem cell niche and promote tumor growth. Cancer Cell 2011 Feb 15;19(2):257-272.

(6) Radisky DC, Kenny PA, Bissell MJ. Fibrosis and cancer: do myofibroblasts come also from epithelial cells via EMT? J Cell Biochem 2007 Jul 1;101(4):830-839.

(7) Klingberg F, Hinz B, White ES. The myofibroblast matrix: implications for tissue repair and fibrosis. J Pathol 2013 Jan;229(2):298-309.

(8) Powell DW, Mifflin RC, Valentich JD, Crowe SE, Saada JI, West AB. Myofibroblasts. I. Paracrine cells important in health and disease. Am J Physiol 1999 Jul;277(1 Pt 1):C1-9.

(9) De Wever O, Mareel M. Role of tissue stroma in cancer cell invasion. J Pathol 2003 Jul;200(4):429-447.

(10) Boussioutas A, Li H, Liu J, Waring P, Lade S, Holloway AJ, et al. Distinctive patterns of gene expression in premalignant gastric mucosa and gastric cancer. Cancer Res 2003 May 15;63(10):2569-2577.

(11) Jiang L, Gonda TA, Gamble MV, Salas M, Seshan V, Tu S, et al. Global hypomethylation of genomic DNA in cancer-associated myofibroblasts. Cancer Res 2008 Dec 1;68(23):9900-9908.

(12) Bhowmick NA, Neilson EG, Moses HL. Stromal fibroblasts in cancer initiation and progression. Nature 2004 Nov 18;432(7015):332-337.

(13) Singer CF, Gschwantler-Kaulich D, Fink-Retter A, Haas C, Hudelist G, Czerwenka K, et al. Differential gene expression profile in breast cancer-derived stromal fibroblasts. Breast Cancer Res Treat 2008 Jul;110(2):273-281.

(14) Grugan KD, Miller CG, Yao Y, Michaylira CZ, Ohashi S, Klein-Szanto AJ, et al. Fibroblast-secreted hepatocyte growth factor plays a functional role in esophageal squamous cell carcinoma invasion. Proc Natl Acad Sci U S A 2010 Jun 15;107(24):11026-11031.

(15) Schor SL, Schor AM, Rushton G. Fibroblasts from cancer patients display a mixture of both foetal and adult-like phenotypic characteristics. J Cell Sci 1988 Jul;90 (Pt 3)(Pt 3):401-407.

(16) Holmberg C, Quante M, Steele I, Kumar JD, Balabanova S, Duval C, et al. Release of TGFbetaig-h3 by gastric myofibroblasts slows tumor growth and is decreased with cancer progression. Carcinogenesis 2012 Aug;33(8):1553-1562.

(17) Direkze NC, Hodivala-Dilke K, Jeffery R, Hunt T, Poulsom R, Oukrif D, et al. Bone marrow contribution to tumor-associated myofibroblasts and fibroblasts. Cancer Res 2004 Dec 1;64(23):8492-8495.

(18) Muerkoster SS, Werbing V, Koch D, Sipos B, Ammerpohl O, Kalthoff H, et al. Role of myofibroblasts in innate chemoresistance of pancreatic carcinoma--epigenetic downregulation of caspases. Int J Cancer 2008 Oct 15;123(8):1751-1760.

(19) Graner F, Glazier JA. Simulation of biological cell sorting using a two-dimensional extended Potts model. Phys Rev Lett 1992 Sep 28;69(13):2013-2016.

(20) De Wever O, Demetter P, Mareel M, Bracke M. Stromal myofibroblasts are drivers of invasive cancer growth. Int J Cancer 2008 Nov 15;123(10):2229-2238.

(21) Kalluri R, Zeisberg M. Fibroblasts in cancer. Nat Rev Cancer 2006 May;6(5):392-401.

(22) Seluanov A, Hine C, Azpurua J, Feigenson M, Bozzella M, Mao Z, et al. Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat. Proc Natl Acad Sci U S A 2009 Nov 17;106(46):19352-19357.

(23) Otranto M, Sarrazy V, Bonte F, Hinz B, Gabbiani G, Desmouliere A. The role of the myofibroblast in tumor stroma remodeling. Cell Adh Migr 2012 May-Jun;6(3):203-219.

(24) Lecca P, Morpurgo D. Modelling non-homogeneous stochastic reaction-diffusion systems: the case study of gemcitabine-treated non-small cell lung cancer growth. BMC Bioinformatics 2012;13 Suppl 14:S14-2105-13-S14-S14. Epub 2012 Sep 7.

(25) Merks RM, Brodsky SV, Goligorsky MS, Newman SA, Glazier JA. Cell elongation is key to in silico replication of in vitro vasculogenesis and subsequent remodeling. Dev Biol 2006 Jan 1;289(1):44-54.

(26) Scianna M. An extended Cellular Potts Model analyzing a wound healing assay. Comput Biol Med 2015 Jul;62:33-54.

(27) Turner S, Sherratt JA, Painter KJ, Savill NJ. From a discrete to a continuous model of biological cell movement. Phys Rev E Stat Nonlin Soft Matter Phys 2004 Feb;69(2 Pt 1):021910.

Brian Granby

SDN-Based Situational Awareness

Brian Granby
Department of Computer Science
Liverpool John Moores University
Liverpool, UK
b.r.granby@2009.ljmu.ac.uk

Bob Askwith
Department of Computer Science
Liverpool John Moores University
Liverpool, UK
r.j.askwith@ljmu.ac.uk

Angelos Marnerides
Department of Computer Science
Liverpool John Moores University
Liverpool, UK
a.marnerides@ljmu.ac.uk

Abstract - The proliferation of ubiquitous computing has played part in drastically changing how individuals and organisations access, process, store and interact with sensitive digital information. These changes, from an endpoint perspective are largely due to the increased presence of internet-connected devices, such as smart devices – equipped with by a diverse variety of networking capabilities.

From a networking perspective, the services consumed by a growing pool of heterogeneous applications that provide rich, extensive functionality towards aforementioned devices. The volume of data produced and consumed by such services introduces an environment whereby existing network border controls that utilized for network security are considered to be somewhat unsuited towards efficiently defending against attacks that target modern society. Throughout this digital evolution, we have in turn also observed an increased number of cyber-attacks against the data centre infrastructures, which host backend services for supporting such applications.

Our project aims to provide a platform for the utilisation of emerging technologies, which can be exploited for gaining situational awareness for cyber security from a global point of view of interconnected computer system architectures. The technologies we utilise within our project are viz: Software-defined Networking, Network Function Virtualization, and Container-based network services. When combined, these technologies are capable of providing network security relative monitoring metadata from both the service layer, and network layer. By utilising this data with techniques such as anomaly detection, network operators are able to tightly integrate security monitoring throughout a system in an abstract way, as well as define security policies in an extensive manner by utilising JSON structured rule sets. These policies enable reactive defence against detected cyber-attacks in an autonomous manner, as well as facilitating network flow control for specific service security needs.

Introduction

Over the previous decade, several technologies have emerged as strong candidates as an attempt to address the diverse characteristics and requirements of modern digital devices. These technologies are: Software-Defined Networking (SDN) [1], Network Function Virtualization (NFV) [2], and more recently – Linux-based Containers (LXC) [3] – all of which have gained considerable attention throughout academia and industry for a variety of domain specific challenges. Each of the above provides functional advantages that facilitate the development of novel techniques required to address the challenges and limitations typically found in managing and securing complex networked environments.

These environments – typically associated with cloud computing, now play a critical role in accommodating growing demand for a broad variety of digital services.

These services have grown to become something that is relied upon globally in order to interact with mission critical- often highly sensitive information[4].

The recent uptake in IT infrastructure reduction strategies, whereby computational resources and services are outsourced into cloud environments introduces management complexities when considering the identification of security-relevant monitoring points within emerging dynamic environments. This is amplified when faced with the large volumes of data that are associated with somewhat borderless ubiquitous environments- which is something introduced by the requirements of heterogeneous accessible services [4].

In parallel, an increased number of reported cyber-attacks have also been observed in recent years. This suggests that the growth of widespread connectivity has enabled cyber-attackers to target environments that are no longer confined to traditional networked environment. Instead, miscreants are now targeting a diverse number of connected systems and devices - examples of these include cloud service provider data centres, smart devices, as well as critical infrastructure systems [5].

Due to this, the security of internet connected systems has grown to become one of the largest and most difficult challenges faced globally[4].

In order to address the above identified security challenges our project combines several benefits found within emerging technologies, and aims to provide a platform which can be exploited for gaining situational awareness for cyber security from a global point of view within dynamic interconnected architectures. The motivation behind our proposed platform aims to provide modular configuration that natively integrates into container-service specification templates, which can also be utilised to design and develop innovative security mechanisms for emerging architectures. Further, by utilising existing network-based intrusion detection techniques – e.g. signature based, and anomaly-based detection[6]

we aim to enable simple configuration of service-specific security deployment. The benefit gain in this can be realised by enabling existing network security mechanisms to become scalable, resilient modules that do not incur the same dependency limitations that are currently found in existing monolithic systems.

Background

This section provides a background on the above-identified technologies. In this, we quickly summarise their origin, potential use cases, the challenges they aim to address, and the benefits provided by each. In section A, we first discuss Network Function Virtualization (NFV). This is followed by Container-based Technologies in section B. Section C provides an overview of Software-Defined Networking (SDN).

Network Function Virtualization

Network-function virtualization (NFV) promotes the consolidation and reduction of specialised network function hardware by enabling network operators to deploy multiple services, commonly deployed as virtual instances, which are run on industry standard servers and merchant silicone [7]. The benefit in this enables network operators to reduce CAPEX and OPEX, whilst gaining a reduction in the time-to-market of new services, as well as removing previous barriers commonly found with ASICS network devices.

Since its proposal in 2013, NFV has gained considerable attention throughout industry and academia. Led by the European Telecoms Standards Institute (ETSI) [8], NFV aims to promote an open ecosystem whereby the decomposition and consolidation of traditional network equipment into virtualized functions provides new avenues of opportunity for accelerating innovation of network services.

It is predicted that the above will enable network operators to break the 'vendor lock-in' paradigm typically associated with proprietary specialized network function equipment. The benefit in this enables rapid deployment and reconfiguration of

network functions, as well as reducing the time to market of new services [9].

Container Based Virtualization

Linux-based containers (LXC) enable the rapid configuration, deployment and management of highly distributed systems that can horizontally scale across large pools of shared compute resource [3]. Containers employ System-level virtualization, and are significantly lightweight in comparison to hypervisor-based virtual instances. This is because they do not require a full operating system installation per virtual instance [10]. Instead, guest containers reside on top of the host systems kernel, and are only allocated access to libraries required to perform their intended function. Hardware resources are allocated via the *control groups (Cgroups)* Linux kernel feature, facilitating resource allocation and isolation of system memory, CPU, and block I/O devices. The *namespaces* feature provides per-process isolation of the host's operating system

The afforded benefit in this offers a reduction in service initiation time, as well as the ability to migrate services amongst hosts without incurring the overhead commonly associated with full system virtualization [11]. Furthermore, as containers run in kernel user-space, we are able to free up system resources that would otherwise be reserved for a hypervisor layer, which typically amounts to around 15% of a host machine's computational resources[12].

Recent developments within Container technologies – namely Docker has enabled the technology to become an attractive alternative for addressing the remaining challenges of NFV orchestration[13] and service composition. Existing research also suggests this alternative approach may provide a means for addressing abnormal latency that is incurred when virtualizing network functions[14]. Furthermore, as network functions are specified as template files, they demonstrate real potential towards becoming a platform agnostic solution. The benefit in this can be realized when considering multiple-cloud provider

situations. To elaborate; in adapting a common template schema, which may be considered as a standardized API –Linux based Containers may provide as a sustainable technology towards preventing the cloud provider lock in scenario.

Software-Defined Networking

In comparison to existing network management protocols and technologies, Software-defined networking (SDN) [15] provides an unrivalled level of flexibility for computer network administration. This is facilitated by the separation of forwarding logic from physical network elements, providing state abstraction in the form on a logically centralised, physically distributed control plane [16]. This separation and abstraction provides a global view of the network state, enabling the ability to control and rapidly reconfigure a network infrastructure in a centralised manner- whilst considering the global state of the whole network. Performing such an activity in existing architectures would have previously been considered impossible.

Whilst SDN is not limited towards any single control protocol- the OpenFlow protocol has already gained standardization [17], making it the most widely utilised at present. The same is yet to be seen when considering standardization of the application plane in SDN virtualized environments. Thus, if the benefits of NFV are to be become a realistic component of the emerging Software-defined Infrastructures, it is essential that a common API or platform be found. More importantly, given the potential network infrastructure's physical distribution, it is imperative that an effective solution is found that enables consistent, reliable orchestration that is capable of enabling global administration of both the physical infrastructure and virtual network device states.

Solution Overview

We are currently in the process of designing and implementing a platform for the utilisation of identified qualities found in emerging technologies. We now provide an overview of required functionalities,

which can be exploited in order to gain situational awareness for cyber security from a global point of view of interconnected computer system architectures.

When combined, these technologies are capable of providing network security relative monitoring metadata from both the service layer, and network layer. By utilising this data with techniques such as anomaly detection, network operators are able to tightly integrate security monitoring throughout a system in an abstract way, as well as define security policies in an extensive manner by utilising JSON structured rule sets. These policies enable reactive defence against detected cyber-attacks in an autonomous manner, as well

as facilitating network flow control for specific service security needs.

We have identified the following capabilities and requirements that facilitate our solution's ability in gaining a better understanding of systems and network architectural situational awareness in large-scale computer infrastructures. Figure 1 illustrates a high-level overview of our proposed platform within a generic architecture.

Network Layer

SDN provided global network view. In this, we utilise OpenFlow protocol [17] statistics for gaining situational awareness of the networks state. Flow statistics are obtained via a physically distributed, logically centralised
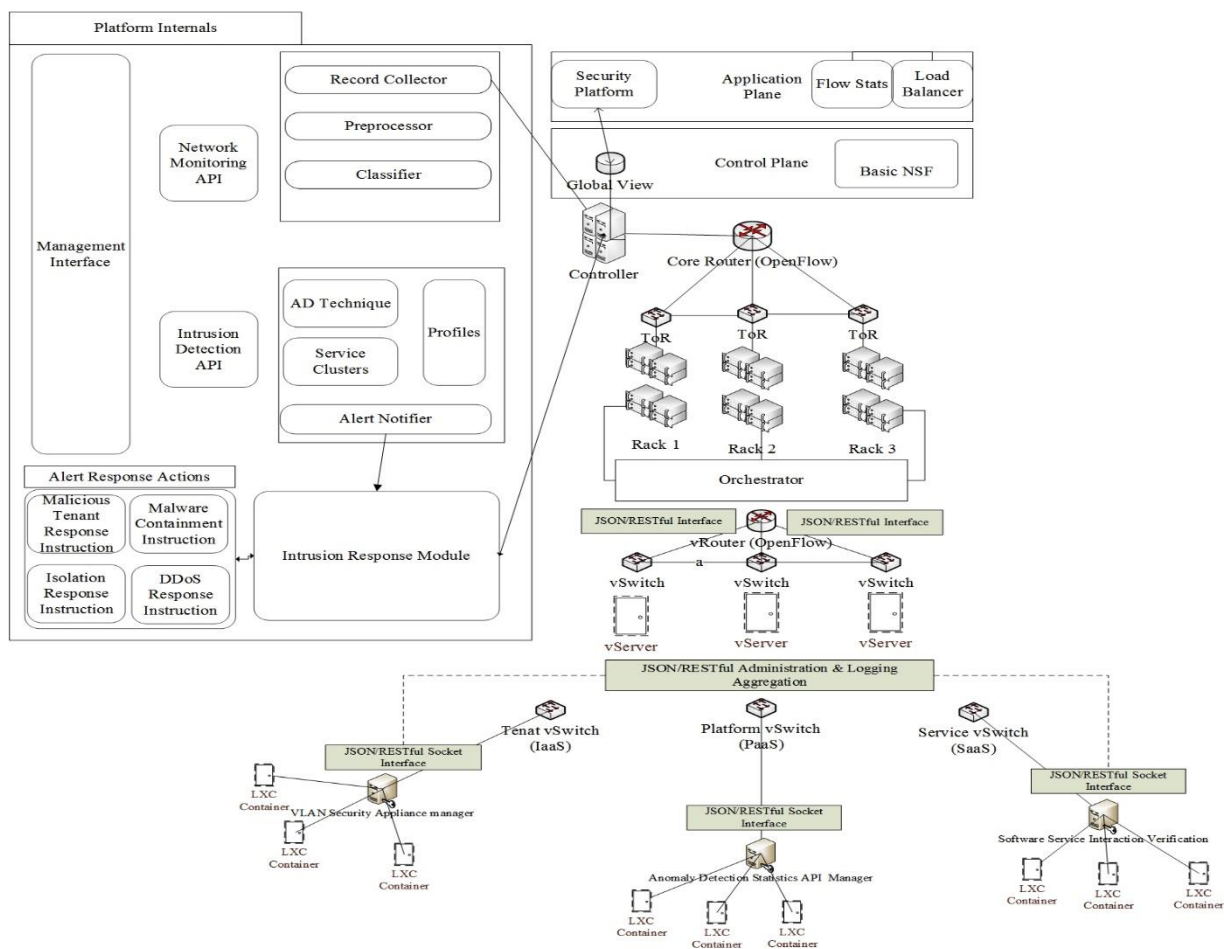
Figure 1- Design Diagram

network controller – which is periodically aggregated from switching infrastructure. These statistics can be utilised as input data for various support algorithms, enabling network operators to make informed decisions when performing network management activities [18].

Compute Technologies

Lightweight container-based technologies for the decomposition and virtualisation of network security appliances and monitoring functions. Existing research shows that containers are quicker to initiate, as well as faster to migrate in comparison to hypervisor-based instances. This suggests they are more suited towards dynamic networked environments.

Unified Schema

Unified schema for the configuration of service-specific monitoring requirements, enabling network operators to define service relevant monitoring data. This will assist with gaining a better understanding of the architectures operational state – providing a mechanism for simple feature selection when faced with the challenges that come with identifying and assessing

appropriate algorithms on a case-to-case basis.

Anomaly Detection API

The Anomaly Detection API is a pluggable interface for rapid deployment and configuration of applicable algorithms. Hence, it is intended to accommodate a broad variety of existing anomaly detection techniques available. By providing a high degree of customisation, operators will be able to deploy new state-of-the-art tools faster, and break away from the vendor lock in paradigm associated with traditional middle boxes. We propose a flexible interface, so that others are free to implement alternative intrusion detection methods such as packed based signature detection engines. By doing so, we intend to provide a foundational grounding for the rapid deployment of, and experimentation with both anomaly detection and packets based intrusion detection techniques. Table 3.1 presents a summary of algorithms that are applicable for anomaly-based intrusion detection [6]

Type	Description
Markov process	Keeps track of the different states the system is in at a specific time interval. The occurrence of a state is low if it corresponds to abnormality.
Operational	An alarm detecting abnormality is raised if count of events that occur is lower than a threshold value or greater than a threshold value
Statistical moments	An event that falls outside the set interval, above or below the moment can be treated as anomalous. Standard deviation and other summaries are known as moments.
Time series	– If the probability of occurrence of a new observation is too low, it is considered an anomaly.

Decision trees	A decision tree is trained with known data before it can classify new or untrained data. After training, it can classify new data.
Artificial neural networks	A dataset of input vectors and corresponding target vectors are used in training the network to associate input with output.
Artificial neural networks	A neural network is created from audit data to determine a string of events as abnormal or normal.
Bayesian networks	In this graphical technique, nodes in the graph represent random variables. The Bayesian network learns causal or dependency relations among attributes in the training data set before classifying new data.
Fuzzy logic	– Such a system learns the characteristics of network traffic by applying fuzzy logic. Signatures are developed by analysing network protocols.
Outlier detection	– If a data point is very different from the rest of the data, it is an outlier. Abnormal traffic that is very different from acceptable traffic based on metrics is identified as outlier.
Finite state machines	– A state contains information about the past, and any changes in the input are noted to identify the abnormality.

Table 1 – Anomaly Detection Techniques

Progress to Date

To date, we have engaged in a critical analysis of applying existing network security approaches towards emerging programmable, dynamic network infrastructures. Our initial step in this was to conduct a broad review of the literature in order to identify the common strengths and weaknesses associated with existing signature-based and anomaly-based intrusion detection. This was followed by preliminary experimentation towards investigating the application of anomaly-based detection by making use of SDN global traffic statistics. This enabled us to develop new skills that will be required when engaging in the analysis of anomaly detection algorithms, as well as technical skills required for the configuration and experimentation with SDN environments. The development of these skills is an on-going process that will prove as a crucial requirement for future experimentation in our project.

During this, we discovered that challenges outside of our initial investigation scope would require further analysis. These challenges are concerned with system scalability and initiation in virtualized environments

Remaining Challenges

At present, we are assessing the problem characteristics associated with dynamic networks. This will enable us to narrow down which anomaly detection technique is required for our research. We have currently identified the application categories, and associated research areas of Anomaly-based detection for computer networks, however we still need to consider what information LXC containers can provide in assisting with the detection of less volumetric behaviour [19].

In a survey by Bhuyan et al. [20] we are presented with a detailed overview on the applicable methods for anomaly-based network intrusion detection. This has assisted in mapping the foundational topics we are investigating. Building on this knowledge, we are in the process of creating an extensive requirements map that will assist to identify LXC relevant problem characteristics that contribute towards gaining complete situational awareness from both a network global overview, as well as service interaction overview. The current version of this requirements map is available for review by visiting the web link provided in the footnote item 1 of this page.

We have identified a recurring presence in an ability to define configuration policies as well as interaction with both SDN and LXC via RESTful interfaces. With this in mind, we still need to investigate the structural requirements in optimizing this feature towards a unified schema for simplified management specification of system configurations, which will prove advantageous when considering autonomous load distribution and pre-emptive initiation of time-sensitive network functions from a security perspective.

To this end, there is still further work we must conduct in the application of our proposed mechanism towards LXC environments

1)   Requirements map - https://www.dropbox.com/s/u53curt6b0hj3qm/Anomaly%20Detection%20spec.jpg?dl=0

Summary

Data centre hosted cloud services have rapidly become a common encounter across a broad socioeconomic landscape. The ability to ubiquitously interact with cloud services via a wealth of heterogeneous, internet-enabled devices has introduced an extremely dynamic environment that current network security mechanisms are unfit to accommodate.

We have discussed our research findings gained throughout the course of our literature review. From this, we have determined the challenges we will face in regards to the scalability limitations of existing network security solutions when applied towards the emerging software-defined infrastructures.

We have presented our current work to date in analysing the functional properties of emerging network architectures, as well as investigating the performance limitations of current hypervisor-based virtualization. From this, we have found that the existing literature surrounding LXC based containers suggests superior scalability, as well as the ability to reduce VM initiation time for time-sensitive virtual network functions.

References

[1]   B. A. a. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 3, pp. 1617–1634, 2014.

[2]   R. Cziva, S. Jouet, and D. P. Pezaros, "GNFC: Towards network function cloudification," in *2015 IEEE Conference on Network Function Virtualization and Software Defined Network (NFV-SDN)*, 2015, pp. 142–148.

[3]   D. Bernstein, "Containers and Cloud: From LXC to Docker to Kubernetes," *IEEE Cloud Comput.*, vol. 1, no. 3, pp. 81–84, 2014.

[4]   M. Ali, S. U. Khan, and A. V. Vasilakos, "Security in cloud computing: Opportunities

and challenges," *Inf. Sci. (Ny).*, vol. 305, pp. 357–383, Jun. 2015.

[5]   R. Piggin, "Cyber security trends: What should keep CEOs awake at night," *Int. J. Crit. Infrastruct. Prot.*, no. December 2015, pp. 1–3, Feb. 2016.

[6]   R. K. Deka, K. P. Kalita, D. K. Bhattacharya, and J. K. Kalita, "Network defense: Approaches, methods and techniques," *J. Netw. Comput. Appl.*, vol. 57, pp. 71–84, Nov. 2015.

[7]   R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network Function Virtualization: State-of-the-art and Research Challenges," *IEEE Commun. Surv. Tutorials*, no. c, pp. 1–1, 2015.

[8]   C. Cui, H. Deng, D. Telekom, and U. Michel, "Network Functions Virtualisation," 2012.

[9]   Open Networking Foundation, "OpenFlow-enabled SDN and Network Functions Virtualization," 2014.

[10]   N. Kratzke, "About Microservices, Containers and their Underestimated Impact on Network Performance," in *Proceedings of CLOUD COMPUTING 2015 (6th. International Conference on Cloud Computing, GRIDS and Virtualization)*, 2015, no. c, pp. 165–169.

[11]   C. Rotter, G. Csatári, L. Farkas, G. Nyiri, L. Janosi, and R. Springer, "Using Linux Containers in Telecom Applications," in *Innovations in Clouds, Internet and Networks, ICIN (2016)*, 2016, pp. 234–241.

[12]   C. Anderson, "Docker [Software engineering]," *IEEE Softw.*, vol. 32, no. 3, pp. 102–c3, May 2015.

[13]   A. Tosatto, P. Ruiu, and A. Attanasio, "Container-Based Orchestration in Cloud: State of the Art and Challenges," in *2015 Ninth International Conference on Complex, Intelligent, and Software Intensive Systems*, 2015, pp. 70–75.

[14]   J. Batalle, J. Ferrer Riera, E. Escalona, and J. A. Garcia-Espin, "On the Implementation of NFV over an OpenFlow Infrastructure: Routing Function Virtualization," in *2013 IEEE SDN for Future Networks and Services (SDN4FNS)*, 2013, pp. 1–6.

[15]   O. Fundation, "Software-defined networking: The new norm for networks," *ONF White Pap.*, 2012.

[16]   D. Kreutz, F. M. V. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," *Proc. IEEE*, vol. 103, no. 1, pp. 14–76, Jan. 2015.

[17]   N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, p. 69, Mar. 2008.

[18]   B. R. Granby, B. Askwith, and A. K. Marnerides, "SDN-PANDA: Software-Defined Network Platform for ANomaly Detection Applications," in *2015 IEEE 23rd International Conference on Network Protocols (ICNP)*, 2015, pp. 463–466.

[19]   V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.

[20]   M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network Anomaly Detection: Methods, Systems and Tools," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 1, pp. 303–336, Jan. 2014.

Emma Beasor

The evolution of red supergiants to supernova in the LMC cluster NGC 2100

E R Beasor and B Davies
Astrophysics Research Institute, Liverpool John Moores University, L3 5RF, UK
E-mail address: e.beasor@2010.ljmu.ac.uk

Abstract. Studying the mass loss rates of red supergiants (RSGs) is important for the understanding of massive star evolution. RSGs end their lives as supernova, and the mass-loss during this phase can dictate the appearance of the resulting explosion. Here we study the mass-loss rates of 19 RSGs in the young massive cluster NGC 2100 in the Large Magellanic Cloud. By fitting mid-IR photometry from WISE and Spitzer/IRAC to DUSTY models it was possible to determine a best-fit inner dust temperature, optical depth and mass loss rate. We find an increase in mass-loss rate with evolution that is well described by the de Jager mass-loss rate prescription, used widely in stellar evolution calculations. We also find that the most evolved stars in the cluster contain excess reddening compared to the other RSGs in the cluster. This is discussed in terms of stellar evolution and supernova progenitors. We argue there is little justification for substantially increasing the mass loss rate during the RSG phase, as has been argued recently in order to explain the absence of high mass Type IIP supernova progenitors.

1. Introduction

Knowledge of the mass loss rates of red supergiants (RSGs) is fundamentally important for understanding stellar evolution. Changing the mass loss rate has effects on the subsequent evolution of the star, as well as the supernova (SN) type and eventual remnant (1)(2).

When a RSG reaches the end of its lifetime, it explodes as a Type IIP core-collapse supernova (CCSN), of which there have been 7 confirmed cases of RSGs as progenitors. Theory predicts that these progenitors can be anywhere in the range of 8.5 to $25M_{solar}$ (3), but so far it seems the stars which explode are of a relatively low mass, with no progenitors appearing in the higher end of the predicted mass range (4)(5).

Are all RSGs exploding as Type IIP SNe? Or does the extreme mass loss affect the final evolution of these massive stars? Stellar evolution models currently rely on observational or theoretical mass loss rate prescriptions which have been based on observations of field stars, rather than coeval, leaving the parameters of initial mass ($M_{initial}$) and metallicity (Z) unconstrained. This could in part explain the large dispersions observed in these trends. It has been suggested that an increase in the mass loss rates during the RSG phase could

lead to the observed maximum mass of observed Type IIP SNe progenitors (6)(7).

In this paper we measure the amount of circumstellar material and estimate mass-loss rates, to investigate whether this is correlated with how close the star is to SN. We model the mid-infrared excess of 19 RSGs in the stellar cluster NGC 2100, each which we assume has the same initial mass and composition, but where the stars are all at slightly different stages of evolution. This allows us to investigate the mass-loss rate behavior with evolution of the RSG.

We begin in Section 2 by describing our dust shell models and choice of input parameters. In Section 3 we discuss applying this to the stars in cluster NGC 2100 and the results we derive from our models. In Section 4 we discuss our results in terms of RSG evolution and progenitors.

2. Dust shell models
The models used in this project were created using DUSTY (10). Stars surrounded by circumstellar dust have their radiation absorbed/re-emitted by the dust particles, changing the output spectrum of the star. DUSTY solves the radiative transfer equation for a star obscured by a spherical dust shell of a certain optical depth (τ_v, optical depth at 0.55 μm), inner dust temperature (T_{in}) at the inner most radius (R_{in}). Below we describe our choices for the model input parameters and our fitting methodology.

2.1 Model parameters
2.1.1 Dust composition
It is necessary to define a dust grain composition when creating models with DUSTY as this determines the extinction efficiency Q_λ, and hence how the dust shell will reprocess the input spectral energy distribution (SED). Observations of RSGs confirm the dust shells are O-rich, indicated by the presence of mid-IR spectral features at 12 and 18μm known to be caused by the presence of silicates. We therefore opted for O-rich silicate dust as described by (11).

2.1.2 Grain size
Recent observations of VY Canis Majoris (12), a nearby dust-enshrouded RSG, estimated the dust surrounding the star to be of a constant grain size of 0.5μm. This is in line with previous observations (13) who found the grain size to be between 0.3 and 1μm. We therefore created models for constant grain sizes of 0.1, 0.2, 0.3, 0.4 and 0.5μm, choosing 0.3μm as our fiducial grain size.

2.1.3 Density distribution
Here, we assumed a steady state density distribution falling off as r^{-2} in the entire shell with a constant terminal velocity. As we do not know the outflow velocity for the RSGs in our sample and observations of outflow velocities for RSGs are limited, we estimate the outflow velocity to be 20km/s. This is consistent with studies that have used maser emission to map the dust shells of RSGs (14)(15). We specify that the shell extends to 1000 times its inner radius, such that the dust density is low enough at the outer limit so that it has no effect on the spectrum. We assume a gas to dust ratio, r_{gd}, of 500,as $r_{gd} \approx 500$ for LMC (16) and a grain bulk density, r_S of 3 g cm^{-3}.

2.1.4 T_{eff}, T_{in} and τ_v
DUSTY requires an input SED to illuminate the dust shell, so that the light can be reprocessed and re-emitted. The SEDs we use are synthesized from MARCS model

atmospheres (17) using TURBOSPECTRUM (18). The average T_{eff} for the stars in NGC 2100 was recently found to be 3890 ± 85K (19) and we have therefore opted for SED temperatures of 3600K, 3900K, 4200K to fully encompass this range and check how robust our results are to input SED.

Finally, DUSTY also allows inner temperature, T_{in}, and the optical depth τ_V to be chosen. T_{in} defines the temperature of the inner dust shell (and hence it's position). The optical depth determines the dust shell mass. As these parameters are unconstrained, in this study we have allowed them to vary until the fit to the data is optimised. This fitting methodology is described in 2.3.

2.1.5 Calculation of mass-loss rate
Mass loss rate is calculated using equation 1

$$\dot{M} = \frac{16\pi \, R_{in} \tau_\lambda \rho_d a r_{gd}}{3Q_\lambda} \qquad (1)$$

where Q_λ is extinction efficiency, R_{in} is the radius at the inner part of the dust shell, ρ_d is the dust grain density, a is the grain size and r_{gd} is the gas to dust ratio.

2.3 Fitting methodology
We first computed a grid of dust shell models spanning a range of inner temperatures and optical depths. For each model we then computed synthetic WISE and Spitzer photometry by convolving the model spectrum with their relevent filter profiles. This synthetic model photometry was compared to each stars mid-IR photometry from WISE, IRAC and MIPS. The grid spanned τ_V values of 0 - 1.3 with 50 grid points, and inner temperature values from 100K to 1200K in steps of 100K. By using χ^2 minimisation we determined the best fitting model to the sample SED. To account for systematic errors we applied a blanket error of 10% to our observations. The errors on our best fitting model parameters were determined by models within our lowest $\chi^2 \pm 10$. This limit was chosen so that the stars with the lowest measured mass-loss rates which were clearly consistent with non-detections, would have mass-loss rate values consistent with 0 (or upper limits only).

3 Application to NGC 2100

Star	T_{in} (K)	τ_V	$\dot{M}$ ($10^{-6}M_\odot$ yr^{-1})	L_{bol}	A_V
1	600^{+200}_{-100}	$0.56^{+0.21}_{-0.14}$	$9.89^{+4.20}_{-3.17}$	5.09 ± 0.09	$0.09^{+0.04}_{-0.02}$
2	600^{+200}_{-100}	$0.64^{+0.26}_{-0.16}$	$9.97^{+4.52}_{-3.19}$	4.97 ± 0.09	$0.10^{+0.05}_{-0.03}$
3	600^{+400}_{-200}	$0.16^{+0.08}_{-0.05}$	$1.98^{+1.07}_{-0.74}$	4.84 ± 0.09	$0.02^{+0.01}_{-0.01}$
4	800^{+400}_{-200}	$0.45^{+0.32}_{-0.16}$	$3.17^{+2.34}_{-1.29}$	4.71 ± 0.09	$0.07^{+0.06}_{-0.03}$
5	700^{+300}_{-200}	$0.29^{+0.16}_{-0.08}$	$2.54^{+1.49}_{-0.86}$	4.73 ± 0.09	$0.04^{+0.03}_{-0.01}$
6	500^{+300}_{-100}	$0.21^{+0.13}_{-0.02}$	$3.25^{+2.12}_{-0.72}$	4.77 ± 0.09	$0.03^{+0.02}_{-0.00}$
7	1200^{+0}_{-500}	$0.27^{+0.07}_{-0.14}$	$0.82^{+0.27}_{-0.46}$	4.68 ± 0.09	$0.04^{+0.01}_{-0.02}$
8	1000^{+200}_{-400}	$0.29^{+0.16}_{-0.10}$	$1.29^{+0.76}_{-0.51}$	4.68 ± 0.09	$0.04^{+0.03}_{-0.01}$
9	1200^{+0}_{-400}	$0.16^{+0.08}_{-0.05}$	$0.48^{+0.26}_{-0.18}$	4.68 ± 0.09	$0.02^{+0.01}_{-0.01}$
10	600^{+500}_{-200}	$0.11^{+0.08}_{-0.03}$	$1.06^{+0.80}_{-0.36}$	4.63 ± 0.09	$0.01^{+0.01}_{-0.00}$
11	1200^{+0}_{-700}	$0.11^{+0.05}_{-0.06}$	$0.28^{+0.14}_{-0.16}$	4.55 ± 0.09	$0.01^{+0.01}_{-0.01}$
12	1200^{+0}_{-600}	$0.16^{+0.08}_{-0.08}$	$0.39^{+0.21}_{-0.21}$	4.51 ± 0.09	$0.02^{+0.01}_{-0.01}$
14	–	< 0.05	< 0.12	4.53 ± 0.10	–
15	–	< 0.03	< 0.08	4.56 ± 0.09	–
16	400^{+300}_{-100}	$0.13^{+0.06}_{-0.02}$	$2.27^{+1.14}_{-0.57}$	4.55 ± 0.09	$0.020^{+0.010}_{-0.000}$
17	1100^{+100}_{-700}	$0.08^{+0.05}_{-0.05}$	$0.23^{+0.15}_{-0.15}$	4.49 ± 0.09	$0.010^{+0.010}_{-0.010}$
18	–	< 0.03	< 0.08	4.43 ± 0.09	–
19	–	< 0.03	< 0.07	4.45 ± 0.11	–

Table 1: Results for stars in NGC 2100. Stars are numbered with #1 having the highest [5.6]-band magnitude and #19 having the lowest. Luminosities quoted are in units of $\log(L_{bol}/L_\odot)$. A_V is the extinction intrinsic to the dust shell.

In this study we apply this dust modeling to a sample of RSGs in a young star cluster. Such clusters can be assumed to be coeval, since any spread in the age of the stars will be small compared to the age of the cluster. Since all the stars currently in the RSG phase will have the same initial mass to within a few tenths of a Solar mass, they will all follow almost the same path across the H-R diagram. Differences in luminosity are caused by those stars with slightly higher masses evolving along the mass-track at a slightly faster rate. It is for this reason that luminosity can be taken as a proxy for evolution. A finding chart for NGC2100 is shown in Figure 1 in which the RSGs are numbered based on [5.6]-band magnitude. Star #13 has been omitted from our analysis due to large disagreements between the MIPS and WISE photometry, as well as WISE and IRAC.

From the photometry alone it was possible to see evidence of mass loss rate evolving with evolution of the RSG. This qualitative evidence is shown in the [8-12] vs. [5.6] CMD, Figure 2. The 5.6-band magnitude can be used as a measure of luminosity as the bolometric correction at this wavelength is largely insensitive to the RSGs temperatures, whilst also being too short a wavelength to be significantly affected by emission from circumstellar dust. The [8-12] colour can be used as a measure of dust shell mass as it measures the excess caused by the broad silicate feature at 10μm. It can be seen from Figure 2 that more

luminous (and therefore, more evolved) RSGs have a larger amount of dust surrounding them (shown by the increasing colour, meaning they appear more reddened), suggesting dust mass increases with age.

Below we discuss our modeling results and compare them to mass-loss rate prescriptions frequently used by stellar evolution groups.

3.1 Modelling results

We ran our fitting procedure for the 19 RSGs located in NGC2100, our results are shown in Table 1. Figure 3 shows an example model fit with observed photometry for star #1. The plot shows our best fit model spectra (green line), the models within our error range (blue dotted lines) and the various contributions to the flux, including scattered flux, dust emission and attenuated flux. It also shows the photometric data (red crosses) and model photometry (green crosses). The 10µm silicate bump can be clearly seen due to dust emission (pink dashed line).
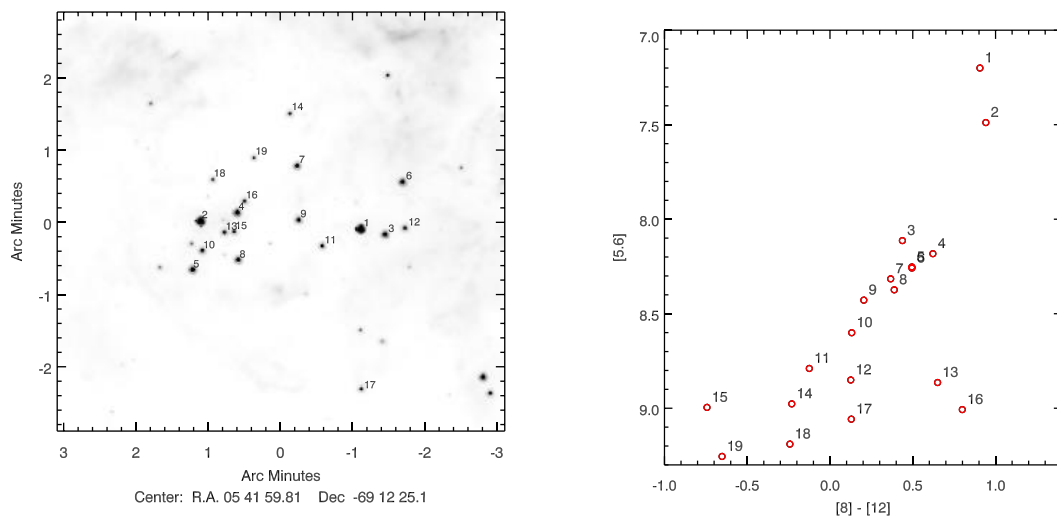
Figure 1: (Left) Finding chart for RSGs in NGC 2100. The stars are numbered based on [5.6]-band magnitude

Figure 2: (Right) Colour magnitude plot of RSGs in the cluster to show increasing dust mass with age. [5.6]-band magnitude is used as an indicator of L_{bol} and the [8-12] colour is used as a measure of dust shell mass. The [8-12] colour is useful as is includes the mid-IR excess and the excess caused by the broad silicate feature.

The plot also shows the significant effect scattering within the dust shell (grey dotted/dashed line), contributing to a large proportion of the optical output spectrum. In Figure 3 we also show a contour plot illustrating the degeneracy between our two free parameters, T_{in} and τ_V, with $\dot{M}$ contours for the best fit $\dot{M}$ and upper and lower $\dot{M}$ contours overplotted. It can be seen that the lines of equal $\dot{M}$ run approximately parallel to the χ^2 contours. This means that despite the degeneracy between τ_V and T_{in} the value of $\dot{M}$ is well constrained and robust to where we place the inner dust rim.

(a)   (b)
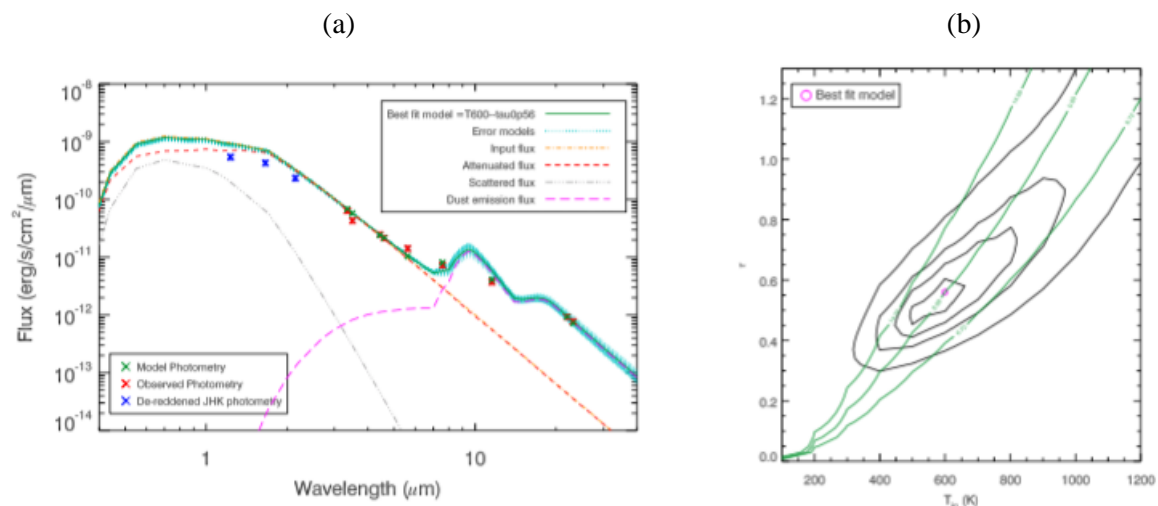
Figure 3: *Left pane (a)l:* Model plot for the star with the highest $\dot{M}$ value in NGC 2100 including all contributions to spectrum. The stars are numbered based on 5.6-band magnitude (#1 being the star with the highest [5.6]-band magnitude). The silicate bump is clearly visible on the spectra suggesting a large amount of circumstellar material. *Right pane (b)l:* Contour plot showing the degeneracy between χ^2 values and best fitting $\dot{M}$ values in units of 10^{-7} $M_\odot$ yr^{-1}. The green lines show the best fit $\dot{M}$ and upper and lower mass loss rate isocontours. It can be seen that while there is some degeneracy between inner dust temperature and optical depth the value of mass loss rate is independent of this.

Fit results for all stars modeled are shown in Table 2. We find a varying T_{in} value for RSGs in this cluster, not a constant value at the dust sublimation temperature of 1200K. L_{bol} is found by integrating under the model spectra with errors on L_{bol} dominated by the uncertainty in T_{eff} . The value of A_V is found from the ratio of input and output fluxes at
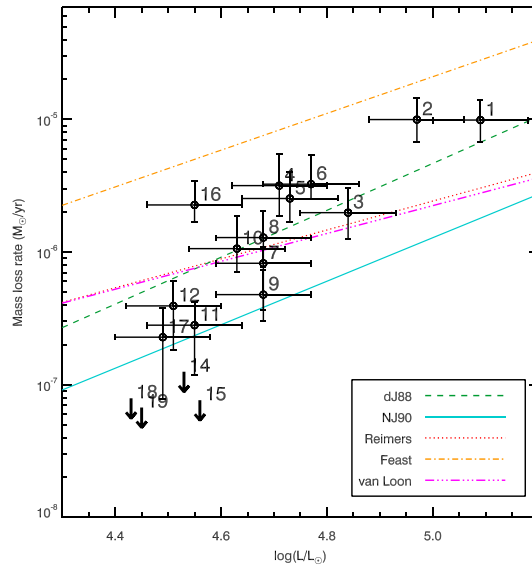
Figure 4: Plot showing mass loss rate versus L_{bol}. A positive correlation can be seen suggesting mass loss rate increases with evolution. This is compared to some mass loss rate prescriptions. The downward arrows show for which stars we only have upper limits on mass loss rate.

0.55μm and is intrinsic to the dust shell. For stars numbered 15, 18 and 19 the value of mass loss rate is so low it can be considered as a non-detection, leaving T_{in} unconstrained.

A positive correlation between mass loss rate and luminosity is illustrated in Figure 4, implying that mass loss rate increases by a factor of 100 during the RSG phase, which according to model predictions should last ap- proximately 10^6 years for stars with initial masses of $15 M_\odot$ (20), see Section 3.2. This plot also shows some mass loss rate prescriptions for comparison (assuming a T_{eff} of 4000K); (21)(hereafter dJ88), Reimer's prescription (22), van Loon's prescription (23), (24)(hereafter NJ90) and Feast's prescription (25). We find our results are best fit by dJ88, van Loon and Reimer's prescriptions, with dJ88 providing a better fit for the more evolved stars (where the mass loss mechanism is stronger).

4 Discussion

We find a clear increase in mass loss rate with RSG evolution, by a factor of ~100 through the lifetime of the star. These results are well described by mass-loss rate prescriptions currently used by some stellar evolution models, particularly dJ88 which matches the mass loss rate of the most evolved RSGs in our study (see Figure 4). We find very little spread of L_{bol} with mass loss rate unlike that observed for field RSGs (e.g. 23). The spread observed in previous results could be due to a varying $M_{initial}$ in the sample stars. By focusing our study on a coeval star cluster we have kept metallicity and initial mass fixed, showing the mass-loss rate prescriptions fit well for LMC metallicity and $M_{initial}$ of $15 M_\odot$.

Mass loss due to stellar winds is a hugely important factor in determining the evolution of the

most massive stars. There is uncertainty about the total amount of mass lost during the RSG phase, and therefore about the exact nature of the immediate SNe progenitors. The impact of mass loss rate on RSG lifetimes, evolution and pre-SNe properties has been studied by computing stellar models for initial masses between 9 and $25M_\odot$ and increasing the mass loss rate by 10 times and 25 times (25). It is discussed in (6) the case for an increased mass loss rate during the RSG phase. By increasing the standard mass loss rate by a factor of 3 in the models, (6) find a blueward motion in the HRD is observed for stars more massive than $25M_\odot$ (non-rotating models) or $20M_\odot$ (rotating models, see (7)).

As can be seen in figure 4 we find the accepted mass loss rate prescriptions commonly used in stellar evolution codes fit well when the variables Z and $M_{initial}$ are fixed. For this $M_{initial}$ ($\sim15M_\odot$) and at LMC metallicity altering the mass loss rate prescriptions seems unjustified. Increasing the mass loss rate by a factor of 10 (as in 25) would result in a strong conflict with our findings.

5 Conclusions

Understanding the nature of the mass loss mechanism present in RSGs remains an important field of study in stellar astrophysics. Here a method of deriving various stellar parameters, T_{in}, τ_V , mass loss rate was presented as well as evidence for an increasing value of mass loss rate with RSG evolution. By targeting stars in a coeval cluster it was possible to study mass loss rate while keeping metallicity, age and $M_{initial}$ constrained. As all stars currently in the RSG phase will have the same initial mass to within a few tenths of a solar mass, it is possible to use luminosity as a proxy for evolution, due to those stars with slightly higher masses evolving through the HR diagram at slightly faster rates. From our study we can conclude the most luminous stars were found to have the highest value of mass loss rate evidenced observationally by colour-magnitude diagrams and also by a positive correlation between bolometric luminosity and mass loss rate.

Our results are well modeled by various mass-loss rate prescriptions currently used by some stellar evolution groups, such as dJ88 and Reimer's, with dJ88 providing a better fit for the RSGs with stronger mass loss rate. We therefore see no evidence for a significantly increased mass loss rate during the RSG phase as has been suggested by various stellar evolutionary groups

 Future work will involve applying this technique to RSGs at solar metallicity to see if the mass-loss rate prescriptions are still appropriate. We also plan to apply this technique to a cluster where the stars have higher initial masses closer to the upper RSG limit.

References

[1] Maeder A., 1981, Astronomy and Astrophysics, 99, 97
[2] Chiosi C., Maeder A., 1986, Annual Review of Astronomy and Astrophysics, 24, 329
[3] Meynet G., Maeder A., 2003, Astronomy & Astrophysics, 404, 975
[4] Smartt S., Eldridge J., Crockett R., Maund J. R., 2009, Monthly Notices of the Royal Astronomical Society, 395, 1409
[5] Smartt S., 2015, Publications of the Astronomical Society of Australia, 32, e016
[6] Georgy C., Ekström S., 2015, arXiv preprint arXiv:1508.04656
[7] Georgy C., 2012, Astronomy & Astrophysics, 538, L8
[8] de Wit W., et al., 2008, The Astrophysical Journal Letters, 685, L75
[9] Danchi W., Bester M., Degiacomi C., Greenhill L., Townes C., 1994, The Astronomical

Journal, 107, 1469

[10] Ivezic Z., Nenkova M., Elitzur M., 1999, Astrophysics Source Code Library, 1, 11001

[11] Draine B., Lee H. M., 1984, The Astrophysical Journal, 285, 89

[12] Scicluna P., Siebenmorgen R., Wesson R., Blommaert J., Kasper M., Voshchinnikov N., Wolf S., 2015, Online Material p, 1

[13] Smith N., Humphreys R. M., Davidson K., Gehrz R. D., Schuster M., Krautter J., 2001, The Astronomical Journal, 121, 1111

[14] Van Loon J. T., Zijlstra A. A., Bujarrabal V., Nyman L.-Å., 2001, Astronomy & Astrophysics, 368, 950

[15] Richards A., Yates J., 1998, Irish Astronomical Journal, 25, 7

[17] Gustafsson B., Edvardsson B., Eriksson K., Jørgensen U. G., Nordlund Å., Plez B., 2008, Astronomy & Astrophysics, 486, 951

[18] Plez B., 2012, Astrophysics Source Code Library, 1, 05004

[19] Patrick L., Evans C., Davies B., Kudritzki R., Hénault-Brunet V., Bastian, N., Lapenna E., Bergemann M., 2016, Monthly Notices of the Royal Astronomical Society, p. stw561

[20] Hirschi R., Meynet G., Maeder A., 2004, Astronomy & Astrophysics, 425, 649

[21] De Jager C., Nieuwenhuijzen H., Van Der Hucht K., 1988, Astronomy and Astrophysics Supplement Series, 72, 259

[22] Reimers D., 1975, Memoires of the Societe Royale des Sciences de Liege, 8, 369

[23] Van Loon J. T., Cioni M.-R., Zijlstra A. A., Loup C., 2005, Astronomy & Astrophysics, 438, 273

[24] Niederhofer F., Hilker M., Bastian N., Silva-Villa E., 2015, Astronomy & Astrophysics, 575, A62

[25] Meynet G., et al., 2015a, Astronomy & Astrophysics, 575, A60

Paul Ross McWhirter

An Intelligent, data analytical framework for Astronomical Light Curve Analysis and Classification

Paul R McWhirter[12], Iain A Steele[2], Dhiya Al-Jumeily[1], Abir Hussain[1] and Paul Fergus[1]

[1]Liverpool John Moores University, Applied Computing Research Group, Faculty of Engineering and Technology, Byrom Street, Liverpool, L3 3AF, UK.

P.R.McWhirter@2014.ljmu.ac.uk
{D.Aljumeily, A.Hussain, P.Fergus}@ljmu.ac.uk

[2]Liverpool John Moores University, Astrophysics Research Institute, IC2, Liverpool Science Park, 146 Brownlow Hill, Liverpool, L3 5RF, UK.
I.A.Steele@ljmu.ac.uk

Abstract. Modern time-domain astronomy is capable of collecting a staggeringly large amount of data on millions of objects in real time. Therefore the production of methods and systems for the automated classification of time-domain astronomical objects is of great importance. The Liverpool Telescope has a number of wide-field image gathering instruments mounted upon its structure. These instruments have been in operation since March 2009 gathering data of large areas of sky around the current field of view of the main telescope. We applied a method designed to extract time-translation invariant features from the time-series light curves of each object for future input into a classification system. These efforts were met with limited success due to noise and uneven sampling within the time-series data. Therefore we propose a methodology capable of distributing the data analysis whilst simultaneously an intelligence service operates to constrain the light curves and eliminate false signals due to noise. This system will be highly scalable whilst maintaining the production of accurate features based on the fitting of harmonic models to the light curves within the initial Structured Query Language (SQL) database.

Introduction

Astronomy is entering a period of unprecedented data gathering capability. Advances in observational, storage and data processing technologies have allowed for extended sky surveys such as the Sloan Digital Sky Survey (SDSS) to be conducted and exploited [1]. Within the next decade a number of even larger surveys are planned such as the Large Synoptic Survey Telescope (LSST) [2]. Technology is now at a point where it has become possible to gather data on wide regions of the sky repeatedly over variable time periods [3]. This data can be analyzed through the use of periodograms to identify periodic structure. Fitted weighted regression learning models can then provide us with valuable knowledge about the presence and classification of astronomical objects that are periodically changing in time [4].

Time domain astronomy is a research area characterized by the large datasets generated by sky surveys containing time-series data [5]. Time-series data contains information on the temporal component of measurements and the whole time-series contains observations across multiple epochs. In Astronomy, it is common for these observations to have an

uneven distribution in time with inconsistent intervals [6]. A major cause of this is weather limitations that can prevent telescope operation for uncertain periods of time. There is a demand for data processing techniques capable of the automated processing of time-series data on individual objects with observations over many days followed by no additional observations for months [7].

In this paper we demonstrate a methodology for the extraction of harmonic features from the light curves and identify where multithreaded operations can be exploited. Additionally, we propose a new intelligence system capable of producing superior features. The rest of this paper is structured as follows. In Section 2, the background of time-domain astronomy is discussed. Section 3 introduces the Small Telescopes Installed at the Liverpool Telescope (STILT) instruments. In Section 4, the feature extraction is described. Section 5 will discuss the new proposed multithreaded machine learning solution. The final conclusions and proposals of future work are provided within Section 6.

Background

Astronomical time-series data is generated through the production of wide-angle images of the sky. By identifying objects in multiple images with different observation times, information on the change of the brightness of these objects can be determined. The resulting brightness-over-time data for each individual object is called the objects light curve [3, 8].

Many astronomical objects exhibit brightness variability due to a large number of differing physical processes that uniquely influence an object's light curve. Therefore, the light curve can be used in the classification of variable objects based on the signature of these physical processes and the detection of unknown candidate objects or even unknown variability phenomena [9]. The first major type of variable astronomical phenomenon is Variable Stars [10]. Variable stars are unstable stars and undergo periods of pulsation where they grow and contract in size [11]. These size oscillations produce changes to the stars temperature and brightness resulting in a measurable change upon the light curves [3, 8]. The light curves of these pulsating stars can be used to produce descriptive features. Models produced from these features can then be used to identify the class of candidate variable objects and the period of their oscillations.

A second important type of variable object is the eclipsing binary [12]. In these systems, two or more stars are in close proximity to each other and execute orbits around a common gravitational center-point. The close proximity of the stars often means that they cannot be distinguished on an image and appear as a single source of light. Variations in these objects are caused by the plane of the orbit aligning with the view from Earth. As a result, one star periodically passes in front of another resulting in a change in the brightness of the source of light in the astronomical images.

Finally, there are also transient events that result in harder-to-predict phenomena [4]. Flare Stars are stars that can undergo occasional outbursts due to magnetic and plasma processes within their atmospheres. These events can be repetitive but not usually with the degree of periodicity of variable stars. For purely transient events, two of the most studied examples are Novae and Supernovae caused by the cataclysmic eruption of stellar material, producing some of the brightness objects in the known universe as the victim star is destroyed or badly disrupted during the event.

The STILT Dataset

The Small Telescopes Installed at the Liverpool Telescope (STILT) dataset is a wide field object SQL database. It contains 1.24 billion separate object observations of 27.74 million independent stellar objects. It was generated through the pre-processing of observational

images gathered by the STILT instruments [3]. The Liverpool Telescope is located at the Observatorio del Roque de los Muchachos on La Palma [13]. The STILT instruments consist of three cameras mounted directly to the body of the main Liverpool Telescope aimed co-parallel with the main telescope's field of view.

As the Skycam images are centered on the view of the main Liverpool Telescope, observations of specific objects are only recorded when they are within the field of view of the camera as the telescope is focused within the vicinity of the objects. This results in time-series with uneven length gaps between observations, greatly increasing the difficulty of identifying variations in the magnitude of the observations.

Feature Extraction

We begin our analysis using a methodology proposed by Debosscher et al. in 2007 and improved upon by Richards et al. in 2011 [14, 4]. The goal is to describe the time-series data for each object as a set of harmonic features that are invariant to the objects mean magnitude and time-translation phase allowing the features to be directly compared to other objects of differing classes. The whole SkycamT database used in this investigation is 180 GB in size with 20 GB of indexes for faster query response times. For each object, a set methodology is applied to generate an associated feature vector. The database is queried for all observations of a specific object. The returned table has its magnitude, date and magnitude error columns retrieved. The identification of the dominant periodic oscillation within the object's time-series is then required. There are a number of possible algorithms that can be deployed on uneven time-series.

Phase dispersion minimization [15] and the String Length Lafler-Kinman statistic can identify how well-aligned data points are placed in phase-space across a sample range of periods [16]. This is accomplished through computing the distance between each data point in phase space. This calculation is performed across a frequency spectrum of candidate frequencies. Upon the alignment of the data points at a frequency close to (or a multiple of) the true frequency, this statistic is minimized [16]. An extension to this idea is a recently proposed periodogram based on the Blum-Kiefer-Rosenblatt statistical independence test [17]. Instead of utilizing the alignment of the data such as in the string-length methods, a rank correlation test is performed on the phase-folded data within each candidate period phase space. As the phase folded light curve aligns at a strong period, the correlation between the magnitude and phase of each data point rises. There also exist Information Theoretic approaches such as slotted correntropy and the improved Correntropy Kernelized Periodogram [18] that have proven to be very effective and are a focus for future initiatives on the Skycam database [8]. However, for this investigation, the methodology proposed by Debosscher et al. in 2007 is utilized [14]. Therefore, a Lomb-Scargle Periodogram is utilized to identify the primary periodic signal within the data.

The Lomb-Scargle Periodogram uses a least-squares spectral analysis. It is a method of estimating the frequency spectrum of time-series data by the fitting of multiple sinusoids to the data using least-squares regression [6, 4]. Like the phase dispersion minimization this method is performed over a frequency range resulting in the statistic normalized power that has a larger value if the fitted sinusoid has a lower chi-squared error with a candidate frequency. When the Lomb-Scargle Periodogram is applied to the time-series observational data, the frequency associated with the maximum power is recorded as the primary frequency which is the reciprocal of the primary period.

Upon the determination of this primary period, a four-harmonic sinusoidal model with a linear trend is then fitted using weighted linear regression with this detected period across the time-series data. This model is then subtracted from the time-series in a process called pre-

whitening. This is done as to eliminate any periodic activity within the time-series based on the previously detected period. This pre-whitened time-series is then used to identify a second period independent of the first dominant period. A harmonic model is then fit for this period and subtracted off in a second pre-whitening phase. Finally, a third period is identified independent to the first two periods. The time-series is then restored to the original time-series archived prior to the pre-whitening operations. A harmonic best-fit is again computed by weighted linear regression for all three periods as shown in Equation 1.

$$y(t) = ct + \sum_{i=1}^{3} \sum_{j=1}^{4} \{a_{ij} \sin(2\pi j f_i t) + b_{ij} \cos(2\pi j f_i t)\} + b_0 \qquad (1)$$

Where the b_0 parameter is the mean magnitude of the light curve and the c parameter is the linear trend of the time-series. The frequencies f_i and the coefficients a_{ij} and b_{ij} provide a good description of the light curve as long as it is periodic and well-described by a sum of sinusoids. These coefficients are not yet time-translation invariant and must be transformed from Fourier coefficients into a set of amplitudes A_{ij} and phases PH_{ij}. The amplitudes and phases are computed using trigonometric identities shown in equations 2 and 3 [14].

$$A_{ij} = \sqrt{a_{ij}^2 + b_{ij}^2} \qquad (2)$$

$$PH_{ij} = arctan(b_{ij}, a_{ij}) \qquad (3)$$

The phases are not yet time-translation invariant and are defined relative to PH_{11}, the phase of the first harmonic of the dominant period using equation 4.

$$PH_{ij}' = arctan(b_{ij}, a_{ij}) - \left(\frac{i f_i}{f_1}\right) arctan(b_{11}, a_{11}) \qquad (4)$$

The phases are then constrained between $-\pi$ to $+\pi$ by the transformation in equation 5. For simplicity the double dash is dropped through the rest of the paper.

$$PH_{ij}'' = arctan\left(sin(PH_{ij}'), cos(PH_{ij}')\right) \qquad (5)$$

The harmonic features include the slope of the linear trend, the three frequencies used in the final harmonic model, the twelve amplitude coefficients and eleven phase coefficients (as PH_{11} is always zero it is discarded) and the ratio of data variance (called variance ratio) between the variance before the pre-whitening of the harmonic model of the primary period and after. This statistic is a strong indicator of the importance of the primary period to the light curve relative to the other periods.

These features have been successfully implemented into classifiers by Debosscher et al in 2007 and Richards et al. in 2011 [14, 4]. These studies made use of well sampled datasets collected by orbital space telescopes and large collaborations. The SkycamT database has very sparse and noisy time-series data. Additionally, the Lomb-Scargle Periodogram is known to strongly identify periodicities for variable stars that are highly sinusoidal such as Mira-class variables. But it can also struggle with less sinusoidal light curves such as eclipsing binaries occasionally missing the period of offering a multiple of the correct period instead of the true period.

Figure 1 demonstrates the model produced by the described method for the SkycamT data collected on the star Mira, the prototype of the Mira class variables showing a clear sinusoidal oscillation. As long as the periodogram returns a result similar to the stars correct period, the linear regression can produce an accurate model. This model is sufficient to generate features within the ranges expected of Mira class variables.

0870−0022951 Light Curve with fitted model
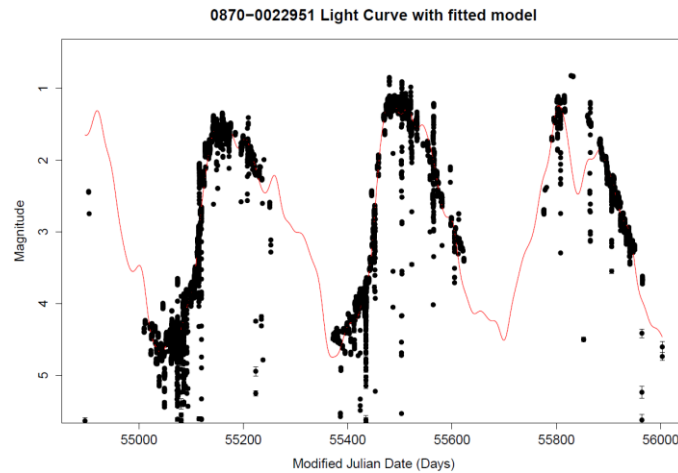
- **Fig. 1: Example model.**

The light curve of the star Mira with a harmonic fit with a primary period of 316 days compared to the period determined by the HIPPARCOS survey of 332 days [19]. Despite this two week discrepancy, the harmonic regression model is a good fit to the data.

Unfortunately this method can result in situations where the model deviates drastically in the non-sampled regions due to 'noisy fringes' in the data. The linear regression is resilient to noise and can evaluate the signals within very poor data. However, coupled with the uneven sampling rate, the noise can result in the linear regression computing a region devoid of data whilst fitting for an incorrect gradient due to the noise of these last data points. As a result the amplitude of the sinusoids can peak beyond a realistic state.

Finally, as the object database exhibits 27.74 million independent stellar objects, the performance of the analysis system is of great concern. Fortunately, the weighted linear regression is implemented using a very efficient normal equation method. Therefore the Lomb-Scargle Periodogram which is $O(N^2)$ in processing complexity is the primary processing component of this analysis method. This is a result of needing to run every observation of an object over a high resolution frequency spectrum to extract the dominate periods otherwise important harmonic variations may be missed. The frequency spectrum can contain tens of thousands of candidate periods to be evaluated. Each period, regardless of proximity to each other, is completely independent. Therefore, this loop can be performed stochastically in parallel as the result is the same whether each candidate period is run on the same CPU core or across thousands of separate CPU cores. This allows the Lomb-Scargle Periodogram to be heavily multithreaded if the hardware exists for considerable performance improvement.

Proposed Solution

In order to improve performance, we propose a system that subdivides the processing tasks between multiple processing cores whilst simultaneously a new 'intelligence service' will constantly monitor the models being generated by the individual agents. This intelligence system will learn stochastically as models are continuously generated for the light curves of different objects.

This system is envisaged to use a form of neural network, operating in the light curves phase space, producing multiple models based on the different light curve profiles discovered. Both Recurrent and Convolutional neural networks have been shown to be potent at predicting time-series [20, 21]. Additionally, these methods can be extended into deep learning through

the addition of more layers if the light curves require the production of more powerful features.

In order to reduce noise, techniques such as wavelet analysis can also be applied [20]. Wavelet analysis has previously been used successfully in the process of creating abstract images of Gamma Ray Burst transient events that retain both temporal and spectral features for classification [22]. It is very difficult to simply filter the noise from the time-series data as the noise has an amplitude very similar to many of the signals. The combination of scalable processing and accurate time-series predictions will lead to high performance processing of the STILT database through the generation of robust features by supervised and unsupervised learning for future multi-class classification analysis.

Conclusion and Future work

The weighted linear regression harmonic best-fit models can be used to produce time-translation invariant features from uneven time-series. The STILT database contains many objects with sufficient noise and uneven sampling to result in poor or physically unrealistic harmonic models. The Lomb-Scargle Periodogram can produce multiples of the correct period and occasionally it misses the periodic signal completely. This problem is exaggerated by light curves exhibiting highly non-sinusoidal signals such as eclipsing binaries. Replacing the Lomb-Scargle Periodogram with a more powerful and less limiting algorithm such as the correntropy kernelized periodogram might alleviate this problem. The proposed solution seeks to introduce scalability and multithreading for the high performance processing of the STILT database.

Our future work will involve the incorporation of the proposed methods into a newly developed data analytics platform. Following this, the models produced can be evaluated through testing previously classified variable objects in the STILT dataset as well as sourcing external datasets for comparative results. These efforts allow the production of robust light curve features that are well placed for future incorporation into a powerful multi-class classification system to rapidly and intelligently perform automated identification of all variable objects within the STILT database.

References

[1] D. G. York, J. Adelman, et al., "The Sloan Digital Sky Survey: Technical Summary," *The Astronomical Journal*, vol. **120**, no. **3**, pp. 1579-2000, 2000.
[2] Z. e. a. Ivezic, "LSST: from science drivers to reference design and anticipated data products," *ArXiv e-prints*, 2011.
[3] N. R. Mawson, I. A. Steele and R. J. Smith, "STILT: System design and performance," *Astronomische Nachrichten*, vol. **334**, no. **7**, pp. 729-737, 2013.
[4] J. W. Richards, D. L. Starr, et al., "On Machine-Learned Classification of Variable Stars with Sparse and Noisy Time-Series Data," *The Astrophysics Journal*, vol. **733**, no. **1**, p. 10, 2011.
[5] S. Vaughan, "Random time series in astronomy," *Philosophical Transactions of the Royal Society*, vol. **371**, no. **20110549**, 2011.
[6] J. D. Scargle, "Studies in Astronomical Time Series Analysis. II. Statistical aspects of spectral analysis of unevenly spaced data," *The Astrophysical Journal*, vol. **263**, pp. 835-853, 1982.
[7] J. S. Bloom and J. W. Richards, "Data Mining and Machine-Learning in Time-Domain Discovery and Classification," in *Advances in Machine Learning and Data Mining for Astronomy*, Taylor & Francis Group, 2011.

[8] P. Huijse, P. A. Estévez, et al., "An Information Theoretic Algorithm for Finding Periodicities in Stellar Light Curves," *IEEE Transactions on Signal Processing*, vol. **60**, no. **10**, pp. 5135-5145, 2012.

[9] P. Protopapas, J. M. Giammarco, et al., "Finding outlier light curves in catalogues of periodic variable stars," *The Royal Astronomical Society, Monthly Notices*, vol. **369**, pp. 677-696, 2006.

[10] L. Eyer and N. Mowlavi, "Variable stars across the observational hr diagram," *Journal of Physics: Conference Series*, vol. **118**, no. **1**, p. 012010, 2008.

[11] J. Percy, Understanding Variable Stars, *Cambridge University Press*, 2007.

[12] D. M. LaCourse, K. J. Jek, et al., "Kepler eclipsing binary stars - VI. Identification of eclipsing binaries in the K2 Campaign o data set," *Monthly Notices of the Royal Astronomical Society*, vol. **452**, no. **4**, pp. 3561-3592, 2015.

[13] I. A. Steele, R. J. Smith, et al., "The Liverpool Telescope: performance and first results.," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 2004.

[14] J. Debosscher, L. M. Sarro, et al., "Automated supervised classification of variable stars I. Methodology," *Astronomy and Astrophysics*, no. **475**, pp. 1159-1183, 2007.

[15] R. F. Stellingwerf, "Period Determination using Phase Dispersion Minimization," *The Astrophysical Journal*, vol. **224**, pp. 953-960, 1978.

[16] D. Clarke, "String/Rope length methods using the Lafler-Kinman statistic," *Astronomy and Astrophysics*, vol. **2**, no. **386**, pp. 763-774, 2002.

[17] S. Zucker, "Detection of Periodicity Based on Independence Tests - II. Improved Serial Independence Measure," *Monthly Notices Letters of the Royal Astronomical Society*, vol. **1**, no. **457**, pp. 118-121, 2016.

[18] W. Liu, P. P. Pokharel and J. C. Principe, "Correntropy: A Localized Similarity Measure," in *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, Vancouver, BC, 2006.

[19] T. R. Bedding and A. A. Zulstra, "HIPPARCOS Period-Luminosity relations for Mira and semiregular variables," *The Astrophysical Journal*, vol. **506**, pp. 47-50, 1998.

[20] M. Langkvist, L. Karlsson and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognition Letters*, vol. **1**, no. **42**, pp. 11-24, 2014.

[21] M. Dalto, "Deep neural networks for time series prediction with applications in ultra-short-term wind forecasting," in *Industrial Technology (ICIT), 2015 IEEE International Conference on*, Seville, 2015.

[22] T. N. Ukwatta and P. R. Wozniak, "Integrating Temporal and Spectral Features of Astronomical Data Using Wavelet Analysis for Source Classification," in *2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Imaging: Earth and Beyond* , Washington DC, 2015.

Gavin Lamb

On-axis orphan afterglow as gravitational wave counterparts

G P Lamb

Astrophysics Research Institute, IC2 Liverpool Science Park, 146 Brownlow Hill, L3 5RF, UK

g.p.lamb@2010.ljmu.ac.uk

Abstract. With the recent detection of gravitational wave (GW) source GW150914 a new non-electromagnetic window on the Universe has been opened. The search for an electromagnetic counterpart to such a GW signal is essential in fully realising the scientific potential of this new type of astronomy. Astrophysical candidates for detectable GW sources include coalescing binary systems with neutron star (NS) or black hole (BH) components. Compact binary mergers of NS-NS or NS-BH systems are thought to be the origin of the collimated jets of short gamma-ray bursts (GRB), and are predicted to form isotropic kilo/macronova. In addition to these counterparts, we consider the on-axis orphan afterglow of failed GRB; by applying the typically observed distribution of Lorentz factor for astrophysical jets to the standard GRB model we predict a significant fraction of compact binary mergers result in failed GRB. On-axis orphan afterglows are a potentially promising counterpart to a detected GW signal from a NS-NS or NS-BH merger. The peak optical flux of an on-axis orphan afterglow within 300 Mpc is detectable in ~70% cases by mid-sized robotic telescopes such as the Liverpool Telescope, and ~100% by next generation telescopes such as the New Robotic Telescope (LT2).

Keywords: high-energy astrophysics, gamma-ray bursts, gravitational waves, electromagnetic counterparts

1. Introduction

Gravitational waves (GW) were first directly detected by the advanced Laser Interferometer Gravitational-wave Observatory (LIGO) with the event GW150914 [1]. The detection of GW opens a new, non-electromagnetic window through which to explore astrophysical phenomena. However, the search for electromagnetic (EM) counterparts to GW sources is essential if we are to fully realise the scientific potential of this new type of astronomy. Analysis of GW150914 indicates that the source of the signal was the inspiral and merger of two comparable-mass black holes (BH) [2], an EM counterpart is not expected from such a BH-BH merger [3]. As LIGO is upgraded to higher sensitivities [4], and additional detectors such as Virgo [5], and KAGRA [6] become operational, we can also expect detectable GW signals from the inspiral and merger of binary neutron star (NSNS) or neutron star-black hole (NSBH) systems [7-8].

The coalescence of NSNS or NSBH systems, when viewed on a rotational axis, are thought to be the origin of short gamma-ray bursts (SGRB) [8], and when viewed off axis, the source of kilo/macronova [9]. SGRB, along with long gamma-ray bursts associated with the core-collapse of massive stars [10], are instantaneously the most luminous objects in the Universe. Accretion onto a compact object is likely to power a collimated, relativistic outflow, where energy is initially dissipated by internal shocks which produce a prompt flash of gamma-rays. At later times the interaction between the outflow and the surrounding medium produces an external shock which generates a bright afterglow [11-12].

SGRB models rely on relativistic motion to explain the observed variability, duration, and energy of the prompt gamma-ray emission and optical afterglow. The precise details of the

acceleration, collimation, and energy content in astrophysical jets, such as those in SGRB, is a major focus of international research efforts. The jets of SGRB are assumed to be baryonic [13-14] with polarisation measurements indicating a contribution from magnetic fields [15-17]. Relativistic velocities and strong magnetic fields are features that are shared by jets associated with gamma-ray bursts, active galactic nuclei (AGN), and microquasars; by comparing the properties of these objects we gain an insight into the processes that govern the formation of relativistic jets [19-20].

Considering the well sampled jets from many different classes of AGN, a power-law distribution for the Lorentz factor (Γ) can be assumed of the form $N(\Gamma) \sim \Gamma^a$, where population synthesis studies indicate $a \sim$ -1.5 [21-22].

Applying what we know about other astrophysical jets to the outflow associated with NSNS mergers, we can maximise the potential EM counterparts that should be considered in follow-up searches to a GW trigger. In addition to the isotropic kilo/macronova and the on-axis SGRB, we propose the on-axis orphan-afterglow of failed GRB from a population of low Lorentz factor jets.

For jets from compact binary mergers with $\Gamma < 100$, the dissipative radius at which gamma-rays are emitted can be below the photospheric radius. Below the photosphere, the medium is optically thick and any prompt gamma-ray emission will become progressively thermalised and adiabatically cooled until the radiation is suppressed. The energy goes back into the outflow until the point where the mass of swept up ambient medium is equivalent to the initial mass in ratio to the outflow Lorentz factor, M / Γ. Here the outflow decelerates as $\Gamma \propto t^{-3/8}$, where t is time, and an external shock forms that produces a bright afterglow via synchrotron emission.

2. Commonly predicted EM counterparts
Here we briefly summarise the commonly predicted counterparts to a GW signal from a NSNS or NSBH merger.

2.1. SGRB
SGRB and their bright multi-wavelength afterglow are theoretically modelled by the rapid accretion of a post-merger, rotationally supported disk of material, onto a newly formed BH (or hyper-massive, rotationally supported, neutron star) that powers an ultra-relativistic and collimated outflow [23-26]. The prompt gamma-ray emission is relativistically beamed with a duration typically < 2 seconds, and a jet opening angle $\theta_j = 16\pm10^o$ [27]. For SGRB associated with a detectable GW signal from a NSNS/NSBH merger the EM counterpart will be detectable if the system is oriented towards an observer.

2.2. Kilo/macronova
Kilo/macronova are rapid-process radioactive-powered transients with isotropic weak supernova-like emission [28]. NSNS mergers are predicted to drive ejecta through dynamically ejected tidal tails, accretion disk neutrino driven winds, or thermonuclear winds [29-31]. The ejected material acts like a heat source and is predicted to peak in the optical or near-infrared with an absolute R-band magnitude of -14 to -17, although precise details are model dependent [32].

3. On-axis orphan afterglow from low Lorentz factor jets
Here we present a model for the on-axis orphan afterglow of failed GRB and the Monte Carlo generated distribution of EM counterparts to a NSNS merger that results in a detected GW signal.

3.1. Relativistic effects and the prompt gamma-ray emission
Relativistic motion is used to explain the optically thin nature of gamma-ray bursts [33] where the source becomes optically thin when $\Gamma > 100$. If there are baryons in a SGRB outflow, the optical depth at given radius R can be estimated by considering the photon scattering due to electrons associated with the outflow baryons, $\tau = \sigma_T E / (4\pi R^2 m_p c^2 \Gamma)$ where σ_T is the

Thomson cross-section, E is the outflow energy, m_p is the mass of a proton, and c the speed of light. The outflow will become optically thin at the photospheric radius,

$$R_s \sim 5.9 \times 10^{13} \, E_{51}^{1/2} \, \Gamma_1^{-1/2} \text{ cm.} \qquad (1)$$

Here we use the convention $Q = 10^x \, Q_x$ and cgs units. The variability timescale of the typical prompt emission, $\delta t \sim 0.1 \text{ seconds}$, can be used with the outflow Lorentz factor to estimate the dissipation radius where gamma-rays are emitted,

$$R_d \sim \Gamma^2 \, c \, \delta t \sim 3.0 \times 10^{11} \, \delta t_{-1} \, \Gamma_1^2 \text{ cm.} \qquad (2)$$

By requiring that the dissipative radius, equation (2), is greater than the photospheric radius, equation (1), gives $\Gamma > 80 \, E_{51}^{1/5} \, \delta t_{-1}^{-2/5}$. If this condition is not met then the prompt emission will be suppressed by adiabatic cooling of the radiation and efficient thermalisation [34-35]; the prompt luminosity is suppressed by the factor $(R_d / R_s)^{2/3}$, if the condition shown in equation (3) is met,

$$\tau \geq m_e \, c^2 / k_B \, T_{BB} \, . \qquad (3)$$

Here $\tau = (R_d/R_s)^2$, m_e is the electron mass, k_B is the Boltzmann constant, and T_{BB} the equivalent blackbody temperature. Additionally, if the condition in equation (3) is met, the peak spectral energy of the prompt emission will be $3\Gamma k_B T_{BB}$, where the energy above this peak exponentially decays. If the condition in equation (3) is not met, the photons will not be efficiently thermalised and the spectra is cut-off at an energy $> m_e \, c^2 \, (\Gamma/\tau)$.

3.2. On-axis orphan afterglow

Even if a jet does not meet the conditions for the emission of gamma-rays, the outflow will eventually collide into the ambient medium and form external shocks; this synchrotron shock radiation is well studied in the context of gamma-ray burst afterglow [36]. The relations for an on-axis orphan afterglow are identical to those for SGRB, the deceleration time of the outflow can be estimated as,

$$t_{dec} \sim 0.5 \, n_{-1}^{-1/3} \, E_{51}^{1/3} \, \Gamma_1^{-8/3} \text{ days.} \qquad (4)$$

Here n is the density of the ambient medium. The afterglow flux can be predicted in the context of the standard fireball model (see [37]).
The jet is considered to break when the sideways spreading is significant [38], this occurs when the Lorentz factor becomes less than the jet opening angle. The break time is,

$$t_b \sim 0.5 \, n_{-1}^{-1/3} \, E_{51}^{1/3} \, \theta_1^{8/3} \text{ days.} \qquad (5)$$

Another consideration is the self-absorption limit; this becomes important when we consider low-frequencies. Radio afterglow is an important component of any afterglow observations because they are long lived and often give additional information about the late time evolution of the outflow. Synchrotron self-absorption is a flux upper limit set by a blackbody with the shock temperature and given by the expression,

$$F_{BB} = 2\pi \, (p-2)/(p-1) \, e_e \, \Gamma^2 \, R_{size}^2 \, D^{-2} \, \nu_{obs}^2 \qquad \nu_a < \nu_m \, . \qquad (6)$$

Here p is the index of the electron Lorentz factor distribution, typically ~ -2.5, e_e is an efficiency parameter ~ 0.1, R_{size} is the apparent size of the fireball [39], D the distance, ν_{obs} the observational frequency, ν_a the frequency below which self-absorption becomes important, and ν_m the synchrotron characteristic frequency.

By assuming that all failed GRB outflow have an opening angle of 20°, we can estimate the jet break time with equation (5), and the limit on low-frequency emission by synchrotron self-absorption with equation (6). From these and the standard fireball afterglow flux estimates (see [37-38]) the peak flux and the peak time can be estimated. For NSNS mergers within 300 Mpc, the distribution of afterglow flux at various frequencies; *g*-band optical, radio at 150 MHz, and sub-mm at 10 GHz, and the corresponding peak time are shown in figure 1.

4. Gravitational wave beaming

Gravitational wave emission from a binary NSNS inspiral is isotropic but due to the superposition of the strain for the two polarisation modes, × and +, the signal is strongest on a rotational axis. A numerical solution for the fraction of events that are inclined towards an observer as a function of opening angle, given a GW detection and an isotropic distribution of mergers, is shown in figure 2. The analytical approximation in [40], and the fraction of events inclined towards an observer for a given opening angle if we consider no GW signal and an isotropic distribution are shown for comparison.
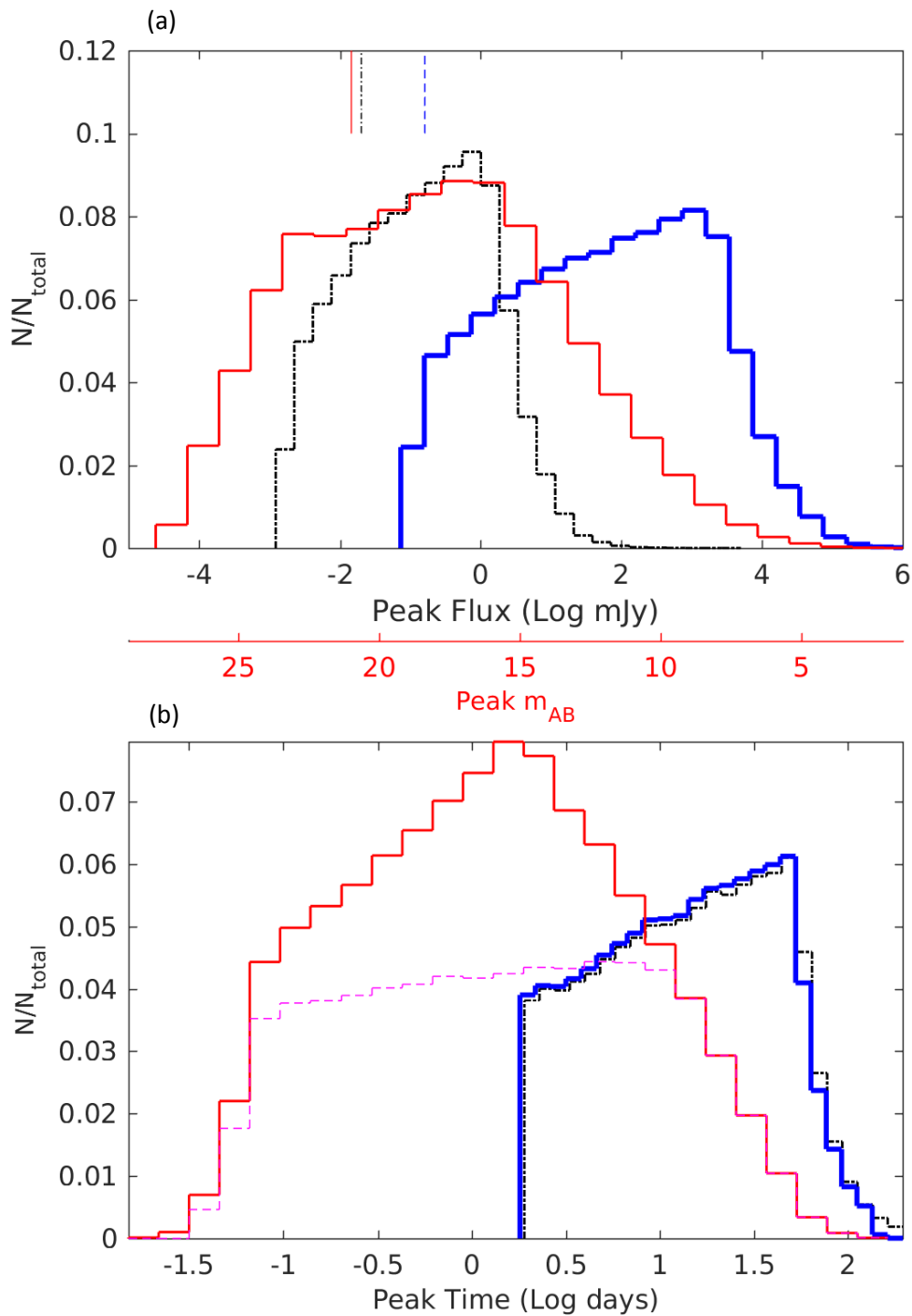
Figure 1: Peak flux/peaktime charts. The peak flux and time for the on-axis orphan afterglow of failed GRB within 300 Mpc. **(a)**: The flux distribution for on-axis orphan afterglow from a synthetic population of NSNS mergers within 300 Mpc. The peak optical flux in the *g*-band is the thin red line, the peak radio flux at 150 MHz is the dashed-dotted black line, and the peak sub-mm flux at 10 GHz is shown in thick blue line. Vertical lines represent sensitivity limits for various radio observatories; the blue dashed line is the sensitivity limit of 48 LOFAR stations at 150 MHz, the dashed-dotted black line is the sensitivity limit of SKA-Low, and the

redline is the limiting optical magnitude of 21. **(b)**: The peak flux time in days for the optical, red line, (optical flux brighter than 21st magnitude is shown as a dashed magenta line), the peak time for sub-mm is shown as a thick blue line, and radio a dashed-dotted line (governed by the jet break time).
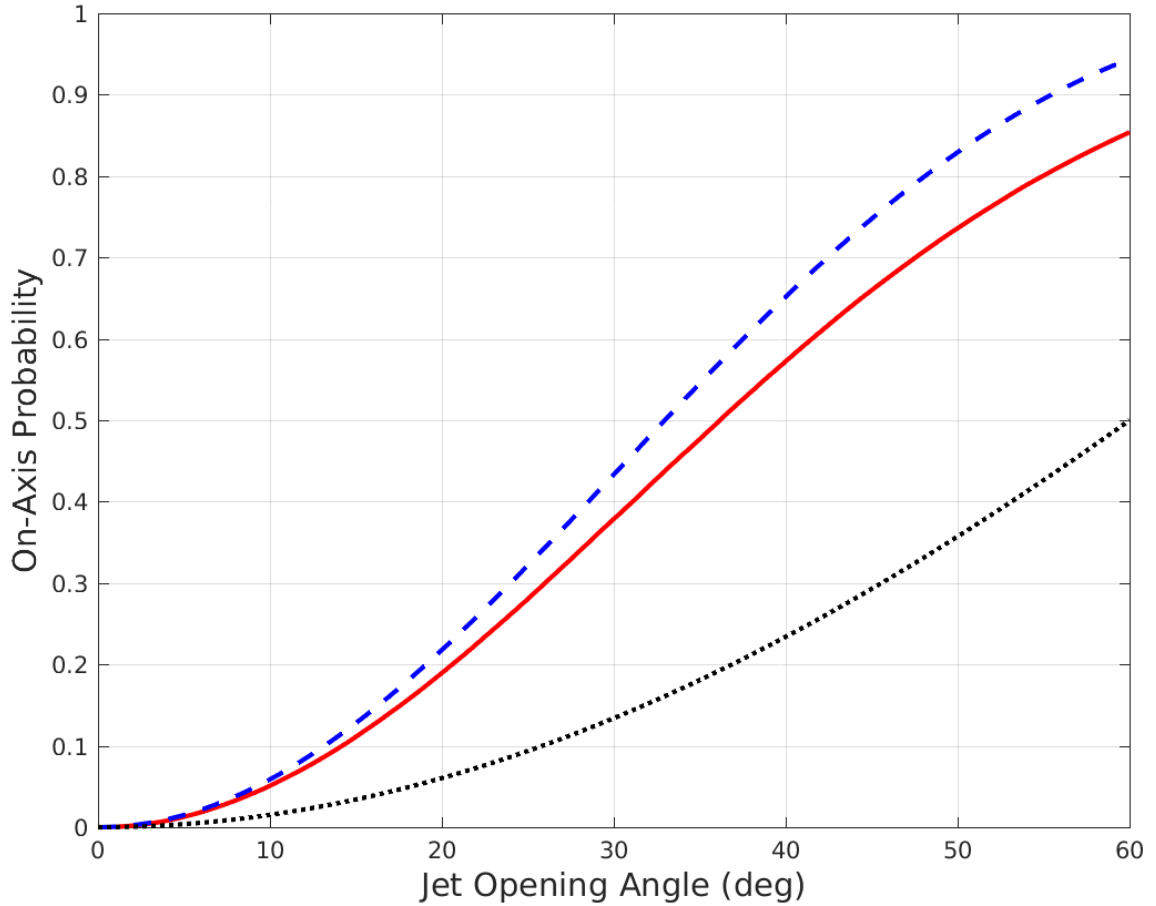
Figure 2: Plot of an axis probability against jet opening angle. For a given jet opening angle the fraction of all GW detected mergers that are inclined towards an observer for an isotropic population of GW detected mergers is shown via an analytical estimate from the results of [40], blue dashed line, and a numerical solution, red solid line. The fraction of the entire isotropic merger population, without GW detection, that is pointing towards the observer for a given opening angle is shown as a dotted black line.

5. Discussion and conclusion

By assuming that the jets of SGRB follow a similar power-law distribution of Lorentz factor as for other astrophysical jets, we have predicted a significant population of undetected failed GRB. A full Monte Carlo simulation predicts 65% of the population of mergers producing jets with a Lorentz factor in the range $3 < \Gamma < 1000$, where the distribution follows an index of -1.5, will result in failed GRB [41]. If the rates of GW detection from NSNS or NSBH binaries exceed that predicted by SGRB population synthesis, ~ 0.3 per year [42], then a hidden population of failed GRB is likely to exist.

EM counterpart searches for GW detection should consider all possible scenarios; SGRB, kilo/macronova, and the optical afterglow of a failed GRB. Within 300 Mpc the afterglow of failed GRB are shown in figure 1 to peak at ~ 18th magnitude with ~ 70% brighter than 21st magnitude. In addition to this optical peak at ~ 30 hours, the sub-mm and radio peaks at ~ 12 days offer additional opportunities to observe the EM counterpart for extended monitoring. Such optical counterparts are an ideal candidate for mid-sized robotic telescopes, such as the Liverpool Telescope (LT), and will be ~ 100% detectable by larger,

4m successor telescopes such as the New Robotic Telescope (LT2).
The detection/non-detection of a population of failed GRB will offer constraints on the Lorentz factor distribution of jets from stellar mass black holes and allow models for jet acceleration and collimation to be further tested. Multi-messenger astronomy offers a unique opportunity to further our knowledge of the fundamental mechanisms that govern the extreme limits of physics within the high-energy Universe.

Acknowledgments

References

[1] Abbott B P, et al. 2016 *Phys. Rev. Lett.* **116** 06112
[2] The LIGO and Virgo Scientific Collaboration 2016 *arXiv* 1602.03840
[3] Loeb A 2016 *ApJL.* **819** L21
[4] Harry G M, et al 2010 *Class. Quantum Grav.* **27** 084006
[5] Accadia T, et al. 2011 *Class. Quantum Grav.* **28** 114002
[6] Aso Y, et al. 2013 *Phys. Rev. D* **88** 043007
[7] Belczynski R E, et al. 2008 *ApJ.* **682** 474-486
[8] Berger E 2014 *ARA & A* **52** 43-105
[9] Metzger B D, et al. 2010 *MNRAS* **406** 2650-2662
[10] Prentice S J, et al. 2016 *MNRAS* **458** 2973-3002
[11] Piran T 2004 *Rev. Mod. Phys.* **76**, 1143-1210
[12] Zhang B, and Meszaros P 2004 *Int. Mod. Phys.* **19** 2385-2472
[13] Paczynski B 1986 *ApJ.* **308** 43-46
[14] Shemi A, and Piran T 1990 *ApJ.* **365** 55-58
[15] Steele I, et al. 2009 *Nature* **462** 767-769
[16] Mundell C, et al. 2013 *Nature* **504** 119-121
[17] Gotz D, et al. 2009 *ApJL.* **695** L208
[18] Yonetoku D, et al. 2011 *ApJL.* **743** L30
[19] Marscher A 2006 RELATIVISTIC JETS: the common physics of AGN, microquasars, and gamma-ray bursts *AIP Conf. Proc.* **856** 1
[20] Nemmen R S, et al. 2012 *Science* **338** 1445
[21] Lister M, and Marscher A 1997 *ApJ.* **476** 572-588
[22] Marscher A 2006 *PoS.* Proc. of the VI Microquasar Workshop: microquasars and beyond (MQW6) 025
[23] Ruffert M, et al. 1997 *A&A* **319** 122
[24] Rosswog S, et al. 2003 *MNRAS* **345** 1077
[25] Shibata M, and Taniguchi K 2008 *Phys. Rev. D* **77** 084015
[26] Rezzolla L, et al. 2011 *ApJL.* **732** L6
[27] Fong W, et al. 2015 *ApJ.* **815** 102
[28] Li L X, and Paczynski B 1998 *ApJL.* **507** L59
[29] Kulkarni S, 2005 *arXiv*:astro-ph/0510256
[30] Metzger B, et al. 2008 *MNRAS* **390** 781
[31] Dessart L, et al. 2009 *ApJ.* **690** 1681
[32] Nissanke S, et al. 2013 *ApJ.* **767** 124
[33] Piran T 1999 *Phys. Rept.* **314** 575
[34] Pe'er A, et al. 2005 *ApJ.* **635** 476-480
[35] Thomson C 2007 *ApJ.* **666** 1012-1023
[36] Kobayashi S, et al. 1999 *ApJ.* **513** 669
[37] Sari R, Piran T, and Narayan R 1998 *ApJL.* **497** L17
[38] Sari R, Piran T, and Halpern J 1999 *ApJ.* **519** L17
[39] Sari R 1998 *ApJ.* **494** L49-L50
[40] Kochanek C, and Piran T 1993 *ApJ.* **417** L17-L20
[41] Lamb G, and Kobayashi S 2016 (in prep)
[42] Metzger B, and Berger E 2012 *ApJ.* **746** 48

John Mackereth

Constraining models of Galactic disk formation with APOGEE and EAGLE: The vertical structure of mono-age populations in the Milky Way

J. Ted Mackereth
Astrophysics Research Institute, 146 Brownlow Hill, Liverpool, L3 5RF
E-mail address: J.E.Mackereth@2011.ljmu.ac.uk

Abstract. Robust models for the formation of galactic disks, and in particular, that of the Milky Way, are a fundamental element in our understanding of the processes which shape the cosmos. Galaxy formation models, in general, have a hard time fitting the numerous observational constraints on the formation of disk components. Using observational data for over 100,000 Milky Way stars, from the Sloan Digital Sky Survey (SDSS) Apache Point Observatory Galactic Evolution Experiment (APOGEE), alongside theoretical numerical simulations of the universe from the Evolution and Assembly of GaLaxies and their Environments (EAGLE) suite, we have a unique opportunity to confront revolutionary observational results with model predictions, to levy novel constraints on models of galactic disk formation. Specifically, we are studying the distribution of stellar chemical abundances and ages with position in the Milky Way, and in the simulated galaxies, which provide an 'archaeological' record of the processes which have taken place during their assembly. Reconstructing the history of galaxies in this way allows us to understand fundamental aspects of their evolution. This particular work presents a new look at the vertical structure of mono-age populations in the Milky Way disk using APOGEE data. We show that the vertical structure is such that the scale heights of mono-age populations smoothly bridge the thin – thick disk 'divide' between 200 and 600pc. We also demonstrate that flaring of disk populations is prevalent in young subsamples. These findings offer constraints for disk formation in the simulations, allowing further study to probe the formation of disks in EAGLE to understand the processes which result in this structure.

Keywords. Galaxy: Disk, Galaxy: Structure, Galaxy: Ages, Stellar Populations

1. Introduction

Galaxy disks form some of the most visually striking structures in the universe. The understanding of their structure is a fundamental element of galactic astronomy, as it can offer an insight into the processes which shape galaxies, and indeed the universe, on a global level. Both observational and theoretical data are now sufficiently advanced such that many of the questions regarding the formation and evolution of galactic disks, which have emerged in the last 50 years or so of research, can now, and are, being confronted.

The Sun's position in Milky Way offers us a unique viewpoint into the internal structure of a typical disk: we find ourselves right at the center. Many years of research, mainly focused on nearby stellar populations, have shown the disk to have a highly complex structure, both spatially and kinematically (e.g. Bovy et al. 2012a,b,c; Haywood et al. 2013),

and as a function of chemical abundances (e.g. Nidever et al. 2014; Hayden et al. 2015) and stellar age (e.g. Haywood et al. 2015). However, working under the assumption that the processes which have lead to these distributions are in effect across most -- if not all -- disk galaxies, we may be able to use MW results as a 'Rosetta stone' for galaxy evolution (see Freeman & Bland-Hawthorn 2002; Rix & Bovy 2013).

The vertical structures of disks are an important diagnostic of their evolutionary history. Early work on external galaxies revealed that vertical luminosity profiles of edge-on disks are such that there is an excess over the single exponential expected if there existed a single, thin disk (Burstein 1979; Tsikoudi 1979). This was later confirmed to be the case in the Milky Way also (Yoshii 1982; Gilmore & Reid 1983). These vertically extended disk components, separate to the galactic halo, are now thought to arise by a variety of competing processes, both internal and external, namely: a) radial migration, explored by Sellwood & Binney (2002) as the redistribution of stellar orbits whilst conserving the vertical angular momentum, invoked to explain the lack of an age—metallicity relation at the solar radius (Edvardsson et al. 1993), b) galaxy mergers, which can both heat the disk dynamically (e.g. Quinn et al, 1993), and also deposit star-forming gas onto it (in gas-rich mergers, see e.g. Brook et al. 2004) and also stars from disrupted satellites, whose orbital inclinations were coincident with the disk prior to merging (e.g. Abadi et al. 2003) and c) that the disk formed upside down in-line with its current thickness (e.g. Bird et al. 2013).

Many of these processes are considered in the context of the formation of a distinct 'thick disk' component, but whether this is actually the case is the subject of debate (Bovy et al. 2012a, 2015; Haywood et al. 2015). Recent results suggest that there may exist a significant divide in chemistry between the disk components (Haywood et al. 2015), but the divide is less clear in spatial structure, and mono-abundance populations (MAPs) seem to bridge the divide, at least in terms of their vertical scale height (Bovy et al. 2012b, 2015, hereafter B15). The analysis in B15 considers that all MAPs are well described by a single vertical exponential by invoking a radially dependent, exponential flaring profile. Under these considerations, it is possible to imagine a scenario where a thick disk is in fact a superposition of many thinner components (as suggested by e.g. Bovy et al. 2012a).

Another important diagnostic of disk formation lies in the understanding of the radial structure. Study of metallicity gradients in the MW disk has shown that they are generally negative, with variations in slope as a function of height above the midplane and $[\alpha/M]$ (Hayden et al. 2014). The metallicity distribution functions (MDFs) in disk radial bins are skewed negatively (positively) at smaller (larger) galactocentric radii (Hayden et al. 2015), which suggests that radial migration is playing an important role in defining the MW disk radial structure (Loebman et al. 2015). A sequence of high $[\alpha/M]$ stars appears homogeneous throughout the disk, which suggests early disk formation from well mixed, molecular gas (Nidever et al. 2014).

The ability of theory to disentangle the processes outlined above is an essential factor in the extended goals of this work. Recent advances in the methods and technology used for simulating galaxies has lead to a revolution of sorts, where hydrodynamical, cosmological (and zoom-in) simulations are beginning to reproduce the key observables of the universe (see e.g. Schaye et al. 2015), and realistic galaxies with extended disk components (e.g. Stinson et al. 2013, Martig et al. 2014). These improvements are due, in part, to the more complete inclusion of radiative feedback in the sub-grid physics of the simulations, which seems to be an important factor in achieving the level of cooling required to form extended stellar disks (see e.g. Brook et al. 2012, but also discussed in Schaye et al. 2015, Crain et al. 2015, in the context of EAGLE).

2. Data
We exploit data from the Apache Point Observatory Galactic Evolution Experiment (APOGEE) survey (Majewski et al. 2015) and theoretical data from the Evolution and Assembly of GaLaxies and their Environments (EAGLE) suite of cosmological, hydrodynamical simulations (Schaye et al. 2015, Crain et al. 2015) to achieve our overarching goals. This dual approach gives an interesting insight into the MW in a

cosmological context, allowing us to probe exactly which processes are dominant in the formation of disks and what the present day structure of the MW disk can tell us about the history of the galaxy. This paper concentrates on results based on APOGEE data, with a view to future work with EAGLE shedding light on the results found.

2.1 APOGEE

SDSS/APOGEE (Majewski et al. 2015) is a high resolution (R ~ 20,000) high signal-to-noise (SNR < 100 per pixel) spectroscopic survey in the near-infrared (NIR) H-band (1.5-1.7μm), targeting red-giant stars in all components of the Milky Way disk. Targets are selected from the 2MASS point source catalogue (Skrutskie et al. 2006), for which the full details of can be found in Zasowski et al. (2013). APOGEE data are generally observed using the 2.5m Sloan Foundation telescope (Gunn et al. 2006), and processed via an automated pipeline (Nidever et al. 2015). The processed spectra are then analysed by the APOGEE Stellar Parameters and Chemical Abundances Pipeline (ASPCAP, García Pérez et al. 2015) to measure stellar parameters and up to 17 chemical abundances. The data for DR12 were then analysed and calibrated via a separate process (Holtzman et al. 2015). This work primarily concerns a subset of the DR12 data which is outlined below.
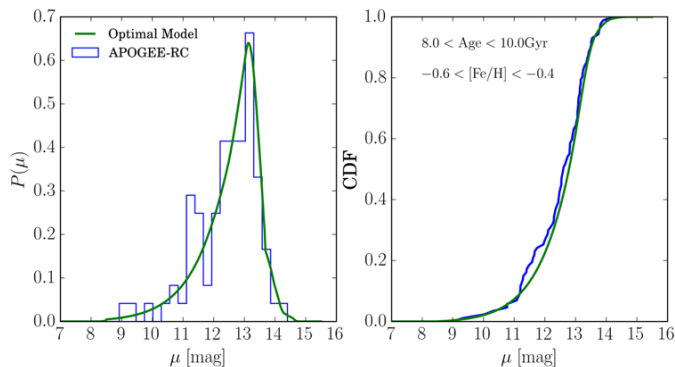
2.2 The APOGEE-RC Catalogue

The APOGEE stars considered here make up a sample of red-clump (RC) stars defined in the DR12 APOGEE-RC catalogue (see Bovy et al. 2014 for a complete description and characterisation). The RC describes a feature of the HR diagram where stars of a given
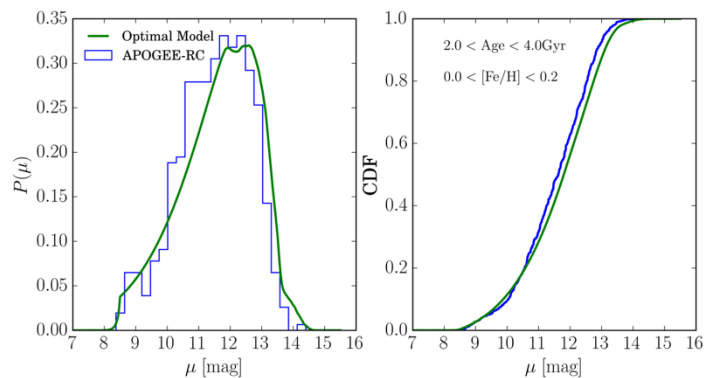
Figure 1 – Comparison between the distance moduli distributions predicted by the best fit model and from the observed data for representative subsets binned by age and [Fe/H]. In both cases the CDF is well modelled. The older subset (top) has smaller number of stars at 88 whereas the younger subset (bottom) has 542 stars. It is clear that the fit is good in both cases, if slightly noisier at lower N.

mass evolve slightly more slowly, creating a 'clump' of stars in the red-giant branch (RGB) for which we then know the absolute magnitude and therefore can make precise distance estimates. After cuts which are outlined fully in B15, a highly pure (~95%) sample of 14,699 RC stars is defined, for which distance estimates are precise to 5%, and for which a detailed reconstruction of the sample selection function can be made. The density fitting methodology used here is extinction corrected through the use of 3D extinction maps of the Milky Way by Marshall et al. (2006), covering the inner-disk and Green et al. (2015) elsewhere. The use of this extinction map is motivated by the exploration of different maps in B15, which demonstrated that these maps gave the best fit to the data.

2.3 Ages for APOGEE DR12

Recent work on abundance analysis has shown that the stellar

Figure 2 – The distribution of the stars in our sample in [Fe/H]—age space, colour coded by [α/Fe]. The overlaid grid demonstrates the binning applied. Most bins are well sampled, with >15 stars in each, although some bins are not uniformly sampled in age.
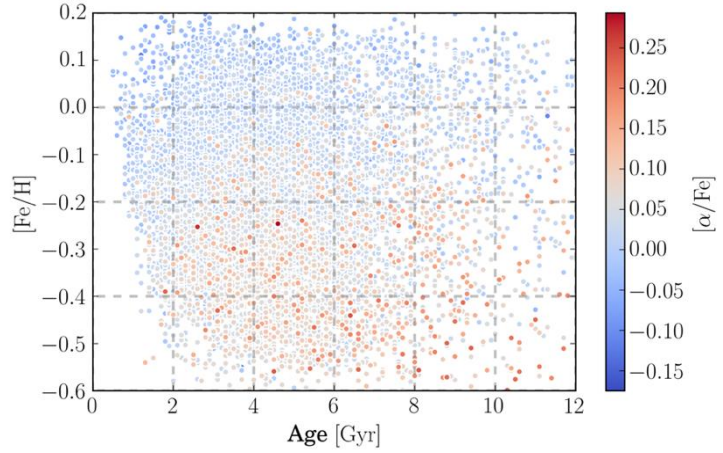
[C/N] ratio is a good proxy for mass (and therefore age), as stars ascending the red giant branch undergo deep mixing to their cores which are enriched in C, N and O by the CNO nucelosynthesis undergone in their main sequence lifetime. As a result of this Martig et al. (2015) (and also, via a novel, data-driven approach, dubbed *The Cannon,* Ness et al., 2015) exploited the [C/N] ratios measured using APOGEE spectra to fit a model for age, generating a catalogue of ages for the majority of the DR12 sample. We cross-match the ages estimated by Martig et al. (2015) with the cleaned sample of RC stars, which gives 10,914 stars in the statistical RC sample with measured ages.

3. Density Fitting

3.1 Method

We model the observed rate of stars over the parameter space of position, magnitude, colour and metallicity using a Poisson process, following methodologies outlined in Bovy et al. (2012, 2015, hereafter B15) and more generally in Rix & Bovy (2013). The code used for this work is based heavily on the publicly available code written for B15 [4]. Best-fit parameters are determined by optimising a likelihood function for the parameters θ (Equation 2 of B15), which is a function of only the density of the observed data points given the density model (parameterised by θ) and the survey effective volume (which is a property of the whole survey given θ). Final parameter and uncertainty estimates are determined via a Markov Chain Monte Carlo (MCMC) sampling of the posterior PDF of the best fit solution (Foreman-Mackey et al. 2015). Following B15, we find that broad sub-samples of stars binned by age and metallicity are well fit by a broken exponential radially, and vertically by a single exponential with an exponentially varying scale height as a function of radius. As such the density is fit with the model:

$$\nu_*(R,\phi,Z) = \Sigma(R)\zeta(Z) = \exp(-h_R^{-1} (R-R_0) + h_Z^{-1}\exp(R_{flare}^{-1}[R-R_0])|Z|) \qquad (1)$$

---

[4] Found at https://github.com/jobovy/apogee-maps

Where h_R is a scale length, which takes a different value before and after R_{peak} (a broken exponential), h_z is a scale height which is varied by the flaring profile, described by R_{flare}^{-1}, which takes a negative value for a scale height which increases with R. These quantities form the set of parameters θ which are fit using the maximum likelihood method above. A comparison between the observed data and a best-fit model for stars with 2<Age<3 Gyr and -0.2<[Fe/H]<0.0 dex and for stars with 10<Age<12 Gyr and -0.6<[Fe/H]<-0.4 dex is shown in Figure 1. This shows the quality of the fit in two of the fine bins used in this analysis, demonstrating that this modelling technique is informative even in bins with low numbers, and where the contamination between bins may be greater than desired in a perfect scenario.
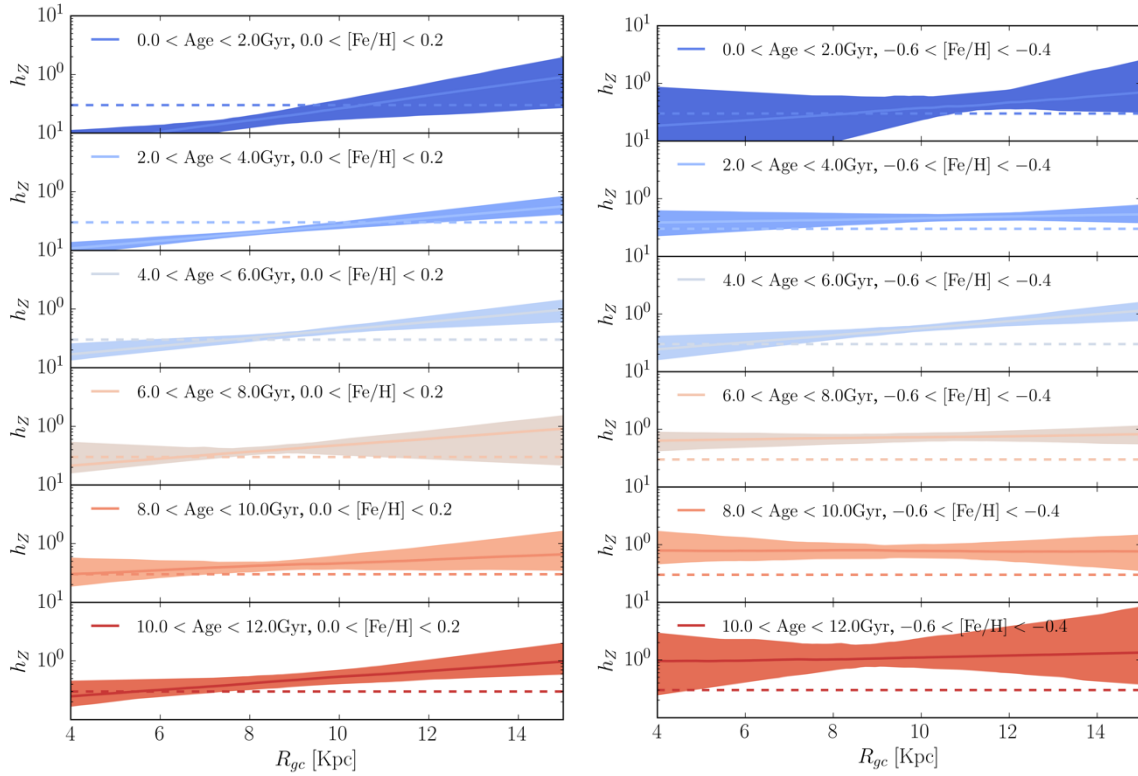
Figure 3- The scale height (h_z) profiles for low and high [Fe/H] stars at different ages. The coloured bands give the 95% uncertainty in the model. It can be seen here that the fit is less well constrained for the low age, low [Fe/H] bin, as a result of the low number of stars. The dashed line is drawn at 300pc to give a feel for the relative normalisation of each profile. Flaring appears stronger at higher [Fe/H] and, at both [Fe/H], lower age. The global thickness of mono age populations appears to be an increasing function of age.
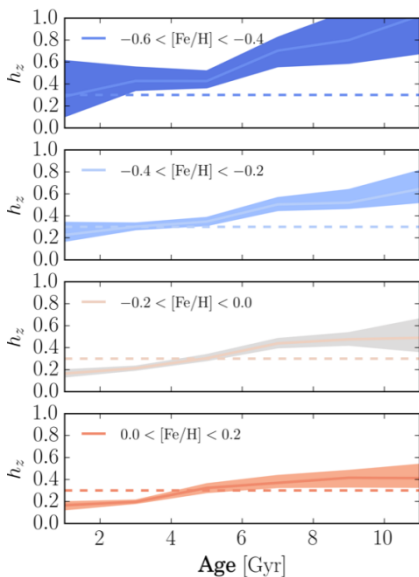
We bin by both age and [Fe/H] when making the fits, to attempt to disentangle any degeneracy between age and metallicity. We show the bins in Figure 2, overlaid on the distribution of stars in age-[Fe/H] space, colour coded by [□/Fe]. It can be seen that the sampling in most bins is sufficient to provide an informative fit, however some bins are non-uniformly distributed in age. We emphasize that this does not affect the quality of fit, but must be taken into account when considering trends in the results.

Figure 4 – the variation of the radius independent component of h_z as a function of age for each [Fe/H] bin. The 300pc dashed line is plotted for reference only. The thickness of populations seems to increase smoothly as a function of age, with increasing amplitude at lower [Fe/H].

3.2 The vertical disk structure

We defer a full discussion of the *radial* surface density profiles of the age-[Fe/H] subsamples to future work, and instead focus here on the *vertical* structure as a function of radius in the fitted models for the samples. Figure 3 shows the scale height as a function of radius, for a slice along the age axis at the highest metallicity bin and the lowest (0.0<[Fe/H]<0.2 dex, and -0.6<[Fe/H]<-0.4 dex). These bins are representative, and are not chosen to match any physical 'tracks' in [Fe/H]—Age space (although this sort of analysis is planned in future work). Regardless of this fact, the general trends between the two sets may hold some insight into the evolution of disk populations.

A notable result, which is seen in both the low and high [Fe/H] bin is the smooth increase in the normalisation of each profile (i.e. the non exponential h_z^{-1} component of the scale height) with age. In both cases, the oldest population is thicker globally, with a scale height greater than 300pc across the majority of the radial range. Figure 4 shows this behaviour better, isolating the fitted h_z for each sequence. It is clear in this plot that for every [Fe/H] bin, the scale height increases with age. It is also worth noting the variation in the amplitude of this increase, as a declining function of [Fe/H].

The second feature of interest in the plots is the level of flaring (the exponentially varying component in the scale height). Each of the high [Fe/H] relations of Figure 3 show noticeable flaring, which appears to have roughly the same gradient as a function of age, with perhaps a slight upturn at lower age. However, in the low [Fe/H] slice, the gradient appears to change with population age, becoming steeper at ~5Gyr, and flattening again at large age. This is qualitatively consistent with the results of B15, which show that high $[\alpha/\text{Fe}]$ sub-samples (shown, for our sample, concentrated to the lower right of Figure 2, consistent with an old, metal poor population) have no significant flaring, and that low $[\alpha/\text{Fe}]$ populations flared more strongly at lower [Fe/H], however a better comparison of these results requires a detailed understanding of the correlations between these three parameters, which will be explored in the complete analysis.

4. Discussion

4.1 Scale Height Transitions

Regardless of the deeper understanding needed of the covariances between $[\alpha/\text{Fe}]$, [Fe/H] and age, we can still make some basic inferences from these (preliminary) results. It seems, at least from the result in Figure 4, that there are no obvious breaks in the h_z – age relation, as might be expected in a dual disk, upside down formation scenario where a thick disk abruptly ceased forming stars, and star formation resumed in a thin disk. Instead, this result suggests that if the disk did form upside down, then the shift from thick to thin was roughly continuous, consistent with theoretical results from e.g. Bird et al. 2013. These results are also consistent with a disk thickening scenario, whereby stars are formed in a thin disk and become heated by some process such as those outlined in the introduction. However, simulation results suggest that radial migration does little for disk thickening (e.g. Minchev et al. 2012)

Until we can ascertain a better model for how the galaxy evolves as a function of (Age, [Fe/H], $[\alpha/\text{Fe}]$), and the radial surface density profiles as a function of these parameters are more well understood, it is difficult to be sure exactly which populations are being traced here. For example, in the analysis of B15, the scale height profiles were studied for tracks in the $[\alpha/\text{Fe}]$—$[\text{Fe/H}]$ plane, which is relatively well understood in terms of galactic chemical evolution. The picture is less clear when relating to age, as the correlation between it and chemical abundances is complex and strongly dependent on the star formation history of the Galaxy.

4.2 The h_z–R_{gc} relation

The results shown in Figure 3 are perhaps more informative at this early stage. It is clear that the mechanism by which a flaring profile arises is acting on timescales <2Gyr, as a clear flare is seen in both 0<Age<2Gyr bins. The flaring is stronger in the high [Fe/H] population. If flaring is a natural result of radial migration, then it perhaps makes sense that the higher

[Fe/H] population is more flared. If a basic model for star formation is assumed, with a radially decreasing rate, such that stars formed in the inner disk have higher [Fe/H] (more enrichment), and are then pushed out by radial migration and flare, whereas the low [Fe/H], outer disk population migrate inwardly and do not flare, then these results may indicate the result of extensive radial migration, particularly in light of very recent simulation results (Grand et al. 2016).

5. Conclusions & Future Work

We have demonstrated here that the density fitting methodology of B15 is fully applicable to APOGEE-RC samples selected by Age and [Fe/H], however careful consideration of the covariant nature of the parameter space of $(\text{Age},[\text{Fe/H}],[\alpha/\text{Fe}])$ must be made. We have made several findings so far. In particular:

- We showed that the assumption of a broken exponential radial and exponentially flaring vertical density profile is sound, given the data, and can be fit to a good level of agreement with the data.
- It is clear that h_z increases with age in all cases considered, and seems to be a relatively smooth transition from ~200pc, to ~600pc, a similar range observed by B15. The relation appears to flatten as a function of [Fe/H]. This suggests that the disk either formed thick and became thin (e.g. Bird et al. 2013), or is heated at all radii by some process which acts on a long timescale (e.g. Minchev et al 2012; Vera-Ciro et al. 2014).
- Flaring in young populations is significant, and appears to roughly decline with age. This seems inconsistent with flaring coaused by a long timescale process such as radial migration, and could suggest that disk populations in the outer disk are heated by external processes such as minor mergers, as suggested by various studies (Quinn et al. 1993; Abadi et al. 2003a,b; Brook et al. 2004).

It is difficult to make clear inferences from these results, which must also be considered in the context of the radial structure to give a full picture. A sure result is the consistency between these findings and B15 in terms of the smooth transition of h_z between ~200pc to ~600pc as the disk gets older. This, considered also in the context of the results of Bovy et al. (2012a), suggests that, structurally at least, the disk is not dual, but a continuum of disks with a smooth variation in thickness.

The ability for simulations to disentangle the processes which have resulted in the observed structure now becomes important. Future work will carry out a similar analysis on a subsample of Milky Way like EAGLE galaxies. This should inform us as to which aspects of the observed Milky Way disk structure are due to large scale external processes such as mergers and which are due to internal, secular evolution such as radial migration. This understanding will make major constraints on models of disk formation. EAGLE also simulates the chemical evolution of galaxies, and so will provide insights into how such processes effect the disk structure as a function of all the parameters considered here: age and abundances such as $[\alpha/\text{Fe}]$ and [Fe/H].

References
Abadi M et al. 2003 Simulations of Galaxy Formation in a Λ Cold Dark Matter Universe. II. The Fine Structure of Simulated Galactic Disks *The Astrophysical Journal* **597(1)** 21
Abadi M et al. 2003 Simulations of Galaxy Formation in a Λ Cold Dark Matter Universe. I. Dynamical and Photometric Properties of a Simulated Disk Galaxy *The Astrophysical Journal* **591(2)** 499

Bird J et al. 2013 Inside out and Upside down: Tracing the Assembly of a Simulated Disk Galaxy Using Mono-age Stellar Populations *The Astrophysical Journal* **773** 43

Bovy J et al. 2014 The APOGEE Red-clump Catalog: Precise Distances, Velocities, and High-resolution Elemental Abundances over a Large Area of the Milky Way's Disk *The Astrophysical Journal* **790** 127.

Bovy J, Rix H-W & Hogg D 2012a The Milky Way Has No Distinct Thick Disk *The Astrophysical Journal* **751** 131

Bovy J et al. 2016 On Galactic Density Modeling in the Presence of Dust Extinction *The Astrophysical Journal* **818(2)** 130

Bovy J et al. 2012b The Vertical Motions of Mono-Abundance Sub-Populations of the Milky Way Disk *The Astrophysical Journal* **755(1)** 1115

Bovy, J. et al. 2012c The Spatial Structure of Mono-Abundance Sub-Populations of the Milky Way Disk *The Astrophysical Journal* **753(2)** 148

Brook C et al. 2012 MaGICC discs: Matching observed galaxy relationships over a wide stellar mass range *Monthly Notices of the Royal Astronomical Society* **424(2)** 1275

Brook C et al. 2004 The Emergence of the Thick Disk in a Cold Dark Matter Universe *The Astrophysical Journal* **612(2)** 894

Burstein D 1979 Structure and Origin of S0 Galaxies. III. The Luminosity Distribution Perpendicular to the Plane of the Disks in S0's *The Astrophysical Journal* **234** 829

Crain R et al. 2015 The EAGLE simulations of galaxy formation: calibration of subgrid physics and model variations *Monthly Notices of the Royal Astronomical Society* **450** 1937

Edvardsson B et al. 1993 The Chemical Evolution of the Galactic Disk – Part One – Analysis and Results *Astronomy and Astrophysics* **275** 101

Foreman-Mackey D et al. 2013 emcee : The MCMC Hammer *Publications of the Astronomical Society of the Pacific* **125(925)** 306

Freeman K & Bland-Hawthorn J 2002 The New Galaxy: Signatures of Its Formation *Annual Review of Astronomy and Astrophysics* **40(1)** 487

García-Pérez A et al. 2015 ASPCAP: The Apogee Stellar Parameter and Chemical Abundances Pipeline *ArXiv e-prints*

Gilmore G & Reid N 1983 New light on faint stars. III - Galactic structure towards the South Pole and the Galactic thick disc *Monthly Notices of the Royal Astronomical Society* **202** 1025

Grand R et al. 2015 Vertical disc heating in Milky Way-sized galaxies in a cosmological context *Monthly Notices of the Royal Astronomical Society* **Advance Access**

Green G et al. 2015 a Three-Dimensional Map of Milky Way Dust *The Astrophysical Journal* **810(1)** 25

Gunn J et al. 2006 The 2.5 m Telescope of the Sloan Digital Sky Survey *The Astronomical Journal* **131(4)** 2332

Hayden M et al. 2015 Chemical Cartography With Apogee: Metallicity Distribution Functions and the Chemical Structure of the Milky Way Disk *The Astrophysical Journal* **808(2)** 132

Hayden M et al. 2014 Chemical Cartography With Apogee: Large-Scale Mean Metallicity Maps of the Milky Way Disk *The Astronomical Journal* **147(5)** 116

Haywood M et al. 2015 Clues to the formation of the Milky Way's thick disk *Astronomy & Astrophysics* **579** A5.

Haywood M et al. 2016 When the Milky Way turned-off the lights: APOGEE provides evidence of star formation "quenching" in our Galaxy *ArXiv e-prints*

Haywood M et al. 2013 The age structure of stellar populations in the solar vicinity *Astronomy & Astrophysics* **560** A109

Holtzmann J et al. 2015 Abundances, Stellar Parameters, and Spectra from the SDSS-III/APOGEE Survey *The Astronomical Journal* **150** 148

Loebman S et al. 2016 Imprints of Radial Migration on the Milky Way'S Metallicity Distribution Functions *The Astrophysical Journal* **818(1)** L6

Majewski S et al. 2015 The Apache Point Observatory Galactic Evolution Experiment (APOGEE) *ArXiv e-prints.*

Marshall D, Robin A, Reylé C, Schultheis M & Picaud S 2006 Modelling the Galactic interstellar extinction distribution in three dimensions *Astronomy & Astrophysics* **453(2)** 635

Martig M et al. 2015 Red giant masses and ages derived from carbon and nitrogen abundances *Monthly Notices of the Royal Astronomical Society* **456** 3655

Martig M, Minchev I & Flynn C 2014 Dissecting simulated disc galaxies - I. The structure of mono-age populations *Monthly Notices of the Royal Astronomical Society* **442(3**) 2474

Martig M, Minchev I & Flynn C 2014 Dissecting simulated disc galaxies - II. The age-velocity relation. *Monthly Notices of the Royal Astronomical Society* **443(3)** 2452

Minchev I et al. 2012 Radial migration does little for Galactic disc thickening *Astronomy & Astrophysics* **548** A127

Ness M et al. 2015 Spectroscopic determination of masses (and implied ages) for red giants *ArXiv e-prints*

Nidever D et al. 2015 the Data Reduction Pipeline for the Apache Point Observatory Galactic Evolution Experiment *The Astronomical Journal* **150(6)** 173

Nidever D et al. 2014 Tracing Chemical Evolution Over the Extent of the Milky Way's Disk With Apogee Red Clump Stars *The Astrophysical Journal* **796(1)** 38

Quinn P, Hernquist L & Fullagar D 1993 Heating of galactic disks by mergers *The Astrophysical Journal* **403** 74

Rix H-W & Bovy J 2013 The Milky Way's stellar disk: Mapping and modeling the Galactic disk *Annual Reviews in Astronomy and Astrophysics* **21**

Roškar R, Debattista V & Loebman S 2013 The effects of radial migration on the vertical structure of galactic discs. *Monthly Notices of the Royal Astronomical Society* **433(2)** 976

Schaye J et al. 2015 The EAGLE project: simulating the evolution and assembly of galaxies and their environments *Monthly Notices of the Royal Astronomical Society* **446** 521

Sellwood J & Binney J 2002 Radial mixing in galactic discs *Monthly Notices of the Royal Astronomical Society* **336(3)** 785

Skrutskie M et al. 2006 The Two Micron All Sky Survey (2MASS) *The Astronomical Journal* **131** 1163

Stinson G et al. 2013 MaGICC thick disc - I. Comparing a simulated disc formed with stellar feedback to the milky way *Monthly Notices of the Royal Astronomical Society* **436(1)** 625

Tsikoudi V 1979 Photometry and structure of lenticular galaxies. I - NGC 3115 *The Astrophysical Journal* **234(1)** 842

Vera-Ciro C et al. 2014 the Effect of Radial Migration on Galactic Disks *The Astrophysical Journal* **794(2)** 173

Yoshii Y 1982 Density Distribution of Faint Stars in the Direction of the North Galactic Pole *Publications of the Astronomical Society of Japan* **34** 365

Zasowski G et al. 2013 Target Selection for the Apache Point Observatory Galactic Evolution Experiment (Apogee). *The Astronomical Journal* **146(4)** 81

Benjamin Mummery

Resolving degeneracy between AGN feedback and neutrino free streaming as structure formation suppression mechanisms in λCDM

Benjamin Mummery, MPhys
Astrophysics Research Institute
B.O.Mummery@2010.ljmu.ac.uk

Abstract. The hierarchical formation paradigm of λCDM with constraints based on high redshift observations overpredicts the degree of observed structure. Both baryonic feedback from AGN and the free streaming of massive neutrinos have been invoked to resolve this discrepancy, however the magnitude of their effects are not well constrained. Using the BAHAMAS and Cosmo-OWLS suites of hydrodynamic cosmological simulations we investigate the physical impact of both mechanisms on large scale structure for a physically motivated range of values for AGN heating and the summed mass of the active neutrino species. For dark matter haloes with masses M200 between 10^{12} and 10^{16} $M_\odot$ in at redshifts between 0 and 3, we investigate the abundance and mass-concentration relation of the halo population. While degenerate at fixed redshift and mass bin, the effects of AGN feedback and neutrino free streaming exhibit markedly different redshift and mass dependences, suggesting that cosmological observations may be used to constrain neutrino and AGN properties.

Keywords. Computational cosmology, etc.

1. Introduction

In the beginning the universe was created. This has made a lot of people very angry and has generally been regarded as a bad move [1]. The growth of structure in the universe since then has been dominated by dark matter (DM), material which makes up 84.5% of the total matter in the universe and interacts only through gravity. Starting from an almost uniform distribution, dark matter clumps together under its own gravity, eventually collapsing into small spherical 'halo' distributions. These haloes subsequently accrete additional DM from their surroundings and merge with other haloes, sequentially forming larger haloes. The gravitational potential wells of these haloes is the environment in which the more familiar 'baryonic' matter forms the stars and galaxies that we observe.

'Geometric' cosmological observations originating at high redshift z, such as the Cosmic Microwave Background (CMB) anisotropy, probe the background cosmology. In the λCDM cosmological model, these observations can be used to constrain the fundamental cosmological parameters. In contrast, non-geometric measurements such as observations

of galaxies and clusters originate at low z and directly probe the matter content of universe and its structure at late times.

Combinations of observations from these categories can be used to break degeneracies between parameters or to provide an independent test of proposed constraints. However, in general predictions from geometric observations require a greater degree of 'clumping' of matter than is indicated by non-geometric observations. A particular example is the overprediction of the number of galaxy clusters by a factor of ~2 compared to observations [2].

Simulation	UV/ X-ray Background	Cooling	Star Formation	Vsn	AGN Feedback	$\log(\Delta T_{heat})$ [Log(k)]	n_{heat}	□	$\Sigma M\nu$ [eV]	M_{DM} 10^9 h^{-1} $M_\odot$	$M_{bar,initial}$ 10^9 h^{-1} $M_\odot$
DMONLY	No	No	No	-	No	-	-	No	-	3.75	-
NOCOOL	Yes	No	No	-	No	-	-	No	-	3.75	7.54
REF				-	No	-	-				
AGN 8.0		Yes	Yes	600		8.0					
AGN 8.5					Yes	8.5	1				
AGN 8.7						8.7					
NU 0.00 (DM)	Yes (No)	Yes (No)	Yes (No)	300 (-)	Yes (No)	7.8 (-)	20 (-)	Yes	0.00	3.85	7.66 (-)
NU 0.06 (DM)									0.06		
NU 0.12 (DM)									0.12		
NU 0.24 (DM)									0.24		
NU 0.48 (DM)									0.48		

Table 1 - Summary of subgrid parameters for Cosmo-OWLS and BAHAMAS simulations

Resolving this discrepancy requires a reduction of the number of predicted haloes. The two primary mechanisms by which this may be achieved are: 1) The inclusion of non-gravitational physics not treated by the linear growth model, and 2) a modification of the underlying cosmological model. The leading contenders for each of these cases are 1) active galactic nuclei (AGN), and 2) the free streaming of one or more active neutrino species.

In each case the strength of the expected effect depends on parameters that are not yet well constrained by observation. Furthermore, since both mechanisms are capable of reducing the total galaxy cluster counts to the observed level, their effects on the cluster population must be at least partially degenerate. We therefore explore the results of a series of simulations with varying parameters for both mechanisms in order to build a physical understanding of the behaviour and interaction of these mechanisms.

This paper is organised as follows: section 2 discusses the BAHAMAS and Cosmo-OWLS simulation suites utilised for this work; sections 3 and 4 provide a brief overview of the relevant AGN and neutrino physics respectively; section 5 presents the results of the investigation; and section 6 summarises our conclusions.

2. Simulations

The Cosmo-OWLS and BAHAMAS simulation suites used here form successors to the Overwhelmingly Large Simulations (OWLS) suite of large, hydrodynamic simulations. Runs in both cases evolve a $(400\ h^{-1}Mpc)^3$ periodic comoving box containing 2×1024^3 particles to $z=0$.

2.1 Cosmo-OWLS

Cosmo-OWLS explores the effect of varying sub-grid physics at fixed cosmology. This suite is therefore well suited to the investigation of the effects of baryonic feedback, including AGN. From the cosmological parameters, initial conditions are computed using N-GenIC1 with the Transfer Function computed from the prescriptions of Eisenstein and Hu [3]. Cosmological parameters are from WMAP7 [4,5].
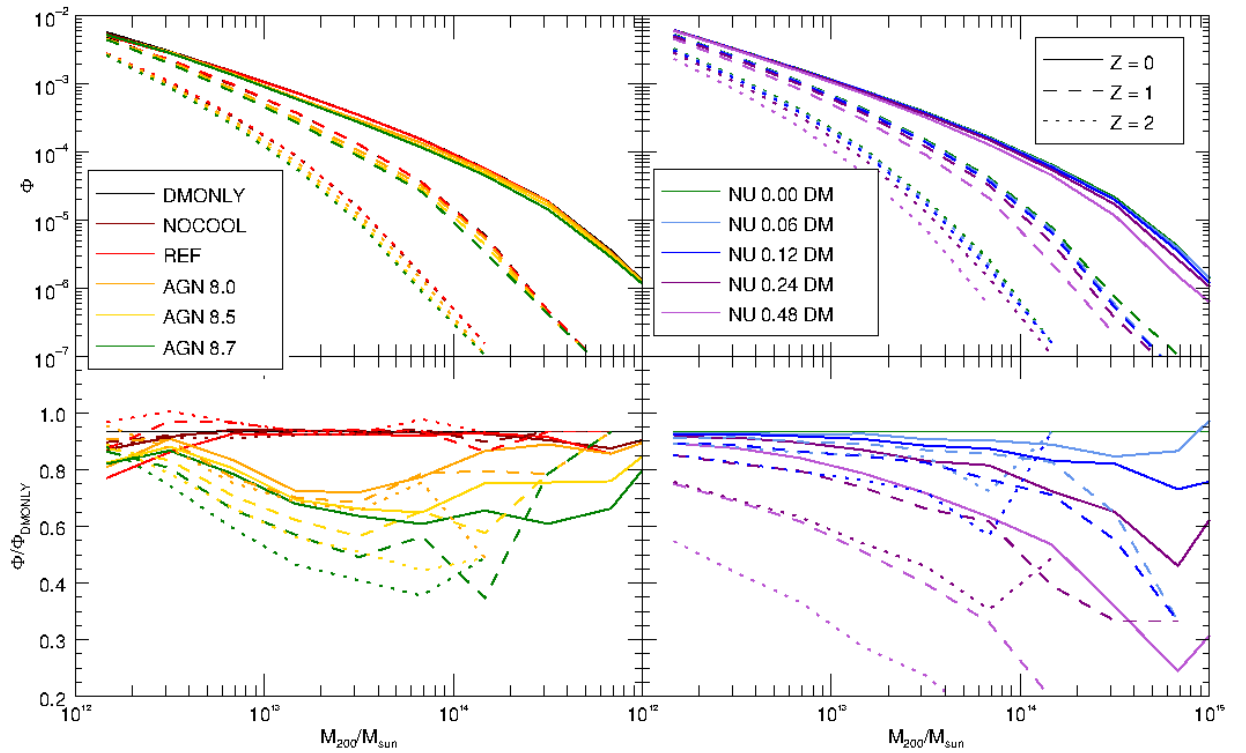
Figure 1- HMFs for varying baryonic physics model (left) and varying neutrino mass (right). Lower panels display the data normalised to the massless neutrino, DM-only case

For this work, the results of six Cosmo-OWLS simulations are used: DMONLY, in which all matter is represented by collisionless particles. This simulation is as a baseline to which to compare the effects of baryonic physics. NOCOOL incorporates the UV/X-ray background according to the prescription of Haardt and Madau [6]. To this, REF adds radiative cooling of gas and stochastically implemented star formation as well as SN feedback implemented using Dalla Vecchia and Schaye's [7] kinetic wind model.

Black holes (BH) are incorporated using Booth and Schaye's [8] subgrid prescription: haloes with of $M_{FOF}>10^{11.6}\ h^{-1}M_\odot$ are seeded with BH sink particles which may then grow via Bondi-Hoyle-Lyttleton accretion and mergers with other BHs. AGN feedback is implemented by storing a fraction ε of the accreted rest mass energy within the BG until sufficient energy is available to heat n_{heat} randomly chosen surrounding particles by a predetermined quantity ΔT_{heat}. Cosmo-OWLS runs take ε=0.015 and n_{heat}=1. The three AGN simulations vary the value of ΔT_{heat} between $10^{8.0}$ and $10^{8.7}$ K.

2.2 BAHAMAS

In contrast to Cosmo-OWLS, the BAHAMAS suite fixes the sub-grid physics and varies cosmology. These runs utilise second order Linear Perturbation Theory rather than the first-order Zel'dovich approximation, and compute transfer functions using the Code for Anisotropies in the Microwave Background [9]. Since the high thermal velocities of neutrinos cause their clustering to remain linear up to late times, their distribution can be computed analytically rather than simulate additional particles.

BAHAMAS implements neutrino physics according to the prescription of Ali-Haımoud and Bird [10], using the full Cold Dark Matter and baryonic mass gravitational potential as a source to compute the neutrino overdensity on the fly, and incorporating this potential back into the particle calculations. The mutual gravitational influence of the neutrino 'field' and the DM therefore influences their simultaneous evolution. For this work, a set of five BAHAMAS runs are utilised wherein the summed neutrino mass ΣMv is varied. Initial cosmological parameters other than ΣM_v are from WMAP9. Four of these runs simulate massive neutrinos with $0.06 < \Sigma M_v < 0.48$ eV. In the fifth, ΣM_v is set to 0 and the simulation is functionally identical to the DMONLY case discussed in §2.1 with the exception of the initial cosmological parameters.

A further set of five runs are used with the same values for ΣM_v alongside the full set of baryonic physics from §2.1. The subgrid physics have been adjusted so as to best reproduce the observed galaxy stellar mass function and hot gas mass fractions of groups and clusters in the absence of neutrinos. These models, alongside those discussed in §2.1, are summarised in table 1.

3. Active Galactic Nuclei

A supermassive black hole (SMBH) is present in the core of every known galaxy. In cases where material is accreting onto the SMBH, the increase in density and velocity leads to the heating of the surrounding medium, resulting in material with very high kinetic energies. Such material can be ejected from the SMBH system with

Figure 2 – Comparison of the multiplicative ''prediction' to the simultaneous hydrodynamic realisation of both effects

sufficient velocity to escape the gravity well of the DM halo itself, reducing the overall mass and lowering the density. This decreases the rate of accretion of matter onto the halo, slowing its growth and effectively partially 'un-collapsing' collapsed structures and thereby decreasing the number of haloes at high masses [11]. The strength of the effect of AGN feedback depends strongly on the subgrid model. We therefore explore a physically motivated range of values for ΔT_{heat}.

4. Neutrinos

Particle physics experiments, as well as astronomical observations, have confirmed the existence of massive neutrinos. Since these constitute a fraction of dark matter which remains relativistic to late times and therefore does not undergo collapse. Overdensities at scales smaller than the neutron free streaming length therefore collapse later. This leads to a reduction in the number counts of cluster mass structures which are still collapsing at z=0 and are therefore sensitive to small changes in collapse time.

The strength of this effect is closely tied to the value of ΣM_v. Atmospheric and solar neutrino oscillation experiments provide a lower limit of ΣM_v=0.06 eV, with astrophysical considerations imposing an upper limit of ~1 eV [12].

5. Results

We investigate the impact of both AGN feedback and neutrino free streaming on three properties of the halo population: The halo mass function (HMF), integrated cluster counts and mass concentration relation (c(M)). In each case we present the results of AGN feedback and neutrino free streaming in isolation before testing a simple multiplicative treatment of the two effects against the full hydrodynamic simulations simultaneously treating both.
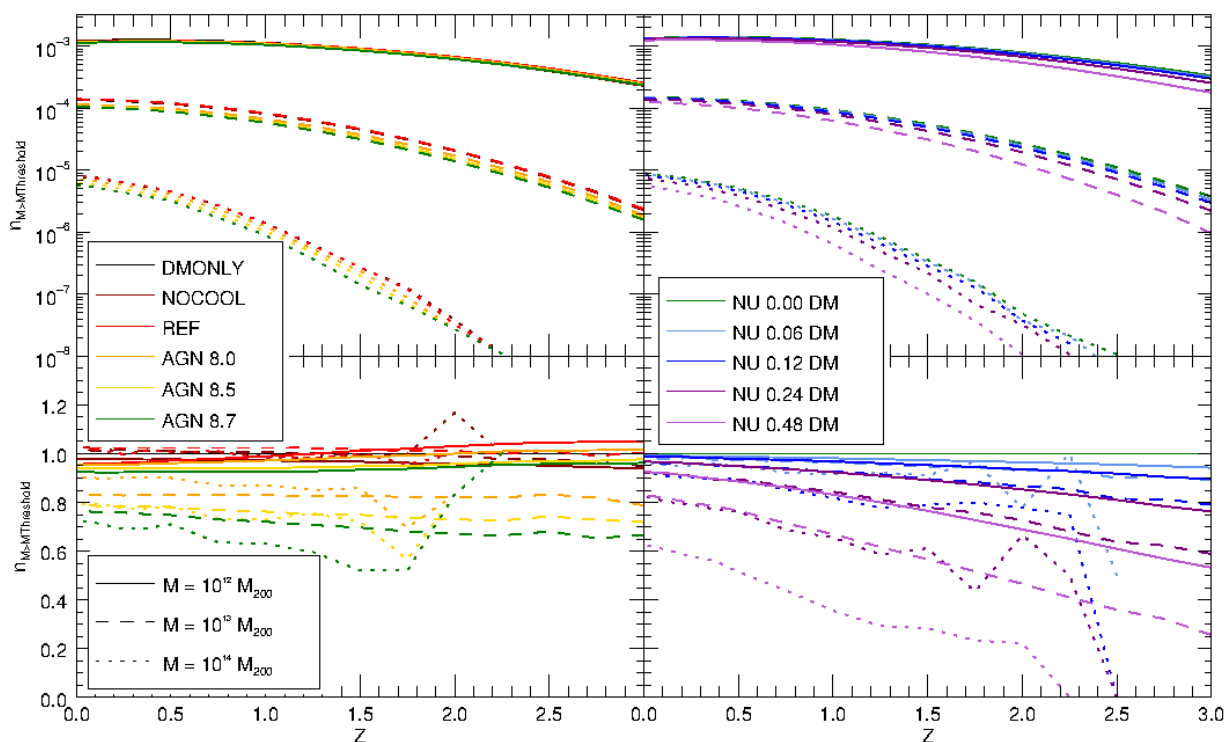
Figure 3 - integrated cluster counts above a threshold mass as a function of redshift for varying baryonic physics (left) and neutrino mass (right).

5.1 The Halo Mass Function

Haloes identified in Cosmo-OWLS and BAHAMAS runs were binned into 10 equally spaced logarithmic mass bins between 10^{12} and 10^{15} $M_{\odot}$, and the resulting counts converted into

number densities per logarithmic unit mass, $\Phi = \frac{dn}{dlog(M)}$. For each run, this process was carried out at snapshots corresponding to z=0, 1 and 2. A side by side comparison of the effects of baryonic feedback and neutrino free streaming on the HMF is given in Figure 1. The upper panel in each case displays Φ as a function of the halo mass, while the lower panels show the data normalised to the relevant DM only sim. Mass ejected from the cluster by AGN feedback results in a $\approx 20 - 30\%$ suppression of the HMF at masses of $\sim 10^{14}\,M_\odot$, similar to the findings of Velliscig et al. [13]. At high masses the gravity wells of haloes become increasingly deep, eventually reaching a situation where baryonic feedback is incapable of ejecting material. Consequently, the behaviour tends back towards the NOCOOL case with the location of the maximum suppression of the HMF varying as a function of the heating efficiency such that sims with a greater value for ΔT_{heat} exhibit peak suppression at higher halo masses

	A	B
NOCOOL	4.878	-0.099
REF	4.65	-0.112
AGN 8.0	4.105	-0.061
AGN 8.5	3.917	-0.073
AGN 8.7	3.842	-0.071
NU 0.00 DM	4.553	-0.072
NU 0.06 DM	4.498	-0.07
NU 0.12 DM	4.411	-0.074
NU 0.24 DM	4.329	-0.07
NU 0.48 DM	4.055	-0.068

Table 2 – Best fit parameters for equation 2

As discussed in §4, Neutrino free streaming preferentially suppresses the high mass end of the HMF. The strength of this suppression varies as a function of ΣM_ν. While the baryonic feedback suppression is largely independent of redshift, a much more significant variation is exhibited by neutrino free streaming. This difference, along with the differences in mass dependence at high masses, offer potential avenues by which the degeneracy of the two mechanism may be broken. We test the assumption that the two mechanisms may be treated multiplicatively by plotting the multiplicative 'prediction' from the separate simulations against the combined simulation results. As can be seen in figure 2, the discrepancy between these treatments is <5% across 3 dex in mass.

5.2 Halo Cluster Counts
The integrated cluster counts above a given threshold mass provide a more direct link to observable quantities than the HMF to which it is closely related. It also provides an opportunity to explicitly evaluate the redshift evolution of the effects of baryonic feedback and neutrino free streaming on the formation of structure. To this end, haloes satisfying $M_{200} > M_{threshold}$ at each value for z were enumerated. Three threshold masses, 10, 10^{13} and $10^{14}\,M_\odot$, were chosen to be loosely analogous to the detection limits for optical, x-ray and S-Z observations respectively.

Following the structure of §5.2, Figure 3 compares the effect of baryonic feedback and neutrino free streaming on the integrated cluster counts number density. Linestyles here denote different $m_{threshold}$ values. The z dependence of the suppression of cluster counts above a given mass due to the two mechanisms vary markedly: while suppression due to baryonic feedback is largely independent of z, the effect of neutrino free streaming increases by a factor of ≈ 3 between z = 0 and 3. The multiplicative test is not carried out for the integrated counts since, as it is the integral of the HMF, good agreement here can be inferred from the good agreement of figure 2.

5.3 Mass Concentration Relations

The internal structure of DM haloes is strongly dependent on the formation history and behaviour of dark matter. While difficult to measure with current techniques, this nonetheless presents a valuable test of our understanding of the formation of structure [14].

As both the baryonic feedback and neutrino free streaming mechanisms affect the time at
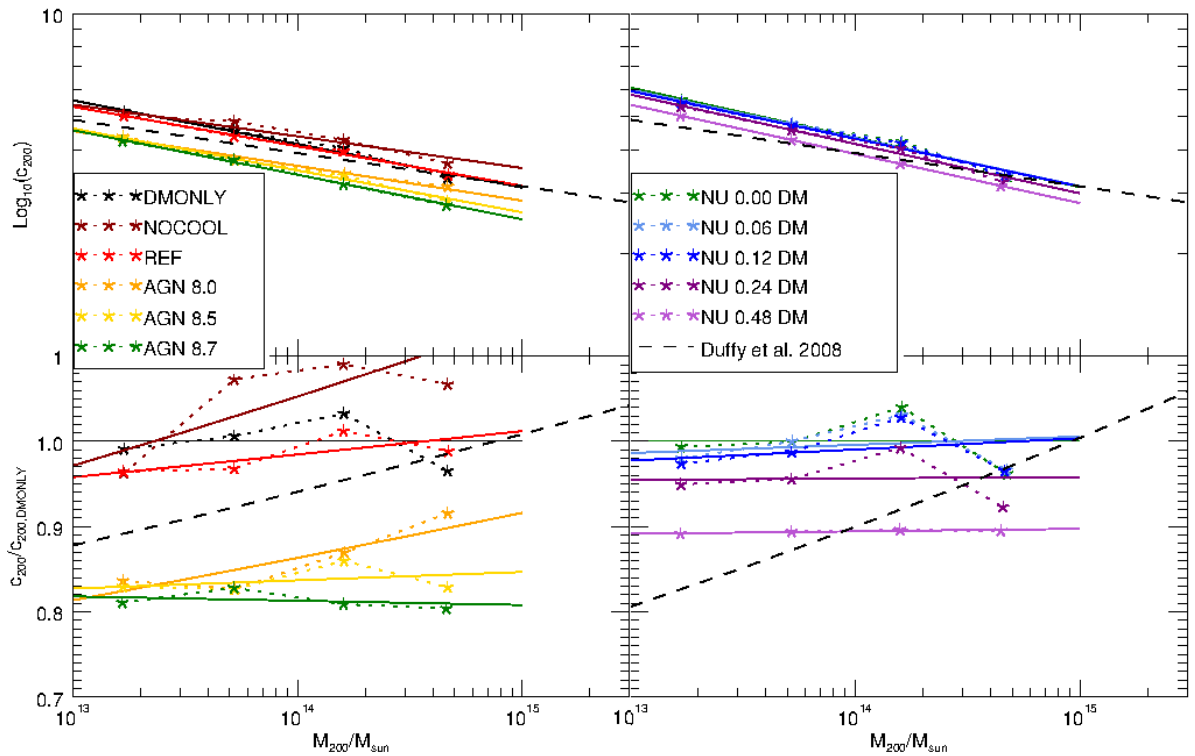
Figure 4 - mass-concentration relations. Stars denote the mean concentration in 0.5 dex mass bins, solid lines display the best fit of equation 2 in each case

which a halo collapses as a function of its mass, an additional outcome is their effect on the internal structure of haloes. Since the central density of a collapsed halo depends on the universal critical density at the time of collapse, haloes affected by either mechanism are expected to have lower central densities for their masses than equivalent haloes in a DM-only universe. This can be parameterised by the concentration parameter c. This parameter is defined as:

$$c_\Delta = \frac{r_\Delta}{r_s}$$

Equation 1

where r_s arises from the Navarro, Frenk and White (NFW) density profile. Navarro et al. [15] found that the densities of DM haloes in N-body simulation were well fitted by a function which ranges from $\rho \propto r^{-1}$ to r^{-3}. The scale radius, r_s, is the radius at which the profile is parallel to $\rho \propto r^{-2}$. The radius r_Δ is the radius within which the mean density of the halo is a factor Δ greater than the universal critical density. We adopt $\Delta=200$ throughout. As the

central density depends on the time of collapse, and therefore the mass, there is a resultant relation between c and the mass of the halo. The alteration of the collapse time as a function of the mass by baryonic feedback or neutrino free streaming will therefore alter the mass concentration relation, c(M).

Neto et al. [16] found that the distribution of mass and concentration values for DM-only haloes in N-body simulations at z = 0 was well fitted by a power law of the form

$$c_\Delta(M_\Delta) = A \left(\frac{M_\Delta}{M_{Fiducial}} \right)^B$$

Equation 2

In order to investigate the effects of neutrino free streaming and AGN feedback on c(M) we construct 1-dimensional radial density profiles for all haloes with masses M200 > 10^{13} M$_\odot$. Values for c are derived in each case by least-squares fitting NFW profiles to the 1D density profiles between $0.1 < r < 1$ R$_{200}$. Mean c_{200} values were then calculated in 6 0.5-dex mass bins between 10^{13} and 10^{16} M$_\odot$.

Figure 4 displays the best fit c(M) relations from equation 2 for different baryon physics models in the absence of neutrino physics in WMAP7 cosmology, and for different ΣM_ν values in the absence of baryonic physics in WMAP9 cosmology at z = 0. The coefficients for these fits are given in table 2. Neutrino free streaming acts primarily to reduce the amplitude of c(M), with a

Figure 5 – best fit mass concentration relations for the simultaneous simulations as well as the multiplicative 'prediction'. The lower panel shows the simultaneous result normalised to the prediction in each case

neutrino mass of ΣM_ν=0.48 eV suppressing the amplitude by a factor of ≈ 13% relative to the DM-only ΣM_ν=0 case. The effect on the gradient is much smaller, of order 5%. Conversely, baryonic feedback alters both the amplitude and gradient by ≈ 20 when fitting to the total matter density profile. As might be expected, c(M) fits to the combined simulations, shown in Figure 5, vary the amplitude by a comparable factor, but exhibit only a ≈ 10% change in gradient relative to nu 0.00 dm. This is predominantly due to baryonic feedback as the gradient variation when varying ΣM_ν is much smaller.

As shown in figure 5, while the multiplicative treatment captures well the amplitude offset of the c(M) relation, it fails to accommodate the change in gradient. This stands in contrast to the good agreement shown in figure 2, implying that bulk properties of the halo population may be treated multiplicatively, but the effect of AGN feedback and neutrino free streaming on the internal structure of individual haloes requires a more in depth analysis.

6. Summary and Conclusions

By exploring the physical impacts of AGN feedback and neutrino free streaming on a simulated halo population we have demonstrated that, while degenerate at a fixed redshift or mass bin, the two effects exhibit markedly different dependences on both of these parameters, allowing the two to be deconvolved by means of observations at multiple z and mass values. This provides a promising avenue into using cosmological observations to further constrain the physical parameters of the mechanisms.

Simulations incorporating both baryonic and neutrino physics are well described by a simple multiplicative treatment of the effects for the bulk properties of the population – the halo mass function and integrated counts. This supports the work of for example Velliscig et al. [13] in developing fitting functions to analytically describe the impacts of baryonic physics. However, the internal structure of the haloes, parametrised by the mass-concentration relation, shows less robust agreement with an evident offset in gradient between the two cases. Investigation of these properties therefore requires a fully hydrodynamic simulation incorporating both neutron and bayonic physics.

7. References

[1] Adams, Douglas. *The Hitchhiker's Guide to the Galaxy*. New York: Harmony, 1980. Print.

[2] Planck Collaboration. Planck 2013 Results XX. Cosmology from Sunyaev-Zel'dovich Cluster Counts. A$A, 517(20), 2014.

[3] Eisenstein, Daniel J., and Wayne Hu. "Power Spectra for Cold Dark Matter and Its Variants." *ApJ The Astrophysical Journal* 511.1 (1999): 5-15.

[4] G. Hinshaw, D. Larson, E. Komatsu, D. N. Spergel, C. L. Bennett, J. Dunkley, M. R. Nolta, M. Halpern, R. S. Hill, N. Odegard, L. Page, K. M. Smith, J. L. Weiland, B. Gold, N. Jarosik, A. Kogut, M. Limon, S. S. Meyer, G. S. Tucker, E. Wollack, and E. L. Wright. Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results. ApJS, 208:19, October 2013. doi: 10.1088/0067-0049/208/2/19.

[5] Amandine M. C. Le Brun, Ian G. McCarthy, Joop Schaye, and Trevor J. Ponman. Towards a realistic population of simulated galaxy groups and clusters. MNRAS, 441(2):1270–1290, 2014.

[6] Francesco Haardt and Piero Madau. Radiative transfer in a clumpy universe: Ii. the utraviolet extragalactic background. ApJ, 461:20, 1995.

[7] C. Dalla Vecchia and J. Schaye. Simulating galactic outflows with kinetic supernova feedback. MNRAS, 387:1431–1444, July 2008. doi: 10.1111/j.1365-2966.2008.13322.x.

[8] C. M. Booth and Joop Schaye. Cosmological simulations of the growth of supermassive black holes and feedback from active galactic nuclei: method and tests. MNRAS, 398(1):53–74, 2009.

[9] Ian G. McCarthy, Joop Schaye, Simeon Bird and Amandine M. C. Le Brun. The BAHAMAS project: Calibrated hydrodynamical simulations for large-scale structure cosmology. In prep.

[10] Y. Ali-Haımoud and S. Bird. An efficient implementation of massive neutrinos in non-linear structure formation simulations. MNRAS, 428:3375–3389, February 2013. doi: 10.1093/mnras/sts286.

[11] I. G. McCarthy, J. Schaye, R. G. Bower, T. J. Ponman, C. M. Booth, C. Dalla Vecchia, and V. Springel. Gas expulsion by quasar-driven winds as a solution to the overcooling

problem in galaxy groups and clusters. MNRAS, 412:1965–1984, April 2011. doi: 10.1111/j.13652966.2010.18033.x.

[12] J. Lesgourgues and S. Pastor. Massive neutrinos and cosmology. Physics Reports, 429:307–379, July 2006. doi: 10.1016/j.physrep.2006.04.001.

[13] Marco Velliscig, Marcel P. Van Daalen, Joop Schaye, Ian G. McCarthy, Amandine M. C. le Brun Marcello Caccioto, and Claudio Della Vecchia. The impact of galaxy formation on the total mass, mass profile and abundance of haloes. MNRAS, 442(3):2641–658, 2014.

[14] J. F. Navarro, E. Hayashi, C. Power, A. R. Jenkins, C. S. Frenk, S. D. M. White, V. Springel, J. Stadel, and T. R. Quinn. The inner structure of ¤CDM haloes - III. Universality and asymptotic slopes. MNRAS, 349:1039–1051, April 2004. doi: 10.1111/j.13652966.2004.07586.x.

[15] Julio F. Navarro, Carlos S. Frenk, and Simon D. M. White. Simulations of x-ray clusters. MNRAS, 275(3):720–740, 1995.

[16] Angelo F. Neto, Liang Gao, Phillip Bett, Shaun Cole, Julio F. Navarro, Carlos S. Frenk, Simon D.M.White, Volker Springel, and Adrian Jenkins. The statistics of λcdm halo concentrations.

Rhana Nicholson

Supernova enrichment of planetary systems in unusual star clusters

Rhana Nicholson & Richard Parker
Astrophysics Research Institute,
IC2 Science Park,
146 Brownlow Hill,
L3 5RF
E-mail address: R.B.Nicholson@2011.ljmu.ac.uk

Abstract. The presence and abundance of short lived radioisotopes (SLRs) ^{26}Al and ^{60}Fe in chondrites means the Sun probably formed in the vicinity of one or more massive stars that exploded as supernovae. Massive stars are more likely to form in massive clusters (>1000 M_{sun}), however the dynamical effects from stellar interactions and photo evaporation from massive stars could be detrimental to the formation of planets. We investigate whether low mass clusters containing one or more massive stars may be the better environment for creating Sun like stars due to being dynamically more quiet.

We analyse the results of N body simulations that follow the evolution of low-mass star clusters that contain several massive stars with a variety of initial conditions. We find that two-body relaxation occurs before the massive stars in the cluster have time to explode as supernovae (SNe). On average ~10 per cent of stars in the cluster are sufficiently enriched, resulting in a lower population of polluted stars from unusual low mass star clusters than high mass star clusters. Even though most enriched stars are unperturbed, the numbers that are enriched are insignificant to the amount produced by high mass clusters. We find that it is unlikely that these unusual low mass clusters represent the environment in which the Sun formed.

1. Introduction

Stars do not form in isolation, but rather in groups varying in number from 10^1 to 10^6 stars (Lada & Lada, 2003)[1]. These star-forming regions have observed densities ranging from 10^0 to 10^6 stars per cubic parsec (pc^3) (Bressert et al, 2010)[2], much higher than the typical galactic density of stars (~0.1 stars/pc^3). Because of this, stars observed in the field, such as the Sun, must have originated in these denser star-forming regions and defused into the galactic disk through processes that are currently poorly understood. Planets and stars are observed to form concurrently (Haisch et al, 2001)[3]. The density of the region where a star forms and the dynamical interactions that a star undergoes can therefore greatly affect planet formation. Stars that are born in regions with more than 100 stars per pc^3 can expect to undergo significant dynamical interactions with other stars in the first 10 million years (Myr) of their lives (Parker, 2014)[4], whereas stars in low density regions may undergo no close interactions. Adverse effects from such interactions can include the destruction or truncation of protoplanetary discs or direct interactions between planets and passing stars (Scally & Clarke 2001, Parker & Quanz, 2012, de Juan Ovelar et al 2012)[5,6,7].

Initial stellar masses vary greatly, with observed masses ranging from ~0.1-200 solar masses and even higher mass stars predicted theoretically. The initial mass function (IMF) describes the initial distribution of masses for a population of stars (Bastian et al, 2010)[8]. Stars greater than 20 solar masses (M_{sun}) are rare and evolve rapidly, in comparison to G type stars (such as the Sun), which are common and have lifetimes of billions of years; each undergoes a very different evolutionary path.

One of the fundamental and outstanding questions in astrophysics is whether our Solar System is a typical planetary system in terms of chemical composition and its dynamically quiet past (Adams, 2010)[9]. Short-lived radioactive species (SLRs) with half-lives less than 10 Myr, such as ^{26}Al and ^{60}Fe, have been detected in chondrites (Marhas et al, 2008)[10]. Chondrites are a type of meteorite dating from the very earliest stages of planet formation and hence are thought to reflect the initial conditions of the solar nebula. These radioactive isotopes appear to be one of the main causes of heating and melting the Earth's interior during its formation and thus are important for physical processes such as plate tectonics. From observations of the interstellar medium (ISM), the Solar system appears to contain an over abundance of these SLRs, although whether an abundance of ^{60}Fe exists and to what degree is still under debate. Because of radio isotopes' short half-lives and homogeneous distribution in the Solar System (Villeneuve et al. 2009)[11], these isotopes must have entered the protoplanetary disc at early epochs.

SLRs can form through several mechanisms, though not all are viable methods for polluting nebulae. Cosmic ray spallation can form ^{26}Al in large enough quantities but cannot produce ^{60}Fe (Lee et al. 1998; Shu et al. 2001)[12,13]. SLRs are also produce in asymptotic giant branch (AGB) stars (Wasserburg et al, 1994, Trigo-Rodríguez et al, 2009)[14,15], however AGB stars are not associated with star-forming regions and a chance encounter between them is very unlikely (Kastner & Myers, 1994)[16]. The most probable scenario is that these SLRs were produced in the cores of massive young stars with initial masses > 20 M_{sun} that polluted the protoplanetary disc when they exploded as supernovae (SNe) (Adams, 2010)[9].

There are two methods by which SNe can pollute protoplanetary discs. The first is through direct pollution, where the forming star is close enough to the SN explosion to receive sufficient enrichment at the time of explosion. However, the region the star needs to inhabit around the SN explosion is small and thus this method has been regarded as improbable and convoluted. At distances smaller than 0.1pc the planetary disc could be severely disrupted or destroyed by the explosion; greater than 0.3pc and the disc would not be sufficiently enriched. The other possible mechanism occurs when a SN pollutes its surrounding environment and triggers star formation, causing stars to form in pre enriched material. The latter method has been analysed using N-body and hydrodynamical simulations (Parker and Dale, 2016)[17] and found to be just as unlikely, if not more so, due to the very specific conditions necessary. Direct pollution requires that the the Sun formed near a high mass star (>20 M_{sun}) and so favours large clusters (>1000 stars) as the birth environment of the Sun. However, stars in massive clusters are likely to undergo significant dynamical interactions and large clusters may in fact be inherently hostile to planet formation. Additionally, N-body simulations of direct pollution in massive clusters have shown that the number of G-dwarfs that are enriched and unperturbed is ~0.5 – 1 per cent (Parker et al, 2014)[18].

N-body simulations of young clusters have shown that close interactions can affect planetary systems dramatically; either liberating planets from their host stars or significantly altering the orbits of the planets, leaving the system dynamically unstable (Parker & Quanz, 2012)[19]. Observations show all planetary orbits lie close to the plane and have small eccentricities (Adams, 2010, Adams et al, 2014)[9,20], confirming there have not been any major stellar interactions in the history of the Solar System, which one might expect from a star born in a dense star cluster. Another detrimental effect is photo evaporation of proto planetary discs due to radiation from massive stars. Combining these two damaging effects, it seems unlikely that the Sun was born in a high mass cluster.

The number of observed clusters has a strong relationship with cluster mass. Low mass clusters (50 - 200 M_{sun}) are far more abundant than high mass clusters (>2000 M_{sun})

due to this relationship (beta = -2) (Battinelli, 1994)[21]. Low mass clusters with high mass stars are unusual, though the number of stars that are contained in these unusual clusters is comparable to the number in high mass clusters. The lower numbers of stars in these unusual low mass clusters means stellar encounters are unlikely. Therefore, providing these unusual low mass clusters are efficient at enriching stars, they could potentially produce more Solar-like Systems than high mass clusters.

We propose that these unusual clusters might be better environments for creating Solar like systems, rather than massive clusters due to lower probabilities of dynamical interactions and weaker external radiation, resulting in an environment that is more conducive to creating Sun like stars. In this paper we run N-body simulations to follow the evolution of low mass clusters (50-200 stars) containing one or more massive stars (>20 M_{sun}) and determine the number of stars that are sufficiently enriched by direct pollution. We also examine the interaction histories of the polluted stars and record any close interactions that occur. A range of initial conditions are used to cover many different potential star formation scenarios, but keep the stellar population constant so that stochastic differences in the clusters' evolution can be identified.

2. Method

We use Monte Carlo codes and N-body simulations to model and evolve clusters containing massive stars and analyse the results to see whether unusual low mass clusters are efficient producers of enriched stars.

2.1. Monte Carlo

Using the Monte Carlo method, a technique that uses repeat random sampling of a function, 10000 cluster masses were randomly drawn from the observed cluster mass relation (Battinelli, 1994)[21]. Once the cluster masses had been generated, clusters with masses between 50 -200 M_{sun} were filled with star masses drawn from the IMF (Figure 1). The star masses were drawn using random sampling (Elmegreen 2006; Parker & Goodwin 2007)[22,23], rather than sorted sampling (Weidner & Kroupa 2006; Weidner et al. 2013)[24,25]. Random sampling draws masses from the IMF based on their probability, and the only upper bound for the most massive star that can form is the mass of the cloud that it forms from. Selected sampling takes into account any direct physical dependence on the mass of the cloud, although there is still debate as to which is the correct method to use. The debate is an important issue with regard to the clusters selected for this paper as sorted sampling would rule out the case where a low mass cluster could produce a massive star, due to the physical dependence on the mass of the cloud. Random sampling does not place limits on low mass clusters, allowing us to select the unusual outcomes of the Monte Carlo simulations.

Star masses were drawn randomly from Maschberger's IMF (Maschberger, 2013)[26] to fill the cluster masses using another Monte Carlo code. After filling the cluster with stars, the median cluster (145 stars) that contained two massive stars was selected for investigation; the presence of two massive stars should result in

Fig. 1: Distribution of number of stars per cluster, separated by number of massive stars that appear in each cluster for all the clusters that were generated between 50-200 M$_{sun}$.

a greater number of enriched stars. To check the assumptions about numbers of stars that

are polluted by SN events, several simulations of clusters containing just one massive star were also run.

2.2. N-body simulations

N-body simulations are one of the most powerful tools used to follow the evolution of clusters. As the evolution of clusters cannot be observed in real time due to the time scales involved, N-body simulations provide a way to model many particles in any cluster-like environment. By running the same simulation multiple times with random initial positions and velocities, we can investigate whether the results of the simulation are statistical anomalies or real predictions as to how clusters evolve.

Observations of young star forming regions show highly sub-structured clusters which are fractal, with no dependence on cluster mass (Cartwright & Whitworth 2004)[27], therefore we set initial conditions which reproduce these conditions. For the median cluster, the initial conditions for the clusters were varied to cover a range of potential star forming environments to see if this affected the percentage of polluted stars. The virial ratio, fractal dimension and initial radius were varied for different simulations of the cluster to understand the effects these conditions had on the production of enriched stars.

The fractal generator from Goodwin & Whitworth (2004)[28] was used to set up the cluster with spatial and kinematic substructure, which is determined from the fractal dimension D. 20 versions of the initial conditions for the same cluster were run, varying only the random number seed used to initialise the stellar positions and velocities. The N-body simulations of the cluster were run for 100 Myr using the kira integrator within the Starlab environment (Portegies Zwart et al. 1999, 2001)[29,30]. We achieve stellar evolution by using the SeBa look-up tables (Portegies Zwart & Verbunt 1996)[31], which are also part of Starlab. During the simulation, 100 snapshots at equally spaced times were output for analysis. Clusters undergo dynamical processes that vary massively in scale and time; close interactions with stars may happen in a matter of years, whereas two body relaxation occurs on time scale of Myrs. Because of this, a Hermite scheme of variable time steps during the simulation was used (Dehnen & Read, 2011)[32]. Any information about close stellar interactions between snapshots was recorded.
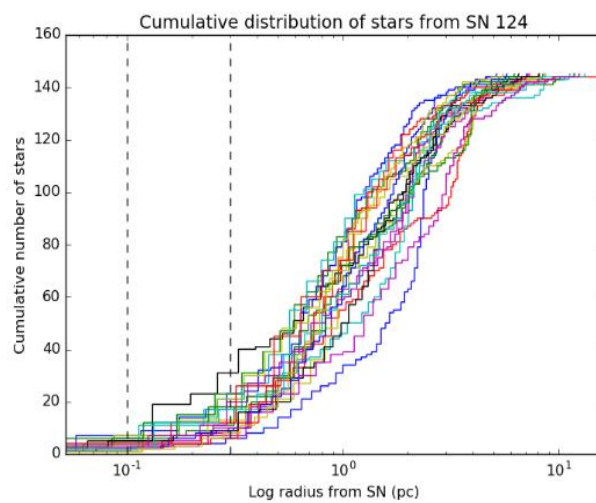
Fig. 2: The cumulative graph of star distance from SN for all 20 runs. SN 124 was the first SN event, occurring at 4.41 Myrs. The average number of stars polluted in this event was 9.
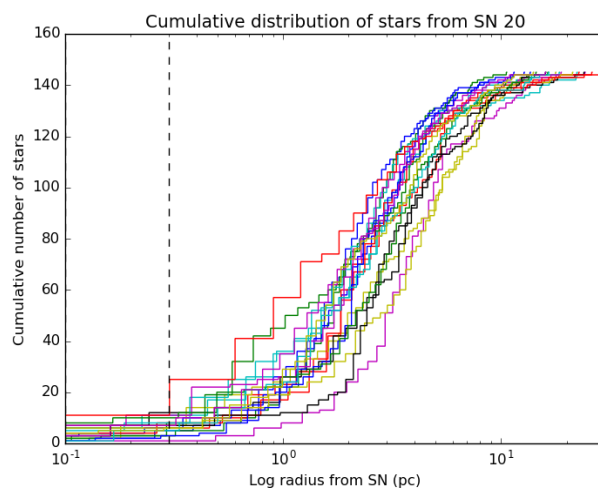
Fig. 3: SN 124 was the second SN event, occurring at 7.78 Myrs. The average number of stars polluted in this event was 9.

3. Results

The low mass cluster selected contained 145 stars, including two massive stars of roughly 22 and 41 solar masses. The 41 solar mass star exploded at 4.41 Myrs, becoming a black hole, and the 22 solar mass star exploded at 7.78 Myrs and became a neutron star. Across the 20 runs of the simulation, on average these two events enriched a total of 13 stars, 8% of the population of the cluster. The first SN event at 4.41 Myrs polluted an average of 9 stars (Figure 2), and the second event at 7.78 Myrs polluted an average of 4 (Figure 3). The number of stars polluted by both SN is negligible. The initial conditions set for this simulation were a radius of 1pc, a fractal dimension of 1.6 and a virial ratio of 0.5, values which are relatively normal for observed star forming regions.

The local and central densities for the cluster were calculated at each snapshot. The local density gives information about the average local density across the whole cluster, as the density from the centre of mass does not account for the very fractal initial conditions in the cluster. The local density was calculated by using the 10 nearest neighbours of each star. The central density, which calculates the density from the central densest point, was calculated by finding the centre of density out to the half mass radius. As the edge of clusters are not well defined, due to dispersion and stars being ejected, the half mass radius gives us a good point of reference when calculating radii and densities.

From the central density and the local density measurements, it is clear that peak density occurs within a few Myr, and the cluster disperses and becomes less dense after this period (Figure 4). As the density decreases in the cluster, the number of stars that can be polluted by SN explosions decreases. As the SNe events occur much later than the peak density, the number of enriched stars remains low.
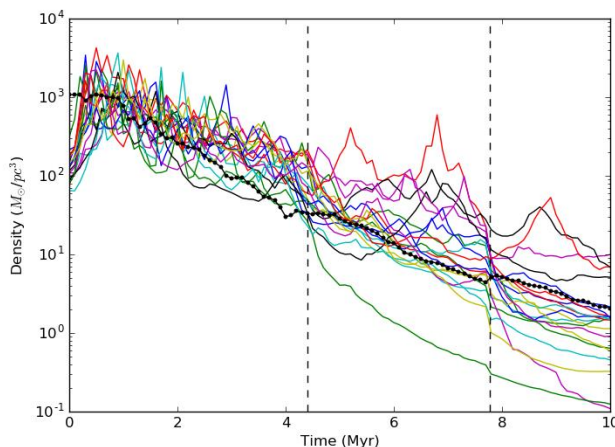
Fig. 4: The density of the cluster with time. The coloured lines show the local density for all 20 runs, the dotted black line shows the average centre of density for all 20 runs. The dashed grey lines indicate the SN event times.

To test whether the peak in density could be changed to a later time by varying the initial conditions and whether there was a set of conditions that were more favourable to creating many polluted stars, a suite of simulations was run using the same median cluster.

The effect of changing the initial conditions was minimal on changing the peak density time and the number of enriched stars. The densest and most fractal clusters, with an initial radius of 0.5 and a fractal dimension of 1.6, polluted less than 15 stars, a much lower number than is produced in massive clusters.

As predicted, of the stars that were enriched, very few underwent stellar interactions within a few hundred AU and most had very quiet dynamical histories (Figure 5). As the initial sizes of proto planetary discs are as yet unknown, interactions that occur at distances of 10^3 AU could possibly have an effect, however as the Solar System is contained within a few hundred AU, it is unlikely these

Fig. 5. The interaction history of all the polluted stars during one simulation run. The dashed lines indicate the SN event times.

will have had a significant effect on planet formation in Solar like systems.

4. Conclusions

The peak in the density of the cluster occurs at the time of formation and decreases almost instantly due to two body relaxation. This is too soon for any stars to have evolved enough that they can explode as SNe. Of the many low mass clusters that we ran simulations for, none produced a high percentage of polluted stars as a result. Even when analysing a system with two massive stars greater than 40 solar masses, the number of stars that were polluted remained low (below 10% of the cluster's population). Changing the initial conditions so the cluster was denser did not significantly increase the number of stars that were polluted either.

With the low rates at which stars were polluted by SNe in these low mass clusters, high mass clusters still seem to be the most likely environment for creating Solar-like systems. With only roughly 10% of the stars being polluted in these unusual low mass clusters, it is unlikely that they represent the environment in which the Solar System formed. Most enriched stars underwent no close stellar interactions and had quiet dynamical histories, as predicted. However, because of the low numbers of stars polluted in these unusual low mass clusters, even though the enriched stars were unperturbed, high mass clusters still produce a higher number of enriched stars overall and so are still better environments for producing Solar like systems.

References:
[1] Lada & Lada, 2003, ARA&A, **41** 57
[2] Bressert et al, 2010, MNRAS, **409** L54
[3] Haisch et al 2001, ApJ, **553** L153
[4] Parker, R. J., 2014, MNRAS, **445** 4037
[5] Scally & Clarke, 2001, MNRAS, **325** 449
[6] Parker & Quanz, 2012, MNRAS, **419** 2448
[7] de Juan Ovelar et al 2012, A&A, **546** L1
[8] Bastian, N. et al 2010, ARA&A, **48** 339
[9] Adams, Fred C. 2010, ARA&A, **48** 47
[10] Marhas, K. K. et al, 2008, ApJ, **689** 622
[11] Villeneuve, J., Chaussidon, M., & Libourel, G. 2009, Science, **325** 985
[12] Lee T., Shu F. H., Shang H., Glassgold A. E., Rehm K. E., 1998, ApJ, **506** 898
[13] Shu F. H., Shang H., Gounelle M., Glassgold A. E., Lee T., 2001, ApJ, **548** 1029
[14] Wasserburg, G. J. et al, 1994, ApJ, **424** 412
[15] Trigo-Rodríguez, J. M. et al, 2009, Meteorit. Planet. Sci., **44** 627
[16] Kastner, J. H., & Myers, P. C. 1994, ApJ, **421** 605
[17] Parker R. J. & Dale J. E., 2016, MNRAS, **456** 1066
[18] Parker et al, 2014, MNRAS, **438** 620
[19] Parker R. J., Quanz S. P., 2012, MNRAS, **419** 2448
[20] Adams, Fred C. et al 2014, ApJ, **789** 86
[21] Battinelli P. et al, 1994, A&AS, **104** 379
[22] Elmegreen B. G., 2006, ApJ, **648** 572
[23] Parker R. J. & Goodwin S. P. 2007, MNRAS, **380** 1271
[24] Weidner C., Kroupa P., 2006, MNRAS, **365** 1333
[25] Weidner C., Kroupa P., Pflamm-Altenburg J., 2013, MNRAS, **434** 84
[26] Maschberger T., 2013, MNRAS, **429** 1725
[27] Cartwright & Whitworth, 2004, MNRAS, **348** 589
[28] Goodwin S. P., Whitworth A. P., 2004, A&A, **413** 929
[29] Portegies Zwart S. F., Makino J., McMillan S. L. W., Hut P., 1999, A&A, **348** 117
[30] Portegies Zwart S. F., McMillan S. L. W., Hut P., Makino J., 2001, MNRAS, **321** 199
[31] Portegies Zwart S. F., Verbunt F., 1996, A&A, **309** 179
[32] Dehnen W. & Read J. I., 2011, EPJP, **126** 55

Alessandro Savino

Constraining the star formation history of dwarf galaxies with horizontal branch stars

Alessandro Savino1,2, Maurizio Salaris1 and Eline Tolstoy2

1Astrophysics Research Institute, LJMU, IC2, Liverpool Science Park, 146 Brownlow Hill, Liverpool

2Kapteyn Astronomical Institute, University of Groningen, Postbus 800, 9700 AV Groningen, The Netherlands

E-mail address: A.savino@2014.ljmu.ac.uk, M.Salaris@ljmu.ac.uk, etolstoy@astro.rug.nl

Dwarf galaxies are considered to be the building blocks of every galaxy in the universe, thus characterising the way they formed and evolved, through the evaluation of the so called star formation history, is crucial to test cosmological models and simulations. Horizontal branch (HB) stars are low-mass, helium burning stars that are found in all old stellar populations. Their intrinsic brightness makes them perfect target to determine the history of galaxies up to 5 Megaparsecs away. Alas, these stars are routinely neglected in star formation history determination due to the number of parameters that affect they photometric properties. The first results of the present research demonstrate that, in spite of these theoretical uncertainties, the inclusion of HB stars is capable of considerably improving the accuracy of the recovered star formation history applying a technique called synthetic HB modelling on the Carina dwarf spheroidal galaxy.

Keywords. galaxies: dwarf – galaxies: evolution – galaxies: stellar content – Hertzsprung-Russell and C-M diagrams – stars: horizontal-branch

1. Introduction

According to current cosmological models, the process of galaxy formation follows a hierarchical mechanism, with every galaxy being assembled from smaller building blocks which are thought to be the progenitor of dwarf galaxies (DGs). In this context, a precise characterization of the processes that led to DGs as we see them today is of paramount importance, as it can be used as a benchmark to test the predictability of different cosmological model prescriptions. To this aim, knowing the star formation history (SFH) of these stellar systems is a great advantage as it sheds light on the time scales of formation and chemical enrichment of these objects.

Unfortunately, a detailed characterization of SFHs is possible only in stellar populations that are close enough to be resolved and, even in this case; the measurement is not exempt from various sources of uncertainties and systematics. A major issue is that the canonical stellar indicators for SFH estimation lie, at least for old populations, in the faint region of the colour magnitude diagram (CMD) and span a relatively small range in colour and magnitude. This characteristic means that these CMD feature are highly sensitive to photometric errors and age-metallicity degeneracy, which can be pronounced as in the case of the main sequence turn-off (TO). Due to these difficulties, the typical time resolution for state of the art SFH determinations hardly is lower than 1-2 Gyr, for old populations (Monelli et al. 2010; de Boer et al. 2012).

Horizontal branch (HB) stars, on the other hand, are considerably brighter and more extended in colour respect to the TO, potentially being interesting targets to recover the

properties of their parent population, and to help break the before mentioned degeneracy when analysed together with other evolutionary phases. Unfortunately, HB is usually neglected in SFH analysis, if not to place very mild constraints. The reason for that is the number of theoretical difficulties that historically prevented to predict from first principle the morphology of the HB of a simple stellar population. One of these difficulties is related to fact that red giant branch (RGB) stars lose mass before reaching the HB phase. The precise amount of this mass loss and its functional dependence is still matter of debate but a number of studies in globular clusters (Gratton et al. 2010; Origlia et al. 2014) seem to hint that a linear relation exists between the integrated mass loss on the RGB and the metallicity of a simple stellar population.

The problem is further complicated in galactic globular cluster by the presence of multiple population that are thought to have a spread in their helium abundance. For long time before the discovery of these populations through the detection of anti-correlation between light elements, which is thought to be a distinctive signature of this phenomenon, HB morphology interpretation was hampered by the presence of this hidden parameter. DGs field stars, on the other hand, seem to lack these peculiar chemical anomalies (Geisler et al. 2007; Fabrizio et al. 2015) and, in spite of them being composite stellar populations, their HB modelling might be noticeably easier, as HB properties will depend on the SFH and the integrated mass loss only.

Salaris et al. (2013) made use of a detailed SFH estimate for the Sculptor dwarf spheroidal (dSph) to model consistently its HB, showing that a high degree of consistency could be reached using a simple mass loss prescription, compatible with previous works on globular clusters, and without the need of a helium enhancement.

Encouraged by this result, we undertake a similar analysis on the Carina dSph, using the SFH from de Boer et al. (2014) and finding that the same mass loss recovered for Sculptor is able to reproduce the general extension of Carina old HB component. At the same time, a number of discrepancies on the detailed HB structure confirmed how, even with the present uncertainties regarding RGB mass loss, a critical analysis of HB properties can help refine state of the art SFH determinations. A brief summary of these results is presented in the next sections, while the complete analysis can be found in Savino et al. (2015).

2. Dataset

Along the line of Salaris et al. (2013), we start from a well determined SFH and we aim to model the HB matching its distribution over an observed CMD and, as a weaker constraint, the properties of variable stars on the HB. The input SFH we use is the one presented in de Boer et al. (2014), which gives the estimated star formation rate on a fine grid of age and metallicities, including an estimate of the $[\alpha/Fe]$ for each bin. Although the SFH is computed for different concentric annuli of the galaxy, we chose to use the global solution inside the tidal radius of the galaxy to maximize our number statistics. It should be noted that the stellar evolution model used in the SFH computation are the same employed in this analysis (BaSTI, Pietrinferni et al. 2005).

As for the CMD, we chose to use the photometric catalogue from Bono et al. (2010) for the numerous advantages it involves. Being a merge of more than 4000 archive images, the photometric accuracy of this dataset is astonishing, with uncertainties of the order of 0.004 mag at the level of the HB. In addition, the problem of contamination by potentially undetected variable stars is strongly mitigated as the averaged magnitude of these stars will be close to their true value. Finally, the impossibility to use a foreground CMD decontaminate the field has been bypassed by means of an exquisite cleaning based on a colour colour diagram, to exclude Milky Way stars and background galaxies.

The variable star modelling has been aimed to match the intrinsic mean colour and magnitude, as well as pulsation properties, of the RR Lyrae stars listed in the Coppola et al. (2015) catalogue. The same catalogue has been used to remove the detected variable from our CMD, thus permitting a further cleaning.

3. Synthetic horizontal branch modelling

Having an available SFH, we can compute different HB synthetic models assuming different mass loss prescriptions. In order to find the model that best matches the observed CMD we could in principle try to match the detailed distribution of HB stars in colour and magnitude. Realistically, this criterion is difficult to fulfil, given the intrinsic limits of the input SFH and the uncertainties in the functional dependence of the integrated mass loss. For this reason, we impose a somewhat weaker constraint that is matching, within the Poisson uncertainty, the relative stellar counts inside three regions that encompass the blue and the red side of the HB and the young red clump (RC), respectively.
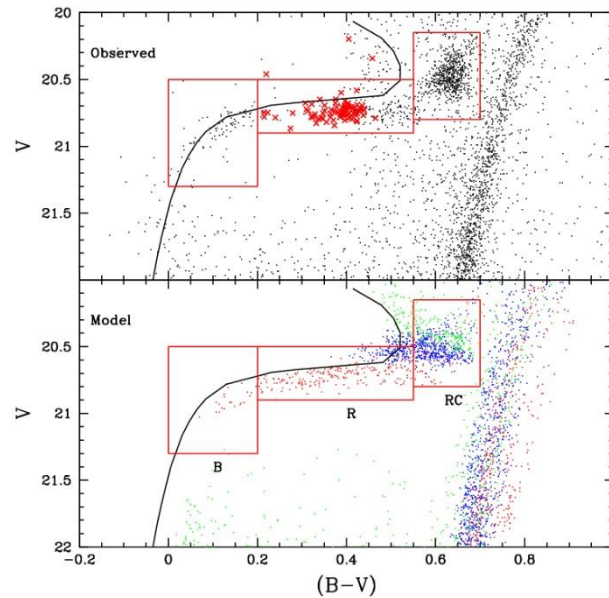
Figure 1: Upper panel: Observed CMD of the HB region of the Carina dSph. Red crosses mark RR Lyrae stars. The red boxes are those chosen for detailed comparison. The solid black line represents the ZAHB, up to a mass of 1.45M for a metallicity of Z=0.0001.

As a natural starting point, we compute the HB model using the mass loss law recovered in Salaris et al. (2013), which is shown in figure 1. It can be noted that the Sculptor-like mass loss is able to reproduce the general colour extension of the old HB, hinting at that some sort of regularity in the mass loss of low mass stars in this range of metallicity. The RC is, instead, manifestly bad reproduced. The presence of a group of synthetic stars blueward of the observed RC may suggest that the mass loss for these young, metal poor stars is too high. This is a reasonable assumption as stars so massive spend a considerably shorter time on the RGB, compared to older stars, and this in turn may affect the total amount of mass lost.

Looking at the zero age horizontal branch (ZAHB) model for this metallicity, however, reveals that there is a critical mass where the initial helium burning locus bends toward bluer colours so there is no mass loss capable of reproducing the RC morphology, given the assumed SFH.

Despite the Sculptor mass law is able to match the colour extension on the HB, the "good fit" criteria are not matched even here, the major cause being a gap in the observed stellar distribution, around (B-V) ~0.3, which is missing in our simulation. As a further evidence of this problem, there is a disagreement between the observed RR Lyrae period distribution and the synthetic distribution, which predicts a fraction of first overtone pulsators which too high. This is naturally explained as the variable stars causing this problem are the stars that in the synthetic HB populate the observed gap.

In spite of the many trials undertaken, we didn't find any way to include this feature in our simulations. The reason for this is that the synthetic stars in the gap region are characterized by different values of metallicity. This implies that there is no mass loss law

(expect maybe a very fine tuned prescription) that can reproduce that gap while preserving the colour extension of the HB.

This issue, together with discrepancy of the RC model is suggesting that there is a noticeable difference between the SFH estimated in de Boer et al. (2014) and the real Carina SFH.

4. Implication for the star formation history

Motivated by the above results, we investigated which sub-population, in the adopted SFH, was responsible for the problem encountered during the HB modelling. The lower panel of figure 2 shows the age and metallicity distribution of the stars in our synthetic HB. The synthetic stars which occupy the region corresponding to the observed gap have been to reside in the two regions enclosed by the blue lines. In particular, we note that old and metal rich of these two components is also responsible for the broad synthetic RGB observed in figure 1. The mismatch on the RC is caused, as already stated, by a population of young and metal poor stars, enclosed by the red line.
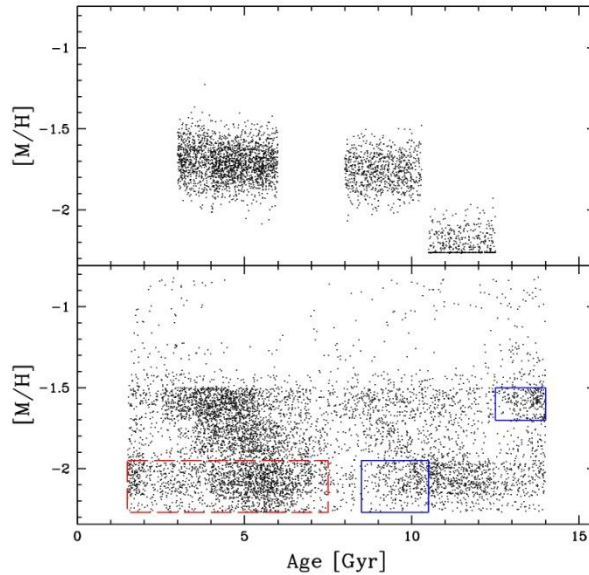
Figure 2: Distribution of HB synthetic stars in the age-metallicity plane. Upper panel: Synthetic generated with our toy SFH. Lower panel: Synthetic generated using the original SFH.

This finding is showing the great constraining power of the HB, as its morphology can point out forbidden ranges of age and metallicities that would result in unobserved features. In the case of Carina, the HB morphology is suggesting that the true SFH of this galaxy is more confined to specific regions of the age metallicity plane. To test this hypothesis, we built a "toy model", more bursty SFH, assuming a Sculptor-like mass loss (conveniently modified for young ages), with the aim of reproducing the the morphology and stellar distribution of Carina helium burning phase. Our best fit model is composed of 4 bursts, which are all consistent with a sub-population of the original SFH and are listed in table 1.

Figure 3 shows the comparison of our simulation with the Carina CMD and the distribution of its HB stars. This toy model is able to match the detailed distribution of the observed HB stars with a precision of the order of the photometric uncertainty. In addition, the morphology of the other evolutionary phases, as the magnitude of the double TO and the colour spread is roughly recovered. Finally, a comparison with RR Lyrae pulsational properties reveals that also the period distribution and the ratio between fundamental pulsators and first overtone pulsators is well reproduced. These results, when considered in light of the differences and similarities between out toy model and the original SFH (as shown in figure 2), suggests that the origin of the difficulties encountered in the HB modelling

is probably related to the finite accuracy of the SFH determination and to the presence in it of sub-populations with are probably absent in the true SFH of Carina.

Table 1: Properties of the star formation bursts present in our toy.

	tmin	tmax	<[M/H]>	SFR
	Gyr	Gyr	dex	
Burst 1	3.0	4.0	-1.69	0.12
Burst 2	4.0	6.0	-1.71	0.39
Burst 3	8.0	10.3	-1.77	0.29
Burst 4	10.5	12.5	-2.23	0.2

Each burst is characterized by a flat age distribution between a minimum and a maximum age and a gaussian metallicity distribution centered at the listed value and with a spread of 0.1 dex. The last column lists the normalized star formation rate for each burst.

a)   **b)**
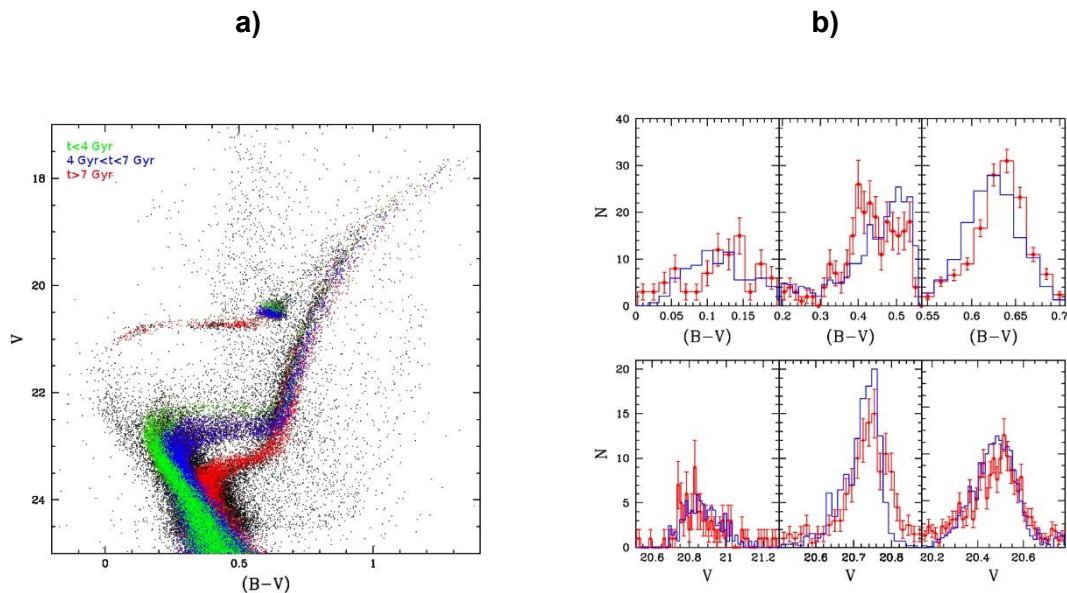
Figure 3: a) Observed CMD of Carina (blackdots), superimposed to the synthetic CMD from our toy SFH (coloured dots). b) Histogram of the stellar distribution in (B-V) colour and V magnitude for the observed (red) and synthetic (blue) HB.

To investigate this issue, and as a final consistency check, we compared the metallicity distribution function (MDF) recovered from the toy model with the observed distribution (Helmi et al. 2006) used in the original SFH determination, with results shown in figure 4. The two distributions don't compare well, even when measurement uncertainties are added to our model MDF (middle panel). The observed MDF is not sharply bimodal as our synthetic MDF and presents an asymmetric broadening toward metal poor values.

The interpretation of this discrepancy may either discourage our model or point to some problem with the spectroscopic determined metallicities. The latter scenario is intriguing, as it would natural explain the presence of the young, metal poor sub-population which is not seen in Carina HB. A possible interpretation for this problem in the metallicity determination may involve the calibration of the calcium triplet CaT-metallicity relation, which has been used to recover the MDF (Starkenburg et al. 2010). This relation assumes a fixed chemical composition pattern with [Ca/Fe] = 0.25. At a fixed metallicity, changing the calcium

abundance will change the equivalent width of this feature, mimicking a different value of metallicity. High resolution spectroscopic analysis have revealed that the stellar population of Carina presents a big spread in calcium abundance, with [Ca/Fe] values ranging from 0.3 to slightly sub-solar values (Shetrone et al. 2003; Lemasle et al. 2012). When this spread is not taken into account, the resultant metallicity distribution inferred from the CaT will be affected, likely underestimated (as the relation was calibrated for high values of [Ca/Fe]). Including an estimate of this effect to our synthetic MDF, adding an additional 0.25 dex spread toward lower metallicities, brings the distribution to a remarkable level of agreement, as shown in the right panel of figure 4.

5. Conclusions

Summarizing, the detailed HB simulation we undertook on the Carina dSph has highlighted the potential of modelling the HB in DGs, at the same time showing the limitation of state of the art SFH determinations based solely on the TO and RGB phases

The overall colour extension of the old HB component, which the Sculptor mass loss law has been able to recover given the input SFH, hints that there is some regularity in the general amount of mass loss that DG stars lose during their RGB evolution.
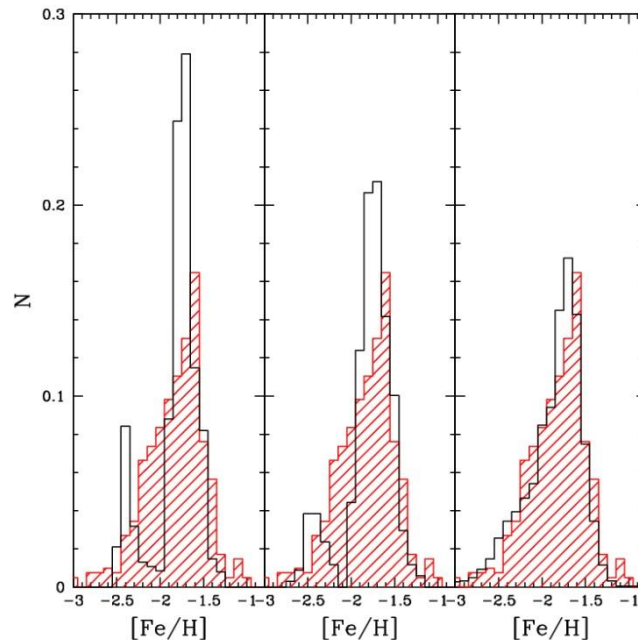
Figure 4: Syntetic (black) and observed (red) MDF for Carina RGB stars. Left panel: original synthetic MDF. Middle panel: after simulating spectroscopic measurement uncertainties. Left panel: after simulating the bias induced by the [Ca/Fe] abundance spread.

On the other hand, a few discrepancies between our model and observations were impossible to overcome changing the mass loss alone. This motivated us to try to establish the constraints that the HB morphology gives on Carina SFH. We then built a toy model SFH that allowed us to reproduce the detailed stellar distribution of the HB, the broad morphology of the other evolutionary phases and the pulsational properties of RR Lyrae stars.

Comparing the metallicity distribution as predicted by our model with the spectroscopic measurements revealed that a potential bias in the metallicity determination, induced by the spread in calcium abundance of Carina stellar population, could in principle reconcile the two MDFs and potentially explain the presence of sub-populations in the original SFH that are not allowed by HB morphology.

In conclusion, these results demonstrate how the inclusion of constraints from the HB when evaluating SFHs will help break the age-metallicity degeneracy and substantially improve the resolution in the final solution. Future work implement this methodology will be undertaken on a sample of local group galaxy, as well as more distant galaxies.

References

Bono G et al. 2010 *PASP* **vol 122** p 651-61

Coppola G et al. 2015 *ApJ* **vol 814** id 71

De boer T J L, Tolstoy E, Hill V, Saha A, Olsen K, Starkenburg E, Lemasle B, Irwin M J and Battaglia G 2012 *A&A* **vol 539** id A103

De boer T J L, Tolstoy E, Lemasle B, Saha A, Olszewski E W, Mateo M, Irwin M J and Battaglia G 2014 *A&A* **vol 572** id A10

Fabrizio M et al. 2015 *A&A* **vol 580** id A18

Geisler D, Wallerstein G, Smith V V and Casetti-dinescu D I 2007 *PASP* **vol 119** p 939-61

Gratton R G, Carretta E, Bragaglia A, Lucatello S and D'orazi V 2010 *A&A* **vol 517** id A81

Helmi A et al. 2006 *ApJ* **vol 651** p L121-4

Lemasle B, Hill V, Tolstoy E, Venn K A, Shetrone M D, Irwin M J, De boer T J L, Starkenburg E and Salvadori S 2012 *A&A* **vol 538** id A100

Monelli M et al. 2010 *ApJ* **vol 722** p 1864-78

Origlia L, Ferraro F R, Fabbri S, Fusi pecci F, Dalessandro E, Rich R M and Valenti E 2014 *A&A* **vol 564** id A136

Pietrinferni A, Cassisi S, Salaris M and Castelli F 2004 *ApJ* **vol 612** p 168-90

Salaris M, De boer T J L, Tolstoy E, Fiorentino G and Cassisi S 2013 *A&A* **vol 559** id A57

Savino A, Salaris M and Tolstoy E 2015 *A&A* **vol 583** id A126

Shetrone M, Venn K A, Tolstoy E, Primas F, Hill V and Kaufer A 2003 *The Astronomical Journal* **vol 125** p 684-706

Starkenburg E et al. 2010 *A&A* **vol 513** id A34

Kate Furnell

Clusters in the Era of eROSITA: The SPIDERS BCG Sample

K Furnell[1*], C A Collins[1], L Kelvin[1], I Baldry[1] and N Clerc[2]

[1] Astrophysics Research Institute, Faculty of Engineering and Technology
IC2,
Liverpool Science Park,
146 Brownlow Hill,
Liverpool,
 L3 5RF
[2] Max Planck Institut für Extraterrestrishe Physik,
Postfach 1312,
85741 Garching bei München,
Germany
* E-mail: astkfurn@2015.ljmu.ac.uk

ABSTRACT
This work represents a sample of 221 X-ray selected, optically confirmed, brightest cluster galaxies from within the SPectroscopic IDentification of eROSITA Sources (SPIDERS) survey across a low to intermediate distance range (2-16 GLyr). Using SDSS DR12 imaging, we identify the BCG using a combination of techniques, including X-ray data and outputs from a robust cluster-finder algorithm (redMAPPER). We also categorise the BCGs morphologically through vigorous visual analysis, and then use these identifications to model their light profile using several pieces of mainstream astronomy software, most notably GALFIT and SExtractor. We discuss the preliminary results of this fitting process and compare the outputs with legacy studies.

Key words: Physics, Astronomy, Cosmology, Galaxy Clusters, Brightest Cluster Galaxies, X-ray

1. INTRODUCTION

Since the era of Zwicky (Zwicky, 1937), astronomers have used clusters of galaxies; namely large groups of galaxies bound together under gravity, as laboratories for galactic evolution (e.g. von der Linden et al., 2007) and as tracers of the large scale structure (LSS) of the Universe (e.g. eBOSS, Tojeiro et al., 2014; REFLEX, Böhringer et al., 2004). They represent the most massive virialised structures in the Universe, with masses between $10^{13} - 10^{15}\, M_\odot$ (1 $M_\odot = 10^{30}$ kg, the mass of the Sun). In general, clusters in the local Universe contain member galaxies which are red in colour and elliptical in morphology (figure 1), indicating that they are no longer forming stars (e.g. Dressler, 1980). This is unlike the general population of galaxies in the Universe as a whole, which tend to be blue (star-forming) and spiral shaped (~70%, e.g. Loveday, 1996). There is no coincidence behind why these

(a)   (b)

Figure 1: Comparison between elliptical (a) and cD-type (b) BCGs at current epochs. The images shown here are false colour composites from the Sloan Digital Sky Survey (section 2).

galaxies group; indeed, their very nature is intrinsically linked to the extremely dense environment in which they reside. Galaxies within clusters are subject to multiple complex interactions as a consequence, which contribute to the 'quenching' of their star formation and a change in their morphology. These include major merger events (where a galaxy comes together with a similarly massive galaxy under gravity), minor merger events (where a more massive galaxy effectively 'consumes' a significantly less massive galaxy) and harassment (where a galaxy is 'disrupted' by a passing neighbour). Toomre (1977), for example, was one of the first to establish that two merging similarly-massive spiral galaxies could result in an elliptical type galaxy.

It is thought that the mechanisms behind star formation in galaxies rely on the presence of the intergalactic medium (IGM). The intergalactic medium refers to a reservoir of superheated (10^5 – 10^7 K), diffuse gas mainly composed of hydrogen that emits X-rays through thermal bremsstrahlung emission (deflection of electrons by charged atomic nuclei). Stars form as the IGM 'cools' and 'condenses', which is a process that becomes disrupted by violent merger events, such as those present in galaxy clusters (e.g. Tremonti et al., 2007). In addition, cluster environments can cause the IGM to be stripped from galaxies falling into them. 'Ram-pressure stripping' (e.g. McCarthy et al., 2008) is such an example, where infalling galaxies feel a 'wind' due to their relative motion with respect to the surrounding intracluster medium (ICM), which causes gas in their IGM to be stripped if it is not sufficiently gravitationally bound. Another example is 'strangulation', where the inflow of gas into the IGM is disrupted due to the presence of the cluster through accretion. Galaxies then use their remaining gas reservoirs which, when used up, effectively terminates their star formation (Peng et al., 2015).

Perhaps the most extreme example of galactic evolution in clusters are the brightest cluster galaxies (or BCGs). These galaxies are by far the most massive in the Universe and can contain hundreds of times more stars than the Milky Way (10^{11} – 10^{13} M$_\odot$). These galaxies reside at the very heart of galaxy clusters, namely the bottom of the potential well, where the majority of interactions occur. In addition, their position often coincides with the cluster X-ray 'peak', otherwise known as the location at which the majority of external gas is flowing into the cluster. Although almost all BCGs which reside in massive clusters are morphologically spheroidal and red in colour, they can appear markedly different. Broadly, BCGs can be split into two further morphological classes: 'elliptical' galaxies, which are compact in appearance and 'cD' galaxies, which have both a central bulge and an extended envelope of stars (figure

1). The majority of studies favour a two-stage process in BCGs: where the 'bulge' in cD galaxies forms first, with the extended stellar halo being formed at a later stage (e.g. Whiley et al., 2008;
Johansson et al., 2012). Elliptical-type BCGs therefore represent the progenitors of cD-type BCGs, having yet to undergo the formation of an extended envelope.

There is much debate in the literature as to how the envelopes in cD galaxies arise. Some studies suggested that due to the presence of BCGs at the peak inflow of gas in clusters, the halo arises from star formation as a result of condensing cluster gas (e.g. Fabian, 1994), which then terminates through heat 'injection' from an energetic source, such as an active galactic nucleus, or AGN (e.g. McNamara et al., 2014). However, other studies have refuted these claims, unable to find enough gas at sufficiently cool a temperature to produce the necessary rates of star formation in the past required to form the envelopes observed in cD-type BCGs that are visible at more local scales (e.g. Peterson & Fabian, 2006).

In more recent years, merger-based theories of envelope formation in cD galaxies have become more popular. Two such theories include 'galactic merging' (e.g. Merritt, 1984) and 'galactic cannibalism' (e.g. Hausman & Ostriker, 1978). Galactic merging, the more violent of the two mechanisms, refers to the partial 'collapse' of a cluster, where a number of comparably-sized member galaxies undergo major mergers in a short timespan to form a single galaxy (e.g. Hopkins et al., 2009), with the outer envelope forming as a result of violent gravitational relaxation. The alternative, galactic cannibalism, suggests that the majority of BCGs already have the bulk of their mass in place at early times, with the envelope element arising at a later stage as a result of mergers with significantly smaller galaxies with respect to the BCG (e.g. Johansson et al., 2012). The latter theorem is more broadly compatible with observations, which find BCGs consistently homogeneous in mass up to large distances (e.g. Collins et al., 2009). For example, Whiley et al., (2008) suggested that BCGs gained the majority of their mass at least 8 billion years ago, with models of their stellar populations indicating the bulk of their intrinsic stars formed roughly 10 billion years ago.

However, the compatibility between observational studies and theoretical simulations modelling BCG formation remains weak. Simulations tend to find growth rates in BCGs which are far steeper than those observed, often by at least a factor of 3-4 (e.g. de Lucia & Blaizot, 2007). To reconcile this, simulators have begun to study the effect of mergers in cluster centres with respect to the intracluster light (or ICL) at late times. The ICL is a faint component of galaxy clusters which is composed of stars dissociated from galaxies through mergers, harassment and tidal stripping, or alternatively stars formed in situ through monolithic collapse of the ICM. The ICL, due to the fact that it is both faint and extended, is notoriously difficult to detect on an observational level and studies attempting to do so have produced a wide range of results, from the ICL making up between 4-50% the total stellar mass of a galaxy cluster (e.g. Burke et al., 2012; Burke, Hilton & Collins 2015; Gonzalez et al., 2007).

Recent models appear to find ICL mass fractions within these limits on a qualitative basis, e.g. (Contini et al., 2014), however it is suggested that the bulk of the mass from these models comes from galactic merging at late times (5-25%), with smaller galaxies contributing little. This contests somewhat with observations; for example, Burke & Collins (2013) found a 50:50 split in mass growth from major and minor mergers in a sample of 14 BCGs within the same timescale of ICL formation. Considering simultaneous growth of the ICL, Burke, Hilton & Collins (2015) then found, using a sample of 25 galaxy clusters within roughly the same epoch, no major mergers present at all in the sample, inconsistent with the notion that the ICL is formed from stars as a result of gravitational relaxation from galactic merging at late times.

Because it is a technical challenge to differentiate the faint ICL component from the faint envelopes of galaxies such as those associated with cD-type BCGs (which actually may well be impossible to do completely without dynamical information), studies looking into parametrising the morphology of cluster galaxies are growing in significance. Zhao et al., (2015), for example, used astronomical pipeline tools (e.g. GALAPAGOS, Barden et al., 2012) on a sample of optically selected clusters from the Sloan Digital Sky Survey (von der Linden et al., 2007; York et al., 2000) with focus on separating the bulge and envelope components of BCGs in their sample. They found that they were able to distinguish sharply between cD and elliptical-type BCGs in a quantitative manner, by comparing remaining flux post-fitting to object extent.

This work focuses on analysing the light profiles of a sample of X-ray selected brightest cluster galaxies from the SPIDERS survey, in order to link their morphological properties with the X-ray and dynamical properties of their host clusters. Parametrisation will be carried out using SIGMA (Kelvin et al., 2011), a software wrapper combining several standard astronomy packages used in image processing to create a seamless routine with minimal user input. Such work may well pave the way for ICL studies on deeper optical images (such as in the Dark Energy Survey; DES Collaboration, 2005), should the methodology used here prove effective. This paper is structured as follows: section 2 introduces the data and sample selection method, section 3 discusses SIGMA and section 4 summarises current findings, including concluding remarks concerning upcoming work.

2. THE DATA

2.2. The SPIDERS Survey

SPIDERS (The SPectroscopic IDentification of eROSITA Sources) is a survey currently running on the Apache Point Observatory 2.5m Sloan Digital Sky Survey Telescope. The Sloan Digital Sky Survey (SDSS; York et al., 2000) is an optical survey which has run for almost two decades, having imaged 35% of the sky in five different wavelength bands, from the near-ultraviolet to the near-infrared (for photometric information, refer to Fukugita et al., 1996). The SDSS has collected imaging data on over 500 million objects to date, is currently on its twelfth data release (DR12; Sloan Collaboration, 2015) and is the most cited survey in the history of astronomy. Alongside SDSS imaging data, SPIDERS uses the multi-object spectrograph on the Sloan telescope (Smee et al., 2013) to take high-quality spectroscopy of galaxy clusters and other X-ray luminous objects (e.g. AGN) identified by the eROSITA satellite (Merloni et al., 2012). eROSITA is an X-ray telescope due for launch in early 2017, with all-sky survey capabilities and much higher angular resolution than previous X-ray missions such as the ROSAT All Sky Survey (RASS; Voges, 1993). It is due to detect all massive clusters in the Universe above $3 \times 10^{14} M_\odot$, which will yield the largest sample of galaxy clusters to date (upwards of 200,000).

As eROSITA is yet to launch, preliminary observations are being carried out using readily-detectable clusters in legacy data. The CODEX catalogue (Finuguenov et al., in prep.) has been the basis of this initial cluster sample. CODEX combines X-ray data from publically available RASS source catalogues with redMAPPER data from SDSS imaging, in order to yield well-characterised galaxy clusters with complete catalogues of identified members. The redMAPPER algorithm is described in detail in Rykoff et al., (2014), but it was essentially designed to optically select galaxy clusters by considering candidate members with similar colours in close proximity to a set of spectroscopically-confirmed red sequence input galaxies (e.g. CMASS, Ahn et al., 2012). Although not perfect by any means, redMAPPER provides a robust estimate for galaxy clusters detectable in shallow optical surveys such as the SDSS, particularly due to the fact that blue cluster members are uncommon at local scales and, if present, are generally far less massive than typical cluster members. Combined with X-ray centroids from RASS, this method of cluster selection is designed to

infer minimal corruption from projection effects and point-like X-ray sources. Input catalogues have been refined by optical analysis (Clerc et al., in press) and observations will continue until 2020.

Spectroscopic information alongside photometric information is desirable for multiple reasons. With spectroscopy, it is possible to analyse the chemical properties of astronomical objects such as galaxies and thereby determine whether any major evolutionary events occurred in the past or are still ongoing, such as a burst of star formation or a phase of AGN activity (e.g. Fernandes et al., 2011). In addition, spectroscopy yields precise distance measurements and astrometry for an object, valuable for constraining cluster membership. Distance measurements using photometry are much more unreliable and often require studies to use statistical tools such as resampling methods in the presence of uncertainty (for an example, see Hartley et al., 2013).

Figure 2: A false colour composite image depicting a full SDSS field. The BCG (green) and the X-ray centroid (red) are visible at the bottom right.

2.3. Sample Selection

BCG candidates were extracted from the 18/11/2015 data release of SPIDERS clusters. In total, the FITS catalogue contained ~1600 clusters with spectroscopic observations. They included information about the location of a cluster on the sky in celestial coordinates, a list of candidate members with spectroscopically-confirmed distances and X-ray information such as luminosity, temperature and the location of peak emission. In addition, the catalogues contained output information from redMAPPER, including the most likely BCG candidate in the lists of observed cluster members. Optical data were obtained from the publically accessible SDSS DR12 science archive (http://dr12.sdss3.org/). The science archive contains pre-processed science-standard imaging data, as well as composite .jpg images for visualisation and optical confirmation. The images are 2048 × 1489 pixels, with an arcsecond to pixel scale of 0.398''/pix.

BCG candidate clusters for the SPIDERS sample were selected via the following criteria:

- Candidate clusters were considered only if they had a signal-to-noise (henceforth SN) value of 3 or greater from RASS. This was to ensure that the measurements represented a robust detection of peak X-ray emission from a cluster.
- BCGs were only considered for a galaxy cluster if there was a candidate above the limiting magnitude of the SDSS survey (21.2 in the near-infrared 'i' band). If there was no such galaxy present, the cluster was not considered.
- To have a member galaxy within a proximity of at least 30' of the X-ray peak. This was to ensure that the clusters selected had a well-defined centre.

Clusters were initially only considered viable if they were detected to have 20 or more members by redMAPPER upon prior recommendation from the CODEX team, to ensure minimal contamination of point sources. This criterion was relaxed in order to increase the number of candidate BCGs to 10, though it was found to make an unsubstantial increase to the final catalogue after manual confirmation (>10 clusters were found to be appropriate). As well as this, if there were no likely BCG candidates present in the catalogue for a cluster, it was discarded. This was revisited at a later stage, but again, the increase in objects was minimal. Cuts were made as conservative as possible in order to maximise the sample size of BCG candidates.
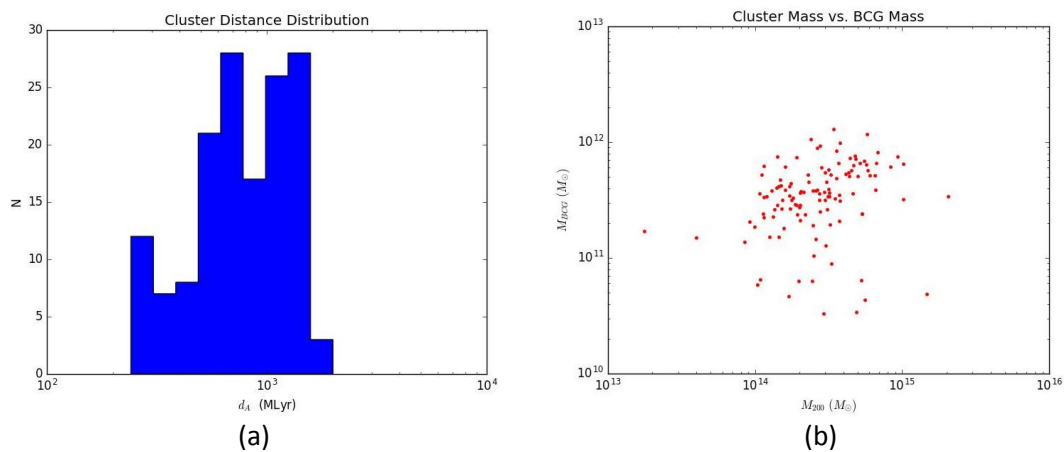
Figure 3: *Distance distribution of the SPIDERS BCGs in millions of light years (a) and (b), scatter plot of BCG stellar mass vs. cluster mass from X-ray scaling relations.*

After initial cuts, the BCG candidate clusters were screened by eye to ensure the identifications listed in the catalogues were accurate. In order to be classified as a BCG, a galaxy was generally required to meet the following criteria:

- Be clearly visible on a logarithmically-scaled image.
- For its host cluster to be clearly visible, with multiple members present.
- To be the brightest galaxy in proximity to the peak of X-ray emission.

An example of this selection is displayed in figure 2. During the screening process, it was noted that ~18% of clusters had a BCG candidate that was clearly present in the images, but absent from the catalogue. This may have been due to the fact that redMAPPER breaks down at close proximities, due to the angular scale of a cluster being large. Comparisons

were drawn between the BCGs selected here and the redMAPPER centroids, with ~95% of galaxies being present in both catalogues.

The separate images were then re-screened to ensure they were appropriate for the study. Truncated BCGs were removed, with no attempt to combine images in order to reduce systematics when dealing with the sky background. In total, this left a remaining sample of 221 BCGs to fit with light profiles. A histogram of their commoving distances can be seen in figure 3, as well as a scatter plot comparing their available stellar masses (152 in total; see Zhao et al., 2015 for an example use of the MPA-JHU catalogue of spectral measurements alongside a detailed description of the available parameters) against
the X-ray gas masses of their host clusters. The BCGs follow a shallow, linear relationship in mass with respect to their host cluster, which is consistent with the literature (e.g. Whiley et al., 2008).

3. STRUCTURAL PARAMETERS

3.1. Initial Analysis

In two dimensions, the light profile of an elliptical galaxy is best modelled by the Sérsic profile (Sérsic, 1964), which has the following form (equation 1):

$$I(R) \propto e^{-kR^{1/n}} \tag{1}$$

(a)

(b)

(c)

(d)
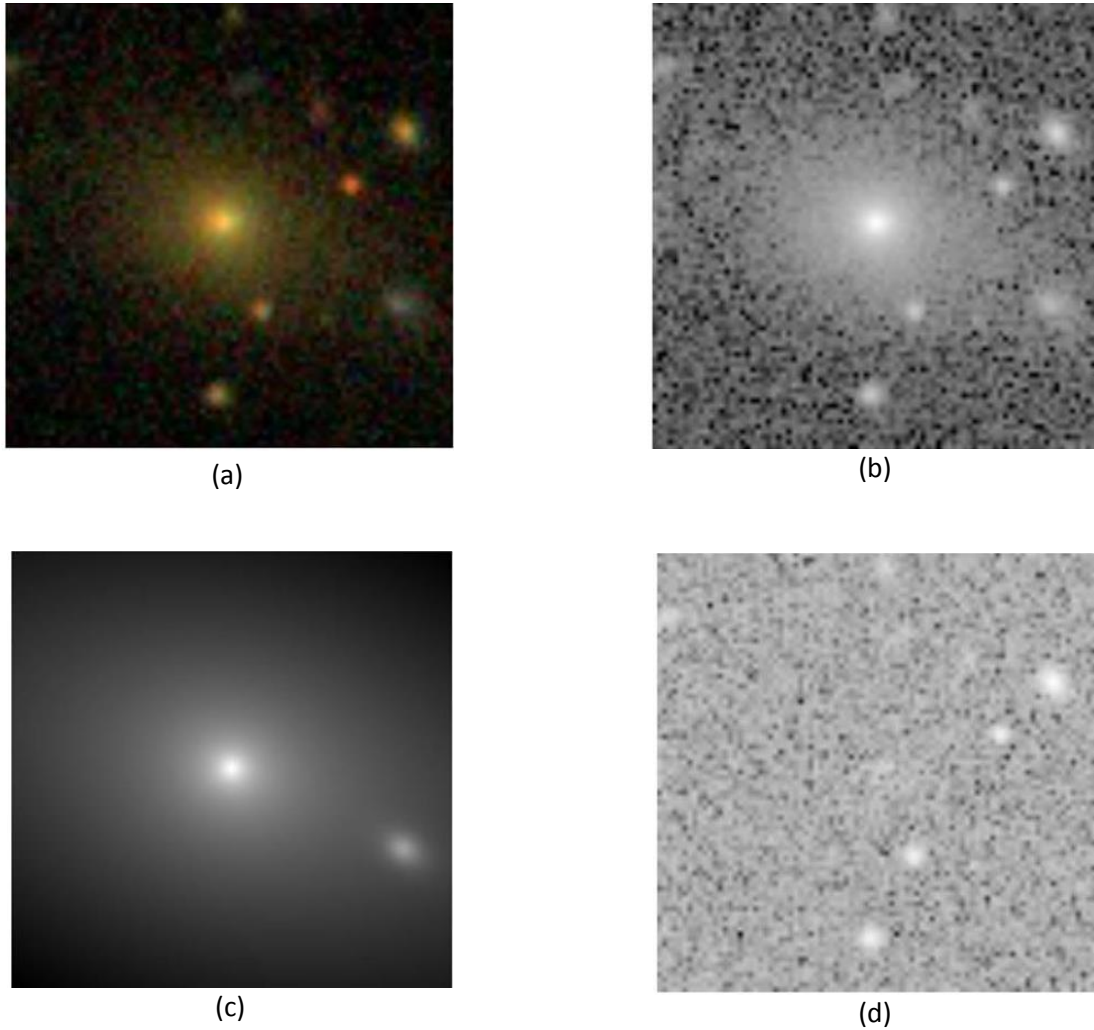
Figure 4: An example of single profile fitting. The original r band data is shown in (b), with the resulting fit shown in (c). The profile subtracted image, or 'residual', is also depicted (d). The original false colour SDSS composite (a) has been included for clarity.

Where I(R) is the intensity of light at a distance R, n is the Sérsic index describing the shape of a light profile and k is a constant (see Caon et al., 1991 for more details). Elliptical galaxies generally follow a profile with a Sérsic index roughly equal to 4.

However, cD galaxies are not fit well with a single component model; instead, multiple studies have favoured a two-component approach, combining both a Sérsic profile for the inner bulge and an exponential profile ($n = 1$) to describe the extended envelope component (e.g. Donzelli et al., 2011). Therefore, in order to produce the best possible fits for the BCGs in this study, the candidate galaxies were broadly classified by eye during the screening process based on halo extent into elliptical and cD-types.

Initial fits were carried out on SDSS *r*-band images using GALFIT (Peng et al., 2002), which is a well-tested and widely used profile modelling software. It runs on a set of user-controlled scripts, which require initial estimations (e.g. position angle, magnitude) and a model with which to fit an object. GALFIT then generates a model based on minimisation of the 'reduced χ^{2}', a statistical indicator of fit quality. The initial estimations for the BCGs in this study were taken from SExtractor outputs (Bertin & Arnouts, 1996), another piece of software broadly used in astronomy for extracting catalogues of

(a)   (b)
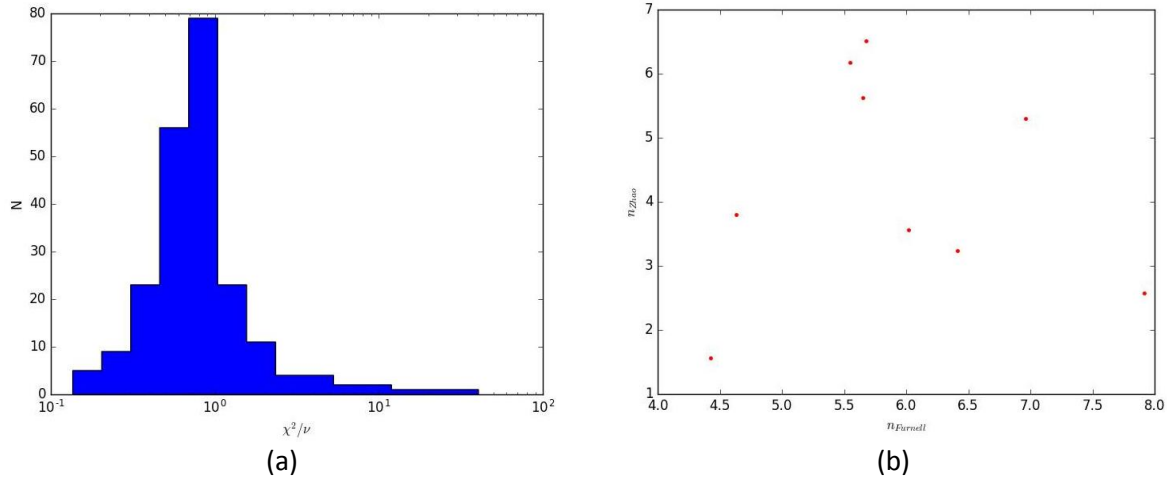
Figure 5: *Reduced χ^2 distribution from initial fits (a). Just under half (~45%) of the BCGs have fits where reduced $\chi^2 \sim 1$, with about 35% being overfitted and 20% being either underfitted or having high reduced χ^2 due to the presence of luminous objects that were not simultaneously fitted. Comparing the Sérsic indices of 9 common BCGs with Zhao et al., (2015), there is no correlation.*

objects from an image. Fitting was opted for over using SExtractor values due to the necessity of producing more accurate geometric and photometric measurements, as SExtractor tends to overestimate the sky background and thereby underestimate true object magnitudes (see Graham & Driver, 2005 for a recent review). Instead, the sky background, for the sake of simplicity, was fit using the inbuilt routines in GALFIT. Figure 4 depicts an example of BCG fitting using this process (reduced χ^2 = 1.024, where reduced χ^2 = 1 indicates a perfect fit). As is visible, the model is an excellent description of the light profile of the object. The code was designed to fit multiple objects simultaneously, as the centre of galaxy clusters tend to be crowded fields (figure 4(c), for example, depicts the final fit of both the BCG and a nearby neighbour).

3.2. SIGMA

However, this simplistic approach has numerous faults. Although some BCGs were indeed fit successfully, there were many failed examples, where a) the fits were too extensive due to an inappropriate model being used (reduced χ^2 < 1), b) the local background could not be represented by the global background fit from GALFIT (attributing an incorrect flux to the desired object) or c), the description of light scatter in the image (i.e. the point spread function, or psf) was too noisy, generating a poor fit. Two example residual images are shown in figure 5, where profile subtraction has been clearly unsuccessful. In addition, having carried out an initial parameter comparison with a sample of 10 BCGs in common with Zhao et al., (2015), there is little consistency (figure 5). Further testing is required in order to fully understand why this is the case, in order to ensure globally accurate fits for the BCG sample in this study.

To rectify these issues, the next step will be to attempt the fits again using a piece of wrapper software. SIGMA, or the Structural Investigation of Galaxies via Model Analysis (Kelvin et al., 2012), contains elegant solutions to some of the problems encountered here. The software has been successfully tested on more than 150,000 galaxies in the GAMA survey (see Driver et al., 2009) and is well suited for the purposes of this study, particularly due to the fact that it favours a 'point and fit' approach through an input coordinate list, which other pieces of wrapper software lack due to being designed to process full surveys rather than single objects (e.g. GALAPAGOS, Barden et al., 2012). SIGMA combines three major pieces of astronomy software in order to generate an automated pipeline, namely SExtractor,

PSFEx (Bertin, 2011) and GALFIT. It contains inbuilt routines that perform masking to remove undesirable objects at the stage of fitting, smooths psfs generated from PSFEx by modelling them as Moffat functions and models the background for objects at local scales to ensure the most accurate fits possible. It is predicated that SIGMA will generate superior fits to carrying out the processes separately without proper masking and background treatment, plus it represents a global, robust method that should be applicable to all of the fields used in this study.

4. SUMMARY

Using data from the SPIDERS survey, we selected a sample of 221 BCGs using a combination of data thresholding and visual analysis. We applied fits to the data using GALFIT, based on parameters extracted by SExtractor, on 221 science-quality pre-processed *r*-band images from the SDSS, without applying any image masking or consideration of the local background. It was then discovered that although this method worked in some cases, a more sophisticated approach was required in order to generate the accurate fits necessary for the BCG sample.

Once accurate fits have been established through SIGMA, there are several main science goals. The first is to establish the relationship between the physical properties of the BCGs (e.g. stellar mass, luminosity, morphology) with the total mass of their host clusters, using proxies for mass such as total ICM gas (based on X-ray luminosities from RASS/CODEX) and from the velocity dispersion of galaxies within the cluster (e.g. Finn et al., 2005). Then, using outputs from SExtractor and GALFIT, estimate the total gain in stellar mass from more a distant cluster population in comparison to a more nearby population by using dynamical friction arguments (e.g. Burke & Collins, 2013). From these measurements, it would be useful to estimate how much stellar matter is 'lost' to the ICL, which would be possible to do using thresholding and stacking methods for a nearby BCG population (e.g. Burke et al., 2012). This then motivates the need for some deep observations on a comparison sample of more distant clusters, which may involve seeking telescope time for infrared imaging (e.g. VLT, VISTA) or by using deeper survey data (e.g. DES). Finally, it would be useful to further analyse the spectroscopy of the SPIDERS clusters, in order to detect the presence of any AGN or star formation activity. Carrying out this approach on such a large X-ray selected sample should help shed light on the mechanisms behind the peculiar growth of BCGs and clusters in general over cosmic time.

REFERENCES

Ahn, C. P., Alexandroff, R., Prieto, C. A., Anderson, S. F., Anderton, T., Andrews, B. H., ... & Bautista, J. (2012), *The Astrophysical Journal Supplement Series*, *203*(2), 21.

Barden, M., Häußler, B., Peng, C. Y., McIntosh, D. H., & Guo, Y. (2012), *Monthly Notices of the Royal Astronomical Society*, *422*(1), 449-468.

Bertin, E., & Arnouts, S. (1996), *Astronomy and Astrophysics Supplement Series*, *117*(2), 393-404.

Bertin, E. (2011), *Astronomical Data Analysis Software and Systems XX* (Vol. 442, p. 435).

Burke, C., Collins, C. A., Stott, J. P., & Hilton, M. (2012), *Monthly Notices of the Royal Astronomical Society*, *425*(3), 2058-2068.

Burke, C., & Collins, C. A. (2013), *Monthly Notices of the Royal Astronomical Society*, *434*(4), 2856-2865.

Burke, C., Hilton, M., & Collins, C. (2015), *Monthly Notices of the Royal Astronomical Society*, *449*(3), 2353-2367.

Böhringer, H., Schuecker, P., Guzzo, L., Collins, C. A., Voges, W., Cruddace, R. G., ... & MacGillivray, H. T. (2004), *Astronomy & Astrophysics*, *425*(1), 367-383.

Collins, C. A., Stott, J. P., Hilton, M., Kay, S. T., Stanford, S. A., Davidson, M., ... & Mann, R. G. (2009), *Nature*, *458*(7238), 603-606.

Contini, E., De Lucia, G., Villalobos, Á., & Borgani, S. (2013), *Monthly Notices of the Royal Astronomical Society*, stt2174.

Dark Energy Survey Collaboration (2005), *arXiv preprint astro-ph/0510346*.

De Lucia, G., & Blaizot, J. (2007), *Monthly Notices of the Royal Astronomical Society*, *375*(1), 2-14.

Donzelli, C. J., Muriel, H., & Madrid, J. P. (2011),*The Astrophysical Journal Supplement Series*, *195*(2), 15.

Dressler, A. (1980), *The Astrophysical Journal*, *236*, 351-365.

Driver, S. P., Norberg, P., Baldry, I. K., Bamford, S. P., Hopkins, A. M., Liske, J., ... & GAMA Team. (2009), *Astronomy & Geophysics*, *50*(5), 5-12.

Fabian, A. C. (1994), *Annual Review of Astronomy and Astrophysics*, Volume 32, 1994, pp. 277-318

Fernandes, R. C., Stasińska, G., Mateus, A., & Asari, N. V. (2011), *Monthly Notices of the Royal Astronomical Society*, *413*(3), 1687-1699.

Finn, R. A., Zaritsky, D., McCarthy Jr, D. W., Poggianti, B., Rudnick, G., Halliday, C., ... & Simard, L. (2005),*The Astrophysical Journal*, *630*(1), 206.

Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., & Schneider, D. P. (1996), *The Astronomical Journal*, *111*, 1748.

Gonzalez, A. H., Zaritsky, D., & Zabludoff, A. I. (2007), *The Astrophysical Journal*, *666*(1), 147.7

Graham, A. W., & Driver, S. P. (2005), *Publications of the Astronomical Society of Australia*, *22*(2), 118-127.

Hartley, W. G., Almaini, O., Mortlock, A., Conselice, C. J., Grützbauch, R., Simpson, C., ... & Dunlop, J. S. (2013), *Monthly Notices of the Royal Astronomical Society*, stt383.Hausman & Ostriker, 1978

Hopkins, P. F., Cox, T. J., Younger, J. D., & Hernquist, L. (2009), *The Astrophysical Journal*, *691*(2), 1168.

Johansson, P. H., Naab, T., & Ostriker, J. P. (2012), *The Astrophysical Journal*, *754*(2), 115.

Kelvin, L. S., Driver, S. P., Robotham, A. S., Hill, D. T., Alpaslan, M., Baldry, I. K., ... & Häussler, B. (2012), *Monthly Notices of the Royal Astronomical Society*, *421*(2), 1007-1039.7

Loveday, J. (1996), *Monthly Notices of the Royal Astronomical Society*, *278*(4), 1025-1048

McCarthy, I. G., Frenk, C. S., Font, A. S., Lacey, C. G., Bower, R. G., Mitchell, N. L., ... & Theuns, T. (2008), *Monthly Notices of the Royal Astronomical Society*, *383*(2), 593-605.

McNamara, B. R., Russell, H. R., Nulsen, P. E. J., Edge, A. C., Murray, N. W., Main, R. A., ... & Kirkpatrick, C. C. (2014),*The Astrophysical Journal*, *785*(1), 44.Merloni, A., Predehl, P., Becker, W., Böhringer, H., Boller, T., Brunner, H., ... & Georgakakis, A. (2012), *arXiv preprint arXiv:1209.3114*.

Merritt, D. (1984), *The Astrophysical Journal*, *276*, 26-37.

Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H. W. (2002), *The Astronomical Journal*, *124*(1), 266.

Peng, Y., Maiolino, R., & Cochrane, R. (2015), *Nature*, *521*(7551), 192-195.

Peterson, J. R., & Fabian, A. C. (2006), *Physics reports*, *427*(1), 1-39.

Ade, P. A. R., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., ... & Bartolo, N. (2015) *arXiv preprint arXiv:1502.01598*.

Rykoff, E. S., Rozo, E., Busha, M. T., Cunha, C. E., Finoguenov, A., Evrard, A., ... & Pierre, M. (2014), *The Astrophysical Journal*, *785*(2), 104.

Sérsic, J. L. (1963), *Boletin de la Asociacion Argentina de Astronomia La Plata Argentina*, *6*, 41.

Smee, S. A., Gunn, J. E., Uomoto, A., Roe, N., Schlegel, D., Rockosi, C. M., ... & Brinkmann, J. (2013), *The Astronomical Journal*, *146*(2), 32.

Tojeiro, R., Ross, A. J., Burden, A., Samushia, L., Manera, M., Percival, W. J., ... & Dawson, K. (2014), *Monthly Notices of the Royal Astronomical Society*, *440*(3), 2222-2237.

Toomre, A. (1977), *Evolution of Galaxies and Stellar Populations* (Vol. 1, p. 401).

Tremonti, C. A., Moustakas, J., & Diamond-Stanic, A. M. (2007), *The Astrophysical Journal Letters*, *663*(2), L77.

Voges, W. (1993), *Advances in Space Research*, *13*(12), 391-397.

Von Der Linden, A., Best, P. N., Kauffmann, G., & White, S. D. (2007), *Monthly Notices of the Royal Astronomical Society*, *379*(3), 867-893.

Whiley, I. M., Aragón-Salamanca, A., De Lucia, G., Von Der Linden, A., Bamford, S. P., Best, P., ... & Noll, S. (2008), *Monthly Notices of the Royal Astronomical Society*, *387*(3), 1253-1263.

York, D. G., Adelman, J., Anderson Jr, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., ... & Boroski, W. N. (2000), *The Astronomical Journal*, *120*(3), 1579.

Zhao, D., Aragón-Salamanca, A., & Conselice, C. J. (2015), *Monthly Notices of the Royal Astronomical Society*, *448*(3), 2530-2545.

Zhao, D., Aragón-Salamanca, A., & Conselice, C. J. (2015), *Monthly Notices of the Royal Astronomical Society*, *453*(4), 4444-4455.

Zwicky, F. (1937), *The Astrophysical Journal*, *86*, 217.

Jade Hind

Association mapping of neuropsychiatric symptoms and socio demographics for predisposition to psychiatric disorders

J. Hind

Faculty of Environment and Technology, Liverpool John Moores University, Byrom Street, L3 3AF

E-mail address: J.Hind@2012.ljmu.ac.uk

Abstract: Psychiatric disorders are common among all populations currently, and as such require a diverging approach to the current myopic practice which can hinder and cause detrimental problems in the treatment of mental disorders. The proposed option is to consider the biological markers indicative of the common underlying symptoms and traits which combine as a diagnosis for many of these disorders; a biological validity of these symptoms could provide a baseline judgement for the predisposition to a disorder or the reclassification of the diagnosis.

1.   Introduction

The focus of this proposal is concerned with psychiatric disorders; disorders affect a patient in various ways such as the way you think, the way you behaviour and what mood you are in. Currently in our society, mental health disorders are becoming more apparent and accepted as we become more politically correct. Given this advance, it is becoming more necessary to find treatments for conditions such as Depression, Schizophrenia, Bipolar Disorder, Obsessive Compulsive Disorder, etc. Psychiatric disorders at present according to recent and past studies prove a prevalence of approximately 46.4% [1] concerning the chances of developing any disorder at any point in a person's life. Anxiety disorders carry a 28.8% chance and substance abuse disorders carry a 14.6% chance. These are concerning figures given their impact on the population. More concerning is the prevalence of mental disorders when compared against some of the world's most debilitating diseases and disorders; at 7.4% of disease burden worldwide [2], mental disorders are more prevalent than HIV/AIDs and Diabetes. Further to this, a study conducted by Kessler, R.C., et al [1], indicated that approximately 50% of mental disorder cases will develop around the age of 14 while 75% of cases will have developed by the age of 24 [1]. Figure 1 outlines the impact of each mental disorder on DALYs based on the number of people and the age.

This study aims to explore the biological markers that are associated with the traits/ symptoms of a patient; it has been clear from the research conducted historically that psychiatric disorders in particular are a challenge for determining genetic markers or combinations. Current research [3] suggests that this is due to the diagnosis of the disorder/ disease; many psychiatric disorders are considered umbrella terms
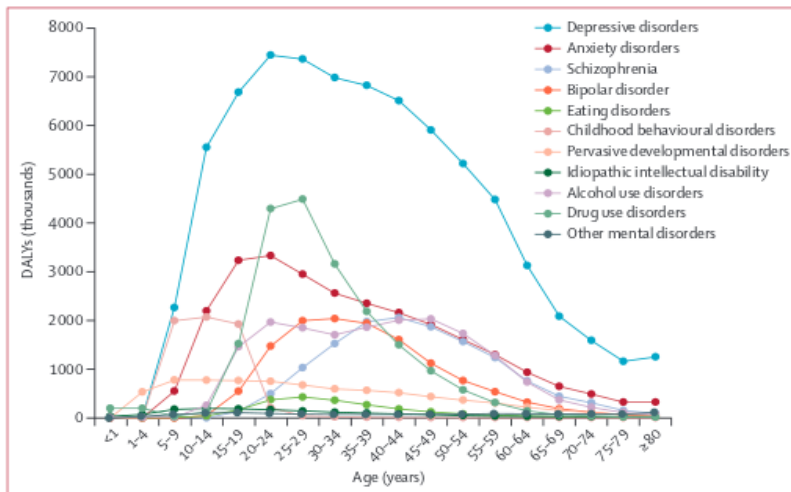
Figure 1: Prevalence by psychiatric Disorder

which collate symptoms of a patient in order to classify their condition into a defined syndrome or disorder. As such, although highly reliable, these diagnoses are not biologically valid. As an example, we can consider research undertaken by Hirschfeld RM, et al. [4] in which 69% of the 600 patients who participated were initially misdiagnosed with unipolar depression rather than the actual diagnosis of bipolar, with the diagnosis not being rectified for a mean of 5.7 years; while another investigation found that this is not an uncommon occurrence[5]; physicians must determine a disorder based on the information given and what they see, which in many cases cannot provide enough to make a valid diagnosis. Bipolar as a disease was not accepted until early 1970's, despite being a proposed disease in 1854 [6], the current psychiatric diagnosis', which are defining terms for many, if not all of our psychiatric disorders, are based on symptoms; You will provide a list of symptoms which will then be used as a reference to apply a defining term to your condition, essentially which compartment do you fit in? To summarise, misdiagnosis is a problem which can lead to severe issues (suicidal risk and social impairment) if untreated or treated using incorrect techniques. More recent research [7] indicates the need for reclassification of mental health disorder based on biological information.

As such, in this proposed study we aim to investigate the underlying traits of psychiatric disorders in order to aid in the stimulation of reclassification based on biologic evidence. This suggests that we should be targeting the underlying traits and symptoms which are used to classify and diagnose patients; instead of using symptoms as a classification variable; a technique which causes misdiagnoses, we suggest the use of genetic markers. Is there a marker, or collection of genetic markers which may be able to identify symptoms that a patient is likely to experience relating to mental health?
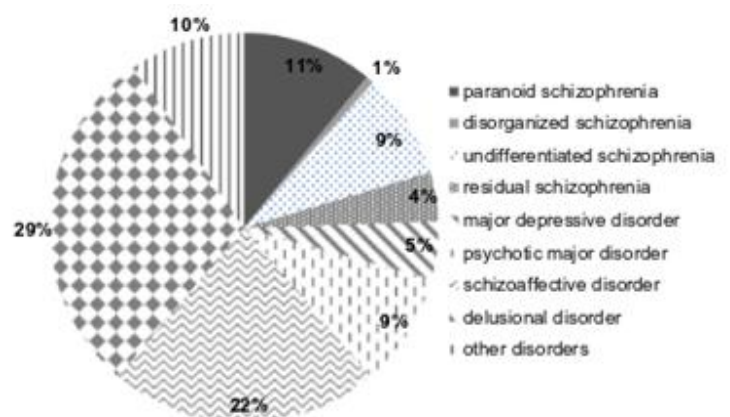
Figure 2: Misdiagnosis of psychiatric disorders

Further to this, is there a common collection of genetic markers which appear together to display very
similar symptoms in patients; could this approach be used to define new disorders which are biologically valid? Further to this case-control study which defines the association between a symptoms or trait and the SNPs, we will consider environmental and socio demographics information which can indicate if particular lifestyle choices, environments or stressors can induce or increase the chance of symptom development. As can be seen within the next section, there has already been studies which have outlined potential variables indicating the associations with certain lifestyle choices.

2.   Literature Review

The current research focuses on the genetic aspect which may specify a biological predisposition based on the very 'building blocks' of our internal code. The following studies provide insight into the research which has already been conducted in this area and the findings which have been determined as a result of the investigation. Research conducted by Kelly, C. et al [8] in 1984 provided evidence which suggested a high correlation in schizophrenic patients and their adoption of **smoking**. With a prevalence of smoking amongst schizophrenic patients being as high as 88% in some areas, this could certainly be an area to explore in regards to the genes or symptoms which may influence the habitual need to smoke; this is likely to correlate highly with patients who abuse substances. Therefore smoking and substance abuse must be considered as both a confounding and exploration variable. Wickman, H. et al [9] produced a study which considered the relation between **common symptoms of schizophrenia** and the **current environment at onset**. Five traits were involved in the study: **Depressive, Manic, Reality Distortion, Disorganisation and Psychomotor Poverty.** The most significant results exposed from this study related to Psychomotor poverty, disorganisation and Reality distortion which yielded that all three traits were associated with deterioration, premorbid functioning and chronic course of the disorder. More interestingly is the relationship between disorganisation and psychomotor poverty with single marital status; and concerning disorganisation, **unemployment,** these findings indicated that **single marital status** and unemployment were significant and related with the traits. [10] is very similar to the proposed research and states that "Concordance rates in monozygotic twins were significantly higher when probands had a greater number of negative symptoms", this outlines that although there no findings which suggested the same pertaining to the positive symptoms, there were however correlative findings based on the negative symptoms. The concluding remark is very similar to the research question posited in this paper: *"The results have important implications for determining diagnostic criteria and understanding the pathogenesis of schizophrenia."* Cardno et al [11], provided a study which looked at the overlapping that occurs in the genetic implications of schizoaffective disorder, schizophrenia and bipolar disorder. The findings suggest that there are **common genetic influences** between these disorders which indicate that the diagnosed terms might be too broadly defined. As such, this supports the posited question in the **need for biologically defined terms** becoming more apparent. Due to the nature of schizophrenia, there is emphasis now being placed on the health implications which are evident in schizophrenic patients whose treatment is causing **weight gain** [12]. Weight in patients with schizophrenia can cause unnecessary stress and general additional health implications; therefore a problem to be addressed is targeting schizophrenic patients who are more likely to gain weight as a result of the treatment

process. A symptom of weight gain could be an issue that is present before treatment is started.

3.   Plan of Action

Considering the posited research, the following discussion will outline and explain the process which will be followed in order to analyse the data to yield results based on the inputted variables.

3.1.   Study Design

The study design will develop in stages, firstly using a GWAS (Genome Wide Association Study), which will provide a list of associative SNPs with the cohort that can be used to for candidate or clustering SNPs, which could indicate either one or multiple symptoms later on in the study process. The next stage is to use a case-control study which will investigate the indicated SNPs outlined in the GWAS to determine whether they apply predominantly in patients afflicted with the symptom/ traits, patients who do not display the symptoms/ traits or to determine whether the SNP does not prove to be significant based on either cohort. When conducting a GWAS and case-control, the cohort will be defined by the symptom/ trait; the symptoms and traits in question are Depression, Anxiety, Obsessive Compulsive Disorder (OCD) and Agoraphobia. These symptoms and traits, while considered disorders in themselves, are common among many disorders, or what is considered in this paper as 'umbrella terms' e.g. Bipolar Disorder, Major Depressive Disorder, Schizophrenia, etc. These study methods are dependent on various methods for the analysis and cleansing of data, below are the steps and methods which are undertaken in order to produce valid markers and variables for analysis. A fundamental step for genetic studies lies in the cleaning and analysis for viability in genetic data. It is important to produce a valid and viable study design in order to reduce systematic bias which can be caused by errors in genotype calling and biases in study design; these issues can cause an increase the number false-positives and false-negative associations

3.2.   Quality Control (QC)

QC involves the analysis of data for the purposes of omitting patients or genetic data which does not meet requirements set out within the study design. The basic steps of this stage include the identification of missing or discordant data which will cause issues during the analysis of the data. For this particular study, the main methods applied consist of: Cochran-Armitage test, MAF (Minor Allele Frequency) > 0.05, Genotype Call (GC) >0.95 and cluster plots, Hardy-Weinberg Equilibrium (HWE), Missing Information, Linkage Disequilibrium and r^2. Population Stratification is another consideration which will be applied if necessary. Another more modern method which is necessary as an applied method in order to produce results which are more reliable is the Q-Values method which will utilise the P-Values of the SNP and provide an estimation of the false-discovery rate which will likely be present in the data. Once the q-values have been calculated, we consider the proportion based on the overall number; the generally accepted level of false-discovery rate is 5%, and this will provide a good stance for the quality of the data.

3.3.   Exploratory Analysis

The next stage which in many cases can be more time consuming is Exploratory analysis, this requires an investigator to familiarise and discover the main characteristic of the data. The techniques which are applied in this section are predominantly visual, allow for easy investigation of the variables and information which could be appropriate for utilization within the investigation of the research. The techniques which will be applied within this research consist of box plot, scatter plot and histograms. Another process in this stage which can

either be considered as part of exploratory analysis, or as a stage of its' own is Principal Component Analysis (PCA). This method is applied as a mathematical algorithm used for the reduction of dimensionality of the data while retaining the majority of variation in the data set; essentially this is a process which can identify the most significant variables which can be applied to the data while eliminating what is usually considered 'noise'; noise is likely to increase the number of false-positive and false-negatives found while analysing data as given enough variables an algorithm is likely to find a correlation based on the information, however this can also be due to confounding variables as mentioned in the previous stage.

3.4. Techniques

With the study analysis, the main method which will apply for many purposes will be logistic regression. Logistic Regression will be used predominantly in this study due the study design providing and utilising discrete information. LR provides a statistical method option which is applied to dataset analysis in cases which involve more than one independent variable; due to the vast amount of information which is necessary to analyse in order to determine the findings based on socio demographics and environmental information, a method which is able to successfully manage and analyse multiple variables is necessary.

3.5. Data Source

The data source which will be analysed within this study is provided by the DBGap resource database [13]. The Genetic Association Information Network (GAIN) study provides 5066 subjects within the full dataset, including subjects for use of schizophrenia case subjects and general use control data subjects. This dataset also provide sociodemographic, environmental and medical information which can be used for the analysis of exploratory and confounding variable information. This study is comprehensive in the additional variables that are associated with the subjects and as a result provides a quality dataset. The GAIN program set out to genotype existing research studies in six major common diseases, combining these results with clinical data to create a new resource with the aim to aid genetic research with a significant resource option.

3.6. Software

The software to be utilised within this study is established in the area of statistics and genomics; with this project being in its' early stages, there is a need to produce results and finding with defined software that is likely to produce findings that are in line with resolved academics and professionals within the area. The software choice that are prominent within the early stages of this project are GBench, R (Bioconductor) and PhenX. GBench [14] is used for the analysis of private data but can be combined with the resources within NCBI database; the use of this software within this project is based on the format of the described data. Visualisation and analysis of the data is more suited to this software and with the additional option of combinatorial datasets from an established resource base, it is a advantageous choice. R(Bioconductor) [15] is an open source framework with a vast amount of libraries which can perform and produce an extensive line of analysis and visual results; further to this there are many libraries for the purposes of data transformation, essential with the varying data formats that are applicable in genetics. PhenX is a toolkit which offers 'high quality, well-established measures of phenotypes and exposures' [16] this particular tool will aid in the discovery of studies for reference and comparison based on the phenotypes and variables chosen. For this research, the focus will be based around the demographics and environmental domains to determine broadly applicable measures.

4.   Conclusion

This paper is a literature review based on the given research question; further study will be conducted following the steps outlined in the plan. As this study is in it's initial stages future work is considered to not only produced results based on the initial study provided but to consider using machine learning, Support Vector Machines (SVM) and K Nearest Neighbour (KNN) have been identified as potential candidates for this process.

5.   References

[1]   Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R. and Walters, E.E., 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Archives of general psychiatry, 62(6), pp.593-602.

[2]   Whiteford, H.A., Degenhardt, L., Rehm, J., Baxter, A.J., Ferrari, A.J., Erskine, H.E., Charlson, F.J., Norman, R.E., Flaxman, A.D., Johns, N. and Burstein, R., 2013. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. The Lancet, 382(9904), pp.1575-1586.

[3]   Singh, T. and Rajput, M., 2006. Misdiagnosis of bipolar disorder. Psychiatry (Edgmont), 3(10), p.57.

[4]   Hirschfeld RM, Lewis L, Vornik LA. Perceptions and impact of bipolar disorder: How far have we really come? Results of the National Depressive and Manic-Depressive Association 2000 survey of individuals with bipolar disorder. J Clin Psychiatry. 2003;64:161–74.

[5]   Phillips, M.L. and Kupfer, D.J., 2013. Bipolar disorder diagnosis: challenges and future directions. *The Lancet*, *381*(9878), pp.1663-1671.

[6]   Altamura, A.C., Buoli, M., Caldiroli, A., Caron, L., Melter, C.C., Dobrea, C., Cigliobianco, M. and Quarantini, F.Z., 2015. Misdiagnosis, duration of untreated illness (DUI) and outcome in bipolar patients with psychotic symptoms: a naturalistic study. Journal of affective disorders, 182, pp.70-75

[7]   Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C. and Wang, P., 2010. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*.

[8]   Kelly, C. and McCreadie, R.G., 1999. Smoking habits, current symptoms, and premorbid characteristics of schizophrenic patients in Nithsdale, Scotland. American Journal of Psychiatry.

[9]   Wickham, H., Walsh, C., Asherson, P., Taylor, C., Sigmundson, T., Gill, M., Owen, M.J., McGuffin, P., Murray, R. and Sham, P., 2001. Familiality of symptom dimensions in schizophrenia. Schizophrenia research, 47(2), pp.223-232.

[10]   Dworkin, R.H. and Lenzenweger, M.F., 1984. Symptoms and the genetics of schizophrenia: Implications for diagnosis. The American Journal of Psychiatry.

[11]   Cardno, A.G. and Owen, M.J., 2014. Genetic relationships between schizophrenia, bipolar disorder, and schizoaffective disorder. Schizophrenia bulletin, p.sbu016.

[12]   Kurzthaler, I. and Fleischhacker, W.W., 2001. The clinical implications of weight gain in schizophrenia. *The Journal of clinical psychiatry*, *62*(suppl 7), pp.1-478.

[13]   Ncbi.nlm.nih.gov. (2016). *dbGaP | phs000021.v2.p1 | Genome-Wide Association Study of Schizophrenia*. [online] Available at: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000021.v2.p1 [Accessed 18 Apr. 2016].

[14] Ncbi.nlm.nih.gov. (2016). *NCBI Genome Workbench*. [online] Available at: http://www.ncbi.nlm.nih.gov/tools/gbench/ [Accessed 18 Apr. 2016].

[15]   Bioconductor.org. (2016). *Bioconductor - Home*. [online] Available at: https://www.bioconductor.org/ [Accessed 18 Apr. 2016].

[16]   Genome.gov. (2016). *Phenotypes and Exposures (PhenX)*. [online] Available at: https://www.genome.gov/27541903/phenotypes-and-exposures-phenx/ [Accessed 18 Apr. 2016].

Phillip Kendrick

Multi-Agent Systems for Cyber Security and Network Forensics

Phillip Kendrick[1], Natalia Criado[2], Abir Hussain[1], Martain Randles[1]

Department of Computer Science, Liverpool John Moores University[1]
Department of Computer Science, King's College London[2]

P.G.Kendrick@2012.ljmu.ac.uk, Natalia.Criado@kcl.ac.uk,
A.Hussain@ljmu.ac.uk, M.J.Randles@ljmu.ac.uk

Abstract. The high number of successful cyber attacks that occur each year evidences the ineffectiveness of conventional centralised security technologies. In this paper we present a Multi-Agent model to overcome some of the limitations of conventional approaches in the collection and analysis of digital evidence during the security process, as well as a simulator for experimentation and future development.

1. Introduction

Providing effective cyber security will require efficient and scalable solutions to meet the ever increasing number of needs found within modern expanded networks. Typical security solutions utilise a number of specialised systems (e.g., Intrusion Detection Systems (IDS) [1], [2] firewalls and forensic tool-kits [3]) often requiring high performance hardware to offset the cost of processing large amounts of data. These solutions often suffer from two problems that we will address in this paper: information overflow, which occurs when systems inefficiently collect all of the available information for bulk processing; and the failure to detect advanced stealthy attacks, which occurs when systems fail to observe the necessary information required for accurate attack analysis. By combining Multi-Agent Systems (MAS) [4] with intelligent information gathering techniques, improvements can be made by taking advantage of automated forensic investigations to proactively gather the necessary data about suspicious activity[5].

A MAS, at its most basic level, can be defined as a collection of intelligent agents. Agents are independent pieces of software often capable of working together to solve problems that could not be solved by a single agent. The following properties are characteristic of intelligent agents [4]: Autonomy, the agent's ability to act independently without any external human operator interaction, agents should not be acted upon by the environment, instead they should affect change within it. Reactivity: the agent's ability to sense environmental changes and react to the situation. Proactivity: the agent's ability to choose actions to achieve goals rather than to procedurally following instructions. Adaptability: the agent's ability to change goals in response to unforeseen circumstances. Communication *&*

---

[5] Suspicious activity is defined as any activity that does not appear to fit the norm of the network.

Coordination: the agent's ability to communicate with other agents to perform more complex tasks together.

IDS are deployed in an attempt to solve the problem of detecting malicious actors who perform actions without authorisation. Problems such as the costly requirement for high performance hardware [5], the inability to monitor all relevant information sources [6], the inefficiency of having to process large amounts of information flowing through a network [5] as well as structural vulnerabilities with centralised technologies make the current era of IDS ineffective. Besides, IDS lack mechanisms for performing automated forensic investigation and typically forensic investigations are performed manually by trained practitioners. The time gap between the detection of an attack and the ensuing forensic investigation can be detrimental to preserving digital evidence, especially when the system remains online during this time. Likewise, deciding which pieces of evidence are important to an investigation can be challenging on expanded networks, especially when evidence can be located remotely. Hence, utilising intelligent agents to respond to security events is more effective than managing teams of forensic practitioners who may be at times slow to respond.

In this paper, intrusion detection and forensic data collection techniques are combined with MAS for more effective and efficient detection in response to cyber attacks. This is performed by gathering the relevant evidence. In particular, we present a Multi-Agent model for the dynamic organisation of agents participating in forensic investigations; an agent architecture endowed with mechanisms for collecting and analysing network data; a protocol for allowing agents to coordinate and make collective decisions on the maliciousness of suspicious activity; and a simulator tool to test the proposed decentralised model, agent architecture and protocol under a wide range of circumstances and scenarios.

The remainder of this paper is organised as follows. Section 2 provides an overview of related research and open issues. Section 3 provides a definition of the proposed model. Section 4 details the simulator being used to test the model. Finally, Section 5 provides a conclusion and details on future work.

2. Related Research

In this section a number of related works are discussed to highlight desirable properties of a MAS as well as areas that will be improved upon.

Shakarian et al. [7] described a cyber attribution system [8], [9] that takes into consideration different data sources and uses MAS to reason about the origin of an attack through the use of agent reasoning. The system uses information gathered about the attack as well as information gathered from a wide range of military sources to reason in-depth about the attribution of an attack. A highlighted danger of relying on external sources of information is the trustworthiness of the source, which must be taken into consideration. The author's system used agents that could both reason about facts and make presumptions by factoring in trustworthiness for each individual source. Facts would, by default, have a higher value, while presumptions would be rated lower based on the trustworthiness of the source of information the presumption was attained from. This use of external information provided an effective way to gain extra contextual information for detected attacks but was heavily reliant on previously collected and catalogued information from military sources.

Haack et al. [10] developed a hierarchical MAS model for monitoring and reporting data within the security environment. The system was composed of a number of agent types, each with a specific task to perform, such as event monitor, alert and report for system operators. The flow of information consisted of a high level policy created by a system operator which would be disseminated to the lower level agents responsible for monitoring networked components. Alerts would be generated by the monitor agents and aggregated by a higher level agent to make decisions about potential security events. The structure of this system is inherently centralised as the information, which is transferred up and down an agent pipeline, gets processed by one dedicated agent rather than having decisions made locally by the individual agents. This system does not take advantage of the structural benefits of MAS whereby data doesn't have to be in a central location but can be processed by the lower level agents.

Jahanbin et al. [11] proposed a MAS framework for forensic information gathering which uses three types of agents for data collection, data analysis and alert generation. The authors note how the MAS paradigm is well suited to the task of forensic data collection as agents can be dispatched to areas of the network to perform collection and analysis of evidence such as log files. This system is structurally similar to Haack et al. [10] with layered agents passing information up the agent pipeline to a central agent for decision making. This central agent structure is similar to an IDS as it collects information and then makes a judgement based on that information, however, if some information is missing, the system would continue processing new information rather than actively searching missing data.

Baig et al. [12] performed a survey of the current application of MAS in a number of critical infrastructure fields including intrusion detection. Emphasis was placed on system resilience so that if the system was attacked, resulting in some agents being forced off-line, the remaining agents should reorganise themselves to continue operation. Having agents specifically designed to adapt to network changes (e.g., hosts being turned off, firewalls restricting access to a subnet or intentional compromise) was shown to be a critical consideration, especially in the security environment. Our system builds upon this requirement by making agents interchangeable such that if one agent is forced to go offline, the entire detection system will continue to function appropriately.

Mees [13] designed a MAS to detect Advanced Persistent Threats by using external data sources to lookup the origins of suspicious connections. Within the framework three agents were described: a consultation agent to evaluate the location of the IP address, an analysis agent to compare the suspect connection with previously seen traffic patterns and a third agent to attempt to distinguish between human and robot connections by performing task-specific analysis. By using agents in this way to gather external information from data sources located beyond the local network perimeter, the agents were able to gather extra information that might not have been available to traditional IDSs which, for security, do not usually make external connections. This system utilises agents capable of performing multiple tasks which doesn't take advantage of having a greater number of more specialised agents for improve scalability. Our model uses a greater number of specialised agents to encourage competition between agents where multiple actions could be taken at any given time.

3. Model Overview

Our architecture consists of a set of agents that can be placed around a network and are capable of monitoring a specific aspect of some technology for intrusion attempts. Each agent is initialised with an *action* which, when performed, can receive data as input and perform data collection to receive some output. These data collection actions are technology-specific scripts which retrieve information such as network information, host data or log entries for analysis. By implementing actions which can be executed on demand by agents, our system differs from current IDS solutions that typically receive a stream of network data to test against either anomaly or signature detection [1]. To implement this system, the concept of *requirements* is introduced to describe the constraints placed on the execution of the agent's action and *collected data* to describe the output of each agent's data retrieval action.

The system is designed to be completely decentralised to avoid introducing a central point of failure as well as improving scalability in large distributed networks. Each agent is implemented with the ability to analyse information that it collects during the data collection process. By keeping the analysis process close to the source of the data, the amount of information being transmitted around a network can be reduced lessening the need for high performance but costly hardware. Each agent can make a *local decision* about the data that it collects to decide whether it believes the data to be malicious or innocuous based on its own analysis of using either signature or anomaly detection.

To enable more complete investigations to take place a communications module is used to enable agent communication for the purpose of performing a series of *extended data collection actions,* characterised by multiple agents working together to gather more information about a network attack. The communication module allows agents to both send data and messages to an individual agent or to broadcast to many agents. This is an improvement over current IDS architectures which will typically analyse a piece of information and then make a decision on the maliciousness of the data without further investigation to find other data points support the analysis.

Following the agent's data collection and analysis process, the agent will combine its ID, the collected data and its local decision about the data into a report that will be sent to other agents. Each agent will generate its own report and combine it with the previous reports, this is then sent between agents gathering additional information at each step. At each stage the agent extracts all data from the aggregated set of reports and broadcasts them out to all agents, each agent listening will then compare its requirements with the set of extracted data to determine if its requirements are fully satisfied. If an agent's requirements are satisfied, the agent will respond by requesting the full set of aggregated reports. At each stage the full set of reports may only be sent to one agent. The agent currently holding the report set must select an agent from the set of agents which responded with a request. This model of broadcasts and requests is used to minimise the amount of data being transferred around a network. Only those agents that have their requirements satisfied will request the full data set keeping its transfer to a minimum.

Finally, after all agents that want to participate in the extended data collection task have done so, characterised by no responses to the broadcast, the last agent will calculate the calmative agent's *global decision* using the information from the aggregated set of agent reports. Various voting protocols for calculating the final global decision from each individual

agent's local decision can be used during this process. By considering the analysis of a wide range of agents we aim to increase accuracy by ensuring a sufficient amount of evidence exists before making decisions about the maliciousness of the event as a whole.

4. Simulator

To test the proposed system a Multi-Agent simulator was developed and the processes described in Section 3 implemented. A variable *analysis* is used to simulate the agent's effectiveness during the data analysis process, a high analysis score is characteristic of an agent that can effectively distinguish between malicious and innocuous security events, while an agent with a low analysis score will be less effective.

Other variables controlling the number of requirements assigned to each agent, the number of agents to be initialised, the number of simulated events and various thresholds for initialising agents can be used to control how the agents respond to events. Profiles can be defined to control agent behaviour at different stages of the security process, for example, the point at which agents must decide which agent to send the aggregated set of reports to next or when calculating the global decision.

Figure 1 shows a visual representation of the simulated extended data collection process. The figure shows a number of agents that have been initialised, with Agent-9 having been selected as the initial agent detecting the attack. The initial agent selection simulates the flagging of some information during the continuous monitoring process which all agents perform against some service or technology on the network. Each agent shown will perform the processes outlined in Section 3, with the exception of the global analysis step, which is reserved for the last agent to perform. After each of the agents has performed its simulated data collection and analysis task, the aggregated set of reports will contain an additional report containing the agent's ID, data and analysis decision about the maliciousness of the collected data. The final agent in this example will use a voting protocol to calculate the global decision for the group using the six decisions from the report.
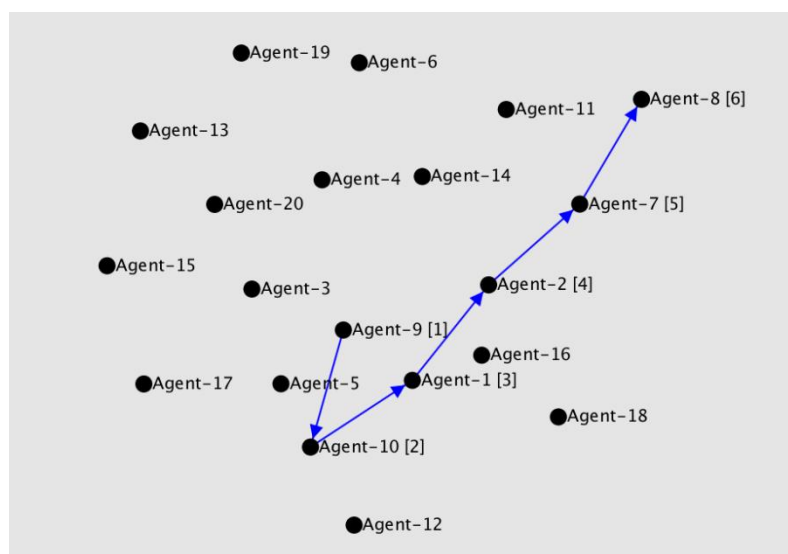
Figure 1: An example of a simulated extended data collection task taking place. Blue edges represent the transfer of the aggregated reports between the agents using the communications protocol. Numbers shown in brackets after the agent ID shows the order of execution.

To increase simulator realism, the concept of *domains* is introduced to emulate properties of the live network. We define a network to be collection of technologies and services, some of which are connected to each other by various protocols, hardware devices of software. The concept of domains is used to group types of technologies such that each type of technology is the basis of one domain. Figure 2 shows a simulated network, with each grouping of coloured nodes representing a different domain and each node representing an agent responsible for monitoring one particular part of that domain.

Within our model, technology or service specific agents will be placed around the network to perform data collection. It is assumed that agents from similar technologies will be more likely to require the data from agents from within the same or a similar technology domain. An example of this is that an agent performing analysis of a Virtual Private Network (VPN) may require information to satisfy its requirements from the general networking domains and the VPN domains, but it is unlikely that the agent would require information from an unrelated technology such as from the mail server domain. This relationship is reflected in the simulator with connected domains representing connections between the underlying protocols that they represent. Various parameters to control the number of domains, the size of the domains and the number of inter-domain links can be controlled to create networks of different shapes and sizes. We are therefore able to test the effectiveness of proposed solutions under a wide variety of scenarios including both simple and complex networks.

Simulated attacks are controlled through the domains concept with variables controlling the penetration of the attack into the network and the detectability of the attack. We are able to simulate the requirements of various real world attack scenarios by controlling these two parameters, for example, to simulate a Denial of Service attack, the penetration of the attack would target a greater number of node and be very detectable, while an Advanced Persistent Threat would spread through a subset of domains and be less detectable.
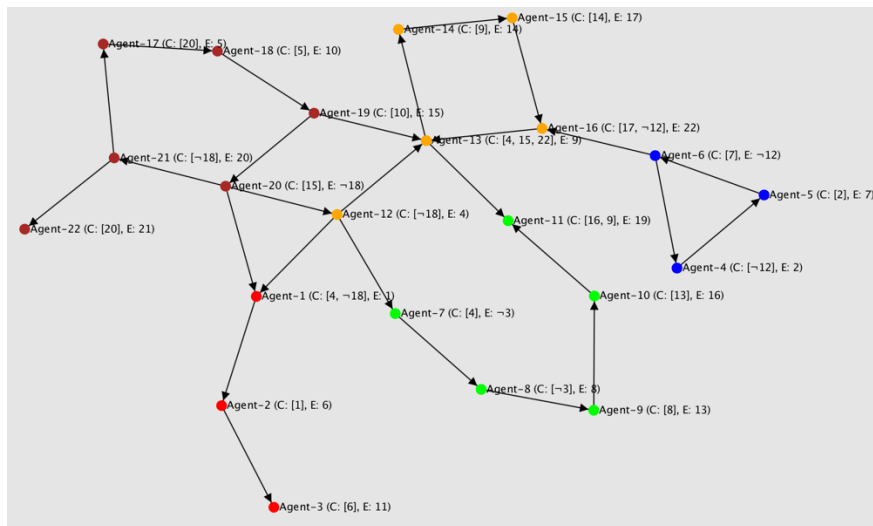
Figure 2: An example of a simulated network with agents placed at various nodes. Each of the coloured nodes represents agents associated with a particular technology or service. Domains are related technologies or services. The *C* and *E* values beside the agent names represent the agent's requirements and collected data. The edges show logical connections between the agents based on these requirements.

5. Conclusion & Future Work

In this paper we have presented an overview of a Multi-Agent model for the decentralized collection of network data as well as a simulation tool to model agent and network characteristics. Moving forward as development of both the agent model and simulator continues, we will focus on improving the simulator realism and identify additional features that could be used as the basis of improving accuracy. Future works will focus on the generation of simulated data to show improvements to the accuracy and efficiency of proposed algorithms as well as using real world datasets to test the system.

[1] B. Mukherjee, L. T. Heberlein, and K. N. Levitt, "Network intrusion detection," *IEEE Network*, vol. 8, no. 3. pp. 26–41, 1994.

[2] T. Verwoerd and R. Hunt, "Intrusion detection techniques and approaches," *Comput. Commun.*, vol. 25, no. 15, pp. 1356–1365, 2002.

[3] M. R. Clint, M. Reith, C. Carr, and G. Gunsch, "An Examination of Digital Forensic Models," *Int. J. Digit. Evid.*, vol. 1, no. 3, pp. 1–12, 2002.

[4] M. Woolridge, *An Introduction To MultiAgent Systems*, 2nd ed. Wiley, 2011.

[5] H.-J. Liao, C.-H. Richard Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 16–24, 2012.

[6] V. Corey, C. Peterman, S. Shearin, M. S. Greenberg, J. Van Bokkelen, and S. Enterprises, "Forensics Analysis," no. December, pp. 60–66, 2002.

[7] P. Shakarian, G. I. Simari, G. Moores, and S. Parsons, "Cyber Attribution : An Argumentation-Based Approach," *arXiv Prepr.*, pp. 151–171, 2015.

[8] P. Shakarian, G. I. Simari, G. Moores, S. Parsons, M. A. Falappa, E. Engineering, C. Science, U. S. M. Academy, and W. Point, "An Argumentation-Based Framework to Address the Attribution Problem in Cyber-Warfare," *arXiv Prepr.*, 2014.

[9] G. I. S. M. Shakarian, Paulo, "Belief revision in structured argumentation," *Found. Inf. Knowl. Syst.*, pp. 324–343, 2014.

[10] J. N. Haack, G. a. Fink, W. M. Maiden, a. D. McKinnon, S. J. Templeton, and E. W. Fulp, "Ant-based cyber security," *Proc. - 2011 8th Int. Conf. Inf. Technol. New Gener. ITNG 2011*, pp. 918–926, 2010.

[11] A. Jahanbin, A. Ghafarian, S. A. H. Seno, and S. Nikookar, "A Computer Forensics Approach Based on Autonomous Intelligent Multi-Agent System," *Int. J. Database Theory Appl.*, vol. 6, no. 5, pp. 1–12, Oct. 2013.

[12] Z. a. Baig, "Multi-agent systems for protecting critical infrastructures: A survey," *J. Netw. Comput. Appl.*, vol. 35, no. 3, pp. 1151–1161, 2012.

[13] W. Mees, "Multi-agent anomaly-based APT detection," *Proc. Inf. Syst. Technol. Panel Symp.*, pp. 1–10, 2012.

Mohammed Khalef

The Utilisation of Machine learning Algorithms for Medical Data Analysis and Deploying Self-Care Management System for Sickle Cell Disease

Mohammed Khalaf[1], Abir Jaafar Hussain[1], Dhiya Al-Jumeily[1], Posco Tso[1], Russell Keenan[2]

[1]Faculty of Engineering and Technology, School of Computing and mathematical sciences, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK.
[2]Liverpool Paediatric Haemophilia Centre, Haematology Treatment Centre, Alder Hey Children's Hospital
M.I.Khalaf@2014.ljmu.ac.uk, {a.hussain, d.aljumeily, F.P.Tso}@ljmu.ac.uk Russell.keenan@alderhey.nhs.uk

Abstract. This paper presents the use of various types of machine learning approaches for the classification of biomedical datasets. The growth of medical information systems has played such an important role in medical domains. Up to this date, the vast majority of hospitals and healthcare sectors continue to use manual approaches for determining the correct medication dosage for sickle cell disease. Such methods depend completely on the experience of medical consultants to determine accurate medication dosages, which can be slow to analyse, time consuming and stressful. The aim of this paper is to provide a robust approach to various applications of machine learning in medical domain problems and keep direct connection between patients and healthcare professionals. The leading motivation for such automated dosage analysis is to enable healthcare organisations to provide accurate therapy recommendations based on previous data The outcomes gained from a range of classifiers during our experiments have shown that the proposed the Levenberg-Marquardt neural network algorithm produced better results over the other range of classifiers. We base the study on different architectures of machine learning in order to maximise accuracy and performance. Using the Receiver Operating Characteristic curve, experiments results showed the following outcomes for our models, in order of best to worst: Levenberg-Marquardt neural network: 0.99, Back-propagation trained feed-forward neural net classifier: 0.993, in comparison to the Random Forest Classifiers with areas of 0.973, Support Vector Classifier: 0.675, and the Trainable Decision Tree Classifier: 0.95. A Linear Neural Network was used as baseline classifier to illustrate the importance of the previous models, producing an area of 0.85, followed by a random guessing model with an area of 0.524.

1. Introduction

Sickle cell disease (SCD) is a genetic disease, which has influence on the life expectancy due to the Red Blood Cells (RBCs) abnormality. This disease is caused through a group of ancestral disorders have impacted in a protein inside the RBC called haemoglobin. SCD disease can be easily inherited by children through the genes responsible for sickle haemoglobin (Hb S), either from both parents or from one of them measured with abnormal haemoglobin [1]. The World Health Organisation (WHO) reported that 7 million babies born each year around the world suffer either from an inherited disease or from a congenital anomaly [2]. Even more worrying is that 5% of the population are affected by haemoglobin disorders, primarily, SCD and thalassemia [3].

Recent research has confirmed the assured effects of a drug called hydroxyurea/hydroxycarbamide in association with modifying the disease phenotype [4]. The medical practice to manage SCD modifying therapy is considerably difficult for medical staff and time consuming. In order to curtail the significant medical variability presented by such difficult crisis, clinicians need to advance adherence to therapy, which is regularly poor and subsequently results in fewer benefits and elevated risks to patients. Up to the present, the new trend of machine learning classifiers is essential for the analysis of data within the healthcare domain. Machine learning offers a large number of services for prognostic and diagnostic problems in clinical societies. ML are being used for the analysis of the importance of medical parameters in terms of their integrations for prognosis, overall patient management, providing therapy and support, and the most important to predict the disease progression [5].

The reminder of this paper is organized as follows. In section 2, we illustrate the system architecture based on web-based system, while Section 3 will discuss the methodology of our experiments. The Results will be shown in Section 4, while the conclusion and future work will be shown in Section 5.

2. System Architecture based on Web-based System

This modern technology provides a proper treatment, preventing test duplication and communicating with patients during emergency situations. Technological solutions ought to be designed based on local realities and to match with local requirements in such a way that measurably contributes and is of practical use in order to achieve the main aims of healthcare development. Figure 1 shows the architecture of the proposed framework for managing and remote monitoring of the patients' diary, based on a web-based application. The proposed system consists of a web application, any smart device or personal computer, and a network coordinator. It shows a general idea of interactions, communication, organisation, channels, and the data flow within the framework. The model will be used to deliver daily feedback to healthcare consultants to provide patients with recommendations and treatments. The proposed framework is divided into two sides. The patient side is used to monitor, to store and collect data, as well as to send feedback messages to the medical specialists in association with high-risk conditions. The high-risk condition could happen when the number of heartbeats is significantly increasing or breathing becomes difficult also known as vaso-occlusive crisis[6]. The hospital side consists of a database and a decision support system. The connection between patients and medical sides is adopted through network communication environments to keep the doctors informed about the patient's condition. The proposed system with its web-based interface, is designed to offer a straight connection between patients and clinicians, it also permits the test results to be gathered from patients that were diagnosed with SCD anywhere, anytime, and on a regular basis.
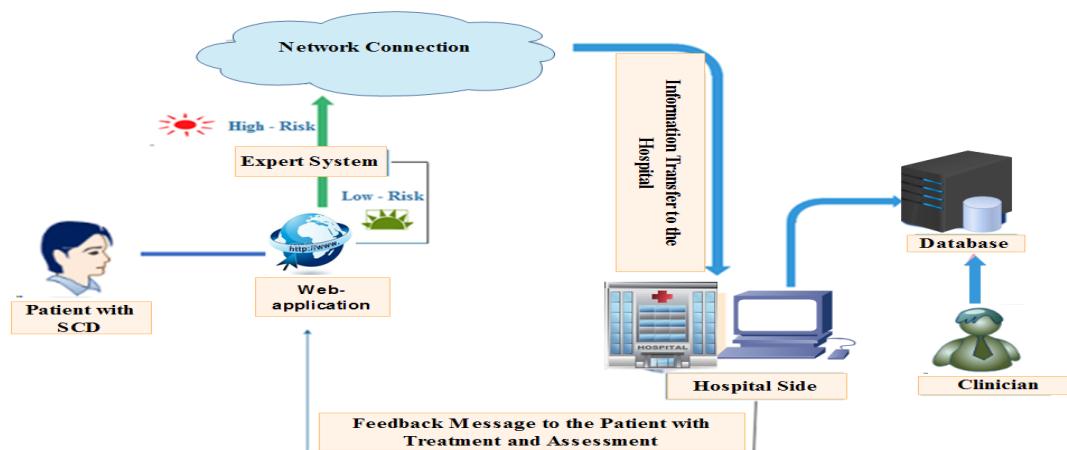
Figure 1.   The proposed System

3. Methodology

Most studies in the field of machine learning models have been built to classify the severe crises of sickle cell disease, rather than using advance predication to provide accurate amounts of a drug named hydroxyurea in modifying the disease phenotype [7]. Up to this moment, there is no standardisation of disease modifying therapy management. Using such the proposed computerised system, the aim is to produce a reproducible and optimised standard of care in different medical settings across the UK and indeed internationally. The main idea behind this study is to use advance models, a type of machine learning algorithm, in order to support the healthcare professionals in offering accurate amounts of medication for each individual patients according to their conditions. In this case, the classification of patient's datasets records at an earlier stage was proposed. This can potentially lower costs, avoiding unnecessary admission to hospitals or special institutions improve patient welfare and mitigate patient illness before it gets worse over time, particularly with elderly people. There are a number of ways we can obtain for better classification. Machine learning approaches can be applied to build strong integrated classifiers to use training and testing datasets, involving past patient observed cases that have been collected from local hospital for Sickle Cell Disease in the city of Liverpool, UK, over the last ten years for 1168 patient records.

3.1 Data Collection

The datasets that are utilised in our study for SCD were collected within a ten-year period. Each sample contains 12 features deemed important factors for predicting the SCD as shown in table 1 [8]. These features have significant effects on the blood test. In order to collect a large amount of data, the local hospital in the city of Liverpool has supported this research with a number of patient records for obtaining better accuracy and services. The resulting dataset comprised 1168 sample points, with a single target variable describing the hydroxyurea medication dosage in milligrams. To facilitate our classification study, the target dosage was discretised into 6 bins, denoted classes 1 through 6. It formed through dividing the output range (in Milligrams) into membership intervals of equal size: Class 1: [145 ≤ Y < 425mg], Class 2: [425 ≤ Y < 655mg], Class 3: [655 ≤ Y ≤ 935mg], Class 4: [935 ≤ Y < 1201mg], Class 5: [1201 ≤ Y < 1435mg], Class 6: [1435 ≤ Y ≤ 1700mg].

Table 1. SCD Datasets Features.

No	Types of Attributes
No	Types of Attributes
1	Weight
2	Haemoglobin(Hb)
3	Mean Corpuscular Volume (MCV)
4	Platelets(PLTS)
5	Neutrophils (white blood cell NEUT)
6	Reticulocyte Count (RETIC A)
7	Reticulocyte Count (RETIC %)
8	Alanine aminotransferase (ALT)
9	Body Bio Blood (BIO)
10	Hb F
11	Bilirubin (BILI)
12	Lactate dehydrogenase (LDH)

3.2 Experimental Setup

The empirical setup includes the design of the test environment utilised in our studies, the configuration of each model the models tested, and. Finally, the performance evaluation used to measure the results of the machine learning models.

This study is composed of trained models using five types of integrated machine learning approaches: the Levenberg-Marquardt neural network algorithm, Back-propagation trained feed-forward neural net classifier, the Random Forest Classifier, the Support Vector Classifier, and Trainable Decision Tree Classifier. These five models are suitable to act as comparators for the purpose of acquiring high performance [9]. The linear model used comprise a linear transformation with a single layer neural network at each class output unit. To obtain performance estimates for the respective models, we ran each simulation 50 times and calculated the mean of the responses. Finally, the random oracle model (ROM) is utilised to found random case performance by the assignment of random responses for each class.

We applied a competing models to the same classification undertaking, In order to provide a comprehensive test environment under consideration. In addition, to posing a (ROM) [10] to provide a baseline indicator to demonstrate the performance formed by random guessing. Additionally, we presented a linear model to observe the differential in performance present between this weak classifier and the non-linear classifiers. The combination of random control baselines, weak and strong, provides an experimental frame of reference through which to gauge the relative performance. We note also that such a set of reference controls is valuable to validate the integrity of the outcomes acquired, since it can be shown that such performance cannot be reached through either the linear model or by random guessing.

Table 2. Performance metric Calculations

Metric Name	Calculation
Sensitivity	TP/(TP+FN)
Specificity	TN/(TN+FP)
Precision	TP/(TP+FP)
F1 Score	2 * (Precision*Recall)/(Precision+Recall)
Youden's J statistic (J Score)	Sensitivity + Specificity − 1
Accuracy	(TP+TN)/(TP+FN+TN+FP)
Area Under ROC Curve (AUC)	0 <= Area under the ROC Curve <= 1

Our classifier evaluation consists of both out-of-sample (testing) diagnostics and in-sample (training), involving sensitivity, specificity, precision, the F1 score, Youden's J statistic, and overall classification accuracy calculated as shown in table 2. Additionally, the classifiers were characterised using the Area Under the Curve (AUC) and Receiver Operating Characteristic (ROC) plots and, where the classification ability across all operating points was ascertained.

4. Results

In this part, we analyse the results from the a number of experiments that have been implemented in this study as listed in Tables 3 and 4, presenting the whole outcomes for training and testing of the models, respectively. We run additional performance visualisations through the use of AUC comparison plots (figures 2 and 3) and the ROC plots as shown in figures 4 and 5. In order to make compassion among classifiers, the LEVNN outperformed the rest of models as demonstrated in Table 4, illustrating capability both in generalising and in fitting the training data to unseen examples. The calculated means of AUCs for the LEVNN model, obtained for six classes during training yielded an area of 1, which considered such an optimal outcomes in comparison to 0.99 over the test sample. Classes one to six were found to illustrate optimal consistent generalisation and performance from the training to the test sets for this classifier. It was discovered that the BPXNC model was able to yield an average AUC of 0.993, outperforming the RF classifier and showing an overall rank of second place. The RFC produced an average AUC of 0.973, ranking third overall, outperforming the lower ranking models by a reasonable margin. The TREEC yield an average AUC of 0.95. Eventually, SVC produced the lowest rate in comparison with other classifiers with 0.666. All four of the top performing models, the LEVNN, BPXNC, RFC and TREEC, obtained nearly ideal AUCs and represent viable candidates for future use. These models produced exceptional results in terms of both training and generalisation.

Table 3. The Performance of Classifiers (Training phase).

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
ROM	0.493	0.542	0.179	0.253	0.0351	0.543	0.488
SVC	0.89	0.636	0.511	0.603	0.526	0.677	0.666
BPXNC	0.991	0.99	0.953	0.972	0.981	0.991	0.998
LEVNN	1	1	1	1	1	1	1
TREEC	1	1	1	1	1	1	1
BPXNC	0.999	0.998	0.99	0.995	0.998	0.998	1
LNN	0.85	0.834	0.544	0.658	0.684	0.841	0.871

Table 4. The Performance of Classifiers (Testing phase).

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
ROM	0.548	0.547	0.197	0.268	0.0949	0.539	0.524
SVC	0.763	0.793	0.521	0.594	0.555	0.807	0.675
BPXNC	0.98	0.984	0.927	0.952	0.964	0.984	0.993
LEVNN	0.977	0.99	0.965	0.97	0.967	0.989	0.99
TREEC	0.913	0.926	0.684	0.737	0.839	0.921	0.95
RFC	0.928	0.931	0.681	0.77	0.859	0.929	0.973
LNN	0.83	0.845	0.543	0.65	0.675	0.84	0.85
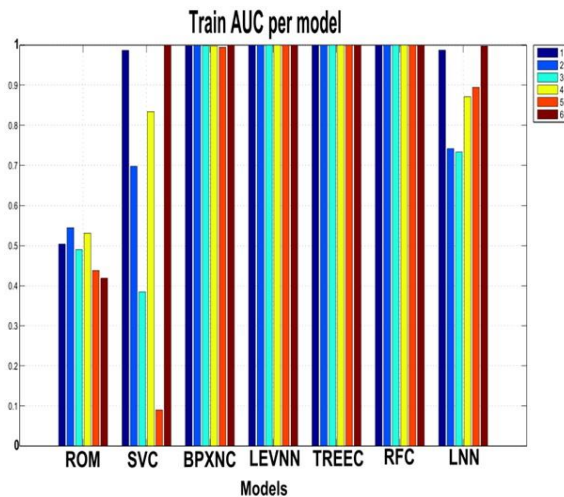
Figure 2. Train AUC per model   **Figure 3**. Testing AUC per model
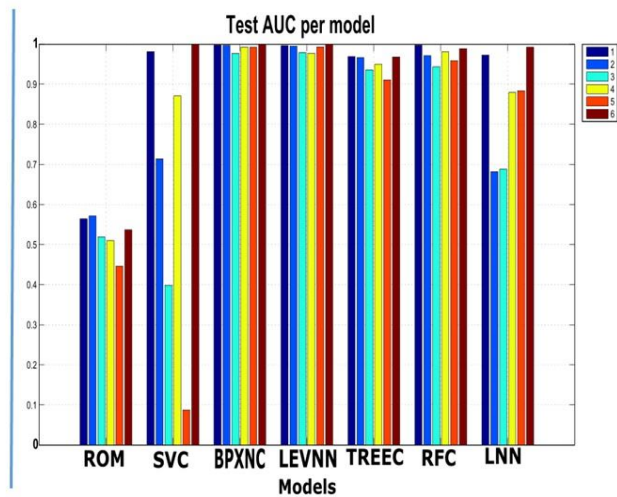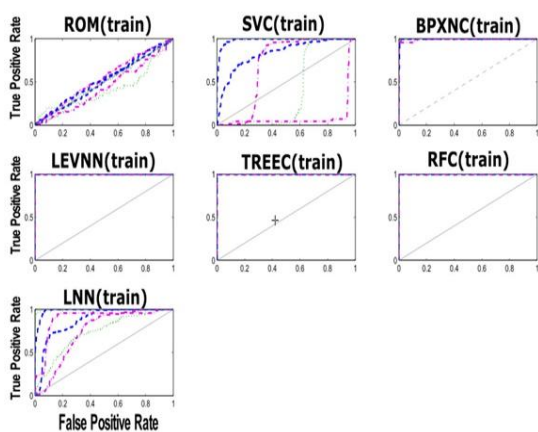
Figure 4. ROC curve (Train) For classifiers   **Figure 5.** ROC curve (Testing) For classifiers
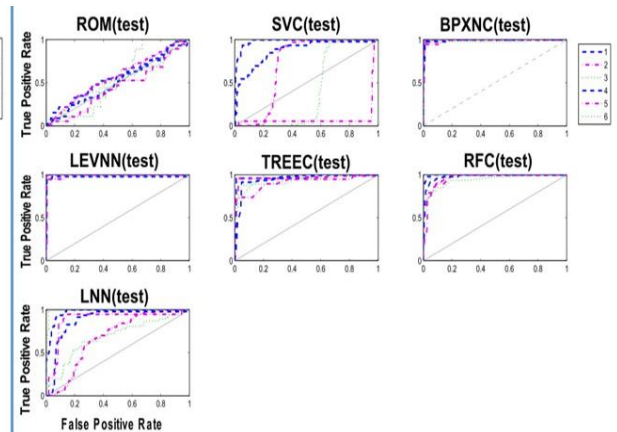
The average test of AUCs for this LNN ranged 0.85, which is seen to demonstrate performance significantly below that of the other models. Moreover, as expected, the average test of AUCs for this ROM ranged 0.524, which is considered the lowest rate among other classifier. The ROM is indicated to follow the diagonal of the ROC plots for all classes (see Figures 2 and 3), demonstrating by compare the importance of the outcomes from the rest of trained classifiers. This process of guessing yields both train and test set AUCs lower than the LNN trained model baseline.

5. Conclusion

This study presents an empirical investigation into the use of various machine learning algorithms to classify the level of dosage for SCD medication. In this paper, various model architectures are used for analysing the medical datasets obtained from SCD patients. The main purpose of this research is to examine the effectiveness of these models in terms of training and testing setting, investigating if such architectures could enhance classification results. It was found through experimental investigation, including the usage of SCD datasets and approaches such as LEVNN, BPXNC, RFC and TREEC that the analysis of medical datasets is viable and produces precise results. The results gained from a range of models during our experiments have demonstrated that the proposed Levenberg-Marquardt neural network classifier produced the best results with the AUC 0.99 in compare with other models. It is found that the outcomes are based on a considerable data sample, containing

more than 1160 sample examples, which supports the significance of the findings. It is recommended that further research will be used to make confirmation on our findings, where a large number of data could be utilised also to advance the performance of the results. In this circumstance, we recommend however that a machine learning algorithms, for instance, support vector machine and k-nearest neighbour classifiers could be used to increase the scale and scope of this research.

References
[1]   D. J. Weatherall, "The role of the inherited disorders of hemoglobin, the first "molecular diseases," in the future of human genetics," *Annual review of genomics and human genetics,* vol. 14, pp. 1-24, 2013.
[2]   D. J. Weatherall, "The importance of micromapping the gene frequencies for the common inherited disorders of haemoglobin," *British journal of haematology,* vol. 149, pp. 635-637, 2010.
[3]   D. J. Weatherall, "The inherited diseases of hemoglobin are an emerging global health burden," *Blood,* vol. 115, pp. 4331-4336, 2010.
[4]   M. Kosaryan, H. Karami, M. Zafari, and N. Yaghobi, "Report on patients with non transfusion-dependent β-thalassemia major being treated with hydroxyurea attending the Thalassemia Research Center, Sari, Mazandaran Province, Islamic Republic of Iran in 2013," *Hemoglobin,* vol. 38, pp. 115-118, 2014.
[5]   G. D. Magoulas and A. Prentza, "Machine learning in medical applications," in *Machine Learning and its applications*, ed: Springer, 2001, pp. 300-307.
[6]   H. R. Yusuf, H. K. Atrash, S. D. Grosse, C. S. Parker, and A. M. Grant, "Emergency department visits made by patients with sickle cell disease: a descriptive study, 1999–2007," *American journal of preventive medicine,* vol. 38, pp. S536-S541, 2010.
[7]   C. Allayous, S. Clémençon, B. Diagne, R. Emilion, and T. Marianne, "Machine Learning Algorithms for Predicting Severe Crises of Sickle Cell Disease," 2008.
[8]   M. Khalaf, A. J. Hussain, D. Al-Jumeily, P. Fergus, R. Keenan, and N. Radi, "A Framework to Support Ubiquitous Healthcare Monitoring and Diagnostic for Sickle Cell Disease," in *Intelligent Computing Theories and Methodologies*, ed: Springer, 2015, pp. 665-675.
[9]   W. Duch, "Towards comprehensive foundations of computational intelligence," in *Challenges for Computational Intelligence*, ed: Springer, 2007, pp. 261-316.
[10]   X.-Y. Jia, B. Li, and Y.-M. Liu, "Random oracle model," *Ruanjian Xuebao/Journal of Software,* vol. 23, pp. 140-151, 2012.

Victor Latorre Garrido

Smart Data Analysis for solving Retail Business Questions

Victor Latorre Garrido

Department of Applied Maths, James Parsons Building, 3 Byrom Street, Liverpool L3 3AF

V.LatorreGarrido@2014.ljmu.ac.uk

Abstract. As customers are the main value for companies, this research in collaboration with Tesco, focuses on answering different business questions in order to gain a better understanding of Tesco customers and identifying differences in behaviour related with their purchase channel.

In order to do this, transaction data from three different data sets have been provided for the study, two of them artificially generated with different customer behaviour and one an anonymised real data. The data used can not be considered as Big Data, but a study framework has been developed in view of applying it in bigger data sets.

K-mean segmentations have been chosen with a stability framework for selecting the best value of "k" in a Big Data environment. In addition, Recency, Frequency and Monetary Value model have been selected for ranking company's customers, obtaining a natural segmentation with this metric.

Keywords. Business Questions, Customer behaviour, Channel, K-means, Big Data, RFM.

1. Introduction

Data is growing every day and companies are storing all the information with the wish of being able to analyse it in the future. This data collected has a few different uses, sometimes companies study what customers think about their products, for international and national security issues; also information is used for customer insight and may be also used for improving profit margins or turnover. This is the starting point in which this PhD will be focused. Certainly, companies store this data for their value but it could also be a value for the people who were the source of this data.

All this data has to be analysed, but their size can make it almost impossible to use usual analytic methodology and algorithms. This data, with their volume, their velocity (creation, storing, and need to analysing) and their variety are defined as Big Data and then we need to adapt as far as we can the existent data analysis methodology in a new way of thinking based on Hadoop Servers and Map Reduce algorithms.

This is collaboration between Tesco and LJMU; where there is a mutual interest in this research. Tesco's current research focus is on the applications for single channel to multi-channel crossing possibilities. When Tesco talk about Channel they refer on their products (retail stores, on-line stores, mobile phone service, insurances services, banking services).

The main point of this study is to understand different customer behaviours depending on their shopping transactions and some cross information in order to gain knowledge and further develop the definition of a Loyalty Metric.

The setting of this multidisciplinary research has now been defined. The research questions arising from this collaboration, belong to marketing field which introduce as mentioned before the need for a definition of a Loyalty Metric. The main objective is to help Tesco better understand their customers and improve interactions between them.

To obtain a Loyalty Metric, mathematical models need to be applied and measured. This is the point that links the marketing question and the mathematical and computing fields exist. To meet our objective further analysis was needed to choose the best model for our needs. A robust segmentation model is needed to allocate customers to a prototype group and act in accordance with the insights generated from them.

The objective is to measure the behaviour of the customers. These measurements are possible because the collection of data relates to Tesco Club Card customers. Due to the size of the data it is not manageable using classical data analysis methods, so different approaches are needed in the analysis. This will be one of the main contributions of this research. All this analysis will be produced in a Big Data environment.

These are the main research questions related with Tesco needs:

• Are all retail customers the same? This question links with the segmentation of the shopping data.

• Have all the retail customers same value for the company? Open a link with the value of the customer and the loyalty measurements.

• What characterised a single channel customer in front of a multi-channel customer? Are they more valuable? [1]

The aim of this paper is to develop and present a methodology for answering these questions.

2. Methodology

In order to answer the research questions, the best approach in the three disciplines implied in this study have to be studied. Bearing in mind the rational, before starting with the marketing issues, the programming language and the segmentation framework related with the Big Data environment have to be selected. Once these decisions have been taken, the next step is to select the appropriate metric used in marketing as Customer Lifetime Value (CLV) [2] [3], Recency, Frequency and Monetary Value model (RFM) [7] [6], etc.

2.1. Programming language

Dealing with the computing language choice, R has been initially chosen as the programming language. This decision was made because the previous knowledge of the language and from some insights from the community showing an intention to support it on the Apache Spark platform.

After some months of dedicated work, the support library which initially looks to make R fit into Spark (SparkR) has been implemented in Spark just as a package to use frameworks but not with full support. Consequently, the decision to select a better supported language in Python has been taken.

2.2. Segmentation framework

For the segmentation framework, K-means has been the algorithm chosen to implement because of its simplicity and because it is already implemented in Spark.

A feature of K-means is that it is more accurate when the clusters are approximately normally distributed and isotropic, with similar within cluster variance/dispersion, which may not be a feature of the data. In addition K-means is initialisation dependent which means that it can find different cluster solutions for different initialisations of the algorithm; this will be addressed by using a separation/concordance framework. These solutions and frameworks need to be delivered within the Big Data environment and underpins the novelty in this research.

In the following subsection, a framework is described which in addition to using Sum-of-Squares error (SSQ) as a separation measure also incorporate a statistical measure of concordance between different k-means solutions that improves the k-means selection based on reproducibility of the solution.

Stability measure of best number of clusters

As mentioned, the idea of the stability measure for the best selection of the number of clusters is given by the paper [5], it introduces a framework to identify a good choice of k and within that k the most stable or reproducible solution.

The analysis is based in 2 variables:

- Measure of best minimum of the objective function (SSQ).

$$\Delta SSQ \;=\; \text{argminS} \sum_{i=1}^{k} \sum_{x_s \, in S_i} (x_j - u_i)^2 \qquad (1)$$

The highest value of ∆SSQ indicates an increase of the distance between different clusters.

- Median Internal Cramers'V which measures the association between the variables inside the clusters. Where:

$$C_v = \sqrt{\frac{\chi^2}{N.min(P-1,Q-1)}} \qquad (2)$$

$$\chi^2 = \sum_{p=1}^{P} \sum_{q=1}^{Q} \frac{(O_{pq} - E_{pq})^2}{E_{pq}} \qquad (3)$$

The method focuses on selecting the values of k which have best minimum and a value of the median internal Cramers' v close to 1. In the Figure 1 an example has been shown, in which a great value of med(CV) for low values of k, which correspond to principal components, a decrees in the concordance value appear for the value of 6 but showing a little increase in the value for k = 7.
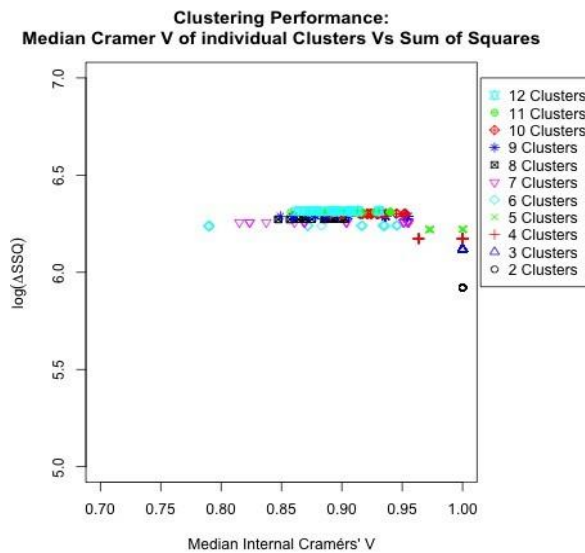
Figure 1: Example of SeCo Map

2.3. Loyalty Measure

Finally, the most relevant aspect of the proposed research to answer customer focused business questions is an RFM model that differentiates Loyal/NonLoyal customers. The main reason for this selection is because with the given information it is an easy model to apply and it's prevalent in the literature. To effectively use this RFM dataspace a simple transformation of the data is required to better implement the methodology. This decision is supported by the results displayed in the next section.

RFM Aggregation, K-means and visualization

With the aim of exploring customer behaviour, the data has been transformed to a loyalty measure in terms of Recency, Frequency and Monetary Value (RFM) for each customer. This is a well-known method in marketing for studying customer value in terms of loyalty and also a way to measure the value per customer. This measure was chosen as it is the straightest forward to obtain from the information provided in the data.

These three variables have been defined as follow:

- Recency: Time lapsed from the last purchase until the end of our window time.

$$R_i = \left| (F - l_i) \right| \qquad (4)$$

Where F is the latest date in the dataset and li is the date of last purchase for each customer.
•   Frequency: Number of times (different days) this customer have purchased on the retail, normalized over the number of days lapsed between their first purchase and the end of our window time.

$$F_i = \frac{d_i}{\left| (i_i - l_i) \right|} \qquad (5)$$

Where d_i is the number of days each customer has purchased over the whole period of time (N_{total}), i_i is the first purchase date and l_i the last purchase date for each customer.
•   Monetary Value: Total amount of money spent over over the number of days between their first purchase and the end of our time window.

$$MV_i- = \frac{Ts_i}{\left| (i_i - l_i) \right|} \qquad (6)$$

Where Ts_i is the total amount spent by each customer over all the transactions.
With our methodology clarified, next step is to define our framework. How will proceed with the data?

The kinds of data in which this study will be developed are transaction data type. As a data analyst the first step when you have a data set is to ensure that you understand what kind of data you have and check for anomalous values, so an Integrity check analysis have been produced as first step of our framework.
Once the data has been cleaned, the next step is to undertake an exploratory analysis of the data to gain some basic insight into the structure of the data. Next, results using the agreed analysis approach to investigate artificially generated data supplied by Tesco will be presented. The aim of the analysis is to define a preliminary analysis framework that can be extended to real basket data.
All the algorithms used have been designed and implemented initially with
R, using a predefined K-means algorithm for doing segmentation and the SeCo (Separation and Concordance map) framework which is used to identify the most stable k-means solution and helps identify an appropriate number of cluster centres.
A preliminary step before the segmentation algorithm is to transform the data into the RFM space previously defined. With the data into this space, and with the segmentation process applied, it will be projected on a three dimensional space with the aim to clarify the meaning of this segmentation. Further comments and interesting results come from these results presented next section.

3. Results
3.1.   Data Introduction
There are three datasets which have been supplied by Tesco. Two artificial generated datasets and a real transaction data set which have been anonymised. The time window for each of the generated data sets is around one year. For the real data set Tesco have provided with a three-month transaction data set.
Following SeCo Map for the two generated data sets are presented with their segmentation plots:
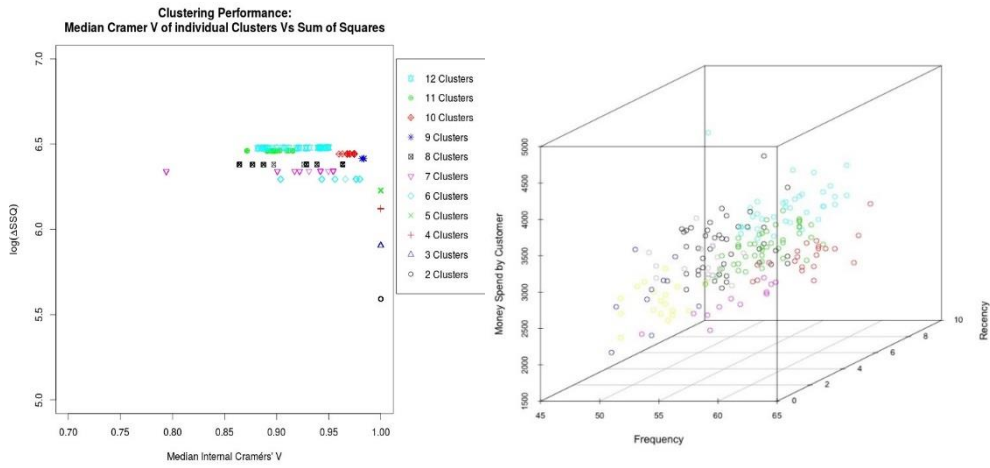
Figure 2: Metric for selection number of clusters in scenario 1 and Axis projections of the RFM with k-mean clustering in dataset 1
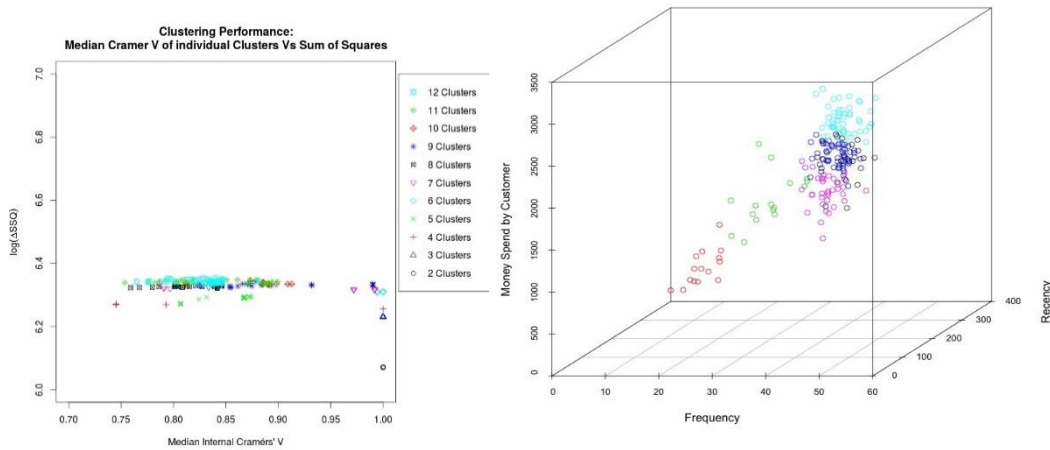
Figure 3: Metric for selection number of clusters in scenario 2 and Axis projections of the RFM with k-mean clustering in dataset 2From the Figure 3, a pattern related with Frequency and Recency was detected. Further analysis comes from this point, defining Loyal/Non-Loyal Customers using RFM model.
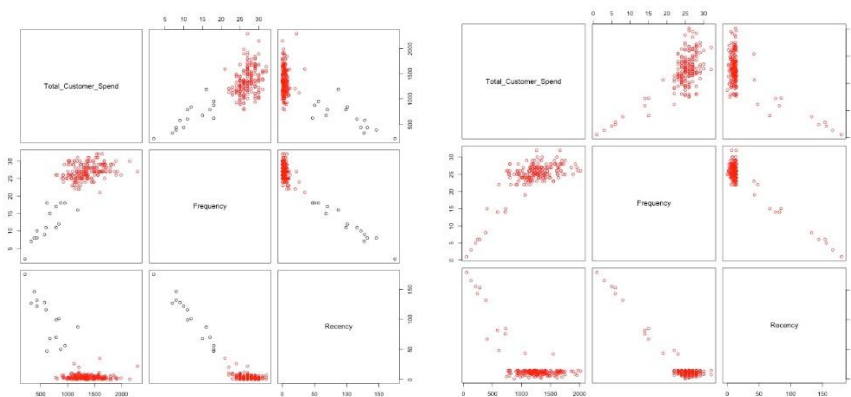
Figure 4: Comparation in dataset 2 of 1st half year customers (left) and customers 2nd half year.

In Figure 4 black points represent Non-Loyal customers and red point represents Loyal customers according our RFM model. A transition over half year is presented following the customers. The right plot presents only red points showing that every customer labelled as Non-Loyal have stop buying in the store. Data is presented in the following contingency Table 1:

Table 1: Loyalty Contingency Table

		2nd Half Year		
		Loyal	Non-Loyal	Lapsing
1st Half Year	Loyal	167	10	6
	Non-Loyal	0	0	15

With this figures conditional probabilities of transition have been defined between one group to the others and Likelihood ratios gives company information about the behaviours of customers depending their frequency and recency.

4. Conclusions

This analysis presents customer insights from Tesco artificially generated customer datasets. The same methodology has also been applied to real-world customer data set giving Tesco insights into customer transitions on transaction behaviour using Likelihood Ratios, a common metric for comparing specific customer behaviour over a baseline population. These results link with the business questions about if every customer is the same within an overall customer population.

On the third data set, in addition to the information given in the previous data, channel information has been provided. This extra information allows a similar analysis using now conditional probabilities same way as used for transition between Loyal/Non-Loyal. This way the customers could be analysed depending the used channels, giving the company the link answer for the profitability question for multi-channel customer in front of single-channel customer.

5. References.

[1]   Jill Avery, Thomas J Steenburgh, John Deighton, and Mary Caravella. Adding bricks to clicks: Predicting the patterns of cross-channel elasticities over time. Journal of Marketing, 76(3):96–111, 2012.

[2]   Yun Chen, Chuan Fu, and Hanhong Zhu. A data mining approach to customer segment based on customer value. In Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on, volume 4, pages 513–517. IEEE, 2008.

[3]   Sunil Gupta, Dominique Hanssens, Bruce Hardie, Wiliam Kahn, V. Kumar, Nathaniel Lin, Nalini Ravishanker, and S. Sriram. Modeling customer lifetime value. 9(2):139–155, 2006.

[4]   Chaoqun Han and Shuiqing Yang. Customer continuance in the context of multi-channels: An internal transfer perspective. In Computer Research and Development (ICCRD), 2011 3rd International Conference on, volume 3, pages 77–81. IEEE, 2011.

[5]   Paulo J. G. Lisboa, Terence A. Etchells, Ian H. Jarman, and Simon J. Chambers. Finding reproducible cluster partitions for the k-means algorithm. BMC Bioinformatics, 14(S-1):S8, 2013.

[6]   Baolong Ma, Fei Li, Gao Wang, and Lun Ran. Targeting valuable customers within a retail reward program database by rfmg model. In Management and Service Science, 2009. MASS'09. International Conference on, pages 1–4. IEEE, 2009.

[7]   Zhang Yihua. Vip customer segmentation based on data mining in mobile-communications industry. In Computer Science and Education (ICCSE), 2010 5th International Conference on, pages 156–159. IEEE, 2010.

John Melthis

Topologies for combining the internet of things and serious games

John Melthis[1], Stephen Tang[2], Martin Hanneghan[3], Chris Carter[4] and Po Yang[5]
Department of Computer Science,
Faculty of Technology and Environment,
Liverpool John Moores University,
James Parsons Building, Byrom Street, Liverpool, L3 3AF, UK.
Email: j.melthis@2015.ljmu.ac.uk[1], o.t.tang@ljmu.ac.uk[2], m.b.hanneghan@ljmu.ac.uk[3],
c.j.carter@ljmu.ac.uk[4], p.yang@ljmu.ac.uk[5]

Abstract—**Serious Games have been established over recent years as a means of utilising gaming for applications other than entertainment. With the emergence of the *Internet of Things* (IoT) paradigm, a new direction for serious games arises, where data gathered from the physical environment can be utilised towards new novel applications.**

This literature survey uncovers existing topologies that can be applied for combining IoT with Serious Games. This paper presents findings from extensive research into IoT, Serious Games, Pervasive Games and Gamification, IoT topologies and *Wireless Sensor Networks* (WSN), to identify the requirements of a topology for Serious Games and IoT. By understanding the topological requirements for combining IoT and Serious Games, the development process is reduced, allowing for the advancement in the mentioned field.

Three topologies are presented for combining IoT with Serious Games and a detailed topology for developing a Serious Game that monitors student attendance is presented. Also included, is an insight into the new paradigm of *Smart Serious Games* (SSGs). This paper will aid future research and development in SSGs determine effective network topologies.

Keywords— IoT; Serious Games; Topology; Smart Serious

Games;

INTRODUCTION

IoT promises a future of interconnectivity between heterogeneous devices and data services. This interconnectivity will allow for better analysis of data driven applications and prospects new software solutions that could not have been achieved without IoT. IoT consists of interconnected devices or Things that operate in Smart Environments and communicate data with virtual identification and/or personalities [1]. In addition, IoT accounts for an ecosystem which is comprised of middle-ware [2], users and interconnected devices.

In a truly interconnected IoT environment, Serious Games could harvest and analyse data from players' physical worlds and present it so behaviour can be positively manipulated. The combination of Serious Games and IoT has recently been termed as Smart Serious Games (SSGs) [3]. SSGs have been defined as the merger of smart technologies, including devices and services, and the principles of Serious Games [3]. This literature details the combination of the advantages of both technologies and its future utilisations including; analytics for industry, a tool for solving serious problems and others.

As this research area is new, literature on SSGs is limited, with some research project beginning to include the term as future works [4], [5].

To develop a SSG, research into IoT, data analytics and Serious Games is required. An area of research that must be addressed are topologies for developing a Serious Game combined with IoT. Research into IoT topologies [6] and game topologies [7] exist, however none has

been conducted on appropriate topologies for SSGs, nor have current topologies been evaluated for this purpose. This paper focuses on this area; specifically, topologies for SSGs, and surveys existing topologies that can be utilised, based on the latest research. By determining the appropriate topologies for combining Serious Games and IoT or SSGs, a foundation for future development in the research field can be achieved, as it is vital to consider the network topology of any research principle that merges with the IoT paradigm.

The paper follows the following outline. Section II presents current topologies, utilised in a variety of applications and analyses the effectiveness of incorporating these topologies for SSGs. Topologies have been sourced from networking games, WSNs and Human - centric WSNs. Requirements for SSG topologies are outlined in Section III. Topologies for SSGs are presented in Section IV, outlining three different types of topologies that can be used for combining Serious Games and IoT. Section V discusses a potential SSG application and illustrates its topology. The conclusive remarks are made in Section VI.

CURRENT TOPOLOGIES

When considering a topology for an SSG, it is not necessary to develop one from the ground up. This section surveys existing topologies that can be utilised or adapted to accommodate for SSGs.

Client Server and Peer-to-Peer Topologies

The client - server architecture provides an example of existing topologies, where principles can be applied for a SSG. In client - server architectures, clients require a service from the server. Clients cannot communicate directly, but rather send messages through the server [7]. A large limitation of the client server architecture is its focus on the server. If the server was to become unresponsive and disconnect, the whole network would become redundant until the server can either be repaired or replaced.

Figure 1 displays a client - server architecture and a peer-to-peer client - server hybrid topology that is adapted for the IoT ecosystem. The proposed architecture will share elements of a peer-to-peer architecture but will fundamentally

differ in the roles clients have within the network. Clients will not be formed solely by computers but rather by a combination of sensors and computational devices. An additional difference is the intercommunication between clients. Clients will need to communicate with each other directly.
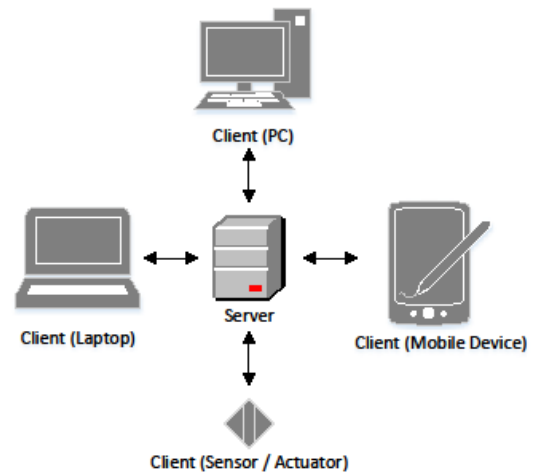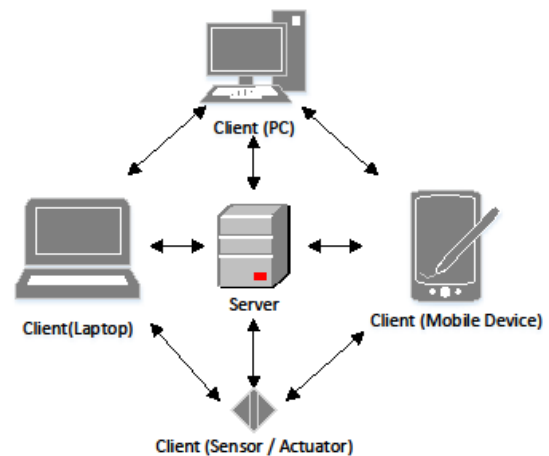
Figure 1: (a) + (b)

(a) Client-server architecture, (b) Peer-to-peer Client- server hybrid

The adoption of peer-to-peer architectures for SSGs is another topology for consideration. Peer-to-peer topologies offer greater scalability over client-server architectures, due to a better distributed network load [8]. IoT is expanding rapidly, with continuous developments in low energy sensor devices and data analytics. To accommodate for the hardware

included in the future of IoT ecosystem, scalability should be considered as a key element. This favours the use of elements from a peer-to-peer topology that harness scalability for a SSG.

The MOPAR project developed a hybrid architecture for online games that focuses on interest management, utilising peer-to-peer mobile technologies [9]. The aforementioned project did not produce a topology but resulted in a hybrid solution with key advantages in scalability and fault tolerance over previous approaches, values that prove beneficial for a SSG game that utilises mobile technologies.

Cloud Topologies

The research of D. Misha et.al. into the utilisation of cloud architecture for online games, suggested a two-tier cloud architecture where Tier 1 is formed by public cloud services and Tier 2 consists of the servers that are closest to any given gamer [10]. The paper of D. Misha et.al. did identify limitations in their proposition related to prising and potential security risks stemming from the cloud. A full cloud integration may prove beneficial for SSGs due to the scalability and decentralised nature of such a topology. Current cost restrictions favour a hybrid architecture, particularly for research purposes.

Cloud-centric IoT presents the idea of cloud computing forming the core of IoT with users, sensor networks, middleware and private clouds completing the paradigm [6]. The framework J.Gubbi et.al. developed creates a scalable IoT for multiple applications. When considering IoT from a scaled back perspective, such a representation becomes accurate. For SSGs consideration must be given to smaller infrastructure also, as not all event driven games that require IoT may require the cloud. For instance, a serious game may operate with a localised server, using sensors and actuators placed in a single room and open source data which fundamentally affect gameplay.

Wireless Sensor Network Topologies

When utilising WSN in an IoT solution, the level of integration required must be determined. C. Alcaraz et. al. have outlined Two approaches for integrating IoT and WSN: Stack-Based and Topology-Based [11].

Focusing on topology based approaches, two solutions were outline: Hybrid and Access Point. Hybrid solutions feature networked nodes that can connect to the Internet autonomously, providing the network with redundancy and intelligence. Access Point Solutions rely on a sink node for connectivity to the Internet, meaning nodes at the edge of the network have less capabilities and specify on the data attempted to be captured e.g. movement [11]. The research of C. Alcaraz et. al. confirms that the potential of WSN is unleashed when integrated into IoT.

Developing a large scale WSN is currently challenging due to the variety in domains, vendors and standard [2]. From the Serious Game context, an interconnected WSN creates new possibilities for gamifying new scenarios and creating pervasive games that interact with the physical and virtual environment.

An effective WSN topology should be scalable, energy efficient and reliable [1]. D. Zhang et. al. introduced a new approach for a weighted network topology in WSN which identified sink holes amongst sensors, and defined a new edge weight [12]. By utilising the results generated, energy holes can be avoided in IoT networks. The scale of a SSG network would form the likelihood of energy holes. For further research a localised hybrid topology will be utilised, detailed in the following section. Larger SSG networks that gather a greater amount of real-world data should utilise a weighted- IoT topology.

Consideration must be given to Human-centric Wireless Sensor Networks (HWSN) when considering a topology for SSGs. Unlike WSNs, HWSNs require that every networked node sends information to a Human Based Sensor (HBS) [13]. This type of topology would be ideal for games that monitor specific body parts or require body scanning, such as brain activity. HWSN topologies can benefit games that require players' current location.

In Section III, a topology is defined for a Serious Game that monitors student attendance using sensory input and interacts with web data services and data sourced from sensors, in order to alter gameplay.

REQUIREMENTS FOR SMART SERIOUS GAMES

A topology for games must consider data packet loss, latency and jitter to

create a fluid gaming experience [7]. The requirements for SSG topologies have not yet been defined in current literature. We propose the following key requirements for an SSG topology, based on the conducted research:

Scalability: Scalability is an ambiguous concept, as it is service specific. For a scalable SSG topology, scalability must consider WSN and gaming network attributes. To incorporate IoT into a game data must be sourced through sensors, either directly or indirectly. By indirect sourcing of data, we consider open datasets available on the Internet. To account for the network demands of a sensor based solution, WSN topology requirements are essential.

A scalable WSN topology allows nodes to be added without major reconstruction or development [1] by accounting for the node density of a given topology [14]. Furthermore, consideration is given to the energy efficiency of an WSN topology. To achieve efficiency, the least amount of power should be utilised for the network to operate reliably, and sleep modes should be operational where possible for nodes that are not constantly active [1].

Aside from the sensory requirements of a SSG topology, the gaming network demands must be met. From a gaming perspective, a scalable network must adapt to peak demand and peer bandwidth [15]. Depending on the size of a player community, demand can soar during off-school hours and significantly affect gameplay. Peer bandwidth will affect data demanding games that require large files to be sent back and forth from a server. Contrary to a traditional peer-to-peer network, a SSG topology must consider mobility in nodes [16] as household devices adopt more sensory equipment.

Quality of Service (QoS): SSG topologies must adhere to Quality of Service (QoS). The required components to achieve QoS for SSGs can be sourced by considering WSNs and networks for online games.

From a WSN perspective, QoS can be achieved by considering the low power consumption attribute of sensor networks [14]. Research by D. Chrstin' et.al, recommends heterogeneous sensors contribute to assure QoS by optimising network resources [17]. When considering the gaming element of a SSG topology, a lack of QoS can detrimentally reduce the effectiveness of an application, for example, latency on realtime, real-world data required to effect gameplay. Networks for online games implement parameter-based and prioritybased QoS [7], to ensure a reduction in latency and provide a fluid gaming experience.

By incorporating the aforementioned elements of WSN and online game networks for QoS, a blend that is sufficient for accommodating future SSG applications that support multiplayer and data from sensor and mobile devices is generated.

Security: The security of a topology can be divided into two categories; physical and network. The physical placement of sensors and actuators must consider vandalism and burglary, particularly when expensive hardware is involved. Security of smart devices must also be considered as they become embedded into SSG solutions. Major manufacturers offer services such as remote wiping to prevent unauthorised access, however SSG solutions should offer players the ability to remove their own devices from a network to maximise security. Allowing users to manipulate their own smart devices on a network can also improve scalability.

From a networking perspective, sensors and actuators with Internet connectivity must utilise protocols and encryption mechanisms that prevent malicious attacks and theft of personal data. Due to the low power nature of sensors and actuators, research has underline the necessity of new security mechanisms that can operate at low power [17]. Further security requirements will be service specific.

IOT TOPPLOGIES FOR SMART SERIOUS GAMES

Interconnected Access Point Topology: For SSGs, three different topologies can be considered. The interconnected access point topology, illustrated in Fig. 2, is better utilised for solutions that involve multiple users and locations such as an augmented reality, Massive Multi-player Online (MMO) game. This is due to the small clusters of sensors and actuators formed with the sink node, allowing a number of low powered nodes to be placed in a variety of locations.

Father.io [6] examples an augmented reality MMO game that could utilise an interconnected access point topology. Focused on merging smart devices with Personal Computer (PC) gaming, Father.io utilises smartphones to augment players' physical perspective and enable them to engage in an MMO shooter. The video trailer showcases a smartwatch utilised in-game alongside a smartphone. To achieve this data must flow between smart devices to allow for a fluid gaming experience. Data could be synced via a sink node, or the smartphone could be utilised as such, however devices could also be connected autonomously, as illustrated in the hybrid topology. Father.io is currently a concept but presents an insight into the future of IoT games.

In all suggested topologies, sensors describe devices that detect player activity e.g. location, movement etc. and transmit it either autonomously or through a sink node. Actuators detail interconnected devices that provide player feedback as a means of transcribing a game event e.g. level completion, confirmation of game event etc. Nodes comprise of actuators or sensors in all suggested topologies.

Fig 2: Interconnected – Access Point Topology.

6
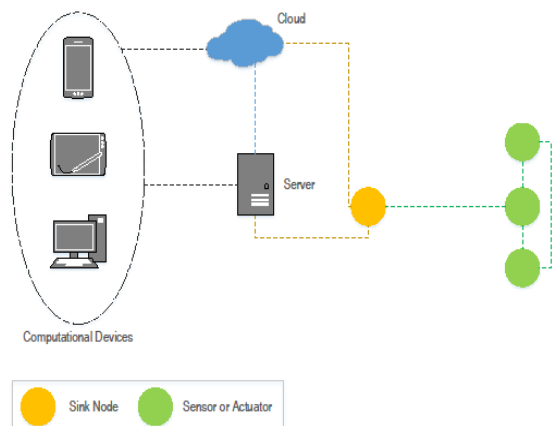
https://www.indiegogo.com/projects/father-io-massive-multiplayer-lasertag/,accessed 30/03/2016

Peer-to-peer Topology:

Game engines provide the core technologies that aid developers in producing a variety of games rapidly. These components are used to handle tasks such as user input from human interface devices and Graphical User Interfaces (GUI), simulating motion, artificial behaviour and physicality in virtual environments, updating internal states of game objects and rendering visuals on the screen [9].

Utilising a peer-to-peer topology for SSGs should consider the size and demand of the network required before being deployed. Dependant on the game requirements, a SSG could utilise this topology with no central server, but rather through computational devices that store and distribute game sections and interact with interconnected sensors and actuators based on a player's location, as illustrated in Fig. 3. Such a topology could be utilised for SSGs that are not resource demanding.

Although the processing power of smart devices is constantly increasing, they are still unable to store and process large virtual world data. Revisiting Father.io as an example, this game would suffer in a peer-to-peer topology due to the vast amount of location data that would need to be stored on smartphones and smartwatches. It could be argued that in the not so distant future, mobile smart devices may well be capable of doing so, however in the near future peer-to-peer topologies in SSGs should consider games with limited virtual environment.
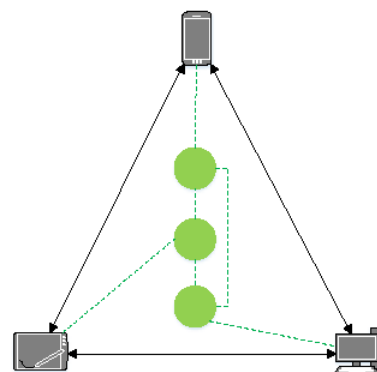
Fig 3: Peer-to-peer Topolgy.

Hybrid Topology:

Digital Content Creation (DCC) tools are used to produce all the assets such as 2D graphics, 3D models and audio for the purpose of virtual world scene

composition. For 2D graphics, tools such as Adobe Photoshop are used to create the GUI and textures that will be used for the 3D models. The creation of 3D assets requires the use of 3D modelling tools such as 3D Studio Max and Maya. These tools also provide the facility to create key-framed 3D animation which can be ported to the game engine. Alternatively, 3D assets can be produced using digital sculpting tools such as Mudbox and ZBrush which provide more fluid manipulation of vertices to ease the creation of form and textures for organic-shaped models [11].

A hybrid WSN topology (Fig. 4) for SSGs would feature autonomous nodes, removing the sink node from the network architecture. By autonomous, this paper describes nodes with Internet connectivity. In this case the location of the nodes is dependent on the game requirements. Numerous locations can be used, providing that all positions provide access to Internet and power. Power management is a key element in hybrid topologies as

the inclusion of connectivity requires more power for each node attached to the

network. Hybrid topologies can be utilised for a variety of SSGs due to its versatility with deployment. As previously mentioned, Father.io could utilise a hybrid topology to provide all smart devices with autonomous connectivity. This improves the network robustness as removes the dependency of a sink node.
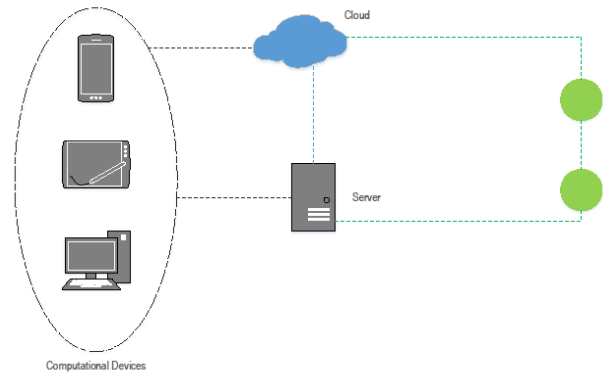
Fig 4: Hybrid Topology.
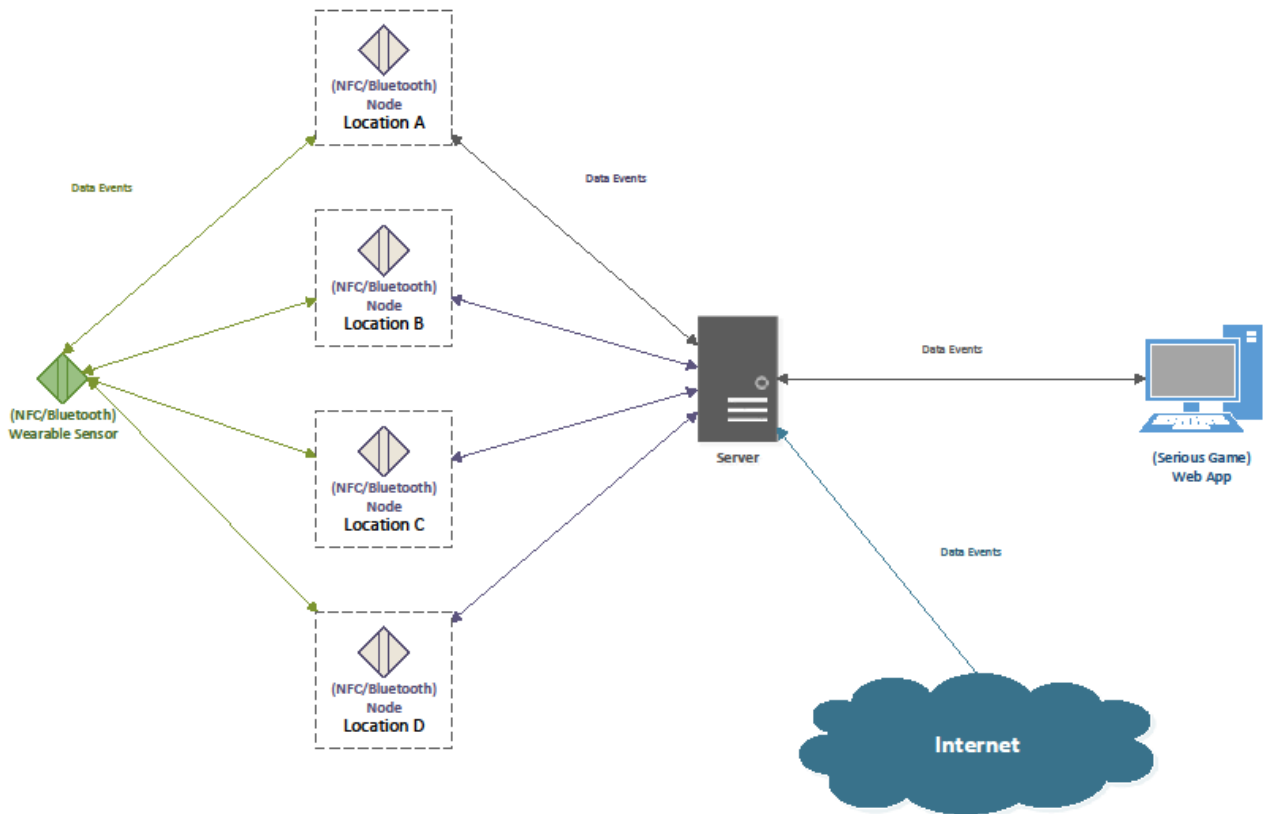
Fig 5: Topology towards an SSG
for behavioural modification

It is important to adhere the requirements discussed including Scalability, QoS and Security, whilst considering the topology's robustness and reliability when utilising any of the aforementioned topologies.

A key difference between the use of topologies for SSGs and other uses such as WSN can be noted in the data that flows through the network. When considering a SSG topology, the data sourced from sensors and sent to actuators can be described as events. This is because any data that derives from a sensor or open source web datasets would inflict a change or an event in gameplay. E.g. a Serious Game that monitors student engagement on a course would input a location check-in and translate it into a gameplay event such as user score, level up, increased inventory etc.

APPLICATION OF SSG TOPOLOGIES

The following section discusses a proposed application for a SSG. This SSG aims to promote positive behavioural manipulation by increasing student engagement and illustrates its accompanied topology (Fig. 5). Games that utilise environmental and user data, sourced through smart devices and services to affect gameplay, could benefit from the proposition. The SSG we propose to develop, to test the effectiveness of the topology, will utilise elements of question and answer to encourage players' attention and focus. The content to be included for this game was influenced by the taxonomy published on Serious Games for Health issues. This taxonomy states for increasing attention and focus, question and answer games are preferred, and for behaviour modification, simulation and virtual games could be utilised [18]. Examples of questions that could be included into the SSG are:

1) Would you like to see how many points you need until the next level?

2) What do you believe is your monthly attendance percentage?

3) You did not attend yesterday's meeting. Would you like to reschedule one in your calendar?

Following, the elements of the topology are outlined and their purpose explained.

Wearable Sensor: WSN is chosen for this topology instead of Radio-Frequency Identification (RFID), to eliminate the need of a reader device [1] and create a more pervasive experience. It is important to aim for a pervasive solution as research has proven an increase in participant's perception for such solutions [19]. To track players location and attendance a wearable sensor will be provided with either Near-Field Communication (NFC) or Bluetooth capabilities. Further research will aim to justify the use of Sensors over RFID or NFC and validate the proposed topology, by cross-examining the two technologies.

Nodes: These nodes will utilise the campuses extended Wi-Fi coverage, relaying information to the central server, located on campus. In the SSG, registering check-ins into locations across the campus will be achieved through nodes with NFC or Bluetooth capabilities. Utilising players smartphones was considered to eliminate the need for static nodes, however due to battery life limitations on smartphones when utilising Bluetooth, we opted away from this solution. Cloud computing was considered for this project, but was not deemed vital in determining the effectiveness of SSGs' for behavioural modification.

Web Datasets: Datasets will be comprised of real world data in an attempt to correlate daily events with players behaviour. If such correlation can be achieved, the game will be able to better predict behaviour and encourage engagement in advance. An example of web datasets to be used include, sun cycle, weather and others. The number of datasets to be included greatly relies on availability of data.

Web Application: For behavioural modification instead of a game, based purely on simulation, we propose an avatar, 2D web application, which levels up and increases inventory based on users performance. As mentioned earlier, this decision derives from the Games for Health taxonomy [18].

Data Events: The data flowing through the illustrated topology is considered as data events. This is because all data generated and utilised (check-in, weather) will cause an immediate effect to the game. E.g. Reactive response: unlock inventory item for customisation of avatar,

Proactive response: encourage players on a rainy day. Figure 4 notes a bi-directional relationship between all components of the uggested topology. By utilising bi-directional communication check-ins can be confirmed and the error percentage of data generated can be reduced. Errors can be reduced by ensuring data events have successfully been read and providing a form of feedback to players, such as haptic.

CONCLUSION

To summarise, this paper presents the topologies that can be utilised for SSGs and a potential application of SSG, accompanied by the proposed topology.

After researching through the relative scientific fields and paradigms, three topologies were suggested; hybrid, interconnected, access point and peer-to-peer. These topologies would cater for most games that utilise sensory input, distributed over a network.

The potential application of SSGs for manipulating behaviour highlights the inclusion of real-world datasets for understanding the effects of everyday events on behaviour. A topology applied towards a SSG for behavioural manipulation is presented.

We accept that this research does not provide a complete illustration of all topologies and topological elements that can be utilised for developing SSGs and encourage further research to enhance this work. It is important to note, due to the new presence of SSGs in research, validation of the propositions is required through experimentation. The suggested application will validate the proposed topology in future works and findings will be published.

REFERENCES

L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey,"Computer Networks, vol. 54, no. 15, pp. 2787–2805, 2010.

Z. Khalid, N. Fisal, and M. Rozaini, "A Survey of Middleware for Sensor and Network Virtualization," Sensors, vol. 14, no. 12, pp. 24046–24097, 2014.

D. Sharma, J. C. Lakhmi, M. Favorskaya, and R. J. Howlett, Fusion of Smart , Multimedia and Computer Gaming Technologies. Springer International Publishing, 1 ed., 2015.

E. Lotfi and B. Mohammed, "Players Performances Analysis based on Educational Data Mining Case of Study : Interactive Waste Sorting Serious Game," International Journal of Computer Applications, vol. 108, no. 11, pp. 13–18, 2014.

L. Elaachak, A. Belahbibe, and M. Bouhorma, "Towards a System of Guidance, Assistance and Learning Analytics Based on Multi Agent System Applied on Serious Games," International Journal of Electrical and Computer Engineering (IJECE), vol. 5, no. 2, pp. 344–354, 2015.

J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," Future Generation Computer Systems, vol. 29, no. 7, pp. 1645–1660, 2013.

G. Armitage, M. Claypool, and P. Branch, Networking and Online Games. Wiley-Blackwell, 2006.

A. Yahyavi and B. Kemme, "Peer-to-peer architectures for massively multiplayer online games," ACM Computing Surveys, vol. 46, no. 1, pp. 1–51, 2013.

A. P. Yu and S. T. Vuong, "MOPAR : A Mobile Peer-to-Peer Overlay Architecture for Interest Management of Massively Multiplayer Online Games," Acm, pp. 99–104, 2005.

D. Mishra, M. El Zarki, A. Erbad, C. H. Hsu, N. Venkatasubramanian, and M. E. Zarki, "Clouds + Games: A multifaceted approach," IEEE Internet Computing, vol. 18, pp. 20–27, 2014.

C. Alcaraz, P. Najera, J. Lopez, and R. Roman, "Wireless Sensor Networks and the Internet of Things : Do We Need a Complete Integration ?," 1st International Workshop on the Security of the Internet of Things (SecIoT'10), no. October 2015, pp. 1–8, 2010.

D.-g. Zhang, Y.-n. Zhu, C.-p. Zhao, and W.-b. Dai, "A new constructing approach for a weighted topology of wireless sensor networks based on local-world theory for the Internet of Things (IOT)," Computers &

Mathematics with Applications, vol. 64, pp. 1044–1055, sep 2012.

S. F. Ochoa and R. Santos, "Human-centric wireless sensor networks to improve information availability during urban search and rescue activities," Information Fusion, vol. 22, pp. 71–84, 2013.

I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," Computer Networks, vol. 38, no. 4, pp. 393– 422, 2002.

G. Schiele, R. Suselbeck, A. Wacker, J. Hahner, C. Becker, and T. Weis, "Requirements of Peer-to-Peer-based Massively Multiplayer Online Gaming," Seventh IEEE International Symposium on Cluster Computing and the Grid CCGrid 07, pp. 773–782, 2007.

"From today's INTRAnet of things to a future INTERnet of things: A wireless- and mobility-related view," IEEE Wireless Communications, vol. 17, no. 6, pp. 44–51, 2010.

D. Christin, A. Reinhardt, P. S. Mogre, and R. Steinmetz, "Wireless Sensor Networks and the Internet of Things : Selected Challenges," Proceedings of the 8th GI/ITG KuVS Fachgespr¨ach Drahtlose Sensornetze, pp. 31–34, 2009.

D. H. Goh, R. P. Ang, and H. C. Tan, "Strategies for designing effective psychotherapeutic gaming interventions for children and adolescents," Computers in Human Behavior, vol. 24, no. 5, pp. 2217– 2235, 2008.

Z. Lv, A. Halawani, S. Feng, S. ur R´ehman, and H. Li, "Touch-less interactive augmented reality game on vision-based wearable device," Personal and Ubiquitous Computing, pp. 551–567, 2015.

Anthony Pich

Education theory informed development of AI assisted serious gaming, in the context of eLearning

Anthony Pich

Department of Computer Science

Abstract

Adaptive learning systems aim to address the individual learner by acknowledging the differences in learning styles and how content delivery should be nuanced accordingly. This study aims to combine the three key concepts of open learning systems (OLSs), serious learning games and education theory driven machine learning into one unified architecture for the most effective delivery of course materials.

Introduction

Effective learning is the ability of a learner to memorise information and cognitive concepts and generalise the information and concepts, so that they can be applied to scenarios that differ from when they were learned. In order to do this, the mind of the learner must be reverse engineered and understood, using an interface that can permit this kind of analysis. The learner can then be categorised based on how their psychological traits correspond to a set of criteria. Once the learner is effectively categorised, the learning resources can be adapted to deliver content according to the defined traits.

The core aims of this study are to formulate an effective communication link between the learner, the game and the learning system, where adaptations need only occur within the system as opposed to the learner. The purpose behind moving the adaptation from the learner to the system is to lessen the learning curve required for the learner to be effectively taught.

The system that this research proposes will require machine learning and statistical analysis for categorisation and rating. For this reason, an effective data mining and analysis element to fit in with the Felder Silverman Learning Style Model ILS (Index of Learning Styles) query structure will be developed, which can both categorise the learner and the learning resources according to the scales; Active/Reflective, Sensing/Intuitive, Verbal/Visual, and Sequential/Global.

Background and Related Works

With current research surrounding the individual benefits of open learning systems, learning through computer games and adaptive learning models, there is a wide range of well refined ideas regarding the best approach to forming solutions. However, pre-existing solutions, such as Moodle or ComputerCraftEDU only address two of the three necessary components (open learning system, learning games and adaptive models). The three concepts of adaptive learning models, learning with games and open learning systems are addressed in this study with the aim of looking into those disciplines and the novel approach of combining and structuring them to not only assist, but drive the absorption of critical materials in the learning context.

Learning Styles

The current state of teaching and existing learning resources mean that the student is responsible for having to adapt to a rigid learning structure, with little to no adaptation in the delivery of the materials. The depths in which learning styles are typically catered to are limited to simply audio, visual and kinaesthetic (VAK), which is a broad and heavily criticised model [4].

Richard M. Felder and Linda K. Silverman have set out a learning style model which adequately accounts for the specificity required for effective teaching. The Felder-Silverman Learning Style Model (FSLSM) has been popularised by the dawn of open learning research, due to the ability to apply machine-learning and artificial neural network data-analysis techniques to extrapolate data from a learner [10].

The learning style profile can be built using a neural network approach, which interpolates each of the 8 elements in varying capacities into the content; Active/Reflective, Sensing/Intuitive, Verbal/Visual, and Sequential/Global [2]. The implementation of this can be structured according to the Index of Learning Styles (ILS) questionnaire, formulated by Felder and Silverman to complement the learning style model. The ILS is a method of categorising learners according to the FSLSM, by gauging answers to a predefined set of questions. The results of this can then be extrapolated into the profile, where the previously mentioned neural network approach can be applied. [7].

The use of experiential techniques, within semi-structured activities (activities whereby a structure is set that facilitates a level of versatility and freedom for the learner to take advantage of) encourages and facilitates learning. In addition, differing experiential learning scenarios can impact the type of information learned, implying that there is a causal link between the presentation of content in a scenario that demands an experiential element and the absorption of information [8].

Intelligent Games and Content Delivery

The concept of using virtual 3D spaces to encourage the learning process is substantiated by the fact that they can offer more collaborative and immersive experiences. The interactivity in 3D environments is nearly limitless when presented in a sandbox manner, meaning that users will be able to present their work to each other and collaborate on creative problems, removing the inhibiting boundaries of standard 2D environment [1, 3].

Learning games, particularly multiplayer games, are effective in the learning process, because they are able to utilise the experiential memory of the player, as opposed to relying on regimented and strictly structured information. It is now known that humans tend to run simulations in their minds about how best to approach a problem, meaning that effective generalisation techniques must be implemented for an effective learning outcome. In other terms, the learner must be able to take the concept from a scenario and apply it to unfamiliar scenarios, where the concept can then become generalised in their minds [6].

Open Learning Systems

There are two crucial elements to implement when providing an effective learning experience to learners. Based on the FSLSM, there must be an identification agent to determine the styles and preferences of the learner (which will then be categorised) and a recommendation agent to implement the preferences into future learning exercises. The profile would be built according to the four scales of the FSLSM, and adaptive exercises would interpret the profile and dynamically determine the best method of content delivery [5].

Research Aims, Objectives and Novelties

The aims and objectives sections will detail the solutions that this project is intended to deliver, along with the objectives that will make a solution possible. The novelties will describe how this PhD research differs from all of the academic materials and implemented systems that are currently available.

Aims

The primary aim of this project is to propose a learning system, which is able to deliver learning content to learners in an optimal manner. This includes teaching the core concepts of particular topics and, in a learner centric manner, allowing the learners to generalise and understand the topic and its fundamental principles by their own volition and from the aid of their peers.

Additionally, an open learning system based on the amalgamation of social networking, gaming and effective learning principles will be proposed to provide a system where a pedagogically driven game with activities engineered with adaptability in mind can perform analysis on the learners' learning style and preferences, and categorise content and delivery methods within a centralised open learning system. The core aims of this system are to formulate an effective communication link between the learner, the game and the learning system, where adaptations need only occur in the system as opposed to the learner. The purpose behind moving the adaptation from the learner to the system is to mitigate the learning curve required for the learner to be effectively taught.

The system must also account for multiple users, where activities or creations can be shared and collaborated on, with an effective line of communication between similarly categorised learners making this possible.

Objectives

The system that this research proposes will require machine learning and statistical analysis for categorisation and rating. For this reason, an effective data mining and analysis element to fit in with the FSLSM ILS query structure will be developed, which can both categorise the learner and the learning resources. Below is a summary of the objectives that will need to be completed:

Game elements:
- Build a 3D voxel based game
- Engineer adaptive exercises for content delivery
- Develop on-the-fly analytics based on the FSLSM framework
- Determine and develop effective communication and analysis with the open learning system

Open learning system elements:
- Build a targeted content delivery framework
- Develop a centralised content classification system, which uses feedback from the game
- Develop communication tools for class structures, achievements and progress tracking

The game and open learning system must have a high level of synchronicity, so that feedback and classifications can be delivered close to instantaneously.

Novelties

There are a number of novelties in this proposal, in both the open learning system and the pedagogical game, which include the following.

1. The use of machine learning techniques to classify and quantify a learner's style and preferences, using in game analytics alone. In essence, this is the ability of the

system to allow on-the-fly adaptations to the learners' style profile. This concept is novel, because analysis typically has a manual questionnaire as a prerequisite, as opposed to in-game analysis alone.

2. The synchronicity between the game and the open learning system, where content is stored in the central system and persistently classified by in game analytics. This means that the learners' dynamically adaptive style index within the game is checked against the respective activities stored on the OLS, which are then fed back into the game. Other research in this area does not attempt to bridge the two concepts of a learning game and an open learning platform into one unified concept.

3. The collaboration of multiple, similarly classed users in an open world environment, so that peers can assist and develop one another, which plays an important role in developing the experiential memory of each learner. This use of the FSLSM is novel, because of its implementation in a social, learning structure to assign matched learners.

4. The duality of having the open learning system, with its ability to deliver content on its own as well as the 3D game for tailored content delivery, so the barebones of the content is made available to view on the OLS, while the related activity is used in the game. Other research focus exclusively on either the OLS or the game system, without accounting for the access control of both.

5. Specially engineered 3D activities and social spaces, which are modified according to adaptive criteria with minimal developer intervention. This means that the game procedurally generates the activities necessary to develop the learner, instead of simply hardcoding every activity. Research relating to this works with pre-set activities, with no means of adapting existing gaming activities on a relatively analogue scale.

Architecture

Game Design and Prototyping

The first objective that should be addressed is the design and implementation of the adaptive pedagogical game. A process of design, prototyping and engineering will be needed to set out the adaptive activities and logical paths that the learner may take, entirely based on the FSLSM. Engineering adaptive three-dimensional tasks will be challenging, which is why multiple iterations may be required.

The technical architecture that will be propose is the use of a voxel engine based structure, whereby the three-dimensional environment will consist of cubic blocks. There will be a multitude of blocks that will be necessary for basic in-game construction and also for the variety of tasks which will need to be implemented. In addition to this, the adaptive activities will require their own blocks, which have varying degrees of interactivity (e.g. wires, logic gates, registers, etc.). The issues that need to be addressed for the implementation of this will be researched, including cross platform capabilities (for better universality) and networking implications.

Ideally, the system would be able to cater to custom exercises and activities from the learner community, so to reduce the total file size; each activity will be implemented as an optional download or expansion pack.

The 3D game itself will be based on Minetest, which is an open source voxel game built in C++ and Lua. The compliance with the GNU LGPL 2.1 means that I can modify and redistribute the code legally and openly, and even affords the ability to leave some elements proprietary where necessary.

Since the machine learning component of the game will need to be monitoring in-game activity, this will possibly need to be bootstrapped to the game, so that there is no layer of abstraction preventing the effective extrapolation and transfer of data. It will not be completely necessary to have a persistent network connection, since a cache of data can be held on the client device before a connection is made, however for optimal synchronicity, a persistent network connection may need to be forced. This will also aid in the authentication process for the learner.

Once the designs, prototypes and approach are fully determined, I can then systematically develop the elements that have been laid out until the unified system is ready.

Open learning system design and prototyping

There are two crucial elements to implement when providing an effective learning experience to learners. Based on the FSLSM, there must be an identification agent to determine the styles and preferences of the learner (which will then be categorised) and a recommendation agent to implement the preferences into future learning exercises. The profile would be built according to the four scales of the FSLSM, and adaptive exercises would interpret the profile and dynamically determine the best method of content delivery [5].

The optimal method of implementation for an open learning system would consist of four different views; the view of the learner (whereby their own progress can be tracked), the view of the parent (whereby the progress of their children can be tracked without compromising the learners' account, the view of the teacher (whereby the progress of all of their students can be tracked and activities can be set) and the necessary administrative view.

The OLS will also have to act as a data repository for in game content, so that activity and level data can be stored in a centralised manner, in accordance with the ILS scores relevant to each.

A PHP based content management system (CMS) will provide a solid framework to develop upon, since the requirement for multi-user interaction will be tackled straight away, especially with the system being loosely social network based. Although a CMS such as Moodle will provide a good framework, some of the more bespoke elements, like messaging, matching learners and classes will need to be manually developed, as well as the lines of communication with the game.

Project Evaluation

Once the system has been completely implemented and tested with tangible metrics, the effectiveness of the system will be evaluated based on:

- Improvement to the learning process
- Effectiveness of technical implementation (debugging and errors)
- Efficiency of the implemented system
- Production feasibility

To accomplish the final evaluation tasks, I will need to deploy the system in a school environment. To do this I will likely approach schools on an ad-hoc basis and discuss the limitations, if any, that may be involved.

With the final report, I will include any additional research that I have conducted, any findings that have been made as a result of the implemented system, the design and implementation logs in full and user responses to the completed system.

References

[1] Allison, C., Miller, A., Sturgeon, T., Perera, I., & McCaffrey, J. (2011). The third dimension in open learning. In *Proceedings - Frontiers in Education Conference, FIE*. doi:10.1109/FIE.2011.6143002.

[2] Ahmad, N.B.H. & Shamsuddin, S.M., 2010. A comparative analysis of mining techniques for automatic detection of student's learning style. In *Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on*. pp. 877–882.

[3] Bourg, D. M., & Seeman, G. (2004). *AI for Game Developers. Environments*. doi:http://proquest.safaribooksonline.com/0596005555?uicode=politicat.

[4] Dunlosky, J., Rawson, K. a, Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. doi:10.1177/1529100612453266.

[5]Fasihuddin, H., Skinner, G., & Athauda, R. (2014). Towards an Adaptive Model to Personalise Open Learning Environments using Learning Styles. *2014 International Conference on Information, Communication Technology and System 2014*, 183–188.

[6] Gee, J. P. (2008). Learning and games. *The Ecology of Games: Connecting Youth, Games, and Learning*, 21–40. doi:10.1162/dmal.9780262693646.021.

[7] Graf, S., Viola, S. R., & Leo, T. (2007). In-Depth Analysis of the Felder-Silverman Learning Style Dimensions. *Journal of Research on Technology in Education*, *40*(1), 79–93. doi:10.1080/15391523.2007.10782498.

[8] Hamer, L. O. (2000). The Additive Effects of Semistructured Classroom Activities on Student Learning: An Application of Classroom-Based Experiential Learning Techniques. *Journal of Marketing Education*, *22*(1), 25–34. doi:10.1177/0273475300221004.

[10] Viola, S., Graf, S., & Leo, T. (2006). Analysis of Felder-Silverman Index of Learning Styles by a Data-Driven Statistical Approach. *Eighth IEEE International Symposium on Multimedia (ISM'06)*, 959–964. doi:10.1109/ISM.2006.30.

Pisit Praiwattana

Development and Analysis of a Games-Based Crisis Scenario Generation System

Pisit Praiwattana, Abdennour El Rhalibi
Department of Computer Science
Liverpool John Moores University,
Byrom Street, L3 3AF, Liverpool UK,
P.Praiwattana @2014.ljmu.ac.uk, a.elrhalibi@ljmu.ac.uk

Abstract.
Crisis is an infrequent and unpredictable event which is challenging to prepare and resolve. Serious-game approach proved to provide potential support in training and simulating event of real-world crisis situation to different stakeholders. Yet in practice, the approach meets with difficulty on how to setup and utilize different core components such as asset management, crisis scenario generation, agent simulation, real-world constraints, and the evaluation process to yield beneficial information upon running the sys-tem. To address this issue, the key question is what can be done to propose a general crisis game-based framework providing necessary core components while generating evaluation result yielding potential analytical data for a crisis management process. Therefore, in this paper, we aim to review and consolidate the existing research on scenario generation techniques then to propose novel solution to derive a desirable scenario content which is also being validated in the simulation framework based on the JADE multi-agent architecture.

Keywords. Crisis simulation framework, Scenario generation, MAS

1. Introduction

Crisis can be defined as any event that is, or is expected to lead to, an unstable and dangerous situation affecting an individual, group, community, or the whole society. They can be caused by man-made actions or natural phenomenon or occasional accidents that are difficult to predict (i.e. earthquake, flooding, terrorism, hazardous outbreak, etc.). Crisis management organizations have to handle available resources and facilities to deal with the crisis disaster. These resources change over time in different situations, which causes challenges in preparing the right countermeasure plan, leading to unintentionally experiment via trial and error. Since crises are infrequent, Walker et al. [1] suggests that training from game with relevant real-world environment and well-defined scenario can be beneficial to crisis management personals to understand and prepare for any upcoming emergency situation due to its cost effectiveness in development compared to the setup-cost of real training   practice.

Scenario may describe a hypothetical situation that provide a necessary environment setup, initial background of event and final goal to represent a sense of real-world situation in the practical training or test. As regard to of crisis scenario, the possible content structure can be separated into context and crisis. Context represents a pre-incident of crisis situation to enable the understanding of specific environment surrounding the area of event while crisis describes a set of specific events which may lead to dangerous, troubled and challenging situations, with potentially cascading effects. To test possible response, the script usually includes unpredictable or probabilistic chain of events that are out of control of any player.

To focus on the process of defining and evaluating a scenario, development of a framework that facilitates the generation of different possible crisis scenario script while simulating its emergency response based on available resources will provide more elaborate understanding of crisis situation and also give a preliminary assessment of crisis scenario and its corresponding plan. The result will be beneficial to the decision support system and to transferring into practical crisis training.

In this paper, we propose the development of an interactive game-based crisis scenario simulation framework developed on top of open multi-agent toolkits such as JADE. The reminder of the paper is organized as follows: In section 2, we introduce some existing automated scenario generation systems; in section 3, we proposed our Crisis Scenario Generation System Framework Design; in section 4, we discussed Framework scalability Evaluation; and, in section 5, we conclude the paper.

2. Automated Scenario Generation Systems

Modern serious games do require a plausible scenario which is often directed from domain experts and it is a very time-consuming process. Automated scenario generation provides major benefits to the manual process by: 1) Content generation can quickly produce on demand scenario with setup and constraints; 2) Computer-generated scenario can be used to supplement human-based scenario quality. For Crisis Management domain, we initially focus on producing a sequence of crisis events while mainly considering: (1) resource management perspective; (2) stake-holder collaboration in solving an emergency situation. Different computational approaches have been proposed for scenario generation system.

Hullett and Mateas [2] apply a planning technique to generate a firefighter rescue training scenario in the collapsing building area. The system use HTNPlanner with building structure data as input while set a goal to create a situation that satisfy crisis final description, domain knowledge must be defined for planner to allow physical consistency and achieving better plausible result. As a result, the system generates a scenario by filling in content as a sequence of event or activities that are expected to occur and which usually manipulate world environment leading from initial state to the desired goal situation. The trainee is given a role and a set of specific skills to per-form in the scenario. The variety of generated scenario is delivered in scale of small, medium and large world setting and it is argued to be better than random probabilistic distribution of element in case of fire situation, damage propagation, and comparable existing systems due to the provision of a domain knowledge consistency model. The main limitation of this system is that the work is tending to encounter a memory shortage during simulating the variation of levels in Medium and Large scale.

Grois et al. [3] developed a SceneGen algorithm for Navy DCTrain System using Noisy Bayesian Network (NOBNs) to search for key-event to satisfying scenario objective using data from knowledge-based in a form of belief network with a penalized likelihood and rejection test to discard non-plausible results. The process is aimed to provide offline-scenario generation. The authors also mention Case-based stochastic perturbation (CBSP) is used to acquire a seed scenario from experts then apply random distribution to manipulate more variation to the original which is likely to bring unreasonable or in-plausible result due to random nature without any testing for plausibility. ScenGen's strength lies in its ability to guarantee the "quality" of each and every scenario it generates through a carefully designed selection bias and it claims to be better in overall performance than Manual design by human subject matter experts (MDHE), Naïve random generation (NRG), and Case-based stochastic perturbation (CBSP) which produce lower-quality scenarios. It defines plausible quality by checking the occurrence of key events according to the desired learning objective. Some learning objective or key event can be occurred and added simultaneously at the same time step. The major drawback of this approach is to require a set of good base human generated scenario as a seed then manipulates them to obtain more variation in the automated results, and in addition the offline generation may lack efficiency in dynamic crisis simulation system when the setting is reflecting the complexity situation from real-world problem.

Zook et al. [4] introduce a combinatorial optimization approach to scenario generation to deliver the requisite diversity and quality of scenarios while tailoring the scenarios to a

particular learner's needs and abilities for military training in virtual environment which is opposite to the planning approach. The main scenario genera-tion based is on a genetic algorithm to search for a best solution; reading in author-specified domain knowledge, the details and the type of possible events and requirement in scenario, and constraint order on events. The process works by considering instant event template for the scenario at random location; mutating the parameter of random chosen scenario; applying cross-over operation, to create new sequence of events to improve the quality of scenario for the next iteration. The authors pro-posed evaluation in an interesting and effective way: 1) quality of solution at run-time; 2.) the diversity of scenario as function of running time, 3) performance of trainee and appropriateness of difficulty level when training on generated scenario. The result from their evaluation is to generate a unique scenario compared to planner generation approach, this technique based on a combinatorial optimization provides lower-quality solution initially, but explores multiple different regions in the solution space containing high-quality solutions and so refines multiple distinct scenario that meet provided learning objectives rather than explore variation on same high-quality scenario. The major drawback is that the generation requires a pre-defined small element of events to be tailored into scenario and initial input of learner model for suitable evaluation fitness function for the genetic algorithm. In this section, we have reviewed some of the proposed techniques on scenario generation system. In next section, we propose a scenario generation framework based on the JADE multi-agent architecture which is aiming to incorporate the scenario generation technique and crisis simulation agent-based model into a unify system.

3. Crisis Scenario Generation System Framework Design

The crisis scenario generation framework is composed of many components. Figure 1 represents a high level architecture of our crisis scenario generation framework. In the following section, we will introduce these components respectively.
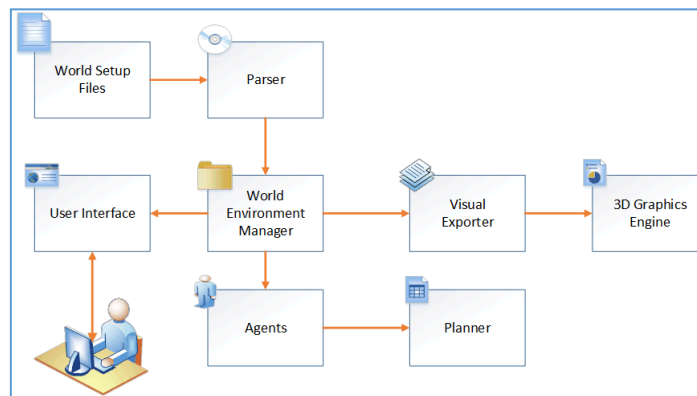
Figure 1 Crisis scenario generation framework architecture.

3.1 Parser Component

Setup information including different scope of scenario world location, available resources, possible set of actions, constraints, possible set of events' description, and related configuration keys. The purpose of this component is to parse the necessary files to initializing the base representation of the system world environment. Moreover, it is possible to create an API for importing a world-setup from different framework. This feature will be revisited in the later stage of development.

3.2 World Environment Manager Component

The world environment manager has the main task to handle the representation of centralized scenario world scene. It is responsible to maintaining the knowledge of the world stage based on any change from agents' actions or event triggering. It also allocates the information regarding the current situation in a scenario to agent manager with planner instance. Moreover, the world environments will coordinate and respond to agent's request for executing an action in the environment. Any conflict from agents' request will be resolved via this module rather than direct negotiation between each agent. In implementation detail,

World Environment Manager may be operating as central agent which stores the world environment data structure and allow the execution of an action to manipulate the existing object variables in environment. The sub-components of world environment manager are described as in the following:

World Scene: The world scene component will hold the list of scenario knowledge and available resources in the current environment state of the world regarding the setup location and being updated by incrementing time steps. It is a central data structure which allows allocating of world stage knowledge into any requested component.

Event Manager: The event manager component utilizes a pool of events with probabilistic model. It fires a sequence of events to conduct the scenario incident, which can be deterministic to representing a cascading effect of specific real-world situation depending on the proper setup files. This component will be directly monitoring the world scene and its incident world event will be allocated with corresponding world-scene information to the requested agent.

Time Manager: This component handles keeping track of the time in the virtual simulation for conveniently synchronizing and calculating the action duration.

Action Capability: All possible agent actions will be stored in this module with their specific precondition to execute such as requester role, available resource condition, and time duration.

Statistic and Evaluation Manager: This component will keep statistic result data regarding each crisis incident in the generated scenario according with its executing plan providing a mean to evaluate the utility from outcome of different plan or decision.

Conflict Manager: The conflict resolver component will have an authority in resolving the conflict between agents' request. Currently, every action is not generated as a partial-order plan so there is likely to have no conflict. However, the necessity to have a dedicated component will become handy in extending the feature of framework.

3.3 User Interface

Current world information sent from the world manager output module will be displayed in this user interface component. User can observe the development of generating scenario step-by-step in a sequence of narrative log text. However, the current design does not focus on visualizing the virtual environment to the user via this channel directly. User may use this component to configure the option of the scenario setup before starting the actual simulation and also assign a preference choice for agent decision making if applicable.

3.4 Agent Component

The framework will elaborate the world scene and scenario by applying a multi-agent framework solution as each agent represents an actor in the scenario. In general, crisis management situation normally includes different tiers of decision making chain of command and, sometimes, crisis scenario can be represented in according to different scope of scenes whether a city-area event or in-a-building floor layout scene. To address this model, agents may be allocated into separated level of abstract representation which can be defined as crisis manager, crisis facility/station, and crisis personal team. The hierarchy in chain of command will be deployed from manager to personal tier respectively. A simplify assumption will be made to indicate an actual sequence of agents' behaviour, which greatly reduces system complexity.

The representation of agents' type is described as in the following:

Decision Making Agent (DMA): This agent acts as a high-level decision making unit, crisis manager. In general, the decision-making agent, crisis manager, will be equipped with an instance of planner to generate plans according to the given world state and event notified from world environment manager. While using utility function preference, High-level plan will focus on allocation of resource, task assignment, optimization then being sent to control agents.

Control Agents (CA): Control agent represents a facility unit, crisis facility. Being allocated the resource and task from DMA, this unit may evaluate resources with the given task in

the case that the additional resource is required. If applicable, CA instantiate a field agent unit with setup of necessary resources and task. CA has main responsibility in assigning sub-goal to each field agent.

Field Agents (FA): This agent represents personal teams with resources. FA will perform direct request to execute an action with a world environment manager component.

3.5 Planner

Planner component allows the generation a plan consisting of sequence actions related to a given world state, allocated resources, and possible actions. This world information must be pre-processed into PDDL format then being stored in the local KBS of the Planner. The external planner instance will be deployed as in the current state to fasten the development of framework. We currently select a planner using forward chaining state space search to reach a goal with heuristic function, FF_Metric [5]. The constraint in resources, agent's actions, and goal's condition will be defined and solved as a constraint satisfaction solution [6].

3.6 Visual Exporter

This component will be used for the frameworks with external graphics engine. The information of world representation must be formatted into a common data extension such as XML or COLLADA, which can be efficiently parsed for visualizing purpose later on. In the early state of current design, exporting data should contain a data structure representing each simulation step of world scene state.

3.7 3D Graphics Engine

This component aims to demonstrate how the frameworks can connect to external 3D graphics engines such as Unity3D and other similar engines. The crisis scenario simulation will be represented as 2D or 3D visual data in which the graphics representation of each element can be mapped from exported data of the world scene. This component is currently an optional extension.

4 Framework scalability Evaluation

Currently, the proposed framework is in an initial development stage, the core components involving world environment, agents, and planning have been implemented. We summarized an   evaluation   in   this   section.
active agents in To test a performance and scalability, we proposed increasing number of the system while measured a total duration to solve the given scenario tasks and a plan generation time For Testing scenario, we implemented a basic firefighter domain which describe fire situation where a world object is set *on-fire*. There are 4 basic operations which are *move-to*, *pick-up-a-supply, extinguish-a-fire* and rest while there are only one type of actor and facility, fireman and fire station.

The number of agents is instantiated with a scene of fixed size world objects in a single event. The time reported are run in a Laptop with Core i7-5500U @2.40GHz, 2 Cores with RAM 16GB (10GB Available), and Window 10.

Our results are illustrated in Figure 3. We test initially with single agent and then 2, 5, 10, 20, 30, 50, 100 respective in simultaneous instantiation to solve a world-event problem with 20 on-fire objects.
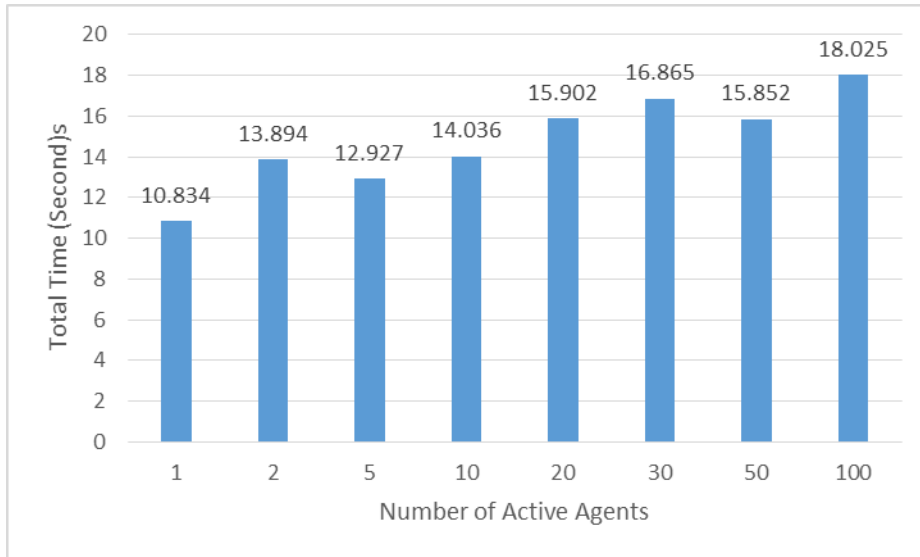
Figure 2 Computation time of simultaneously solving a world event by agents.
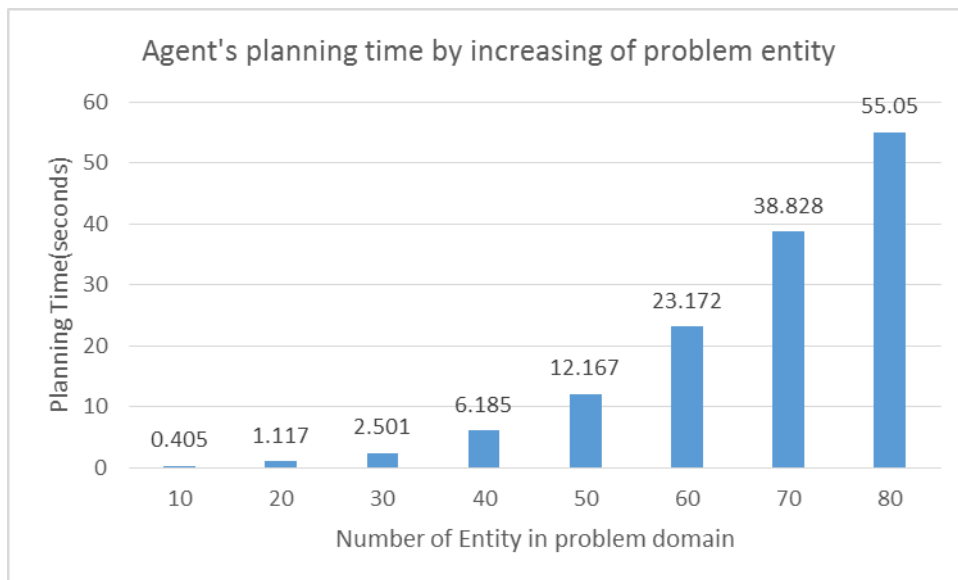
Figure 3 Planning time of single agent by increasing of entity instance in problem

The result from figure 3 indicates that increasing the number of agents will require more time due to the fact that they generated more collision in action execution in the environment such as trying to execute an action on the same object. Only one agent will suceed while another fails triggering replanning process. This situation may lead to studying of opportunistic competition between each agent and invention of more efficient planning and replanning algorithm for multi-agent system. Figure 4 represents required time of planning process during the event. Increasing more entities during planning will result in increasing of computation time exponentially. The partial solution is to divide a problem into sub-goal with lesser entities which may reduce the individual planning in some case. Currently, the evaluation is based on a single machine rather than hosting an agent on different machines communicated on JADE environment which could improve stability and reduce execution time in larger world setup. The re-evaluation will perform after the implementation of command-tier agent such as fire station to simulate resource management environment in future version.

5 Conclusion

In this paper, we discuss a review on existing work regarding automated scenario generation technique then present a design of crisis scenario simulation framework developed on top of open multi-agent framework such as JADE to provide a new solution incorporating techniques based on planning in both area. This approach has a potential to provide a large variety of study and test environment to validate theory and plan in different situation on across domain especially crisis management. Further, it will allow the evaluation of trade-off between each decision on different perspective as stated in separated tiers of agent. While in regarding the scalability problem for simulation, the multi-agent framework enables a larger scale of computation with connecting more platforms as a container to the environment.

In the future work, we will continue the implementation of base simulation framework and provide more evaluation on result of different scenario setup with different tiers of agent. We will also develop an automated scenario generation on top of the simulation framework to validate and provide diversity of representation of crisis scenario.

6 Reference

[1] Walker, W. E., Giddings, J., & Armstrong, S. (2011). Training and learning for crisis management using a virtual simulation/gaming environment. *Cognition, Technology & Work Cogn Tech Work*, 13, 3, 163-173.

[2] Hullett, K., Mateas, M. (2009). Scenario generation for emergency rescue training games. *4th International Conference on Foundations of Digital Games* - FDG '09.

[3] Grois, E., Hsu, W. H., Voloshin, M., & Wilkins, D. C. (1998). Bayesian network models for generation of crisis management training scenarios. In *AAAI/IAAI*, 1113-1120.

[4] Zook, A., Riedl, M. O., Holden, H. K., Sottilare, R. A., and Brawner, K. W. (2012). Automated scenario generation: Toward tailored and optimized military training in virtual environments. *7th International Conference on the Foundations of Digital Games.*

[5] Hoffmann, J. (2003). The Metric-FF Planning System: Translating ``Ignoring Delete Lists'' to Numeric State Variables. *Journal of Artificial Intelligence Research*, 291-341.

[6] Yokoo, M. (2001). Constraint Satisfaction Problem. In *Distributed Constraint Satisfaction, Springer Berlin Heidelberg*, 1-45.

Taybia Mohammed

The effects of acoustic vibration on fibroblast cells

T Mohammed, MF Murphy, F Bezombes, F Lilley and D Burton
General Engineering Research Institute, Liverpool John Moore's University, Byrom Street, Liverpool, L3 3AF, UK. Email:t.mohammed@2013.ljmu.ac.uk

Abstract. The human body is a complex system and we have made great progress in understanding the biochemical mechanisms. In contrast, we know very little about how the body responds to the physical forces which it is exposed to on a daily basis and play an equally important role in helping the body to function normally. Recent work has shown that externally applied mechanical stimulation, can result in changes to the biochemistry of cells and this can result in changes to cellular processes including; apoptosis (cell death), differentiation (a change in cell specialization) and proliferation (cell multiplication).

The aim of this research is to investigate the effects of mechanical stimulation upon the migratory properties of fibroblast cells. Fibroblast cells play an important role in wound healing. A reduced wound healing response is known to be associated with many conditions where ulcers are common, including: diabetes, ischemia, pressure wounds and old age. Currently, acceleration of wound healing involves the use of drugs. However, what if we could accelerate or promote wound healing simply through physical mechanical stimulation? Using a speaker-based system of low frequency low amplitude (LFLA) vibration was applied to cells in vitro. Human lung fibroblast cells (LL24) were exposed to mechanical stimulation for 5 minutes at frequencies between 0 and 1600Hz. Time-lapse microscopy was used to determine cell migration distance. Results presented show that mechanical stimulation can enhance LL24cell migration and this is coupled with a change in actin organisation. This leads a promise for vibration-assisted wound healing.

Keywords: Mechanical stimulation; low frequency low amplitude (LFLA);

1.0   Introduction

Cells and tissues, by their very nature, have evolved to sense their physical surroundings. This is known as mechanotransduction and involves mechanosensitive feedback controls that help regulate important cellular functions such as development, homeostasis, apoptosis, differentiation and proliferation [1]. Very little research has been carried out to understand the effects of mechanical stimulus on cell/tissue behaviour. However, we do know that changes in a cells/tissues ability to respond to forces are associated with certain disease states including; muscular dystrophies, cardiomyopathies, cancer progression and metastasis [1, 2].

Recently mechanical stimulation of cells has shown great promise from a medical standpoint. For example, Wang *et al.*, applied low-magnitude stimulation (32-37 Hz) to mesenchymal progenitor cells (MPCs) for 10 and 20 minutes, respectively and found that secreted proteins (or vibe-proteins), which have the capacity to promote bone formation, were up-regulated. The authors suggest that by virtue of their secretory status, some vibe proteins may be candidates for pre-clinical development as anabolic agents for the treatment of osteoporosis [2]. Similarly, Kulkarni *et al.*, used a modular piezoelectric actuator to vibrate cells at 4Hz for 1

hour on 3 consecutive days and found that bone reabsorption (breakdown of bone) was inhibited [3].Whilst Wu *et al.*, found that LMHF vibrations (45Hz for 15 min per day) was anabolic for bone by inhibition of osteoclast differentiation [4]. In contrast, Kim *et al*, found that human exposure of mesenchymal stromal cells (hMSC) to low-magnitude high-frequency (LMHF) vibrations (30-40Hz) resulted in an increase in proliferation. The authors suggest that LMHF could enable the osteogenic process [5].

Such studies suggest that mechanical stimulation may have great potential as a novel therapeutic treatment for medical intervention for a range of conditions. This work is concerned with investigating the effects of the effects of mechanical stimulation on fibroblast cell migration. Fibroblast cells play an important in wound healing and we hypothesised that mechanical stimulation may be used to enhance the wound healing process.

2.0 Methods

2.1. Mechanical stimulation via acoustic vibration

To mechanically stimulate cells we used a speaker-based system. The speaker-based system was built using a 0.2W super-thin, waterproof Mylar speaker (45mm) and an Arduino microcontroller board for control (Fig. 1). This system enables mechanical stimulation to be applied (via a sinusoidal waveform) to the underside of the cell culture dish to which the cells adhere, at frequencies ranging between 100-1600Hz.

2.2. Calibration of mechanical stimulation system

To calibrate the frequency and amplitude generated by the speaker a laser vibrometer (Polytech Ltd.) was used. Briefly, a cell culture dish (35mm) containing cell culture growth medium was rested upon the speaker (40mm diameter). The laser spot from the vibrometer was focused, through a ×10 microscope objective lens, onto the inner bottom surface of the dish (the surface to which the cells would adhere). Next, the speaker was set to vibrate at either 100, 200, 400, 800 or 1600Hz. Vibration frequency (Hz) and amplitude of displacement (μm) were obtained though measurement of the displacement of the laser spot, which is software driven in the Polytech system.

2.3. Cell culture

Human lung fibroblast cells LL24 were obtained from the European Collection of Animal Cell Cultures (ECACC). The cells were maintained in Dulbecco's Modified Eagles Medium (DMEM) supplemented with 2mM L-glutamine (Invitrogen), 10% foetal bovine serum (FBS) (sigma, UK) and 1% penicillin- streptomycin in T75 cell culture flasks, at 37°C in a humidified atmosphere containing 95% air and 5% CO2.

2.4 Cell Imaging

All images obtained during this study were from a Zeiss 510 Meta laser scanning microscope, mounted on an Axiovert 200M BP computer-controlled inverted microscope. This microscope is equipped with the following laser lines; blue diode 405nm, Argon ion 458, 477, 488 and 514nm, He-Ne 543nm. For cell migration studies, Differential Interference Contrast (DIC) microscopy was carried out using the He-Ne 543nm laser. For imaging filamentous actin the He-Ne 543nm laser was employed.

2.5. Mechanical stimulation of cells

Cells were seeded into 35mm cell culture dishes at a density of 2×10^4 cells/cm^2 and left to attach to the dishes for 24 hours at 37°C in a humidified atmosphere, containing 95% air and 5% CO_2. After 24 hours the dish was removed from the incubator and mechanically stimulated for 5 minutes at either 0 (control), 100, 200, 400, 800, or 1600Hz. Next, the dish was immediately placed into the environmental chamber (37°C humidified atmosphere with 95% air and 5% CO_2) of the laser scanning microscope and imaged every 5 minutes for 4 hours using DIC time-lapse microscopy.

Each experiment was repeated three times and ImageJ software was used to manually track the migration of single cells (n=30 cells from each frame in the time-lapse sequence) in each population after 4 hours so as to determine mean migration distance (μm).

2.6. Mechanical stimulation and actin organisation
Given that cell morphology and the actin cytoskeleton undergo reorganisation prior to and during cell migration, fluorescence microscopy was used to document the effects of mechanical stimulation upon cell morphology and actin organisation. The LL24 cells were seeded into 35mm cell culture dishes at a density of 2×10^4cells/cm^2 and left to attach for 24 hours at 37˚C in a humidified atmosphere containing 95% air and 5% CO_2. Next, the cells were mechanically stimulated for 5 minutes at 0, 100, 200, 400, 800, or 1600Hz and then chemically fixed to preserve their morphology.

For fixation the cells were washed (×1) in phosphate buffered saline solution (PBS) for 5 minutes and then fixed at room temperature in 10% paraformaldehyde for 10 minutes. Following fixation the cells were washed with PBS and permeabilised for 20 minutes at room temperature using 0.5% Triton-X 100. The cells were then washed (x1) with PBS and the actin filaments labelled by staining with Rhodamine-Phalloidin according to the manufacturer's instructions (Cytoskeleton Inc.). After staining, the dishes were thoroughly washed with PBS to eliminate any background staining and the cells were imaged using an excitation wavelength of 543nm. Images were analysed to determine changes to the morphology of cells.

3.0 Results
3.1 Speaker Device

To assess the effects of mechanical stimulation upon fibroblast cell migration, a system was developed that can deliver low-frequency-low-amplitude acoustic mechanical stimulation, via a sinusoidal waveform, to cells growing in a cell culture dishes (Fig. 1 left). This system allows the frequency of mechanical stimulation to be controlled via an Arduino controller. In order to determine that the vibration frequency was accurate and stable, a laser vibrometer was used to calibrate the system. It was found that below 100Hz and above 1600Hz the system was unstable (in terms of frequency), as laser vibrometry recorded multiple harmonics outside of this frequency range (data not shown). Therefore, frequencies between 100 and 1600Hz were used, so as to accurately deliver stable, low-frequency-low-amplitude mechanical stimulations to cells in a continuous manner. Laser vibrometry also recorded amplitude of displacement (μm), which can be seen to decrease as frequency is increased (Fig. 1 right).
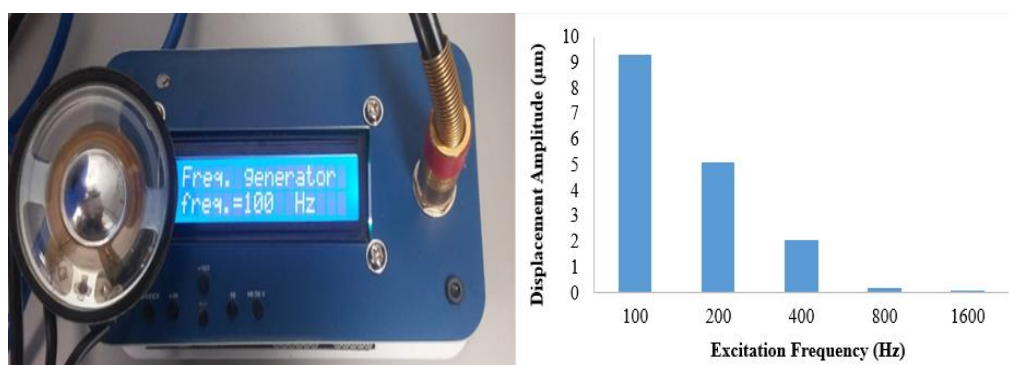
Figure. 1. Speaker-based system (left) and calibration of frequency *versus* amplitude of displacement of the inside bottom surface of a 35mm cell culture dish to which the cells adhere (right)

3.2 The effect of Mechanical Stimulation on Single Cell Migration
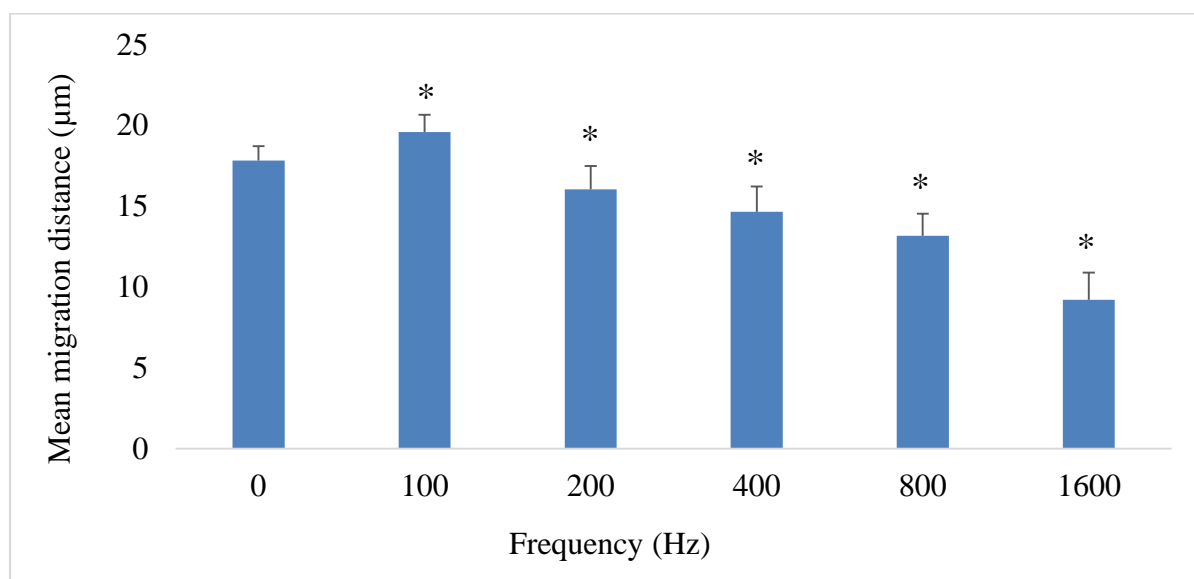
Figure 2. Analysis of migration over a 4-hour period shows that 100Hz vibration (5 minutes) enhanced cell migration. However, there was found to be a steady decline in overall mean migration distance (μm) over the 4-hour period from 200-1600Hz.

In order to investigate if mechanical stimulation has any effect on the migration of individual fibroblast cells, time-lapse microscopic imaging and subsequent cell tracking using ImageJ software was carried out. Cells were mechanically stimulated using 100, 200, 400, 800, or 1600Hz for 5 minutes and migration was subsequently recorded over a 4-hour period. As can be seen from Figure 2, mechanical stimulation for 5 minutes at 100Hz significantly enhanced (approx. 10% increase) the mean migration distance (μm) of human lung fibroblast (LL24) cells, when compared to the control population ($p<0.05$). In contrast, mechanical stimulation for 5 minutes for all frequencies above 100Hz resulted in a decrease in the mean cell migration distance (μm), when compared to the control ($p<0.05$). This decrease can be seen to occur in a steady manner as mechanical stimulation frequency increased (and amplitude decreased) resulting in a decrease in the mean migration distance (μm) of approximately 50% (at 1600Hz), compared to the control population.

These results show that fibroblast cell migration can be controlled using acoustic vibration and that migration distance is dependent upon vibration frequency. The trend observed was repeatable for both cell lines, particularly the increase in mean cell migration distance (μm) seen at 100Hz. On all occasions a steady decrease in mean cell migration was observed for the lung cells (LL24), with 1600Hz always resulting in lowest mean cell migration distance (μm).

3.2 Effects of mechanical stimulation on actin organisation in cells following mechanical stimulation.
Actin filament organisation was examined following vibration. As can be seen from figure 3 vibration was found to increase the levels of actin at the leading edge of the cells (yellow arrows) compared to the control. Overall, there was no obvious change in actin stress fiber formation.
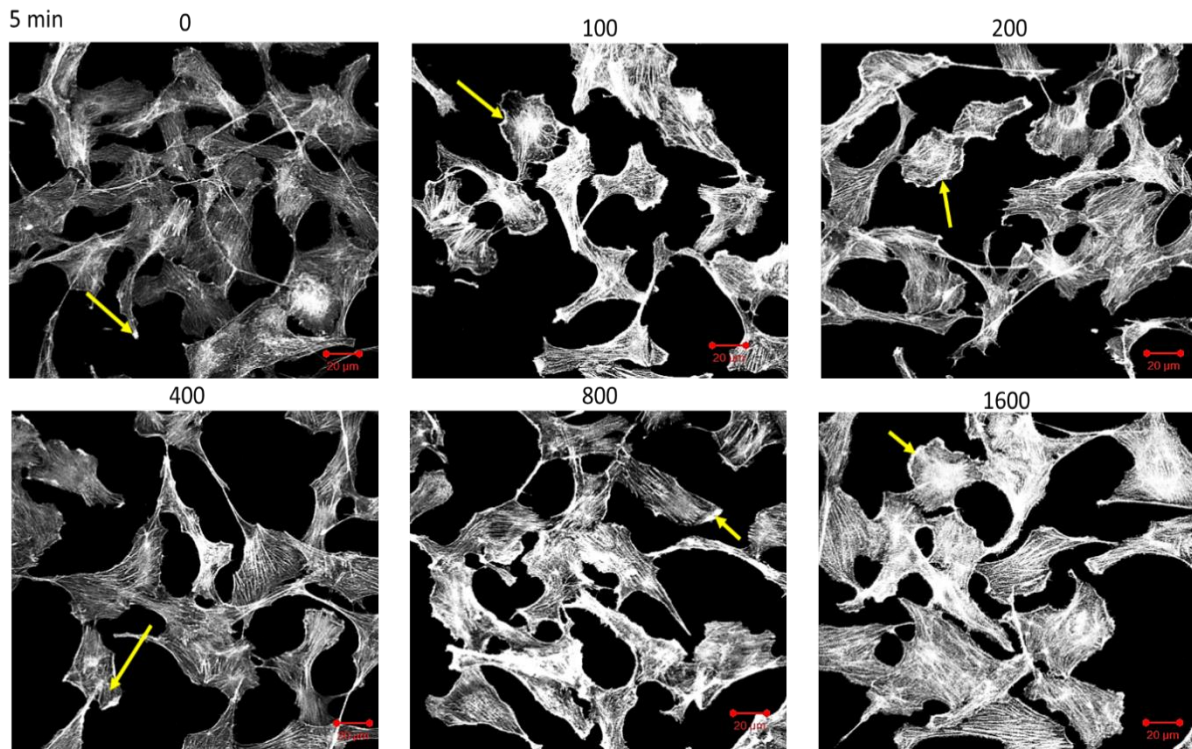
Figure 3: F-actin organisation in LL24 cells following mechanical stimulation (0Hz to 1600Hz) for 5 minutes. Yellow arrows highlight lamellipodia/ruffling.

Cell migration is associated with well characterised changes to the cell. Many of these changes are macro-scale structural and morphological changes and include; the formation of lamellipodia/membrane ruffling, membrane blebs and actin filopodia, the latter of which has been shown to stimulate cell migration. Given that fibroblast cells have been shown to be mechanosensitive and that actin remodelling is associated with cell migration, the effect of mechanical stimulation on actin organisation was investigated. Following mechanical stimulation of LL24 cells at 0-1600Hz, the cells were fixed and stained using Phalloidin, which specifically labels actin filaments (F-actin). Confocal microscopy revealed that mechanical stimulation encouraged actin remodelling it was observed that there was an increase in lamellipodia/membrane ruffling (identified by the arrows in figure 3) and an increase in stress-fiber formation/density, compared to the control. However, there were no notable differences in the F-actin organisation as frequency increased (figure 3).

4.0 Conclusion

We have shown that mechanical stimulation (100-1600Hz) applied via acoustic vibration, can affect cell fibroblast cell migration in a frequency-dependent manner. In particular, mechanical stimulation at 100Hz significantly increased the mean cell migration distance for both cell types studied here. For the LL24 lung fibroblast cells, mechanical stimulation at frequencies above 100Hz resulted in a steady decrease in the mean cell migration distance without affecting cell viability. Mechanical stimulation was also found to affect cell morphology and actin organisation, with acoustic vibration increasing the formation of lamellipodia and filopodia.

5.0 References

[1] Ingber DE. (2006). Cellular mechanotransduction: putting all the pieces together again. FASEB J. 20 (7): 811-27.

[2] Wang H., Brennan T.A, Russell E., Kim J.H., Egan K.P., Chen Q., Israelite C., Schultz D.C., Johnson F.B. and Pignolo R.J. (2013). R-Spondin 1 promotes vibration-induced bone formation in mouse models of osteoporosis. J Mol Med (Berl). 91 (12): 1421-9.

[3] Kulkarni R.N., Voglewede P.A and Liu D. (2013). Mechanical vibration inhibits osteoclast formation by reducing DC-STAMP receptor expression in osteoclast precursor cells. Bone. 57(2): 493-498.

[4] Wu S.H., Zhong Z.M and Chen J.T. (2012). Low-Magnitude High-Frequency Vibration Inhibits RANKL-Induced Osteoclast Differentiation of RAW264.7 Cells. Int J Med Sci. 9 (9): 801-807.

[5] Kim I.S., Song Y.M., Lee B. and Hwang S.J. (2012). Human Mesenchymal Stromal Cells are Mechanosensitive to Vibration Stimuli. J Dent Res. 91 (12): 1135-40.

Temitope Sam Odusina

Development of An Adaptable Airborne 3D Stereovision System

Sam-Odusina T.J[1], Burton D.R.[1], Lilley F[1] and Bezombes F[1]
[1] General Engineering Research Institute, Liverpool John Moores University

Abstract. In recent years, Unmanned Aerial Vehicles (UAVs) are being used for many new and exciting applications such as autonomous monitoring, inspection and surveillance of buildings. In addition, real time state estimation, obstacle detection and avoidance, 3-D measurements and mapping are also crucial research topics in the field of robotics and mechatronics with the aid of these platforms. In 3-D data acquisition sensors like RADAR, scanning LIDAR and other active sensing methods are becoming increasingly difficult to mount on these platforms as a result of weight constraints. The design of an adaptable airborne 3-D stereovision system is discussed to address these problems and provide potential solutions. The UAV needs to fly through pre-assigned waypoints to satisfy the working principles necessary to achieve this system structure. This paper focuses on implementing Dronekit API to communicate with vehicles over the MAVLink protocol. This is currently being applied to guide UAVs to specific Global Positioning System (GPS) positions in Software in the Loop Simulation (SITL). The next phase of this research would involve implementing this architecture on real hardware devices.

Keywords

GPS, UAV, 3D, Stereo-Vision, LIDAR, RADAR, SITL.

1. Introduction

Unmanned Aerial Vehicles (UAV) have several advantages in exploration and reconnaissance mission since they can operate in highly clustered environments or confined spaces. However in other to operate these platforms safely, fast and accurate pose estimation from external sensors like GPS is needed for control. UAV platforms equipped with GPS can be applied in a large variety of scenarios, and supply a testbed to investigate several unsolved problems. For example in the area of autonomous monitoring, they could be used to autonomously search a given area such as river or coastal line for a specified structure, identify the structure and also map the coordinates of the structure. In surveillance missions for military and civil applications, object identification and tracking with aid of UAV platforms are imperative to detect and localise targets of interest in a dynamic and unknown environment. Considerable work has been done in the past in developing vision based control methods to provide lots of information on these target of interest with the use of passive sensors embedded on UAVs (Teulière et al., 2011).

In the area of 3-D data acquisition, the integration of 3-D optical measurement systems with UAVs enable on-demand very high resolution data collection which can be customised efficiently. This has led to an extensive amount of research in the past. However current solutions suffer from various limitations such as weight constraints and cost as discussed in Section (2) below. This work proposes a novel airborne stereo setup implemented on a fleet of UAVs as a potential solution where the proposed system is flexible, cost effective and relatively accurate in comparison to existing measurement technology of a similar class. This paper discusses the use of open

source software and hardware to control a fleet of UAVs as a first step towards achieving the stated goal.

2 .Literature Review

2.1. Multi-sensor integration

A multi sensor integration of laser scanner, cameras, Inertial Measurement unit (IMU)and a Global Position System for a UAV has been proposed for 3-D mapping(Nagai et al., 2009). The integration of GPS and IMU data is effective for high-accuracy positioning of the UAV platform. 3-D shape is acquired by the laser scanner and texture information is acquired by the digital cameras. However there are downsides to such system design, the integration of so many sensors on-board the platform leads to an overall increase in complexity of the overall system. There is also an increase in payload in the UAV platform which can possibly lead to reduced flight time of the vehicle.

2.2. Scanner-less LIDAR systems

A new generation of airborne scanner-less laser LIDAR systems designated for UAVs under low altitude operations, which could provide strong technical support for 3-D data acquisition has been proposed (Zhou et al., 2013). Lidar systems work on the principle of emitting a pulse and measuring how long it takes for the pulse to be reflected from the target with the use of photodiodes. It is noteworthy to highlight that the technique LIDAR systems adopt for obtaining 3-D data requires a time resolution of picoseconds for deriving a millimetre level resolution, hence it would be difficult to implement these systems at low-cost. The use of LIDAR systems does not provide cost effective solution for 3-D data acquisition.

2.3 Structure from Motion

A structure from motion (SfM) approach has also been used to obtain 3-D information. This technique exploits the relative motion between camera and scene. It implements the use of only one camera and no special lighting to compute depth from the spatial and temporal changes occurring in an image sequence. Hence given a defined motion, depth can be determined from observing apparent motion of the camera in 2D space. SfM operates under the assumption that 3-D structure can be resolved from a series of overlapping images. Key features which are invariant to image scaling, rotation and partially invariant to changes in illumination are identified in individual images and are then used for image correspondence. Once the image correspondence problem is solved, depth estimation can be derived by matching similar points between the different frames of the motion sequence.(Westoby et al., 2012). Structure from motion techniques have been adopted in UAVs for 3D reconstruction and measurements (Obanawa et al., 2014, Clapuyt et al., 2015, Genchi et al., 2015). The traditional setup in UAV photogrammetry involves mounting one camera on a single drone and consequently taking overlapping images from different views at pre-defined waypoints in such a way that image acquisition is optimised. A single camera taking images at different known positions allows for a wide base length configuration. However the UAV motion between camera positions in this configuration introduces two new problems. One of the issues is that the UAV estimates the relative positions of the camera at two consecutive locations innacurately, due to the fact these positions are dependent on the positioning accuracy of the UAV. The other issue is

regarding the wide-baseline itself, resulting in a larger change in view-point and thus making the two images less similar, which makes the stereo matching process more difficult (Olson and Abi-Rached, 2010). Furthermore the Structure from Motion approach assumes a static environment and may break down in the presence of a dynamic environment such as trees moving in the wind (Stefanik et al., 2011)

2.4 Shape from Stereo

A shape from stereo (Stereovision) design has also been explored in the past. This technique is quite similar to SfM but instead of a sequence of monocular images, a calibrated rigid set of two cameras separated by a fixed distance is utilized to capture 3-D information at a single time. Stereovision setup implemented in UAVs provides a unique advantage in comparison to Structure from motion because UAV motion is not required to extract 3D data. This implies 3D measurements could be performed with this setup from a hovering aircraft. The current stereovision technique in UAVs involves a stereo setup where two cameras were rigidly mounted on an autonomous helicopter for real-time generation of 3-D terrain models (Stefanik et al., 2011). The setup however had an accuracy ranging from 56cm to 65cm across the field of view at an attitude of 40m. This is as a result of the spatial accuracy being dependent on camera resolution, baseline distance between the cameras and the altitude above ground level (AGL). The wider the base length and focal length, the higher the system precision (Okutomi and Kanade, 1993).There are of course limitations on how far apart the cameras could be attached on the vehicle due to the physical constraints imposed on mounting the cameras on the vehicle.

From the literature cited so far, it soon becomes apparent that existing 3D measurement technologies are limited in their application for use on UAVs, as a result of their various restrictions. This paper is concerned with investigating new techniques to provide more flexible and robust airborne measurement systems for deployment on UAVs that could deliver a spatial resolution and measurement accuracy comparable with existing methods for ground-based 3-D measurements.

3. Contributions

An airborne flexible stereo setup is proposed, whereby a small fleet of UAVs each equipped with a camera can be used to provide a configurable base-length between stereo pair leading to an adjustable accuracy for measurement. Implementation of these multi-view stereo systems on GPS-based UAV platforms would be completely novel. It would have the potential to provide a higher level of flexibility and redundancy/robustness than existing systems and would allow 3-D measurements from unconventional viewpoints in comparison to existing ground-based 3-D optical measurement techniques. However in other to achieve this, the control of UAVs from a codebase needs to be addressed. The use of open-source autopilot systems which run the MAVLink protocol is presented as a solution. Dronekit API and SITL have been implemented with the aid of this protocol to successfully guide a simulated vehicle to preassigned GPS coordinates

4. UAV Control

4.1. MAVLink Protocol

In order to enable deployment of these robust, scalable GPS-based UAV platforms the MAVLink protocol is required. MAVLink protocol is a communication protocol that allows control of MAVLink-enabled aerial vehicles. In this paper, MAVLink enabled vehicles are referred to as open-source autopilot systems that support the MAVLink communication protocol. MAVLink was first released in 2009 by Lorenz Meier (QGroundControl, 2014). This protocol is a very lightweight header protocol that has only 8 bytes overhead per packet and has built in packet-drop detection, among other features. The Pixhawk autopilot (flight controller) is an example of an autopilot system that supports the MAVLink protocol as shown in figure1 below. This hardware device runs the open-source ArduCopter firmware. ArduCopter is one of the most mature and popular open-source flight software systems in use today. In addition to the standard manual controls found in flight control software, ArduCopter also has a number of useful autonomous features such as *Loiter,* which allows the UAV to hold its current position using GPS and altitude data, *ReturntoLaunch* which calls the UAV back to its home position, and *GUIDED,* which allows the UAV to follow GPS coordinates assigned to it. The codebase implemented for UAV Control in this application utilizes the *GUIDED* mode to enable a vehicle follow preassigned GPS coordinates.
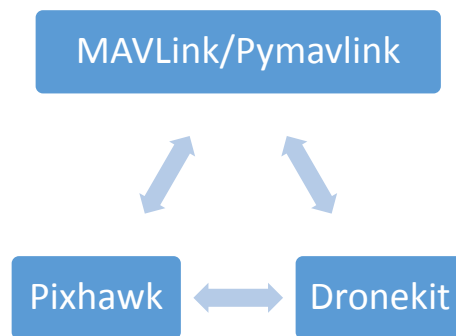
Figure 1: Software stack for communicating with MAVLink-enabled vehicles

By default, MAVLink provides binding to the C programming language which would enable the GCS to transmit C structures encoding commands. However a python a library named Pymavlink built upon the C library provides Python bindings and allows for the transmission of Python like-objects with MAVLink. The use of a scripting language like Python provides the advantage of an extremely simple and friendly syntax that can easily display structures without having to write a lot of formatting code for each of the messaging types. The Pymavlink library is not directly implemented in this application, but however provides the first layer of abstraction in the software stack as shown in figure1 above.

4.2. Dronekit API

Dronekit is a set of open source application program interfaces (API) developed by 3D Robotics (Dronekit, 2015). It is built upon the Pymavlink module as shown in figure1 and provides APIs for Python and Android. The API is compatible with vehicles that communicate using the MAVLink protocol such as the Pixhawk flight controller and provides classes and methods to:

- Connect to a vehicle or multiple vehicles from a python script
- Get and set vehicle state and obtain telemetry information

- Guide UAVs to a specified GPS position in GUIDED mode
- Send arbitrary custom messages to control UAV movement
- Override RC channel settings.

The customised GCS implements the use of Dronekit API to connect to a vehicle and extract telemetry information which contains the present GPS coordinates of the vehicle and also changes the mode of a vehicle to a GUIDED mode which permits a vehicle to follow GPS coordinates assigned to it.

4.3. Software in the Loop testing of vehicle

Before deploying any code written for our UAV on actual hardware it is important to test the code using Software in the Loop (SITL) simulation/modelling to identify and fix any bugs in the code that could lead to unexpected behaviour of UAVs in flight. The SITL is a build of the autopilot code using an ordinary C++ compiler which runs as an executable in a Linux environment and also uses a full sensor emulation. It essentially allows the running of the UAV Copter firmware without any hardware being present (ArduPilot Dev Team, 2016).

The SITL simulator can be started with the command **sim_vehicle.sh** after installing the required packages needed to run the SITL simulator and also adding the necessary directories to the search path in a Linux environment (ArduPilot Dev Team, 2016). Once the SITL simulator is spawned, it opens up and waits for Transmission Control Protocol (TCP) connections on local IP address using port 5760 (*127.0.0.1:5760).* Additional TCP ports 5762 and 5763 are also open for connection.

Dronekit API then provides functionality in python to connect to the SITL vehicle on the target port 5760. Once connection to the SITL vehicle is established, the *VehicleMode()* class is called to obtain the current mode of the vehicle and also set the vehicle's mode to GUIDED which permits vehicle to follow assigned GPS coordinates. Once in GUIDED mode, a call to the *LocationGlobalRelative*() class is made which expects an argument of desired GPS latitude and longitude relative to the World Geodetic System (WGS84) coordinate system and also altitude relative to the home position of the vehicle. These classes mentioned above are provided by the Dronekit API and have been used to write code to enable the SITL vehicle to follow preassigned GPS coordinates as shown in figure2 and figure3 below.

Cursor: -35.363048 149.165339 (S 55 696729 6084542) 584.5m
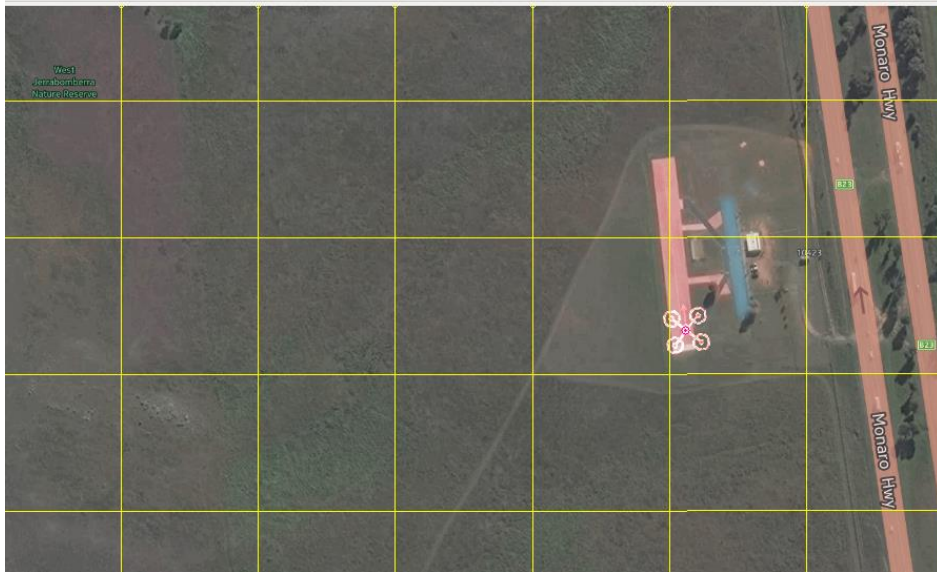
Figure 2: Simulated Vehicle located at home location GPS Latitude: -35.363048 and Longitude 149.1653339

Cursor: -35.363552 149.163824 (S 55 696590 6084490) 585.0m

Figure 3: Simulated Vehicle located at GPS latitude: -35.363552 and Longitude 149.163824

5. Conclusion

This paper has addressed the various limitations of existing 3-D measurement methods. This work then proposes an airborne adaptable 3-D stereovision system as a potential solution which would potentially open up many new application areas for 3-D precision scanning. In order to achieve this, open source software and hardware has been heavily utilized to communicate with simulated vehicles in SITL to for code debugging before

they are deployed on actual hardware. The next phase of this research will involve deploying the code on actual hardware devices with the aid of Dronekit API and the MAVLink protocol to guide UAV pairs to preassigned GPS coordinates. This setup would allow for the development of a fleet of UAV with adjustable base-length properties between each pair, which can potentially lead to the development of a novel airborne stereo setup.

6. References

ARDUPILOT DEV TEAM. 2016. *Setting up SITL on Linux* [Online]. Available: http://ardupilot.org/dev/docs/setting-up-sitl-on-linux.html.

CLAPUYT, F., VANACKER, V. & VAN OOST, K. 2015. Reproducibility of UAV-based earth topography reconstructions based on Structure-from-Motion algorithms. *Geomorphology*.

DRONEKIT. 2015. *Dronekit-Developer tools for drones* [Online]. Available: http://dronekit.io/.

GENCHI, S. A., VITALE, A. J., PERILLO, G. M. E. & DELRIEUX, C. A. 2015. Structure-from-Motion Approach for Characterization of Bioerosion Patterns Using UAV Imagery. *Sensors (14248220),* 15**,** 3593-3609.

NAGAI, M., CHEN, T., SHIBASAKI, R., KUMAGAI, H. & AHMED, A. 2009. UAV-Borne 3-D Mapping System by Multisensor Integration. *Ieee transactions on geoscience and remote sensing,* 47**,** 701-708.

OBANAWA, H., HAYAKAWA, Y. S. & GOMEZ, C. 2014. 3D modelling of inaccessible areas using uav-based aerial photography and structure from motion. *Chikei/Transactions, Japanese Geomorphological Union,* 35**,** 283-294.

OKUTOMI, M. & KANADE, T. 1993. A Multiple-Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 15**,** 353-363.

OLSON, C. F. & ABI-RACHED, H. 2010. Wide-baseline stereo vision for terrain mapping. *Machine Vision & Applications,* 21**,** 713-725.

QGROUNDCONTROL. 2014. *MAVLink Micro Air Vehicle Communication Protocol - QGroundControl GCS* [Online]. Available: http://qgroundcontrol.org/mavlink/start.

STEFANIK, K., GASSAWAY, J., KOCHERSBERGER, K. & ABBOTT, A. 2011. UAV-based stereo vision for rapid aerial terrain mapping. *GIScience and Remote Sensing,* 48**,** 24-49.

TEULIÈRE, C., ECK, L. & MARCHAND, E. Chasing a moving target from a flying UAV. 2011 / 01 / 01 / 2011. 4929-4934.

WESTOBY, M. J., BRASINGTON, J., GLASSER, N. F., HAMBREY, M. J. & REYNOLDS, J. M. 2012. 'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology,* 179**,** 300-314.

ZHOU, G., YANG, B., ZHANG, W., TAO, X., ZHAO, W., YUE, T., ZHOU, X. & YANG, C. Simulation study of new generation of airborne scannerless LiDAR system. 01 / 01 / 2013. 524-527.

Adeniyi Olumide

Initial Investigations into Through-Life Monitoring of Solder Joints

Olumide A.Adeniyi, Guang-Ming Zhang, David M.Harvey, Derek R. Braden

Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK.

E-mail address: O.A.Adeniyi@2015.ljmu.ac.uk.

Abstract. In this study, solder joints were monitored using ultrasonic transducers. It is well documented that solder interconnections are the weakest link in terms of a circuit board assembly's (CBA) reliability. This is due in part to the coefficient of thermal expansion mismatch in materials used in the construction of components found on CBA when exposed to thermal cyclic, environmental conditions which in turn lead to fatigue failures. This paper presents a method to monitor solder joints in area array packaging using non-destructive techniques. Test boards with organic substrate thickness of 0.8mm and 1.6mm, containing 6 flip-chips and 8 BGA chips were subjected to an accelerated thermal cycling test (ATC) that ramps between 125°C and -40°C.Test boards were monitored at regular intervals by Acoustic Micro Imaging. Experimental work to show difference image qualities for transducers with different frequencies were performed.

Keywords: Accelerated Thermal Cycling, Reliability.

1. Introduction

Increasingly high connection density area array packages are used to deliver the complex features and functionality required by modern consumers. The ever reducing package size results in smaller interconnect size (solder bumps) and unsurprisingly results in allied reliability issues experienced at the interconnect boundaries. Undeniably this is a major disadvantage of area array packages such as Ball Grid Array (BGA) and flip chips, since the solder joints cannot flex in the way that longer leads can, as they are not mechanically as compliant.

Solder joints form a mechanical as well as an electrical connection in electronic devices. During usage circuit board assemblies are exposed to cyclic changes in temperature. The rate of change, exposure time and thermal excursion limits are dependent upon product application and usage sometimes known as 'Mission Life'. However, the package to printed circuit board (PCB) solder joint interconnections are extremely important since they represent, in many cases, the weakest link in terms of product reliability. Solder joints however do not only provide excellent electrical connection, but they mechanically affix the component to the PCB, provide conduits for heat dissipation into the circuit board and help match expansion differences between PCB and populated components.

The response to temperature change is a result of environmental and power dissipation of individual components. All of which create stress related thermal cycling conditions on solder joints. Heat can be generated by the device itself as the current flows through it, but may also be generated by the external environment to which it is exposed. For example an engine mounted electronics control unit (ECU) in a vehicle application because of the different thermal properties of the materials involved, the temperature increase causes coefficient of thermal expansion mismatches. Due to this mismatch deformation between the

organic substrate and silicon die occur during environmental operation. The joints then start to deform and creep occurs inside them. Over time the creep deformation induced by the temperature variations causes cracks to initiate and to grow.

The level of cracking impairs the electrical functionality of the joint. Thus one important quality measurement is the number of temperature cycles an electronic device is able to withstand, can be used as a metric which defines joint reliability. Accelerated Thermal Cycling (ATC) is used to predict quality during the time a product will used. For example, the design life of a car is typically 15 years. Therefore it is not unrealistic to expect 8000 hours or 5000 to 1000 thermal cycles over 15 years. If a thermal cycle in the field is 2 hours on average then we are looking at 10000 to 20000 hours of testing. This is too long compared to design life cycles hence we have to accelerate the testing.

During the useful life of a modern electronic circuit board, its reliability is a function of various stresses like thermal, mechanical and electrical over time. Pahl et al [1] explain that an increase in thermal stresses of an electronic circuit board increase the failure rate and decrease the reliability through electrochemical degradation processes. High temperature can also cause melting of the solder joints of the flip chip on PCB and also show progressive impairment of the performance levels due to degradation effects [2].

But in this project, quantitative ATC tests will be performed, to quantify the reliability of solder joints under different thermal conditions and produce adequate data that will be required for future data analysis.

During the last decade, Non-Destructive Evaluation (NDE) has played a role in the development of better processes and process control measures used in manufacturing. However it is still mainly used in high reliability products to verify and validate through test particularly a given design. For companies deploying electronics in their products, particularly in harsh environments such as automotive, the failure of the electronics components can be expensive in terms of warranty issues resulting in product recall and damage to brand image. One of the most effective ways to demonstrate solder joint integrity and reliability is through reliability testing. Therefore, using a non-destructive method to assess through-life performance during thermal excursions and to quantify product reliability becomes a desired tool, for products designers. It also allows engineers to evaluate manufactured Circuit Board Assembly (CBA) without the need to physically cross-section components which destroys them and instead leaves products intact.

Several non-destructive methods techniques have been proposed to investigate the defects in microelectronics devices; these include X-ray microscopy [3], Optical inspection, Scanning Acoustic Microscopy (SAM) [4], Scanning electron microscopy and Laser ultrasound coupled to interferometry.

Inspection is performed with X-ray by using the variation in the transmission absorption of X-ray energy due to differing materials used in the construction of a component. A material absorbs X-rays proportional to its atomic mass and density [5]. This makes X-ray suitable for the investigation of volumetric type defects. Unfortunately, X-rays system resolution can be no better than the diameter of the X-ray source or spot size of the transmission tube, in which the present tubes are capable of achieving the spot sizes around 1μm. Braden et al [6], proposed a technique on through life non-destructive evaluation in which a key solder feature, nucleation at the bump to silicon interface was measured. When this propagated as a laminar crack plane captured was attempted using Acoustic Micro Imaging (AMI). The measurement data was compared to Finite Element Analysis (FEA) studies in order to understand the differences in reliability prediction. Furthermore, it was found in this study that FEA simulations which take into account floor plan layout and constraint points

(resulting from gluing or screwing CBA's to a metal housing) showed a difference in predicted reliability outcome compared to traditional simulation methods. From the literature reviewed so far, various techniques have been used for locating and characterizing defects in microelectronics device having different materials structures.

It soon becomes apparent that existing non-destructive evaluation and life monitoring of modern electronic circuit board are limited in their application for use in electronic industries. This work is concerned with developing and performing through-life monitoring of solder joints in an electronics systems subjected to thermal fatigue cycling environments.

2. Methodology

The lifetime prediction of solder joints in area array packaging applications is a quantitative means of predicting reliability. This section provides details of how to use non-destructive inspection tools to monitor the solder joints through life performance during thermal cycling as shown in Figure 1:
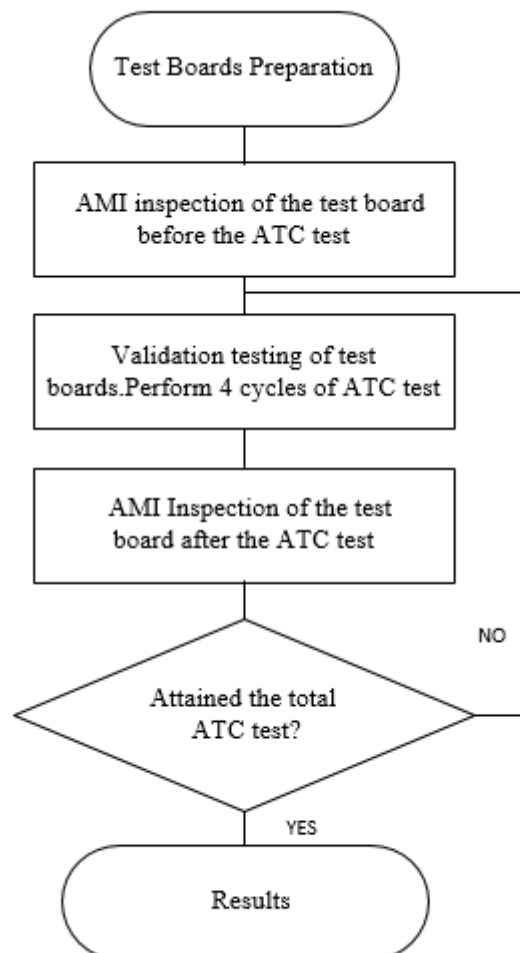
Figure 1: Flowchart showing Methodology for the project work

2.1. Experimental testing:

2.1.1. Accelerated Temperature Cycling Test of Solder Joints

Test boards with organic substrate thickness of 0.8mm and 1.6mm, containing 6 flip-chips and 8 BGA chips were designed. The flip-chip packages contained 109 solder joints. Which were assembled without underfill in order to age the flip-chips in a short period of time. The minimum temperature used was -40°C and the maximum was 125°C. The dwell time used was 35minutes because the longer the dwell time, creep damage accumulates faster. The test was performed for a period of 4 thermal cycles. After that the test boards were removed from the temperature chamber for acoustic microscopy imaging to determine if any defects were on the boards. This cycle will be repeated until all solder joints fail and their reliability profiles have been recorded.

2.1.2. Acoustic Microscopy Imaging

Acoustic Microscopy uses high-frequency ultrasonic energy, typically from 10 MHz to 300 MHz pulsed from a focused transducer through a coupling medium such as distilled or deionized water into the unit to be tested to determine air gap type defects such as cracks, voids, or delaminations, as well as material changes such as porosity and inclusions. The reflected pulses were used to generate ultrasound C-scan images. Considering the penetration and resolution, a number of transducers with different frequencies 50MHz, 100MHz and 230MHz were used to scan the flip-chips on the test boards.

3. Results

3.1. ATC Result

Figure 2 shows the thermal profiling results obtained for a test sample when subjected to four thermal cycles. With the thermal chamber set to a mission temperature from -40 degree C to 125 degree C. The test board showed a thermal response in terms of being able to track the ATC profile. The maximum temperature of the PCB reached was 125°C with the minimum temperature reached being -40°C. It well found that dwell times at which the test objects reached the minimum and maximum temperatures varied across the samples and sample positions. Moreover, the dwell time of 125°C of the top center of the PCB was 15 minutes and that of the bottom center and the top and bottom corner was 10 minutes. The dwell time of -40°C of the top center of the PCB is about 20 minutes and that of the bottom center and the top and bottom corner is about 13 minutes. The longer the dwell time, the more accumulated creep damage there will be. The ramp time in going between the temperature extremes at 5 degrees C per minute was 33mins.
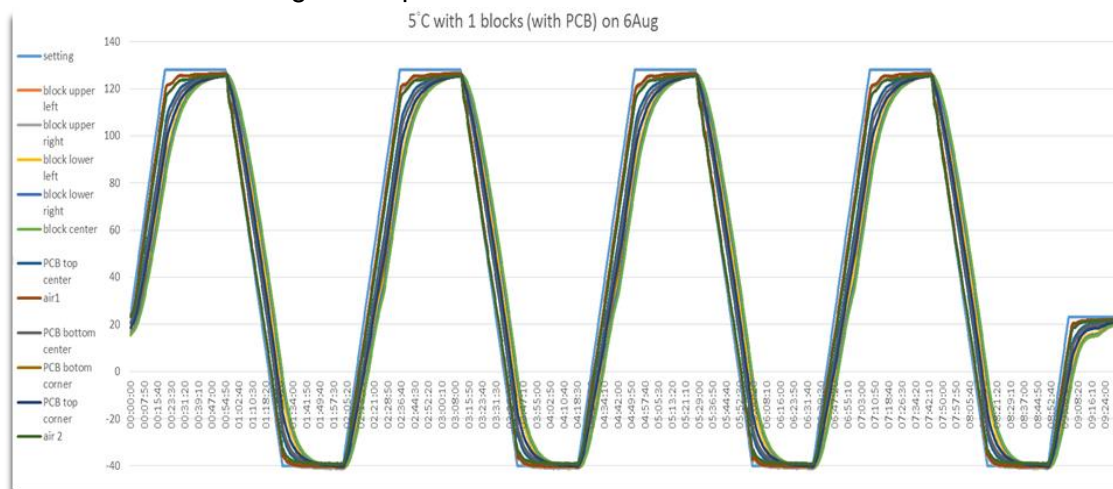
Figure 2: Example of ATC profile chamber temperature-40°C to 125°C, dwell time 30mins.

3.2. AMI Result

In this project, a 50MHz piezoelectric transducer was used to scan the test boards with 0.0520mm resolution, as shown in Figure 3a. The image acquired was blurred, because lower frequencies give less resolution but greater penetration, so are unable to provide the required lateral and axial resolution. While in Figure 3b, a 100MHz transducer was used to scan the same test board with 0.0260mm resolution. it gives better resolution compare to 50MHz because a higher frequency transducer has higher spatial precision. Finally, a 230MHz transducer with 0.0113mm resolution was used to scan the test boards, in order to provide the highest possible axial resolution as shown in Figure 3c.
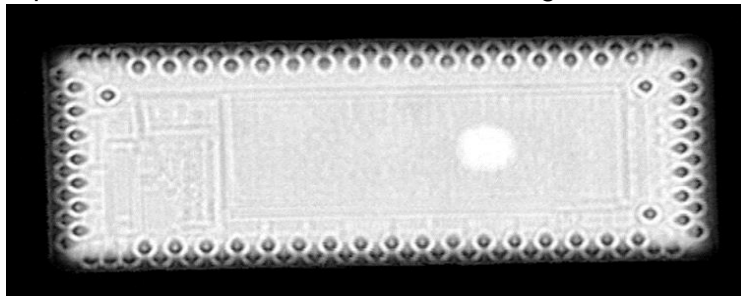
Figure 3a: C-scan image of flip chip using 50MHz Transducer.
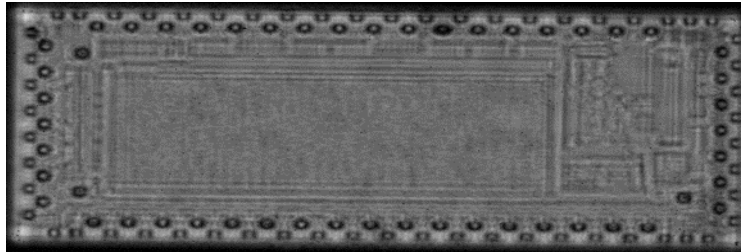
Figure 3b: C-scan image of flip-chipUsing100MHz Transducer.
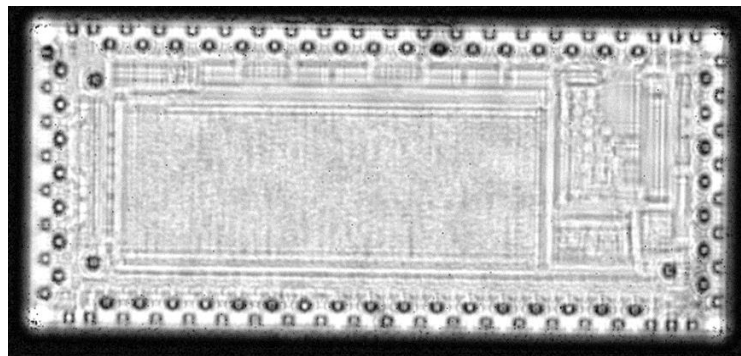
Figure 3c: C-scan image of flip-chip using 230MHz Transducer.

As shown in Figure 3d, the solder joint appears as a black shape, with a light grey area in the middle which represents the joints behaviour and reaction during the solder joint reliability test. This light grey region in the centre indicates the quality of the bonding and that a connection between the die and the bump exists. Due to the intrinsic properties of the ultrasound signal striking the materials results in the signal being scattered at the solder joint edges causing a loss of information, which forms the darker pixels surrounding the central lighter pixel area. The number of pixels that form the diameter of the grey region of interest in Figure 3d was 41 pixels. Both the light grey and dark regions will provide useful information to aid the defect analysis of the solder joint which is going to enables to classify and measure the defects. The wider the recorded area of interest, the more accurate the defect measurement resolution will be.
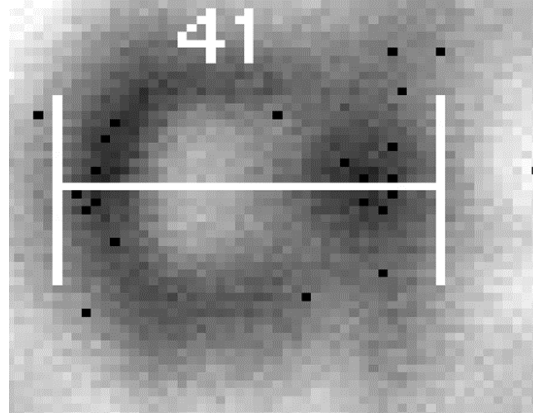
Figure 3d: Single solder joint image using 230MHz

4. Transducer Characterization

Various transducer with different frequencies was used in this work, the characteristics of low and high frequency in AMI inspection is illustrated in Table 1.

Table 1: Resolution, scan Time and focal length of the transducer frequencies used

Frequency(MHz)	Focal Length (Inches)	Resolution (pixel)	Resolution (mm)	Scan Time
50	0.50	424*424	0.0520	58s
100	0.50	836*836	0.0260	1 min 4s
230	0.25	2252*2252	0.0113	7 min 4s

5. Conclusion

In conclusion, the solder joints in area array packaging were monitored using various ultrasonic transducers. There were clear differences in the image qualities, including solder joint and pixel resolution key knowledge has been gained in the non-destructive ultrasound measurement of solder joints. Particular attention was paid to improving image resolution and quality through taking more accurate measurements, optimizing machine setups and transducer selection. ATC tests were carried out to understand thermal profiling and will be used to generate rapid ageing of the test boards, so that the effects of ageing can be studied in a shorter period of time. These results are encouraging and suggest that the approach proposed in our project shows a line of enquiry worth pursuing.

6. References

[1] Pahl B., Kallmayer C., Aschenbrenner R., and Reichl H, 2004.Long time reliability study of soldered flip chips on flexible substrates, Microelectronics Reliability, vol.44, no.2, pp.309-314.

[2] Turek S., Ibrahim T.O., Dababneh A., Elgaali O., Zhang Y., and Marler T, 2012. A model based enterprise approach in electronics manufacturing. Computer-Aided Design Applications, vol.9, no.6, pp.847–856.

[3] Moore T.D, 2002. Three dimensional X-ray laminography as a tool for detection and characterization of BGA package defects. IEEE Trans CPT, vol.25, no.2, pp.224-229.

[4] Semmens J.E., Martell S.R., and Kessler L.W, 1996. Analysis of BGA and other area array packaging using acoustic micro imaging. Proceeding of SMTA Pan Pacific Microelectronics Symposium, pp.285-290.

[5] Bernard D, 2003.Selection Criteria for X-ray Inspection systems for BGA and CSP solder joints analysis, NEPCON, Shanghai.

[6] Braden D.R, Yang R.S.H., Zhang G.M., and Harvey D.M, 2010. Investigation into impact of component floorplan layout on the overall reliability of electronics system in harsh environments. 3rd Electronics System Integration Technology (ESTC), pp.13-16, September, Berlin.

Ali Alzeyadi

Evaluate the performance of phosphate adsorption by materials packed in upflow filter

Ali Alzeyadi [1, 3], **Edward Loffill** [2], **Rafid Alkhaddar** [2]

[1] Postgraduate Research Student, Liverpool John Moores University, Civil Engineering Department, Henry Cotton Building,15-21 Webster Street, Liverpool, L3 2ET, UK
[2] Liverpool John Moores University, Civil Engineering Department, Peter Jost Enterprise Centre, Byrom Street, Liverpool, L3 3AF, UK
[3] Al Qadissiya University, IRAQ.
E-mail address: A.T.Alzeyadi@2013.ljmu.ac.uk

Abstract. All over the world, the standards of environmental protection became more rigorous. Hence, the secondary sewage treatment SST became no longer enough to guarantee the effluent quality. Advanced tertiary treatment is required for further decreasing of the residual constituents from SSE. Phosphate is released into the aquatic environment because of many human activities; generally, wastewater treatment works WTWs represents one of the major phosphate sources. A variety of technologies have been developed and applied for the phosphate removal such as chemical precipitation, Enhanced biological Phosphate removal EBPR and Constructed wetlands CW. Continuous, Up-flow filter CoUF is one of the promising technology in tertiary treatment; this study seeks to develop CoUF for phosphate removal. This work have been focus on improve the type of filter media; several selected media subjected to examination for demonstrate their affinity, capacity and reactivity for Phosphate. The chemical and physical characteristics are the key factors that help to understand the removal process. Thus, the material have been selected according to the metals sort they contain them (Fe/Al hydroxide or containing soluble Ca) were they offer a good tendency for Phosphorus sorption.

The results show that the Phosphate removal efficiency is highly dependent on the effluent contact time, particle size distribution PSD and chemical composition. In conclusion, the metals types played the crucial role in the treatment process in comparison with other factors. The filter media that contain Fe/Al hydroxides exhibited better phosphate removal than media contain soluble Ca in terms of contact time and PSD.

Keywords: continuous, Up-flow filter, Fe/Al hydroxide, phosphate, tertiary treatment, wastewater treatment.

1. Introduction

Over the last decades, phosphorus P reduction from discharging into the aquatic environment is representing a serious and costly challenge in many parts of Europe and other countries around the world (Delgado and Scalenghe, 2008, Ballantine and Tanner, 2010, Buda et al., 2012). The effluent from wastewater treatment plants WWTPs is the predominant point sources of P, in Europe it is contribute in more than 50% of contaminant (Valsami-Jones, 2004, Farmer, 2001).

Despite sand and gravel have a limited capability for P removal but they are utilized as P removal media for long time. Currently, searching for materials with a significant affinity for P is required to achieve the necessary water quality as declared in the EU Water Framework Directive WFD (Kaasik et al., 2008). Based on this ground, many researchers considered the investigation on alternative materials has become essential as one of the sustainable technologies for P removal. According to Johansson Westholm (2006) these materials can be classified into three groups: natural materials, industrial by-products, and manufactured products. Recently, large number of Phosphorus sorbing materials PSMs have examined to determine their capability to remove P from wastewater. The most studies are discussed the materials removal capacity for high P concentrations (Cucarella and Renman, 2009).

Many authors' states that the efficiency of the material that act as P filter is depend on the chemical composition of these materials. Especially, if they contain hydroxides and oxides such as Al, Fe, Ca and Mg (Vohla et al., 2011). In addition, the physical properties that characterize the material play a vital role in mechanism of P removal (Mann, 1996). The size, shape and porosity of the materials have a considerable influence on the surface reactivity between the materials and the effluent. Hence, the increase of specific surface area SSA for filter media effect positively on P sorption (Mann, 1996). The main target of this work is to investigate proper materials to act as P-filter media. The laboratory work include packing material in a lab scale upflow filter. The good filter material must interact rapidly and have a tendency to P sorption and high capacity to P retain.

2. Materials and method
2.1 Filter material

Timber bottom ash TBA and limestone have been selected to evaluate their eligibility to act as P-filter media according to their chemical composition and availability. The chemical composition is responsible on define the interaction between the materials and P ions. According to the literature review as mentioned in introduction, the materials that contain hydroxides and/or oxides such as Al, Fe, Ca and Mg is related to the P removal efficiency. Based on this fundamental, the materials were selected according to their chemical content from hydroxides and oxides. X-ray fluorescence analyser XRF was performed to show that TBA content 2.168 Aluminium oxide, 1.397 Iron oxide, 7.615 calcium oxide and the rest are silica and fines. While the limestone contain 84.162 calcium oxide. Phosphate adsorbs to surfaces rich with calcium oxide; this is followed by precipitation of Ca-phosphorus ions. Thus, the limestone in this work called Ca-P. In addition, the sand capability for P removal has been examined. The sand result was compared with TBA and Ca-P to demonstrate the impact of the characteristics of the selected material on P removal mechanism. Figure 1 illustrates the total particle size distribution for each selected media; the material also fractionated to 1 and 2 mm particle size groups to obtain better investigation in determining the influence of physical and chemical characteristics on P sorping by PSMs.
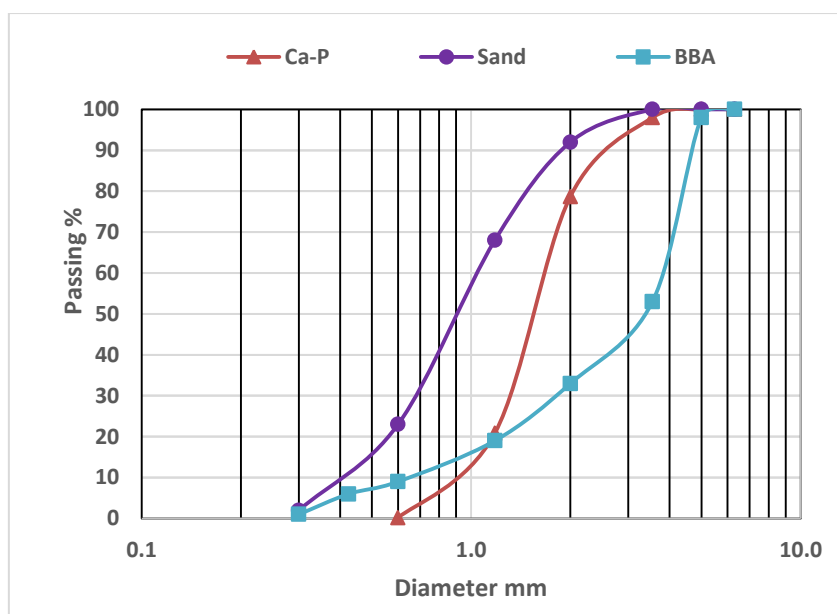
Figure 1: Particle size distribution PSD for the selected material

2.2 The synthetic influent

Phosphorus in wastewater exists in various forms such as particulate, organic and inorganic Phosphorus. The predominant form is the inorganic Phosphorus; especially, in form of orthophosphate. Therefore, monopotassium phosphate KH_2PO_4 is used as influent solution that pumped to the filtration system that set up in this work. The salt was dissolved in deionized water to prepare the phosphate solution. Commonly, the P concentrations in wastewater range from 5 to 10 mg P/l. However, in some cases they can be high as 20 or 30 mg P/l. The initial P concentration in this work was 10 mg P/l as to simulate the maximum P concentration in wastewater that is most expectable.

2.3 Lab scale filter

Three acrylic cylinders were used as filter columns in the experiments. The dimensions of each filter were 10 cm diameter and 30 cm depth. The depth of the packed media is 13 cm placed over 10 cm gravel layer lies at the bottom to maintain the influent transfer through the filter. The flow was operated as in upflow mode using submersible pumps at flow rate 1 l/min.

2.4 Operation condition

The filtration process time should be adequate to reach the equilibrium point for P sorption at each selected media. Moreover, P sorption mechanism is sophisticated and passes over different patterns. The process can be started with rapid sorption reactions followed by gentle processes that are then subdivided into relatively rapid and very slow reactions. Therefore, the P sorption equilibrium can be achieved after minutes, hours, days or months based on the sort of material that set as filter media, initial P concentration and temperature. The filter systems have been operated for 24 hours at room temperature as shown in table 1.

Table 1: the operating conditions for the experiments

Column	Packed material	PSD mm	Temperature C	flow rate l/min	operating time hr
1	TBA	0.28 to 5	19 to 21	1	24
2	Ca-P	0.6 to 4.2	19 to 21	1	24
3	Sand	0.3 to 3.5	19 to 21	1	24
4	TBA	1	20 to 22	1	24
5	Ca-P	1	20 to 22	1	24
6	Sand	1	20 to 22	1	24
7	TBA	2	21 to 22	1	24
8	Ca-P	2	21 to 22	1	24
9	Sand	2	21 to 22	1	24

2.5 Effluent sampling and measuring

The effluent samples were collected in plastic containers over the operating time of the experiments. At first the samples were collected each 5 minutes then each 15 min, 30 min, 1, 2 and 4 hr because the sorption rate decrease with the time. The collected samples then filtered with filter paper 0.45 μm and the phosphate concentration was measured according to the Amino Acid method by utilizing HACH LANGE DR 2800 spectrophotometer.

3. Results and discussion

The packet filter materials contain three types of minerals Al, Fe and Ca. According to XRF analysis, the TBA consists of Al, Fe and Ca in different percentages, while the Ca-P contain calcium oxide. On the other hand, sand poorly consists of Al, Fe and Ca; the silica is the main constituent in sand.

As illustrated in the table 1, the first run for each media has performed where the media packed according to their PSD as shown in the figure 1. Obvious variation in phosphate removal were observed over the course of the experiment. Figure 2 shows that phosphate removal of Ca-P and sand were less efficient than TBA. The TBA was removed 72.17% from the phosphate at the end of the experiment duration, while the phosphate removal by Ca-P and sand were 59.33% and 56% respectively.
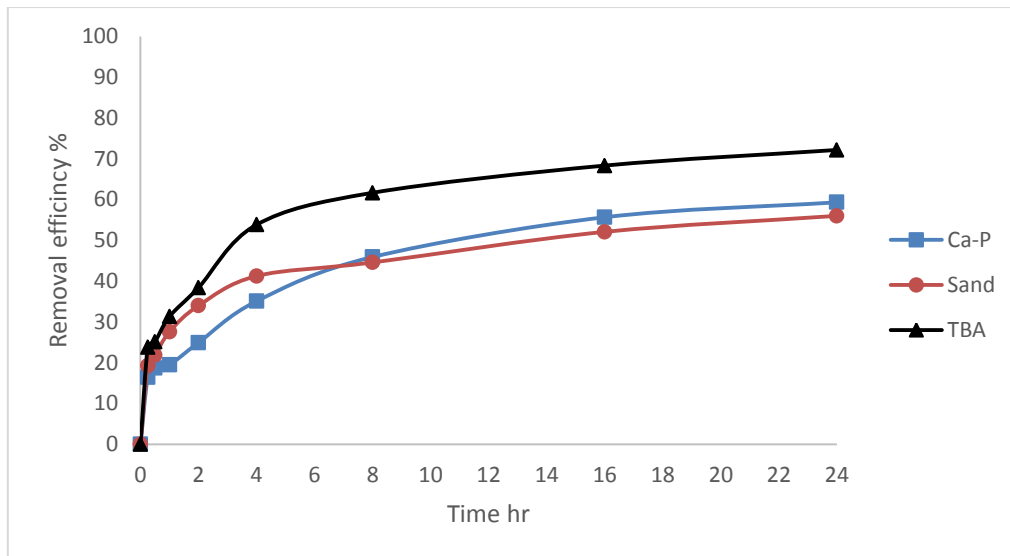
Figure 2: phosphate removal efficiency for the selected materials

It is noticeable that the phosphate removal pattern for Ca-P and sand has changed during the course of the experiment. The sand showed a faster phosphate removal for first 6 hr in comparison with Ca-P. Afterward the Ca-P starts to be efficient than sand. This indicates that the removal over the Ca-P is slow and need to long contact time with the influent to retain the phosphate. While the findings revealed that even though sand-filter has a relatively good phosphate sorption but it might have difficulties in retaining the Phosphate; may be attributed to the vacancy of sand from the metals ions which able to interact with phosphate ions. At the start of the experiment for less than 30 minutes, the phosphate retention efficiency was almost convergent for all materials (roughly 20% P removal). Afterward the phosphate retention efficiency for TBA increased dramatically in comparison with other two materials. Therefore, the strongly phosphate-retaining material is TBA. Ca-P was slightly more efficient than sand.

Particle size is one of the main factors which maintain the treatment efficiency. Furthermore, the durability and reliability of the filtration system. Each filter material was fractionated into two groups of particle size 1 and 2 mm by sieving. Then, the experiments were conducted at operating conditions as illustrated in table 1, for experiments 4 to 9. As in the previous experiment, submersible pumps in these experiments pumped synthetic phosphate solution concentration of 10 mg P/l. the results of phosphate binding per unit mass of media for both particle size were compared with the materials at the total particle size as shown in figure 3. Obviously, at all particle sizes the phosphate uptake by TBA is better than Ca-P and sand. Likewise, the materials at particle size 1 mm offers better sorption than other particle sizes because the small particle size correlates to the surface area positively. Small particle size lead to increase the available surface area for contact between the materials and the impurities. Thus, increase the chance of removal achievement. There is slight difference in phosphate sorption over the particle size groups for the Ca-P and sand. The maximum sorption which achieved by Ca-P for the particle sizes (1mm, 2mm and PS total) are (0.037, 0.033 and 0.033 mg p/g) respectively. While, the maximum sorption that achieved by sand for the same order of particle size groups are (0.037, 0.033 and 0.035 mg P/l). On the other hand, there is a significant difference in phosphate sorption can be observed between the particle size groups for TBA; the smallest particle size media shows high affinity for phosphate retention. As shown in figure 3 theTBA at particle size (1 mm) is retained the 0.75 mg phosphate per each g of media. While the phosphate retained by particle size (2 mm and PS total) were 0.042 and 0.048 respectively.
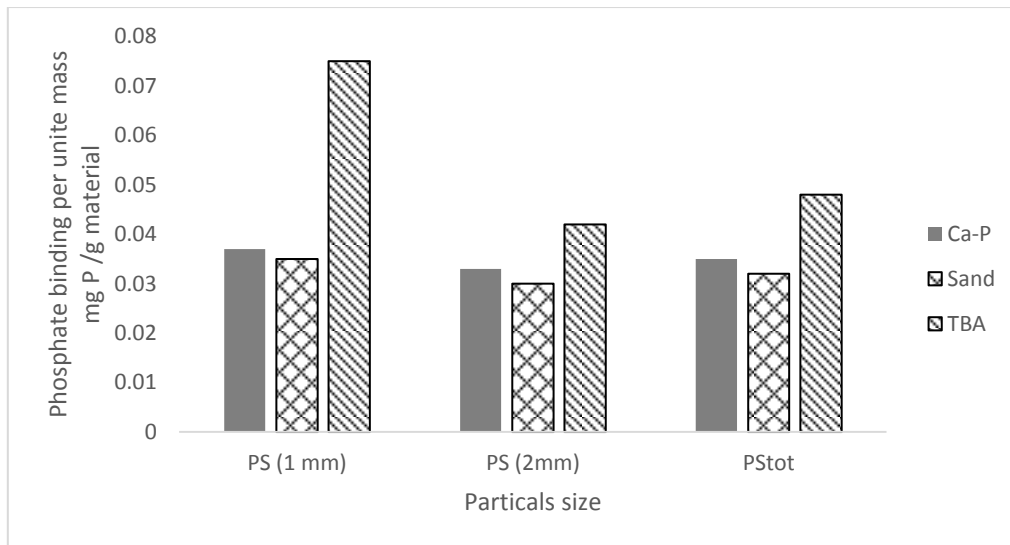
Figure 3: influence of particle size on the phosphate sorption

Advanced physical, chemical and biological processes have utilized to decrease the total phosphorus to achieve the requirements of the legislation of aquatic environment protection. 0.1 mg/l is the acceptable phosphate concentration, which discharged from wastewater into the water bodies. This standard value is obtained when the phosphate retention rate reach 0.05 mg P/l. As illustrated in figure 4 all materials have achieved a phosphate removal rate less than standard rate except the TBA at particle size 1mm. In short time from the experiment starts, exactly after less than 4 hr the TBA at PS 1 mm exceeded the removal rate 0.05 mg P/g. However, the TBA showed an outstanding phosphate removal efficiency in comparison with Ca-P and sand. The results suggest as well that the chemical composition for materials is the essential factor in retain the phosphate from the effluent. Especially, the material that contain the metals such as Fe/Al.
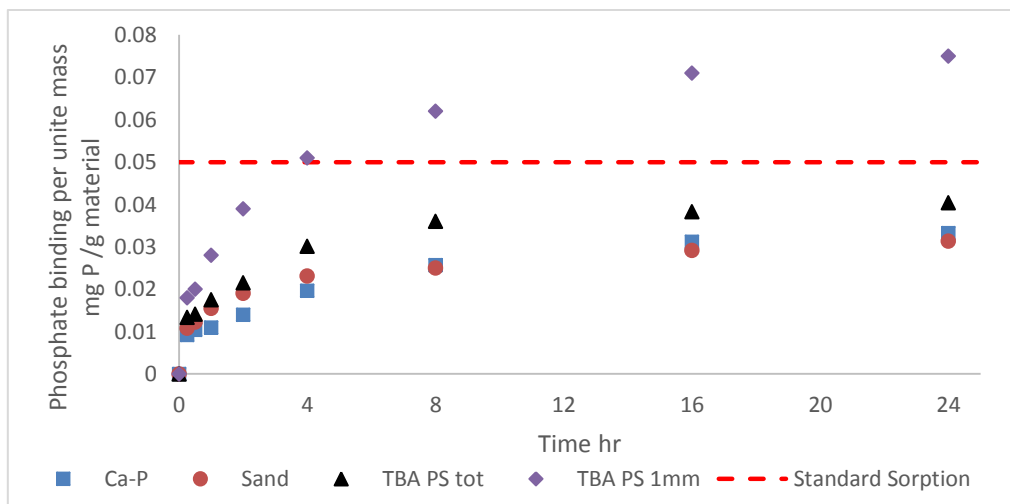
Figure 4: phosphate sorption rate in comparison to the standard phosphate rate

4. Conclusion

Sorption studies have carried out to define the phosphorus uptake by several material were selected according to their chemical composition. Their chemical composition investigated by X-ray fluorescence analyser and the finding revealed that the TBA contain a Fe/Al and limestone contain high percentage of Ca metal. TBA in comparison with Ca-P and sand achieved high performance for phosphate removal. This finding may be related to the type of

metals in the materials composition. The calcium ions need a plenty of time to absorb the phosphate ions. While, the Fe/Al hydroxides are adsorb them rapidly. According to the results, the particle size is one of the key factors in phosphate removal process but its influence will not be valuable for phosphate removal without present the Fe/Al hydroxides and soluble Ca oxides in the material composition.

References

Ballantine, D. J.;Tanner, C. C. (2010) Substrate and filter materials to enhance phosphorus removal in constructed wetlands treating diffuse farm runoff: a review. *New Zealand Journal of Agricultural Research,* 53**,** 71-95.

Buda, A. R.;Koopmans, G. F.;Bryant, R. B.;Chardon, W. J. (2012) Emerging technologies for removing nonpoint phosphorus from surface water and groundwater: Introduction. *Journal of Environmental Quality* 41**,** 621-627.

Cucarella, V.;Renman, G. (2009) Phosphorus sorption capacity of filter materials used for on-site wastewater treatment determined in batch experiments – A Comparative Study. *Journal of Environmental Quality,* 38**,** 381-392.

Delgado, A.;Scalenghe, R. (2008) Aspects of phosphorus transfer from soils in Europe. *Journal of Plant Nutrition and Soil Science,* 171 552-575.

Farmer, A. M. (2001) Reducing phosphate discharges: the role of the 1991 EC urban wastewater treatment directive. *Water Sci. Technol,* 44**,** 41-48.

Johansson Westholm, L. (2006) Substrates for phosphorus removal—Potential benefits for on-site wastewater treatment. *Water Research,* 40**,** 23-36.

Kaasik, A.;Vohla, C.;Mõtlep, R.;Mander, Ü.;Kirsimäe, K. (2008) Hydrated calcareous oil-shale ash as potential filter media for phosphorus removal in constructed wetlands. *Water Research,* 42**,** 1315-1323.

Mann, R. A. (1996) *Phosphorus removal by constructed wetlands: Substratumadsorption.* Ph.D, Univ. of Western Sydney.

Valsami-Jones, E. (2004) *Phosphorus in environmental technologies: Principlesand applications,* London, IWA Publ.

Vohla, C.;Kõiv, M.;Bavor, H. J.;Chazarenc, F.;Mander, Ü. (2011) Filter materials for phosphorus removal from wastewater in treatment wetlands—A review. *Ecological Engineering* 37**,** 70-89.

Abdelkarim Ertiame

Adaptive Nonlinear Observer Based Fault Detection in Nonlinear Multivariable System Using a Learning Methodology

Abdelkarim m Ertiame, L Yu, and J B Gomm

Control System Research Group, School of Engineering, Liverpool john Moores University, Byrom Street, Liverpool, L3 3AF, UK.

A.M.Ertiame@2011.ljmu.ac.uk

D.Yu@ljmu.ac.uk

J.B.Gomm@ljmu.ac.uk

Abstract. In the recent engineering issues the task of the detection and diagnosis of the system failures are being intensively more significant. This paper presents a robust fault diagnosis scheme for abrupt and incipient faults in nonlinear dynamic system. A general framework is developed for model-based fault detection and diagnosis using on-line approximators and adaptation/learning schemes. In this framework, neural network models constitute an important class of on-line approximators. The changes in the system dynamics due to fault are modelled as nonlinear functions of the state, while the time profile of the fault is assumed to be exponentially developing. The changes in the system dynamics are monitored by an on-line approximation model, which is used for detecting the failures. A systematic procedure for constructing nonlinear estimation algorithm is developed, and a stable learning scheme is derived using Lyapunov theory. Simulation studies are used to illustrate the results and to show the effectiveness of the fault diagnosis methodology.

Keywords. Robust fault detection; model-based fault detection; nonlinear observer; on-line   approximators; nonlinear adaptive estimator.

Introduction

With associate increasing demand for higher performance moreover as for a lot of safety and reliability of dynamic systems, fault diagnosis has received a lot of attention. The matter of on-line fault detection and isolation has become a serious issue in chemical engineering. Several fault diagnosis (FD) approaches have been proposed for processes operating mainly in steady-state conditions e.g., continuous reactors. Due to the high nonlinear dynamics and unsteady operating conditions of batch chemical systems, the application of these techniques are very challenging task to implement. [2], [5], [7], [8] and [10].The application of observer based fault detection (FD) has been intensively studied over the last two decades. In [5], detecting faults in nonlinear dynamic systems using observer model based approach was proposed. Another method was studied in [6] used a learning methodology for failure detection and accommodation. In [2] authors studied the design and analysis of a general framework for model-based fault detection and diagnosis of a class of incipient faults. An automated fault diagnosis architecture using nonlinear online approximators with an adaptation scheme is designed and analysed. In [7] researchers presented in their paper a robust fault diagnosis scheme for detecting and approximating state and output faults occurring in a class of nonlinear multiinput–multioutput dynamical systems.. Another study was conducted in [9] described a fault diagnosis algorithm for a class of nonlinear dynamic systems with modelling uncertainties when not all states of the system are measurable. In [4] authors developed a nonlinear observer-based approach for

distributed fault detection of a class of interconnected input–output nonlinear systems. In [10] researchers studied a robust fault diagnosis scheme for abrupt and incipient faults in nonlinear uncertain dynamic systems.

In this work a fault diagnosis methodology for incipient and abrupt faults is developed. We consider nonlinear dynamical systems whose dynamics change at some unknown time due to a failure. This change is modelled as an unknown nonlinear function of the state and input variables with a time-varying failure profile. In order to capture the nonlinear characteristics of faults, we design a nonlinear estimator using the online approximation (OLA) approach with an adaptive scheme for the adjustable parameters or weights. The stability and performance properties of the fault diagnosis scheme are rigorously established under the assumption of full state measurement. These results are obtained in the presence of *approximation errors,* that is, errors arising as a result of imperfect modelling of the system Deviations due to faults by the online approximator. From an adaptive theory viewpoint, the objective of this section is to develop a *learning* methodology for incipient and abrupt failure detection. In this framework, online approximators (such as neural networks, spline functions, wavelets, etc.) are used to monitor the system for any deviations due to faults. By using the adaptively capabilities of online approximators, they can be used not only to detect the occurrence of System failures, but also to provide an online estimate of the fault characteristics (diagnosis [2],[5],[7],[8] and [10].

Modelling A non-adiabatic Continues Stirred Tank Reactor (CSTR)

A common chemical system encountered in the process industry is the Continuously Stirred Tank Reactor (CSTR). Here we will study a jacketed diabetic (i.e., non-adiabatic) tank reactor described extensively in Bequette's book "Process Dynamics: Modelling, Analysis and Simulation", published by Prentice-Hall, 1998. The vessel is assumed to be perfectly mixed, and a single first-order exothermic and irreversible reaction. A Schematic diagram of the vessel and the surrounding cooling jacket is shown in Figure 1.
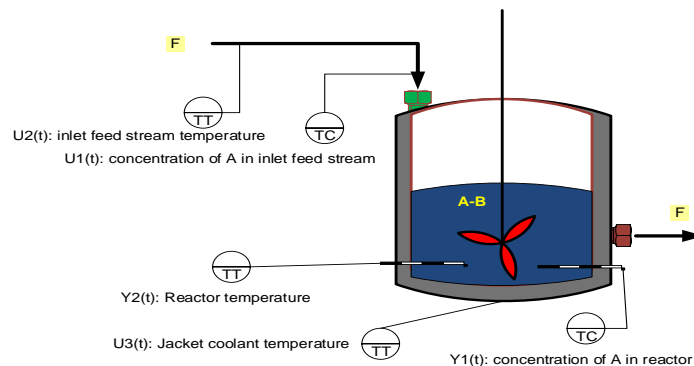
Figure 1 Schematic diagram of a CSTR.

Modelling the CSTR

The CSTR system is modelled using basic accounting and energy conservation principles. The change of the concentration of reagent A in the vessel and the temperature change in the reactor per time unit can be modelled as:

$$\frac{dCA}{dt} = \frac{F}{V} * (CAf(t) - CA(t)) - r(t) \tag{1}$$

$$r(t) = k * \exp\left(\frac{-E}{(R*T(t))}\right) * CA(t) \tag{2}$$

$$\frac{dT(t)}{dt} = \frac{F}{V}(Tf(t) - T(t)) - \left(\frac{H}{cp*rho}\right) * r(t) - \frac{(U*A)}{(cp*rho*V)} * (T(t) - T_j(t)) \tag{3}$$

General Formulation
Representation of Failures

The class of dynamical systems under study is described by

$$\dot{x}(t) = \xi(x(t), u(t)) + \beta(t-T) f(x(t), u(t)) \tag{4}$$

Where(x) is the state vector,(u) is the input vector, (T>0)is the beginning time of the failure, and β is a square function representing the time profiles of failures. We consider incipient faults that are modelled by

$$\beta(t-T) = diag(\beta_1(t-T), \beta_2(t-T), ..., \beta_n(t-T))$$

Where

$$\beta_i(\tau) = \begin{cases} 0, & \text{if } \tau < 0, \\ 1 - e^{-\rho i \tau}, & \text{if } \tau \geq 0, \end{cases} \quad i = 1, 2, ..., n \tag{5}$$

The "healthy" system in the absence of any faults is described by

$$\dot{x}h(t) = \xi(xh(t), u(t)) := \xi^*(xh(t), u(t)) + \tilde{\xi}(xh(t), u(t)) \tag{6}$$

Where ξ^* represents the nominal dynamics (known) and $\tilde{\xi}$ characterizes any discrepancy between the actual plant and nominal model that may occur due to modelling errors. In this work, we first consider the ideal case where $\tilde{\xi} \equiv 0$ and then the case where $|\tilde{\xi}(x,u)| \leq \xi_0$ for all $(x,u) \in (x*u)$, where $\xi_0 \geq 0$ is a known constant. In general, the design and analysis of robust fault diagnosis architectures based on nonlinear modelling techniques requires further investigation.

Nonlinear Estimator

The failure representation described by (4) provides a framework for characterizing a wide class of faults. In general, the magnitude of faults in practical applications depends on the state of the system as well as the system input. An adaptive nonlinear estimator is given by:

$$\hat{x} = W(s)[z] \tag{7}$$

$$z = \xi(x, u; \hat{\theta}) \tag{8}$$

$$\dot{\hat{\theta}} = \eta(x, u, \hat{x}; \hat{\theta}) \tag{9}$$

Where $W(s)$ is a stable filter matrix, (7) and (8) represents an observer-based nonlinear estimation scheme, and (9) is the adaptive law of the adjustable parameters.

Now we consider the construction of a nonlinear estimator for modelling deviations in system dynamics due to failure. Based on the system representation described by (l), we choose an estimated model of the form

$$\dot{\hat{x}} = A\hat{x}(t) + \xi^*(x(t), u(t)) + \hat{f}(x(t), u(t); \hat{\theta}(t)) - Ax(t) \tag{10}$$

Where $\hat{x} \in \Re^n$ is the *estimated state vector,* $\hat{f}$ represents an on-line approximation (*OLA*) model $\hat{\theta} \in \Re^q$ is a vector of *adjustable parameters* or *weights,* and A is a stable constant square matrix of dimension $n*n$, whose eigenvalues lie in the left-half complex . The initial value of the estimated parameter vector for the estimated model (10) is $\hat{\theta}(0) = \hat{\theta}^0$ chosen such that $\hat{f}(x,u,\hat{\theta})=0$ for all (x,u), corresponding to the case of no failure (healthy condition), while the initial value of the estimated state vector is selected as $\hat{x}(0)=x(0)$. Starting from

these initial conditions, the main objective is to adjust (using input/output information) the parameter estimate $\hat{\theta}(t)$ at each time t so that $\hat{f}(x,u,\hat{\theta})$ approximates the unknown function $\beta(t-T)f(x,u)$ as closely as possible. Once this is achieved then the output of the on-line approximator $\hat{f}$ can be used to detect and diagnose as well as to accommodate any system failures. The on-line approximator, denoted by $\hat{f}$ represents the adjustable component of the estimation model. A block diagram representation of the estimated model (10) is depicted in Fig. 2.
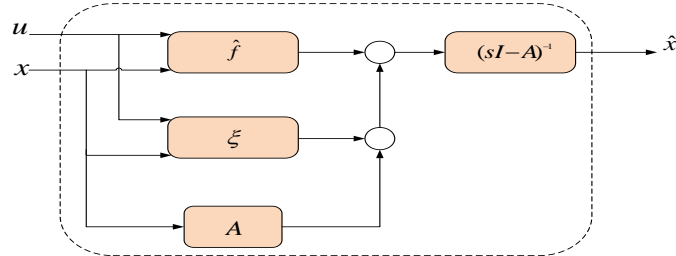
Figure 2 Block diagram representation of the estimation model.

On-Line Approximators

In this work a class of radial basis function (RBF) neural network is used as an online approximator for detection. The output of RBF networks is of the form

$$\hat{f}_n(z;\hat{\theta}):=\left\{\sum_{i=1}^{n}\hat{\theta}_i w_i(z):\hat{\theta}_i \in \Re\right\} \tag{11}$$

Learning Schemes

We use the Lyapunov synthesis approach to derive the adaptive law for updating the parameter estimates. In our case, the Lyapunov synthesis approach yields the following adaptive law for updating the parameter estimates:

$$\dot{\hat{\theta}}=\mathrm{P}\{\Gamma Ze\} \tag{12}$$

Where

$$Z:=\left[\frac{\partial \hat{f}(x,u;\hat{\theta})}{\partial \hat{\theta}}\right]^T , \quad \Gamma = \Gamma^T > 0 \tag{13}$$

Where e is the estimation error, Γ is the positive definite matrix $\Gamma \in \Re^{q*q}$ is known as the *learning rate matrix*, while $z \in \Re^{q*n}$ is the *sensitivity function* between the output of the network approximator and the adjustable parameters or weights. Thus, the overall adaptive law for updating the parameter estimates of the on-line approximator, using the Lyapunov synthesis approach and the projection modify cation, is given by:

$$\dot{\hat{\theta}} = \Gamma Ze - X^*\Gamma \frac{\hat{\theta}\hat{\theta}^T}{|\hat{\theta}|^2}\Gamma Ze$$

Where X^* denotes the *indicator* function given by

$$X^* = \begin{cases} 0, & if\,(|\hat{\theta}|<M_{\hat{\theta}}) \quad or \quad (|\hat{\theta}|=\overline{M}_{\hat{\theta}} \ and \ \hat{\theta}^T\Gamma Ze\leq 0) \\ 1, & if\,(|\hat{\theta}|<\overline{M}_{\hat{\theta}} \quad and \quad \hat{\theta}^T\Gamma Ze>0) \end{cases} \tag{14}$$

Simulation Example

Using the methodology described in previous sections, an estimated model is constructed. This estimated model is described by the following state-space representation as following:

$$\dot{x} = \begin{bmatrix} F/V*(u(1)-x(1))-k0*\exp(-E/R*x(2)))*x(1) \\ F/V(u(2)-x(2))-(H/HD)*k0*\exp(-E/R*x(2)))*x(1)-(HA/(HD*V))*(x(2)-u(3)) \end{bmatrix}$$
$$+ \begin{bmatrix} 0 \\ \hat{f}(x_2,\hat{\theta}) \end{bmatrix} - p \begin{bmatrix} \hat{x}_1-x_1 \\ \hat{x}_2-x_2 \end{bmatrix}$$

Where $[x_1,x_2]^T = [y_1,y_2]^T$ is the state vector of the system; $\hat{x}=[\hat{x}_1,\hat{x}_2]^T$ is the estimated state vector; $p > 0$ is the pole location of the filter; $\hat{f}$ is the on-line approximation model used to monitor the system; and $\hat{\theta}$ is a vector of adjustable parameters.

The classes of failures considered in this work are strictly related to the dispositioning of the hot and cool water valves of the cooling system. This leads to a temperature of the fluid entering the jacket different with respect to the commanded value. Hence, an actuator fault results in a faulty input temperature given by

$$u^*(t)=u(t)+f_a(t)$$

Where $u(t)$ is the commanded value. The time profile adopted for the fault function f_a is

$$f_a(t)=u_0(1-e^{-\mu(t-t_0)}), \quad t\geq t_0$$

Where u_0 is the maximum amplitude, t_0 is the fault occurrence time, and μ is the fault evolution rate. Parameter μ is used to simulate a desired time evolution: small value characterize slowly developing faults (incipient faults); large values are used to model step-like behaviours of the fault (abrupt faults).

The reactor temperature sensor fault is superimposed with 10% change of the measured reactor temperature and simulated from time 10 s to 15 s.

In this simulation example we use a class of neural networks, known as Radial Basis Function (RBF) networks, as the on-line approximator model. Specifically we use Gaussian RBF networks which are described by

$$\hat{f}(x_2,\hat{\theta})= \sum_{i=1}^{N} \hat{\theta}_i \exp\left(-\|x_2-c_i\|^2 \Big/ \sigma^2\right)$$

We choose a uniform width $\sigma = 0.6$ for the basis functions, and $N = 19$ fixed centres c_i, which are evenly distributed in the interval $[-9\ 9]$.

The standard adaptive law in this case is

$$\dot{\hat{\theta}} = P\{\Gamma Z e\}$$

$$\dot{\hat{\theta}}=\Gamma Z e - X^* \Gamma \frac{\hat{\theta}\hat{\theta}^T}{|\hat{\theta}|^2}\Gamma Z e$$

Where

$$Z := \left[\frac{\partial \hat{f}(x,u\,;\hat{\theta})}{\partial \hat{\theta}}\right]^T \quad , \quad \Gamma = \Gamma^T > 0$$

And x^* denotes the *indicator* function given by

$$X^* = \begin{cases} 0, & if\,(|\hat{\theta}| < M_{\hat{\theta}}) \quad or \quad (|\hat{\theta}| = \overline{M}_{\hat{\theta}} \ and \ \hat{\theta}^T \Gamma Z e \leq 0) \\ 1, & if\,(|\hat{\theta}| < \overline{M}_{\hat{\theta}} \quad and \quad \hat{\theta}^T \Gamma Z e > 0) \end{cases}$$

The OLA output norm and state error norm may be used to monitor the system for failure detection. In Fig.6 the evolution of the output state estimation error norm is represented for

the time interval $[0,60]s$. The rapid jump at $t=10$ provides a measure for detecting the system's failure. The evolution of the output estimation error norm is given by

$$e_N(t) := \left[(x1(t)-\hat{x}1(t))^2 + (x2(t)-\hat{x}2(t))^2\right]^{1/2}$$

$$\phi_N(t) = \left|\hat{f}(x1(t),\hat{\theta}(t))\right|$$
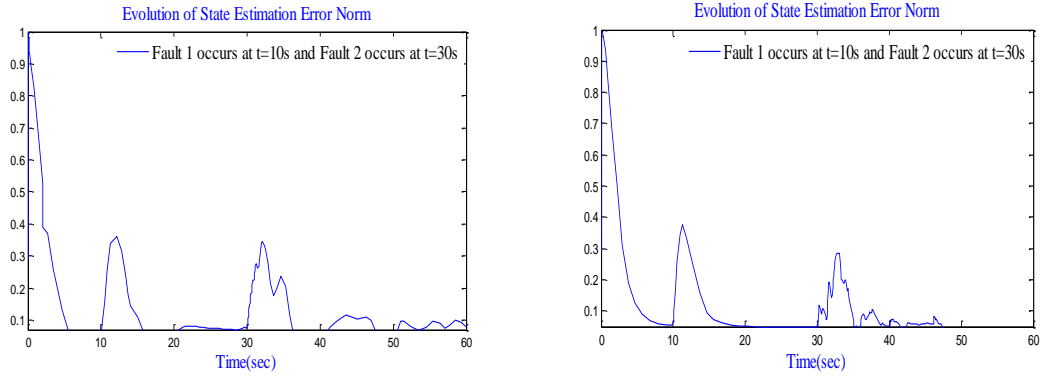
Fig.3. Evolution of the output estimation error norm $e_N(t)$ for a simulated example with sensor fault and an actuator incipient fault occurred at $t=10s$ and $t=30s$ respectively.
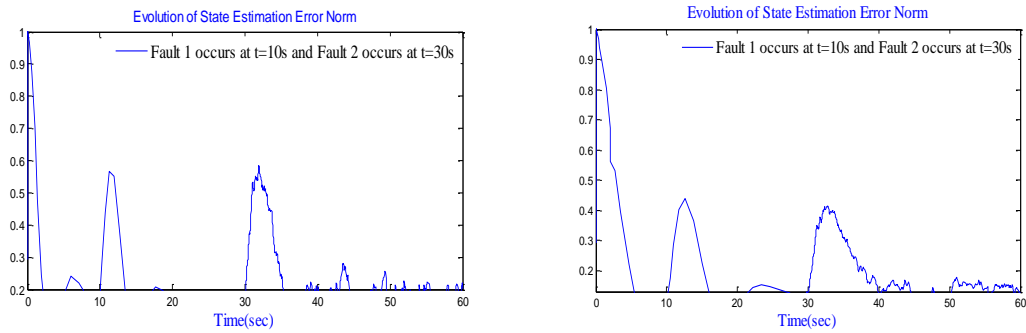
Fig.4. Evolution of the output estimation error norm $e_N(t)$ for a simulated example with sensor fault and an actuator abrupt fault occurred at $t=10s$ and $t=30s$ respectively.

In this simulation we use the adaptive law, where the learning arte is chosen as $\Gamma=\gamma I$ and $\gamma=10$. the projection operator P is used to constrain $\hat{\theta}$ within $M_{\hat{\theta}} = \left\{\hat{\theta} \in R^{19} : |\hat{\theta}| \leq 100\right\}$. Finally, the filter pole is set to $p=1$, and the initial parameter estimation vector is chosen as $\hat{\theta}(0)=0$, which corresponds to modelling a no-failure situation.

The online approximator (OLA) output norm is used as an indicator of a system failure. It can be seen from Fig 2-3 all faults have been clearly detected.

Conclusion

In this work we have presented a general learning methodology and some preliminary analytical results concerning the use of neural networks and other on-line approximation models for diagnosis of failures in dynamical systems. Nonlinear modelling techniques have been employed for monitoring the dynamical system and for estimating any changes that may occur due to a failure. The estimated model is used for detecting failures in nonlinear systems. The main advantage of using nonlinear estimation techniques in FDA is the ability to model a larger and more practically realistic class of failures. The methodology developed in this work is based on analytical redundancy techniques. One of the important issues avoided in this work is the effect of modelling uncertainty on the FDA learning scheme. In particular, we have assumed that the nominal model provides an accurate description of the physical system in the absence of any failures. In the presence of modelling uncertainty, the FDA learning scheme may perceive this uncertainty as a change in the system dynamics,

thus confusing the effect of faults and possibly leading to false alarms. Another assumption that we have made is the availability for measurement of all the states. Relaxation of these assumptions, and investigation of the effect of modelling uncertainty on the performance of the learning scheme, are topics for future work.

References

1. Caccavale, F., et al. (2009). "An integrated approach to fault diagnosis for a class of chemical batch processes." Journal of Process Control **19**(5): 827-841.

2. Demetriou, M. A. and M. M. Polycarpou (1998). "Incipient fault diagnosis of dynamical systems using online approximators." Automatic Control, IEEE Transactions on **43**(11): 1612-1617.

3. Frank, P. M. and X. Ding (1997). "Survey of robust residual generation and evaluation methods in observer-based fault detection systems." Journal of Process Control **7**(6): 403-424.

4. Keliris, C., et al. (2015). "A robust nonlinear observer-based approach for distributed fault detection of input–output interconnected systems." Automatica **53**: 408-415.

5. Polycarpou, M. M. and A. J. Helmicki (1995). "Automated fault detection and accommodation: a learning systems approach." Systems, Man and Cybernetics, IEEE Transactions on **25**(11): 1447-1458.

6. Polycarpou, M. M. and A. T. Vemuri (1995). "Learning methodology for failure detection and accommodation." Control Systems, IEEE **15**(3): 16-24.

7. Trunov, A. B. and M. M. Polycarpou (2000). "Automated fault diagnosis in nonlinear multivariable systems using a learning methodology." Neural Networks, IEEE Transactions on **11**(1): 91-101.

8. Vemuri, A. T. and M. M. Polycarpou (1997). Input-output systems robust nonlinear fault diagnosis. American Control Conference, 1997. Proceedings of the 1997.

9. Venkatasubramanian, V., et al. (2003). "A review of process fault detection and diagnosis: Part I: Quantitative model-based methods." Computers & Chemical Engineering **27**(3): 293-311.

10.   Xiaodong, Z., et al. (2002). "A robust detection and isolation scheme for abrupt and incipient faults in nonlinear systems." Automatic Control, IEEE Transactions on **47**(4): 576-593.

Engku Ahmad Rafiqi Engku Ariff

A comparison between two- and three-level six-phase drive

E A R Engku Ariff
School of Engineering, Technology and Maritime Operations, James Parsons Building, Byrom Street, L3 3AF, Liverpool
E.A.EngkuAriff@2013.ljmu.ac.uk

Abstract. The introduction of multilevel inverters has been slowly affecting the interest of researchers and industry towards the conventional two-level inverter in recent years. This is because, two-level converter demonstrates significantly lower performance than multilevel inverters and it is unable to meet the demands for higher power applications. On the other hand, the capability of the multilevel inverters to sustain higher dc-link voltage while still using the same ratings of power semiconductors as in the conventional two-level inverter offers the ability to the multilevel inverters to operate in higher power applications. However, as the number of levels increases, the number of required power semiconductors also increases, so does the complexity in controlling the inverter. In this paper, a comparison between the two-level and multilevel i.e. three-level inverter driving a symmetrical six-phase machine, in terms of power circuit topology, modulation approach and performance, is discussed. Both conventional and multilevel drives are accurately modelled using Vector Space Decomposition (VSD) method and controlled using the developed space vector modulation technique. The harmonics which contribute to the machine losses are identified and controlled by setting them to zero. Furthermore, the applied switching sequences are chosen based on several requirements and conditions such that, the produced output phase voltages are sinusoidal, on average, which results in lower THD, hence better performance. Both drives are developed, controlled and simulated in Matlab/Simulink environment. Obtained results i.e. comparison between the calculated voltage and current THD of both drives, prove that multilevel inverter does offer better performance than two-level inverter but requires more complex modulation approach.

Keywords. Space vector modulation, multilevel multiphase drives, symmetrical six-phase induction machine, vector space decomposition, total harmonic distortion

1. Introduction

The conventional two-level voltage source inverter is considered as a mature technology and has been dominating drive applications for decades. However, due to limitation of the rating of available power semiconductors, the conventional two-level inverter was not able to meet the increasing power demands in drive applications. Since the development of power semiconductors with much higher rating was still in infant stage, this limitation was overcome with the birth of multilevel i.e. three-level inverter topology in early 1980's known as neutral point clamped topology (NPC) [1]. This new inverter topology can offer much higher power while still using the same rated power semiconductors used in conventional two-level inverter [2]. However, the total number of power semiconductors will be higher and the controllability of the inverter becomes more complex. As the name suggests, multilevel inverter also offers higher number of inverter output voltage levels. As the number of inverter output voltage levels increases, the summation of the voltage levels will produce a staircase-shape waveform which approximates towards a sinusoidal waveform, thus, reducing the

voltage THD [3]. Therefore, multilevel inverter has gradually been replacing the domination of the conventional two-level inverter in many industrial applications [4, 5]. In this paper, a comparison between the conventional two- and three-level inverter driving a symmetrical six-phase machine with single neutral point is discussed, in terms of power circuit topology, modulation strategy (based on Space Vector PWM, SVPWM approach) and the performance of the inverter using voltage and current THDs as figure of merits. Both drives are modelled using VSD method and simulated in Matlab/Simulink environment.

2. Power circuit topology

The power circuit of conventional two- and three-level (based on NPC topology) six-phase inverter are shown in Figure 1 and Figure 2, respectively. The switching states, corresponding to the possible output leg voltage levels of both inverters are also shown.

Switching States e.g. leg A		Output leg voltage
S_A	S_A'	
off	on	0
on	off	V_{dc}
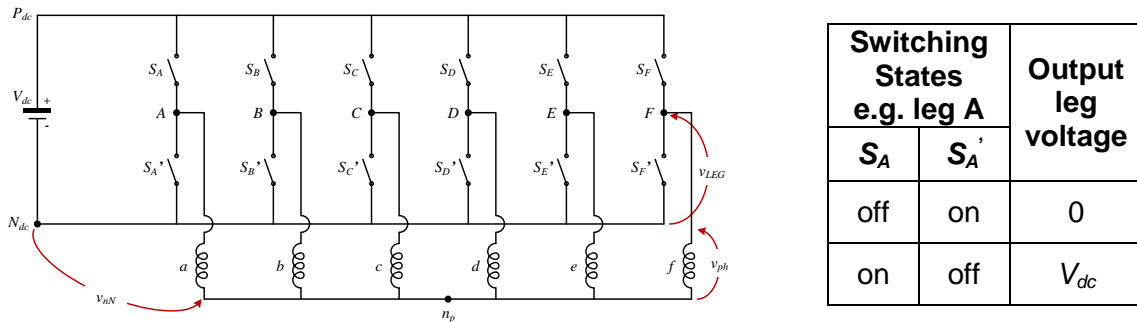
Figure 1: Power circuit of two-level symmetrical six-phase drive.

Switching States e.g. leg A				Output leg voltage
S_{A2}	S_{A1}	S_{A2}'	S_{A1}'	
off	off	on	on	0
off	on	on	off	$V_{dc}/2$
on	on	off	off	V_{dc}
on	off	off	on	High impedance
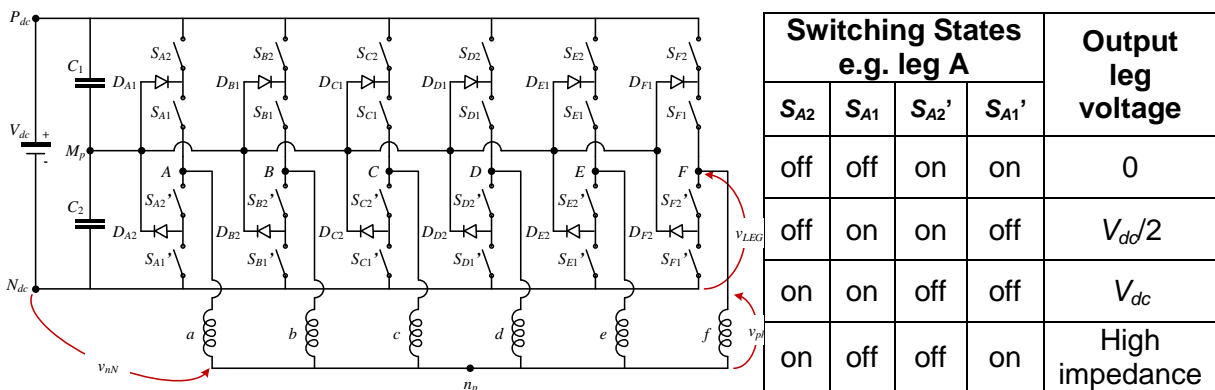
Figure 2: Power circuit of three-level symmetrical six-phase drive.

As can be seen in Figure 2, the dc-bus voltage is divided between two dc-bus capacitors for three-level case. If the power semiconductors i.e. switching devices used in three-level inverter have the same rating as in conventional two-level inverter, they only require to sustain half of the dc-bus voltage. Therefore, three-level inverter offers the capability to potentially operate at higher dc-bus voltages. However, the number of switching devices increases to four with the additional two clamping diodes per inverter leg.

3. SVPWM control scheme for two- and three-level six-phase inverter

In two-level six-phase inverter, there are 2^6 = 64 possible switching state configurations (from 000000 till 111111 in binary, normalised, representation). These possible switching states which defines 63 space vectors, can be projected into two mutually orthogonal two-dimensional planes, α-β and x-y, and onto two zero-component axes, 0^+ and 0^- as shown in Figure 3a , using VSD method. Note that the axes in Figure 3a are normalised to V_{dc}. The VSD transformation for symmetrical six-phase case is given with:

$$
\begin{bmatrix} v_\alpha \\ v_\beta \\ v_x \\ v_y \\ v_{0^+} \\ v_{0^-} \end{bmatrix} = \frac{2}{6} \cdot \begin{bmatrix} 1 & \cos(\alpha) & \cos(2\alpha) & \cos(3\alpha) & \cos(4\alpha) & \cos(5\alpha) \\ 0 & \sin(\alpha) & \sin(2\alpha) & \sin(3\alpha) & \sin(4\alpha) & \sin(5\alpha) \\ 1 & \cos(2\alpha) & \cos(4\alpha) & \cos(6\alpha) & \cos(8\alpha) & \cos(10\alpha) \\ 0 & \sin(2\alpha) & \sin(4\alpha) & \sin(6\alpha) & \sin(8\alpha) & \sin(10\alpha) \\ 1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & -1/2 & 1/2 & -1/2 & 1/2 & -1/2 \end{bmatrix} \cdot \begin{bmatrix} v_a \\ v_b \\ v_c \\ v_d \\ v_e \\ v_f \end{bmatrix}
\tag{1}
$$

It should be noted that, the VSD transformation in (1) is shown for the phase voltages but it is also applicable to leg voltages and phase currents.

The corresponding low order harmonics of the order of $6k \pm 1$ ($k = 0, 1, 2, 3 \ldots$), $6k \pm 2$ ($k = 0, 1, 2, 3 \ldots$) and $3k$ ($k = 1, 3, 5\ldots$) are mapped into α-β, x-y planes and onto 0^- axis, respectively. From the load perspective, only low order harmonics mapped onto α-β plane contribute to torque production while other low order harmonics mapped onto different plane and axis contribute to machine losses [6]. On the other hand, the projection of space vectors onto 0^+ axis represents common mode voltage (CMV) for the leg voltage space vectors, while for the phase voltage space vector projections, it is zero. Since the purpose of SVPWM is to obtain the desired phase voltages, the 0^+ axis is not considered in the space vector modulation methods.
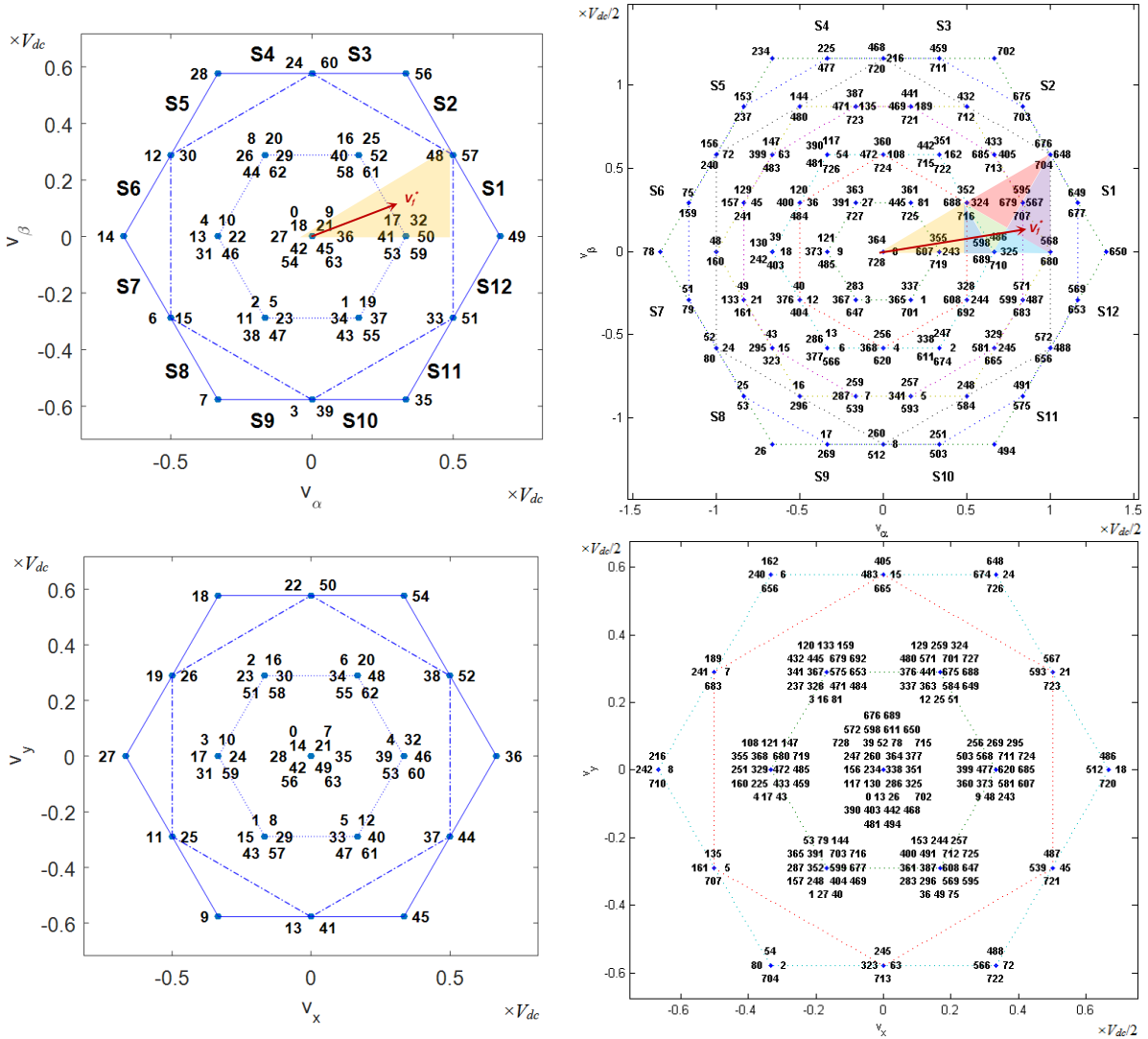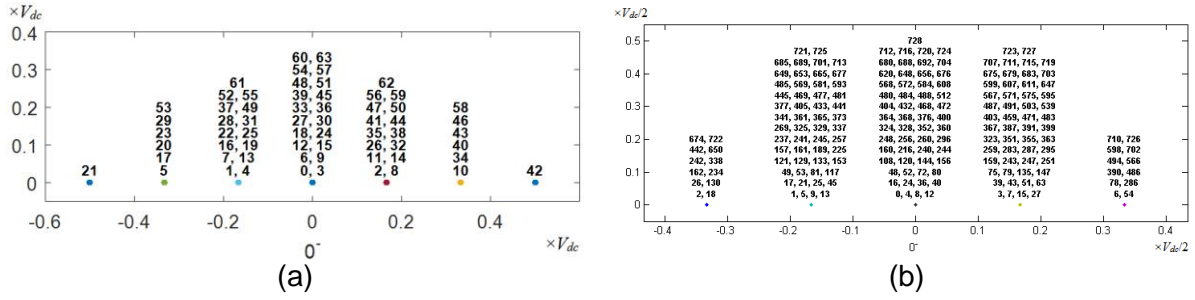
Figure 3: Projection of phase voltage space vectors (top to bottom) in *α-β* plane, *x-y* plane and 0⁻ axis for (a) two- and (b) three-level symmetrical six-phase drive.

In order to obtain sinusoidal output phase voltages, the potential switching sequences are chosen such that, they meet several requirements and conditions. The potential switching sequences should be zero on average for *x-y* and 0⁻ components, i.e. the low order harmonics corresponding to *x-y* plane and 0⁻ axis do not exist. Moreover, the number of the chosen space vectors ought to be the same as the number of phases of the machine, which is six [7]. In order to minimise losses and lessen d*v*/d*t*, the transition of output leg voltages also needs to be taken into consideration where the desirable level of transition is one, either increasing or decreasing in transition [8]. Also, it is desirable that the transition of the output leg voltages in one switching period is symmetrical i.e. in the first half of the switching period, the output leg voltages increase in transition level and in the other half, they decrease. Thus, the potential starting switching states should only consist values which are less than the maximum representation of normalised output voltage level (i.e. 0s in two-level case, 0s and 1s in three-level case) [9].

As a result, there is only one starting switching state i.e. 000000 in two-level case, that meets the condition for all sectors. By meeting the rest of the stated requirements, one finds that, there is only one switching sequence per sector. For example, the switching sequence for the first sector, S1 is [0 32 48 49 57 59 63] or 000000-100000-110000-110001-111001-111011-111111 in binary (normalised) representation. However, the switching sequences are valid only in their respective region of application (e.g. in S1, the region of application is denoted by yellow triangle as shown in Figure 3a top) which limits the maximum achievable modulation index, m_i to 1 (m_i is defined as ratio of fundamental voltage over $V_{dc}/2$). Then, based on the location of desired fundamental phase voltage, V_f^* (refer to Figure 3a top), the corresponding switching sequence is identified and the dwell times are calculated using time matrix approach as in:

$$\begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \\ T_6 \end{bmatrix} = \begin{bmatrix} v_{\alpha,1} & v_{\alpha,2} & v_{\alpha,3} & v_{\alpha,4} & v_{\alpha,5} & v_{\alpha,6} \\ v_{\beta,1} & v_{\beta,2} & v_{\beta,3} & v_{\beta,4} & v_{\beta,5} & v_{\beta,6} \\ v_{x,1} & v_{x,2} & v_{x,3} & v_{x,4} & v_{x,5} & v_{x,6} \\ v_{y,1} & v_{y,2} & v_{y,3} & v_{y,4} & v_{y,5} & v_{y,6} \\ 1 & 1 & 1 & 1 & 1 & 1 \\ v_{0^-,1} & v_{0^-,2} & v_{0^-,3} & v_{0^-,4} & v_{0^-,5} & v_{0^-,6} \end{bmatrix}^{-1} \begin{bmatrix} v_\alpha^* \\ v_\beta^* \\ v_x^* \\ v_y^* \\ 1 \\ v_{0^-}^* \end{bmatrix} \cdot T_s \qquad (2)$$

As the desired output phase voltage is a sinusoidal waveform without low order harmonics, the reference voltage for *x*, *y* and 0⁻ components (v_x^*, v_y^* and $v_{0^-}^*$) in (2) should be set to zero.

For three-level six-phase inverter, there are three possible output leg voltage levels (0, $V_{dc}/2$ and V_{dc}), as shown in Figure 2, which results to $3^6 = 729$ possible switching state configurations. Note that, high impedance state for output leg voltage is not considered in the space vector modulation method. The space vectors can also be projected into *α-β*, *x-y* planes and onto 0⁺, 0⁻ axes using VSD method (1). However, not all space vectors are involved in switching sequence selection process. In order to minimise number of switching states that should be considered, order-per-sector law as in [10] is implemented first. The number of switching states is reduced in this way to 189. The projection of the remaining

switching states into α-β, x-y planes and onto 0⁻ axis is shown in Figure 3b. Note that the axes in the Figure 3b are normalised to $V_{dc}/2$.

By following the requirements and conditions for switching sequence selection, one finds that, there are seven out of 28 switching states that can be chosen as potential starting switching states in each sector. From these seven starting switching states, 64 potential switching sequences are identified per sector. The possible transitions of space vectors i.e. in α-β, x-y planes and 0⁻ axis for these 64 switching sequences are determined. Hence, the identified switching sequences are categorised based on the transition patterns which result to 32 unique transition patterns. The transition patterns which do not provide voltage cancellation in x-y plane and 0⁻ axis are identified and discarded, thus significantly reducing the number of transition patterns to six per sector. The remaining six transition patterns also correspond to six regions of application i.e. sub-sectors (denoted by different colour of triangles as shown in Figure 3b top) [11]. Similarly, the maximum achievable m_i is 1. For the selected six transition patterns, there are still twenty potential switching sequences left associated to them, thus causing redundancy in switching sequences. The final switching sequences are selected further in terms of capability of balancing capacitors voltage and minimising switching transitions between sub-sectors [7, 10], which narrows them to six switching sequences per sector. In summary, there are six switching sequences which correspond to six sub-sectors per sector.

Since the application of the sub-sectors is reliant on the location of desired fundamental phase voltage, V_f^* (refer to Figure 3b top), a method on which sub-sectors are identified is applied [11]. This identification method uses the borders of the sub-sectors as limits in determining the location of V_f^* in the α-β plane at any given time. Finally, the dwell times of applied switching sequence are calculated using (2).

4. Simulation of two- and three-level symmetrical six-phase drives

The SVPWM control scheme for two- and three-level symmetrical six-phase drives are developed and simulated using MATLAB/Simulink. The dc-bus voltage, V_{dc} is set to 200 V. The frequency of sinusoidal symmetrical six-phase reference voltages, V_{ref} is set to $f_{ref} = 50$ Hz, and the magnitude is controlled by m_i, which varies from 0.1 to 1 with 0.05 increments. The switching frequency, f_s of the inverter is set to 2 kHz. The output legs of the inverter are connected to a model of symmetrical six-phase machine with single isolated neutral point. The machine parameters are: stator and rotor resistance, $R_s = R_r = 3$ Ω, stator and rotor leakage inductance, $L_{ls} = L_{lr} = 0.005$ H, mutual inductance, $L_m = 0.185$ H, number of pole pairs, $P = 2$, and rotor inertia, $J_r = 0.03$ kgm². The machine is operating at no-load. The simulation is run until the speed of the machine has reached steady state and THDs are calculated for duration of one period for phase 'a' voltage and current as:

$$\text{THD}_v = \sqrt{\sum_{k=2}^{h} v_k^2 / v_1^2} \qquad \text{THD}_i = \sqrt{\sum_{k=2}^{h} i_k^2 / i_1^2} \qquad (3)$$

where v_k i.e. i_k represent k-th voltage i.e. current harmonic in spectrum, and h is 420 (hence, first ten sidebands are considered for THD calculation).

Obtained simulation results of phase 'a' voltage and current THDs for full linear range of m_i are shown in Figure 4. As one might expect, the values of voltage THD of three-level inverter are lower compared to two-level inverter, as shown in Figure 4a. This is because the number of output phase voltage levels of three-level inverter is higher than for two-level inverter, which leads that its output phase voltage waveforms approximate more towards sinusoidal waveforms (refer to the top waveforms in Figure 5), thus reducing the THD. This result is also in agreement with the results of analytical expression in [12].

The values of current THD of three-level inverter are higher than for two-level inverter at lower values of m_i; while for higher m_i, it is the other way round, as shown in Figure 4b. This behaviour is expected, since in three-level inverter, the pivot states are shared between nearest pivot vectors which offers reduced current harmonic distortion at higher m_i; however, at lower m_i i.e. in the region equivalent to two-level's region of operation, the sharing of

nearest pivot vectors are not the best modulation option. On the contrary, modulation which uses zero vector as pivot vector, as in the modulation for two-level inverter offers lower current ripple at lower m_i, thus lower current THD [13].

(a)   (b)
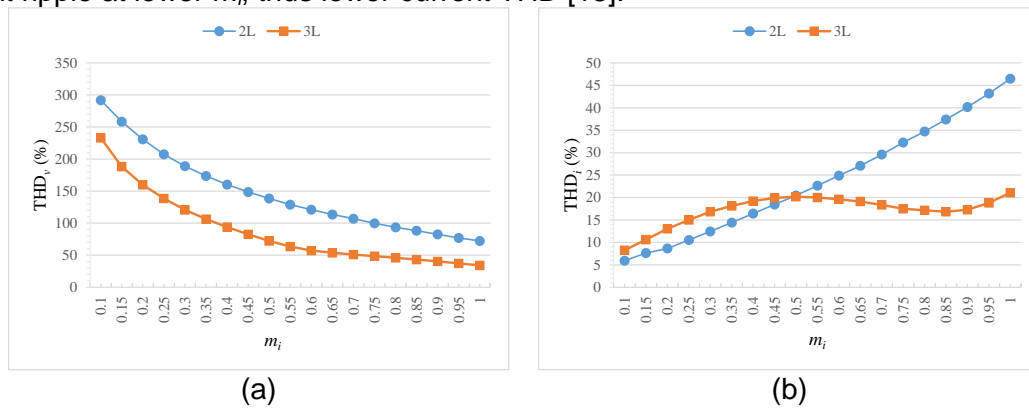
Figure 4: Comparison two- and three-level inverter phase 'a': (a) voltage THD and (b) current THD for m_i ranging from 0.1 to 1.

Simulation results when $m_i = 1$ are shown in Figure 5. One can see that the voltage step of the phase voltage and current ripple for three-level inverter is lower compared to two-level inverter. These contribute to lower magnitudes of harmonics in voltage and current spectra (refer to Figure 5). In addition, the validity of the developed SVPWM for two- and three-level symmetrical six-phase inverter can also be observed from the spectra, where the low order harmonics do not exist while the fundamental satisfies the required value (for $m_i = 1$, $V_1 = 100$ V).

(a)   (b)
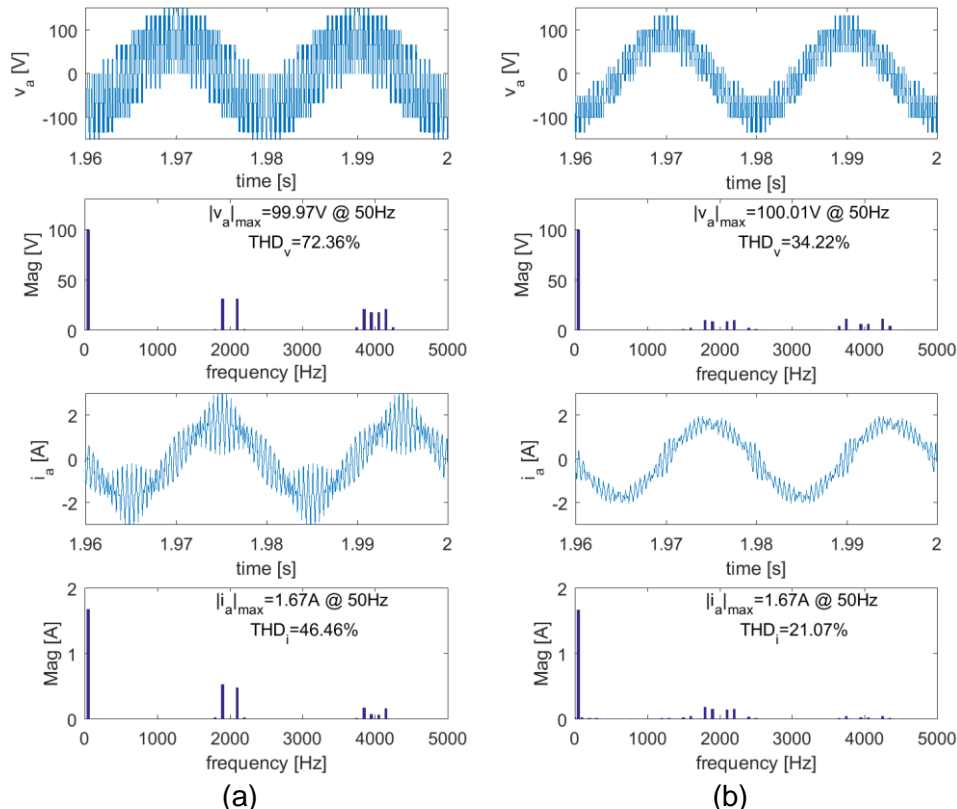
Figure 5: Phase 'a' voltage and current waveforms with corresponding spectra.

5. Conclusion

In this paper, a comparison between two- and three-level symmetrical six-phase drive is presented. The power circuit topologies of both drives are compared by complexity of number of components. In addition, the space vector modulation techniques for both drives, are explained in details. It is shown that the offline calculation and analysis in development

of modulation algorithm for three-level inverter is more complex compares to two-level inverter. Obtained simulation results show the characteristic of voltage and current THDs. Shown results indicate that three-level inverter has better performance than conventional two-level inverter.

References
[1] Nabae A, Takahashi I, Akagi H 1981 *IEEE Trans. on Ind. Applications* **17** 518-523
[2] Franquelo L G, Rodriguez J, Leon J I, Kouro S, Portillo R, Prats M A M 2008 *IEEE Ind. Elec. Magazine* **2** 28-39
[3] Lai J S, Peng F Z 1996 *IEEE Trans. on Ind. Applications* **32** 509-517
[4] Rodriguez J, Bernet S, Steimer P K, Lizama I E 2010 *IEEE Trans. on Ind. Electronics* **57** 2219-2230
[5] Rodriguez J, Lai J S, Fang Zheng P 2002 *IEEE Trans. on Ind. Electronics* **49** 724-738
[6] Dujic D, Iqbal A, Levi E 2007 A space vector PWM technique for symmetrical six-phase voltage source inverters *12th European Conf. on Power Electronics And Adj. Speed Drives* (*Aalborg, Denmark*) p 24-32
[7] Kelly J W, Strangas E G, Miller J M 2003 *IEEE Trans. on Energy Conversion* **18** 259-264
[8] van der Broeck H W, Skudelny H C, Stanke G V 1988 *IEEE Trans. on Ind. Applications* **24** 142-150
[9] Dordevic O, Levi E, Jones M 2013 *IEEE Trans. on Power Electronics* **28** 637-649
[10] Gao L, Fletcher J E 2010 *IEEE Trans. on Ind. Electronics* **57** 2332-2343
[11] Engku Ariff E A R, Dordevic O, Jones M 2016 Space vector PWM technique for a three-level six-phase drive *8th IET Int. Conf. on Power Electronics, Machines and Drives* (*Glasgow, Scotland*) p 1-6
[12] Dordevic O, Jones M, Levi E 2015 *IEEE Trans. on Power Electronics* **30** 1645-1656
[13] Grandi G, Loncarski J, Dordevic O 2015 *IEEE Trans. on Ind. Electronics* **62** 2721-2730

Ivan Zoric

Analysis of the unbalance in phase voltage harmonics of the asymmetrical multiphase machines with single neutral point

Ivan Zoric

Faculty of Engineering and Technology, James Parsons Building, Byrom Street, L3 3AF, Liverpool

i.zoric@2014.ljmu.ac.uk

Abstract. In safety critical applications such as more electric aircraft, or where the maintenance is not easily accessible i.e. offshore wind farms, the capability to operate under faulted condition is of extreme importance. With their inherited fault-tolerant operation, multiphase machines are the current state-of-the art in high power and high reliability drives. It has been shown that machines with single neutral point offer better fault tolerance capability than the ones with multiple neutral points. When a high power machine is inverter supplied, low order harmonics are present due to the low switching frequency and nonlinear inverter properties, such as dead time or voltage drop at the inverter switches. Moreover, low order harmonics can be deliberately injected to increase dc bus voltage utilisation. This paper shows that when asymmetrical multiphase machines with a single neutral point are used, the presence of balanced low order harmonics in leg voltages will produce unbalanced phase voltage harmonics and consequently unbalanced phase current harmonics. This can lead to the unbalanced power losses per phase and uneven strain on the inverter switches. By analysing the neutral point voltage harmonics, unbalance in the phase voltage harmonics is explained. Analytical expressions for neutral point voltage harmonics and phase voltage harmonics are provided for all possible asymmetrical machine configurations with a single neutral point. The proposed theory is verified using simulations in the case of the asymmetrical twelve-phase induction machine with a single neutral point.

Keywords. Multiphase machines, phase voltage harmonics, unbalanced operation, low-order harmonics.

1. Introduction

With variable speed drives reaching multi-MW power rating, multiphase machines offer numerous advantages over three-phase ones. Lower current and power per phase allows use of reduced rating switches, torque ripple is at the higher frequencies, and fault-tolerant operation is highly desirable in safety critical application [1, 2]. Some examples where advantages of the multiphase machines are of particular importance include fault-tolerant wind power generation systems [3], increasing the dc link voltage by series connection of the voltage source inverters (VSI) [4], multi-motor applications [5], and more reliable ship propulsion [6] and generation [7].

Fault tolerance being by far one of the largest benefits of the multiphase machines, has been researched to a great extent. In many cases used machine was with multiple three-phase windings. In these cases, fault tolerance strategy is very simple and can be achieved by simple switching off the winding set with the faulted phase [8]. On the other hand, research shows that switching off an entire winding set is not an optimal solution and that configurations with the single neutral point provide superior performance in fault-tolerant operation. By utilising flux/torque non-producing planes or by changing the transformation matrix, very small drop in available power under open-faulted phase can be achieved while optimising operation for torque ripple or stator resistive losses [9].

In addition to the advantages of the single neutral point in terms of better fault tolerance, additional benefits of torque ripple being at the higher frequencies and easier manufacturing of stator windings can be utilised if asymmetrical or split-phase machines are used [10].

The problems related to the use of the asymmetrical configurations with single neutral point are a non-standard vector space decomposition (VSD) transformation matrices and unbalance in phase voltage harmonics, and consequently phase current harmonics. Extensive research has been carried out on modelling and finding appropriate decoupling matrices for split-phase machines with both multiple and single neutral points [11-14]. Analysis of the torque ripple harmonic has been also carried out [10, 15]. Nevertheless, the presence of low order harmonics in this particular type of multiphase machines has not been explored to the full extent.

In this paper it is shown that balanced low-order leg voltage harmonics, usually present in high power low switching frequency multiphase drives, can produce unbalanced phase voltage harmonics in the case of the asymmetrical machines with a single neutral point. Unequal phase voltage harmonics may lead to unequal thermal losses among phases and uneven strain on switching devices, which should be taken into consideration during the drive design stage. Starting from the analysis of the neutral point voltage harmonics, analytical expressions for phase voltage harmonics are provided and verified by simulation. Furthermore, the proposed theory covers all cases of the asymmetrical machines with a single neutral point.

2. Asymmetrical machine and leg voltage harmonics
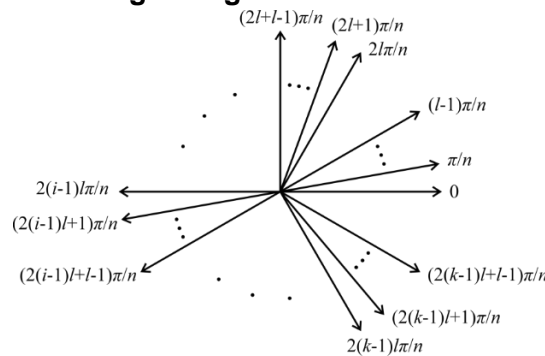
Figure 1 - Phase propagation angles of the asymmetrical multiphase machine.

Firstly, a definition of asymmetrical machine and phase propagation angles in the case of arbitrary number of phases is in order. Namely, stator windings of the n-phase asymmetrical machines considered here consist of k-phase winding sets where k is a prime number. Number of winding sets is equal to l hence the total number of phases is $n=lk$. Phase propagation within each sub-winding set is equal to $2\pi/k$. In an asymmetrical machine the phase propagation between sub-winding sets is π/n. A schematic of the phase shifts for the asymmetrical n-phase machine consisted of l k-phase winding sets is shown in the figure 1. Where i represents the phase number within the winding sets, hence it varies from 1 to k. j is introduced to denote the winding set number, and it varies from 1 to l, then the phase shift angle of i-th phase in the j-th winding set can be defined as:

$$\theta_{j,i} = \frac{\pi}{n}(2l(i-1)+j-1), \quad j=1,2,...,l, \ i=1,2,...,k \quad (1)$$

It should be pointed out that for all future equations indices j and i are varied as in (1), if not explicitly specified otherwise. It is considered that the machine is inverter supplied and that the leg voltages may contain harmonics. All leg voltage harmonics of the given harmonic order are considered balanced; they have the same amplitude. Hence, by using (1), leg voltages for the harmonic order h have amplitude A_h and are defined as:

$$v_{j,i,h}^l(t) = A_h \cos\left(h\left(\omega t + \frac{\pi}{n}(2l(i-1)+j-1)\right)\right) \quad (2)$$

The fundamental is considered to be the 1st harmonic. Without loss of generalisation, harmonic order h is a positive integer. Firstly, the neutral point voltage is analysed, and

afterwards the expression for phase voltages is proposed. In what follows it is assumed that the machine phases are balanced, so the impedance of every phase is the same.

3. Neutral point voltage analysis

The single neutral point voltage is found by firstly establishing the neutral point voltage of each winding set. From equation (2), the neutral point voltage for the harmonic order h of the winding set j can be found as a sum of all leg voltages of that winding set and dividing the result with the number of phases in the winding set. After solving the equations, expression for neutral point voltage for the harmonic order h of the sub-winding set j is defined as:

$$v_{j,h}^{NP}(t) = \begin{cases} A_h \cos\left(h\left(\omega t + \dfrac{\pi}{n}(j-1) \right) \right), & h = h_n k \\ 0, & h \neq h_n k \end{cases}, \quad h_n = 1, 2, \ldots \quad (3)$$

In the previous equation $h_n = h/k$ represents the normalised harmonic order. This is introduced to simplify the equations that follow and provide a tool for derivation of the phase voltage harmonic in the general case.

Finding the value of the single neutral point voltage for the harmonic order h is achieved by summing the values of individual neutral points of each winding set (3) and dividing the results with the number of winding sets l. Since harmonics of the order $h \neq h_n k$ cannot exist in any neutral point of the winding sets, they also do not exist in the single neutral point. While keeping in mind that in this analysis the phase voltage is defined as a difference between leg and single neutral point voltage, it follows that leg and phase voltage harmonics are identical for the cases where $h \neq h_n k$. Consequently, there is no unbalance and these cases are omitted from future analysis. Neutral point voltage for the harmonic order h is defined as follows:

$$v_h^{NP}(t) = \frac{A_h}{l} \frac{\sin\left(h_n \dfrac{\pi}{2} \right)}{\sin\left(h_n \dfrac{\pi}{2l} \right)} \cos\left(k h_n \left(\omega t + (l-1)\dfrac{\pi}{2n} \right) \right) \quad (4)$$

Expression (4) shows that the amplitudes of the neutral point voltage normalised harmonics are not dependant on the number of phases in the winding set. This means that every normalised neutral point voltage harmonic $h_n = h/k$ has the same amplitude for any asymmetrical machine with a single neutral point that has the same number of winding sets (l). Visualisation of the neutral point voltage normalised harmonics, h_n, amplitudes in the case of the machines with 2, 3, 4, and 5 winding sets is shown in figure 2 (all values are per-unit).

For the sake of clarity, example can be made with the comparison of the nine-phase (k=3, l=3) and the 15-phase (k=5, l=3) machines, top right plot in figure 2. For example, cases of the 1st and 3rd normalised harmonics ($h_n = h/k$) are the 3rd and the 9th harmonics of the 9-phase machine, while in the 15-phase case they are the 5th and the 15th harmonics. Since both machines have the same number of winding sets, their normalised harmonics are the same. Therefore, 3rd harmonic of the 9-phase has the same amplitude as the 5th harmonic of the 15-phase. Same applies for the 9th and 15th harmonics of the 9-phase and 15-phase machines respectively.
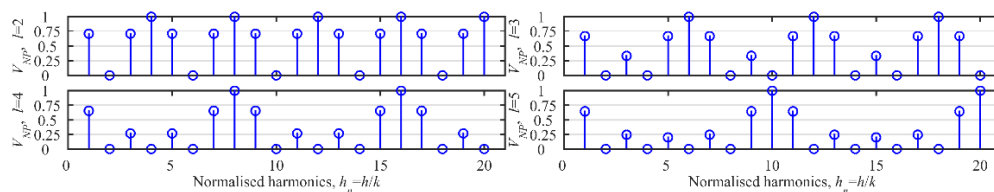
Figure 2 - Amplitudes of the neutral point voltage normalised harmonics for the machines with 2, 3, 4, and 5 winding sets.

Even order normalised harmonics are either zero or one, but since even harmonics cannot exist in symmetrical supply, they are not analysed here. It should be pointed out that

(4) will become limit equation when h_n becomes integer multiple of $2l$ (divide by zero). One way of solving this is to use L'Hôpital's rule, so instead of $\sin(\pm\pi), \cos(\pm\pi)$ is obtained.

4. Phase voltage analysis

Phase voltage is calculated as a difference between leg voltage (2) and neutral point voltage (4). After solving the equation and replacing h_n is with h/k, the expressions for amplitude and phase of the phase voltage harmonics are as follows:

$$A_{j,i,h}^{ph}=A_h\sqrt{1+\left(\frac{1}{l}\frac{\sin\left(h\frac{\pi}{2k}\right)}{\sin\left(h\frac{\pi}{2n}\right)}\right)^2-\frac{2}{l}\frac{\sin\left(h\frac{\pi}{2k}\right)}{\sin\left(h\frac{\pi}{2n}\right)}\cos\left(\frac{h}{n}j\,\pi+\frac{h}{k}i\,2\pi-\frac{h(5l+1)}{2n}\pi\right)} \quad (5)$$

$$\theta_{j,i,h}^{ph}=\text{atan2}\left\{\frac{\dfrac{\sin\left(h\frac{\pi}{2k}\right)}{\sin\left(h\frac{\pi}{2n}\right)}\sin\left(h\frac{\pi}{2n}(l-1)\right)-l\sin\left(h\frac{\pi}{n}(2l(i-1)+j-1)\right)}{\dfrac{\sin\left(h\frac{\pi}{2k}\right)}{\sin\left(h\frac{\pi}{2n}\right)}\cos\left(h\frac{\pi}{2n}(l-1)\right)-l\cos\left(h\frac{\pi}{n}(2l(i-1)+j-1)\right)}\right\}+\pi \quad (6)$$

Where, again, j and i designate the winding set and the phase within that set, respectively. When the expression for phase voltage amplitude (5) is analysed it can be seen that all phases within one winding set have the same amplitude of the voltage harmonics, i is multiplied by 2π which is the period of cosine function. On the other hand, voltage harmonic amplitudes of the phases in different winding sets are different.

For the sake of clarity, this has been demonstrated in the example of the nine-phase machine. Per-unit amplitudes of phase voltages 3rd harmonic, normalised with the amplitude of the leg voltage 3rd harmonic, are shown in the figure 3. Amplitudes of the 3rd harmonic of the phase voltages within each winding set are the same. On the other hand, amplitudes of the 3rd harmonic of the phase voltages of the second winding set (2nd, 5th, and 8th phase) are different than the ones of the first winding set (1st, 4th, and 7th phase).
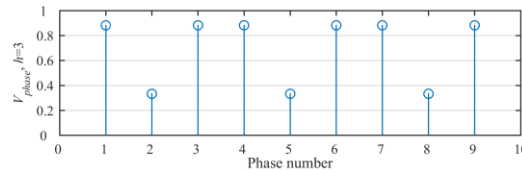
Figure 3 - 3rd harmonic phase voltage amplitudes of the asymmetrical nine-phase machine with a single neutral point.

From the equation (4) can be concluded that when the phase angle of the neutral point voltage is divided by the harmonic order, it is always equal to half of the phase shift angle of the first phase in the last winding set, $(l-1)\pi/2$. Hence, the neutral point voltage vector is always positioned in the middle of the first group of the leg voltage vectors, as shown in figure 4. In this figure phase angles of the leg voltage vectors are divided by the harmonic order, thus this analysis holds true for all harmonic orders.
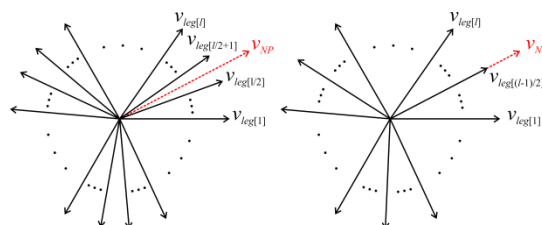
Figure 4 - Influence of the neutral point voltage vector on the phase voltages in the case of the even (left) and odd (right) number of winding sets.

The phase voltages are calculated as a difference between leg voltages and neutral point voltage. It can be seen that leg voltage vectors are located an equal angular distance from the neutral point voltage vector. Therefore, the corresponding phase voltages are equally influenced by the neutral point voltage and so have the same amplitude. In the case of the even number of winding sets, figure 4 (left), each two pairs of the phase voltages have the

same amplitude. On the other hand, in the case of the odd number of winding sets, figure 4 (right), vector of the neutral point voltage coincides with one of the leg voltage vectors, $(l-1)/2$.

As stated before, phase voltage harmonics of the same order in one winding set have the same amplitude. Hence, the number of phases per winding set is irrelevant for the analysis. Consequently, only the first phases of all the winding sets will be analysed, while the 2nd, 3rd, or any other phase in the winding set has the same amplitude of the considered phase voltage harmonic.

Amplitudes of the first 4 odd normalised phase voltage harmonics in the case of the machines with 2, 3, 4, and 5 winding sets is given in the figure 5. In the ideal case with a symmetrical supply, even harmonics do not exist. If even harmonics are present they are balanced and their amplitude is equal to zero or to the amplitude of the corresponding leg voltage harmonics. Hence, they are not of interest in the analysis and are omitted from the figure 5.

Looking at figure 5, it can be seen that for the same number of winding sets l, phase voltage harmonic unbalance is symmetrical around the imaginary line $l/2$, red dotted line. If the line coincides with a phase number, l is odd; phase voltage harmonic amplitude of that phase has a unique value, while phase voltage harmonic amplitudes of the other phases are the same if they are at the same distance from $l/2$. Alternatively, if the line $l/2$ is between two phases, l is even, then phases equidistant from the line $l/2$ have the same value of the phase voltage harmonic amplitudes.
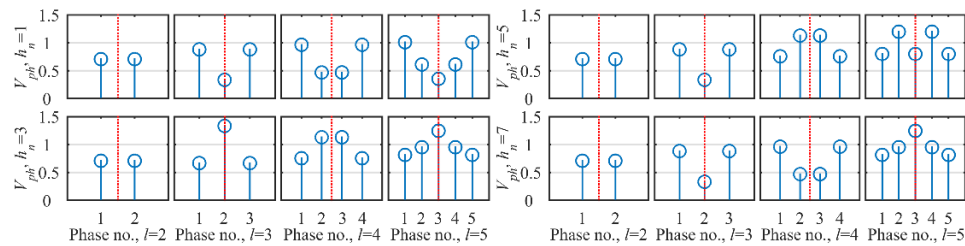
Figure 5 - Normalised amplitudes of the 1st, 3rd, 5th, 7th and 9th normalised phase voltage harmonics. 1st phases of all winding sets in the case of the machines with 2, 3, 4, and 5 winding sets are shown.

It should be pointed out that in the case when there are 2 winding sets, i.e. a 6-, 10-, 14-phase machine, unbalance does not exist due to the fact that the winding set phases are both affected in the same way by neutral point voltage as shown in figure 4 (left) and figure 5 (1st and 5th columns, l=2). However, the amplitudes of the phase voltage harmonics are different than the corresponding amplitude of the leg voltage harmonics.

5. Simulation results

To verify the analysis simulations have been carried out for a 12-phase asymmetrical induction machine with a single isolated neutral point. The machine is considered ideal and balanced and is supplied by 12-phase two-level inverter in 180° conduction mode. As a consequence, the supply generates all odd-order harmonics. Dc link voltage is set to 600V. The fundamental frequency is set to 50Hz. Leg and phase voltages with the corresponding FFTs are shown in figure 6. Only the first phase is shown in time domain (plots on the left), but all phases are included in the FFT analysis (plots on the right). Only the first 21 harmonics are shown due to the very small impact of higher order harmonics.
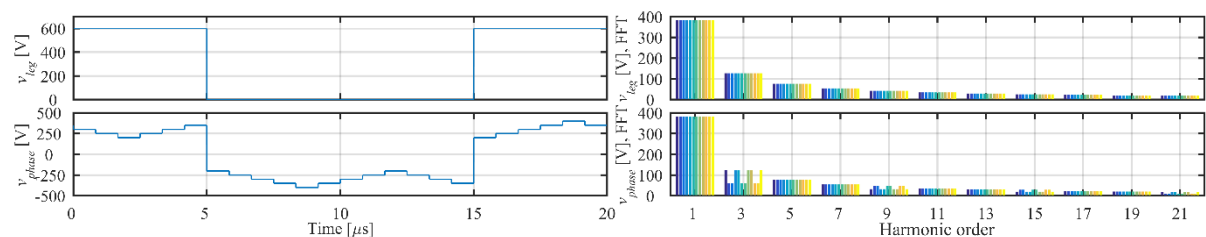
Figure 6 – Leg and phase voltages with corresponding FFTs.

For the sake of clarity, even harmonics are omitted from the plot, since they are equal to zero, along with the dc component. It can be seen that harmonics of the leg voltages are balanced, meaning that amplitudes of the leg voltages for the considered harmonic order have the same value. On the other hand, the unbalance exists in the case of some harmonics, i.e. 3rd, 9th, 15th, and 21st, or in terms of the normalised harmonics $h_n=h/k$, 1st, 3rd, 5th, and 7th.

The theory is verified by comparing the values of the phase voltage harmonics obtained by the simulation and equation (5). Leg voltage harmonic amplitudes obtained by FFT are used as input to the equation (5). The simulated and calculated amplitudes of phase voltage harmonics, for the harmonic orders where unbalance exists, are shown in the figure 7. It can be seen that both simulation and calculation of the unbalanced phase voltage harmonics produce the same results, confirming the proposed theory.
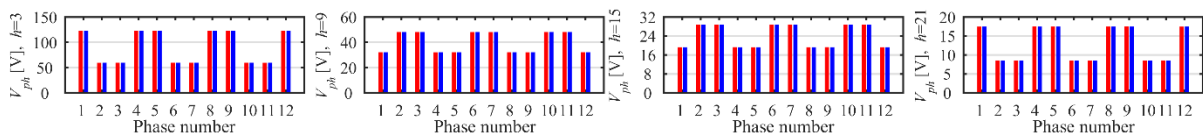
Figure 7 - Phase voltage harmonics amplitudes of the 3rd, 9th, 15th, and 21st harmonics (red – simulation, blue – calculation).

6. Conclusion

It has been found that in the case of asymmetrical multiphase machines with a single neutral point even if the leg voltage harmonics have the same amplitudes; there is an unbalance in the phase voltage harmonics, and consequently phase current harmonics. Based on the definition of phase propagation angles of an asymmetrical multiphase machine a general expression for neutral point and phase voltages has been found for arbitrary number of phases and arbitrary harmonic order. By analysing the neutral point voltage angle, a simple explanation has been presented for the unbalance that is seen in the phase voltage harmonics. Obtained equations are valid for any harmonic order in the case of any asymmetrical machine with a single neutral point.

Verification of the presented theory has been undertaken using simulations in Simulink environment for the case of the twelve-phase inverter supplied machine in 180º conduction mode. Obtained results are in are in good agreement with the proposed theory.

7. References

[1] E. Levi, R. Bojoi, F. Profumo, H. A. Toliyat, and S. Williamson, "Multiphase induction motor drives - a technology status review," *IET Electric Power Applications*, vol. 1, no. 4, pp. 489-516, 2007.

[2] E. Levi, "Multiphase electric machines for variable-speed applications," *IEEE Trans. on Industrial Electronics*, vol. 55, no. 5, pp. 1893-1909, 2008.

[3] M. J. Duran, I. G. Prieto, M. Bermudez, F. Barrero, H. Guzman, and M. R. Arahal, "Optimal fault-tolerant control of six-phase induction motor drives with parallel converters," *IEEE Trans. on Industrial Electronics*, vol, 63, no. 1, pp. 629-640, 2016.

[4] H. S. Che, W. P. Hew, N. A. Rahim, E. Levi, M. Jones, and M. J. Duran, "A six-phase wind energy induction generator system with series-connected DC-links," in *Proc. IEEE Int. Symposium on Power Electronics for Distributed Generation Systems PEDG*, Aalborg, Denmark, pp. 26-33, 2012.

[5] M. Jones, S. N. Vukosavic, and E. Levi, "Parallel-connected multiphase multidrive systems with single inverter supply," *IEEE Trans. on Industrial Electronics*, vol, 56, no. 6, pp. 2047-2057, 2009.

[6] C. Sun, S. Ai, L. Hu, and Y. Chen, "The development of a 20MW PWM drive for advanced fifteen-phase propulsion induction motors," *Journal of Power Electronics*, vol. 15, no. 1, pp. 146 – 159, 2015.

[7] G. Sulligoi, A. Tessarolo, V. Benucci, M. Baret, A. Rebora, and A. Taffone, "Modeling, simulation, and experimental validation of a generation system for medium-voltage DC integrated power systems," *IEEE Trans. on Industry Applications*, vol. 46, no. 4, pp. 1304-1310, 2010.

[8] L. Alberti, and N. Bianchi, "Experimental tests of dual three-phase induction motor under faulty operating condition," *IEEE Trans. on Industrial Electronics*, vol. 59, no. 5, pp. 2041-2048, 2012.

[9] A. Tani, M. Mengoni, L. Zarri, G. Serra, and D. Casadei, "Control of multiphase induction motors with an odd number of phases under open-circuit phase faults," *IEEE Trans. on Power Electronics*, vol. 27, no. 2, pp. 565-577, 2012.

[10] E. A. Klingshirn, "High phase order induction motors - Part I-description and theoretical considerations," *IEEE Trans. on Power Apparatus and Systems*, vol. PAS-102, no. 1, pp. 47-53, 1983.

[11] A. Tessarolo, "On the modeling of poly-phase electric machines through vector-space decomposition: theoretical considerations," in *Proc. IEEE Int. Conf. on Power Engineering, Energy and Electrical Drives POWERENG*, Lisbon, Portugal, pp. 519-523, 2009.

[12] R. H. Nelson, and P.C. Krause, "Induction machine analysis for arbitrary displacement between multiple winding sets," *IEEE Trans. on Power Apparatus and Systems*, vol. PAS-93, no. 3, pp. 841-848, 1974.

[13] Y. Zhao, and T.A. Lipo, "Space vector PWM control of dual three-phase induction machine using vector space decomposition," *IEEE Trans. on Industry Applications*, vol. 31, no. 5, pp. 1100-1109, 1995.

[14] A. Rockhill, and T.A. Lipo, "A simplified model of a nine phase synchronous machine using vector space decomposition," in *Proc. IEEE Power Electronics and Machines in Wind Applications PEMWA*, Lincoln, NE, USA, pp. 1-5, 2009.

[15] K. Wei, D. Wang, X. Zheng, and S. Cheng, "Research on relationship between harmonic currents and resultant harmonic magnetomotive forces in multiphase machines," in *Proc. IEEE Conference and Expo Transportation Electrification Asia-Pacific*, Beijinng, China, pp. 1-4, 2014.

Ahmed Fayas Najib

Selection of the Most Important Factor That Influences the Opening of the NSR Using an Analytic Hierarchy Process (AHP)

A F Ahmad Najib, D B Matellini, Jin Wang and S Bonsall
Liverpool Logistics, Offshore and Marine Research Institute, School of Engineering, Technology and Maritime Operations, Liverpool John Moores University, James Parson Building, Byrom Street, Liverpool L3 3AF, UK.
E-mail address: A.F.AhmadNajib@2013.ljmu.ac.uk

Abstract. The Northern Sea Route or NSR is a new shipping route in the Arctic. It can save 40% of the distance by comparing with the Suez Canal route. However, there are still a lot of uncertainties and issues in the NSR. This research is trying to analyse what is the criteria that influences the opening of the NSR by using an AHP technique. All criteria was gathered through literature review and extensive brainstorming with the experts. There are 4 steps involved in the AHP technique and at the end all the criteria is rank based on their weight. The sensitivity analysis is used to check the strength of the model.

1.0 Introduction

The NSR has been used as Russia's internal shipping transit since 1932 and in 1991; it was open for international use for the first time (Blunden, 2012). During that time, all of the countries did not show as much attention as it is today. Many believe it is because of the diminishing of ice in the Arctic due to the climate change that leads to the opening of the NSR. However, there are many other factors that also affect the use of the NSR. These 8 main factors are the political factor, legal factor, economic factor, environmental factor, social factor, technological factor, safety factor and the advantages of the NSR in comparison to other alternatives.

The NSR also has attracted lots of researchers all over the world with different fields and backgrounds to study this new shipping route. Some of the studies are, geopolitics and policies (Kaczynski, 2013; Blunden, 2012), the legal issues (Molenaar, 2009; Franckx, 2009), economic feasibility (Verny & Grigentin, 2009; Liu & Kronbak, 2010; Schoyen & Brathen, 2011), social (Meschtyb et al, 2005), technology (Kaczynski, 2012) and many other fields. All factors that influence the opening of the NSR are gathered from these different published papers, institutional reports, position papers, news and through interview with the experts. Despite all the research mentioned above, there is no research yet that using the Multi-Criteria Decision Making (MCDM) approach that analyse the NSR. Therefore, this chapter intends to analyse the most important factor that influence the opening of the NSR. An Analytic Hierarchy Process (AHP) which is one of the MCDM approach will be used to analyse each level of criteria that has been determined.

2.0 Background of an AHP method

The Analytic Hierarchy Process (AHP) is a problem-solving framework and a systematic procedure for representing the elements of any problem (Saaty, 1983). It has been developed by Thomas Saaty (1980,) and it has increased in popularity and most widely used multi-criteria decision making (MCDM) approaches (Merwe, 2008; Saaty 1980). The AHP is

based on the experience gained by Saaty, while he was directing research projects in the US Arms Control and Disarmament Agency (Bhushan & Rai, 2004). This technique is suitable for dealing with complex systems that involve making a choice from several alternatives and providing a comparison of the considered options. It is also capable of taking large quantities of decision making criteria of quantitative and qualitative nature into consideration and at the same time facilitating the construction of a flexible hierarchy to address a decision making problem (Cheng, 2002). Saaty established a consistent way of converting such pairwise comparisons (for example 'A' is more important than 'B') into a set of numbers representing the relative priority of each of the criteria. In addition, the AHP incorporates a useful technique for checking the consistency of the decision maker's evaluations, thus reducing the bias in the decision making process (Saaty, 1980). AHP has been applied to support decision making in different fields, for example, in engineering (Katarne et al., 2013; Triantaphyllou et al., 1995), healthcare (Pecchia et al., 2011), marketing (Wickramasinghe et al., 2009) and accounting (Apostolou & Hassell, 1993). The AHP also has been accepted as a leading multi-attribute decision model both by practitioners and academics (Presley, 2006; Saaty, 2007).

3.0 Methodology

This research will be conducted by using an AHP methodology which has 4 steps, while step 5 is a sensitivity analysis for a validation process.

3.1 Modelling of the important factors that influence the opening of the NSR

All 4 steps of the AHP technique mentioned earlier will be demonstrated here with the proposed model. The purpose of this test case is to select and ranked the important factors that influence the opening of the NSR.

Step 1: Define the problem and determine the kind of knowledge sought.

The first step was to identify the critical factors or criteria that considered affecting the use of the NSR. Therefore, the goal of the study is to select the most important factor that influences the opening of the NSR. All the factors or criteria are gathered through literature review and extensive brainstorming with the experts. The decision analysts must have a clear understanding of the problem under investigation. Lack of understanding and wrong information will affect the whole structure of the problem

Step 2: Set up the hierarchical system by decomposing the problem into a hierarchy of interrelated elements.

This is the most creative and important part of decision making. All the factors gathered were decomposed into a hierarchy of goal, main criteria, sub-criteria and sub-sub criteria. This process was done together with the active involvement and discussion among the research team and an expert of the NSR. In this stage, the weak and unimportant factors were also eliminated. Appendix A shows the entire hierarchical model in the AHP framework consisting of the goal, main criteria, sub criteria and sub-sub criteria.

Step 3: Construct a set of pairwise comparison matrices to determine the weight.

The third step is the identification of a preference or priority for each criterion in terms of how it contributes to the upper level event. Each element in an upper level is used to compare the elements in the level immediately below with respect to it. The process involves the employment of the pairwise comparison method to each group in the hierarchy to form a matrix and comparing each of the paired elements in the matrices. To construct the measurement, a set of questionnaires will be sent to a number of experts for analysing the priority of each evaluation parameter to another by incorporating the ratio scale of the pair-wise comparison. During this process, the experts are expected to specify how their judgements on a lower level criterion contribute to the formulation of the upper level criteria or top level event. Consequently, there are 7 experts involved in this data collection. The 7

experts are from the shipping industry and the academicians. Hence, the criteria are ranked based on the weight values obtained from the calculation.

Table 1: The ranking of the main criteria

Criteria	Weight value	Ranking
EF	0.2139	1
AF	0.1828	2
FF	0.1528	3
PF	0.1336	4
TF	0.1175	5
VF	0.0881	6
LF	0.0582	7
SF	0.0530	8

Based on Table 1 above, the Economic Factor (EF) is the most important factor that influences the opening of the NSR. Ranking number 2 is the Advantages of the NSR in comparison to other alternatives (AF), followed by Safety Factor (FF). Political Factor (PF) is in ranking number 4. Technological Factor (TF), Environmental Factor (VF), Legal Factor (LF) and Social Factor (SF) are in ranking number 5, 6, 7 and 8 respectively.

Step 4: The calculation of the consistency of the pairwise judgements.

This involves carrying out a consistency measurement to screen out the inconsistency of responses. Comparisons made by this method are subjective and the AHP tolerates inconsistency through the amount of redundancy in the approach. If the consistency index fails to reach a required level then answers to comparisons may be re-examined. The weight values obtained in the pair-wise comparison matrix are checked for consistency purpose using a Consistency Ratio (*CR*). The *CR* value is computed using the equations provided by Saaty, (1990). Consequently, all the experts' judgement is consistent and reliable.

Step 5: Sensitivity analysis
The solution to a decision problem, the global ranking of criteria or alternatives, may not provide enough information to the decision maker to make a final decision. There are several reasons why a sensitivity analysis should be conducted on the results. For instance, the judgements for some criteria may be subjective or there may be uncertainty in the data that leads to the preference value. In addition, the preference judgements may come from a group decision where there are different opinions. Moreover, different prioritisation methods may yield different results for the same pair-wise comparison matrix. A sensitivity analysis provides more understanding about the problem and in this way the decision maker should be able to make a more informed decision. According to Lucia and Mark (2001), parameter sensitivity is usually performed as a series of tests in which the modeller sets different parameter values to see how a change in the parameter causes changes in the model. Also, sensitivity analysis helps to build confidence in the model by studying uncertainties that are often associated with parameters in the model (Lucia and Mark, 2001). There are several methods to perform sensitivity analysis on AHP problems, but this research only focuses on the numerical incremental analysis (Chen and Kocaoglou, 2008). This approach involves changing the weight values and calculating the new solution. The method, also known as One-at-a-time (OAT) (Leonelli, 2012), works by incrementally changing one parameter at a time, calculating the new solution and graphically presenting how the global ranking changes. This is the most commonly used method in associated software tools and, according to Chen and Kocaoglou, is also the most popular in the literature where AHP is used to solve problems. A sensitivity analysis was undertaken to test the outcome of the application of the

AHP model, with 16 different weighting scenarios. For each scenario indicate the outcome of the sensitivity analysis against the base case for comparative purposes. From the original weight values, a new set of weight values is obtained using the percentage increase of 10% and 30%.The sensitivity analysis is important because the ranking of the criteria may not provide enough information for decision maker to make a final decision. Therefore, based on the sensitivity analysis that has been conducted, it can be concluded that the proposed AHP modelling study is reliable and consistent.

4.0 Conclusion

An AHP modelling study has been conducted in this chapter with the purpose of demonstrating the issues or uncertain situations faced by shipping companies in the NSR. The developed model is dynamic and enables it to be used in different situations faced by shipping companies. The sensitivity analysis confirmed the consistency of ranking of the test case. Nevertheless, the selection of evaluation of criteria, sub-criteria and sub-sub-criteria can be improved from time to time based on the changes of situation of the NSR in the future. Thus, the output will be different from this test case. The Economic Factor is classified as the most important factor that influences the opening of the NSR.

5.0 References:

1. Blunden, M., 2012. Geopolitics and the Northern Sea Route. International Affairs, 88(1), pp. 115-129.
2. Kaczynski, V.M., 2013. Russian Arctic resource development and related policy considerations. Georgetown Journal of International Affairs, pp. 181-191.
3. Molenaar, E.J., 2009. Arctic marine shipping: overview of the international legal framework, gaps and options. Journal of Transnational Law & Policy, 18.2, pp. 289-325.
4. Franckx, E., 2009. The legal regime of navigation in the Russian Arctic. Journal of Transnational Law & Policy, 18.2, pp. 327-342.
5. Verny, J., & Grigentin, C., 2009. Container shipping on the Northern Sea Route. International Journal Production Economics, 122, pp. 107-117.
6. Liu, M., & Kronbak, J., 2010. The potential economic viability of using the Northern Sea Route (NSR) as an alternative route between Asia and Europe. Journal of Transport Geography, 18(3), pp.434-444.
7. Schoyen, H. and Brathen, S., 2011. The Northern Sea Route versus the Suez Canal: cases from bulk shipping. Journal of Transport Geography, 19(4), pp. 977-983.
8. Meschtyb, N.A., Forbes, B.C. and Kankaanpaa, P., 2005. Social impact assessment along Russia's Borthern Sea Route: Petroleum Transport and the Arctic Operational Platform (ARCOP). InfoNorth, 58(3). pp. 322-327.
9. Kaczynski, V.M., 2012. Russian Arctic shipping and icebreaking. International Seminar of the Graduate School of Logistics (GSL) University of Incheon.
10. Saaty, T. L., (1983), Priority setting in complex problems, IEEE Trans. Engineering Manage., vol. EM-30, pp. 140-155.
11. Saaty, T.L., 1980. The Analytic Hierarchy Process. New York: McGraw-Hill Book Co.
12. Merwe, A.V.D., 2008. Ideal mode Analytic Hierarchy Process pairwise comparison model. Final MCDM land use report. Annex E, Model description.
13. Bushan, N.; Rai, K., 2004. Strategic Decision Making Applying the Analytic Hierarchy Process. London: Springer.
14. Cheng, E.W.L., 2002. Analytic Hierarchy Process (AHP): A defective tool when used improperly. Measuring Business Excellence, 6(4). pp. 33-37.

15. Katarne, R. and Negi, J., 2013. Determination of importance criteria: Analytic Hierarchy Process (AHP) in technological evolution of automobile steering. International Journal of Industrial Engineerig Research and Development (IJIERD), 4(1). pp. 10-18.

16. Triantaphyllou, E. and Mann, S.H., 1995. Using the Analytic Hierarchy Process for decision making in engineering applications: some challenges. International Journal of Industrial Engineering: Applications and Practice, 2(1). pp. 35-44

17. Pecchia, L., Bath, P.A., Pendleton. N. and Bracale, M., 2011. Analytic Hierarchy Process (AHP) for examining healthcare professionals' assessment of risk factors. The relative importance of risk factors for falls in community-dwelling older people. Methods of information in Medicine. 50(5). pp 435-444.

18. Wickramasinghe, V. and Takano, S., 2009. Application of combined SWOT and Analytic Hierarchy Process (AHP) for tourism revival strategic marketing planning: a case of Sri Lanka tourism. Journal of the Eastern Asia Society for Transportation Studies, 8.

19. Apostolou, B. and Hassell, J.M., 1993. An overview of the Analytic Hierarchy process and its use in the accounting research. Journal of Accounting Literature, 12, pp. 1-28.

20. Presley, A., 2006. ERP investment analysis using the strategic alignment model. Management Research News, 29(5), pp. 273-284

21. Saaty, T.L., 2007. Time dependent decision-making; dynamic priorities in the AHP/ANP: Generalizing from points to functions and from real to complex variables. Mathematical and Computer Modelling, 46(7-8), pp. 860-891.

22. Saaty, T.L., 1990. How to make a decision: The Analytic Hierarchy Process. European Journal of Operational Research, 48, pp. 9-26.

23. Lucia, B. and Mark, C., 2001. An introduction to sensitivity analysis. The Massachusetts Institute of Technology.

24. Chen, H. and Kocaoglu, D.F., 2008. A sensitivity analysis algorithm for hierarchal decision models. European Journal of Operational Research, 185, pp. 266-288.

25. Leonelli, R.C.B., 2012. Enhancing a decision support tool with sensitivity analysis. University of Manchester.

Appendix A

The 3 level of criteria that influence the opening of the NSR

	Level 1	Level 2	Level 3
	Criteria	**Sub-criteria**	**Sub-sub-criteria**
The most important factors that influence the opening	Political Factor (PF)	Campaign Effort (PFA)	Promotion by the Russians (PFAA)
			Collaboration with other countries (PFAB)
			Level of Russian state investment on the infrastructure (PFAC)
		Administration Procedures (PFB)	No ship deviation without Russian permission (PFBA)
			Ship owners need to submit their request to use the NSR 4 months in advance (PFBB)
			Mandatory local inspection of the vessel even though the vessels fulfils the requirements (PFBC)
		Foreign Affairs (PFC)	Political risks and uncertainties because the NSR is in Russian territorial water (coastal route) (PFCA)
			Increasing militarization of the Arctic by the

			Russian Government (PFCB)
			Changes in international political/strategic configuration and relations between major world actors and Arctic ocean coastal states (PFCC)
			Unpredictable behaviour of the Russian Government in relation to selected prospective users of the NSR (PFCD)
	Legal Factor (LF)	Legal status of the NSR. Full Russian jurisdiction or some international status (LFA)	
		Border disputes in the Arctic (LFB)	
		Legal status of vessels and flags when transiting the NSR (LFC)	
		No international legally binding requirements for ship designs & ice class ship (LFD)	
	Economic Factor (EF)	Shipping fixed costs (EFA)	Capital costs (ice strengthened vessels) (EFAA)
			The NSR Insurance costs (EFAB)
			Ship depreciation (EFAC)
			Manning costs (EFAD)
		Shipping variable costs (EFB)	Fuel costs (EFBA)
			The NSR fees (Meteorological forecast & ice breaking) (EFBB)
			Ice pilot fees (EFBC)
		Commercial Aspect (EFC)	Shifts in economic geography (EFCA)
			Lack of major economic centre along the route (EFCB)
			Status of natural resources in Arctic (EFCC)
			Tourism industry (EFCD)
	Environmental Factor (VF)	Disappearing of summer sea ice (VFA)	More navigable days for shipping operations (VFAA)
			Possible extinction of Polar bears (VFAB)
			Some Arctic fisheries will be affected (VFAC)
		Challenges to operation (VFB)	Operational conditions like wind chills, darkness in winter, sea ice & ice bergs, high latitudes and etc. (VFBA)
			Seasonality of operations (Navigable for 2 to 4 months in eastern part of the NSR :without ice breaking assistance) (VFBB)
			Shallow seas & straits (Vessel size restriction in coastal route) (VFBC)
		Impact on the marine environment and marine biodiversity (VFC)	Accidental discharges of polluting substances (cargo or fuel) (VFCA)
			Operational discharges (cargo residues, fuel residues),garbage and sewage and emissions (CO_2, NO_2 SO_2) (VFCB)
			Navigation impacts (noise pollution and interference with marine species that cause disruption of behaviour and etc.)(VFCC)
			Introduction of alien organisms through ballast water exchanges or attachment to vessel hulls. (VFCD)
	Social Factor (Indigenous People) (SF)	Loss of food source (SFA)	
		Loss of housing (SFB)	
		Disease (SFC)	
		Loss of culture (SFD)	
		Stimulation of economic	

		activity of people in the north region (SFE)	
	Technological Factor (TF)	Advanced ice breaking technology (TFA)	
		New ship technology/design (TFB)	
		Aerial drones will be used to spot free and fast ice (TFC)	
	Safety Factor (FF)	Status of shipping and port infrastructure (FFA)	Status of search and rescue facilities (FFAA)
			Status of availability of international port along the route (FFAB)
			Status of ships repair and maintenance facilities (FFAC)
		Status of navigational aids facilities (FFB)	Charting and monitoring (FFBA)
			Radio and satellite communications and emergency response (FFBB)
			Observational networks and forecast for weather, icing, waves and sea ice(FFBC)
		Training for crew for Arctic operations (FFC)	
	Advantages of the NSR in comparison to other alternatives (AF)	Shorter route (AFA)	Saving in time(AFAA)
			Saving in expenses (AFAB)
			Increase the number of roundtrips (AFAC)
			Reduced air emissions from ships (AFAD)
		No piracy/terrorism threat (AFB)	
		No vessel size restriction for further north route of the NSR (AFC)	

Ambisire Usman

An overview on resilience modelling for oil terminal operations

AY Usman, J Ren, J Wang, K Jones
Liverpool John Moores University, UK
E-mail: A.Y.Usman@2012.ljmu.ac.uk, J.Ren@ljmu.ac.uk,
J.Wang@ljmu.ac.uk, K.O.Jones@ljmu.ac.uk.

Abstract. The maritime industry by nature is very conventional and diversifying from the norm and has often proven to be difficult. Oil and gas terminal have continually been adapted to new requirements and regulation in design and operation. Risk impacts associated with increase accidents such as the Mumbai high North disaster (2005) which ruptures the flexible riser pipe system causing severe explosion, damage to infrastructure, environment and loss of life with high consequences. The challenges of these risks are so enormous that it requires flexible yet robust risk analysis methods in order to tackle the unpredictable events. It is important to investigate the uncertainties of oil terminals operations in order to monitor, mitigate and control the operational risks. The important questions to be considered are: (a) will the proposed framework and models affect the operations in oil terminals (b) What are the risks involved in oil terminal operations (c) What are the values of the risk models to oil terminal operations (d) Should operators attribute loss to normal errors in measuring hazards, other than improved techniques to rectify measuring of errors. This paper tends to discuss the regular operations of oil terminals from existing literatures and to identify different risk models for oil terminal operations.

Keywords: Resilience, Oil terminal operation, Risk assessment, Safety management, Uncertainty.

1. Introduction

Seaport and offshore terminals have an important role in maritime transportation. Maritime critical infrastructure and transportation (MCIT) systems are exposed to diverse risks in their optimised standard of operations due to the continual growth in system complexity. The interfacing and interaction of component factors to collectively provide the functionality needed by the system to attain its goal is under scrutiny due to high uncertainties.

Over the past few years, events such as the Mumbai high north disaster (2005) which ruptured the flexible pipe system, the Bunga Alpinia disaster (2012) which caused severe explosion, New York harbour oil terminal disaster (2012) which damaged infrastructures, environment and loss of life and the Santos, Brazil (2015) disruption of container operations caused by fire at the ultra-cargo liquid bulk terminal. The result of such events led to major economic loss and as such, the challenges of these risks are so enormous that it requires flexible yet robust risk analysis methods in order to tackle the unpredictable outcomes.

Terminals play an important role in freight transportation. When an accident occurs, the terminal cannot perform its desired functions thus resulting to a failure (Baublys 2007).When dealing with oil terminal and ship operations, there are a lot of uncertainties in such

processes, thus optimization under high uncertainties is essential within operational research (Hess and Hess 2010). Terminal operations could be in the form of a deterministic terminal behaviour i.e. from ship navigation to the idle state, preparatory state, transhipment state, closing state, repair and maintenance state or a stochastic terminal behaviour i.e. where the order of state and transition do not follow a logical workflow due to unforeseen influences on regular operations.

The hazard components covered are in the vicinities of highly active reputable ports such as the current and change in probability of ship navigation, (un)docking, pipe trestles, vapour handling, pumping, (un)loading, consequences resulting from terminal operation accidents such as environmental consequences, consequences to the ship & port, and third-party consequences, as well as other issues such as effects on Search and Rescue operations, oil spill monitoring, surveillance, security and risk reduction measures. Important questions to be considered includes: What are the hazards and risks involved in oil terminal operations? What are the values of the proposed novel risk models to oil terminal operations? Should operators attribute loss to normal errors in measuring hazards, other than improved techniques to rectify measuring of errors? This paper tends to discuss the existing problems in oil terminal platforms and on the other hand investigate the uncertainties in oil terminal operations.

2. Current practices on maritime safety

Following the public inquiry into the Piper Alpha accident (1988) which claimed 167 lives, the response to the accepted findings of the Piper Alpha inquiry launched a review of all offshore safety legislation and implemented changes (Wang 2002). In 1993, Formal Safety Assessment (FSA) was adopted by the International Maritime Organisation (IMO) and became the norm for a strategic oversight of safety and pollution prevention in the maritime sector; the processes are to identifying hazards, evaluating risks and decide on an appropriate course of action to manage these risks and this provides a good platform for safety management practices (Trbojevic and Carr 2000). More so, a Safety Management System (SMS) approach was developed through a quantified risk assessment and it provides a detailed assessment of risks and the identification of the most appropriate management strategies. This was in conformity to the requirements of the Port Marine Safety Code which applies to all UK ports and for which the policy is laid down by the department for Transport in consultation with the industry and monitored by the Maritime and Coastguard Agency. The purpose of a safety management system is to ensure that the organization is achieving their goals safely, efficiently, and without damaging the environment (Wang 2002) as such, high risk areas are investigated in detail and the approach for risk prioritization discussed by management. Figure 1 shows the main percentage of the principal causes of safety problems in maritime transportation and Figure 2 shows distribution of potential annual loss in Maritime Transport System (MTS).
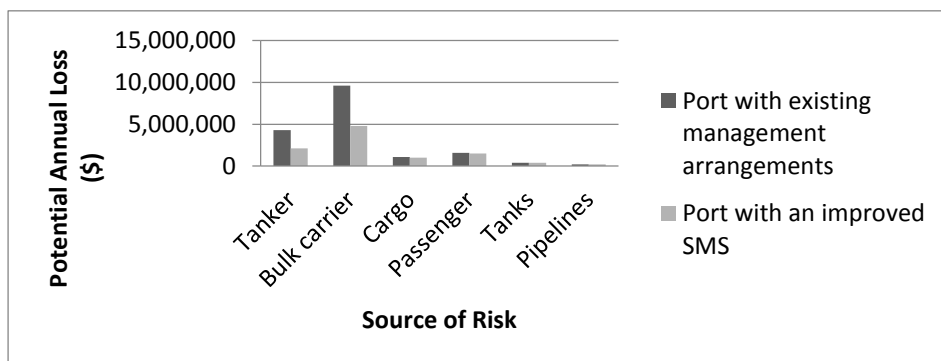
Figure 1: Comparison of potential annual loss in MTS as adapted in Trbojevic and Carr (2000)
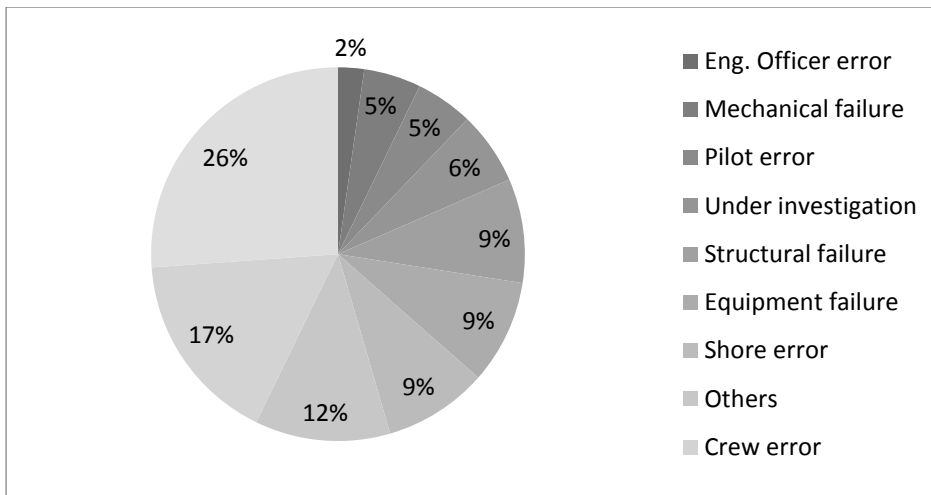
Figure 2: Percentage of the principal causes of safety problems as adapted in Soares and Teixeira (2001)

2.1 Oil Terminal Operations (OTOs) on onshore and offshore terminal platforms

Few literatures reveals that operation is probably the main contributor of accidents in maritime platforms and increasingly more attention should be given to the study of these activities and formalised risk assessment should gradually become more common in complex systems (Soares and Teixeira, 2001, Ronza et al, 2005, Ronza et al, 2007). The offshore facilities identified as maritime platforms are as follows: (1) trans-shipment stations, (2) a maritime floating port terminal, (3) drilling platforms, (4) crude oil recovering platforms, (5) crude oil production platforms, (6) linking platforms, (7) water injection platforms, (8) pumping platforms, (9) shelter platforms, (10) telecommunication platforms, (11) crude oil measurement platforms, and (12) flaring platforms(Villasenora et al. 2003). Tankers, usually large ships are used for carrying oil products in bulk and this has been the method of transporting large quantities of oil around the world. Figure 3 represents generic ship/terminal interface for OTOs and Figure 4 represents a general activity in oil terminal. The classifications of tankers according to their capacity to carry oil are as follows:

1) GP (general purpose) tankers, under 38,000 tonnes

2) MR (medium range) tankers, 38,000 to 50,000 tonnes

3) Panamax which are ship that can access a Panama Canal, 50,000 to 79,000 tonnes

4) Aframax, 80,000 to 125,000 tonnes

5) Suezmax, 125,000 to 200,000 tonnes and also can access a Suez Canal

6) VLCC (Very Large Crude Carriers), over 20,000 tonnes and the ULCC (Ultra Large Crude Carrier).These are generally considered as super tankers because they are ships capable of carrying above 250,000 tonnes.
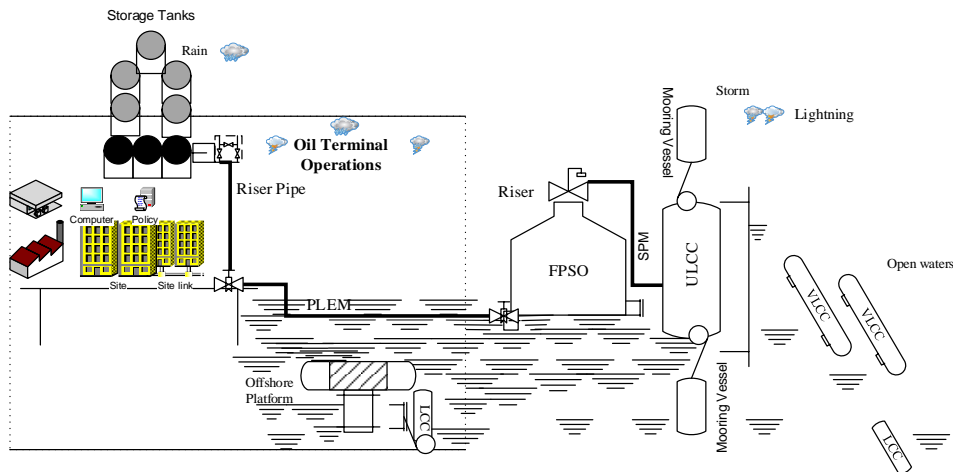
Figure 3: A generic environment for OTOs.
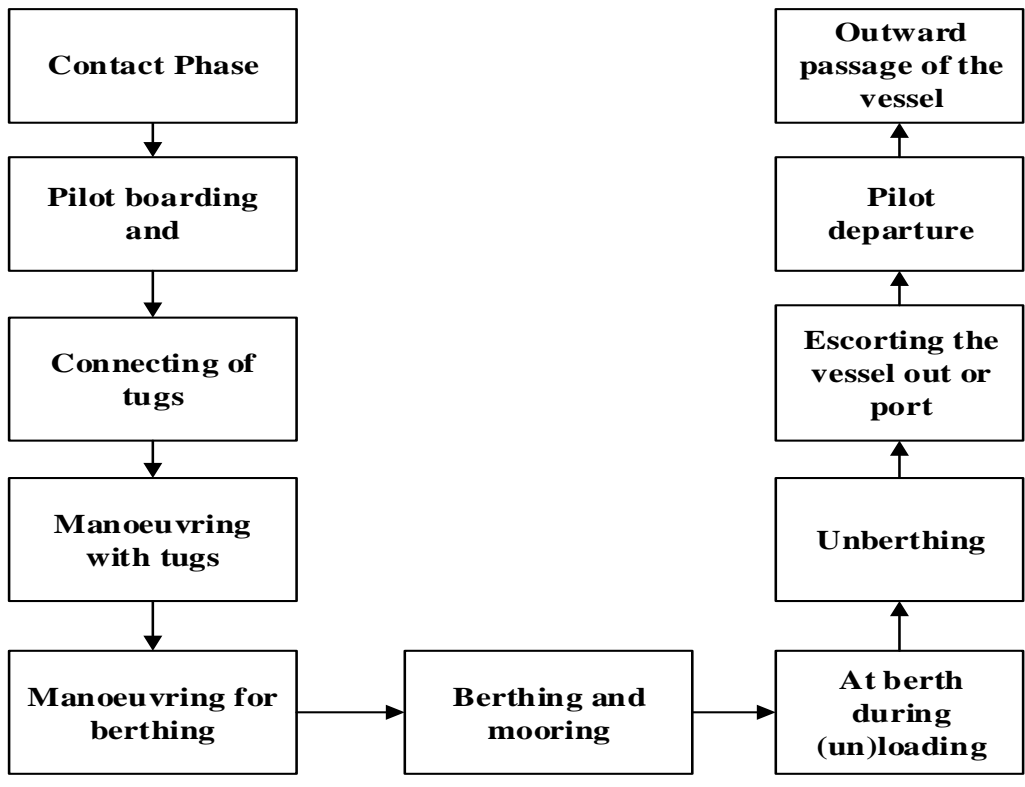
Figure 4: A generic oil terminal operation as adapted from Trbojevic and Carr (2000)

2.2 A Resilience Engineering (RE) review relevant to OTOs

There are different definitions of resilience in a resilience engineering context (Madni and Jackson 2009, Mansouri *et al.* 2010, Steen and Aven 2011, Dinh *et al.* 2012), but a generalized definition of resilience

by Wreathall's, 2006 come into context in this research; he defines resilience as management's ability to recognise, adapt or recover a system rapidly to a stable condition, enabling it to continue its activities during and after a major accident rather than prevent incidents from occurring. In the context of port infrastructure system (PIS), resilience can be defined as a function of systems vulnerability against potential disruption with adaptive capacity in recovering in the face of major shocks within a reasonable timeframe after a major accident. Therefore, resilience as an additional safety measure is needed in complex operational designs (Dinh et al. 2012). Different research review have stated that for a

system or an organisation to be resilient enough, it must have the following qualities: the ability to: 1) Respond to regular and irregular threats in a robust, yet flexible manner, 2) Monitor what is going on, including its own performance, 3) Anticipate risks (risk events) and opportunities, and 4) Learn from experience.

2.3 Approaches on how to measure system resilience

The following are ways to measure resilience: 1) Vulnerability assessment using performance metrics for evaluating resilience (McManus et al. 2008): it contributes to increased situation awareness, promotes the development of adaptive capacity and gives the organisation something to work on such as using preparedness and criticality information for both response and recovery phase of a crisis, 2) Design and retrofit strategies for resilience: resilience heuristics comes to effect where experience, judgement and intuition for which result can infrequently be quantified or verified (Madni and Jackson 2009) and 3) Risk-based assessment and decision methods for achieving resilience that are supported by a cost/benefit analysis and performance-based methods for codes, standards, and practices (Wang 2002, Berle *et al.* 2011).

2. Research methodology

The proposed methodology for hazard evaluation process in the most generic case should follow 4 major steps:

Step 1– Define the aim of the evaluation.
Step 2– Identification of the hazard factors.
Step 3– Developing a dynamic and sub-dynamic hierarchical structure based on the hazard factors.
Step 4– Conduct an empirical study

3.1 Existing hazards on oil terminals platform

Step 1: The aims of the evaluation are to investigate the uncertainties of oil terminals operations in order to monitor, mitigate and control the operational risks.

Step 2: A review on academic journal articles and conference papers, accident reports (MARSH 2011, MARSH 2013, ITOPF 2015) seminars and brainstorming sessions on OTOs and oil terminals platform has identify various hazards in the maritime industry. The identified hazard for OTOs can be categorised into (a) Man related hazard (b) Machine related hazard (c) Communication and correspondence related hazard (d) Management related hazard and (e) Nature related hazard.

Step 3: Complex Marine Operations (CMOs) are often associated with high level of uncertainty due to the ever-changing environment leading to a range of possible accident. Uncertainty in OTOs requires method that combines expert's judgement and human knowledge to propose a reliable dynamic hierarchical structure. A pilot study was conducted with specialist and academicians to determine the consistency of the identified unexpected events for further expert's judgement. Fig 5 shows a dynamic hierarchical structure for OTOs. In the process of developing a sub-hierarchical structure, the clustering process involve three major steps; namely (a) identification of elements of similar kind in the hierarchical structure (b) grouping of observable elements together into smaller group on the basis of self-similarities (c) developing the sub-hierarchical structure(Lu and Shang 2005, John and Ai 2010, Xin and Huang 2013).

Step 4: Empirical study provides a set of values and consistency to simulate a proposed resilience decision strategy for OTOs. More so, it is an investigative platform, where there is no extent research on significant hazards causing unexpected events on CMOs. This study was executed in 3 phases: (1) Questionnaire formulation and pilot study, (2) Choosing the right experts, (3) Survey data collection and description. The questionnaire was web-based and a link was e-mailed to targeted expert participants. . Experts with relative (onshore/offshore oil fields) and vast (Academic, maritime domain, oil and gas refineries) experience related to this research where drawn at random. Another factor being considered as criteria for choosing experts was the region/country; it has to be a crude oil producing region or where refined crude oil is being consumed in large amount thereby leading to large

import of these refined products. To identify these experts (Hammitt and Zhang 2013), publicly available directories, LinkedIn and recommended oil terminal professional where targets. Fig 5 shows a graph representing the empirical evaluation.

3. Data collection results and discussion

From the questionnaire, the research revealed that 100% of the participants show that there is a good awareness and knowledge about the Hazard Factors (HFs), as well as the possible consequences of these HFs to OTOs. In addition, Table 1 shows the consistency of participant's responses. The participants are all senior employee with experience all over the world, which also represents around 90% of the whole sample. In order to eliminate the presence of response bias, participants were chosen from the United Kingdom, Iran, United Arab Emirate, China and Nigeria. These represents the major oil importing and exporting countries were OTOs activities are being performed in a larger scale. According to participants, the probability of these HFs occurring on oil terminal platforms span across very unlikely (1) to moderately likely 6).

Table 1: Consistency of HFs based on experts' responses

Hazards/Failures	HFs	Arithmetic mean
Man related Hazard	Major Accident Hazards	4.40
	Duty Holders Error	4.80
	Personnel issues	4.40
	Sabotage	3.40
	Indirect Contributing Factors	3.80
Machine related Hazard	Maintenance Event Hazards	3.80
	Equipment Failure	3.80
	Well Control System Failure	4.00
Coordination and correspondence related Hazard	Platform Communication Misinterpretation	4.80
	Situation Awareness	4.40
	Lack of Proper Crew Interaction	4.20
Management related Hazard	Human Resources Error	4.00
	Job Safety Rules and Regulation	3.80
	Latent Error	4.00
	Design Error	3.60
	Business Risk	3.60
Nature related Hazard	Hydrologic	3.40
	Seismic	3.40
	Atmospheric	3.60
	Epidemic	3.60
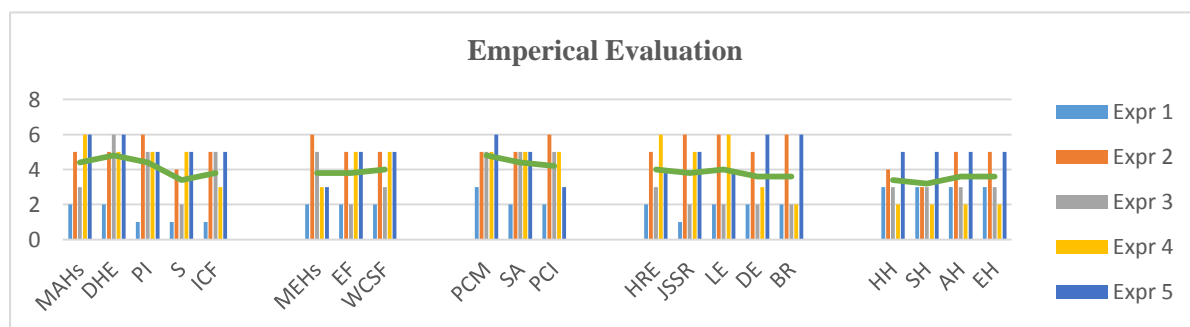
Figure 5: A graph showing the Empirical evaluation of HFs

5. Conclusion

Oil terminal platforms are vital complex marine operations, thereby facilitating global trade. The assessment of significant events disrupting OTOs tend to suffer from a lack of data to explain such events, and also from their infrequent occurrences. The risk within these operational infrastructures requires attention in respect to their identification, assessments, mitigation to facilitate the process involved to significantly tackle uncertainties on oil terminal platforms for resilience improvement. Onshore/offshore oil industry is distinctive because of the convergence of several hazardous factors. There is a continual risk of organisational and fatal accidents resulting from unanticipated actions from employees on offshore installations.

References

Baublys, A. (2007) 'Probability models for assessing transport terminal operation.', *journal of Transport*, 22(1), 3-8.

Berle, Ø., Asbjørnslett, B. E. and Rice, J. B. (2011) 'Formal Vulnerability Assessment of a maritime transportation system', *Journal of Reliability Engineering & System Safety*, 96(6), 696-705.

Dinh, L. T. T., Pasman, H., Gao, X. and Mannan, M. S. (2012) 'Resilience engineering of industrial processes: Principles and contributing factors', *Journal of Loss Prevention in the Process Industries*, 25(2), 233-241.

Hammitt, J. K. and Zhang, Y. (2013) 'Combining experts' judgments: comparison of algorithmic methods using synthetic data', *Risk Analysis*, 33(1), 109-20.

Hess, S. and Hess, M. (2010) 'Predictable uncertainty about terminal operations in the sea', *Transport*, 25(2), 148-154.

Hetherington, C., Flin, R. and Mearns, K. (2006) 'Safety in shipping: the human element', *Journal of Safety Research*, 37(4), 401-11.

ITOPF (2015) *Oil tanker spills statistics*.

John, H. R. and Ai, S. (2010) 'Revisiting our 1991 paper on fire risk assessment', *Fire Technology*, 46(1), 789-801.

Lu, C.-S. and Shang, K.-C. (2005) 'An empirical investigation of safety climate in container terminal operators', *jounal of Safety Research*, 36(3), 297-308.

Madni, A. M. and Jackson, S. (2009) 'Towards a conceptual framework for resilience engineering', *Systems Journal, IEEE*, 3(2), 181-191.

Mansouri, M., Nilchiani, R. and Mostashari, A. (2010) 'A policy making framework for resilient port infrastructure systems', *Journal of Marine policy*, 34(6), 1125-1134.

MARSH (2011) *"Large property damage losses in the hydrocarbon industry"* 22nd Edition.

MARSH (2013) *"Large property damages in the hydrocarbon industry"* 23rd Edition.

Sii, H., Ruxton, T. and Wang, J. (2002) 'Novel risk assessment and decision support techniques for safety management systems.', *Proceedings of IMarEST Part A- Journal of Marine Engineering and Technology*, 2002(1), 15 - 27.

Soares, C. G. and Teixeira, A. P. (2001) 'Risk assessment in maritime transportation', *journal of Reliability Engineering and System Safety*, 74(3), 299 - 309.

Steen, R. and Aven, T. (2011) 'A risk perspective suitable for resilience engineering', *Journal of safety Science*, 49(2), 292-297.

Trbojevic, V. M. and Carr, B. J. (2000) 'Risk based methodology for safety improvements in ports.', *J Hazard Mater*, 71, 467- 480.

Villasenora, R., Magdalenoa, M., Quintanara, A., Gallardoa, J. C., Lopeza, M. T., Juradoa, R., Mirandaa, A., Aguilara, M., Melgarejoa, L. A., Palmerına, E., Vallejoa, C. J. and Barchetb, W. R. (2003) 'An air quality emission inventory of offshore operations for the exploration and production of petroleum by the Mexican oil industry', *Journal of Atmospheric Environment*, 37(26), 3713-3729.

Wang, J. (2002) 'Offshore safety case approach and formal safety assessment of ship.', *J Safety Res*, 33(1), 81-115.

Xin, J. and Huang, C. (2013) 'Fire risk analysis of residential building based on scenario clusters and its application in fire risk management', *Fire Safety*, 62(1), 72-78.

Azraa Ab Rahim

Experimental method for groundwork modeling a two-mass rotational (TMR) system with backlash – A comparison between PRBS and PRMLS excitation.

A A A Rahim, C Matthews
Faculty of Engineering and Technology,
Liverpool John Moores University,
Liverpool, UK
A.A.AbRahim@2013.ljmu.ac.uk, C.Matthews@ljmu.ac.uk

Abstract. This paper intends to lay the groundwork for modeling backlash which is a discontinuous nonlinearity in a two-mass rotational (TMR) system. It involves experiments and practical considerations prior to carrying out a system identification (SID) of the system. The SID is carried out in MATLAB. A linear configuration of the TMR plant was used to identify the linear representation which produced an AutoRegressive eXogenous (ARX) model. Different levels of backlash are then added to the TMR and the Hammerstein model is employed to identify the system inclusive of backlash. A comparison between the Pseudorandom Multilevel Sequence (PRMLS) and the Pseudorandom Binary Sequence (PRBS) as the input excitation is carried out to rule out speed dependency of the backlash nonlinearity in the system. PRMLS can invoke the nonlinearity of a system when the nonlinearity is amplitude dependent. The results obtained are intended for use in the design of active vibration control.

Keywords. Backlash, System Identification (SID)

1. Introduction

A two-mass rotating (TMR) system is a good representation of the driveline system in which it incorporates a drive, a transmission with some speed reduction mechanism, and an inertial load which is subjected to some external loading. For this TMR system, it is desired to focus on the transient response and vibration because the effect of backlash phenomena tends to make it difficult to control the vibrating system.

A lot of effort can be spent into trying to estimate and physically represent the backlash as performed by Lagerberg in [1] however, the question that remains unanswered is, whether it is an efficient way of modeling, on the basis that there are certain information about the backlash that are not available. Meanwhile, another alternative of looking at it is from the black box perspective, by producing a mathematical description that allows the presence of the backlash to be considered without explicitly modelling it.

There are a number of ways to represent the backlash in systems depending on the degree of complexity and detail desired. However, it has to be kept in mind that, the more complicated a model is, for example by explicitly modeling the nonlinearity, the less applicable it tends to be when it comes to designing a controller for the system. If the model is nonlinear, then many classical and modern robust control techniques are not suitable since they are based on the linearity assumption.

1.1. Backlash

Backlash is a non-smooth nonlinearity with multivalued mapping. Backlash arises when there are two moving parts that should move together and produce an output movement but due to some gap between them, the output does not materialize. It is also described as

excessive play between mechanical parts and most commonly found in gears where there exist a gap between mating gears [2][3][4][5][6][7].

Backlash belongs to the hard nonlinearity group due to its' discontinuous and non-differentiable character and it is widely acknowledged that controlling a system with backlash is complicated especially when high precision is required [9], [10]. Even though efforts have been made to control systems with backlash, steady state errors and limit cycles can still surface [3], [11]. Other than that, backlash can contribute to excite high-order modes in a driveline, further degrading the driveline performance [12].

As aptly described in [14], the best model is one which is simple, meaning that it does not include every complexity present in the system, but rather just enough to capture the important features and behavior of a system.

According to [5], [9], [10], [15] an ideal representation of backlash as illustrated in figure 1 with input signal, u(t) and output variable x(t) can be mapped as in (1)

$$x(t) = \begin{cases} m_L[u(t) + c_L] & u(t) \leq z_L \\ x(t-1) & z_L \leq u(t) \leq z_R \\ m_R[u(t) - c_R] & u(t) \geq z_R \end{cases} \tag{1}$$

The parameters that characterize the backlash are the dead zone constants, $c_L > 0$ and $c_R > 0$ and as well as the slopes m_L and m_R.
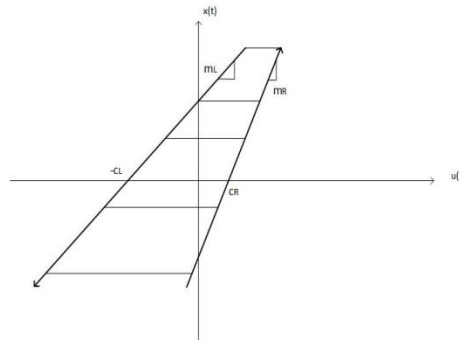
Figure 1. Ideal Backlash characteristic

2. System Identification

A methodology for performing a systematic design of a controller for system with backlash for a TMR system needs to be formulated which require a plant model.

This article considers the use of a system identification (SID) approach where the SID test is run inside certain bounded operating regimes and then the linear models in each of those regimes can be identified. That will provide a set of linear identified models which are all representative of the local behaviours. Those set of models can then be used to represent the behaviour of the system over its whole operating range and the nonlinearity can be treated as uncertainty in the context of the way the system behaves.

2.1. Hammerstein Model

In this article, the backlash is a nonlinearity that precedes the dynamic linear system and to reduce the complexity of identifying a nonlinear model from empirical data, model structures which contain a static nonlinearity in series with a linear dynamic system have been extensively studied [10], [16], [17]. The location of the nonlinear element results in different responses. The Hammerstein model is for systems where the nonlinear element precedes the linear block as is the case in this article. In cases where the linear element precedes the nonlinear block, it is represented by the Wiener model.

The Hammerstein model of a nonlinear system is shown in figure 2. All of the dynamic in the two mass rotational system are linear dynamics and by intentionally adding backlash to the system, the static nonlinearity is superimposed onto it and that is a nonlinear mapping of a linear system.
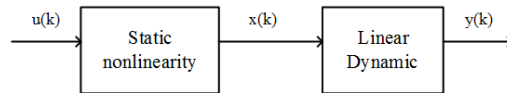
Figure 2. Hammerstein model

3. Application

A vehicle driveline is one application in which there exist discontinuous nonlinearities such as backlash. The driveline system is known to be nonlinear especially when there is a change in torque delivery from the engine specifically during launch phase in which the vehicle is currently stationary and the driver depresses the accelerator pedal. Backlash is present in the driveline and when the entire backlash is traversed, the resulting impact load will excite the torsional modes of the driveline even more. This torsional oscillation in the driveline is transmitted through the tyres so the vehicle and the driver experience it as a longitudinal oscillation or a superimposed oscillation on top of the low frequency acceleration of the car. This is one of the main concerns of noise, vibration and harshness (NVH) issues that affect driveability.

4. Experimental Setup

In this paper, a physical representation of the system is used but on a different scale to a vehicle system. The TMR system is going to be used as a basis for exploration into possible ways to characterize its behavior and use the characterization to produce useful control objectives for the design of active vibration control.

4.1. ECP220 Industrial Emulator

For the purpose of experimental work, the ECP Industrial Emulator (Model 220) which will be referred to as ECP220 from here onwards, in figure 3 was used, which is a two mass rotating plant. The TMR system has a pair of inertias which are connected via a drive belt system with speed reduction. The drive belt system can be rigid, or it can be modified to add a compliant interconnection hence the ECP Industrial Emulator is adjustable so that it can be used to represent a number of different variations. This rig is capable of incorporating backlash by adjusting the position of a grub screw depending on how much backlash is desired.

Figure 3. ECP Industrial Emulator Model 220 (ECP220)

Accordingly, this article then uses ECP220 to establish different sets of configuration setup as a simplification of the driveline system. It is possible to incorporate the configuration as in figure 4, in which the symbols are defined in table 1. Figure 4 illustrates the representation of a TMR system which is a rigid model. The TMR is sued as an analogue to a car driveline on a smaller scale where the engine and transmission inertia are lumped at the left hand side and then a lossless gearbox that then drives the load inertia which is the combination of inertia of the drive shafts, the wheels, the hub assemblies. It also incorporates a backlash through an adjustable gap via a grub screw located under the idle gears in figure 3.
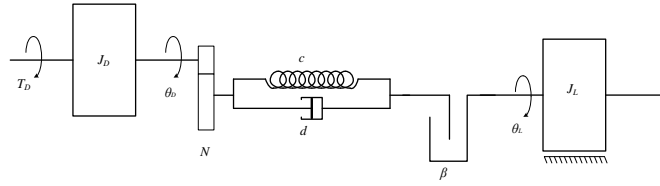
Figure 4. Rigid TMR system with backlash

Table 1. Parameter Definitions for TMR System Configuration.

Symbol	Name	Unit
T_D	Drive Torque	Nm
J_D	Drive Inertia	$kg.m^2$
J_L	Load Inertia	$kg.m^2$
θ_D	Drive Angle	rad
θ_L	Load Angle	rad
N	Gear Ratio	-
d	Damping Constant	Nm.s/rad
c	Spring constant	Nm/rad
β	Backlash	rad

Initially, in order to perform characterization experiments, a set of experiments were carried out according to the following setup:

i. Nominal plant with no backlash.

ii. Backlash level 1 (the grub screw was loosened by a 90˚ counter clockwise (CCW) turn to create a gap).

iii. Backlash level 2 (the grub screw was loosened by an additional 90˚ CCW in addition to backlash level 1).

iv. Backlash level 3 (the grub screw was loosened by an additional 90˚ CCW in addition to backlash level 2).

v. Backlash level 4 (the grub screw was loosened by an additional 90˚ CCW in addition to backlash level 3).

4.2. Stimulus

Pseudo-random binary sequence (PRBS) as depicted in figure 5 is a well-known input excitation used for identification. It is a periodic, deterministic signal with properties similar to white noise auto-correlation function contributing to the high frequency content [11]. According to several authors [12], [11], [13], [14], PRBS is a practical choice of input and provides sufficient excitation for linear system identification. It is produced by having two states, either ON or OFF for a specific length and the sequence can be repeated multiple times.
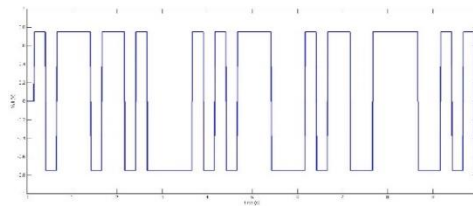
Figure 5. PRBS input excitation

Consequently, [15] highlights the characteristics of PRBS by the following:

i. The duration of the switching sequence.

ii. The switching time, T_{SW} which is indicated by the shortest time change between the levels of signal (multiples of sampling time).

There is no doubt that the PRBS is a popular test signal owing its popularity to its persistently exciting characteristics as abundantly reported in the literature [7], [8],[14],[16],

however, there are others who casted some doubt about its persistency for identifying nonlinear systems as suggested in [14] and [17] to which an alternative signal was suggested, the Pseudorandom Multilevel Sequences (PRMLS) to be used in nonlinear system identification. PRMLS as illustrated in figure 6 is similar to PRBS other than having more than just binary signal amplitudes [18]. This feature of having multilevel amplitudes can invoke the nonlinearity of the system when the nonlinearity in the system is amplitude dependent [17].
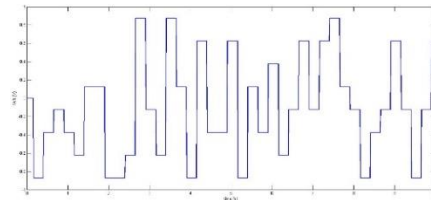
Figure 6. PRMLS input excitation

One method to investigate whether there exist any amplitude dependent nonlinearity is to take a set of data collected using the PRBS input excitation with a limited amplitude range, and identify a linear model and the model is then validated against unseen data. The second round of validation of the linear model is then done against data collected using the PRMLS input excitation there is a wider range of amplitude above and beyond the amplitude used in the original identification. If the validation gives a consistent quality of fit then it is possible to rule out is that there exist an amplitude dependent element of the nonlinearity.

5. Result
The backlash nonlinearity is not a speed dependent nonlinearity; rather it tends to be position dependent. In terms of the PRMLS, it gives the ability to do go through a series of speed range but for the backlash nonlinearity, where the focus is on traversing the gap, PRMLS provides very limited benefit. However, if the two mass rotational system includes friction, which is velocity dependent, then it is expected that the PRMLS will make a big difference. Therefore, for the backlash nonlinearity, its response should not be affected by speed albeit the traversal of the backlash maybe slightly faster than or slower. Due to the nature of backlash, the speed of the traversal is not going to be directly proportional to the speed of the load of the system.

Since it is desired to explore the nonlinearity of the system, using a PRMLS input excitation allows the determination of whether there exists any significant amplitude dependent nonlinearity. If the system is not speed dependent, the result of the system identification using PRMLS and the PRBS excitation will not vary too much.

If the intention is to identify a system that has an amplitude dependent nonlinearity using the PRBS input signal, it would be possible to establish a good linear model of the system for a range of amplitudes where the system behaves linearly. However, if the model were to be subjected to bigger amplitudes, then the linear models no longer provides a good fidelity of the system, hence exposing the amplitude dependency of the nonlinearity in the system. This is because most physical systems have nonlinear characteristics outside a limited linear range. This behavior can only be observed by comparing the quality of the model, with PRBS and with PRMLS, where there is a wider range of amplitudes.

As shown in figure 8 and figure 9, the validation fit of the model are consistent hence it is possible to say that the backlash nonlinearity is not speed dependent.
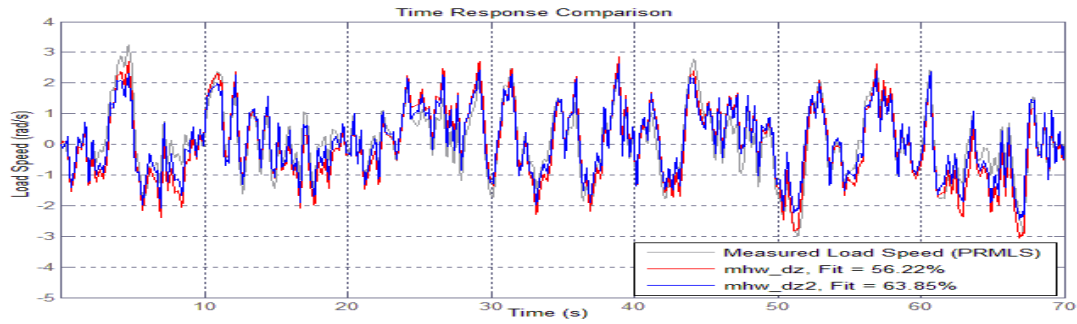
Figure 7. Validation of model obtained using PRBS excitation with data collected with PRMLS excitation.
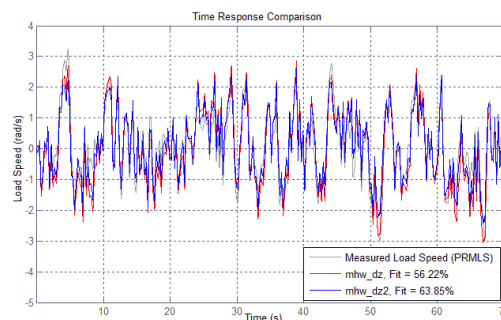
Figure 8. Validation of model obtained using PRBS excitation with data collected with PRMLS excitation.
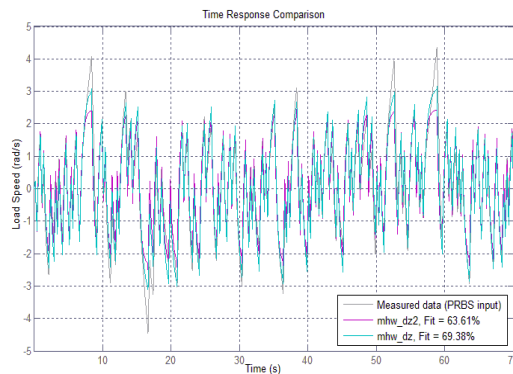
Figure 9. Validation of model obtained using PRMLS excitation with data collected with PRBS excitation.

6. Conclusion

This paper establishes the groundwork for modeling a two-mass rotational system with backlash using the Hammerstein model. It elaborates on the steps required to perform a practical SID on a real plant. The guidelines to perform experiments on the physical plant and practical considerations will be used as a base for future work in characterizing discontinuous nonlinearities in a two mass rotational system application.

7. References

[1]   A. Lagerberg, "Control and Estimation of Automotive Powertrains with Backlash," Chalmers University of Technology, Sweden, 2004.

[2]   M. Nordin, J. Galic', and P.-O. Gutman, "New Models for Backlash and Gear Play," *Int. J. Adapt. Control Signal Process.*, vol. 11, pp. 49–63, 1997.

[3]   M. Nordin and P.-O. Gutman, "Controlling mechanical systems with backlash—a survey," *Automatica*, vol. 38, no. 10, pp. 1633–1649, Oct. 2002.

[4]   R. M. R. Bruns, J. F. P. B. Diepstraten, X. G. P. Schuurbiers, and J. A. G. Wouters, "Motion

Control of Systems with Backlash," 2006.

[5]   J. Vörös, "Parametric Identification of Systems with General Backlash," *Informatica*, vol. 23, no. 2, pp. 283–298, 2012.

[6]   P. Durdevic, "Balancing a Robot Using Intelligent Control," Esbjerg Institute of Technology, Aalborg Universitet Esbjerg, 2012.

[7]   N. Sarkar, R. E. Ellis, and T. N. Moore, "Backlash Detection in Geared Mechanisms: Modeling, Simulation, and Experimentation," *Mech. Syst. Signal Process.*, vol. 11, no. 3, pp. 391–408, May 1997.

[8]   Z. Yumrukcal, "Dynamic Modeling of High Precision Servo Systems with gear backlash," Middle East Technical University, 2013.

[9]   G. Tao and P. V. Kokotovic, "Adaptive Control of Systems with Backlash *," *Automatica*, vol. 29, no. 2, pp. 323–335, 1993.

[10]   V. Cerone, D. Piga, and D. Regruto, "Bounded Error Identification of Hammerstein Systems with Backlash," in *Block Oriented Nonlinear System Identification*, Berlin, Heidelberg: Springer Verlag, 2010, pp. 367–382.

[11]   E. Mohammadiasl, "Vibration Detection and Backlash Suppression in Machine Tools," in *RSJ nternational Conference on Intelligent Robots and Systems*, 2009, pp. 972–977.

[12]   J. A. De Marchi, "Modeling of Dynamic Friction, Impact Backlash and Elastic Compliance Nonlinearities in Machine Tools, with Applications to Asymmetric Viscous and Kinetic Friction Identification," Rensselaer Polythechnic Institute, Tro, New York, 1998.

[13]   G. Ferrari-trecate and M. Gati, "Observability analysis and state observers for automotive powertrains with backlash : a hybrid system approach," vol. 33, no. 0, pp. 1–14, 2005.

[14]   K. Velten, *Mathematical Modeling and Simulation: Introduction for Scientists and Engineers*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2008.

[15]   K.-H. Han, G.-O. Koh, J.-M. Sung, and B.-S. Kim, "An Adaptive Control Approach for Improving Control Systems with Unknown Backlash," *Int. J. Aeronaut. Sp. Sci.*, vol. 12, no. 4, pp. 360–364, Dec. 2011.

[16]   E. Elayan, "Nonlinear system identification of hystersis-backlash in hammerstein models," *2013 5th Int. Conf. Model. Simul. Appl. Optim.*, no. 2, pp. 1–6, Apr. 2013.

[17]   K. Elleuch and A. Chaari, "Modelling and Estimation of Hammerstein System with Preload Nonlinearity," no. 16, pp. 13–20, 2010.

Chengpeng Wan

Resilience of transportation systems: literature review on its different dimensions

Chengpeng Wan
Liverpool Logistic Offshore & Marine Research Institute,
Liverpool John Moores University, Liverpool L3 3AF, Merseyside, UK.
E-mail address: c.wan@2015.ljmu.ac.uk

Abstract. As a crucial part in global supply chain and international logistics networks, a transportation system plays a key role in transporting cargos among different regions and countries. Safety is one of the issues with great importance in transportation research that should not be ignored. However, its foci have being expanded from traditional risk through security, and to resilience and sustainability. Resilience, which was initially introduced as the ability of a system to resist disturbances, maintain functions and recover from disruptions, has attracted considerable interests from both academics and industries in recent years. Various studies have been carried out on transportation resilience from different perspectives. Consequently, different definitions have been developed to describe resilience. This paper presents a systematic review of the literature relating to transportation resilience with emphases on different dimensions of it. It aims to figure out what kind of essential characters it should have. This paper will provide comprehensive insights into understanding the transportation resilience, as well as establish new horizons for relevant research topics.

1. Introduction

Since the start of systematic research on transportation resilience a decade ago, different terms have been used to measure the resilience and to describe its characteristics, including but not limited to vulnerability (e.g. Omer et al. 2012; Zhang et al. 2015), adaptability (e.g. Becker & Caldwell 2015), robustness (e.g. Blockley 2012), preparedness (e.g. Miller-Hooks et al. 2012), redundancy (e.g. Berle 2011), response (e.g. DiPietro et al. 2014) and recovery (e.g. Adams et al. 2012; Ashok & Banerjee 2014). It is quite often the case that the same term is explained from different perspectives to address the requirements in the development of specific resilience frameworks. However, there are few studies to analyse and derive the similarity and difference of the application of such terms in different transport sectors. In view of this, this paper will conduct a comprehensive literature review on transportation resilience from different dimensions, in order to provide reference for its further study.

2. Review of dimensions of transportation systems

2.1 Vulnerability

Vulnerability analysis has been increasingly important in network related transport studies. It has been evidenced by many previous research findings in various transport segments such as road networks, railway networks, maritime transportation systems (e.g. Mansouri et al. 2009), as well as those from a higher level, such as comprehensive transportation systems or integrated supply chain networks (e.g. Klibi & Martel 2012). Furthermore, emphases on definitions of vulnerability vary with respect to its specific application context. In a study of resilient infrastructure, vulnerability was defined as the susceptibility to damage or perturbation – especially where small damage or perturbation leads to disproportionate consequences (Blockley et al. 2012). From a mathematic perspective, vulnerability can be measured as a combination of likelihood of an event and its consequences (Pettit et al. 2010;

Faturechi & Miller-Hooks 2013). In the risk management of maritime supply chains (Berle et al. 2011), vulnerability was regarded as the property of a transportation system that may weaken or constrain its ability to endure, handle and survive threats and disruptive events that originated both within and outside the system boundaries. The concept of vulnerability has also been applied in analysis of port resilience from both infrastructural and operational aspects. Hsieh et al. (2014) developed fourteen vulnerable factors of ports from four dimensions, and conducted a case study of four international commercial ports in Taiwan. The results revealed that the port vulnerability was importantly affected by capacity and efficiency. Other research relating to the evaluation of port vulnerability can also be found in, such as Berle et al. (2011), Becker et al. (2014), Nursey-Bray et al. (2013), Becker & Caldwell (2015) and Mansouri et al. (2010).

2.2 Adaptability

Adaptability (also known as adaptive capacity) is defined as one of the functions of a resilient system, which reflects its flexible ability to response to new pressures (Bhamra et al. 2011). Similar definitions are presented by Pettit et al. (2010) as the ability to modify operations in response to challenges or opportunities. It has also been encompassed in the definition of resilient countries as timely adaptation in response to a changing environment. Such definitions indicate that its main features lies in response to changes to reflect the dynamic nature of complex systems. Azadeh et al. (2014) claimed that a supply chain with adaptability should have an ability to properly perform in all periods, that is, prior to, during and after an event. Regarding the applications of adaptability in road transportation, five methods were summarised by Rendall et al. (2011), including modifying travel time to avoid network peaks, changing trips, changing fuel types, shifting to a more efficient mode, and changing destinations.

Besides, it should be noted that adaptability and vulnerability have been considered in pairs in a few studies (e.g. Wang 2015). Omer et al. (2012) considered vulnerability and adaptability as two main elements that need to be addressed in order to build the resilience of a system. Furthermore, to make an infrastructure system less susceptible to disruptions, several schemes that can reduce the vulnerability of a system were proposed, including redundancy, diversity, hardening, capacity tolerance and modularity. While, other schemes, such as proper resources allocation, contingency preparation, collaboration and system cognition, were put forward to increase the adaptive capacity of the system, hence it can recover from disruptions and return back to its original service or close to it. A more conceptual study by Wang (2015) provided evidence to support the idea that both vulnerability and adaptability had linkages to resilience in many different ways, and viewed them as two key characteristics, reflecting positive and negative sides of system resilience, respectively. Based on that, the concept of "comprehensive resilience in transportation" was further proposed referring to the quality that led to "recovery, reliability and sustainability" (Wang 2015).

2.3 Flexibility

The flexibility of a system represents its ability to respond to shocks (Cox et al. 2011) and adjust itself to changes through contingency planning after disruptions (Faturechi & Miller-Hooks 2014b). It is also referred to as an ability to reconfigure resources (Berle et al. 2013) as well as to cope with uncertainties. Due to its property to adapt to changing circumstances and demands (Chen & Miller-Hooks 2012), it has been considered as the same as adaptability in some research (e.g. Faturechi & Miller-Hooks 2014) in terms of measuring performance of a system. As such, connotations of flexibility are opposite to that of robustness which emphasises the ability to endure these changes rather than adapt to them (Faturechi & Miller-Hooks 2014a). As one of the twenty-four resilience strategies in supply chain management summarised by Tukamuhabwa et al. (2015), increasing flexibility has being repeatedly discussed in recent decades. In this context, flexibility is understood as "the ability of a firm and supply chain to adapt to changing requirements with minimum time and effort" (Tukamuhabwa et al. 2015). In another research, the concept of flexibility applied in supply chains in terms of mitigation strategy are divided into system, process and product flexibility (Ivanov et al. 2014). Similar ideas are those that describe the flexibility of

transportation networks as node, link and temporal flexibility. Regarding its applications in transportation fields, Ishfaq (2012) built the flexibility in transportation operations through the cooperation of diversifying modes of transportation. Cox et al. (2011) proposed a series of metrics to evaluate the resilience of passenger transportation systems on the basis of the flexibility and other resilience features.

2.4 Reliability

Being a crucial parameter of resilience (e.g. Wang 2015; Baroud et al. 2014), reliability is generally defined as the probability that a network remains operative given the occurrence of a disaster or disruption event (Faturechi & Miller-Hooks 2014b). In this way, reliability to some extent is a metric for measuring the post-disaster performance of a system. According to Berdica (2002), reliability can be viewed as the complement of vulnerability. Vulnerability accounts for potential losses or degradations, while reliability considers the functionality left after a disruption (Edrissi et al. 2015). However, according to Barker et al. (2013), reliability decides a system's performance during the time period before the strike of an external disturbance, in which, it becomes a kind of characteristics describing the pre-disruption performance of a resilient system, and it is able to provide a baseline for the performance of service when the system operates at the stable state (Baroud et al. 2014). Various reliability measures have been applied in transportation systems to reflect certain functions for different purposes, such as connectivity reliability, travel-time reliability, capacity reliability and emergency response reliability (Edrissi et al. 2015). One of the well-studied topics of reliability lays on the determination of the component influencing the performance of the whole system most (Barker et al. 2013).

2.5 Robustness

According to Faturechi & Miller-Hooks (2014a), robustness is a measure of strength and can be seen as a complement to vulnerability, which is similar to the reliability (Faturechi & Miller-Hooks 2014a). However, in terms of construction of the conceptual framework for resilience, Steen & Aven (2011) considered robustness from a risk perspective as a two-dimensional combination of consequences and associated uncertainties, given the occurrence of an initiating event. In this context, it became an antonym of the vulnerability. Similar ideas can be found in the predictive measure taxonomy given by Cox et al. (2011). Robustness is the property of being strong, healthy and hardy (Blockley et al. 2012).Thus, it is generally defined as the ability to withstand or absorb disturbances and remain intact when exposed to disruptions (Faturechi & Miller-Hooks 2014b). It is among the four elements that contribute to the resilience of infrastructure systems (Bruneau et al. 2003). Based on the similar ideas, a R4 framework is proposed to describe determinants of a transportation system's resilience. Robustness is also one of the aspects that should be covered during the vulnerability analysis of road networks. In a recent study (Blockley et al. 2012), resilience and robustness are thought to have some similarities, and the authors point out that robustness is necessary but not sufficient for resilience. No matter it is in the four-dimensional resilient system proposed by the Multidisciplinary Center for Earthquake Engineering Research (MCEER), in the three-component based Resiliency Index defined by the National Infrastructure Advisory Council, or in a recent study (Chang et al. 2014) which estimates the resilience of city infrastructure systems from only two aspects, robustness has always been involved as one of the significant components of resilience.

It should be noted that the above introduced elements, that is, vulnerability, adaptability, flexibility, reliability and robustness, as well as resilience itself, are also considered as several commonly used indicators for performance measuring when analysing disasters' impacts on transportation systems (Faturechi & Miller-Hooks 2014b).

2.6 Recoverability

Recoverability (or the ability to recover) has been discussed the most in the research on transportation resilience. Baroud et al. (2014) considered resilience to be a time-dependent proportional measure, and calculated its recoverability according to the speed at which a system recovered. Similar applications of this factor as one of the metrics in quantitatively measuring the transportation resilience can also be found in other research work (e.g. Baroud et al. 2014b; Pant et al. 2014; Baroud et al. 2015). However, instead of using the

concept of recoverability, other research has been conducted regarding the ability to recover. It has been regarded to as a feature of secure and highly functioning transport networks. Definitions of resilience with emphases on the ability to recover can be seen in following research. Chen & Miller-Hooks (2012) held the opinion that resilience is an ability depending on network structure and those activities that can preserve or restore service when disruptive events occur. The authors thus carried out some comparative studies on four typical network structures, as well as studies of their performance with and without recovery activities involved. The results highlighted the importance of recovery activities in enhancing the recovery ability of transportation networks.

2.7 Redundancy

Redundancy indicates the ability of certain components of a system to assume the functions of failed components without adversely affecting the performance of the system itself (Haimes 2009). In brief, it reflects an availability of alternative choices (Tukamuhabwa et al. 2015), through which parallel systems can be utilised to provide alternative operations in case of fails of the original system (Omer et al. 2012). Due to the fact that redundancy usually involves duplication of capacities in order to continue operations during failures, it can thus be considered as a way to increasing system flexibility through the adaptable deployment of resources. In transportation fields, redundancy is also viewed as the existence of optional routes between origins and destinations, which can help to mitigate adverse impacts of disasters to a transportation network. It is commonly understood that a system with more redundancy has a relative higher resilience (Fiksel 2003), leading to its longer term development (O'Kelly 2015). The redundancy of routes is of great significance especially under emergency situations, but at the same time, over-pursue of redundancy will inevitably lead to an exorbitant cost. Thus a tradeoff between the system resilience and operation cost is necessary before design or strategy development of transportation networks. Although Sheffi (2007) regards building redundancy as one of the most straightforward approaches to create resilience, however, a recent study by Kim et al. (2015) reveals that the overall resilience of a network may not be improved through only establishing redundancy in nodes or arcs of a network, which conflicts with some of conventional views on network resilience.

2.8 Survivability

Survivability is generally defined as the ability to withstand sudden disturbances to functionality while meeting original demands (Faturechi & Miller-Hooks 2014b), and it is also one of the four components of resilience highlighted by Baroud et al. (2014b). Survivability techniques have been considered as an access to mitigating the vulnerability of a network system (e.g. Baroud et al. 2014; Barker et al. 2013). Thus survivability approaches can help to reduce the adverse impacts on a system from unexpected disruptive events. An early study claimed that the survivability of a network can be enhanced by extra connections between nodes, and similar ideas have been applied in a recent study (Ip & Wang 2011) to measure the resilience of a node within a network. Although telecommunication networks have been recognised as a main filed in terms of the application of survivability measures (Faturechi & Miller-Hooks 2014b), there are still some other research have that been conducted to explore models for measuring survivability with special attention being paid on transportation networks.

2.9 Preparedness

Preparedness refers to "prepare certain measures before disruption happens" (Jin et al. 2014), with the object to enhance the resilience of a system by lessening potential negative impacts from disruptive events. It can be subdivided as emergency preparedness and response preparedness, favoured by different industrial sectors (Berle et al. 2011). Preparedness is the second step belonging to the pre-disruption stage in the framework of disaster management (which is mitigation, preparedness, response, and recovery) proposed by Altay and Green (2006). Preparedness activities such as system hardening can mitigate the initial impacts on the system, so as to reduce its vulnerability (Barker et al. 2013). In addition, preparedness strategies like training response teams, support a more efficient response, which contributes to post-disaster stages as well (Faturechi & Miller-Hooks 2014b).

Preparedness has been widely applied in the area of stochastic programming modelling when measuring the resilience of a transportation system. Faturechi & Miller-Hooks (2014b) developed a three-stage Stochastic Mathematical Program to measure and optimise the resilience of roadway networks considering the effect of preparedness. Other studies of system resilience incorporating preparedness actions into transportation network modelling process can be found in, for example, Zhang et al. (2015) and Miller-Hooks et al. (2012).

2.10 Resourcefulness

Resourcefulness has appeared in various literature for different use. It is defined as "the availability of materials, supplies, and crews to restore functionality" in a study of resilience of transportation systems (Adams et al. 2012). In another research relating to transport security policy, resilience is discussed in terms of its applications at the economic level. In that context, resourcefulness is treated as one of stabilizing measures in terms of resilience (Reggiani 2013). Similar ideas can be found in the Global Risks (2013) report, where resourcefulness is regarded as a composition of the resilience (Dias 2015). In addition, it is involved as one of the elements compromising the Resiliency Index proposed by the National Infrastructure Advisory Council (Omer et al. 2012), as well as an element of the R4 framework defined by Bruneau et al. (2003). In a review work of resilience analysis of engineered and infrastructure systems, Francis & Bekera (2014) define resourcefulness as the level of preparedness in effectively resisting an adverse event.

2.11 Responsiveness

In a study of supply chain networks (Klibi et al. 2010), responsiveness is described as the capability of a supply chain system to respond positively to disturbances, and the development of this capability can be either redundancy or flexibility based. Generally, responsiveness policies are developed to provide a sufficient response to short-term variations in process of both supply and demand. They provide a barrier against threats and risks, so as to increase the expected value of supply chain networks (Klibi et al. 2010). Thus, responsiveness is an important factor to the resilience of supply networks (Klibi & Martel 2012). However, like redundancy, responsiveness factors of a supply chain system may also increase the costs while improving its service level (Ivanov et al. 2014). From a systemic perspective, responsiveness of a system can be considered from two distinctive aspects, that is, before and after a disruption, which are also referred to as phases of prevention and recovery (Mansouri et al. 2010).

2.12 Rapidity

Rapidity is a well-studied concept in the "resilience triangle" that has been applied in civil infrastructures for decades, with an emphasis on the speed to recover. It belongs to one of the properties in the abovementioned R4 framework (Bruneau et al., 2003), the performance of which depends on other elements within the same framework, which are robustness, redundancy, and resourcefulness. Rapidity is defined as "the capacity to restore functionality in a timely manner by identifying and mobilizing material, monetary, informational, and human resources" (Adams et al. 2012). Similar definitions of rapidity can be found in the research of disaster management as "the capacity to restore functionality in a timely way, containing losses and avoiding disruptions", which has an impact on the duration of reduced system performance (Dorbritz 2011).

References

Adams, T.M., Bekkem, K.R. & Toledo-Durán, E.J., 2012. Freight Resilience Measures. *Journal of Transportation Engineering*, 138(11), pp.1403–1409.

Altay, N., and W. G. Green. 2006. OR/MS Research in Disaster Operations Management. *European Journal of Operational Research,* 175, pp. 475-493.

Ashok, V. & Banerjee, S., 2014. Enhancing resilience of highway bridges through seismic retrofit Ashok. *Earthquake Engineering & Structural Dynamics*, 43, pp.1173-1191.

Azadeh, a. et al., 2014. Modelling and improvement of supply chain with imprecise transportation delays and resilience factors. *International Journal of Logistics Research and Applications*, 17(4), pp.269-282.

Barker, K., Ramirez-Marquez, J.E. & Rocco, C.M., 2013. Resilience-based network component importance measures. *Reliability Engineering & System Safety*, 117, pp.89–97.

Baroud, H., Barker, K., et al., 2014. Importance measures for inland waterway network resilience. *Transportation Research Part E: Logistics and Transportation Review,* 62, pp.55–67.

Baroud, H. et al., 2015. Inherent Costs and Interdependent Impacts of Infrastructure Network Resilience. *Risk Analysis*, 35(4), pp.642–662.

Baroud, H., Ramirez-Marquez, J.E., et al., 2014. Stochastic Measures of Network Resilience: Applications to Waterway Commodity Flows. *Risk Analysis*, 34(7), pp.1317–1335.

Becker, A. et al., 2014. Towards seaport resilience for climate change adaptation: Stakeholder perceptions of hurricane impacts in Gulfport (MS) and Providence (RI). *Progress in Planning*, 99, pp.1–49.

Becker, A. & Caldwell, M.R., 2015. Stakeholder Perceptions of Seaport Resilience Strategies: A Case Study of Gulfport (Mississippi) and Providence (Rhode Island). *Coastal Management*, 43(1), pp.1–34.

Berle, Ø. Norstad, I. & Asbjørnslett, B.E., 2013. Optimization, risk assessment and resilience in LNG transportation systems. *Supply Chain Management: An International Journal*, 18(3), pp.253–264.

Berle, Ø. Rice Jr., J.B. & Asbjørnslett, B.E., 2011. Failure modes in the maritime transportation system: a functional approach to throughput vulnerability. *Maritime Policy & Management,* 38(6), pp.605–632.

Bhamra, R., Dani, S. & Burnard, K., 2011. Resilience: the concept, a literature review and future directions. *International Journal of Production Research*, 49(18), pp.5375–5393.

Blockley, D., Godfrey, P. & Agarwal, J., 2012. Infrastructure resilience for high-impact low-chance risks. *Proceedings of the ICE - Civil Engineering*, 165(CE6), pp.13–19.

Chang, S.E. et al., 2014. Toward Disaster-Resilient Cities: Characterizing Resilience of Infrastructure Systems with Expert Judgments. *Risk Analysis*, 34(3), pp.416–434.

Chen, L. & Miller-Hooks, E., 2012. Resilience: An Indicator of Recovery Capability in Intermodal Freight Transport. *Transportation Science*, 46(1), pp.109–123.

Cox, A., Prager, F. & Rose, A., 2011. Transportation security and the role of resilience: A foundation for operational metrics. *Transport Policy*, 18(2), pp.307–317.

Dias, P., 2015. Is toughness a better metaphor than resilience? *Civil Engineering and Environmental Systems*, 32(1-2), pp.68–76.

DiPietro, G.S., Scott Matthews, H. & Hendrickson, C.T., 2014. Estimating economic and resilience consequences of potential navigation infrastructure failures: A case study of the Monongahela River. *Transportation Research Part A: Policy and Practice*, 69, pp.142–164.

Dorbritz, R., 2011. Assessing the resilience of transportation systems in case of large-scale disastrous events. In Proceedings of the 8th International Conference on Environmental Engineering. pp. 1070–1076.

Edrissi, A., Nourinejad, M. & Roorda, M.J., 2015. Transportation network reliability in emergency response. *Transportation Research Part E: Logistics and Transportation Review*, 80, pp.56–73.

Faturechi, R. & Miller-Hooks, E., 2014a. A Mathematical Framework for Quantifying and Optimizing Protective Actions for Civil Infrastructure Systems. *Computer-Aided Civil and Infrastructure Engineering*, 29, pp.572–589.

Faturechi, R. & Miller-Hooks, E., 2014b. Measuring the Performance of Transportation Infrastructure Systems in Disasters : A Comprehensive Review. *ASCE Journal of Infrastructure Systems*, 21(1), pp.1–15.

Faturechi, R. & Miller-Hooks, E., 2014c. Travel time resilience of roadway networks under disaster. *Transportation Research Part B: Methodological*, 70, pp.47–64.

Francis, R. & Bekera, B., 2014. A metric and frameworks for resilience analysis of engineered and infrastructure systems. *Reliability Engineering & System Safety*, 121, pp.90–103.

Haimes, Y.Y., 2009. On the Definition of Resilience in Systems. Risk Analysis, 29(4), pp.498–501.

Hsieh, C.-H., Tai, H.-H. & Lee, Y.-N., 2014. Port vulnerability assessment from the perspective of critical infrastructure interdependency. *Maritime Policy & Management*, 41(6), pp.589–606.

Ip, W.H. & Wang, D., 2011. Resilience and friability of transportation networks: Evaluation, analysis and optimization. *IEEE Systems Journal*, 5(2), pp.189–198.

Ishfaq, R., 2012. Resilience through flexibility in transportation operations. *International Journal of Logistics Research and Applications*, 15(4), pp.215–229.

Ivanov, D., Sokolov, B. & Dolgui, A., 2014. The Ripple effect in supply chains: trade-off "efficiency-flexibility-resilience" in disruption management. *International Journal of Production Research*, 52(7), pp.2154–2172.

Jin, J.G. et al., 2014. Enhancing metro network resilience via localized integration with bus services. *Transportation Research Part E: Logistics and Transportation Review*, 63, pp.17–30.

Kim, Y., Chen, Y.-S. & Linderman, K., 2015. Supply network disruption and resilience: A network structural perspective. *Journal of Operations Management*, 33-34, pp.43–59.

Klibi, W. & Martel, A., 2012. Modeling approaches for the design of resilient supply networks under disruptions. *International Journal of Production Economics*, 135(2), pp.882–898.

Klibi, W., Martel, A. & Guitouni, A., 2010. The design of robust value-creating supply chain networks: A critical review. *European Journal of Operational Research*, 203(2), pp.283–293.

Mansouri, M., Nilchiani, R. & Mostashari, A., 2010. A policy making framework for resilient port infrastructure systems. *Marine Policy*, 34(6), pp.1125–1134.

Mansouri, M., Sauser, B. & Boardman Dr., J., 2009. Applications of Systems Thinking for Resilience Study in Maritime Transportation System of Systems. In IEEE SysCon 2009 -3rd Annual IEEE International Systems Conference, 2009.

Miller-Hooks, E., Zhang, X. & Faturechi, R., 2012. Measuring and maximizing resilience of freight transportation networks. *Computers & Operations Research*, 39(7), pp.1633–1643.

Nursey-Bray, M. et al., 2013. Vulnerabilities and adaptation of ports to climate change. *Journal of Environmental Planning and Management*, 56(June 2015), pp.1021–1045.

O'Kelly, M.E., 2015. Network Hub Structure and Resilience. *Networks and Spatial Economics*, 15(2), pp.235–251.

Omer, M. et al., 2012. A framework for assessing resiliency of maritime transportation systems. *Maritime Policy & Management*, 39(7), pp.685–703.

Pant, R. et al., 2014. Stochastic measures of resilience and their application to container terminals. *Computers & Industrial Engineering*, 70, pp.183–194.

Pettit, T.J., Fiksel, J. & Croxton, K.L., 2010. Ensuring Supply Chain Resilience: Development of a Conceptual Framework. *Journal of Business Logistics*, 31(1), pp.1–21.

Reggiani, A., 2013. Network resilience for transport security: Some methodological considerations. *Transport Policy*, 28, pp.63–68.

Rendall, S. et al., 2011. Quantifying Transport Energy Resilience. *Transportation Research Record: Journal of the Transportation Research Board*, 2242, pp.72–80.

Shafieezadeh, A. & Ivey Burden, L., 2014. Scenario-based resilience assessment framework for critical infrastructure systems: Case study for seismic resilience of seaports. *Reliability Engineering & System Safety*, 132, pp.207–219.

Steen, R. & Aven, T., 2011. A risk perspective suitable for resilience engineering. *Safety Science*, 49(2), pp.292–297.

Tukamuhabwa, B.R. et al., 2015. Supply chain resilience: definition, review and theoretical foundations for further study. *International Journal of Production Research*, 7543(May), pp.1–32.

Wang, J.Y.T., 2015. "Resilience thinking" in transport planning. *Civil Engineering and Environmental Systems*, 32(1-2), pp.180–191.

Zhang, X., Miller-Hooks, E. & Denny, K., 2015. Assessing the role of network topology in transportation network, pp.33–62.

Minho Ha

A hybrid approach to the modelling of port performance measurement

Min-Ho Ha
Liverpool Logistics Offshore and Marine Research Institute (LOOM)
m.ha@2012.ljmu.ac.uk

Abstract. This study develops a new conceptual port performance measurement model by taking the perspectives from different port stakeholders. The novelty lies in its capability of dealing with interdependency among the performance measures as well as accommodating both qualitative and quantitative evaluations on the measures simultaneously. This is achieved by introducing a hybrid approach in order to deal with the inherent data uncertainties and the interdependencies among the port performance indicators (PPIs). An analysis of four major container ports in South Korea is conducted to demonstrate and validate the proposed framework. The empirical results indicate that the hybrid approach attempting to use quantitative modelling for dealing with the uncertainties and interdependency problems can be successfully implemented. The hybrid model represents an effective performance measurement tool and offers a diagnostic instrument to ports/terminals performance evaluation and/or monitoring so as to satisfy different requirements of various groups of port stakeholders in a flexible manner.

1. Introduction

Modern container ports are part of complex systems operating in an uncertain logistics environment. They also are places where a number of port stakeholders provide products and services and create value together. The interests of different port stakeholders, i.e., port authorities, port users, service providers and related communities, in economic, social, and environmental issues, are sometimes in conflict (Notteboom and Winkelmans, 2003). Port authorities increasingly rely on stakeholder management practices to secure long-term relations with key stakeholders (Dooms and Verbeke, 2007). Performance measurement has become an important tool in stakeholder management, while at the same time the challenging multi-stakeholder environment complicates port performance measurement. Little research has been done on addressing the multi-stakeholder dimension in port performance measurement and the interdependency of port performance indicators (PPIs) in a more quantitative way. Also, the existing literature tends to focus on limited dimensions or specific areas of ports and terminals. Such fragmented approaches may fail to take into account new issues and challenges faced by ports.

This study argues that there is a need for a new performance measurement framework not just to meet the needs of port stakeholders, but also to enrich the diagnostic tools available to support decision-making in complex port/terminal systems operating in an uncertain environment. This framework needs to involve multiple dimensions with both quantitative and qualitative port performance indicators (PPIs) in order to offer diagnostic instruments to decision makers. The decisions are usually made on multiple uncertain attributes. Consequently, this study needs to deal with the inherent data uncertainties which are sometimes unavoidable in port/terminal operational contexts. Furthermore, it needs to identify interdependencies among the PPIs. Given complex port activities and operations, decision makers may require an essential understanding of the interdependency among the PPIs and develop appropriate solutions to improve port/terminal performance. In order to tackle these problems, this study proposes a new conceptual PPI measurement model using a hybrid approach of a fuzzy logic based evidential reasoning (FER) method (Yang and Xu, 2002), a decision making trial and evaluation laboratory (DEMATEL) tool (Gabus and Fontela, 1973) and an analytic network process (ANP) technique (Saaty, 1996). An analysis of four major container ports in South Korea is conducted to demonstrate and validate the proposed framework.

In the next section, the research framework is presented and a case study of Korean container port performance measurement is conducted to demonstrate and validate the proposed framework. Finally, the paper concludes with a discussion of results and recommendation for further research in section 3.

2. A Hybrid Methodology for Port Performance Measurement

Step 1: Identifying PPIs, setting assessment grades to each PPI and collecting data. The PPIs crucially needed for measuring port performance were identified. Previous studies suggest that port performance should cover the range of port activities to cope with new evolutionary changes (Woo et al., 2011). In addition, the PPIs should allow the ports to measure and communicate their impacts on society, economy and environment (ESPO, 2010) and to be consistent with their goals (Kaplan and Norton, 2004). The selection of PPIs has been done through literature review and industrial practices in a pre-selection phase and then confirmed by a panel of ten experts[7] to assess the suitability of the potential indicators and to test the feasibility of the selected indicators (Okoli and Pawlowski, 2004). Through the content validation, 60 PPIs in the lowest level are particularly significant as representing indicators for container port performance measurement under 16 principal-PPIs and 6 dimensions. Next, assessment grades are allocated to all PPIs. For assessing a qualitative PPI, different sets of linguistic terms such as {very low, low, medium, high, very high} for "commitment and loyalty of port employees" were defined by domain experts (Yang, 2001). If PPI is of a quantitative nature, it can be assessed using numerical grades (Yang, 2001) based on various data (i.e. consulting reports, journal papers and internal data of terminal operators). From this perspective, a set of quantitative grades, for example, {≤0%, 5%, 10%, 15%, 20%, ≥25%} for "throughput growth" are developed based on a list of the top 50 world container ports.

The PPIs include various types of numeric and subjective data to reflect the complexity of port/terminal business environments. The quantitative data (i.e. CA and FS) were collected directly from terminal operating companies and information systems/databases managed by port authorities, government and credit rating agencies. The qualitative PPIs were collected using questionnaire results obtained from three groups of terminal operators (TO), users (i.e. shipping lines and freight forwarders, PU) and administrators (i.e. port authority and government, AD) to assess their own associated PPIs to measure each port's performance. The detailed responses of the survey are listed in
Table **1**.

Table 1. Response details

	Busan North			Gwangyang			Incheon			Busan New		
	TO	PU	AD	TO	PU	AD	TO	PU	AD	TO	PU	AD
Total distributed	100	200	40	75	200	40	75	200	40	125	200	40
Total received	2 (2)	38 (31)	0	40 (40)	26 (26)	10	0	15 (11)	0	4 (4)	38 (31)	0
Online received	30 (29)	20 (12)	9 (6)	0	5 (3)	0	41 (39)	26 (17)	11 (6)	26 (24)	20 (12)	9 (6)
Usable response	(31)	(43(2))	(6)	(40)	(29 (8))	(10)	(39)	(28)	(6)	(28)	(43(2))	(6)
Used for analysis	(31)	(127)	(18)	(40)	(85)	(30)	(39)	(84)	(18)	(28)	(213)	(18)
Judgement on:	SA, TSCI SSS, EVS	US, TSCI	SG	SA, TSCI, SSS, EVS	US, TSCI	SG	SA, TSCI, SSS, EVS	US, TSCI	SG	SA, TSCI, SSS, EVS	US, TSCI	SG

---

[7] The group included 6 industrial experts who have been working in shipping and port industry for more than 15 years with PhD (1 expert from a shipping line), MSc (3 experts from terminal operators, a shipping line and a forwarder) and BA (1 from a terminal operator and a forwarder, respectively) degrees participated in the judgements. 2 professors who have more than 15 years teaching and research experience participated in the survey. Lastly, 2 experts from government/port authority (1 department manager and 1 managing director) who have been working for port logistics departments participated in the survey.

Step 2: Evaluate each port/terminal based on the lowest PPIs.

Each PPI can be assessed using a belief degree represented by judgements (Yang, 2001). The judgements can be presented by degrees of belief (DoB) which belong to either linguistic terms (for the qualitative PPIs) or numerical values (for the quantitative PPIs). The former can be obtained by expert judgements and the latter needs to be calculated through various location measurement techniques (Yang *et al.*, 2009). From this perspective, both the DoB belonging to linguistic terms and the numerical grades are calculated by applying different methods. The degree of belief associated to numerical grades using a location measurement technique can be transformed using Eq. (1). Any quantitative number $h_{j,i}$ (with an evaluation grade H_j) is evaluated between $h_{j-1,i}$ (with an evaluation grade H_{j-1}) and $h_{j+1,i}$ (with an evaluation grade H_{j+1}).

$$If \; h_{j-1,i} < h_{j,i} < h_{j+1,i} \;\; then \; B_{j+1,i} = \frac{h_{j,i} - h_{j-1,i}}{h_{j+1,i} - h_{j-1,i}}, B_{j-1,i} = 1 - h_{j+1,i} \qquad (1)$$

where $B_{j+1,i}$ represents the degree of belief associated quantitative number with the grade H_{j+1} and $B_{j-1,i}$ represents the degree of belief associated quantitative number with the grade H_{j-1}.

For example, a set of quantitative grades $H = \{\text{leq } 0\%(H_1), 5\%(H_2), 10\%(H_3), 15\%(H_4), 20\%(H_5), geq \; 25\%(H_6)\}$ for "throughput growth" is already defined in the second step. If throughput growth, for example, is 2%, throughput growth set is assessed as follows:

$H^{TG} = \{(\text{less than } 0\%, 0.6), (5\%, 0.4), (10\%, 0), (15\%, 0), (20\%, 0), (more \; than \; 25\%, 0)\}$

Step 3: Mapping process –Transform the evaluation from the lowest level PPIs to top level PPI

PPIs for port performance measurement include different numbers and linguistic terms of assessment grades in the lower-level PPIs and the associated upper-level PPI. The defined grades, thus, need to be interpreted and transformed into a unified format for assessment of the associated upper level PPIs (Yang *et al.*, 2009). This can be done using a fuzzy IF-THEN rule based belief structure. Yang (2001) developed the rule based utility techniques that can be easily applied for transforming qualitative and quantitative data, hence the techniques have already been proven by many scholars (Yang, 2001; Yang *et al.*, 2009). The core of this technique is a fuzzy mapping technique to transform fuzzy inputs to fuzzy outputs. As shown in Fig., I^i ($\sum_{i=1}^n I^i \leq 1$) indicates the fuzzy input associated with a lower-level PPI and O^j ($O^j = \sum_{i=1}^n I^i \beta_i^j$) represents the fuzzy output transformed from I^i. B_i^j ($\sum_{j=1}^n \beta_i^j = 1$) denotes the DoBs assigned by experts for presenting the relationship between assessment grades of different levels. For example, the upper level principal-PPI "output (OPC)" can be expressed using linguistic terms as "very low (OPC1)", "low (OPC2)", "medium (OPC3)", "high (OPC4)" and "very high (OPC5)". The numerical grades used to assess the lowest level PPI "throughput growth (TG)" can be expressed "leq 0% (TG1)", "5% (TG2)", "10% (TG3), "15% (TG4)", "20% (TG5)" and geq 25% (TG6). It is noteworthy that a throughput growth of the "leq 0%" means that the output is said to be equivalent to a grade "very low" using fuzzy rules. Based on R^5 and R^6, 2% throughput growth can be transformed into (30% OPC2 ($O^2 = 0.4 \times 0.75$) and 10% OPC1 ($O^1 = 0.4 \times 0.25$)) and (60% OPC1 ($O^1 = 0.6 \times 1$)) respectively. It can be equally described as 70% OPC1 and 30% OPC2 (Fig. 1). This mapping process can be conducted from the lowest level PPIs to the top level goal in a similar manner.
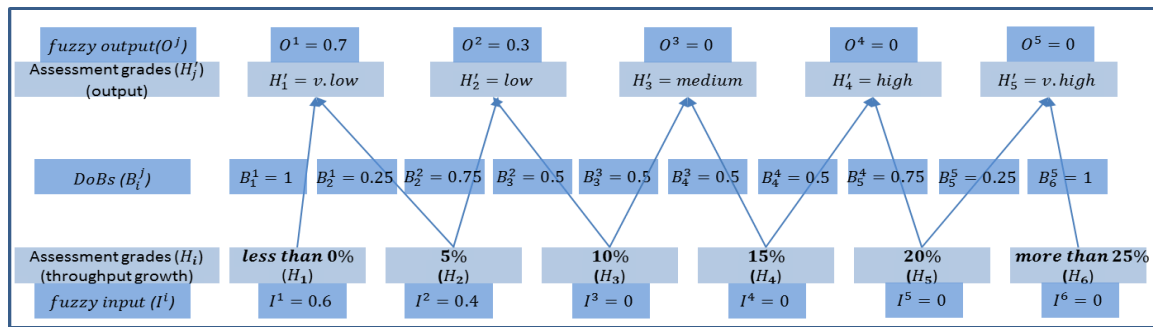
Fig.1. Fuzzy Mapping Process

Step 4: Identify PPIs interdependency and evaluate their weights using DEMATEL and ANP. The same panel of the ten experts were asked to determine the interdependency among six categories. The initial direct-relation 6×6 $matrix\ Z$ is obtained by pairwise comparisons in terms of influences and directions. Then, the normalised direct-relation matrix D is calculated and the total-relation matrix T and sum of influence given and received by each category are obtained (Table 2). A threshold value of 0.82 (=29.63/36) is calculated in order to determine direct relationships between the 6 dimensions. Based on the direct influence matrix, the experts[8] were asked to determine the interdependency among the principal-PPIs. The same process is carried out to obtain a direct influence matrix for principal-PPIs. A threshold value of 0.11 (=28.99/256) is calculated. Only the PPIs whose influence values in each cell are higher than the threshold value can be chosen and converted into a digraph (

Table **3**). Then, the ANP method is used to obtain the adjusted weights of the 16 principal-PPIs. Based on

Table **3**, the experts[9] are asked to respond questions, for example, "which principal-PPI influences on 'productivity' more: 'output' or 'lead-time'? and how much more?" In terms of this process, a number of comparison matrices can be formed, and an unweighted super-matrix can be obtained. By calculating the limiting power of the weighted super-matrix a limit super-matrix can be generated (

Table **4**). The results in the limit super-matrix can be used as weights of sixteen principal-PPIs. The final step is to obtain local weights of 60 PPIs. The local weights of 60 PPIs obtained by AHP are shown in

Table **5**. Using an integrated method of DEMATEL and ANP, this study identified PPIs' interdependency and their relative importance in all hierarchical levels. The next step is to synthesise the evaluations of each terminal against all PPIs together with their weights using intelligent decision system (IDS; Yang and Xu, 2000) incorporating the ER algorithm.

Table 2. The total influence matrix

	CA	SA	FS	US	TSCI	SG	pr_i^+	pr_i^-
CA	**0.88**	**1.01**	**0.97**	**1.08**	**0.92**	**0.93**	11.07	0.53
SA	**0.94**	0.77	**0.87**	**0.93**	**0.84**	**0.86**	10.32	0.07
FS	**0.85**	**0.84**	0.65	0.80	0.73	**0.84**	9.50	(0.10)
US	**0.92**	**0.89**	0.83	0.74	0.80	0.80	10.08	(0.12)
TSCI	**0.85**	**0.83**	0.78	**0.83**	0.61	0.75	9.22	0.08
SG	**0.83**	0.79	0.70	0.73	0.67	0.59	9.09	(0.46)

Table 3. The total influence matrix of principal PPIs

		CA			SA			FS		US		TSCI		SG			
		OPC	PDC	LTC	HCS	OCS	ICS	PFF	LSF	SFU	SCU	ITST	VAST	ICIT	SSS	EVS	SES
CA	OPC	0.17	0.24	0.23	0.17	0.18	0.19	0.18	0.16	0.19	0.19	0.14	0.12	0.14	0.15	0.13	0.13
	PDC	0.26	0.18	0.26	0.19	0.19	0.20	0.19	0.16	0.20	0.20	0.15	0.13	0.15	0.14	0.14	0.13
	LTC	0.25	0.26	0.18	0.18	0.18	0.20	0.18	0.16	0.22	0.19	0.15	0.12	0.16	0.15	0.14	0.12
SA	HCS	0.21	0.22	0.21	0.10	0.10	0.11	**0.15**	**0.14**	0.17	0.16	0.11	0.12	0.13	0.14	0.12	0.12
	OCS	0.21	0.21	0.20	0.10	0.10	0.11	**0.16**	0.13	0.18	0.15	0.12	0.11	**0.13**	0.13	0.12	0.12

[8] 8 experts (2 terminal operators, 1 shipping line, 1 forwarder, 2 academics and 2 government representatives) from among the 10 experts.
[9] 4 experts (1 terminal operator, 1 shipping line, 1 forwarder, 1 academic) from among the 10 experts.

	ICS	**0.23**	**0.23**	**0.24**	0.11	0.11	**0.12**	**0.16**	**0.14**	**0.18**	**0.16**	**0.13**	**0.12**	**0.14**	**0.14**	**0.12**	0.11
FS	PFF	**0.14**	**0.14**	**0.14**	0.11	0.11	0.11	0.06	0.05	0.07	0.06	0.05	0.04	0.05	0.11	0.10	**0.13**
	LSF	**0.14**	**0.13**	**0.13**	0.11	0.11	**0.11**	0.06	0.05	0.07	0.06	0.05	0.04	0.05	0.11	**0.12**	0.11
US	SFU	**0.20**	**0.20**	**0.19**	**0.15**	**0.16**	**0.16**	**0.15**	**0.14**	0.10	0.09	0.07	0.06	0.07	0.08	0.07	0.07
	SCU	**0.18**	**0.18**	**0.17**	0.11	0.11	**0.12**	**0.15**	**0.14**	0.08	0.07	0.05	0.05	0.06	0.07	0.06	0.06
TSCI	ITST	**0.19**	**0.18**	**0.19**	**0.13**	**0.13**	**0.16**	0.09	0.08	**0.16**	**0.13**	0.06	0.06	0.07	0.07	0.06	0.06
	VAST	**0.17**	**0.16**	**0.16**	**0.12**	**0.12**	**0.13**	0.08	0.07	**0.15**	**0.14**	0.05	0.05	0.06	0.06	0.05	0.05
	ICIT	**0.20**	**0.20**	**0.20**	**0.15**	**0.15**	**0.18**	0.10	0.09	**0.17**	**0.15**	0.07	0.06	0.07	0.07	0.07	0.06
SG	SSS	0.11	0.11	**0.12**	0.04	0.04	0.04	0.04	0.03	0.04	0.04	0.03	0.02	0.03	0.03	0.03	0.03
	EVS	0.07	0.08	0.07	0.02	0.02	0.03	0.02	0.02	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02
	SES	0.06	0.06	0.06	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.01

Table 4. Limited super matrix

	CA			SA			FS		US		TSCI			SG		
	OPC	PDC	LTC	HCS	OCS	ICS	PFF	LSF	SFU	SCU	ITST	VAST	ICIT	SSS	EVS	SES
	0.12	0.14	0.12	0.06	0.05	0.10	0.08	0.06	0.10	0.05	0.03	0.02	0.02	0.02	0.02	0.02
Ranking	(2)	(1)	(3)	(8)	(9)	(5)	(6)	(7)	(4)	(10)	(11)	(15)	(12)	(13)	(16)	(14)

Table 5. Local weights for 60 PPIs

PPIs	OPC1	OPC2	PDC1	PDC2	PDC3	PDC4	PDC5	PDC6	LTC1	LTC2	LTC3	HCS1	HCS2	HCS3	HCS4
LW	0.696	0.304	0.158	0.132	0.107	0.345	0.103	0.155	0.602	0.185	0.213	0.246	0.243	0.354	0.157
PPIs	OCS1	OCS2	OCS3	OCS4	ICS1	ICS2	ICS3	PFF1	PFF2	PFF3	LSF1	LSF2	LSF3	SFU1	SFU2
LW	0.175	0.296	0.198	0.330	0.364	0.301	0.335	0.318	0.328	0.354	0.342	0.349	0.309	0.361	0.147
PPIs	SFU3	SFU4	SFU5	SCU1	SCU2	SCU3	ITST1	ITST2	ITST3	ITST4	VAST1	VAST2	VAST3	VAST4	ICST1
LW	0.134	0.188	0.170	0.549	0.315	0.137	0.466	0.159	0.197	0.178	0.369	0.172	0.197	0.262	0.291
PPIs	ICST2	ICST3	ICST4	SSS1	SSS2	SSS3	SSS4	EVS1	EVS2	EVS3	EVS4	EVS5	SES1	SES2	SES3
LW	0.261	0.232	0.216	0.298	0.206	0.231	0.265	0.158	0.145	0.248	0.149	0.300	0.578	0.272	0.150

Step 5: Synthesise DoBs and weights of PPIs using ER algorithm.

The transformed results from the lowest level PPIs to the top level goal and their interdependent weights can be synthesised by IDS incorporating the ER algorithm and utility technique in a hierarchical decision structure. The Window-based tool, IDS, facilitates the process of making decisions from collecting information to building up a model, defining alternatives and criteria and different assessments (Yang and Xu, 2000). As shown in
Table **6**-
Table **8**, the performance scores of each port can be easily compared and decision makers can straightforwardly identify their strengths and weaknesses. These benchmarking results provide an important contribution for decision makers to enhance their port/terminal performance based on any necessary comparisons. Furthermore, it can be used for a longitudinal study to investigate the improvement of ports within different timeframes.

Table 6. Performance score on 16 principal-PPIs

16 principal-PPIs	Busan North	Gwangyang	Incheon	Busan New	Ranking
Output (OPC)	0	0.2360	0.2979	0.6718	BN>I>G>B
Productivity (PDC)	0.5759	0.2835	0.3714	0.5029	B> BN>I>G
Lead time (LTC)	0.9835	0.9615	0.9627	0.9947	BN>B>I>G
Human capital (HCS)	0.6642	0.6287	0.7943	0.7255	I> BN>B>G
Organisation capital (OCS)	0.7223	0.6770	0.7627	0.7522	I> BN>B>G
Information capital (ICS)	0.7363	0.6453	0.7552	0.6978	I>B> BN>G
Profitability (PFF)	0	0.2287	0.6101	0.6849	BN>I>G>B
Liquidity & Solvency (LSF)	0.6569	0.5847	0.9466	0.3364	I>B>G>BN
Service fulfilment (SFU)	0.6738	0.7504	0.6493	0.7186	G> BN>B>I
Service costs (SCU)	0.6010	0.6214	0.5970	0.5648	G>B>I>BN
Intermodal transport systems (ITST)	0.6780	0.6865	0.6928	0.7289	BN>I>G>B
Value-added services (VAST)	0.6312	0.6702	0.5940	0.6951	BN>G>B>I
Information/communication integration (ICIT)	0.6839	0.6921	0.6961	0.7495	BN>I>G>B
Safety and security (SSS)	0.8875	0.9020	0.8818	0.9549	BN>G>I>B
Environment (EVS)	0.6514	0.4537	0.5110	0.7635	BN>B>I>G
Social engagement (SES)	0.6583	0.4654	0.4647	0.7318	BN>B>G>I

Table 7. Performance score on 6 dimensions

6 dimensions	Busan North	Gwangyang	Incheon	Busan New	Ranking
Core activities (CA)	0.5313	0.4716	0.5274	0.7146	BN>B>I>G
Supporting activities (SA)	0.7305	0.6601	0.7848	0.7306	I>BN>B>G
Financial strength (FS)	0.2432	0.3488	0.7707	0.5527	I>BN>G>B
User satisfaction (US)	0.6667	0.7347	0.6458	0.6973	G>BN>B>I
Terminal supply chain Integration (TSCI)	0.6822	0.6987	0.6858	0.7442	BN> G>I>B
Sustainable growth (SG)	0.7744	0.6633	0.6705	0.8580	BN>B>I>G

Table 8. Performance score of each port

Ports	Performance	Ranking index	Ranking
Busan North	VP 0.23; P 0.1; M 0.03; G 0.22; VG 0.42	0.61	4
Gwangyang	VP 0.21; P 0.14; M 0.03; G 0.21; VG 0.40; UK 0.01	0.61	3
Incheon	VP 0.11; P 0.14; M 0.04; G 0.22; VG 0.48; UK 0.01	0.70	2
Busan New	VP 0.10; P 0.11; M 0.04; G 0.25; VG 0.51	0.74	1

Note: 1) VP, very poor; P, poor; M, medium; G, good; VG, very good; UK, unknown.
2) UK has arisen due to unavailable quantitative data.

3. Discussion and conclusion

Previous studies on port performance, port selection and port competitiveness generally treat PPIs as independent from each other and focus mainly on sea-side operations only. Moreover, existing studies typically lack a structured approach to performance measurement in a multi-stakeholder environment. In this regard, we introduced a new framework based on the combination of DEMATEL and ANP and incorporating the ER algorithm and utility technique to capture the interdependency among PPIs and to incorporate multiple objectives of key stakeholders. This study identifies the overall PPIs with respect to different stakeholders as well as the weights of the interdependency PPIs and synthesise the evaluations of quantitative and qualitative PPIs with their weights through an IDS tool.

The result suggests that Busan New Port shows the best results, followed by Incheon Port. Busan North Port is assessed to be the least competitive port with the lowest performance especially in terms of output and profitability. Busan New Port outperforms the other ports in terms of output, lead-time, profitability, intermodal transport systems, etc. but is less competitive at the level of two principal-PPIs such as liquidity & solvency and service costs. This is because the five terminal operators in Busan New Port started up operations from 2005 to 2011 respectively. The rather recent heavy initial capital spending for port superstructure, state-of-the-art systems and equipment and the capital is generally raised from financial institutions and investors through project finances. With regard to the service costs, the adjacent port, Busan North Port, lowered its service price to secure its market share from the moment Busan New Port started operations (source from interview with terminal operators in Busan Port). The 'lower price' strategy is the more preferential strategy than others when port operators adjust themselves in a changing business environment characterised by intense port competition. Derived from the results of interdependency among the six dimensions, core activities are affected by all dimensions and itself (inner dependency) as well as affect all dimensions including itself. This directly influences the interdependency among the 16 principal-PPIs. Amongst 16 principal-PPIs, productivity, output and lead-time in core activities are found to be the most influential whilst environment is the least influential principal-PPI (

Table **3**). A plausible explanation would be that in the context of the container port industry, container throughput, berth-yard operation, mode turnaround time, labour productivity, competency of human resource and information technology are important criteria for port performance measurement. However, being cost and price competitive is crucial but not sufficient for port performance measurement. This finding is in line with the general argument in port selection/ competitiveness research that a shipping line is likely to choose a port due to the port's cargo generation and hinterland connectivity (Yeo *et al.,* 2008). Ports should not only take into account internal competency of core and supporting activities, but

also be aware of the tangible and intangible integration with stakeholders to sustain themselves in a highly competitive environment. The results yielded by the hybrid approach present the ranking of the ports in terms of their overall performance with respect to multiple PPIs as well as a PPI's ranking with a single performance value. This feature provides both port operators and port authorities with valuable insights as this framework allows them (1) to recognise current strengths and weaknesses of each port; (2) to better understand the conditions and status of their competitive ports; (3) to prioritise investment to improve competitiveness and customers' satisfaction by adjusting their strategies based on the relative importance of PPIs.

Nevertheless, further studies for identifying factor correlation and result validity are to be conducted. Based on the research findings, further empirical study to benchmark port performance in different regions/areas and for different timeframes can be carried out to identify the best practices/solutions of the leading performers in view of an improvement of weaker PPIs.

6. References

Dooms, M., Verbeke. A., 2007. Stakeholder Management in Ports: A Conceptual Framework Integrating Insights from Research in Strategy, Corporate Social Responsibility and Port Management. International Association of Maritime Economists (IAME), July 4–6, Athens, Greece.

ESPO, 2010. Port Performance Indicators Selection and Measurement Work Package 1(WP1): Pre-Selection of an initial set of indicators.

Gabus, A., Fontela, E., 1973. Perceptions of the World Problematique: Communication Procedure, Communicating With Those Bearing Collective Responsibility. DEMATEL Report No. 1, Battelle Geneva Research Centre, Geneva, Switzerland.

Kaplan, R.S., Norton, D.P., 2004. Measuring the strategic readiness of intangible assets. Harvard business review 82(2), 52-63.

Notteboom, T., Winkelmans. W., 2003. Dealing with Stakeholders in the Port Planning Process. In: Dullaert, W. and Jourquin, B. (eds.), Across the Border: Building Upon a Quarter of Century of Transport Research in the Benelux, Antwerp: De Boeck, 249–265.

Okoli, C., and Pawlowski, S.D. (2004). The Delphi method as a research tool: an example, design considerations and applications. Information & management, 42(1), 15-29.

Saaty, T.L., 1996. Decision making with dependence and feedback, the analytic network process. Pittsburgh: RWS Publication.

Woo, S.-H., Pettit, S., Beresford, A.K., 2011. Port evolution and performance in changing logistics environments. Maritime Economics & Logistics 13(3), 250-277.

Yang, J.-B., 2001. Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainties. European Journal of Operational Research. 131(1), 31-61.

Yang, J., Xu. D., 2000. An Introduction to an Intelligent Decision System: IDS.' Proceedings of the 42nd Annual Conference of the UK Operational Research Society, September 12–14, Swansea, UK.

Yang, J.-B., Xu, D.-L., 2002. On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 32(3), 289-304.

Yang, Z., Bonsall, S., Wang J., 2009. Use of hybrid multiple uncertain attribute decision making techniques in safety management. Expert System with Applications 36(2), 1569-1586.

Yeo, G.T., Roe, M., Dinwoodie, J., 2008. Evaluating the competitiveness of container ports in Korea and China. Transportation Research Part A: Policy and Practice 42(6), 910-921.

Hani Al Yami

Risk Management in Maritime Container Port

Hani Alyami, Zaili Yang, Ramin Riahi, Stephen Bonsall & Jin Wang
Liverpool Logistics Offshore and Marine (LOOM) Research Institute
H.M.Al-Yami@2010.ljmu.ac.uk

Abstract. Sea ports and maritime terminals are essential economic elements of any coastal State, as a result, research and subsequent improvements in related areas including operational, organisational, economical, business, and natural conditions that affect seaports and marine terminals have been occurring for many years. Therefore, safety aspects in these areas need to be addressed appropriately with an effective risk management methodology.

Introduction

Safety performance in maritime port plays an important role in enhancing port operation reliability, maritime transportation safety and supply chain distribution resilience. However, the task is not straightforward given the challenges, including that risks in seaports operations need to be appropriately addressed to ensure economic wealth, operational efficiency, personnel safety, and terminals security. Port operations safety affected by multiple factors related to design, installation, operation and maintenance that need to be addressed and evaluated with effective risk management methodology. This paper describing the most widely applied risk based methodologies by various industries and briefly introducing the application of suitable method on container terminal operations safety. The proposed approach can also be tailored for wider application in other engineering and management systems, especially when instant risk ranking is required by the stakeholders to measure, predict and improve their system safety and reliability performance.

Maritime Port and Container Terminal Port Safety Research

Safe and adequate port operation has a profound impact on service quality, productivity cost, and lifestyle. Therefore, system safety evaluation including the early detection of hazards is critical in avoiding performance degradation and damage to human life or machinery. The effect of accidents and/or disasters that jeopardise terminal operations can be eliminated or reduced, if a robust forecasting mechanism is developed and effectively enforced (Hu *et al.,* 2010). Safety of seaports is playing an increasingly important role in ensuring supply chain resilience and sustainability, thus attracting some risk related research (Legato and Monaco, 2004; Marlo and Casaca, 2003; Trbojevic and Carr, 2000; Fabiano *et al.,* 2010; Mokhtari *et al.,* 2011; Soares and Teixeira, 2001) from different operational, organizational and economic perspectives. A review by Pallis *et al.,* (2010) on 395 port-related journal papers published between 1997 and 2008 disclosed that, despite the criticality of safety and security in efficient supply chains and international trade, risk analysis persistently occupied a backseat role within port research being overwhelmed by other aspects involving efficiency analysis, port competition, geographical analysis and spatial evolution, policy and governance, to name but a few. In addition, a review of 984 papers published in Maritime Policy & Management (MPM) by Notteboom *et al.,* (2013) reveals that a core theme in seaport studies over the past 40 years of its existence discloses that research in ports has

evolved its research frameworks and techniques in many fields such as geography, econometrics, welfare economics, operations research, logistics/supply chain management and strategic management. In the last 5 years the themes of ports in transport and supply chains, port governance and port competition and competitiveness were dominating port research. On the other hand, in the first two decades of MPM, regulatory issues referring to competition, pricing, financing, environmental, safety and security related policy practices were research themes attracting much attention. Since then, port terminals including container terminals have been developed rabidly and aggressively, creating a growing interest in examining the prospects and limits of safety aspects in such growth and complex activities in port operations.

Maritime infrastructure such as container terminals is critical and costly engineering systems that enable economic activities through the transfer of goods and services between national and international destinations. Given the significance of their operations to the world economy, they face a variety of operational and environmental uncertainties that make them vulnerable to hazards related to transport safety, shipping efficiency; supplies chain distribution reliability, and loss prevention (Mansouri *et al.,* 2009). Many researchers and some leading international bodies such as the International Maritime Organization (IMO), United Nations Conferences on Trade and Development (UNCTAD), World Bank, European Commission, the Asian Development Bank, and Academic Bodies have been the main parties having strong tendencies to carry out research concerning maritime ports and terminals aiming to enhance the efficiency and effectiveness. Hence, safety issues are of significant importance to the maritime ports including terminals. However, the research on safety and security analysis of container terminal operations in a quantitative way has not yet been well conducted and the issue has been recently raised in Yang, Ng, and Wang (2014). If the risk result cannot be quantified, it may well not motivate the industrial stakeholders to take control measures confidently. Many researchers have carried out studies related to the business for marine ports and terminals including reviews of the development of seaports via case studies to enhance the efficiency (Goss, 1990; UNCTAD, 1996; Palmer, 1999; World Bank, 2001). The increasing number of publications in the last decade indicates the importance of the aforementioned areas of research, which may have been sufficient. However, dynamic and enforced changes that occur in marine ports and offshore terminals operational, organisational, economic, business, and natural factors related environments, makes it imperative to look into safety aspects to include the recently emergent risk management related issues with respect to externally and internally driven elements.

Container Terminal Port Safety related Existing Methodologies

Globalization of trade has led to a rapid increase in container vessel's movements in many seaports in which most busy seaports, as this trade continues to grow, will face many risk factors related to economic wealth, operational efficiency, personnel safety, and terminals security. While significant academic effort is devoted to port centred logistics and port operational optimization (e.g., Liu *et al.,* 2002; Vis and Koster*,* 2003; Steenken *et al.,* 2004; Güntheret *et al.,* 2006), there are relatively few studies on port safety and risk (e.g., Yip *et al.,* 2002; Yang *et al.,* 2010), revealing a research gap to be fulfilled. In a broader scope, few paid attentions to the performance reliability of port's transportation system. The beginning of container transportation initiated five decades ago, however, while containerization has developed rapidly since then from ships and ports to patterns of global trade, its impact on safety analysis has not been explored properly in scientific research (Fabiano *et al.,* 2009).

Safety has been an issue form the beginning of shipping industry for hundreds of years and most likely it will be, as the shipping industry growth continues. Tracking the industry history, the need for measures implementation that can control the risks are often recognised after catastrophic accidents. For instance, the Titanic disaster in 1912 led to the first International Conference on Safety of Life at Sea (SOLAS). The United States introduced to the International Safety Conference, the concept that ship safety should be measured as the

extent of damage a ship in 1960 after the capsizing of the liner Andrea Doria. The International Convention for the Prevention of Pollution (MARPOL) from Ships during the 1970s due to several oil tanker accidents. The Exxon Valdez disaster in 1990 resulted in the use of double hull tankers mandated. All aforementioned incidents and the subsequent acts from member states or international bodies indicate the continual necessity for introducing modern risk management methods in the maritime industry.

In respect of risk based methodologies, a systematic concern with risk based methodology began in the aerospace sector following the fire of the Apollo test AS-204 in 1967 as the basic methods of probabilistic risk assessment. The nuclear industry developed probabilistic safety assessments in the 1970s, introducing the first full scale application of this method, including an extensive analysis of the accident consequences which was undertaken in the reactor safety study WASH-1400 (Jensen, 2002). The chemical industry used quantitative risk assessment (QRA) in the 1970s which was developed and applied in the 1980s by the offshore industry in Norway after the Alexander accident (Bai and Jin, 2016). The development of Safety case in the UK after the Piper Alpha accident in 1988 (HSE, 1992). Moreover, after Lord Carver's report on the investigation of the capsize of the Herald of Free Enterprise published in 1992, UK proposed a particular type of risk management framework for the safety regime to the IMO in 1993 that was referred to as the Formal Safety Assessment (FSA). Various industries have been performing risk based methodologies either quantitative or qualitative or integration of both. Eventually, all are processing in three main categories, to identify, evaluate and mitigate the impact of risk factors that are threatening the industry. As risk-based methodologies it may look quite similar in the procedures. However, the process on conducting each step, as well as the techniques and tools used, may be different (EMSA, 2012).

1.1. Quantitative Risk Analysis (QRA)

QRA also called Probabilistic risk analysis (PRA) or probabilistic safety analysis have been adopted as part of the schematic framework by relevant authorities. In other sectors the analytic QRA methodology is intensively applied to ensure claims for safety or to determine the need for further improvement by providing numerical evidence (Haigh, 2003). Essentially, QRA provides answers to the following three questions (Farquharson and McDuffee, 2003):

- What can happen? Or what are the initiators or initiating events (i.e. undesirable starting events) that lead to adverse consequences?
- How likely is it to happen? Or what are their frequencies?
- Given that it occurs, what are the consequences? Or what and how severe are the potential determinants?

The identification of the possible hazards, the quantification of the frequencies that the hazard materializes and, finally, the severity of the consequences are the parameters that define risk (Tim and Roger, 2001).

1.2. Safety Case

A safety case merely aims to ensure an adequate level of safety for a particular facility, based upon the management and control of the risks associated with that facility. A central feature of a safety case is that the designers and operators (i.e. duty holders) of a facility takes responsibility for assessing the risks associated with their facility, and for documenting how the safety management system limits those risks to an acceptable risk level (Sutton, 2014b). The key elements of the safety case concepts are described as follows (Wang, 2002; Sutton, 2014b).

1. Hazards identification is the first step.
2. Risk estimation for all hazards once they have been identified to determine the associated risks. All hazards can generally be classified in accordance with the ALARP principle (HSE, 1992).
3. Risk reduction is required to reduce the risks associated with significant hazards and the mitigation measures are in place should a major incident occur.

4. Risk Control to take the most appropriate action in the event that a hazard becomes a reality to minimize its effects on operators, facility and the environment.

5. Develop and implement a Safety Management System (SMS) to ensure that the SMS meets the pertinent health and safety rules and regulations. The effectiveness of the SMS system is usually monitored and verified by means of regular audits, and compliance with the requirements of the safety case checked by means of inspections.

1.3. Formal Safety Assessment (FSA)

The FSA is a tool for decision makers to make the decision process more rational with a view to achieving a balance between maritime safety or protection of the marine environment and costs (IMO, 2002). This type of framework is due to the unique features of the maritime industry as there is no single regulator; no single culture; and no uniformed education and qualification system that exists in the global maritime industry. The functional components in a FSA process are listed as follows:

1. Hazards identification;
2. Risk analysis;
3. Risk control options;
4. Cost benefit assessment; and
5. Recommendations for decision making.

The adoption of FSA for shipping represents a fundamental cultural change, from a largely reactive approach, to a structured and systematic methodology integrated, proactive and effective based on the evaluation of risk (Pillay and Wang, 2003). The main features of FSA are a formalized procedure, an audible process, communicated safety objectives, and priorities based on cost effectiveness (Bai and Jin, 2016). In a complex systems and operational situations such as container terminal, quantitative and global assessments only, are not able to cope with all the technical and operational specifics, have not only created the need for qualitative risk assessments but also for systematic and transparent risk management approaches.

Risk Management Application on Container Terminal Port

The public pressure on one side and the pressure from the scientific community have raised the attention of the maritime industry to the need of risk based methodology. In order to distinguish from the earlier risk management approach introduced in this industry for many years, FSA is the term which is most often used today to describe this type of management. Despite the fact that other industries use different terms (i.e. QRA, Safety Case, and FSA), the main features of any risk based methodology are the same. Independently of the names adopted by the industries, the risk assessment procedure consists of similar phases that identify, evaluate and mitigate the impact of risk factors affecting an industry and addressing those three phases in different risk based processes. However, FSA has some advantages than other methodologies as follows (Bai and Jin, 2016):

- Hazards are identified proactively through the hazard identification process with reference to many different hazard identification approaches may be applied.
- Risks associated with various hazards are described and analysed. The risk, conventionally, is a composite of the likelihood and consequences of the potential undesirable events arising from a hazard, but other risk attributes can be assigned to the conventional risk composition in order to improve the system safety taking into account the possible uncertainty involved. The risk analysis covers a certain time span, that is, the operational life, and may involve various quantitative or qualitative tools to perform risk composition calculations.
- The cost-benefit analysis may be followed to compare the costs for preventive/protective measures with the benefits.
- The basic elements are integrated into a risk model with information about the hazards, their associated risks and the cost effectiveness of alternative risk control

options is provided, where the objective is to recommend the most cost-effective, preventive, and mitigating measures for risk management strategy.

1.4. Risk Management Application Process

1. Identify hazardous events (HEs) (i.e. failure modes) in container operational system.
HEs can be identified through quantitative or qualitative methods in hierarchal structure. However, depending on the terminal characteristics (such as: operations complexity, activity magnitude, work force, and equipment) being evaluated and the resources accessible, the process used to conduct HEs identification can differ.

2. Risk assessment for each HEs at the bottom level individually
Due to possible uncertainty involved, failure inputs can be evaluated by using defined linguistic grades with degree of belief (DoBs) which means that multiple IF-THEN rules will be employed in risk evaluation for a particular HE, requiring an appropriate tool capable of synthesizing the associated DoBs in the THEN parts of different involved rules. Probabilistic graphical methods can be used for encoding a complete distribution over a multi-dimensional attributes such as Bayesian networks and Markov networks.

3. Risk assessment for the HEs globally
After obtaining each HEs evaluation (i.e. linguistic grades with DoBs) at the bottom level, it can be further analysed for the safety estimate of a whole container operational system and the risk impact analysis of each HE on the whole system by using each HEs evaluation obtained. Therefore, it requires a tool that has the capability of synthesising evaluations in a hierarchical structure such as Evidential Reasoning method that shows a potential in aggregation process for multi-levels of hierarchy structure.

4. Risk control options
Once the risks have been evaluated, next phase is to identify the options available for controlling the risks. These can range from doing nothing to introducing risk control options in order to eliminate the cause of the risk altogether, or to reduce it to secure certain benefits to people, environment and property of the system with confidence that the risk is one that is worth taking and that it is being properly controlled. There are two methods for controlling risk, namely:
•   Reactive approach: to reduce the severity of the failure.
•   Proactive approach: to reduce the frequency of an initiating event
A reactive approach or a mitigating option approach refers to risk mitigation actions initiated after risk events occur and can be seen as initiation of contingency plans. On the other hand, a proactive approach or a preventive approach refers to actions initiated based on chance of a risk event occurring such as marine insurance domain that have been using risk mitigation options strategies extensively in the marine industry (Kartam and Kartam, 2001; Bai and Jin, 2016). A combination of these two approaches is also applied to risk management to reduce the likelihood of risk, reduce the risk consequences, and can be extends further to, for example, risk avoidance, risk transfer, or to retain the risk (RMS, 2005). The basic aim is risk reduction and the key test is one of reasonable practicability; that is, the cost to reduce the risk further is grossly disproportionate to the bene-fit gained. This will involve cost benefit analysis.

5. Cost benefit assessment
The risk management objective in this phase is the avoidance of accidents and this can be measured by evaluating the avoidance of harm to people, damage to property, environment, and other costs. To achieve a balance, the benefits of a risk control action must be considered and compared to the implementation cost which can be accomplished through a cost benefit analysis (Bai and Jin, 2016). The risk control measure/s costs are estimated taking into account both enforcement cost such as personal work force including inspection, auditing and, installation, and regulatory costs cost such as capital cost and, compliance

cost (EUC, 2013). Similar to costs estimation process is undertaken to estimate benefits, the net worth value of each option can be calculated by deducting the benefits from the costs. Sensitivity analysis can be carried out to estimate the level of confidence that can be attached to the calculated net worth value of each option, then, risk control options can be ranked based on their cost effectiveness (HSE, 2001; Ahmed *et al.,* 2007). There are several methods exist for cost benefit analysis, usually referred to as Multi-Criteria Decision Making (MCDM) techniques. Among these techniques and the widely used in academic research are TOPSIS (Kahraman *et al.,* 2007), Analytic Hierarchy Process (AHP) (Golec and Taskin, 2007), Analytic Network Process (ANP) (Yuksel and Dagdeviren, 2007), ELECTRE (Wang and Triantaphyllou, 2005), PROME-THEE (Dagdeviren, 2008), and Axiomatic Design (AD) (Kulak and Kahraman, 2005). There is no single correct technique for a particular decision problem but some techniques are more suitable than others (Mergias *et al.,* 2007).

6. Recommendations for decision making
Finally, based on the above analytical process the result to take appropriate risk measures and select the suitable advised recommendations are given to decision maker.

7. An Integrated container port system risk analysis
The complexity of handling a large amount of data dealing with two different methodologies for HEs risk assessment either individually or globally with reference to their software, would burden the stakeholders by going through all that not user-friendly processes to measure, predict, and improve their system safety and reliability performance. In addition, due to the optimization required for the risk assessment processes (i.e. the inputs and outputs for individual or global HEs risk assessments) with evaluations tasks cause a tremendous amount of calculations, however, it can be overcome by applying a computational modelling tool that capable of simulate such processes and provide accurate results as the original. Many types of technologies can be introduced to develop such models. However, numerous applications of Artificial Neural Networks (ANNs) have been successfully applied for real time risk prediction on various sectors in the past decade, due to their reliable, robust, and salient characteristics in capturing the non-linear relationships between variables (i.e. multi-input/output) in complex systems. It has many egregious functions such as adaptive learning, real time operation and self-organization, thinking and reasoning, judging and memory, and fault tolerance (Widrow *et al.,* 1994; Kumar and Ravi, 2007). ANN approach. integrate

Conclusion

Technological effort and development in risk management have resulted in strengthening safety standards by mitigating the risks driven HEs which is evidenced by the practical applications and research over the years. In order to achieve an optimum evaluation, control, and safety management performance in maritime port operation, new concepts involving efficient operation and design should be developed and implemented. High quality representative models in terms of accuracy and reliability assurance can provide a favourable solution in the risk evaluation process that helps to predict the risk magnitude, explain the real safety performance, and develop a continuous risk management strategy for complex systems. Therefore, this paper focused on the risk management methodology for container terminal port on operational aspects including technical and personal factors. Other risk concerns influencing container terminal port safety such as managerial, policy implications, natural and political issues can similarly be investigated in order to provide a panoramic view on terminal risk analysis. In addition, the proposed method highlights its potential in facilitating risk analysis of system design and operations in a wide context when being appropriately tailored to study other seaports.

Muhammad Irfan Nawaz

The requirement for new generation surveyors in the offshore industry

M Irfan Nawaz
Liverpool Logistics Offshore and Marine Research Institute, James Parson Building, Byrom Street.
E-mail address: M.I.Nawaz@2012.ljmu.ac.uk

Abstract.

The tragic accident of Piper Alpha in 1988 has forced the offshore industry to redefine the regulatory requirements and the role of classification societies. Traditionally, classification societies rely on the competency and experience of the workforce (surveyors) to carry out inspections. Present-day classification societies' surveyors act as an Independent Competent Person (ICP) under the new proactive safety case regime. However, little is known about the competency requirements of ICP and various studies confirm that operators/owners of the offshore installations also lack understanding of their role. This paper, therefore, highlights the importance of new generation surveyors, to complete a transition from prescriptive regime to a proactive regime. This is the first study that critically examines the role of ICP and its significance. This paper begins with the research of a current understanding of the ICP role; it then goes on to develop a hypothesis by identifying the improvement areas and proposing solutions. Finally, the hypothesis is tested by a survey of ICPs from two different classification societies. This aids in developing understanding and identifies key factors that should be considered to complete the transition to a proactive regime. The practical implications are explored and guidelines are synthesized for ICPs.

Keywords: *Keywords:* Safety case regime; Independent Competent Person; Classification societies

List of Acronyms and Abbreviations

DHs: Duty Holders
HSE: Health Safety Executive
ICPs: Independent Competent Persons
IVBs: Independent Competent Body
LR: Lloyd's Register
MAH: Major Accident Hazards
SCEs: Safety Critical Elements
SCR: Safety Case Regulations
WSVs: Written Schemes of Verification

Introduction

The Health and Safety Executive (HSE) carried out a comprehensive study of asset integrity on UK offshore assets (HSE, 2007). The findings and recommendations of the study developed a perception that verification was not delivering the benefits (Oil & Gas UK, 2012). Before, evaluating the effectiveness of the verification process, it is worth mentioning the brief background and role of Independent Verification Bodies (IVBs) in this process. Classification societies were appointed by the Secretary of State for Energy in viewing their independence, competency, experience in certification work, specialist resource in the United Kingdom and abroad, and recognition in the industry (Saebo, 1991). Classification societies are non-governmental organizations and originally formed by insurers to access the quality of ships. Present-day classification societies' surveyors act as an Independent Competent Persons (ICPs) under the Safety Case Regulations (SCR) 2005. Traditionally, classification societies rely on the competency and experience of the workforce (surveyors) to carry out inspections. However, little is known about the competency requirements of ICPs and various studies confirm that operators/owners of the offshore installations also lack understanding of their role.

It is the responsibility of duty holders to select the independent competent person (HSE, 2005). ICPs can be selected through three different channels, certification authorities in a new role of IVBs, in house team, and independent competent person. Different ICPs can also be chosen at the same time. Competence is judged prior to selection and independence is shown by management hierarchy within the system. There is no specific way to define or express competence in quantitative way, although adequate competence and experience is desired. Competencies include technical expertise, knowledge, and experience to carry out tasks related to verification activity. Further, it is desired that adequate number and suitable persons are also available to direct and harmonize verification activities (LR, 1996).

Verification is not the change of logo of certification. It covers much broader area than certification. Under certification hardware, integrity was checked and it has its own limitation such as its ability to thoroughly address hazards. Whereas, verification process draw up verification scheme that not only directs to check the integrity of hardware but also ensures hazards are managed and performance is achieved as intended (LR, 1996). This suggests that in-order to implement the verification process effectively verification scheme should be clear and precise. It is the responsibility of Duty Holders (DHs) to chalk out the verification scheme with the Independent Competent Person (ICP). Verification scheme is considered suitable if it ensures that safety critical elements such as emergency shutdown system, lifeboats, helideck, active firefighting facilities etc., remain in good repair and condition (HSE, 2006). The first serious discussion and analysis of different facets of the verification regime was conducted by Wang et al. (2012), which strengthened the concept of risk-based verification. The authors argue that available schemes are identical; thus, they proposed a risk-based verification framework. The identical Written Schemes of Verification (WSVs) schemes contradict the objectives of the proactive regime, where it is recognised that each installation is different and system based verification scheme is desired. Safety

case regulations also emphasise the proactive role of duty holders and ICPs activities are a mere confirmation of the effectiveness of the process. ICPs activities will not only verify that SCEs are healthy, but also give assurance that the system is able to foresee the development of Major Accident Hazards (MAHs). Otherwise, there will not be a notable difference between a classification regime and a verification regime, where surveyors had a leading role, and owners had a lagging role. Based on literature review, the hypothesis is developed that written scheme are unclear or vague, therefore exiting WSVs cannot achieve the objectives of proactive verification process. The hypothesis is tested by a survey of ICPs from two different classification societies. This aids in developing understanding and identifies key factors that should be considered to complete the transition to a proactive regime. The practical implications are explored and guidelines are synthesized for ICPs.

Demographics Characteristics of the participants

Thirty questionnaires were distributed among ICPs from two different classification societies, 25 were returned, which shows a response rate of 83%. High response rate was achieved due to fact that the questionnaires were mostly handed over face-to-face during the planned meetings.

The demographic characteristics of the respondents such as age, education, years of experience, role of ICPs were asked in the questionnaire. Demographic characteristics of the participants (Table 1) show that:

a) Out of 25, 40% (n=10) of participants were aged between 31-40, 44% (n=11) of participants were aged between 41-50, 16% (n=4) of participants were aged between 51-60. The result reflects that ICPs are generally in a middle age group.
b) In terms of educational levels, about 60% (n=15) participants have a bachelor's degree, 32% (n=8) participants have a diploma, 8% (n=2) participants have a master's degree.
   The implication is that majority of the respondents are educated. This result is understandable and covers the different levels of education. However, it can be argued that in order to achieve the recognisable-chartered status, masters degree is becoming a prerequisite.
c) The participants have different years of experience, about 36% (n=9) participants have one year or less experience, about 28% (n=7) participants have two to five years' experience, about 36% (n=9) participants have two to five years' experience. This is a good indication that IVBs utilises different level of experience in order to develop the expertise in the field.
d) The participants have different level of roles, about 8% (n=2) participants are assistant surveyors, about 40% (n=10) participants are surveyors, about 44% (n=11) are senior surveyors and about 8% (n=2) are team leads whose role is to manage and provide support to team of surveyors.

Table 1 Demographic Characteristic of Participants

Demographic	Category	Frequencies	Percentage
Age	20 or Less	0	0
	21- 30	0	0
	31- 40	10	40
	41- 50	11	44
	51- 60	4	16
Education	High School	0	0
	Diploma	8	32
	Bachelor	15	60
	Masters	2	8
	PhD	0	0
Years of Experience	1 or Less	9	36
	2- 5	7	28
	5 or More	9	36
Role of ICPs	Assistant Surveyor	2	8
	Surveyor	10	40
	Senior Surveyor	11	44
	Surveyors Team lead	2	8

Survey Results

The survey was based on ICPs experience with different duty holders and their attitude on performance standards and written scheme of verification. Results shown in table 2 suggest that about 56% (n=14) ICPs agree that performance standards and written scheme of verification are vague/unclear, about 32% (n=8) ICPs are neutral that performance standards and written scheme of verification are vague/unclear, about 12% (n=3) ICPs disagrees that performance standards and written scheme of verification are vague/unclear. The results validate the hypothesis of the study. In order to further understand, the route cause of the issue, the participants were questioned; do they consider that written schemes of verification use qualitative data? The results are shown in table three, which suggest that use of quantitative data shall contribute in achieving the measurable performance.

Table 2 ICPs view on vague and unclear PS/WSVs

	Frequency	Percent	Valid Percent
disagree	3	12.0	12.0
neutral	8	32.0	32.0
agree	14	56.0	56.0
Total	25	100.0	100.0

Table 3 ICPs view on quantitative data used in performance standards and written schemes of verification.

	Frequency	Percent	Valid Percent
disagree	4	16.0	16.0
neutral	18	72.0	72.0
agree	3	12.0	12.0
Total	25	100.0	100.0

In the event of failure or degradation of the SCEs, the risk assessments are carried out. The survey inquired ICPs on their involvement, and their knowledge of different risk assessment techniques. The below results shown in table 4 and 5 suggests that duty holders should involve ICPs in the process. In the same vein, IVBs should train their workforce on risk assessment techniques.

Table 4: ICPs involvement in risk assessment exercises

	Frequency	Percent	Valid Percent
strongly disagree	1	4.0	4.0
Disagree	19	76.0	76.0
Neutral	4	16.0	16.0
Agree	1	4.0	4.0
Total	25	100.0	100.0

Table 5: ICPs familiarisation with techniques such as HAZOP, HAZID, Bow tie models, Fault tree analysis, and Event tree analysis

	Frequency	Percent	Valid Percent
disagree	4	16.0	16.0
neutral	9	36.0	36.0
agree	12	48.0	48.0
Total	25	100.0	100.0

Discussion

Verification is a continuous process and in order to ensure that a benefit of proactive regime is achieved, all stakeholders should communicate effectively. Stakeholders such as duty holders and IVBs should work together to train their work force. IVBs can consider investing in continued education of ICPs. The results show that only two ICPs have a master's degree that can be improved. Although, this does not indicate that other ICPs are less qualified, they may have vast experience in the industry, which is always useful to make pragmatic decisions. Similarly, familiarisation with hazard analysis techniques is another area where support can be provided to ICPs that is outside of the traditional role of surveyors.

Performance standards and written schemes are drawn in consultation with the ICPs. Still, the results suggest that these documents are vague or unclear and inclusion of quantitative data can improve assurance of SCEs. Possible future work is to drawn up the proactive assurance framework that will have following features:

1. Monitor and advise on the health of SCEs.
2. Foresee SCEs failure and development of MAHs.
3. Provide valuable input and data for internal audit.
4. Improve data management and analysis.
5. Consider the expert judgement in the analysis.
6. The BN model will identify the area of improvement in the safety critical system.
7. Propose changes in the maintenance management system.
8. Assist in planning and resource allocation in high-risk areas to achieve proactive planning instead of reactive planning.

References

HSE, (2005). *Computer model for the design and validation of directed water deluge systems for the protection of plant containing pressurised flammable materials against fire*. Report Prepared by London South Bank University. Health and Safety Executive

HSE, (2006). *A guide to the Offshore Installations (Safety Case) Regulations 2005*. 3rd ed. Richmond: Health and Safety Executive

HSE, (2007). *Key Programme 3 - Asset Integrity Programme*. A report by the Offshore Division of HSE's Hazardous Installations Directorate. Health and Safety Executive

LR (1996), Experience with Verification Schemes under new UK Offshore Regulations, LR Technical Association

Oil & Gas UK, (2012). *Assurance & Verification Practitioner's Guide*. [online] Available at: https://www.stepchangeinsafety.net/safety-resources/publications/assurance-and-verification-practitioners-guide [Accessed 16 Feb. 2014].

Saebo E (1991) "The Certifying Authority after Cullen" Offshore Operations Post Piper Alpha, IMarE (C), Volume 103, Marine Management (Holdings) Ltd, ISBN 0-907206-39-5

Wang, J., Matellini, B., Wall, A. and Phipps, J. (2012). Risk-based verification of large offshore systems. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, 226(3), pp.273-298

Ran Wang

Benchmarking dynamic three-dimensional bin packing problems using discrete-event simulation

R Wang, T T Nguyen, S Kavakeb, Z Yang, and C Li
Liverpool Logistics Offshore and Marine Research Institute (LOOM),
School of Engineering, Technology and Maritime Operations,
Liverpool John Moores University, L3 3AF, United Kingdom
E-mail address: R.Wang@2015.ljmu.ac.uk

Abstract. In this paper a framework is developed to generate benchmark problems for dynamic three-dimensional (3D) bin packing problems (BPPs). This framework is able to generate benchmark problems for different variants of BPPs by taking into account potential uncertainty in real-world BPPs, which are uncertainties in dimensions, costs, weights of upcoming items. This paper has three main contributions. First, a benchmark generator framework is developed for the first time using an open source discrete-event simulation platform. This framework generates benchmark problems for BPPs by reproducing uncertainty in real-world BPPs. Second, this framework can be integrated with any dynamic BPP algorithm so that the optimisation algorithm can be run alongside the simulation to solve dynamic BPPs. Third; various performance measures from the literature are included in the framework to evaluate the optimisation algorithms from different perspectives. Thanks to the 3D visualisation feature of this framework, the optimisation results can also be observed visually. Finally, empirical experiments on a real-world BPP are conducted to verify these contributions.

Keywords. Benchmarking, bin packing problem, dynamic optimisation, simulation

1. Introduction

Three-dimensional (3D) bin packing problems (BPPs) are NP-hard optimisation problems in which a number of boxes are packed into one or multiple 3D bins. Depending on the characteristics of the problem, different objectives can be defined for BPPs such as input minimisation or output maximisation. The size of bins can be either identical [1-3] or varied [4-8]. The aim of output maximisation is maximising the volume or number of packed boxes given the limited number of bins [9-12]. Some BPPs have multiple objectives: minimising the cost while packing items with preferences [8]; minimising the cost and also packing items with the same destinations together [13]. Regarding constraints on BPPs, [14] introduced various constraints on potential containers, items, loading and allocation.

In the literature the BPPs are normally considered in an ideal situation with no uncertainty. In reality, however, uncertainty is a frequent feature of real-world BPPs. Hence, we need to take uncertainty into account in academic research. This paper is structured as follows. Section 2 reviews benchmark problems for BPPs, possible uncertainties in BPPs and the performance measures used in the literature. Section 3 describes the structure and functions of the proposed framework. It then proposes a group of test cases for BPPs. Section 4 carries out an experiment

for a real-world BPP using the proposed framework including a dynamic BPP algorithm. The experimental results are then provided. Section 5 concludes this paper.

2. Literature review

Real-world optimisation problems are subject to uncertainty and unknown changes. As changes occur in an optimisation problem, the dynamic optimisation algorithm need to react to the changes to produce a new optimal solution in regards to the changes. Due to the complexity of uncertainty in real-world problems, it is difficult to evaluate dynamic optimisation algorithms effectively. In academic research on dynamic benchmark problems, attempts are made to capture changes that reflect different characteristics of real problems. These benchmarks can then be used as a basis to evaluate algorithms. However, generating such benchmark problems is not trivial due to difficulty in capturing and reproducing uncertainty in real-world problems. This difficulty arises due to the lack of enough knowledge about uncertainty and also due to the difficulty in formulating/modelling uncertainty. In general, as in [15], a decent benchmark problem for uncertain optimisation problems should encompass the following characteristics: 1) the flexibility of setting objectives, problem uncertainties and dimensions; 2) simplicity and efficiency of problem implementation and evaluation; 3) allowing conjectures to real-world problems. In the literature, there have been some attempts to provide benchmark problems for uncertain optimisation problems. The paper [15] reveals that most existing academic benchmark problems focused only on some of the above-mentioned criteria due to complexity of these problems. Nevertheless, there have been few publications that attempted to consider some of these criteria by taking into account changeable constraints or the situation where future events depend on previous solutions [16-24]. These attempts, however, still have some limitations to capture properly the uncertainty in real-world problems. Thus, these limitations in benchmark problems make it very difficult to evaluate robust/dynamic algorithms properly and hence it is not very clear that the developed robust/dynamic optimisation algorithms in the literature can effectively tackle real-world optimisation problems.

One of the most common uncertainties in literature is the uncertain characteristic of upcoming items. It was mentioned in [25] that the volumes of items and the capacities of bins are only approximately known due to economic restrictions or technical limitations. Also in [26] the item profits were considered as random variables because of the handling operations. Another uncertainty proposed in [27] is the unknown future demand of items. It may generate extra cost or loss profit that the planned containers are insufficient. Thus, due to the variety of uncertainty, it is difficult to generate test case depending on different problem requirements.

According to the literature, many performance measures used in existing experiments are specific for each algorithm. It means that while tackling the same problem, it is hard to compare the performance of different algorithms. With the purpose of providing a framework that different algorithms of BPP can be compared, common performance measures are essential. The volume utilisation which is the total volume of packed items in this bin divided by the volume of the bin is mentioned in several experiments [28-30] which is either total volume utilisation or the average. In addition, the running time of the algorithm is another measurement that is commonly referenced [31, 13].

3. BPP framework

3.1 Features of the framework

The framework provides three objectives that optimisation algorithms can choose to optimise: 1) minimise the number of bins; 2) minimise the cost of bins; 3) maximise the profit of packed items.

Given the objectives above, the framework can generate uncertainty in: size of items, weight of items, profit of items, and cost of bins. The uncertain values can be generated under any distribution (e.g. uniform distribution, normal distribution). Both two-dimensional (2D) and 3D (cuboid items) BPP can be visualised. The framework can generate problems with single bin, multiple identical bins, or multiple bins of different type.

The framework is developed in Java. To illustrate how the framework can be used to test algorithms that solve the static BPP, we use a static bin packing algorithm [3] which solves the BPP in an offline way, assuming that all items are available beforehand. This algorithm uses heuristic approaches on initially sorted items by non-increasing volume. Although it is not the latest method, it is one of the few available 3D bin packing algorithms whose source code is accessible and detailed algorithm description is available. Because the purpose of this paper is to provide a proof of concept, we feel the decision of choosing this algorithm is justified.

To illustrate how the framework can be used to test dynamic algorithms that solve the dynamic BPP in an online way, we implement a new online algorithm. It should be noted that the dynamic BPP is very new to the academic community, and while there has been a few research that proposes solving algorithms [32, 33], these research have not provided any experimental details to prove that these algorithms work. Because of that, here we just provide a simple algorithm as a proof of concept. The algorithm works by packing upcoming items layer by layer under the same assumption as the static algorithm. The algorithm assumes that there is no information of upcoming items and hence the problem needs to be solved online. At the time an item comes, the information of the item, the current bin and a packing location are passed to the algorithm as parameters and the algorithm goes on packing the new item into the current available bins. Items are packed in a bin in "layers" (see figure 1). The size of each layer is the maximum size of the packed items. The algorithm will check whether the current vertical layer in a bin has enough space for this item. If there is enough space on the current layer, the item is packed to the location and returns the updated packing location. If there is no enough space on this layer, it will check the next layer till this bin is full and then open a new one. The simulation process of online bin packing is shown in figure 2.
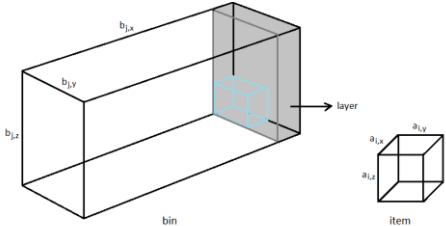
Figure 1: Example of item, bin, and layer in 3D.
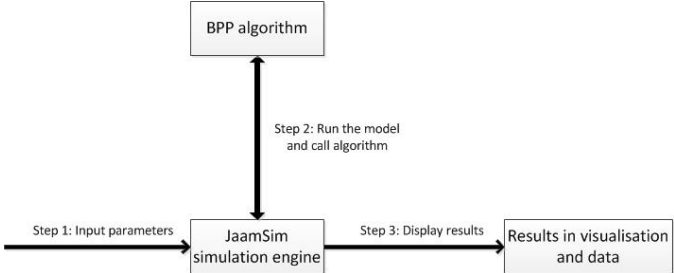
Figure 2: The process flow of simulation.

As mentioned in section 2, performance measures are necessary to be introduced due to the lack of common measures. To evaluate the effectiveness and efficiency of the algorithms, we use the following measures: average utilisation, number of bins, and running time. The average utilisation shows how much the bin capacity is used on average. It provides a reference to see how close the items are packed in each bin. The average utilisation of the bins is the total volume of packed items dividing by the total volume of used bins. Moreover, the number of bins is used as another performance measure regarding the effectiveness of algorithms. It represents how an algorithm performs with an objective of input minimisation such as number of bins minimisation or cost of bins minimisation. The running time generally means the process time of an algorithm, for example, how fast or slow the algorithm identified the optimum solution.

3.2 Generating test problems

This subsection proposes test cases that we suggest to use as default. However, our framework is not limited to them. Users can customise test cases based on their own needs. Here we considered three levels of test cases: easy, medium, and hard which represent the level of difficulty of packing items. The time that an algorithm takes normally depends on the difficulty of the problem. Due to the lack of a proper 3D benchmark, the following instances are extended from [34] which provides instances in 2D. Another dimension is added for bins and items. They are displayed in table 1 based on following types of cuboids that are defined in terms of the width W, height H and length L of the bins. The input used for generating the data set is listed below.

Type 1: w_j uniformly random in $[\frac{2}{3}W, W]$; h_j uniformly random in $[1, \frac{1}{2}H]$; l_j uniformly random in $[1, \frac{1}{3}L]$; Type 2: w_j uniformly random in $[1, \frac{1}{2}W]$; h_j uniformly random in $[\frac{2}{3}H, H]$; l_j uniformly random in $[\frac{1}{2}L, L]$; Type 3: w_j uniformly random in $[\frac{1}{2}W, W]$; h_j uniformly random in $[\frac{1}{2}H, H]$; l_j uniformly random in $[1, \frac{1}{2}L]$; Type 4: w_j uniformly random in $[1, \frac{1}{2}W]$; h_j uniformly random in $[1, \frac{1}{2}H]$; l_j uniformly random in $[\frac{2}{3}L, L]$.

Table 1: Example of the problem instances generated by the framework

Data set No.	Category	Bin size (W*H*L)	Total number of items	Item (w j * h j * l j)
I_20	Easy	30*30*30	20	Uniformly random in [1, 10]
I_40	Easy	30*30*30	40	Uniformly random in [1, 10]
I_60	Easy	30*30*30	60	Uniformly random in [1, 10]
I_80	Hard	30*30*30	80	Uniformly random in [1, 10]
I_1000	Hard	30*30*30	1000	Uniformly random in [1, 10]
II_20	Easy	100*100*100	20	Uniformly random in [1, 35]
II_40	Medium	100*100*100	40	Uniformly random in [1, 35]
II_60	Medium	100*100*100	60	Uniformly random in [1, 35]
II_80	Hard	100*100*100	80	Uniformly random in [1, 35]
II_1000	Hard	100*100*100	1000	Uniformly random in [1, 35]
III_20	Medium	100*100*100	20	Uniformly random in [1, 100]
III_40	Medium	100*100*100	40	Uniformly random in [1, 100]
III_60	Hard	100*100*100	60	Uniformly random in [1, 100]
III_80	Hard	100*100*100	80	Uniformly random in [1, 100]
III_1000	Hard	100*100*100	1000	Uniformly random in [1, 100]
IV_40	Hard	100*100*100	40	Type 1 with probability 70%, Type 2, 3, 4 with probability 10% each
IV_1000	Hard	100*100*100	1000	Type 1 with probability 70%, Type 2, 3, 4 with probability 10% each

.

4. Case study

4.1 Experimental design

The optimal solution that a dynamic BPP algorithm could find in the online dynamic case would always be worse than or at best equal to the optimal solution found in the static case. Due to that, to evaluate the efficiency of a dynamic BPP algorithm, we can compare its solution with that of an established static BPP algorithm. To demonstrate this type of comparison, in this experiment we are going to compare our online algorithm with the static algorithm (section 3.1) to evaluate the effectiveness/efficiency of the online algorithm. The simulation runs on an Intel

Core 2, 3.06 GHz computer with 4.0 GB RAM. The test set in table 1 is applied to both algorithms. In each group, we run the algorithm for ten replications.

4.3 Performance analysis

Table 2 shows the average results of ten replications. The average numbers of bins determined by the online and static algorithms do not have a significant difference when the number of items is low or the sizes of items are small in proportion to bin sizes, i.e. I_20 and II_20. However, for I_40, I_60, II_40, II_60, III_20 and III_40 the static solution provides a smaller number of bins with a larger average utilisation rate. This reflects the fact that in most cases the solution found online would be worse than the solution found offline in a static way.

In terms of the running time, the online algorithm has significantly shorter process times in comparison with the static algorithm. According to the results, the static algorithm takes much longer to achieve a packing plan for hard level of test cases. By the time we submit this paper it has not even been able to find a solution for all the large scale cases. The online algorithm, on the contrary, is much faster and can solve all the problems, including the hard, large-scale instances. For example, in the results of I_1000, II_1000, III_1000 and IV_1000 shown in table 2, in the online algorithm it takes only 0.002 seconds at most. This solving time includes both computational time of the proposed online algorithm and the simulation time to visually display the loading process. It means the simulation is also fast and our framework is capable of handling problems with a large number of items which is common in the real-world BPPs. Figure 3 shows an example of how the process of packing bins online is displayed in 3D in our proposed framework.

Table 2: Bin packing results. N/A represents that an algorithm has not finished the job by the time of submission.

Data set No.	Total No. of items	Online			Static		
		No. of bins	Average utilisation	Running time (s)	No. of bins	Average utilisation	Running time (s)
I_20	20	1	11.994%	0.001	1	11.994%	0.129
I_40	40	1.1	24.054%	0.001	1	25.752%	0.136
I_60	60	2	19.037%	0.001	1	38.077%	0.116
I_80	80	5	37.43%	0.001	N/A	N/A	N/A
I_1000	1000	22.3	29.16%	0.001	N/A	N/A	N/A
II_20	20	1	11.619%	0.001	1	11.619%	0.226
II_40	40	1.4	17.442%	0.001	1	22.273%	0.146
II_60	60	2	16.92%	0.001	1	33.83%	0.19
II_80	80	3.2	13.71%	0.001	N/A	N/A	N/A
II_1000	1000	26	22.46%	0.002	N/A	N/A	N/A
III_20	20	8.8	29.77%	0.001	4.5	58.43%	2.001
III_40	40	15.7	29.633%	0.001	6.6	70.107%	14.214
III_60	60	24.3	32.13%	0.001	N/A	N/A	N/A
III_80	80	37	33.11%	0.001	N/A	N/A	N/A
III_1000	1000	382	32.77%	0.001	N/A	N/A	N/A
IV_40	40	2	25.54%	0.001	N/A	N/A	N/A
IV_1000	1000	81	42.44%	0.001	N/A	N/A	N/A

5. Conclusion

This paper for the first time proposes a framework to generate benchmark problems for dynamic BPPs using discrete-event simulation. The developed benchmarks provide a test bed for evaluation of the applicability of dynamic optimisation algorithms to real-world BPPs. A set of test cases were generated using the framework to evaluate the online algorithm. The results of the online algorithm on these test cases were then compared with those of a static algorithm from the literature.

The future directions for further improvement of the proposed framework are as follows. First, in this paper only limited algorithms were evaluated, but this framework is capable of being integrated with other algorithms such as evolutionary algorithms. Second, state-of-the-art online bin packing algorithms from the literature can be reproduced in the framework to achieve competitive results. Third, the impact of uncertainty on each dimension of items can be investigated.
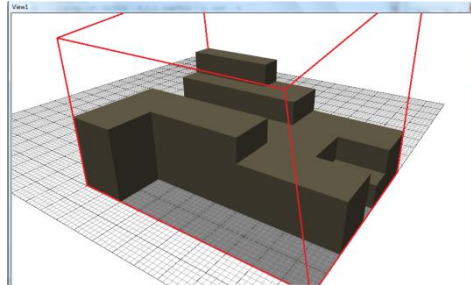

Figure 3: A 3D view of the online bin packing process, as displayed by our framework.

Acknowledgement

References

[1] Crainic, T.G., Perboli, G., Tadei, R.: Ts 2 pack: A two-level tabu search for the three-dimensional bin packing problem. European Journal of Operational Research 195(3) (2009) 744–760

[2] Feng, X.: Hybrid genetic algorithms for the three-dimensional multiple container packing problem. Flexible Services and Manufacturing – Springer (2013)

[3] Martello, S.: Algorithm 864: General and robot-packable variants of the three-dimensional bin packing problem. ACM Transactions on Mathematical Software (2007)

[4] de Almeida, A.: A particular approach for the three-dimensional packing problem with additional constraints. Computer and Operation Research (2010)

[5] Alvarez-Valdes, R.: A grasp/path relinking algorithm for two- and three-dimensional multiple bin size bin packing problems. Computers & Operations Research (2013)

[6] Che, C.H.: The multiple container loading cost minimization problem. European Journal of Operational Research (2011)

[7] Eley, M.: A bottleneck assignment approach to the multiple container loading problem. OR Spectrum (2003)

[8] Tian, T., Zhu, W., Lim, A., Wei, L.: The multiple container loading problem with preference. European Journal of Operational Research (2015)

[9] Lim, A.: The single container loading problem with axle weight constraints. International Journal of Production Economics (2013)

[10] Liu, J.: A novel hybrid tabu search approach to container loading. Computers & Operations Research (2011)

[11] Junqueira, L.: Three-dimensional container loading models with cargo stability and load bearing constraints. Computers & Operations Research (2012)

[12] Costa, M.d.G., Captivo, M.E.: Weight distribution in container loading: a case study. International Transactions in Operational Research (2014)

[13] Ceschia, S., Schaerf, A.: Local search for a multi-drop multi-container loading problem. Journal of Heuristics 19(2) (2013) 275–294

[14] Bortfeldt, W.: Constraints in container loading - a state-of-the-art review. European Journal of Operational Research (2012)

[15] Nguyen, T.T., Yang, S., Branke, J.: Evolutionary dynamic optimization: A survey of the state of the art. Swarm and Evolutionary Computation 6 (2012) 1–24

[16] Nguyen, T.T., Yao, X.: Continuous dynamic constrained optimization – the challenges. IEEE

Transactions on Evolutionary Computation 16(6) (2012) 769–786

[17] Jin, Y., Sendhoff, B.: Constructing dynamic optimization test problems using the multi-objective optimization concept. In: Applications of Evolutionary Computing. Springer (2004) 525–536

[18] Richter, H.: Memory design for constrained dynamic optimization problems. In: Applications of Evolutionary Computation. Springer (2010) 552–561

[19] Weicker, K., Weicker, N.: Dynamic rotation and partial visibility. In: Proceedings of the 2000 Congress on Evolutionary Computation. (2000) 1125–1131

[20] Nguyen, T.T., Yao, X.: Dynamic time-linkage problems revisited. In: Applications of Evolutionary Computing. Springer (2009) 735–744

[21] Nguyen, T.T., Yao, X.: Benchmarking and solving dynamic constrained problems. In: Evolutionary Computation, 2009. CEC'09. IEEE Congress on, IEEE (2009) 690–697

[22] Nguyen, T.T.: Continuous dynamic optimisation using evolutionary algorithms. PhD thesis, University of Birmingham (2011)

[23] Nguyen, T.T., Yao, X.: Dynamic time-linkage evolutionary optimization: Definitions and potential solutions. In: Metaheuristics for Dynamic Optimization. Springer (2013) 371–395

[24] Kavakeb, S., Nguyen, T.T., Yang, Z., Jenkinson, I.: Evolutionary fleet sizing in static and uncertain environments with shuttle transportation tasks - the case studies of container terminals. IEEE Computational Intelligence Magazine, in press (2016)

[25] Peng, J., Zhang, B.: Bin packing problem with uncertain volumes and capacities (2012)

[26] Perboli, G., Tadei, R., Baldi, M.M.: The stochastic generalized bin packing problem. Discrete Applied Mathematics 160(7) (2012) 1291–1297

[27] Crainic, T.G.: Bin packing problems with uncertainty on item characteristics: an application to capacity planning in logistics. Social and Behavioral Sciences (2014)

[28] Lim, A., Zhang, X.: The container loading problem. In: Proceedings of the 2005 ACM symposium on Applied computing, ACM (2005) 913–917

[29] Zhu, W., Lim, A.: A new iterative-doubling greedy–look-a-head algorithm for the single container loading problem. European Journal of Operational Research 222(3) (2012) 408–417

[30] Jiang, J., Cao, L.: A hybrid simulated annealing algorithm for three-dimensional multi-bin packing problems. In: Systems and Informatics (ICSAI), 2012 International Conference on, IEEE (2012) 1078–1082

[31] Egeblad, J., Pisinger, D.: Heuristic approaches for the two-and three-dimensional knapsack packing problem. Computers & Operations Research 36(4) (2009) 1026–1049

[32] Burcea, M., Wong, P.W.H., Yung, F.C.C.: Online multi-dimensional dynamic bin packing of unit-fraction items. In: Algorithms and Complexity. Springer (2013) 85–96

[33] Epstein, L., Levy, M.: Dynamic multi-dimensional bin packing. Journal of Discrete Algorithms 8(4) (2010) 356–372

[34] Lodi, A., Martello, S., Vigo, D.: Heuristic and metaheuristic approaches for a class of two-dimensional bin packing problems. INFORMS Journal on Computing 11(4) (1999) 345–357

Shudong Li

Numerical modelling of soft material systems with embedded stiffer layers

Shudong Li[1*], Mark Lake[2], Yaodong Gu[3], and James Ren[1]

[1] Department of Maritime and Mechanical Engineering, Liverpool John Moores University.

[2] School Sport and Exercises Science, Liverpool John Moores University.

[3] School of Sport Science, Ningbo University, China.

*E-mail address: S.Li2@2014.ljmu.ac.uk

Abstract. Soft material systems with embedded stiffer layers represent a new material system increasingly used in medical, sport and electronic systems. A typical example is a rubber insole embedded with a rigid reinforce phase (e.g. plastics). The behaviour of such a system under uniform or localised loading is different from traditional layered composites. However its mechanics is not well defined and a detailed study is required. This paper reports recent research on the simulation and prediction of the deformation of thin stiff shell embedded within a soft material system under localised loading conditions represented by indentation with a flat ended indenter. A new modelling method of embedded 2D structure has been developed and showed a comparable trend with analytical solution. Parametric FE models of different embedded structures have been developed to predict the deformation of embedded shells at different locations underneath the indenter.

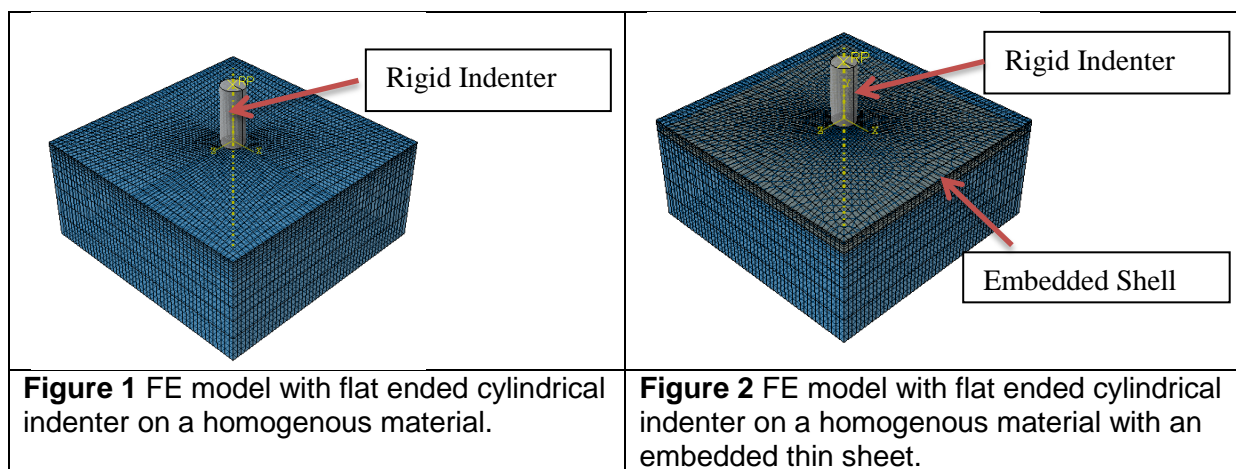**Keywords:** Indentation, embedded system, rubber

1. Introduction

Soft materials with embedded structures are widely used in many applications, particularly in sport and medical systems (*Patton et al, 2015; Heidner and Adams, 2016*). This material type refers to systems with a stiffer structure (either in 3D or 2D form) being fully enclosed in a matrix for improving mechanical performance or achieving other functions. For example, a thin stiffer plate could be inserted in a soft insole to increase the strength or pressure distribution (*Kajtaz et al, 2015*). Heating element can be used in insole or midsole for thermal management of shoes. Recently, there has been considerable work in wearable technology in footwear for both research purpose and technological development, in which a thin electronic device or chip is embedded in the sole system (Crea *et al*, 2014). A detailed understanding the mechanics of these material systems is important for improving the structure integrity and reliability of systems.

Typically, embedded system can be made of a deformable material (i.e. the insert is not significant stiffer than the matrix) or an inextensible material (where the insert is much stiffer than the matrix). In service, the system could be under different loading mode, such as torsion, bending and indentation (*Oyen and Cook, 2003; Siivola et al, 2016*), which poses a new challenge in material selection and design. Studying the deformation of the system under an indentation process is particularly important for such a system. In an indentation process, an indenter is pressed onto the sample surface and deforms material. This is a

process similar to a situation when a sharp object (such as stone) deforming a shoe sole system, which may directly influence the embedded structures. Many works have studied the indentation of soft material such as rubbers (*Wu et al, 2016; Needleman et al, 2015*), foams (*Siivola et al, 2016; Duncan et al, 2015*) or composites (*Zhang et al, 2016*) with a range of indenter shapes such as spherical or flat indenters. Indentation is a typical localized loading, in which the stress and strain condition is not well defined. The deformation of an embedded system in a matrix is even more complex.This paper reports recent research on the simulation and prediction of the deformation of thin stiffer shells embedded within a soft material system under indentation load. A new modelling methodology of embedded 2D structure has been developed. The modelling results were compared to the prediction of analytical solution. Parametric FE models have been developed to study different embedded structures, which potentially could be used in materials design with balanced functions and structure integrities.

2. Finite Element Model and Simulation Approach of Embedded Systems

Figure 1 FE model with flat ended cylindrical indenter on a homogenous material.	**Figure 2** FE model with flat ended cylindrical indenter on a homogenous material with an embedded thin sheet.

Figures 1 and Figure 2 shows the FE model for indentation of a homogenous samples and a sample with an embedded shell, respectively. In both models, the indenter is modelled as a rigid body as it is much stiffer than the sample. Finer meshes have been applied in the region around and underneath the indenter. The matrix is modelled with solid elements (element type: C3D8R) and the embedded shell is modelled with shell elements as its in–plane dimension is much higher than the thickness. The procedure to build the embedded shell is much more complicated than the homogenous model. A thin shell is firstly built, and then embedded in a square block sample. A procedure using constraint type of "Embedded Region" in interaction has been developed to model the thin shell within the "whole model" *(ABAQUS 6.14 Theory Manual)*. The thickness of the shell can be modified by assigning different thicknesses in the section module in ABAQUS. With this procedure, the position and thickness of the shell can be modified within the system without the need of applying major change to the meshing of the matrix. Using such an approach, the model can be modified flexibly to simulate indentation tests of different conditions/situations, such as different shell thicknesses and shell positions from the surface. With a 3D FE model, the orientation or angle of the indenter to the sample surface can also be assessed. The sample is sufficient large to avoid the sample size effects. In both FE models, the bottom surface of the sample is fixed in all degree of freedom (DOFs). A fixed displacement is applied on the indenter to move it onto the sample. As shown in the model, the top surface of the sample and the shell is partitioned into four regions to facilitate the assigning of element and ensure that there is no lateral movement of indenter from the central points. Mesh sensitive tests have been performed by altering the element size in the model until there is no significant change of the modelling results (the indentation force is controlled within 5%) with mesh sizes.

3. Numerical and Analytical Results and Analysis

In this work two conditions have been studied to establish the validity of the model. The first model (Designated as Homogenous model) is a model with no thin shell embedded. This allows a direct comparison between the numerical results and a known analytical solution for indentation of an elastic half space with a circular flat indenter. The second model (Designated as inextensible shell model) consists of a sample with an embedded shell, which is much stiffer than the matrix.
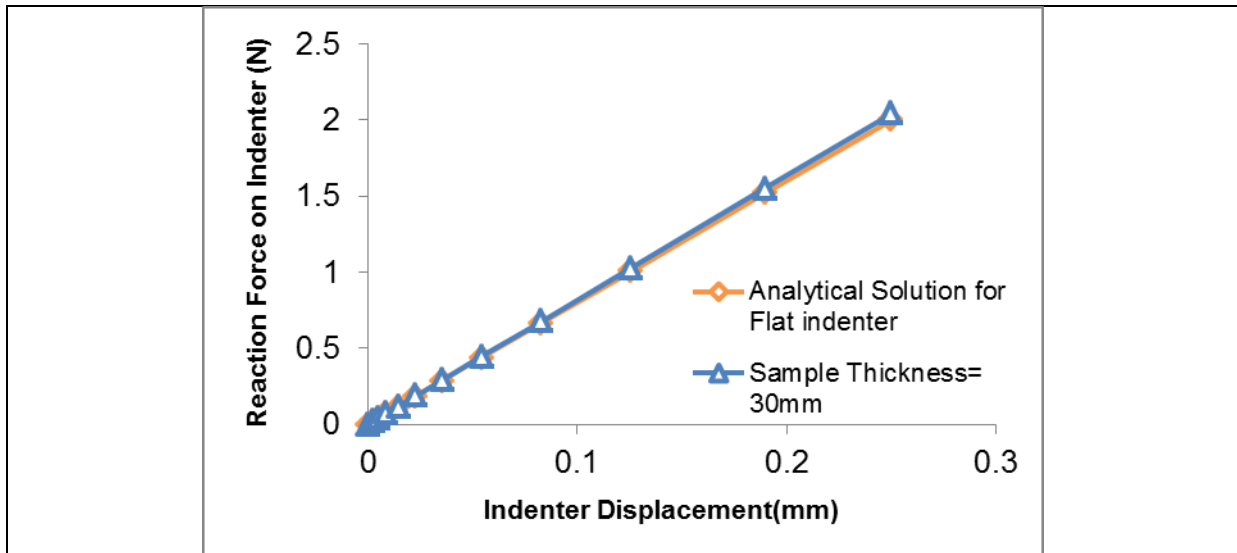

Figure 3 Comparison between indentation force-displacement data from numerical and analytical solution on a homogenous sample.

Figure 3 shows force-indentation depth data of the homogenous model in comparison with a known analytical solution (Equation 1) for flat-ended cylindrical indentation *(Riccardi, 2004)*:

$$P = \frac{2aEd}{1-v^2}$$ (1)

Where 'P' is the load on indenter, 'a' is the radius of the cylindrical indenter; 'E' is the Young's Modulus of the sample. 'v' is the Poisson's ratio of the matrix. 'd' is the displacement of the indenter. As shown in the figure, the numerical data is in a very good agreement with the analytical solution. Similar work has been conducted on other models with different indenter sizes, all the numerical results showed a reasonable agreement with the analytical solutions. This suggests that the FE model is accurate.
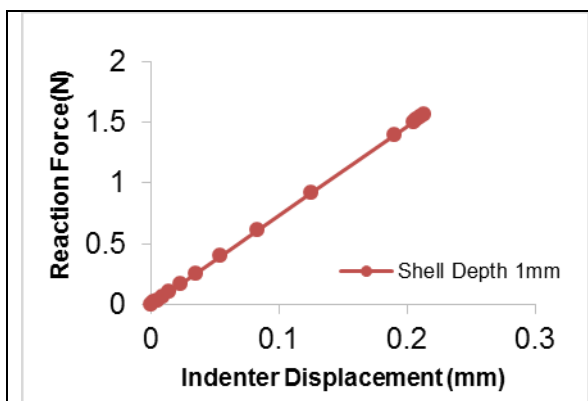

Figure 4 Typical force displacement data of a sample with a thin embedded shell.
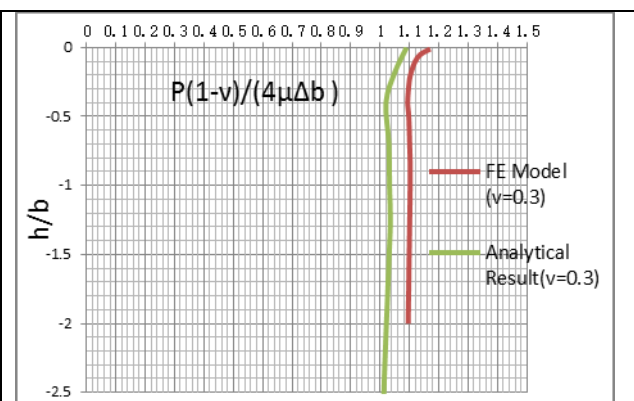

Figure 5 Comparison between numerical data and analytical data.

Figure 4 shows the force displacement data on a sample with a thin embedded shell (0.1mm in thickness). The data shows that the force-displacement follows a linear relationship. For a thin inextensible shell under an idealised boundary condition, the relationship between force and shear modulus has been established by Selvadurai (2009) through a dimensionless term:

$$M = \frac{P(1-v)}{4\mu\Delta b}$$ (2)

Where 'P' is the indentation load, 'v' is the Poisson's ratio of matrix, 'μ' is the Shear Modulus of matrix, 'Δ'is the displacement of indenter; "b" is the radius of cylindrical indenter. This offers a mean to assess the accuracy and validity of the FE model under comparable conditions. Figure 5 shows this dimensionless parameter (M) *vs.* the ratio between the depth of the shell and the radius of the indenter (h/b). As shown in the figure 5, the FE data shows a reasonable agreement with the prediction of the analytical solution with different shell positions represented by different h/b values. The numerical data and the analytical solution show a similar trend, but there are some slight differences between the values. This is reasonable as the analytical solution did not consider the thickness of the embedded shell. Both the analytical and numerical data shows a progressive decrease of force when the embedded thin shell is located further away from the indenter and sample surface. When the shell is close to the top sample surface, the force (as reflected by the value of M) is much higher representing more significant effects of the shell on the indentation resistance. With increasing shell depth, the indentation force reduces and eventually becomes insignificant. When 'h/b' is over 0.5, the "M" becomes close to 1 within the indentation depth range of this work. This suggests that the embedded shell no longer has any significant influence on the indentation resistance within small elastic deformation.

(a) Force displacement data of sample with embedded system when the shell is at different locations.

(b) Deformation of the shell (Shell close to the surface)

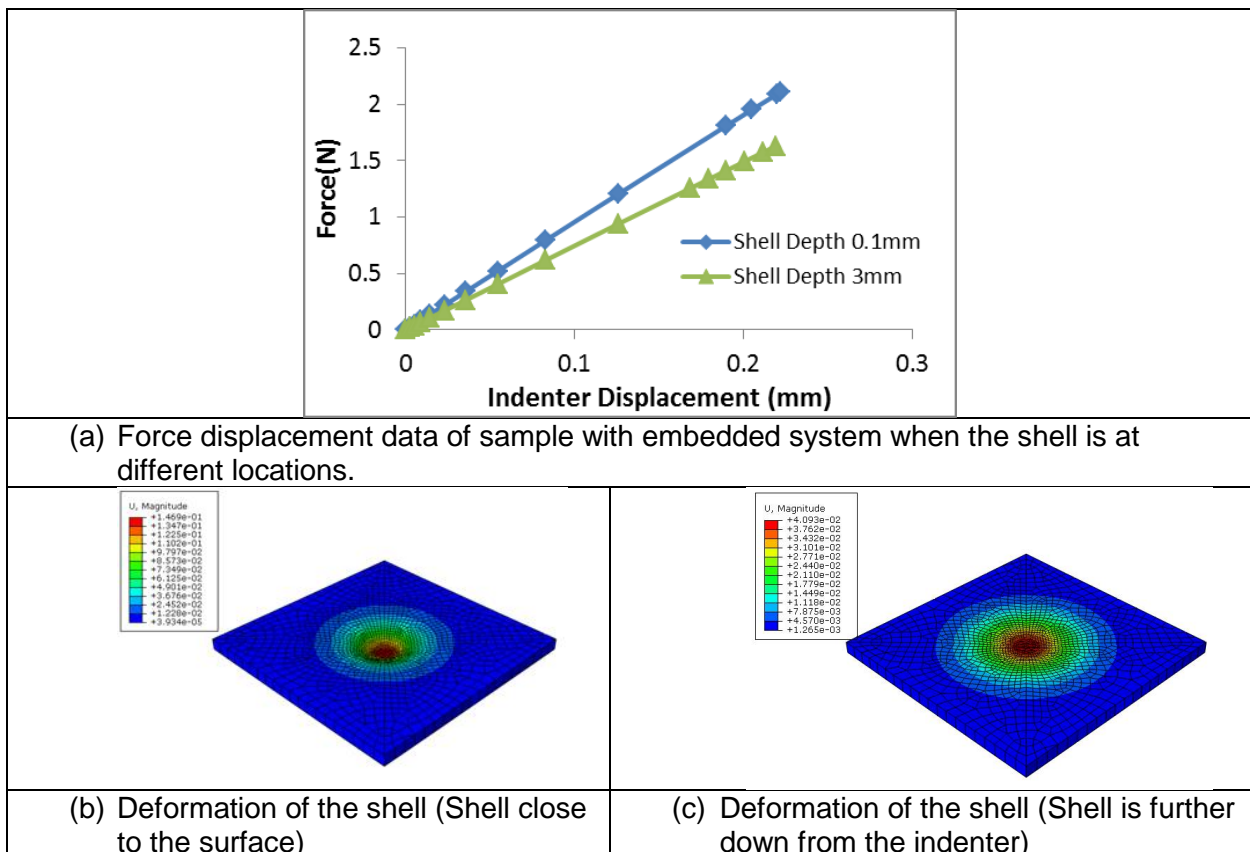(c) Deformation of the shell (Shell is further down from the indenter)

Figure 6 (a), (b) and (c) Force displacement data and shell deformation of sample with the embedded stiff shell at different locations underneath the indenter.

Figure 6a shows the indentation force-displacement data with a thicker embedded shell (1mm in thickness) at different locations underneath the indenter. It is clearly shows that, when the shell is very close to the sample surface, the force is apparently higher than that for the case when the shell is further away from the indenter. Figures 6b&c plot the displacement field to show the deformation of the shell. It is interesting to note that the shell deeper in the sample has exhibited a much larger deformed volume/area even though the overall displacement level is lower. This will provide a much better protection to the embedded system. Future work is to systematically quantify, with the FE model established, the effects of shell thickness and the ratio between the stiffness of the shell and the matrix and establish an optimum structure to protect the embedded system though parametric studies and design of experiments (DOEs).

4. Summary

This paper reports recent research on the simulation and prediction of the deformation of thin stiff shell embedded within a soft material system under indentation loading conditions. FE models of a flat ended indenter on homogenous sample showed a very good agreement with a known analytical solution. A new modelling methodology of embedded 2D structure has been developed and used to study the deformation of embedded systems. The numerical results show a comparable trend with analytical solution. The results showed that the effect of embedded shell on the indentation resistance is associated with the shell thickness and the ratio between the shell position and the indenter diameter.

References:
Abaqus Theory Manual, 2014 Version 6.14-1 Dassault Systems.
Braun B J, Veith N T, Hell R, Dobele S, Roland M, Rollmann M, Holstein J and Pohlemann T 2015. Validation and reliability testing of a new, fully integrated gait analysis insole *J. foot ankle res.* 8(1), pp 1-7.
Crea S Donati, M Rossi S M M D, Oddo C M, and Vitiello N 2014 A wireless flexible sensorized insole for gait analysis *Sensors* 2014(1), pp1073-93.
Duncan O, Foster L, Senior T, Alderson A and Allen T 2016 Quasi-static characterisation and impact testing of auxetic foam for sports safety applications *Smart Mater. Strut.* 25(5), p054014.
Heidner M C and Adams D O, Aga Medical Corporation 2016 Multi-layered medical device for treating a target site and associated method *U.S. Patent* 9,232,992.
Kajtaz M, Karren J and Subic A 2015 Experimental investigation into suitability of smart polymers as an impact-absorbing material for an improved rugby headgear *Int. Conf. Mech. Mat. Mech. Chem. Eng. (Barcelona) pp* 62-74.
Needleman A, Tvergaard V and Van der Giessen E 2015 Indentation of elastically soft and plastically compressible solids *Acta Mech.* 31(4), pp 473-80.
Oyen M L and Cook R F 2003 Load–displacement behavior during sharp indentation of viscous–elastic–plastic materials *J. Mat. R. 18*(01), pp 139-50.
Patton D A and McIntosh AS 2015 Considerations for the performance requirements and technical specifications of soft-shell padded headgear *Proc. Inst. Mech. Eng. P J. Sports Eng. Technol.*p 1754337115615482.
Riccardi B and Montanari R 2004 Indentation of metals by a flat-ended cylindrical punch *Mater. Sci. Eng. R-Rep.* A *381*(1), 281-91.
Selvadurai A P S 2009 Boussinesq indentation of an isotropic elastic halfspace reinforced with an inextensible membrane *Int. J. Eng. Sci.Int.* 47(11), pp 1339-45.
Siivola J T, Minakuchi S and Takeda N 2016 Unloading response prediction of indentation loaded foam core sandwich structures using extended foam material model with tensile hardening. *Compos.Part B-Eng. 84*, 71-82.

Wu C E, Lin K H and Juang J Y, 2016 Hertzian load-displacement relation holds for spherical indentation on soft elastic solids undergoing large deformations. *Tribol. Int.* 97, pp 71–6

Zhang X, Wang P, Neo H, Lim G, Malcolm A A, Yang E H and Yang J 2016 Design of glass fiber reinforced plastics modified with CNT and pre-stretching fabric for potential sports instruments. *Mater. Des.* 92, pp 621-31.

Salifu Osman

Verification of Olive Oil using Optical and Microwave Spectroscopy Sensors

S B Osman*, J D Cullen, O Korostynska, A Mason, A Shaw, A I Al-Shamma'a

BEST Research Institute, School of the Built Environment,
Liverpool John Moores University,
Liverpool, L3 3AF, UK
s.b.osman@2013.ljmu.ac.uk*

Abstract. The verification of vegetable oils is seemingly important when it comes to the examination of authenticity and the detection of possible adulteration of extra virgin olive oils with seed oils or low-quality olive oils. Classical methods such as, Gas Chromatography, High Performance Liquid Chromatography, Fourier Transform Infrared and Nuclear Magnetic Resonance, are hereby too expensive for widespread industrial use and require samples to be analysed in dedicated laboratories thus incurring a significant time penalty. This paper demonstrates that the use of optical and microwave sensors could offer real-time measurement of olive oils adequate for determining product authenticity. It does this by evaluating a sensor system used to measure the dielectric properties against the characteristics of extra virgin olive oil and refined olive oils. In particular, the capability of the system to distinguish between these oils, even when mixed, or dyed is demonstrated. An extra virgin olive oil and a refined olive oil were tested using the microwave cavity sensor, the IDE sensor and a trial capacitive digital sensor. The VNA was set to span from 10 kHz to 15GHz to read the S_{11} parameter. The results show promising peaks of interest in all the spectra. In addition a spectrophotometer was used to test the chemometric character of the mixture and the results show similar patterns emerged using this optical means. These experiments show that a combination of these sensing techniques can be used for olive oil verification.

Keywords - microwave spectroscopy, spectrophotometer, sensor, vector network analyzer, olive oils, adulteration and real time analysis.

1. Introduction

The use of olive oil as an economically viable nutrient is well spread internationally (Bendini et al, 2002). The high costs of extra virgin olive oil, when compared to other commonly used vegetable oils make it prone to adulteration with less expensive oils in order to increase profits. The most commonly used adulterants found in virgin olive oil are mainly seed oils, such as sunflower, soy, corn and rapeseed oils, and some nut oils, such as hazelnut and peanut oils (Firestone, 2001). Several commercially known categories of olive oil are legally defined by the European Community Council of Regulation (EC, 2007). There is also the possibility of mixing less expensive commercially known categories such as refined or mixed olive oil and pomace oil with the highest quality product, like extra virgin olive oil, for economic reasons. Detection of these two types of adulteration is often complicated with no single test available that can alone accomplish the task, especially when oils with similar chemical compositions to extra virgin olive oil are utilised (García-González and Aparicio, 2006).

The detection and verification of the adulteration of extra virgin olive oil are not the easiest; traditional ways demand monitoring of several organic compounds and to establish a comparison with typical unadulterated oils in order to identify significant change of composition that could be related to adulteration (Fragaki et al, 2005). The following topics discuss brief exposure of some of the methods and state of arts as used to work along these purposes.

The use of microwave spectroscopy (Korostynska et al, 2013) in conjunction with spectrophotometry has allowed for simplification of these processes to give way for a cost effective solution to this problem. It is for this purpose that three different microwave sensors are trialled and spectrophotometric readings taken to verify changes in the spectra of extra virgin olive oil as against refined olive oil.

2. Methodology

To verify the applicability of the sensors, a sample of extra virgin olive oil (EVOO) and that of refined olive oil (ROO) was measured using three different sensors. The Rohde and Schwarz ZVA24 VNA were utilized, and range of frequencies from 10 MHz to 15 GHz selected.

For the cavity sensor (fig. 1), two ports were connected to the VNA, and a test tube samples of the oils were inserted into the sensor read the reflected powers of S11 and S21. Areas of interest were spanned out for data analysis using Microsoft Excel.

For the IDE sensor (fig. 2), one port was connected to the VNA, and the samples of the oils were pipetted into the sensor to read the reflected powers of S11. Areas of interest were spanned out for data analysis using Microsoft Excel.

For the trial capacitive sensor (fig. 3), two ports were connected to the microwave cavity sensor, and the samples of the oils were placed into the sensor to read the reflected powers of S11. Areas of interest were spanned out for data analysis using Microsoft Excel.

For the spectrophotometric measurements (fig. 4), the samples were placed in cuvettes. Air (vacuum) was used as the Blank. The wavelength spanned between 650 and 700nm to read the *absorbance*.

TABLE 1: Table of measurements

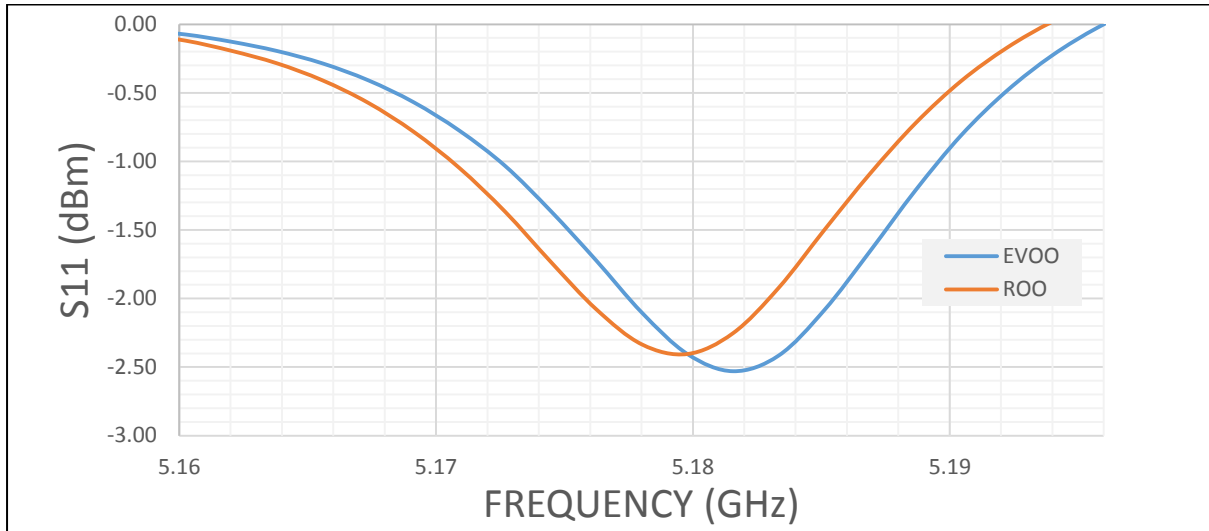| Sample	Description	Weight (g)	Percentage (%)
1	EVOO	50	100
2	ROO	50	100

3.   Results

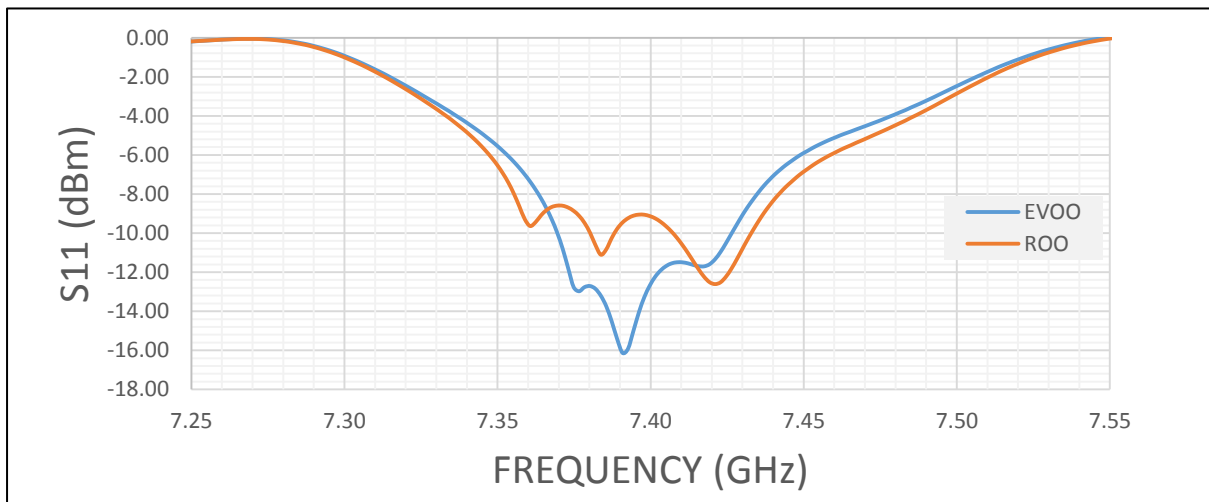*Figure1: Microwave spectra of cavity sensor*

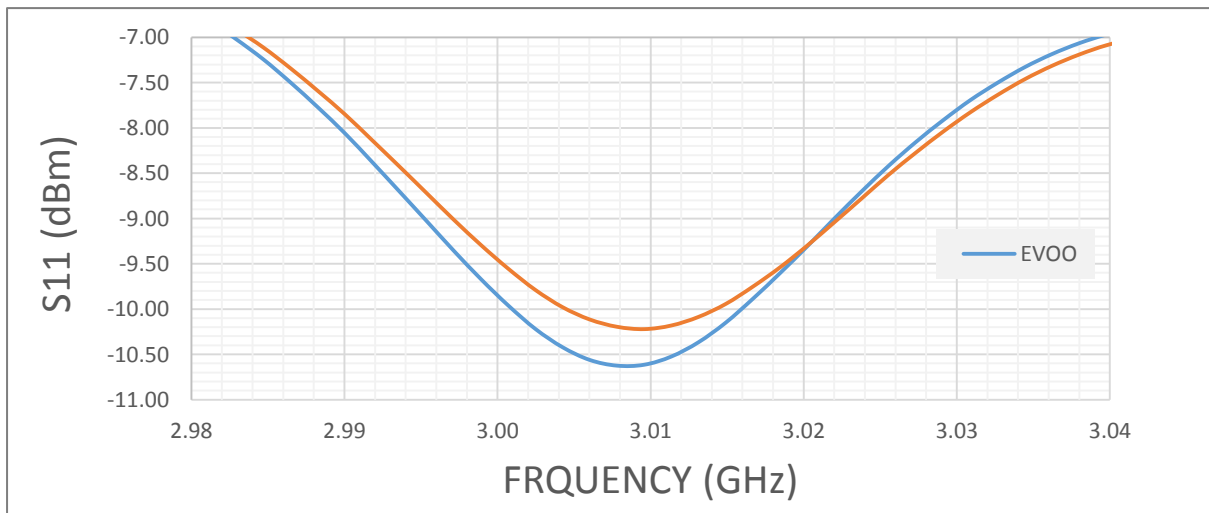*Figure 2: Microwave spectra using the IDE sensor*

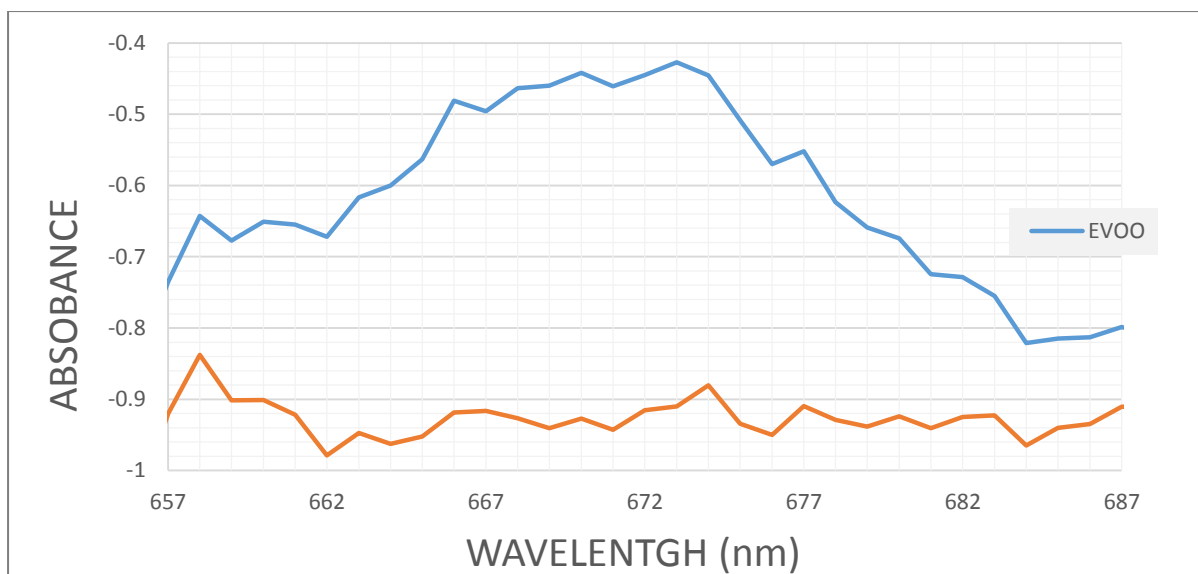*Figure 3: Microwave spectra using trial capacitive IDE sensor*

Figure 4: Optical Spectra

Conclusion

In this experiment a verification of olive oil was carried using various sensors. Three microwave sensors: made up of the cavity and two different IDE sensors and an optical spectrophotometer were utilized. The ability of these systems to distinguish between refined olive oil and extra virgin olive oil was probed. Arriving at the conclusion that any of these techniques could be used as a means and/or could complement each other for the verification of olive oil. It is important to hereby also clarify that for the case of unadulterated oil (EVOO, ROO) either of these tests could be used, but if adulterated oil is used the use of dye for example, and then a combination system incorporating different sensors would be better since it would become less susceptible to circumvention.

REFERENCES

Bendini A, Cerretani L, Carrasco-Pancorbo A, Gómez-Caravaca A M. Segura-Carretero A and Fernández-Gutiérrez, (2007). Phenolic molecules in virgin olive oils: a survey of their sensory properties, health effects, antioxidant activity and analytical methods. An overview of the last decade. Molecules, 12, 1679–1719.

European Community, Commission Regulation 702/2007 of 21 June 2007 amending Commission Regulation No. 2568/91/EEC, Official Journal of the European Communities L161 (2007) 11–27.

Firestone D. (2001). Assuring the integrity of olive oil products. Journal of AOAC International, 84, 176–180.

Fragaki G, Spyros A, Siragakis G, Salivaras E, and Dais P. (2005). Detection of extra virgin olive oil adulteration with lampante olive oil and refined olive oil using nuclear magnetic resonance spectroscopy and multivariate statistical analysis. Journal of Agriculture and Food Chemistry, 53, 2810–2816.

García-González D L and Aparicio R (2006). Olive oil authenticity: The current analytical challenges. Lipid Technology, 18, 81–85.

Korostynska O, Blakey R, Mason A and Al-Shamma'a A. Novel method for vegetable oil type verification based on real-time microwave sensing, Sensors

and Actuators A: Physical, vol. 202, pp. 211-216, DOI: 10.1016/j.sna.2012.12.011, 2013.

Keyur Joshi

Determining overall quality of milk products using microwave spectroscopy

K H Joshi*, **A Mason, O Korostynska, A Shaw, J D Cullen and A Al-Shamma'a**
Built Environment and Sustainable Technologies (BEST) Research Institute,
Department of Built Environment,
Liverpool John Moores University,
Liverpool,
L3 2ET
*K.H.Joshi@2014.ljmu.ac.uk

Abstract: In this article, a novel approach is introduced for quality testing of milk products using microwave spectroscopy. Electromagnetic wave sensors are used to record the spectral signatures from the most widely consumed categories of milk in the commercial market; namely, skimmed, semi-skimmed and whole milk. This database is then analysed to determine the overall quality of given milk products in terms of milk type i.e. milk contents, and adulteration i.e. aging and spoilage. This methodology attempts to resolve the limitations of milk quality testing schemes being employed currently in practice to give a milk quality monitoring system which is real-time, more cost-effective, and less complex than the other existing standards. With further optimization in the design of this type of sensors and their selectivity, this system, in situ, has the potential to be applied at certain levels of milk supply chain hierarchy - specifically by medium or large scale retailers to test the milk products that they are receiving after production from the dairy industries and several stages of transport before selling to the small-scale retailers and the end consumers.

Key Words - *adulteration, dairy industries, electromagnetic wave sensors, microwave spectroscopy, milk quality testing, milk supply chain, milk type, real-time, spectral signatures, spoilage detection.*

1. Introduction

Milk and dairy products play an important part not only in the diets we follow daily but also in overall economy of many developing as well as developed nations. It is an integral part of the daily meal, especially, for those who follow a vegetarian diet due to its nutritious values. The dairy sector is the fastest growing industry, specifically in Asian, Latin American and Caribbean parts of the globe with the increasing consumption of milk products in recent decades [1]. More than 6 billion people worldwide consume milk and milk products, most of which belong to the developing nations of the world, and this number is rapidly growing [2]. Due to this, milk processing, after milk production, becomes second very important phenomenon in the dairy industry. Alongside diligent handling, milk also requires processing due to its short shelf-life because of its nature to provide the medium for microorganisms like bacterial pathogens to grow which causes spoilage of milk and also brings consumers to various diseases.

It is important to note that the primary goal of milk processing is to help maintain the actual quality of raw milk and thereby to ensure its preservation. Milk processing enables its

preservation for much longer periods and helps in reducing or avoiding food-borne illnesses resulting from the lack of its quality or spoilage [3]. The world milk production in the year 1982 was 482 Million Tons which drastically increased to 754 Million Tons in the year 2012, making it one of the primary sources of income around the world, with more than 50% rise in just 30 years and having engaged around 150 million families all over the world in milk production [4].

Microwave sensors are used for a wide variety of applications including measurement of distance, movement, shape, and particle size, but the largest group of applications is concerned with the measurement of material properties [5]. The fundamental principle of EM wave sensors that operate at microwave, radio frequencies (RF), is based on the interaction of signals with the medium of material under the test. This interaction between microwave signal and the medium material can be in the form of amplitude attenuation or phase shift, and it determines the relative permittivity and permeability [6].

Here, spectral responses, from the three commonly used types of milk (whole milk, semi-skimmed milk and skimmed milk), are captured using EM wave sensors and recorded through Vector Network Analyzer (VNA). For all the three categories, adulteration process of milk was examined with the help of resonator cavity sensors, as microwave sensing devices, in conjunction with VNA to enable the measurements.

Classification among the three milk types and the ageing of milk could be shown in this work, besides spoilage, comprehensively with the resulting graphs from relevant practical measurements comprising the detection and analysis of spectral signatures for S_{11} and S_{21} parameters, connectively known as scattering parameters, which are the measures of reflected back energy (power) towards the input Port-1 and transferred energy (power) towards the input Port-2, respectively, in the EM wave sensor assembly. The work explained in this article lays a foundation for a novel, real-time, less complicated, less expensive yet state-of-the-art method for milk quality control that can be applied within and outside the premises of laboratory restrictions, in situ. The text may be divided into sections, subsections and, where necessary, subsubsections.

1.1. Background

In improving the milk preservation techniques, spoilage detection plays a vital role. Milk spoilage can be defined in many ways and its measure could be different each time depending upon the context it is discussed in. Adulteration check has also been one of the sustained interests in milk quality control process, considering the ambiguity of "Use by Date" and its relevance to the actual spoilage of the milk as this vagueness troubles both, the consumers as well as the manufacturers. Buyers are mostly reluctant to purchase products close to their dates of expiry. This ultimately imposes a negative financial impact on dairy industry [7]. Milk is a perishable product and spoils very quickly. Its nature of having lower acidity and higher nutrients make it the perfect breeding platform for bacterial micro-organisms, including those which are responsible for food poisoning, known as pathogens. Bacteria contaminate the milk, and their destruction is the main reason for milk processing which enables milk preservation. This can be achieved by fermentation, heating, cooling, and removal of water or concentration or separation of various components from milk, to produce foods like butter and cheese [8].

1.2. Difficulties in milk supply chain

To understand where the significant amounts of challenges lie in the system, it is necessary to study the milk supply chain. Milk is tested at both, procurement and packaging levels, by the industries for required level of quality and its assurance, but after each phase of transportation it is most likely that the quality of milk product is altered, especially when the distance is longer and the quantity is large. Fragmented Milk Production, where milk production takes place in an unorganized sector, is a big challenge and raises concerns when it comes to milk quality and milk procurement in many economies like India – which is

the largest producer and consumer of milk, having the largest livestock worldwide [9]. The research report by the team also adds that milk supply chains face difficulties in transporting milk, especially in large quantity, under desired conditions over longer distances(>200km) while ensuring the quality of milk products.

One of many, news stories that keep arising on regular basis regarding the contamination found in raw milk, mentioned that the California Department of Food and Agriculture inspectors confirmed the bacterial presence following to product testing as a routine inspection and sample collection at the facility. All the products of the said batch were ordered to be quarantined, followed by their removal and disposal from the store shelves, including those already belonging to customer refrigerators [10].

2. Aim and Objectives

2.1. Aim of the project

The aim of this project is to determine the overall quality of milk products using a novel electromagnetic (EM) wave sensor based technique, which is low-cost, real-time and that which integrates all the objectives given below in a single prototype design.

2.2. Objectives

Figure 1, highlights the objectives for this particular article to achieve the aim of the project as above.
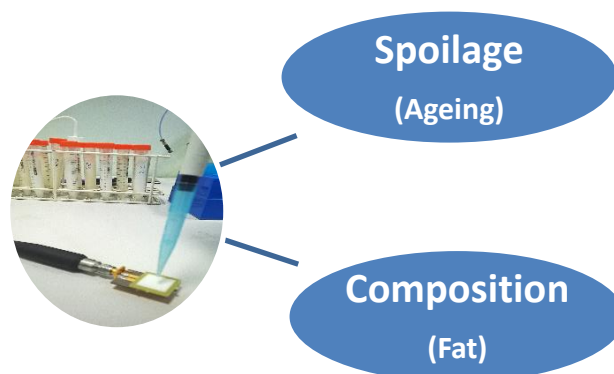
Figure 1. Milk samples are tested for their ageing and fat composition.

3. Current Advancements

Table 1 lists the main features of currently used advancements specifically for milk spoilage detection or protein and fat count at experimental level, with their limitations and advantages based on [7]. This work is an attempt to overcome the drawbacks that these existing standards have, with an alternative.

Table 1. Milk Quality Testing Techniques.

Techniques [a]	Attributes		
	Feature(s)	Limitation	Advantage
pH Indicators	pH Level Detection	Inaccuracy	Spontaneous
Disposable Taste Sensors	Lipids Based Spoilage Detection	Technical Complexity	Disposable Use
Gas-Sensor	Oxides Based	Costly and	Detects Specific

Techniques [a]	Attributes		
	Feature(s)	Limitation	Advantage
Arrays	Bacterial Detection	Complicated	Bacterial Growth
IR/NIR Spectroscopy	Spectral Based pH Level Detection	Expensive Setup	Very Accurate
Electrical Method	Amperometry Based Coliform detection	Continuous Supervision Required	Relatively Less Complex
Remote-Query Sensors	Wireless Bacterial Detection	Sensitive to Viscosity	Passive / Remote Detection
Optical Sensors	Photonic Spoilage Determination	Complex and Costly Setup	Platform for Real-Time Use
Protein/Fat Count	Protein / Fat Fluctuations Count	Different for Each Type of Fat/Protein	Wide Sensitivity Range

[a] Current State of the Art

4. Design Methodology and Measurements
4.1 Electromagnetic Wave Sensors
The antennas used for wireless and mobile communication operate on the EM wave transmission and reception. The EM wave based sensors are the same devices customized for different type of applications of sensing and detection. Wide range of variety exists within this group of sensors based on the way they are made, type of operation or range of frequency they work upon.

The material to be tested with this methodology is exposed to the electromagnetic field generated by the sensor fed through a source. The changes captured in the sensor's spectral responses, through a network analyser, then become indication of the differences among samples under consideration [11]. These differences could be based on the composition, i.e. milk types, or could be changes between the adulterated and fresh samples of the same milk product. Contaminated samples, e.g. by detergents, are also a subject of study in this work. Figure 2 shows microwave resonator cavity with a test-tube holder attached to its surface allowing the 15ml test-tube carrying milk sample to be held in place while quality monitoring takes place.

Figure 2. Microwave Resonator Cavity.

The resonant frequencies are different for TM and TE modes for a cylindrical cavity [12]:

- For TM Mode:

$$f_{mnp} = \frac{c}{2\pi\sqrt{\mu_r \epsilon_r}} \sqrt{\left(\frac{X_{mn}}{R}\right)^2 + \left(\frac{p\pi}{L}\right)^2}$$

(1)

- For TE Mode:
$$f_{mnp} = \frac{c}{2\pi \sqrt{\mu_r \epsilon_r}} \sqrt{\left(\frac{X'_{mn}}{R}\right)^2 + \left(\frac{p\pi}{L}\right)^2}$$
(2)

Where, R and L are radius and length for a cylindrical cavity, respectively. X_{mn} represents n^{th} zero of mth Bessel function and X'_{mn} represents n^{th} zero of derivative of mth Bessel function.

4.2. Sample Preparation and Measurement Conditions

As shown in figure 3(a), the three milk-types; skim, semi-skim and whole milk were used to make 5 samples each from and 15ml size polypropylene plastic test-tubes were used for sample preparation. The sample making was done at room temperature and the storage was governed by temperature controlled climate with minimum variations. To allow more number of points in the graph to avoid possible out of range errors, and to check linearity within varying fat values, two milk types were mixed in order of fat values.

For example, 50% skim milk was mixed with another 50% of the semi-skim milk, similarly, 50% of semi-skim milk was mixed with other 50% of whole milk. The storage conditions along with the measurement specifications are depicted in table 2. For the better accuracy of captured spectral data, more than 3 repetitions were made and then aggregated, for each of the 5 samples for all 5 categories of milk samples.

Table 2. Measurement Specifications / Storage Conditions.

No. of Measurements	5 Samples x 5 Repetitions = 25
Size of Samples	15 ml
Period of Supervision	1 week
Temperature	24 °C ± 2 °C
Frequency Sweep	10 MHz – 15 GHz
Channel Base Power	10 dBm
Milk Fat Contents (Grams/100ml)	Whole = 3.6g, Semi-Skim = 1.8g, Skim = 0.1g Skim + Semi-skim ≈ 0.95g, Semi-skim + Whole ≈ 2.7g

For the measurements a resonance cavity sensor was used as shown in figure 3(b) along with a VNA. A wide range of frequency sweep from 10 MHz – 15 GHz was applied. This allows more spectral data to be captured before sending to further analysis ensuring comparison of more spectral signatures within ultra-wide band of frequency for better results. The distinguishing frequencies were identified for all 3 types of milk to enable their classification. Same goes for the detection of spoilage and adulteration for a given type of milk and its comparison with previous day measurements to the following days in 1 week period. S_{11} and S_{21} parameters plots were captured, with 10dBm channel base power of VNA.
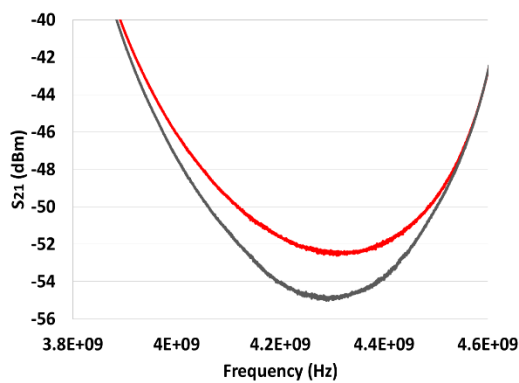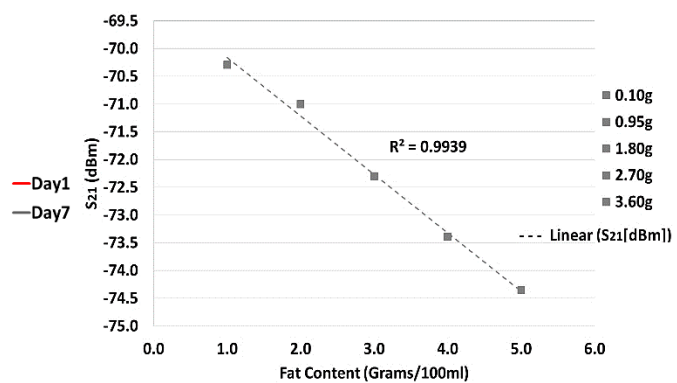
(a)

(b)

Figure 3. (a) 5 Sample Categories Prepared using 3 Milk Types. (b) Quality Test using EM wave Cavity

5. Quality Analysis Results

(a)

(b)

Figure 4. (a) Frequency vs. S_{21} plot for three milk types Day-1 (Red) and Day-7 (Dark Grey). (b) Linear Regression for 5 Milk Sample Categories with Varying Fat Content and Correlation 99.39%.

6. Discussion and Conclusion

Figure 4(a) shows clear distinction between sample from day-1 and day-7 of milk, it is interesting to note that the spoilage pattern regardless of fat content in each milk type is similar, which also compliments the studies made in the past and discussed in this paper. Figure 4(b), on the other hand, distinguishes among the 5 various samples made from 3 categories of milk. The figure shows spectral graph of S_{21} against the fat content of milk. The biggest advantage of this novel technique is, it does not only detect ageing but also distinguishes quite well among the types of milk as can be seen.

Hence, this technique establishes a promising platform for design and development of a milk quality testing and control system as it has several advantages over the other already existing standard practices, such as less complexity, higher accuracy, time and cost-effectiveness and universal approach to milk types and quality detection, unlike fat and protein count methods or gas sensor arrays where the design is selective to any given type of lipids/proteinases/bacteria.

This work presents a very good prototype model for a novel, online EM wave based quality testing system that can be made standalone and applied to large or medium scale retailers'

level of milk supply chain hierarchy, which should aid towards the problem solving that the dairy industries and end consumers face due to either complexity or costliness of existing standard practices for milk quality control.

7. References

[1]   Food and Agriculture Organization of the United Nations, Milk and dairy products in human nutrition. Rome: FAO Publications, 2013.

[2]   FAO, "Milk and milk products," FAO Publications, 2015. [Online]. Available: http://www.fao.org/agriculture/dairy-gateway/milk-and-milk-products/en/#.Vc_BJpXH_mQ. [Accessed: 01-Apr-2016].

[3]   FAO, "Milk processing," FAO Publications, 2015. [Online]. Available: http://www.fao.org/agriculture/dairy-gateway/milk-processing/en/#.Vc-_vpXH_mQ. [Accessed: 01-Apr-2016].

[4]   FAO, "Milk production," FAO Publications, 2015. [Online]. Available: http://www.fao.org/agriculture/dairy-gateway/milk-production/en/#.Vc_CHZXH_mQ. [Accessed: 01-Apr-2016].

[5]   E. Nyfors, "Industrial Microwave Sensors - A Review," Subsurf. Sens. Technol. Appl., vol. 1, no. 1, pp. 23–43, 2000.

[6]   S. H. H. Alhajeri, "Online Electromagnetic Wave Monitoring System for Petroleum Industry Applications," Liverpool John Moores University, 2010.

[7]   M. Lu, Y. Shiau, J. Wong, R. Lin, H. Kravis, T. Blackmon, T. Pakzad, T. Jen, A. Cheng, J. Chang, E. Ong, N. Sarfaraz, and N. S. Wang, "Milk Spoilage: Methods and Practices of Detecting Milk Quality," Food Nutr. Sci., vol. 04, no. 07, pp. 113–123, 2013.

[8]   P. Fellows and A. Hampton, Small-scale food processing - A guide for appropriate equipment. London: FAO, Intermediate Technology Publications, 1992.

[9]   NPCS Team, Market Research Report on Milk Processing & Dairy Products in India (Butter, Yogurt, UHT Milk, Cheese, Ice Cream, Ghee & Other Products). NIIR PROJECT CONSULTANCY SERVICES, 2014.

[10]   News Desk, "Raw Milk Recalled in California for Campylobacter Contamination," Food Safety News, 2015. [Online]. Available: http://www.foodsafetynews.com/2015/10/raw-milk-recalled-in-california-for-campylobacter-contamination/#.VnJ4MFduTOs.

[11]   A. Mason, S. C. Mukhopadhyay, K. P. Jayasundera, and N. Bhattacharyya, "Flexible Electromagnetic Wave Sensors," in Sensing technology : current status and future trends I, Cham ; New York: Springer, 2014.

[12]   T. P. Wangler, RF Linear Accelerators, 2nd and Complete Edition, Wiley, 2008.

Kannangara Wijekoon

Information requirement for managing built environment facilities

Christaline Wijekoon, Anupa Manewa, Andrew Ross and Matthew Tucker

Department of Built Environment,

Liverpool John Moores University,

United Kingdom

K.A.Wijekoon@2015.ljmu.ac.uk

R.M.Manewa@ljmu.ac.uk

A.D.Ross@ljmu.ac.uk

M.P.Tucker@ljmu.ac.uk

Abstract.
Facilities Management (FM) is a significant process, by which an organisation ensures that its facilities support core operations while contributing to organisation's strategic objectives. In fact, effective information management is one of significant drivers for enhancing continuous facilities management of built assets. However, the complex adaptive nature of construction influences the information management somewhat challenging. Most of the construction facilities are designed for long lifecycles with lengthy 'in-use' (FM) phases. The information flow within this phase is complex however highly necessary for effective facilities management.
Literature reveals Building Information Modelling (BIM) as an innovative method for managing information in construction facilities. However its application is considerably limited in the facilities management stage. This creates number of inefficiencies in the FM process. Therefore this research aims to identify how BIM can be used as an effective vehicle to manage facilities information which will enhance the value of built asset. This preliminary paper focuses on information requirement of facilities managers and their perceptions on usability of that information to complete FM functions.

Data were collected from 13 semi-structured interviews with construction professionals, who have over 5 years of industry experience. The data were analysed through coding (Open, Axial and Selective). Findings reveal three main categories of information based on their frequency of use. They are essential information, performance enhancing information and information with a potential future application. Moreover, it was highlighted that facilities managers tend to demand every item of information by assuming that information will help them to manage the facility in future. Cost of information management is another critical issue however it is not investigated within this paper. In summary, facilities managers believe BIM as an

effective and efficient tool to enhance the value of facilities information management.

1. Introduction

Facilities Management (FM) is a significant process, by which an organisation ensures that its facilities support core operations while contributing to organisation's strategic objectives. In fact, effective information management is one of the significant drivers for enhancing continuous facilities management of built assets.   However, the complex adaptive nature of construction influences the information management somewhat challenging. Most of the construction facilities are designed for long lifecycles with lengthy 'in-use' (FM) phases. The information flow within this phase is complex however highly necessary for effective facilities management.   Literature reveals Building Information Modelling (BIM) as an innovative method for managing information in construction facilities. However its application is considerably limited in the facilities management stage.   This creates number of inefficiencies within the FM process. Therefore this research aims to identify how BIM can be used as an effective vehicle to manage facilities information which will enhance the value of built asset. However this preliminary paper focuses on information requirement of facilities managers and their perceptions on usability of those information to complete the FM functions

2. Facilities information management

Facilities Management (FM) is the centre point of responsibility which ensures services of an organisation perform up to the agreed standards to support the core business performance to achieve business objectives (British Institute of Facilities Management, 2015). To be successful, a business should understand the impact of rising cost on building occupancy, services and workplace management over the business life cycle (Codinhoto and Kiviniemi, 2014). FM is dealing with large amount of building information; acquiring, updating and analysing (Wang et al., 2013).   On the other hand; facilities managers' spend lot of time in searching the required information (Jylhä and Suvanto, 2015). Therefore, capturing and storing required information related to the building is the initial success factor for a well-planned facility management (Akcamete et al., 2011).   In order to be successful in continuously growing, complex built environment, FM requires to manage the information produced by different stakeholders throughout building life cycle (Pittet et al., 2014).

This has become a more complex and challenging task due to increasing volume, continuous changes take place in information and variety of parties interested or using the same information (Zhao et al., 2008). as a solution, Wang et al. (2013) suggest the early engagement of FM construction stages of a built asset through BIM by explaining its positive impact over reduced operational cost. Yet, the early engagement of FM is rarely practiced in the reality. As a much practical solution, BIM is capable of facilitating initial data for FM which enhances the potentials of FM (Becerik-Gerber et al., 2012). BIM as a platform coming from the early stages of a building is a perfect solution for FM data management. What is more important is, BIM allows to communicate FM needs in early stages (British Institute of Facilities Management, 2012). The positive contribution of adopting BIM in

facilities information management is identified as a significant value addition associates with BIM (Gu and London, 2010). Eadie et al. (2013) highlight that facilities managers and client benefit the most out of BIM implementation. Conversely, from clients perspective, a considerable effort should be given to define client's FM needs at the early construction stages (Becerik-Gerber et al., 2012).

3. Methodology

Literature review was undertaken to identify the facilities information management in construction and also to establish the information value. 13 semi structured interviews (5 facility managers, 2 estate managers, 2 contractors, 1 architect, 1 surveyor, 1 BIM manager and a CAFM service provider) were conducted among construction industry professionals. The key purpose of these interviews was to identify the facilities information requirements and flows (in and out) during the facilities management stage and further to explore how different stakeholders with different interests recognise 'value' of information. Data were analysed qualitatively based on 3 stem coding process - Open Coding, Axial Coding and Selective Coding. Accordingly, themes generated within data were identified during Open Coding and similar themes were categorised based on their common features at the Axial Coding process. Finally, relationships between themes and categories were revealed while identifying information as the core category at the Selective Coding.

4. Data collection and analysis

The interview transcripts were analysed through coding process adopted from grounded theory research methodology. The first step of analysis engaged with Open Coding where the researcher looked for the themes generated from the data with an open mind. All possible themes were captured and categories were made grouping the similar themes together. These categories were coded by giving a name which represents the similarity in the themes (Figure 1). In Axial Coding, the properties and dimensions of categories were defined and relationship among categories were realised by going through the data once more. Memos were written during the analysis. Figure 1 is an extract from the first two steps of the analysis explaining the development of "Types of information" category.
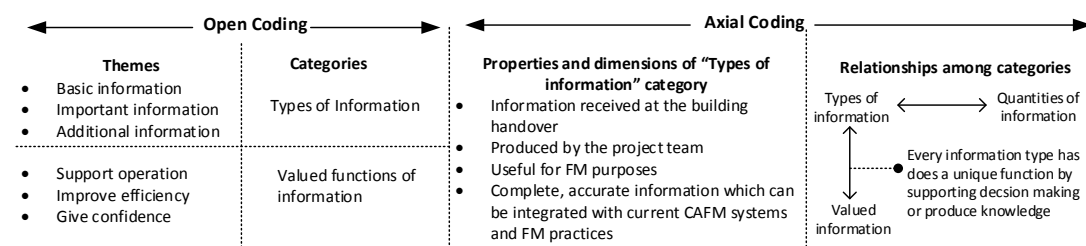
Figure 1: Extract from Data Analysis - Open and Axial Coding

Three kinds of information were highlighted by the respondents when referring to the construction information. A category named "types of information" was developed grouping these 3 types of information. Going through the interview transcript and memos highlighted the common and extreme characteristics such "*the lead contractor is liable to provide necessary documents in softcopies at the handover*" which was stated by a facilities manager who was used to determine the properties and dimensions of the category. For example, this quotation emphasise a dimension of the "types of information" category through the code "*at the handover*" by limiting the amount of factors that fall under the

category. Another example for a characteristic of the category was derived from a surveyor's statement *"not all information is accurate most of the time"* which made the point that information should be accurate to perceive its value. Once the categories were structured, key relationships were noted such as each type of information has a predictable quantity and a function. Then at the selective coding, "information" was selected as the core category and the relationship of this to other categories was formed to explain the developed concept.

To explain the three categories revealed under "types of information"; the first was coded as "essential" contains the information which are mandatory to have in order to run the facility. This is more towards adhering to the statutory compliance. Second category of information coded as "performance enhancement" is the group of information which are not essential to have with to operate the facility but having this additional information will help to provide an efficient service. The final category is "future potential". This includes the "*nice to know*" information which is explained by an information provider and also by the BIM manager. This does not have any practical application under current system but may have a potential contribution in the future with the system and technology changes.

5. Discussion

Current BIM standards and guidelines declares that its clients' responsibility to define necessary information to the project team through an Employer Information Requirement (EIR). Through the data analysis it was revealed that the responsible persons for defining EIR prefer to use a standard format and customising it as required rather listing out the project specific documents. BIM guidelines such as Construction Operations Building Information Exchange (COBie) and Industry Foundation Classes (IFC) provide such templates to understand the information requirement. However, facilities managers do not highly appreciate these templates. The negative attitude of users and complexity of these data templates can be highlighted as the key reasons for poor application of the currently available structured models. On the other hand, information users believe that they need all the information of a particular built asset to operate the building efficiently. Therefore, the core reason behind poor application of BIM in FM is lack of experience and understanding over information value. However, this is hidden behind the complexity of BIM and hard efforts to be made on changing the traditional practices.

Evidently, the information users maintain a positive relationship with the value and quantity of information. They are pleased to have more information in terms of volume, robustness and quality rather having less. However, getting to details of why users prefer to have more information revealed that the complete set of information would do nothing much more than giving confidence to the information users. As a respondent answered "*we never know what information will be needed with the possible future changes….. Never thought it will be necessary to know the content of asbestos*" Based on this fact, they are willing to acquire every possible item of information about the building, which assures their readiness for unforeseeable futures. Contradictorily, information users are not very much sure about the application of this "future potential" information as they currently possess no experience of handling such a rich information base about the built asset.

Although, information users theoretically prefer to have a complete set of information about the building none of them had such database or poor readiness to adopt such in practice. Currently, the buildings are operated and maintained up to a satisfactory level with the

available "essential" information. Therefore, they find no interest to put much effort on gathering more information although there are inefficiencies due to lack of information. The reason given for this difference in what facility managers prefer to have theoretically and what information they practically have is the heavy work load beyond information management and difficulties in maintaining the information.

The interesting point made through this study was the experience of information users when they had to develop their own information database for old buildings which had no information on hand by the time of purchase. It was accepted that without handover/as-built information facilities managers face many difficulties in operating and maintaining the building. Therefore, at the initial stage they make use of their experience and make the decisions over the building and ultimately develop the building information database as they get to know the features of the considered building.

BIM practitioners hold evidence for the value addition gained through BIM. If the same amount of information is handed over through BIM, building operators are capable of overcoming the basic problems faced in traditional method. It will bring in many opportunities by having digitized information. Value addition through digitized information will enable the long term use of information, inter-operability within Computer Aided Facilities Management (CAFM) systems and make it easy to update and manage information. However, value of information through BIM exceeds more as it is capable of handling more information and brings value through graphical demonstration of information. Also, with the future expectations of integrating BIM enhances its capabilities. This takes the value of information to the next level as this one way information feeding loop will become a cycle going beyond feeding information to BIM but also generating information through BIM.

Conclusion

Findings reveal three main categories of information based on the purpose and frequency of use. Therefore, the construction information categories in FM are essential information, performance enhancement information and information with a future potential application. Moreover, it was highlighted that facilities managers tend to demand for every item of information which includes all three types by assuming that information will help them to manage the facility in future. Although an exact application of most of the information is not clear at the moment this demand is made. Cost of information management is another critical issue coming along when deciding which information to take forward to FM stage. However, it is not investigated within this paper.

BIM is an emerging trend in construction industry bringing many advantages and improvements. With the growing number of research under this topic and real case studies, there is no doubt that positive contribution of BIM overrides its negatives (Akhurst and Gillespie, 2006). Having identified the potentials of BIM, construction industry is now pushing towards a whole life application. Considered to it, the potential uses of BIM for FM is very much beneficial. However, there is not much interest within facilities managers to adopt BIM. It is due to the lack of experience and understanding over information value. The importance of knowing the value of information will promote the adoption of BIM beyond construction and limits its application at the economical point rather overloading information to fill the BIM capabilities.

Reference

Akcamete, A., Akinci, B. & Garrett, J. H. 2011. Potential utilisation of building information models for planning maintenance activities. *In:* Tizani, W. (ed.) *International Conference on Computing in Civil and Building Engineering.* Nottingham University Press.

Akhurst, P. & Gillespie, N. 2006. Sydney Opera House: Facilities Management exemplar project. *FMA Conference.* Australia: CRC for Construction Innovation.

Becerik-Gerber, B., Jazizadeh, F., Li, N. & Calis, G. 2012. Application areas and data requirements for BIM-enabled facilities management. *Journal of Construction Engineering and Management,* 138(3)**,** pp 431-442.

British Institute of Facilities Management. 2012. BIM and FM: Bridging the gap of success, British Institute of Facilities Management (Herts).

British Institute of Facilities Management. 2015. *Facilities Management Introduction* [Online]. Available: http://www.bifm.org.uk/bifm/about/facilities [Accessed 05 May 2015].

Codinhoto, R. & Kiviniemi, A. 2014. BIM for FM: A Case Support for Business Life Cycle. *In:* Fukuda, S., Bernard, A., Gurumoorthy, B. & Bouras, A. (eds.) *Product Lifecycle Management for a Global Market.* Springer Berlin Heidelberg.

Eadie, R., Browne, M., Odeyinka, H., McKeown, C. & McNiff, S. 2013. BIM implementation throughout the UK construction project lifecycle: An analysis. *Automation in Construction,* 36(145-151.

Gu, N. & London, K. 2010. Understanding and facilitating BIM adoption in the AEC industry. *Automation in Construction,* 19(8)**,** pp 988-999.

Jylhä, T. & Suvanto, M. E. 2015. Impacts of poor quality of information in the facility management field. *Facilities,* 33(5/6)**,** pp 302-319.

Pittet, P., Cruz, C. & Nicolle, C. 2014. An ontology change management approach for facility management. *Computers in Industry,* 65(9)**,** pp 1301-1315.

Wang, Y., Wang, X., Wang, J., Yung, P. & Jun, G. 2013. Engagement of Facilities Management in Design Stage through BIM: Framework and a Case Study. *Advances in Civil Engineering.*

Zhao, Y., Tang, L. C. M., Darlington, M. J., Austin, S. A. & Culley, S. J. 2008. High value information in engineering organisations. *International Journal of Information Management,* 28(1)**,** pp 246-258.

Phillip Taylor

A review of the capabilities and limitations of general experimental techniques, for the observation of drag reducing mechanisms, and the determination of skin friction in turbulent boundary layers over textured surfaces.

Phillip Stephen Taylor
General Engineering Research Institute, LJM University, Byrom Street, Liverpool, L3 3AF
P.Taylor@2011.ljmu.ac.uk
Professor Xun Chen
General Engineering Research Institute, LJM University, Byrom, Street, Liverpool, L3 3AF
X.Chen@ljmu.ac.uk

Abstract.
The fuel consumption of a moving vehicle, for example, an aircraft, boat or car, is dependent on the drag experienced by the external surface of the vehicle. As fluid flows over a solid surface, the fluid in the near-wall region experiences high shear forces. This region, known as a boundary layer, contributes significantly to the drag force acting on the surface. Past research has shown that manufacturing textured geometries onto a smooth surface can reduce the drag force that is induced by the boundary layer. Changes in the boundary layer structure that are caused by the presence of the textured geometries can be observed in the near-wall region. This paper investigates the suitability of two common velocity measurement techniques, hot-wire anemometry and pressure probes, for experimental analysis of boundary layers over drag-reducing textured surfaces. The potential drag reducing mechanisms, and key fluid properties relating to the changes in drag, are compared with the accuracy and precision of the considered measurement techniques. Additional sources of error, due to the presence of the textured roughness geometries, are identified.

Keywords. Turbulent boundary layer, Riblets, Roughness, Drag reduction, Hot-wire anemometry, Pressure probes, Roughness function.

1.0 Introduction

In the case of fluid flowing over a solid surface, friction based drag forces (viscous forces) act on the surface. The forces are generated by the fluid in a thin region adjacent to the wall, referred to as a turbulent boundary layer. Textured surfaces, with roughness elements of certain geometries, have been found to reduce the viscous force, and hence the drag force acting on the surface [1]. The resultant drag reduction can significantly reduce the energy and fuel consumption in a wide range of engineering applications.

The properties of the flow in the boundary layer can be used to obtain values of the viscous forces acting on the surface. Changes in the turbulent boundary layer between a smooth wall and a textured wall could provide an insight into the fluid mechanisms which lead to drag reduction on a textured surface. A turbulent boundary layer is divided into two regions; the inner region and the outer region [2]. The properties of the flow in the inner region are independent of the properties of the main body of the flow, far away from the wall. The outer

region is a fully turbulent region which makes up 80% of the total boundary layer thickness. The properties of the fluid in this region are typically unaffected by the presence of roughness on a surface. Many properties in the inner region, such as velocity u, and distance from the wall y, can be scaled using the value of skin friction at the wall τ_0, which is a parameter representative of the viscous drag. The velocity and wall distance can be represented as dimensionless values, u^+ and y^+ which are defined by (1). Although dimensionless, the quantity of a scaled variable in a turbulent boundary layer is referred to in terms of 'viscous units'.

$$u^+ = \frac{u}{\sqrt{\tau_0/\rho}} \,, \quad and \quad y^+ = \frac{y}{v}\sqrt{\frac{\tau_0}{\rho}}, \quad (1)$$

Where $\Box$ is the density of the fluid and v is the viscosity of the fluid. The inner region is further divided into the viscous sub-layer and the logarithmic region. The viscous sub-layer is located adjacent to the wall, and has a linear relationship between the scaled velocity and scaled wall-distance, (2) [3]. The logarithmic region is governed by the linear relationship given in (3) [4].

$$u^+ = y^+, \quad (2)$$

$$u^+ = C_1 \ln(y^+) + C_2 - \Delta U^+, \quad (3)$$

The roughness function ΔU^+ in (3), represents the difference in the value of u^+ between a smooth and a rough surface, at a point in the logarithmic region. For a smooth wall boundary layer $\Delta U^+ = 0$. A single velocity value which is obtained in the viscous sub-layer can be used to determine a value of skin friction, the accuracy of which depends only on the accuracy of the velocity and wall-distance measurements. Alternatively, the skin friction can be determined from the log-law of the wall relation (3), using a series of measured points in the logarithmic region. The latter method is often used due to the difficulty of obtaining measurements in the thin viscous sub-layer. However, this method contains additional uncertainty as the constant C_1 must be known, and can be flow dependant. The constant C_2 is determined from the boundary layer over a smooth surface, however, the accuracy of C_2 depends on the accuracy of the known value of C_1.

Early drag reducing surfaces consisted of 2-dimensional riblets with cross-sections of various shapes, aligned in the streamwise direction of the flow [5]. Textured surfaces with triangular riblets, aligned in the streamwise direction, have produced drag reductions of up to 8% [6]. Drag reduction occurred for cases in which both the height and spacing of the riblets were less than 25 viscous units. A greater drag reduction of 10% has been observed for thin, blade riblets with a spacing of 18 viscous units [7]. Berchet et al. [8] demonstrated the potential for drag reduction over a surface with 3-dimensional roughness geometries. The surfaces consisted of staggered blades of a finite length. A significant drag-reduction of 7% was achieved. It was noted that, despite the significant drag reduction, these riblets were unsuitable for practical application and manufacture due to their complex geometry. Experiments on simple blade riblets observed an 'upwards shift' of the velocity profile, providing a negative roughness function [9]. A similar roughness function was observed by Baron and Quadrio [10] for triangular shaped riblets. These experimental studies attributed the behaviour of the roughness function to an increase in the viscous sublayer thickness. This idea was later confirmed by Wang et al. [11], when experiments performed on surfaces with longitudinal v-shaped grooves observed that the viscous sub-layer thickness increased by up to 10%. This was accompanied by a negative roughness function.

Experimental analysis of a boundary layer requires accurate determination of the origin of the velocity profile y_0, at which $u = 0$. For smooth wall boundary layers, the origin of the

velocity profile is naturally located at the solid wall, $y = 0$. In the case of 2-dimensional, longitudinal riblets, the origin of the velocity profile is not located at the plane of the roughness crests. Instead, the origin is offset by a small distance below the crests, at a location known as the 'virtual origin' [5]. The concept of a virtual origin is shown in figure 1.
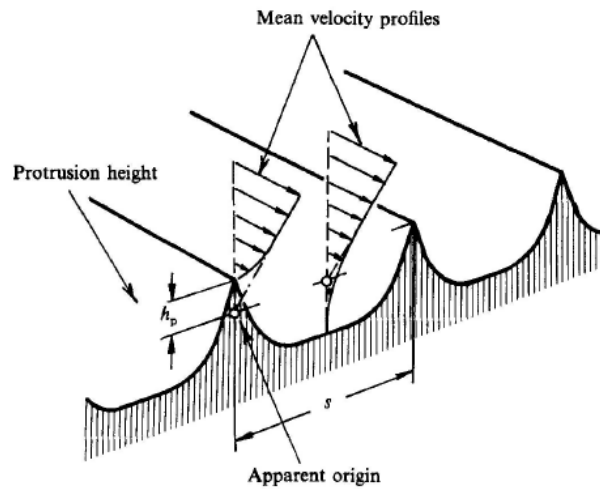

Figure 1. A turbulent boundary layer velocity profile for flow over grooved, blade riblets. The virtual origin of the velocity profile (apparent origin) is located below the riblet tips (Bechert and Bartenwerfer, 1987).

This review aims to identify key properties which must be obtained, and important sources of error, in the experimental analysis of drag reducing riblet surfaces. Past literature and published results are considered for cases of both drag reducing and drag increasing surfaces. The observed turbulent mechanisms which could potentially lead to drag reduction, are identified. Additional sources of experimental error resulting from the geometries of the textured surfaces, are discussed.

2. Mechanisms of drag reduction
2.1 Causes of drag reduction
Past numerical analyses for a range of drag reducing surfaces have observed important fluid mechanisms, which can be attributed to the decrease in viscous drag. Numerical simulations for triangular riblets suggested that the riblets restricted fluid flow and reduced the movement of fluid in the spanwise direction [12]. For straight blade riblets, Martin and Bhushan [13] found that, for drag reducing regimes, turbulent vortices were lifted off the surface by the riblet tips. The surface area of the wall which was exposed to the turbulent vortices was significantly decreased. When the riblet spacing was greater than the diameter of the vortices, the vortices were located in the gaps between riblets, resulting in an increased skin friction. In disagreement with the result by Choi et al. [12], no significant changes were observed in the spanwise movement of fluid between the drag reducing and drag increasing regimes. Garica-Mayoral and Jimenez [14] examined longitudinal blade riblets with various values of riblet spacing. For cases of drag reduction, recirculating vortices were located in the grooves between riblets. For cases of drag increase, which were observed for increased spacing, these vortices broke down and larger vortices appeared above the riblet crests, as illustrated in figure 2. These larger vortices generated significant amounts of additional stress in the boundary layer.
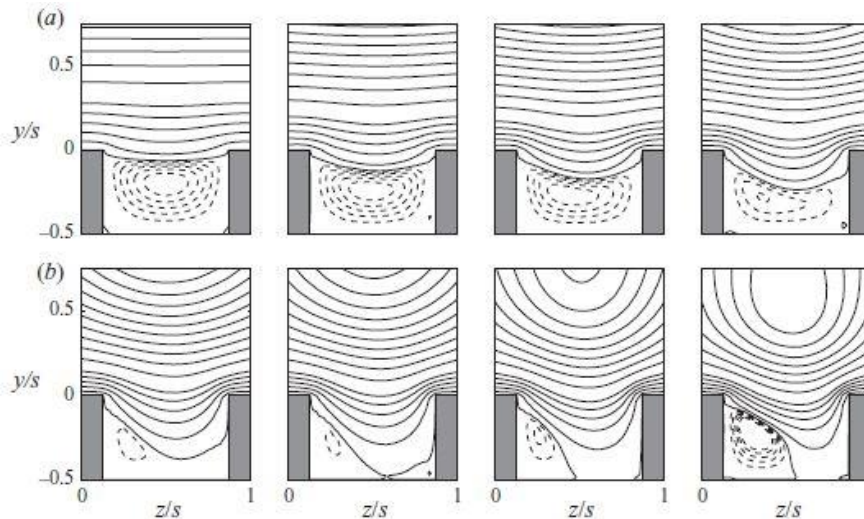
Figure 2. A cross-sectional view of the velocity streamlines for a) a drag decreasing regime and b) a drag increasing regime. The streamlines illustrate the growth of turbulent structures during the breakdown of drag reducing regime. (Garcia-Mayoral and Jimenez, 2010).

2.2 Prediction of changes in drag

The design and development of drag reducing surfaces requires parameters which can be used to predict the resulting change in drag. A logical first step was to compare the dimensions of individual roughness geometries (i.e. spacing, height, width), with the observed roughness function. Orlandi et al. [15] performed numerical analysis of transverse and longitudinal riblets of various shapes. There was no observed correlation between the dimensions of the textured geometries and the roughness function, indicating that a pre-known geometry was insufficient for predicting drag reduction.

Expanding analysis to a roughness of 3-dimensional staggered cube arrays, Orlandi and Leonardi [16] observed that significant fluctuations of the wall-normal velocity v, occurred in the plane of the roughness crests. In this region the scaled, mean value of wall-normal velocity $\bar{v}_0^+$, was constant. Plotting $\bar{v}_0^+$ against the roughness function produced similar trends for the 3-dimensional and 2-dimensional geometries considered. Furthermore, Lee et al. [17] found that $\bar{v}_0^+$ could be used to predict the effect of riblets in the outer region of a turbulent boundary layer. There existed a critical value of $\bar{v}_0^+$, below which the properties of the flow in the outer layer were unaffected by the presence of the roughness. Above this critical value, the turbulent stresses in the outer region increased. In agreement with Orlandi and Leonardi [16], Lee et al. [17] observed linear relationship between the roughness function and $\bar{v}_0^+$, further indicating that $\bar{v}_0^+$ is a suitable parameter for predicting the drag reducing or drag increasing properties of textured surfaces.

2.3 Form drag

The viscous drag is one of two components which contribute to the total drag. The other component is form drag, which results from a difference in pressure across a solid body in the flow. In the case of 2-dimensional riblets, the cross-section of the surface is constant in the stream wise direction, resulting in a zero value of form drag. For 3-dimensional riblets, the contribution of the form drag can be significant. Leonardi and Castro [18] investigated the viscous drag and form drag components for 3-dimensional cubic roughness with a large roughness height. The viscous drag was found to be only 7% of the total drag. Similar cube arrays were studied by Claus et al. [19], with varying angles of orientation to the flow. In all angles, there was significant form drag, with the minimum value occurring for an array with the cubes aligned (as opposed to staggered) in the streamwise direction. Moving on from straight edged cubes, Chatzikyriakou et al. [20] investigated the effects of hemispherical roughness elements in an aligned array, which had the potential to reduce form drag. No

overall drag reduction was observed, with the skin friction increasing significantly by up to 36%.

3. Comparison of Experimental Techniques

3.1 Hot-wire anemometry

Hot wire anemometers measure the instantaneous flow velocity using the convection properties of fluids. A thin wire, supported between two prongs, is placed into the flow. The temperature of the wire, and hence its resistance, is dependent on the velocity of the fluid. Velocity values can be obtained by measuring the current which is passing through the wire and converting these values into velocity values through a calibration relation. Importantly, hot-wire anemometers can be used to determine values for the both the time-averaged velocity, and time dependant velocity fluctuations of a flow. The velocity fluctuations can be used to obtain the turbulent properties of the flow. Accurate measurement of velocity fluctuations requires the wire length to be smaller than 20 viscous units $(L^+ < 20)$, to observe the behaviour of the small scale turbulent eddies [21]. Increasing the wire length beyond this value can cause error in the measured friction velocity of up to 30% [22]. The small-scales of textured geometries place additional constraints on wire length. Past literature, discussed in section 3.2, suggested that drag reducing or drag increasing properties of a textured surface can be categorised by the scaled wall-normal velocity, at the plane of the riblet crests. To obtain an accurate value of $\bar{v}_0^+$ experimentally, the length of the wire must be significantly smaller than the spanwise spacing between roughness elements. Some drag reducing regimes have been found to occur for a spacing as small as 5 viscous units [7].

On a smooth wall surface, a hot wire anemometer in close proximity to the wall will experience significant error. The presence of the wire in the fluid causes a wake to appear behind the probe. When this wake makes contact with the solid surface there is additional heat loss into the wall [23]. In the case of conducting walls, this 'wall effect' can occur within the outer limit of the viscous sub-layer, $y^+ \approx 5$. Additional error is introduced from the wall distance y, which occurs due to an inaccurate determination of the wall location [24]. Methods have been developed, which exploit the wall effect to accurately determine the location of the surface [25]. As the prongs make contact with the wall there is a sudden change in the measured voltage, due to conduction of heat into the wall. This method was found to predict wall location up to an accuracy of ± 0.018 viscous units. Even in the case of complex textured geometries, the voltage changes will still occur when a prong contacts any part of the roughness surface. Since the geometries of the roughness will be pre-known, the method by Hutchins and Choi [25] can be easily adapted for use on textured surfaces. Ideally, the hot-wire anemometer should have an accuracy and precision which can observe the presence of turbulent vortices located in the grooves between riblets [13], or near to the roughness crests [14].

3.2 Pressure probes

The pressure probe is an alternative, and well established, method of measuring mean velocity. A tube is placed in the flow, with its opening aligned with the direction of the flow. The pressure of fluid flowing into the tube is measured and compared to the pressure of static fluid to obtain the stream wise velocity. Pressure probes have several key advantages over hot-wire probes. The sensitivity of a pressure probe to an error in the tube alignment is very low and, provided that the tube is sufficiently small, the effects of spatial averaging across the opening of the tube are often negligible [26]. In addition, pressure probes can be placed in direct contact with the wall, without experiencing physical damage or conduction effects. Pressure probes are accompanied by additional error, due to their physical presence in the flow. This disadvantage is offset by detailed correction equations which account for flow disturbance (viscous) effects and turbulence effects [27]. The sources of error are easier account for or eliminate, compared to errors for hot-wire probes.

Corrections for turbulence effects require an accurate estimation of the turbulent intensity at the point of measurement in the flow. Bailey et al. [28] showed the importance of turbulence corrections for a pressure probe placed in the inner-region. Applying a viscous correction produced velocity measurements with an error of 5% when located in the near-wall region. Applying a turbulence correction significantly reduced this error to approximately 3%. For velocity measurements obtained in the outer region, the correction equations can produce values with a significantly reduced error of less than 1%. Whilst pressure probes cannot be used to determine the time-dependant properties of a fluid, such as the instantaneous velocity fluctuations, their high accuracy outside the near-wall region makes than suitable for the validation of the results of hot-wire probes.

3.3 Skin friction

The skin friction, and hence the viscous drag, is often determined from velocity measurement taken in the logarithmic region using the relation in (2). Any error or uncertainty in the values of u and y, or the flow dependant constants C_1 and C_2 will contribute to the error in the skin friction value [29]. Alternatively, a single pressure measurement, obtained from a pressure probe located adjacent to the wall can be converted into a value for skin friction [30]. As a pressure probe is often unable to obtain an accurate measurement in the viscous sub-layer, this method relies on a pre-known relation in the log-law region. For accurate analysis of a surface with an unknown log-law relation, as is the case with an unknown roughness function, the skin friction must be determined by a method which is independent of the relation in the logarithmic region [31]. Hutchins and Choi [25] demonstrated that a single velocity value taken near the outer limit of the viscous region $3.5 < y^+ < 5$, could determine the skin friction to an accuracy of +/-1.04%. This is provided that the wall location is determined using the method outlined in Section 3.1. In the case of textured surfaces, the accuracy and validity of this method will depend on the location of the virtual origin. If the virtual origin lies less than 1.5 viscous units below the roughness crests, then it is possible to determine the skin friction from a velocity measurement obtained in the region $3.5 + h_0 < y^+ < 5$, where h_0 is the protrusion height (see figure 1). However, additional error will be introduced as the accuracy will depend on accurate measurement of both the location of the roughness crests and the virtual origin.

A direct measurement of total drag can be obtained by a force balance supporting the surface, or a section of the surface. For 2-dimensional riblets, a direct value of skin friction would allow accurate scaling of the inner region. The accuracy of the determined roughness function would be increased, as it would depend only on the accuracy of the u and y measurements, and the accuracy of the flow dependant constant C_1. In the case of 3-dimensional roughness, the difference between the measured total drag, and the viscous drag, determined from (2) or (3), would provide a value of the form drag. Clearly, this value would be more accurate if the skin friction can be determined from the viscous sub-layer, (2).

4. Conclusions

The capabilities of hot-wire anemometers and pressure probes have been discussed with regards to the observable drag reducing and drag increasing mechanisms in turbulent boundary layers. Past numerical studies have connected drag reducing effects to turbulent vortices in the near-wall region. These vortices can be located either in the grooves between riblets, at the roughness crests or, in some drag increasing cases, above the roughness crests. A detailed experimental analysis would require accurate measurements of instantaneous, time-dependant velocity within the grooves, and in the region immediately above the roughness crests. Furthermore, past results have indicated that the roughness function, and hence viscous drag reduction, correlates to the mean wall-normal velocity in the plane of the roughness crests. The accuracy of a hot-wire probe located in this plane will the reduced by the wall effect, resulting from the close proximity of the wire to the roughness crests. If possible, samples of textured surfaces should be constructed from non-conducting materials. In this case, the error in near-wall velocity measurements, due to the wall effect

could be considered negligible.

A hot-wire anemometer has been shown to produce a skin friction value to an accuracy of approximately 1%. This value can be obtained from a single velocity measurement in the viscous sub-layer. Methods which exploit the conduction effects of the hot-wire probe are capable of determining the location of the solid wall to within +/-0.018 viscous units. With this method, the error in the wall distance could be considered negligible. It can be expected that the accuracy will be less for flows over a textured surface, due to the need to determine the location of the virtual origin. To avoid significant inaccuracies from the wall effect, a hot-wire probe should be used to obtain the velocity value in the upper part of the viscous sub-layer, but above the roughness crests. Pressure probes, with a predicted error of 1%, are suitable for the validation the mean velocity measurements from the hot-wire anemometer, in the outer region. Proper correction of the turbulent effects in this region will be possible, provided that the roughness does not affect the turbulence intensity in this outer region, as is the case with certain cube arrays. When pressure probes are placed within the inner region, the corrected velocity measurements can only be obtained to an accuracy of 3%. The unknown turbulent structures and turbulent intensity distributions near a textured surface will make accurate turbulence correction difficult in the inner-region of the boundary layer. It is clear that reliable measurements of skin friction from the velocity profile can only be obtained using a hot-wire anemometer, placed in the viscous sub-layer. The presence of textured geometries on the surface can have a variety of effects on the properties of turbulence in the boundary layer. Corrected pressure probe measurements alone cannot be relied upon to produce an accurate mean velocity profile for either the inner region of outer region of the boundary layer. However, as the sources and magnitudes of errors in pressure probes are often well understood, it is recommended that pressure probes are used alongside hot-wire anemometers where possible. Provided that the wire is of a suitable length $L^+ < 20$, the turbulence properties from the hot-wire anemometers could provide an approximate turbulence correction needed for the pressure probe. In turn, the corrected pressure probe readings could be used to assess the accuracy of the mean velocity values obtained from the hot-wire anemometer.

For 2-dimensional riblets, the total drag can be determined from viscous sub-layer measurements alone. However, for a reliable analysis drag reducing surface, the total drag measurements should be obtained independently to account for any uncertainty from measurements of the boundary layer. In the case of 3-dimensional roughness geometries, the form drag can reduce the total drag reduction, or even create a total drag increase. Total drag measurements must be measured directly to allow clear identification of the form drag component.

References

[1] Dean B and Bhushan, B 2010 Shark-skin surfaces for fluid-drag reduction in turbulent flow: a review. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* **368** 4775--806.
[2] Bradshaw P and Huang G P 1995 The law of the wall in turbulent flow *Proc.: Mathematical and Phys. Sciences* **451** 165--88
[3] Schlichting H 1979 *Boundary Layer Theory* ed F J Cerra (McGraw-Hill) p 603
[4] Schultz M P and Flack K A 2007 The rough-wall turbulent boundary layer form the hydraulically smooth to the fully rough regime *J. of Fluid Mech.* **580** 381-405
[5] Bechert D W and Bartenwerfer M 1989 The viscous flow on surfaces with longitudinal ribs *J. of Fluid Mech.* **206** 105--29
[6] Walsh M J 1982 Turbulent boundary layer drag reduction using riblets *20th Aerospace Sciences Meeting*
[7] Bechert D W, Bruse M, Hage D W, Hoeven J G T and Hoppe G 1997 Experiments on drag-reducing surfaces with their optimization with adjustable geometry *J. of Fluid Mech* **338** 59--87
[8] Bechert D W, Bruse M and Hage W 2000 Experiments with three-dimensional riblets as an idealized model of shark skin *Experiments in Fluids* **28** 403--12
[9] Choi K 1989 Near-wall structure of a turbulent boundary layer with riblets *J. of Fluid Mechanics*

208 417--58

[10] Baron A and Quadrio M 1993 Some preliminary results on the influence of a turbulent boundary layer over riblets *International J. of Heat and Fluid Flow* **14** 223--30

[11] Wang J-J, Lan S-L and Chen G 2000 Experimental study on the turbulent boundary layer flow over riblets surfaces. *Fluid Dynamics Research.* **27** 4, 217--29

[12] Choi H, Moin P and Kim J 1993 Direct numerical simulation of turbulent flow over riblets *J. of Fluid Mechanics* **255** 503--39

[13] Martin S and Bhushan B 2014 Shark skin inspired low-drag microstructure *J. of Fluid Mech.* **756** 5-29

[14] Garcia-Mayoral R and Jimenez J 2014 Hydrodynamic stability and breakdown of the viscous regime over riblets *J. of Fluid Mechanics* **678** 317--47

[15] Orlandi P, Leonardi S and Antonia R A 2006 Turbulent channel flow with either transverse or longitudinal roughness elements on one wall *J. of Fluid Mechanics* **561** 279-305

[16] Orlandi P and Leonardi S 2006 DNS of turbulent channel flows with two- and three-dimensional roughness *J. of Turbulence* **7** N73

[17] Lee J H, Seena A, Lee S and Sung H J Turbulent boundary layers over rod- and cube-roughened walls *J. of Turbulence* **40** 1-26

[18] Leonardi S and Castro I P 2010 Channel flow over large cube roughness: a direct numerical simulation study *J. of Turbulence* **651** 519--39

[19] Claus J, Krogstad P-A and Castro I P 2012 Some measurements of surface drag in urban-type boundary layers at various wind angles. *Boundary-Layer Meteorology* **145** 407--22

[20] Chatzikriakou D, Buongiorno J, Caviezel, D and Lakeha D 2015 DNS and LES of turbulent flow in a closed channel featuring a pattern of hemispherical roughness elements *International Journal of Heat and Fluid Flow* **53** 29--43

[21] Ligrani P M and Bradshaw P 2000 Spatial resolution and measurement of turbulence in the viscous sublayer using sub miniature hot-wire probes *Experiments in Fluids* **5** 407--17

[22] Orlu R and Alfredsson P H 2010 On spatial resolution issues related to time-averaged quantities using hot-wire anemometry. *Experiments in Fluids* **49** 101--10

[23] Bhatia J C, Durst F and Jovanovic J 1982 Corrections of hot-wire anemometer measurements near walls *J. of Fluid Mech.* **122** 411--31

[24] Orlu R, Fransson J H M and Alfredsson, P H 2010 On near wall measurements of wall bounded flow: the necessity of an accurate determination of the wall position. *Progress in Aerospace Sciences* **46** 8 353--87

[25] Hutchins N and Choi K-S 2002 Accurate measurements of local skin friction coefficient using hot-wire anemometry *Progress in Aerospace Sciences* **38** 421--46

[26] Chu S H 1975 Pressure probes for fluid measurement *Progress in Aerospace Sciences* **16** 147-223

[27] McKeon B J, Li J, Jiang W, Morrison J F and Smits A J 2003 Pitot probe corrections in fully developed turbulent pipe flow *Measurement Science and Technology* **14** 1449---58

[28] Bailey S C C et al 2013 Obtaining accurate mean velocity measurements in high Reynolds number turbulent boundary layers using Pitot tubes *J. Fluid Mech.* **715** 642--70

[29] Alfredsson P H, Imayama S, Lingwood R J, Orlu R and Segalini 2013 Turbulent boundary layers over flat plates and rotating disks – The legacy of the Von Karman: A Stockholm perspective *European J. of Mech. B/Fluids* **40** 17-29

[30] Patel V C 1965 Calibration of the Preston tube and limitations on its use in pressure gradients *J. of Fluid Mech.* **23** 185-208

[31] DeGraaf D B and Eaton J K 2000 Reynolds-number scaling of the flat-plate turbulent boundary layer *J. of Fluid Mech.* **422** 319--46

James Wharton

A Numerical Investigation into the Frictional Performance of Ground, Micro-structured Surfaces

J T Wharton, X Chen and D R Allanson
Advanced Manufacturing Technology Research Laboratory, General Engineering Research Institute, Liverpool John Moores University, Byrom St., Liverpool, L3 3AF
E-mail address: j.wharton@2009.ljmu.ac.uk

Abstract. Previous to this investigation, a study was carried out to prove that micro-scale textures could be manufactured onto a bearing surface using the cylindrical grinding process. A numerical investigation was performed to model oil flow over these ground surface patterns in order to optimize the surface geometry for texture-induced lift and frictional reduction properties. This report describes a newly developed python based script that links the computer aided design SALOME$^{©}$ GEOM module with OpenFOAM$^{©}$'s snappyHexMesh hexahedral meshing software and the SimpleFOAM iterative-based solver. This script enables the user to analyse the performance of a particular textured surface, automatically. A parametric study was performed using the newly developed python control script to investigate the performance of a bearing operating with a linear speed of 5ms^{-1}. The parametric study was based on a set of simulations which were designed using the Taguchi method. The results of the simulations were then used to produce a set of direct effect charts so that the effect of surface texture parameters on bearing performance indicators could be clearly identified.

Keywords. Functional surfaces, textured surfaces, grinding, hydrodynamic bearing, friction, surface roughness, computational fluid dynamics.

1. Introduction

One of the biggest challenges facing today's engineers is the need to reduce our day-to-day energy consumption and the use of the planet's limited fossil fuel supply. Transport infrastructure is one of the biggest consumers of fossil fuels and 33% of the energy lost in combustion is due to friction between mechanical components, this equated to 180,000 million litres of fuel in 2012 worldwide [1]. So the need to find efficiencies in engineering systems is of great importance and the reduction of losses associated with friction using surface engineering technologies may provide some of the answer. Due to recent developments in the fields of metrology and manufacturing, micro-scale engineered surfaces are possible to create, repeatedly, using modern technologies and processes [2]. However, further research needs to be carried out in order to identify the potential benefits of these surfaces and also improve the manufacturing processes, to make them more accessible to industry. Chen et. al. (2015) [3] showed a cylindrical grinding process that could produce micro-scale textures onto a surface using a novel dressing operation. The shape of these pockets was investigated and was found to be ellipsoidal. This investigation aims to study

the interaction of oil flow over these ground surfaces and optimise the surface geometry in order to see if there are any potential benefits in relation to surface drag reduction for a hydrodynamic bearing application.

2. The Modelling Process

This section of the report discusses how the model was setup and how the python script interacts with OpenFOAM© and SALOME© to produce a set of results.

2.1. Couette Flow

In order to simplify the problem, the whole bearing was not modelled, only a finite section of the surface interaction is investigated, which was simplified as couette flow (figure 1). This type of flow makes the following assumptions:

- No flow in y or z-direction;
- Steady state;
- Distance is infinitely long in x-z plane;
- Fully developed flow;
- No pressure drop in the x-direction;
- At y=0, $\boldsymbol{u}$ = (0,0,0), i.e. no-slip boundary condition;
- At y=H, $\boldsymbol{u}$ = (U$_x$,0,0), i.e. moving wall boundary condition.

Therefore, if dp/dx =0, the Navier-Stokes equation simplifies to equation 1.
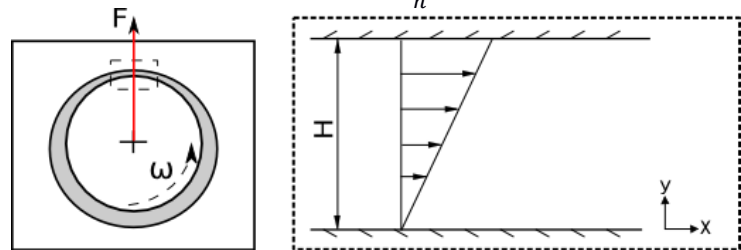
$$u(h) = \frac{u \cdot y}{h} \qquad (1)$$

Figure 1. Couette flow schematic.

Equation 1 indicates that for a non-textured surface, a linear velocity gradient exists across the film thickness and the velocity gradient is proportional to the shear stress acting on the wall. The wall shear stress was approximated using equation 2. The results from the modelling process were validated against this value to make sure the setup of the model was correct.

$$\tau_w = \frac{du}{dy}\mu \qquad (2)$$

The couette flow was modelled with the inlet and outlet of the domain connected by a translational periodic boundary condition (BC). The symmetry faces in the z-direction were also connected by a periodic BC. The top surface was represented by a moving wall with a velocity 5ms^{-1}. The bottom surface had a non-slip wall BC applied (zero velocity). All BCs that were not periodic had a zero gradient pressure BC applied.

Patankar and Spalding's (1972) [4] Semi-Implicit Method for Pressure Linked Equations (SIMPLE) iterative-based solver was used to solve for pressure and velocity fields. The algorithm works by initially guessing a pressure and velocity field. The momentum equations are then solved to find an estimated velocity field. Using the continuity equation, a pressure correction equation is generated and solved to attain a pressure correction. The pressure correction (multiplied by an under-relation factor) is then applied to the original guessed pressure field. The velocity is then also corrected, again subject to some under-relaxation. This process is repeated until the convergence criteria have been satisfied.

Initial runs of the calculations identified that default settings for under-relaxation were ill suited to the problem, generated divergent behaviour and failure to achieve the target residual error of 1e-8. The under-relaxation factors were lowered to 0.1 for the pressure solver and 0.3 for the momentum solver to guarantee stability in the parametric study.

2.2. Python Script

There are three python scripts used to automate this modelling process. One of the scripts is purely for generating a stereolithography (STL) geometry file, using the GEOM module within SALOME©. Another script is used to setup the file structure, write the trial conditions to each of the setup files and then run all the appropriate programs. The third script is responsible for the parametric study and requires a .CSV file input, detailing the values of the important controlling parameters. Due to the extensiveness of these scripts, this report shall only discuss the overall processes implemented by the scripts and not discuss in detail how every line of the script interacts with these higher level processes.

The simplest script is the parametric study script. Before script execution, a comma separated value file must have already been created. Each line in the file is one trial in the study and the parameter values are specified in the following order: angle, pocket length, pocket width, pocket depth, x-direction spacing, y direction spacing, stagger distance, clearance size. When the script is run, a numpy command allows the file to be read and stored as an array. Starting at the first iteration, the script reads the first line of variables from the array and then runs the overall python script. The python script is run with a set of arguments (the variables taken from the array) and these are parsed and then passed to the running script. Once the script has finished, it moves the generated data to a new folder away from the home directory allowing for the next study to be run.

The overall python script shall now be explained with the aid of the process flow diagram in figure 2a. Now that the arguments have been passed the script can create the folder structure and copy all the files from the source folder. It then writes to the files which are needed to run the modelling programs. The python script that generates the STL file using SALOME© is run straight after this operation. For each file written to, the script first opens the original file in read only mode and then opens a temporary file, a duplicate of the original, in write mode. It searches for the appropriate line in the original file and replaces that exact line in the temporary version. Once all the lines have been replaced with the correct parameters, it replaces the original file with the temporary file. Once a STL file has been created using SALOME©, the surfaceFeatureExtract OpenFOAM© program is then executed. This wraps surfaces around the STL file in order to create a geometry that is able
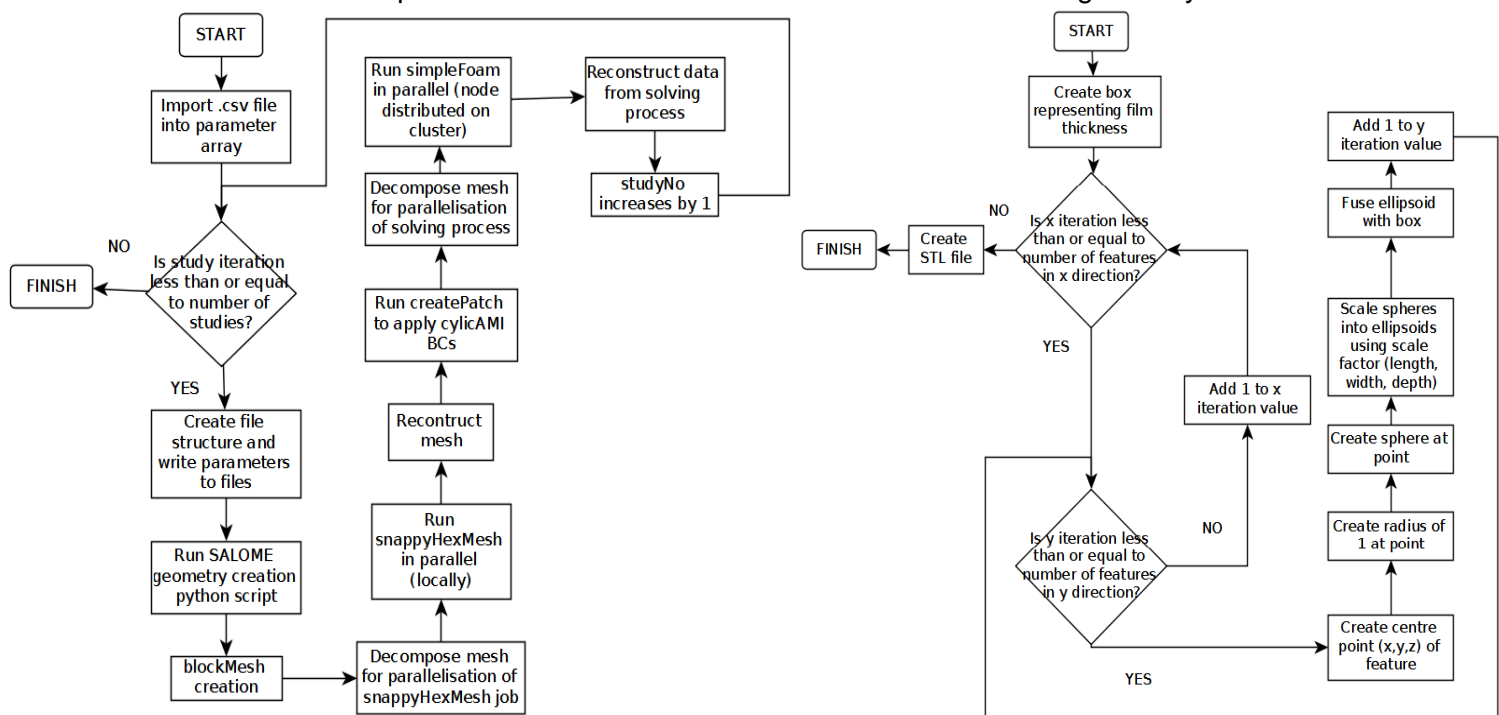
to work with the snappyHexMesh program. Before the meshing process can begin though, a background mesh has to be created.

Figure 2a. (left) Flow diagram of overall python script. Figure 2b. (right) Flow diagram of GEOM SALOME python script.

The background mesh is created using blockMesh which creates a hexahedral only mesh. At this point, temporary boundary conditions are applied to each of the 6 faces. The sliding wall on the top face is the only BC that is not subsequently replaced. The meshing procedure is computationally demanding so to speed up the process, the meshing process is parallelised. In order to do this, the background mesh is sub-divided into a number of regions, with each region being assigned to a physical CPU core, using the decomposePar program. SnappyHexMesh is now run. Using the extracted surface geometry as a reference, the meshing program refines the cells that intersect the surface. A mesh independence study has been performed using some initial trial runs in order to identify the mesh level suitable to refine the surface cells to (each level refines the cell by a factor of 2^3). The meshing procedure then removes any cells which have less than 50% volume inside the fluid domain. Any cells that do not fit the surface intersection correctly after the removal process undergo "snapping". This is where the hanging vertices of the cells are moved in order to fit the surface. Any poor quality cells that result from this operation are identified automatically and undergo further iterations of refinement. After the mesh has been created successfully, the mesh regions are assembled again to create the full mesh. This is done for two reasons, one so that it can be distributed on a larger number of cores than used for meshing but more importantly, to run the createPatch process. This program (createPatch) applies the translational periodic boundary condition to the inlet/outlet and symmetry faces. Now the model is ready for solving using simpleFoam, the script launches simpleFoam and displays a plot of the residuals for each solver using gnuplot. This continues until the script encounters the word "End" in the log file. This happens when the solving process completes successfully. If the solver crashes, it will cancel the command and then move onto the next study.

The last part of the program to be discussed is the python script (figure 2b) which automates the creation of the STL file using the GEOM computer aided design (CAD) module built within SALOME©. The first step in the program is to create a Cartesian box representing the film thickness. Using the spacing, clearance and pocket sizes, a minimum and maximum point is calculated. These points are the minimum information required to create this shape. The next command in the script creates the central point of the first pocket, which is placed at the origin. After the first point has been created, a radius of 1 is generated central to the point; this is then revolved to create a sphere. The reason for unity radius is that the sphere can now be multiplied by scale factors which are actually the required dimensions of the final surface feature. The sphere is multiplied by scale factors (length, width, depth), stretching the shape into an ellipsoid. If the pocket is not aligned with the flow direction, the ellipsoid is rotated by required angle. Now that the geometry is of the right size and orientation, half the ellipsoid shape is fused with the film thickness box, leaving half of the ellipsoid protruding from the surface (this represents the pocket of fluid). The next feature will be created at the required y-increment, away from the origin. This is repeated until the required number of pockets in the y-direction equals the y-increment number. The process is then repeated at the required x-distance away from the origin. Once the target number of x-increments has been met, SALOME© exports the geometry as an STL file.

3. Parametric Study

The next part of this report concentrates on the parametric study which has been carried out in order to investigate and optimise the features for this particular flow.

3.1. Taguchi Array Design

In order to study the effect of each parameter effectively and efficiently, the array was designed using Taguchi methods. This method significantly reduces the number of trials needed for the parametric study whilst ensuring that each of the parameters has been analysed so that a comparison of each level can take place. The trial conditions have been designed so that the orientation is tested to two levels (parallel and perpendicular to the flow) and all the remaining parameters are tested to three levels (table 1b). If a full factorial study would consist of 4374 trials. As each trial takes approximately 2 hours to run, it would roughly take a whole year (364.5 days) to complete this analysis. The minimum mixed level Taguchi table needed to cover the trial conditions is a L18 ($2^1 3^7$) orthogonal array (table 1a);

the 18 trials in this fractional factorial array only need approximately 36 hours to run. In order to implement this study, a comma separated value (CSV) file was created using the L18 array, replacing the level number with the actual parameter value. The CSV file formed the input, as previously described.

3.2. Results of Parametric Study

Once the parametric study had finished, a set of direct effect charts (figures 3a and 3b) were produced in order to visualise the effect that each parameter level had on the shear stress and the lift to drag ratio. The effect of clearance was the most prominent and completely outweighed the other effects, so these were not included on the figures. The effects of reducing the clearance increased the shear stress, as expected, according to the relationship described in equation 2. Firstly, looking at the effect of angle (A) on both shear stress and lift to drag ratio, it is clear that having streamwise textures are more beneficial, for both reducing drag but also generating a higher texture-induced lift per unit of shear stress. From a manufacturability perspective, streamwise textures are easier to apply using a cylindrical grinding process. The merit of each remaining parameter level will be based on the lift to drag ratio performance indicator, as this evaluates both enhanced load carrying capacity and the effects of surface drag, both of which need to be balanced to ensure an optimal surface texture design for a hydrodynamic bearing application. The largest length (B) and width (C) levels together with the middle depth (D) and stagger (G) levels produced the best performance. The largest texture density generated by the smallest pitch in both the x (E) and y (F) axis, produced a greater performance benefit which agrees with Ramesh et. al.'s (2013) [5] modelling conclusions.

Table 1a. (left) L18 ($2^1 3^7$) orthogonal array. Table 1b. (right) trial conditions.

Test No.	A	B	C	D	E	F	G	H
0	1	1	1	1	1	1	1	1
1	1	1	2	2	2	2	2	2
2	1	1	3	3	3	3	3	3
3	1	2	1	1	2	2	3	3
4	1	2	2	2	3	3	1	1
5	1	2	3	3	1	1	2	2
6	1	3	1	2	1	3	2	3
7	1	3	2	3	2	1	3	1
8	1	3	3	1	3	2	1	2
9	2	1	1	3	3	2	2	1
10	2	1	2	1	1	3	3	2
11	2	1	3	2	2	1	1	3
12	2	2	1	2	3	1	3	2
13	2	2	2	3	1	2	1	3
14	2	2	3	1	2	3	2	1
15	2	3	1	3	2	3	1	2
16	2	3	2	1	3	1	2	3
17	2	3	3	2	1	2	3	1

Level	1	2	3
Angle (A)	0	90	
Length (B)	200	250	300
Width (C)	25	50	75
Depth (D)	1	3	5
X Pitch (E)	300	400	500
Y Pitch (F)	50	100	150
Stagger (G)	0	0.25 pitch	0.5 pitch
Clearance (H)	3	5	7

4. Confirmation of Optimal Parameters

To confirm the results of the parametric study, the best and worst performing levels of each parameter were used in an analysis, to verify their effects. The results of the analysis were also compared to the analytic prediction of couette flow between "smooth" surfaces. Table 2 shows the two trials, using the best and worst performing levels of each parameter according to the lift drag ratio performance indicator
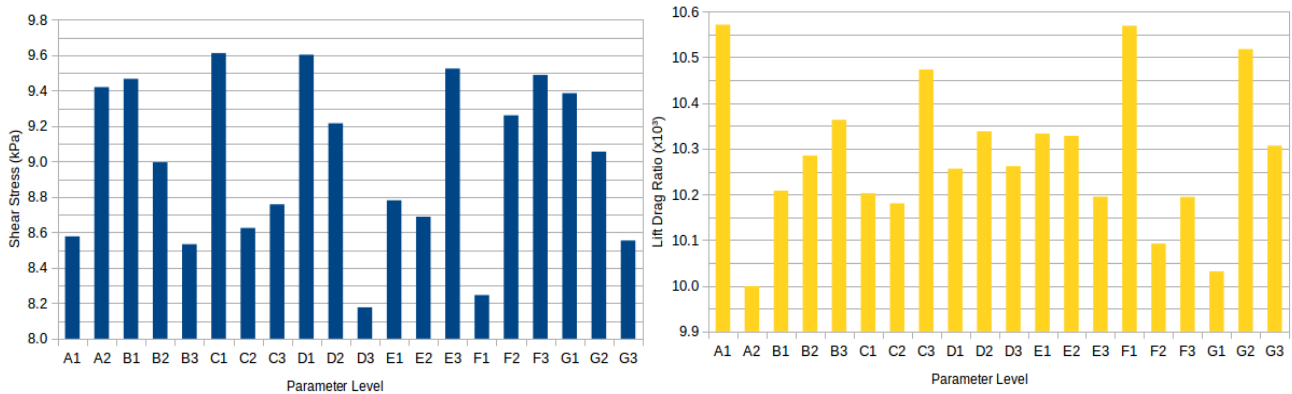
Figure 3a. (left) Shear stress direct effect chart. Figure 3b. (right) Lift to drag ratio direct effect chart.

Trident 15W40 engine oil properties were used for this computational study. The dynamic viscosity of this oil is 0.0883Pas. According to equation 2, for a clearance of 3μm, a shear stress of 14.7kPa would be generated at the bearing surface. According to the results of the final modelling study, a shear stress of 9.3kPa was produced by the best surface parameters and a shear stress of 14.4kPa for the worst.

Table 2. Best and worst parameter sets.

	Angle (deg.)	Length (μm)	Width (μm)	Depth (μm)	X Pitch (μm)	Y Pitch (μm)	Stagger
Best	0	300	75	3	300	50	0.25
Worst	90	200	50	1	500	100	0

The comparisons of these figures allows for some degree of validation. However, these surfaces also have a pressure-based drag contribution as well. So to obtain the correct drag reduction, the total drag force needs to be taken into account and compared to the non-textured surface (couette flow). The results of these calculations can be seen in figure 4.
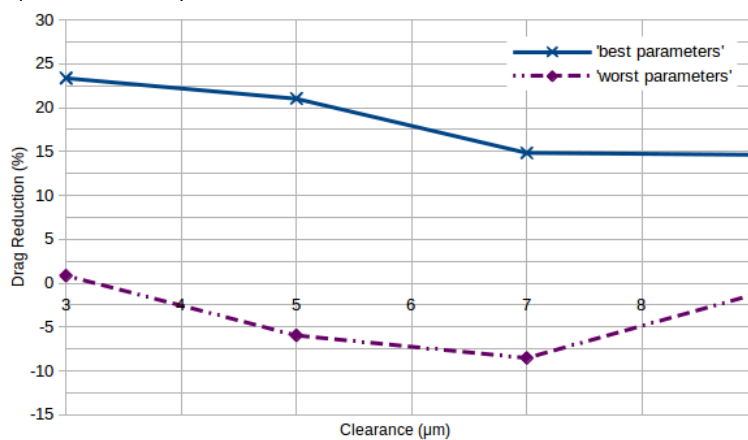
Figure 4. Drag reduction comparisons for best and worst parameter combinations.

5. Conclusive Remarks and Recommendations for Further Work

The first point to discuss is in relation to the successful creation of a python-based script used to automate the geometry creation, mesh generation and computational fluid dynamics (CFD) based modelling of surface textures. All of the software dependencies (OpenFOAM© and SALOME©) used in this script are totally open source, which allow for further research to be carried out without the costs and restrictions of licensing which comes with commercial based software. Another point to raise in relation to the script, is that it has been designed in a way, so that only one line of script needs to be edited in order to allow for other shapes to be analysed. The script creates an internal flow study at the moment but this is only due to the boundary conditions applied. These can be replaced in the OpenFOAM files to suit an external flow scenario. So, by accident, what has been created is a script that allows for CFD simulations of any three-dimensional textured surface. This is beneficial to the community because generating a structured mesh round three-dimensional patterns is time-consuming, complex and potentially expensive.

The parametric study carried out with the use of Taguchi methods has resulted in an optimum set of parameters which show significant drag reduction across a range of clearances (figure 4). What is also interesting in this result, is the worst parameters because they show a significant increase in drag. This result aligns with the conclusions of Syed and Sarangi (2014) [6]. If the surface textures are not optimised, then the surface can actually be of no benefit and the use of a "smooth" surface can actually prove to be superior. For the purpose of this report, only the modelling strategy and functionality of software tools are being evaluated. Further work will be required to study and investigate the flow field around a surface that reduces drag and compare that to one that increases drag. This should provide some further insight into the cause of surface drag reduction mechanisms.

6. References

[1] Holmberg K, Andersson, P, Nylund N-O, Makela, K and Erdemir A 2014 *Global Energy Consumption Due to Friction in Trucks and Buses* Tribology International **78** 94-114

[2] Bruzzone A A G and Costa H L 2013 *Functional Characterization of Structured Surfaces for Tribological Applications* Procedia CIRP **12** 456-461

[3] Chen X, Li H, Cao H, Wharton J, Allanson D R and Hu Z 2015 *Grinding for Microstructural Functional Surface* Applied Mechanics and Materials **806** 48-53

[4] Patankar S V and Spalding D B 1972 *A Calculation Procedure for Heat, Mass and Momentum Transfer in Three-dimensional Parabolic Flows*. Int. J. Heat Mass Transfer **15** 1787

[5] Ramesh A, Akram W, Mishra S P, Cannon A H, Polycarpou A A and King W P 2013 *Friction Characteristics of Microtextured Surfaces Under Mixed and Hydrodynamic Lubrication* Tribology International **57** 170-176

[6] Syed I and Syrangi M 2014 *Hydrodynamic Lubrication with Deterministic Micro Textures Considering Fluid Inertia Effect* Tribology International **69** 30-38

Leverage a Trust Service Platform for Data Usage Control in Smart City

Nguyen Binh Truong & Gyu Myoung Lee
Computer Science Department, School of Mathematics & Computing
Liverpool John Moores University (LJMU), Liverpool, UK
n.b.truong@2015. & g.m.lee@ljmu.ac.uk

Abstract. In the Internet of Thing, data is almost collected, aggregated and analyzed without human intervention by machine-to-machine communications resulting in raising serious challenges on access control. Particularly in Smart City ecosystem in which multi-modal data comes from heterogeneous sources, data owners cannot imagine how their data is used to extract sensitive information. Thus, there is a critical need for a novel access control method that minimizes privacy risks while increasing ability of access control for data owners. Our solution is to build a trust-based usage control mechanism called TUCON that enables stakeholders to personally set access control policies based on their trust relationships with data consumers. In this study, we introduce two novel components integrated in the Smart City shared platform: a Trust Service Platform and a Data Usage Control then combine them to form TUCON. We describe the TUCON mechanism including conceptual model, architecture, formalization, and practical development in detail. Roles and interactions of TUCON components in Smart City system are also described. We believe the study provides better understanding on both trust and usage control in the Internet of Things as well as opens several important research challenges and directions in the future.

1. Introduction

In recent years, we have been witnessing a novel network paradigm– the Internet of Things (IoT) which has imposed various research areas and applied to many types of network environment. In IoT infrastructure, billions of electronic devices are connected ranging from small and low computation capability devices such as Radio Frequency Identification tags (RFIDs) to complex devices such as smartphones, smart appliances and smart vehicles. The increases in quantity and connectivity result in dramatically increasing flow of data. In IoT, data is almost collected, aggregated and mined without human intervention by machine-to-machine communications that could lead to difficulties in complying privacy and security. And most data owners are not aware of how their data is used. Particularly, in an environment like Smart City in which multi-modal information coming from heterogeneous sources such as location, traffic, weather, gasoline and electronic usages [1], data can be aggregated and mined by a malicious participant to infer sensitive information. Moreover, stakeholders also have less opportunity to learn about data-usage practices. These reasons will aggravate problems on data privacy and data sharing in Smart City, thus, a new access control model for Smart City ecosystem is an urgent need.

Our previous studies have proposed a conceptual model based on Usage Control (UCON) and a handling mechanism for data access control in Smart City in which stakeholders can put their preferences in forms of constrains and obligations on the use of data [2]. However,

the proposed model cannot cope with many complex scenarios. The success of any data sharing platform depends on the compliance on data protection regulations and, beyond legal obligations, on trust relationships between stakeholders and data consumers. Our solution is to integrate a trust service platform to a UCON mechanism called Trust-based Usage Control (TUCON) that can guarantee data is only permitted to access and obligate by trusted sources. TUCON offers several benefits such as policy enforcement based on attributes of stakeholders and consumers, based on obligation actions, and based on trust. It offers data abstraction and data monetization features, and offers on-the-go usage control decision that adapts with environment changes. The main contributions in this paper are following: (i) a novel trust service platform with a trust model, a system architecture, and a trust computation procedure. (ii) TUCON: a novel usage control conceptual model and architecture for Smart City that considers three basic UCON factors: authorizations, obligations and conditions regarding to the trust platform. (iii) We provide formalization and prototype for both trust service platform and UCON including data abstraction, data annotation, semantic and reasoning mechanism.

The rest of the paper is organized as follows. Section II provides background and related work on trust and usage control. Section III introduces trust model, trust architecture and trust computation and practical development of the trust service platform. Section IV is dedicated to TUCON including a conceptual model and a system architecture. Section V focuses on practical expression and prototype implementation of TUCON. This concludes our research work and outlines future research directions in the last section.

2. Background and related work

2.1. Trust in IoT

Trust plays an important role in supporting both human and services to overcome perception of uncertainty and risk when making decision. It has been accepted that trust interplays among human, social sciences and computer science, affected by both objective factors (direct information) and subjective factors (third party information) from physical properties to social relations [4]. Trust can be considered as a computational value depicted by relationships among trustor, trustee and other entities, described in a specific context, measured by several Trust Metrics (TMs); and evaluated by a trust computation mechanism. In IoT, trustor and trustee can be human or machine, and context is as being under a service in a specific environment. Our previous research has investigated trust between a user and a vehicle, in a context of Car Sharing service in Smart City environment [5]. Trust has been considered as one of key factors for enhancing user privacy and networking security, and for establishing seamless connectivity and reliable services in IoT.

2.2. Usage Control

UCON is a new model of access control and initially proposed by Sandhu and Park [6] with a purpose of being addressed to emerging digital environments, allowing to apply in various access control situations. UCON enables two advanced features to cope with dynamic networking environment: (i) mutability of attributes, and (ii) continuity of an access decision. Basically, UCON keeps track of changes of attributes and policies when access is in progress, resulting in being able to change permission decision. Then an authorization system revokes granted rights or terminates resource usages accordingly. The permission decision is determined based on three factors called Authorizations, Obligations and Conditions. Authorizations are predicates over subjects (data consumers) and/or objects (stakeholders, data) attributes and put constrains on them to judge and grant the subjects a certain right on the objects. Obligations is a novel component in UCON model that examines the accomplishment of compulsory tasks that subjects have being done to objects before, during and after access period. Conditions are constrains from environment attributes, not related to both subjects and objects but affect the usage decision process [7].

2.3. Related Work

There is much research literature working on UCON for data sharing in some emerging network environments such as Social Network, Cloud Computing, IoT and Smart Cities. UCON features and research challenges have been well studied by in a survey conducted by A. Lazouski and his colleagues in [8]. Authors in [7] have extended traditional access control

models for providing obligations and conditions when accessing enterprise resources, forming a simple usage control mechanism. A simple accountability model and a platform have proposed in [9] allowing participants to explore consequences of different usage control policies. A privacy model is proposed in [10] in which semantic web technologies are utilized for supplying a privacy model and offering users to impose their preferences and control over data in Smart Grid environment. We have continued previous studies on data usage control [2] by integrating with our trust platform introduced in [5] and believe TUCON will open some approaches for trust-based usage control model in IoT ecosystems.

3. Trust Service Platform

The platform cooperates with applications and services to compute trust relationships among entities in IoT, working as a core service to offer securer transactions and better quality of service and experience.

3.1. RRK Trust Model

Despite the large number of trust-related research works, a standard, completed trust model is still under investigation [11]. In social science, people base on three sources of information to judge trust: public evidences (as reputation), opinions from surroundings (as recommendation), and their own understandings (as knowledge). We believe the process can be applied in IoT system. Based on the conceptual model proposed in [5], a trust model called RRK is proposed comprised of Reputation, Recommendation, and Knowledge as three TMs. Each TM is derived from other sub-TMs or TAs that reflect trust aspects in Social-Cyber-Physical (SCP) world (Fig. 1a).
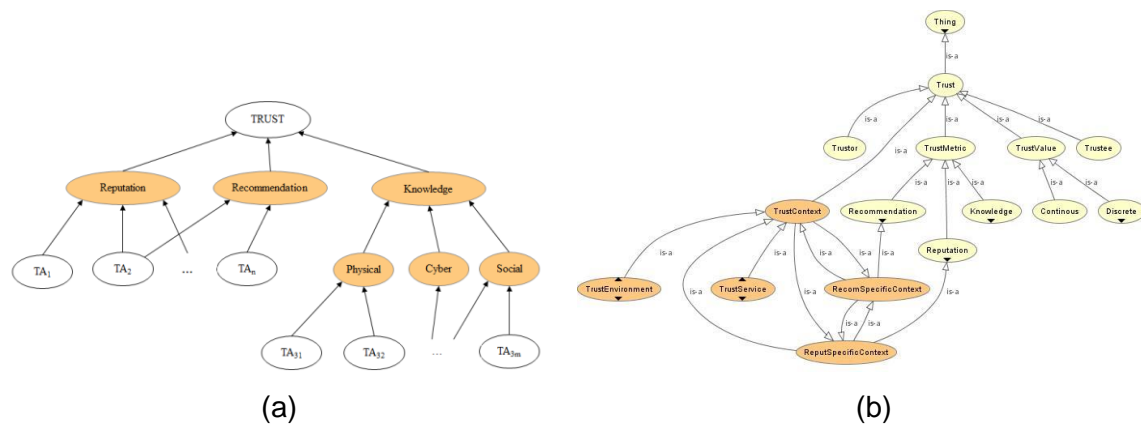

(a)   (b)

Figure 1:The proposed RRK Trust Model (a) and the associated upper ontology (b)

There are two methods to derive TMs, TAs and trust. The first uses mathematical models such as weighted sum, Bayesian neutron networks, reputation heuristic or PageRank algorithms for computing some sorts of TMs such as Recommendation and Reputation. The second uses inference engines for inferring new trust-related knowledge from a knowledge base for obtaining some TAs and trust levels.

3.2. Trust Computation System

To calculate trust, sufficient data about trustor, trustee and trust context needs to be collected, aggregated, processed and annotated in order to create a set of semantic information which is a part of a trust knowledge base. The rest of the trust knowledge base is in form of rules acquired from a knowledge acquisition mechanism. The knowledge base is as an input of an inference engine to infer new knowledge and then to reason a trust value. Base on the trust value, access control decisions could be made accordingly. The conceptual trust computation procedures are illustrated in Fig. 2.
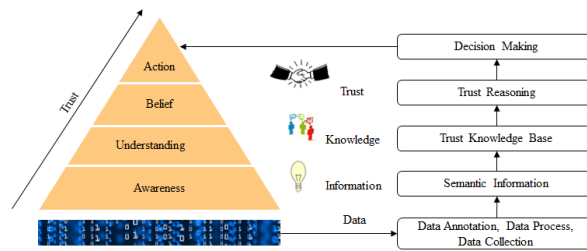
Figure 2: Conceptual Trust Computation Processes

From Data to Semantic Information: Various types of trust-related data are collected from many sorts of sources in Smart City. The integration of SCP data enables the incorporation of situation and context-awareness, thus enables intelligent autonomous applications and services *[12]*. This leads to a need for a data integration and annotation framework with a data model and knowledge representation. The framework is also required for enhancing semantic interoperability for handling semantic information. Several semantic web technologies could be used for trust modelling, data integration, and data query. For example, ontology is used to represent trust and TUCON domain-specific models. Fig. 1b shows the upper ontology for RRK used in our implementation. Based on the ontologies, trust data is annotated accordingly using RDF schema (RDFS) as metadata and semantic information.

From Semantic Information to Trust Knowledge Base: The trust knowledge base comprises of structured and unstructured information represented in a machine-interpretable language. The creation of the trust knowledge base includes the creation of facts about trust (declarative knowledge) and the creation of logics among concepts described in the facts (procedural knowledge) *[14]*. In the trust prototype, we use a knowledge representation formalism combining both rule-based language and ontology for supplying reasoning capabilities. Specifically, semantic information extracted in the first process is converted into facts in form of Description Logics *[15]*. Rules can be monotonic or non-monotonic to express knowledge about ontologies such as classes, sub-classes, instances and relations. The rules are the most important part of the knowledge base. They interpret meanings and describes relationships of the concepts in the facts. The process to create rules for a knowledge base is called knowledge acquisition, a part of knowledge engineering. It is a complicated process that acquires knowledge from many resources such as user preference, domain experts, documents, Internet resources, etc., using various kinds of methods such as interview with human; data mining and machine learning with data and Internet resources [16].

Trust Reasoning Mechanism: A semantic reasoner is used for inferring new knowledge related to trust including facts about TAs, TMs and trust values. The final goal is to compute trustworthiness based on the trust knowledge base. In this study, the trustworthiness is simplicity defined in three levels: low, medium and high meaning as distrust, normal and trust, respectively. The reasoner takes the trust knowledge base as its input and infers new knowledge as new facts, as a result, additional rules in the knowledge base are triggered; new other facts could be created. This process would iterate until a goal has reached or no rules can be matched. In the demonstraiton, a Jena generic reasoner with forward chaining mode is used to infer new facts and trust levels.

Trust Service Platform Architecture in Smart Cities

The trust platform architecture is called Trust Analysis and Management Platform (TAMP). TAMP comprises of four components namely Reputation System, Trust Agent, Trust Broker and Trust Engine and is introduced and described in detail in our previous work [5]. In the next section, we will describe how the component of the trust service platform and TUCON are incorporated in the Smart City shared platform.

4. Trust-based Usage Control Mechanism

4.1. TUCON Conceptual Model

The initial step in the design of any UCON mechanism is to identify objects to be protected, subjects that request to access and perform actions on objects. Actions are obligations describing how the objects are exploited by the subjects. It is needed to define Access rights associated with each of the obligations and Authorizations that predicate the access rights based on attributes (ATT(O)), subjects attributes (ATT(S)) and the environment

attributes as Conditions. In TUCON, objects are dataset owned by stakeholders, subject are data consumers, conditions are trust relationship between data owners and data consumers as illustrated in Fig. 3a. Details of the conceptual model are clearly described in the section V.
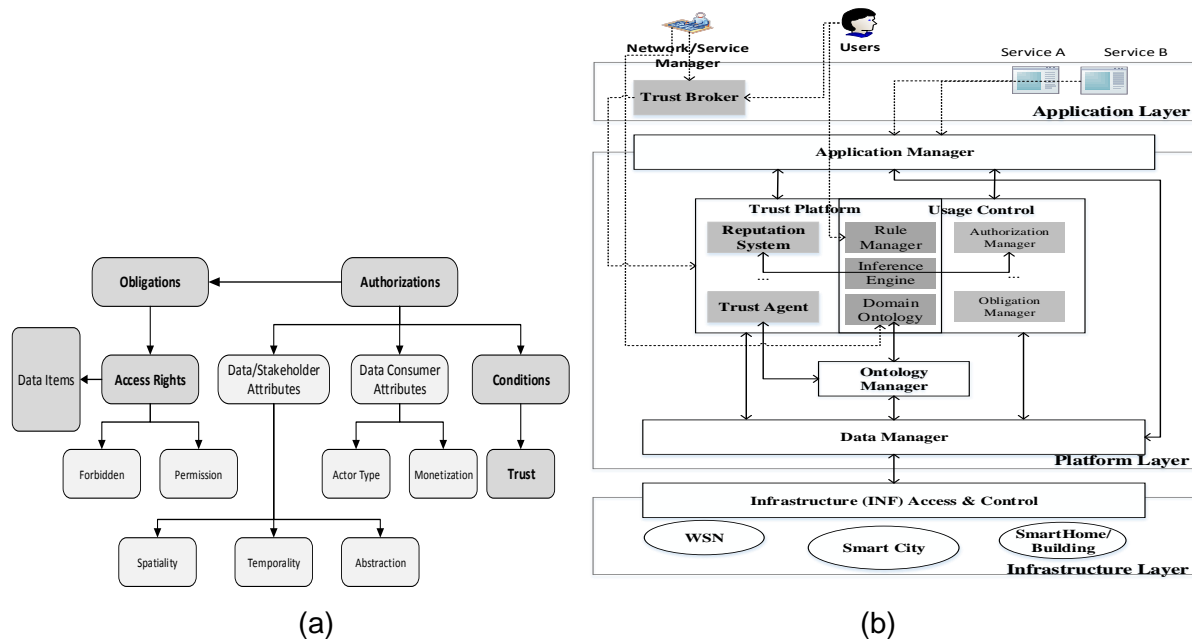

(a)   (b)

Figure 3: TUCON conceptual model (a) and architecture in the Smart City shared platform (b)

4.2. TUCON System Architecture

TUCON architecture is built under context of the 3-layered Smart City shared platform proposed in [17] [18]. The three layers are Infrastructure Layer (INF), Platform Layer (PLA) and Application Layer (APP). The platform is to deal with data acquisition and data annotation from deployed sensors exploited by multiple applications and services. The Data Manager (DM) is to work with IoT data and resources from INF whereas the Application Manager (AM) is as an interface between application and PLA. An Ontology Manager (OM) is also introduced for data annotations and for supporting semantic-based Wireless Sensor Networks (WSN). TUCON architecture is created by incorporating the TAMP and usage control components into the 3-layered shared platform. As illustrated in Fig. 3b, three mutual components are shared between TAMP and TUCON: Rule Manager (RM), Inference Engine (IE) and Domain Ontology (DO).

- RM is for handling rules in the trust knowledge base in TAMP and authorization policies (rules) in TUCON. Note that the rules are express the relationships among classes and individuals of the ontologies, thus, incurring interactions among RM, DO and OM. RM directly interacts with Users for acquiring user preferences in the form of rules in case of TUCON. RM also interacts with Trust Brokers for the user preferences in case of TAMP.

- IE implements some reasoners for inferring new facts and trustworthiness in TAMP as well as inferring access rights for TUCON. TAMP and TAMP can use same or different reasoners depending on their formalization types. In this study, we use Description Logics with Ontology for trust and Defeasible Logics (DLs) for usage control formalizations, resulting in different reasoning mechanisms are used.

- DO is a manager for handling domain-specific ontologies, and for cooperating with OM for data annotation and data abstraction in both TAMP and TUCON. DO works with Network/Service Manager for ontology update.

5. Practical Expression and Prototype

In this section is to explain elements and formalization of the TUCON conceptual model by implementing a prototype based on some semantic-web technologies and DLs.

5.1. DataItems:

A Data Item is an individual of Context Element container proposed in the NGSI 9/10 Information Model[10] used to exchange information about an entity, including entity ID, context attributes, related attribute domains, and metadata for all of the attribute values of the given domain. DataItem is formally defined in XML DTD syntax.

5.2. Authorizations:

TUCON policies represent constrains based on object attributes, subject attributes and conditions. The Authorizations optionally contains the following expressions: (i) ATT(O): Temporal Constraints for temporal granularity, Spatial Constraints for spatial granularity, and Abstraction Constraints for masking of certain information. (ii) ATT(S): Actor Type such as institutional, commercial operators, equipment manufacturers, or service providers, Monetization as the purpose of using data such as selling, training, or providing customer supports. (iii) Conditions: trust value between data owner (trustor) and data consumers (trustee). Authorization is also represented in XML DTD format.

5.3. Obligations:

This is a set of actions on DataItems such as Full Access, Partly Access, Dissemination, Storage, and Analysis. The Obligations are associated to Data Monetization, for example, if a data consumer requests for selling data, then Obligation action should be Dissemination; or if data consumer requests for statistical training, then Obligation action should be Partly Access with constrains Temporality = {Weekly} & Abstraction = {Statistical }.

5.4. Access Rights:

Access Control decision is associated with Obligations and Authorization value. We simply define as Permission and Forbidden representing whether DataItems are allowed to "*share*" or not. The word "*share*" now can be more specifically understood in the TUCON context as: "*allow to conduct appropriate obligation actions*":

<AccessRight(Rule*)> && <Rule(Obligation?, Authorization?)>

5.5. TUCON Formalization and Expression

The approach for TUCON formalization is based on DLs, a non-monotonic formalism with normative conflicts-solving ability and low computational complexity [19]. Particularly, an extension of DLs formalism enriched with model and deontic operators are used as a formal model for TUCON policies due to its representational capability of Obligations and Authorization factors. Below examples show how DLs is applied for TUCON formalization:

- *Facts:* Facts in DL represent the ATT(O), ATT(S) and Condition (in terms of trust level). For example, two institutional organizations (IO1 and IO2) with "High" and "Low" trust value ,respectively, are represented as FACTs below:

---

[10] https://forge.fiware.org/plugins/mediawiki/wiki/fiware/index.php/NGSI-9/NGSI-10 information model

FACTS
$F_1{}^{TUCON}$(IO1): {ActorScope(Institutional)}
$F_2{}^{TUCON}$(IO1): {TrustScope(High)}
$F_1{}^{TUCON}$(IO2): {ActorScope(Institutional)}
$F_2{}^{TUCON}$(IO2): {TrustScope(Low)}

RULES
$R_1{}^{TUCON}$(IO): {X[OB] => SpatialScope(Street)}
$R_2{}^{TUCON}$(IO1): {IO1[OB] => SpatialScope(any)}
$R_5{}^{TUCON}$(IO2): {IO2[OB] => SpatialScope(Zone)}
$\qquad R_2{}^{TUCON}$(IO2) > $R_1{}^{TUCON}$(X)
$\qquad R_3{}^{TUCON}$(IO2) > $R_1{}^{TUCON}$(X)

- *Rules and Superiority Relations:* All constrains among stakeholders, data, actors, conditions and TUCON AccessRight are represented in DL rules. Note that in DL, there are three different rules types which have different meanings. The strict rules can never be defeated, while defeasible rules can be defeated by contrary evidences. Strict rules and defeasible rules are used for drawing conclusion whereas defeater rules are only used to prevent from making conclusions. Superiority relations of rules are used to set the priority among these rules. The rules above is an example of defeasible rules and superiority relations of the two instituional actors IO1 and IO2. X represents any institutional actor. OB shorts for Obligations action, is a modal operator of DL extension. The example can be explained as follow: by default, any institutional organization is allowed to conduct OB on data at spatial street level. However, this policy can be overruled when considering trust relationship between the actor and the data owner. For example, if trust value is high, then the actor can access all spatial level of data (actor IO1) or if trust value is low, then only zone level of data is permitted.

- *DL Inference Engine and TUCON request:* An example of a consumer X requests for data with Obligation action OB will be expressed in DL as following:

$R_{req.}{}^{TUCON}$(X[OB]):{SpatialScope(Street), TemporalScope(daily),
  AbstractScope(detail) =>X[OB]}

A DL inference engine is used to get the conclusion that whether RreqTUCON is defeasible proven in the DL theory or not. The inference algorithm is based on DL Proof Theory mentioned in *[19]*. Several candidate of DL reasoners can be applied and we choose Spindle11 for our demonstration. The conclusion is as following:

Conclusions
====================
-D $R_{req.}{}^{TUCON}$(X[OB])
-d $R_{req.}{}^{TUCON}$(X[OB])

meaning that the Rreq.TUCON(X[OB]) request is Defeasible Provable in the DL theory. That means thedata consumer satisfies all the authorization policies to obligate the action OB on the stakeholder's data, the AccessRight now is Permission. The characteristic of the DLs formalism is suitable for any usage control mechanism since the facts, the rules are defeasible and can be overruled by supplying more facts, rules, and superior relations in DLs, resulting in conclusion changes. This feature enables the ability of continuity of an access decision in TUCON.

6. Conclusion and Future Work

In this research, we have introduced a novel usage control mechanism that leverages a trust platform called TUCON in order to provide securer data access control based on trust relationship between data owners and data consumers in Smart City environment. Firstly, we present the trust service platform TAMP in accordance with RRK trust model, trust architectures and trust

---

[11] http://spin.nicta.org.au/spindle/

formalization. An implementation prototype of trust computation based on semantic-web technologies is also described. Secondly, we introduce the TUCON conceptual model and TUCON architecture considering trust components in the three-layered Smart City shared platform. Finally, the practical expression for TUCON and the prototype are clearly characterized using DLs.

There are two main research directions could be taken to fulfil TUCON mechanism. The first direction is the improvement of TAMP by developing an automated intelligent rules creation for the trust knowledge base instead of being predefined by using machine learning techniques for rules pattern recognitions. A verification mechanism is also needed to investigate and develop to check the quality of the knowledge base such as consistency and redundancy. The second direction requires a large number of research studies need to be done in order to strengthen TUCON by improving UCON conceptual model, formal model, architectures and enforcement mechanisms as well as enhancing mutuality of attributes and continuity of access decision.

References

[1] T. Payam Barnaghi *et al.,* "Challenges for Quality of Data in Smart Cities," *Journal of Data and Information Quality (JDIQ)*, vol. 6, no. 23, 2015.

[2] Q. Cao, G. Madhusudan, R. Farahbakhsh and N. Crespi, "Usage Control for Data Handling in Smart Cities," *IEEE Global Communications Conference (GLOBECOM)*, San Diego, CA, 2015.

[3] B. Alcalde, "Towards a Decision Model based on Trust and Security Risk Management," *Seventh Australasian Conference on Information Security*, pp. 61-67, 2009.

[4] N. B. Truong, T. Won and G. Lee, "A Reputation and Knowledge Based Trust Service Platform for Trustworthy Social Internet of Things," *Innovations in Clouds, Internet and Networks (ICIN)*, France, 2016.

[5] J. Park and R. Sandhu, "Towards usage control models: Beyond Traditional Access Control," *Seventh ACM Symposium on Access Control Models and Technologies*, New York, USA, 2002.

[6] J. Park and R. Sandhu, "The UCON ABC Usage Control Model," *ACM Transactions on Information and System Security (TISSEC)*, vol. 7, no. 1, pp. 128-174 , 2004 .

[7] A. Lazouski *et al.*, "Usage control in computer security: A survey," *Computer Science Review*, vol. 4, no. 2, pp. 81-99, 2010.

[8] J. Pato *et al.,* "Aintno: Demonstration of Information Accountability on the Web," *IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT)*, Boston, MA, 2011.

[9] S. Speiser *et al.*, "Web Technologies and Privacy Policies for the Smart Grid," *Annual Conference of the IEEE Industrial Electronics Society (IECON 2013)*, Vienna, Austria, 2013.

[10] Y. Sun, W. Yu, Z. Han and K. Liu, "Information Theoretic Framework of Trust Modeling and Evaluation for Ad Hoc Networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, pp. 305 - 317, Feb. 2006.

[11] A. Sheth, P. Anantharam and C. Henson, "Physical-Cyber-Social Computing: An Early 21st Century Approach," *IEEE Intelligent Systems*, vol. 28, no. 1, pp. 78 - 82, Feb., 2013.

[12] S. Russell and P. Norvig, "Knowledge and Reasoning," *Artificial Intelligence: A Modern Approach*, New Jersey, Prentice Hall, 2014, pp. 149-297.

[13] F. Baader, D. Calvanese, D. McGuinness, D. Nardi and P. Patel-Schneider, "The description logic handbook: theory, implementation, and applications", New York, NY, USA: Cambridge University Press, 2003.

[14] K. Simon and C. Malcolm, "An Introduction to Knowledge Engineering", New York, NY, USA: Springer-Verlag, 2006.

[15] I. Khan, F. Belqasmi, R. Glitho, N. Crespi, M. Morrow and P. Polakos, "Wireless sensor network virtualization: early architecture and research perspectives," *IEEE Network*, vol. 29, pp. 104-112, June 2015.

[16] I. Khan, R. Jafrin, F. Errounda, R. Glitho, N. Crespi, M. Morrow and P. Polakos, "A data annotation architecture for semantic applications in virtualized wireless sensor networks," *IFIP/IEEE International Symposium on Integrated Network Management (IM 2015)*, Ottawa Canada, May

2015.

[17] D. Nute, "Defeasible logic: Handbook of Logic in Artificial Intelligence and Logic Programming", New York, USA: Oxford University Press, 1994.

Christopher Wren

Developing a Missing Persons Search and Prioritization Toolset (MPSP)

Chris Wren
School of Computer Science
Liverpool John Moores University
Liverpool UK
c.wren@ljmu.ac.uk

Abstract—**Missing person cases consume time and effort on what are already over-stretched Police resources. Across the UK Police forces have to deal with approximately 1000 enquiries a day relating to missing persons. Current approaches are inherently manual and rely on Police domain knowledge to relate a missing person's behavior to the surrounding geography and terrain. Searches are highly probabilistic, based on a missing person's profile, mental and spatial behavior patterns. Computer support is available, but this is largely in the form of data entry and database systems. This paper describes the proposed toolset which provides computational support for missing person cases, specifically those at suicide risk. The toolset is essentially a database fronted by a public domain map application. Its potential lies in the provision of 'what-if' search scenarios and potential search locations. The paper describes the development of the toolset using missing person/suicide data for the Merseyside region in the UK and describes the research challenges that must be surmounted to bring the work to fruition.**

Keywords-missing persons; Police; suicide; sucide risk; suicide hotspot mapping; Google Map; database; GIS; mental health; Merseyside

Introduction

Every day throughout the UK people go missing, either of their own volition or due to circumstances beyond their control. When a person disappears it sets a range of possible scenarios for what has happened and many questions can arise: Are they lost somewhere? Did the person suffer mental health? Did they take their own life? Could they have been taken? These are just some of the questions asked and there can be a lot of speculation and information that needs to be dealt with.

A. Missing persons and the Police

Missing persons span all age ranges, different sexes, social classes and geographical areas. To 'go missing' is often regarded as selfish however, missing behaviour is often indicative of other underlying problems that are not always immediately obvious. This type of behaviour can lead to a variety of consequences both for the missing person and those left behind.

It is important to gain an understanding of the amount of people who go missing, when developing any system that proposes to assist in their location. The National Policing Improvement Agency [1] analyzed missing person data in 2009/2010 and concluded that in the UK the Police deal with in the region of 350,000 incidents a year. They also found that in dealing with missing person enquiries on a day to day basis, there are around a 1000 enquiries per day across the UK. In a separate study, it was found that approximately two-thirds of missing person cases concern people under the age of 18 [2].

The overwhelming majority of cases are resolved quickly, most without any drastic consequences. Research shows that three-quarters of disappearances reported to the Police are resolved within two days [3]. However, a significant minority of 8% (approximately 20,000) last longer than one week, and 1% (2,500) last in excess of a year. Adults are more likely to remain

missing for longer periods than young people [4]. Unfortunately, there are also a large number of people that do not return safely.

The *Association of Chief Police Officers (ACPO) Guide on the Management Recording and Investigation of Missing Persons* [5] defines a missing person as "anyone, whose whereabouts are unknown, whatever the circumstances of disappearance and they are considered missing until their location has been established". 'Missing person' is a broad term that covers runaways, disaster victims, missing hikers, crime victims, overdue travelers, people suffering dementia, people suffering from depression people who intend to commit suicide and those who voluntarily walk away from their established lives and relationships.

The Human Rights Act 1998 places a positive obligation on Police forces around the country along with supporting agencies to provide the service of finding missing persons, safeguarding vulnerable people and improving on public safety [5]. When a person goes missing the local Police force is responsible for initiating the investigation and subsequent reporting. The early stage of the investigation is crucial as time is of the essence and it is essential to have the correct information, experience and available resources to deal with the situation in an efficient manner. Any fault or delay in the way the incident is handled can be the difference between life and death. With this mind Police forces are keen to make use of systems, intelligence and any other resources, which can assist the task of finding missing persons.

B. GIS

Geographic Information Systems (GIS) provide an interactive way of illustrating geographic information through a map. The primary function of GIS is to handle large quantities of geographically referenced data by storing, editing, organizing, processing and analyzing for visualization purposes related to a variety of human activities. Within GIS, the information collected for an area is divided into layers. Each layer represents spatial datasets of real world locations that may include spatial objects such as roads, bus routes, shopping centres and industrial parks. GIS also provides an ability to combine a variety of datasets in many different ways; for example a view showing just bus routes, shopping centres and public houses.

Many law enforcement agencies worldwide already use GIS technology for the purpose of 'intelligence-led' policing in crime analysis. ESRI, a large US GIS software company, has produced a GIS based solution that is applied to crime analysis worldwide [6]. This system is also becoming an integrated part of UK policing, with many Police forces across England and Wales adopting this solution. Research into the use of GIS has shown it to be an effective tool for crime analysis and subsequently for fighting crime. The increasing availability of geo-coded crime data leads to a greater understanding of crime and a means to visualize and target repeated crime patterns.

The argument of this paper is that missing person cases could also benefit from the use of GIS. Missing persons are characterized by spatial behavior patterns which could be visualized via GIS [7]. Along with providing the Police with an efficient tool to analyze this information and gain a better understanding of incidents, GIS would also be particularly useful when dealing with repeated cases, as more information is collected and accumulated with each case. In particular the paper describes how the web-based GIS technology of Google Maps can be used to assist missing person cases in the Merseyside region.

The remainder of this paper is structured as follows: Section II describes the current approaches used to finding missing persons. Section III then describes the requirements analysis in terms of function and data. Section IV describes the prototype implementation. Section V discusses some related work and section VI draws overall conclusions and identifies the specific research challenges and work focus.

CURRENT APPROACH

The investigation of missing persons is a difficult area of Police work and potentially one that can cause significant problems if the enquiry is not conducted in a satisfactory manner. Clearly there is little margin for error the need to "get it right first time" often results in considerable time and effort expended in missing person cases. This section provides a generalised review of the procedures used by Police forces in the UK.

C. Initiation

The initiation of a missing person report is crucial and often has a bearing on the overall outcome of the case. The 'golden hour' in a murder investigation is recognised as being of vital importance and this applies to missing person cases. It is also worth noting that many murder investigations are first received as missing person cases [5].

The first step in any missing person inquiry starts with the initial reporting, which is usually:

- A telephone call to a Police station
- A visit to a Police station
- Contact with a Police officer/Police staff away from a Police station

The responsibilities of the first officer receiving the initial report are:

- Establish the facts and keep accurate records of what was said and by whom.
- Make an assessment of the circumstances of the disappearance in order to make a judgement regarding the risks to which that person or the community are likely to be exposed. The decision, the evidence supporting the decision should be recorded.
- Gather sufficient information about the missing person to enable an effective and thorough investigation to be conducted. The depth of that information will vary according to the assessment of risk. Very detailed information and a lifestyle profile will be needed in high-risk cases.
- Input information onto a missing person IT management systems.

According to *ACPO Guide* [5] the minimum information that is gathered when taking the initial report is:

- Name
- Age
- Ethnicity
- Description of person
- Description of clothing
- Home address
- Location missing from
- Circumstances of going missing
- Whether this behaviour out of character
- Details of any vehicle or other transport used
- Assessment of the person reporting
- Name, address and telephone number of person reporting

This information, and other information that arises as the investigation proceeds, are recoded on a missing persons IT management system local to the investigating Police force. The COMPACT (Co-Ordinated Missing Person and Case Tracking) computer system is used by several UK Police forces for recording missing person information [8].

D. Risk assessment

After the initiation of the case it is necessary to make an informed Risk Assessment of the circumstances of the individuals disappearance from the information provided, and subsequently the time and resources spent on the case will depend on this risk assessment. Every reported missing person incident should be assessed as to whether it should be declared a critical incident. A critical incident is defined as 'any incident where the effectiveness of the Police response is likely to have a significant impact on the confidence of the victim and/or the victim's family and/or the community [5].

Risk assessment procedures are particularly important for the Police because of the high volume of missing person's reports they receive. The correct categorization of risk is extremely important as this has an influence on the level of policing response and subsequent management of the inquiry. The risk assessment is always subject to on-going and regular review, as any change in the circumstance can alter the risk level.

The *ACPO Guide* [5] state the national approved risk categories for missing persons as:

- High risk - the risk posed is immediate and there are substantial grounds for believing that the subject is in danger through their own vulnerability, or may have been the victim of a serious crime, or the risk posed is immediate and there are substantial grounds for believing that the public are in danger. High risk requires the immediate deployment of Police resources and possible a press/media strategy and/or close contact with outside agencies.
- Medium risk - The risk posed is likely to place the subject in danger or they are a threat to themselves or others. Medium risk requires some level of proactivity by Police and other agencies
- There is no apparent threat of danger to either the subject or the public. Other than initial checks with local hospitals, it will not require any level of proactive involvement by Police. Note however, that Low risk does not equate to No risk; many murder cases initially reported as missing persons fit the low risk classification.

The risk category can change as fresh information is available and there is a need to continuously evaluate the category for each missing person case.

E. Scenario analysis

An approach used by many UK Police forces when seeking to explain a disappearance is to examine the "scenario" of the situation [9]. Creating scenarios is already recommended as a basis for developing of Scenario 'testing' which requires an analysis of the (relative) frequency with which each scenario occurs. For example, analysis could be undertaken of the frequency with which young adult males last seen on a night out return of their own accord or are found deceased in water [9].

The *ACPO Guide* [5] describes that in the management of locating missing persons, scenario building is widely recommended as a basis for developing search strategies. In the context of risk assessment, scenarios might be considered as the narrative accounts of a disappearance which provide a suggested outcome of a disappearance based on a selection of factors about the individual or the circumstances of the episode. This approach looks at the kind of story surrounding the missing person and what they did after they left the point they were last known to be [10]. This gives an idea where to search, how to search and what investigation needs carrying out.

The first step is to adopt a logical approach, which reflects on what is known about the person after the initial risk assessment. This includes the demographic profiling of the person along with assessing the circumstances surrounding their absence. Questions such as: How long the person has been missing? Have they ever gone missing before? Are all important factors to be considered in the scenario.

According to [10] there are generally four categories of missing persons:
- Lost
- Voluntary Missing
- Forced Missing
- Missing due to accident or illness

A scenario based evaluation can assist in establishing which one of these outcomes the subject falls under. At the initiation it is often impossible/difficult to determine the appropriate category. As the investigation progresses and with the aid of any witness interviewing and intelligence led inquiries officers are usually able to eliminate several categories [9].

Scenario analysis also pays close attention to the person involved. Every missing person is individual; no two are the same. Officers attempt to find out as much as possible about the subject before initiating a particular response. As mentioned already there are many varied reasons why a person goes missing. However, recorded data has shown that approximately 80% of adults who go missing are known to have some form of mental illness and the overall majority of young people who go missing simply run away from home [9]. These two crude but reliable metrics provide a good starting point for any scenario analysis. For instance if the subject is known to have mental health problems this is likely to have an impact on their behaviour, which in turn is likely to influence the places they are likely to visit. This intelligent reasoning can prove extremely useful and in particular

help in focusing the search by identifying both high probability and high probability search areas as considered next.

<p style="text-align:center;">*F.* *Conducting the search*</p>

Most searches are conducted using a map for visual aid and computer-based maps are advantageous as they allow large areas to be considered by scrolling and viewing different segments of a map covering a large area. However, the search process still requires the evaluation of the information gained from the Initiation, Risk Assessment and Scenario Analysis in conjunction with the geographic data and terrain displayed on the map. This evaluation at present is largely carried out by the investigation team; for example missing persons with suicidal tendencies may focus the search on areas with large trees, waterways, bridges and railways. Officers aware of the local geography will be aware of such features and know of their location on the map. The research described in this paper intends to ease this interpretation by providing a map application which can use information from the investigation in conjunction with information from previous cases to display probable areas and geographic features to search.

Searching is mainly a probabilistic process since it is not possible to do a thorough search of everywhere all at once in a limited time. Decisions need to be made on how to best utilize the available resources. A probability distribution map is often used to mark the most likely areas in which the missing person can be located. The profile of the person in conjunction with geographic features, terrain and weather conditions may be used to mark the map with high areas of probability. These areas of high probability are the focus of the initial search after which the search may move on to areas lower probability [7].

To start the initial search the Police need to determine the co-ordinates of the Last Known Point (LKP)[12] [7], where the person was last seen to establish where to begin. From the LKP they go on to outline a Probability Of Area (POA)[13], which is the region on the map which contains the highest likelihood of the person being found [7]. To assist in this task they utilize various numeric descriptions of the missing person and travel data accumulated from previous missing person's cases. This aggregated data is then used to identify behaviours and patterns formulated from missing person categories and their spatial behaviour. This method is widely used and proves to be an efficient means of estimating a search area.

Learning more about the whereabouts and travel habits of missing persons will enable Police to link the distances which missing persons travel or the mode of transportation to their demographic background. Another important factor taken into consideration is how far a person can travel within the time frame they are missing. Time is crucial in an investigation as the shorter the time period they are absent the higher the probability that they will successfully found, as is the case with missing children. A radius surrounding the POA is drawn from a likely time-frame within there is reasonable expectation of tracing the individual. As the time frame increases so does the radius and the probability of them being successfully found decreases in most circumstances [11].

It is clear to see from the search process the importance of the use of maps and more significantly the intelligence and decision making, focused around the map. This provides one of the main motivators for this paper: "How can we develop an interactive GIS which provides timely geographic information and supports 'what-if' scenario elaboration".

<p style="text-align:center;">REQUIREMENTS</p>

MPSP is intended to be a software prototype system that comprises a database and an interactive GIS front-end application. The two are to be connected, but the database may still be used for analysis purposes in isolation from the GIS front-end. The development of the MPSP system requires that certain functionally be available to the users and certain data requirements are met, as considered below.

---

[12] Last known point, also referred to as last known position: (LKP) point on the map where the missing person was last seen or known to be.

[13] Probability of area: (POA) refers to the likelihood of the missing person being searched for actually being in the area.

The MPSP system will combine idiosyncratic information about missing persons together with trend data from past cases to help identify and prioritize search areas in missing person cases. Primarily the data will be visualized through a GIS front-end application based on Google Map technology, which will provide functionality through which users can at best locate missing persons or at least narrow the search and make educated decisions. The following is an outline of the functionality required of the Map application.

1. Trend data
   1.1 Display average search distances for key sub-groups (missing children, ADHD, depression, suicide, dementia, psychosis/schizophrenia, bipolar disorder, alcohol and drug abuse, gender differences)
   1.2 Display Location of local suicide and attempted suicide hotspots
   1.3 Display Knowledge about suicide methods for key sub-groups

2. Idiosyncratic information
   2.1 Entry of Last Known Whereabouts - these can be entered onto the map and the Map application develops a search prioritization based on function 1.1 above.
   2.2 Entry of Areas of Significant Interest - work, relatives, friends, places known to visit e.g. pubs, college. These can be entered onto the Map and the system develops a search prioritization area based on function 1.1 above
   2.3 Display High, Medium, Low risk for suicide - display suicide intent scales and other questions about behaviours which are known to be warning signs for suicide

3. Timing - is essential to missing person enquiries (recall - three-quarters of disappearances reported to the Police are resolved within two days). MPSP will be required to forecast widening search areas across a weekly time scale.

MPSP is also intended to support strategic decision making around the allocation of resources, implementing and monitoring action at suicide 'hotspot' locations and reporting to the National Suicide Prevention strategy. To support multiagency working and reporting MPSP has the following additional functions:

4. Places of safety and local support services will be displayed on the map.
5. A layer of IMD inequality profiles (Index for Multiple Deprivation) will be mapped onto the Merseyside geography.

H. Data requirements

The analysis of missing persons, attempted suicide and completed suicide presents a multi-variable problem. As part of the ongoing research into missing persons, attempted and completed suicides, access has been granted to various datasets from the Merseyside region:

HM Coroner – access to completed suicides
North West Ambulance Service – access to attempted suicides
Merseyside Police – access to missing person cases

The MPSP system will maintain a dataset relating to missing persons, attempted and completed suicides, which can be used in its own right. Additionally, the data will be used by the MPSP Map application to assist the location of missing persons. A variety of data is required for MPSP from several different sources, which are summarized in Table 1.

TABLE 1: MPSP DATA

Function	Data Source
Average search distance for key subgroups	Grampian Police document Police data
Known suicide and attempted suicide hotspots	Highway Agency and Police data Public health data
Method of suicide for key subgroups	Mortality files data
Last known whereabouts	Police data
Significant areas	Police data
Suicide risk	Police data Suicide intent scales
Timing	Police data

Due to the sensitivity of the data it was necessary to obtain ethical approval and to ease the collection of the various data a data protocol has been established, which allows data entry according to various codes and options to select from. As MPSP is intended as an ongoing support system in the search for missing persons, each new case will present new data which will in turn feature in MPSP. As mentioned previously, the data will be stored in a database that can be queried in its own right and used for subsequent multivariate analysis and correlation studies. The database will also support the Map application by revealing the locations in a geographic format that can be used to assist a sear

IMPLEMENTATION

The MPSP system will be implemented using a Windows based Apache/MySQL/PHP (WAMP) platform in conjunction with Google Maps technology as considered below.

I. *MySQL/PHP*

MySQL is a popular open source database system. One of the main benefits of MySQL is that it can be scaled down to support embedded database applications and equally be scaled up to support huge datasets. MySQL is the de-facto standard database for web sites that support both huge volumes of data and users (e.g. Friendster, Yahoo and Google).

PHP is an open source server side scripting language (like ASP) that supports many databases (MySQL, Informix, Oracle, Sybase etc.). The MySQL/PHP combination provides an attractive database development environment due to the cross-platform characteristic (i.e. develop in Windows and serve on a Unix platform).

PHP script was then used to export the table data in XML format for use in conjunction with the Google Maps application. When location data is available in XML format the Map application can be developed using JavaScript.

J. *Google Maps technology*

Google Maps technology allows developers to embed Google Maps in web pages and make use of a number of services for customizing the map and adding useful content to the map. Originally Google Map development relied on the use of JavaScript, however, more recently Google Maps has expanded to support the development of Map applications using Adobe Flash. Google Maps technology is essentially a series of Maps APIs and the two main APIs used for the implementation of MPSP are:

- Maps JavaScript API allows pages to be developed using JavaScript, which is an object-oriented scripting language based on the Java programming language

- Maps API Web Services provide an interface for requesting Maps API data from external services and using them within Maps applications. The Web Services may be used for performing geocoding, generating driving directions, and obtaining elevation profiles

RELATED WORK

Surprisingly there is little research focused on the use of interactive GIS to assist missing person cases, or more specifically to identify suicide location correlated to a potential victim. One relevant work worthy of mention is that of [12] which considers the problem of finding a lost person from the spatial thinking of the search teams themselves. The work describes how similar use of hotspots and LKP are utilised but it tends to focus on those who are genuinely lost, such as children and missing hikers and less so on persons suffering mental health problems.

It is important to mention that in addition to the Police there are a number of voluntary organizations that help trace missing people. *Missing People* (formerly *National Missing Persons Helpline – NMPH*) www.missingpeople.org.uk is a charity to help find missing persons and offer support to relatives. Similarly, look4them.org.uk is a website which coordinates activities from a number of the organizations in the UK (including the Police and Salvation Army) to assist missing persons cases.

CONCLUSIONS AND FUTURE WORK - THE RESEACH CHALLENGES

The paper has highlighted the increasing problem of missing person cases faced by society and highlighted the burden placed on an over-worked Police force. The current approach to dealing with missing persons has been considered and the role that can be played by interactive GIS maps was highlighted. The remainder of the paper describes the direction and research challenges to be addressed in the development of the MPSP toolset. The research challenges are:

Development of appropriate algorithms to determine and map possible search locations of persons who are a suicide risk. Specifically, with non-domestic suicides individuals gravitate to localised suicide "hotspots". It is essential that we are able to develop the requisite algorithms based on local information matched to a potential victim's profile to identify such locations. That his develop a "look here" search strategy

Mapping services such as Google are open to the public, in addition to this much information about suicides, for example coroners data is confidential. However, for applications to function there will be a requirement to map or encode and layer information about deaths at specific locations with a high degree of accuracy and also display this information. It will be also necessary then to encode or enable a mapping service to access this information or provide a means of correlating specific locations with a high degree of accuracy, to specific events that obscures the details from public domain. The ability for this is twofold, to maintain confidentiality and to prevent unnecessary interest of highlighting specific locations with "ghoulish" connotations or for that matter highlighting the locations that may draw potential suicides to that location on a public map interface.

The current research and development is based on the Merseyside region, thus current data is localised. Additionally, police forces across the UK not only record missing persons using a variety of systems and do not record standard facts uniformly across force regions. Thus, the localised map is a thematic map. This is a type of map especially designed to show a particular theme connected within a specific geographic area. However, for a generalised solution it will be necessary to devise an approach that enables representation of local features to local data by police region. The problem is compounded by the varying ways in which localised data is recorded and formatted. It is necessary to devise a standard classification in order that this information can be recorded and formatted and indeed can be formatted across mapping software.

National Police Improvement Agency, Missing Persons Beureau, "Missing Persons: Data and Analysis 2009/2010"
http://www.npia.police.uk/en/docs/Missing_Persons_Data_and_Analysis_2009-10.pdf

PACT, "Every 5 Minutes: A review on the available data on missing children in the UK", 2006
http://uk.missingkids.com/en_GB/docs/Every_5_Minutes_PACT.pdf

R. Tarling, J. Burrows, "The nature and outcome of going missing: the challenge of developing effective risk assessment procedures", International Journal of Police Science and Management, 2004.

G. Newiss, "Understanding the risk of going missing: estimating the risk of fatal outcomes in cancelled cases", Policing: An International Journal of Police Strategies & Management, 2006.

National Police Improvement Agency , "ACPO Guide on the Management, Recording and Investigation of Missing Persons", 2011
http://www.acpo.police.uk/documents/crime/2011/201103CRIIMP02.pdf

ESRI, "When Every Second Counts", 2009, http://www.esri.com/news/arcuser/0609/yosar.html

M. McDaniel, "Agent-Based Modeling of Lost Person Wayfinding", MA Thesis, University of California, 2010 http://www.geog.ucsb.edu/graduates/masters-theses/pdf/McDanielM_MastersThesis.pdf

WPC Software, "COMPACT Police Missing Persons System", 2010, http://www.wpcsoft.com/products/compact.aspx

Grampian Police, " Missing persons: understanding planning and response", Aberdeen, 2007.

N. Biehal, F. Mitcell, J. Wade, "Lost from view: Missing persons in the UK", Social Work Research and development Unit, University of York, 2003.

K. Shalev, M. Schaefer, A. Morgan "Investigating missing person cases: how can we learn where they go or how far they travel", International Journal of Police Science and Management, 11(2), pp 123-129, 2009.

C.D. Heth, E.H. Cornell, "A eographic Information System for Managing Search for Lost Persons", In G. Allen (Ed.), *Applied spatial cognition: From research to cognitive technology*. Mahwah, NJ: Lawrence Erlbaum Associates,