DEFENCE S&T TECHNICAL BULLETIN

VOL. 13 NUM. 2 YEAR 2020 ISSN 1985-6571

CONTENTS

BERSEKUTU BERTAMBAH MUTU

Ministry of Defence
Malaysia

SCIENCE & TECHNOLOGY RESEARCH
INSTITUTE FOR DEFENCE (STRIDE)

ISSN 1985-6571
9 771985 657008

AIMS AND SCOPE

The Defence S&T Technical Bulletin is the official journal of the Science & Technology Research Institute for Defence (STRIDE). The journal, which is indexed in, among others, Scopus, Index Corpenicus, ProQuest and EBSCO, contains manuscripts on research findings in various fields of defence science & technology. The primary purpose of this journal is to act as a channel for the publication of defence-based research work undertaken by researchers both within and outside the country.

WRITING FOR THE DEFENCE S&T TECHNICAL BULLETIN

Contributions to the journal should be based on original research in areas related to defence science & technology. All contributions should be in English.

PUBLICATION

The editors' decision with regard to publication of any item is final. A manuscript is accepted on the understanding that it is an original piece of work that has not been accepted for publication elsewhere.

PRESENTATION OF MANUSCRIPTS

The format of the manuscript is as follows:

a) Page size A4
b) MS Word format
c) Single space
d) Justified
e) In Times New Roman, 11-point font
f) Should not exceed 20 pages, including references
g) Texts in charts and tables should be in 10-point font.

Please e-mail the manuscript to:

1) Gs. Dr. Dinesh Sathyamoorthy (dinesh.sathyamoorthy@stride.gov.my)
2) Dr. Mahdi bin Che Isa (mahdi.cheisa@stride.gov.my)

The next edition of the journal (Vol. 14, Num. 1) is expected to be published in April 2021. The due date for submissions is 13 January 2021. **It is strongly iterated that authors are solely responsible for taking the necessary steps to ensure that the submitted manuscripts do not contain confidential or sensitive material.**

The template of the manuscript is as follows:

TITLE OF MANUSCRIPT

Name(s) of author(s)

Affiliation(s)

Email:

ABSTRACT

Contents of abstract.

Keywords: *Keyword 1; keyword 2; keyword 3; keyword 4; keyword 5.*

1. TOPIC 1

Paragraph 1.

Paragraph 2.

1.1 Sub Topic 1

Paragraph 1.

Paragraph 2.

2. TOPIC 2

Paragraph 1.

Paragraph 2.

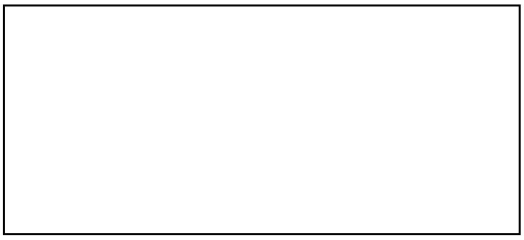


Figure 1: Title of figure.

Table 1: Title of table.

Content	**Content**	**Content**
Content	Content	Content
Content	Content	Content
Content	Content	Content

Equation 1 (1)

Equation 2 (2)

REFERENCES

Long lists of notes of bibliographical references are generally not required. The method of citing references in the text is ‘name date’ style, e.g. ‘Hanis (1993) claimed that...’, or ‘…including the lack of interoperability (Bohara *et al.*, 2003)’. End references should be in alphabetical order. The following reference style is to be adhered to:

<u>Books</u>

Serra, J. (1982). *Image Analysis and Mathematical Morphology*. Academic Press, London.

<u>Book Chapters</u>

Goodchild, M.F. & Quattrochi, D.A. (1997). Scale, multiscaling, remote sensing and GIS. *In* Quattrochi, D.A. & Goodchild, M.F. (Eds.), *Scale in Remote Sensing and GIS*. Lewis Publishers, Boca Raton, Florida, pp. 1-11.

<u>Journals / Serials</u>

Jang, B.K. & Chin, R.T. (1990). Analysis of thinning algorithms using mathematical morphology. *IEEE T. Pattern Anal.*, **12**: 541-550.

<u>Online Sources</u>

GTOPO30 (1996). *GTOPO30: Global 30 Arc Second Elevation Data Set*. Available online at: http://edcwww.cr.usgs.gov/landdaac/gtopo30/gtopo30.html (Last access date: 1 June 2009).

<u>Unpublished Materials (e.g. theses, reports and documents)</u>

Wood, J. (1996). *The Geomorphological Characterization of Digital Elevation Models*. PhD Thesis, Department of Geography, University of Leicester, Leicester.

X-RAY PHOTOELECTRON SPECTROSCOPY (XPS) AND CURRENT CAPACITY STUDY OF Al-Zn AND Al-Zn-Sn ALLOYS DISSOLUTION BEHAVIOUR IN TROPICAL SEAWATER

Mahdi Che Isa*, Nik Hassanuddin Nik Yusoff, Mohd Subhi Din Yati, Mohd Moesli Muhammad & Hasril Nain

Science & Technology Research Institute for Defence (STRIDE), Ministry of Defence, Malaysia

*Email: mahdi.cheisa@stride.gov.my

ABSTRACT

The passivation of the Al-Zn-xSn alloys can be explained by the spontaneous formation of a protective oxide film that further impedes the reaction of aluminium with aggressive environments. The properties of corrosion product and oxide layer on Al-Zn-xSn alloys in tropical seawater were studied at room temperature using current capacity measurements and X-ray photoelectron spectroscopy (XPS) techniques. The results showed that alloying addition was observed to influence the electrochemical behaviour of this alloy. The current capacity measurement showed that 0.5 Sn (wt.%) improved the anode performance by increasing its current efficiency. However, further increase of the wt.% of Sn produced a negative impact on the alloys by reducing the value of its current efficiency. The presence of 1.5 Sn (wt.%) in the alloy increased the formation of local cathodic sites and thus, reduced anode efficiency to only 70%. The study also showed that the alloys with low Sn content and impurities (Fe) produce anode efficiency of up to 85%. Based on the XPS analysis, the oxide film formed on the Al-Zn-xSn alloy consists of a mixture of SnO and SnO_2, which play a key role in creating oxide layer defect, reducing the electrical resistance at metal-electrolyte interface and activating the electrochemical dissolution on the alloy surface.

Keywords: *Al-Zn-Sn alloy; oxide layer; x-ray photoelectron spectroscopy (XPS); electrochemical; current capacity.*

1. INTRODUCTION

Al-Zn-Sn alloy is among the preferred materials to be used as sacrificial anodes for cathodic protection of steel structures in marine environments due its reasonable cost, availability and high theoretical current capacity (2,900 Ah/kg) (Salinas & Bessone, 1991; Tamada & Tamura, 1993; Kamarudin *et al*., 2010; Isa *et al*., 2012; Khireche *et al*., 2014). However, the electrochemical behaviour of this type of alloy is strongly influenced by the formation of a passive oxide film on the surface when exposed to oxidising conditions, such as water, air, or other oxygen-containing fluids and gases (El Shayeb *et al*., 2001; Gudić *et al*., 2001; Kyung-Keun & Kwang-Bum, 2001; Munoz *et al*., 2002; Gudić *et al*., 2005; Evertsson *et al*., 2015; Ferdian *et al*., 2017; Šekularac & Milošev, 2018).

In order to promote surface activation, pure aluminium is usually alloyed with Zn (5 wt.%) and small quantities of elements (In, Ga, Sn, Bi), as well as applied with oxide deposition (RuO_2) and catalytic coating (RuO_2, IrO_2) (Bessone *et al.,* 2005; Shibli & George, 2007; Gudić *et al*., 2010; Anshuman *et al.,* 2011; Mohammad & Ahmad, 2011; Xiong *et al*., 2011; Flamini & Saidman, 2012; Din Yati *et al.,* 2014). When an alloying element is added, the properties of its oxide layer will change and the alloy can provide adequate electrochemical reaction for specific applications, such in marine environments (El Shayeb *et al*., 2001; Jun-guang *et al*., 2011; Senel & Nisancioglu, 2014). Alloying elements are

used in order to shift the corrosion potential towards sufficiently electronegative values and to produce a more uniform attack morphology (Isa *et al*., 2010; Khireche *et al*., 2014).
The combination of electrochemical methods with surface analytical techniques offers very useful methods to study charge the transfer process, oxide composition of corrosion products, oxidation states, as well as thickness and structure of the oxide layer (Forget *et al*., 2003; Lei *et al*., 2010; Duchoslav *et al*., 2014; Winiarski *et al*., 2016; Ramya *et al*., 2018). It has been reported elsewhere that X-ray photoelectron spectroscopy (XPS) has been used to study surface chemistry, bonding structure and composition of surfaces and interfaces of materials (Gredelj *et al*., 2001; Moffitt *et al*, 2001; Natishan *et al*., 2002; Xingwen & Guoqiang, 2004; Zähr *et al*., 2012).

The aim of this work is to understand the effects of Sn addition on the aspect of Al-Zn alloy dissolution behaviour and characterise the passive oxide layer properties subjected to current capacity test in tropical marine seawater with the aid of the XPS technique.

2. MATERIALS & METHODS

2.1 Alloy Preparation and Anode Performance

Four alloy compositions were fabricated through the conventional casting method under inert atmosphere, with the method being explained in detail in Isa *et al*. (2010). The nominal composition of the alloy was Al-5.5Zn-xSn (%wt.) while Sn additions were made from 0.5 to 1.5 %wt. Part of the samples were cut, cleaned and submitted for wet chemical analysis using a Varian Spectra AA-10 Atomic Absorption Spectrometer (AAS). The analysis was based on procedures stated in ASTM E 34, (2011). Both anode (Al-Zn-Sn alloys) and cathode (steel) specimens were polished to 1,200 grit, washed and finally rinsed with acetone. The individual dimensions for the round shape anode were 0.3 cm in thickness and 1.8 cm in diameter. The cathode with rectangular shape had dimensions of 0.8 x 4.1 x 16.0 cm. The separation or distance between the anode and cathode was 10 cm and the anode weight was recorded before the test was started. Both the anode and cathode were immersed in a lightproof Perspex tank containing 30 L of filtered tropical seawater medium for about 72 h. A current density was supplied to the anode at amount of 0.5 mA/cm^2 and charge transfer reading was recorded using coulometer for 72 h for anode capacity determination. The anode specimen was then washed with water, soft brushed and dried. For weight loss determination, the anode specimen was dipped in cleaning solution (50 g/L chromic acid) for 20 s. The sample was then washed with water, rinsed with acetone and dried to determine their weight loss (ASTM G 97, 2018). Anode specimens that undergone several stages of grinding with SiC paper grit 600 to 1200 will then be polished with ultrafine cloth paper and diamond paste. Finally, in order to reveal the material microstructure, the anode specimen will undergo the etching process using Kellers solution.

2.2 X-Ray Photoelectron Spectroscopy (XPS) Study

Surface analysis was carried out to study the nature and composition of the corrosion product layers after the samples were exposed to natural seawater through galvanic coupling for a period of three days during the current capacity test. The chemical composition of the corrosion product on the alloy surface after immersion in seawater was examined using XPS. The data was presented as XPS spectrums that were collected using a Kratos XSAM-HS spectrometer located at the Centre for Research & Instrumentation Management (CRIM), Universiti Kebangsaan Malaysia (UKM). The Al-Zn and Al-Zn-Sn alloy samples were mounted on a copper sample holder with double sided carbon tape and were kept under vacuum in the spectrometer for 1 h to remove physisorbed water. The experiments were conducted in an ultra-high vacuum (UHV) system, where monochromatic Mg K$_a$ X-rays (1,253.6 eV) were used as the excitation source and a pressure of 4 x 10^{-9} Torr was maintained in the spectrometer chamber. An accelerating voltage of 14 kV and an anode current of 10 mA were employed during analyses. The analyser was operated in the fixed analyser transmission mode at a pass energy of 160 eV. High resolution scans were recorded for the following XPS lines; O 1s, Al 2p, C 1s, Sn 3d, Cl 2p, Mg 2p and Zn 2p 3/2 using a 1 eV step size and dwell time of 0.1 s per step with recommended range of binding energy with pass energy of 20 eV. The spectrometer was calibrated using clean and pure Ag plates. The XPS data consisted of survey scans over the entire binding energy range (30 – 1,153.6 eV) and selected scans on the core-level photoelectron peaks of interest. In order to minimise experimental error, all the samples were measured using the same sensitivity settings, intensity of X-ray radiation and analyser pass energy. The charge correction involved the use of the C 1s peak (284.5 eV) (Warret *et al*., 1996). XPS data analysis (i.e., smoothing, background subtraction and curve fitting) was undertaken using the accompanied software according to procedures reported in the literature (Chastain, 1992; Briggs & Seah, 1996; Bluhm, 2011).

3. RESULTS & DISCUSSION

3.1 Chemical Compositions & Anode Performance

The chemical composition results obtained using the wet analysis method (AAS) for the fabricated Al-Zn-xSn alloys are shown in Table 1. This method was chosen because it has the capability to detect low levels of element concentration and is recommended as one of the quality control methods for alloy production (Lenar, 2000). The amount of Sn was found to be less than the nominal compositions. This could be as a result of evaporation or loss during the melting process due to the differences in the melting point of the individual elements. It has been reported that the presence of impurities such as Fe in these types of alloys are found to be deleterious or can cause a detrimental effect towards anode performance (Salinas & Bessone, 1991; Breslin *et al*., 1993). In order to overcome this weaknesses, elements such as zinc and stanum are be added to control the adverse effects of the impurities (Munoz *et al.,* 2002; Senel & Nisancioglu, 2014).

Alloying is made for the purposes of improving the performance of the aluminium alloy for specific applications. The alloying element is also useful in producing better metallurgical properties, such as smaller grain size in anode materials (Salinas *et al*., 1999; Mathiyarasu *et al*., 2001). Besides that, the addition of alloying elements is capable of mitigating the adverse effects of impurities presence in the alloy. Previous researchers reported that through the addition of elements such as Sn, In or Ga, the alloy has better surface activity, dissolution morphology and dissolution products (Lin & Shih, 1987; Talavera *et al*, 2002; Isa *et al*., 2010; 2012).

Table 1: Chemical composition of the fabricated Al-Zn-xSn alloys analysed using AAS.

Bal : Balance ; ND : Not Detected

Sample (Nominal Composition, wt.%)	Contents (wt.%)			
	Zn	Sn	Fe	Al
Al-5.5Zn	5.52	-	0.003	Balance
Al-5.5Zn-0.5Sn	5.49	0.45	0.003	Balance
Al-5.5Zn-1.0Sn	5.52	0.92	0.011	Balance
Al-5.5Zn-1.5Sn	5.45	1.33	0.018	Balance

The anode performance or current capacity test is designed to simulate the service operating condition of sacrificial anodes. The basic anode properties obtained from this test are anode capacity and efficiency. Anode efficiency is defined as the ratio between the actual and theoretical values of current capacity, and is evaluated using the following equation (ASTM G97, 2018).

$$\text{Anode efficiency (E \%)} \quad = \quad \frac{\text{Actual anode capacity}}{\text{Theoretical anode capacity}} \text{ X } 100\% \qquad (1)$$

where:

$$\text{Actual anode capacity (Ah/kg)} = \frac{\text{Ah}}{\text{kg}} = \frac{\text{Ah}}{(M_1 - M_2)} \text{ X } 1000 \qquad (2)$$

where M_1 is the initial mass of the aluminium alloy and M_2 is the final mass (in grams). The theoretical anode capacity is:

$$\text{Theoretical anode capacity (Ah/kg)} = \frac{(96{,}480 \text{ C} / 3{,}600 \text{ S})}{\text{Equivalent weight of aluminium alloys (kg)}} \qquad (3)$$

The values of anode capacity and efficiency were calculated and listed in the Table 2. It shows that a small amount of Sn content (0.5 wt.%), which has been classified as activator in this type of alloy, can cause a significant increase in anode efficiency. The presence of impurities in the alloy, usually heavy metal (Fe), and secondary phases, such as low solid solubility alloying element (Sn), in the aluminium led to the formation of a local cathodic micro-cell on the anode surface, thus giving detrimental effect on the anode performance. For example, the Al-Zn-Sn sample with 0.5Sn (wt.%) content produced anode capacity of 2,787.56 Ah/kg, while the Al-Zn-Sn sample with 1.5Sn (wt.%) content produced anode capacity of 2,756.71 Ah/kg. These samples recorded decreasing efficiency with increasing Sn contents.

The lower values of anode capacity indicate that the anodes are able to supply only a partial amount of current for protection. The missing percentage represents the amount of charge lost, which is no longer useful for cathodic protection. The principal factors affecting anode wastage are hydrogen evolution due to local cathodic action as a result of alloying addition and also mechanical loss from the anode surface known as chunk effect. The decrease of the anode efficiency for Al-Zn-Sn alloys is mainly caused by the presence of high Sn content which is more than 0.5Sn (wt.%) and impurity, that can badly influence anode performance (Breslin *et al*, 1993; Bruzzone *et al.*, 1997; Barbucci *et al.*, 1998).

Table 2: Anode efficiency for Al alloys in aerated tropical seawater at 27 °C and pH 8.1.

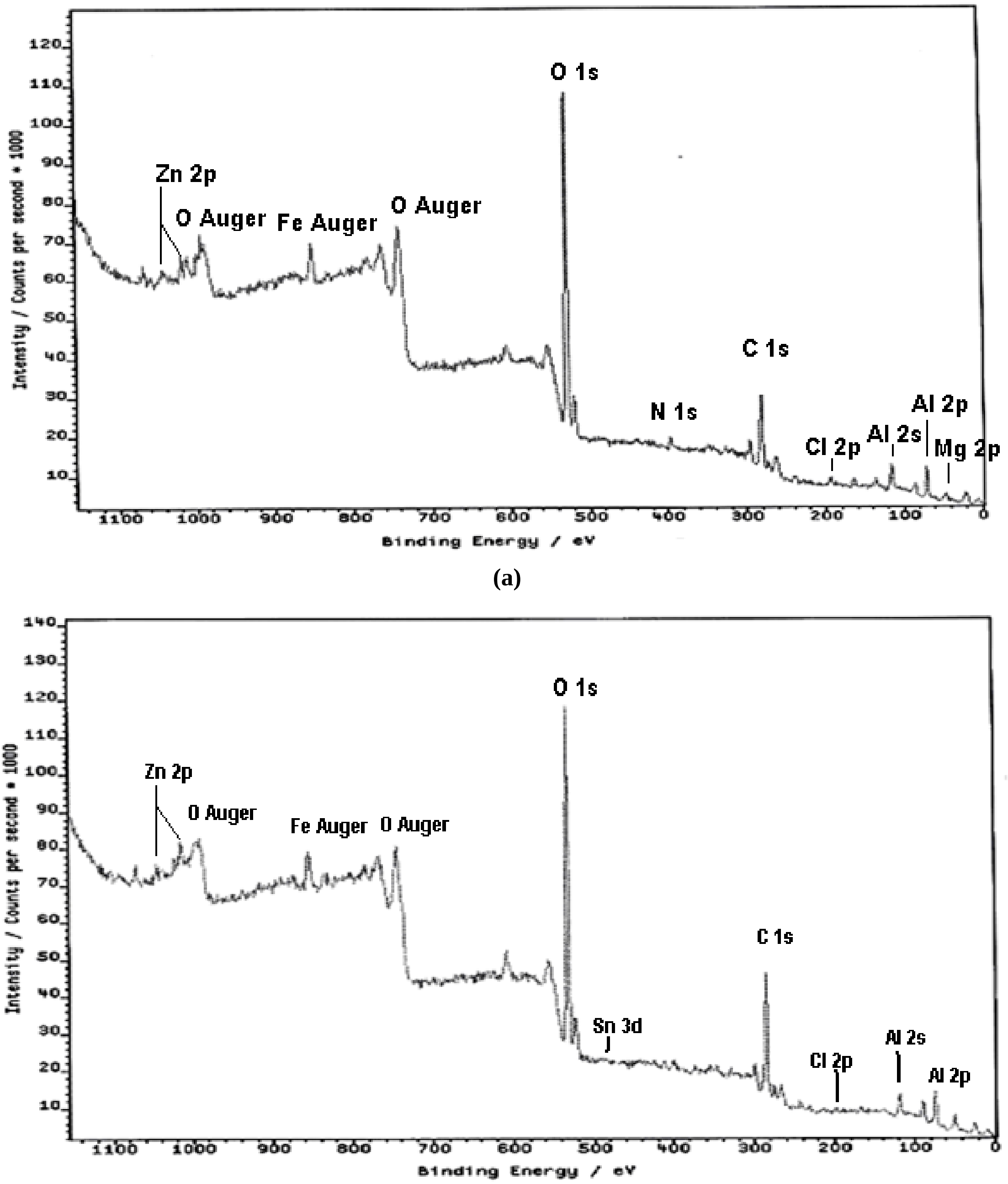
Samples	Actual current capacity (Ah/kg)	Theoretical current capacity (Ah/kg)	Anode efficiency (%)
Al-5.5Zn	2012.52	2798.92	71.90
Al-5.5Zn-0.5Sn	2368.32	2787.56	84.96
Al-5.5Zn-1.0Sn	2195.69	2770.24	79.14
Al-5.5Zn-1.5Sn	1950.10	2756.71	70.74

3.2 XPS Study

The XPS study was carried out to differentiate the chemistry of the corrosion product and oxide film on the surface of the alloys. The XPS survey spectra of the Al-5.5Zn and Al-5.5Zn-0.5Sn alloys after the current capacity test in natural tropical seawater of pH 8.1 are presented in Figure 6. The peaks of oxygen, carbon, aluminium, zinc and tin are visible in the spectrum. The XPS spectra of Al-Zn-Sn alloy is very similar to those of Al-Zn, except that the XPS spectra of Al-Zn-Sn shows a pronounced peak of Sn (Sn 3d) at binding energy 483.6 – 488.2 eV. In these spectra, characteristic peaks are recorded for aluminium (Al 2p) at 74.4-75.3 eV, oxygen (O 1s) at 531.2 - 533.2 eV, auger at 750.8

eV, chloride (Cl 2p) at 199.2 eV, and carbon (C 1s) at 284.5 eV as residual from the oil vapours of the diffusion pump (Werret *et al.*, 1996).

Table 3 shows the binding energy (BE) values for the different elements such as Al, Zn and Sn in both alloys after the high-resolution scanning and fitting process. These finding are in line results reported by Chasstain (1992), Wöll (2007), Uhart *et al.* (2016), Tardio & Cumpson (2018) and Greczynski & Hultman, (2020). The XPS survey spectrum of the Al-Zn-Sn in Figure 7 shows clear Sn peaks (Sn 3d 483.6 eV) especially after surface etching. The appearance of Sn peaks in the XPS spectra of the Al-Zn-Sn alloy indicates that the alloy surface contains Sn that is responsible for oxide layer modification as shown in the electrochemical impedance spectroscopy (EIS) study conducted by Isa *et al.* (2010). The presence of Sn in the alloy increases the number of flawed regions on the surface and initiates the surface activation process as indicated by the presence of inductive loops in the EIS diagrams (Venugopal & Raja, 1997; Isa *et al.*, 2010).



(a)

(b)

Figure 6: XPS survey scan for the (a) Al-5.5Zn and (b) Al-5.5Zn-0.5Sn alloy surfaces before etching.

Table 3: Binding energy values of the proposed components for Al alloys after the current capacity test in tropical seawater at pH 8.1.

Samples	Binding energy, corrected (eV)		
	Element (shell)	Before etching (Proposed species)	After 6 min of etching (Proposed species)
Al-5.5%wt.Zn (Al-1)	Al 2p	(i) 75.4 (Al_2O_3)	(i) 75.3 (Al_2O_3) (ii) 72.7 (Al)
	Zn 2p	(i) 1022.4 (ZnO) (ii) 1045.7 (ZnO)	(i) 1022.3 (ZnO) (ii) 1045.6(ZnO)
	O 1s	(i) 533.2 (M-OH)	(i) 531.7 (M-O) (ii) 533.2(M-OH)
	Cl 2p	(i) 198.0 (Cl) (ii) 199.5 (M-Cl_2)	(i) 198.5 (Cl) (ii) 200.3 (M-Cl_2)
Al-5.5%wt.Zn-1.5%wt.Sn (Al-2)	Al 2p	(i) 75.3(Al_2O_3)	(i) 74.9 (Al_2O_3) (ii) 72.9 (Al)
	Zn 2p	(i) 1022.0 (ZnO) (ii) 1045.6 (ZnO)	(i) 1022.3(ZnO) (ii) 1045.3 (ZnO)
	O 1s	(i) 533.2(M-OH)	(i) 531.6(M-O) (ii) 533.2 (M-OH)
	Cl 2p	(i) 199.2 (Cl) (ii) 199.8 (M-Cl_2)	(i) 198.5 (Cl) (Cl) (ii) 199.8 (M-Cl_2)
	Sn 3d	(i) 483.6 (Sn) (ii) 486.2 (Sn^{2+}) (iii) 488.2 (Sn^{4+})	(i) 483.8 (Sn) (ii) 486.2 (Sn^{2+}) (iii) 488.3 (Sn^{4+})

Possible modifications in the chemistry of the oxide film that are responsible for active dissolution of Al-Zn as a result of Sn addition were further examined using high resolution scanning XPS. The Al (2p) spectra obtained from these two alloys are shown in Figure 8. In the case of Al-Zn, the spectra revealed only one peak at 74.4 eV arising from Al^{3+} in the oxide film (Figure 8(a)). However, after the 6 min etching process, the XPS spectra for this alloy revealed two peaks at 72.7 and 75.3 eV (Figure 8(b)) arising from Al in the substrate and Al^{3+} in the oxide film respectively, dependent on oxidation state and coordination number of the Al atom (Barr *et al*, 1994; 1997; Zähr *et al*, 2012). The spectra measured for the Al-Zn-Sn sample also shows one peak (75.3 eV) corresponding to the Al^{3+} in the oxide film (Figure 9(a)), and reveals peaks at 72.9 and 74.9 eV, also arising from Al in the substrate and Al^{3+} in the oxide film respectively (Figure 9(b)). This indicates the presence of the metallic element in both alloys after 6 min of etching. The oxidation shift between the metallic and oxidised Al peaks is around 2.0 to 2.6 eV and agrees well with reported peak separation values (Arranz & Palacio, 1996; Gredelj *et al.*, 2001; 2002).

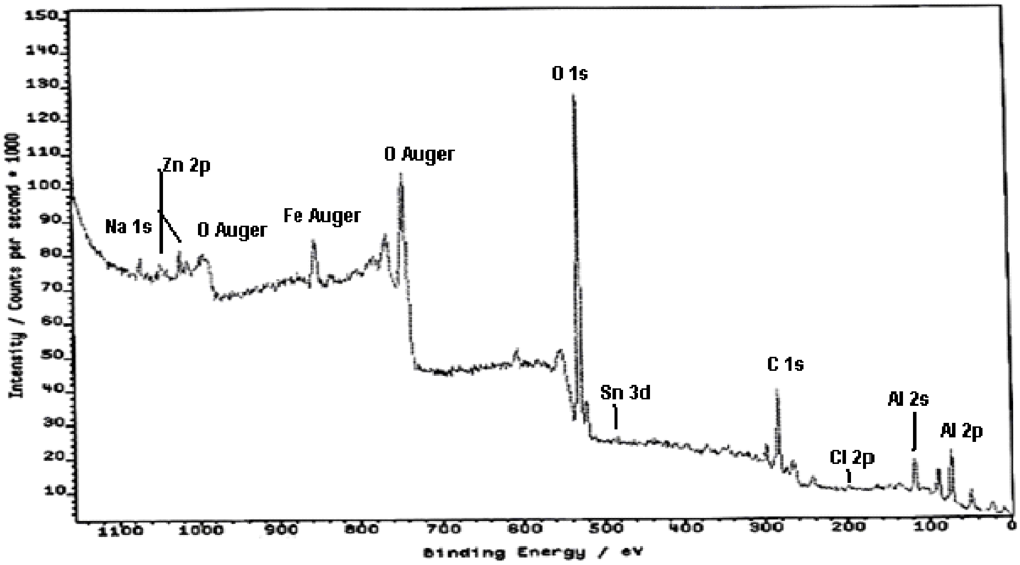
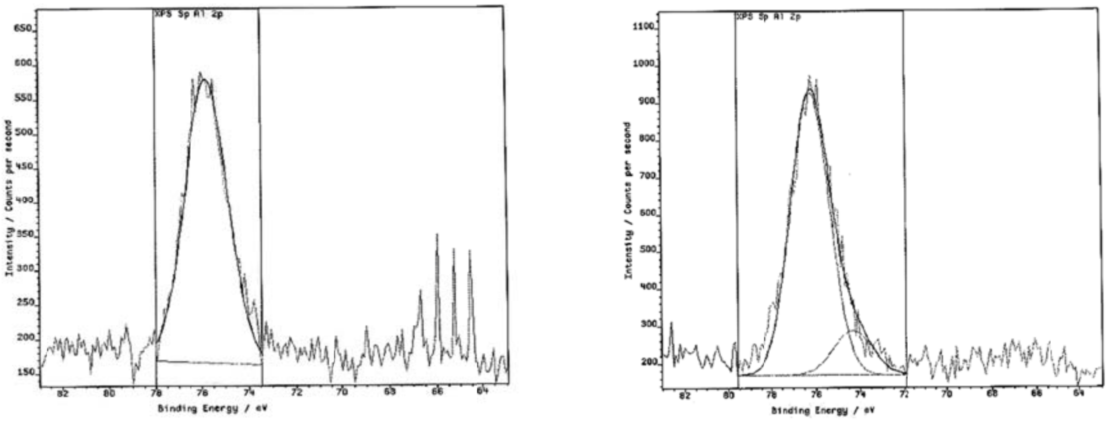
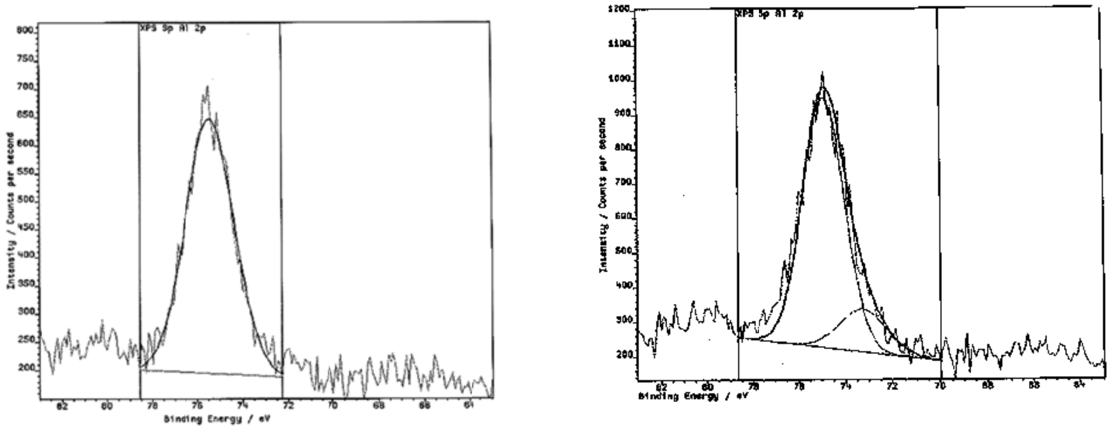


Figure 7: XPS survey scan for the Al-Zn-Sn alloy surface after 6 min of etching.



(a) (b)

Figure 8: High resolution XPS spectra and peak deconvolution for Al 2p in the Al-Zn alloy: (a) Before etching (b) After etching.

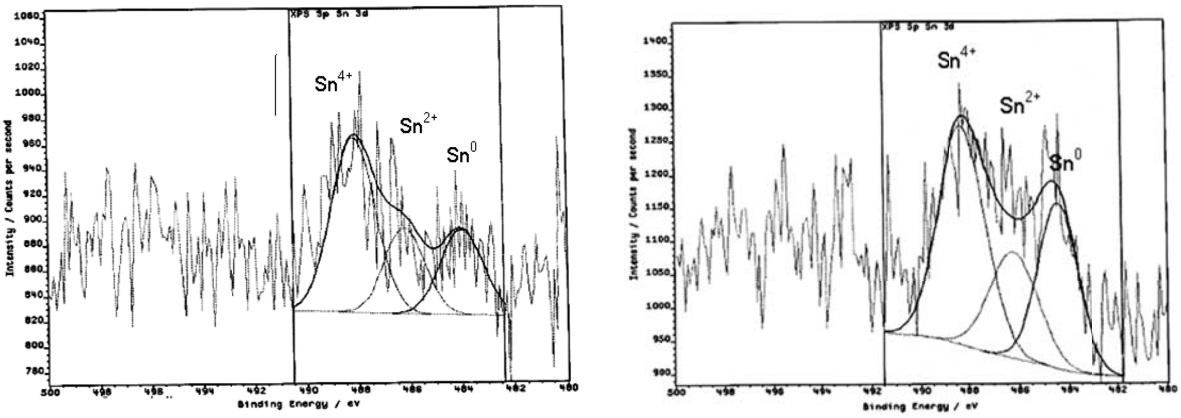


(a) (b)

Figure 9: High resolution XPS spectra and peak deconvolution for Al 2p in the Al-Zn-1.5Sn alloy: (a) Before etching (b) After etching.

Figure 10 shows the Sn 3d spectra of the Al-Zn-Sn alloy, which clearly shows the existence of $3d_{5/2}$ energy level, becomes broader after sputtering. The peak contains a shoulder and is deconvulated for further analysis. From the deconvoluted Sn 3d 5/2 peak, one can determine that it is built-up as a mixture of three components. The peaks at 488.2, 486.2 and 483.6 eV are attributed to Sn^{4+}, Sn^{2+} and Sn^{0} respectively. These values are in accordance with binding energies values presented by other authors (Sistiaga *et al.*, 1998; Batzill & Diebold, 2005; Aragón *et al*, 2015).

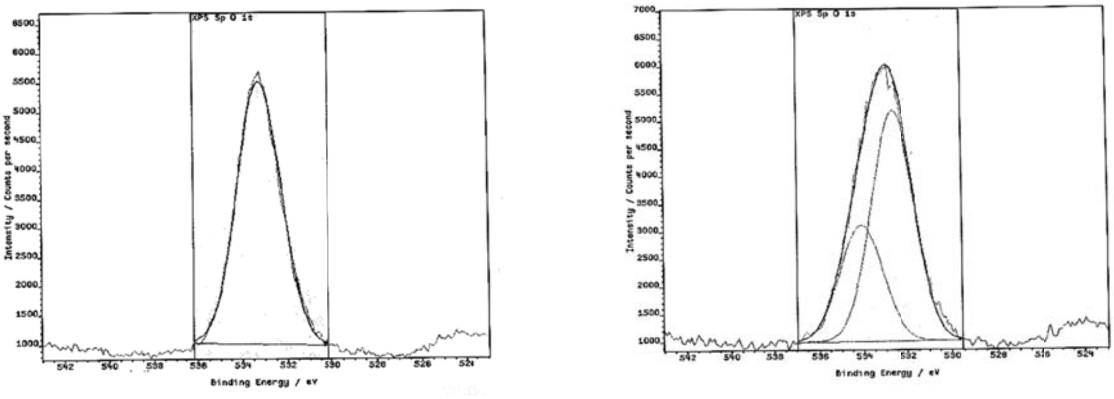
The rising intensity of the shoulder after 6 min of sputtering indicates that there was an apparent increase in the concentration of Sn^{4+} and Sn^{2+}, which suggests that Sn^{4+} coexists with Sn^{2+} in the oxide film (Figure 10(b)). The binding energy (486.2 eV) corresponds to the value reported for oxidised tin (Chastain, 1992; Aragón *et al*, 2015). On the basis of these results, it can be concluded that oxidised tin (most likely in the form of SnO and SnO_2) can coexist, as the potential difference between these species in only 2.0 eV (Szuber *et al.*, 2001; Neudachina *et al.*, 2005; Wu *et al.*, 2007). The present results are supported further by the fact that Sn^{2+} and Sn^{4+} have been found in oxide films of Al-Sn, Pb-Sn and Sn-Ag alloys (Kudo *et al.*, 1996; Hung *et al.*, 2006; Neveu *et al.*, 2006; Zhang *et al.*, 2009).



(a) **(b)**

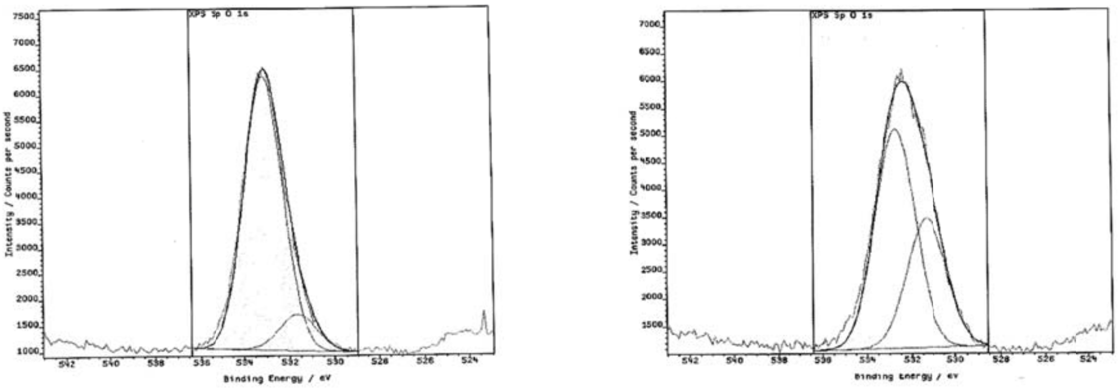
Figure 10: High resolution XPS spectra and peak deconvulation for Sn 3d in the Al-Zn-1.5Sn alloy: (a) Before etching (b) After etching.

Figures 11 and 12 show the measured O1s signal recorded for the Al-Zn and Al-Zn-Sn alloys respectively. The peak in the O1s spectra is centred at 533.2 eV for both the alloys, which could suggest that the surfaces of the alloys before etching are mainly covered with hydroxide, and the presence of adsorbed or coordinated water molecules (Chasstain, 1992; Wöll, 2007; Uhart *et al.*, 2016; Tardio & Cumpson, 2018). The O1s spectra after etching consisted of a peak that is attributable to the presence of two components, which were not quite resolved from each other but were deconvulated using a Gaussian function centred at 531.7 and 533.2 eV for the Al-Zn alloy. Meanwhile, for the Al-Zn-Sn alloy, the Gaussian function is centred at 531.6 and 533.2 eV.



(a) **(b)**

Figure 11: High resolution XPS spectra and peak deconvulation for O1s in the Al-Zn alloy: (a) Before etching (b) After etching.

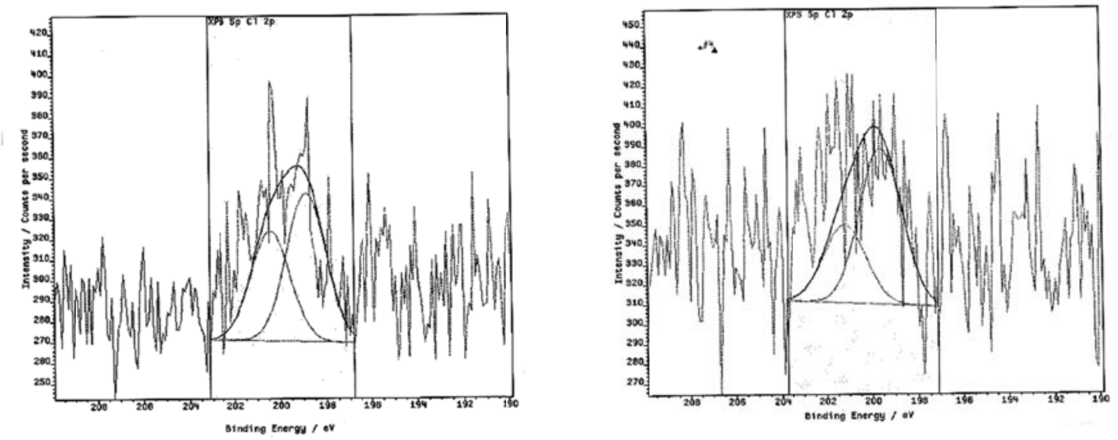


(a) **(b)**

Figure 12: High resolution XPS spectra and peak deconvulation for O1s in the Al-Zn-1.5Sn alloy: (a) Before etching (b) After etching.

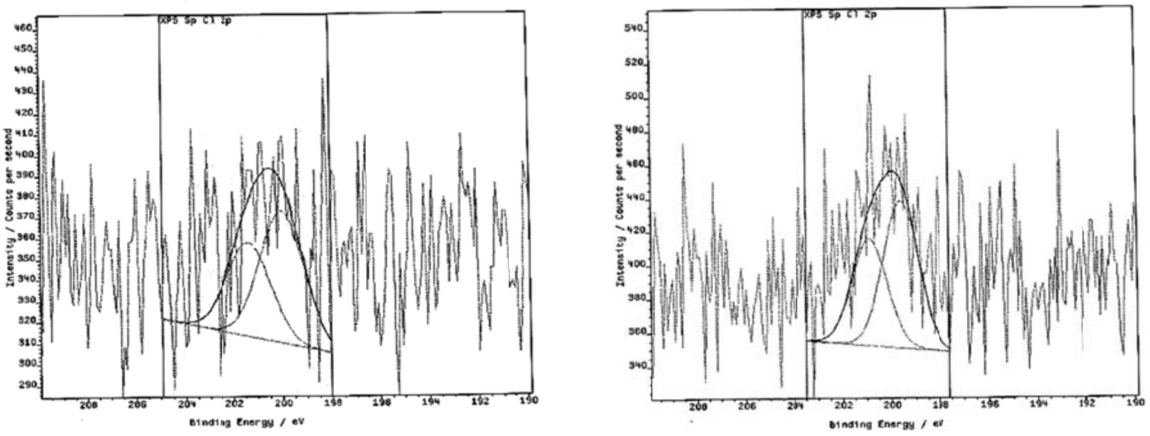
The results from the fitting analysis reveals two peaks with binding energies of 531.7 and 533.2 eV, indicating that oxygen was present in different forms. The higher binding energy peak came from oxygen in the M-OH bond, while the lower binding energy peak arose out of the M-O bond. Therefore, the O1s peak at 531.7 eV could be assigned to O in SnO and SnO_2. The binding energy at 533.2 eV, which is higher than those for tin oxides, may have originated from the OH groups, which is a result of water adsorption on the surface of SnO_x film (Kwoka *et al.*, 2005, 2006).

The influence of chloride on the chemical composition of the corrosion product and / or passive layers formed on the alloy surfaces after the current capacity test in natural seawater was also studied. The presence of Cl 2p peak in the spectrum of both samples (Figures 13 and 14) before and after etching indicates that chloride ions incorporated into the corrosion product and oxide film of the alloys. The Cl (2p) spectrum measured for the Al-Zn and Al-Zn-Sn alloys before etching reveals that the intensity of the peak is more intense for the Al-Zn-Sn alloy. Furthermore, the intensity of the chloride peak from the Al-Zn alloy decreased after etching. The Al-Zn-Sn alloy showed more intense chloride peaks for even 6 min of etching of the corrosion product layer.



(a) (b)

Figure 13: High resolution XPS spectra and peak deconvulation for Cl2p in the Al-Zn alloy: (a) Before etching (b) After etching.



(a) (b)

Figure 14: High resolution XPS spectra and peak deconvulation for Cl2p in the Al-Zn-1.5Sn alloy: (a) Before etching (b) After etching.

This confirmed that chloride ions migrated quickly from the film-solution interface to the metal-film interface when Al-Zn was alloyed with Sn. In the case of the Al-Zn alloy, the migration mechanisme of chloride ion was not the same as in the Al-Zn-Sn alloy. The higher content of OH compound and intensity of chloride peak measured for the Al-Zn-Sn alloy indicates that the top surface of the alloy could be in the form of hydroxychlorite complex, whereas the inner layer mainly consists hydroxide. Therefore, it can be suggested that the higher hydroxide content and increased hydroxychlorite formation were responsible for the oxide layer breakdown in the Al-Zn-Sn alloy. This observation is in agreement with the current capacity measurement, which showed that the performance of Al-Zn alloy was lower as compared to Al-Zn-Sn alloy. Based on the results of the study, it can confirmed that the oxide layer of the Al-Zn-Sn alloy is less stable due to the incorporation of Sn oxides, which increases the ionic activity at the layer.

4. CONCLUSION

In the present study, the properties of oxide layers formed on Al-Zn and Al-Zn-Sn alloys in tropical seawater solution was studied using XPS as well as current capacity measurement. The dominant oxides formed on the surfaces of both the Al-Zn and Al-Zn-Sn alloys after the current capacity test in seawater were Al_2O_3 and hydrated oxide. After 6 min of etching and high resolution XPS scanning, the oxide layer of Al-Zn-Sn contained a small amount of suboxide SnO and SnO_2, suggesting that Sn^{2+} and Sn^{4+} coexist in the oxide layer, and are responsible for surface defects as well as activation process. The XPS results showed that the existence of Sn^{2+} and Sn^{4+} in the oxide film would be expected to generate more defect sites for chloride attacks to the oxide layer of the base alloy.

Current capacity or anode performance was explored to determine the charge movement or ion mobility of these alloys and proved to be useful to indicate the influence of alloying elements in enhancing the dissolution process at the alloy-solution interface. The higher performance of Al-Zn-Sn alloys can be correlated to the formation of more defects in the oxide layer, which can allow for more reactions, aggressive anion attacks and ionic species movement in the electrolyte. Furthermore, the presence of less stable SnO and SnO_2, as shown by XPS analysis is ascribed to the incorporation of Sn^{2+} and Sn^{4+} in the destabilised Al_2O_3 passive layer and at the same time modified the chemical characteristic of the film. A strong correlation between the current capacity and changes in oxidation states of Sn element within the passive film was established

ACKNOWLEDGMENT

The authors wish to thank to Government of Malaysia for the financial support to carry out this project. The authors also would like to thank STRIDE for providing the facilities to conduct EIS experiments and gratefully acknowledge Prof. Dr. Mohd Ambar Yarmo from UKM for helping in analysing the XPS spectra.

REFERENCES

Anshuman, S., Chuan, Z., Austin, C., Ray, K. & Dane, M. (2011). Ab initio and thermodynamic modelling of alloying effects on activity of sacrificial aluminium anodes. *Corr. Sci.*, **53:** 1724-1731

Aragón, F.H., Gonzalez, I., José. A. H. Coaquira, J.A.H, Hidalgo, P., Brito, H.F., Ardisson, J.D., Waldemar A. A. Macedo, W.A.A. & Paulo C. Morais, P.C. (2015). Structural and surface study of Praseodymium-Doped SnO_2 nanoparticles prepared by the polymeric precursor method. *J. Phys. Chem. C.*, **119:** 8711-8717

Arranz, A. & Palacio, C. (1996). Characterisation of the surface and interface species formed during the oxidation of aluminum. *Surf. Sci.*, **355:** 203-213.

ASTM (American Society for Testing and Materials) (2011). *ASTM E34: Test Methods for Chemical Analysis of Aluminum and Aluminum-Base Alloys*. American Society for Testing and Materials (ASTM). Philadelphia, USA.

ASTM (American Society for Testing and Materials) (2018). *ASTM G97: Standard Test Method for Laboratory Evaluation of Magnesium Sacrificial Anode Test Specimens for Underground Applications*. American Society for Testing and Materials (ASTM). Philadelphia, USA.

Barbucci, A., Cabot, P.L., Bruzzano, G. & Cerisola, G. (1998). Role of intermetallics in the activation of Al-Mg-Zn alloys. *J. Alloys Comp*., **268:** 295-301

Barr, T.L., Seal, S., Wozniak, K. & Klinowski, J. (1997). ESCA studies of the coordination state of aluminium in oxide environments. *J. Chem. Soc., Faraday Trans.,* **93:** 181-186

Barr, T.L., Seal, S., Chen, L.M. & Kao, C.C. (1994). A new interpretation of the binding energies in X-ray photoelectron studies of oxides. *Thin Solid Films.,* **253:** 277-284

Batzill, M. & Diebold, U. (2005). The surface and materials science of tin oxide. *Prog. Surf. Sci.,* **79:** 47-154

Bessone, J.B., Flamini, D.O. & Saidman, S.B. (2005). Comprehensive model for the activation mechanism of Al–Zn alloys produced by Indium. *Corr. Sci.,* **47:** 95-105

Bluhm, H. (2011). X-ray photoelectron spectroscopy (XPS) for in situ characterization of thin film growth. Woodhead Publishing Series in Electronic and Optical Materials,

Breslin, C.B., Friery, L.P. & Carroll, W.M. (1993). Influence of Impurity Elements on Electrochemical Activity of Aluminum Activated by Indium, *Corr.,* **49**:895-902.

Briggs, D. & Seah, M.P. (1996). *Practical surface analysis, Auger & X-ray Photoelectron Spectroscopy, Vol. 1, 2nd Ed.* Wiley, New York

Bruzzone, G., Barbucci, A. &. Cerisola, G. (1997). Effect of intermetallic compounds on the activation of aluminium anodes. *J. Alloys Comp*., **247:**210-2 16

C. Wöll. (2007). The chemistry and physics of zinc oxide surfaces. *Prog. in Surf. Sci*., **82:** 55-120

Chastain, J. (1992). *Handbook of X-Ray Photoelectron Spectroscopy*. Perkin Elmer Corp., Minnesota:.

Din Yati, M.S., Derman, M.N., Isa, M.C., Ahmad, M.Y., Yusoff, N.H.N., Muhammad, M.M. & Nain, H. (2014). The effect of metallic oxide deposition on the electrochemical behaviour of Al-Zn-Mg-Sn alloy in natural tropical seawater. *IOP Con. Ser.: Mat. Sci. Eng*., **60**: 012051

Duchoslav, J., Steinberger, R., Arndt, M. & Stifter, D. (2014). XPS study of zinc hydroxide as a potential corrosion product of zinc: Rapid X-ray induced conversion into zinc oxide. *Corr. Sci.,* **82:** 356-361

El Shayeb, H. A., Abd El Wahab, F. M. & Zein El Abedin, S. (2001). Electrochemical behaviour of Al, Al–Sn, Al–Zn and Al–Zn–Sn alloys in chloride solutions containing stannous ions. *Corr. Sci*., **43:** 655-669

Evertsson, J., Bertram, F., Zhang, F., Rullik, L. & Lundgren, E. (2015). The thickness of native oxides on aluminum alloys and single crystals. *App. Sur. Sci*., **349**: 826-832

Ferdian, D., Pratesa, Y., Togina, I. & Adelia, I. (2017). Development of Al-Zn-Cu alloy for low voltage aluminum sacrificial anode. *Proc. Eng.,* **184:** 418-422

Flamini, D.O. & Saidman, S.B. (2012). Electrochemical behaviour of Al–Zn–Ga and Al–In–Ga alloys in chloride media. *Mat. Chem. Phy.,* **136:** 103-111

Forget, L., Wilwers, F., Delhalle, J. & Mekhalif, Z. (2003). Surface modification of aluminum by n-pentanephosphonic acid: XPS and electrochemical evaluation. *App. Surf. Sci*., **205:** 44-55

Greczynski, G. & Hultman, L. (2020). X-ray photoelectron spectroscopy: Towards reliable binding energy referencing. *Prog. Mat. Sci.,* **107:** 1-46.

Gredelj, S., Andrea, A.R., Kumar, S. & Cavallaro, G.P. (2001). Characterization of aluminium surfaces with and without plasma nitriding by X-ray photoelectron spectroscopy. *App. Sur. Sci.,* **174:** 240-250

Gredelja, S., Gersona, A.R., Kumara, S. & McIntyreb, N.S. (2002). Plasma nitriding and in situ characterisation of aluminium. *App. Sur. Sci*., **199:** 234-247.

Gudić, S., Radošević, J., Krpan-Lisica, D. & Kliškić, M. (2001). Anodic film growth on aluminium and Al–Sn alloys in borate buffer solutions. *Elec. Acta.,* **46:** 2515-2526

Gudić, S., Radošević, J., Smoljko, I. & Kliškić, M. (2005). Cathodic breakdown of anodic oxide film on Al and Al–Sn alloys in NaCl solution. *Elec. Acta.,* **50:** 5624-5632

Gudić, S., Smoljko, I. & Kliškić, M. (2010). Electrochemical behaviour of aluminium alloys containing Indium and Tin in NaCl solution. *Mat. Chem. Phy*., **121**: 561-566

Hung, F.Y., Lin, H.M., Chen, P.S., Lui, T.S. & Chen, L.H. (2006). A study of the thin film on the surface of Sn-3.5Ag/Sn-3.5Ag-2.0Cu lead-free alloy. *J. Alloys Comp*., **415:** 85-92

Isa, M.C., Ahmad, M.Y., Daud A.R. & Daud, M. (2010). The effect of Sn on the impedance behaviour of Al-Zn alloys in natural chloride solution. *Key Eng. Mat*., **442**: 322-329

Isa, M.C., Daud, A.R., Ahmad, M.Y., Daud, M., Shamsudin, S.R., Hassanuddin, N., Din Yati, M.S. & Muhammad, M.M. (2012). An electrochemical impedance spectroscopy study of Al-Zn and Al-Zn-Sn alloys in tropical seawater. *Key Eng. Mat*., **510-511**: 284-292

Jun-guang, H., Jiu-ba, W., Xu-dong, L., Guo-wei, W. & Chun-hua, X. (2011). Influence of Ga and Bi on electrochemical performance of Al-Zn-Sn sacrificial anodes. *Tran. Nonferro. Met. Soc. China.,* **21:** 1580-1586

Kamarudin, S.R.M., Daud, M., Sattar, S. & Daud, A.R. (2010). A study of Al-5.5Zn-1.5Sn alloy sacrificial anode cathodic protection requirements for structure used in seawater. *AIP Conf. Proc.,* **1202:** 181

Khireche, S., Boughrara, D., Kadri, A., Hamadau, L. & Benbrahim, N. (2014). Corrosion mechanism of Al, Al-Zn and Al-Zn-Sn alloys in 3% NaCl Solution. *Corr. Sci*., **87:**504-516.

Kudo, M., Ishijima, A. & Morohashi, T. (1996). Ion induced alteration at Pb-Sn alloy surface investigated by Auger electron spectroscopy and X-ray photoelectron spectroscopy. *App. Surf. Sci.,* **100-101:** 134-137

Kwoka, M., Ottaviano, L., Passacantando, M., Santucci, S. & Szuber, J. (2006). XPS depth profiling studies of L-CVD SnO_2 thin films. *App. Surf. Sci.,* **252:** 7730–7733

Kwoka, M., Ottaviano, L., Passacantando, M., Santucci, S., Czempik, G. & J. Szuber. (2005). XPS study of the surface chemistry of L-CVD SnO_2 thin films after oxidation. *Thin Solid Films.,* **490:** 36-42

Kyung-Keun, L. & Kwang-Bum, K. (2001). Electrochemical impedance characteristics of pure Al and Al–Sn alloys in NaOH solution. *Corr. Sci*., **43:** 561-575

Lei Wang, L., Tadashi Shinohara, T. & Bo-Ping, Z. (2010). XPS study of the surface chemistry on AZ31 and AZ91 magnesium alloys in dilute NaCl solution. *App. Surf. Sci.,* **256:** 5807-5812

Lenar, J. (2000). *Laboratory Evaluations in Anode Manufacturing*. Corrosion paper 00679, NACE, Houston, Texas, USA

Lin, J.C. &. Shih, H.C. (1987). Improvement of the current efficiency of an AlZnIn anode by heat□ treatment. *J. Elec. Soc.,* **134:** 817-822

Mathiyarasu, J., Nehru, L.C., Subramaniam, P., Palaniswamy, N. & Rengaswamy, N.S. (2001). Synergistic interaction of Indium and Gallium in the activation of aluminium alloy in aqueous chloride solution. *Anti-Corr Meth. & Mat*., **48:** 324-329

Moffitt, C.E., Wieliczka, D.M. & Yasuda, H.K. (2001). An XPS study of the elemental enrichment on aluminum alloy surfaces from chemical cleaning. *Sur. Coat. Tech.,* **137:** 188-196

Mohammad, R.S. & Ahmad, K. (2011). Optimization of Manganese and Magnesium contents in As-cast Aluminum-Zinc-Indium Alloy as sacrificial anode. *J. Mat. Sci. Tech.,* **27:** 785-792

Munoz, A.G., Saidman, S.B. & Bessone, J.B. (2002). Corrosion of an Al–Zn–In alloy in chloride media. *Corr. Sci*., **44**: 2171–2182

Natishan, P.M., Yu, S.Y. O'Grady, W.E. & Ramaker, D.E. (2002). X-ray absorption near edge structure and X-ray photoelectron spectroscopy studies of chloride in passive oxide films. *Elec. Acta.,* **47:** 3131-3136

Neudachina, V.S., Shatalova, T.B., Shtanov, V.I., Yashina, L.V., Zyubina, T.S., Tamm, M.E. & Kobeleva, S.P. (2005). XPS study of SnTe(1 0 0) oxidation by molecular oxygen. *Surf. Sci.,* **584:** 77-82.

Neveu, B., Lallemand, L., Poupon, G. & Mekhalif, Z. (2006). Electrodeposition of Pb-free Sn alloys in pulsed current. *App. Surf. Sci.,* **252**: 3561-3573

Ramya, S., Krishna, N.G.D. & Mudali, K.U. (2018). In-situ Raman and X-ray photoelectron spectroscopic studies on the pitting corrosion of modified 9Cr-1Mo steel in neutral chloride solution. *App. Surf. Sci.,* **428**: 1106-1118

Salinas, D.R. & Bessone, J.B. (1991). Electrochemical behavior of Al-5%Zn-0.1%Sn sacrificial anode in aggressive media: Influence of its alloying elements and the solidification structure. *Corr.,* **47:** 665-674

Salinas, D.R., Garcia, S.G. & Bessone, J.B. (1999). Influence of alloying elements and microstructure on aluminium sacrificial anode performance: case of Al–Zn. *J. App. Elec.,* **29:** 1063-1071

Šekularac, G. & Milosevic, I. (2018). Corrosion of aluminium alloy AlSi7Mg0.3 in artificial sea water with added sodium sulphide. *Corr. Sci,*. **144**: 54-73

Senel, E. & Nisancioglu, K. (2014). Anodic activation of aluminium containing small amounts of Gallium and Tin. *Corr. Sci*., **88:** 280-290.

Shibli, S. M. A. & George, S. (2007). Electrochemical impedance spectroscopic analysis of activation of Al–Zn alloy sacrificial anode by RuO_2 catalytic coating. *App. Surf. Sci*., **253**: 7510-7515

Sistiaga, M., Cuesta, A., Pierna, A.R. & Gutiérrez, C. (1998). Characterization by electrolyte electroreflectance and X-ray photoelectron spectroscopy of amorphous Ni59Nb40Pt1-xSnx alloys and their activation by HF solutions. *Surf. Sci.,* **410**: 312-320

Szuber, J., Czempik, G., Larciprete, R., Koziej, D. & Adamowicz, B. (2001). XPS study of the L-CVD deposited SnO. *Thin Solid Films.,* **391:** 198-203

Talavera, M.A., Valdez, S., Juarez-Islas, J.A., Mena, B. & Genesca, J. (2002). EIS testing of new aluminium sacrificial anodes. *J. Appl. Elec*., **32:** 897-903

Tamada, A. & Tamura, Y. **(1993).** The electrochemical characteristics of aluminum galvanic anodes in an Arctic seawater. *Corr. Sci.*, **34:** 261-277

Tardio, S. & Cumpson, P.J. (2018). Practical estimation of XPS binding energies using widely available quantum chemistry software. *Surf. Interface Anal.,* **50:**5–12.

Uhart, A., Ledeuil, J.B., Gonbeau, D., Dupin, J.C. & Esteban, J. (2016). An Auger and XPS survey of cerium active corrosion protection for AA2024-T3 aluminum alloy. *App. Surf. Sci*., **390:** 751-759.

Venugopal, A. & Raja, V.S. (1997). AC impedance study on the activation mechanism of aluminium by indium and zinc in 3.5% NaCl medium. *Corr. Sci*., **39:** 2053-2065

Werrett, C.R., Bhattacharya, A.K. & Pyke, D.R. (1996). The validity of Cls charge referencing in the XPS of oxidised Al-Si alloys. *App. Surf. Sci*., **103:** 403-407

Winiarski, J., Tylus, W. & Szczygieł, B. (2016). EIS and XPS investigations on the corrosion mechanism of ternary Zn–Co–Mo alloy coatings in NaCl solution. *App. Surf. Sci*., **364**: 455-466

Wu, Q., Song, J., Kang, J., Dong, Q.F., Wu, S.T. & Sun, T.S. (2007). Nano-particle thin films of tin oxides. *Mat. Lett*., **61:** 3679-3684

Xingwen, Y. & Guoqiang, L. (2004). XPS study of cerium conversion coating on the anodized 2024 aluminum alloy. *J. Alloys & Comp*., **364:** 193-198

Xiong, W., Qi, G.T., Guo, X.P. & Lu, Z.L. (2011). Anodic dissolution of Al sacrificial anodes in NaCl solution containing Ce. *Corr. Sci*., **53:** 1298-1303.

Zähr, J., Oswald, S., Türpe, M., Ullrich, H.J. & Füssel, U. (2012). Characterisation of oxide and hydroxide layers on technical aluminum materials using XPS. *Vacuum.,* **86:** 1216-1219

Zhang, S., Zhang, Y. & Wang, H. (2009). Effect of oxide thickness of solder powders on the coalescence of SnAgCu lead-free solder pastes. *J. Alloys & Comp*., **487**: 682-686.

OPTIMISATION OF COLOURIMETRIC APTASENSOR FOR DETERMINATION OF METHYLPHOSPHONIC ACID USING RESPONSE SURFACE METHODOLOGY

Fellyzra Elvya Pojol[1], Soleha Mohamat Yusuff[2], Keat Khim Ong[2,3*], Jahwarhar Izuan Abdul Rashid[3], Nor Laili- Azua Jamari[3], Siti Aminah Mohd Noor[3], Noor Azilah Mohd Kasim[2,3], Norhana Abdul Halim[3], Wan Md Zin Wan Yunus[4], Victor Feizal Knight[2] & Chin ChuangTeoh[5]

[1]Department of Defence Science, Faculty of Defence Science and Technology, National Defence University of Malaysia (UPNM), Malaysia
[2]Centre for Chemical Defence, National Defence University of Malaysia (UPNM), Malaysia
[3]Department of Chemistry and Biology, Centre for Defence Foundation Studies, National Defence University of Malaysia (UPNM), Malaysia
[4]Centre for Tropicalisation, National Defence University of Malaysia (UPNM), Malaysia
[5]Engineering Research Center, Malaysian Agricultural Research and Development Institute (MARDI), Malaysia

[*]Email: ongkhim@upnm.edu.my

ABSTRACT

*Methylphosphonic acid (MPA) is a chemical marker of organophosphorus chemical warfare agents (CWAs) and thus, the detection of MPA is important to examine the exposure of CWAs to humans and the environment. In this study, citrate capped gold nanoparticles (cit-AuNPs) and DNA aptamer were used to detect MPA based on colourimetry. In addition, image processing was utilised to convert colour images of the mixture into red, green and blue (RGB) values in order to improve the accuracy of detection. The independent variables (concentration of cit-AuNPs, concentration of DNA aptamer and incubation period) influencing the detection were optimised using faced-centred central composite design (FCCD) with the response surface methodology (RSM). Analysis of variance (ANOVA) showed that the concentration of cit-AuNPs and incubation period significantly affected the detection. From the analysis, a good correlation between the experimental and predicted responses was found as indicted by high R^2 value (*95.48%*). The optimum conditions of the detection were suggested as the following conditions: concentration of cit-AuNPs of 155 nM, concentration of DNA aptamer of 1 μM and incubation period of 24.85 min.*

Keywords: *Aptamer; colourimetry; image processing; methylphosphonic acid; response surface methodology (RSM).*

1. INTRODUCTION

The development of sensitive, rapid and cost-effective sensors have been vigorously reported for detection of nerve agents (NAs) (Vernekar *et al*., 2016), as NAs can cause paralysis and mortality (Sathe *et al*., 2018). However, NAs can be hydrolysed to methylphosphonic acid (MPA) in the environment and human body, and its half-life can reach up to 18 years (Mill & Gould, 1979; Katagi *et al*., 1997). Therefore, the detection of MPA is very important, as MPA is one of the chemical markers of exposure to organophosphorus chemical warfare (Savel'eva *et al*., 2001).

Colourimetry is an attractive method because of its simplicity, short detection time and low cost (Gan *et al*., 2020). Gold nanoparticles (AuNPs) draw great attentions from scientists due to its superior advantages (Gan *et al*., 2020). Numerous studies used AuNPs in colourimetric detection of organophosphate pesticides (Wang *et al*., 2016; Bala *et al*., 2017). For example, the detection of omethoate in water samples was based on the coordination between omethoate molecules (Wang *et al.*, 2016), while the determination of malathion was based on the interaction between peptides of peptide-modified AuNPs and malathion (Bala *et al.*, 2017). Among different recognition molecules, aptamer is an extremely ideal recognition molecule, which can lead to the change in colourimetric signals by specific binding to organophosphorus pesticides (Xu *et al.*, 2019). Aptamer is a synthetic single-stranded oligonucleotide or peptide that can bind specifically to toxins, cells, small compounds, etc. (Hassani *et al.*, 2018). In addition, aptamers also have numerous beneficial features, including high sensitivity and stability in the complex environments (Florea *et al.*, 2013; Khavani *et al.*, 2019).

In this study, a low cost, rapid and sensitive method using thiolated DNA aptamer-citrate capped AuNPs (DNA-cit-AuNPs) as the aptasensor were developed to detect MPA. In order to obtain the highest detection sensitivity of MPA, the optimum experimental conditions need to be determined. Hence, three parameters (concentration of cit-AuNPs, concentration of DNA aptamer, and incubation period) were optimised using the response surface methodology (RSM) based faced-centred central composite design (FCCD) for MPA detection.

2. EXPERIMENTAL SETUP

2.1 Synthesis of Citrate Capped Gold Nanoparticles

Citrate capped gold nanoparticles (Cit-AuNPs) were synthesised using the following steps. A colourless solution of tri-sodium citrate dihydrate ($C_6H_5Na_3O_7.2H_2O$) (Na-cit) (100 mL, 0.34 mM) was firstly boiled for 15 min prior to the addition of gold (III) chloride trihydrate ($HAuCl_4.3H_2O$) solution (0.5 mL, 21.57 mM). Then, the mixture was continuously stirred at 1,500 rpm for 20 min at 100 °C before cooling down to room temperature. The solution was then centrifuged for 30 min at 12,000 rpm to remove excess sodium citrate. The supernatant layer was removed and kept at 4 °C for further use.

2.2 Parameters Investigated

In this study, three parameters were studied to optimise the detection of MPA: (A) concentration of cit-AuNPs, (B) concentration of DNA aptamer and (C) incubation period. The Minitab 16 software was used to design the experiments and analyse the results obtained for this study. The experimental range and levels of independent variables investigated are presented in Table 1.

Table 1: Experimental range and levels of independent variables.

Independent variables	Symbol	Unit	Low (-1)	Mid (0)	High (+1)
Concentration of cit-AuNPs	A	nM	127	165	204
Concentration of DNA aptamer	B	µM	1	5	9
Incubation period	C	min	0	30	60

2.3 Detection of Methylphosphonic Acid

The prepared DNA-cit-AuNPs was used as a chemical sensor to detect 15 mM of MPA based on the conditions presented in Table 1. Briefly, cit-AuNPs was added to 1 µL of DNA aptamer and incubated at 37 °C for a certain incubation period. After the incubation, MPA solution was added to the mixture (total volume of 1,000 µL) and immediately, the mixture was shaken gently. After detection, the colour of the mixture was captured using a smartphone at a fixed distance of 9 cm. The conditions for capturing the images were kept constant for all experimental runs. The images were then analysed using the imageJ software to obtain digital values of red (R), green (G) and blue (B). An average RGB was used to calculate the ΔRGB value, which determines the magnitude of the vector between the blank RGB value (R_0, G_0, B_0) and the RGB value of sample (R_1, G_1, B_1) (Murdock *et al*., 2013). The similar procedure was applied for detection in the absence of MPA, which served as a blank.

3. RESULTS AND DISCUSSION

3.1 Detection of Methylphoshonic Acid

The experimental conditions of (A) concentration of cit-AuNPs, (B) concentration of DNA aptamer and (C) incubation period were optimised using RSM based FCCD in order to achieve a stable sensing performance. The experimental design matrix, as well as experimental and predicted values of ΔRGB are presented in Table 2. The results revealed that there was a good agreement between experimental and predicted values of ΔRGB. Furthermore, ΔRGB has a variation from 5 to 16, which shows the significance of optimisation.

Table 2: Experimental and predicted values of ΔRGB.

Experimental run	Concentration of cit-AuNPs (nM)	Concentration of DNA aptamer (μM)	Incubation period (min)	ΔRGB	
				Experimental	Predicted
1	165	5	30	16	16
2	127	9	0	12	12
3	165	5	30	16	16
4	127	1	0	12	12
5	204	9	0	6	6
6	165	5	30	16	16
7	127	9	60	10	10
8	165	5	30	16	16
9	165	5	60	13	14
10	204	9	60	5	5
11	165	5	30	16	16
12	165	9	30	13	14
13	127	5	30	14	14
14	165	5	30	16	16
15	127	1	60	10	10
16	204	1	0	6	6
17	165	5	0	15	15
18	204	5	30	9	9
19	165	1	30	14	14
20	204	1	60	6	6

3.2 Analysis of Variance (ANOVA)

The ANOVA results of the full regression model for ΔRGB are presented in Table 3. The input variables of B, AB and BC were eliminated due to their insignificant effects on the detection as their *p*-values were larger than 0.05, and thus, ANOVA was reanalysed. The results of ANOVA for the reduced model is shown in Table 4. The results revealed that the *F*-value (59.11) of the reduced regression model was large, the *p*-value (0.000) of the reduced regression model was lower than 0.05 and the *p*-value (0.449) of the lack of fit was larger than 0.05. Hence, these ANOVA results proved that the reduced regression model was highly significant and well fitted to the experimental data as shown by the large *F*-value of the model. The accuracy of the reduced regression model was indicated by the low *p*-value (<0.05) of the regression and non-significant value of the lack of fit. In addition, the coefficients of determination, R^2, $R^2_{predicted}$ and $R^2_{adjusted}$ play important roles to evaluate the goodness of the developed reduced regression model. The calculated R^2, $R^2_{predicted}$ and $R^2_{adjusted}$ were 0.9548, 0.9261 and 0.9386 respectively, demonstrating that the developed reduced regression model is able to explain 95.48, 92.61 and 93.86% of the total variation in the variables investigated in the detection of MPA. The coefficient determination values were high and close to 1. The high value of R^2 is satisfactory to describe the relationship between the response and variables, while a high value of $R^2_{predicted}$ indicates that the responses could be predicted accurately using the developed reduced regression model.

Only the linear terms of concentration of cit-AuNPs (A) and incubation period (C), as well as quadratic terms of A^2 and C^2 significantly affected MPA detection as shown by $p < 0.05$. The *F*-test can be used to evaluate the statistical significance of all the terms in the regression model. The results also revealed that concentration of cit-AuNPs (A) exhibited the largest effect on ΔRGB as indicated by the largest of *F*-value of 72.41, followed by incubation period (C) with *F*-value of 5.23.

Table 3: Analysis of variance (ANOVA) of ΔRGB (*Full model*).

Source	Degree of freedom	Adj Sum of squares	Adj Mean squares	*F*-Value	*p*-Value
Model	9	287.529	31.9477	224.80	0.000
Linear	3	72.888	24.2959	170.96	0.000
A	1	67.600	67.6000	475.68	0.000
B	1	0.401	0.4013	2.82	0.124
C	1	4.886	4.8863	34.38	0.000
Square	3	211.180	70.3935	495.33	0.000
AA	1	43.258	43.2583	304.39	0.000
BB	1	11.000	11.0000	77.40	0.000
CC	1	6.188	6.1875	43.54	0.000
2-Way Interaction	3	1.379	0.4596	3.23	0.069
AB	1	0.123	0.1231	0.87	0.374
AC	1	1.131	1.1308	7.96	0.018
BC	1	0.125	0.1250	0.88	0.370
Error	10	1.421	0.1421		
Lack-of-Fit	5	1.421	0.2842	*	*
Pure Error	5	0.000	0.0000		
Total	19				

$R^2 = 99.51\%$, $R^2_{predicted} = 96.53\%$, $R^2_{adjusted} = 99.07\%$

Table 4: Analysis of variance (ANOVA) of ΔRGB (*reduced model*).

Source	Degree of freedom	Adj Sum of squares	Adj Mean squares	*F*-Value	*p*-Value
Model	5	275.881	55.176	59.11	0.000
Linear	2	72.486	36.243	38.82	0.000
A	1	67.600	67.600	72.41	0.000
C	1	4.886	4.886	5.23	0.038
Square	2	200.180	100.090	107.22	0.000
AA	1	71.173	71.173	76.24	0.000
CC	1	16.200	16.200	17.35	0.001
2-Way Interaction	1	1.131	1.131	1.21	0.290
AC	1	1.131	1.131	1.21	0.290
Error	14	13.069	0.934		
Lack-of-Fit	9	2.694	0.898	0.95	0.449
Pure Error	5	10.375	0.943		
Total	19				

$R^2 = 95.48\%$, $R^2_{predicted} = 92.61\%$, $R^2_{adjusted} = 93.86\%$

3.3 Development of Second-Order Polynomial Model for MPA Detection

Each independent variable (concentration of cit-AuNPs, concentration of DNA-aptamer and incubation period) was investigated at three levels: -1, 0 and +1, as shown in Table 1. The behaviour of the colourimetric detection of MPA can be explained using an empirical second-order polynomial as expressed in the following equation:

$$\boldsymbol{Y = \beta_o + \sum \beta_i X_i + \sum \beta_{ij} X_i X_j + \sum \beta_{ii} X_i^2} \qquad (1)$$

where Y is the predicted response (ΔRGB), β_o is the intercept, β_i is the effect of the linear terms, β_{ii} is the effect of the quadratic terms, β_{ij} is the effect of the interaction terms, and X_i are coded values of the corresponding ith factors.

As mentioned in Section 3.2, B, AB and BC were eliminated due to their insignificant effects on the detection. Thus, the constant and coefficients were recalculated using the Minitab software, with the results presented in Table 5. These constant and coefficients were substituted into Equation 1 to develop Equation 2.

Table 5: Estimated regression coefficients for ΔRGB (*reduced regression model*).

Term	Coef
Constant	-60.41
A	0.976
C	0.0728
AA	-0.003
CC	-0.0025
AC	0.000326

$$\Delta RGB = -60.41 + 0.976A + 0.0728C -0.003A^2 - 0.0025C^2 + 0.000326AC \qquad (2)$$

Equation 2 is the reduced regression model of the empirical relationship between the input variables of A and C with ΔRGB, where ΔRGB represents the predicted response, while A, B and C are the coded values of the investigated independent variables. The input variable of concentration of DNA aptamer (B) was eliminated due to its insignificant effect on the detection as mentioned in the ANOVA results. The positive and negative signs of the coefficients indicate positive and negative effects on the response, respectively. The equation shows that only A, C and AC showed positive effects on the response.

3.4 Three-Dimensional (3D) Response Surface Plot

The interaction between independent variables (AC) with ΔRGB was evaluated using a three-dimensional (3D) response surface plot (Figure 1). Figure 1 portrays the increments of ΔRGB as the concentration of cit-AuNPs decreases from 204 to 127 nM and the incubation period decreases from 60 to 0 min, while other parameters are constant.

The 3D surface plot demonstrates that increasing the incubation period, resulted in decrease of the ΔRGB values. The binding between cit-AuNPs and DNA aptamer occurred rapidly near the mid-level of incubation period as the highest value of ΔRGB was obtained at that level. Fast detection using AuNP aptamer-based was also reported by Li *et al.* (2019). Besides that, the optimum ΔRGB can be achieved using cit-AuNPs with concentration around mid-level. Further analysis on the optimum conditions for the MPA detection was conducted in the following sub-section using the response surface optimiser.

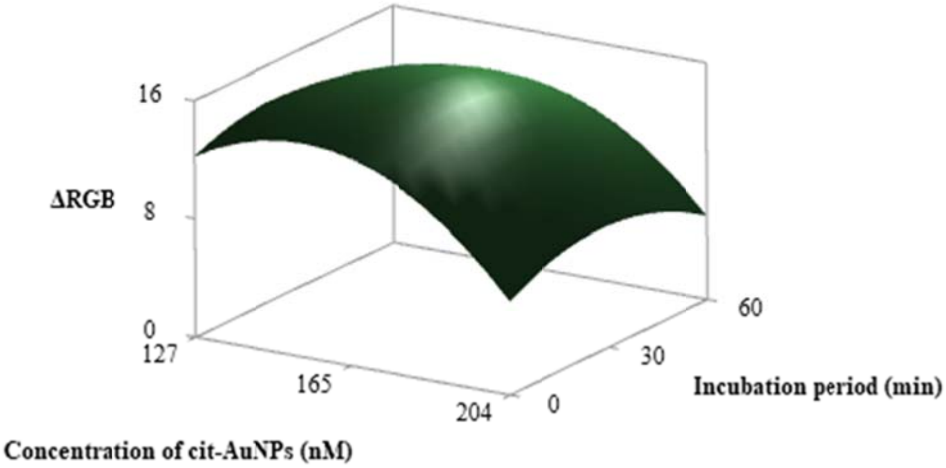


Figure 1: 3D surface plot of the effect of concentration of cit-AuNPs and incubation period (AC) on ΔRGB.

3.5 Optimisation Using Response Surface Optimiser

The response surface optimiser analysis was used to predict the optimum conditions to achieve the optimum ΔRGB values. The optimum conditions to achieve 15.95 of ΔRGB with composite desirability of 0.995 were suggested with the following conditions (Figure 2): concentration of cit-AuNPs of 155 nM, concentration of DNA aptamer of 1 μM and incubation period of 24.85 min.

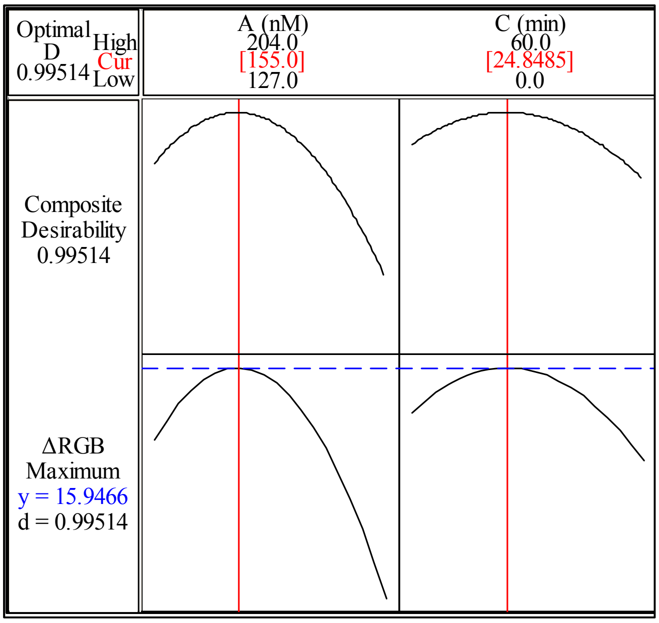


Figure 2: Optimisation plot of ΔRGB.

4. CONCLUSION

The faced-centred central composite (FCCC) design successfully developed a second-order polynomial model to optimise the MPA detection based on ΔRGB values as the response of the colourimetric aptasensor. The ANOVA results showed that the concentration of cit-AuNPs and incubation period were independent variables with high significant effects on ΔRGB. The optimum conditions for the detection of MPA were found to be as the following conditions: concentration of cit-AuNPs of 155 nM, concentration of DNA aptamer of 1 µM and incubation period of 24.85 min. Concentration of cit-AuNPs showed greater impact on ΔRGB followed by incubation period and concentration of DNA aptamer.

ACKNOWLEDGEMENT

The project was financial supported by the Ministry of Higher Education Malaysia through the Fundamental Research Grant Scheme (FRGS/1/2017/STG01/UPNM/02/1) and ChemDef Grant (UPNM/2018/CHEMDEF/ST/2). The authors also would like to acknowledge the National Defence University of Malaysia (UPNM) for providing the research facilities for this study.

REFERENCES

Bala, R., Dhingra, S., Kumar, M., Bansal, K., Mittal, S., Sharma, R.K. & Wangoo, N. (2017). Detection of organophosphorus pesticide – Malathion in environmental samples using peptide and aptamer based nanoprobes. *Chem. Eng. J.,* **311**: 111-116.

Florea, A., Taleat, Z., Cristea, C., Mazloum-Ardakani, M. & Săndulescu, R. (2013). Label free MUC1 aptasensors based on electrodeposition of gold nanoparticles on screen printed electrodes. *Electrochem. Commun.,* **33**: 127-130.

Gan, Y., Liang, T., Hu, Q., Zhong, L., Wang, X., Wan, H., & Wang, P. (2020). In-situ detection of cadmium with aptamer functionalized gold nanoparticles based on smartphone-based colorimetric system. *Talanta,* **208**: 120231-120237.

Hassani, S., Akmal, M.R., Salek-Maghsoudi, A., Rahmani, S., Ganjali, M.R., Norouzi, P. & Abdollahi, M. (2018). Novel label-free electrochemical aptasensor for determination of diazinon using gold nanoparticles-modified screen-printed gold electrode. *Biosens. Bioelectron.,* **120**: 122-128.

Katagi, M., Nishikawa, M., Tatsuno, M. & Tsuchihashi, H. (1997). Determination of the main hydrolysis products of organophosphorus nerve agents, methylphosphonic acids, in human serum by indirect photometric detection ion chromatography. *J. Chromatogr. B,* **698**: 81-88.

Khavani, M., Izadyar, M. & Housaindokht, M.R. (2019). Theoretical design and experimental study on the gold nanoparticles based colorimetric aptasensors for detection of neomycin B. *Sensor Actuat. B-Chem.,* **300**: 126947-126955.

Li, D., Wang, S., Wang, L., Zhang, H. & Hu, J. (2019). A simple colorimetric probe based on anti-aggregation of AuNPs for rapid and sensitive detection of malathion in environmental samples. *Anal. Bioanal. Chem.*, **411**: 2645-2652.

Mill, T. & Gould, C.W. (1979). Free-radical oxidation of organic phosphonic acid salts in water using hydrogen peroxide, oxygen and ultraviolet light. *Environ. Sci. Technol.,* **13**: 205-208.

Murdock, R.C., Shen, L., Griffin, D.K., Kelley-Loughnane, N., Papautsky, I. & Hagen, J.A. (2013). Optimization of a paper-based ELISA for a human performance biomarker. *Anal. Chem.,* **85**: 11634-11642.

Sathe, M., Ghorpade, R., Merwyn, S., Agarwal, G.S. & Kaushik, M.P. (2018). Direct hapten-linked competitive inhibition enzyme-linked immunosorbent assay (CIELISA) for the detection of O-pinacolyl methylphosphonic acid. *Analyst,* **137**: 406-413.

Savel'eva, E.I., Radilov, A.S., Kuznetsova, T.A., & Volynets, N.F. (2001). Determination of methylphosphonic acid and its esters as chemical markers of organophosphorus chemical warfare agents. *Russ. J. Appl. Chem.*, **74**: 1722–1727.

Vernekar, A.A., Das, T., & Mugesh, G. (2016). Vacancy-engineered nanoceria: Enzyme mimetic hotspots for the degradation of nerve agents. *Angew. Chem. Int. Ed. Engl.,* **55**: 1412-1416.

Wang, P., Wan, Y., Ali, A., Deng, S., Su, Y., Fan, C. & Yang, S. (2016). Aptamer-wrapped gold nanoparticles for the colorimetric detection of omethoate. *Sci. China Chem.,* **59**: 237-242.

Xu, M., Obodo, D. & Yadavalli, V.K. (2019). The design, fabrication, and applications of flexible biosensing devices. *Biosens. Bioelectron.,* **124-125**: 96-114.

EFFECT OF CITRATE CAPPED GOLD NANOPARTICLES VOLUME ON ACEPHATE DETECTION USING COLORIMETRIC ASSISTED IMAGE PROCESSING TECHNIQUE

Mohd Junaedy Osman[1], Wan Md Zin Wan Yunus[2], Ong Keat Khim[*1,3], Jahwarhar Izuan Abdul Rashid[1],
Buong Woei Chieng[3] & Chin Chuang Teoh[4]

[1]Centre for Defence Foundation Studies, National Defence University of Malaysia (NDUM), Malaysia
[2]Centre for Tropicalisation, National Defence University of Malaysia (NDUM), Malaysia
[3]Centre for Chemical Defence, National Defence University of Malaysia (NDUM), Malaysia
[4]Engineering Research Centre, MARDI, Headquarter Serdang, 43400 Serdang, Selangor, Malaysia

Email: ongkhim@upnm.edu.my

ABSTRACT

Acephate (Ac) is an organophosphate (OPP) that is widely used as an insecticide. It is highly toxic, thus simple, and rapid detection method of Ac is in urgent demand. Citrate capped gold nanoparticles (Cit-AuNPs) are excellent materials for colorimetric detection of OPPs. The effect of Cit-AuNPs and Ac volume ratio (Cit-AuNPs: Ac) on the colorimetric sensor capability was evaluated for detection of Ac in this study. Cit-AuNPs were synthesised via reduction of chloroauric acid ($HAuCl_4$) using tri-sodium citrate dihydrate and used to detect Ac. Colour images of all complexes formed were captured and digitised using ImageJ software to obtain red values (RVs) as the response of the detection. Analysis using Ultraviolet-Visible (UV-Vis) spectrophotometer and High Resolution-Transition Electron Microscope (HRTEM) were employed to examine the red shift and Cit-AuNPs aggregation, respectively, whereas Fourier Transform Infra-red (FTIR) was used to investigate chemical interaction between Cit-AuNPs and Ac after the detection. The results of this study demonstrated that the lowest volume ratio of Cit-AuNPs: Ac (1:9) was sufficient to produce obvious colour change for the detection of Ac.

Keywords: *Acephate; colorimetric detection; gold nanoparticles; organophosphate; image processing.*

1. INTRODUCTION

Organophosphorus compounds (OPs) are organic compounds containing phosphorus centre bonded to at least one organic group (R) (alkyl or aryl) and double bonded to a sulphur or oxygen atom (Koskela, 2010; Osman *et al.,* 2019). OPs are also main components in insecticides, herbicides, and nerve agents (Ali *et al.,* 2019; Reynoso *et al.,* 2019). OPs are toxic to human, insects, animals and environment (Bartelt-Hunt *et al.,* 2008; Zehani *et al.,* 2015; Jain *et al.,* 2019). This characteristic make OPs as potential chemical warfare agent (CWA) simulants (Yang, 1999; Joshi *et al.,* 2006; Romano *et al.,* 2009; Koskela, 2010).

Acephate (Ac) is an organophosphate (OPP) (a class of OPs) that is widely used as insecticides (Kumar & Upadhay, 2013). OPPs are very toxic due to their abilities to deactivate the activity of acetylcholinesterase (AChE) enzyme through phosphorylation process *(*Čolović *et al., 2013;* Phugare *et al*., 2012; Wei *et al*., 2011). AChE catalyses the hydrolysis of the neurotransmitter acetylcholine (ACh) into choline and acetic acid, which allows the activation of cholinergic neurons. The deactivation of AChE leads to accumulation of ACh in synaptic cleft, which interrupts the nerve impulse transmission and consequently creates 'jam' to the nervous system (Čolović *et al*., 2013). Thus, rapid detection and reliable quantification of Ac is necessary for human health safety.

Trace levels of OPPs are commonly determined by the conventional detection methods which requires high specification instruments including chromatography and mass spectroscopy (Sulaiman *et al.,* 2020) which are expensive and time consuming. Thus, rapid and low-cost analytical methods to detect OPs are highly demanded (Songa & Okonkwo, 2016).

Colorimetric sensors have been commonly utilised for OPPs detection due to their low-cost, easy read-out and rapid visual determination with naked eyes (Rebollar-Pérez *et al.,* 2016). Nanoparticles (NPs) such as citrate capped gold nanoparticles (Cit-AuNPs) (Fahimi-Kashani & Hormozi-Nezhad, 2016) and silver nanoparticles (AgNPs) (Hsu *et al.,* 2017) are often employed as a probe in colorimetric sensor for OPPs detection, exploiting the unique capability of NPs to change the colour of suspension due to aggregation (Nilam *et al.,* 2017). This colour change can be observed by naked eyes even at lower concentrations of the analyte. In addition, Cit-AuNPs are more preferable due to its properties such as distinct colour change and have good stability compared to AgNPs (Yeh *et al.,* 2012). Cit-AuNPs are usually synthesised through reduction of gold chloride using sodium citrate (La Spina *et al.,* 2017).

Several studies such as effects of pH (Mohseni *et al.,* 2017; Rastogi *et al.,* 2017; Li *et al.,* 2018; Rana *et al.,* 2018; Duenchay *et al.,* 2019), ions addition (Rawat *et al.,* 2016; Zhang *et al.,* 2018) and AuNPs to analytes volume ratio (Nietzold & Lisdat, 2012; Yadav *et al.,* 2018; Duenchay *et al.,* 2020) on detection capability of AuNPs as colorimetric sensors were conducted. The results revealed that the volume ratio was an important factor that influenced the detection performance of the colorimetric sensors. Thus, in this work, effect of Cit-AuNPs volume on detection of Ac using Cit-AuNPs as a chemical sensor was studied.

2. MATERIALS AND METHODOLOGY

2.1 Materials

Gold (III) chloride trihydrate ($HAuCl_4.3H_2O$; $\geq$ 49 % of Au content) of ACS grade and acephate ($C_4H_{10}NO_3PS$) of pestanal analytical grade were purchased from Sigma Aldrich, USA. Tri-sodium citrate dihydrate ($C_6H_5Na_3O_7.2H_2O$) (Na-Cit) of ACS grade was procured from Merck KGaA, Germany. All solution preparations and dilutions were performed using MilliQ water with resistivity of 18.2 MΩ.cm.

2.2 Synthesis of Cit-AuNPs

Cit-AuNPs was synthesised based on the reported procedure (Bala *et al.,* 2015) with some modifications. The details of the procedure used was as follows: 100 mL of 0.25 mM of gold (III) chloride solution was heated until boiling at constant stirring of 300 rpm followed by rapid addition of 2 mL of 34 mM Na-Cit. The reaction mixture was then further boiled at same stirring speed for 20 minutes. The produced suspension was labelled as Cit-AuNPs. The suspension were cooled down to room temperature and stored in a dark bottle at 5 °C.

2.3 Colorimetric Detection of Ac Using Cit-AuNPs

The effect of volume ratio of Cit-AuNPs to Ac (final concentration of 80 mM) was investigated to determine the optimum Cit-AuNPs to Ac volume ratio for detection of Ac. Different volume ratios of Cit-

AuNPs : Ac were used as follow: 9:1, 8:2, 7:3, 6:4, 5:5, 4:6, 3:7, 2:8 and 1:9, with a total volume of 1000 µL for each mixture. The colour images of the Cit-AuNPs-Ac complexes formed were recorded using Samsung S7 edge handphone at a fixed distance of 10.5 cm from the complexes formed. The red values (RVs) of all cropped images were digitised using the ImageJ software.

2.4 Characterisation

2.4.1 Laser Test

A laser test was performed using a Logitech laser pointer to confirm the formation of Cit-AuNPs.

2.4.2 Ultraviolet-Visible Spectroscopy

UV-Vis analysis of samples was performed using a Genesys 6 UV-Vis spectrophotometer at a scanning rate of 1 nm/sec from 400 nm to 750 nm. Absorbance versus wavelength was recorded for each sample before and after detection of Ac using Cit-AuNPs.

2.4.3 High Resolution Transmission Electron Microscopy

High-Resolution Transmission Electron Microscopy (HRTEM) analysis of samples was performed using a JEOL JEM 2100F HRTEM. All samples were dropped on the viewing grids and left overnight to dry. The HRTEM image of Cit-AuNPs before and after detection using optimum volume ratio of Cit-AuNPs to Ac were obtained.

2.4.4 Fourier Transform Infra-Red Spectroscopy

Fourier Transform Infra-Red (FTIR) spectra were recorded using a Perkin Elmer Spectrum 100 equipped with Attenuated Total Reflection (ATR) accessory to analyse Ac, synthesised Cit-AuNPs, and Cit-AuNPs-Ac complex formed.

3. RESULTS AND DISCUSSION

Cit-AuNPs were successfully synthesised using Na-Cit as a reducing agent as well as a capping agent to stabilise AuNPs. A simple laser test was also conducted to study the existence of surface plasmon resonance (SPR) for Cit-AuNPs (Pluchery *et al.,* 2013; Pradeep & Ashok 2017). A successful formation of Cit-AuNPs was distinguished by the colour change from yellow (gold (III) chloride trihydrate) to dark red and occurrence of Cit-AuNPs SPR was observed from the laser test (Figure 1) and UV-Vis absorbance spectrum (Figure 2), respectively.

Figure 2 shows the UV-Vis absorption spectrum of Cit-AuNPs. The absorbance of Cit-AuNPs at λ_{max} 526 nm was 0.9902, resulting from SPR phenomenon of Cit-AuNPs. Figure 3 depicts a HRTEM micrograph of synthesised Cit-AuNPs. Clearly, the results showed that Cit-AuNPs particles have spherical shape and were well dispersed in the aqueous solution with the average particle sizes of 16 nm. Well dispersed Cit-AuNPs indicates that there is a strong repulsion force existed between particles due to electrostatic repulsion, steric exclusion, or a hydration layer on the Cit-AuNPs surface (Jazayeri *et al.,* 2016; Jin *et al.,* 2016).

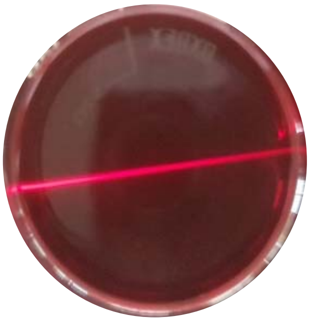


Figure 1: Irradiation of laser beam pass through Cit-AuNPs suspension.

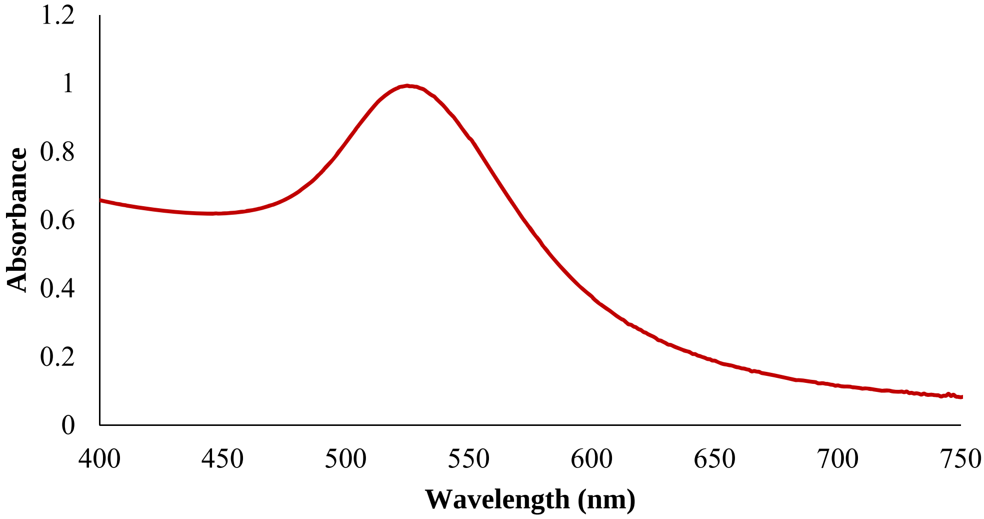


Figure 2: UV-Vis absorption spectrum of Cit-AuNPs suspension.

Table 1 shows the effect of Cit-AuNPs volume on detection of Ac, by using different volume ratios of Cit-AuNPs to Ac. In general, formation of purple to dark purple of complexes were recorded when volume ratios of Cit-AuNPs to Ac used decreased. The cropped images of the complexes were digitised to RGB (red, green and blue) colour values to eliminate false positives and false negatives by observing the results with naked eyes. The RVs was found to be the most dominance compared to green and blue colour values; thus, only RVs was reported in this study. The results also revealed that RVs decreased when intensities of purple colour complexes increased. The darkest purple colour complex with the lowest mean of RVs was obtained when 9:1 of Cit-AuNPs: Ac volume ratio was applied. In contrast, the highest mean of RVs (the lowest purple colour intensity) was achieved when using the lowest volume ratio of Cit-AuNPs: Ac *i.e.* 1:9 and it was sufficient to produce obvious colour change for detection of Ac.

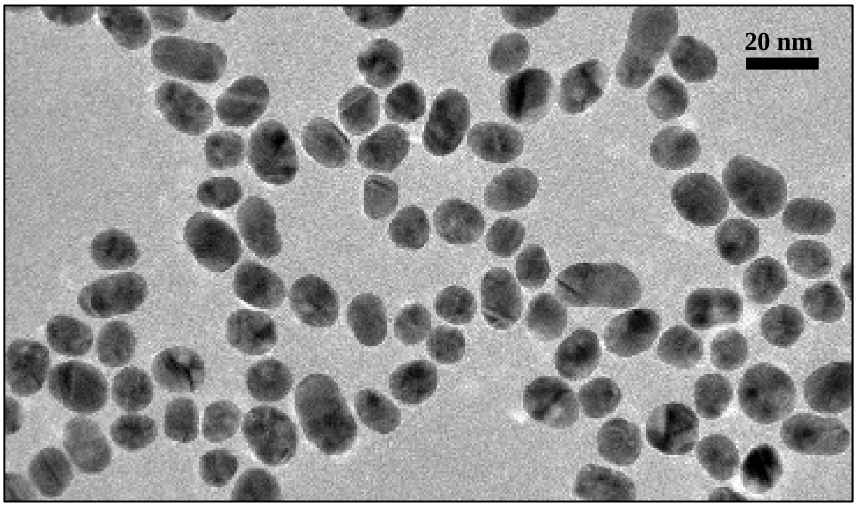


Figure 3: HRTEM micrograph of synthesised Cit-AuNPs suspension at a magnification of 20 nm.

Table 1: Cropped images and RVs of the Cit-AuNPs-Ac complexes formed after detection of Ac using various Cit-AuNPs: Ac volume ratios.

Cit-AuNPs: Ac volume ratio	Blank	9:1	8:2	7:3	6:4	5:5	4:6	3:7	2:8	1:9
Cropped image	●	●	●	●	●	●	●	●	●	●
Mean RVs	150	32	36	39	40	52	52	54	75	99
Standard Deviation of RVs	0.49	0.05	0.24	0.12	0.12	0.04	0.02	0.18	0.31	0.37

Figure 4 shows UV-Vis spectra of all Cit-AuNPs-Ac complexes formed after detection of Ac using various Cit-AuNPs: Ac volume ratios. The characteristic absorbance of AuNPs at 526 nm which correspond to the SPR of Cit-AuNPs. The negatively charged citrate ions, together with the counter-ions in the medium, formed a repulsive electric double layer to stabilise AuNPs (Xu *et al.*, 2011; Lang *et al.*, 2013). However, the existence of Ac reduces the SPR absorbance peak might be due to weakening of the repulsive forces in Cit-AuNPs (Rios-Corripio *et al.*, 2013). Conversely, a new absorption band at a higher wavelength (650 -700 nm) gradually increased due to formation of the aggregates (Xu *et al.*, 2011; Mazurek *et al.*, 2013). According to Xu *et al.* (2011), the aggregation is due to the decrease of the particles' repulsion force that allows them to aggregate. The Cit-AuNPs-Ac complexes (dark purple complex) formation due to the aggregation was further confirmed by HRTEM analysis as shown in Figure 5.

FTIR spectra show the presence of representative functional groups in the Cit-AuNPs-Ac complexes (Figure 6). Broad O-H peaks at 3,350 cm^{-1} were found in all spectrum due to the presence of water that used as a medium of the reaction. The Cit-AuNPs spectrum (Figure 6 a) shows double peaks of antisymmetric, *As* (1,741 cm^{-1} and 1,638 cm^{-1}) and symmetric, *S* COO$^-$ stretching (1,366 cm^{-1}), C-O stretching (1216 cm^{-1}) and COO$^-$ bending (694 cm^{-1}) due to the presence of citrate ions on the surface of AuNPs (Park & Shumaker-Parry, 2014). The FTIR spectrum of Ac (Figure 6 b) shows indicative peaks of amide N-H (3,200 cm^{-1} and 1,666 cm^{-1}), aliphatic C-H (3,000 cm^{-1}), C=O (1,700 cm^{-1}), *As* CH_3 deformation (1,432 cm^{-1}), P=O (1,222 cm^{-1}), *S* CH_3 rocking (around 934 cm^{-1}) and P-S (687 cm^{-1}).

FTIR spectrum of the Cit-AuNPs : Ac complex (Figure 6 c) shows widening of *As* COO$^-$ 1,638 cm^{-1}, a shift of *S* COO$^-$ 1,361 cm^{-1} and reduced transmittance percentage of P=O peak (1,217 cm^{-1}), indicating that Ac has interacted to citrate COO$^-$ site via physical interaction and/or a possible hydrogen bond formation (Zhou *et al.*, 2020). The disappearance of the peak at around 3200 cm^{-1} (N-H stretching) and the reduction of the peak intensity around 1,230 cm^{-1} (P=O and C-N stretching) suggested S_N2 type interaction occurs at phosphorus atom where citrate ion of Cit-AuNPs replaces the amide group and breaks P-N bond in the Ac structure (Dyguda-Kazimierowicz *et al.*, 2014).

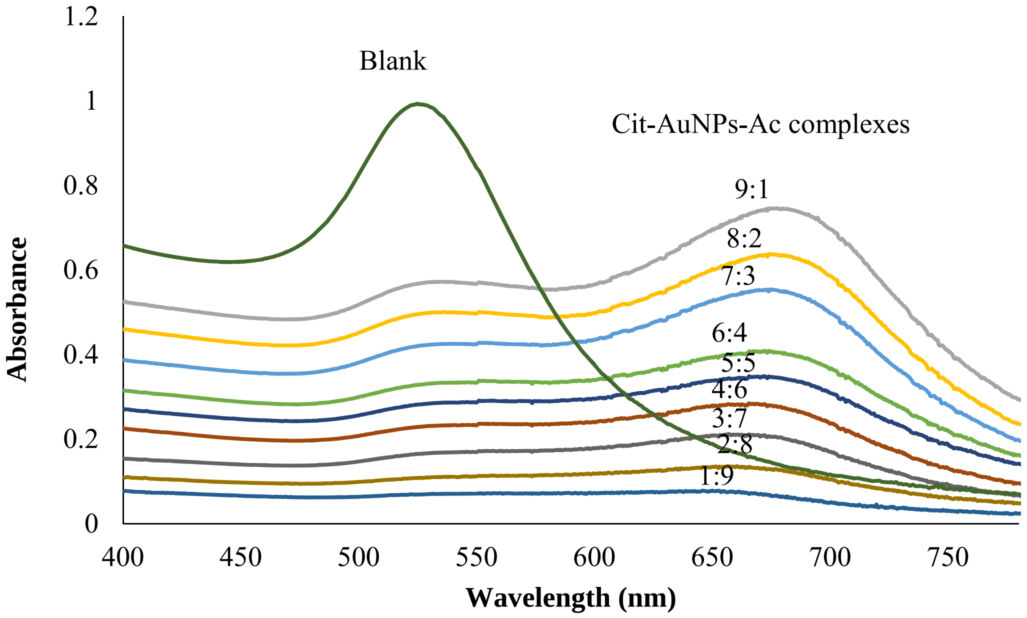


Figure 4: UV-Vis absorption spectra of blank and Cit-AuNPs-Ac complexes formed after detection of Ac using various Cit-AuNPs: Ac volume ratios.

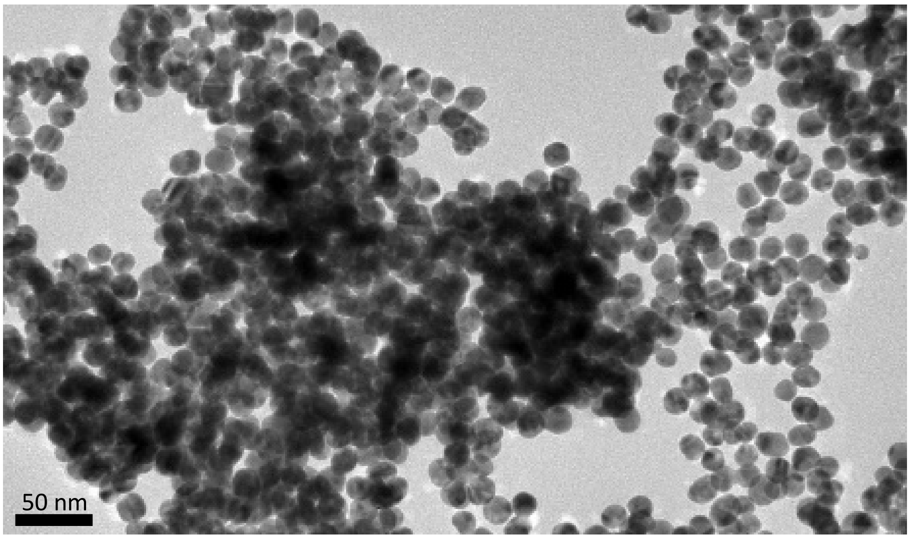


Figure 5: HRTEM micrograph of Cit-AuNPs-Ac complexes formed after detection of Ac using 9:1 Cit-AuNPs: Ac volume ratio at a magnification of 50 nm.

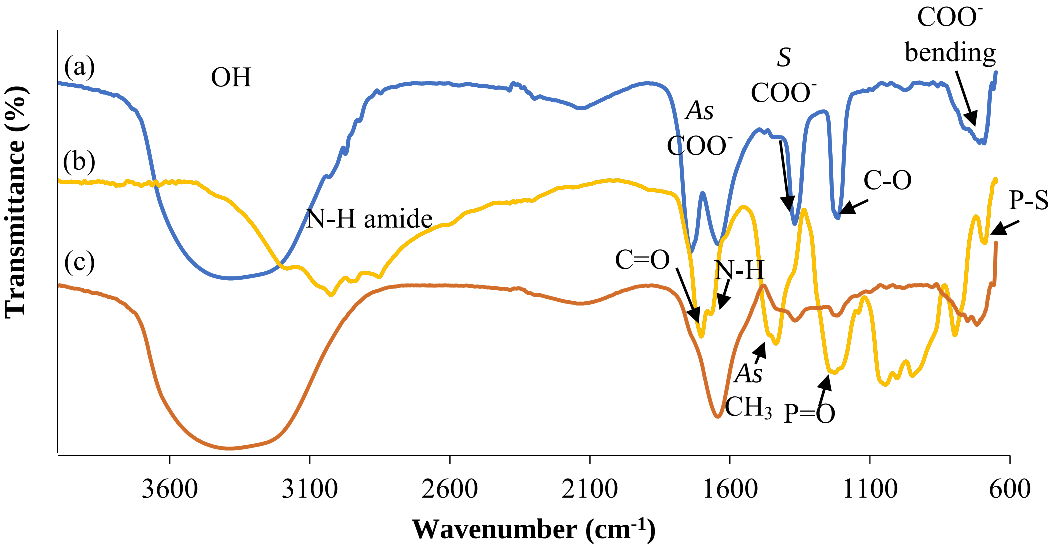


Figure 6: FTIR spectra of (a) Cit-AuNPs, (b) Ac, and (c) Cit-AuNPs-Ac complexes formed after detection of Ac using 9:1 Cit-AuNPs: Ac volume ratio.

4. CONCLUSION

The effect of Cit-AuNPs and Ac volume ratio (Cit-AuNPs: Ac) on the colorimetric detection of Ac was investigated in this study. The interaction between Ac and Cit-AuNPs resulted in colour change from red to purple / dark purple using different Cit-AuNPs: Ac volume ratios, due to the aggregation of Cit-AuNPs. The results also showed that RVs decreased when intensities of purple colour complexes increased. The darkest purple colour complex with the lowest mean of RVs was obtained when 9:1 of Cit-AuNPs: Ac volume ratio was applied. Contrary, the highest mean of RV (the lowest purple colour intensity) was achieved when applying the lowest volume ratio of Cit-AuNPs : Ac *i.e.* 1: 9. In conclusion, the lowest volume ratio of Cit-AuNPs : Ac was sufficient to produce obvious colour change for detection of Ac.

ACKNOWLEDGEMENT

This study was financial supported (UPNM/2016/GPJP/3/SG/2) by the National Defence University of Malaysia (NUDM). The authors also would like to acknowledge the university for providing the research facilities required for this study.

REFERENCES

Ali, S. S., Gangopadhyay, A., Pramanik, A. K., Guria, U. N., Samanta, S. K., & Mahapatra, A. K. (2019). Ratiometric sensing of nerve agent mimic DCP through in situ benzisoxazole formation. *Dye. Pigment.* **170**:107585–94.

Bala, R., Sharma, R. K., & Wangoo, N. (2015). Development of gold nanoparticles-based aptasensor for the colorimetric detection of organophosphorus pesticide phorate. *Anal. Bioanal. Chem.* **408**:333–38.

Bartelt-Hunt, S. L., Knappe, D. R. U., & Barlaz, M. A. (2008). A review of chemical warfare agent simulants for the study of environmental behavior. *Crit. Rev. Environ. Sci. Technol.* **38**:112–36.

Čolović, M. B., Krstić, D. Z., Lazarević-Pašti, T. D., Bondžić, A. M., & Vasić, V. M. (2013).

Acetylcholinesterase Inhibitors : Pharmacology and Toxicology. *Curr. Neuropharmacol.* **11**:315–35.

Duenchay, P., Chailapakul, O., & Siangproh, W. (2019). A transparency sheet-based colorimetric device for simple determination of calcium ions using induced aggregation of modified gold nanoparticles. *Int. J. Mol. Sci.* **20**:2954–66.

Duenchay, P., Kaewjua, K., Chailapakul, O., & Siangproh, W. (2020). Application of Modifier-free Gold Nanoparticle Colorimetric Sensing for Rapid Screening and Detecting Vitamin B1. *New J. Chem.* **44**:9223–29.

Dyguda-Kazimierowicz, E., Roszak, S., & Sokalski, W. A. (2014). Alkaline hydrolysis of organophosphorus pesticides: The dependence of the reaction mechanism on the incoming group conformation. *J. Phys. Chem. B* **118**:7277–89.

Fahimi-Kashani, N., & Hormozi-Nezhad, M. R. (2016). Gold-nanoparticle-based colorimetric sensor array for discrimination of organophosphate pesticides. *Anal. Chem.* **88**:8099–8106.

Hsu, C. W., Lin, Z. Y., Chan, T. Y., Chiu, T. C., & Hu, C. C. (2017). Oxidized multiwalled carbon nanotubes decorated with silver nanoparticles for fluorometric detection of dimethoate. *Food Chem.* **224**:353–58.

Jain, M., Yadav, P., Joshi, A., & Kodgire, P. (2019). Advances in detection of hazardous organophosphorus compounds using organophosphorus hydrolase based biosensors. *Crit. Rev. Toxicol.* **49**:387–410.

Jazayeri, M. H., Amani, H., Pourfatollah, A. A., Pazoki-Toroudi, H., & Moghadam, B. S. (2016). Various methods of gold nanoparticles (GNPs) conjugation to antibodies. *Sens. Bio-Sensing Res.* **9**:17–22.

Jin, N. Z., Anniebell, S., Gopinath, S. C. B., & Chen, Y. (2016). Variations in Spontaneous Assembly and Disassembly of Molecules on Unmodified Gold Nanoparticles. *Nanoscale Res. Lett.* **11**:339–49.

Joshi, K. A., Prouza, M., Kum, M., Wang, J., Tang, J., Haddon, R., Chen, W., & Mulchandani, A. (2006). V-type nerve agent detection using a carbon nanotube-based amperometric enzyme electrode. *Anal. Chem.* **78**:331–36.

Koskela, H. (2010). Use of NMR techniques for toxic organophosphorus compound profiling. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **878**:1365–81.

Kumar, V., & Upadhay, N. (2013). Chemical and Biochemical Mechanistic Fate of Acephate. *Int. J. Sci. Eng. Res.* **4**:2674–78.

Lang, N. J., Liu, B., Zhang, X., & Liu, J. (2013). Dissecting Colloidal Stabilization Factors in Crowded Polymer Solutions by Forming Self-Assembled Monolayers on Gold Nanoparticles. *Langmuir* **29**:6018–24.

Li, X., Cui, H., & Zeng, Z. (2018). A simple colorimetric and fluorescent sensor to detect organophosphate pesticides based on adenosine triphosphate-modified gold nanoparticles. *Sensors* **18**:4302–14.

Mazurek, S., Mucciolo, A., Humbel, B. M., & Nawrath, C. (2013). Transmission Fourier transform infrared microspectroscopy allows simultaneous assessment of cutin and cell-wall polysaccharides of Arabidopsis petals. *Plant J.* **74**:880–91.

Mohseni, N., Bahram, M., & Baheri, T. (2017). Chemical nose for discrimination of opioids based on unmodified gold nanoparticles. *Sensors Actuators, B Chem.* **250**:509–17.

Nietzold, C., & Lisdat, F. (2012). Fast protein detection using absorption properties of gold nanoparticles. *Analyst* **137**:2821–26.

Nilam, M., Hennig, A., Nau, W. M., & Assaf, K. I. (2017). Gold Nanoparticle Aggregation Enables Colorimetric Sensing Assays for Enzymatic Decarboxylation. *Anal. Methods* **9**:2784–87.

Osman, M. J., Wan Yunus, W. M. Z., Ong, K. K., & Abd Rashid, J. (2019). Defence Science , Engineering & Technology Recent Advances Techniques for Detection of Organophosphates : A Review. *J. Def. Sci. Eng. Technol.* **2**:49–70.

Park, J. W., & Shumaker-Parry, J. S. (2014). Structural study of citrate layers on gold nanoparticles: Role of intermolecular interactions in stabilizing nanoparticles. *J. Am. Chem. Soc.* **136**:1907–21.

Phugare, S. S., Gaikwad, Y. B., & Jadhav, J. P. (2012). Biodegradation of acephate using a developed bacterial consortium and toxicological analysis using earthworms (Lumbricus terrestris) as a model animal. *Int. Biodeterior. Biodegrad.* **69**:1–9.

Pluchery, O., Remita, H., & Schaming, D. (2013). Demonstrative experiments about gold nanoparticles and nanofilms: An introduction to nanoscience. *Gold Bull.* **46**:319–27.

Pradeep, W. P., & Ashok, P. J. (2017). A Review on Gold Nanoprticles Synthesis and Characterization. *Univers. J. Pharm. Res.* **2**:65–69.

Rana, K., Bhamore, J. R., Rohit, J. V., Park, T. J., & Kailasa, S. K. (2018). Ligand exchange reactions on citrate-gold nanoparticles for a parallel colorimetric assay of six pesticides. *New J. Chem.* **42**:9080–90.

Rastogi, L., Dash, K., & Ballal, A. (2017). Selective colorimetric/visual detection of Al3+ in ground water using ascorbic acid capped gold nanoparticles. *Sensors Actuators, B Chem.* **248**:124–32.

Rawat, K. A., Majithiya, R. P., Rohit, J. V., Basu, H., Singhal, R. K., & Kailasa, S. K. (2016). Mg2+ ion as a tuner for colorimetric sensing of glyphosate with improved sensitivity: Via the aggregation of 2-mercapto-5-nitrobenzimidazole capped silver nanoparticles. *RSC Adv.* **6**:47741–52.

Rebollar-Pérez, G., Lima-Zambrano, F., Bairán, G., Rodríguez-Enríquez, A., Ornelas-Soto, N., Méndez, E., & Torres, E. (2016). Colorimetric Assay for Detection of Organophosphorus Pesticides by Decrease of Standard Catalytic Activity of Chloroperoxidase. *Environ. Eng. Sci.* **33**:951–61.

Reynoso, E. C., Torres, E., Bettazzi, F., & Palchetti, I. (2019). Trends and perspectives in immunosensors for determination of currently-used pesticides: The case of glyphosate, organophosphates, and neonicotinoids. *Biosensors* **9**:1–20.

Rios-Corripio, M. A., Garcia-Perez, B. E., Jaramillo-Flores, M. E., Gayou, G. L., & Rojas-Lopez, M. (2013). UV – Visible intensity ratio (aggregates / single particles) as a measure to obtain stability of gold nanoparticles conjugated with protein A. *J. Nanoparticle Res.* **15**:1624–32.

Romano, J. A. Jr., Lukey, B. J., & Salem, H. (2009). Chemical Warfare Agents: Chemistry, Pharmacology, Toxicology, and Therapeutics. *Int. J. Toxicol.* **28:**132–34.

Songa, Everlyne A., & Okonkwo, Jonathan O. (2016). Recent approaches to improving selectivity and sensitivity of enzyme-based biosensors for organophosphorus pesticides: A review. *Talanta* **155**:289–304.

La Spina, R., Spampinato, V., Gilliland, D., Ojea-Jimenez, I., & Ceccone, G. (2017). Influence of different cleaning processes on the surface chemistry of gold nanoparticles. *Biointerphases* **12**:031003.

Sulaiman, I. S. Che, Chieng, B. W., Osman, M. J., Ong, K. K., Rashid, J. I. A., Yunus, W. M. Z. Wan, & Noor, S. A. M. (2020). A review on colorimetric methods for determination of organophosphate pesticides using gold and silver nanoparticles. *Microchim. Acta* **187**:1–22.

Wei, C., Zhou, H., & Zhou, J. (2011). Ultrasensitively sensing acephate using molecular imprinting techniques on a surface plasmon resonance sensor. *Talanta* **83**:1422–27.

Xu, Q., Du, S., Jin, G. D., Li, H., & Hu, X. Y. (2011). Determination of acetamiprid by a colorimetric method based on the aggregation of gold nanoparticles. *Microchem. Acta* **173**:323–29.

Yadav, R., Patel, P. N., & Lad, V. N. (2018). High selective colorimetric detection of Cd2+ ions using cysteamine functionalized gold nanoparticles with cross-linked DL-glyceraldehyde. *Res. Chem. Intermed.* **44**:2305–17.

Yang, Y. C. (1999). Chemical detoxification of nerve agent VX. *Acc. Chem. Res.* **32**:109–15.

Yeh, Y. C., Creran, B., & Rotello, V. M. (2012). Nanoscale Gold nanoparticles : preparation, properties, and applications in bionanotechnology. *Nanoscale* **4**:1871–80.

Zehani, N., Kherrat, R., Dzyadevych, S. V., & Jaffrezic-Renault, N. (2015). A microconductometric biosensor based on lipase extracted from Candida rugosa for direct and rapid detection of organophosphate pesticides. *Int. J. Environ. Anal. Chem.* **95**:466–79.

Zhang, J., Zheng, W., & Jiang, X. (2018). Ag+-Gated Surface Chemistry of Gold Nanoparticles and Colorimetric Detection of Acetylcholinesterase. *Small* **14**:1–8.

Zhou, Y., Li, C., Liu, R., Chen, Z., Li, L., Li, W., He, Y., & Yuan, L. (2020). Label-Free Colorimetric Detection of Prothioconazole Using Gold Nanoparticles Based on One-Step Reaction. *ACS Biomater. Sci. Eng.* **6**:2805–11.

A REVIEW OF CALIXARENE LANGMUIR–BLODGETT THIN FILM CHARACTERISTICS FOR NANOSENSOR APPLICATIONS

Mohd Syahriman Mohd Azmi*, Muhammad Faiz Farhan Noorizhab, Faridah Lisa Supian & Syed A Malik

Department of Physics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris (UPSI), Malaysia

*Email: syahriman@fsmt.upsi.edu.my

ABSTRACT

Calixarene is a vase-like supramolecule structure with many useful characteristics. For several decades, many studies have been conducted about the characteristics of calixarene that can be applied in fields such as material science, biotechnology and nanosensors. The Langmuir–Blodgett (LB) technique is a powerful method in obtaining a homogenous monolayer or several layers of thin film by controlling the molecular orientation on the air–water surface. Novel characteristics of calixarene have been discovered in applications such as sensing material, water treatment and catalysts. Calixarene also has the ability to form a stable complex with biomolecules, making it able to be developed as a biosensor. The discussion in this review paper focuses on the characteristics of calixarene thin films made using the LB method that are suitable for nanosensor applications.

Keywords: *Calixarenes; Langmuir–Blodgett (LB); thin film; nanosensors; Langmuir monolayer.*

1. INTRODUCTION

Calixarene is a vase cavity-shaped group bound by cyclic molecules. This nano-basket molecule is a well-known supramolecule along with cyclodextrins and crown ester (Gutsche, 2008). Calixarene has a unique structure, and its lower and upper rims are functionally modifiable with other functional groups or ligands. Due to this trait, calixarene is an excellent platform for numerous host–guest interactions (Figure 1) for countless anions, cations, and organic or inorganic molecules with high selectivity (Gutsche, 1989; Supian, 2010; Mokhtari & Pourabdollah, 2012). Calixarene is also suitable for water treatment applications, metal recovery from solutions and nuclear waste management, and can be used as a catalyst (Mandolini & Ungaro, 2000; Bingol *et al.*, 2010; Tabakci & Yilmaz, 2014; Lim & Supian, 2017; Acikbas *et al.*, 2017).

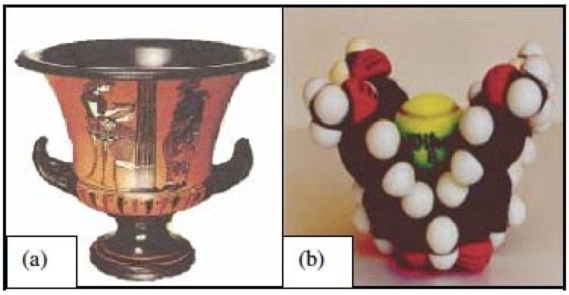


Figure 1: Calixarene is known as a host–guest material: (a) Calix crater. (b) CPK model carrying a guest. (Source: Supian, 2010)

The molecular structure of calixarene resembles a basket shape with a distinct upper rim, which is hydrophobic, hydrophilic lower rim and central annulus, as shown in Figure 2 (McMohan *et al.*, 2003; Gorbunov *et al.*, 2020). A number is placed between calix and arene, which results in calix[n]arene.

The [n] denotes the quantity of the aryl group. Hence, a cyclic tetramer is called calix[4]arene, a cyclic hexamer is called calix[6]arene, and a cyclic octamer is called calix[8]arene (Eddaif *et al.*, 2019a), as displayed in Figure 3 (McMohan *et al.*, 2003).

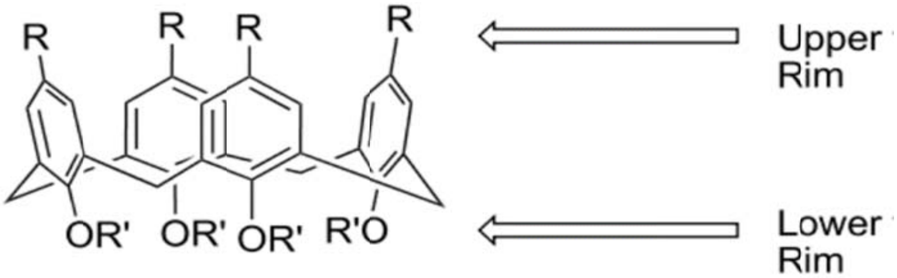


Figure 2: Schematic representation of calix[4]arene structure.
(Source: McMohan *et al.*, 2003)

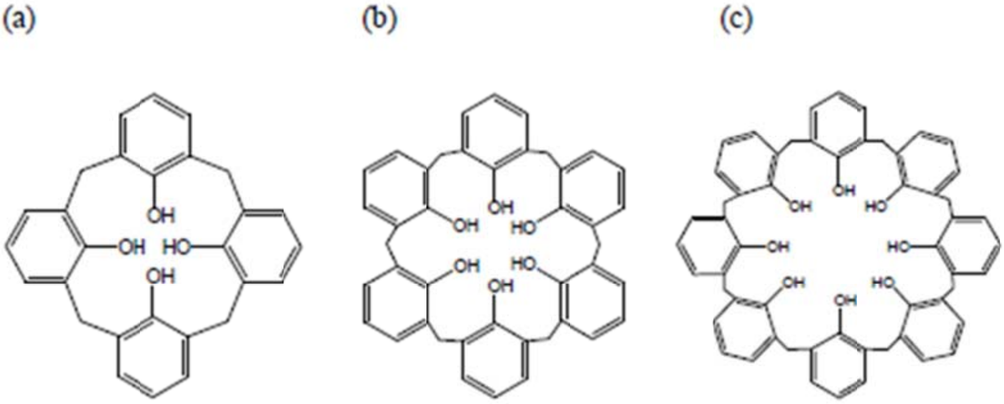


Figure 3: Chemical structures of (a) calix[4]arene, (b) calix[6]arene and (c) calix[8]arene.
(Source: McMohan *et al.*, 2003)

Numerous research works have been conducted on the properties of calixarene, leading to innumerable applications. Calix[4]arenes Langmuir–Blodgett (LB) alternate layers (containing either carboxylic acid or amino substituents) have been proven as one of the effective methods with a highly non-linear I-V behaviour that is suitable for electrical applications (Ozbek *et al.*, 2019). Flores-Sanchez *et al.* (2020) found that the presence of calixarene may enhance gas sensor response against ammonia and amine vapours due to the higher accessibility of volatiles to disaggregated porphyrins in LB films. Acikbas *et al.* (2018) highlighted that calix[4]arene thin film has high selectivity as chemical sensor material with large response to dichloromethane vapour.

The objective of this review is to discuss on the properties of calixarene thin films made mainly using the LB method that are suitable for nanosensor applications. In this paper, a brief outline of the basic structures of calixarenes, the LB film and its properties, as well as its optical and electrical conductivities are studied. Lastly, reports on hybrids of calixarenes with assorted materials using LB are summarised.

2. CALIXARENE LANGMUIR–BLODGETT (LB) FILMS

A LB film is defined as a highly ordered monolayer or multilayered film transferred onto a solid substrate using a LB trough. The LB film contains one or more monolayers of an amphiphile. The suitable molecule for LB should possess an amphiphilic characteristic, which must have both hydrophilic and hydrophobic characters, as shown in Figure 4 (Supian, 2010). A study on surface potential and temperature effect of amphiphilic polysiloxane using the LB technique has been successfully accomplished by Supian *et al.* (2019).

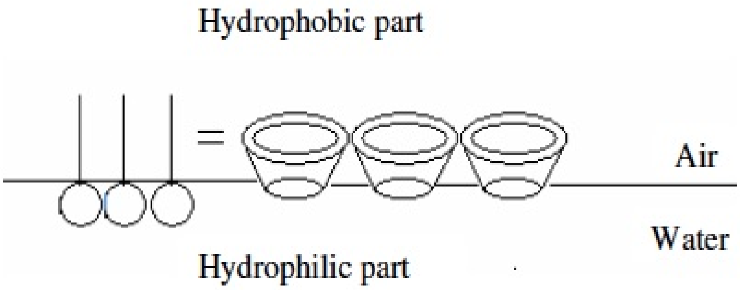


Figure 4: Hydrophilic and hydrophobic parts of a simple amphiphile (left) and calixarene (right). (Source: Supian, 2010)

A Langmuir monolayer can be transferred onto solid substrates by dipping a solid substrate, such as glass or silicon (Si), up and down through the monolayer, as shown in Figure 5. There are three types of LB depositions:

- X-type: Head-to-tail assembly to the substrate
- Y-type: Head-to-head alternating with tail-to-tail assembly to the substrate
- Z-type: Tail-to-head assembly to the substrate.

When the surface pressure isotherm (Π) is high enough (solid phase) to ensure a homogeneous multilayer, the attraction between the molecules in the monolayer provides lateral strength to prevent the monolayer from falling apart during the transfer to the solid substrate. The film deposition depends on the amphiphile itself, but it usually ranges from 10 to 40 mN/m for calixarenes (Supian, 2010).

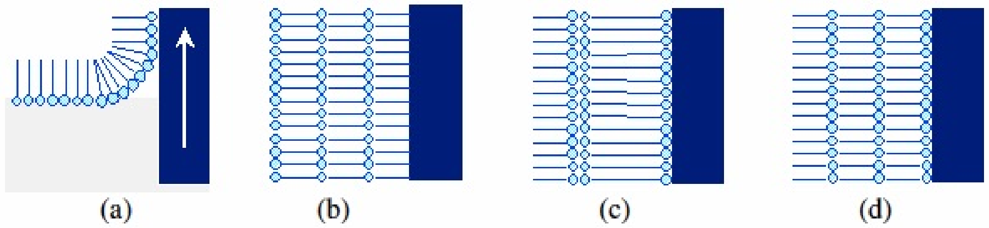


Figure 5: Deposition of (a) a floating monolayer on a solid substrate using (b) X-type, (c) Y-type and (d) Z-type depositions. (Source: Supian, 2010)

Calix[4]arene and calix[8]arene can form a highly uniform monolayer and well-ordered Langmuir film on water subphase transferred LB multilayer solid substrates (Sen *et al.*, 2019). This type of film has been applied in areas of gas sensing, ion binding (Hassan *et al.*, 1998) and pyroelectric heat detection (McCartney *et al.*, 1997). Among small organic molecules, calixarene is unique due to its high thermal stability, with a melting point normally around 250–280 °C (Boehmer *et al.*, 1987). This is exceptional for an LB-forming material that usually has melting point in the range of 60–80 °C (Robert, 1990).

Weis *et al.* (2007) conducted an interaction study of a calix[4]resorcinarene in a LB form monolayer with dopamine using surface potential (Δ*V*), while Nabok *et al.* (1995) discovered the behaviour of calix[4]resorcinolarene using Π and surface potential–area (Δ*V*-*A*) data in order to investigate the adsorption mechanism of some different upper rims of calix[4]resorcinolarene. Supian *et al.* (2010) conducted a study of the interaction of the LB films of two calixarenes (Figure 6) with aqueous copper (Cu^{2+}) and lithium (Li^{+}) ions. The results of the study showed the binding interaction between Cu^{2+} and Li^{+} with the two calix[4]arenes, which were named 5,11,17,23-tetra-tert-butyl-25,27-diethoxycarbonyl methyleneoxy-26,28-ihydroxycalix[4]arene (material A) and 5,17-(9H-fluoren-2-yl)methyleneamino)-11,23-di-tert-butyl-25,27-diethoxycarbonylmethyleneoxy-26,28-dihydroxycalix [4]arene (material B).

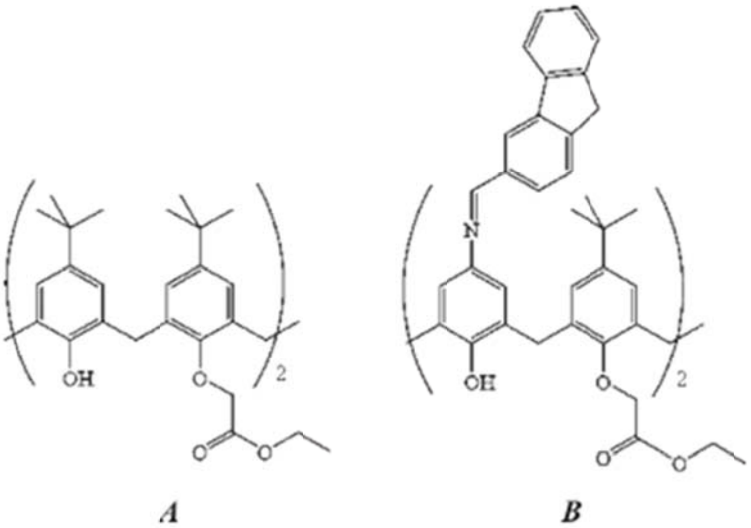


Figure 6: The two calixarenes used in Supian *et al. (*2010): (A) 5,11,17,23-tetra-tert-butyl-25,27-diethoxycarbonylmethyleneoxy-26,28-dihydroxycalix[4]arene and (B) 5,17-(9H-fluoren-2-yl) methyleneamino)-11,23-di-tert-butyl-25,27-diethoxycarbonyl)methyleneoxy-26,28-dihydroxycalix[4]arene.

Figure 7 shows a perspective view of the calix-Schiff B 1/2CHCl$_3$. Hydrogen bonds are shown as dashed lines. Hydrogen atoms and minor components of the disorder have been omitted for clarity. The unsubstituted phenols at the lower rim each diverge at the hydrogen bonds with adjoining phenyl ether and carbonyl oxygen atoms (Supian *et al.*, 2010).

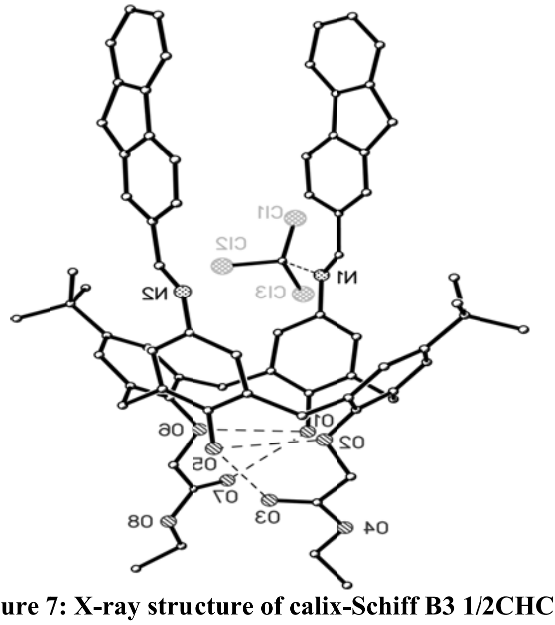


Figure 7: X-ray structure of calix-Schiff B3 1/2CHCl$_3$.
(Source: Supian *et al.*, 2010)

Azahari *et al.* (2014) studied calix[4]arene LB film as a potential compound for molecular recognition as ionophores on capturing lead in water. Calix[4]resorcinarene macrocycle properties have been extensively studied from biological controls to heavy metal ion sensing, such as interacting with Cd^{2+}, Pb^{2+}, Hg^{2+} and Cu^{2+} via Π (Maya *et al.*, 2017; Eddaif *et al.*, 2019b).

Figure 8 shows the isotherm behaviour of material B in Supian *et al.* (2010) using three different Li^+ concentrations. A calixarene undertakes two principal state transitions prior to compression. The first transition is the gas–liquid transition, which starts at ~3.2 nm^2/molecule (for the pure water subphase). The second transition is the liquid–liquid or liquid–quasi-solid transition, which starts at ~2.0 nm^2/molecule. The second transition is not a liquid–solid transition due to the compressibility of the high-pressure phase that is lower than normally predicted for a two-dimensional solid phase. The highly amphiphilic nature of a calixarene leads to a well-defined isotherm, which is known to yield a well-ordered Langmuir film on air–water interface (Supian *et al.*, 2010; Lim & Supian, 2019).

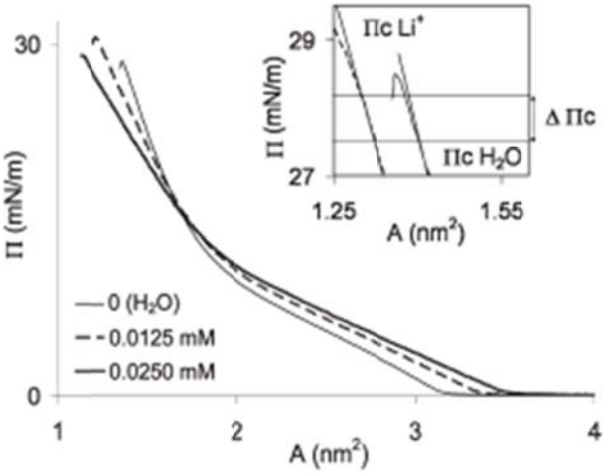


Figure 8: Surface pressure isotherm (Π) using three different concentrations of Li^+ ions in the subphase. The inset diagram shows that ΔΠc = ΠcH2O-Π ions, showing the difference of the collapse pressure of water and Li^+ ion.
(Source: Supian *et al.*, 2010)

3. OPTICAL PROPERTIES OF CALIXARENE

The electron transport properties in polycrystalline films are strongly affected by intergrain properties. Absorption spectra are more appropriate in determining the energy band gap. In the fundamental absorption edge region, the absorption coefficient (*a*) depends on the energy of the incident photon (*hv*), as in the following equation (Leontie *et al.*, 2018):

$$ahv = A(hv - E_{gh})^{\,n} \qquad (1)$$

Leontie *et al.* (2018) investigated the absorption spectra and optical band of calixarene thin film, as shown in Figures 9 and 10 respectively. Figure 9 shows the absorption spectra in the photon energy range for ultraviolet, visible and near infrared (0.71–4.03 eV). Absorbance decreased with increase in wavelength. The highest absorption decreases were in the ultraviolet region (4.03–3.50 eV), while other samples in the study by Yadav *et al.* (2010) showed a decrease in absorbance for the visible region (1.54–2.94 eV).

Using Equation 1, the curves can be fitted to any standard dependence, in particular $(ahv)^2 = f(hv)$, to allow a direct transition. The best fit is obtained for $(ahv)^2 = f(hv)$ dependence, as consequently shown in Figure 10. The curves indicate the estimation of the optical band gap values. The value of the direct optical gap corresponds to the band-to-band transition, where the activation energies (E_{a2}) were determined by the electronic transport mechanism in actual organic films (Leontie *et al.*, 2018).

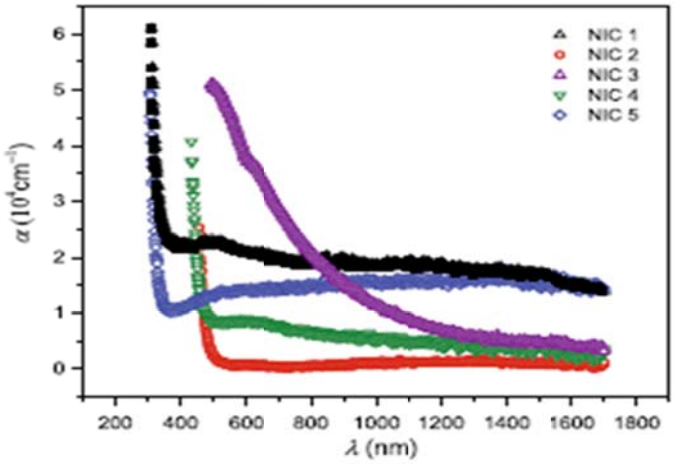


Figure 9: Plot of absorption coefficient vs. photon energy for the investigation in Leontie *et al.* (2018).

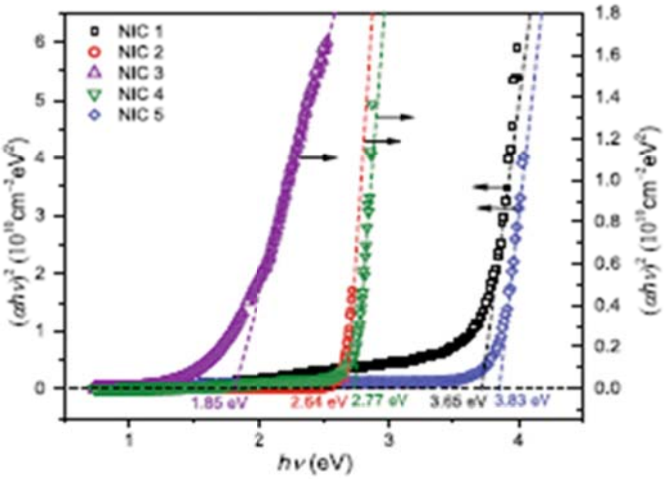


Figure 10: Optical absorption spectra of present organic films.
(Source: Leontie *et al.*, 2018)

4. CONDUCTIVITY ENHANCEMENT OF CALIXARENE / MWCNT FILM

Calixarene is the most promising candidate among supramolecules to be used as nanosensors, because of its host–guest properties that can be modified through repetition of the upper and lower rims. Different conductivities obtained showed results in different conformations of calixarenes (Csokai, 2006; Botha *et al.*, 2014), providing highly selective binding energy towards certain ionic or molecular substances (Chen *et al.*, 2000). Some have even shown chromogenic sensing properties (Yang *et al.*, 2000; Capan *et al.*, 2010). Calix[8]arene has been reported to provide a fast and reversible sensing response to chloroform (Qureshi *et al.*, 2008; Capan *et al.*, 2010). These supramolecules have poor conductivity despite having superior sensing ability. However, several studies have been conducted with calixarene carbon nanotubes (CNTs), leading to the discovery of better results in guest detection (Wang *et al.*, 2012; Gaicore & Srivastava, 2012; Mermer *et al.*, 2012). The addition of conducting polymers has also been used in the approach to increase the sensing properties of calixarene (Viglok & Swager, 2002; Gokoglan *et al.*, 2015).

Table 1 presents the conductivity values of calixarenes. The combination of calixarenes and multi-walled carbon nanotubes (MWCNTs) resulted in the development of sensitivity in the calix[8]arene thin films and increase in conductivity of the thin films by a great ratio. This is because MWCNT is a conductor that eases electron transfer (Supian *et al.*, 2017). For a calix[8]arene / MWCNT thin film ratio of 2:1, a decreased conductivity was observed as compared to calix[8]arene / MWCTN with ratio of 1:1 (Chaabane *et al.*, 1994; Wang *et al.*, 2012; Ozbek *et al.*, 2013; Wang *et al.*, 2015). This is due to increasing levels of calix[8]arene in the film.

Table 1: Resistivity, conductivity and thickness of C[8]1/MWCNTs and C[8]2/MWCNT thin films. (Source: Supian *et al.*, 2017)

Ratio in sample calix[8]arene/MWCNTs	Resistivity, ρ (Ωm)		Conductivity, σ ($\mu\Omega^{-1\ m-1}$)		Thickness, h (μm)	
	C[8]1/MWCNTs	C[8]2/MWCNTs	C[8]1/MWCNTs	C[8]2/MWCNTs	C[8]1/MWCNTs	C[8]2/MWCNTs
1:0	-	-	-	-	12.76	12.20
1:1	16554.36	13870.49	60.407	72.096	16.48	14.36
1:2	3320.39	1071.55	301.169	933.228	22.66	20.18
2:1	24483.97	14642.58	40.843	68.294	20.23	16.75

Supian *et al.* (2013) and Razali *et al.* (2015a, b) visualised the morphologies of calix[8]arene / MWCNT composites. Film ratio of 1:2 gave higher value of conductivity as compared to film ratio of 1:1. This led to the assumption that the results obtained were caused by the preference of MWCNTs contacting better with calix[8]arene with a ratio of 2, resulting from the type of orientation for the attachment of calix[8]arene with the ratio of 2 to MWCNTs.

5. HYBRIDS OF CALIXARENE THIN FILM

While calixarene itself has many advantages in molecular recognition, hybridising them enhances and extends their potential through extensive chemical versatility of the modified upper and lower rims, as well as various sizes of cavities. The discovery of a new macrocyclic building block that enables the build-up of a functional architecture on its semirigid scaffold started a new direction for research in supramolecular chemistry and related fields (Steed & Gale, 2012; Zhu & Fang, 2014).

Versatile macrocycles such as calix[n]arenes were discovered decades ago and have been widely exploited since then (Leontie *et al.*, 2018). The functionalisation of the existing scaffold is currently in focus rather than finding a new one, because the existing scaffold offers almost unlimited possibilities. For example, pillar[5]arene is a very useful macrocylic building block (Ogoshi *et al.*, 2008).

5.1 Calixarene and Gold Nanoparticles

Gold nanoparticle properties include high chemical stability, surface functionalisation, biocompatibility and tunable optics. Its electronic properties have led to many promising applications (Cioffi *et al.*, 2011; Li & Yang, 2013). A calixarene cavity provides a side binding for an organic guest molecule. The selectivity to particular analytes of a calixarene can be controlled by altering the size of the cavity and by peripheral substituted groups (Ozmen *et al.*, 2014a, 2015). A calixarene molecule is excellent for use as a thin film-forming material, which can then be used as a host molecule for both organic (Nabok *et al.*, 1997; Ozbek *et al.*, 2011) and inorganic gasses (Richardson *et al.*, 2006; Ohira *et al.*, 2009).

Figure 11 shows the UV–visible spectra of thin films with one gold nanoparticle (AuNP) layer and two calixarene layers, as well as hybrid films consisting of one AuNP layer and two calixarene molecule layers on top (Capan *et al.*, 2017). The AuNP monolayer shows a broad absorption band of UV–visible spectra at room temperature (Heriot *et al.*, 2006; Hardy & Richardson, 2008; Capan *et al.*, 2017). The absorbance of a calixarene bilayer is significantly low (Ozbek *et al.*, 2011; Ozmen *et al.*, 2014b; Capan *et al.*, 2017). On the other hand, hybridisation greatly increases the absorbance of the film (Manera *et al.*, 2008; Capan *et al.*, 2017).

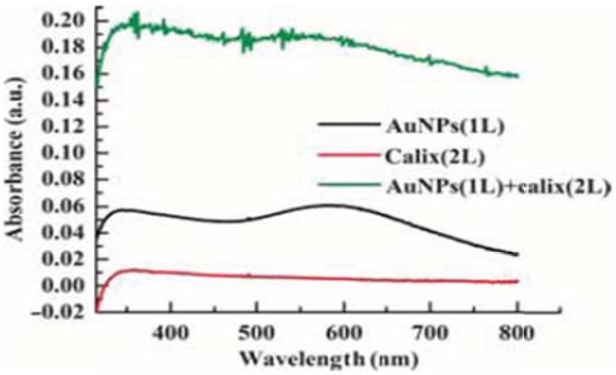


Figure 11: UV–visible spectra of thin films with one AuNP layer, and two calixarene layers and hybrid films.
(Source: Capan *et al.*, 2017)

5.2 Calix[4]naphthalene

Calix[n]naphthalene is a macrocycle consisting of naphthalene units joined through methylene bridges at metaposition. This can provide advantages as macrocyclic scaffolds originate from an enlarged cavity with electron-rich aromatic surfaces (Boinski *et al.*, 2016). In addition, the naphthalene geometry causes the macrocycles to be uneven and thus display interesting stereochemical characteristics such as inherent chirality. Although the extended cavity features of calix[n]naphthalene have been fully studied (Ashram *et al.*, 2001), the number of family members of calix[n]naphthalene remains small, including 1-hydroxynaphthalene (Georghiou *et al.*, 1995; Georghiou *et al.*, 1999; Mizyed *et al.*, 2001) and 2-hydroxynaphthalene (Chowdhury & Georghiou, 2002), as well as various macrocycles obtained from 2,7-dialkoxynaphthalenes (Shorthill & Glass, 2001; Shorthill *et al.*, 2002) or chromatic acid (Poh *et al.*, 1989). Pillar[5]arenes and hybrid[4]arenes can be synthesised through the reaction of Bronsted acid-mediated directly with alkoxybenzenes with formaldehyde (Boinkski & Szumna, 2012; Boinski *et al.*, 2015).

6. CONCLUSION

Calixarene, despite having poor electrical conductivity, is still the most promising candidate among supramolecules to be used as nanosensors, due to its unique host-guest properties that can be modified through repetition of the upper and lower rims, which leads to superior sensing element. A calixarene cavity provides a side binding for an organic guest molecule. The selectivity to particular analyses of a calixarene can be controlled by altering the size of the cavity and by peripheral substituted groups. Researchers have found that hybridisation of calixarene thin film with nanomaterials tremendously increases its conductivity. The combination of calixarenes and MWCNTs results in the development of sensitivity in calix[8]arene thin films and an increase in conductivity by a great ratio. This is as MWCNT is a conductor that eases electron transfer. On top of that, hybridisation of calix[8]arene thin film with inorganic nanoparticles greatly increases the absorbance of the device. Calixarenes are typically used in nanosensors by coating nanomaterials with gold and then tethering the calixarenes to the gold.

REFERENCES

Acikbas, Y., Bozkurt, S., Halay, E., Capan, R., Guloglu, M. L., Sirit, A. & Erdogan, M. (2017). Fabrication and characterization of calix[4]arene Langmuir-Blodgett thin film for gas sensing applications. *J. Incl. Phenom. Macrocycl. Chem.*, **89**: 77-84.

Acikbas, Y., Bozkurt, S., Erdogan, M., Halay, E., Sirit, A. & Capan, R. (2018). Optical and vapor sensing properties of calix[4]arene Langmuir-Blodgett thin films with host–guest principles. *J. Macromol. Sci. – Pure App. Chem.*, **55**: 526-532.

Ashram, M., Mizyed, S. & Georghiou, P.E. (2001). Synthesis of hexahomotrioxacalix[3]naphthalenes and a study of their alkali-metal cation binding properties. *J. Org. Chem.*, **66**: 1473-1479.

Azahari, N.A., Supian, F.L., Richardson, T.H. & Malik, S.A. (2014). Properties of calix4-lead (Pb) films using Langmuir-Blodgett (LB) technique as an application of ion sensor. *Adv. Mater. Res.*, **895**: 8-11.

Bingol, H., Kocabas, E., Zor, E. & Coskun, A. (2010). A novel benzothiazole based azocalix[4]arene as a highly selective chromogenic chemosensor for Hg^{2+} Ion: A rapid test application in aqueous environment. *Talanta*, **82**: 1538-1542.

Boehmer, V., Marschollek, F. & Zetta, L. (1987). Calix[4]arenes with four differently substituted phenolic units. *J. Org. Chem.*, **52**: 3200-3205.

Boinski, T. & Szumna, A. (2012). A facile, moisture-insensitive method for synthesis of pillar[5]arenes—the solvent templation by halogen bonds. *Tetrahedron*, **68**: 9419-9422.

Boinski, T., Cieszkowski, A., Rosa, B. & Szumna, A. (2015). Hybrid [n]Arenes through thermodynamically driven macrocyclization reactions. *J. Org. Chem.*, **80**: 3488-3495.

Boinski, T., Cieszkowski, A., Rosa, B., Leśniewska, B. & Szumna, A. (2016). Calixarenes with naphthalene units: calix[4]naphthalenes and hybrid[4]arenes. *New J. Chem.*, **40**: 8892-8896.

Botha, F., Jan, B., Václav, E., Oldřich, H., Lukáš, V., Ivana, C. & Pavel, L. (2014). Recognition of chiral anions using calix[4]arene-based ureido receptor in the 1,3 alternate conformation. *Tetrahedron*, **70**: 477-483.

Capan, R., Ozbek, Z., Goktas, H., Şen, S., İnce, F.G., Ozel, M.E., Stanciu, G.A. & Davis, F. (2010). Characterization of Langmuir-Blodgett films of a Calix[8]arene and sensing properties towards volatile organic vapors. *Sens. Act. B: Chem.*, **148**: 358-365.

Capan, I., Hassan, A.K. & Abbas, R.R. (2017). Fabrication, characterization and gas sending properties of gold nanoparticles and calixarene multilayer. *Bull. Mat. Sci.*, **40**: 31-36.

Chaabane, R., Ben, M., Gamoudi, G., Guillaud, C., Jouve, F., Gaillard & Lamartine, R. (1994). Elaboration and characterization of thin calixarene films. *Synthetic Metals*, **66**: 49-54.

Chen, L., Zeng, X., Ju, H., He, X. & Zhang, Z. (2000). Calixarene derivatives as the sensory molecules for silver ion-selective electrode. *Microchem. J.*, **65**: 129-135.

Chowdhury, S. & Georghiou, P.E. (2002). Synthesis and properties of a new member of the calixnaphthalene family: A C2-symmetrical endo-calix[4]naphthalene. *J. Org. Chem.*, **67**: 6808-6811.

Cioffi, N., Colaianni, L., Ieva, E., Pilolli, R., Ditaranto, N., Angione, M.D., Cotrone, S., Buchholt, K., Spetz, A.L., Sabbatini, L. & Torsi, L. (2011). Electrosynthesis and characterization of gold nanoparticles for electronic capacitance sensing of pollutants. *Electrochimica Acta*, **56**: 3713-3720.

Csokai, V., Grün, A., Balázs, B., A., Gábor Tóth, S. & Bitter, I. (2006). Functionalized thiacalix- and calix[4]arene-based Ag+ Ionophores: Synthesis and comparative NMR study. *Tetrahedron*, **62**: 10215-10222.

Dragoman, D. & Dragoman, M. (2002). *Optical Characterization of Solids.* Springer, Berlin Heidelberg.

Eddaif, L., Shaban, A. & Telegdi, J. (2019a). Sensitive detection of heavy metals ions based on the calixarene derivatives-modified piezoelectric resonators: a review. *Int. J. Env. Analy. Chem.*, **99**: 824-853.

Eddaif, L., Shaban, A. & Telegdi, J. (2019b). Application of the Langmuir technique to study the response of C-dec-9-en-1-ylcalix[4]resorcinarene and C-undecylcalix[4]resorcinarene ultra-thin films' interactions with Cd^{2+}, Hg^{2+}, Pb^{2+}, and Cu^{2+} cations present in the subphase. *Water Air Soil Pollut.*, **230**, 279.

Flores-Sanchez, R., Gamez, F., Lopes-Costa, T. & Pedrosa, J.M. (2020). A calixarene promotes disaggregation and sensing performance of carboxyphenyl porphyrin films. *ACS Omega*, **5**: 6299-6308.

Gaichore, R.R. & Srivastava, A.K. (2012). Multiwalled carbon nanotube-4-tert-butyl calix[6]arene composite electrochemical sensor for clenbuterol hydrochloride determination by means of differential pulse adsorptive stripping voltammetry. *J. App. Electrochem.*, **42**: 979-987.

Georghiou, P.E., Ashram M., Li, Z. & Chaulk, S.G. (1995). Syntheses of calix[4]naphthalenes derived from 1-naphthol. *J. Org. Chem.*, **60**: 7284-7289.

Georghiou, P.E., Mizyed, S. & Chowdhury, S. (1999). Complexes formed from [60]fullerene and calix[4]naphthalenes. *Tetra. Lett.*, **40**: 611-614.

Gokoglan, T.C., Soylemez, S., Kesik, M., Unay, H., Sayin, S., Yildiz, H.B., Cirpin, A. & Toppare, L. (2015). A novel architecture based on a conducting polymer and calixarene derivative: Its synthesis and biosensor construction. *RSC Adv.*, **5**: 35940-35947.

Gorbunov, A., Kuznetsova, J., Deltsov, I., Molokanova, A., Cheshkov, D., Bezzubov, S., Kovalev, V. & Vatsouro, I. (2020). Selective azide-alkyne cycloaddition reactions of azidoalkylated calixarenes. *Org. Chem. Front.*, **7**: 2432-2441.

Gutsche, C.D. (1989). *Calixarenes*. The Royal Society of Chemistry, Cambridge, Great Britain.

Gutsche, C.D. (2008). *Calixarene: An Introduction.* The Royal Society of Chemistry, Cambridge, Great Britain.

Hardy, N.J. & Richardson, T.H. (2008). Temperature effects on the optical properties of thiol encapsulated gold nanoparticle thin films. *Coll. Surf. A: Physicochem. Eng. Asp.*, **321**: 285-291.

Hassan, A.K., Nabok, A.V., Ray, A.K., Davis, F. & Stirling, C.J.M. (1998). Complexation of metal ions with Langmuir–Blodgett films of novel calixarene azo-derivative. *Thin Solid Films*, **327**: 686-689.

Heriot, S.Y., Zhang, H.L., Evans, S.D. & Richardson, T.H. (2006). Multilayers of 4-methylbenzenethiol functionalized gold nanoparticles fabricated by Langmuir–Blodgett and Langmuir–Schaefer deposition. *Coll. Surf. A: Physicochem. Eng. Asp.*, **278**: 98-105.

Leontie, L., Danac, R., Carlescu, A., Doroftei, C., Rusu, G. G., Tiron, V., Gurlui, S. & Susu, O. (2018). Electric and optical properties of some new functional lower-rim-substituted calixarene derivatives in thin films. *App. Phy. A*, **124**: 1-12.

Li, H. & Yang, Y.W. (2013). Gold nanoparticles functionalized with supramolecular macrocycles. *Chin. Chem. Lett.*, **24**: 545-552.

Lim, D.C.K. & Supian, F.L. (2017). Elucidation of properties in Langmuir monolayers and Langmuir-Blodgett (LB) films formed by calix[6]arene. *Jurnal Fizik Malaysia*, **38**: 10034-10043.

Lim, D.C.K. & Supian, F.L. (2019). Calix[4]arene and calix[8]arene Langmuir films: surface studies, optical and structural characterizations. *Int. J. Inno. Tech. Exp. Eng.*, **8**: 80-85.

Mandolini, L. & Ungaro, R. (2000). *Calixarenes in Action*. Imperial College Press, London.

Manera, M.G., Spadavecchia, J., Buso, D., Fernández, C.D.J., Mattei, G., Martucci, A., Mulvaney, P., Pérez-Juste, J., Rella, R., Vasanelli, L. & Mazzoldi, P. (2008). Optical gas sensing of TiO_2 and TiO_2/Au nanocomposite thin films. *Sens. Act. B: Chem.*, **132**: 107-115.

Maya, R.J., Krishna, A., Sirajunnisa, P., Suresh, C.H., & Varma, R.L. (2017). Lower rim-modified calix[4]arene–bentonite hybrid system as a green, reversible, and selective colorimetric sensor for Hg^{2+} recognition. *ACS Sustain. Chemi. Eng.*, **5**: 6969–6977.

McCartney, C.M., Richardson, T., Greenwood, M.B., Cowlam, N., Davis, F. & Stirling, C.J.M. (1997). The effect of pendant chain structure on the pyroelectric behaviour of calix[8]arene Langmuir- Blodgett films. *Supramol. Sci.*, **4**: 385-390.

McMahon, G., O'Malley, S., & Nolan, K. (2003). Important calixarene derivatives – their synthesis and applications. *Arkivoc,* **7**: 23-31.

Mermer, O., Okur, S., Sumer, F., Ozbek, C., Sayın, S. & Yılmaz, M. (2012). Gas sensing properties of carbon nanotubes modified with calixarene molecules measured by QCM technique. *Acta Phy. Pol. A*, **121**: 240-242.

Mizyed, S., Ashram, M., Miller, D.O. & Georghiou, P.E. (2001). Supramolecular complexation of [60]fullerene with hexahomotrioxacalix[3]naphthalenes: a new class of naphthalene-based calixarenes. *J. Chem. Soc., Per. Trans. 2*, 1916-1919.

Mokhtari, B. & Pourabdollah, K. (2012). Extraction of alkali metals using emulsion liquid membrane by nano-baskets of calix[4]crown. *Kor. J. Chem. Eng.*, **29**: 1788-1795.

Nabok, A.V., Lavrik, N.V., Kazantseva, Z.I., Nesterenko, B.A., Markovskiy, L.N., Kalchenko, V.I. & Shivaniuk, A.N. (1995). Complexing properties of calix[4]resorcinolarene LB films. *Thin Solid Films*, **259**: 244-247.

Nabok, A.V., Hassan, A.K., Ray, A.K., Omar, O. & Kalchenko, V.I. (1997). Study of adsorption of some organic molecules in calix[4]resorcinolarene LB films by surface plasmon resonance. *Sens. Act. B: Chem.*, **45**: 115-121.

Ogoshi, T., Kanai, S., Fujunami, S., Yamagishi, T. & Nakamoto, Y. (2008). Para-bridged symmetrical pillar[5]arenes: Their lewis acid catalyzed synthesis and host–guest property. *J. Am. Chem. Soc.*, **130**: 5022-5023.

Ohira, S.I., Wanigasekara, E., Rudkevich, D.M. & Dasgupta, P.K. (2009). Sensing parts per million levels of gaseous NO_2 by an optical fiber transducer based on calix[4]arenes. *Talanta*, **77**: 1814-1820.

Ozbek, Z., Capan, R., Goktas, H., Sen, S., Ince, F.G., Ozel, M.E. & Davis, F. (2011). Optical parameters of calix[4]arene films and their response to volatile organic vapors. *Sens. Act. B: Chem.*, **158**: 235-240.

Ozbek, C., Culcular, E., Okur, S., Yilmaz, M. & Kurt, M. (2013). Electrical characterization of interdigitated humidity sensors based on CNT modified calixarene molecules. *Acta Phy. Pol. A*, **123**: 461-463.

Ozbek, Z., Davis, F., & Capan, R. (2019). Electrical properties of alternating acid and amino substituted calixarene Langmuir-Blodgett thin films. *J. Phys. Chem. Solids*, **136**, 109146.

Ozmen, M., Ozbek, Z., Bayrakci, M., Ertul, S., Ersoz, M. & Capan, R. (2014a). Preparation and gas sensing properties of Langmuir–Blodgett thin films of calix [n] arenes: investigation of cavity effect. *Sens. Act. B: Chem.*, **195**: 156-164.

Ozmen, M., Ozbek, Z., Buyukcelebi, S., Bayrakci, M., Ertul, S., Ersoz, M. & Capan, R. (2014b). Fabrication of Langmuir–Blodgett thin films of calix [4] arenes and their gas sensing properties: Investigation of upper rim para substituent effect. *Sens. Act. B: Chem.*, **190**: 502-511.

Ozmen, M., Ozbek, Z., Bayrakci, M., Ertul, S., Ersoz, M. & Capan, R. (2015). Preparation of Langmuir–Blodgett thin films of calix [6] arenes and p-tert butyl group effect on their gas sensing properties. *App. Surf. Sci.*, **359**: 364-371.

Poh, B.L., Lim, C.S. & Khoo, K.S. (1989). A water-soluble cyclic tetramer from reacting chromotropic acid with formaldehyde. *Tetra. Lett.*, **30**: 1005-1008.

Qureshi, I., Memon, S. & Yilmaz, M. (2008). Extraction and binding efficiency of calix[8]arene derivative toward selected transition metals. *Pak. J. Ana. Env. Chem.*, **9**: 96-100.

Razali, A.S., Supian, F.L., Salleh, M.M. & Bakar, S.A. (2015a). Characterization and detection of cadmium ion using modification calixarene with multiwalled carbon nanotubes. *Int. J. Chem., Mol., Nuc., Mat. Metal. Eng.*, **9**: 304-307.

Razali, A.S., Supian, F.L., Bakar, S.A., Richardson, T.H. & Azahari, N.A. (2015b). The properties of carbon nanotube on novel calixarene thin film. *Int. J. Nano. Mat.*, **8**: 39-45.

Richardson, T.H., Brook, R.A., Davis, F. & Hunter, C.A. (2006). The NO_2 gas sensing properties of calixarene/porphyrin mixed LB films. *Coll. Surf. A: Physicochem. Eng. Asp.*, **284**: 320-325.

Roberts, G.G. (1990). *Langmuir-Blodgett Films*. Plenum Press, New York.

Sen, S., Capan, R., Ozbek, Z., Ozel, M.E., Stanciu, G.A. & Davis, F. (2019). Langmuir–Blodgett film properties of based on calix[4]resorcinarene and the detection of those against volatile organic compounds. *Applied Physics A*, **125**.

Shorthill, B.J. & Glass, T.E. (2001). Naphthalene-based calixarenes: unusual regiochemistry of a Friedel-Crafts alkylation. *Org. Lett.*, **3**: 577-579.

Shorthill, B.J., Granucci, R.G., Powell, D.R. & Glass, T.E. (2002). Synthesis of 3,5- and 3,6-linked calix[n]naphthalenes. *J. Org. Chem.*, **67**: 904–909.

Steed, J.W. & Gale, P.A. (2012). *Supramolecular Chemistry: From Molecules to Nanomaterials*. Wiley, New Jersey.

Supian, F.L. (2010). *Sensing Interactions Within Nanoscale Calixarene and Polysiloxane Langmuir-Blodgett Films*. PhD Thesis, Department of Physics and Astronomy, University of Sheffield, Sheffield.

Supian, F.L., Richardson, T.H., Deasy, M., Kelleher, F., Ward, J.P. & McKee, V. (2010). Interaction between Langmuir and Langmuir-Blodgett films of two calix[4]arenes with aqueous copper and lithium ions. *Langmuir*, **26**: 10906-10912.

Supian, F.L., Bakar, S.A., Azahari, N.A. & Richardson, T.H. (2013). Characteristics of a novel calix[8]arene modified with carbon nanotubes thin films for metal cations detection. *AIP Conf. Proc.*, **1528**: 260-265.

Supian F.L., Lim, D.C.K. & Razali A.S. (2017). Conductivity comparison of calix[8arene]-MWCNTs through spin coating technique. *Sains Malaysiana*, **46**: 91-96.

Supian, F.L., Lim, D.C.K., Azmi, M.S.M. & Darus, M.M. (2019). Surface potential and temperature effect studies of homo- and copolysiloxanes through Langmuir Blodgett technique. *Defence S&T Tech. Bull.*, **12**: 218-225.

Tabakci, B. & Yilmaz, A. (2014). Amine-derivatized calix[4]arenes for sensitive extraction of cupric ion and formation of amine radical cation. *J. Mol. Struc.*, **1075**: 96-102.

Vigalok, A. & Swager, T.M. (2002). Conducting polymers of tungsten(VI)-oxo calixarene: Intercalation of neutral organic guests. *Adv. Mat.*, **14**: 368-371.

Wang, L., Wang, X., Shi, G., Cheng, P. & Ding, Y. (2012). Thiacalixarene covalently functionalized multiwalled carbon nanotubes as chemically modified electrode material for detection of ultratrace Pb^{2+} ions. *Ana. Chem.*, **84**: 10560-10567.

Wang, N., Chang, P.R., Zheng, P. & Ma, X. (2015). Carbon nanotube-cyclodextrin adducts for electrochemical recognition of tartaric acid. *Dia. Rel. Mater.*, **55**: 117-122.

Weis, M., Janicek, R., Cirak, J. & Hianik, T. (2007). Study of the calix 4 resorcinarenedopamine interactions in monolayers by measurement of pressure-area isotherms and Maxwell displacement currents. *J. Phy. Chem. B*, **111**: 10626-10631.

Yadav, B.C., Yadav, R.C. & Dwivedi, P.K. (2010). Sol-gel processed (Mg–Zn–Ti) oxide nano-composite film deposited on prism base as adopt-electronic humidity sensor. *Sens. Act. B: Chem.*, **148**: 413-419.

Yang, Q., Qin, X., Yan, C. & Zhu, X. (2015). A novel fluorescent chemosensor for safranine T based on calixarene-1,3-diacyl hydrazone. *Sens. Act. B: Chem.*, **212**: 183-189.

Zhu, C. & Fang, L. (2014). Mingling electronic chemical sensors with supramolecular host-guest chemistry. *Cur. Org. Chem.*, **18**: 1957-1964.

REVIEW OF ANALYTICAL METHODS FOR DETERMINATION OF PARABENS IN COSMETIC PRODUCTS

Faris Rudi

Science and Technology Research Institute for Defence (STRIDE), Ministry of Defence, Malaysia
Chemistry Department, Faculty of Science, University of Malaya (UM), Malaysia

Email: faris.rudi@stride.gov.my

ABSTRACT

Parabens are well known preservatives that are widely used in cosmetic and personal care (CPC) products as they have a broad spectrum of antimicrobial properties, as well as being non-sensitising and non-irritating. Over the years, parabens have been considered to be a relatively safe compound. However, a few studies found that parabens might have estrogenic properties and there is an ongoing debate regarding the potential risk of cancer from consuming this product. This paper presents an overview of sample preparation methods, as well as instrumental analysis of current and past researches on the determination of parabens in cosmetic products. It first reviews on sample preparation methods that effectively eliminate complex matrices in cosmetics products, including liquid-liquid extraction (LLE), solid phase extraction (SPE), solid phase microextraction (SPME), dispersive liquid-liquid microextraction (DLLME) and supercritical fluid extraction (SFE). Then, the analytical techniques for the analysis are reported, mainly using high-performance liquid chromatography (HPLC), gas chromatography (GC) and electrophoresis. Research gaps and suggestions for future studies on the detection of parabens in cosmetic products are also given.

Keywords: *Parabens; cosmetic and personal care (CPC) products; complex matrices; sample preparation; analytical techniques.*

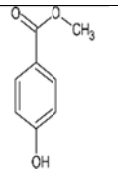
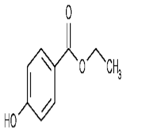
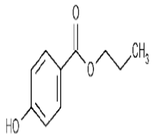
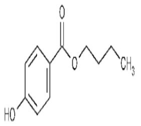
1. INTRODUCTION

Parabens is a group of para-hydroxybenzoic acid esters that is made up of methylparaben (MP), ethylparaben (EP), propylparaben (PP), butylparaben (BP), isobutylparaben (iBP), isopropylparaben (iPP), benzylparaben (BzP) and their sodium salts (Chao *et al*., 2020). Parabens is typically added to cosmetic and personal care (CPC) products as it has broad antimicrobial properties. Furthermore, parabens have characteristics of being odourless, tasteless, inexpensive, stable over a wide range of pH, non-decolourising and adequately soluble in water, which explains the wide range of applications of parabens as preservative (Cabaleiro *et al*., 2014a). While parabens have been known to have low toxicity effect, the increase of alkyl chain in the ester group of parabens contributes to increase in toxicity (Kolatorova *et al*., 2018). In addition, long-chain compounds are not commonly applied to cosmetic products due to lack of solubility (Matwiejczuk *et al*., 2020). The physiochemical properties of the most common parabens used in cosmetic products are shown in Table 1.

Routledge *et al*. (1998) found that a group of parabens, namely MP, EP, PP and BP, have mild estrogenic effects for humans and wildlife. Libei *et al*. (2016) further studied the estrogenic effects of parabens. In addition, findings of parabens in human breast tumours have raised concern on the toxicity of this preservative that leads to the risk of breast cancer (Darbre *et al*., 2004, 2013; Amin *et al*., 2019).

Presently, the European Union legislation has set limits on parabens and their salts for up to 0.4% (w/w) for MP and EP, 0.14% (w/w) for the sum of PP and BP, and 0.8% (w/w) for total parabens concentration in CPC products. Hence, the development of simple and reliable analytical methods to determine parabens is crucial to ensure that CPC products in markets contain parabens values that align with the regulation. A review by Wang & Liu (2007) summarised the sample preparation and instrumental techniques used in determination of parabens in cosmetic products between 1980 to 2007. In addition, Cabaleiro *et al.* (2014) summarised sample preparation methods for up to 2014.

Table 1: Physiochemical properties of the most common parabens used in cosmetics (Matwiejczuk *et al.*, 2020).

SPECIFICATION	MP	EP	PP	BP
Chemical formula				
Molecular weight, g/mol	152.16	166.18	180.21	194.23
Cas no	99-76-3	120-47-8	94-13-3	94-26-8
pKa value (indicate the strength of an acid)	8.17	8.22	8.35	8.37
Solubility at 25% (m/m)				
Water	0.25	0.11	0.04	0.02
Propylene glycol	26	20	29	49
Ethanol	32	41	50	68
Melting point, °C	131	117	97	68
Boiling point, °C	275	297	-	-

The objective of this paper is to review the analytical methods that are used for sample preparation and instrumental analysis for determination of parabens in cosmetic products. In addition, it emphasises on the research gaps and suggestions for future studies.

2. TECHNIQUES USED FOR SAMPLE PREPARATION FOR PARABENS DETERMINATION

Generally, cosmetic products contain very complex matrices that demand the use of long steps of sample preparation. There are a number of methods for sample preparation for parabens determination based on simple dilution and homogenisation with a suitable organic solvent, such as methanol (Grzeskowiak *et al.*, 2016a), ethanol (Huang *et al.*, 2013) and propanol (Memon *et al.*, 2005). However, the extracted samples still contain a lot of matrices that might affect chromatography sensitivity. Recently, a number of methods have been developed to eliminate the matrix effect. Liquid-liquid extraction (LLE) is a method with single or combination of different solvents to extract the parabens that have been used (Gabriella *et al.*, 2016a; Nourolhoda *et al.*, 2019a). LLE provides simplicity for the extraction process, but there are some disadvantages for this technique, such as emulsion formation and being environmentally unfriendly due to high consumption of organic solvent (Cabaleiro *et al.*, 2013). Solid phase extraction (SPE) (Figure 1) is an alternative for LLE that is used for the isolation and concentration of analytes from liquid flowing sample stream by their transfer to and retention (sorption) in a solid phase. This solid phase is then isolated from the sample and the analytes are recovered using liquid elution (Poole *et al.*, 2012). Among the SPE cartridges used are C18 (Shen *et al.*, 2007a; Uysal *et al.*, 2008a) and C8 (Han *et al.*, 2008a). However, the drawback of this technique is that it requires a lot of solvent for extraction (Cabaleiro *et al.*, 2014b).

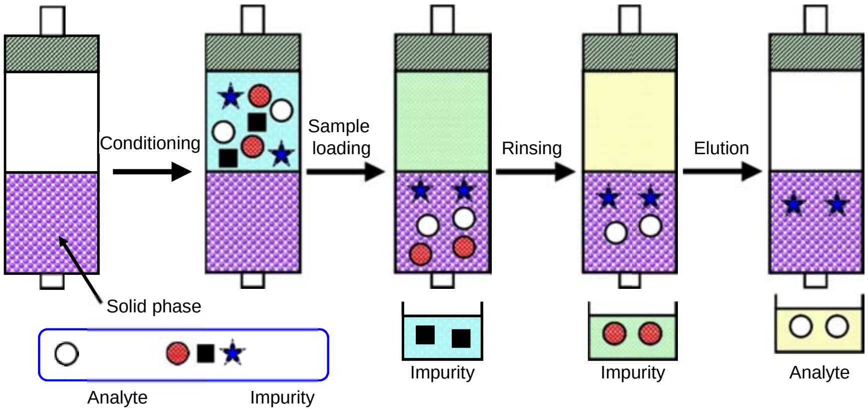


Figure 1: Schematic diagram of the SPE technique (Hiroyuki, 2017).

An alternative for SPE is solid-phase microextraction (SPME) (Figure 2), which reduces solvent consumption and operation time. SPME has two different techniques, which are headspace and direct SPME. For headspace SPME, the sample needs to be brought to equilibrium and the fibre exposed to the headspace of the sample for a period of time. Meanwhile, direct SPME involves immersing the fibre directly into the sample matrix. Direct SPME has been applied mostly for the non-volatile character of parabens (Fei *et al.*, 2011a), while we found only one published application of headspace SPME for parabens determination (Yang *et al.*, 2010a). The drawback of SPME is the limited number of stationary phases as it covers mostly non-volatile compounds, which poses a problem for volatile compounds (Ocana *et al.*, 2015a).

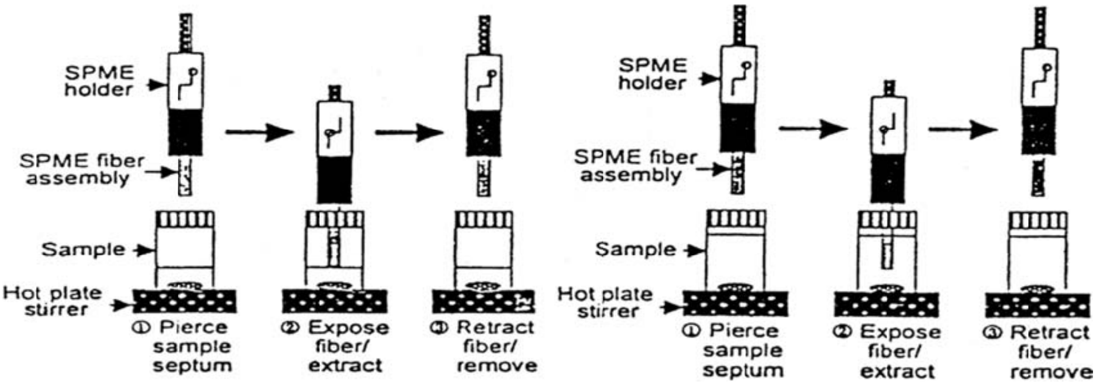


Figure 2: Schematic diagrams of headspace (left) and direct SPME (right) (Jatinder *et al.*, 2005).

Dispersive liquid-liquid microextraction (DLLME) is an alternative technique to SPE and SPME, in which sorbent is dispersed into the sample matrix (or its extract), with the close contact that is obtained between the sorbent particles and the analyte favouring the kinetics of the sorption (Alberto *et al*., 2019). DLLME (Figure 3) involves the extraction and simultaneous concentration of the desired analyte from an aqueous solution using a small amount of organic solvent. It is an extraction technique that offers higher efficiency with only microvolume of solvent (Hongmin *et al*., 2014; Hwang *et al*., 2018). Dyia *et al.* (2018) used DLLME for extraction of methyl parabens from cosmetic products. Additionally, supercritical fluid extraction (SFE) is another extraction method for parabens determination that does not need any sample pretreatment (Lee *et al*., 2006a). SFE is well known as green technology by using CO_2 as the extraction solvent (Figure 4). Despite the benefit of SFE requiring no solvent to be used at all, it needs to be operated at high pressure (Puah *et al*., 2005).

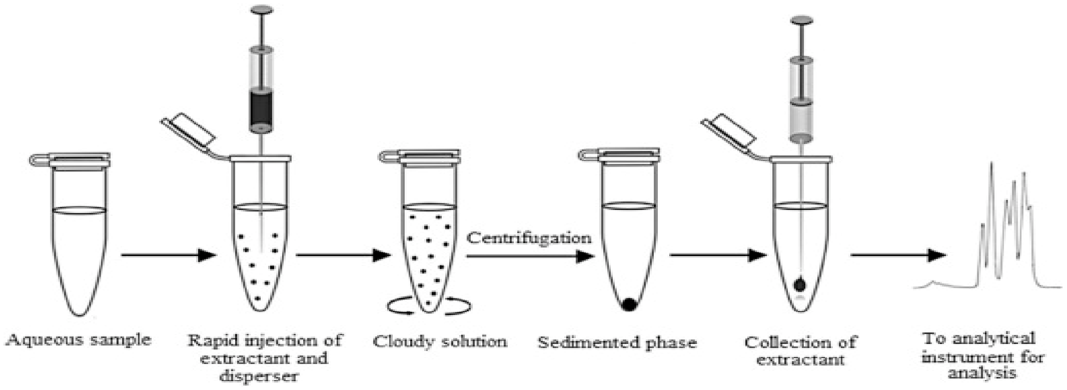


Figure 3: Steps in the DLLME technique (Ahmad et *al*., 2015).

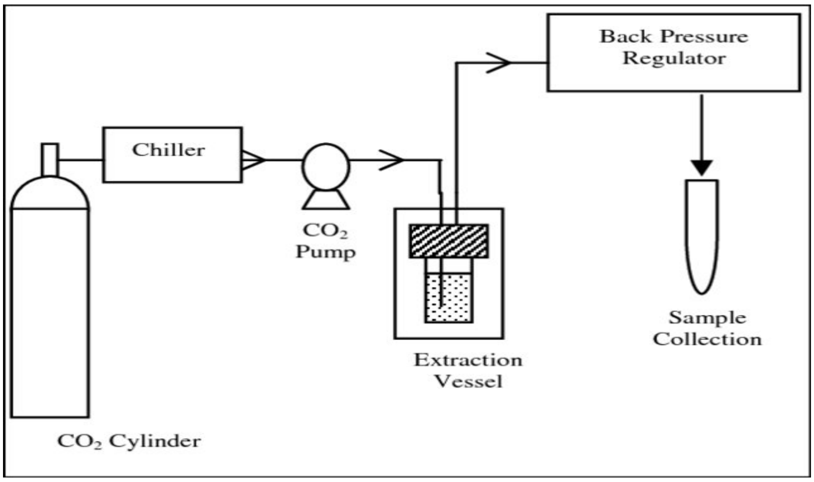


Figure 4: Simplified flow diagram of SFE (Puah *et al*., 2005).

3. SUMMARY OF ANALYTICAL TECHNIQUES USED AND RESULTS FROM LITERATURE

The last few years have seen a number of determination and analytical techniques being developed for parabens determination. High-performance liquid chromatography (HPLC) is the most common analytical technique used, followed by gas chromatography (GC) and electrophoresis.

HPLC is a specific form of column chromatography that is generally used in chemical analysis to identify, separate and quantify analytes. It operates at high pressure that pumps the sample (analyte) dissolved in a solvent (mobile phase) through a column with an immobilised stationary phase (Figure 5). The properties of the sample, mobile phase and stationary phase determine the retention time of the analyte, whereby analytes that have strong interactions with the stationary phase will elute for a longer period (Olga & Karin, 2017). HPLC allows for samples to be analysed without any derivatisation step since it is suitable for non-volatile compounds. This advantage contributes to the determination of parabens in cosmetic products in a short time as compared to other analytical techniques. Table 2 sums up the techniques and results for determination of parabens in cosmetic products using HPLC. Until now, HPLC ultraviolet detector (HPLC-UV) is the most used analytical technique for this type of parabens determination. Memon *et al.* (2005) successfully detected parabens with range of detection limit between 25-250 ng mL^{-1} using this technique. In addition, other detectors used for this technique are fluorescent detector (HPLC-FD) (Grzeskowiak *et al*., 2016b), diode array detector (HPLC-DAD) (Fei *et al*., 2011b; Nourolhoda *et al*., 2019b), mass spectrometer single quadrupole (HPLC-MS) (Lee *et al*., 2006b) and mass spectrometer triple quadrupole (HPLC-MS/MS) (Gabriella *et al*., 2016b).

Gas chromatography (GC) is a separation technique used to isolate volatile components of analytes in the mixture based on the differences in the mode of partitioning between mobile and stationary phases. The injected sample is then vaporised and transferred to a column by a mobile phase (Figure 6). The column that is packed with a finely divided solid or coated film is a stationary phase. The analyte interaction that occurs at the column results in the separation of the analyte of interest. The analyte then is eluted to the detector for signal generation (Rahman et al., 2015). Table 3 summarises the GC techniques with flame ionisation detector (GC-FID) (Hongmin *et al*., 2014; Dyia *et al*., 2018) and mass spectrometer detector (GC-MS) (Shen *et al*., 2007b; Yang *et al*., 2010b). Generally, GC-MS and HPLC-MS offer the same advantages, such as identification of analytes with low detection limit, which allows for the study of very low level of parabens in cosmetic products. GC-MS has some advantages as compared to HPLC-MS, in particular higher resolution, lower solvent of waste production and lower cost. However, GC-MS requires longer sample preparation time as the sample needs to be derivatised, which also raises the possibility of errors occurring during the derivatisation (Ocana *et al*., 2015b).

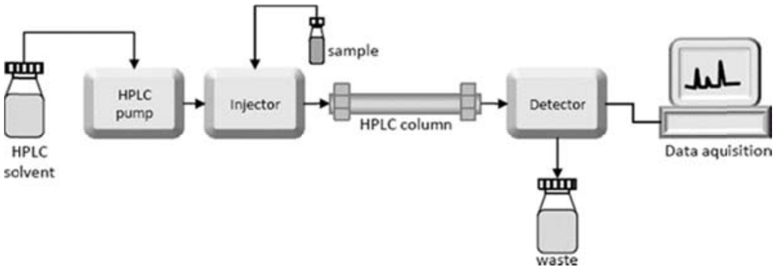


Figure 5: Schematic diagram of HPLC (Sylwester, 2013).

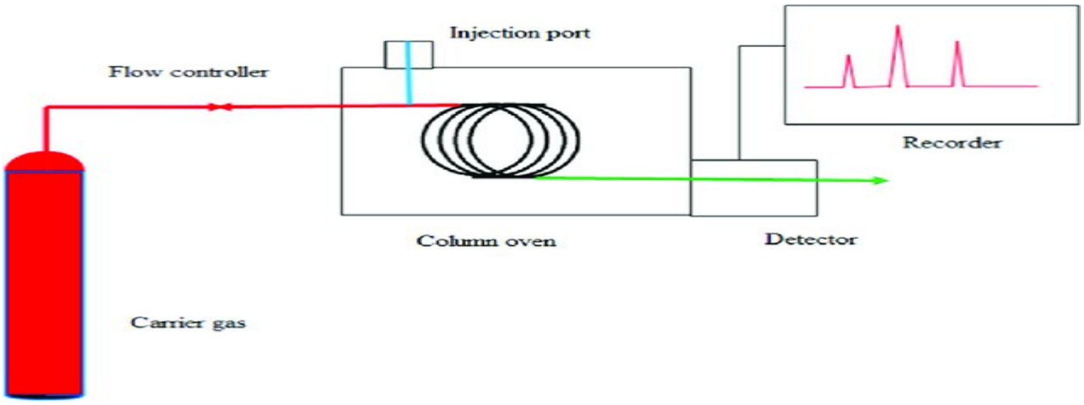


Figure 6: Schematic diagram of GC (Mallaiah, 2018)

Table 2: HPLC techniques for the determination of parabens in cosmetic products.

Technique	Parabens	Sample	Extraction method	Mobile phase	Stationary phase	Detection limits	Reference
HPLC-FD	MP, EP, PP, BP	Tonics & micellar water	LE	Gradient methanol / water	C18	0.007-0.014 µg mL^{-1}	Grzeskowiak *et al*. (2016)
HPLC-UV	MP, EP, PP, BP	Shampoo, hand lotion, creams & bath foam	LE	Isocratic water / propanol	C18	25-250 ng mL^{-1}	Memon *et al*. (2005)
HPLC-DAD	MP, EP, PP, BP	Sunscreens, lotions & creams	SPME	Acetonitrile /water	C18	0.12-0.15 µg mL^{-1}	Fei *et al*. (2011)
HPLC-DAD	MP, EP, PP, BP	Toothpaste & mouthwash	LE	Gradient methanol / water	C18	0.0004-0.001 µg mL^{-1}	Nourolhoda *et al*. (2019)
HPLC-MS/MS	MP, PP	Serum	LLE	Gradient methanol / acetonitrile	C18	1-20 ng mL^{-1}	Gabriella *et al*. (2016)
HPLC-MS	MP, EP, PP, BP	Lanoline cream, skin milk & cream	SFE	Gradient methanol / water	C18	4.7-19.3 ng/g	Lee *et al*. (2006)

Table 3: GC techniques for the determination of parabens in cosmetic products.

Techniques	Parabens	Sample	Extraction method	Detection limits	Reference
GC-FID	MP	Sunscreens	DLLME	0.082 µg mL^{-1}	Dyia *et al*. (2018)
GC-FID	MP, EP, PP, BP	Face masks, moisture cream, face cream & hair cream	DLLME	2.0 – 9.5 µg g^{-1}	Hongmin *et al*. (2014)
GC-MS	MP, EP, PP, BP	Homemade cream	SPME	0.001-0.015 µg L^{-1}	Yang *et al*. (2010)
GC-MS	MP, EP, PP, BP	Cosmetics	SPE	0.1-5.0 µg Kg^{-1}	Shen *et al*. (2007)

Electrophoresis refers to the movement of particles through a stationary fluid under the influence of an electric field. The principle of electrophoresis is the existence of charge separation between the surface of particle and fluid that surrounding it. An applied electric field acts on the resulting charge density, resulting in the particles migrating and the fluid around particles to flow (Scott & Curtis, 2000). Electrophoresis provides a simple, fast and sensitive technique for parabens determination in cosmetic products. However, for BP, it can only be separated by varying the pH value or methanol percentage in the buffer solution (Labat *et al*., 2000). There are two techniques for electrophoresis, which are capillary electrophoresis (CE) and micellar electrokinetic chromatograph (MEC). The set up for CE is shown in Figure 7. A microcapillary is stretched between two reservoirs that are filled with buffer solution. The analyte is introduced at one end of the capillary in the form of a plug that travels down the capillary due to electrophoretic mobility. The difference in the electrophoretic mobility of the analyte causes it to separate and travel to the detector (Sandip, 2005). For MEC (Figure 8), the surfactant is added to the buffer solution with concentration above the critical micellar concentration. The separation is based on the differences in the distribution constant between two phases migrating at different velocities due to the electrokinetic effects, resulting in micelles being formed (Hancu *et al*., 2012). Table 4 shows a summary of CE and MEC techniques used for determination of parabens in cosmetic products.

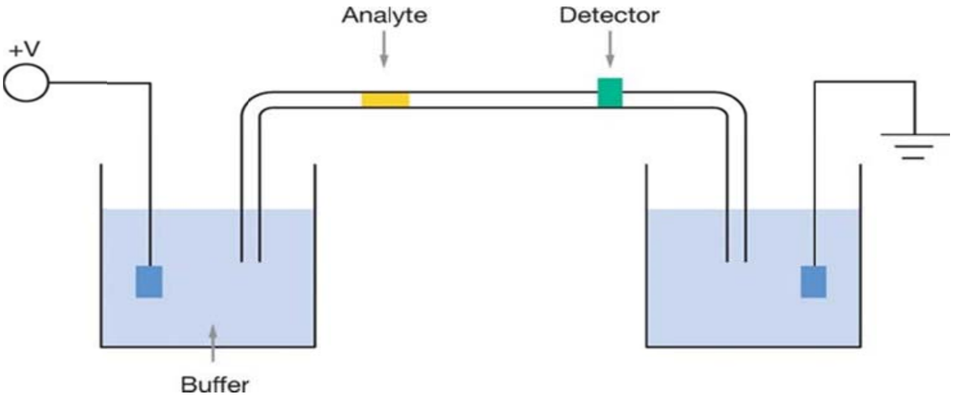


Figure 7: Schematic diagram of CE (Sandip, 2005).

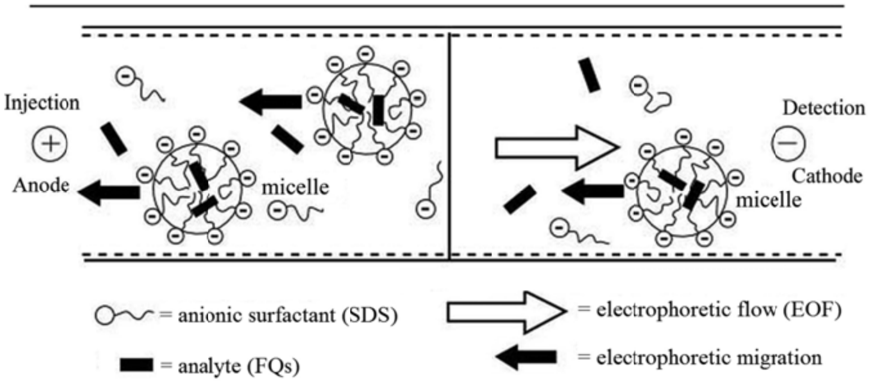

Injection
Anode
micelle
Detection
Cathode
micelle
= anionic surfactant (SDS)
= analyte (FQs)
= electrophoretic flow (EOF)
= electrophoretic migration

Figure 8: Schematic diagram of MEC (Hancu *et al*., 2012).

Table 4: Electrophoresis techniques for the determination of parabens in cosmetic products.

Techniques	Parabens	Sample	Extraction method	Carrier buffer	Voltage	Detection limits	Reference
CE	MP, EP, PP, BP	Shampoo, hair dyes, tooth paste & gel	SPE	20 mM borate, 10 % (v/v) MeOH	20 kV	1.42-2.86 µM	Uysal *et al.* (2008)
MEC	MP, EP, PP, BP	Cream, lotion, moisturiser	LE	1.0 mM phosphate buffer, 10% (v/v) ethanol	12.5 kV	0.48-1.52 µg mL^{-1}	Huang *et al.* (2013)
MEC	MP, EP, PP, BP	Cream, lotion, gel	C8	20 mM sodium tetraborate (pH 9.3, 100mmol/L SDS)	15 kV	0.07-0.1 µg mL^{-1}	Han *et al.* (2008)

4. CONCLUSION

This paper reviewed the sample preparation methods and analytical techniques for determination of parabens in cosmetic products. A number of sample preparation methods have been developed to effectively eliminate complex matrices in cosmetics products, including SPE, SPME, DLLME and SFE. SPE techniques have the potential to be used as an extraction technique as it offers fast and effective results. Meanwhile, SPME offers low solvent use and fast extraction time. DLLME, on the other hand, has the potential to be chosen as an effective technique as it is suitable for very low limit of parabens in cosmetic products. Finally, SFE provides an extraction technique without any use of solvent. However, it must operate at high pressure.

HPLC, GC and electrophoresis are among the analytical techniques used for parabens determination in cosmetic products. HPLC offers sample treatment without any derivatisation steps and is suitable for most non-volatile compounds. Hence, it is the commonly used analytical technique for parabens determination in cosmetic products. GC, on the other hand, is more suitable for volatile compounds, as well as provides simplicity and fast analysis time as compared to HPLC. Finally, electrophoresis offers fast and simple analysis for parabens determination in cosmetic products.

Suggestions for future studies to improve the sensitivity of parabens determination in cosmetics are:

(1) Possibility to apply nanomaterials as an extraction technique that allows for matrix clean up without decrease in sensitivity.
(2) Developing an electrochemical sensor for selective detection of parabens.

REFERENCES

Ahmad, W., Al-Sibaai, A.A., Bashammakh, A.S., Alwael, H. & El-Shahawi, M.S. (2015). Recent advances in dispersive liquid-liquid microextraction for pesticide analysis. *Trends Anal. Chem*, **72:** 18-192.

Alberto, C., Soledad, C. & Rafael, L. (2019). Dispersive micro-solid phase extraction. *Trends Anal. Chem*, **112:** 226-233.

Amin, M.M., Maryam, T., Asfane, C., Elham, A., Majid, H., Karim, E., Roya, K., Sedigheh, K. & Marjan, M. (2019). *Biomed Environ Sci*, **32:** 893-904.

Cabaleiro, N., Calle, I. D., Bendicho, C. & Lavilla, I. (2013). Current trends in liquid-liquid and solid-liquid extraction for cosmetic analysis: a review. *Trends Anal. Chem.*, **5:** 323-340.

Cabaleiro, N., Calle, I. D., Bendicho, C. & Lavilla, I. (2014). An overview of sample preparation for the determination of parabens in cosmetics. *Trends Anal. Chem.*, **57:** 34-46.

Chao, L., Xinyi, C., Yi, C. & Chunyang, L. (2020). Parabens concentrations in human fingernail and its association with personal care product use. *Ecotoxicol. Environ. Saf.*, **202:** 110933.

Darbre, P.D. (2013). Underarm cosmetics and breast cancer. *J. Appl. Toxicol.*, **23:** 89-95.

Darbre, P.D., Aljarrah, A., Miller, W.R., Coldham, N.G., Sauer, M. J. & Pope, G.S. (2004). Concentrations of parabens in human breast tumours. *J. Appl. Toxicol.*, **24:** 5-13.

Doron, S., Friedman, M., Falach, M., Sadovnic, E. & Zvia, H. (2001). Antibacterial effect of parabens against planktonic and biofilm streptococcus sobrinus. *Int. J. Antimicrob. Agents*, **18:** 575-578.

Dyia, S.M., Nurul, A.H., Shakirah, H. & Izhan, W.N. (2018). Extraction of methylparaben in cosmetic using dispersive liquid-liquid microextraction based on solidification of floating organic drop coupled with gas chromatography flame ionization detector. *Malayisan J. Anal. Sci*, **21:** 1289-1298.

European Commission. (2014). *Commission regulation (EU) No. 1004/2014 of 18 September 2014 amending annex V to regulation (EC) No. 1223/2009 of the European Parliament and of the Council on Cosmetic Products (Text with EEA Relevance).* Publication Office of the EU, Brussels, Belgium.

Fei, T., Haifeng, L., Ding, M., Ito, M. & Lin, J.M. (2011). Determination of parabens in cosmetic products by solid-phase microextraction of poly (ethylene glycol) diacrylate thin film on fibers and ultra-high-speed liquid chromatography with diode array detector. *J. Sep. Sci*, **34:** 1599-1606.

Gabriella, P.T., Nayara, K.S.S., Ana, C.A. & Isarita, M. (2016). Determination of parabens in serum by liquid chromatography tandem mass spectrometry:Correlation with lipstick use. *Regul. Toxicol. Pharmacol*, **79:** 42-48.

Golden, R., Gandy, J. & Vollmer, G. (2005). A review of the endocrine activity of parabens and implications for potential risks to human health. *Crit. Rev. Toxicol*, **5:** 435-458.

Grzeskowiak, A.Z., Werner, J., Skowron, M.J. & Goslinska, B.C. (2016). Determination of parabens in cosmetic products using high performance liquid chromatography with fluorescene detection. *Anal. Methods*, **8:** 3903-3909.

Han, F., He, Y.Z. & Yu, C.Z. (2008). On-line pretreatment and determination of parabens in cosmetic products by combination of flow injection analysis, solid-phase extraction and micellar electrokinetic chromatography. *Talanta*, **5:** 1371-1377.

Hancu, G., Rusu, A., Simon B., Boia, G. & Gyeresi, A. (2012). Simultaneous separation of ciprofloxacin, norfloxacin and ofloxacin by micellar electrokinetic chromatography. *J. Braz. Chem. Soc.*, **10:** 1889-1894.

Hiroyuki, K. (Eds.) (2017). *Liquid Chromatography, 2nd Ed.* Elsevier, Amsterdam, Netherlands.

Hongmin, W., Jinjuan, Y., Hanqi, Z. & Yuhua, S. (2014). Ultrasonic nebulization extraction assisted dispersice liquid-liquid microxtraction followed by gas chromatography for the simultaneous determination of six parabens in cosmetic products. *J. Sep. Sci.*, **37**: 2349-2356.

Huang, J.Q., Hu, C.C. & Chiu, T.C. (2013). Determination of seven preservatives in cosmetic products by micellar electrokinetic chromatography. *Int. J. Cosmet. Sci.;*, **35:** 346-353.

Hwang, T.Y., Kin, C.M. & Shing, W.L. (2018). A review on extraction solvents in the dispersive liquid-liquid microextraction. *Malaysian J. Anal. Sci.*, **22:** 166-174.

Jatinder, S.A., Ashok, K.M., Varinder, K. & Phillipee, S.K. (Eds). (2005). *Critical Review in Analytical Chemistry*. Taylor & Francis, Oxfordshire, UK

Kolatorova, L., Vitku, J., Hampl, R., Adamcova, K., Skodova, T., Simkova, M., Parizek, A., Starka, L. & Duskova, M. (2018). Exposure to bisphenols and parabens during pregnancy and relation to steroid changes. *Environ. Res.*, **163:** 115-122.

Labat, L., Kummer, E., Dallet, P. & Dubost, J.P, (2000). Comparison of high-performance liquid chromatography and capillary zone electrophoresis for the determination of parabens in cosmetic product. *J. Pharm. Biomed. Anal.*, **23:** 763-769.

Lee, M.R., Lin, C.Y. & Tsai, T.F. (2006). Simultaneous analysis of antioxidants and preservatives in cosmetic by supercritical fluid extraction combined with liquid chromatography-mass spectrometry. *J. Chromatogr. A*, **1120**: 244-251.

Libei, S., Tong, Y., Jilong, G., Zhaobin, Z., Ying, H., Xuan, X., Yingli, S., Han, X., Junyu, L., Desheng, Z., Linlin, S. & Jun, L. (2016). The estrogenicity of methylparaben and ethylparaben at doses close to the acceptable daily intake in immature Sprague-Dawley rats. *Sci. Rep.,* **6:** 25173.

Mallaiah, M. & Venkat, R.G. (2018). Data on acetic acid-methanol-methyl acetate-water mixture analysised by dual packed column gas chromatography. *Data Brief,* **18:** 947-960.

Matwiejczuk, N., Gallicka, A. & Brzoska, M.M. (2020). Review of the safety of application of cosmetic products containing parabens. *J. Appl. Toxicol.*, **40:** 176-210.

Memon, N., Bhanger, M.I. & Khuhawer, M.Y. (2005). Determination of preservatives in cosmetics and food samples by micellar liquid chromatography. *J. Sep. Sci*, **28:** 635-638.

Nourolhoda, R. & Zarrin, E. (2019). Curcumin loaded magnetic graphene oxide solid-phase extraction for the determination of parabens in toothpaste and mouthwash coupled with high performance liquid chromatography. *Microchem. J.*, **148:** 616-625.

Ocana-Gonzalez, J.A., Villar-Navaro, M., Ramos-Payan, M., Fernandez-Torres, R. & Bello-Lopez, M.A. (2015). New developments in the extraction and determination of parabens in cosmetic and environmental samples. A review. *Anal. Chim. Acta,* **858:** 1-15.

Olga, E.P. & Karin, S. (2017). High performance liquid chromatography (HPLC)-based detection and quantification of cellular c-di-GMP. *Methods Mol Biol*, **1657: 33**-43.

Poole, C.F. & Poole, S.K. (Eds). (2012). *Comprehensive sampling and sample preparation*. Academic Press, Cambridge, Massachusetts.

Puah, C.W., Choo, Y.M., Ma, A.N. & Cheng, H.C. (2005). Supercritical fluid extraction of palm carotenoids. *Am. J. Environ. Sci,* **4:** 264-269.

Rahman, M.M., Abd El-Aty, A.M., Choi, J.H., Shin, H.C., Shin, S.C. & Shim, J.H. (Eds) (2015). *Analytical Separation Science.* Wiley-VCH Verlag GmbH & Co. KGaA.

Routledge, E.J., Parker, J., Odum, J., Ashby, J. & Sumpter, J.P. (1998). Some alkyl hydroxyl benzoate preservatives (parabens) are estrogenic. *Toxicol. Appl. Pharmacol.*, **153:** 12-19.

Sandip Ghosal. (2005). Electrokinetic flow and dispersion in capillary electrophoresis. *Annu. Rev. Fluid. Mech.*, **38:** 309-338.

Scott, R. R. & Curtis, A.M. (2000). Electrophoresis techniques. *Sep. Purif. Methods,* **29:** 129-148.

Shen, H., Ying, L., Cao, Y., Pan, G. & Zhou, L. (2007). Simultaneous determination of phthalates and parabens in cosmetic products by gas chromatography/mass spectrometry coupled with solid phase extraction. *Sepu,* **25:** 272-275.

Sylwester Czaplicki. (Eds). (2013). *Column chromatography*. IntechOpen,

Uysal, U.D. & Gurayb, T. (2008). Determination of parabens in pharmaceutical and cosmetic products by capillary electrophoresis. *J. Anal. Chem*, **63:** 982-986.

Wang, P. & Liu, Y. (2007). Cosmetic preservatives and analysis methods used in China. J. *Environ. Health,* **24:** 557-559.

Yang, T.J., Tsei, F.J., Chen, C.Y., Cherng, T.C. & Lee, M.R. (2010). Determination of additives in cosmetic by supercritical fluid extraction on-line headspace solid-phase microextraction combined with gas chromatography-mass spectrometry. *Anal. Chim Acta,* **668:** 188-194.

VIBRATION ANALYSIS OF FLEXIBLE COUPLING BY CONSIDERING SHAFT MISALIGNMENT

Yogeswaran Sinnasamy*, Noor Aishah Sa'at, Hasril Nain & Khairul Anuar Ahmad

Science & Technology Research Institute for Defence (STRIDE), Ministry of Defence, Malaysia

*Email: yoges.sinnasamy@stride.gov.my

ABSTRACT

In reciprocating-rotating machines, flexible couplings are widely used in many applications, such as marine gear box and diesel engine. Vibration monitoring is highly capable of detecting any abnormalities on flexible couplings before failure occurs. In this paper, vibration measurement is utilised on a rotor dynamic test apparatus to predict vibration spectrums for various motor revolutions. A self-designed simplified type flexible coupling was used in the experiments. Vibration spectrums captured on two units of healthy rolling element bearings in three different directions at various speeds were analysed, with the measurements found to have high level of accuracy. Based on the procedure developed in this experimental work, condition monitoring (CM) for rotor-flexible coupling-bearing systems can be developed in the future.

Keywords: *Flexible coupling; vibration monitoring; fundamental frequency; vibration spectrums; accuracy.*

1. INTRODUCTION

Vibration is one of the most common parameters that are used for monitoring the health condition of equipment and certain type of machineries based on movement of associated components and supporting bases as per operational requirements. Condition monitoring (CM) of bearing faults is typically implemented using experimental based vibration analysis data (Mehdi *et al.,* 2011; Desavale *et al.*, 2013; Vishwakarma *et al.*, 2017; Malla & Panigrahi, 2019). Gani & Salami (2004), Vishwakarma *et al*. (2017) and Malla & Panigrahi (2019) demonstrated the various types of common rotating machinery faults that can be detected using vibration analysis.

CM is defined as the continuous evaluation of the health of a plant and its equipment throughout its service life. It is important to be able to detect faults while they are still developing. This is called incipient failure detection (Li *et al*., 2012; Elamin, 2013; Shi *et al*., 2020). Meanwhile, failure is the termination of the ability to perform the required function, and fault is defined as a situation that exists after a failure (Elamin *et al*., 2010). Incipient detection of diesel engine failures provides a safe operating environment and thus, it is becoming increasingly important to use comprehensive CM schemes for continuous assessment of the combustion and mechanical conditions of reciprocating machineries (Gu *et al*., 2006).

Couplings are machine parts that perform the connection between two consecutive elements of a kinematic chain. The coupling element transmits the torque between rigid coupling components that are situated concentrically one inside the other. Coupling are widely used in naval and merchant ships to support main and auxiliary propulsion systems. In rotating machineries, these couplings are subject to unbalanced forces generated in machines (Mihaela & Silviu, 2014).

A coupling fault can result in unscheduled maintenance and in extreme cases, plant shutdown. These faults may have developed during longer runs of coupling under certain conditions. Severe vibrations of coupling can even cause the entire system to function incorrectly and subsequently result in downtime of the system and financial loss to the user (Nistane & Harsha, 2016).

Shweta & Tuljapure (2015) discussed on the causes of flexible coupling failures, such as human errors, corrosion, wear, fatigue, hardware failure and shaft failure. The majority of the problems described are caused by vibrations, and the phenomena of vibration are complex as they are misunderstood. Misalignment is the principle source of most vibrations, whereby excessive misalignment can cause severe damage to flexible couplings. There are several symptoms indicating misalignment, such as excessive radial and axial vibrations, high casing temperature at or near the bearings, high temperature of discharge oil, excessive amount of oil leakage at the bearing seal, as well as loose foundation bolts, and broken or loose coupling bolts.

Pallavi (2014) conducted an experimental investigation to detect unique vibration signatures for bent shaft. Experimental studies were performed on a rotor dynamic test apparatus to predict the vibration spectrum for shaft bending when a shaft is bent at the coupling end. A four pin type flexible flange coupling was used in the experiment. The rotor shaft vibrations were measured using fast Fourier transform (FFT) analyser. The acquired spectrums were compared with available literature for confirmation of experiment results. The results showed that bent shaft could be characterised primarily using 1X and 2X shaft running speed.

Khot & Pallavi (2015) conducted experimental investigation of faults such as parallel and angular misalignments with the help of FFT analyser. For identification of faults, frequency spectrums were obtained using an experimental setup developed for rotor system. A simulation study of misalignment effect in rotary system was also conducted using ANSYS. The results were compared to validate simulation study with experimental results, which were found to be in close agreement.

Grega *et al.* (2016) demonstrated the effects of different types of flexible couplings on the size of vibration in a gearbox that forms part of a mechanical system based on the experimental measurements. The experiments were carried out in an operating mode with rotation speed between 200/min and 1,000/min. Four types of coupling were examined; pneumatic flexible coupling, Hardy coupling, Periflex coupling and claw coupling. In this operating mode, the values of vibrations in the gearbox were obtained. The measured value of the monitored vibrations, which is known as root mean square (RMS), is regarded as an effective value of vibration speed. Nikhil *et al.* (2017) compared analytical and simulation methods for dynamic torsional stiffness of couplings, and found that both the results match with deviation of approximately 2%.

Danish *et al.* (2019) observed the vibration signatures of two different coupling types (beam and split muff couplings) and the effect of increased shaft speed on rotor vibration at driving and non-driving ends. A signal-based approach was chosen using machinery fault simulator with the help of Vibra-Quest software package and with 2X amplitude of running shaft speed. Beam coupling was concluded to be the most suitable for the said condition for minimal vibration in context with parallel misalignment, while split muff coupling was showing much higher vibration amplitudes.

Shamin & Pallavi (2014) conducted an experimental investigation of unbalance with the help of FFT analyser and its unique vibration spectrum for different types of couplings, such as jaw coupling, flexible flange coupling and rigid coupling. They reported that unbalance related faults show dominant peak at 1×. For experimental identification of unbalance, an experimental setup was constructed and frequency spectrums were acquired for jaw coupling, flexible flange coupling and rigid flange coupling. The experimental results were found to be in close agreement with results available in literature.

In this paper, vibration measurement is utilised to detect the locations on rotor dynamic system that record the highest levels of vibration. Based on the measurement procedure developed in this experimental work, CM for rotor coupling-bearing systems can be developed in the future.

2. METHODOLOGY

2.1 Experimental Setup

This study is conducted on a rotor dynamic system, which is an apparatus for simulation that can be utilised for the study vibration behaviour of rotating elements, such as flexible couplings and bearings. Figure 1 and Table 1 show the description of the rotor dynamic system, which consists of 1/7 HP alternating current (AC) induction motor, flexible coupling and rotor shaft of 335 mm, which is supported by two identical ball bearings (pillow blocks). The gap between these bearings is 165 mm. The diameter of the rotor shaft is 10 mm. The rotor shaft is driven by a motor that is controlled using a variable frequency drive (VFD) as shown in Figure 2, which is mainly used to increase or decrease the speed of motor for up to 11,500 rpm. The flexible coupling used in this study is 44 mm in length and 30 mm in diameter. It consists of two portions, one from the driving unit, which is the motor, and the other end is the driven unit, which is the bearing.

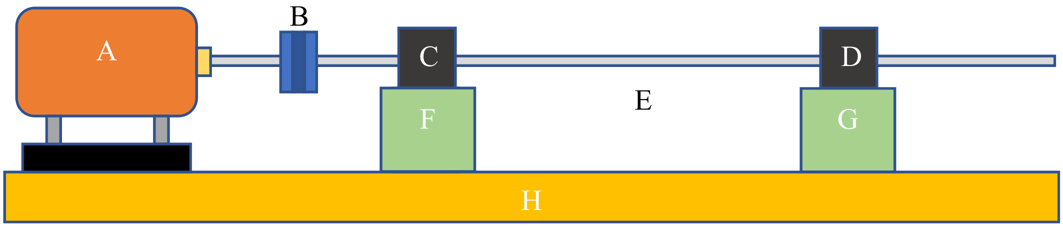
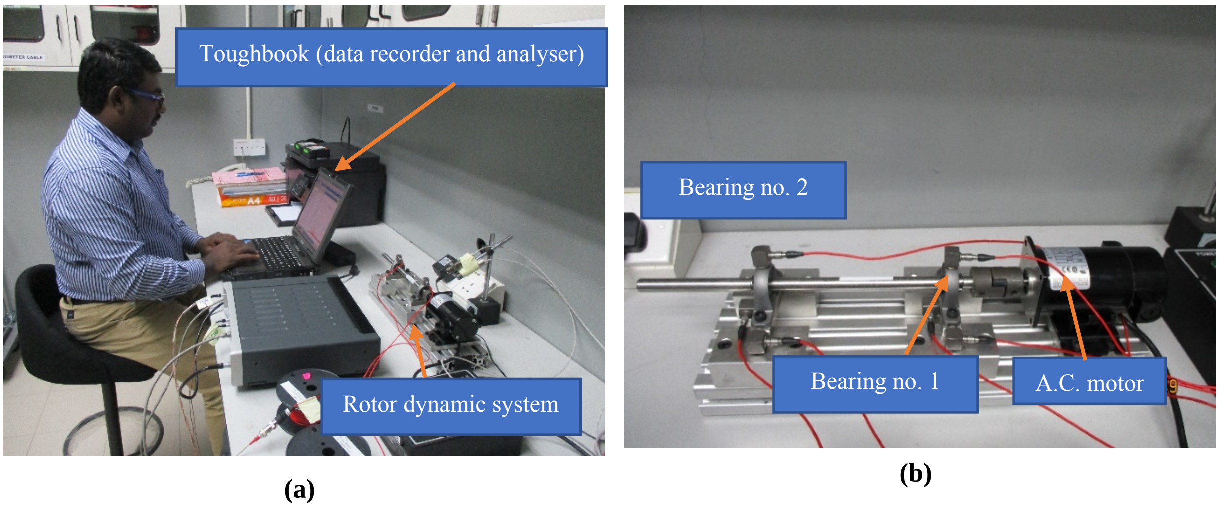


Figure 1: Diagram of the rotor dynamic. The descriptions for Locations A-H are given in Table 1.

Table 1: Description for each location on the rotor bearing shown in Figure 1.

Location	Description
A	AC Induction motor
B	Coupling
C	Pillow block no. 1
D	Pillow block no. 2
E	Rotor shaft
F	Bearing support no. 1
G	Bearing support no. 2
H	Main support board

Figure 3 shows the acquisition platform, which consists of a Toughbook (data recorder and analyser), data acquisition unit (SCADAS DAQ), RPM probe, accelerometers and the rotor bearing. A total of five accelerometers (vibration sensors) are used in this study. These accelerometers are mounted on the housings for both bearing and motor using glue, as shown in Figure 4. This technique is used for temporary installation because the rotor bearing surface is not adequately prepared for stud mounting. Three sensors are mounted on the surface of the bearing in three directions (axial, vertical and transverse) and two sensors are mounted on the surface of the driving end of the motor. These sensors are connected to a data acquisition unit using high quality measurement cables integrated with the LMS Test.Lab software platform.



(a) (b)

Figure 3: Acquisition platform: (a) Measurement in progress. (b) Main components of the rotor dynamic system.

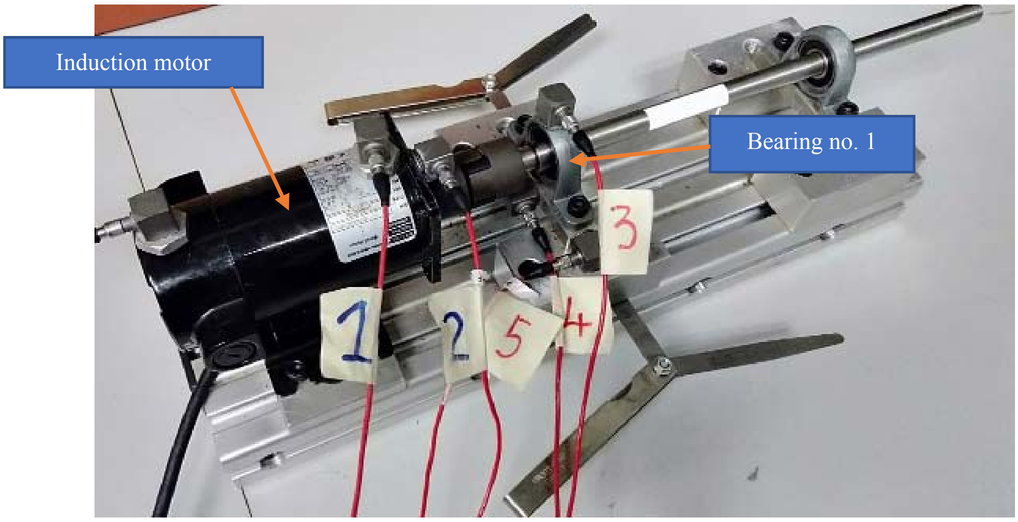


Figure 4: Accelerometers and measurement locations on the rotor dynamic system.

2.2 Measurement Procedure

Figure 4 shows the locations of the accelerometers, and their directions on the motor and rotor bearing. Between these two parts of the rotor dynamic system, there is a unit of flexible coupling as shown in Figure 1. First, the rotor dynamic is run for a few minutes to settle down all minor vibrations. Before creating unbalance, the shaft is visually checked for any misalignment and unbalance. The two dial gauge method is used to check the proper alignment and balancing.

The measurements were performed for the well balanced and unbalanced rotor systems at both the drive end (DE) of the motor and housing of the bearing, with vibration signals recorded for seven revolutions (motor speeds) - approximately at 550, 750, 950, 1,150, 1,350, 1,550 and 1,750 rpm. The duration for each measurement is five minutes. Table 2 shows the running rpm and frequency that is obtained by dividing the rpm value by 60 min, as shown in the following equation:

$$\text{Fundamental frequency}, f = \frac{Shaft\text{ rpm}}{60} \quad (1)$$

Table 2: Calculated fundamental frequencies using Equation 1.

Shaft speed (rpm)	Calculated fundamental frequency (Hz)
550.0	9.17
750.0	12.5
950.0	15.83
1150.0	19.17
1350.0	22.50
1550.0	25.83
1700.0	28.33

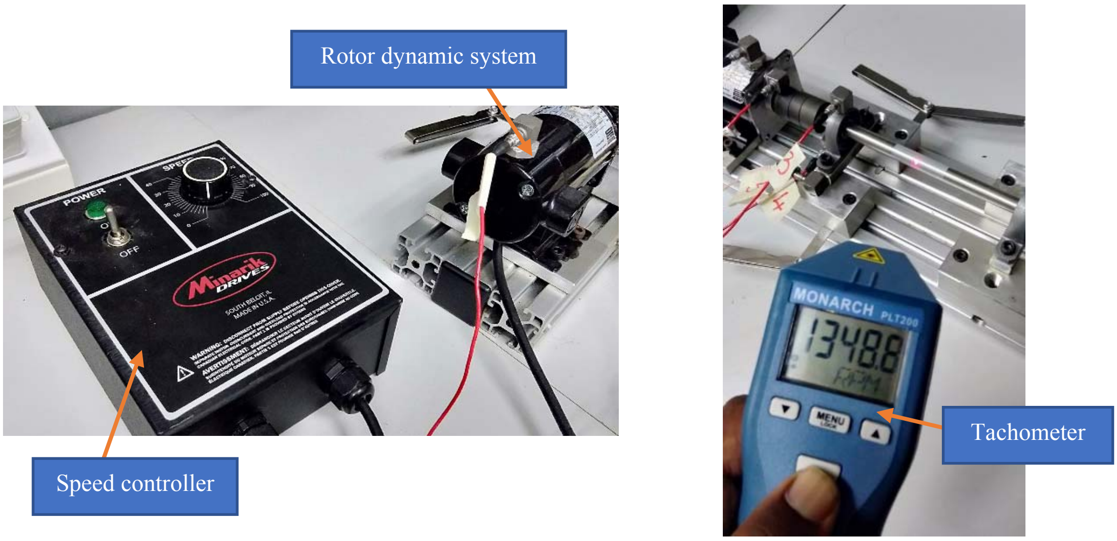
2.3 Measurement Matrix

Table 3 shows the measurement matrix in this study, which covers five different levels or heights of support block of the bearing. The list of different heights also includes normal height, which is the original height without any increment. At each different height, vibration measurement is conducted at seven various revolutions of the rotor dynamic system, starting from the minimum revolution until it reaches maximum, which could be considered as high-speed level rotation.

Table 3: Measurement matrix.

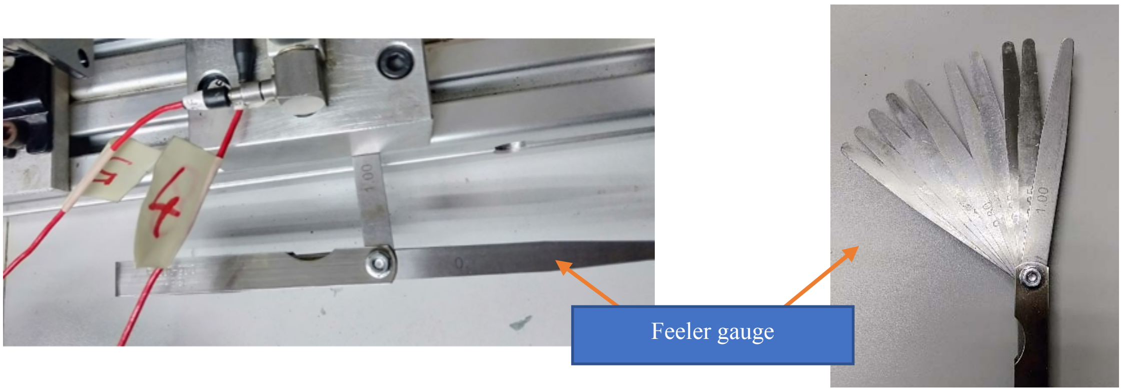
Revolution (Motor Speed) (rpm)	Increment of height of bearing no. 1 support block (mm)				
	Normal height	0.4	0.6	0.8	1.0
550 (±5.00)	Section 1-run 1,2,3	Section 2-run 1,2,3	Section 3-run 1,2,3	Section 4-run 1,2,3	Section 5-run 1,2,3
750 (±5.00)	Section 1-run 4,5,6	Section 2-run 4,5,6	Section 3-run 4,5,6	Section 4-run 4,5,6	Section 5-run 4,5,6
950 (±5.00)	Section 1-run 7,8,9	Section 2-run 7,8,9	Section 3-run 7,8,9	Section 4-run 7,8,9	Section 5-run 7,8,9
1150 (±5.00)	Section 1-run 10,11,12	Section 2-run 10,11,12	Section 3-run 10,11,12	Section 4-run 10,11,12	Section 5-run 10,11,12
1350 (±5.00)	Section 1-run 13,14,15	Section 2-run 13,14,15	Section 3-run 13,14,15	Section 4-run 13,14,15	Section 5-run 13,14,15
1550 (±5.00)	Section 1-run 16,17,18	Section 2-run 16,17,18	Section 3-run 16,17,18	Section 4-run 16,17,18	Section 5-run 16,17,18
1750 (±5.00)	Section 1-run 19,20,21	Section 2-run 19,20,21	Section 3-run 19,20,21	Section 4-run 19,20,21	Section 5-run 19,20,21

In this measurement, a Minarik Drives (USA) speed controller is used to control the revolution of the rotor dynamics. At normal height condition, the vibration measurement is started at 550 rpm. The revolution in rpm is verified using a handheld tachometer as shown in Figure 5 (b). The measurement is conducted for 7 min. for each revolution, three runs are obtained. Once completed at 550 rpm, the revolution is increased to approximately 750 rpm and subsequently to 950, 1,150, 1,350, 1,550 and 1,750 rpm. At each revolution, the data recording was maintained for 5 min using time domain format. Once completed, the height of the supporting block of the bearing support is increased to 0.4 mm using two units of feeler gauges at the right and left sides of the bearing support as shown in Figures 6(a). Figure 6(b) shows one of the feeler gauges with various thicknesses of shims that was used in this study.



(a) **(b)**

Figure 5: Speed controller and tacho for rpm measurement.



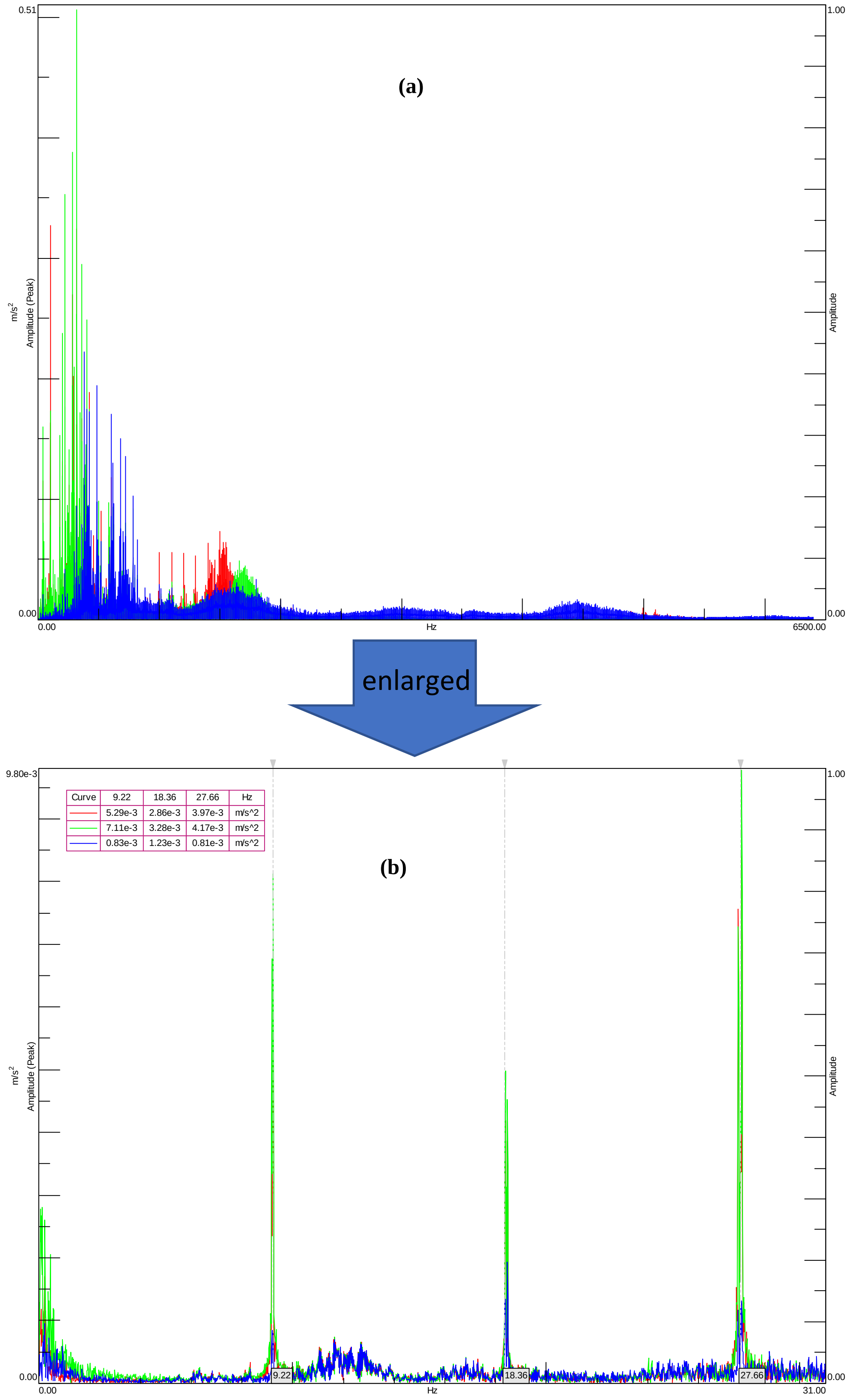
(a) **(b)**

Figure 6: Feeler gauge with various thickness of shims.

Once again, the measurement commences at the minimum revolution, which is approximately 550 rpm and above procedures are repeated until the speed reaches the maximum revolution, which is approximately 1,750 rpm. By referring to Table 3, it shows that for each revolution, certain levels of tolerances are set because speed adjustment using this controller is done manually by tuning the speed knob and thus, it is difficult to get exactly the same revolution in rpm as previous measurement.

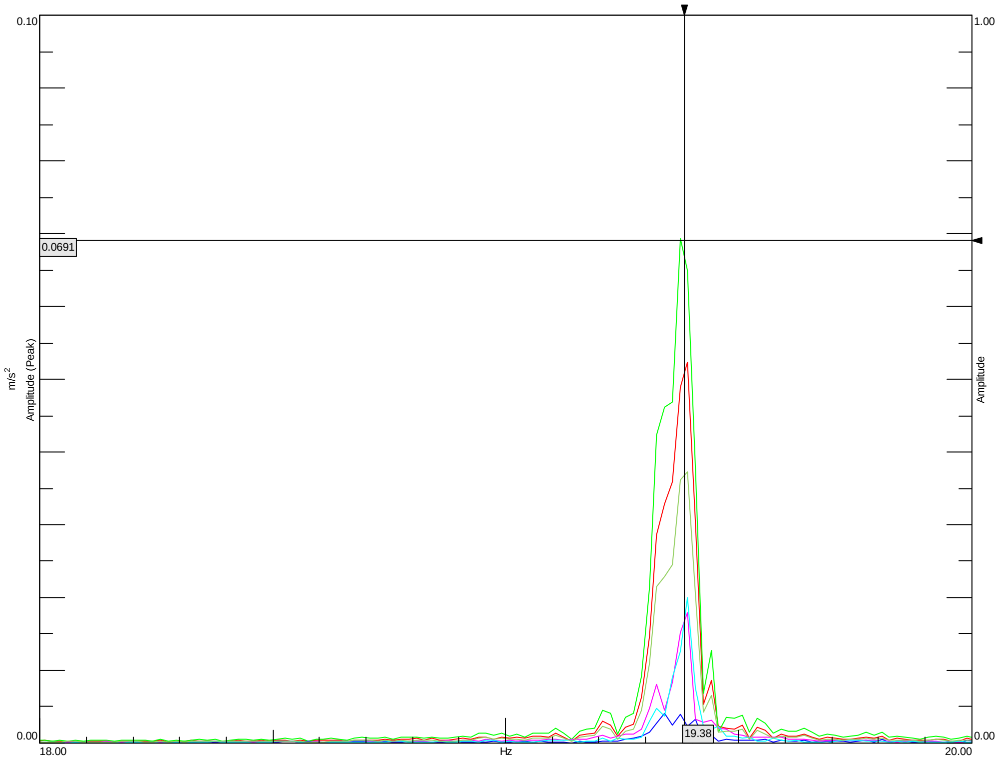
3. RESULTS AND ANALYSIS

Figure 7(a) shows the vibration spectrum in frequency domain for up to 6,400 Hz. The spectrum is zoomed to a smaller scale as shown in Figure 7(b). The next step is the identification process of the amplitude values at fundamental or running frequency at five different locations. For example, Figure 8 is recorded at revolution of 1,550 rpm and at bearing support height of 0.5 mm. The fundamental frequency is 19.38 Hz and acceleration is 0.0691 m/s^2 (peak) by referring to the intersection point on the y-axis. By utilising this method, other amplitude values at various revolutions and heights are obtained to form vibration trends as shown in Figures 9-13.



Curve	9.22	18.36	27.66	Hz
🟥	5.29e-3	2.86e-3	3.97e-3	m/s^2
🟩	7.11e-3	3.28e-3	4.17e-3	m/s^2
🟦	0.83e-3	1.23e-3	0.81e-3	m/s^2

Figure 7: (a) Vibration spectrum in frequency domain. (b) View of the frequency spectrum (enlarged).



“ — ”	Location num 1	“ — ”	Location num 3	“ — ”	Location num 5
“ — ”	Location num 2	“ — ”	Location num 4		

Figure 8: Identifying vibration amplitudes of different locations.

Figures 9 – 13 show the trending of accelerations at five different locations on the rotor dynamic system; two on the motor and three on the surface of bearing housing. These amplitudes are measured at various revolutions; starting from the minimum revolution of 550 rpm until the maximum revolution of 1,750 rpm with increment of 200 rpm at each measurement stage. Each pattern shows the trending of vibration amplitude at different heights of bearing support, including the one during normal height condition.

Figure 9 shows the vibration amplitudes at location number 1, which is located on the surface of driving end of the motor in the vertical direction at different heights. Based on this figure, it is found that the vibration amplitudes are increasing from minimum revolution until it reaches maximum revolution. This pattern of vibration trending is almost similar at all conditions of bearing support heights.

In addition, based on the vibration trending pattern for location number 1, we observe that whenever the bearing support height is increased, the vibration amplitudes also increase consistently except during support heights of normal and 0.4 mm. It shows that lower increment of height contribute to less effect on the vibration amplitude changes due to mild existence of misalignment condition.

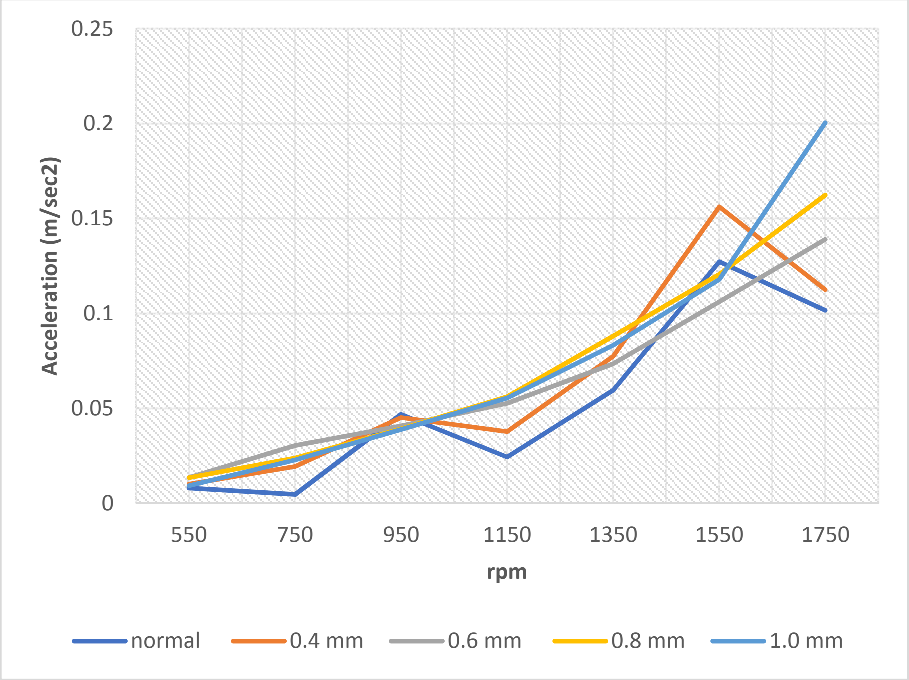


Figure 9: Comparison of measured vibration amplitudes at location 1.

Figure 10 shows the vibration amplitudes at location number 2, which is located on the surface of driving end of the motor in the axial direction. Almost a similar trending pattern as per location number 1 is observed at location number 2, even though both are measured in different directions. This similarity shows that the locations of measurement that very close to each other do not contribute to any major changes in the vibration trending. These can be observed based on the amplitudes measured at different revolutions and bearing support heights in this study.

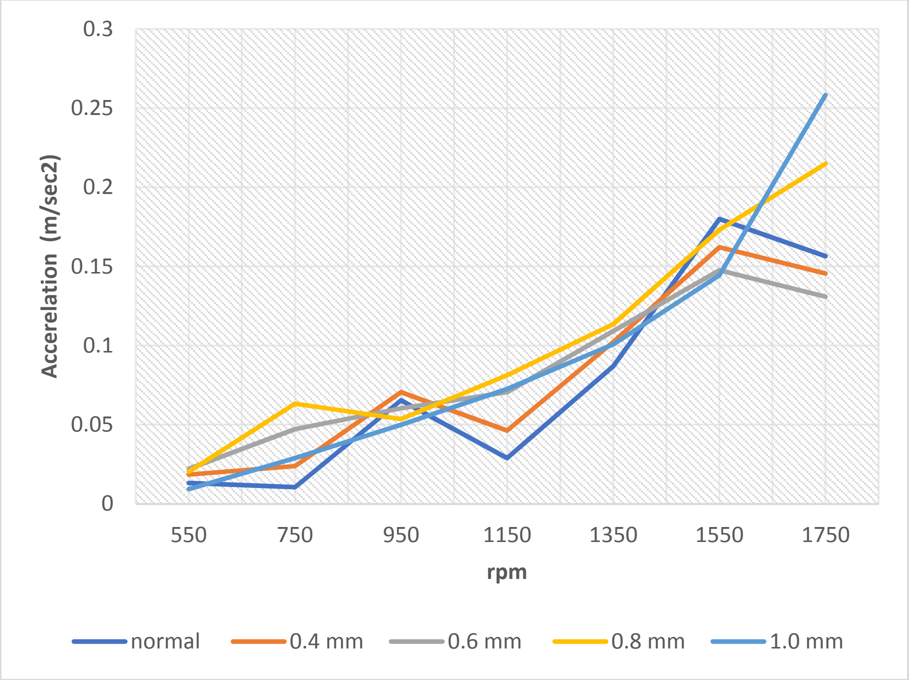


Figure 10: Comparison of measured vibration amplitudes at location 2.

Figures 11-13 show the trending of vibration amplitudes at location number 3, 4 and 5 on the bearing housing, which is supported by the bearing support. In comparison with Figures 9 and 10, these three figures show obvious differences in term of measured vibration amplitudes and their trending with different revolutions and heights.

Figure 11 shows vibration amplitudes at location number 3, which is located on the surface of the bearing housing in the vertical direction. This is considered as the driven side of the flexible coupling. Based on this figure, it shows lower range of vibration trending at revolutions between 550 and 1,150 rpm at all levels of support heights.

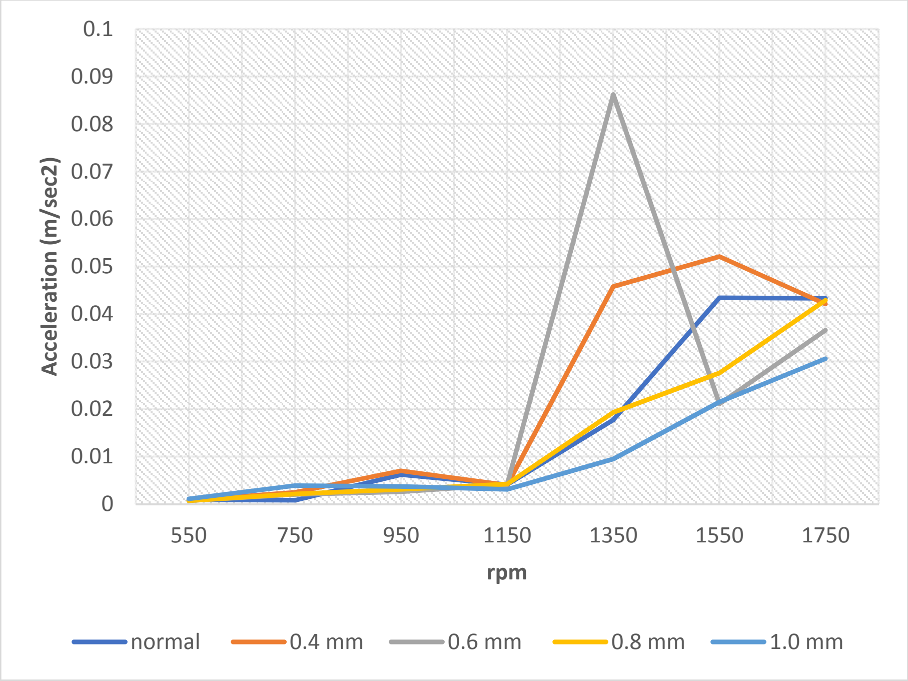


Figure 11: Comparison of measured vibration amplitudes at location 3.

At 1,350 rpm and above revolutions, inconsistent conditions are observed in terms of vibration trending, especially at height of 0.6 mm, which recorded a very high value of vibration amplitude. This value suddenly dropped to lower amplitude value once the revolution increased to 1,550 rpm. This abnormal condition could be due to the existence of critical speed or resonance of the rotor dynamic system. An almost similar condition is observed at height of 0.4 mm, but the measured amplitude is slightly lower than at height of 0.6 mm once the revolution increased to 1,550 rpm. Meanwhile, at heights besides 0.4 and 0.6 mm, the vibration amplitudes changes are consistent with increasing revolutions and bearing heights.

Figure 12 shows the vibration amplitudes at location number 4, which is located on the surface of bearing housing in the axial direction. At this location, the vibration amplitude trending is almost same for different heights during lower range of revolutions except at heights of 0.8 mm and 1.0 mm. At 1.00 mm, the vibration amplitudes are increasing when the revolutions are increased towards maximum revolutions.

Figure 13 shows the vibration amplitudes at location number 5, which is located on the surface of bearing housing in the transverse direction. At lower range of revolutions, which is between 550 and 1,350 rpm, it is found that the vibration amplitudes are not consistent and shows signs of increment at lower rate after 950 rpm. In addition, at all heights, the vibration amplitudes increased when the revolution is increased gradually from 1,350 to 1,550 rpm. The same pattern of consistent and gradual increment of vibration amplitudes are also observed when the revolution is increased from 1,550 to 1,750 rpm. The highest vibration amplitude is recorded at height of 0.6 mm, far larger in comparison than the other heights.

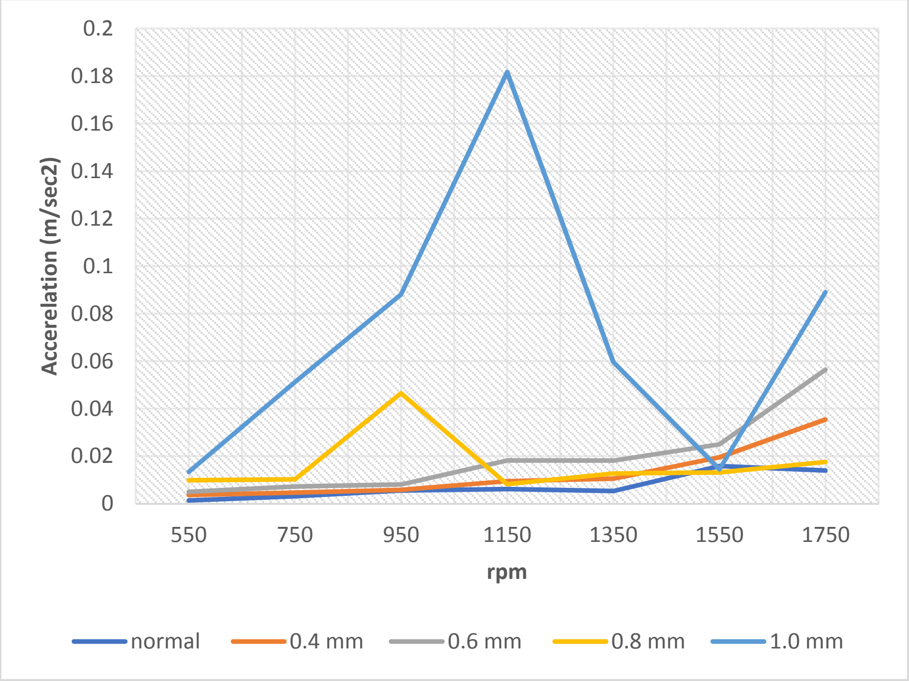


Figure 12: Comparison of measured vibration amplitudes at location 4.

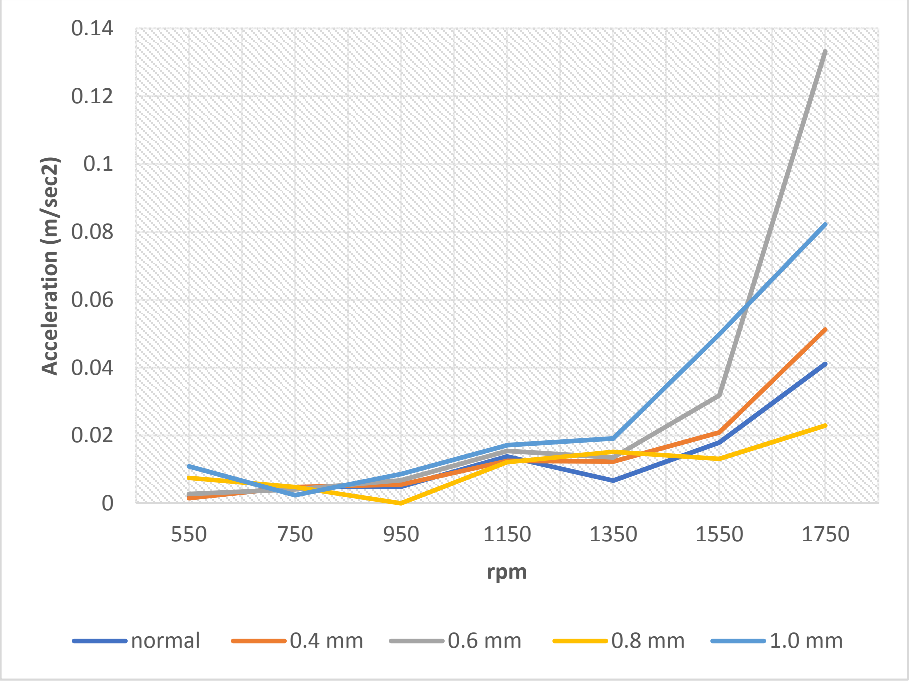


Figure 13: Comparison of measured vibration amplitudes at location 5.

4. CONCLUSION

Based on this study, we can conclude that the misalignment symptom, which simulated in experimental mode study by imposing different heights of bearing support, have generated increments in terms of accelerations at different revolutions. This shows that the presence of misalignment along the ship propulsion shafting line consisting of heavy machineries, such as diesel engine and gearbox, would contribute certain levels of vibrations at both sides of driving and driven ends. These both ends are connected using flexible coupling, which is subject to transmit power and compensates for some degree of misalignment due to torsional vibration.

Indirectly, severe misalignment and torsional vibration could damage the flexible coupling. In the long run, internal components of coupling, such as gears and gaskets, which are subject to tear and wear, would fail and contribute to catastrophic failure. For example, whenever a gasket that prevents oil leaking inside the coupling is dislocated from the original location, this will force leaking of lubricant oil that is basically for the purpose of reducing higher friction inside the coupling. Knocking sounds and wobbling are some of indications showing that the coupling needs to be dismantled and sent for overhaul.

Therefore, identifying failure mode at an early stage could benefit the user and prevent catastrophic failures. A CM programme could be utilised in detecting symptoms relating to coupling failure, and give warning or alarm to the user. This would save time and money, as well as being important for the safety of operators because in certain cases, there is possibility for the coupling to fail horribly and fly off from the connection, which could cause injuries to nearby operators (Elamin *et al*., 2010).

REFERENCES

Danish, I., Mohd, A. & Dr. Ahmad, A.K., (2019). Experimental study of vibration signature of various couplings using MFS lite. *Int. J. Eng. Appl. Sci. Tech.,* **4**: 86-93.

Desavale, R.G., Venkatachalam, R. & Chavan, S.P. (2013). Antifriction bearings damage analysis using experimental data based models. *ASME J. Tribology,* **135**: 041105.

Elamin, F. (2013). *Fault Detection and Diagnosis in Heavy Duty Diesel Engines Using Acoustic Emission.* PhD thesis, The University of Huddersfield, Huddersfield, UK.

Elamin, F., Gu, F. & Ball, A. (2010). Diesel engine injector faults detection using acoustic emissions technique. *Modern Appl. Sci.,* **4**: 3-13.

Gani, A. & Salami, M.J.E. (2004). Vibration faults simulation system (VFSS): A lab equipment to aid teaching of mechatronics courses. *Int. J. Eng. Ed.,* **20**: 61-66.

Grega, R., Homišin, J., Krajňák, J. & Urbanský, M. (2016). Analysis of the impact of flexible couplings on gearbox vibrations. *Sci. J. Silesian Univ. Tech.,* **91**: 43-50.

Gu, F., Yesilyurt, I., Li, Y., Harris, G. & Ball, A. (2006). An investigation of the effects of measurement noise in the use of instantaneous angular speed for machine diagnosis. *Mech. Syst. Signal Proc.,* **20**: 1444-1460.

Khardersab, A. & Shivakumar, S. (2018). Experimental investigation of the excitation forcing function in rotating machinery. *Procedia Manuf.*, **20**: 247-252.

Khot, S.M. & Pallavi, K. (2015). Simulation and experimental study for diagnosis of misalignment effect in rotating system. *J. Vib. Anal., Meas. Contr.*, **3**: 165-173.

Kiran Kumar, B., Diwakar, G. & Satynarayana, M.R.S. (2012). Determination of unbalance in rotating machine using vibration signature analysis. *Int. J. Mod. Eng. Res.*, **2**: 3415-3421.

Li, Z., Yan, X., Guo, Z., Zhang, Y., Yuan, C. & Peng, Z. (2012). Condition monitoring and fault diagnosis for marine diesel engines using information fusion techniques. *Electron. Electr. Eng.,* **123**: 109-112.

Malla, C. & Panigrahi, I (2019). Review of condition monitoring of rolling element bearing using vibration analysis and other techniques. *J. Vib. Eng. Tech*, **7**: 407-414.

Mehdi, A., Rohani, B. & David, A. (2011). A new model for estimating vibration generated in the defective rolling element bearings. *J. Vib. Acoustics,* **133**: 1 – 8.

Mihaela, U. & Silviu, B. (2014). consideration about elastic coupling. *Appl. Mech. Mater.,* **657**:760-764.

Nikhil, G., Shivaji, O. & Maruti, M. (2017). Torsional dynamic stiffness of flexible coupling by analytical and simulation method. *Int. Eng. Res. J.*, **2**: 121-123.

Nistane, V.M. & Harsha, S.P. (2016). Failure evaluation of ball bearing for prognostics. *Procedia Tech.*, **23**: 179 – 186.

Pallavi, K. (2014). Experimental study to identify the vibration signature of bent shaft. *International J. Eng. Res. Tech.*, **3**: 214-216.

Shamim, P. & Pallavi, K. (2014). Experimental study to identify the effect of type of coupling on unbalance using frequency spectrum analysis. *IOSR J. Mech. Civil Eng.*, **11**: 13-16.

Shi, H., Guo, J., Bai. X., Guo, L., Liu, Z. & Sun, J. (2020). Gearbox incipient fault detection based on deep recursive dynamic principal component analysis. *IEEE Access*, **8**: 57646-57660

Shweta, S.P. & Tuljapure, S.B. (2015). Causes of coupling failures and preventive actions. *Int. J. Res. Appl. Sci. Eng. Tech.*, **3**: 412-415.

Vishwakarma, M., Purohit,R, Harshlata, V. & Rajputa, P. (2017). Vibration analysis & condition monitoring for rotating machines: A review. *Mater. Today Proc.*, **4**: 2659-2664.

FRACTURE FAILURE ANALYSIS OF A MARINE PROPELLER SHAFT

Mohd Moesli Muhammad*, Mohd Subhi Din Yati, Nik Hassanuddin Nik Yusoff & Mahdi Che Isa

Marine Technology Research Group, Maritime Technology Division (BTM), Science & Technology Research Institute for Defence (STRIDE), Ministry of Defence, Malaysia

*Email: moesli.muhammad@stride.gov.my

ABSTRACT

In this paper, root cause analysis of the failure of a marine propeller shaft was performed using the standard procedure for failure analysis. The propeller shaft failed during cruising and it was completely broken into two separate pieces. The fracture surface was examined using macroscopic and fractography techniques. It was discovered that the failure of the shaft was due to fatigue failure. Beach marks were found on the fracture surface, with the directions of crack growth being from two points of origin at the sharp corner of the keyway. Further investigation revealed that fretting marks and micro-cracks were present on the keyway surface, which contributed to the crack initiation. Both material composition and hardness results showed that the propeller shaft complied with 316 Stainless Steel grade, which is suitable for marine applications.

Keywords: *Failure analysis; fretting; propeller shaft; beach marks; keyway.*

1. INTRODUCTION

A propulsion system is a mechanical mechanism that produces thrust to move a ship across the sea. This system consists an engine, which transmits the power connected to a shaft to drive the propeller. Since the propeller shaft is a medium to transmit the power from the engine by the rotational movement, most naval architects design this shaft in solid cylindrical shape and fabricate it from metallic materials that are able to operate under high strength and broad range of loads. Due to the major role of the propeller shaft, failure of this component can lead the catastrophic failure to the propeller system (Pantazopoulos & Papaefthymiou, 2015; Vardhan *et al.*, 2019).

In this paper, a study was carried out on a failed ship propeller shaft (Figure 1). It was reported that the failure occurred when the ship was on normal cruising, where the fractured shaft at the end portion with the propeller fell into the sea. From the background information, no abnormality of severe vibration or underwater accident was recorded. The propeller shaft was collected and sent to STRIDE for detailed failure investigation. Figure 2 shows the as received condition of the propeller shaft. The failed sample was completely broken into two separate pieces.

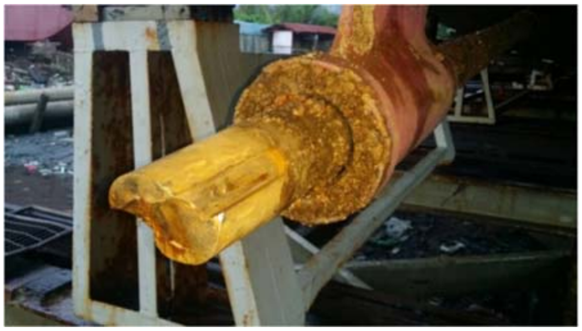


Figure 1: Propeller shaft failure.

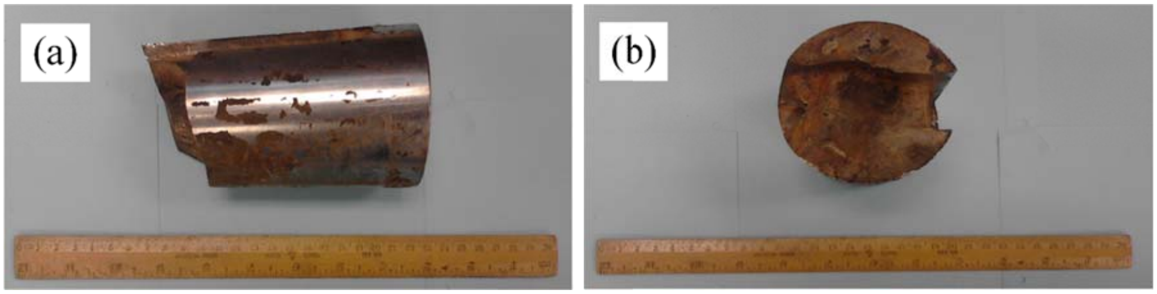


Figure 2: As received propeller shaft: (a) Side view (b) Top view.

The main goal of this failure investigation is to collect as much as possible data on how and when the failure may have initiated in the propeller shaft. The investigation will also determine whether other parameters such as mechanical properties of materials were insufficient for the purposes of the propeller shaft.

2. METHODOLOGY

According to the design specification, the propeller shaft should be safe under normal operation conditions. However, the failure still occurred. Therefore, there should be important factors that were not considered during maintenance works or alignment process. Investigations were carried out to reveal the cause of the failure from metallurgical considerations (ASM, 1986; 1992; 1993). Before any laboratory work was carried out, all available data concerning the specifications, drawings, component design and history of maintenance record were gathered and analysed. The failed propeller shaft was inspected visually using a Zeiss Stemi DV4stereomicroscope. Next, the failed samples were cut using an abrasive cutter for detailed metallurgical analysis. Before that, care was taken to avoid damage especially on the fractured surfaces. The fractured surfaces were ultrasonically cleaned with 100 ml nitric acid (HNO_3), 20 ml hydrofluoric acid (NaOH) and reagent water (ASTM, 1999) to remove corrosion products and calcium deposited (Figure 3). After cleaning, the failed sample was further observed in high magnification using a COXEM EM30 tabletop scanning electron microscope (SEM). The chemical compositions were analysed using a Bruker S4 Pioneer wave dispersive X-ray fluorescent (WDXRF). For the hardness, the samples were measured using a Shimadzu HMV-2T microhardness tester at five different locations.

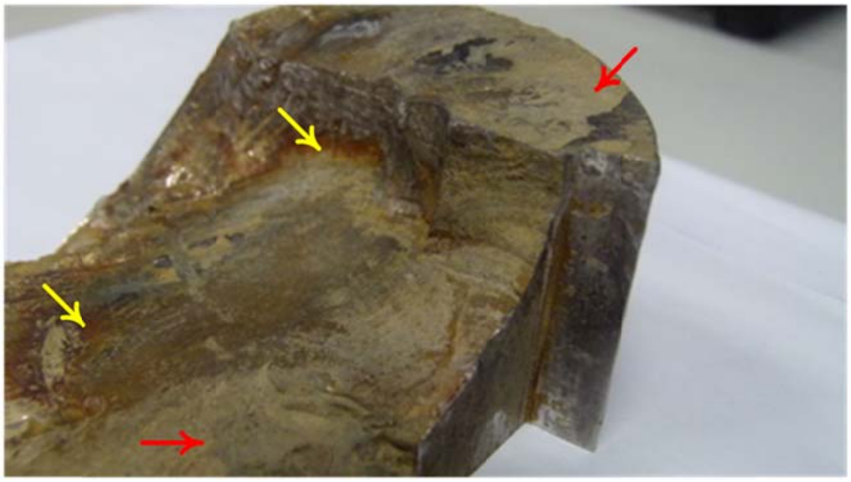


Figure 3: The fracture surfaces (before cleaning) indicate the existence of a calcium deposited layer (red arrows) and corrosion products (yellow arrows).

3. RESULTS & DISCUSSION

3.1 Macroscopic Examination

Figure 4 shows the fracture surface of the broken propeller shaft. The examination revealed wave like lines on the fracture surfaces. In failure analysis, these marks on the fracture surface are known as beach marks or progression marks (Atxaga & Irisarri, 2010; Zangeneh *et al.*, 2013; Pantazopoulos & Papaefthymiou, 2015). These beach marks are an extremely important indication to describe that the failure is caused by fatigue. The direction of the beach marks can be clearly seen from two points of origin, which are points A and B, which are the weakness points due to both fillets of the keyway corner being sharp. According to Huang *et al.* (2019), sharp corners of a fillet keyway is a high stress concentration point that can cause high local stress to be as much as ten times than the average nominal stress. The examination also revealed that the propagation region of the beach marks from point A is larger than from point B. This area is approximately two third of fractured surfaces. The large area of beach marks from initiation point A indicates the direction of shaft rotation (arrow Figure 4) before rupture. Then, both propagation directions of beach marks A and B collide in region C, which is the final rupture. This final rupture region exhibits rough surface and fast fracture, with the area being estimated as only one third of the fracture surface.

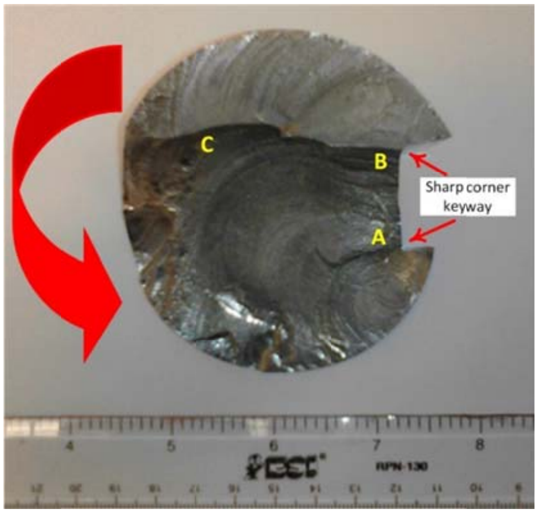


Figure 4: Macrograph of fracture surface of propeller shaft.

Further analysis was conducted to determine the cause of initiation points A and B. Figure 5(a) shows the fracture surface being cut into three pieces. Sections 1 and 2 were viewed under high magnification, which represents the beach marks from point A and B respectively (Figure 5(b) and 5(c)). Examination of the fracture surface at the keyway points A and B (Figure 6(a) and 6(b)) found fretting, indicating that slippage had occurred. The fretting occurred on the keyway surface due to the mismatch between lock size and keyway. The combination of the fretting and cyclic stress on the keyway produced the small cracks (points A & B). This continuous repetitive stress caused the cracks to enlarge, propagate and finally rupture. This fretting greatly reduced the fatigue limit of the shaft metal and resulted in initiation of fatigue cracks (Sitthipong *et al.*, 2017).

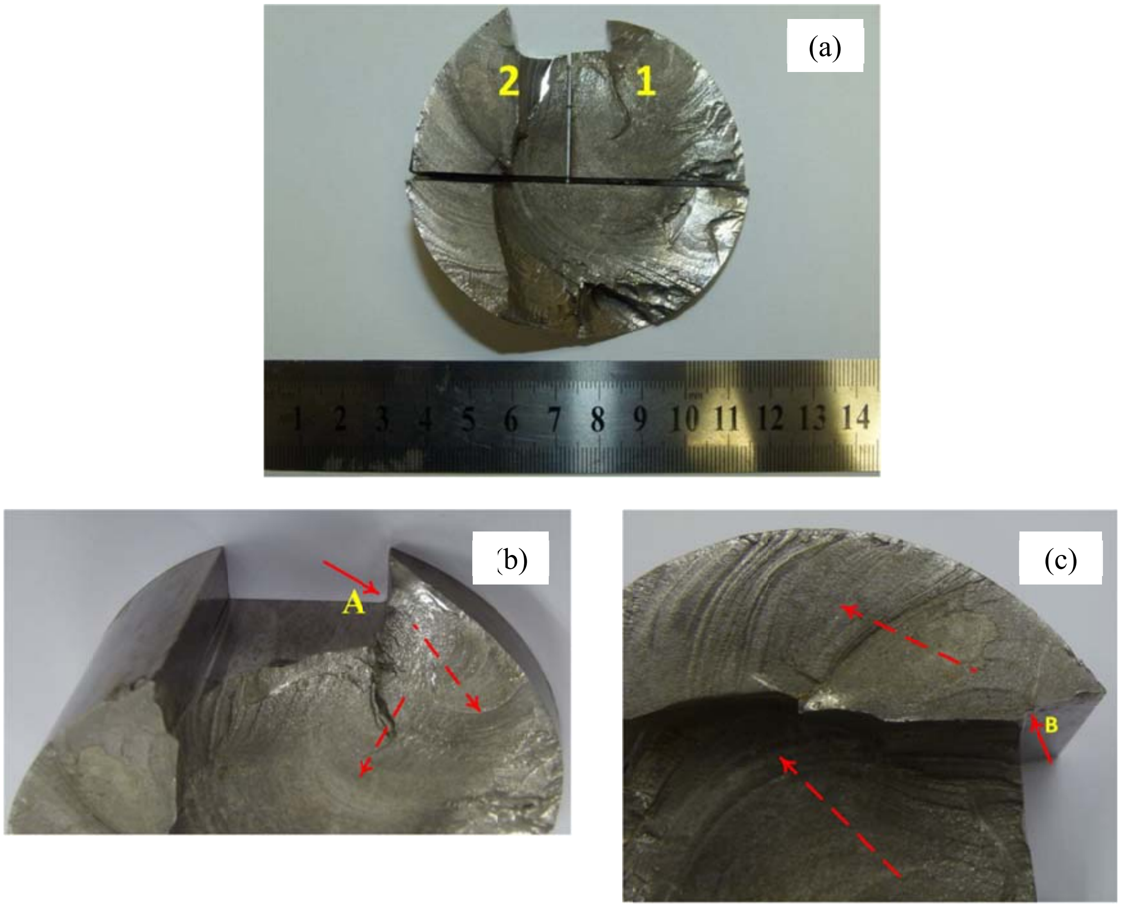


Figure 5: (a) The fracture surface was cut into three pieces. The red arrow indicates the starting point and the directions of the propagation. (b) Point A. (c) Point B.

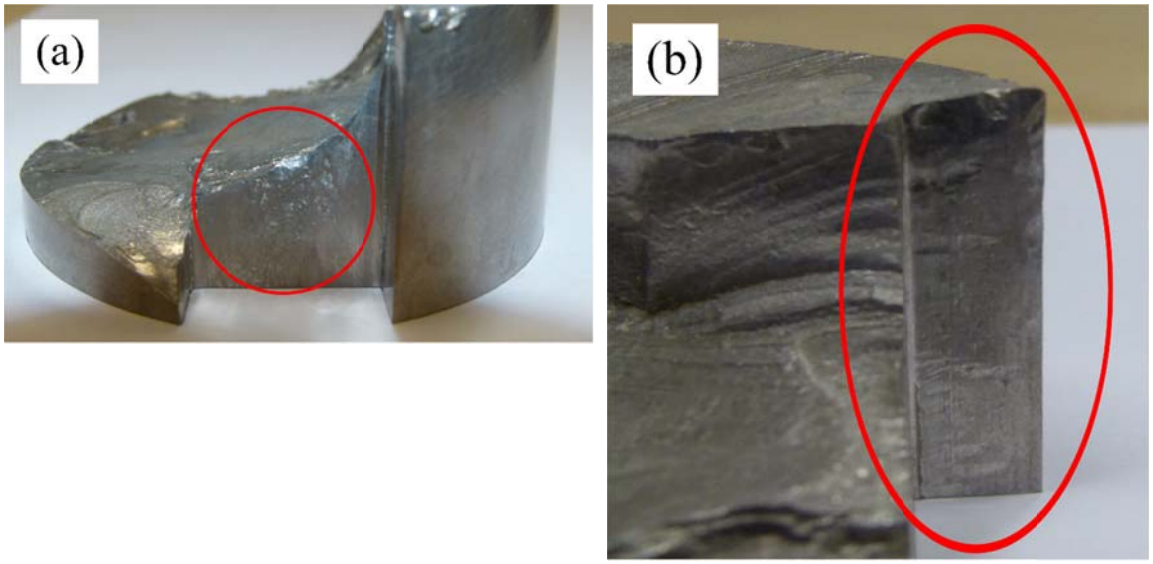


Figure 6: Fretting was found on the keyway surface, indicating mismatch between lock size and keyway, which also contributed to the failure of the propeller shaft: (a) Point A (b) Point B.

3.2 Fractography Examinations

The examination on the surface morphologies at points A and B are shown in Figures 7 and 8 respectively. The observation was performed using the SEM to reveal the fracture features. The SEM images in the Figures 7(a) and 8(a) clearly indicate the presence of the cracks on the sharp corner fillet of the keyway. With further increase of the magnification, micro-crack and striation marks appear at point A (Figure 7(b)), and only striation marks present at point B (Figure 8(b)). The fatigue striation marks appear adjacent to the crack initiation point for both points A & B, strongly indicating that the fracture failure was due to fatigue failure. The micro-cracks on the surface are contributed by improper machining during grinding process. This defect caused to crack initiation that can be the potential point of failure (Zerbst & Klinger, 2019). Failure of this kind can be avoided by using a half-round keyway, which permits the use of a round key or by using a generous fillet radius in the keyway. A half-round keyway produces a local stress that is only twice the average stress (Atxaga & Irisarri, 2010).

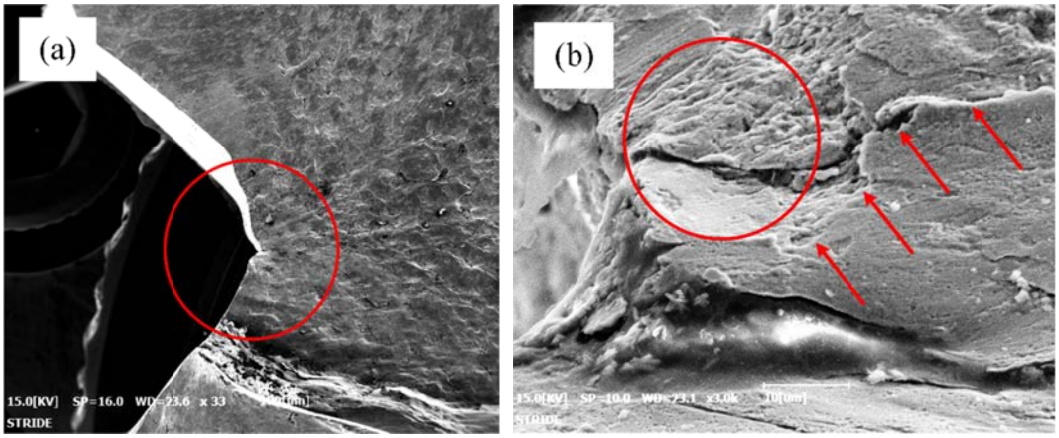

(a)
15.0[KV] SP=16.0 WD=23.6 x 33
STRIDE
(b)
15.0[KV] SP=10.0 WD=23.1 x3.0k 10[um]
STRIDE

Figure 7: Shaft surface morphology at point A: (a) Crack at sharp corner area of keyway (b) Striation marks and micro-cracks adjacent to the initiation point.

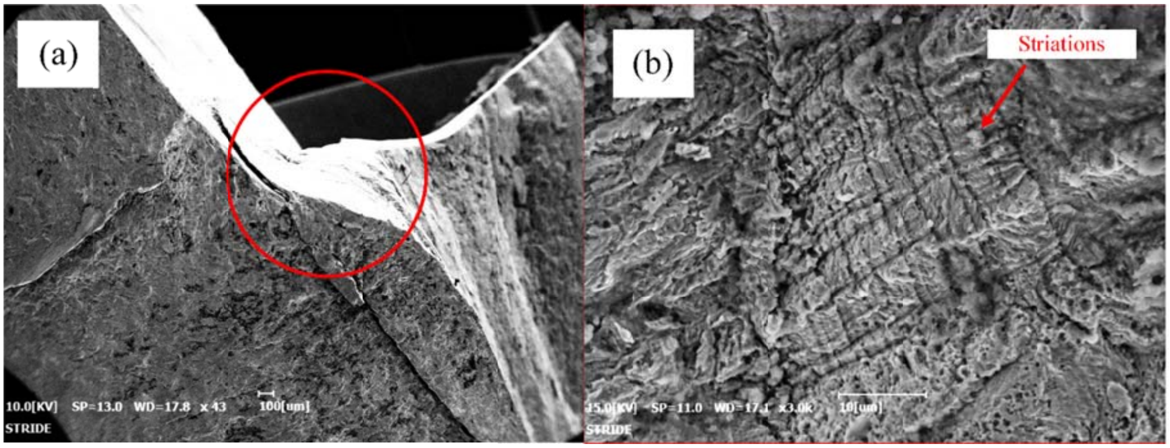

(a)
10.0[KV] SP=13.0 WD=17.8 x 43 100[um]
STRIDE
(b)
Striations
15.0[KV] SP=11.0 WD=17.1 x3.0k 10[um]
STRIDE

Figure 8: Shaft surface morphology at point B: (a) Crack at sharp corner area (b) Striation marks.

3.3 Materials Composition

Chemical analysis on the failed sample was carried out using the WDXRF to determine the elemental composition presence in the propeller shaft material. Based on the results obtained in the Table 1, the propeller shaft was found to contain high percentage of Cr (Chromium), Ni (Nickel) and Mo

(Molybdenum) as major alloying elements with the Fe (Iron) as a matrix in this alloy. High concentration of Cr (almost 16%) with Ni and Mo content indicate that the shaft material is Type 316 Stainless Steel or Austenitic Chromium-Nickel Stainless Steel (ASM, 1993). The addition of Cr, Ni and Mo elements in steel provide excellent corrosion resistance and strength at elevated temperature. Therefore, the selection of this material is suitable for use in propeller shaft application and is not a cause of the failure.

Table 1: Chemical composition of the failed propeller shaft analysed using the WDXRF.

Element	Cr	Ni	Mo	Mn	C	Cu	Si	Co	V	P	Fe
Composition (%)	15.80	9.07	1.88	1.07	0.99	0.40	0.29	0.16	0.06	0.03	Remainder

3.4 Materials Hardness

The hardness of the failed shaft was measured by using the microhardness tester. The hardness test was carried out on five different locations, with the results presented in Table 2. Based on the hardness value obtained, the shaft hardness complied with the properties of stainless steel, whereby the hardness must be above than 170 HV, and thus is not a contribution of failure (Pantazopoulos & Papaefthymiou, 2015).

Table 2: Hardness values of failed shaft.

Bolt samples	1	2	3	4	5	Average
Hardness Value (HV)	177	181	208	218	188	194

4. CONCLUSION

This investigation was carried out on the failed propeller shaft that was used for a marine propulsion system. Macroscopic examination on fracture surface showed that beach marks appear, with directions from two points of origin, which were at the sharp corner of the keyway. The examination on the keyway found that fretting occurred, which was caused by mismatch between the lock size and keyway. Further examination using SEM for higher magnification on the fracture surface revealed that the striation and micro-cracks appeared on the fracture surface. The chemical composition showed that the shaft was 316 Stainless Steel and the average hardness was 194 HV. Based on macroscopic and fractography examinations, it can be concluded that the failure of propeller shaft was due to fatigue failure.

ACKNOWLEDGEMENT

The authors would like to thank the officers and staff from the various laboratories at the Maritime Technology Division (BTM), Science & Technology Research Institute for Defence (STRIDE), Ministry of Defence,Malaysia for their technical support during investigation works.

REFERENCES

ASM (American Society for Metals) (1986). *ASM Metals Handbook, Volume 11: Failure Analysis and Prevention*. American Society for Metals (ASM)

ASM (American Society for Metals) (1992). *ASM Handbook of Case Histories in Failure Analysis, Vol. 1 and Vol. 2*. American Society for Metals.

ASM (American Society for Metals) (1993). *Handbook of Case Histories in Failure Analysis, Vol. 2*. American Society for Metals.

ASTM (American Society for Testing and Materials (ASTM)) (1999). Standard Practice for Preparing, Cleaning and Evaluating Corrosion Test Specimens. American Society for Testing and Materials (ASTM).

Atxaga, G. & Irisarri, A.M. (2010). Failure analysis of the end of a shaft of an engine. *Eng. Fail. Anal.*, **17**:714–721.

Huang, Z., Zhang, Z., Teng, Z., Khan, M. K., Wang, Q. & Wang, J. (2019). Effect of fretting damage on characteristics of high strength bearing steel up to very high cycle fatigue. *Eng. Fract. Mech.*, **217**: 106526.

Pantazopoulos, G. & Papaefthymiou, S. (2015). Failure and fracture analysis of austenitic stainless steel marine propeller shaft. *J. Fail. Anal. Prev.*, **15**:762–767.

Sitthipong, S., Towatana, P. & Sitticharoenchai, A. (2017). Failure analysis of metal alloy propeller shafts. *Mater. Today Proc.*, **4**:6491–6494.

Vardhan, D.H., Ramesh, A. & Reddy, B.C.M. (2019). A review on materials used for marine propellers. *Mater. Today Proc.*, **18**:4482–4490.

Zangeneh, S., Ketabchi, M. & Kalaki, A. (2013). Fracture failure analysis of AISI 304L stainless steel shaft. *Eng. Fail. Anal.*, **36**:155–165.

Zerbst, U. & Klinger, C. (2019). Material defects as cause for the fatigue failure of metallic components. *Int. J. Fatigue*, **127**:312–323.

DEVELOPMENT OF BLAST RESISTANT MATERIALS USING GREEN MANUFACTURING PROCESS

Mohammed Alias Yusof[1*], Norazman Mohamad Nor[1], Muhamad Azani Yahya[1], Vikneswaran Munikanan[1], Arifin Ismail[2] & Ho Chong Choai[3]

[1]Faculty of Engineering, National Defence University of Malaysia (NDUM), Malaysia
[2]Faculty of Management Studies, National Defence University of Malaysia (NDUM), Malaysia
[3]Secuglass Sdn. Bhd., Malaysia

[*]Email: alias@upnm.edu.my

ABSTRACT

For many years glass has been one of the most widely used construction materials for building facades. However, glass as a building material is very brittle, when explosion occurs, the air blast pressure, typically fractures glass windows, which might kill people and damage the surrounding areas. Evidence obtained from several terrorist attacks on building~~s~~ facades worldwide support this view. Currently, blast resistant glass panels are produced using polyvinyl butyral (PVB) interlayers and manufactured using the lamination process. This process requires heating and also an autoclave system. Using this procedure is expensive and is also not environment-friendly. Thus, researchers from the National Defence University of Malaysia (NDUM) collaborated with a glass manufacturer in Malaysia, Secuglass Sdn. Bhd. to develop a blast resistant glass panel, using green manufacturing process, which is more cost effective and environment~~-~~friendly. This process does not require heating as well as an autoclave for the lamination. With this new manufacturing process uses polyurethane resin as an interlayer to provide protection against blast explosions. In this process, the polyurethane resin is poured into the cavity between two sheets of glasses that are held together until the resin cures. In this research, the glass panel produced from this process was tested for resistance against blast loading by using actual explosives. A field blast test was conducted on the laminated glass with polyurethane resin as an interlayer. The test was carried out in accordance with the ASTM F 1642-04 international testing standard for blast resistant glazing. The blast test results showed that using the laminated glass with polyurethane resin interlayer by using the green manufacturing process, the blast was able to withstand the peak overpressure of 305.09 kPa and reflected pressure of 4,688.43 kPa. In conclusion there is a potential for this manufacturing process to be used in the production of blast resistant glazing with polyurethane resin as an interlayer.

Keywords: *Green manufacturing technology; Polyurethane resin: Polyvinyl butyral (PVB); laminated glass; air blast loading.*

1. INTRODUCTION

Terrorists attack on buildings and infrastructures have become a global phenomenon. In most cases, the terrorists used explosives hidden in vehicles and blew it up at a close distance from the intended targets. With intensive shock waves created by this explosion which propagates outward at a supersonic velocity accompanied by heat and light ~~that~~ induces pressure on the building structures and causes significant damage to the structure and loss of life. Another method is to protect the buildings from damages by incorporating a blast resistance design, blast resistance materials and also retrofitting of the existing structure (Alias *et al*., 2015). This area of research is currently receiving more attention from many structural engineers as they began to consider the blast resistance materials in their designs in order to protect important buildings and structures from such attacks (Ghani Razapur *et al*., 2007). Glass is one of the most widely used construction material for building facades. When explosions occur, the air blast pressure typically fractures windows that might kill the people. This can be seen from several terrorist attacks on buildings.

Very often, annealed glass is used in windows due to its low cost. However, the material is brittle in nature, and thus it offers only little resistance to the blast pressure produced by explosions. When the annealed glass is subjected to explosions, it will break into very sharp fragments that can travel at a very high velocity and cause~~s~~ injuries to human beings. Historically, the majority of injuries from bomb blasts have been from flying glasses. This can be seen from the damaged glass building in Norway, 2011. The bomb was placed inside a van next to the tower block, housing the office of the Prime Minister, Jens Stoltenberg. The explosion killed eight people and injured at least 209 people (Guardian, 2011). The damage to the building was due to the terrorist attack as shown in Figure 1.

Figure 1: Damaged to the building in terrorist attack in Norway (Guardian, 2011).

Currently, the blast resistant glass used in most buildings is made from laminated glass. Laminated glass consists of two or more panes of glasses with one or more layers of polyvinyl butyral (PVB) sandwiched in between them and heated in autoclave. PVB is the most common interlayer used and it is bonded between the glass layers by the application of pressure and heat (Hopper *et al*., 2012). The typical laminated glass is shown in Figure 2.

The manufacturing process for laminated glass requires both heating and autoclave. The lamination process comprises of glass loading, glass surface preparation, assembly of glass/pvb/glass, de-airing and glass-sealing, unloading and autoclave. This process is expensive and not environment-friendly. The current process of producing laminated glass is shown in the Figure 3.

Alias *et al*. (2015) conducted field blast test on the laminated glass panel using polyurethane resin as an interlayer and also annealed glass panel which acted as a control sample. In his research a total of four glass panels which consisted of one 7.52 mm thick annealed glass panel and three 7.52 mm thick laminated glass panels were subjected to air blast loading with explosive charge weight, ranging from 225 to 600 g at a standoff distance of 1,500 mm. The blast test results showed that the 7.52 mm thick annealed glass was damaged when it was subjected to peak overpressure up to 250 kPa resulting from 225 g of explosive. Meanwhile, the 7.52 mm thick laminated glass with polyurethane resin interlayer survived the peak pressure up to 650 kPa resulting from 600 g of explosive. The *polyurethane* resin interlayer has successfully retained the glass fragments and also reduced the risk of cutting injuries from the glass breakage. The results showed that the laminated glass panel using polyurethane resin as an interlayer has a potential to be used as a blast resistant laminated glass panel.

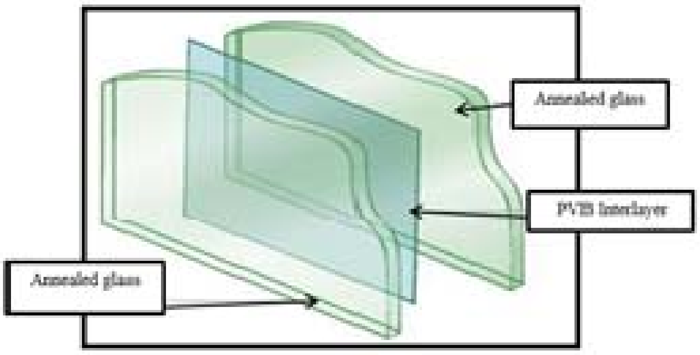


Figure 2: Laminated glass.

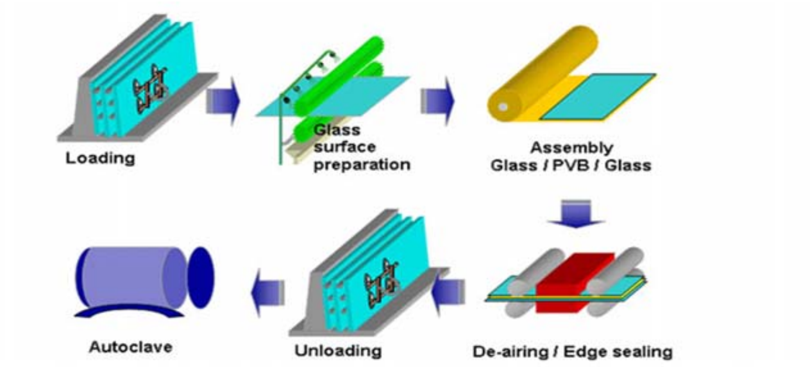


Figure 3: Existing process to produce laminated glass.

Larcher *et al.* (2016) carried out a numerical stimulation to assess blast loaded laminated glass windows with PVB interlayer and to be used under certain circumstance to determine the hazard levels. The researches were based on the ISO 16933 standard. The results showed that the hazard levels are normally established through experimental work. The fragmentation of laminated glass still cannot be represented very well by numerical simulations. Therefore, the researchers concluded that the numerical simulation can only be used as a supplement to the experimental investigation. Biolzi *et al.* (2018) have investigated on laminated glass under static and impact loading. The structural performance of laminated glass for different glass thickness and interlayer was used in this research. The static test was performed with a loading system made of four pneumatic jacks attached to the strong floor that applied the load. As for the impact test, the researchers used 1 kg of hard and semi rigid tempered steel ball as impactor. The laminated glasses were repeated for three times at three different contact spots. The researchers concluded that the stiffness of the SGP interlayer was better than PVB as interlayer.

The objective of this research is to develop a blast resistant glass panel using green manufacturing process which is more cost effective and environment-friendly. This method has the potential to save the manufacturing cost up to 30% and also reduce the emission of CO2 as it does not use the autoclave equipment which is very expensive and also contributes to the emission of the CO2 gas.

2. METHODOLOGY

The methodology of this research is divided into three major parts. The first part is to produce the glass, using the polyurethane resin with green manufacturing process. The green manufacturing process is referred to the process and does not require heating and also autoclave for the lamination. This new manufacturing technology uses polyurethane resin as the glass interlayer. In this new process the resin is poured into the cavity in between two sheets of annealed glasses that are held together until the resin cures. The injection of resin in between the glass interlayer is shown in Figure 4.

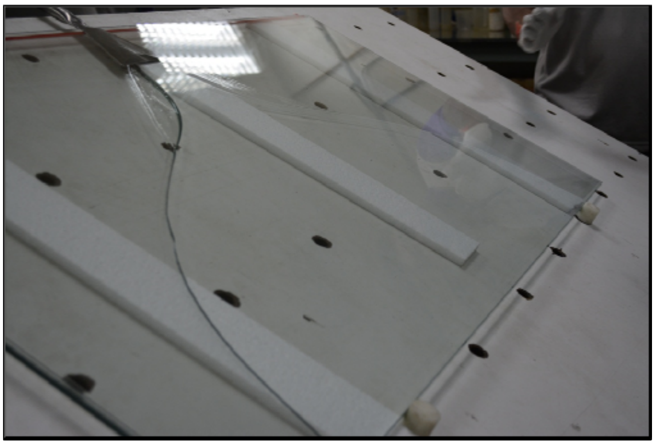


Figure 4: Injection of resin in between the glass.

This process does not require heating and also autoclave for the lamination process. This new manufacturing process involves glass surface preparation, edge fastening, injection of resin, curing and finally unloading. The new green manufacturing process is to produce blast resistant laminated glass as shown in Figure 5.

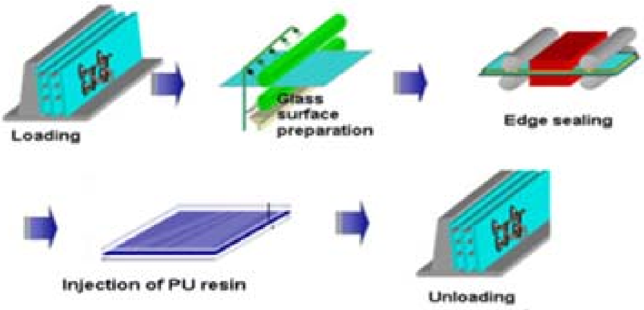


Figure 5: Green manufacturing process for blast resistant laminated glass.

In the second part of the research, the mechanical properties of the glass were obtained using several international testing standards. Modulus of rupture of laminated glass was obtained based on ASTM

C158-02 and the compressive strength was determined based on ASTM C39. The results of the test are tabulated in Table 1.

Table 1: Mechanical properties of laminated glass panel.

Laminated Glass	Strength
Stress-Strain Diagram (MPa)	2.95
Modulus of Rupture (MPa)	70.90
Compressive Strength (MPa)	926.33

In the last part of the research, a field blast test was conducted using ASTM F 1642 04 (2010). This is the standard test method for glazing system which is subjected to air blast loading. In this experiment, a total of three glass panels consisting of 6.76 mm of laminated glass with polyurethane resin interlayer with the size of 900 mm x 1,100 mm were prepared by Secuglass Sdn Bhd for testing. These glass samples were mounted to a frame in a manner, consistent with the installation in the field. A witness glass panel was placed at a distance of 3 m from the window test panel. The witness panel consists of a 2.5 cm thick layer of aluminium faced extruded Styrofoam insulation. The witness panel serves to record the presence of fragments that impinges upon its surfaces. Instrumentation to record the blast wave pressure time loading was required to assure the desired loading was achieved. A cross-sectional representation of the testing facilities can be seen in Figure 6.

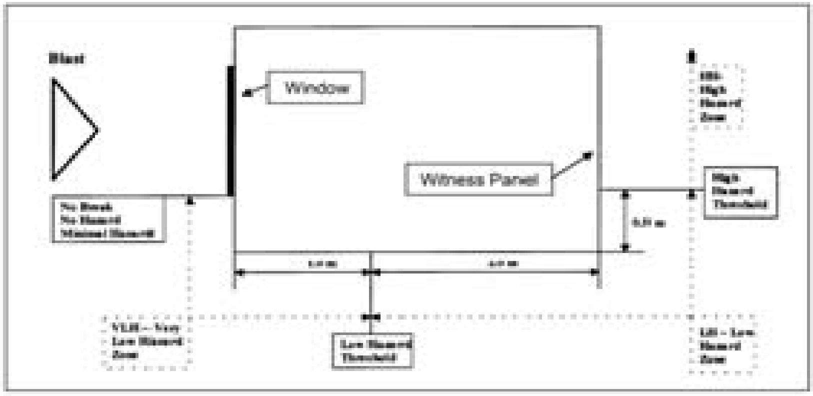


Figure 6: ASTM F 1642 04 (2010) standard test method for glazing system subjected to air blast loading.

A blast testing frame was fabricated according to ASTM F 1642 04 (2010). In this field blast test, there were three types of sensors that have been used to capture the blast related data, which are Piezoelectric ICP$^{®}$ pressure sensor, Piezoelectric ICP$^{®}$ reflected pressure sensor, and also pressure probes. Two numbered, Piezoelectric ICP$^{®}$ pressure sensors were mounted on both sides of the glass panels and number one of Piezoelectric ICP,$^{®}$ the reflected pressure sensor was mounted on the top middle of the glass panel. In addition to this, two of pressure probes were located at a distance of 1380 mm. These sensors were connected to the signal conditioning module and then to the high speed data acquisition system. The blast results were displayed using LabVIEW program. This glass panels were subjected to 400 g of C-4 explosive at a standoff distance of 1,380 mm. The testing frame and set up is shown in Figure 7.

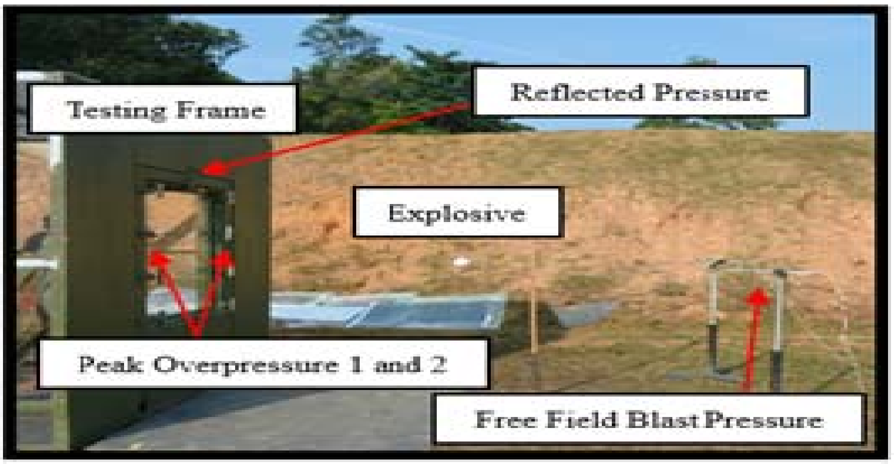


Figure 7: Blast testing frame.

4. RESULTS & DISCUSSION

A total of three sample glass panels were tested under the blast loading procedure. In this experiment, 6.76 mm of laminated glass with polyurethane resin interlayer were used. The size of the glass panel was 900 mm x 1,100 mm. These glass panels were subjected to 400 g of C-4 explosive at a standoff distance of 1,380 mm. The results for peak overpressure were recorded as 44.25 Psi (305.09 kPa) as shown in Figure 8.

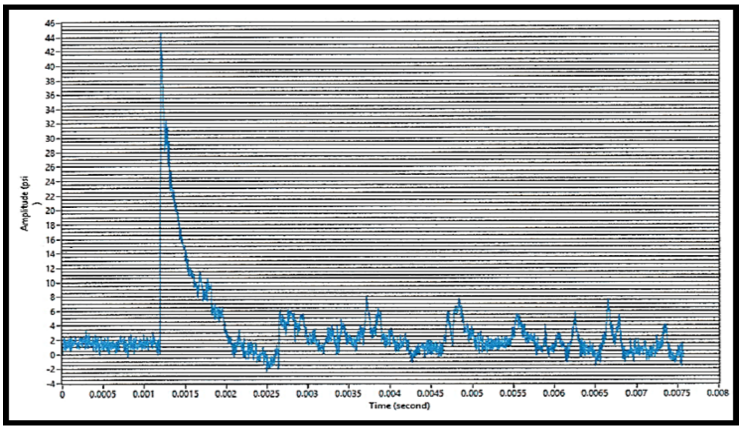


Figure 8: Results for peak overpressure.

The reflected pressure was measured using reflected pressure gauge located at the top of the glass panel frame. The reflected pressure recorded was 680 Psi (4,688.43 kPa) as shown in Figure 9.

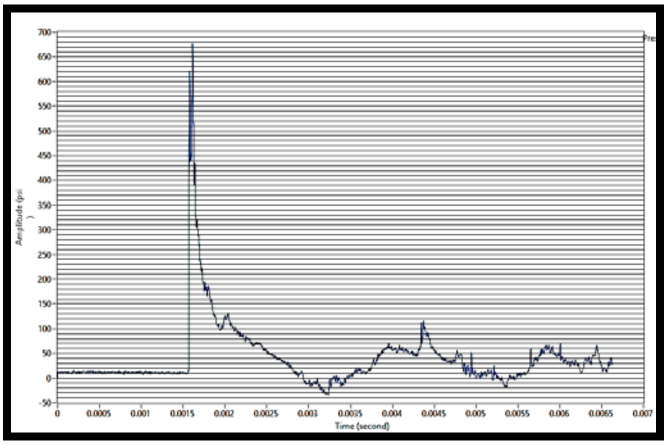
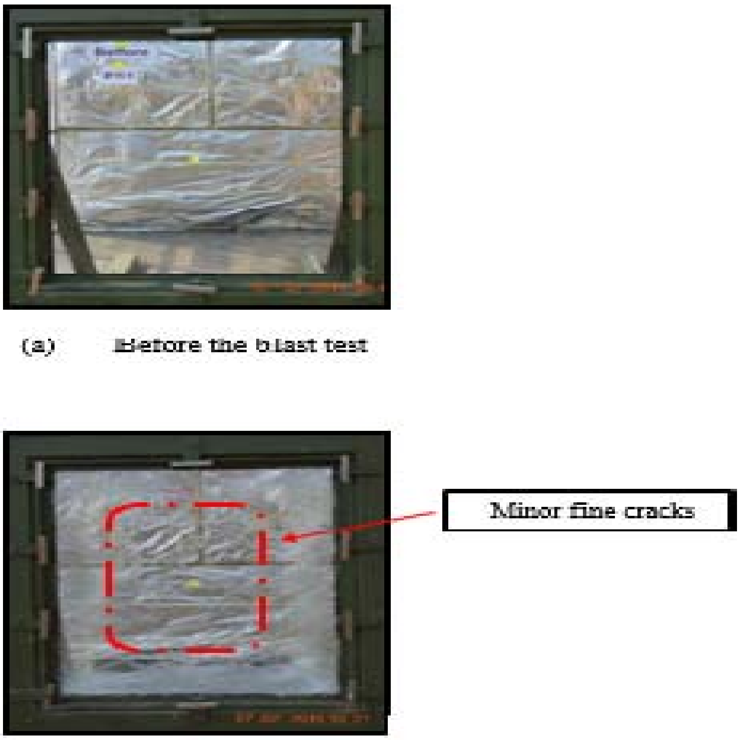


Figure 9: Results for reflected pressure.

From the field blast test conducted, it was observed that the polyurethane resin samples did not break into pieces. The laminated glass with polyurethane resin interlayer remained intact to the frame and was able to withstand the peak overpressure of 305.09 kPa and reflected pressure of 4,688.43 kPa resulting from a charge weight of 400 g of C-4 explosive at a standoff distance of 1,380 mm. There were only minor fractures observed on the surface of the laminated glass panel. This is shown in Figure 10. In addition to this, the deflection measured was 70 mm. Based on the ASTM F 1642-04 standard, the laminated glass panel with polyurethane resin interlayer fell into category B which is not hazardous. The photographs of the glass panels before and after the blast test are shown in Figure 10.



(a) Before the blast test

(b) After the blast test

Figure 10: Photo of glass panel before and after the blast test.

5. CONCLUSION

The blast test results revealed that laminated glasses with polyurethane resin are able to withstand the peak overpressure of 305.09 kPa and reflected pressure of 4,688.43 kPa. The results showed that the laminated glass with polyurethane resin interlayer is in category B of the standard which is referred as not hazardous. In conclusion, laminated glass with polyurethane resin interlayer is strongly recommended for blast resistant application in the structures of buildings and the green manufacturing process, the laminated glass is adequate.

REFERENCES

Alias, M.Y., Norazman, M.N. & Arifin, I. (2015). Behaviour of laminated glass panel using polyurethane resin as an interlayer subjected to air blast loading. *J. Sci. Res. Dev.*, **2:** 14-18.

ASTM C 158-02 (2017). *Standard Test Methods for Strength of Glass by Flexure*. American Standard Testing Method (ASTM), US.

ASTM C39 (2018). *Standard Test Method for Compressive Strength*. American Standard Testing Method (ASTM), US.

ASTM F 1642-04 (2010). *Standard Test Method for Glazing and Glazing Systems Subject to Air Blast Loadings*. American Standard Testing Method (ASTM), US.

Hooper, P.A, Sukhram, A.M., Blackman, B.R.K. & Dear, J.P. (2012). On the blast resistance of laminated glass. *Int. J.Solid Struc.*, **49**: 899-918.

Ghani Razapur, A., Tolba, A. & Contestabila, E. (2007). Blast loading response of reinforced concrete panels with externally bonded CFRP laminates. *J. Comp.*, **38:** 535-546.

Guardian (2011). *Norway attack.* Available online at : https://www.theguardian.com/world/2011/jul/23/norway-attacks (Last access date : 25 May 2020)

Biolzi, L., Bonati, A. & Cattaneo, S. (2018). Laminated glass cantilever plates under static and impact loading. *Adv. Civil Eng.*. **3**:1-11

Larcher, M., Solomos, G., Casadei, F. & Gebbeken, N. (2016). Experimental and numerical investigations of laminated glass subjected to blast loading, *Int. J. Impact Eng.*, **38**: 42-50.

OPTICAL OBSERVATION OF DETONATION OF SHALLOW BURIED CHARGE IN SANDY SOIL

Zulkifli Abu Hassan, Aniza Ibrahim* & Norazman Mohamad Nor

Department of Civil Engineering, Faculty of Engineering, National Defense University of Malaysia (NDUM), Malaysia

*Email: aniza@upnm.edu.my

ABSTRACT

Small-scale blast tests were carried out to observe and measure the influence of sandy soil on explosive blast intensity. The tests were carried out to simulate the blast impact imparted by an anti-vehicular landmine on a vehicle with reference to light armoured vehicles. A steel apparatus with a scale factor of 1:10 and weighing about 22 kg was used to represent the size and weight of a vehicle, and a mass of 20 g high explosive charge was used as a surrogate landmine. The observations and measurements were made by optical method using a high-speed video camera. The time of occurrence of the three phases of detonation in soil for upward translation time of the test apparatus was recorded. The recorded flight time and peak height reached by the apparatus was used to determine the energy transfer and initial velocity. The data for detonation in sandy soil was compared with those for air blast detonation. At an identical stand-off distance, the blast intensity of detonation in sandy soil is higher than that of air blast detonation. Based on the optical observation and quantified data, the effect of soil in amplifying blast intensity is apparent and may be attributed to the effect of soil funnelling on blast wave and from the impact of soil ejecta.

Keywords: *Buried explosive; small-scale experiment; tropical soil; blast intensity; soil ejecta.*

1. INTRODUCTION

Light armoured vehicles (LAV) are designed to have reliable protective systems that enable them to operate effectively in the battlefield and landmine infested areas and ensure the safety of occupants as well as to prevent severe injuries or casualties. Most of the LAVs in commission are certified based on blast test regulatory standards such as AEP 55 (NATO, 2006) and RSA-MIL-STD-37 (Nell, 2000), or are verified in accordance with internationally accepted test standards. Each regulatory standard has its own testing requirements; for instance, in AEP 55, simulation of landmine blast effects is performed by detonating explosive charge either in a steel pot or in a saturated gravel test bed. Detonations in these mediums are highly repeatable and the imparted overpressure is expected to be consistent.

LAVs are designed to conform to a certain level of protection, but this does not necessarily mean that they are able to withstand similar landmine threat, especially when they are deployed in areas with different ground conditions. Previous researches have shown that the magnitude of a blast effect is affected by soil conditions (Clarke *et al*. 2017, 2020; Ambrosini & Luccioni, 2019), and that the impact of a similar landmine explosion on the undercarriage of an LAV could be intensified if it was not abated by ground conditions. LAVs must be agile in terms of performance and have a protective system that ensures better survivability. To achieve these objectives, data on the effects of various ground conditions is of critical importance in the effort to design optimal protective systems through refined numerical simulation.

In this study, small-scale blast tests were carried out using portable test apparatus to measure the effects of sandy soil on shallow-buried explosive blast intensity. A high-speed video camera was used to record the movements of the apparatus during the tests, and the imparted energy was measured using an optical method. This paper presents the findings of experimental tests carried out to determine the effects of blast loading from the detonation of shallow-buried explosive on energy transfer and the initial velocity of the test apparatus in relation to the detonation phase of an explosion in the soil.

2. EFFECT OF SOIL PROPERTIES

Soil properties influence the way a blast load concentration, which consists of detonation products and momentum of ejecta mass, is forced upwards (Tremblay *et al*., 1998; Deshpande *et al*., 2009). The redirection of blast load is also influenced by the formed crater, which in turn is influenced by soil properties and depth of burial of explosives (Bergeron *et al*., 1998; Ambrosini *et al.*, 2003). Although soil properties and depth of burial of explosives have a separate influence on the magnitude of a blast load, their impact is usually interrelated and interdependent.

The moisture content, characteristics, and particle size distribution of soils play an important role in altering the effects of landmine blast. Soils with higher moisture content or saturation level can increase the blast impulse multiple folds (Hlady, 2004; Fourney *et al*., 2005). The blast impulse from the detonation of an explosive charge buried in fine-grained soils is higher than that buried in coarse-grained soils (Hlady, 2004; Ehrgott *et al*., 2011). Nevertheless, some research suggested that the consistency of blast impulse intensity is determined by the grain size distribution of soils, where soils with uniform characteristics typically contribute towards consistent intensity of blast impulse compared to non-uniform soils (Clarke *et al*., 2015). Therefore, more in-depth investigation on the influence of fine-grained and coarse-grained soils has to be carried out to prove that the variation in blast effect is distinctly influenced by the particle size distribution of soils.

The grain size of soil particles influences the way blast load concentration is forced upward. Unlike fine-grained clayey soils that are usually cohesive, the coarse-grained and cohesionless properties of sandy soils influences the soil interaction phase during the detonation of an explosive charge. The concentration of blast load consists of detonation products and momentum of ejecta mass (Tremblay *et al*., 1998; Deshpande *et al*., 2009). Crater formation and shape are influenced by the concentration and redirection of blast load as well as by depth of burial of explosives and soil properties (Bergeron *et al*., 1998; Ambrosini *et al*., 2003).

3. EXPERIMENTAL TEST

3.1 Test Apparatus

The experiment used an apparatus with a steel target plate to receive the detonation impact of an explosive. It was modelled at scale factor of 1:10 to represent the undercarriage of a vehicle; however for the purpose of blast impulse measurement and a full ejecta impact on the target plate, the target face was designed to have a square shape instead of a rectangle. The apparatus weighs about 21.74 kg, including the steel jig and four adjustable legs, target plate, and sensor adapter. Figure 1 shows the schematic drawing of the apparatus.

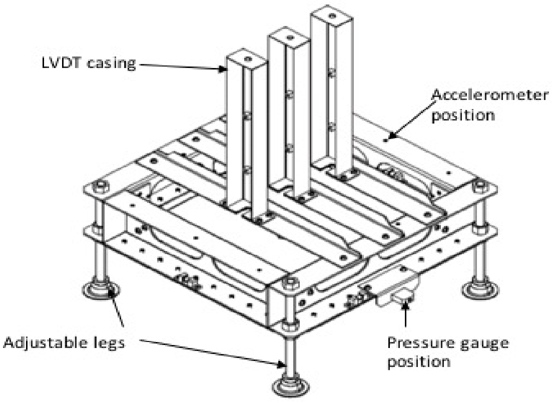


Figure 1: Schematic drawing of the test apparatus.

The steel test jig consists of 500 x 500 mm C-section steel structural members bolted to form a jig frame with a 390 x 390 mm square opening. It stands on four adjustable legs that allows for stand-off distance (SOD) variation. A 500 x 500 x 5 mm ASTM M41 steel grade sacrificial target plate is bolted to the bottom face of the steel jig and is free to move under blast load at the frame square opening.

3.2 Explosive Charge

Each blast test in this experiment was conducted by detonating a mass of 20 g ammonium nitrate (AN) emulsion high explosive commercial grade explosive charge. Figure 2 shows the charge assembly.

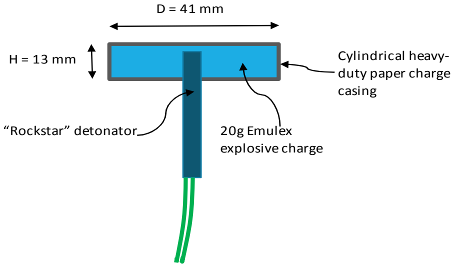

D = 41 mm
H = 13 mm
Cylindrical heavy-duty paper charge casing
"Rockstar" detonator
20g Emulex explosive charge

Figure 2: Explosive charge assembly.

The explosive charge was moulded into a disc-shaped charge using heavy-duty paper casing with a height-to-diameter (H/D) ratio of approximately 0.33. This ratio conforms to the surrogate anti-tank landmine shape described in AEP-55. It has a density of between 1.13 g/cc and 1.24 g/cc and relative bulk strength of 109. The velocity of detonation (VOD) of Emulex is between 4500 and 5500 m/s with explosion energy of around 2.85 MJ/kg. A detonator containing a secondary charge mass of 720 mg pentaerythritol tetranitrate (PETN) was inserted from the bottom at the centre of the disc-shaped charge to approximately half the depth of the charge.

3.3 Test Bed

The explosive was detonated in a soil bed of loosely compacted silica sand contained within a 1 m x 1 m x 1 m test bed pit in the ground. Table 1 presents the properties of the silica sand.

Table 1: Silica sand bed properties.

Test bed	Properties
Silica sand	Appearance: natural rounded particles. Particle size range: 0.06 - 1.0 mm, uniformly graded Specific gravity: 2.65 g/cm^3 Bulk density: 1.4 g/cm^3 Moisture content: 1.0 %

3.4 Test Setups

The experimental blast tests consist of two setups, air-blast setup (Figure 3) and buried-blast setup (Figure 4).

3.4.1 Air-Blast Setup

The apparatus was placed horizontally on a 450 mm high concrete stand. The explosive charge was attached 60 mm from the end of a polystyrene spacer block while the other end of the polystyrene spacer block was positioned and mounted at the centre of the target plate.

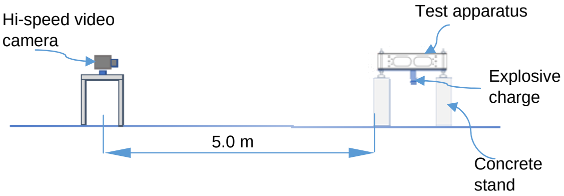


Figure 3: Air-blast setup.

3.4.2 Buried-Blast Setup

The apparatus was placed horizontally on the ground (on the sand bed). The depth of burial (DOB) was kept constant at 10 mm from the top surface of the explosive charge to the ground level, and the stand-off distance (SOD) is 50 mm from the ground level to the surface of the target plate.

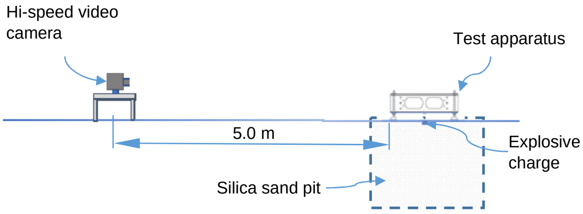


Figure 4: Buried-blast setup.

3.5 MEASUREMENT METHOD

The response of the test jig was measured by optical method using a Hi-Speed Video Camera with maximum speed of 680000 frame per second. This method allows the determination of initial velocity, v_o, by using two typical equations to calculate the v_o values (Denefeld *et al.*, 2017). The initial velocity, v_o, based on half-flight translation is given by the following equation:

$$v_o = \frac{(s_2 - s_1) + {}^1\!/_2 g(t_2^2 - t_1^2)}{t_2 - t_1} \qquad (1)$$

where s_1 is the jig position at the start of the test, s_2 is jig position at peak height, t_1 is recording time at the beginning of the test, and t_2 is time as the jig reaches peak height.

Initial velocity was also calculated based on the full-flight of the jig to narrow the observation error in the optical method. It was determined from the jig's initial position to its final position before falling back to its initial location. If t_f is the time when the jig starts to move vertically upwards to the time when it falls back to its initial position, g is gravitational constant equals to 9.81 m/s^2, and θ is angle of upwards translation, and therefore v_o is a result from the following equation:

$$v_o = \frac{t_f g}{2Sin\theta} \qquad (2)$$

4. RESULTS AND DISCUSSION

Five air-blast tests and three buried-in-sand blast tests were carried out. The response of the jig was observed from the time when it launched vertically upwards to reach peak height until it fell freely back to the point where it started. Figure 5 shows the average flight-time history of the jig in both air-blast test and buried-in sand blast test. For both tests, the time when the jig is about to move upwards is taken as 0 *s*. In the air-blast test the average peak height occurred at 56 mm, the average half-flight time was at 0.095 *s*, and the average total full-flight time is 0.19*s*. In the buried-in sand blast test, the average peak height is 245 mm while the average half-flight and full-flight time are 0.21 s and 0.4 s, respectively.

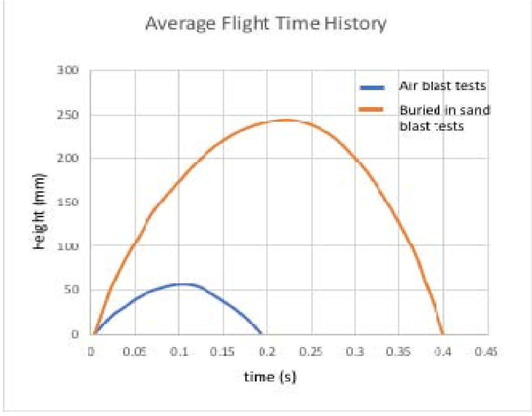


Figure 5: Average flight-time histories in air-blast test and buried-in sand test.

4.1 Detonation Phase in the Soil

Figures 6a-6b, 7a-7c, and 8 show that the flight time history coincides with the detonation phases in the soil.

4.1.1 Phase 1: Detonation and Early Interaction with the Soil

Upon ignition, the explosive was consumed by detonation wave, and the initial expansion sent pressure wave through the soil. A fireball appeared at this stage (Figure 6a), and shrank at 0.133 ms (Figure 6b); the shock wave hit the target plate at 0.233 ms and caused deformation.

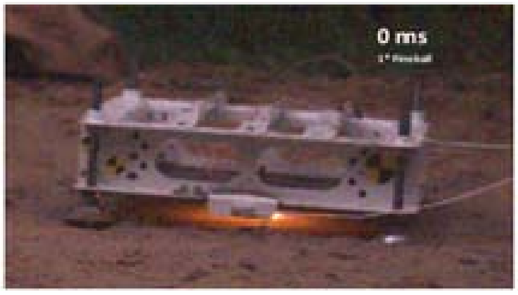


Figure 6a: Detonation of the explosive.

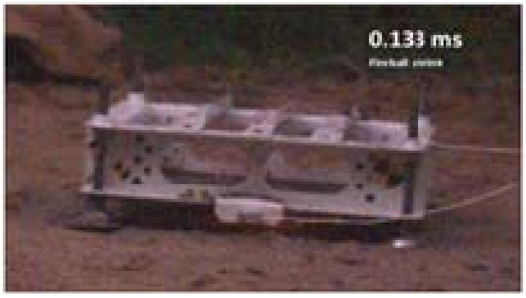


Figure 6b: The shock wave hits the target plate.

4.1.2 Phase 2: Gas Expansion

The target plate reached maximum local deformation at 0.266 ms (Figure 7a), which is the gas expansion phase. The detonation wave was reflected from the air-to-soil interface. The high-pressure gasses continued to expand and compressed the soil beneath it, and forced the soil to evacuate at high speed. The clouds of soil cap ejected from the soil surface spread around the jig (Figure 7b), and the test jig began to move upwards at 1.565 ms. In the flight-time history plot this point is considered as the beginning of flight time. Phase 2 progressed until 7.000 ms (Figure 7c) during which the test jig moved further upwards while the high-pressure gas of detonation products continue to shear the region of soil surrounding the jig and the soil on the cavity walls**.**

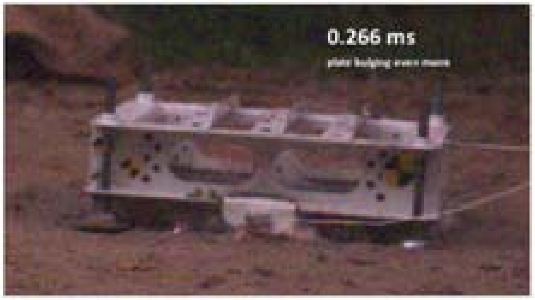


Figure 7a: Local deformation reaches maximum.

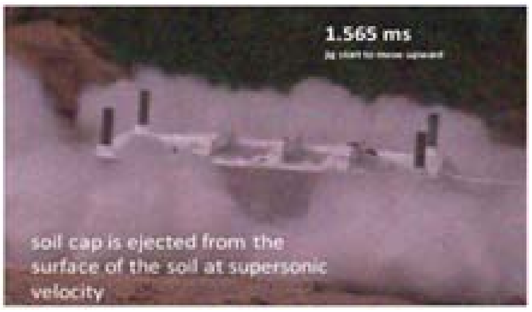

1.565 ms
jig start to move upward
soil cap is ejected from the surface of the soil at supersonic velocity

Figure 7b: Impulse transfer has ended.

Figure 7c: Apparent global movement.

4.1.3 Phase 3: Soil Ejecta

Phase 3 began after 7.00 ms and involved a complex interaction between compression wave, rarefaction wave, and the soil within the cavity. This phase is characterised by a very wide spread of the ejecta as the surrounding soil was further eroded and ejected upwards at high velocity. The observation at 160.00 ms showed a wide spread of the ejecta as the test jig moved upwards to approach its peak height (Figure 8).

With the exception of response time, the sequence of events when the target plate deformed is in agreement with the results obtained in a full-scale numerical simulation of a 17-ton vehicle subjected to 6 kg explosive blast load detonated in a steel pot (Zakrisson, 2010).

Concurrent with this sequence of events, the transfer of energy from explosive blast loading was completed after the energy has been converted into the work done to deform the target plate. This means that the energy which causes the jig to move upwards is the energy exceeding the work done; however, in the buried-in-sand test, the additional impact from soil ejecta during the upward translation of the jig continues to transmit energy.

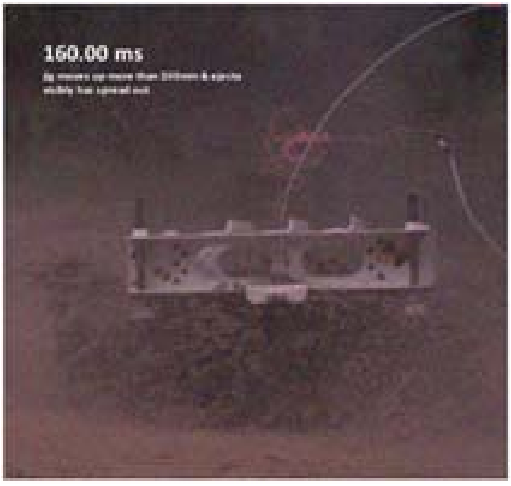


Figure 8: Global movement approaching maximum.

4.2 ENERGY TRANSFER

At a constant gravitational acceleration, g, the potential energy of a mass is the function of its height. The energy transferred from an explosion can be determined from the peak height reached by the test jig during global movement. Figure 9 shows the calculated energy transfer in both test setups. It shows that, in the buried-in-sand explosion, the average energy transfer that impacted the test jig is 52.2 J of the excess energy from the explosive; this causes the jig to move vertically upwards 4 times higher than the height in air-blast tests, which has an average energy transfer of 12 J.

4.3 INITIAL VELOCITY AND IMPULSE TRANSFER

The two calculation methods are for the half-flight and full-flight of the jig. Figure 10 shows the initial velocities of air-blast test calculated using Equation 1 for half-flight and Equation 2 for full-flight. The half-flight has a higher initial velocity value compared to the initial velocity for full-flight.

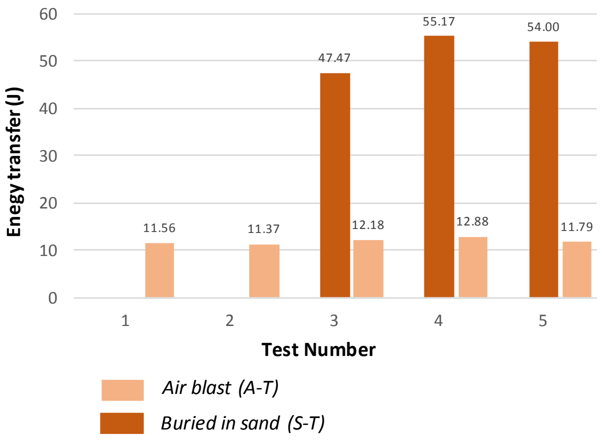


Figure 9: Average excess energy transfer in air-blast test and buried-sand test.

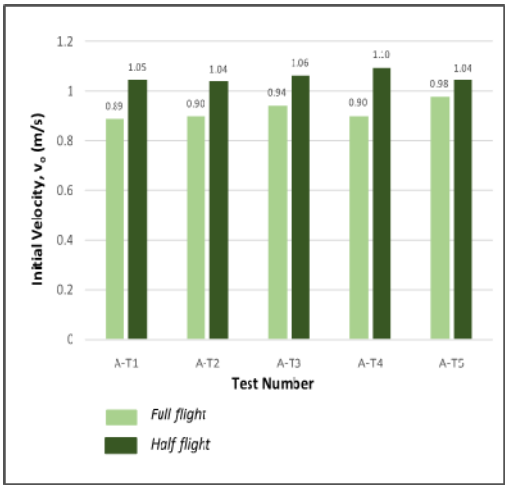


Figure 10: Initial velocity for full-flight and half-flight in the air-blast test series.

The calculated initial velocity for half-flight is higher than the calculated initial velocity of the full-flight in both test series. In half-flight, the fight time is from the moment the jig started to move upward until it reached peak height. The impulse transferred from the blast load forced the apparatus to displace vertically, and the apparatus begins to lose momentum as it reached peak height. At the peak height, the jig was observed to float in the air for about 20 ms before falling to the ground. The floating time is accounted for in the total flight time but not in the half-flight time, and this influences the value of initial velocity in the half-flight calculation. Figure 11 shows the average initial velocity obtained in both the air-blast and buried-in-sand tests.

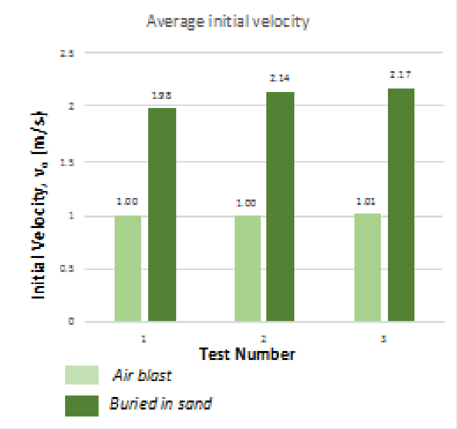


Figure 11: Average initial velocity in air-blast and buried-in-sand test.

Since initial velocity is a function of flight distance and time, the initial velocity values of the buried-in-sand test will be twice the initial velocity in the air-blast test. The average initial velocity gives a mean initial velocity of 1.0 m/s for the air-blast test and 2.1 m/s for the buried-in-sand test.

Figure 12 shows that the impulse transfer value derived from the initial velocity has a similar ratio difference between the two tests series. However, the derived impulse does not imply a total impulse transfer to the apparatus since the blast load impacted a non-rigid target. The test jig moved upwards as a result of the residual impulse from the first impact of blast wave on the target plate, and in the case of buried-in-sand test the additional impulsive load could come from the impact of soil ejecta.

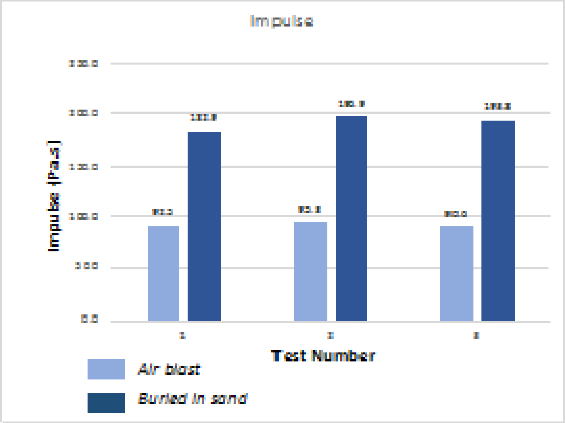


Figure 12: Impulse transfer in air-blast and buried-in-sand test.

5. CONCLUSION

The optical observation in small scale experimental tests was used to measure the influence of sandy soil on explosive blast intensity. It is important to note that the collision from the impact of buried explosive blast loading in these experimental tests is a non-rigid, or elastic, collision. During the blast test, the test plate underwent peak and residual deformation, where some of the energy was absorbed when the blast wave hit the target plate before the jig began to move upwards.

Although the measured global movement or velocity does not indicate a total energy transfer from the explosion, measurement through observation has made it possible to determine the influence of the medium where the explosive detonated. Quantified data from the experiment has demonstrated the way the bed properties of silica sand affect the change in the blast intensity of an explosion. The energy transfer is 4 times higher, and the transmitted impulse is twice greater when the explosive charge detonated in silica sand compared to in air-blast tests. During the blast phase in the soil, the impact of soil funnelling, and the impact of the ejecta may have contributed to a greater magnitude of the explosive blast output.

ACKNOWLEDGEMENT

The authors would like to express their gratitude for the financial support provided by National Defense University of Malaysia (NDUM) through UPNM/2016/GPJP/4/TK/7.

REFERENCES

Ambrosini, D., Luccioni, B. & Danesi, R. (2003). Craters produced by explosions on the soil surface. *Mecanica Computacional,* **22**: 679-692.

Ambrosini, D., & Luccioni, B. (2019). Effects of underground explosions on soil and structures. *Underground Space*, In press.

Bergeron, D., Walker, R. & Coffey, C. (1998). *Detonation of 100-gram Anti-Personnel Mine Surrogate Charges in Sand: A Test Case for Computer Code Validation*. Technical Report 668. Defence Research Establishment Suffield, Ralston, Alberta, Canada.

Clarke, S. D., Fay, S. D., Warren, J.A., Tyas, A., Rigby, S.E., Reay, J.J. & Liversey, R. (2015). Geotechnical causes for variation in output measured from shallow buried charges. *Int. J. Impact Eng.*, **86**: 274-283.

Clarke, S. D., Fay, S. D., Warren, J. A., Tyas, A., Rigby, S. E., Reay, J. J., Liversey, R. & Elgy, I. (2017). Predicting the role of geotechnical parameters on the output from shallow buried explosives. *Int. J. Impact Eng.,* **102**: 117-128.

Clarke, S., Rigby, S., Fay, S., Barr, A., Tyas, A., Gant, M. & Elgy, I. (2020). Characterisation of buried blast loading. *Proc. Royal Soc. A: Math., Phys. Eng. Sci.,* **476**: 2236.

Denefeld, V., Heider, N. & Holzwarth, A. (2017). Measurement of the spatial specific impulse distribution due to buried high explosive charge detonation. *Defence Tech.,* **13**: 1-9.

Deshpande, V.S., McMeeking, R.M., Wadley, H.N. & Evans, A.G. (2009). Constitutive model for predicting dynamic interactions between soil ejecta and structural panels. *J Mech Phys. Solids*, **57**: 1139-1164.

Ehrgott, J.Q., Akers, S.A., Windham, J.E., Rickman, D.D. & Danielson, K.T. (2011). The influence of soil parameters on the impulse and air blast overpressure loading above surface-laid and shallow-buried explosives. *Shock Vib.,* **18**: 857-874.

Fourney, W.L., Leiste, U., Bonenberger, R. & Goodings, D.J. (2005). Mechanism of loading on plates due to explosive detonation. *Fragblast*, **9**: 205-215.

Hlady, S.L. (2004). Effect of soil parameters on land mine blast. *18th Int. Symp. Military Aspects Blast Shock (MABS18)*, Bad Reichenhall, Germany.

NATO (2006*.) Procedures for Evaluating the Protection Level of Logistic and Light Armoured Vehicles*. NATO, Brussels, Belgium.

Nell, S. (2000). *Design, Development and Evaluation of Landmine Protected Wheeled Vehicles.* RSA-MIL-STD-37, Issue 3, U.S. Department of Defense (DoD), Washington D.C.

Tremblay, J.E., Bergeron, D.M. & Gonzalez, R. (1998). *Key Technical Activity 1-29: Protection of Soft-Skinned Vehicle Occupants from Landmine Effects.* Technical Cooperation Program, Val-Belair, Canada.

Zakrisson, B. (2010). *Numerical and Experimental Studies of Blast Loading*. Licentiate Thesis, Lulea University of Technology, Sweden.

CHARACTERISATION OF SHIP MAGNETIC SIGNATURES USING OVERRUN MAGNETIC RANGING IN THE EQUATORIAL REGION

Abdul Rauf Abdul Manap[1*], Hasril Nain[2], Mahdi Che Isa[2], Mohd Hambali Anuar[1], Mohd Hazri Rahmat[3], Roslan Slamatt[1] & Muhammad Syauqat Abd Khalid[2]

[1]Degaussing Centre
[2]Maritime Technology Division
[3]Research, Innovation and Strategic Direction Unit
Science & Technology Research Institute for Defence (STRIDE), Ministry of Defence, Malaysia

[*]Email: rauf.manap@stride.gov.my

ABSTRACT

Malaysia is located in the equatorial region, which experiences the largest horizontal component of the Earth's magnetic field as compared to other regions. The presence of any ferromagnetic materials will alter the existing magnetic strength distribution in surrounding region. In this study, overrun magnetic ranging procedure was conducted on six ferromagnetic ships of different classes in order to characterise their magnetic signatures. The operating histories of these ships exceed 10 years. The results from the study showed that two different types of ship magnetic signatures were identified. The first type is a mixture of longitudinal and vertical components, while the second type is a vertical component of magnetic signature. Based on the results, it can be summarised that the history of where the ship is exposed to the Earth's magnetic field along operation plays a major role in contributing to its magnetic signature characteristics.

Keywords: *Equatorial region; Earth’s magnetic field; ship magnetic signature; overrun magnetic ranging; ferromagnetic ships.*

1. INTRODUCTION

Malaysia is located in equatorial region, which experiences the largest horizontal component of the Earth's magnetic field as compared to other regions (Chulliat *et al*., 2020). Studies on the Earth’s magnetic field is very important in order to understand the oceanic magnetic environment that is experienced by ships. Besides that, understanding the molecular theory of magnetism is also important to explain the changes of material properties that are exposed to the Earth's magnetic field (Coey, 2009).

Almost all ship structures are made out of welded steel plates. Due to the ferromagnetic behaviour of steel, the ship structure is magnetised in the presence of the Earth’s magnetic field. This is due to the fact that steel has a number of complex magnetic characteristics, including hysteresis, magneto-mechanical effects and inhomogeneous magnetisation (Berti *et al*., 2015; Aydin *et al*., 2017; Kachniarz *et al*., 2018; Mahdi *et al*., 2019).

Magnetic ranging is the procedure to perform magnetic measurements (MOD, 1944; Daya *et al*. 2005). Currently, there are two types of ranging that are practiced, which are aerial and underwater ranging. Aerial ranging is a new invention of portable magnetic ranging system, whereby it uses airborne drone embedded with magnetometer and digitiser to perform the magnetic mapping in the horizontal plane (NR, 2017; Mahdi *et al*., 2019). Underwater ranging can be classified into two classes, which are static and overrun ranging. The difference between these two types of measurement is based on the operational process. The overrun system depends on the manoeuvring capability of the ship’s crew, as well as the size and time required for the ranging and deperming procedure. In terms of technical efficiency, the static system has the highest efficiency. Even though the overrun system is less efficient, its shortcomings can be controlled using other systems, including a sensor array that consists of five tri-axis sensors, and tracking system (Kama, 2019).

The aim of this paper is to characterise the magnetic signatures of six different ferromagnetic ships that operate in different territories around the world. The operating histories of these ships exceed 10 years. The ships underwent overrun magnetic ranging to capture and collect their magnetic signatures. The data collected was analysed to categorise the character of each ship and finally to identify the relationship between ships' magnetic signatures and effect of local Earth magnetic field on the ships' magnetisation.

2. METHODOLOGY

Overrun magnetic ranging procedure was performed to acquire the initial magnetic field signatures of the ships. Each ship went through two cardinal headings, which are towards the north and south directions. The ranging procedure was conducted in Malaysian waters, which has recorded a total Earth magnetic field of approximately 41,787.8 nanoTesla (nT) (NOAA, 2018). The water depth during ranging was about 18 m with sea condition at sea state 1 by referring to the Beaufort Scale, whereby the wind is described as a light air with speed in the range of 1 to 3 knots. The probable wave height is up to 0.1 m, while the state of sea is visualised as having ripples with the appearance of scales but without foam crests (NWS, 2020).

The ranging was conducted at ranging facilities where the sensor and control station are placed at specific areas. Sensor arrays where been placed underwater and connected via underwater cable to the control equipment at the shore station. The ship under test (SUT) was moved above the sensor arrays from north to south and south to north headings with a constant speed of 4 knots in order to get the magnetic signature.

The magnetic field signature was captured and recorded using a vertical axis sensor (*z*-axis) for further analysis. The reference for sensor orientation was standardised using the right-handed rectangular coordinate system (*x*, *y*, *z*). The orientation of the vertical axis sensor (*z*) was positive downwards following the North-East-Down (NED) convention and standard (Lucas & Richards, 2015; DOD, 2019). The six ships, named as M1, M2, M3, M4, M5, and M6, have their own histories in terms of location of construction, as well as operational time and area. The setup of the ship magnetic ranging is shown in Figure 1.

3. RESULTS & DISCUSSION

The magnetic ranging procedure was performed in two cardinal headings; north and south directions. The data collected for the two different headings for the six ships was plotted and analysed. In this study, only recorded data captured by the vertical axis sensor (z-axis) is used. The ship's magnetic signatures for the north and south headings are shown in Figures 2 and 3 respectively. It is observed that there are generally two types of ship's magnetic signature that can be identified. The first type is a mixture of longitudinal and vertical, while the second type is a vertical magnetic signature.

Ships M1 and M2 operated in the equatorial region and were exposed to the dominant horizontal Earth magnetic field, where the strength was: north - 41,444.7 nT, west - 79.9 nT and upward - 4,501.9 nT (NOAA, 2018). Both ships have a mixed type magnetic signature, consisting of longitudinal magnetisation (red pole at the bow) in the north heading, as well as vertical magnetisation (positive downward / red pole along the keel) in the south heading. During the north heading measurement, the effect of the horizontal Earth magnetic field towards the north contributed to the longitudinal magnetic field signature formation, which changed to the vertical magnetic field signature during the south heading measurement as illustrated in Figure 4.

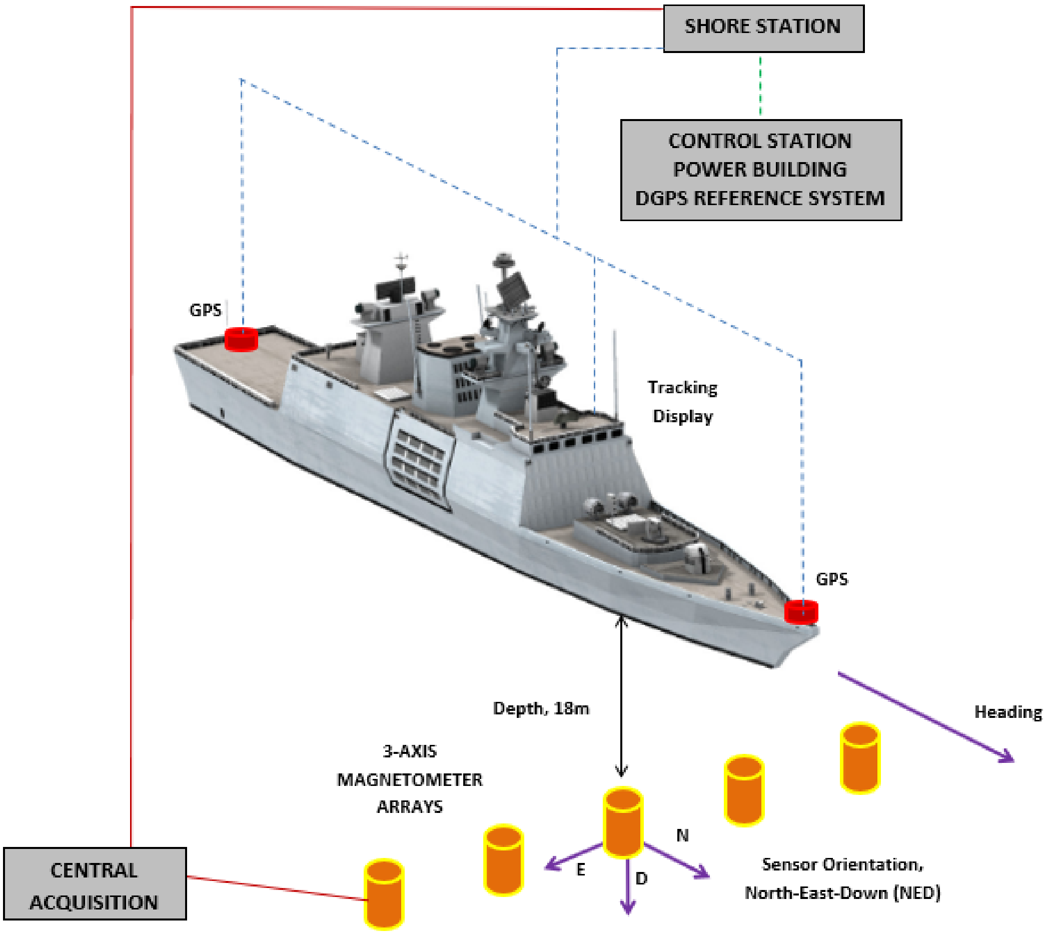


Figure 1: Typical magnetic ranging of a ship.

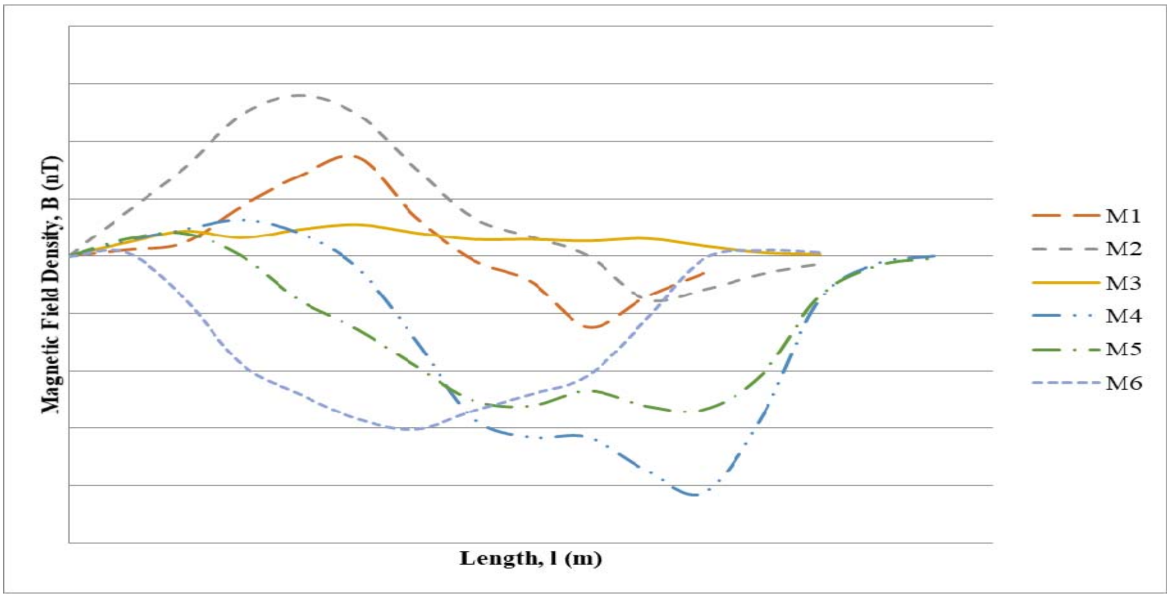


Figure 2: Recorded magnetic signatures for the north heading (*z*-axis) for the six different ships.

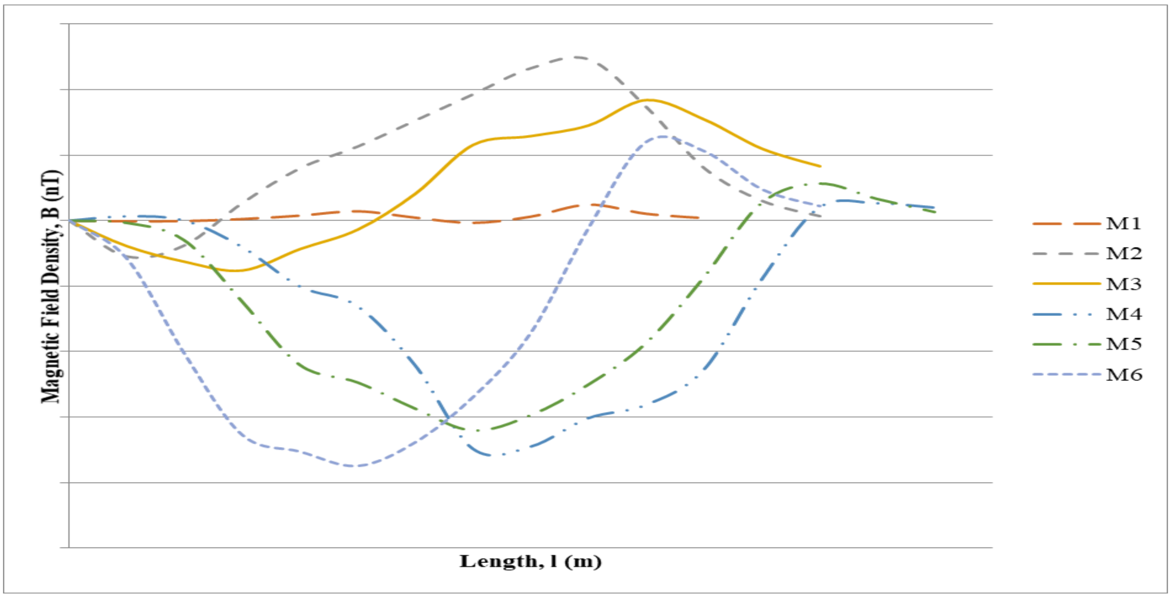


Figure 3: Recorded magnetic signatures for the south heading (*z*-axis) for the six different ships.

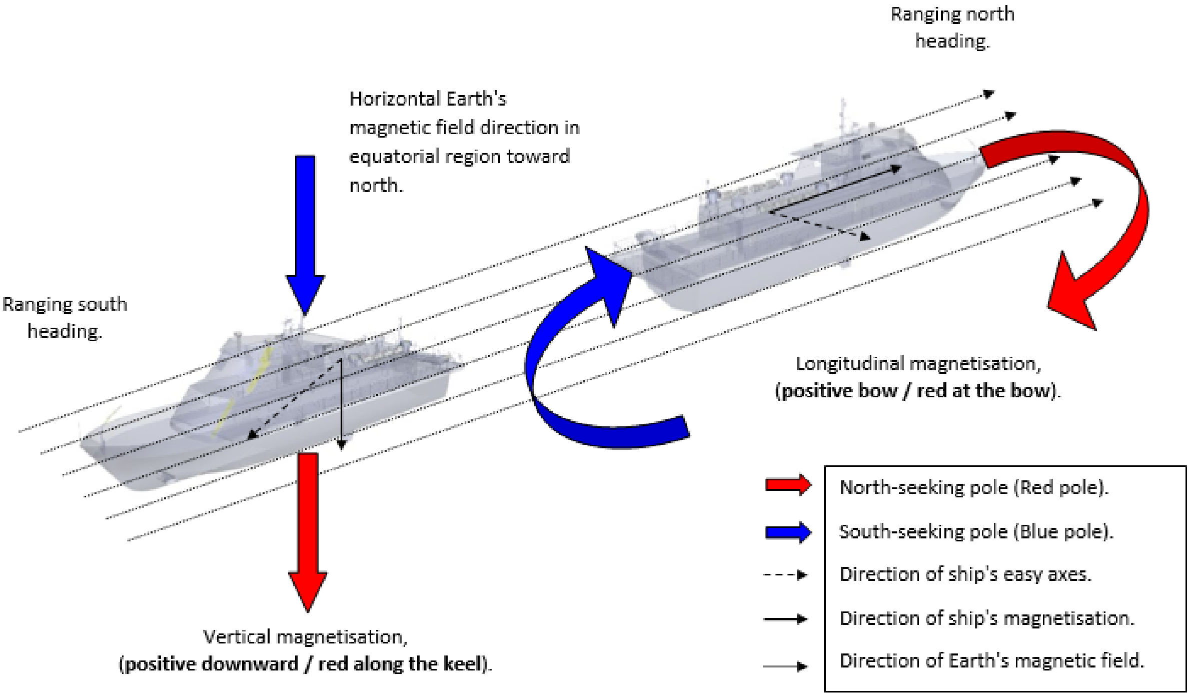

Ranging north heading.
Horizontal Earth's magnetic field direction in equatorial region toward north.
Ranging south heading.
Longitudinal magnetisation, (positive bow / red at the bow).
Vertical magnetisation, (positive downward / red along the keel).
North-seeking pole (Red pole).
South-seeking pole (Blue pole).
Direction of ship's easy axes.
Direction of ship's magnetisation.
Direction of Earth's magnetic field.

Figure 4: Magnetisation for ships M1 & M2.

Ship M3 also operated in the equatorial region. Figure 2 and 3 shows that M3's magnetisation was induced by the Earth's magnetic field where the strength was: north - 37,602.8 nT, west - 240.8 nT and downward - 28,005.8 nT (NOAA, 2018). M3's magnetisation is the opposite to ships M1 and M2, whereby it consists of vertical magnetisation (positive downward / red pole along the keel) in the north heading and longitudinal magnetisation (red pole at the stern) in the south heading, as illustrated in Figure 5.

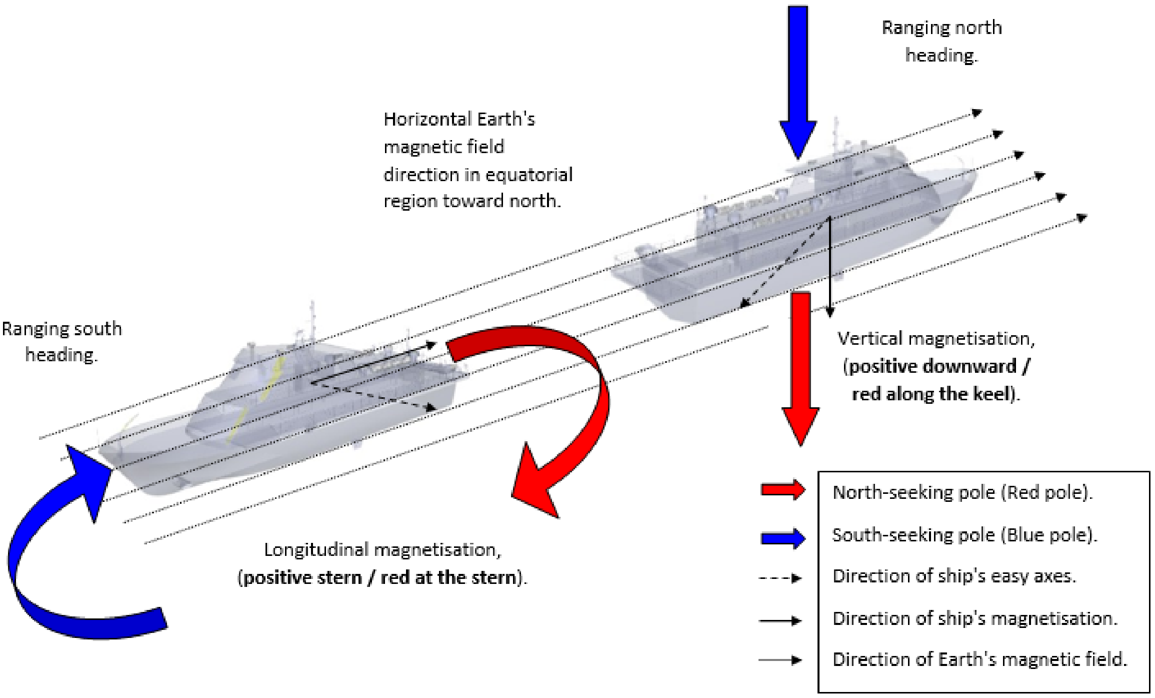

Ranging north heading.
Horizontal Earth's magnetic field direction in equatorial region toward north.
Ranging south heading.
Vertical magnetisation, (positive downward / red along the keel).
Longitudinal magnetisation, (positive stern / red at the stern).
North-seeking pole (Red pole).
South-seeking pole (Blue pole).
Direction of ship's easy axes.
Direction of ship's magnetisation.
Direction of Earth's magnetic field.

Figure 5: Magnetisation for ship M3.

Based on the magnetic signature of ships M1, M2, and M3, we are able to determine or predict the magnetic signature of any ship with the information from two headings (north and south). M1, M2 and M3 consist of longitudinal magnetisation with induced longitudinal magnetisation approximately equal to the permanent longitudinal magnetisation. The vertical magnetisation of ships can be assumed to be mostly permanent because at the ranging location, the direction of the ships' magnetisation is vertical (positive downward / red pole along the keel), which is the opposite to the vertical component of the Earth's magnetic field (positive upward). The magnitude of the ships' vertical magnetisation is significantly smaller than the longitudinal magnetisation because the ships experienced a larger horizontal component of the Earth's magnetic field.

The operational histories of ships M4, M5 and M6 show that they operated somewhere in the southern region, thus making it possible to be exposed to the maximum upward vertical Earth magnetic field for up to 67,000 nT (Chulliat *et al.*, 2015). The value of the vertical component of the Earth's magnetic field is proportional to the cosine of the angle of magnetic latitude or 'Dip' (inclination), being maximum at the magnetic poles and zero at the magnetic equator (MOD, 1975). M4, M5 and M6 were exposed to the dominant upward vertical Earth magnetic field, where the strength was: north - 24,153.6 nT, east - 5,389.3 nT and upward - 51,405.9 nT (NOAA, 2018).

The three ships produced the same magnetic field signature for both headings. They have vertical magnetisation (positive upward / red pole in the superstructure) at the north and south headings, as illustrated in Figure 6. Even though the magnetic ranging was performed in the equatorial region, it did not affect the ships' magnetic field signature. This can be explained through the hysteresis loop, whereby when a ship is frequently exposed to the highest strength of the Earth's magnetic field in southern hemisphere, the constraints holding some of the domains are overcome by the magnetic coupling acting on each domain, with the domains aligning themselves parallel to the magnetic field. The ship thus acquires a resultant magnetic moment in the upward direction of the Earth's magnetic field. Upon sailing to the equator, the magnitude and direction of the Earth's magnetic field changes and thus, certain domains return to their original random orientation, known as easy axes, while others continue to align themselves in the original or previous direction of the Earth's magnetic field. Hence, part of the ship's magnetic moment remains when the field is decreasing, which is known as remnant magnetism. A ship that sails from the southern region to the equatorial region will experience decreasing Earth magnetic field strength but retains its magnetisation condition, which indicates that the ship has permanent dominant upward vertical magnetisation (MOD, 1944; Holmes, 2008). Hence, the vertical magnetisation of the ship can be assumed as a dominant component of its magnetic signature.

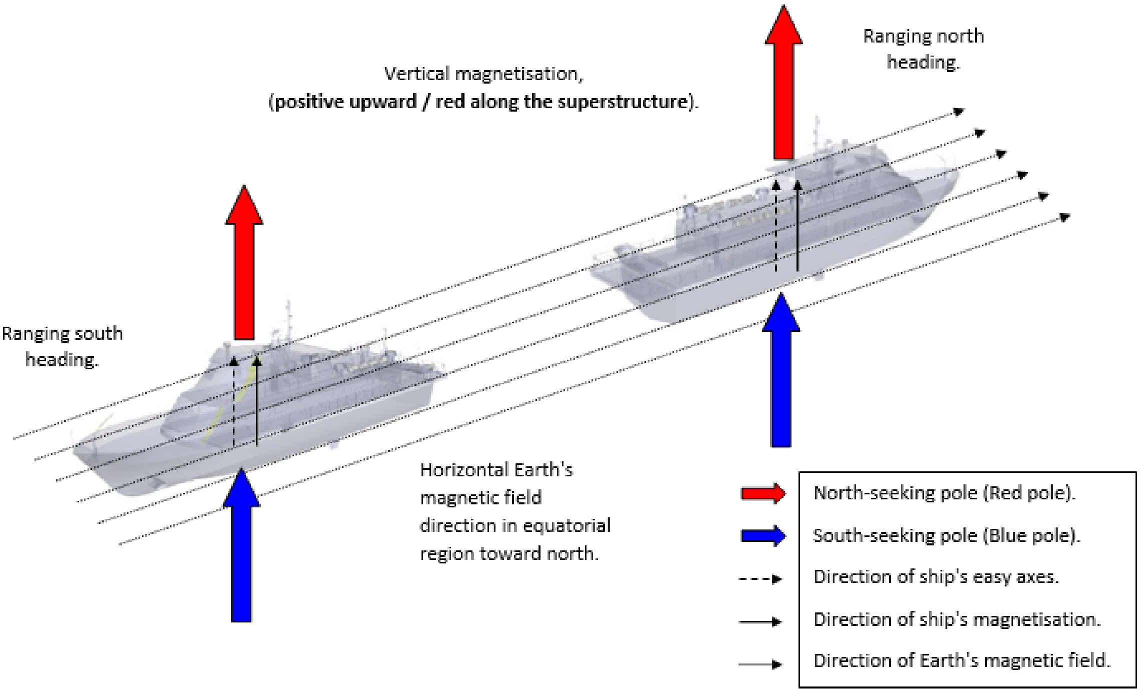

Vertical magnetisation,
(positive upward / red along the superstructure).
Ranging north heading.
Ranging south heading.
Horizontal Earth's magnetic field direction in equatorial region toward north.
North-seeking pole (Red pole).
South-seeking pole (Blue pole).
Direction of ship's easy axes.
Direction of ship's magnetisation.
Direction of Earth's magnetic field.

Figure 6: Magnetisation for ships M4, M5 and M6.

Ships with ferromagnetic hull material have large positive susceptibility to the Earth's magnetic field. Therefore, it can be magnetised much more easily than other materials with widely different degrees of ease for different values of magnetising force. It can also retain magnetisation when the magnetisation force is removed and tend to oppose a reversal of magnetisation after being magnetised (Aibangbee & Onohaebi, 2018). Due to the varying Earth magnetic field that is constantly encountered during voyages, the strength of ships' magnetisation keeps changing. The Earth's natural magnetic fields between the North and South magnetic poles are being crossed routinely while the ships are sailing. The traversing of these natural fields and ships lying dormant for certain extended periods of time during scheduled maintenance also contribute to the changes of their magnetic signatures (Mahdi *et al.*, 2019). Effect of the Earth's magnetic field show the polarity of the vertical component of a ship's permanent is somewhat more predictable than the horizontal components (Holmes, 2008).

4. CONCLUSION

Based on the results of the magnetic ranging and signature analysis conducted, the ships' magnetic signatures were found to be dependent on their operational area. Recorded data from the ranging of ships M4, M5 and M6 showed that the equatorial Earth magnetic field gave no effect to the ship that operated in the southern region because the Earth's magnetic field strength in the equatorial region is the weakest. However, for ships M1, M2 and M3 ships that operated in the equatorial region, it has a unique characteristic of magnetic signature. They have a mixed signature with longitudinal and vertical magnetic signatures or vice versa, and can easily be induced or influenced by the Earth's magnetic field. In future research, our team will apply mathematical algorithms to predict the magnetic signatures and compare the real magnetic measurements with simulation data, with aim building more appropriate algorithms to predict a ship's area of operation based on its magnetic signature obtained from magnetic ranging.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support from STRIDE's management in supporting this work. We also thank our colleagues from STRIDE's Maritime Technology Division (BTM), who provided valuable insight and expertise that greatly assisted the study, as well as comments that helped to improve the manuscript.

REFERENCES

Aibangbee, J.O. & Onohaebi, O. (2018). Ferromagnetic materials characteristics: Their application in magnetic core design using hysteresis loop measurements. *Am. J. Eng. Res.*, **7**: 113-119.

Aydin, U., Rasilo, P., Martin, F., Singh, D. & Arkkio, A. (2017). Magneto-mechanical modelling of electrical steel sheets. *J. Mag. Mag. Mat.*, **439**: 82-90.

Berti, A., Giorgi, C. & Vuk, E. (2015). Hysteresis and temperature-induced transitions in ferromagnetic materials. *Appl. Math. Model.*, **39**: 820-837.

Chulliat, A., Macmillan, S., Alken, P., Beggan, C., Nair, M., Hamilton, B., Woods, A., Ridley, V., Maus, S. & Thomson, A. (2015). *The US/UK World Magnetic Model for 2015-2020*. Technical Report, National Geophysical Data Center, National Oceanic and Atmospheric Administration (NOAA), Maryland, US.

Chulliat, A., Brown, W., Alken, P., Beggan, C., Nair, M., Cox, G., Woods, A., Macmillan, S., Meyer, B. & Paniccia, M. (2020). *The US/UK World Magnetic Model for 2020-2025: Technical Report*, National Centers for Environmental Information, National Oceanic and Atmospheric Administration (NOAA), Maryland, US.

Coey, J.M.D. (2009). *Magnetism and Magnetic Materials*. Cambridge University Press, New York.

Daya, Z.A. Hutt, D.L. & Richards, T.C. (2005). *Maritime Electromagnetism and DRDC Signature Management Research*. Defence R&D Canada (DRDC), Canada.

DOD (Department of Defense) (2019). *MIL-PRF-89500B: Performance Specification World Magnetic Model (WMM)*. Department of Defense, USA.

Kachniarz, M., Kołakowska, K., Salach, J. & Nowak, P. T. (2018). Magnetoelastic Villari effect in structural steel magnetized in the Rayleigh region. *Acta Phy. Pol. Ser. A*, **133:**660-662.

Kama, S. (2019). STL overview of deperming ranges. *Undersea Defence Tech. (UDT) Conf.* Stockholmsmässan, Sweden 13-15 May 2019.

Holmes, J.J. (2008). *Reduction of a Ship's Magnetic Field Signatures*. Morgan & Claypool, US

Lucas, C.E. & Richards, T.C. (2015). *A Novel Technique for Modeling Ship Magnetic Signature*. Defense Research and Development Canada.

Mahdi, C.I., Hasril, N., Nik Hassanudin, N.Y., Abdul Rauf, A.M., Roslan, S. & Mohd Hambali, A. (2019). An overview of ship magnetic signature and silencing technologies. *Defence S&T Tech. Bull.*, **12(2)**: 176-192.

MOD (Ministry of Defence). (1944). *BR 825(1)/ CB 3139 Manual of Degaussing: Part 1*. Ministry of Defence, UK.

MOD (Ministry of Defence). (1975). *BR 825(4) Degaussing by Magnetic Treatment*. Ministry of Defence, UK.

NOAA (National Oceanic and Atmospheric Administration) (2018). *Magnetic Field Estimated Values* Available online at: https://www.ngdc.noaa.gov/geomag-web/?model=igrf#igrfwmm (Last access date: 21 September 2018).

NWS (National Weather Service) (2020). *Beaufort Scale – Estimating Wind Speed and Sea State* Available online at: https://www.weather.gov/pqr/beaufort (Last access date: 5 August 2020).

NR (Navy Recognition) (2017). *ECA Group to Supply Degaussing & UAV IT180 Based Magnetic Ranging System to Asian Customer*. Available online at: http://navyrecognition.com/index.php/news/defence-news/2017/february-2017-navy-naval-forces-defense-industry-technology-maritime-security-global-news/4931-eca-group-to-supply-degaussing-uav-it180-based-magnetic-ranging-system-to-asian-customer.html (Last access date: 15 March 2018).

INTER-PULSE PARAMETER ESTIMATION AND RECOGNITION FOR RADAR WARNING RECEIVER

Ahmad Zuri Sha'ameri*, Boodhoo Kirish & Taha Mahmoud Al-Naimi

School of Electrical Engineering, Faculty of Engineering, Universiti Teknologi Malaysia (UTM), Malaysia

*Email: ahmdzuri@utm.my

ABSTRACT

A radar pulse train demodulated from a received signal is characterised by a stream of pulses separated at time intervals referred to as the pulse repetition interval (PRI). Estimation and recognition of the PRI type from the received radar pulse train could be used by a radar warning receiver (RWR) on an aircraft that allows the pilot to decide on the appropriate course of action. A methodology is presented to implement an RWR processing element that estimates parameters and recognises the PRI types, such as constant, 2-level staggered, wobulation and jitter. Verification of the RWR processing element based on Monte Carlo simulation shows that the recognition rate obtained at signal-to-noise ratio (SNR) above 8 dB is close to 100% for constant PRI type and between 75 to 90% for the rest of the PRI types. The complex pulse train structure of the other PRI types contributes to lower recognition rate.

Keywords: *Time of arrival (TOA); pulse repetition interval (PRI) recognition; Neyman-Pearson principle; statistical signal processing; radar warning receiver.*

1. INTRODUCTION

Since World War II, radar is widely used for both civilian and military applications. The employment of radar in for military applications has led to the development electronic warfare (EW) to counter its use (Wiley, 2006). There are three main subdivisions in EW (Genova, 2018): electronic attack (EA), electronic support (ES) and electronic protection (EP). Between these three areas, ES, which is the acquisition of the adversary's electromagnetic profile, is the key step before EA through the use of countermeasure to deny access and EP to enable the continual use of the electromagnetic spectrum. Radar warning receiver (RWR) is a specialised form of ES that is installed on aircrafts specifically designed to provide warning of possible threats from radar guided weapons (Guo & Tracey, 2020). With the warning issued, the next course of action for the aircraft's pilot is to distract the guided weapon using either evasive manoeuvres or activate countermeasures. Due to space constraints on aircraft, the RWR should be compact, but still provide results in real-time.

The major components of an ES system or RWR are signal reception, parameter estimation and recognition. Various works have been reported in this area, including instantaneous frequency measurement (Lin *et al.*, 2020), error reduction (Jawad *et al.*, 2020), deinterleaving for PRI estimation (Ge *et al.*, 2019; Lin et al., 2020), detection of low probability intercept radar waveforms (Neves *et al.*, 2016), wavelet based parameter estimation (Gençol *et al.*, 2016)' and PRI recognition (Jordanov *et al.*, 2016; Ahmed *et al.*, 2018, 2019; Cain *et al.*, 2018). The first stage in an ES system is to capture the signal over a broad frequency range covering from 0.5 to 20 GHz (Numlk, 2000). A channelised receiver is among the methods used for this purpose by utilising multiple narrowband receivers allocated to specific frequency bands (Gong *et al.*, 2013). However, an alternative approach using photonics was proposed by Lin *et al.* (2020), which allows the use of a single wideband receiver. For low probability of intercept (PRI), where the peak power is low, time-frequency analysis is used based on the wavelet transform to improve signal detection (Neves *et al.*, 2016). Once the radar pulse train is detected, errors could occur due to noise that results in missing or spurious pulses. Improved detection of pulses is achieved by performing compensation before actual PRI estimation (Jawad *et al.*, 2020). Receiving over a wide frequency range results in the reception of multiple radar pulse trains that has to be separated through a process of deinterleaving to allow the estimation of PRI parameters for each radar type. Various methods described includes improvement of the sequential difference histogram (Liu & Zhang, 2018) and pulse correlation to compensate for the location of missing or spurious pulses (Ge *et al.*, 2019). Another work on time-frequency analysis employs the wavelet transform to improve parameter estimation of radar pulse trains (Gençol *et al.*, 2016).

Recognition forms the last stage in the function of an ES system. Conventional methods for recognition are less complex as compared to machine learning and do not require training data. Recent work by Ahmed *et al.* (2018) proposed a threshold based cascaded PRI recognition scheme where recognition is performed according to the possible signal types. A review of related work can be found in Ahmed *et al.* (2019). Machine learning approaches were described using convolutional neural network (Cain *et al.*, 2018), as well as neural networks, support vector machines and random forests by Jordanov *et al.* (2016). Despite their complexity, machine learning approaches give better recognition accuracy as compared to conventional approach. However, complexity would be an issue if implemented on RWR instead of ES due to space constraint and the emphasis on real-time implementation.

An algorithm suitable for implementation in an RWR processing element is presented in this paper that recognises the various PRI types such as constant, 2-level staggered, wobulation and jitter. Thus, this paper is organised as follows to cover the description of the algorithm and its verification starting with Section 2 that describes the architecture of the RWR, followed by Section 3 that defines the signal model, detection and problem statement. Parameter estimation and PRI recognition are discussed in Section 4, while the performance of the algorithms with application in a practical scenario are presented in Section 5. Finally, the paper is concluded in Section 6.

2. SYSTEM ARCHITECTURE

Figure 1 shows the block diagram of single channel architecture from an RWR channelised receiver, where the signals are received at the RF front end with respect to their frequency and direction of arrival. At the output, the signal is demodulated with its carrier removed to produce a train of pulses that will be referred to in this paper as a radar pulse train. The different stages that are performed are signal detection, pre-processing, parameter estimation and PRI type recognition, which constitutes the RWR processing element. Details on the function of each stage, which is the main contribution of this paper, is described in Figure 2.

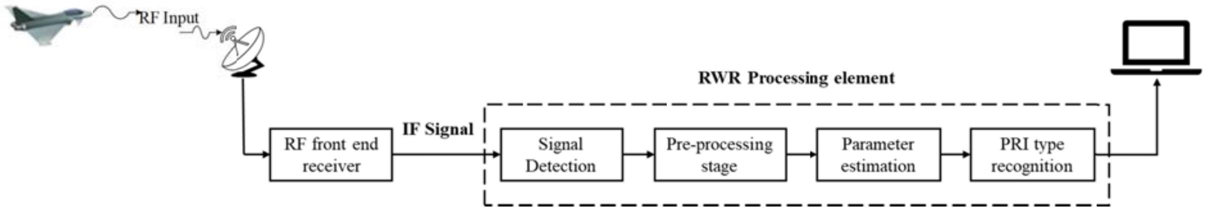


Figure 1: Single channel RWR channelised receiver architecture (Neri, 2006).

The signal detection is optimised by minimising the probability of false alarms according to the Neyman-Pearson criteria. Once the radar pulse train is detected, the pre-processing stage first performs deinterleaving followed by time of arrival (TOA) and PRI estimation. Next, histogram analysis (Ata'a & Abdullah, 2007; Liu & Zhang, 2018) determines the distribution of PRIs from the radar pulse train, which is the preliminary step to discriminate between constant PRI type with the other types. The parameter estimation stages contain more functions relative to the pre-processing stage by performing statistical analysis, such as the computation of average and variance for PRI and instantaneous TOA difference, as well as the estimation of wobulation frequency from the power spectrum of the instantaneous TOA difference. All of these parameters are used as input vector to the PRI type recognition stage. Once the signal is recognised, the results could be passed to the pilot to decide the next course of action or integrated as part of a weapon system to activate the countermeasures. Further details on the various stages of the RWR processing element is described in Section 3.

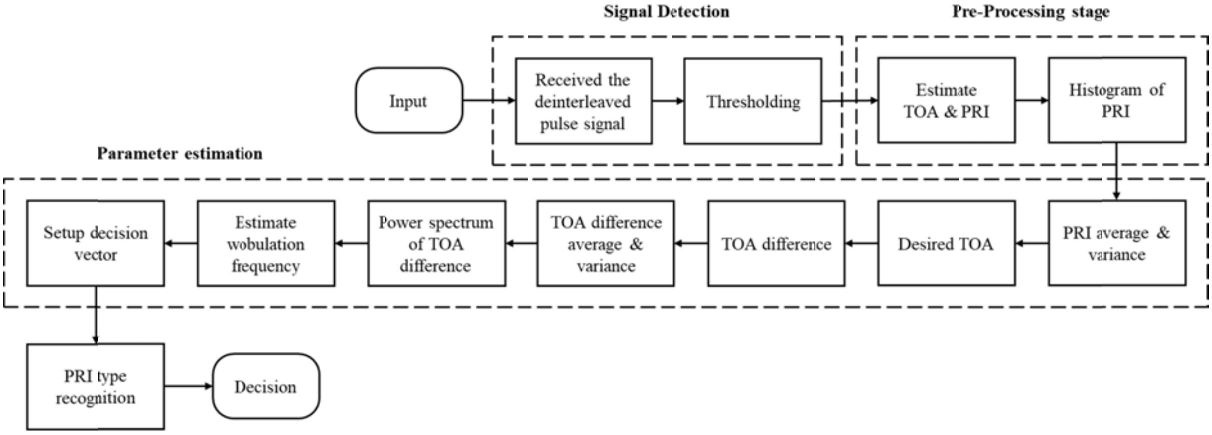


Figure 2: Block diagram of the RWR processing element showing the various stages starting with signal detection, pre-processing, parameter estimation and PRI type recognition.

3. SIGNAL MODEL AND DETECTION

This section introduces the various PRI types that are used in this paper, the methodology for detecting the radar pulse based on the Neyman-Pearson criteria, and the problem statement.

3.1 Signal Model

A major feature of an RWR is to determine the timing parameters using the inter-pulse analysis of an intercepted radar pulse train, such as the time of arrival (TOA), and consecutively recognise the PRI type, such as constant, staggered, wobulation (also referred as sinusoidal) and jitter. The function of a radar's performance is related to the PRI type (Wiley, 2006). Modern radars could have common PRI – single PRI type within a pulse train - or complex PRI – a combination of PRI types within a pulse train (Ghani *et al*., 2017). Radars used in applications such as maritime and air traffic control tend to use common PRI to reduce the cost and to perform basic target detection functions. For military applications, specifically for air defence or early warning, a complex PRI is used to minimise ambiguities, improve detection and reduced probability of intercept (POI) as well as provide target detection in three dimension (Wolff, 2020). Table 1 shows the various PRI types, as well as their characteristics and applications.

Table 1: Examples of PRI types, as well as their characteristics and applications (Wiley, 2006; Ghani *et al.*, 2017).

Pulse train	Application	Description
Constant	Search radars	The average PRI value of such signal has a variation of extremely small value (usually less than 1%), hence the average PRI here has a constant PRI.
Staggered	Eliminate blind speed in moving targets and integration radar systems	It involves usually ≥ 2 PRIs in a pulse sequence, hence the sequence will have multiple pulses and after a period the sequence will repeat.
Wobulation (Sinusoidal Variations)	Conical scan tracking system for missile guidance technique.	The PRI here is modulated in a sinusoidal variation within a period. Variation is up to about 5%.
Jitter	EP systems to overcome certain types of jamming.	The average PRI undergoes large random variations (up to 30%).

The focus of this paper is on the radar pulse train with common PRI types. With reference to Figure 2, the received radar signal at RF is removed of the carrier after demodulation and contains only the radar pulse and interference modelled as additive white Gaussian (AWGN). The four PRI types presented with respect to the timing characteristics are shown in Figure 3 and their parameters are defined in Table 2.

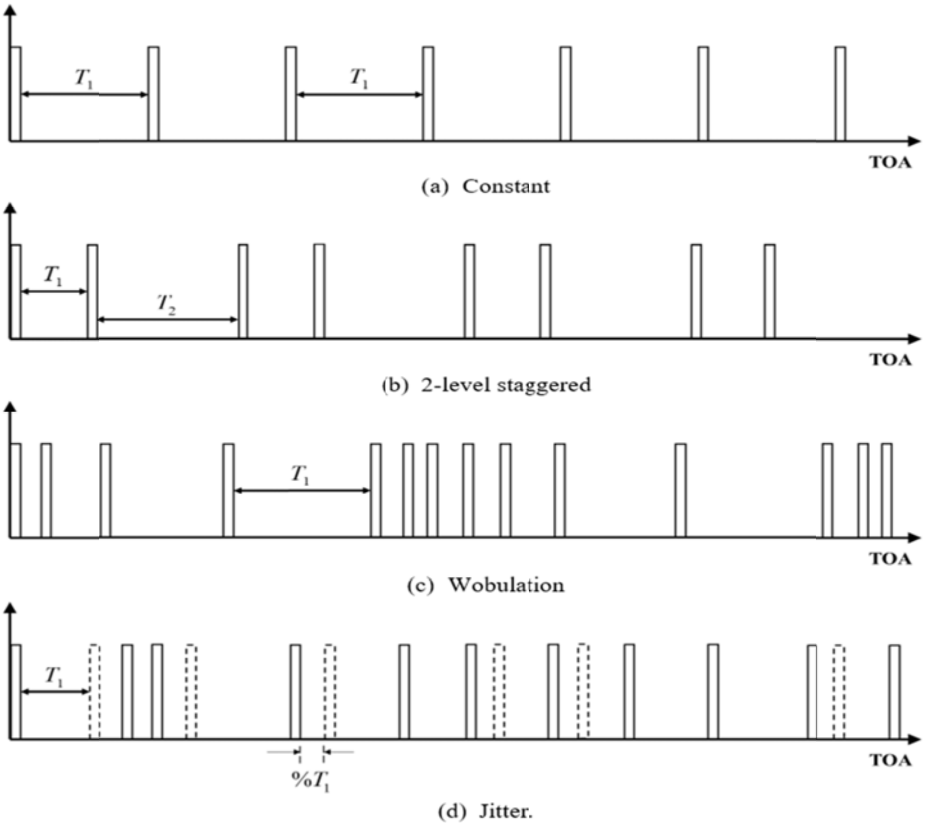


Figure 3: Example of radar pulse trains for various PRI type (Wiley, 2006; Ghani et al., 2017)

Table 2: Time parameters for various pulse train types [Pulse width τ is 1 sample, T_1 – PRI, T_2 – the second PRI for 2-level staggered PRI type, f_w - wobulation frequency]

PRI type	PRI name	PRI parameters (samples)
Constant pulse	SP1	$T_1 = 4$
	SP2	$T_1 = 8$
	SP3	$T_1 = 16$
2-level staggered	2SP1	$T_1 = 4$, $T_2 = 12$
	2SP2	$T_1 = 2$, $T_2 = 14$
Wobulation	WB1	$f_w = 1/4$, $T_1 = 16$
	WB2	$f_w = 1/8$, $T_1 = 8$
Jitter	Jitter1	$T_1 = 8 \pm 1$
	Jitter2	$T_1 = 8 \pm 2$

Radar systems in practice, such as TRS 2201, Jawor-M2 and TPS-830KE, have larger PRI than those described in Table 2 since the objective of this paper is more towards developing an analysis and recognition algorithm suitable for use in RWR. For example, TRS 2201 is a medium power radar (MPR) operating in the S-band that has a PRI of about 4 ms, while the Jawor-M2 is an L-band on board weapon control radar with PRI of 2.5 ms. TPS-830KE is a low-altitude surveillance radar that operates in the X-band with PRI of 0.4 ms. This radar could be used for coastal surveillance, air defence gap filler or point defence (Wolff, 2020). Further extension of this work is applicable to radar pulse trains similar to real radar systems.

3.2 Signal Detection

For a given time instance *n*, the received signal $x[n]$ can be expressed as:

$$x[n] = s[n] + v[n] \qquad 0 \le n \le N-1 \tag{1}$$

where *N* is duration of the received signal, *s*[*n*] is the actual signal and *v*[*n*] is interference modelled as AWGN. Two possible hypotheses defined for detecting a radar pulse are: null hypothesis H_0 indicates only noise is present while the alternate hypothesis H_1 indicates both signal and noise are present. In order to simplify the notations, a single realisation of the received signal *x* is assumed instead of *x*[*n*] and can be assigned to the hypotheses as (Srinath & Rajasekaran, 1996):

$$\begin{aligned} H_0 &: x = v \\ H_1 &: x = A + v \end{aligned} \tag{2}$$

where *A* is the amplitude of the detected radar pulse *s*[*n*]. Each of the hypotheses that represents a possible realisation is conditioned to a probability density function and the objective is to find the threshold value that will separate the signal from the noise. The resulting conditional probability density function can be expressed as (Srinath & Rajasekaran, 1996):

$$\begin{aligned} p_x\left(x \mid H_0\right) &= \frac{1}{\sqrt{2\pi\sigma_v}} \exp\left(-\frac{x^2}{2\sigma_v^2}\right) \\ p_x\left(x \mid H_1\right) &= \frac{1}{\sqrt{2\pi\sigma_v}} \exp\left(-\frac{(x-A)^2}{2\sigma_v^2}\right) \end{aligned} \tag{3}$$

where σ_v is the standard deviation of noise.

For digital communication, the priori probabilities for the transmitted symbols are equal and known (Sklar, 2001). However, this is not true for radar and thus, the Neyman-Pearson criteria is used as a detection criteria (Mahafza, 2017; Ghani *et al.*, 2017). Hence, the detection threshold γ is designed to meet the desired false alarm rate according to the following conditions (Srinath & Rajasekaran, 1996):

$$P_{FA} = \int_{\gamma}^{\infty} p_X(x \mid H_0)dx \tag{4}$$

By maximising the probability of detection P_D (Wickens, 2002), the likelihood ratio (*L*) verifies the appropriateness of the threshold according to the following:

$$L(x) = \frac{p_X(x \mid H_1)}{p_X(x \mid H_0)} \ge \gamma \tag{5}$$

From the desired threshold derived from Equation 4, the probability of missed P_M and detection P_D can be calculated as follows:

$$P_M = \int_{-\infty}^{\gamma} p_X(x \mid H_1)dx \tag{6}$$

$$P_D = 1 - P_M \tag{7}$$

The relationship between false alarm rate with the probability of detection and miss is shown in Figure 4 and in tabulated form in Table 3.

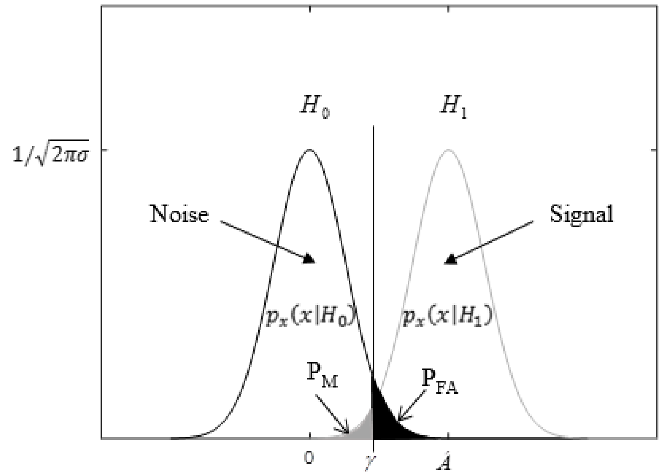


Figure 4: Probability density functions of AWGN and the signal (Srinath & Rajasekaran, 1996).

Based on Equation 4, the selected threshold depends only on the AWGN variance rather than the signal-to-noise ratio (SNR). With reference to Table 3, lower probability of false alarm P_{FA} results in a lower probability of detection P_D and higher probability of miss P_M. For example, a false alarm rate difference between 10^{-6} and 10^{-3} results in a difference in the probability of detection P_D by about 24%. The results in Table 3 show that the probability of detection increases with the false alarm rate. Typically, the false alarm rate used in practical radar systems is 10^{-6} (Mahafza, 2017). From an operational perspective, low false alarm rate is desired at the expense of reduced detection rate. This is because a higher false alarm rate would result in unnecessary deployment of assets to intercept targets that may not be there in the first place and that will contribute to a higher operation cost.

Table 3: Threshold level for various false alarm rates and their respective probability of detection and miss at SNR of 15 dB.

Probability of false alarm (P_{FA})	*Threshold γ*	*Probability of detection (P_D)*	*Probability of miss (P_M)*
10^{-6}	4.97	0.75	0.25
10^{-5}	4.26	0.918	0.082
10^{-4}	3.72	0.973	0.027
10^{-3}	3.08	0.994	0.006
10^{-2}	2.33	1	0

3.3 Problem Statement

Error to detect the pulse from a received radar pulse train in the presence of noise could result in difficulty to deinterleave a mix of radar pulse trains coming from multiple sources (Liu & Zhang, 2018) or the estimation of PRI parameters (Ahmed *et al.*, 2018). Detection error could result in two possibilities: missing or spurious pulse. Spurious pulse occurs during the interval between consecutive radar pulses. Noise that is present within the interval when exceeding the threshold as defined in Section 3.2 is considered as a true radar pulse. Depending on the SNR condition, there could be more than one spurious level within the interval. Missing pulse occurs when amplitude of the true pulse is pulled down by the noise present in the signal such that it falls below the threshold value. Similar to spurious pulses, missing pulses also depend on the SNR. In conclusion, both missing and spurious pulses depend on SNR, with a high SNR ensuring error free detection.

ES requires the detection, localisation, parameter estimation and recognition of PRI type over wide frequency band (Neri, 2006). Specifically for RWR, all of these functions have to be conducted in real-time given the constraint of space and processing on board an aircraft to enable the pilot to take evasive action or activate a countermeasure (Numlk, 2000). Thus, the RWR has to be simpler in design as compare to an ES system but at the expense of a reduction in accuracy. For example, having many channels in a channelised receiver to cover a given frequency band will be better for accuracy since the sensitivity per receiver could be made higher and improve accuracy (NAWCWD Avionics Department, 2013). However, space constraints on an aircraft would not be possible to install such as a system. Therefore, the design of an RWR has to consider the operational requirement that provides a compromise between accuracy and complexity.

4. PARAMETER ESTIMATION AND PRI TYPE RECOGNITION

This section describes the time parameters for various PRI types together with the implementation of the recognition system.

4.1 Parameter Estimation

The signal parameter of interest in this work is the timing information that is estimated from the received radar pulse train. Typically, this is derived from the PRI that is of periodic intervals between each detected pulse. The estimated PRI and its variations are then estimated from statistical quantities such as average and variance to classify the various radar pulse trains. As shown in Figure 2, the received signal is removed of its modulation parameters such as frequency and modulation type. With reference to the threshold γ defined in Equation 4, the estimated TOA is:

$$\text{if } x[n] \geq \gamma, \quad T_{OA}[m] = n, \quad 0 \leq m \leq M-1, \quad M << N \tag{8}$$

where *m* being the sequence of the instantaneous T_{OA}, *M* is the maximum number possible pulses and *N* is the time duration of the radar pulse train. From the estimated T_{OA}, the instantaneous PRI $T_p[k]$ is estimated by taking the difference between two consecutive T_{OA}s as:

$$T_p[k] = \left|T_{OA}[m] - T_{OA}[m-1]\right|, \quad 0 \leq m \leq M-1, \quad 1 \leq k \leq K, \; K << M \tag{9}$$

where *k* is the identifier for the sequence of the estimated PRI, and *K* is the total number of estimated PRIs.

From the estimated PRI, histogram analysis is used to differentiate the various PRI types (Liu & Zhang, 2018). A histogram can be defined as a function that counts the number of occurrences that falls into each of the disjoint categories (known as bins). For data where each bin can be represented as a single value rather than a range of values, the histogram can be expressed as:

$$H_X\left[x\right] = \sum_{l=1}^{L} n_l \delta\left[x-l\right] \tag{10}$$

where *L* is the possible bins representing the PRI value and n_l is the number of occurrences for a given PRI value. In relation to the instantaneous PRI from Equation 9, the total number of occurrences N_h that should be related to the number of estimated PRIs is given as:

$$\sum_{l=-L}^{L} n_l = N_h \tag{11}$$

Assuming no detection error for constant PRI type, the number of occurrences presented on the vertical axis is N_h, which corresponds to the estimated PRI value on the horizontal axis. Detection error in the radar pulse train results in the appearance of other PRI values in the histogram. The actual PRI could still be estimated provided the occurrence of the erroneous PRIs do not dominate the occurrence of the actual PRI estimate.

The histogram analysis has a drawback in representing the dominant PRI values in the radar pulse train and not the PRI values order of appearance. Essentially, this is true to differentiate between staggered with wobulation PRI type. Further details of this problem will be described in the next section. Thus, processing steps are required in addition to histogram analysis to differentiate these PRI types.

From all instantaneous PRIs estimated from Equation 9, the PRI average is:

$$\operatorname{avg}\left[T_p\right]=\frac{1}{K}\sum_{k=1}^{K}T_p\left[k\right] \tag{12}$$

The instantaneous average TOA, avg[T_{OA}[m]] obtained by multiplying the time index m for the estimated TOA from Equation 8 with the PRI average from Equation 12:

$$\operatorname{avg}\left[T_{OA}[m]\right]=m\left[\operatorname{avg}[T_p]\right] \qquad 1\le m\le M \tag{13}$$

The instantaneous TOA difference, $\varDelta T_{OA}[m]$ derived from Equations 8 and 13 is:

$$\Delta T_{OA}\left[m\right]=T_{OA}\left[m\right]-\operatorname{avg}\left[T_{OA}[m]\right] \qquad 1\le m\le M \tag{14}$$

Once obtained, the average and variance of instantaneous TOA difference are calculated as follows:

$$\operatorname{avg}\left[\Delta T_{OA}\right]=\frac{1}{M}\sum_{m=1}^{M}\Delta T_{OA}\left[m\right] \tag{15}$$

$$\operatorname{var}\left[\Delta T_{OA}\right]=\frac{1}{M}\sum_{m=1}^{M}\left[\Delta T_{OA}\left[m\right]\right]^2-\left[\operatorname{avg}\left[\Delta T_{OA}\right]\right]^2 \tag{16}$$

Both average and variance of the instantaneous TOA difference described in Equations 15 and 16 can be used to differentiate the various PRI types such as constant pulse, staggered and wobulation. Since the instantaneous average TOA, avg[T_{OA}[m]] is derived from the PRI average in Equation 12, the interval between each avg[T_{OA}[m]] will be constant. Thus, the instantaneous TOA difference, $T_{OA}[m]$ in Equation 14 should be able to detect the changes in the instantaneous TOA. For constant PRI type, the instantaneous TOA difference $\varDelta T_{OA}[m]$ is always zero since the interval between each TOA estimate is constant, which will result in zero values for the average and variance instantaneous TOA difference - avg[$\varDelta T_{OA}$] and var[$\varDelta T_{OA}$] - defined in Equations 15 and 16. Variation in the interval between each estimated TOA for staggered PRI type consecutively produces variation in the instantaneous TOA difference $\varDelta T_{OA}[m]$. resulting in non-zero average avg[$\varDelta T_{OA}$] and variance var[$\varDelta T_{OA}$]. Similar to staggered PRI, the wobulation PRI type will exhibit instantaneous TOA difference $\varDelta T_{OA}[m]$ with non-zero average and variance. Since the PRI variation is sinusoidal as shown in Table 2, variation in the instantaneous TOA difference $\varDelta T_{OA}[m]$ would correspondingly be sinusoidal. Thus, the variation characteristics of the instantaneous TOA difference could be used to determine if the PRI type is either staggered or wobulation.

The wobulation frequency is estimated directly from the period of the instantaneous TOA difference $\varDelta T_{OA}[m]$ described in Equation 14. Since noise is present, some of the received pulses could either be missing or inserted during the process of detection described in Section 3.2. Thus, it is more appropriate to apply spectrum analysis techniques to estimate the wobulation frequency instead of direct estimation from the instantaneous TOA difference. Since the instantaneous TOA difference is similar to a discrete-time signal, it would be appropriate to apply the discrete Fourier transformation (DFT). The DFT and the subsequent power spectrum of the instantaneous TOA difference are:

$$F\left[k\right]=\begin{cases}\displaystyle\sum_{m=0}^{M}\varDelta T_{OA}\left[m\right]exp\left(-j\frac{2\pi}{M}mk\right) & 0\le k\le M-1\\ 0 & \text{elsewhere}\end{cases} \tag{17}$$

$$S_F\left[k\right]=\frac{1}{M}\left|F\left[k\right]\right|^2 \tag{18}$$

Hence, the wobulation frequency f_w estimated from the peak of the power spectrum is:

$$f_w = \arg\left(\max_k\left(S_F\left[k\right]\right)\right) \tag{19}$$

4.2 Recognition Parameter Selection

The timing parameters defined in the previous section are then used to design the PRI type recognition function. First, the various PRI types - constant, staggered, wobulation and jittered - are generated and their parameters are then estimated. An in-depth analysis will be done to determine the best parameters to represent each PRI type.

4.2.1 Constant PRI Type

The timing parameters defined in Table 2 are used to generate a radar pulse train for constant PRI type SP2. Table 4 describes the methodology for estimating the timing parameters. For a constant PRI type, whether it has PRI 4, 8 or 16, it is clearly seen that for all constant PRIs, the average and variance of the instantaneous TOA difference at the right most column are all zero. This is because the PRI does not change, which results in zero values for the instantaneous TOA difference defined in Equation 12. Thus, both average and variance could be used as parameters to identify constant PRI type. However, noise in the received signal as described by the detection methodology in Section 3.2 could either cause missing or spurious pulses in the radar pulse train that consequently results in error in TOA estimation. Thus, both the average and variance for the instantaneous TOA difference could have non-zero values.

Table 4: Analysis of radar pulse train for constant PRI type SP2:

(a) No errors at 7 dB SNR.

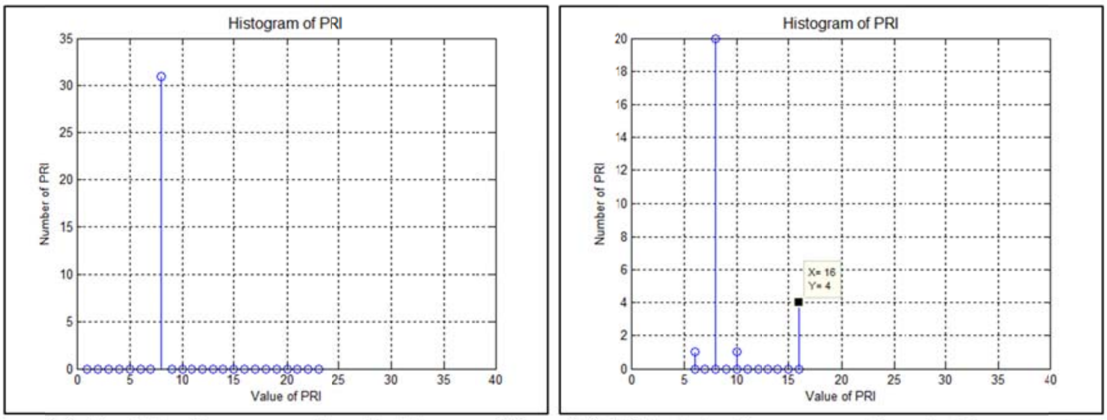
n	$T_{OA}[m]$	m	k	$T_P[k]$	$avg[T_{OA}[m]]$	$\Delta T_{OA}[m]$
1	8	1	1	8	8	0
2	16	2	2	8	16	0
3	24	3	3	8	24	0
4	32	4	4	8	32	0
5	40	5	5	8	40	0
6	48	6	6	8	48	0
.	56	7	7	8	56	0
.	64	8	8	8	64	0
256						
				average =8		average= 0 variance= 0

(b) With errors at 4 dB SNR.

n	$T_{OA}[m]$	m	k	$T_P[k]$	$avg[T_{OA}[m]]$	$\Delta T_{OA}[m]$
1	16	1	1	16	15	-1
2	24	2	2	8	30	-6
3	40	3	3	16	45	-5
4	48	4	4	8	60	-12
5	64	5	5	16	75	-11
6	72	6	6	8	90	-18
.	80	7	7	8	105	-25
.	88	8	8	8	120	-32
256	.					
				average= 15		average= -13.75, variance= 112.5

Besides using the instantaneous TOA difference and its statistical properties, histogram analysis can be used to estimate the PRI distribution and assess its performance when detection error occurs. The histogram analysis results in Figure 5, which shows the number of occurrences for all possible estimated PRI values. For constant PRI type, the histogram should show only one PRI value. In Figure 5(a), a peak is observed with 31 occurrences with an estimated PRI of 8 samples since the radar pulse train is constant PRI SP 2 as defined in Table 2.

When detection error occurs at SNR of 4 dB, other PRI values are estimated besides the actual values as shown in the histogram in Figure 5(b). Besides the 20 peak occurrences at estimated PRI of 8 samples, other PRIs are also estimated for lower occurrences (less than four) at 6, 10, and 16 samples contributed by detection error. For this purpose, the histogram threshold used to determine the actual PRI of a radar pulse train from the histogram is $\gamma_{HIST} = S_{max}$, where S_{max} is the peak occurrence on the histogram. For no detection error shown in Figure 5(a), the histogram threshold is 31 at PRI value of 8 samples. The peak occurrence is lower if the detection error is present in the radar pulse train as shown in Figure 5(b). For both cases, the total number of occurrences is 31, with the peak occurrence being lower at S_{max}=26 at PRI value of 8 samples due to the presence of low occurring spurious peaks in the histogram at PRI of 6, 10 and 16 samples. Thus, the histogram threshold value will vary according to the peak occurrence on the histogram and the presence of detection error in the received radar pulse train. However, most important is to be able to estimate the PRI of a given radar pulse train.



(a) No detection error, threshold γ_{HIST} =31 **(b) With detection errors, threshold γ_{HIST} = 26**

Figure 5: Histogram for constant PRI type.

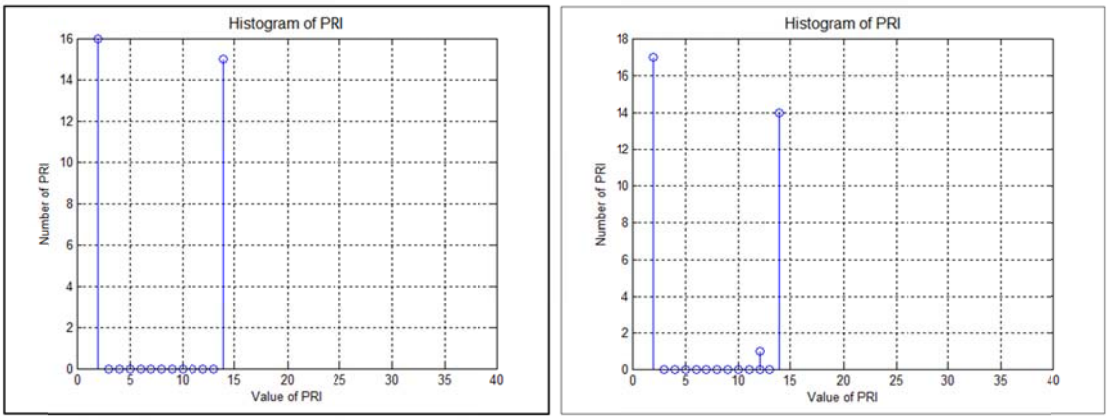
4.2.2 2-level Staggered PRI Type

Similar to constant PRI type in the previous sub-section, a 2-level staggered PRI type is generated for radar pulse train 2SP2 based on the timing parameters defined in Table 2. The resulting signal is presented in Table 5 together with the analysis performed using instantaneous TOA and instantaneous TOA difference at SNR 10 dB. At this SNR level, the detection error is minimal in the radar pulse train. Unlike the constant PRI type, the average and variance in the instantaneous TOA difference as shown in the right most column Table 5 is non-zero since the PRI has two values – 2 and 14 samples – and the resulting instantaneous TOA difference appears as a constant sequence of samples (-6, 0, -6, 0, …..) shown in the last column of Table 5. Thus, the average and variance of the instantaneous TOA difference should be non-zero and could be used to decide if the pulse train is a constant or 2-level staggered PRI type. However, this method fails at lower SNR due to the detection error as demonstrated in the previous sub-section.

Table 5: Analysis of radar pulse train for 2-level staggered PRI type 2SP2.

n	$T_{OA}[m]$	m	k	$T_P[k]$	$avg[T_{OA}[m]]$	$\Delta T_{OA}[m]$
1	0	1	1	2	8	-6
2	2	2	2	14	16	0
3	16	3	3	2	24	-6
4	18	4	4	14	32	0
5	32	5	5	2	40	-6
6	34	6	6	14	48	0
.	48	7	7	2	56	-6
.	50	8	8	14	64	0
.	64	9				
256				average= 8		average= -3.0, variance= 3.05

Histogram analysis could be used to determine the number of PRI present in this radar pulse train. From the histogram in Figure 6(a), the peak occurrences in the histogram at 16 and 15 corresponds to PRI of 4 and 14 samples, respectively. The suitable histogram threshold is 15 for this radar pulse train using the method described for constant PRI type. Detection error at SNR of 3 dB introduces low occurring spurious pulses at PRI of 12 samples as shown in Figure 6 (b). A histogram threshold of 14 will be able to estimate the true PRI of 4 and 14 samples respectively. Thus, a histogram threshold value that is relatively high compare to the lower level occurrences due to detection error will still be able to estimate the PRI of the 2-level staggered PRI type.



(a) with no errors having γ_{HIST} =14 **(b) with errors having γ_{HIST} = 14**

Figure 6: Histogram for 2-level staggered PRI type.

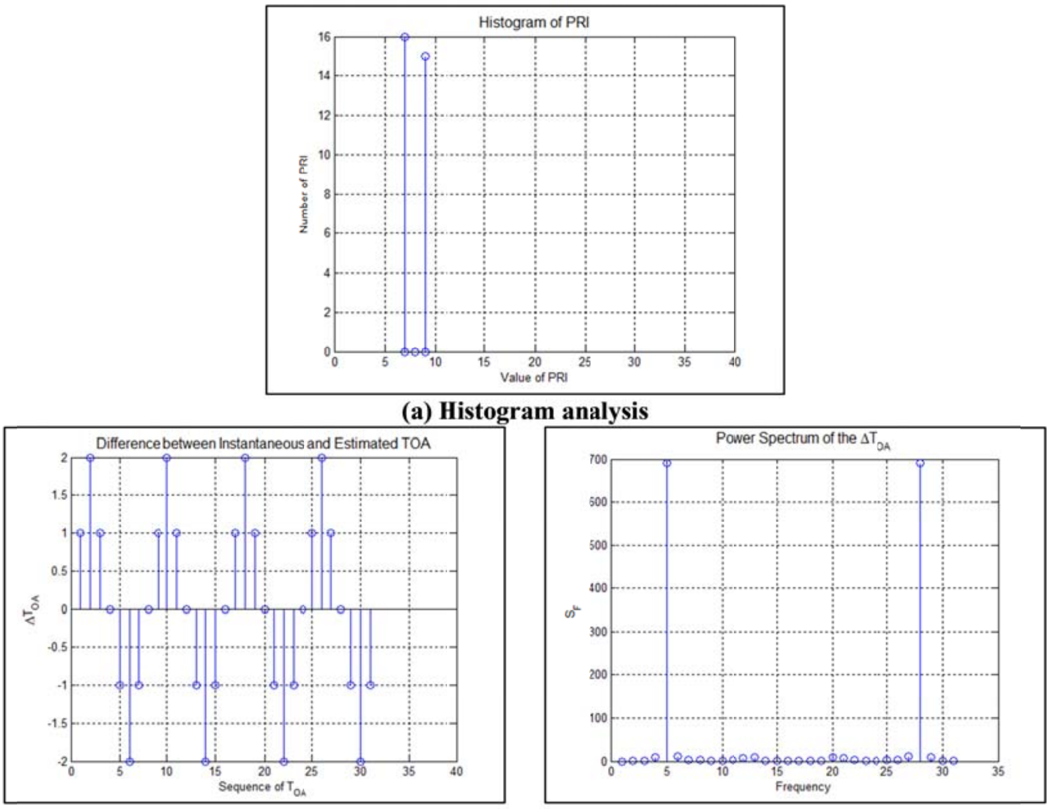
4.2.3 Wobulation PRI Type

For wobulation PRI type, the timing parameters defined in Table 2 is used to generate a radar pulse train WB2 at SNR of 10 dB. The resulting radar pulse train is shown in Table 4 together with the methodology for estimating the timing parameters. As expected, the average PRI estimated is 8 samples, which represents the true signal parameters but varies within a range of values. However, observing the instantaneous PRI in the fifth column does not describe how the PRI varies for a given time instant. However, the variation in the instantaneous PRI is more obvious in the instantaneous TOA difference shown in the last column of the table. A sinusoidal sequence with a wobulation period T_w of 8 samples and its corresponding frequency f_w of 1/8 Hz is observed. Similar to any sinusoidal waveform, the instantaneous TOA difference has zero value and a non-zero variance of 1.6. The estimation of the wobulation period and frequency can be improved further in the presence of detection error by applying either the histogram or spectrum analysis shown in Equations 17 to 19.

Table 6: Analysis of radar pulse train for wobulation PRI type WB2 at SNR of 10 dB.

n	$T_{OA}[m]$	m	k	$T_P[k]$	$avg[T_{OA}[m]]$	$\Delta T_{OA}[m]$
1	0	1	1	9	8	1
2	9	2	2	9	16	2
3	18	3	3	7	24	1
4	25	4	4	7	32	0
5	32	5	5	7	40	-1
6	39	6	6	7	48	-2
.	46	7	7	9	56	-1
.	55	8	8	9	64	0
.	64	9				
256				average = 8 (rounded)		average= 0 variance = 1.6

Figure 7 shows the histogram analysis and the instantaneous TOA difference with its corresponding power spectrum. Starting with the histogram analysis in Figure 7(a), the occurrence peaks of 16 and 15 are observed corresponding to PRI of 7 and 9 samples respectively. Direct interpretation of the radar pulse train characteristics could misclassify this signal as 2-staggered PRI type even though the radar pulse train is wobulation PRI type. Thus, histogram analysis on its own could only estimate the number of PRIs and their estimated values, and could not differentiate between a multi-level staggered or wobulation PRI type. Thus, differentiating the two PRI types requires additional analysis.



(a) Histogram analysis

(b) Instantaneous TOA difference

(c) Power spectrum of instantaneous TOA difference

Figure 7: Histogram and spectrum analysis of wobulation PRI type WB2.

Figure 7(b) graphically represents the instantaneous TOA difference for the wobulation PRI type WB2 that was shown earlier in Table 6. The sinusoidal nature of the instantaneous TOA difference is noticeably obvious in the plot and the frequency can be estimated by measuring the period of the waveform. Due the possibility of detection error, this may not be the best way to estimate the frequency due to missing pulses or spurious pulses. Figure 7(c) shows the power spectrum obtained using Equation 18. The wobulation frequency estimated from the peaks of the power spectrum using Equation 19 is (5-1) = 4 frequency samples. Since the total number of sample points is 32, the wobulation frequency is the frequency sample divided by the total number of sample points to obtain 4/32 = 1/8 Hz. This is correct since the actual PRI type is WB2 whose parameters are defined in Table 2.

4.2.4 Jitter PRI type

The timing parameters for jitter PRI type Jitter1 shown in Table 2 is used to generate a radar pulse train. Table 7 presents analysis of the radar pulse train together with the estimation of the timing parameters. Starting with most basic parameters, the average PRI is as expected at 8 samples but the instantaneous PRI shown in the fifth column does not exhibit a constant set of values similar to constant PRI type or 2-level staggered PRI types. This is also true for the instantaneous TOA difference in the last column of the table with non-zero values for the average and variance at 2.5 and 2.29. The analysis results show the PRI is centred at 8 samples but varies within a given time.

Table 7: Analysis of radar pulse train for jitter PRI type Jitter1 at SNR of 10 dB.

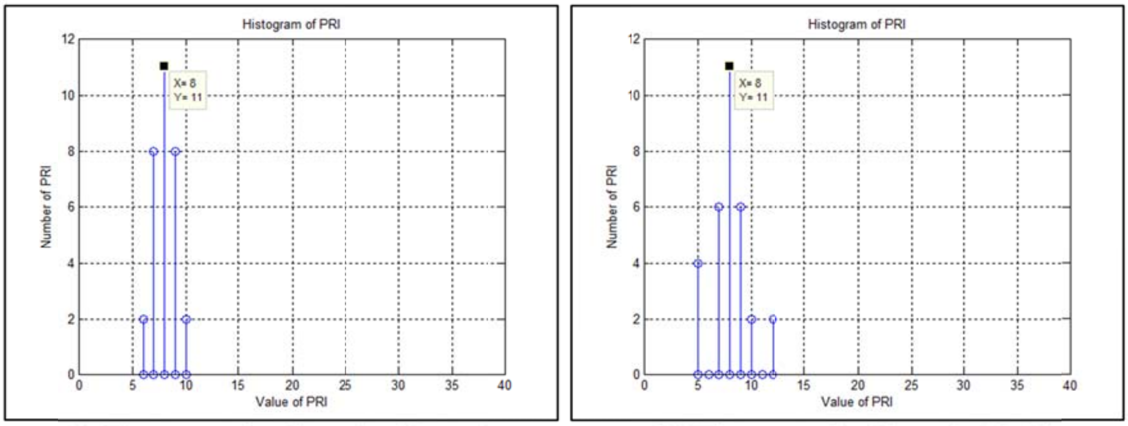
n	$T_{OA}[m]$	m	k	$T_P[k]$	$avg[T_{OA}[m]]$	$\Delta T_{OA}(m)$
1	8	1	1	12	8	0
2	20	2	2	7	16	4
3	27	3	3	9	24	3
4	36	4	4	5	32	4
5	41	5	5	11	40	1
6	52	6	6	6	48	4
.	58	7	7	8	56	2
.	66	8	8	7	64	2
.	73	9	9	-	-	-
256	.			average = 8		average = 2.5 variance= 2.29

Besides the direct estimation of the radar pulse train parameters as shown in Table 7, histogram analysis could also be used for this purpose. Figure 8 shows the histogram analysis for Jitter1 and Jitter2 PRI types. Between the two PRI types, it appears that the PRI is centred at 8 samples while the estimated PRI is spread out over a broader range for Jitter2 (5 to 12 samples) as compared to Jitter1 (6 to 10 samples). The centre value and spread for the estimated PRI obtained by the average and variance from the histogram are:

$$\operatorname{avg}\left(T_{p,h}\right)=\frac{1}{N_h}\sum_{k=1}^{N_h}x_k H_X(x_k) \qquad (20)$$

$$\operatorname{var}(T_{p,h})=\frac{1}{N_h}\sum_{k=1}^{N_h}x_k^2 H_X(x_k)-\left[\operatorname{avg}(T_{p,h})\right]^2 \qquad (21)$$

where N_h is the total number of occurrences. By using Equations 20 and 21, the average and variance for Jitter1 are 8 and 1 respectively, and for Jitter2 are 8 and 2.84 respectively. Thus, variance can be used to differentiate between the two jitter radar pulse trains with the same average PRI but differs in the PRI spread that is estimated by the variance. This last step is included in the classification to prevent the misclassification of jitter PRI type with other PRI types.



(a) Histogram for Jitter 1 with smaller variation

(b) Histogram for Jitter 2 with bigger variation

Figure 8: Histogram for Jitter1 and Jitter2 radar pulse train with PRI average of 8 samples.

4.3 PRI Type Recognition

Based on the analysis of results performed Section 4.2, the analysis method and corresponding parameters that can uniquely identify the various PRI types are as follows:

1. **Constant PRI type:** Single peak detected for the histogram that is used to as the PRI estimate.
2. **Wobulation PRI type:** Two or more peaks detected from the histogram, the average instantaneous TOA is zero while the variance is non-zero, and wobulation frequency estimated from the peak of the power spectrum of the instantaneous TOA difference.
3. **2-level staggered PRI type:** Two peaks detected from the histogram, the average and variance instantaneous TOAs are non-zero, and the PRIs are estimated from the peaks of the histogram.
4. **Jitter PRI type:** More than two peaks detected from the histogram, and the average and variance PRI are estimated from the histogram.

The parameters described for identifying the various PRI types are used to implement a rule based PRI type recognition function shown in Algorithm 1. From the analysis methods described in Section 3.3, the parameters that are used by the PRI type recognition function to differentiate between the various PRI types for an input radar pulse train are: peak value of the histogram S_{max}, array of estimated PRIs, average and variance instantaneous TOA difference, average and variance PRI, power spectrum of instantaneous TOA difference, and estimated wobulation frequency. The recognition process begins with a set of rules that classifies the estimated radar pulse train parameters starting with constant PRI type, wobulation PRI type, 2-level staggered PRI type and jitter PRI types. An input radar pulse train that does not match with any one of parameters defined in the PRI type recognition function is considered unknown.

Algorithm 1: Algorithm for PRI type recognition function.

```
Function [signal no] = PRI type recognition($S_{\max}$, PRIarray, No of PRI, avg[$\Delta T_{OA}$], var[$\Delta T_{OA}$], $S_{F\text{array}}$,$f_w$ avg[$T_p$], var[$T_p$]
// PRIarray is an array of estimated PRI from the histogram
// SFarray is the array of the power spectrum of the instantaneous $\Delta T_{OA}$
// arranged in descending order.
// $f_w$ is the wobulation frequency
// $S_{\max}$ peak occurrence on the histogram
if $\gamma_{HIST} \geq 0.85 S_{\max}$          // Find the peak of the histogram that matches the PRI estimate
                                      // to determine if radar pulse train constant PRI
    if PRIarray(1) == 4 , signal no = 1                               //SP1
    elseif PRIarray (1)  == 8 , signal no = 2                         // SP2
    elseif PRIarray (1) == 16 , signal no = 3                         // SP3
    else signal no = unknown                          // No match is found for constant PRI type
else                                                  // the radar pulse train has multiple PRIs
    if (1 < No of PRI < 4) and var[$\Delta T_{OA}$] = 0
            // PRI array should have between 1 and 4 elements
            // while variance of $\Delta T_{OA}$ [$m$] should not be zero.
        if  avg[$\Delta T_{OA}$]=0 and max($S_{\text{Farray}}$) = $S_{\text{Farray}}$ (1) or max($S_{Farray}$)=$S_{Farray}$(2))
            // The radar pulse train is wobulation PRI type
                if  0.2 <$f_w$ < 0.3, signal no = 6                        // WB1
                elseif $f_w \leq$ 0.2, signal no = 7                        // WB2
                else signal no = unknown …// No match is found for wobulation PRI type
        elseif average[$\Delta T_{OA}$]< 0             // The radar pulse train is 2-level staggered PRI type
                if  PRIarray (1) = 12 and PRIarray (2) == 4, signal no = 4          //  2SP1
                elseif PRIarray (1) = 14 and PRarray (2) == 2, signal no = 5        //  2SP2
                else signal no = unknown…// No match is found for 2 level PRI type
        else signal no = unknown               // No match is found among wobulation or staggered
    elseif (No of PRI $\geq$ 4) and avg[$\Delta T_{OA}$]~=0 and var[$\Delta T_{OA}$]~= 0
        // The radar pulse train is jitter PRI type
        If avg[$T_{p,h}$]=8                  //average of PRI estimate in histogram analysis is 8
                If var[$T_{p,h}$], < 1.5, signal no = 8                      //  Jitter 1
                                    // If variance of PRI pulses in histogram is less than >1.5
                else signal no = 9                                          //  Jitter 2
        else signal no = unknown                        // No match is found for jitter PRI type
    else signal no = unknown                            // No match is found all the PRI types
EndFunction
```

5. RESULTS & DISCUSSION

The complete performance verification of RWR processing element described in Figure 2 is presented in this section starting with the recognition rate followed by confusion chart analysis. The final evaluation considers how the RWR processing element is expected to perform when used in a real operational scenario.

5.1 Recognition Rate Performance

Since the received radar pulse train is corrupted by interference modelled as AWGN as shown in Equation 1, it is necessary to access the performance of the complete RWR processing element shown in Figure 2 for all realisations at a given SNR range. Monte Carlo simulation is utilised for this purpose based on 100 realisations for various PRI types as shown in Figure 2 at SNR range from 0 to 20 dB. The results presented in Figure 9 plots the recognition rate against the SNR. On average, correct recognition a rate of 90 % is achieved for SNR above 9 dB. Above this range, the simplest type, which is the constant PRI type, is classified close to 100%, while the more complex types, which are 2-level staggered, wobulation and jitter PRI types, have lower recognition rate between 80 to 90%. Unknowns recognised at less than 5% means correct recognition each radar pulse train within its PRI type.

The cut-off SNR that describes a sharp transition from high to low recognition rate depends on the PRI type. Due to the detection error, difficulty arises to estimate the actual PRI since missing pulses extend the PRI length while spurious pulses reduce the PRI length. This is true for radar pulse trains with low PRI where the cut-off SNR is higher as compared to pulse train with high PRI. The lowest cut-off SNR at about 2 dB are for radar pulse train SP3 and WB1 with average PRI of 16 samples while the highest is 9 dB for SP1 with average PRI of 4 samples. The jitter PRI types –Jitter1 and Jitter2- and WB2 have average PRI of 8 samples, which fell in the middle of the range with cut-off SNR of 5 dB. Similarly, this also true for the 2-level staggered PRI type. Although there are two PRIs for each 2-level staggered PRI type, the higher PRI compensates for the sensitivity of the lower detection error resulting in a cut-off SNR of 5 dB. Therefore, it is expected that the cut-off SNR will be lower if the PRIs used have values much higher than what is described in this paper.

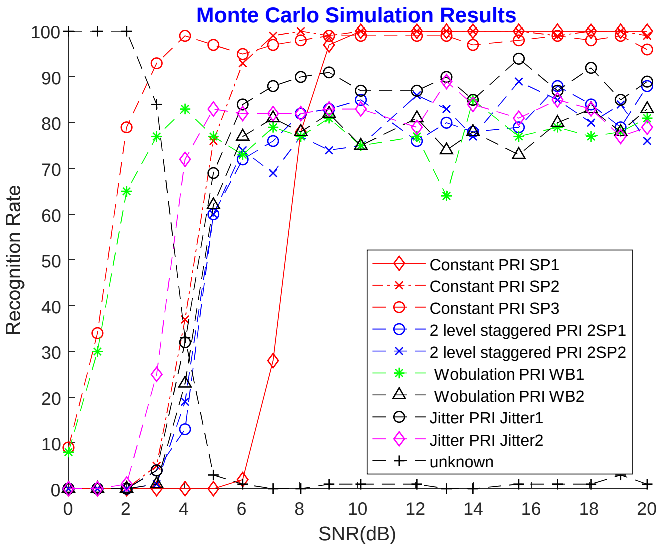


Figure 9: Recognition rate for SNR range from 0 to 20 dB.

5.2 Confusion Chart Analysis

The confusion chart is used to present how the PRI type is correctly recognised and misrecognised at a given SNR (Ahmad & Sha'ameri, 2015). Unlike Figure 9, the results only show the recognition rate for a given SNR range. Two confusion charts are presented in this section: the first for SNR of 10 dB as shown in Table 8(a) and the second in Table 8(b) for SNR of 5 dB. The main purpose of this analysis is to evaluate the effect of miss-recognition at two different SNR conditions.

As shown in Table 8(a) for SNR 10 dB, constant PRI types in general are correctly classified. Similar to the recognition rate in Figure 8, the recognition rate for more complex radar pulse trains, such as 2-level staggered, wobulation and jitter, are lower. Almost 15% of 2-level staggered PRI type is recognised as jitter PRI type while 5% is recognised as unknown. Within the jitter PRI type, misrecognition occurs between Jitter1 and Jitter2, with the higher occurrence when Jitter2 is misclassified as Jitter1 at 18%. At lower SNR of 5 dB, the recognition rate is downgraded for all PRI types as shown in Table 8(b). For constant PRI type, all of SP1s are misclassified as SP2 (18 %), SP3 (8 %) and unknown (74%), while SP3s are correctly classified at 97%. Similar results are also shown in Figure 8, which is due to lower PRI of 4 samples for SP1. The recognition rate for 2-level staggered drops to within the 58% range and most recognition errors are categorised into the unknown range at 30%. Similar results are also observed for wobulation in terms of the drop in the recognition rate and consecutively recognised as unknown. Within the jitter PRI type, the recognition rate is lower for Jitter1 at 66% as compared to Jitter2 at 87%, which is similar to the rate at higher SNR.

Table 8: Confusion chart: (a) SNR 10 dB, (b) SNR 5 dB.

(a) SNR 10 dB

	PRI Name		*Constant Pulse*		*2-level staggered*		*Wobulation*		*Jitter*		*Unknown*
PRI type		SP1	SP2	SP3	2SP1	2SP2	WB1	WB2	Jitter 1	Jitter 2	
Constant	SP1	**100**	0	0	0	0	0	0	0	0	0
	SP2	0	**100**	0	0	0	0	0	0	0	0
	SP3	0	0	**98**	0	0	0	0	0	0	2
2-level staggered	2SP1	0	0	0	**79**	0	0	0	0	15	6
	2SP2	0	0	0	0	**83**	0	0	0	12	5
Wobulation	WB1	0	0	0	0	0	**82**	0	0	0	18
	WB2	0	0	0	0	0	0	**84**	0	0	16
Jitter	Jitter1	0	0	0	0	0	0	1	**85**	9	6
	Jitter2	0	0	0	0	0	0	0	18	**82**	0

(b) SNR 5 dB

	PRI Name		*Constant*		*- level staggered*		*Wobulation*		*Jitter*		*Unknown*
PRI type		SP1	SP2	SP3	2SP1	2SP2	WB1	WB2	Jitter 1	Jitter 2	
Constant	SP1	**0**	18	8	0	0	0	0	0	0	74
	SP2	0	**73**	23	0	0	0	0	0	0	4
	SP3	0	0	**97**	0	0	0	0	0	0	3
2 level staggered	2SP1	0	0	0	**59**	0	0	0	0	12	29
	2SP2	0	0	0	0	**58**	0	0	0	12	30
Wobulation	WB1	0	0	0	0	0	**73**	0	0	0	27
	WB2	0	0	0	0	0	0	**79**	0	0	21
Jitter	Jitter1	0	0	0	0	0	0	0	**66**	22	12
	Jitter2	0	0	0	0	0	0	0	11	**87**	2

5.3 Operational Scenario

With the completed RWR processing element and performance verification, the next step to evaluate its performance under a real operational condition. The main purpose of this section is to determine the recognition range for an RWR when an aircraft flies within range of hostile territory protected by a multi-layered air defence infrastructure consisting of air defence radar and surface to air missiles for point and wide area defence (Grant, 1998).

The architecture of an RWR is described in Numlk (2000) while most publications tend to cover the broad topic of ES (Adamy, 2008; Neri, 2006). Since some parameters are not clearly defined, such as bandwidth and the resulting sensitivity, the best that can be done is to make a reasonable assumption to allow real-time analysis and recognition of PRI type over a broad range of frequency, such as from 0.5 to 20 GHz.

The specifications for the RWR are shown in Table 9. Due to space constraint on an aircraft, the number of wideband receivers is limited to four and the frequency band is allocated per receiver according to the typical configuration of an RWR (Numlk, 2000). Coverage over all directions in addition to localised electromagnetic sources such as ground based radar is achieved using multiple directional antennas. Each antenna should have a beam-width of 90° and a horn antenna is assumed since its radiation pattern meets this requirement (Balanis, 2016). Furthermore, information on the actual antenna used in an RWR is manufacturer dependent with no details described.

Table 9: Specifications of the wideband receiver for used in the RWR.

Frequency coverage	0.5 to 20 GHz
Number of wideband receivers	4
RWR band 0 (0.5 to 2 GHz)	1.5 GHz
RWR band 1 (2 to 4 GHz)	2 GHz
RWR band 2 (4 to 8 GHz)	4 GHz
RWR band 3 (8 to 20 GHz)	12 GHz
Main lobe antenna gain, G_R	10 dBi
Side lobe antenna gain, G_{RS}	–10 dBi
Noise figure	13 dB

Typical long range air surveillance ground based radars operate either in the L- or S-band (Adamy, 2008), while medium or short range radars of similar function operate at a higher frequency, such as the X-band, due to lower power requirements and the need for mobility. Tracking radars for gunnery or missile guidance operate at an even higher frequency in the Ku-band due to better directivity of the electromagnetic waves and size reduction at this frequency band (Neri, 2006). For this analysis, the radar considered is a current generation long range air surveillance since this is the first line of defence that will detect an approaching aircraft. The transmitter parameters of the radar defined in Table 10 are based on the Lockheed-Martin TPS 77 L-band ground based air surveillance radar (Lockheed Martin Corporation, 2020). Since the antenna aperture size is provided, the mainlobe gain can be calculated as follows (Mahafza, 2017):

$$G_T = \frac{4\pi f^2 A_e}{c^2} \tag{22}$$

where f is the carrier frequency of the radar, A_e is the antenna aperture size in m^2 and c is the speed of light in m/s. Since the antenna is parabolic, the sidelobe and backlobe gain is taken as -30 dB difference with relation to the mainlobe gain (Balanis, 2016). It should be noted that this radar has a low peak transmitted power as compared with legacy radars, where the peak transmitted power is normally in the excess of 150 kW (Wolff, 2020). This is achieved by having a longer PRI and employing pulse compression signaling. For comparison purposes, a legacy radar is assumed with a peak transmitted power of 150 kW while the other transmitter parameters follow the description in Table 10.

Table 10: Transmitter specifications for current generation a long range air surveillance radar

Frequency Band (L-band)	1.215 to 1.400 GHz
Instrumented range (km)	470
Peak transmitted power (kW)	19.9 kW
Antenna aperture size	27.1 m^2
Main lobe antenna gain (Calculated)	30 dBi
Side lobe antenna gain (Calculated)	0 dBi

To determine the quality of reception of the RWR, the sensitivity is to be determined from basic principles since this parameter is not provided. Assuming temperature is at 300 K, the sensitivity of a receiver can be calculated as (NAWCWD Avionics Department, 2013):

$$S = -174 + NF + 10\log_{10}\left(BW\right) \tag{23}$$

where *NF* is the noise figure and *BW* is the bandwidth of the receiver in Hz. *NF* ranges from 5 to 15 and very much depends on the quality of the receiver (NAWCWD Avionics Department, 2013). For this evaluation, a NF of 13 dB is considered and the sensitivity calculated using the parameters in Table 9 and Equation 23 for the RWR 0 band receiver is -69 dBm. The calculated sensitivity is comparable to that described for an RWR implemented on a channelised receiver (Neri, 2006).

For link budget analysis, free space propagation can be utilised due to the availability of line of sight path between the radar and an aircraft flying at high altitude. The received power in dBm can be expressed as (Adamy, 2008):

$$P_R = P_T + G_T + G_R - 20\log_{10}\left(R\right) - 20\log_{10}\left(f\right) - L - 32 \tag{24}$$

where P_T is the transmit power in dBm, G_T is the transmit antenna in dBi, G_R is the receive antenna in dBi, *R* is distance in km, *f* is frequency in MHz, and *L* is losses due to cabling, connectors and other sources in dB. From the receiver sensitivity obtained from Equation 23, the SNR in dB can be calculated as:

$$SNR = P_R - S \tag{25}$$

From the RWR specifications in Table 9 with sensitivity of -69 dBm and radar transmitter specifications in Table 10, the received SNR obtained from Equation 25 for ranges from 100 to 400 km, as presented in Table 11. In general, the received SNR at the RWR for the new radar is lower as compared to the legacy radar. This means that the RWR has to be at a closer range to the new radar as compared to the legacy radar to correctly recognise the PRI type. At a distance of 300 km, there is a difference of about 9 dB between the received SNR at the RWR for the new and legacy radars. The received SNR is 3.4 dB for the new radar, while for the legacy radar, it is 15 dB. By referring to the recognition rate in Figure 8, the received SNR of 3.4 dB at 300 km range means that the recognition rate for all PRI types is less than 50%. At closer range of 200 km, a received SNR of 7 dB improves the recognition rate to close to 100% for constant PRI, and between 80 to 90% for the other PRI types. For legacy radars, similar recognition rates is possible even at a range of 400 km with a received SNR of 10 dB. As a conclusion, the aircraft has to be closer to the new radar as compared to the legacy radar to be able to correctly recognise the PRI type.

Table 11: Received SNR comparison between the new and legacy air surveillance radars for various ranges.

Range (km)	Received SNR (dB) new radar	Received SNR (dB) legacy radar
100	13	22
200	7	16
300	3.4	13
400	0.93	10

6. CONCLUSION

A RWR processing element is presented in this paper that performs PRI type analysis and recognition for constant, 2-level staggered, wobulation and jitter radar pulse train. In order to determine the parameters suitable to recognise the different PRI types, analysis is done on the actual PRI sequence and its derived sequences, such as instantaneous average TOA, instantaneous TDOA, histogram analysis, power spectrum of the instantaneous TDOA and the estimated statistical parameters derived from their average. The relevant parameters are identified, which is then used as input for a rule-based recognition algorithm. Verification of the RWR processing element is performed using Monte Carlo simulation where PRI recognition is performed for 100 realisations per SNR for a range from 0 to 20 dB. At SNR above 8 dB, constant PRI are correctly recognised close to 100%, while the recognition rate is between 75 to 90% for the rest. The cut-off SNR where the recognition rate drastically drops depends on the PRI, where the cut-off is lower for the high PRI radar pulse train. Performance analysis was also conducted based on parameters of an actual RWR, and compared with a legacy air surveillance radar.

REFERENCES

Ata'a, W.A. & Abdullah, S. N. (2007). Deinterleaving of radar signals and PRF identification algorithms. *IET Radar Sonar Navigation,* 1: 340–347.

Adamy, D.L. (2008). *EW 103: Tactical Battlefield Communications Electronic Warfare*. Artech House, Norwood, Massachusetts.

Ahmad, A.A. & Sha’ameri, A.Z. (2015). Classification of airborne radar signals based on pulse feature estimation using time-frequency analysis. *Defence S&T Tech. Bull.*, 8: 103–120.

Ahmed, U.I., Aziz, I. & Rehman, T.U. (2019). Comprehensive review of pulse repetitions interval (PRI) classification schemes. *6th Int. Conf. Aerospace Sci. Eng. (ICASE 2019)*, pp. 6–11.

Ahmed, U.I., Ur Rehman, T., Baqar, S., Hussain, I. & Adnan, M. (2018). Robust pulse repetition interval (PRI) classification scheme under complex multi emitter scenario. *22nd Int. Microwave Radar Conf. (MIKON 2018)*, pp. 597–600.

Balanis, C.A. (2016). *Antenna Theory Analysis and Design*. Wiley & Sons Inc., Hoboken, New Jersey.

Cain, L., Clark, J., Pauls, E., Ausdenmoore, B., Clouse, R., & Josue, T. (2018). Convolutional neural networks for radar emitter classification. *2018 IEEE 8th Annual Comput. Commun. Workshop Conf. (CCWC 2018)*, pp. 79–83.

Ge, Z., Sun, X., Ren, W., Chen, W. & Xu, G. (2019). Improved Algorithm of Radar Pulse Repetition Interval Deinterleaving Based on Pulse Correlation. *IEEE Access*, 7: 30126–30134.

Gençol, K., At, N., & Kara, A. (2016). A wavelet-based feature set for recognizing pulse repetition interval modulation patterns. *Turkish J. Electr. Eng. Comp. Sci.*, 24: 3078–3090.

Genova, J. (2018). *Electronic Warfare Signal Processing*. Artech House, Norwood, Massachusetts.

Ghani, K.A., Sha’Ameri, A.Z., Dimyati, K., & Daud, N.G.N. (2017). Pulse repetition interval analysis using decimated Walsh-Hadamard transform. *2017 IEEE Radar Conf. (RadarConf 2017)*, pp. 58–63.

Gong, S.-X., Wei, X.-Z., & Li, X. (2013). Review of wideband digital channelized receivers. *Acta Electronica Sinica*, 5: 949–959.

Grant, R. (1998). *The Radar Game: Understanding Stealth and Aircraft Survivability*. IRIS Independent Research, Arlington.

Guo, S., & Tracey, H. (2020). Discriminant Analysis for Radar Signal Classification. *IEEE T. Aerosp, Electr. Syst.*, 56: 3134-3148

Jawad, M., Iqbal, Y. & Sarwar, N. (2020). PRI characteristics analysis under complex environment of spurious and missing observations. *Proc. 2020 17th Int. Bhurban Conf. Appl. Sci. Tech. (IBCAST 2020)*, 1: 617–621.

Jordanov, I., Petrov, N. & Petrozziello, A. (2016). Supervised radar signal classification. *Proc. Int. Joint Conf. on Neural Networks (IJCNN 2016)*, pp. 1464–1471.

Lin, T., Zou, C., Zhang, Z., Zhao, S., Liu, J., Li, J., Zhang, K., Yu, W., Wang, J. & Jiang, W. (2020).

Differentiator-based photonic instantaneous frequency measurement for radar warning receiver. *J. Lightwave Tech.*, 38: 3942–3949.

Liu, Y. & Zhang, Q. (2018). Improved method for deinterleaving radar signals and estimating PRI values. *IET Radar Sonar Nav.*, 12: 506–514.

Lockheed Martin Corporation. (2020). AN/TPS-77 Long-Range Air Surveillance Radars. Available online at: https://www.lockheedmartin.com/en-us/products/ground-based-air-surveillance-radars.html (Last access date: 25 July 2020)

Mahafza, B.B. (2017). *Introduction to Radar Analysis, Second edition*. CRC Press, Taylor & Francis Group, Boca Raton, Florida.

NAWCWD Avionics Department. (2013). *Electronic Warfare and Radar Systems*. Avionics Dept AIR-4.5, Washington DC.

Neri, F. (2006). *Introduction to Electronic Defense Systems*. Artech House Radar Library, Paverstone Dr Ste B, Raleigh, North Carolina.

Neves, S.R., De Oliveira, A., Serra, R., Segadilha, L.E., Monteiro, F. & Lopez, J. M. (2016). Using wavelet packets to analyze FM LPI radar signals. *Proc. IEEE Sensor Array Multichannel Signal Proc. Workshop (SAM 2020)*, pp. 1–4.

Numlk, N. (2000). *Electronic Warfare Fundamentals*. Det 8, ACC TRSS. Nellis AFB NV.

Sklar, B. (2001). *Digital Communications: Fundamentals and Applications*. Prentice-Hall PTR, Upper Saddle River, New Jersey.

Srinath, M. D., & Rajasekaran, P. K. (1996). *An Introduction to Statistical Signal Processing with Applications*. Prentice Hall, Englewood Cliffs, New Jersey.

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. Oxford University Press, Madison Avenue, New York.

Wiley, R. G. (2006). *ELINT: The Interception and Analysis of Radar Signals*. Artech House, Norwood, MA.

Wolff, C. (2020). Radartutorial. Available online at: http://www.radartutorial.eu/index.en.html (Last access date: 25 July 2020)

SYNTHETIC APERTURE RADAR (SAR) SIGNAL SIMULATION USING RANGE DOPPLER ALGORITHM (RDA)

Chan Yee Kit*, Lee Yung Chong & Koo Voon Chet

Centre for Remote Sensing & Surveillance Technologies, Faculty of Engineering & Technology,
Multimedia University, Malaysia

*Email: ykchan@mmu.edu.my

ABSTRACT

In this paper, a brief working principle of synthetic aperture radar (SAR) will be covered. The details of the range Doppler algorithm (RDA) will be studied including the mathematical models of the involved transmitted and received signals as well as the algorithm. The implementation and algorithm development of the RDA will be carried out in MATLAB. The processing steps of the RDA will be illustrated in detail together with the SAR raw signal simulation as point targets and its mathematical models. Several SAR signal properties will be highlighted as these parameters are relatively important in order to develop the algorithm for focusing SAR data. Analytical expressions of baseband SAR signal that consist of point targets are derived and they are used to generate the SAR raw data in MATLAB for further SAR processing. The derivation of the expressions allows the design of the matched filters that are used in range compression and azimuth compression in the RDA in order to process the SAR raw data. The SAR raw data will be processed by using the developed RDA and will be illustrated in the form of SAR images.

Keywords: *Synthetic aperture radar (SAR); range Doppler algorithm (RDA); signal modelling; SAR simulation; SAR image formation.*

1. INTRODUCTION OF SYNTHETIC APERTURE RADAR (SAR) SIGNALS

SAR is an active sensor transmitting its own illumination and is capable of obtaining an image based on the reflectivity of a target or earth surface (Chan & Koo, 2008; Bovenga, 2020). Generally, the working principle of SAR involves several operations and processes. Firstly, the radar in the SAR system will transmit an electromagnetic pulse through a transmitting antenna. The scattered energy pulse from targets will be received by the receiving antenna and sampled by the system. The received echo at a particular point along the flight path is called the SAR slant range sample. This action is repeated to collect range samples at various along-track (azimuth) locations, which synthesises a very long antenna aperture size. By doing so, this will gain a fine resolution in the flight path direction via SAR signal processing (Chan & Koo, 2008; Curlander & McDounough, 1991; Bovenga, 2020). With the proper signal processing done on the raw data, SAR is capable to calculate the distance, detect a target it tracks from the SAR system and form a radar image based on the measure of the scene reflectivity. Figure 1 depicts the geometric model of a SAR with a beam footprint pointed on the earth.

SAR signals are acquired as raw data in the two dimensional time domain and these data are often transformed into other domains for processing efficiency purposes (Cumming & Wong, 2005). The domains involved typically are range Doppler domain or two dimensional frequency domain. SAR raw data itself does not visualise and represent any useful information on the capture scene and is not an image in any form. Instead, it consists of the signal properties of the illuminated scene reflectivity and has to be processed via image formation algorithm to produce a radar image and subsequently applications such as targets detection and classification can be performed (Curlander, & McDounough, 1991; Cumming & Wong, 2005).

The main objective of SAR data processing is the identification of the range and azimuth coordinates for the targets that are lying in the strip-map. In radar systems, the received echo signal looks very similar to the transmitted pulse but with phase differences. It is a one-dimensional signal with voltage as a function of time as shown in Figure 2. Each of the segment in the signal indicates the ground echo received during each pulse cycle. The essential information is located in the phase of the received signal. Thus, further processing is required in order to retrieve the phase information for the formation of a focused image (Vant *et al.*,1978; Wu *et al.*, 1982).

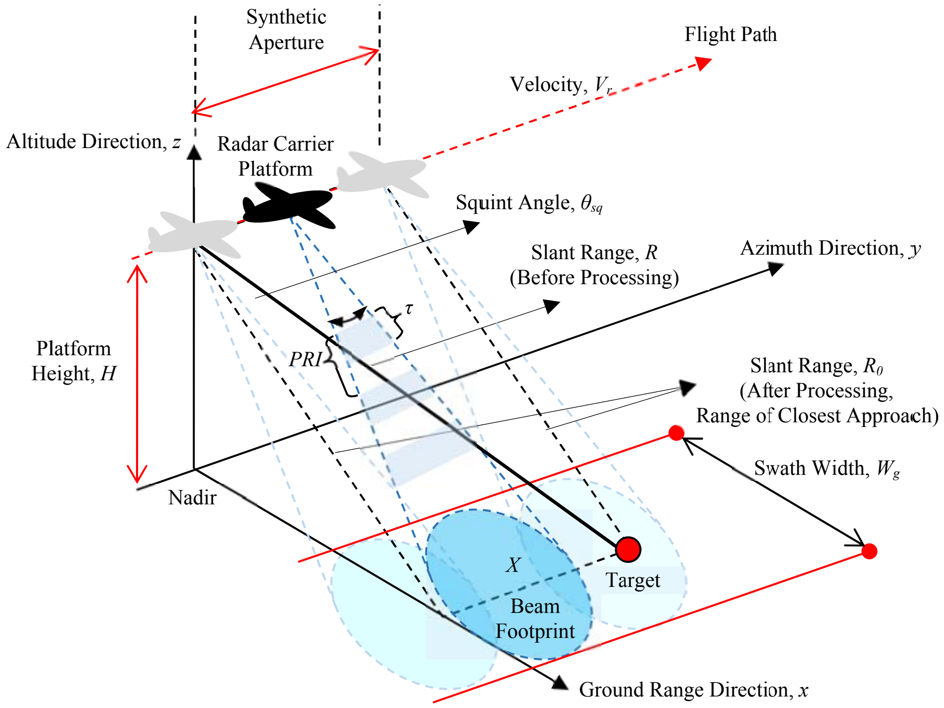


Figure 1: Three-dimensional ideal stripmap SAR imaging geometry model.

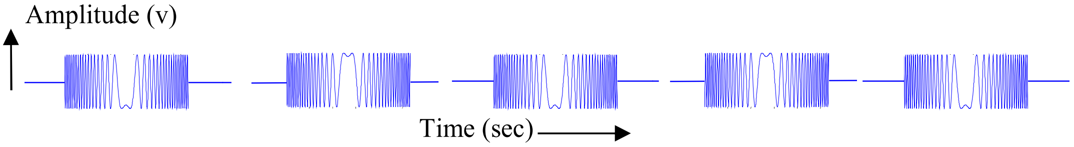


Figure 2: Samples of the received radar signal.

It is necessary to understand the characteristics of the received SAR signal and the important processing parameters prior to the SAR signal processing algorithms. For SAR data, they are acquired and stored in two-dimensional time domain, which are range time and azimuth time (Chan & Koo, 2008; Zhu *et al.*, 2020) The received echo signal will form a two-dimensional complex data matrix. Each of the complex sample consists of the real and imaginary parts, which represent the amplitude and phase values of the SAR raw data. The range direction is the dimension with a range line that comprises the complex echo signals after being down converted to baseband and digitised for memory storage. The SAR sensor captures a range line when it is traveling at a distance of $V_r \cdot PRI$ where V_r is the velocity of the SAR platform and *PRI* is the pulse repetition interval of the transmitting pulse. Hence, this forms the other dimension of the data matrix called azimuth. Figure 3 portrays the two-dimensional SAR data memory placement. The target as in the figure is just entering and leaving the radar beam footprint when the SAR sensor is at points *A* and *B* respectively. The received signal from the target is stored into a row of the SAR data memory, whereby the memory is the input memory to the SAR signal processor. More pulses are being transmitted as the sensor moves forward according to the flight path. These associated echoes are then acquired and written into successive rows in the data memory. The whole process is repeated until the required synthetic length is achieved. Several assumptions are made in Figure 3 in order to simplify the explanation of SAR data memory placement. Only a single target is assumed in the capture scene and the radar beamwidth is assumed to be finite in azimuth. However, naturally the data memory captured from the received signal contains more than one target. The azimuth beamwidth also is not finite in the actual world. Thus, the scattered energy received from the azimuth sidelobes based on each target is also recorded before point *A* and after point *B*.

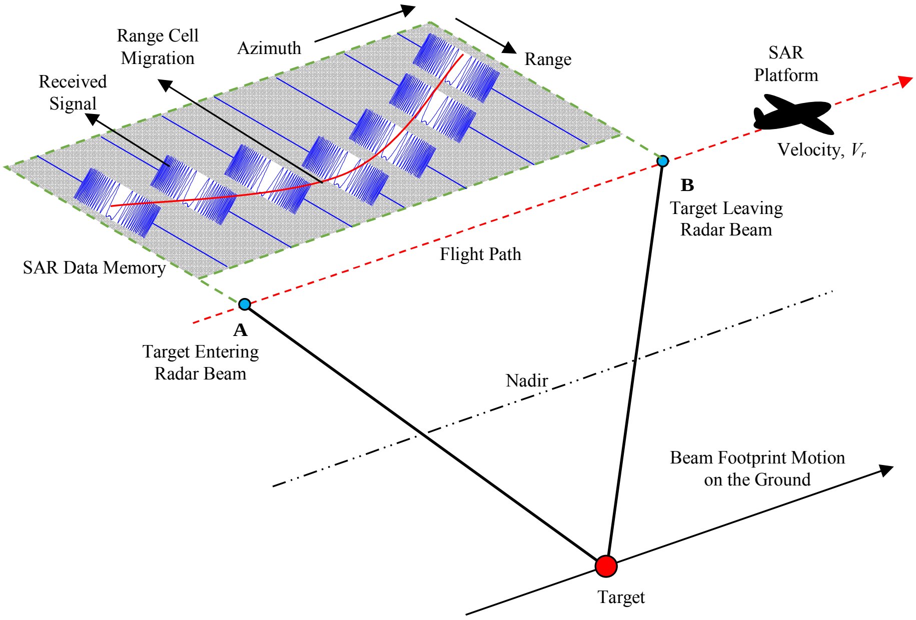


Figure 3: Two-dimensional SAR data memory placement.

2. PROPERTIES OF SAR SIGNAL

In order to develop the SAR processing algorithm, it is relatively important to study and understand the parameters and characteristics of the SAR signal prior to the SAR processing. These signals are referred to as transmitted signal and received signal in both carrier band and baseband. In this paper, airborne SAR is considered and therefore, the geometrical parameters that arise from satellite borne SAR can be neglected. The following section provides a detailed discussion on the important parameters to be considered during the processing of SAR raw data, as well as the SAR processing algorithm to be used.

2.1 Transmitted Signal

The SAR signal is firstly considered in range or beam direction then followed by azimuth direction. The signal coupling in between the range and azimuth coordinates will be clearly visible in the later part. In the direction of range of airborne SAR, the radar transmits a linear frequency modulated (LFM) pulse $s_{TX}(t)$ as denoted by:

$$s_{TX}(t) = \omega_r(t)\cos\left[2\pi f_0 t + \pi\beta t^2\right] \quad (1)$$

The transmitted LFM pulse is a function of range time *t* with a carrier frequency f_0. The pulse duration and the chirp sweep rate are labelled as τ and β respectively, whereas the bandwidth of the pulse is equivalent to the product of τ and β. The pulse envelope of the transmitted LFM pulse $\omega_r(t)$ is the transmit window that defines the transmission duration. This pulse envelope is approximated by a rectangular function such that $\omega_r(t) = rect\left[\frac{t}{\tau}\right]$. The chirp signal is transmitted in an interval called pulse repetition interval (PRI) or its frequency form called pulse repetition frequency (PRF). When the radar is not transmitting any signal, it stays in listening mode and receives echoes scattered from surfaces or objects on the ground (Sommer & Ostermann, 2019). In airborne SAR, the system will receive each echo directly after the transmitted pulse and before the next pulse is being transmitted, as illustrated in Figure 4.

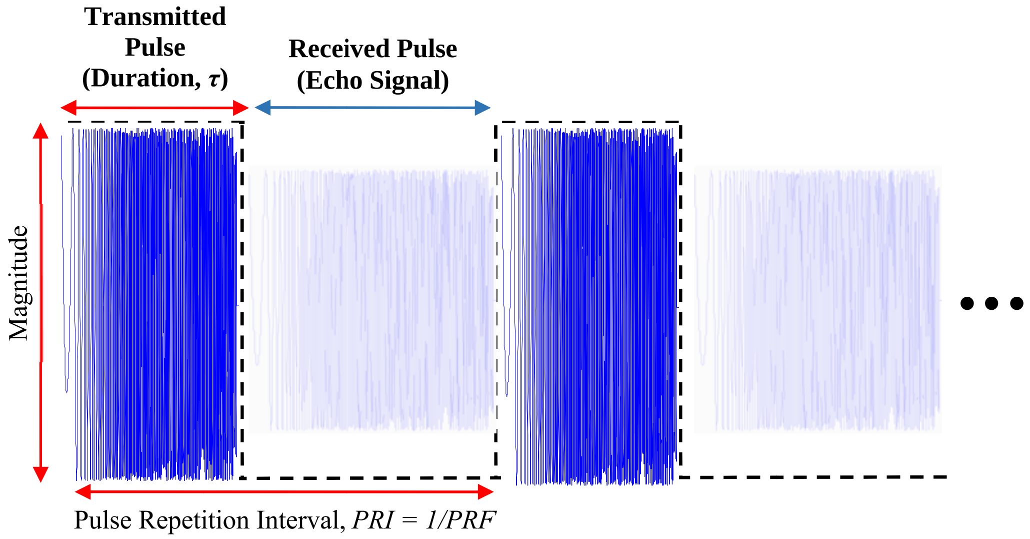


Figure 4: Transmitting and receiving cycles of a pulsed radar.

2.2 Received Signal (Raw Data)

In the point of view of the received SAR signal, the received signal needs to be demodulated into a baseband signal before it is treated as raw data. The demodulation process is done through the radio frequency (RF) subsystem in the radar system. The RF subsystem will perform quadrature demodulation in order to filter out the carrier frequency signal and maintain the baseband signal as SAR raw data for further processing. The quadrature demodulation is illustrated in Figure 5.

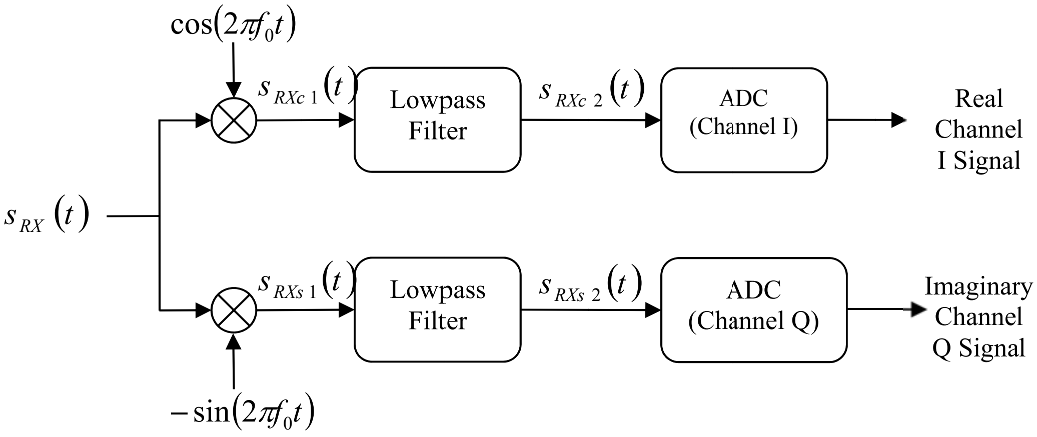


Figure 1: Quadrature demodulation of the received SAR signal.

The quadrature demodulation process demodulates the received signal, $s_{RX}(t)$ from carrier band into baseband and produces a complex value output with real channel (I) and imaginary channel (Q). These two channels will be sampled and digitised by the analogue to digital converter (ADC) so that digital SAR signal processing algorithm can be further performed in order to obtain the SAR image. The mathematical expression of the received signal in carrier band $s_{RX}(t)$ is given by:

$$s_{RX}(t) = \cos[2\pi f_0 t + \varphi(t)] \tag{2}$$

Where f_0 is the carrier frequency (GHz) with various orders of magnitude that is much higher than the bandwidth of the desired modulation $\varphi(t)$ (MHz). In the upper channel, the signal is firstly multiplied by a term of $\cos(2\pi f_0 t)$ using the following trigonometric identity:

$$\cos A \cos B = \frac{1}{2}\cos(A+B) + \frac{1}{2}\cos(A-B) \tag{3}$$

This will lead to the following multiplication result:

$$s_{RXc1}(t) = \frac{1}{2}\cos[4\pi f_0 t + \varphi(t)] + \frac{1}{2}\cos[\varphi(t)] \tag{4}$$

The first cosine term in Equation 4 is the signal with carrier frequency centred around $2f_0$ while the second cosine term is the baseband signal governed by the bandwidth of $\varphi(t)$. Hence, the first carrier frequency cosine term can be removed using a low pass filter and the resultant of the filtered signal is given by:

$$s_{RXc2}(t) = \frac{1}{2}\cos[\varphi(t)] \tag{5}$$

In the lower channel, as in Figure 5, the signal is firstly multiplied by a term of $-\sin(2\pi f_0 t)$ using another trigonometric identity given by:

$$\cos A \sin B = \frac{1}{2}\sin(A+B) + \frac{1}{2}\sin(A-B) \tag{6}$$

This multiplication will express the signal as the sum of low and high frequency components denoted as:

$$s_{RXs1}(t) = \frac{1}{2}\sin[4\pi f_0 t + \varphi(t)] + \frac{1}{2}\sin[\varphi(t)] \tag{7}$$

Similarly, the first sine term in Equation 7 is the signal with carrier frequency centred around $2f_0$ while the second sine term is the baseband signal governed by the bandwidth of $\varphi(t)$. Thus, the first carrier frequency sine term can be removed using a low pass filter, with the resultant of the filtered signal being:

$$s_{RXs2}(t) = \frac{1}{2}\sin[\varphi(t)] \tag{8}$$

The signals $s_{RXc2}(t)$ and $s_{RXs2}(t)$ will be sampled by the ADC at a rate that fulfils the Nyquist sampling theorem. The cosine and sine terms multiplications lead to the two signals to be in phase and quadrature phase, hence the signal can be rewritten as a complex signal in the form of:

$$s_{RX3}(t) = s_{RXc2}(t) + j s_{RXs2}(t) = \frac{1}{2}e^{j\varphi(t)} \tag{9}$$

The signal in Equation 8 will be the required baseband signal used in the processing of SAR data. It is noticeable that the sampled $s_{RXc2}(t)$ and $s_{RXs2}(t)$ signals are I and Q channels for in phase and quadrature phase. The phase term $\varphi(t)$ of the demodulated baseband received SAR signal is denoted as:

$$\varphi(t) = -\frac{4\pi f_0 R(\eta)}{c} + \pi\beta\left(t - \frac{2R(\eta)}{c}\right)^2 \tag{10}$$

By combining Equations 9 and 10, the demodulated baseband radar signal received from a point target can be rewritten and modelled in the form of:

$$s_{RX}(t,\eta) = A_0 \omega_r\left(t - \frac{2R(\eta)}{c}\right)\omega_a(\eta - \eta_c)e^{-j\left\{\frac{4\pi f_0 R(\eta)}{c} + \pi\beta\left(t - \frac{2R(\eta)}{c}\right)^2\right\}} \tag{11}$$

where A_0 is an arbitrary complex constant, t is the range time, η is the azimuth time referenced to the closest approach, η_c is the beam centre offset time, $\omega_r(t)$ is a rectangular function for the range envelope, $\omega_a(\eta)$ is a sinc-squared function for the azimuth envelope, f_0 is the radar carrier frequency, β is the range chirp FM rate and lastly $R(\eta)$ is the instantaneous slant range. The $R(\eta)$ can be given as:

$$R(\eta) = \sqrt{R_0^2 + (V_r\eta)^2} \tag{12}$$

The mathematical model of the received signal in Equation 11 will be used to generate all point targets reflections in MATLAB so that SAR processing algorithm can be implemented in order to process the final compressed SAR data or SAR image. In the following sections, all the processing steps involved in the range Doppler algorithm (RDA) will be discussed and explained using a simulated airborne C-band SAR raw data. These parameters are tabulated in Table 1.

Table 1: Simulation parameters of C-band airborne SAR.

Name of Parameters	Symbols	Value (Units)
Slant Range Center	$R(\eta_c)$	30 *km*
Transmitted Pulse Duration	T	2.5 *μs*
Chirp Signal Bandwidth	B	50 *MHz*
Range Chirp Signal Sweep Rate	β	20 *MHz/ μs*
Range Resolution	δ_R	3 *m*
Range Sampling Frequency	f_s	100 *MHz*
Carrier Signal Frequency	f_0	5.3 *GHz*
Azimuth Sampling Frequency (PRF)	f_a	350 *Hz*
Radar Platform Velocity	V_r	150 ms^{-1}
Real Antenna Aperture Length	L_a	1 *m*

3. IMPLEMENTATION AND DEVELOPMENT OF RANGE DOPPLER ALGORITHM (RDA) ON SAR RAW DATA

3.1 Introduction to SAR Image Formation Algorithm

Early SAR systems adopted optical processing instead of computer digital processing due to the high volume of SAR raw data and the complicated image formation algorithm, SAR signal processing was not able to be performed using a computer because the processing and memory capability could not meet the minimum requirement at that time. Well-focused images were able to be produced using optical processors although it required precise alignment of lenses. In 1957, Michigan University, USA used an optical processor to obtain the first focused X band stripmap sidelooking SAR image (Cutrona *et al.*, 1966; Brown & Porcello, 1969). In the 1950s and 1960s, the lack of fast computing machines and advanced digital processing algorithms discouraged the development of wavefront reconstruction based SAR image formation techniques.

Compared with optical processing, digital processing is more accurate and more flexible. Digital processing can adopt all kinds of methods to correct range migration, phase errors due to motion of the platform, and realise accurate image formation. With the advancement of semiconductor technology and digital signal processing techniques, SAR signal processing gradually migrated from optical to digital processing. At the end of the 1960s, the Environmental Research Institute of Michigan successfully developed a dual frequency, dual polarised airborne SAR system (Rawson & Smith, 1974), which employed a digital processor for non-real-time formation of images. During the 1970s, the rapid development of semiconductor and computer technologies offered the possibility of SAR digital signal processing. Concentrated effort in the researches of digital SAR image formation algorithms and processing was made, which led to various SAR projects and SEASAT programmes (John & Kirk, 1975; Cumming & Bennett, 1979; Vant *et al.*, 1979; Bennett *et al.*, 1980; Wu, 1980; Cumming & Wong, 2005).

The SAR return echo are transformed into other domains such as the range Doppler domain or two dimensional frequency domain for better processing efficiency. The phase characteristic of targets can be determined by the relationship of the instantaneous range of the targets with respect to the sensor. When the range of the received echo is visualised versus time, its curve passes through several range cells, with this range variation property being referred to as range cell migration (RCM) and it imposes a frequency modulation (FM) characteristic on the signal in the azimuth direction (Long *et al.*, 2019). It complicates the signal processing, but it is a significant feature of SAR. Existing and commonly used high precision SAR processing algorithms include RDA, chirp scaling algorithm (CSA) and wave number / Omega-K algorithm (ωkA) (Cumming & Wong, 2005).

RDA is the most widely used satellite SAR image formation algorithm. This algorithm was developed to process SEASAT SAR data in 1976 – 1978 (Wu *et al.*, 1982; Jin & Wu, 1984). RDA utilises frequency domain operation in both range and azimuth directions to achieve block processing efficiency. All processes of RDA are done in simple one-dimensional operations. Range cell migration correction (RCMC) is implemented with interpolation between the two one-dimensional operations i.e. the range-Doppler domain, subsequently simplifying the two-dimensional azimuth reference to one-dimension, thus improving the quality of the image. However, the interpolation increases the amount of computation. When the squint angle increases, its azimuth phase history is no longer a linear FM-like signal, and greatly influences the precision of the image. In order to solve this problem, RDA with secondary range compression (SRC) (Jin & Wu, 1984; Curlander & McDounough, 1991) was proposed. Other algorithms were also proposed to improve RDA shortly afterwards, such as the squint imaging mode method (Chang *et al.*, 1989) and the subaperture approach algorithm (Moreira, 1992). The basic RDA implementation and its shown in Figure 6.

ωkA (Caforio, *et al.*, 1991; Li & Lofeld, 1991; Scheuer & Wong, 1991; Bamler, 1992; Prati & Rocca, 1992) and CSA (Raney, 1992; Moreira & Huang, 1994; Raney *et al.*, 1994; Davidson *et al.*, 1996; Moreira *et al.*, 1996) are two main two-dimensional image formation methods. ωkA is a precision SAR image formation algorithm where only multiplication is employed for the matched filtering process and no approximation takes place in the image reconstruction process. However, an unavoidable shortcoming of this algorithm is the increased number of computation operation. ωkA uses a special operational in the two-dimensional frequency domain that corrects the range dependence of the range-azimuth coupling and the azimuth frequency dependence. It is able to process high squint and wide apertures with efficient and accurate processing. The bulk compression in two dimension frequency domain partially focuses the targets since the reference function is computed for a selected range. The key step in ωkA, i.e. the Stolt interpolation, is applied to focus the remaining targets. However, the Stolt interpolation is very computationally intensive. It can be replaced by an approximation where a phase multiply is used after the bulk compression to correct the differential azimuth compression. Nevertheless it cannot reduce the error of differential RCMC. The implementations of ωkA is shown in Figure 7.

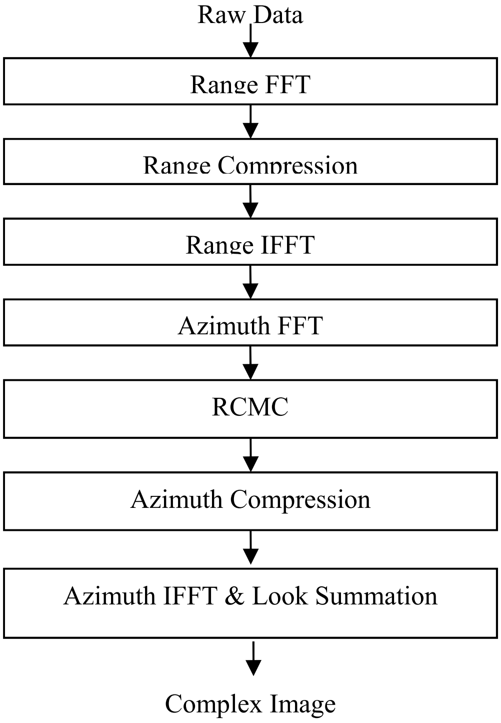


Figure 6: Implementation of RDA.

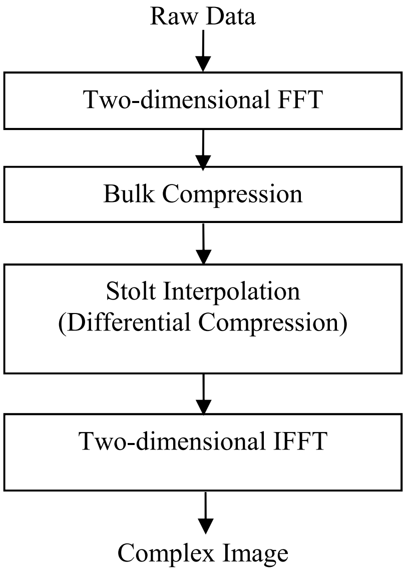


Figure 7: Implementation of ωKA.

In CSA, a frequency modulation is employed to “chirp scale” the received echo in range Doppler domain. Thus, RCMC can be applied in frequency domain instead of using time domain interpolation. The chirp scaling process produces the same RCM for all ranges. Therefore, RCM can be easily corrected using the phase multiply of the reference function. The bulk compression, including RCMC, range compression (RC) and SRC, is performed in a two-dimensional frequency domain. CSA eliminates the interpolation that is required in the RDA RCMC. In the range doppler domain, the range dependent residual correction is performed. A description of CSA is shown in the functional block diagram in Figure 8.

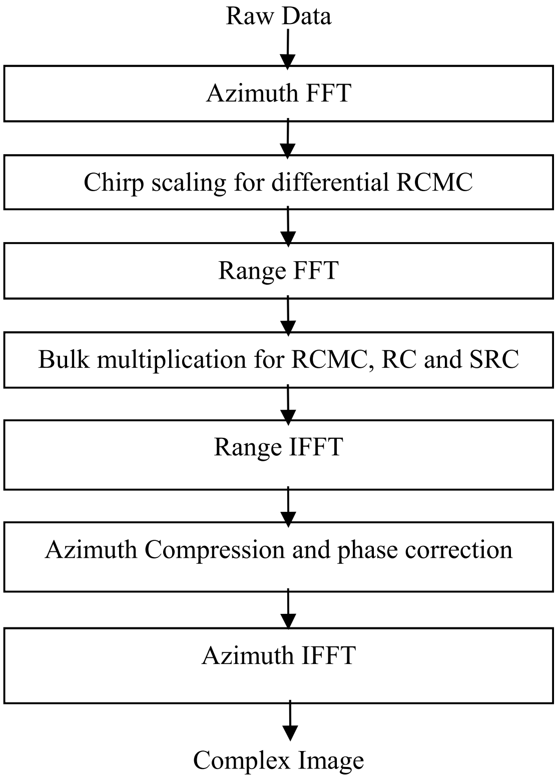

Raw Data
Azimuth FFT
Chirp scaling for differential RCMC
Range FFT
Bulk multiplication for RCMC, RC and SRC
Range IFFT
Azimuth Compression and phase correction
Azimuth IFFT
Complex Image

Figure 8: Implementation of CSA .

3.2 RDA Implementation and Simulations

Most of the SAR processing algorithms work in the frequency domain for reasons of efficiency. In the context of SAR, one of the suitable domains for efficient processing is the range Doppler domain (range time versus azimuth frequency). The distinguishing feature of RDA is the ability to achieve block processing efficiency with the use of frequency domain operations in both range and azimuth directions while maintaining the simplicity of one dimensional operations. This effectively decouples range and azimuth dependencies in the processing operations and simplified SAR processing problem as it allows the two dimension operations to be processed separately. The computation load of RDA is relatively low as compared to other algorithms. Therefore, the RDA is preferable to be implemented in SAR signal processing and image formation.

Generally, RDA can be treated as a two-dimensional correlating procedure operation. The two dimensions of the correlation processing are implemented as two separate one-dimensional matched filter operations. Three principal steps of the algorithm are range compression, RCMC and lastly azimuth compression. Both range and azimuth compressions are the one-dimensional matched filter operations. In signal memory, often the received data consists of the Doppler shift due to the change in range of the target when the SAR sensor is synthesising the aperture. These trajectories migrate through the range cells during the exposure time of the target, with this phenomenon being known as RCM, which complicates the processing. Therefore, RCMC as an explicit processing operation is required in order to correct such migration. However, the necessity of performing RCMC relies on the fineness of the range samples. If the change is in range or the RCM is lesser than one sample size (one pixel) or the range resolution, RCMC is unnecessary as the effect arising from RCM is not significant and unobservable in the compressed data (Smith, 1991). Figure 9 graphically portrays the functional block diagram of RDA in detail. A SAR sensor transmits a modulated carrier band signal towards an area of interest in a specified PRI and the echo scattered from the area of interest will be received via antenna. The received signal will be demodulated into the baseband signal by the SAR sensor internal RF subsystems. The demodulated received signal will be sampled and digitised by the ADC with a sampling rate that fulfils the Nyquist sampling theorem. The digitised received baseband signal is also known as SAR raw data. In RDA, there are several processing steps involved, as shown in Figure 9, which are range compression, RCMC and azimuth compression. Each processing step will be discussed and implemented in detail in the following sections.

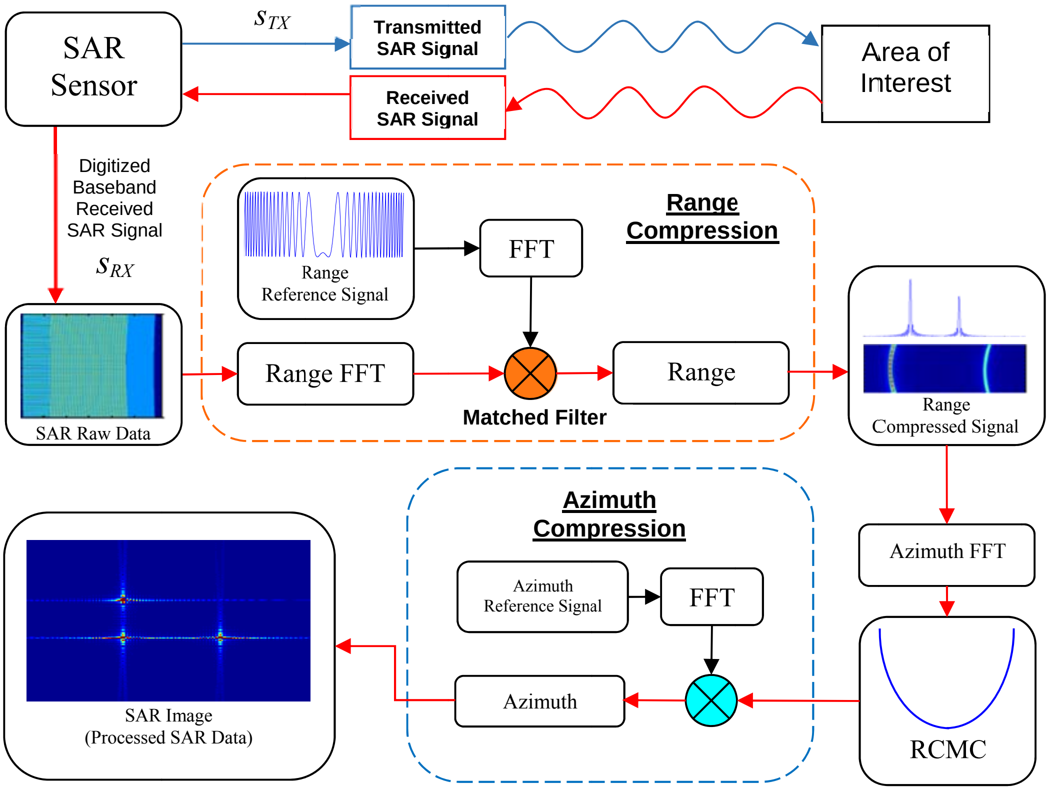


Figure 9: Detailed functional block diagram of RDA.

The range compression step compresses the data into a narrower pulse by performing a correlation of the data with a range reference function. It consists of three operations, which are range FFT, range matched filtering and lastly range IFFT. The digitised received SAR raw data is firstly retrieved from the memory in order to start the processing. Since the SAR raw data is sampled in time domain, the data is required to be transformed into frequency domain using range FFT in order to perform the range matched filtering operation. This operation is performed in frequency domain for speeding up the processing time. The matched filter is designed based on the transmit signal of the SAR system. The impulse response of the match filter is simply the scaled, time reversed and delayed version of the transmitting signal. The range matched filtering is done in the domain called range-azimuth domain, which is range frequency domain and azimuth time domain.

After filtering by the matched filter, the received signal in frequency domain is then transformed back into time domain using IFFT in order to obtain the range compressed data, which divulges information about the relative distance between the radar and ground of the observed scene. For the azimuth compression, it follows the same basic reasoning as in range compression. The only difference is the frequency domain azimuth reference function. Since azimuth compression is performed in frequency domain, performing azimuth FFT on the output data from range compression is required. RCMC can be carried out in this domain according to the amount of RCM of the trajectories. These data are then multiplied with the frequency domain matched filter. After the matched filtering, the compression is done by performing IFFT on the output of the matched filtering, resulting in the SAR compressed data or SAR image.

In order to implement RDA for SAR raw data processing, point targets are simulated in MATLAB. The geometric position of the scene with the simulated point targets is illustrated in Figure 10, where there are three point targets with overlapped range and azimuth in the received signals. Target A is overlapped with Target C in range direction, which means that these two targets have the same slant range of the closest approach. As for Targets B and C, both targets have the same azimuth time as they are crossed by the beam centre of the SAR sensor. In the range profile of the range compressed signal, it is expected to observe that there are only two compressed pulses with one pulse having higher magnitude while the other with lower magnitude. The higher magnitude compressed pulse is the one consisting of reflections from targets A and C, and they have the same slant range, which causes the target energy to overlap with each other, while target B is separated from the other two due to different slant range. Targets A and C can be separated only after azimuth compression as they cross the centre beam with different azimuth times.

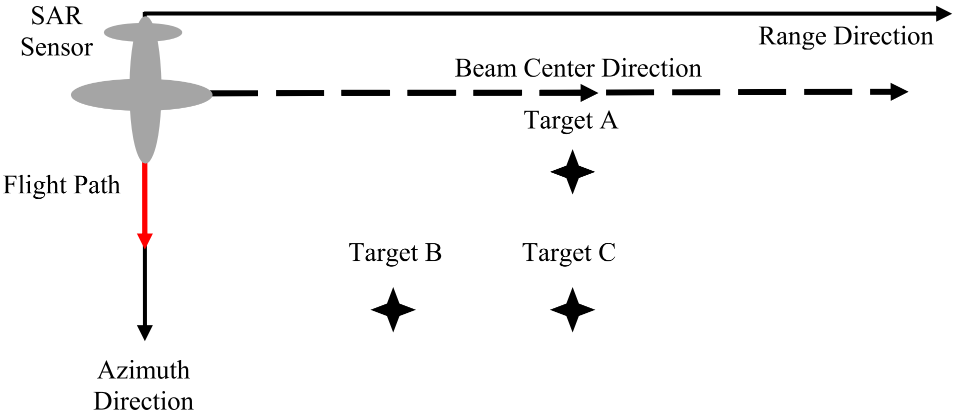


Figure 10: Point targets simulation scene for RDA processing.

The SAR raw data that is modelled in Equation 11 is used to simulate the received signal that consists of the three point targets as depicted in Figure 10. The simulated raw data with three targets is illustrated in Figure 11 using jet scale maps with the real and imaginary parts of the echo signal in complex form. The cooler (darker) colours represent lower reflections from the targets, while hotter colours (brighter) indicate stronger reflections from the targets. Figure 12 illustrates the magnitude and the phase of the simulated raw data generated from Figure 10.

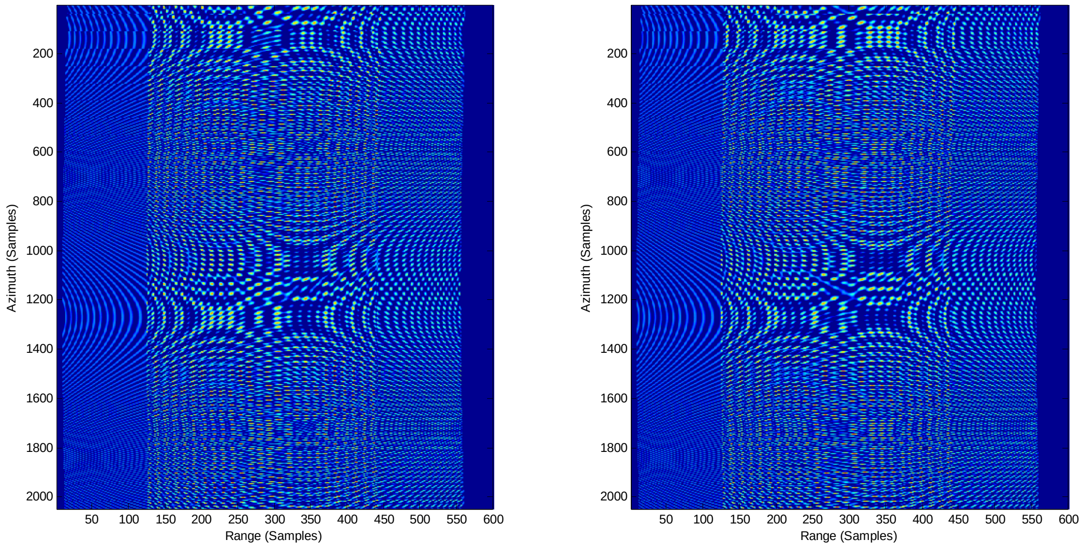


Figure 11: Real (left) and imaginary (right) parts of the simulated raw data.

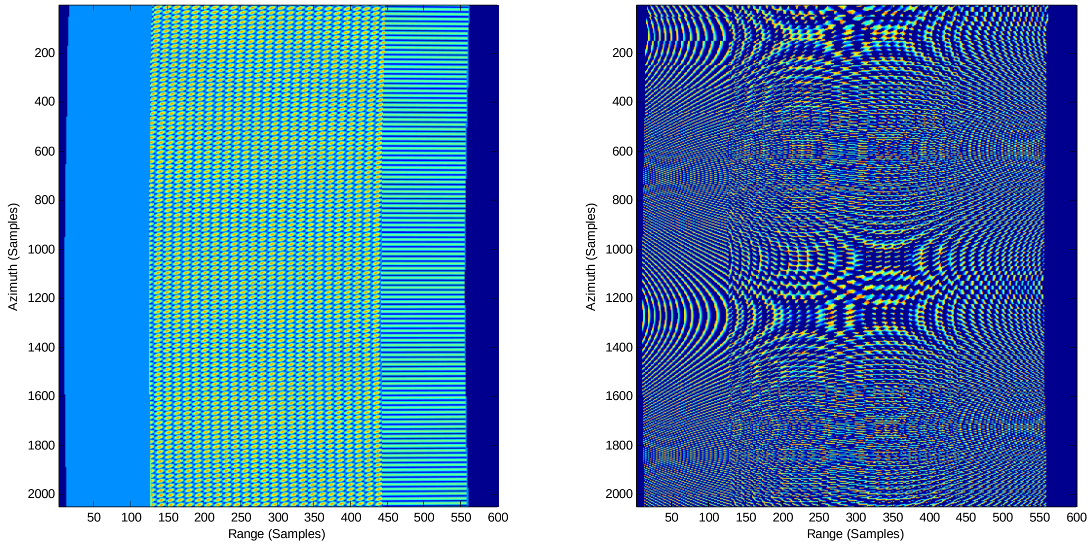


Figure 12: Magnitude (left) and phase (right) of the simulated raw data.

3.2.1 Range Compression

The range profile of the SAR raw data can be retrieved by performing the range compression processing steps in RDA. Matched filtering is the key technique used in this processing step in order to compress the raw data in range direction. The transmitted signal of the SAR stated in Equation 1 will be used as the range reference function for the design of the matched filter. The filter is generated by taking the complex conjugate of the replica in Equation 1 and transforming it into the frequency domain using the range Radix-2 FFT algorithm. In order for the matched filter to have the same length of the raw data, zero padding is performed with the frequency domain filter by adding zeros to the end of the filter. Figures 13 and 14 show the reference function of the matched filter in both time and frequency domains.

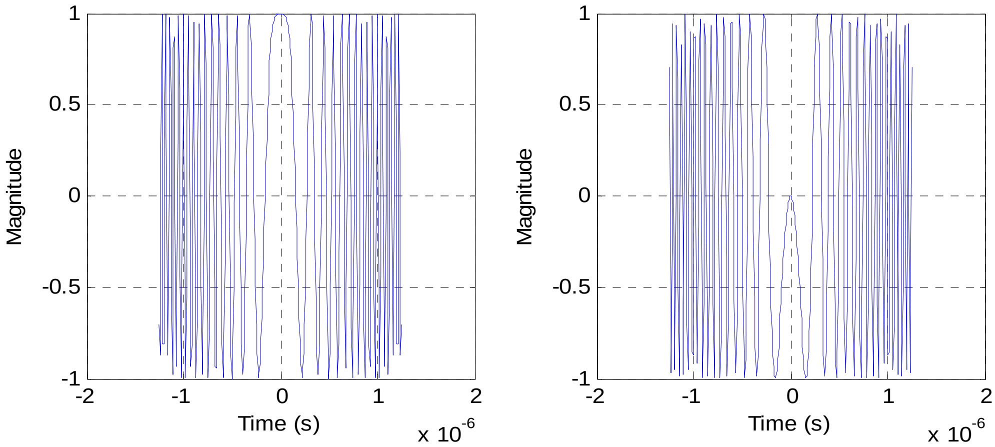


Figure 13: Real (left) and imaginary (right) parts of the matched filter reference function.

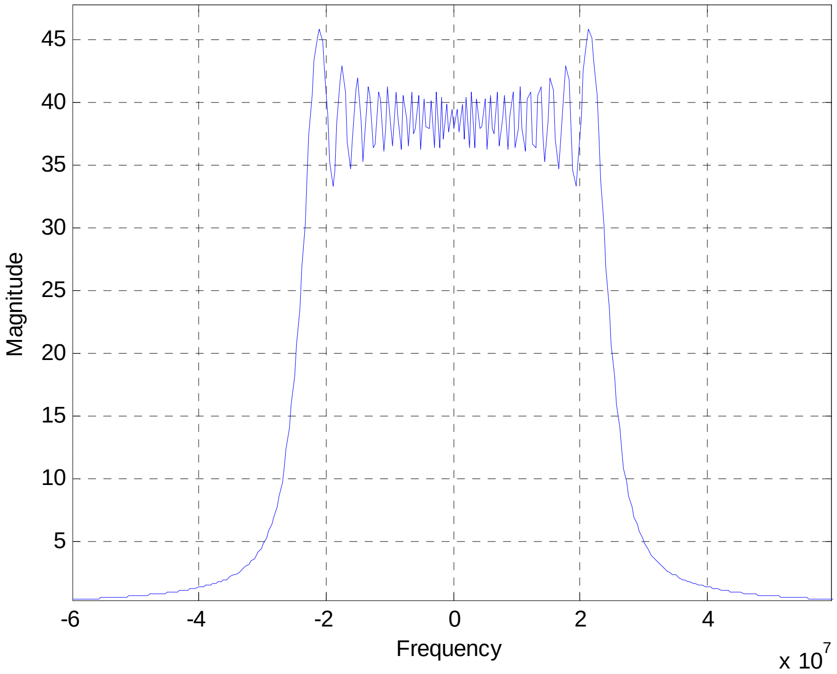


Figure 14: Matched filter reference function in the frequency domain.

After the transformation, the azimuth direction of the data still remain in the time domain, with only the range direction being transformed into the frequency domain. The data and the generated filter now both will be in the frequency domain. With $S_{RX}(f,\eta)$ being the range FFT signal of $s_{RX}(t,\eta)$ as in Equation 11 and the matched filter in frequency domain being denoted as $H_r(f)$, the output of the matched filter $s_{RC}(t,\eta)$ can be expressed in the form of:

$$\begin{aligned} s_{RC}(t,\eta) &= IFFT_t\{S_{RX}(f,\eta)\times H_r(f)\} \\ &= A_0 p_r\left(t-\frac{2R(\eta)}{c}\right)\omega_a(\eta-\eta_c)e^{-j\frac{4\pi f_0 R(\eta)}{c}} \end{aligned} \tag{13}$$

From Equation 13, A_0 is the overall gain that includes the scattering coefficient, which is kept as unity in this simulation, while $p_r(t)$ is a sinc-like range envelope that incorporates the target range migration via the azimuth varying parameters $\frac{2R(\eta)}{c}$. The parameters $\omega_a(\eta\text{-}\ \eta_c)$ and $e^{-j\frac{4\pi f_0 R(\eta)}{c}}$ are the azimuth envelope and phase respectively, which are not affected by the range compression process.

The matched filtering operation is a correlation in the time domain or equivalent as multiplication operation in the frequency domain. As a resultant, the transmitted signal will be removed and filtered out by the matched filter and the output of the filter will only contain the scattered targets received by the SAR sensor. The matched filter output will be transformed back into the time domain using a range IFFT so that the range compressed profile can be observed. Figures 15 illustrates the range compression results of the simulated point targets as mentioned in Figure 10. It can been seen that targets A and C both are having the same slant range of the closest approach, whereas target B is at different range that is closer to the SAR sensor. Thus, from the range profile shown in Figure 15, two compressed pulses can be observed. The first compressed pulse is scattered from target B while the second compressed pulse is scattered from targets A and C. The amplitude of the second pulse is slightly higher than the first one because both targets A and C are located at the same range location. Similar conditions can be seen by viewing the range compressed data in range and azimuth directions as illustrated in Figure 16. The point targets in the range compressed data can be noticed as the trajectories, with the range compressed image showing the interference between the two targets A and C, which are coincident with each other in the range direction. It is important to take note that the RCM effect in the time domain can be observed in Figure 16 as the compressed trajectories that are in curve form instead of straight form. The next processing steps will be the azimuth FFT in order to transform the range compressed data into range Doppler domain for the operations of RCMC and azimuth compression.

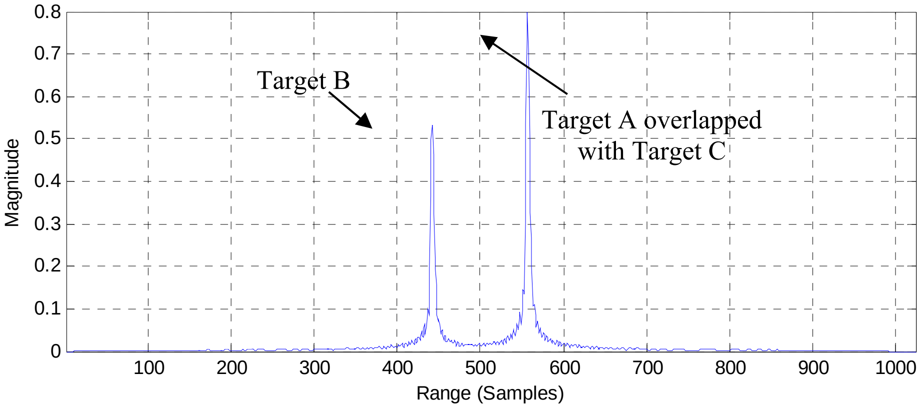


Figure 15: Range profile results of range compression.

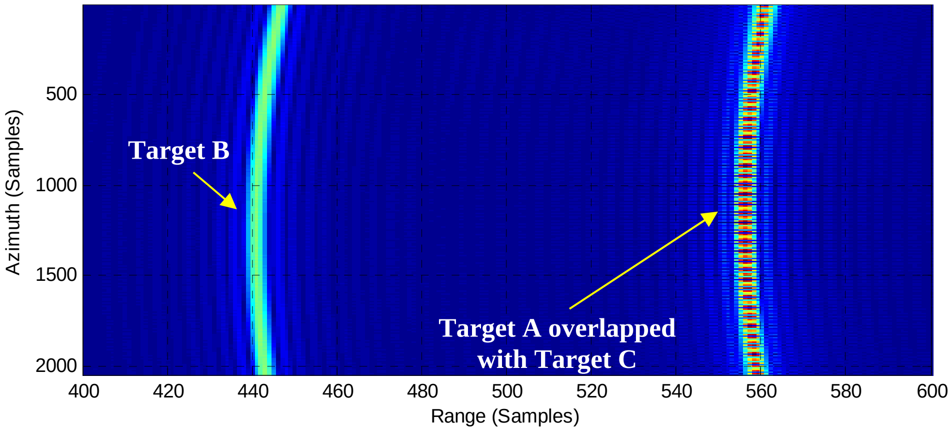


Figure 16: Magnitude SAR image of the range compressed data.

3.2.2 Azimuth FFT

Similar to the range FFT, the FFT algorithm used in the azimuth direction is the same as the range direction, which the Radix-2 FFT. The azimuth FFT will transform the range compressed data into range Doppler domain, which are the range time and azimuth frequency domains respectively. In this low squint airborne C-band SAR, the beam of the antenna pointing in the direction is close to the zero

Doppler direction. Thus, the range equation can be approximated by the parabolic equation as in Equation 12 by justifying from the assumption that R_0 is relatively much greater than $V_r\eta$ in the low squint airborne SAR case:

$$R(\eta)=\sqrt{R_0^2+(V_r\eta)^2}\approx R_0+\frac{(V_r\eta)^2}{2R_0} \tag{14}$$

By substituting Equation 14 into Equation 13, the range compressed data can be rewrite as:

$$s_{RC}(t,\eta)=A_0p_r\left(t-\frac{2R(\eta)}{c}\right)\omega_a(\eta-\eta_c)e^{-j\left\{\frac{4\pi f_0R_0}{c}+\frac{2\pi(V_r\eta)^2}{R_0\lambda}\right\}} \tag{15}$$

The second exponential term as in Equation 15 is the azimuth phase modulation. It can be seen that the phase term is a function of azimuth time square η^2, thus the signal also consists of the linear FM characteristics with the azimuth LFM rate of:

$$K_a=\frac{2V_r^2\cos^2\theta_{r,c}}{R(\eta_c)\lambda}\approx\frac{2V_r^2}{R_0\lambda} \tag{16}$$

From Equation 16, the $cos^2\theta_{r,c}$ term is equal to unity for a small squints angle. The relationship between time and azimuth frequency can be expressed as:

$$f_\eta=-K_a\eta \tag{17}$$

The azimuth FFT will be performed on each range gate in order to transform range compressed data into range Doppler domain. The transformed data can be expressed as:

$$\begin{aligned}S_1(t,f_\eta)&=FFT\{s_{RC}(t,\eta)\}\\&=A_0p_r\left(t-\frac{2R_{rd}(f_\eta)}{c}\right)W_a\left(f_\eta-f_{\eta_c}\right)e^{-j\left\{\frac{4\pi f_0R_0}{c}-\frac{\pi f_\eta^2}{K_a}\right\}}\end{aligned} \tag{18}$$

with the azimuth envelope $\omega_a(\eta-\eta_c)$ being transformed into $W_a(f_\eta-f_{\eta c})$, which is in the range time azimuth frequency domain. The first exponential term in Equation 18 is the inherent phase information, which is meaningful for applications of interferometry, but is unimportant in application of intensity images. As for the second exponential term, it is the azimuth modulation with linear FM characteristics in f_η. The term of $R_{rd}(f_\eta)$ is the range cell migration in the range enveloped, which is expressed in the range Doppler domain. The migration of the trajectories can be easily obtained by substituting Equations 16 and 17 into Equation 14, which yields:

$$R_{rd}(f_\eta)\approx R_0+\frac{V_r^2\left(-\frac{f_\eta}{K_a}\right)^2}{2R_0}=R_0+\frac{R_0\lambda^2f_\eta^2}{8V_r^2} \tag{19}$$

It is important to take note that the targets with the same slant range of the closest approach will follow the same trajectory in the range Doppler domain. This property will allow the correction of range migration of one trajectory has the effect of correcting the whole family of trajectories with the same range of closest approach. Hence, RCMC can be carried out conveniently in the range Doppler domain.

3.2.3 Range Cell Migration Correction (RCMC)

The RCM from Equation 19 can be significantly observed in Figure 13. The amount of RCM to be corrected is expressed in the second term of Equation 19:

$$\Delta R(f_\eta)=\frac{R_0\lambda^2 f_\eta^{\ 2}}{8V_r^{\ 2}} \tag{20}$$

where RCM equals to the target displacement in the function of azimuth frequency f_η and also function of range variant R_0. The amount of RCM as in Equation 20 is illustrated in Figure 17.

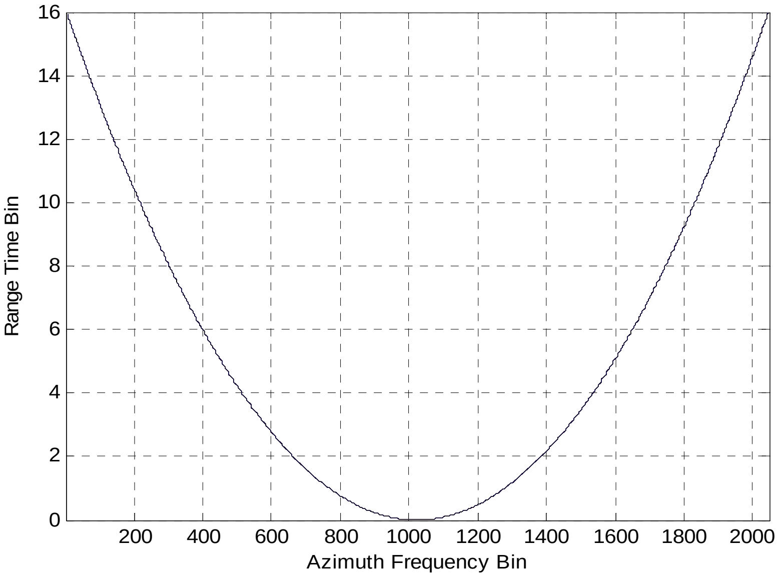


Figure 17: Total RCM.

RCMC can be performed using the range interpolation operation on Equation 18 and the corrected signal now becomes:

$$S_2(t,f_\eta)=A_0 p_r\left(t-\frac{2R_0}{c}\right)W_a\left(f_\eta-f_{\eta_c}\right)e^{-j\left\{\frac{4\pi f_0 R_0}{c}-\frac{\pi f_\eta^{\ 2}}{K_a}\right\}} \tag{21}$$

From Equation 21, it is noticed that the range envelope p_r no longer depends on the azimuth frequency as the RCM is corrected. The range energy now is centred at the position of $t=2\frac{R_0}{c}$ as the range closest approach.

As portrayed in Figure 18, the trajectories in the range compressed data after RCMC are now in straightened form. Similar to the range compressed data without RMCM, only two targets are distinguishable in the figure although in total there are three point targets. This is because Targets A and C are overlapped with each other and collocated at the same slant range. Figure 19 illustrates the range profile of the point targets with RCMC. By comparing Figures 15 and 19, which are the range profiles without and with RCMC respectively, the amplitude for all point targets in Figure 19 is much higher than Figure 15 due to RCMC having concentrated the energy of the targets to the centre position of the point targets.

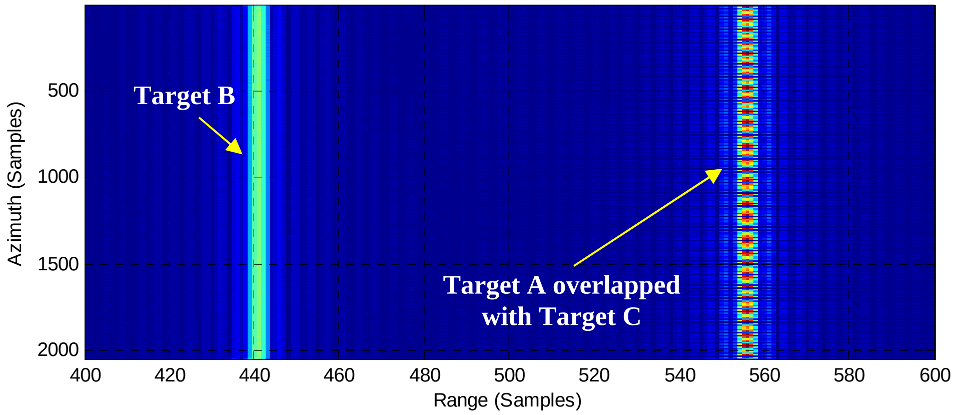


Figure 18: Magnitude SAR image of the range compressed data with RCMC.

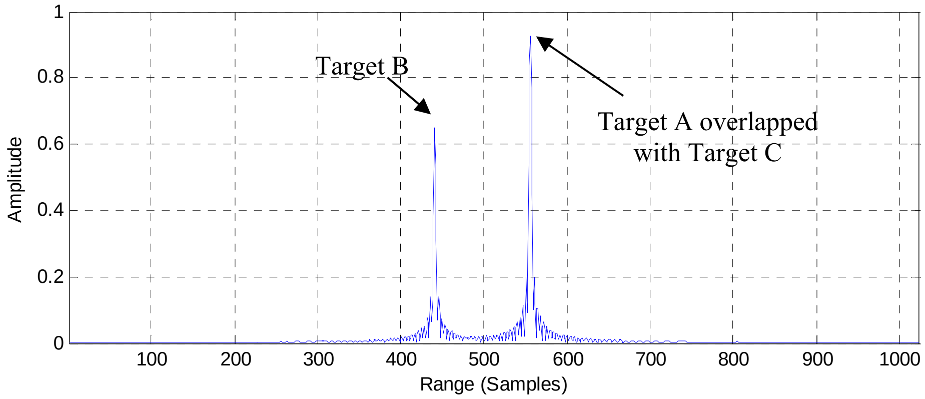


Figure 19: Range profile results of range compression.

3.2.4 Azimuth Compression

Azimuth compression is the final step in RDA for the processing of SAR raw data. The results of the azimuth compression will be the compressed complex SAR image that can be used for further detection and look summation purposes. Similar to the range compression, in azimuth compression, matched filtering technique is applied as well. The difference between the range and azimuth matched filtering is the involved domain for the operations and the reference function of the matched filter. Azimuth matched filtering is performed in the range time azimuth frequency domain. Since the operation of RCMC is performed in range Doppler domain as well, it is convenient to perform azimuth matched filtering immediately after the RCMC operation. The matched filter reference function is derived based on the second exponential term as in Equation 21. The azimuth matched filter is generated by taking the complex conjugate of the exponential term and it is expressed in terms of azimuth frequency f_η and slant range of R_0, where K_a is a function of R_0:

$$H_{az}(f_\eta) = e^{-j\frac{\pi f_\eta^2}{K_a}} \tag{22}$$

The data after RCMC from Equation 21 is multiplied with the frequency domain matched filter in Equation 22 as the processing step of azimuth matched filtering. The result after the azimuth matched filter is denoted as:

$$\begin{aligned} S_3(t,f_\eta) &= S_2(t,f_\eta)\times H_{az}(f_\eta) \\ &= A_0 p_r\left(t-\frac{2R_0}{c}\right)W_a\left(f_\eta - f_{\eta_c}\right)e^{-j\left\{\frac{4\pi f_0 R_0}{c}\right\}} \end{aligned} \tag{23}$$

After the azimuth matched filter, an IFFT operation is required in order to transform the data from the range Doppler domain back into the range time azimuth time domain as the completion of the azimuth compression:

$$\begin{aligned} s_{ac}(t,\eta) &= IFFT\left\{S_3(t,f_\eta)\right\} \\ &= A_0 p_r\left(t-\frac{2R_0}{c}\right)p_a(\eta)e^{-j\left\{\frac{4\pi f_0 R_0}{c}-2\pi f_{\eta_c}\eta\right\}} \end{aligned} \tag{24}$$

where p_a is a sinc-like function similar to p_r. The first exponential term in Equation 24 is the target phase due to its range position R_0, while the second exponential term is the linear phase term arising from $f_{\eta c}$.

Figure 20 illustrates the azimuth compression result of the simulated point targets as in Figure 10. It can been seen that Targets B and C both have the same azimuth time, whereas Target A is at a different azimuth time when crossed by the SAR sensor. Thus, from the azimuth profile shown in Figure 20, two compressed pulses can be observed. The first compressed pulse is Target A, while the second compressed pulse is Targets B and C. The amplitude of the second pulse is slightly higher than the first one because both Target B and C have the same azimuth time.

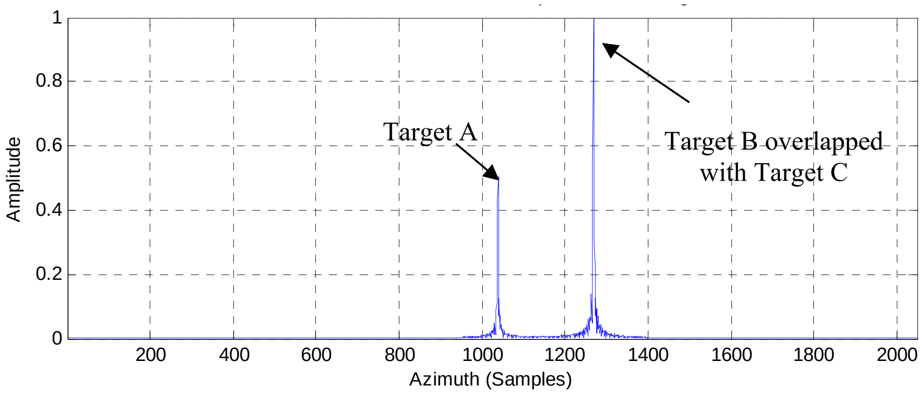


Figure 20: Azimuth profile results of azimuth compression.

3.2.5 Compressed Complex SAR Image

The processed SAR data after the azimuth compression processing step is also known as compressed complex SAR image. The compressed data is often store in complex form format while the image is commonly referred to as a single-look complex product. Figure 21 illustrates the zoomed in version of the azimuth compression result. It is shown that the three targets are compressed and fulfil to the scene as in Figure 10 for the positions of all point targets.

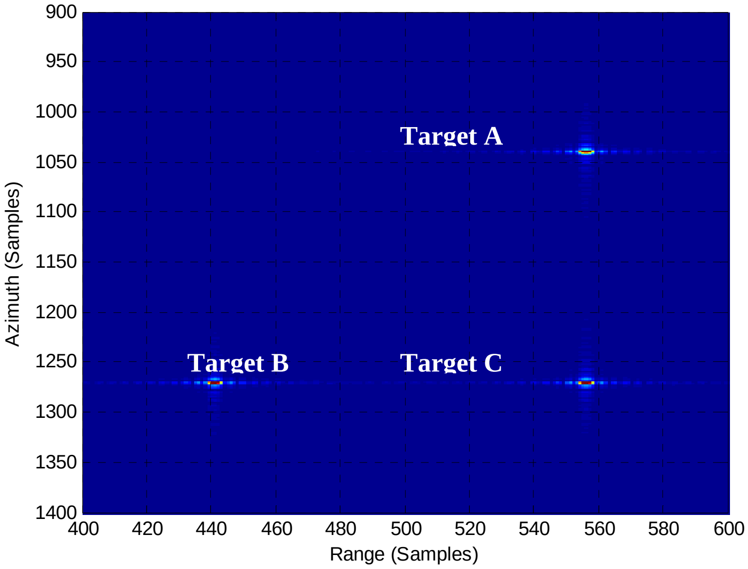
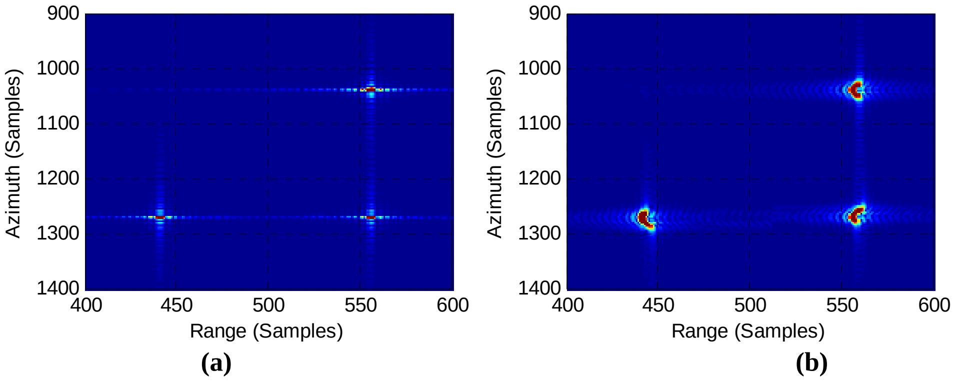


Figure 21: Compressed complex SAR image of azimuth compression.

The effect of RCMC also can be seen significantly in the compressed SAR image. Figure 22 shows the side by side comparison of the compressed SAR images. Figure 22(a) is the compressed SAR image with the application of RCMC, while Figure 22(b) without the RCMC process. From the later figure, it is noticeable that if the RCM is greater that the resolution of the range, and no RCMC is performed, the final compressed SAR image will be affected. The energy from the point targets are spread out, causing the compressed SAR image to look unfocused. As a result, the point targets look like curvature "c" shape spots instead of focused spots as in Figure 22(a).



(a)

(b)

Figure 22: Compressed Complex SAR images (a) with and (b) without RCMC.

4. CONCLUSION

In this paper, the properties of the SAR transmitted and received signals are presented together with their mathematical models. The implementation of RDA is discussed and explained in detail, which involves range compression, RCMC and azimuth compression. Eventually, the compressed complex SAR images are simulated together with the effect of RCMC in the final output of the RDA.

REFERENCES

Bamler, R. (1992). A comparison of Range-Dopper and Wavenumber Domain SAR Focusing Algorithms. *IEEE Trans. Geosci. Remote Sens.*, **30** (4): 706-713.

Bennett, J.R., Cumming, I. & Deane, R. (1980). The digital processing of SEASAT synthetic radar data. *Proc. IEEE Int. Radar Conf. 1980*, pp. 168-175.

Bovenga, F. (2020). Special issue: Synthetic aperture radar (SAR) techniques and applications. *Sensors*, **20**: 1851.

Brown, W.M. & Porcello, L.J. (1969). An introduction to synthetic aperture radar. *IEEE Spectrum*, **6**: 52-62.

Cafforio, C., Prati., C. & Rocca., F. (1991). SAR focusing using seismic migration techniques. *IEEE Trans. Aerosp. Electron. Syst.*, **27**: 194-207.

Chan, Y.K. & Koo, V.C. (2008). An introduction to synthetic aperture radar (SAR). *Prog. Electromagn. Res. B*, **2**: 27-60.

Chang, C.Y., Jin, M. & Curlander, J.C. (1989). Squint mode processing algorithm. *Proc. IGARSS'89*, Vancouver, pp. 1702-1706.

Cumming, I.G. & Bennett, J.R. (1979). Digital processing of SEASAT SAR data. *Proc. IEEE Int. Radar Conf.*, pp. 45 -47.

Cumming, I.G. & Wong, F.H. (2005). *Digital Processing of Synthetic Aperture Radar Data*. Artech House, Norwood, Massachusetts.

Curlander, J.C. & McDounough, R.N. (1991) *Synthetic Aperture Radar, Systems and Signal Processing*. John Wiley & Sons, New York.

Cutrona, L.J., Leith, E.N., Porcello, L.J. & Vivian, W.E. (1966). On the application of coherent optical processing techniques to synthetic aperture radar. *Proc. IEEE*, **54**: 1026-1032.

Davidson, G.W., Cumming, I.G. & Ito, M.R. (1996). A chirp scaling approach for processing squint mode SAR data. *IEEE Trans. Aerosp. Electron. Syst.*, **32**: 121-133.

Jin., M.Y. & Wu, C., (1984). A SAR correlation algorithm which accommodates large range migration. *IEEE Trans. Geosci. Remote Sens.*, **22**: 592-597.

John, C. & Kirk, J.R. (1975). A discussion of digital processing in synthetic aperture radar. *IEEE Trans. Aerosp. Electron. Syst.*, **11**: 326-337.

Li, A. & Lofeld, O. (1991). Two-dimensional SAR processing in the frequency domain. *Proceedings of IGARSS'91*, pp. 1065-1068.

Long, T., Zeng, T., Hu, C., Dong, X., Chen, L., Liu, Q. & Wang, Y. (2019). High resolution radar real-time signal and information processing. *China Commun.*, **16**: 105-133.

Moreira, A. (1992). Real-time synthetic aperture radar (SAR) processing with a new subaperture approach. *IEEE Trans. Geosci. Remote Sens.*, **30**: 714-722.

Moreira, A. & Y. Huang. (1994). Airborne SAR processing of highly squinted SAR data using a chirp scaling approach with integrated motion compensation. *IEEE Trans. Geosci. Remote Sensing*, **32**: 1029-1040.

Moreira, A., Mitermayer, J. & Scheiber, R. (1996). Extended Chirp scaling algorithm for air and spaceborne SAR data processing in Stripmap and ScanSAR imaging modes. *IEEE Trans. Geosci. Remote Sens.*, **34**: 1123-1136.

Prati, C. & Rocca, E. (1992). Focusing SAR data with time-varying Doppler centroid. *IEEE Trans. Geosci. Remote Sens.*, **30**: 550-558.

Raney, R.K. (1992). An exact wide field digital imaging algorithm. *Int. J. Remote Sensing*, **13**: 991-998.

Raney, R.K., Runge, H., Bamler, R., Cumming, I.G. & Wong, F.H. (1994). Precision SAR processing using chirp scaling. *IEEE Trans. Geosci. Remote Sens.*, **32**: 786-799.

Rawson, R. & Smith, F. (1974). Four channel simultaneous X-L band imaging SAR radar. *Int. Symp. Remote Sens. Environ.*, pp. 251-270.

Scheuer, T.E. & Wong, F.H. (1991). Comparison of SAR processor based on a wave equation formulation. *Proc. IGARSS'91*, pp. 635-639.

Smith, A.M. (1991). A new approach to range Doppler SAR processing. *Int. J. Remote Sens.*, **12**: 235-251.

Sommer, A. & Ostermann, J. (2019). Backprojection subimage autofocus of moving ships for synthetic aperture radar. *IEEE Trans. Geosci. Remote Sens.*, **57**: 8383-8393.

Vant, M.R., Hasiam, G.E. & Royer, G.M. (1978). A digital signal processing approach for satellite – borne synthetic aperture radar (SAR*). Proc. Int. Conf. Radar*, pp. 251-256.

Vant, W.R., Herring, R.W. & Shaw, E. (1979). Digital processing techniques for satellite-borne synthetic-aperture radars. *Can. J. Remote Sens.*, **5**: 67 -73.

Wu, C. (1980). A digital fast correlation approach to produce SEASAT SAR imagery. *IEEE Int. Radar Conf.*, pp. 153-160.

Wu, C., Liu, K.Y. & Jin, M. (1982). Modelling and a correlation algorithm for space borne sar signals. *IEEE Trans. Aerosp. Electron. Syst.*, **18**: 563-574.

Zhu, Z., Zhang, H. & Xu, F. (2020). Raw signal simulation of synthetic aperture radar altimeter over complex terrain surfaces. *Radio Sci*, **55** : e2019RS006948.

ATMOSPHERIC ATTENUATION AND ENVIRONMENTAL INFLUENCE IN AERONAUTICAL MOBILE SATELLITE COMMUNICATIONS (MSC)

Dimov Stojce Ilcev

Space Science Center (SSC), Durban University of Technology (DUT), Durban, South Africa

E-mail: ilcev@dut.ac.za

ABSTRACT

This paper describes the effects of atmospheric attenuation and environmental influence as very important particulars for aeronautical Mobile Satellite Communications (MSC), because such factors generally tend to impair the performance of satellite links, although signal enhancements are also occasionally observed. As a radio frequency (RF) signal radiates through an Earth-to-sky communication link, its quality degrades as it propagates through the satellite link caused by the atmospheric attenuation, special propagation effects, environmental influence and many other interference considerations in space. This degradation significantly affects satellite transmission links, particularly the extent of degradation depends on the link, atmosphere, transmitted signal and receiver antenna parameters. The specific effects of clear sky, transionospheric propagation and path depolarization causes are examined and explained, focusing on important propagation characteristics in aeronautical MSC, reflection, fading, interference from adjacent satellite systems and specific local environmental influence caused by aircraft onboard superstructures.

Keywords: *Clear-sky attenuation; wave-front incoherence; scintillation and multipath influence; Faraday rotation; ionospheric scintillation; depolarization and polarization.*

1. INTRODUCTION

In aeronautical transport strategy and transmission systems, stable propagation of radio signals and determination of atmospheric attenuation and environmental influence are necessary to achieve safe and reliable aeronautical mobile satellite communication networks. In fact, accurate propagation information is required to support the design, implementation and operation of most modern Mobile Satellite Communications (MSC) systems and in particular Aeronautical MSC (AMSC) networks. Thus, the propagation behavior of radio waves, in the ionosphere and troposphere, near the Earth's surface, or upon reflection from the surface, is of concern to telecommunication system designers intending to use an atmospheric propagation medium for the transmission of electromagnetic energy between antennas in the system. Signal degradations that occur with sufficient frequency and intensity to affect the performance and availability objectives must be estimated and accounted for in the link budget as part of the system design. Methods are thus required to predict the magnitude and occurrence of relevant propagation impairments with sufficient accuracy for engineering applications (Ilcev, 2013, 2018).

The general objectives of this paper are to introduce the various types of impairments to radio wave propagation concerning the effects of atmospheric attenuation and environmental influence as very important particulars for design and development of an aeronautical MSC model. The main scope of the study is to provide an understanding of the basic propagation mechanisms in relations to the particular propagation characteristics of the ionosphere for development of predictive models for the quantitative evaluation of propagation effects in the bands allocated for the aeronautical MSC networks. Thus, special research and measurements programs have been conducted for over a decade in the study of radio wave propagation on ground-to-aircraft and opposite paths, which are the specific effects of clear-sky effect, transionospheric propagation, reflection, fading, interference from adjacent satellite systems and specific local environmental influence (Ilcev, 2013, 2018).

Radio propagation is the behaviour of radio signals as they are propagated by radio transmitters into various part of atmosphere via transmitting antenna, from one fixed or mobile point to another. Radio signals traveling through the atmosphere layers suffer attenuation even during fine weather without clouds. Thus, the clear-sky attenuation is mainly the result of absorption of energy from the transmission by water vapor and oxygen molecules. However there are other modes of clear-sky affect that have influences on propagation, such as follows (ITU, 1996):

1. Defocusing and Wave-Front Incoherence Contribution – Several expressions have already been evaluated and are provided in No 2.3.2 of the International Communication Union (ITU)-R P.618 recommendation to estimate defocusing (beam-spreading) losses on paths at very low elevation angles. The loss is implicitly accounted for in the prediction methods for low-angle fading found in articles No 2.4.2 and 2.4.3 of the recommendation. Hence, small-scale irregularities of the refractive index structure of the atmosphere cause incoherence in the wave front at the receiving antenna. In any case, this will result in both rapid radio signal fluctuations and an antenna-to-medium coupling loss that can be described as a decrease of the antenna gain. In practice, signal loss due to wave-front incoherence is probably only significant for large-aperture antennas, high frequencies and elevation angles below 5°. Measurements made in Japan with 22 m antenna suggest that at 5° elevation angle, the loss is about 0.2 to 0.4 dB at 6 / 4 GHz, while measurements with a 7 m antenna at 15.5 and 31.6 GHz gave losses of 0.3 and 0.6 dB respectively at 5^{o} elevation angle.

2. Scintillation and Multipath Influence – Small-scale irregularities in the atmospheric refractive index cause rapid amplitude variations. In such a way, different tropospheric effects in the absence of precipitation are unlikely to produce serious fading in radio and space telecommunication systems operating at frequencies below 10 GHz and at elevation angles above 10°. In addition, at low elevation angles and at frequencies above 10 GHz, tropospheric scintillations can, on occasion, cause serious degradations in performance. Atmosphere scintillation measure models that include frequency, elevation angle and antenna diameter, as well as including meteorological parameters, can be used to account for regional and seasonal dependencies.

3. Propagation Delays – Additional propagation delays superimposed on the delay due to free space propagation are produced by refraction through the troposphere precipitation and the ionosphere. Therefore, at frequency above 10 GHz, the ionospheric time delay is generally less than that for the troposphere.

4. Angle of Arrival Values – The gradient of the refractive index of the atmosphere causes bending of the radio ray, with the angle of arrival varying from that calculated based on the geometry of the path. Since the relative index varies largely with altitude, the angle-of-arrival variation is much greater in the elevation than in the azimuth angle. In addition, turbulent irregularities of the refractive index can give rise to angle-of-arrival scintillations. Both of these effects decrease markedly with elevation angle and are generally insignificant for elevation angles above 10^{o}. The effects are independent of frequency (Davies, 1965; Collin, 1985; ITU, 1996; Ilcev, 2016).

2. TRANSIONOSPHERIC PROPAGATION

Radio waves at frequencies of VHF and above are capable of penetrating the ionosphere and therefore, they provide transionospheric telecommunications. The ionosphere consists of a layer somewhere between 80 and 150 km altitude, where the density of the atmosphere is very low. Radiation from the Sun ionizes some molecules and it takes a long time for them to be neutralized by other ions. The concentration of ions varies with height, time of day, season and the part of the 11-year sunspot cycle that the Sun happens to be. Transionospheric radio wave propagation occurs when a radio wave travels from the surface of the Earth into space and vice versa. The ionospheric plasma is an anisotropic, depressive medium that also highly nonlinear. Waves can be scattered into other modes (at different frequencies) by density irregularities, for instance, and the presence of the wave itself can in turn affect the propagation of the wave (a principle known as "self-action") (ITU, 1996).

Many types of irregularities occur at the *E* and *F*-region altitudes of the ionosphere. However, the current research aims to experimentally quantify the nonlinearities that occur in the *D*-region ionosphere, which is responsible for the vast majority of radio wave attenuation. Understanding the physical processes that determine the level of attenuation is key to predicting radio wave signal intensities in space (for ground-based transmissions). For instance, without considering nonlinearities, a high-power 30 MHz radio wave might be attenuated by 20 dB as it traverses the *D*-region ionosphere. Thermal nonlinearities (produced by the radio wave heating the plasma) could produce an additional 5 dB of attenuation. In addition, other nonlinearities have yet to be fully investigated in order to experimentally quantify their values (Ilcev, 2016).

The results of study show that the ionosphere can provide a duct for radio waves, but most radio waves above the frequency of 50 MHz will penetrate the ionosphere and escape. Consequently, it is these higher frequencies that are used for satellite communications, radio astronomy and satellite navigation. However, as the frequency increases, the effect of the ionosphere decreases, it can have a significant impact on the operation of Earth-to-space systems. In particular, irregularity in the ionosphere can cause fluctuations in signals (scintillation) that can severely degrade the operations of such systems (Ilcev, 2002a, 2016, 2018).

2.1 Faraday Rotation and Group Delay

Ionospheric effects are significant for frequencies up to about 10 GHz, and are particularly important for GEO and non-GEO satellite constellations operating below 3 GHz. At the frequencies used for satellite transmission, signals pass right through and are subject to negligible refraction, less than 0.01° at 30° elevation. The total electron content (TEC) accumulated through the transionospheric transmission path results in the rotation of the linear polarization of the signal carrier and a time delay in addition to the anticipated propagation path delay. This delay is known as the group delay, while the rotation of the linear polarization of the carrier is known as Faraday rotation. Given knowledge of the TEC, Faraday rotation and group delay can be estimated for communication applications. The TEC (N_T) can be evaluated by the following formula:

$$N_T = \int_s N_e(s)\, ds \quad [\text{electrons/m}^2] \tag{1}$$

where N_e = electron density [electrons/m^2] and s = propagation path length through the ionosphere [m]. Typically, $\mathrm{N_T}$ varies from 1 to 200 TEC units (1 TEC unit = 10^{16} el/m^2). Thus, N_T has typical values in the range of 10^{16} and 10^{18} el/m^2. Even when the precise propagation path is known, the elevation of N_T is difficult to determine because N_e is highly variable in space and time.

When propagating through the ionosphere, a linearly polarized wave will suffer a gradual rotation of its plane of polarization due to the presence of the geomagnetic field and the anisotropy of the plasma medium. Namely, this trend slows down the signal because the Earth's magnetic field penetrates the ionosphere when ions (charged particles), subject to the alternating electric field of a signal, tend to gyrate around the local line of force. The magnitude of Faraday rotation will depend on the frequency of the radio waves, geomagnetic field strength and electronic density (concentration) of the plasma, such as given in the following relation:

$$\Phi = N_T\,(KM/f^2) = 2.36 \text{ x } 10^2\, B_E N_T f^2 \quad [\text{radians}] \tag{2}$$

where K = 2.36 x 10^4 [meter-kilogram-second (MKS) for units]; M = (B_E secϕ) at 420 km of height; B_E = longitudinal component of the Earth's magnetic induction along the ray path [Tesla]; ϕ = zenith angle of the ray; and f = frequency [Hz]. Typical values of Φ as a function of frequency for representative TEC values are shown in Figure 1 (Ilcev, 2016, 2017).

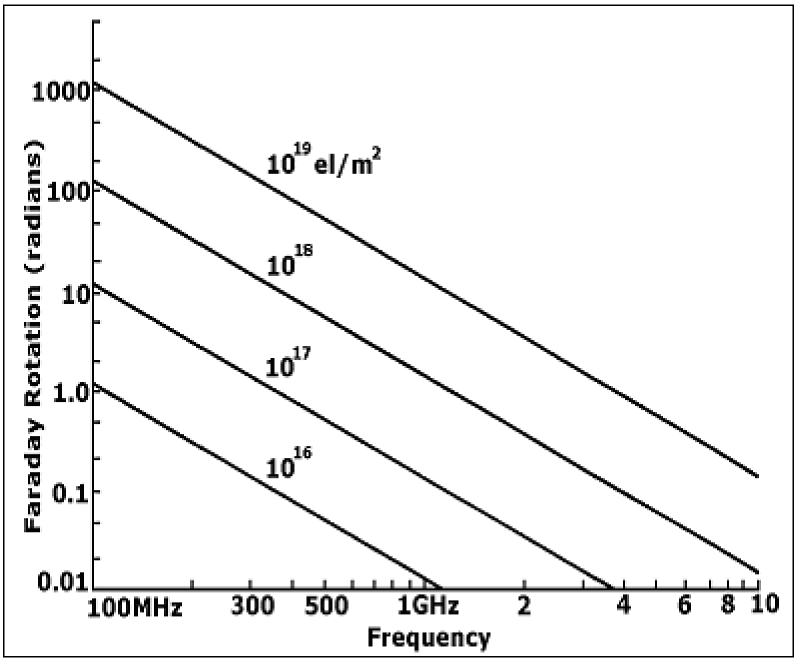


Figure 1: Faraday rotation as a function of TEC and RF (Source: ITU, 1996).

Hence, the occurrence of Faraday rotation is well understood, and can be predicted with a high degree of accuracy and compensated for by adjusting the polarization tilt angle at the GES terminal. Global Positioning System (GPS) and similar satellite navigation systems, which use the 1 to 2 GHz frequency spectrum and depend on measuring the travel time of EM signals, has to correct for this effect. The presence of charged particles in the ionosphere slows down the propagation of radio signals along the path and produces a phase advance. In such a way, the time delay in excess of the propagation time in free space is called the group delay (T_g) and is given by the following relation ITU, 1996:

$$T_g = 1.34 \text{ x } 10^{-7} N_T/f^2 \text{ [sec]} \qquad (3)$$

Thus, the time delay with reference to radio wave propagation in vacuum is an important factor to be considered for satellite communication, tracking and navigation systems (ITU, 1996; Ilcev, 2016).

2.2 Ionospheric Scintillation

Ionospheric effects are important at frequencies below 1 GHz, although they may even be important at frequencies above 1 GHz and are dependent on location, season, solar activity (sunspots) and local time. At this point, ionospheric scintillation occurs as short-term, rapid signal fluctuations and is mainly caused by irregularities in the ionosphere ranging from altitudes of 200 to 600 km. In fact, the frequency-dependence depends on the ionospheric conditions, but the attenuation varies approximately at the same rate as the square of the wavelengths. The effect is greater for lower frequencies and at lower latitudes, while high latitude areas near the Arctic polar region bounded between ±20° are susceptible to intense scintillation activity. In the L- and S-band, this effect can be ignored at medium latitudes except during periods of solar activity. When the Sun is very active, L-band enhancement and fading of 6 and –36 dB respectively are observed even at 37° latitude. Scintillation activity is at the maximum during the night, lasting from 30 min to a number of hours (ITU, 1996).

2.3 Other Ionospheric Effects

Other ionospheric effects include (ITU, 1996):

1. Dispersion – When transionospheric radio signals occupy a significant bandwidth of the propagation delay, being a function of frequency, it introduces dispersion. The differential delay across the bandwidth is proportional to the integrated electron density along the ray path. Hence, for an integrated electron content of 5 x 10^{17} el/m^2, a signal with a pulse length of 1 μs will sustain a differential delay of 0.02 μs at 200 MHz, while at 600 MHz, the delay would be only 0.00074 μs.

2. Refraction – When radio waves propagate obliquely through the ionospheric layer, they undergo refraction, which produces a change in the direction of arrival of the ray.

3. Absorption – For equatorial and mid-latitude regions, radio waves of frequencies above 70 MHz will assure penetration of the ionosphere without significant absorption, while for frequencies below 70 MHz, the ionospheric absorption loss is significant.

4. Doppler Frequency Shift – This is the special effect of frequency change due to the temporal variability of the ionospheric layer upon the apparent frequency of the carrier, which is the Doppler shifted carrier. For example, at $f = 1.6$ GHz, (GPS system), the observed frequency change Δf at high latitude is: $\Delta f/f < 10^{-9}$ (Del Re *et al*., 2008; Ilcev, 2013).

2.4 Sky Noise Temperature Contributions

The mechanism that causes absorption of energy from a wave passing between space and the Earth also causes the emission of thermal noise at RF. Some radio noise is added to the emission reaching the receiver, whereas the Earth itself radiates noise, which can enter the transmission path via satellite or the GES receiving antennas. Sources of radio noise of interest on Earth-to-space paths are the atmosphere, clouds, rain, extraterrestrial sources and noise from the surface of the Earth. Prediction methods are given in the ITU-R P.372 Recommendation. The thermal noise power N, available from a blackbody having a noise temperature of the source (T [K]), measured in bandwidth (B [Hz]), is given by:

$$N = kTB \quad [\mathrm{W}] \tag{4}$$

where k = Boltzmann's Constant. The special power density N_0 of noise from source is:

$$N_0 = N/B = kT \quad [\mathrm{WHz}^{-1}] \tag{5}$$

Thus, in considering the level of noise received at a GES or satellite from sources external to the environment, it is convenient to identify the brightness temperature (T_B) for each separate source and a coefficient (η), which represents the efficiency, with which the receiving antenna captures noise from that source. Then, the noise temperature component due to the identified source is given by:

$$t = \eta T_B \quad [\mathrm{K}] \tag{6}$$

Thus, the total noise entering the system from all of these sources, expressed as a noise temperature, can be obtained by summing all the component noise temperatures (ITU, 2002a; Ilcev, 2005, 2016).

2.5 Atmospheric Noise Temperature Elements

The noise temperature of a satellite-based antenna is dominated by the high temperature emitted by the Earth, which fills, or mostly fills, the main beam of the antenna array. Additional noise from precipitation or other variables is insignificant in this case. In fact, for a global beam, the noise temperatures are dependent both on the frequency and position of the satellite with relation to the major landmasses of the Earth.

The ground-based antenna observes the relatively cool sky and therefore, the presence of clouds and rain can significantly raise the noise temperature of the antenna. In general, the brightness temperature of the atmosphere due to permanent gases and rain is given as (ITU, 1996):

$$T_B = T_m\,(1 - 10^{-A/10}) \quad [\mathrm{K}] \tag{7}$$

where T_m = effective temperature of the attenuating medium (atmosphere, clouds, rain), which is about 270 K, and A = total attenuation due to the medium. The effect of rain on the satellite downlink is not just attenuation but also decrease in carrier-to-noise ratio (C/N) due to the higher noise temperature seen in rainy conditions as compared to clear sky conditions. In some cases, the noise temperature increase can provide more effect on the link than the attenuation itself (ITU, 1996; Ilcev, 2005).

3. PATH DEPOLARIZATION CAUSES

The atmosphere behaves as an anisotropic medium for radio propagation. Consequently, power from one polarization is coupled to its orthogonal component, causing interference between the channels of a dual polarized system. In this sense, depolarization or cross-polarization may occur when EM waves that propagate through media are anisotropic, namely asymmetrical with respect to the incident of polarization.

Meanwhile, depolarization in the form of Faraday rotation of the plane of linear polarization occurs in the ionosphere because of presence of the Earth's magnetic fields. At this point, the resulting impairments are typically circumvented by using circular polarization at frequencies below 10 GHz, for which the effect can be significant. Depolarization is often the most significant path impairment for 6 / 4 GHz satellite systems and can be the limiting performance factor for some 14 / 11 GHz satellite paths, especially at lower path elevation angles in moderate rain climates (Ilcev, 2016).

On the other hand, depolarization in precipitation is caused by differential attenuation and phase shifts that are induced between orthogonal components of an incident wave by anisotropic hydrometeors. Orthogonally polarized radio waves propagating in a medium that causes only differential phase shift are depolarized but maintain orthogonality. If the medium induces differential attenuation, the waves are also deorthogonalized (Novik, 1987; ITU, 1996; Ilcev, 2018).

3.1 Depolarization and Polarization Components

The importance of depolarization for satellite communications systems depends on a few components: frequency of the signal (f), geometry of path (θ = elevation angle and τ = tilt angle of the received polarization), local climatic factors (severity of the rain) and sensitivity to cross-polar interference (whether the system employs frequency reuse) (Ilcev, 2016).

The electromagnetic (EM) waves comprise both the electric and magnetic field vectors. Therefore, these two components travel in the direction of the transmission path and are orthogonal, while the orientation of the electric field vector defines the polarization of the transmitted waves. In general, as the wave progresses in time, the tip of the electric vector traces an ellipse in a plane perpendicular to the propagation direction. A representative polarization ellipse of representative elliptically polarized radio wave is displayed in Figure 2. Two important parameters in this figure are the axial ratio and inclination tilted angle with respect to the reference axis (τ). The polarization ellipse may be tilted at an angle τ with respect to the particular coordinate frame. Thus, the general form of a polarized wave, when viewed perpendicular to the direction of travel, is elliptical in shape. The polarization state of a wave is completely specified by its polarization ellipse, i.e., the amplitudes of the major axis (E_{max}), minor axis (E_{min}) and the sense of rotation of the vector also defines the axial ratio using the following expression (ITU, 1996):

$$A_R = 20\log (E_{max}/ E_{min}) \quad [\text{dB}] \qquad (8)$$

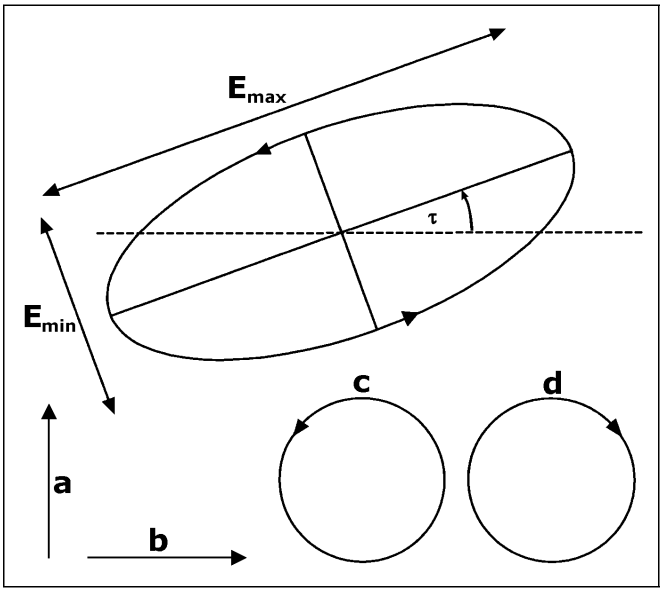


Figure 2: Generalized elliptical waveform (Source: Sheriff, 2001).

In MSC, four types of polarization are employed, shown in Figure 2 (a) vertical linear polarization (VLP); (b) horizontal linear polarization (HLP); (c) left-hand circular polarization (LHCP) and (d) right-hand circular polarization (RHCP). The direction of the travel is symbolical "into the paper". Horizontal and vertical polarizations are defined with respect to the horizon, RHCP has a clockwise rotation and RHCP has an anticlockwise rotation when viewed from the antenna in the direction of travel. If E_{max} and E_{min} are equal in magnitude, the polarization state can be RHCP or LHCP, depending on the sense of rotation and in the case where E_{max} is nonzero and E_{min} is zero, the value of electric vector maintains a constant orientation defined by E_{max} and the polarization state is said to be linear (Sheriff, 2001).

The polarization quantity of interest for frequency re-use communications is the cross polarization isolation (XPI), defined as the decibel ratio of the (desired) co-polar power received in a channel to the (undesired) cross-polar power received in that same channel. However, in practice, XPI is difficult to measure because the cross-polarized components cannot be distinguished from noise in the co-polar channel. The quantity usually measured is the cross polarization discrimination (XPD), defined as the ratio of the co-polarized power received in one channel to the cross-polarized power detected in the orthogonal channel, both arising from the same transmitted signal. The theory predicts that XPD and XPI components are equivalent for most practical situations. The polarized wave will comprise the wanted polarization together with some energy transmitted on the orthogonal polarization. The degree of the coupling of energy between polarizations is determined by relation (a), while for the coexisting circular polarized waves (b), XPD can be determined from the axial ratio are given by (ITU, 1996):

$$\text{(a) XPD} = 20\log |E_{cpr}/E_{xpr}| \quad [\text{dB}] \quad \text{and} \quad \text{(b) XPD} = 20\log (A_R + 1/A_R - 1) \quad [\text{dB}] \qquad (9)$$

where E_{cpr} = received co-polarized electric field strength and E_{xpr} = received cross-polarized electric field strength (ITU, 1996; Sheriff, 2001; Ilcev, 2016).

3.2 Relation Between Depolarization and Attenuation

ITU provides a step-by-step method for the calculation of hydrometer-induced cross- polarization, which is valid for frequencies within the range 8 to 35 GHz and for elevation angles less than 60°. In addition, the attenuation due to rain exceeding the required percentage of time (p) and polarization tilt angle τ with respect to the horizontal also need to be known. XPD due to rain is given by:

$$\mathrm{XPD}_{\mathrm{rain}} = C_f - C_A + C_\tau + C_\theta + C_\sigma \quad [\mathrm{dB}] \tag{10}$$

where C_f = frequency-dependent term or 30 logf [GHz]; C_A= rain-dependent term or $V(f)$ logA_p; C_τ = polarization improvement factor or –10log [1–0.484 (1+cos4τ)]; C_θ = elevation angle-dependent term or –10log (cosθ) and C_σ = canting angle term or 0.0052σ. In the above, the canting angle refers to the angle at which a falling raindrop arrives at the Earth with respect to the local horizon. The terms τ, θ and σ are expressed in degrees, while σ has values of 0, 5, 10 and 15° for 1, 0.1, 0.01 and 0.001% of time respectively. Taking ice into account, XPD not exceeding p% of time is given by:

$$\mathrm{XPD_p} = \mathrm{XPD}_{\mathrm{rain}} - C_{ice} \quad [\mathrm{dB}] \tag{11}$$

where C_{ice} = ice depolarization or $\mathrm{XPD}_{\mathrm{rain}}$ x (0.3+0.1 log p)/2 (ITU, 2006; Stacey, 2008; Ilcev, 2016).

4. PROPAGATION EFFECTS IMPORTANT FOR MSC SYSTEMS

A Mobile Earth Station (MES) terminal operates in a dynamic and often unsuitable environment in which propagation conditions are constantly changing. Namely, satellite transmission path profile in MSC varies continuously while the mobiles are in motion. The MES uses relatively broad beam antenna systems, which have only a limited discrimination against signals reflected from scattered objects and surfaces. Due to the inherent random nature of disturbances, radio signals are usually characterized statistically. In general, a signal arriving at the antenna of a MES consists in the vector sum of a direct component and diffused components arising from multipath reflection. The resultant effects of these additional considerations are (Ilcev, 2016):

1. Signals suffer attenuation whenever the satellite path to mobiles is shadowed.
2. Signals fluctuate randomly because reflected and scattered random signal components arriving at the mobile antenna are picked up, which is known as multipath.
3. The power spectral density of multipath radio noise is a function of the mobile's speed and the environmental conditions.

Depending on the environment in which a mobile terminal operates, the satellite channels in the MSC may be categorized as maritime, land and aeronautically based. Each category has its typical channel characteristics because of the different propagation environments. The three different categories of MES, which are Ship Earth Station (SES), Vehicle Earth Station (VES) and Aircraft Earth Station (AES), each have their own distinctive channels that need to be considered separately. Thus, the local operation variable environment has a significant impact on the achievable quality of transmission service in MSC. On the contrary, the Ground Earth Station (GES) or gateway terminals serving in a mobile satellite network can be optimally located in fixed positions with constant channel characteristics to guarantee good visibility to the satellite at all times, eliminating or reducing the negative and undesirable effect of the local environment to a minimum (ITU 1996; Ilcev, 2016).

4.1 Propagation in Aeronautical MSC System

Aeronautical telecommunication requirements and propagation conditions are superior to those in maritime and land mobile applications because there are not many difficulties and obstacles between a satellite and aircraft. However, at low elevation angles and when a low-gain antenna is used, multipath fading caused by reflection from the sea or ground surface occurs, although it is less than in the maritime case.

Otherwise, in an aeronautical MSC an aircraft demodulator must track the received signal and remove the Doppler Effect due to the high speed of flight, using digital signal processing or using a pilot signal. Thus, the ability of communications to aircraft via satellite is becoming increasingly important. Due to the safety regulations of the International Civil Aviation Organization (ICAO), communication channels to an aircraft need to be specified to a high degree of reliability. A typical flight on board an aircraft comprises the following phases:

1. The aircraft taxies to a position on the runway, ready to take-off.
2. Take-off of aircraft and ascension to usual cruise altitude at a constant height (altitude) above the cloud layer.
3. Descent from cruise altitude to a landing on the runway.
4. Taxiing from the runway to a stand-by place in the airport for a new flight.

Each of the above phases can be considered to have particular channel characteristics. For example, while the aircraft is in the airport, a land MSC channel type environment could be subject to sporadic shadowing due to adjacent buildings, airport structures, and other aircraft and close obstacles. The aeronautical MSC channel is further complicated by the maneuvers performed by an aircraft during the course of a flight, which could result in the aircraft's structure blocking the line-of-sight to the satellite.

The body of the airplane is also a source of multipath reflection, which also needs to be considered. In fact, as was discussed earlier, the speed of an aircraft introduces large Doppler spreads as an additional effect. In fact, the effect of multipath reflections from the sea for circular polarized L-band transmissions will be explained later, together with the maritime case. Thus, ITU provides a methodology similar to that for the maritime link, for determining the multipath power resulting from specular reflection from the sea. In such a way, the ITU Recommendation includes a methodology to derive the mean multipath power as a function of elevation angle and mobile antenna gain (Sheriff, 2001).

In this sense, by applying this method for an aircraft tracking position 10 km above the sea surface and for a minimum elevation angle of 10^{o}, the relative multipath power will be in the range of approximately -10 to -17 dB, for antenna gains varying from 0 to 18 dBi, respectively (ITU, 1996).

Presently, AMSC systems are served by the L-band spectrum. Due to the bandwidth restrictions at this frequency spectrum, services are limited to voice and low data rate applications. Inmarsat recently improved speed of transmission with a new generation of satellite constellation. The need to provide broadband multimedia satellite services, akin to those envisaged by new satellite UMTS/IMT-2000 will require the move up in frequency to the next suitable bandwidth, the K and Ka-bands (Ilcev, 2005).

At these frequencies, tropospheric effects will have an impact on link availability during the time when the aircraft is below the cloud layer. At this point, the transmission channel characteristics for a K-band satellite communication system have been investigated in Europe and the USA. Thus, here a Rice-factor of 34 dB is reported for line-of-sight operation, while shadowing introduced by the aircraft's wing during a turning maneuver resulted in a fade of 15 dB (ITU 1996; Ilcev, 2013).

4.2 Surface Reflection and Local Environmental Effects

Surface reflections and local environmental effects are important for aeronautical and other MSC systems because such factors generally tend to impair the performance of satellite communications links, although signal enhancements are also occasionally observed. Local environmental effects include shadowing and blockage from objects and vegetation near the MES terminal. At this point, surface reflections are generated either in the immediate vicinity of the MES terminals or from distant reflectors, such as mountains and large industrial infrastructures. The reflected transmission signal can interfere with the direct signal from the satellite to produce unacceptable levels of signal degradation. In addition to fading, signal degradations can include inter symbol interference, arising from delayed replicas (Ilcev, 2013).

The impact of the impairments depends on the specific application. In the case of typical land MSC links, all measurements and theoretical analysis indicate that the specular reflection component is usually negligible for path elevation angles above 20^{o}. Moreover, for handheld terminals, specular reflections may be important as the low antenna directivity increases the potential for significant specular reflection effects. For MSC system links, design reflection multipath fading, in combination with possible shadowing and blockage of the direct signal from the satellite, is generally the dominant system impairment (Novik, 1987; ITU 1996; Ilcev, 2005).

4.3 Reflection from the Earth's Surface

Prediction of the propagation impairments caused by reflections from the Earth's surface and from different objects (buildings, hills, mountains, vegetation) on the surface is difficult because the possible impairment scenarios are quite numerous, complex and often cannot be easily quantified. For example, the degree of shadowing in land MSC satellite links frequently cannot be precisely specified. Therefore, impairment prediction models for some complicated situations, especially for land MSC links, tend to be primarily empirical, while more analytical models, such as those used to predict sea reflection fading, have restricted regions of applicability. Nevertheless, the basic features of surface reflections and the resultant effects on propagating signals can be understood in terms of the general theory of surface reflections, as summarized in the following classification (Ilcev, 2016):

1. **Specular Reflection from a Plane Earth** – Here, the specular reflection coefficient for vertical polarization is less than or equal to the coefficient for horizontal polarization. Thus, the polarization of the reflected waves will be different from the polarization of the incident wave if the incident polarization is not purely horizontal or purely vertical. For example, a circularly polarized incident wave becomes elliptically polarized after reflection.

2. **Specular Reflection from a Smooth Spherical Earth** – Here, the incident grazing angle is equal to the angle of reflection. The amplitude of the reflected signal is equal to the amplitude of the incident signal multiplied by the modules of the reflection coefficient.

3. **Divergence Factor** – When rays are specularly reflected from a spherical surface, there is an effective reduction in the reflection coefficient, which is actually a geometrical effect arising from the divergence of the rays.

4. **Reflection from Rough Surface –** In many practical cases, the surface of the Earth is not smooth. Namely, when the surface is rough, the reflected signal has two components: one is a specular component, which is coherent with the incident signal, while the other is a diffuse component, which fluctuates in amplitude and phase with a Rayleigh distribution.

5. **Total Reflected Field –** The total field above a reflecting surface is a result of the direct field, the coherent specular component and the random diffuse component.

6. **Reflection Multipath –** Owing to the existence of surface reflection phenomena signals may arrive at a receiver from multiple apparent sources. Thus, the combination of the direct signal (line-of-sight) with specular and diffusely reflected waves causes signal fading at the receiver. The resultant multipath fading, in combination with varying levels of shadowing and blockage of the line-of-sight components, can cause the received signal power to fade

severely and rapidly for MES and is really the dominant impairment in the MSC service (ITU 1996; Ilcev, 2016; Ghasemi *et al.*, 2016).

4.4 Fading in MSC Systems Due to Sea Surface Reflection

Multipath fading due to sea reflection is caused by interference between direct and reflected radio waves. The reflected radio waves are composed of coherent and incoherent components, namely specular and diffuse reflections respectively that fluctuate with time due to the motion of sea waves. The coherent component is predominant under calm sea conditions and at low elevation angles, whereas the incoherent component becomes significant in rough sea conditions. If the intensity of the coherent component and the variance of the incoherent component are both known, the cumulative time distribution of the signal intensity can be determined by statistical consideration.

In any event, a prediction model for multipath fading due to sea reflection was first developed for Maritime MSC (MMSC) systems at frequency near 1.5 GHz. Although the mechanism of sea reflection is common for MSC systems, only with the difference that fading characteristics for AMSC are expected to differ from those for MMSC. This is because the speed and altitude of aircrafts are so much greater than those of ships. At this point, the effects of refractions and scattering by the sea surface become quite severe in the case of MSC, particularly where antennas with wide beam widths are used. The most common parameter used to describe sea condition is the significant wave height (H), defined as the average value of the peak-to-trough heights of the highest one-third of all waves. Empirically, H is related to the root mean square (rms) height (h_o) by:

$$H = 4h_0 \tag{12}$$

Hence, at 1.5 GHz, the smaller-scale waves can be neglected and the rms value of the sea surface slopes appears to fall between 0.04 to 0.07 in the case of wave heights less than 4 m. Thus, with diminishing satellite elevation angle, the propagation path increases, causing a decrease of signal power at the receiver (R_x) side. The noise level is initially constant, but upon reaching some critical value of the elevation angle, sea-reflection signals appear at the R_x input, which begins to affect the C/N ratio.

In order to include the effect of multipath interference caused by sea-refracted signals, the reception quality would be more properly described by C/N plus M, where M is an interfering sea-reflected signal acting as a disturbance. Thus, sea-reflected signals differ in structure and can be divided into two categories:

1. Radio signals with rapid continuous fluctuations of amplitudes and phases, and with a possible frequency shift due to the motion of small portions of the specular cross-section relative to the source of signals (noise or diffused components).
2. Radio waves with relatively slowly changing phase close to the phase of the basic signal and with amplitude correlating with that of the basic signal (specular component).

Consequently, within the overall specular cross-section, an angle of arrival reflected radio signals relative to the horizontal plane may be regarded as constant and can be described by the following expression:

$$\alpha = 90^{\circ} - \gamma \tag{13}$$

where α = angle of radio signals arrival in accordance with Figure 3 and γ = reflection angle. The modulus of sea reflection factor for L-band signals is within 0.8 and 0.9, which means that the amplitude of the specular reflected signal is nearly the same as that of the direct signal. As measurements have shown, the noise component depends only upon an elevation angle and a wave height. Decreasing the elevation angle and increasing the wave height result in an increase in the total amplitude of the noise, which includes the noise component.

At elevation angles below 5^o, the amplitude component reaches a peak value and is no longer affected by the wave height. Now an increase of the wave height causes primarily more frequent variations in the noise component. The corresponding deviation of C/N measured in 1 kHz bandwidth amounts from 4.5 to 5 dB. The specular component that appears at the R_x input together with the direct signal causes fading in the direct signal due to both the minor difference between their phases and the slow change of the parameters of the reflected signals. The ratio of the direct to specular reflected signal can be described as:

$$C/M = (C + G_\varepsilon) - [C - G_{(\alpha\ +\ \varepsilon)}] \qquad (14)$$

where value C = direct signal; M = specular reflected signal power from the sea; G_ε = maximum gain of the receive aircraft antenna pointing towards the satellite; ε = elevation angle; and α = angle of radio signals, as is depicted in Figure 3. In addition, keeping accuracy sufficient for practical purposes, the previous relation is given the following equation:

$$C/M = \beta_{C/N} + [G_\varepsilon - G_{(\alpha\ +\ \varepsilon)}] \qquad (15)$$

where $\beta_{C/N}$ = deviation of C/N ratio. With decreasing elevation angle, C/M diminishes monotonically, except for the elevation angle range of 5 to 8°, within which a rise in C/N is observed (Ilcev, 2013).

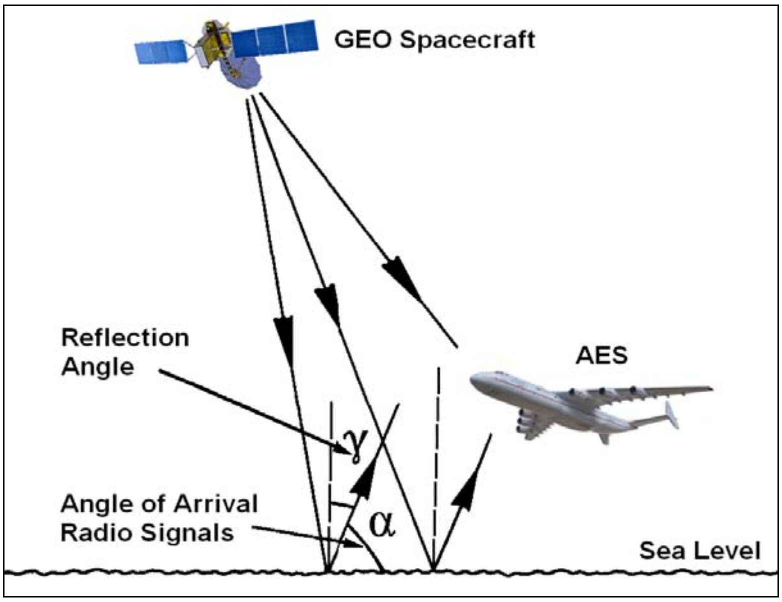


Figure 3: Geometry of sea reflection of satellite radio signals (Sources: Ilcev, 2018).

This is obviously due to the fact that at the said angles, the difference in path between the direct and specular signal becomes negligible, so conditions appear close to the summation of the similar signals at the receiver input. In such a way, an increase of the C/N value plus M ratio is observed simultaneously due to reaching a peak value of amplitude in the noise (diffused) component. In fact, experimental measurements show that as the elevation angle decreases from 10^o to 1^o, the mean C/N plus M diminishes from 22–24 dB to 17–18 dB, with the deviation increasing from 1.5–2 dB to 4.5 - 5.0 dB (Sheriff, 2001; ITU 2002b; Stacey, 2008; Ilcev, 2016, 2018).

4.5 Interference from Adjacent Satellite Systems

In MSC systems for ships, vehicles and aircraft installations, small mobile antennas are essential for operational and economic reasons. As a result, a number of low gain-to-noise temperature (G/T) value MES terminals with smaller MSC antennas have been developed. However, such satellite antenna systems are subject to the restriction of frequency utilization efficiency, or coexistence between two or more satellite systems in the same frequency band and / or an overlap area where both satellites are visible.

For mutual coordination between two different mobile or fixed satellite communication systems in the same frequency band, a highly reliable interference evaluation model covering both interfering and interfered with conditions is required. Investigation into this area has been undertaken in particular by ITU-R Study Group 8. The advancement of such a model is an urgent matter for the ITU-R considering the number of MSC systems that are being developed in for mobile communications.

In MSC systems, the desired signal from the satellite and interfering signal from an adjacent satellite independently experience amplitude fluctuations due to multipath fading, necessitating a different treatment from that for fixed satellite systems. The main technical requirement is a formulation for the statistics of differential fading, which is the difference between the amplitude of the two satellite signals (ITU, 1996).

At this point, the method given in No 5 of ITU-R P.680 Recommendation presents a practical prediction method for signal-to-interference ratio, where the effect of thermal noise and noise-like interference is taken into account, assuming that the amplitudes of both the desired and interference signal affected by the sea reflected multipath fading follow the Nakagami–Rice distributions. In fact, this fading situation is quite probable in the maritime environment (Ilcev, 2013).

The basic assumptions of the intersatellite model are shown in Figure 4, as an example of interference between adjacent satellite systems, where (Left) is downlink interference on the MES terminal side and (Right) is uplink interference on the satellite side. This applies to multiple systems sharing the same frequency band. It is anticipated that the interference causes an especially severe problem when the interfering satellite is at a low elevation angle viewed from the ship presented in this figure because the maximum level of interference signal suffered from multipath fading increases with decreasing elevation angle. Another situation is interference between beams in multi-spot-beam operation, where the same frequency is repeatedly allocated (ITU 1996; Sheriff, 2001; Ilcev, 2017).

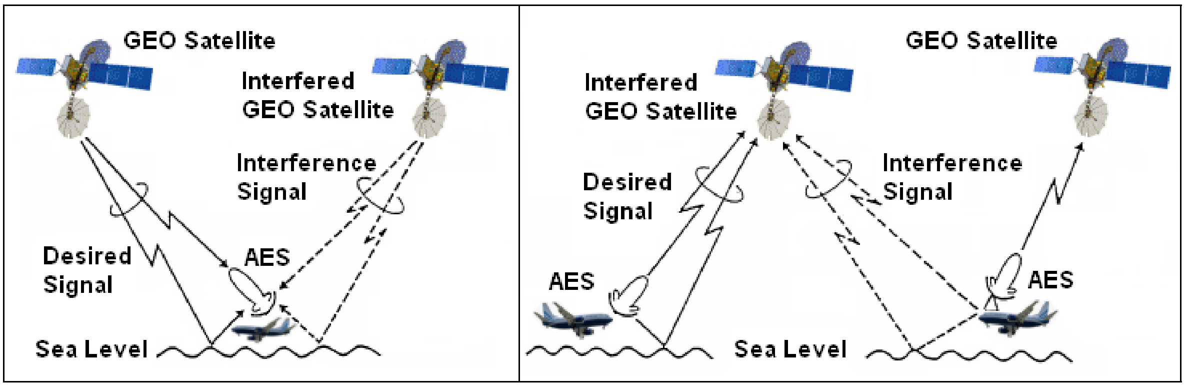


Figure 4: Basic model for intersatellite interferences phenomena (Sources: Ilcev, 2013, 2017).

5. LOCAL ENVIRONMENTAL INFLUENCE IN AERONAUTICAL MSC (AMCS) SYSTEMS

Local environmental influence is important for AES equipped with beam width antenna. In fact, many factors, with different kinds of noise sources, tend to make disturbances in AMSC channels. Another factor that affects communication links is RF emission from different noise sources in the local environment. Specific local aircraft environmental factors can be noise contributions from various sources in the vicinity of the AES and the influence of the aircraft's superstructure in the operation of aeronautical mobile terminals (Ilcev, 2013).

Some of these local environmental factors can affect AES when an aircraft is flying nearby the ground and, some of these are permanent noise sources. The environmental sources include broadband noise sources, such as electrical equipment and motor vehicles and out-of-band emission from powerful transmitters (T_x) such as onboard radars and aircraft HF transmitters (ITU 1996; Ilcev, 2016).

5.1 Noise Contribution of Local Aircraft' Environment

Some of the significant noise contributions from the local aircraft environment are as follows (Ilcev, 2013):

1. **Atmospheric Noise from Absorption** – Absorbing atmospheric media, such as water vapor, precipitation particles and oxygen emit thermal noise, that can be described in terms of antenna noise temperature. These effects were discussed at the beginning of this paper.

2. **Industrial Noise** – Heavy electrical equipment tends to generate broadband noise that can interfere with sensitive receivers. Therefore, a high percentage of this noise originates as broadband impulsive noise from ignition circuits. Namely, the noise varies in magnitude by as much as 20 dB, depending on whether it is measured on a normal working day or on weekends and holidays when it is lower in magnitude.

3. **Out of Band Emission from Radar** – Airborne and surveillance radars operating in pulse mode can generate out of band emissions that can interfere with aircraft receivers. However, such emissions can be suppressed by inserting waveguide or coaxial filters at the radar transmitter output.

4. **Interference from High Power Communication Transmitters** – High power aircraft radio and satellite transmitters, including powerful terrestrial transmitters, for example HF aircraft radio transceivers; HF radio diffusion and TV broadcasting can interfere with AES terminals.

5. **Interference from Vehicles** – Under certain operational conditions, RF emissions from different service vehicles in airport may impair R_x sensitivity. The noise measurement emanating from heavy traffic is about –150dB (mW/Hz) within RF bands of 1.535 to 1.660 MHz (ITU 1996; Ilcev, 2016).

5.2 Blockages Caused by Aircraft Superstructures

Aircraft superstructures can produce both reflection multipath and blockage in the direction of the satellite. For the most part, reflections from the aircraft's superstructure located on the deck can be considered coherent with the direct signal. The fading depth due to these reflections depends on a number of construction parameters including the shape of the aircraft; location of the aircraft's antenna; antenna directivity and sidelobes level; axial ratio and orientation of the polarization ellipse; and azimuth and elevation angles towards the satellite. Antenna gain has a significant influence on the fading depth. In this case, low gain antennas with broader beam widths will collect more of the reflected radio signals, producing deeper fades (Ilcev, 2013).

Blockage can be caused by aircraft superstructures, such as fuselage, wings, propellers, jet engines, tails and insignificant various types of antennas deployed on the aircraft. Signal attenuation depends on several parameters including diameter of wings, size of antenna, and distance between antenna and obstacles. Thus, estimated attenuations due to blocking by an aircraft's obstacle structures is shown in Figure 5 for antenna gains of 20 dB (Left) and 14 dB (Right) respectively (ITU 1996; Ilcev, 2005).

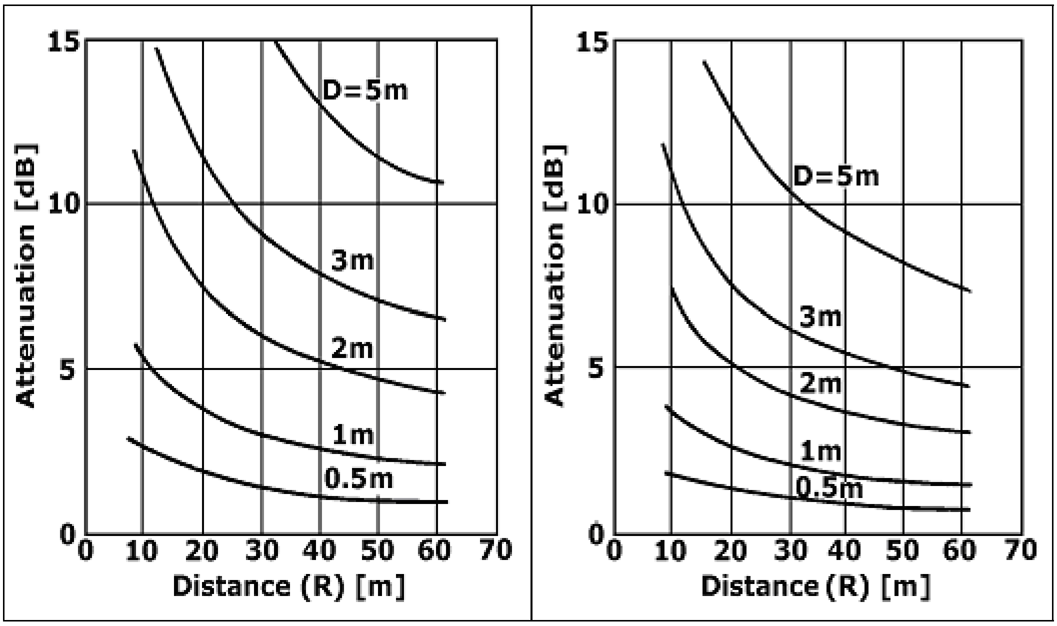


Figure 5: Estimated attenuation due to blocking (Source: ITU, 1996; Ilcev, 2013).

5.3 Blockage Motion of Aircraft's Antenna

The motion of mobile satellite antennas is an important consideration in the design of the AMSC systems. The received signal level is affected by the antenna off-beam gain because the antenna motion is influenced by the aircraft's motion. The random aircraft motion must be compensated by a suitable stabilizing mechanism to keep the antenna properly pointed towards the satellite. This is normally achieved either through a passive gravity stabilized platform or an active antenna tracking system. In either case, the residual antenna pointing error can be significant enough to warrant its inclusion in the overall link calculation.

Thus, earlier experimental evidence suggests that the roll motion of aircraft approximates a zero mean Gaussian distribution over the short-term of the sea waves. The standard deviation of the distribution (σ_s) is a function of the aircraft characteristics, and the state of the rolling and pitching. Figure 6 illustrates the distribution of the instantaneous roll angle of an aircraft under moderate to rough drift conditions during the flight. The solid curves in the figure represent measured values, while the dashed curves show the calculated values for the stabilized antenna motion over the rough drift conditions. Thus, the distribution of the aircraft approximates to a Gaussian standard deviation of distribution with $\sigma_s = 5.42$ value (Ilcev, 2018).

Also shown in Figure 6 is the distribution of roll angle of a passively stabilized antenna under the same conditions, which also follows a zero mean Gaussian distribution with a quantum of $\sigma_s = 0.99$. The solid curves represent measured values while the dashed curves show the calculated values for the stabilized antenna motion over rolling and pitching conditions of the aircraft. The standard deviation of the distribution is a function of the aircraft's drift characteristics in rolling and pitching conditions (Ilcev, 2013).

Otherwise, the relation between the standard deviations of the two distributions depends on the design of the passive stabilizer. Although the aircraft motion is much reduced, depending on the antenna beam width, the residual pointing error may be large enough to produce appreciable signal fluctuations. Over a long period of sea waves time, σ_s varies as a function of the sea surface conditions and its distribution can be approximated either as a log normal or Weibull distribution. The same context can be used to explain the distribution of the instantaneous roll angle of an aircraft under moderate to rough drift conditions in the air (ITU 1996; Ilcev, 2013, 2018).

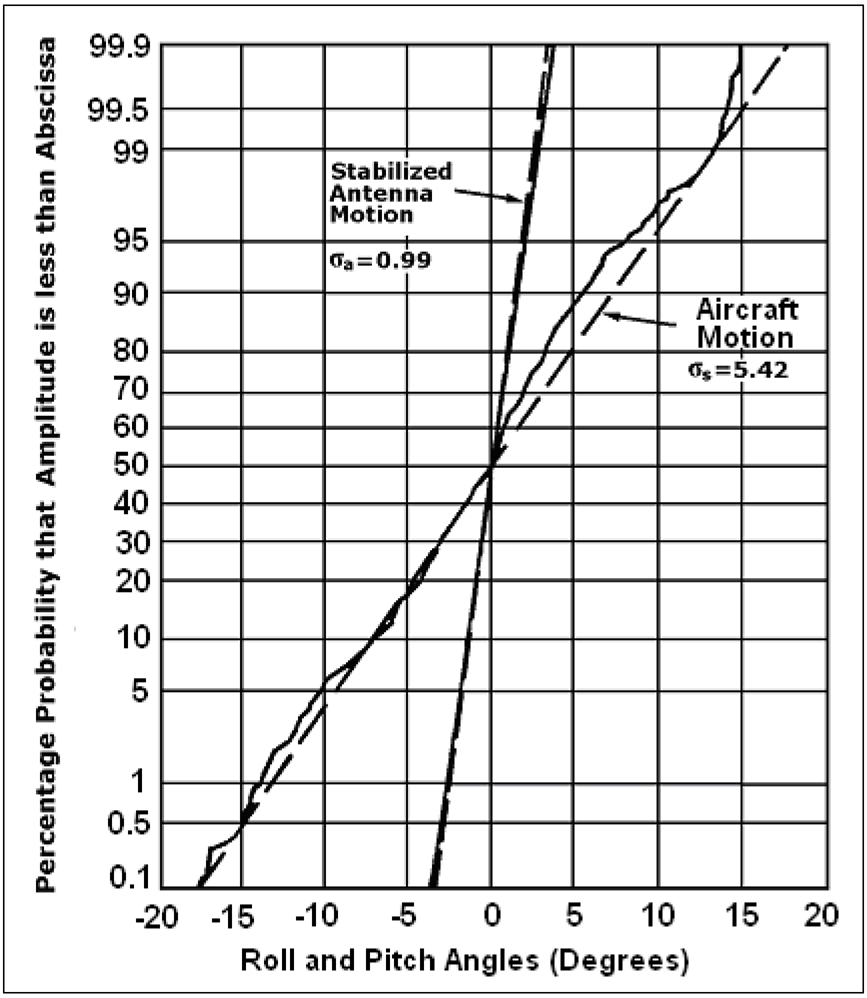


Figure 6: Measured stabilized antenna motion (Source: ITU, 1996).

6. CONCLUSION

The common satellite channel environment affects radio wave propagation in many ways. The different parameters influenced are mainly path attenuation, polarization and noise. The factors to be considered are gaseous absorption in the atmosphere, absorption and scattering by clouds, fog, all precipitation, atmospheric turbulence, and ionospheric effects. In this sense, several measurement techniques serve to quantify these effects in order to improve reliability in the system design. As these factors are random events, MSC system designers usually use a statistical process in modeling their effects on radio wave propagation. In order to design an effective MSC model, it is necessary to consider the quantum of all propagation characteristics, such as signal strength loss in normal environments, path depolarization causes, transionospheric contribution as well as propagation effects important for mobile systems, including reflection from the Earth's surface, fading due to sea and land reflection, signal blockage, and the different local environmental interferences for all mobile and handheld applications. At any rate, the local propagation characteristics on the determinate geographical position have very specific statistical proprieties and results for ships, vehicles and aircrafts.

REFERENCES

Collin, R.E. (1985). *Antenna and Radiowave Propagation*. McGraw-Hill, New York, US.

Davies, K. (1965). *Ionospheric Radio Propagation*. U.S. Department, Washington, DC, US.

Del Re, E. & Ruggieri, M. (2008). *Satellite Communications and Navigation Systems*. Springer, New York, US.

Freeman, R. L. (1987). *Radio Systems Design for Telecommunications (1-100 GHz)*. John Wiley, Chichester, UK.

Ghasemi, A. & Abedi, A. (2016). *Propagation Engineering in Wireless Communications.* Springer, Boston, US.

Ilcev D.S. (2005). *Global Mobile Satellite Communications for Maritime, Land and Aeronautical Applications*. Springer, Boston, US.

Ilcev, D.S. (2013). *Global Aeronautical Communications, Navigation and Surveillance (CNS) – Theory and Applications*. Volume 1 & 2, AIAA, Reston, Virginia, US.

Ilcev, D.S. (2017). *Global Mobile Communications, Navigation and Surveillance (CNS)*. Durban University of Technology (DUT), Durban, South Africa.

Ilcev D. S. (2016). *Global Mobile Satellite Communications for Maritime, Land and Aeronautical Applications*. Volume 1 & 2, Springer, Boston, US.

Ilcev, D.S. (2018). *Mobile Satellite Antenna and Propagation*. Durban University of Technology (DUT), Durban, South Africa.

ITU (1996). *Radiowave Propagation Information for Predictions for Earth-to-Space Path Communications*. International Telecommunication Union (ITU), Geneva, Switzerland.

ITU (2002a). *Handbook – Mobile Satellite Service (MSS)*. International Telecommunication Union (ITU), Geneva, Switzerland.

ITU (2002b). *Handbook on Satellite Communications*. International Telecommunication Union (ITU), Geneva, Switzerland.

Novik, L.I. (1987). *Sputnikovaya Svyaz Na More*. Sudostroenie, Leningrad, Russia.

Sheriff R.E. (2001). *Mobile Satellite Communication Networks*. Wiley, Chichester, UK.

Stacey D. (2008). *Aeronautical Radio Communication Systems and Networks*. Wiley, Chichester, England.

EVALUATION OF THE VULNERABILITIES OF UNMANNED AERIAL VEHICLES (UAVS) TO GLOBAL POSITIONING SYSTEM (GPS) JAMMING AND SPOOFING

Dinesh Sathyamoorthy*, Zainal Fitry M Amin, Esa Selamat, Shahrudin Abu Hassan, Ahmad Firdaus Ahmad Kazmar & Zaherudin Zaimy

Science & Technology Research Institute for Defence (STRIDE), Ministry of Defence, Malaysia

*Email: dinesh.sathyamoorthy@stride.gov.my

ABSTRACT

This study is aimed at evaluating the vulnerabilities of unmanned aerial vehicles (UAVs) to Global Positioning System (GPS) jamming and spoofing. It is conducted for two commercial UAVs: Osman X28 (UAV A) and CSJ S167 (UAV B). The results obtained demonstrate that UAVs are susceptible to GPS jamming and spoofing even at relatively low power levels of interference / spoofing signals. Given the increasing proliferation of UAVs in various applications, attention should be given to addressing the security vulnerabilities of UAVs.

Keywords: Unmanned aerial vehicles (UAVs); Global Positioning System (GPS) jamming and spoofing; GPS L1 coarse acquisition (C/A) signal; signal power level; position fix loss.

1. INTRODUCTION

Unmanned aerial vehicles (UAVs), also known as drones, are aircrafts that do not carry any crew, but rather are operated remotely by human operators, or autonomously via preprogrammed software or robots. UAVs vary widely in size and capacity, and have become increasingly prevalent. Their use has increased exponentially over the last decade for a broad range of applications, including cartography and mapping, inspection of remote power lines and pipelines, delivery services, telecommunications relay, police surveillance, traffic monitoring, border patrol and reconnaissance, as well as emergency and disaster monitoring (Dinesh, 2010, 2015, Bhattacharjee, 2015; Kille, *et al*., 2019).

From a military perspective, UAVs, which can be recoverable or expendable, are generally used to operate in dangerous or hostile territories, without endangering the operators. It is employed for surveillance and reconnaissance, information collection, detection of mines, and for combat purposes. UAVs hold many attractions for the military. They are generally smaller, lighter and cheaper as compared to manned aerial vehicles as they do not need equipment to support a crew. The recent commercial availability of a new generation of small UAVs, often quadcopters or some other form of rotorcraft, has seen UAVs being increasingly used for various civilian applications (Dinesh, 2010, 2015; Bhattacharjee, 2015; Kille, *et al*., 2019).

Virtually all modern commercial UAVs capable of autonomous flight are navigated using Global Navigation Satellite Systems (GNSS). Civilian GNSS signals are weak, rendering them susceptible to jamming (Dinesh, 2009), as well as unencrypted and unauthenticated, rendering them susceptible to spoofing (Dinesh, 2013). The weak security of GNSS can be taken advantage of to confuse or commandeer UAVs (Gettinger, 2015; Humpreys, 2015, 2017; Westbrook, 2019; Harrison *et al*., 2020).

Jamming is defined as the broadcasting of a strong signal that overrides or obscures the signal being jammed (NAWCWD, 2013; Poisel, 2013; Adamy, 2015; Li *et al.*, 2018). Since GNSS satellites, powered by photocells, are approximately 20,000 km above the Earth surface, GNSS signals that reach the Earth have very low power levels (approximately -160 to -130 dBm), rendering them highly susceptible to jamming (Dinesh, 2009; Jones, 2011; Last, 2016; Parkinson, 2016; Faria, *et al.*, 2018). For example, a simple 1 W battery-powered jammer can block the reception of GNSS signals approximately within a radius of 35 km from the jammer (Papadimitratos & Jovanovic, 2008; Borio *et al.*, 2016). Given the various incidents of intentional and unintentional jamming of GNSS signals, including military GNSS signals (Adams, 2001; Jewell, 2007; Seo & Kim, 2013; Buesnel, 2016; GPS World, 2019; Harrison *et al*., 2020), the development of various GNSS anti-jamming technologies has received significant attention (Jones, 2011; Bar-Sever, 2016; Falleti *et al*., 2020).

Spoofing refers to forging and transmission of false navigation messages in order to manipulate the navigation solutions of GNSS receivers. Spoofing signals can be generated by GNSS simulators, equipment which is commercially available today. The received power of the spoofing signal should exceed that of the legitimate signal, this being essentially a form of jamming. The receiver then operates with the forged signal as the input and computes the location induced by the spoofer. Spoofing is more sinister than intentional jamming because the targeted receiver cannot detect a spoofing attack and hence, cannot warn users that its navigation solution is untrustworthy. While spoofing is more difficult to achieve than jamming, in many cases even if a spoofer is not fully successful, it can still create significant errors and jam GNSS signals over large areas (Volpe, 2001; Scott, 2012; Dinesh, 2013; Ruegamer & Kowalewski, 2015; Liu *et al*., 2018). Hanlon *et al.* (2009) and Montgomery *et al.* (2009) classified GNSS spoofers into three categories, simplistic, intermediate and sophisticated, depending on their complexity and level of robustness required to the associated counter-spoofing measures. While GNSS spoofing was initially considered as an emerging risk, recent incidents have deemed it to be a major threat (Shepard *et al.*, 2012; Tucker, 2015; Bhatti & Humphreys, 2017; C4ADS, 2019; Goward, 2020; Harrison *et al*., 2020)

Many current GNSS receiver evaluations concentrate on radio frequency interference (RFI) and spoofing operability (ION, 1997; Gautier, 2003; Boulton *et al*., 2011; Glomsvoll, 2014; Johnson *et al*., 2018). In previous studies conducted by the Science & Technology Research Institute for Defence (STRIDE), the effect of RFI on Global Positioning System (GPS) performance was studied via field evaluations (Dinesh *et al*., 2009, 2010a, b; Ahmad Norhisyam *et al*., 2013a, b) and GPS simulation (Dinesh *et al*., 2012a, 2014a, 2017a). Field evaluations were also used to study the effect of GPS spoofing (Dinesh *et al*., 2012b).

This study is aimed at evaluating the vulnerabilities of UAVs to Global Positioning System (GPS) jamming and spoofing. The study focuses on the GPS L1 coarse acquisition (C/A) signal, which is the most commonly used civilian GPS signal. It is an unencrypted signal with a fundamental frequency of 1,575.42 MHz, and a code structure that modulates the signal over a 2 MHz bandwidth (DOD, 2001; USACE, 2011; Kaplan & Hegarty, 2017). The study is conducted for two commercial UAVs: Osman X28 (UAV A) and CSJ S167 (UAV B).

2. GPS JAMMING

2.1 Methodology

The evaluation of the effect of GPS jamming on UAVs is conducted using GPS simulation, which will allow for the tests to be held with various repeatable conditions, as defined by the users. As the tests are conducted in controlled laboratory environments, they will not be inhibited by unintended signal interferences and obstructions (Aloi *et al*., 2007; Kou & Zhang, 2011; Pozzobon *et al*., 2013). In our previous studies, in addition to RFI, GPS simulation was used to evaluate the vulnerabilities of GPS to multipath (Dinesh *et al*., 2013, 2014b), GPS satellite clock error (Dinesh *et al*., 2015a, 2019), varying speeds (Dinesh *et al.*, 2015b), power consumption (Dinesh *et al.*, 2016) and GPS antenna orientation (Dinesh *et al.*, 2017b).

The apparatus used in the study are an Aeroflex GPSG-1000 GPS simulator (Aeroflex, 2010), an Advantest U3751 spectrum analyser (Advantest, 2009), an IFR 2023B signal generator (IFR, 1999) and a Hyperlog 60180 directional antenna (Aaronia, 2009). The study is conducted in STRIDE's semi-anechoic chamber (A. Faridz, 2010) to avoid external interferences signals and multipath errors. The test setup employed is as shown in Figure 1. Simulated GPS signals are generated using the GPS simulator and transmitted via the coupler, while interference signals are generated using the signal generator and transmitted via the directional antenna. The following assumptions are made for the tests:

i) No ionospheric or troposheric delays
ii) No clock and ephemeris error
iii) No multipath fading or unintended obstructions
iv) No unintended interference signals.

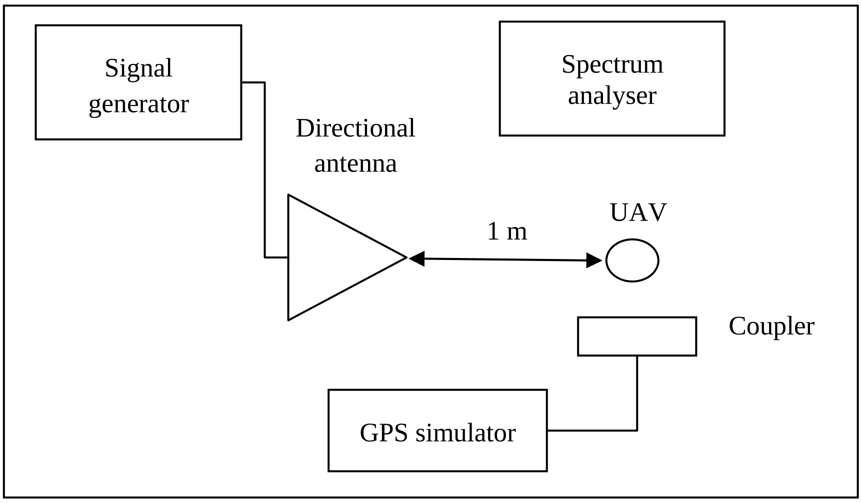


Figure 1: The test setup employed for the study.

The date of simulation is set at 8 January 2020. The almanac data for the period is downloaded from the US Coast Guard's web site (USCG, 2020), and imported into the GPS simulator. This study is conducted for GPS signal power levels of -130, -135, -140 and -145 dBm. The test procedure is conducted for coordinated universal time (UTC) of 0000 at Kajang, Selangor, Malaysia (N 2° 58', E 101° 48').

The interference signal used is a frequency modulated (FM) signal with information frequency of 5 kHz and bandwidth of 2 MHz. The carrier frequency is varied from 1,475 to 1,675 MHz at intervals of 5 MHz. Interference signal transmission is started at power level of -140 dBm. The power level is increased by increments of 3 dBm until the UAV's GPS readout is jammed.

2.2 Results & Discussion

The interference signal power levels at which the location fix for both UAVs is lost are shown in Figure 2. It is observed that with decreasing power levels of the GPS signal, the power levels of interference signals required to jam the GPS signal decrease, and the range of frequencies of interference signals that affect the GPS signal increase. This is as decreasing GPS signal power level results in reduced carrier-to-noise density (C/N_0) levels for GPS satellites tracked by the receivers, which is the ratio of received GPS signal power level to noise density. Lower C/N_0 levels result in increased data bit error rate when extracting navigation data from GPS signals, and hence, increased carrier and code tracking loop jitter. This in turn results in more noisy range measurements and thus, the GPS signal is more susceptible to interference (DOD, 2001; Petovello, 2009; USACE, 2011; Kaplan & Hegarty, 2017). In addition, UAV A is found to have better RFI operability as compared to UAV B. This is as the GPS receiver for UAV A has higher receiver sensitivity, allowing it to track increased C/N_0 levels.

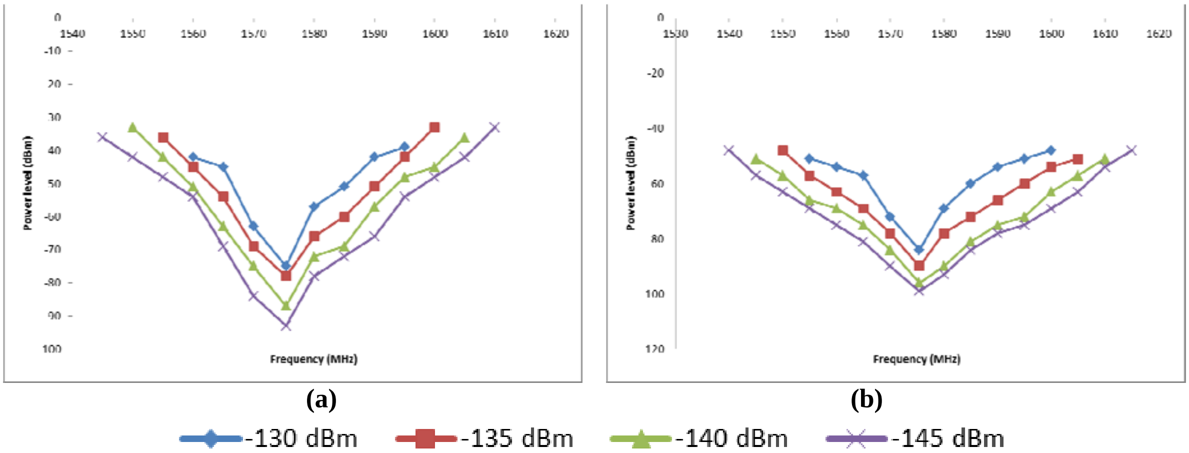


Figure 2: Interference signal power levels at which the location fix is lost for: (a) UAV A (b) UAV B.

It is observed that the interference signal power levels required to jam the GPS signal are significantly high as compared to the corresponding GPS signal power levels. The noise-like C/A code structure, which modulates the L1 signal over a 2 MHz bandwidth, allows for the signal to be received at low levels of interferences. The P(Y) code (restricted to the US military) has a more robust structure, modulating the L1 and L2 signals over 20 MHz bandwidths, and has better resistance to interference. In addition, the absence of other error parameters, including ionospheric and tropospheric delays, satellite clock, ephemeris and multipath errors, and unintentional signal interferences and obstructions, resulted in the required minimum jamming power levels in this study to be significantly higher as compared to field evaluations conducted in Dinesh *et al*. (2009a, 2010a, b) and Ahmad Norhisyam *et al*. (2013a, b).

It should also be noted that, as demonstrated in Dinesh (2010b, 2012a, 2017a), interference signals with power levels that are lower than the minimum jamming power level could still cause disruption to the accuracy of the GPS readings. However, this could not be investigated in this study as the GPS readouts for the UAVs did not provide accuracy assessment.

3. GPS SPOOFING

3.1 Methodology

The UAVs' GPS performance is evaluated during simplistic GPS spoofing attacks, whereby spoofing is conducted using a standalone GPS simulator, which at present poses the greatest near-term threat. In this type of spoofing attack, the spoofing signal is not synchronised (in terms of power level, phase, Doppler shift and data content) with the genuine GPS signals received by the target GPS receiver. This could cause the target GPS receiver to temporarily lose position fix lock first, before being taken over by the spoofing signal. Even if the unsynchronised attack could avoid causing loss of lock, it could still cause an abrupt change in the target GPS receiver's time estimate. Rudimentary counter-spoofing measures, such as amplitude and time-of-arrival discrimination, and loss of lock notification, could be used to detect simplistic spoofing attacks. However, many of present civilian GNSS receivers are not equipped with such measures, and hence, are vulnerable to simplistic spoofing attacks (Humphreys *et al.*, 2008; Hanlon *et al.*, 2009; Montgomery *et al.*, 2009; Ruegamer & Kowalewski, 2015; Liu *et al*., 2018).

The study is conducted via field evaluations held at the STRIDE Block B car park (Figure 3) on 10-11 June 2020. Trimble Planning's (Trimble, 2020) estimate of GPS satellite coverage indicated that the periods of the tests (UTC 0100 - 0400) coincided with periods of good GPS coverage, with high satellite visibility (generally between 8 to 12 satellites) and low position dilution of precision (PDOP) (generally between 1.6 to 2.2) values. Nevertheless, Trimble Planning only takes into account estimated satellite positions and geometry, and does not consider other sources of GNSS errors. Furthermore, the parameters of elevation cutoff and obstacles are estimated from 30 m resolution terrain models, which do not take into consideration man-made structures, and thereby, are subject to errors. It should also be noted that the PDOP values provided by Trimble Planning are best case estimates of GNSS coverage in a particular region. The actual PDOP values obtained by a particular GNSS receiver are dependent on its sensitivity and the strength of GNSS signals received.

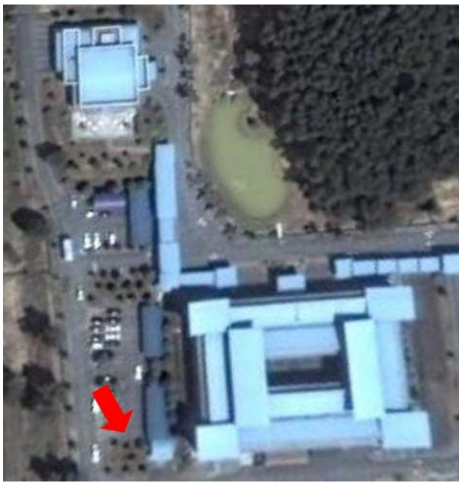


Figure 3: Test area located at N 2º 58.056' E 101º 48.586'
(Source: Screen capture from Google Earth)

The test is conducted using the setup shown in Figure 4. The spoofing signal generated by the GPS simulator is transmitted via a GPS Source A11XLV GPS amplifier (GPS Source, 2006) and a GPS Source L1P GPS passive antenna (GPS Source, 2007). The spectrum analyser is employed to monitor for possible unintentional interference signals. The almanac data for the period of the test is imported into the GPS simulator via its internal GPS receiver. The spoofing signal is set for position of S 16° 55', E 145° 46' (Cairns, Queensland, Australia -approximately 5,000 km from the test area), while the time is set at the simulator's GPS receiver's time. Once a position fix is obtained with the UAV's GPS receiver, transmission of the spoofing signal is started at power level of -140 dBm and increased by increments of 3 dBm. The power level at which loss of position fix occurs and the time required for spoofing to take place are noted.

3.2 Results & Discussion

The results of the tests are shown in Tables 1 and 2. For both the UAVs, varying minimum spoofing signal power levels and times between position fix lost and spoofing are observed for the different dates and times. No clear correlation is observed between these two parameters and the corresponding PDOP values. The buildings and trees in the vicinity of the test area could have resulted in the actual PDOP values being significantly different from the estimated values. Furthermore, other GNSS error parameters, including ionospheric and tropospheric delays, satellite clock, ephemeris and multipath errors, and unintentional signal interferences and obstructions, could have affected GPS coverage during the periods of the tests.

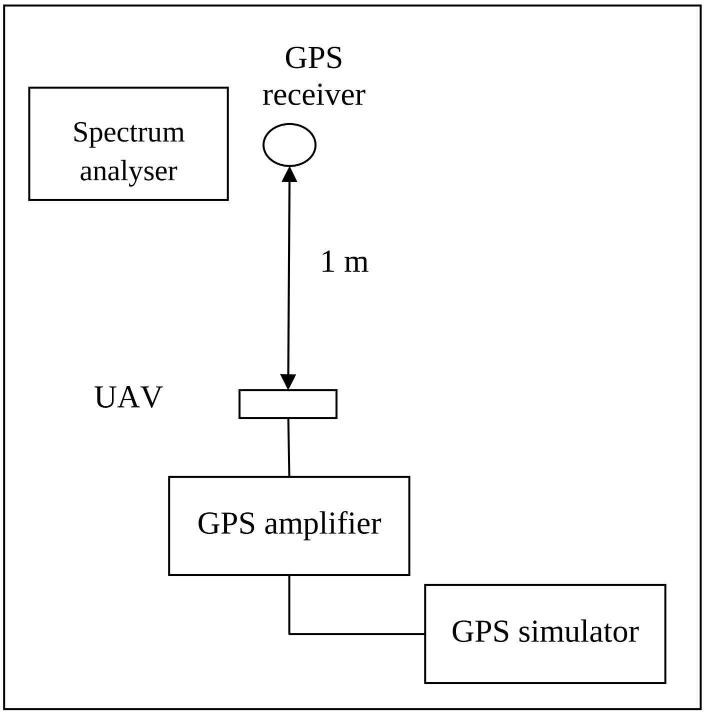


Figure 4: The test setup employed.

Table 1: The effect of spoofing attacks on UAV A.

Reading	Date (2020)	Time	PDOP	Minimum spoofing signal power level (dBm)	Time between position fix loss and spoofing (s)
1	10 June	0938	2.01	-98	83
2	10 June	1121	1.67	-104	26
3	10 June	1306	1.61	-92	11
4	10 June	1514	1.75	-104	16
5	11 June	0943	1.83	-101	33
6	11 June	1201	1.94	-104	24

Table 2: The effect of spoofing attacks on UAV B.

Reading	Date (2020)	Time	PDOP	Minimum spoofing signal power level (dBm)	Time between position fix loss and spoofing (s)
1	10 June	1025	1.64	-104	151
2	10 June	1219	1.93	-101	69
3	10 June	1433	1.96	-104	35
4	10 June	1612	1.78	-107	51
5	11 June	1107	1.62	-104	138
6	11 June	1256	2.16	-101	104

The minimum spoofing signal power level required to cause position fix loss and, subsequently, spoofing is dependent on the received GPS signal power during the tests. During periods of poor coverage, when the received GPS signal power levels are lower, the required minimum spoofing signal powers levels would be lower, and vice-versa. Similar to the GPS jamming evaluation, it is also observed that the minimum spoofing power levels are significantly higher as compared to the received GPS signal power level. However, the required minimum spoofing power levels required to cause position fix loss is lower as compared to required minimum interference signal power levels observed during the GPS jamming evaluation. This occurred as the difference in synchronisation between the genuine and spoofing GPS signals forced the GPS receiver to recompute its position fix at relatively lower spoofing signal power levels.

At the minimum spoofing power level, the time between position fix loss and spoofing is dependent on the level of synchronisation, in terms of power level, phase, Doppler shift and data content, between the genuine and spoofing GPS signals. When the both signals are closely synchronised, spoofing occurs very quickly. However, when the signals are largely unsynchronised, position fix loss occurs for a longer period of time as the GPS receiver has to recompute is position fix.

It is generally observed that spoofing occurred faster for UAV A as compared to UAV B. This occurred as the GPS receiver for UAV A has significantly better performance characteristics, particularly receiver sensitivity and reacquisition time, which allows it to recompute its position fix faster as compared to the GPS receiver for UAV B. In other words, the improved design that allows the GPS receiver for UAV A to have better performance has, ironically, made it more vulnerable to spoofing.

Similar to GPS jamming, as demonstrated in Dinesh (2012b), spoofing signals with power levels that are lower than the minimum spoofing power level could still cause disruption to the accuracy of the GPS readings. Furthermore, once spoofing has taken place, as the spoofing signal has higher power level than the actual GPS signal, it results in higher GPS reading accuracy, albeit for a wrong position / time.

4. CONCLUSION

This study has demonstrated that the UAVs are susceptible to GPS jamming and spoofing even at relatively low power levels of interference / spoofing signals. Given the increasing proliferation of UAVs in various applications, attention should be provided to addressing the security vulnerabilities of UAVs.

It should be noted that the tests conducted in this study were for only two UAVs. Additional tests using a wider range of UAVs are needed to further validate the findings of this study. Furthermore, a limitation faced in this study was that the GPS simulator used only allows the transmission of the GPS L1 C/A signal. The proposed future work is for the procurement of a GNSS simulator that will allow transmission of other GPS signals, in particular L2C and L5, along with signals of other GNSS systems (GLONASS, BeiDou and Galileo).

REFERENCES

A. Faridz, A.G., M. Razali, M.Y. & W. Salwa, W.H. (2010). STRIDE's 3 meters EMC semi-anechoic chamber: Design considerations and compliance to standards. *2010 IEEE Asia-Pacific Conf. Appl. Electromagnetics (APACE 2010)*, 9-11 November 2010, Port Dickson, Negeri Sembilan, Malaysia.

Aaronia (2009). *Precompliance Test Antenna Series HyperLOG® 60xxx: Span 680 MHz to 18 GHz*. Aaronia AG, Strickscheid, Germany.

Adams, T.K. (2001). GPS vulnerabilities. *Mil. Rev.*, 1: 10-16.

Adamy, D.L. (2015). *Electronic Warfare Against a New Generation of Threats*. Artech House, Boston.

Advantest (2009). *U3741/3751 Spectrum Analyzers*. Advantest Corporation, Chiyoda-ku, Tokyo.

Aeroflex (2010). *Avionics GPSG-1000 GPS / Galileo Portable Positional Simulator*. Aeroflex Inc., Plainview, New York.

Ahmad Norhisyam, I., Dinesh, S. & Azman, M.S., 2013. Effect of radio frequency interference (RFI) on the performance of Global Positioning System (GPS) static observations. *9th IEEE Colloq. Signal Process. Appl. (CSPA 2013)*, 8-10 March 2013, Kuala Lumpur.

Ahmad Norhisyam, I., Dinesh, S. & Azman, M.S., 2013. Effect of radio frequency interference (RFI) on the precision of GPS relative positioning, *8th Int. Symp. Digital Earth 2013 (ISDE 2013)*, 26-29 August 2013, Kuching, Sarawak.

Aloi, D.N., Alsliety, M. & Akos, D.M. (2007). A methodology for the evaluation of a GPS receiver performance in telematics applications. *IEEE T. Instrum. Meas.*, 56: 11-24.

Bar-Sever, Y. (2016). Networks for robust civil signal performance monitoring & RFI detection. *17th Space-Based PNT Advisory Board Meet.*, 18-19 May 2016, National Harbor, Maryland.

Bhattacharjee, D. (2015). *Unmanned Aerial Vehicles and Counter Terrorism Operations*. Indian Council of World Affairs, New Delhi.

Bhatti, J.A. & Humphreys, T.E. (2017). Hostile control of ships via false GPS signals: Demonstration and detection. *Nav.*, **64**: 51-66.

Borio, D., Dovis, F., Kuusniemi, H. & Lo Presti (2016). Impact and detection of GNSS jammers on consumer grade satellite navigation receivers. *Proc. IEEE*, 104: 1233-1245.

Boulton, P., Borsato, R., Butler, B. & Judge, K. (2011). GPS interference testing: Lab, live, and LightSquared. *Inside GNSS*, 5: 32-45.

Buesnel, G. (2016). *GPS Jamming Incident at Cairo Airport Highlights Growing Risks to Navigation Systems*. Available online at: https://www.linkedin.com/pulse/gps-jamming-incident-cairo-airport-highlights-growing-guy-buesnel (Last access date: 9 September 2016).

C4ADS (2019). *Above Us Only Stars: Exposing GPS Spoofing in Russia and Syria*. C4ADS, Washington D.C.

Dinesh, S. (2009). Vulnerabilities of civilian Global Navigation Satellite Systems (GNSS) signals: A review. *Defence S&T Tech. Bull.*, 2: 100-114.

Dinesh, S. (2010). Key defence R&D fields to develop the national defence industry: Focus on C4ISR in support of network centric operations and unmanned vehicles. *Defence S&T Tech. Bull.*, **3**: 43-60.

Dinesh, S., 2013. Global Navigation Satellite System (GNSS) spoofing: A review of growing risks and mitigation steps. *Defence S&T Tech. Bull.*, **6**: 42-61.

Dinesh, S. (2015). A review of security threats of unmanned aerial vehicles and mitigation steps. *J. Defence Secur.*, **6**: 81-97.

Dinesh, S., Wan Mustafa, W.H., Mohd Faudzi, M., Kamarulzaman, M., Nor Irza Shakhira, B., Siti Robiah, A., Norhayaty, Z., Aliah, I., Lim, B.T., Arumugam, P., Zainal Fitry, M.A., Mohd Rizal, A.K., Azlina, B. & Mohd. Hasrol, H.M.Y. (2009). Evaluation of the effect of radio frequency interference (RFI) on Global Positioning System (GPS) receivers. *Defence S&T Tech. Bull.*, 2: 115-129.

Dinesh, S., Wan Mustafa, W.H., Mohd Faudzi, M., Kamarulzaman, M., Hasniza, H., Nor Irza Shakhira, B., Siti Robiah, A., Shalini, S., Jamilah, J., Aliah, I., Lim, B.T., Zainal Fitry, M.A., Mohd Rizal, A.K., Azlina, B. & Mohd Hasrol, H.M.Y. (2010a). Evaluation of power levels required by interference signals at various distances to jam the Global Positioning System (GPS) L1 coarse acquisition (C/A) signal. *Defence S&T Tech. Bull.*, 3: 14-28.

Dinesh , S., Wan Mustafa, W.H., Mohd Faudzi., M., Kamarulzaman, M., Hasniza, H., Nor Irza Shakhira, B., Siti Robiah, A., Shalini, S., Jamilah, J., Aliah, I., Lim, B.T., Zainal Fitry, M.A., Mohd. Rizal, A.K., Azlina, B. & Mohd. Hasrol, H.M.Y. (2010b). Evaluation of the effect of radio frequency interference (RFI) on Global Positioning System (GPS) accuracy. *Defence S&T Tech. Bull.*, 3: 100-118.

Dinesh, S, Mohd Faudzi, M. & Zainal Fitry, M.A. (2012a). Evaluation of the effect of radio frequency interference (RFI) on Global Positioning System (GPS) accuracy via GPS simulation. *Defence. Sci. J.*, **62**: 338-347.

Dinesh, S., Mohd Faudzi., M., Nor Irza Shakhira, B., Siti Robiah, A., Shalini, S., Aliah, I., Lim, B.T., Zainal Fitry, M.A., Mohd. Rizal, A.K., & Mohd Hasrol, H.M.Y. (2012b). Evaluation of Global Positioning System (GPS) performance during simplistic GPS spoofing attacks. *Defence S&T Tech. Bull.*, **5**: 99-113.

Dinesh, S., Shalini, S., Zainal Fitry, M.A. & Siti Zainun, A. (2013). Evaluation of the repeatability of Global Positioning System (GPS) performance with respect to GPS satellite orbital passes. *Defence S&T Tech. Bull.*, **6**: 130-140.

Dinesh, S., Mohd Faudzi, M., Rafidah, M., Nor Irza Shakhira, B., Siti Robiah, A., Shalini, S., Aliah, I., Lim, B.T., Zainal Fitry, M.A., Mohd Rizal, A.K. & Mohd Hasrol Hisam, M.Y. (2014a). Evaluation of the effect of radio frequency interference (RFI) on Global Positioning System (GPS) receivers via GPS simulation. *ASM Sci. J.*,**8**: 11-20.

Dinesh, S., Shalini, S., Zainal Fitry, M.A., Siti Zainun, A., Siti Robiah, A., Mohd Idris, I. & Mohd Hasrol Hisam, M.Y. (2014b). Evaluation of the effect of commonly used materials on multipath propagation of Global Positioning System (GPS) signals via GPS simulation. *Adv. Mil. Tech.*, **9**: 81-95.

Dinesh, S., Shalini, S., Zainal Fitry, M.A., Asmariah, J. & Siti Zainun, A. (2015a). Evaluation of the effect of Global Positioning System (GPS) satellite clock error via GPS simulation. *Defence S&T Tech. Bull.*, **8**: 51-62.

Dinesh, S., Shalini, S., Zainal Fitry, M.A., Asmariah, J. & Siti Zainun, A. (2015b). Evaluation of the accuracy of Global Positioning System (GPS) speed measurement via GPS simulation. *Defence S&T Tech. Bull.,* **8**: 121-128.

Dinesh, S., Shalini, S., Zainal Fitry, M.A., Asmariah, J. & Siti Zainun, A. (2016). Evaluation of trade-off between Global Positioning System (GPS) accuracy and power saving from reduction of number of GPS receiver channels. Appl. Geomatics, **8**: 67-75.

Dinesh, S., Zainal Fitry, M.A. & Shahrudin, A.H. (2017a). Evaluation of Global Positioning System (GPS) adjacent band compatibility via GPS simulation. *Defence S&T Tech. Bull.*, **10**: 229 – 235.

Dinesh, S., Shalini, S., Zainal Fitry, M.A., Mohamad Firdaus, A., Asmariah, J. & Siti Zainun, A. (2017b). Evaluation of the effect of Global Positioning System (GPS) antenna orientation on GPS performance. *Defence S&T Tech. Bull.*, **10**: 33-39.

Dinesh, S., Zainal Fitry, M.A & Esa, S. (2019). Evaluation of the effectiveness of receiver autonomous integrity monitoring (RAIM) in Global Positioning System (GPS) receivers. *Defence S&T Tech. Bull.*, **12**: 295-300.

DOD (Department of Defence) (2001). *Global Positioning System Standard Positioning Service Performance Standard, Command, Control, Communications, and Intelligence*. Department of Defence (DOD), Washington D.C.

Falleti, E., Gamba, M.T. & Pini, M. (2020). Design and analysis of activation strategies for adaptive notch filters to suppress GNSS jamming. *IEEE T. Aero. Elec. Sys.*, In press

Faria, L.A., Silvestre, C.A.M, correia, M.A.F & Roso, N.A (2019). GPS jamming signals propagation in free-space, urban and suburban environments. *J. Aerosp. Technol. Manag.*, **10**: e0618.

Gautier, J. (2003). *GPS/INS Generalized Evaluation Tool (GIGET) for the Design and Testing of Integrated Navigation Systems*. Ph.D. Thesis, Harvard University, Cambridge, Massachusetts.

Gettinger, D. (2015). *Domestic Drone Threats*. Available online at: http://dronecenter.bard.edu/what-you-need-to-know-about-domestic-drone-threats (Last access date: 14 July 2015).

Glomsvoll, O. (2014). *Jamming of GPS and GLONASS Signals: A Study of GPS Performance in Maritime Environments Under Jamming Conditions, and Benefits of Applying GLONASS in Northern Areas Under Such Conditions*. Master’s Thesis, University of Nottingham, Nottingham.

GPS World (2019). *Year-Long Ocean Cruise Finds GNSS Interference…Everywhere*. Available online at: https://www.gpsworld.com/year-long-ocean-cruise-finds-gnss-interference-everywhere (Last access date: 19 April 2020).

GPS Source (2006). *L1P GPS Antenna. GPS Source*. GPS Source Inc., Pueblo West, Colarado.

GPS Source (2007). *A11XLV Digital Variable Gain GPS Amplifier*. GPS Source Inc., Pueblo West, Colarado.

Goward, D. (2020). *GPS Circle Spoofing Discovered in Iran*. Available online at: https://www.gpsworld.com/gps-circle-spoofing-discovered-in-iran (Last access date: 21 April 2020).

Hanlon, B.O., Ledvina, B., Psiaki, M.L., Kintner. P.M. & Humphreys, T.E. (2009). *Assessing the Spoofing Threat*. Available online at: http://www.gpsworld.com/defence/security-surveillance/assessing-spoofing-threat-3171?page_id=1 (Last access date: 4 November 2009).

Harrison, T., Johnson, K., Roberts, T.G. & Way, T. (2020). *Space Threat Assessment 2020*. Center for Strategic and International Studies (CSIS), Washington, D.C.

Humphreys, T.E., Ledvina, B.M., Psiaki, M.L., & Kintner, J. (2008). Assessing the spoofing threat: Development of a portable GPS civilian spoofer. *ION GNSS 2008*, 16-19 September 2008, Savannah International Convention Center, Savannah, Georgia.

Humpreys, T. (2015). *Statement on the Security Threat Posed by Unmanned Aerial Systems and Possible Countermeasures*. Statement to the Subcommittee on Oversight and Management Efficiency of the House Committee on Homeland Security, 18 March 2015, Washington D.C.

Humphreys, T. (2017). Counter-UAV challenges: Is GNSS spoofing effective? *ION GNSS+ Hostile MAV Threats, Detection and Countermeasures*, September 2017.

IFR (1999). *2023A/B, 2025 Signal Generators*. IFR Americas Inc., Wichita, Kansas.

ION (Institute of Navigation) (1997). *Institute of Navigation Standard 101 (ION STD 101): Recommended Test Procedures for GPS Receivers, Revision C*. Institute of Navigation (ION), Manassas, Virginia.

Jewell, J. (2007). *GPS Insights: JNC Briefing on Jamming Incident*. Available online at: http://gpsworld.com/defensegps-insights-april-2007-8428 (Last access date: 9 September 2016).

Johnson, C., Lee, C. & Ascencio, M. (2018). Developmental test NAVFEST: A large-scale, multi-aircraft, GPS jamming test event. *2018 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, 23-26 April 2018, Monterey, California.

Jones, M. (2011). The civilian battlefield: Protecting GNSS receivers from interference and jamming. *Inside GNSS*, **6**: 40-49.

Kaplan, E.D. & Hegarty, C.J. (2017). *Understanding GPS/GNSS: Principles and Applications, 3rd Ed*. Artech House, Norwood, Massachusetts.

Kille, T., Bates, P.R. & Lee S.Y. (2019). *Unmanned Aerial Vehicles in Civilian Logistics and Supply Chain Management*. IGI Global, Hershey, Pennsylvania.

Kou, Y. & Zhang, H. (2011). Verification testing of a multi-GNSS RF signal simulator. *Inside GNSS*, **6**: 52-61.

Last, D. (2016). GNSS vulnerability to jamming and spoofing. *Nordic Navigation Forum*, 2 February 2016, Bodø, Norway.

Li, T., Song, T. & Liang, Y. (2018). *Wireless Communications under Hostile Jamming: Security and Efficiency*. Springer, Singapore.

Liu, Y., Li, Sihai, L., Fu, Q. & Liu, Z. (2018). Impact assessment of GNSS spoofing attacks on INS/GNSS integrated navigation system. *Sensors*, **18**: 1433.

Montgomery, P., Humphreys, T.E. & Ledvina, B.M. (2009). A multi-antenna defence receiver-autonomous GPS spoofing detection. *Inside GNSS*, **4**: 40-46.

NAWCWD (Naval Air Warfare Center Weapons Division) (2013). *Electronic Warfare & Radar Systems Engineering Handbook*. Raleigh, North Carolina.

Papadimitratos, P. & Jovanovic, A. Protection and fundamental vulnerability of GNSS. *Int. Workshop Satell. Space Commun.* 2008 (IWSSC'08), France, 1-3 October 2008, Institut Supérieur de l'Aéronautique et de l'Espace (ISAE), Toulouse, France.

Parkinson, B. (2016). GPS spoofing & jamming: Uncovering vulnerability truths & myths. *17th Space-Based PNT Advisory Board Meet.*, 18-19 May 2016, National Harbor, Maryland.

Petovello, M. (2009). Carrier-to-noise density and AI for INS / GPS integration. *Inside GNSS*, **4**: 20-29.

Poisel, A.R. (2013). *Information Warfare and Electronic Warfare Systems*. Artech House, Boston.

Pozzobon, O., Sarto, C., Chiara, A.D., Pozzobon, A., Gamba, G., Crisci, M. & Ioannides, R. (2013). Developing a GNSS position and timing authentication testbed: GNSS vulnerability and mitigation techniques. *Inside GNSS*, **8**: 45-53.

Ruegamer, A. & Kowalewski, D. (2015). Jamming and spoofing of GNSS signals –An underestimated risk. *FIG Working Week 2019*, 17-21 May 2015, Sofia, Bulgaria

Seo, J. & Kim, M. (2013). Loran in Korea: Current status and future plans. *Eur. Navig. Conf. (ENC 2013)*, 23–25 April 2013, Vienna, Austria.

Scott, L.S. (2012). Spoofs, proofs & jamming: Towards a sound national policy for civil location and time assurance. *Inside GNSS*, 7: 42-53.

Shepard, D.P., Bhatti, J.A. & Humphreys, T.E. (2012). Evaluation of smart grid and civilian UAV vulnerability to GPS spoofing attacks. *ION GNSS 2012*, 17-21 September 2012, Nashville Convention Center, Nashville, Tennessee.

Tucker, P. (2015). *DHS: Drug Traffickers Are Spoofing Border Drones*. Available online at: https://www.defenseone.com/technology/2015/12/DHS-Drug-Traffickers-Spoofing-Border-Drones/124613 (Last access date: 19 April 2020).

Trimble (2020). *Trimble's Planning Software*. Available online at:

http://www.trimble.com/planningsoftware.shtml (Last access date: 20 April 2020).

USACE (US Army Corps of Engineers) (2011). *Engineer Manual EM 1110-1-1003: NAVSTAR Global Positioning System Surveying.* US Army Corps of Engineers (USACE), Washington D.C.

USCG (US Coast Guard) (2020). GPS NANUs, Almanacs, & Ops Advisories. Available online at: http://www.navcen.uscg.gov/?pageName=gpsAlmanacs (Last access date: 20 April 2020).

Volpe (2001). Vulnerability *Assessment of the Transport Infrastructure Relying on the Global Positioning System*. John A. Volpe National Transportation Systems Center, Department of Transport, Washington D.C.

Westbrook, T. (2019). The Global Positioning System and military jamming: The geographies of electronic warfare. *J. Strategic Stud.*, **12**: 1-16.

CORONAVIRUS SPATIAL BIG DATA PREDICTIVE ANALYSIS FOR THE SOUTHEAST ASIAN REGION

Arun Kumar Verma[1*], Anjul Verma[2] & Aditi Verma[3]

[1]Vidyadaan Institute of Technology and Management, Aryabhatta Knowledge University, India
[2]Business School, University of Liverpool, Liverpool, England
[3]Qualcomm India Pvt. Ltd., India

[*]Email:arun@vidyadaan.org

ABSTRACT

The outbreak of the 2019 novel coronavirus disease (COVID-19) spread geospatially to more than 200 countries across the globe causing more than 12.39 mil people of the global population to be infected and 0.55 mil deaths (as of 10 July 2020), which is exponentially increasing and spreading in a spatiotemporal way to new geographical locations. This has led to a serious threat to human health and life, posing challenges to control the severity of the coronavirus spectrum. In the southeast Asian region, the outbreak of COVID-19 first arrived in Thailand on 13 January 2020, followed by South Korea on 20 January 2020, Vietnam and Taiwan on 22 January 2020, Hong Kong and Singapore on 23 January 2020, Malaysia on 25 January 2020, and Philippines on 30 January 2020, before reaching India on 31 January 2020. This has resulted in the imposition of national lockdowns / recommendations to control the outbreak of the coronavirus spectrum. The global spread of the coronavirus spectrum has highlighted the need for big data analysis to help decision makers in the design of lockdown measures. In this paper, coronavirus spatial big data predictive analysis has been carried out for the southeast Asian region along with countries located in different latitudes with varying populations from 4 to 150 mil where coronavirus first reached 100, 1,000 and 5,000 reported cases. The coronavirus predictive models have been developed for different stages of the outbreak based on big data analysis of 5-day averaging of daily new coronavirus spectrum for these countries, which acts as knowledge classifier for predicting the trend of coronavirus spectrum for new geographical locations. In this paper, the impacts of latitude on mortality (death per 1 mil populations) from coronavirus have also been discussed based on big data analysis to understand the alarming phase of the outbreak.

Keywords: *2019 novel coronavirus disease (COVID-19); spatial big data; southeast Asian region; predictive analysis; coronavirus spectrum model.*

1. INTRODUCTION

The outbreak of the 2019 novel Coronavirus disease (COVID-19) emerged from Wuhan and spread throughout the Hubei province of China, and further spatially transmitted in an exponential spread to more than 200 countries (Pranab *et al.*, 2020). The global spread of the coronavirus spectrum created a public health emergency of international concern, which caused more than 12.39 mil people of the global population to be infected and 0.55 mil deaths (as of 10 July 2020). The outbreak of coronavirus is still increasing exponentially and spreading in a spatiotemporal way to new geographical locations, which seriously threatens the human health and life of the people as well as affecting economic and social development (Rajesh & Priya, 2020; Coronavirus, 2020; Prakash, 2020; Chenghu *et al.*, 2020). The spatial spreading of coronavirus due to large scale migration from the Hubei province resulted in an outbreak in the southeast Asian region, covering the latitude of 38°N to 6°S, with the first reported coronavirus case in ~~at~~ Thailand on 13 January 2020, followed by South Korea on 20 January 2020 as well as Vietnam and Taiwan on 22 January 2020. The first coronavirus case was then reported in Hong Kong and Singapore on 23 January 2020, which was followed by Malaysia on 25 January 2020 and Philippines on 30 January 2020, prior to reaching India on 31 January 2020. The spread of coronavirus in the southeast Asian region posed challenges to control the severity of the spectrum of the outbreak (BBC, 2020; WHO, 2020).

The Hong Kong, Vietnamese and South Korean governments imposed national lockdowns / recommendations to control the exponential rise of the spectrum of coronavirus from 8, 13 and 20 February 2020 respectively, after 16, 22 and 31 days of the first reported coronavirus case. The governments of Singapore, Malaysia, Philippines, Thailand, Taiwan and India imposed these measures from 6, 13, 15, 20, 24 and 25 March 2020 respectively, whereas, the Indonesian government imposed a national lockdown from 15 March 2020, after 13 days of the first reported coronavirus case (BBC, 2020; Rajesh, 2020; George & Rebecca , 2020; Tanzin *et al*., 2020). The Myanmarese government executed a national lockdown on 13 March 2020, prior to the arrival of first coronavirus case on 27 March 2020, which was found to be most effective to control its outbreak and keep the country in the most safe zone (BBC, 2020).

The first 100 coronavirus cases were reported in 14 countries from 12 to 15 March 2020, covering the latitude of 60°N to 6°S, which includes Finland, Ireland, Poland, Czech Republic, Romania, Israel, Egypt, Portugal, Brazil and Saudi Arabia, along with India, Philippines, Thailand and Indonesia of the southeast Asian region (Coronavirus, 2020). Similarly, the first 1,000 coronavirus cases were reported in 11 countries from 28 March to 1 April 2020, covering the latitude from 60°N to 35°S, which includes Finland, Russia, Serbia, Mexico, Panama, Columbia, Peru and Argentina, along with India, Philippines and Singapore of the southeast Asian region. Similarly, first 5,000 coronavirus cases were reported in 10 countries from 5 to 9 April 2020, covering the latitude of 56°N to 12°S, which includes Denmark, Russia, Ireland, Poland, Czech Republic, Romania, Japan and Peru, along with India and Malaysia of the southeast Asian region (Coronavirus, 2020; Pranab *et al*., 2020).

The design of different spatial measures of the national lockdown/ recommendations/ health emergency measures, such as surveillance, identification, testing, tracking, segregating the different zones and medical treatment to control the outbreak of coronavirus spectrum, differs from populations, geographical area, communication and medical infrastructure of the country depending on their socio-economic conditions (Rajesh & Priya, 2020; Poonam, 2020; George & Rebecca , 2020). Coronavirus spatial big data predictive analysis plays an important role for predicting the trend of the spectrum and suggesting different measures including supply chain infrastructure to control the outbreak by using geographical information systems (GIS) technologies for geospatial big data analysis (Chenghu *et al*., 2020). Coronavirus spatial big data predictive analysis is helpful for decision and policymakers in taking suitable measures based on the understanding the current spectrum of the coronavirus outbreak. For example, when the mortality per million is plotted against the latitude for 15 April 2020, low population mortality from coronavirus have been observed in the countries below 35 °N, which supports vitamin D as a factor for determining the severity (Panarese & Shahini, 2020; Jonathan *et al*., 2020).

In this paper, coronavirus spatial big data predictive analysis for southeast Asian countries and other countries where coronavirus cases reached the first 100, 1,000 and 5,000 cases, have been carried out to understand the severity based on total coronavirus cases and daily new coronavirus data (COVID-19 India, 2020; Coronavirus, 2020). In the present work, the 5-day moving average techniques has been considered for big data predictive analysis for the purpose of developing coronavirus spectrum models for predicting the different stages of outbreak, which acts as knowledge classifiers for predicting the trend of the spectrum of the outbreak. This paper further presents the impact of latitude on the mortality (death per million populations) and variability of the mortality from coronavirus for the southeast Asian region.

2. CORONAVIRUS SPECTRUM IN THE SOUTHEAST ASIAN REGION

2.1 Basic Details of the Coronavirus Outbreak

Table 1 describes basic details such as latitude, coronavirus tests per 1 mil population, death per 1 mil population and global rank as of 6 May 2020 for the southeast Asian region (BBC,2020; Coronavirus, 2020).

Table 1: Basic details of coronavirus for southeast Asian countries.

Country	Latitude (N/S)	First coronavirus case	National lockdown / recommendations	Global rank 6 May 2020	Test per 1 mil population 3 May 2020	Death per 1 mil population 3 May 2020
South Korea	38	20 January 2020	20 February 2020	38	12488	5.0
India	29	31 January 2020	25 March 2020	14	864	1.0
Taiwan	25	22 January 2020	24 March 2020	122	2727	0.3
Hong Kong	22	23 January 2020	08 February 2020	90	22448	0.5
Myanmar	22	27 March 2020	13 March 2020	142	211	0.1
Vietnam	21	22 January 2020	13 February 2020	131	2681	0.0
Philippines	15	30 January 2020	15 March 2020	39	1151	6.0
Thailand	14	13 January 2020	20 March 2020	65	3264	0.8
Malaysia	3	25 January 2020	13 March 2020	51	6352	3.0
Singapore	1	23 January 2020	06 March 2020	26	24600	3.0
Indonesia	-6	2 March 2020	15 March 2020	36	427	3.0

2.2 Variability of the Spectrum of Total Coronavirus Cases

Figure 1 depicts the spectrum of coronavirus cases from 19 February to 5 May 2020 for southeast Asian countries, which shows that the spectrum of total number of coronavirus cases increased very sharply in South Korea from 1 March 2020 and in Malaysia from 20 March 2020, but the spread of the outbreak of coronavirus in other countries of the region became significant after 1 April 2020 onwards. The spectrum of coronavirus cases increased exponentially and non-linearly for India, Singapore, Indonesia and Philippines. The spectrum in South Korea, Malaysia, Thailand, Hong Kong, Taiwan and Vietnam attained the flattening of the curve, which indicates control of the outbreak due to their effective and successful operational lockdown models.

2.3 Variability of the Spectrum of Daily New Coronavirus Cases

The 5-day moving average technique was applied on the coronavirus spatial big data to understand the various stages of the outbreak. Figure 2(a) depicts the growth of the spectrum of the 5-day moving average for daily new coronavirus cases from 19 February to 31 March 2020 for the southeast Asian region. It is observed that the spectrum increased exponentially for South Korea and attained a maximum value in February 2020, then decreased exponentially until attaining stability to control the outbreak from 15 March 2020 onwards. The spectrum increased non-linearly for Malaysia and attained a maximum value in March 2020, then decreased exponentially until attaining the stability to control the outbreak from 15 April 2020. Other countries in the southeast Asian region, in particular Thailand, Hong Kong, Taiwan, Vietnam and Myanmar, successfully managed to control the growth of daily new coronavirus cases, whereas non-linear exponential growth of the spectrum was observed to continue for India after 25 March 2020. Figure 2(b) depicts the exponential rise of the spectrum of daily new coronavirus cases in

April 2020 for Singapore prior to reaching the peak of the spectrum and further decreasing with fluctuations in the spectrum. Furthermore, the spectrum of new cases for South Korea, Malaysia, Hong Kong, Vietnam, Taiwan and Thailand decreased to the minimum in April 2020, whereas Philippines and Indonesia were struggling to stabilise the growth of new cases.

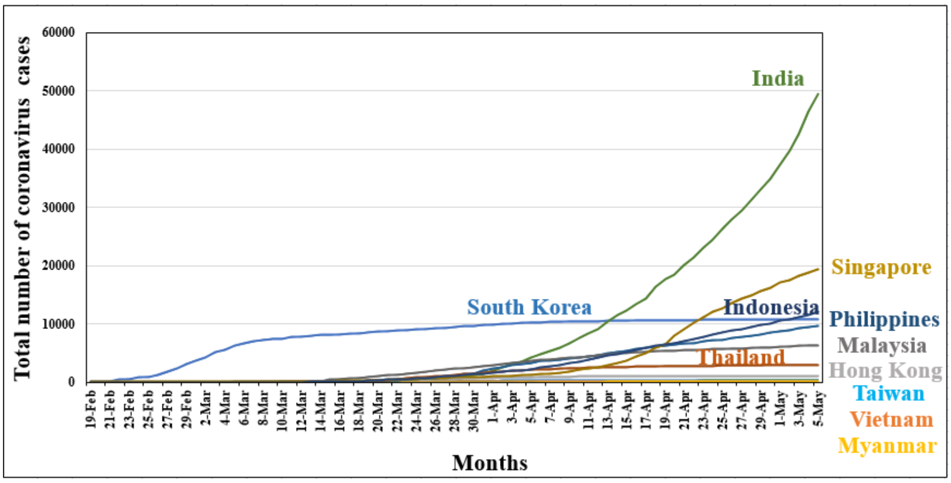
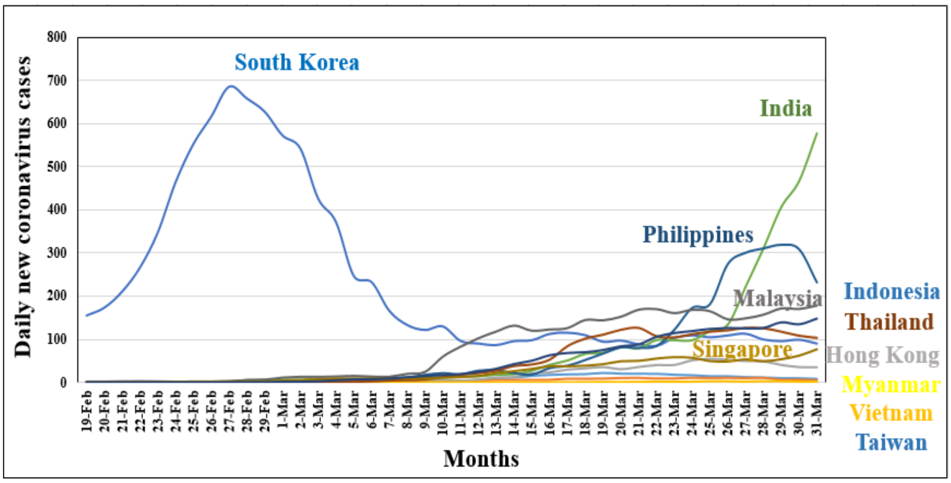
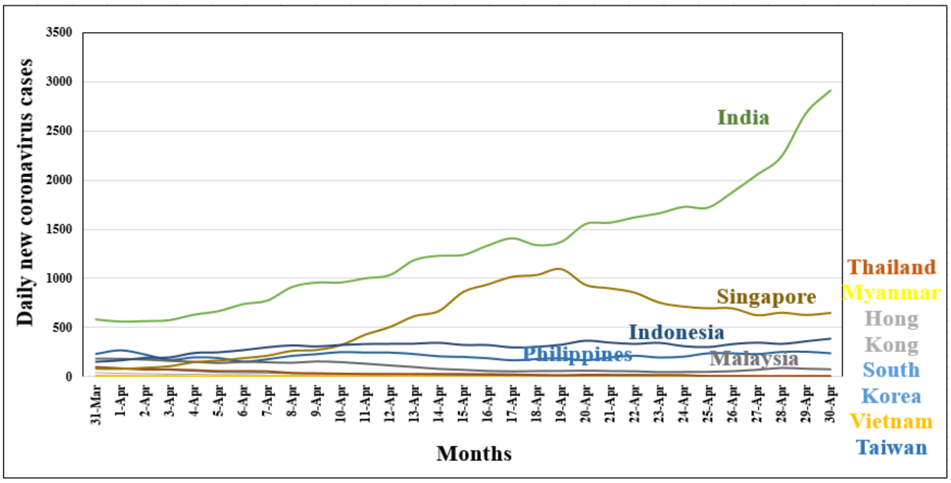


Figure 1: Variation of the spectrum of total number of coronavirus cases for 19 February to 5 May 2020.



(a)



(b)

Figure 2: Daily new coronavirus case spectrum for: (a) 19 February to 31 March 2020 ; (b) 1 -30 April 2020.

2.4 Variability of the Spectrum of Daily New Coronavirus Cases for Countries Reaching the First 100 Cases

Figure 3 depicts the variability of the spectrum of daily new coronavirus cases for countries crossing the first 100 cases on 12-15 March 2020, which shows that the spectrum increased exponentially for Brazil, India and Saudi Arabia from 31 March 2020 onwards and was still continuing to increase with higher slopes to reach beyond the critical stage of the outbreak. It is observed that the spectrum of new cases for Brazil increased exponentially, varying from 2,000 daily new cases to more than 10,000 daily new cases after attaining an unstable oscillatory stage from 25 April 2020 prior to reaching beyond the critical stage. Furthermore, the trend of new cases for India and Saudi Arabia were critical due to continued non-linear exponential rise, whereas the trends of the spectrum for Portugal, Israel, Ireland, Poland, Thailand and Romania confirm successful measures to control the outbreak, with continued exponential decrease of trends of new cases. The trends of the spectrum of other countries, in particular Finland, Egypt, Czech Republic, Philippines and Indonesia, were decreasing to achieve stability to control and reach a safer stage of the outbreak.

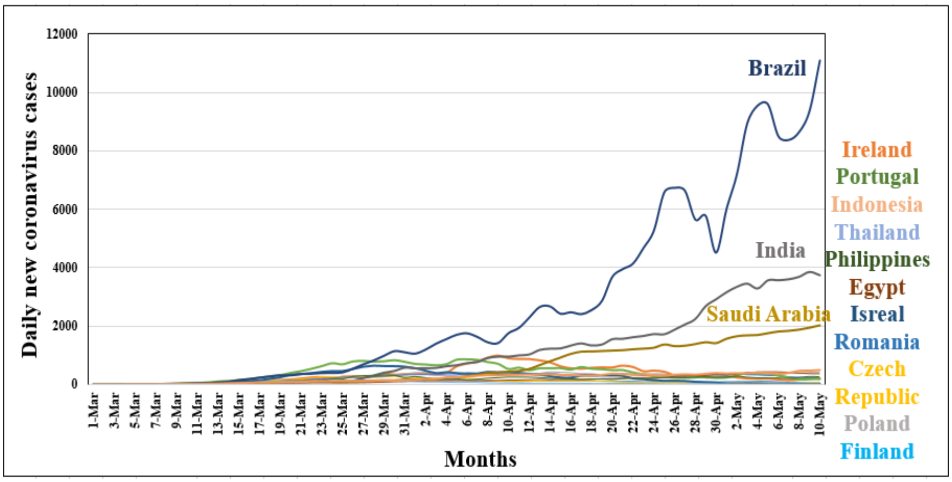
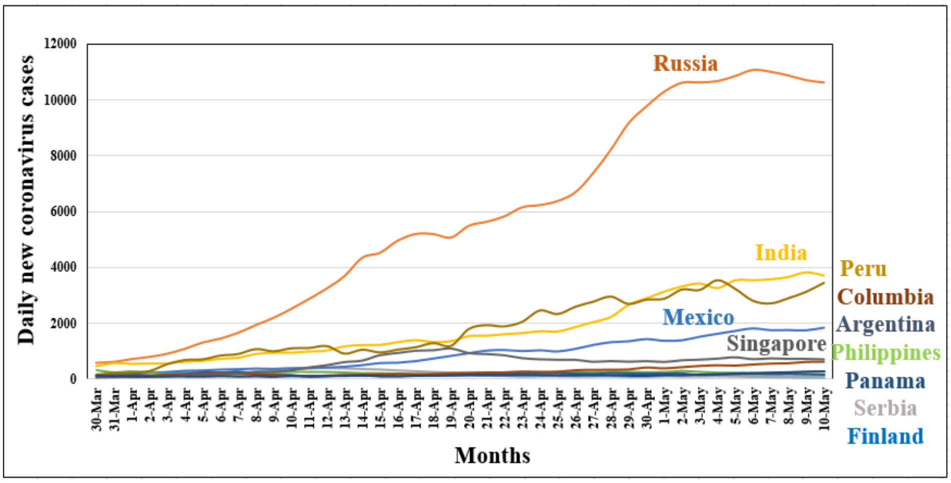


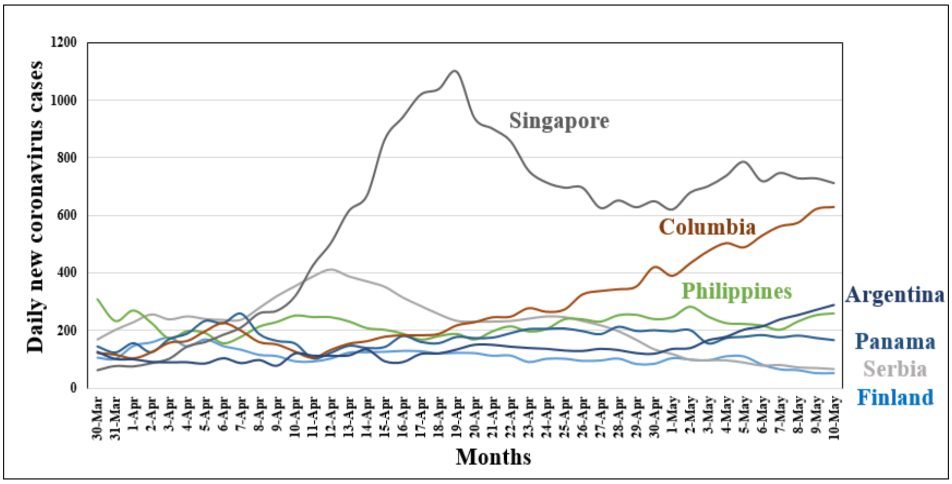
Figure 3: Variation of daily new coronavirus cases for 1 March to 10 May 2020.

2.5 Variability of the Spectrum of Daily New Coronavirus Cases of Countries Reaching the First 1,000 Coronavirus Cases

Figure 4 depicts the variability of the spectrum of daily new coronavirus cases for countries crossing 1,000 reported cases on 22-28 March. Figure 4(a) shows the growth of the spectrum for India and Peru varying very closely to each other with non-linear rise from 25 March 2020, whereas Russia follows a non-linear exponential increase in the spectrum to reach beyond the critical phase of the outbreak. The spectrum of new cases for Mexico and Colombia shows linear increase in its growth to reach towards the critical phase of the outbreak, whereas the spectrum for Singapore, Argentina, Panama and Philippines shows decreasing trend of new cases after reaching a maximum peak and approaching towards the recovery of the outbreak. Figure 4(b) shows the spectrum of new cases for Serbia and Finland almost approaching towards the recovery stage due to continuous decrease for more than 30 days, whereas Singapore, Argentina, Panama and Philippines were still struggling to control the number of new cases.



(a)



(b)

Figure 4: Variation of daily new coronavirus cases from 1 March to 10 May 2020.

2.6 Variability of the Spectrum of Daily New Coronavirus Cases for Countries Reaching the First 5,000 Coronavirus Cases

Figure 5 depicts the variation of the spectrum of daily new coronavirus cases for countries crossing first 5,000 cases on 3-5 April 2020. The variation of the spectrum from 30 March to 10 May 2020 shows steep exponential rise of daily new cases for Russia, India and Peru, reaching beyond the critical conditions of the outbreak. Linear decrease of the spectrum is observed after attaining its peak for Ireland, Poland, Japan, Denmark, Czech Republic, Romania and Malaysia, reaching the stage of recovery of the outbreak.

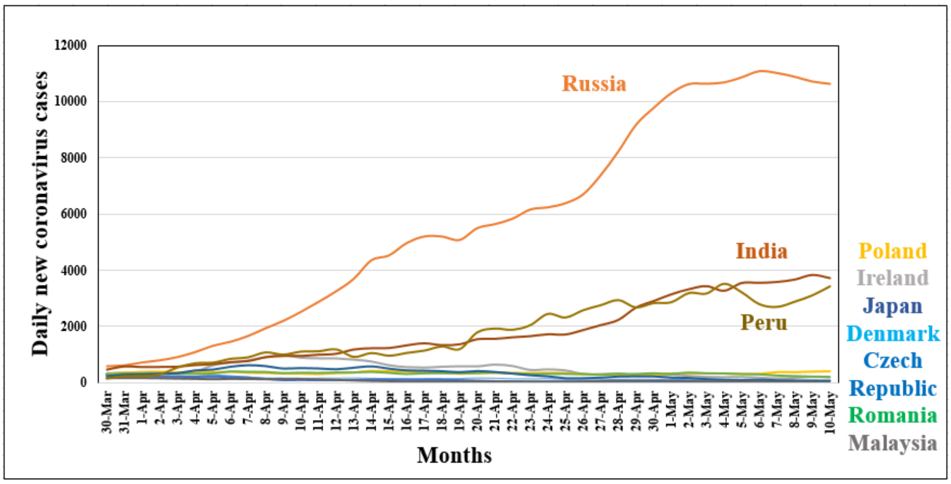
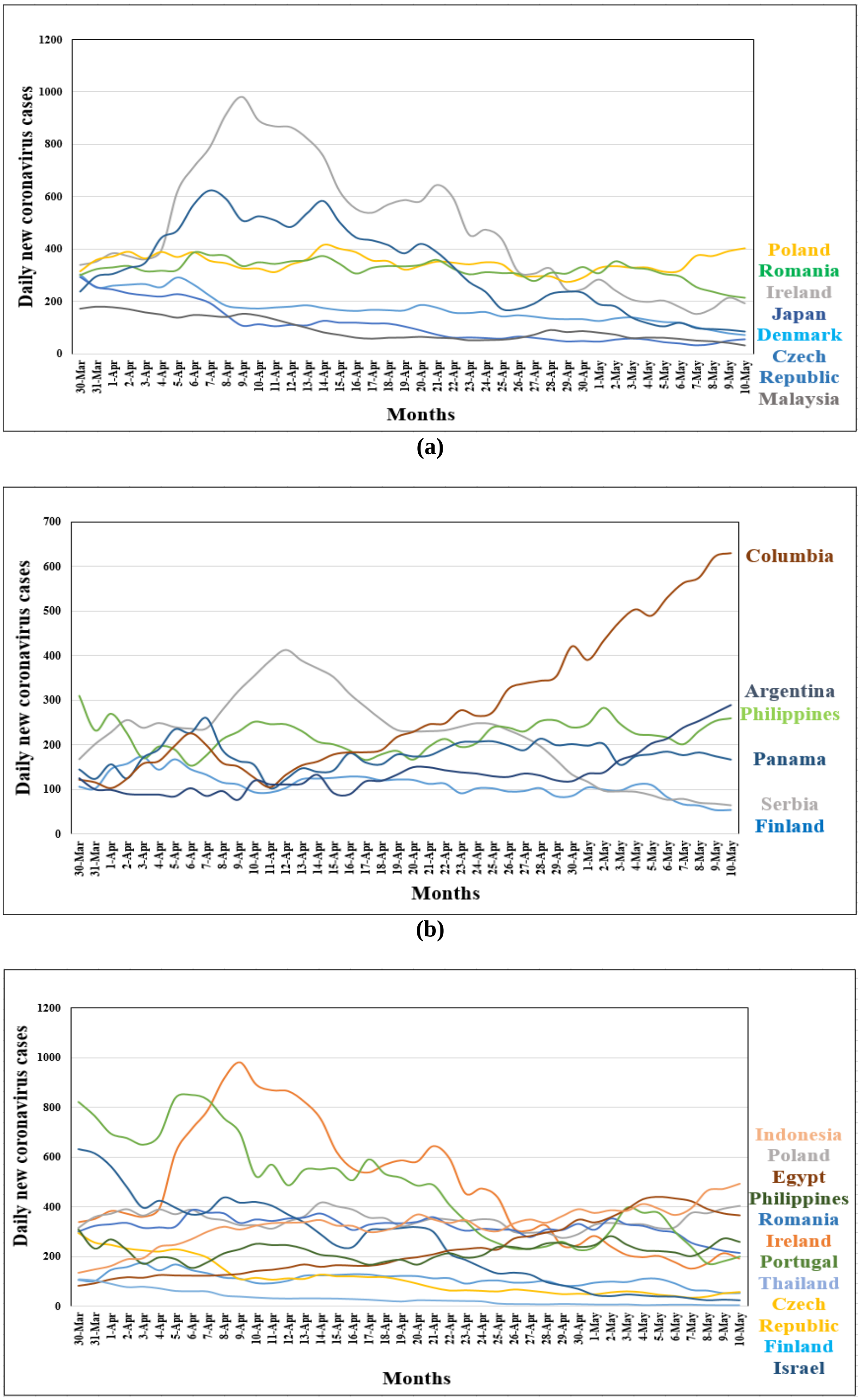


Figure 5: Variation of daily new coronavirus cases from 30 March to 10 May 2020.

2.7 Spatial Big Data Predictive Analysis Using Knowledge Classifier for Recovery Trend

The spectrum of 5-day moving average of daily new coronavirus cases for different countries acts as knowledge classifier for predicting the trend of the spectrum of coronavirus for any geographic locations due to the expected similarity in the spectrum. Figure 6(a) depicts the similar characteristics of the spectrum of coronavirus for Ireland and Japan. The spectrum increased exponentially and attained the peak at approximately 1,000 and 600 cases, prior to decreases following a log-normal spectrum distribution. Similar characteristics of the spectrum are observed for Malaysia, Denmark and Czech Republic, attaining peak value between 200 and 300 cases, prior to non-linear decrease for more than 45 days. Figure 6(b) depicts the spectrum of coronavirus cases for Serbia and Finland, attaining peak values of 400 and 200 cases respectively, prior to decrease following a log-normal spectrum distribution. Figure 6(c) depicts similar spectrum distribution model for Ireland, Portugal and Israel after attaining maximum value of daily new cases of approximately 1,000, 800 and 600 respectively, whereas Finland, Czech Republic and Thailand followed a similar spectrum of coronavirus after attaining peak values ranging from 100 to 300 cases. These spectrum distributions of attaining towards its recovery from the outbreak can be used as knowledge classifier for developing predictive spectrum distribution models for predicting the trend of coronavirus spectrum for other geographical locations.



(a)

(b)

(c)

Figure 6: Coronavirus spectrum distribution models as a knowledge classifier.

3. SPATIAL BIG DATA PREDICTION MODELS FOR DIFFERENT STAGES OF THE CORONAVIRUS OUTBREAK

The empirical models developed based on the coronavirus spectrum to successfully control the outbreak and recover from the first outbreak can be used for any other geographical locations for predicting the recovery from the expected spectrum by extrapolating the models developed by polynomial regression techniques. The envelope of the spectrum of empirical models shall vary with the different phases, such as Beyond the Critical Stage (Phase 5), Critical Stage (Phase 4), Stabilising Stage (Phase 3), Safe Stage (Phase 2), Recoverable Stage (Phase 1), and Complete Recovery Stage (Phase 0) of controlling the outbreak.

3.1 Coronavirus Spectrum Model of the Complete Recovery Stage

Figure 7 depicts the spectrum distribution model for complete recovery from the first outbreak of coronavirus based on the spectrum of daily new coronavirus cases in Serbia and Japan using polynomial regression analysis of the daily new coronavirus cases from 10 March to 15 May 2020, where the maximum values of daily new coronavirus cases were 300 and 500 respectively.

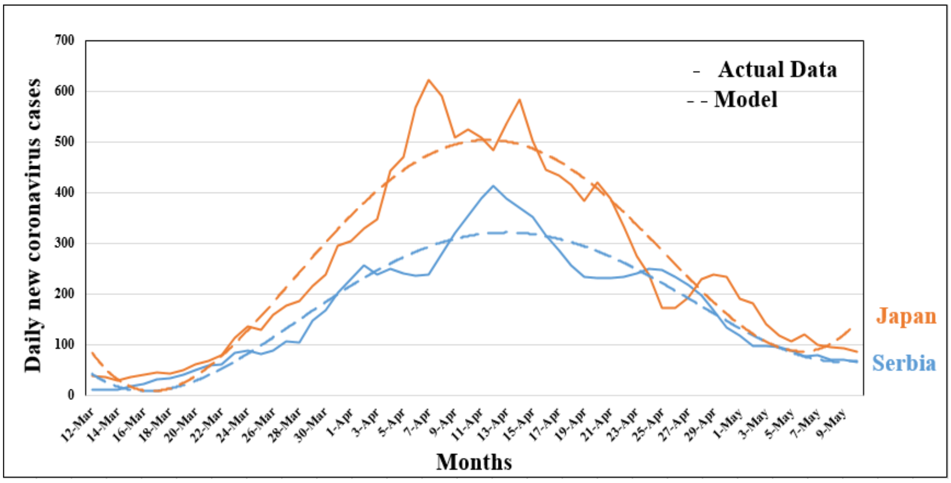
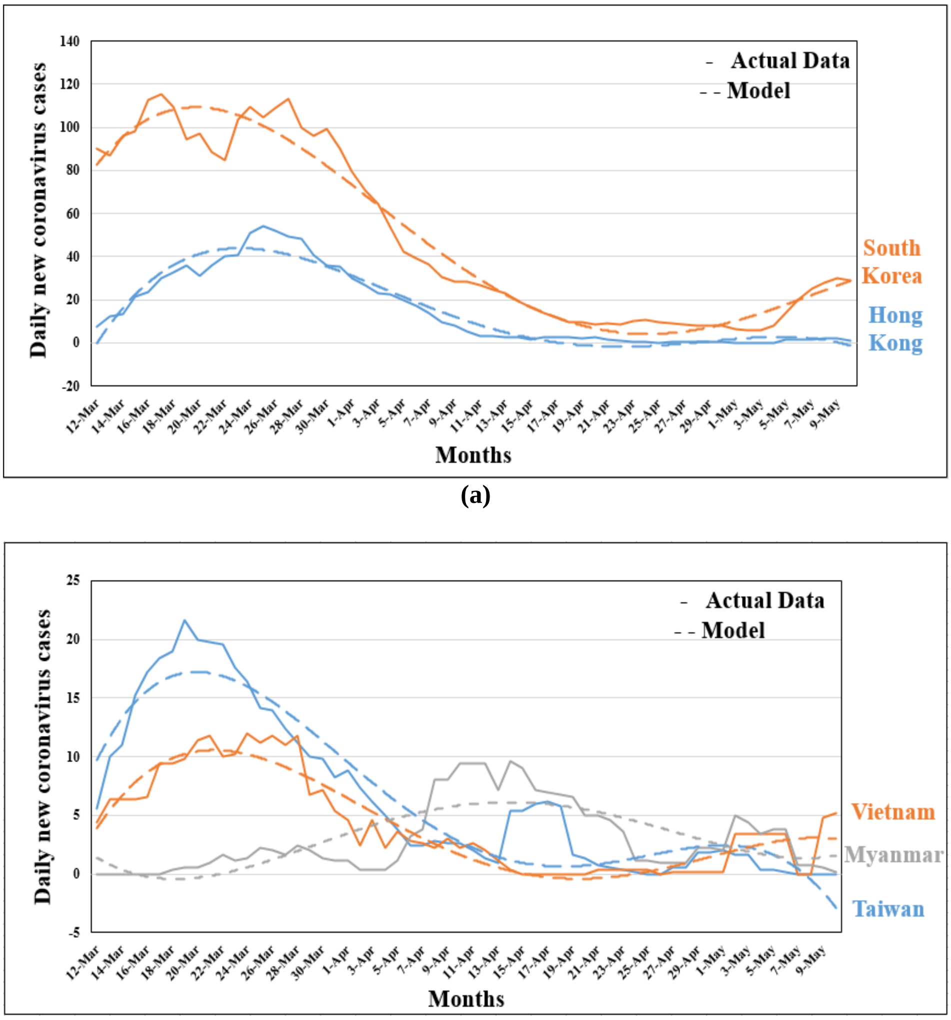


Figure 7: Coronavirus spectrum model for the complete recovery stage.

3.2 Coronavirus Spectrum Model for the Recovery Stage

Figure 8 shows the spectrum distribution model for moving towards recovery from the first outbreak of coronavirus for South Korea, Hong Kong, Taiwan, Vietnam and Myanmar, where the maximum values of daily new cases reached 110, 45, 20, 10 and 6 respectively. The coefficients of the spectrum distribution model vary based on the maximum values of daily new cases for the case of Japan, Serbia, South Korea, Hong Kong, Taiwan and Vietnam. These models can be used as a knowledge classifier for predicting the trend of the recovery cycle, where the maximum values of daily new cases vary from 6 to 500.



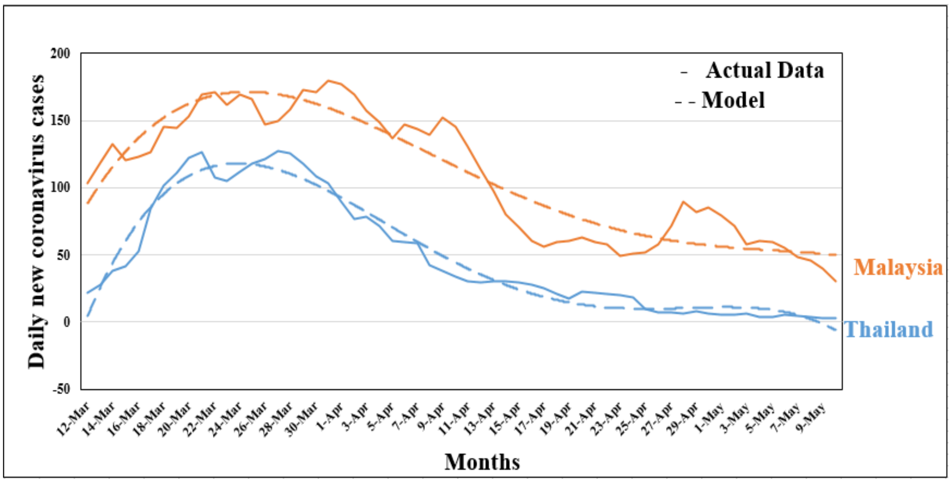
(a)

(b)

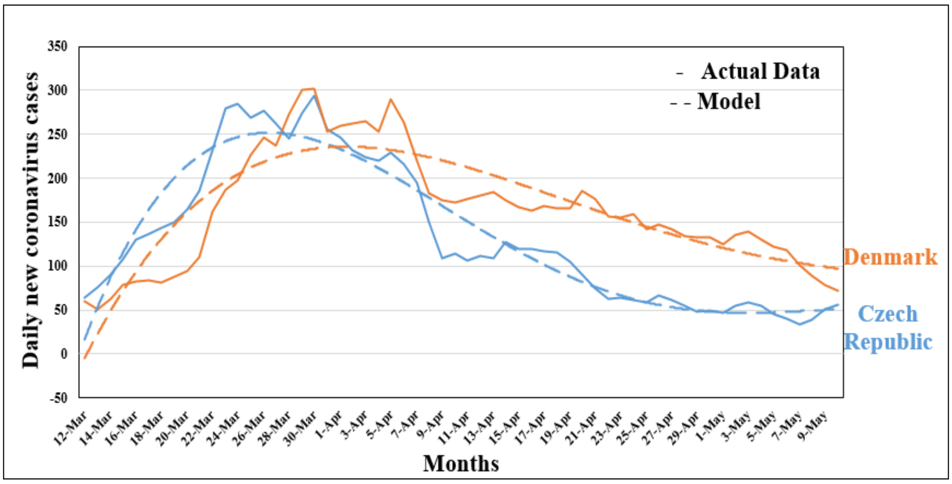
Figure 8: Coronavirus spectrum model for the recovery stage.

3.3 Coronavirus Spectrum Distribution Model for the Safe Stage

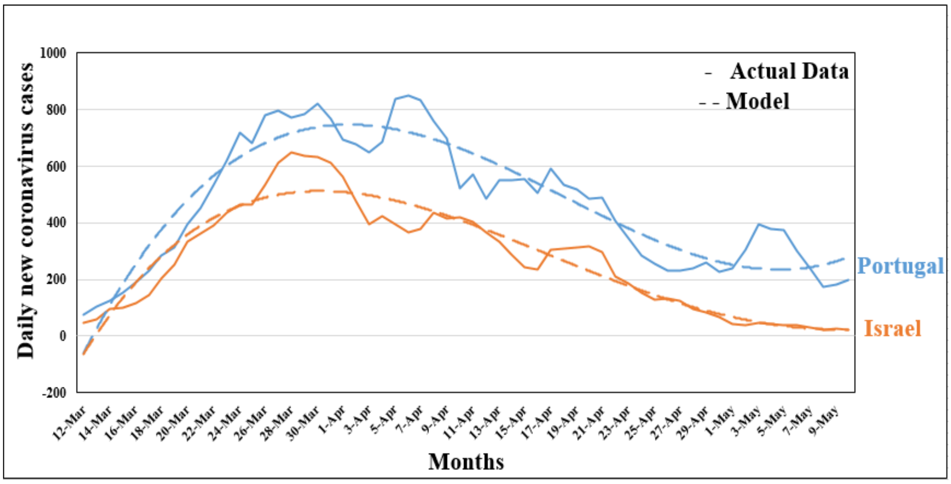
Figure 9 shows the spectrum distribution models for successfully controlling daily new coronavirus cases for the safe stage for Thailand, Malaysia, Czech Republic, Denmark, Israel, Portugal, Ireland and Singapore, where the maximum values of daily new cases reached to 120, 180, 230, 250, 600, 800, 900 and 1,000 cases respectively. It is found that the coefficients of spectrum distribution models also vary with the maximum values of daily new cases. These spectrum models show exponential decrease of the coronavirus spectrum after reaching the maximum value prior to moving towards recovery from the first outbreak, whereas a second peak of the first outbreak is observed for Singapore after attaining the first peak of daily new cases.


Actual Data
Model
Daily new coronavirus cases
200
150
100
50
0
-50
Malaysia
Thailand
12-Mar
14-Mar
16-Mar
18-Mar
20-Mar
22-Mar
24-Mar
26-Mar
28-Mar
30-Mar
1-Apr
3-Apr
5-Apr
7-Apr
9-Apr
11-Apr
13-Apr
15-Apr
17-Apr
19-Apr
21-Apr
23-Apr
25-Apr
27-Apr
29-Apr
1-May
3-May
5-May
7-May
9-May
Months

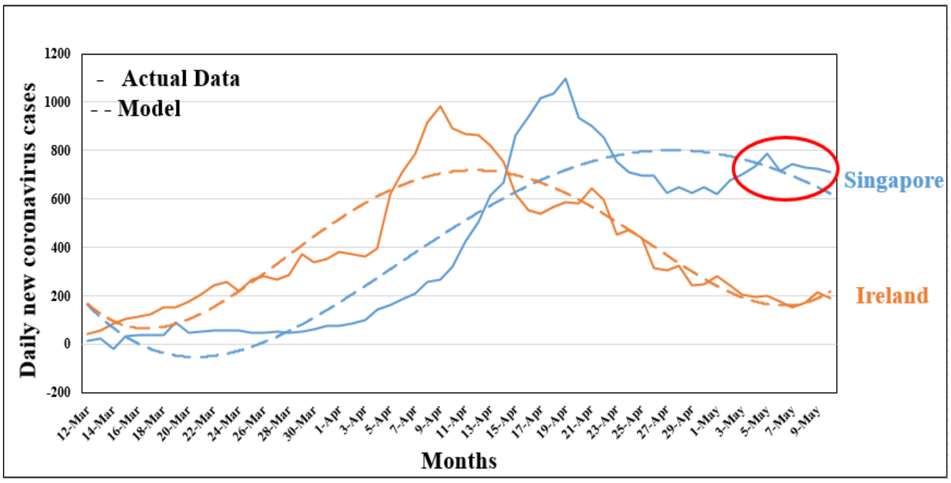
(a)


Actual Data
Model
Daily new coronavirus cases
350
300
250
200
150
100
50
0
-50
Denmark
Czech Republic
12-Mar
14-Mar
16-Mar
18-Mar
20-Mar
22-Mar
24-Mar
26-Mar
28-Mar
30-Mar
1-Apr
3-Apr
5-Apr
7-Apr
9-Apr
11-Apr
13-Apr
15-Apr
17-Apr
19-Apr
21-Apr
23-Apr
25-Apr
27-Apr
29-Apr
1-May
3-May
5-May
7-May
9-May
Months

(b)


Actual Data
Model
Daily new coronavirus cases
1000
800
600
400
200
0
-200
Portugal
Israel
12-Mar
14-Mar
16-Mar
18-Mar
20-Mar
22-Mar
24-Mar
26-Mar
28-Mar
30-Mar
1-Apr
3-Apr
5-Apr
7-Apr
9-Apr
11-Apr
13-Apr
15-Apr
17-Apr
19-Apr
21-Apr
23-Apr
25-Apr
27-Apr
29-Apr
1-May
3-May
5-May
7-May
9-May
Months

(c)

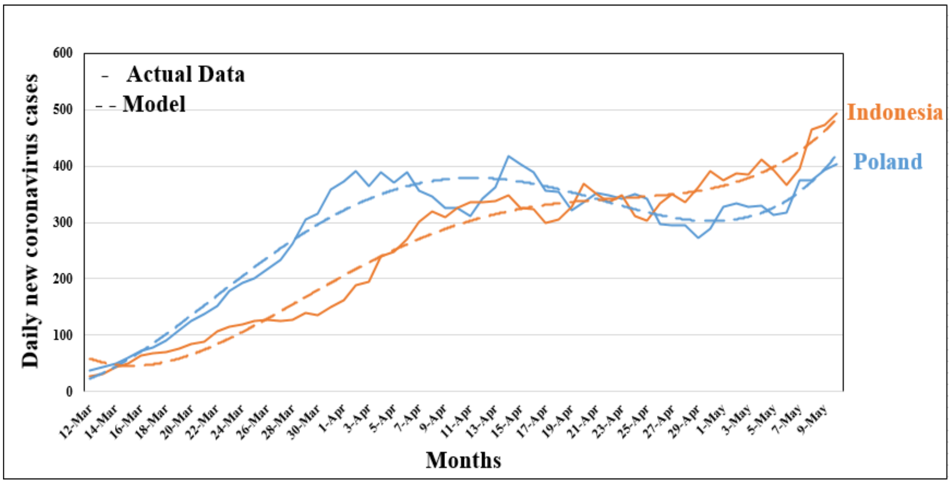


(d)

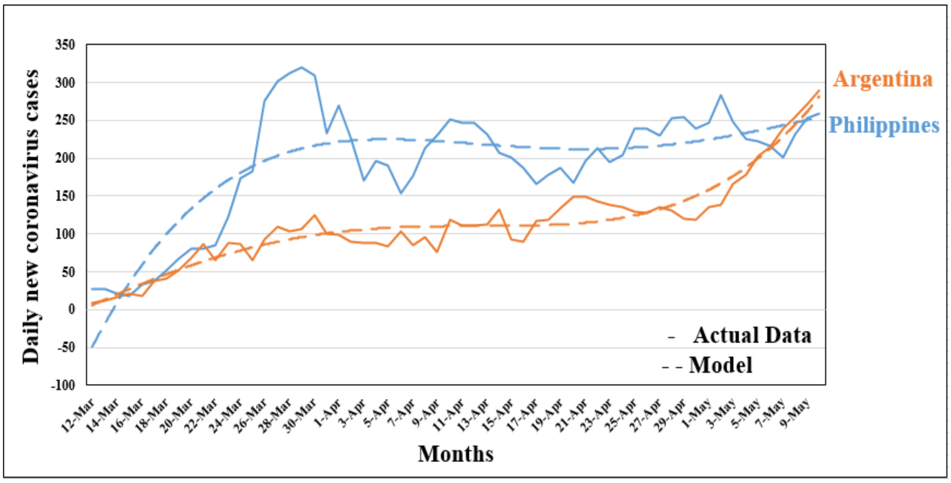
Figure 9: Coronavirus spectrum models for the safe stage.

3.4 Coronavirus Spectrum Distribution Model of the Stabilising Stage

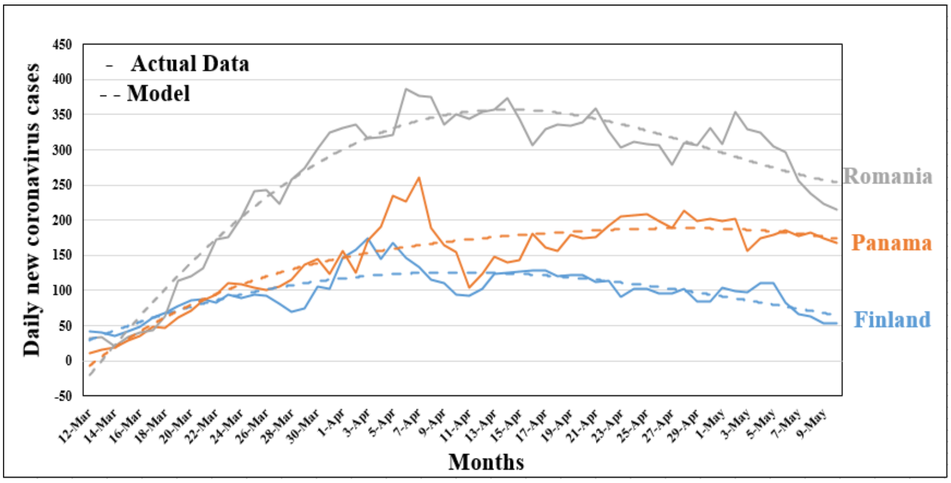
Figure 10 shows the oscillatory spectrum distribution models for Indonesia, Poland, Argentina, Philippines, Romania, Panama and Finland prior to reaching the stabilising stage for the first outbreak, where the maximum values of daily new cases vary from 100 to 500 cases. The spectrum distribution model for Indonesia and Poland moves towards the second higher peak of first outbreak with Gaussian distribution, as depicted in Figure 10(a), while the spectrum distribution models for Philippines and Argentina shows increasing exponential trend after attaining the stabilising stage to control the outbreak, as depicted in Figure 10(b). The spectrum distribution models for Romania, Panama and Finland follows continued exponential decrease of the cases after attaining the stabilising stage to control the outbreak and move towards the recover cycle of the outbreak as depicted in Figure 10(c). The oscillation in the coronavirus spectrum distribution becomes dominant for higher values of daily new cases and leads to moving towards the critical stage of the outbreak.



(a)



(b)

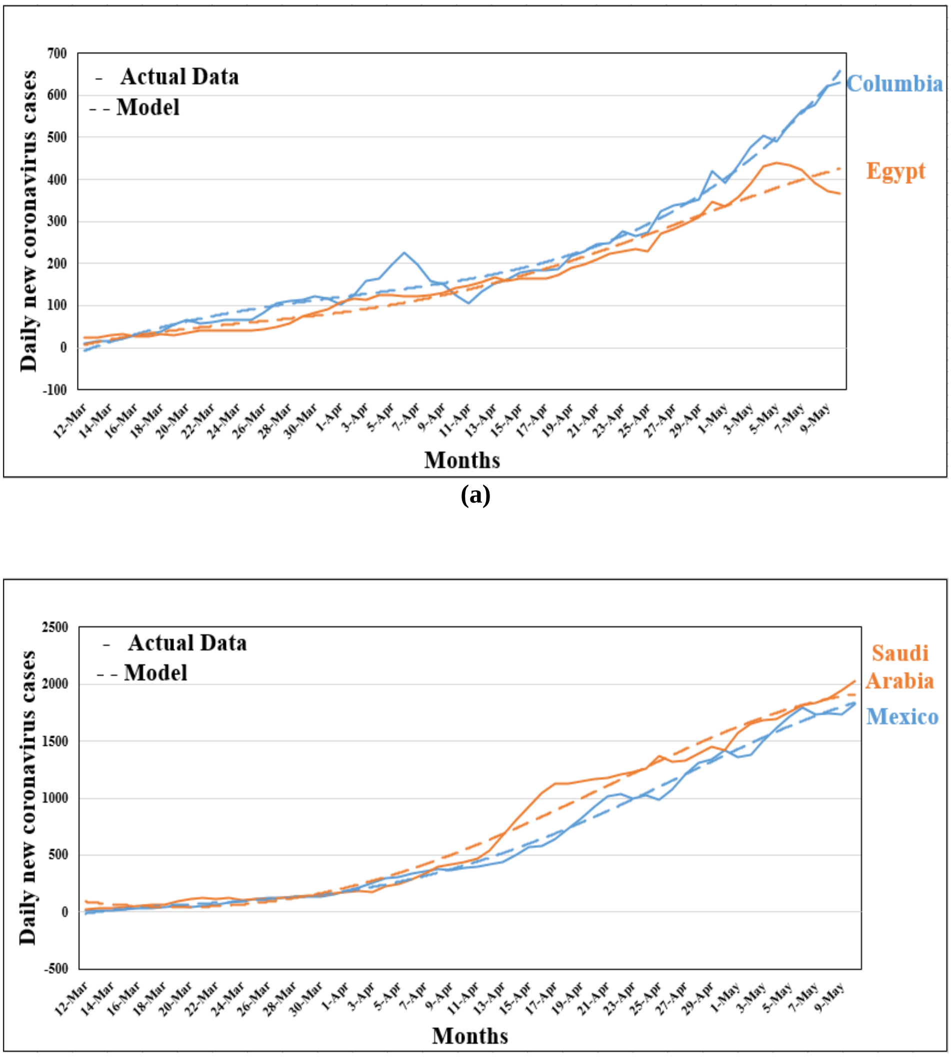


(c)

Figure 10: Coronavirus spectrum models for the stabilising stage.

3.5 Coronavirus Spectrum Distribution Model for the Critical Stage

Figure 11 depicts the positive exponential spectrum distribution model with varying daily new coronavirus cases for Egypt, Colombia, Mexico and Saudi Arabia. The spectrum shows the continuity of exponential increase of daily new cases to reach the critical stage of the outbreak. The trend of the recovery spectrum from the first outbreak can be predicted from the analysis of big data predictive model based on attaining the peak of the spectrum and continued decrease of the spectrum thereafter, which is not depicted for Egypt, Colombia, Mexico and Saudi Arabia.



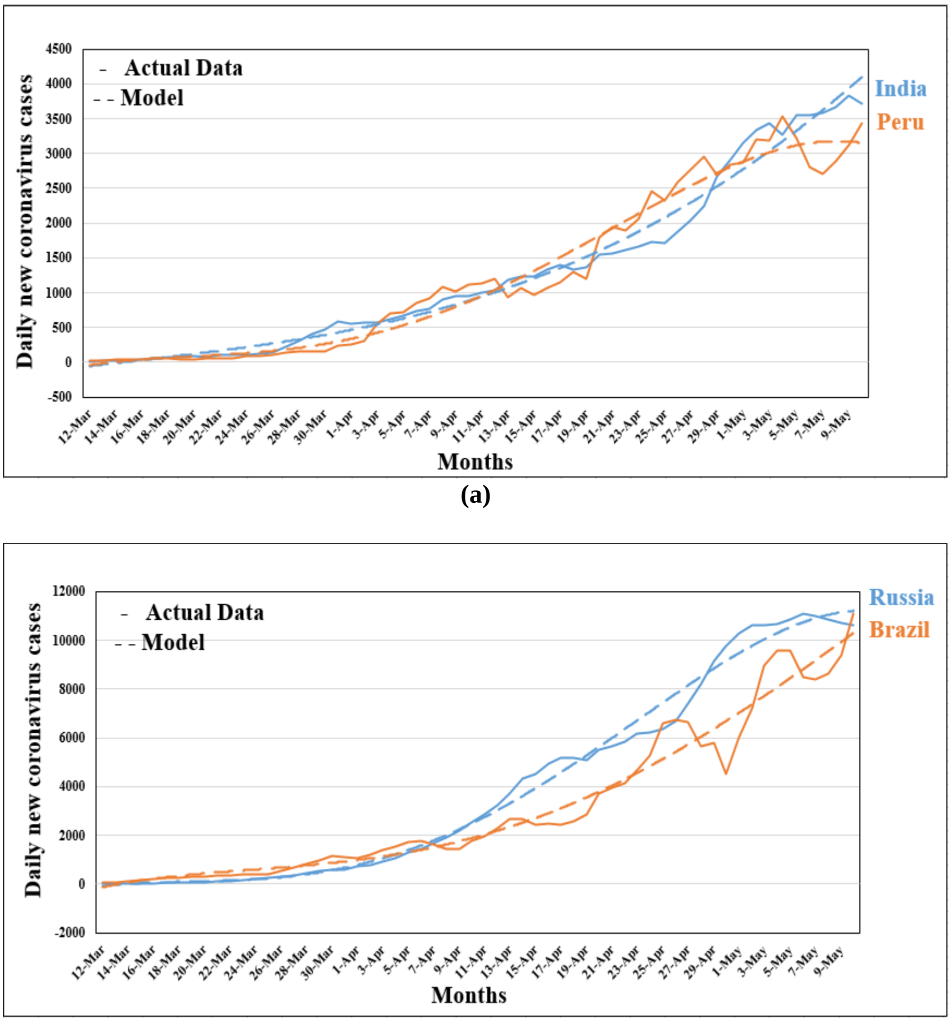
(a)

(b)

Figure 11: Coronavirus spectrum models for the critical stage.

3.6 Coronavirus Spectrum Distribution Model for Beyond the Critical Stage

Figure 12 depicts the coronavirus spectrum distribution models for beyond the critical stage for India, Peru, Russia and Brazil, with varying slopes of exponential increase of more than 2,000 daily new cases. The slope of exponential spectrum initially varies by more than 20% of increase of daily new cases until reaching 500 daily new cases, thereafter the spectrum varies with more than 10% increase of daily new cases until reaching 1,000 daily new cases. The spectrum reaches 2,000 daily new cases with more than 7% increase of daily new cases after attaining 1,000 daily new cases. Figure 12 depicts daily 5% increase of new coronavirus cases after attaining 2,000 daily new cases, until reaching 8,000 daily new cases for beyond the critical stage of the outbreak.



(a)

(b)

Figure 12: Coronavirus spectrum models for beyond the critical stage.

3.7 Coefficients of Coronavirus Spectrum Distribution Models for Different Stages

The polynomial spectrum distribution models developed for predicting the trend of the spectrum and predicting recovery from the outbreak, based on 5-day moving average of coronavirus data for different countries with varying maximum coronavirus spectrum and its characteristics for different stages of the spread of the outbreak is expressed as:

$$y(x) = A_0 + A_1 x + A_2 x^2 + A_3 x^3 + A_4 x^4 \qquad (1)$$

where x is the number of days, $y(x)$ is daily new coronavirus cases, and A_0, A_1 , A_2 , A_3 and A_4 are the coefficients for different stages of the outbreak for different countries. Table 2 describes the coefficients of different models developed for different countries in the present work, which are Beyond the Critical Stage (Phase 5), Critical Stage (Phase 4), Stabilising Stage (Phase 3), Safe Stage (Phase 2), Recoverable Stage (Phase 1), and Complete Recovery Stage (Phase 0)for controlling the outbreak.

Table 2: Coefficients of spatial big data predictive models.

Stages	Country	Model Coefficients					Maximum Spectrum
		A_0	A_1	A_2	A_3	A_4	
Complete Recovery	Serbia	61.041	-22.28	2.6674	-0.0714	0.0006	412
	Japan	124.006	-45.493	5.1747	-0.1425	0.0011	623
Recovery	Vietnam	2.1623	1.8872	-0.1313	0.0029	-0.00002	12
	Taiwan	7.4031	2.5238	-0.1988	0.0047	-0.00004	22
	Hong Kong	-10.051	10.544	-0.6546	0.0137	-0.00009	54
	South Korea	74.429	8.9373	-0.6811	0.0138	-0.00009	115
Safe Stage	Ireland	219.82	-59.825	6.7487	-0.1845	0.0015	980
	Singapore	222.56	-64.114	4.414	-0788	0.0004	1097
	Israel	-140.13	80.599	-3.1442	0.0405	-0.0002	648
	Portugal	-151.73	95.967	-3.0798	0.0258	0.00001	851
	Czech Republic	-23.734	42.837	-2.1351	0.0367	-0.0002	295
	Denmark	-34.291	30.989	-1.1649	0.0159	-0.00008	302
	Malaysia	72.181	17.17	-0.935	0.0167	-0.0001	180
	Thailand	-19.553	25.68	-1.5332	0.0314	-0.0002	127
Stabilising Stage	Finland	22.512	7.1932	-0.1301	-0.00003	0.000007	174
	Panama	-21.145	14.613	-0.4402	0.0069	-0.00005	260
	Romania	-43.184	22.48	-0.2056	-0.0055	0.00007	387
	Argentina	-1.5349	7.4822	-0.0565	-0.0047	0.00008	>290; 110
	Philippines	-82.065	34.332	-1.3451	0.0214	-0.0001	>320; 240
	Poland	15.264	6.8053	1.0701	-0.0417	0.0004	>417; 390
	Indonesia	68.763	-11.488	1.594	-0.0425	0.0004	>493; 350
Critical Stage	Egypt	-0.1347	7.4195	-0.3679	0.0118	-0.0001	>440
	Colombia	-18.293	11.646	-0.3353	0.0041	0.00002	>630
	Mexico	-30.37	19.207	-1.3705	0.055	-0.0005	>1825
	Saudi Arabia	102.57	-13.131	0.5756	0.0212	-0.0003	>2025
Beyond the Critical Stage	Peru	-74.901	34.724	-3.0551	0.1405	-0.0014	>3525
	India	-80.4	23.035	-0.2979	0.0222	-0.00007	>3835
	Brazil	-240.5	121.11	-6.9816	0.2224	-0.0015	>11100
	Russia	-131.68	69.939	-8.0564	0.4085	-0.004	>11100

4. ANALYSIS

4.1 Spatial Big Data Spectrum Models as a Knowledge Classifier

Coronavirus spatial big data predictive spectrum models have been developed based on polynomial regression techniques using the spectrum of daily new coronavirus cases for different countries, for predicting future trends of the outbreak as well as a knowledge classifier to predict the coronavirus spectrums to new geographical locations. The predictive spectrum model for estimating the recovery period for beyond the critical stage is difficult without attaining the first peak of the spectrum, but the predictive models developed as explained in Table 2 can be applied as a knowledge classifier for predicting the spectrum trend for different stages of the outbreak. Figure 13 depicts the spectrum of 5-day moving average for daily new coronavirus cases (solid line) and the spectrum model (dash line) for Bihar and Delhi, India for the first 60 days of the outbreak. It shows the non-linear exponential rise of the spectrum of daily new cases for Delhi, where the spectrum reaches 100 daily new cases within 22 days from the first reported case, which further reduces after attaining 300 daily new cases prior to attaining the critical stage of the outbreak. It also shows very slow linear rise of the spectrum of daily new cases for Bihar, reaching 50 daily new cases after 50 days from the first reported case, prior to reach the stabilising stage of the outbreak. The coefficients of the spectrum models based on polynomial regression techniques using the spectrum of daily new coronavirus cases for Bihar and Delhi are described in Table 3.

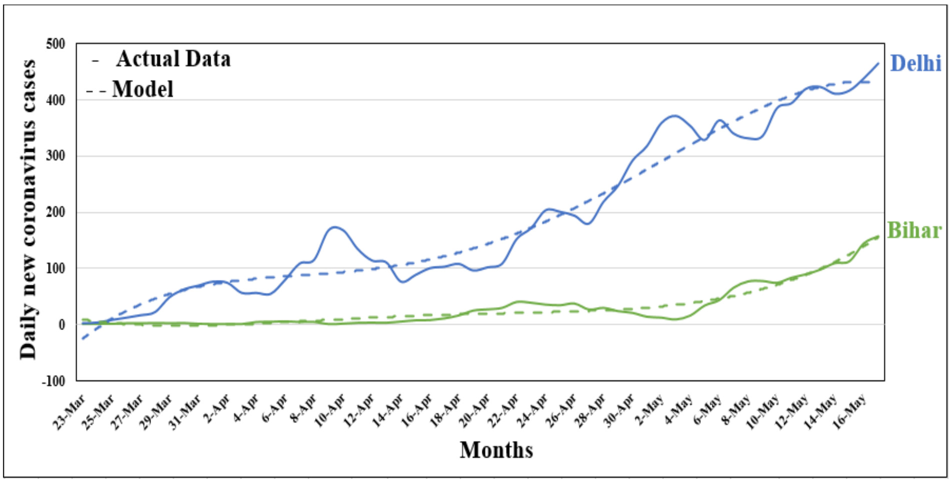
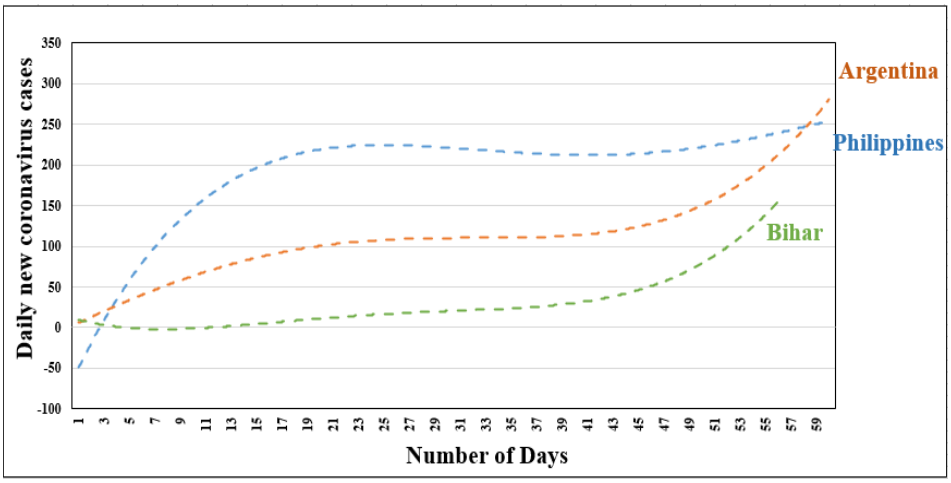


Figure 13: Spectrum models of daily new coronavirus cases for Bihar and Delhi.

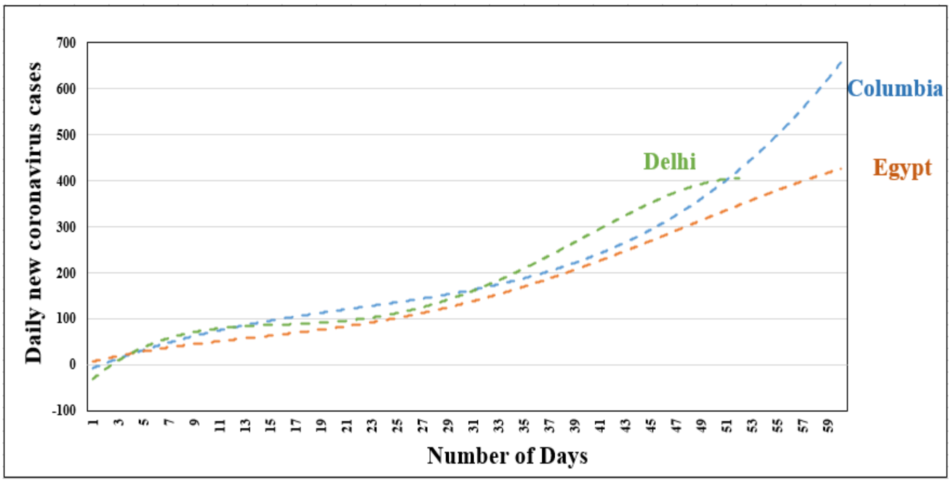
Table 3: Spectrum model coefficients for Bihar and Delhi.

State	Model Coefficients					Maximum Spectrum
	A_0	A_1	A_2	A_3	A_4	
Bihar	13.503	-4.6193	0.4375	-0.0134	0.0001	156.6
Delhi	-46.134	23.131	-1.5341	0.0444	-0.0004	465.2

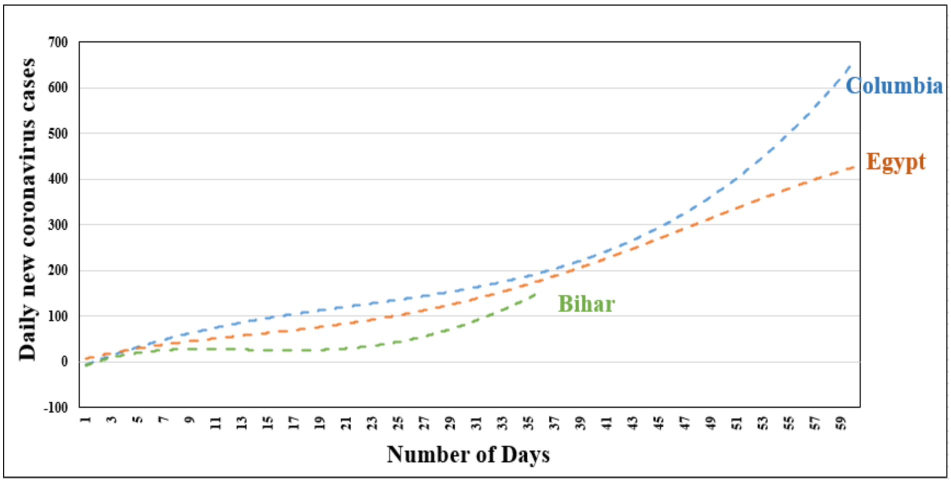
Predictive spectrum models of other countries developed as knowledge classifier have been used to predict the trend of daily new cases based on similarity analysis and matching of the spectrum model of Bihar and Delhi. Figure 14 (a) & (b) depicts the similarity analysis of the spectrum model of daily new coronavirus cases for Bihar and Delhi with the predictive spectrum models of other. The spectrum model of daily new cases for Bihar follows a similar trend of the predictive spectrum models for Argentina and Philippines after 25 days of the outbreak due to non-significant rise of daily new cases, whereas the spectrum model of daily new cases for Delhi follows a similar trend of the predictive spectrum models for Colombia and Egypt from the beginning of the outbreak. Figure 14(c) & (d) further depicts the matching analysis of the spectrum models of daily new cases for Bihar and Delhi with predictive spectrum models for predicting the future trends of the spectrum of the outbreak. Figure 14(c) shows the spectrum model for Bihar with significant number of daily new cases matching with predictive spectrum models for Columbia and Egypt, resulting into shift of the spectrum of Bihar. Thus, the predictive spectrum models for Columbia and Egypt can be used as a knowledge classifier to predict the trends of the spectrum of daily new cases for Bihar. Figure 14 (d) shows the spectrum model for Delhi with significant number of daily new cases matching with the predictive spectrum models for Saudi Arabia and Mexico, which can be used for predicting the trends of the spectrum. The predictive spectrum models matched with the spectrum models of new geographic locations can be used to predict the trend of spectrums and help the decision maker to take suitable measures to control the outbreak.


Daily new coronavirus cases
350
300
250
200
150
100
50
0
-50
-100
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59
Number of Days
Argentina
Philippines
Bihar

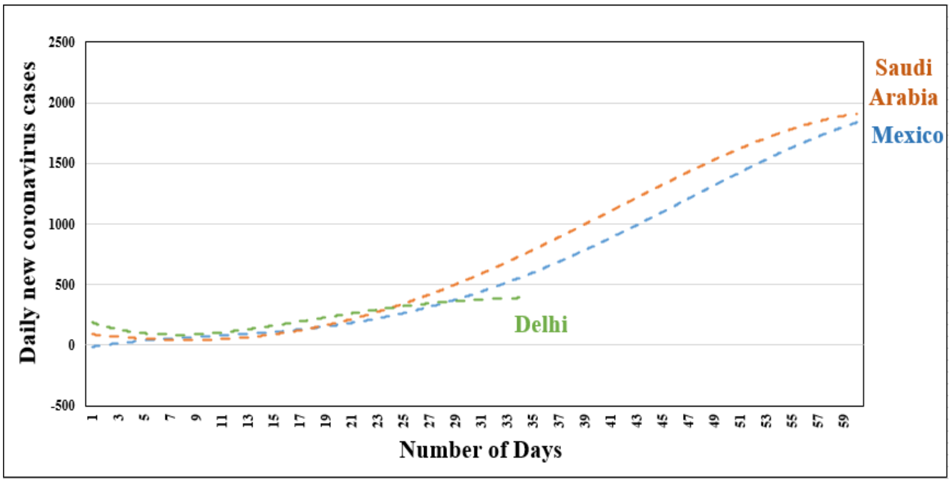
(a)


Daily new coronavirus cases
700
600
500
400
300
200
100
0
-100
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59
Number of Days
Columbia
Delhi
Egypt

(b)


Daily new coronavirus cases
700
600
500
400
300
200
100
0
-100
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59
Number of Days
Columbia
Egypt
Bihar

(c)



(d)

Figure 14: Coronavirus predictive spectrum model for Bihar and Delhi.

4.2 Impact of Latitude on Mortality from Coronavirus

In Jonathan *et al*. (2020), when mortality per million population was plotted against latitude for 120 countries located in the Northern and Southern Hemispheres based on the mortality data for 15 April 2020, it showed marked variation in mortality among different countries due to the potential impact of immune-modulating therapies, highlighting the importance of nutrition and vitamin D. In countries situated beyond the latitude of 35 °N and 35 °S, people do not receive sufficient sunlight to retain adequate vitamin D levels during the winter. It showed relatively low population mortality for countries situated at latitude below 35 °N and 35°S with correlation coefficient of 0.53 between mortality and latitude.

Figure 15 depicts the variation of mortality (death per 1 mil population) from coronavirus with latitude varying from 35 °N to 6 °S based on the mortality data from 15 April to 6 June 2020 for the southeast Asian region, which confirms relatively low mortality in the region as compared to the countries situated at latitude above 35°N, which is depicted in Figure 16. Significant variation of mortality from coronavirus is not observed for South Korea, Taiwan, Hong Kong, Thailand and Malaysia during these periods, while there is significant variation in mortality for India, Indonesia, Singapore and Philippines during the same period, which supports mortality as factor determining severity of the outbreak depending on their variability factor from 15 April 2020 as depicted in Figure 17. The variability factor of mortality from coronavirus varied by2.0, 3.0, 3.5 and 16.5 times for Singapore, Philippines, Indonesia and India respectively within 52 days between 15 April and 6 June 2020 as described in Table 4. This justifies that the variability factor of mortality for any country is a significant factor for determining the severity of the outbreak as compared to the mortality value due to coronaviruses depicted in Figures 16 and 18 for other countries situated at latitude below 60°N. Thus, higher value of variability factor of mortality from coronavirus for India, Peru, Columbia, Mexico and Russia indicates the severity of the outbreak, whereas very low population mortality from coronavirus observed in Figures 15 and 16 confirm earlier findings of low population mortality for the countries below the latitude of 35°N (Jonathan *et al*., 2020) along with Peru, Panama, Mexico, Portugal, Luxembourg and Ireland with high population mortality. Furthermore, lower variability factor of mortality from coronavirus during the same period for Singapore shows that the severity of outbreak in in the safe stage, whereas moderate variability factor of mortality for Philippines and Indonesia shows that the severity of outbreak is in the stabilising stage.

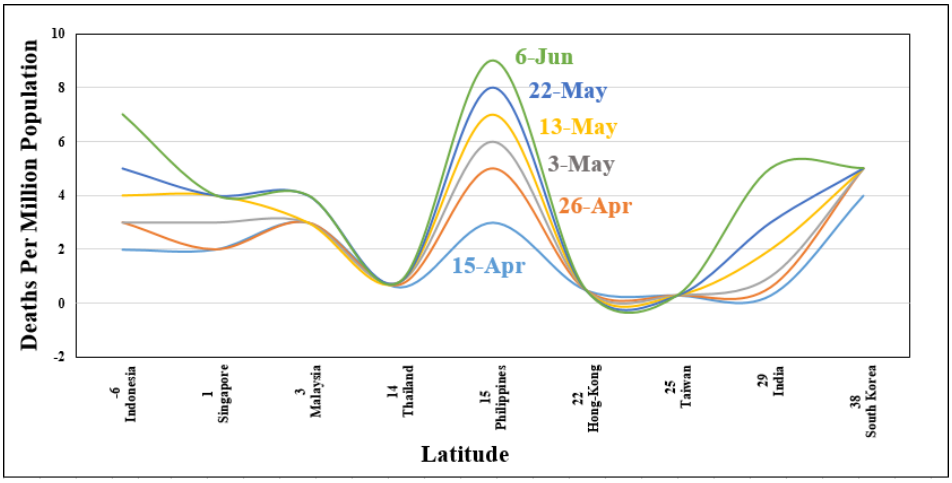


Figure 15: Variation of death per mil with latitude for the southeast Asian region.

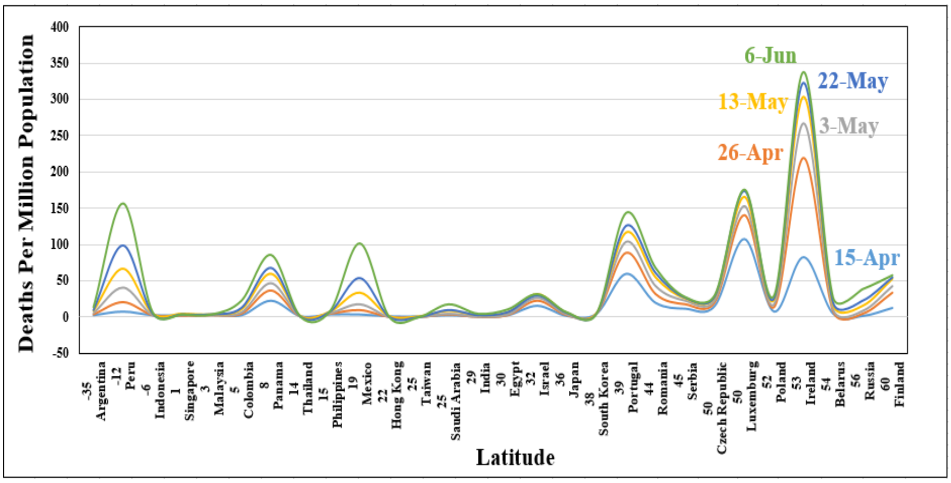
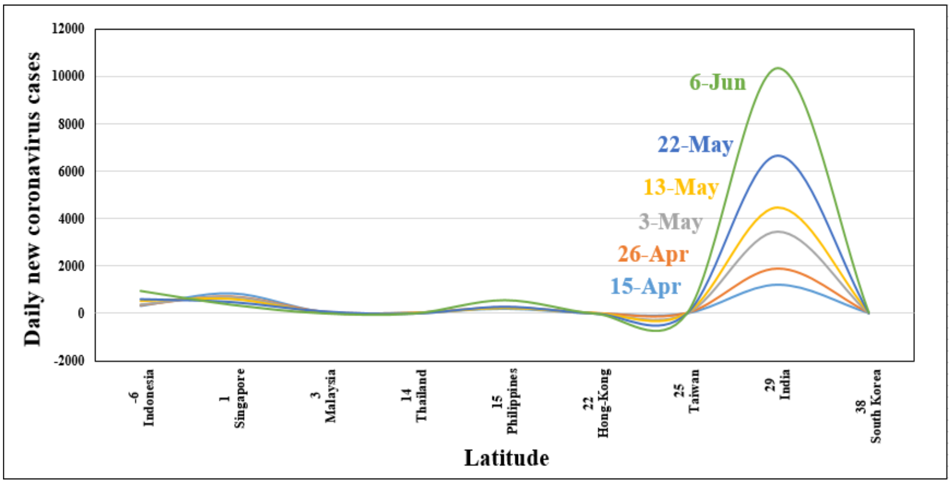
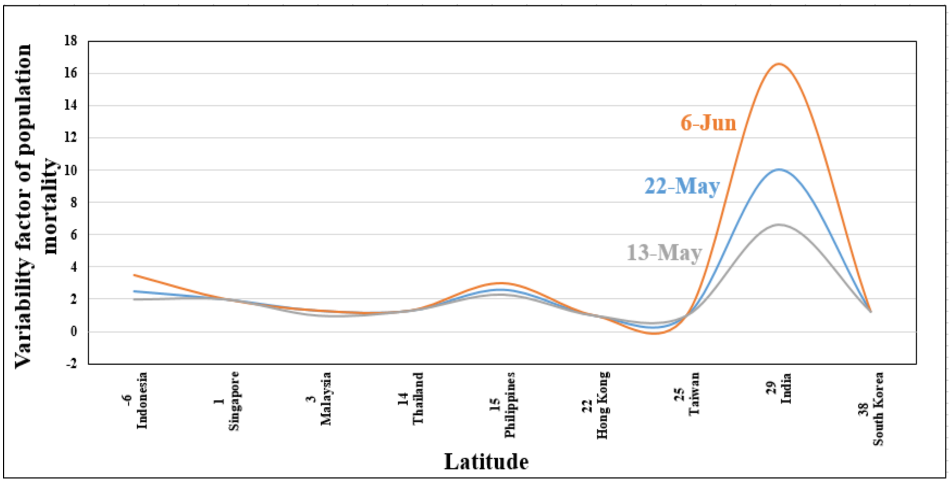


Figure 16: Variation of death per mil for 28 countries between latitude 60 °N and 35 °S.



(a)



(b)

Figure 17: (a) Daily new cases and (b) variation of variability factor of mortality with latitude for the southeast Asian region.

Table 4: Population mortality and variability factor of population mortality.

Country	**Latitude**	**Deaths per mil population**						**Variability factor of population mortality with respect to 15 April 2020 on**		
		15Apr	**26Apr**	**3 May**	**13May**	**22May**	**6Jun**	**13May**	**22May**	**6Jun**
Indonesia	-6	2	3	3	4	5	7	2	2.5	3.5
Singapore	1	2	2	3	4	4	4	2	2	2
Malaysia	3	3	3	3	3	4	4	1	1.3	1.3
Thailand	14	0.6	0.7	0.8	0.8	0.8	0.8	1.3	1.3	1.3
Philippines	15	3	5	6	7	8	9	2.3	2.6	3
Hong Kong	22	0.5	0.5	0.5	0.5	0.5	0.5	1	1	1
Taiwan	25	0.3	0.3	0.3	0.3	0.3	0.3	1	1	1
India	29	0.3	0.6	1	2	3	5	6.6	10	16.6
South Korea	38	4	5	5	5	5	5	1.25	1.25	1.25

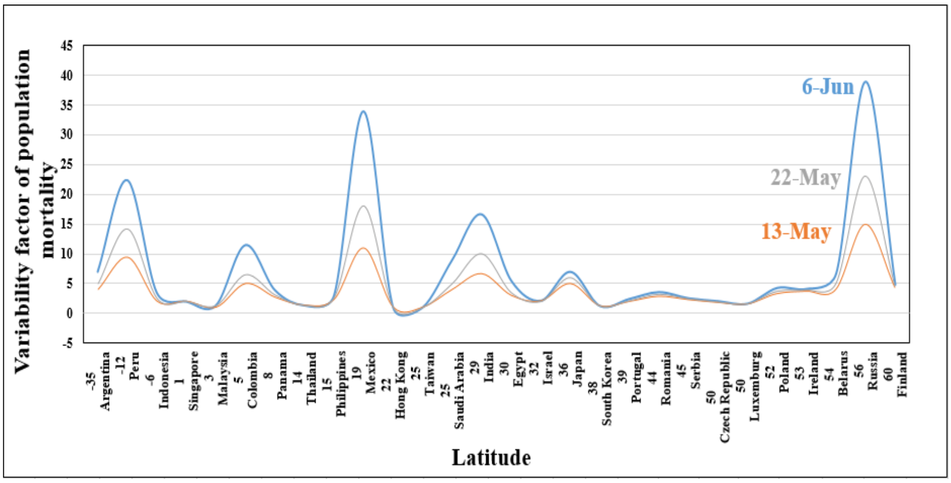


Figure 18: Variation of variability factor of population mortality for 28 countries between latitude 60 °N and 35 °S.

5. CONCLUSION

The present study of spatial big data predictive analysis of the daily new coronavirus case spectrum was carried out for 30 countries varying from latitude 60°N to 35°S of Hemisphere, which includes 11 countries in the southeast Asian region, which are India, Thailand, Malaysia, Indonesia, Singapore, Philippines, South Korea, Hong Kong, Taiwan, Vietnam and Myanmar, as well as 19 other countries, which are Russia, Brazil, Mexico, Peru, Egypt, Columbia, Saudi Arabia, Finland, Panama, Romania, Argentina, Poland, Ireland, Denmark, Portugal, Israel, Czech Republic, Japan and Serbia. Coronavirus spectrum distribution models have been developed for these countries by classifying the different stages of the outbreak, which are beyond the critical stage, critical stage, stabilising stage, safe stage, recovery stage and complete recovery stage. It is observed from the coronavirus spectrums that the outbreaks in India, Russia, Brazil and Peru reached beyond the critical stage, followed by Mexico, Egypt, Colombia and Saudi Arabia in critical stage. The spectrums for South Korea, Hong Kong, Vietnam, Taiwan, Japan and Serbia were in the process of recovering from the first outbreak, whereas Malaysia, Thailand and Singapore attained the safe stage to recover from the outbreak.

These predictive spectrum distribution models are used as a knowledge classifier for predicting the coronavirus trends for the Bihar and Delhi provinces in India, based on similarity and matching analysis of the spectrum model developed with 5-day moving average data of daily new cases. Furthermore, the predictive spectrum models as knowledge classifier can be used for predicting the recovery trends after attaining of the peak of the spectrum model based on daily new cases of new geographic locations. The high accuracy of prediction of trends by using knowledge classifier is observed due to initial matching with the spectrum model based on daily new cases.

The analysis of the impact of latitude on population mortality from coronavirus from 15 April to 6 June 2020 shows low mortality in the southeast Asian region below the latitude 35 °N, which validates to lower population mortality observed by Jonathan *et al.* (2020) based on mortality data of 120 countries below the latitude 60°N for 15 April 2020. It is also shown that the variability factor of population mortality for a geographic location is an appropriate determining factor for the severity of the coronavirus outbreak, whereas lower population mortality for countries between the latitude 35 °N and 35 °S supports the theory of vitamin D as a factor for determining the severity of the outbreak.

REFERENCES

BBC (2020). *Coronavirus: The world in lockdown in maps & charts/ Source*: *Oxford Covid-19 Government Response Tracker /BBC Research*. Available online at: https:// www.bbc.com/news/world-52103747 (Last access date: 10 July 2020).

Chenghu, Z., Fenzhen, S., Tao, P., Zhang, A., Yunyan, D., Bin, L., Zhidong, C., Juanle, W., Wan, Y., Yunqiang, Z., Ci, S., Jiechen, J.X., Li, F., Ma, T., Lili, J., Fugqin, Y., Jiewei, Y., Yunfeng, H. & Yilan, L. (2020). Covid-19: Challenges to GIS and Big Data. *Geogr.Sustainability*,**1:**77-87.

COVID-19 India (2020). *Covid-19 India Data.* Available on line at: http://www.covid19india.org (Last access date: 10 July 2020).

Coronavirus (2020). *Global Covid-19 Data for All Countries.* Available online at: https://www. worldmeters.info/coronavirus (Last access date: 10 July 2020).

George, M.V. & Rebecca, J. (2020). Covid-19 in India: moving from containment to mitigation. *Indian J. Med. Res.*, **151**: 136-139.

Jonathan, M.R., Sreedar, S., Eamon, L.& Rose, A.K. (2020). Low population mortality from COVID-19 in countries south of latitude 35 degree north supports vitamin D as a factor determining severity. *Aliment. Pharm. Therap.*,**51**: 1438-1439.

Panarese, A. & Shahini, E. (2020). Covid-19 and vitamin D. *Aliment. Pharm. Therap.*.51: 993-995.

Poonam, K.S. (2020). The research community must meet to coronavirus disease 2019 challenges. *Indian J. Med. Res.,* **151**: 116-117.

Prakash, K.T. (2020). Impact of health of people and wealth of nation. *Indian J. Med. Res.,***151**: 121-123.

Pranab, C., Nazia, N., Anup, A., Bebtosh, D., Sayantanu, B., Swarup, S., Gupta, N. & Raman, R.G. (2020). The 2019 novel coronavirus disease (COVID-19) pandemic: A review of the current evidences. *Indian J. Med. Res.,***151**: 147-159.

Rajesh, B. & Priya, A. (2020). Lesson learnt during the first 100 days of Covid-19 pandemic in India. *Indian J. Med. Res.,***151**: 387-391.

Rajesh, B. (2020). Public engagement is the key to contain Covid-19 Pandemic. *Indian J. Med. Res.,***151**: 118-120.

Tanzin, D., Sushma, C., Kapil, G., Preeti, P., Rajesh, S., Tripurari, K., Jain, S.K., Singh, S.K. & Jai, P.N. (2020). Responding to COVID-19 pandemic: why a strong health system is required. *Indian J. Med. Res.,***151**: 140-145.

WHO (World Health Organization) (2020). *Coronavirus Disease (COVID-2019) Situation Reports:* Available online at:https//www.who.int/emergencies/disease/novel-coronavirus-2019/situation-reports (Last access date:8 May 2020).

APPLICATION OF SPATIO-TEMPORAL EPIDEMIOLOGICAL MODELER (STEM) TO AN ANTHROPIC SMALLPOX DIFFUSION SCENARIO

Federico Baldassi[1,*], Orlando Cenciarelli[2], Andrea Malizia[3] & Pasquale Gaudio[1]

[1]Department of Industrial Engineering
[2]International CBRNe Master Courses
[3]Department of Biomedicine and Prevention
University of Rome Tor Vergata, Italy

[*]Email: federico.baldassi@gmail.com

ABSTRACT

The use of mathematical models to simulate the diffusion of biological agents represents an essential tool to understand the dynamics of epidemic spread. In particular, mathematical models can be applied to scenarios of deliberate release of biological warfare agents, e.g., during simulations of a terrorist attack, to evaluate their potential effects and to study possible strategies to implement effective countermeasures. In this paper, an open-source software named Spatio-Temporal Epidemiological Modeler (STEM) has been applied to a possible scenario of deliberate release of smallpox virus by an unknown terrorist group in Italy. By providing boundary conditions derived from the literature, and making conservative preliminary assumptions, it was possible to recreate a reference scenario for the voluntary diffusion of smallpox, while providing an insight into the application of user-friendly tools for the implementation of epidemiological models as a support for decision makers in the field of biosecurity.

Keywords: *Epidemiological modelling; smallpox; bioterrorism; countermeasures; Susceptible / Exposed / Infectious / Recovered (SEIR) model.*

1. INTRODUCTION

According to current opinions, the devastating effects of the spread of genetically modified microorganisms can be potentially more dangerous for the population than an attack with a nuclear fission weapon (Henderson, 1999; Olson & Shchelkunov, 2017). In addition, while nuclear weapons (as well as conventional ones) cause damage to both physical infrastructure and people, the use of a biological weapon would only have effects on the population, animals and even materials. The biological warfare agents (BWA), microorganisms used for warfare or terrorism purposes, are defined as organisms or substances derived directly from living organisms, whose use can result in deaths or serious injuries to people, animals and plants (Dudley & Woodford, 2002a). To be effective as a BWA, a biological agent must possess several characteristics, including high mortality and transmissibility rates, ability to survive for a long period in the environment, as well as resistance to methods of air, water and food purification (Cenciarelli *et al*., 2014). In order to carry out an effective bioterrorist attack, it is crucial that the attackers possess the means (e.g. vaccines and other tools for prophylaxes) not available for the victims. Moreover, modern bioengineering and molecular biology techniques can potentially improve the virulence and resistance to treatments for selected microorganism (Utgoff, 1993; Olson & Shchelkunov, 2017; Meyer *et al.*, 2020).

Mathematical models constitute an essential tool to assess the potential spread of epidemics. In the event of a deliberate release of biological agents, early estimation of potentially affected areas and individuals is essential. The risk assessment that can be implemented through these tools is therefore a valuable help to the emergency management. In this paper, we used the Spatio-Temporal Epidemiological Modeler (STEM) to understand and evaluate how a smallpox outbreak caused by a bioterrorist attack would spread, and also evaluate the application of specific physical countermeasures. We built a deterministic Susceptible / Exposed / Infectious / Recovered (SEIR) model to simulate the course of the epidemic, and evaluate the effectiveness of social distancing, shutting down air travel and preventing mixing of infected individuals across borders.

Assuming a reference scenario with 1,000 index cases and considering the implementation of the intervention 25 days after the attack, the model forecasts an epidemic of thousands of cases with an outbreak duration of 180 days.

STEM is an open source tool is designed to provide support to decision makers in the evaluation of how many people will be involved in an epidemic and in the visualization of the dynamics of the spread of infectious diseases, independently on whether they are the result of a natural epidemic or non-conventional human activity (Baldassi *et al*., 2015). In particular, STEM uses mathematical models of diseases (based on differential equations) to simulate the development or evolution of a disease in space and time.

2. POXVIRUS AS A BWA

Consciously or unconsciously, mankind has been using biological agents as lethal weapons for centuries. Microorganisms and biological toxins have been used from ancient times for direct attacks against the people. Among the several episodes recalled by history, the use of smallpox virus as a biological weapon was among the most sensational. In the 18th century, during the French-Indian war (1754-1767), the British colonial army used smallpox-contaminated blankets to disseminate the disease among native American tribes (Dudley & Woodford, 2002b); the smallpox outbreak killed more than 50% of the affected people and the recrudescence of the virus among the indigenous lasted for more than 200 years (Riedel, 2005; Cenciarelli *et al.*, 2013). More recently, it was reported (although not officially confirmed) that during the Cold War, several attempts for the genetic modification of the smallpox virus were carried out by Soviet scientists to create a new and more destructive BWA (Henderson, 2009).

Humans are the only known hosts of the smallpox virus; this aspect has allowed, through two global vaccination campaigns carried out by the World Health Organization (WHO), the total eradication of the pathogen (Wolfe *et al.*, 2007). The last natural case in the world was in Somalia in 1977, and at this date, the vaccination campaign was already stopped (Henderson, 2009).

The poxviruses (*Poxviridae*) comprise a family of enveloped DNA viruses that replicate within the cytoplasm of vertebrates and invertebrates' cells. Only members of the genus Orthopoxvirus, which includes smallpox, can infect humans. Smallpox is readily transmitted from person to person via saliva or nasal secretion droplets and contaminated objects (Moss, 2007; Sulaiman *et al.*, 2007; Cenciarelli *et al.*, 2013; Meyer *et al.*, 2020). The virus is highly infective and the mortality rate can reach 40%. During the infection, the smallpox virus enters the respiratory tract and spreads between mucous membranes, moving quickly into local lymph nodes (Breman & Henderson, 2002). An incubation period that lasts from seven to 17 days follows the infection of the host. The early symptoms are common to other diseases (e.g., cold and flu); this period is called the prodromal phase. During this stage, the mucous membranes in the throat and mouth are infected. The virus then invades the capillary epithelium of the dermis in the skin, leading to the development of lesions (Breman & Henderson, 2002). Currently, no treatments for smallpox infection are known; the therapy approach involves supportive care using antipyretic and anti-inflammatory treatments (Esposito & Fenner, 2001; Bhalla & Warheit, 2004; Cenciarelli *et al.*, 2013; Meyer *et al.*, 2020).

According to the severity of the disease, and to the possibility to be used as a bioterrorism agent, the U.S. Centers for Disease Control and Prevention (CDC) classified *Variola major* as a "Category A" biological agent (CDC, 2015); posing a global threat for biosafety and biosecurity (Henderson, 1999; Whitley, 2003). Due to the stop of the vaccination campaigns, a large part of the world population is no longer immune. Therefore, smallpox is once again an ideal candidate to be used as a biological weapon (Mahy, 2003). Following the official eradication, the remaining stocks of smallpox were collected by two WHO approved Biosafety Level 4 (BSL-4) laboratories. Particularly, 451 smallpox virus stocks are located at the CDC in Atlanta (USA) and 120 stocks in State Research Center of Virology and Biotechnology (SRCVB) in Koltsovo (RUS) (DHHS, 2009). Furthermore, the possibility that samples containing the DNA of the smallpox virus have been left, even unintentionally in remote parts of some laboratories cannot be completely ruled out, posing the risk that one or more of these samples might be recovered with malevolent intentions. The recent discovery of Variola virus in old specimens at the National Institutes of Health (NIH) in Bethesda, Maryland highlights this risk (CDC, 2014; Kaiser, 2014; Reardon, 2014).

3. METHODOLOGY

3.1 Spatio-Temporal Epidemiological Modeler (STEM)

STEM is an open-source software built on JAVA™ platform that allows to create spatial and temporal patterns of the spread of infectious diseases. The software comes with some existing compartment models, as Susceptible / Infectious (SI), Susceptible / Infectious / Recovered (SIR) and Susceptible / Exposed / Infectious / Recovered (SEIR) models pre-coded with both deterministic and stochastic engines, as well as a new model building framework that allows users to rapidly extend existing models or create entirely new models (Eclipse Foundation, 2020).

The STEM software integrates data from geographical information systems (GIS) from across the world, including information about national borders, populations, shared borders, highways, airports, etc. STEM organizes the world as a graph into a modular and hierarchical modeling structure; from bottom to top, this structure has three basic levels: graphs, models and scenarios. In particular, graphs represent spatial entities with defined shape and geospatial location, or carries information relevant for the scenario to be modeled (e.g., population, area). Models consist of at least one graph and a combination of other information regarding the population and disease states (compartments). In the end, scenarios may contain a variety of additional components, but they consist at least of a model and a time sequencer to easily create, run, and visualize the experiments (Eclipse Foundation, 2020).

3.2 Scenario

As described in Section 2, smallpox is among the most dangerous organisms that might be used by bioterrorists and is not widely available. An attack using this virus would involve relatively sophisticated strategies and would deliberately seek to sow public panic, social disruption, discredit official institutions, and shake public confidence in government.

In this paper, the authors created a possible scenario that considers a release of aerosolized smallpox agent in a public building placed in a big city in central Italy. The number people initially involved and heavily infected is 1,000. The airborne nature of smallpox allows the disease to spread quickly (O'Toole, 1999; Bozzette *et al.*, 2003; Olson & Shchelkunov, 2017).

The scenario designed in this work intends to provide food for thoughts and elements for discussion in the field of mathematical models as decision support tools not only for classical epidemiology, but also for planning and response in the framework of bioterrorism and biological related threats.

3.3 SEIR Epidemic Model

Several mathematical models can be used to fit epidemic data (Lavine *et al.*, 2008; Bachinsky & Nizolenko, 2013; Ndanguza *et al.*, 2013; Eclipse Foundation, 2020). In this work, smallpox epidemiological data was analyzed through a simple deterministic (continuous time) SEIR epidemic model (Hethcote, 2000). Almost all existing literature (Chowell *et al.*, 2004; Lekone & Finkenstädt, 2006; Legrand *et al.*, 2007) on smallpox epidemic prediction is based on this model. Typically, these types of models of the behavior of an infectious disease in a large population of people consider each individual as being in a particular epidemiological state. These states are often called compartments, and the corresponding models are called compartment models. The analyzed population (N) is classified into four epidemiological states: Susceptible (S), Exposed (E), Infectious (I) and Recovered (R). These compartments are described in Table 1, while the model flowchart is represented in Figure 1.

Table 1: Epidemiological states of the population considered for the SEIR epidemic model and compartments of *S*(*t*), *E*(*t*), *I*(*t*) and *R*(*t*) of the whole population size (*N*).

EPIDEMIOLOGICAL STATES	**POPULATION (*N*)**	**DESCRIPTION**
Susceptible (*S*)	*S*(*t*)	Susceptible individuals at the time *t*
Exposed (*E*)	*E*(*t*)	Exposed individuals at the time *t*, not yet infectious; incubation period of $\mathbf{1/\mathcal{E}}$ days
Infectious (*I*)	*I*(*t*)	Infected and infectious individuals at the time *t*; they move to the *R* class at the per-capita rate $\mathbf{1/\gamma}$
Recovered (*R*)	*R*(*t*)	Individuals who recovers from the disease or die as consequence of the disease at the time *t*

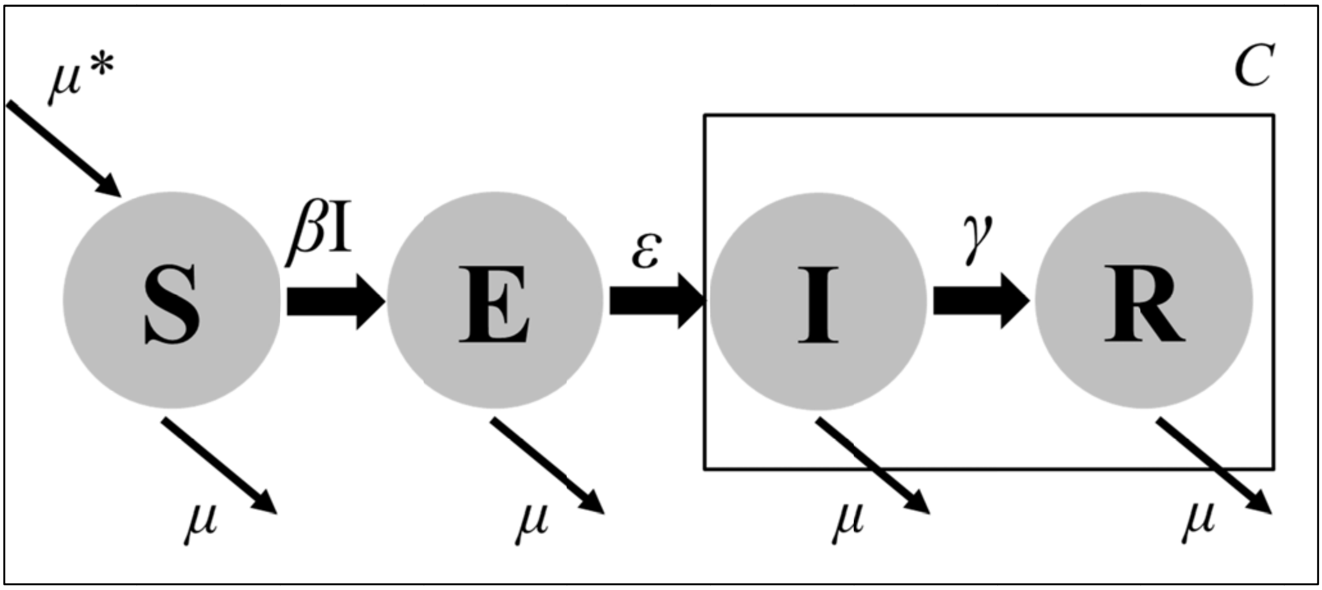


Figure 1: SEIR compartment model. This model shows the SEIR epidemiological states, the transmission rate (*β*), the incubation rate (*ε*) , the recovery rate (*γ*), the population birth rate (*μ), and the population death rate (*μ*). *C* is not a compartment but is a cumulative number of disease cases occurred (*I* + *R*).**

The transmission process is modeled by the system of the following nonlinear ordinary differential equations (Lavine *et al.*, 2008; Eclipse Foundation, 2020): Susceptible individuals at time *t* (1); exposed individuals at time *t* (2); infected and infectious individuals at time *t* (3); and individuals who recover from the disease or die as consequence of the disease at time *t* (4).

$$\frac{dS(t)}{dt} = -\frac{\beta S(t)I(t)}{N}\ \alpha R(t) + \mu\big(N - S(t)\big) \tag{1}$$

$$\frac{dE(t)}{dt} = \frac{\beta S(t)I(t)}{N - \varepsilon E(t)} - \mu E(t) \tag{2}$$

$$\frac{dI(t)}{dt} = \varepsilon E(t) - \gamma I(t) - \mu I(t) \tag{3}$$

$$\frac{dR(t)}{dt} = \gamma I(t) - \alpha R(t) - \mu R(t) \tag{4}$$

The model takes into account both the people infected by direct contact with an infected person, and the people infected by indirect contact with the equation $\beta S(t)I(t)/N$. Individuals who pass on to the infectious stage and show the symptoms of the disease are denoted by εE, where ε is the per-capita infectious rate. $1/\varepsilon$ represents the average time for a latent individual to become infectious. It will be denoted by γI, where γ is the per-capita recovered rate.

The first assumption to apply the SEIR model in this work considers the population as closed; meaning that the effects of demographic changes are theoretically minimized during the epidemic. Both birth rate and death rate are taken to be zero, and thus, the compartment model is said to be closed, i.e., the population is static with no new individuals coming or going.

When the population is approximately constant over the time of the epidemic and the disease has a high mortality rate (such as smallpox), a new parameter called infectious mortality rate (δ) can be assigned to track the deaths caused by the disease. In this case, the total mortality rate for the infectious state is represented by δ. Consequently, $1/\delta$ is the average time it takes for an individual to die once having entered the infectious stage.

Without medical treatment, the R-class is removed because the outcome of the disease will surely be death (D) (Lekone & Finkenstädt, 2006; Ndanguza *et al.*, 2013). C is not an epidemiological parameter and does not represent a compartment, rather it is used to keep track of the cumulative number of smallpox cases at the onset of symptoms (sum of I and R).

The smallpox virus epidemiological data obtained from a deep review of the scientific literature available is reported in Table 2.

Table 2: **Smallpox virus epidemiological data provided by scientific literature.**

EPIDEMIOLOGICAL FEATURES	VALUE	REFERENCES
Incubation rate ($\mathcal{E}$)	0.0667 day^{-1}	Gani & Leach (2001), Legrand *et al*. (2004), Del Valle *et al*. (2005) and Adivar & Selen (2011)
Recovery rate (γ)	0.0625 day^{-1}	
Infectious mortality rate (δ)	0.0268 day^{-1}	
Transmission rate (β)	0.1 day^{-1}	
Size of the population (N)	57.756.988	Eclipse Foundation (2020)
Number of index cases	1,000	Section 3.2
Population density	206/km^2	Worldometer (2020)

3.4 Interventions - Physical Countermeasures

The physical countermeasures are intended three different type of interventions: social distancing, shutting down air travel (for a county, state or the whole country), and preventing mixing of infected individuals across borders. These kinds of interventions are used, together with a vaccination program and isolation of infected individuals (medical countermeasures), to control an outbreak (Olson & Shchelkunov, 2017; Meyer *et al.*, 2020). STEM uses triggers, predicates and modifiers to implement interventions. A trigger contains a predicate which, when satisfied, invokes one or more modifiers that change some aspect of a running simulation.

3.4.1 Social Distancing

Social distancing (e.g., distributing face masks, closing public buildings) reduces the transmissibility of the pathogen, so changing the transmission rate in the disease makes sense. In this work, the control of the outbreak by implementing social distancing was applied 26 days after the attack; consequently, the transmission rate (β), according to previous studies (Ahmed *et al.*, 2018), was reduced by 50% from 0.1 day^{-1} to 0.05 day^{-1}.

3.4.2 Shutting Down Air Travel

Shutting down the air transportation in a region is one of the options that can be carried out to control an outbreak. To model this in STEM, modifiers that change the total number of passengers traveling to / from the major Italian regions (Lazio, Tuscany, Lombardy, Piemonte, Emilia-Romagna, Sicily and Sardinia) were applied. In particular, the air traffic was blocked in these regions after 26 days from the attack by bringing the air travel rate value to 0.

3.4.3 Preventing Mixing of Infected Individuals Across Borders

The contacts of infected and healthy individuals across borders can either be shut down globally, or for a particular border only. The common borders (Figure 2) were modified in the simulated scenario. In particular, the common boundaries between the main Italian regions (Lazio, Tuscany, Lombardy, Piemonte, Emilia-Romagna) were closed 26 days after the attack.

3.5 Assumptions

To run the simulations, the following assumptions were considered:

i) The whole population was considered susceptible at t=0; at the start of the outbreak $N = S(t)$.

ii) During the epidemic simulation, the population was considered as constant. In a constant population, no deaths due to outside factors are taken into account. Moreover, the number of births that occurred is so small that is negligible. For this reason, in the simulation the parameters μ* and μ were unconsidered.

iii) All the observed cases (expect index cases) were assumed to be related to human-to-human transmission only.

iv) The time from the bioterrorist attack to the intervention of the authorities was considered 25 days. If an attack with smallpox occurred, the first case would develop the first symptoms around 12–14 (7–17) days (mean and range of the duration of the latency period) after the attack (Legrand *et al.*, 2003). Since smallpox was eradicated in 1979, lack of experience could lead physicians to misdiagnose smallpox initially, delaying intervention (O'Toole, 1999; Legrand *et al.*, 2003; Madeley, 2003). Thus, we assumed that the time to intervention ranged between 7 and 45 days, and was fixed at 25 days (half).

v) The three types of interventions were applied together in the simulation; the time of intervention was fixed at 26 days after the attack.

vi) The duration of the outbreak was considered as 180 days (about 25 weeks); at that simulated time, the number of exposed (E) and infectious (I) was near to zero, and the number of disease deaths and recovered (R) remained constant.

vii) The prodromal phase was combined with the incubation phase. Thus, the mean time in the non-infectious stages denoted by E, here corresponding to the incubation period plus the prodromal phase, was assumed to be 15 days. The incubation rate ($\mathcal{E}$) was consequently 1/15.

viii) Since smallpox vaccination programs ended about 30 years ago, and the effectiveness of a smallpox vaccine is assumed to last for 10 to 30 years (Kaplan, 2003), it was assumed that the population has no immunity.

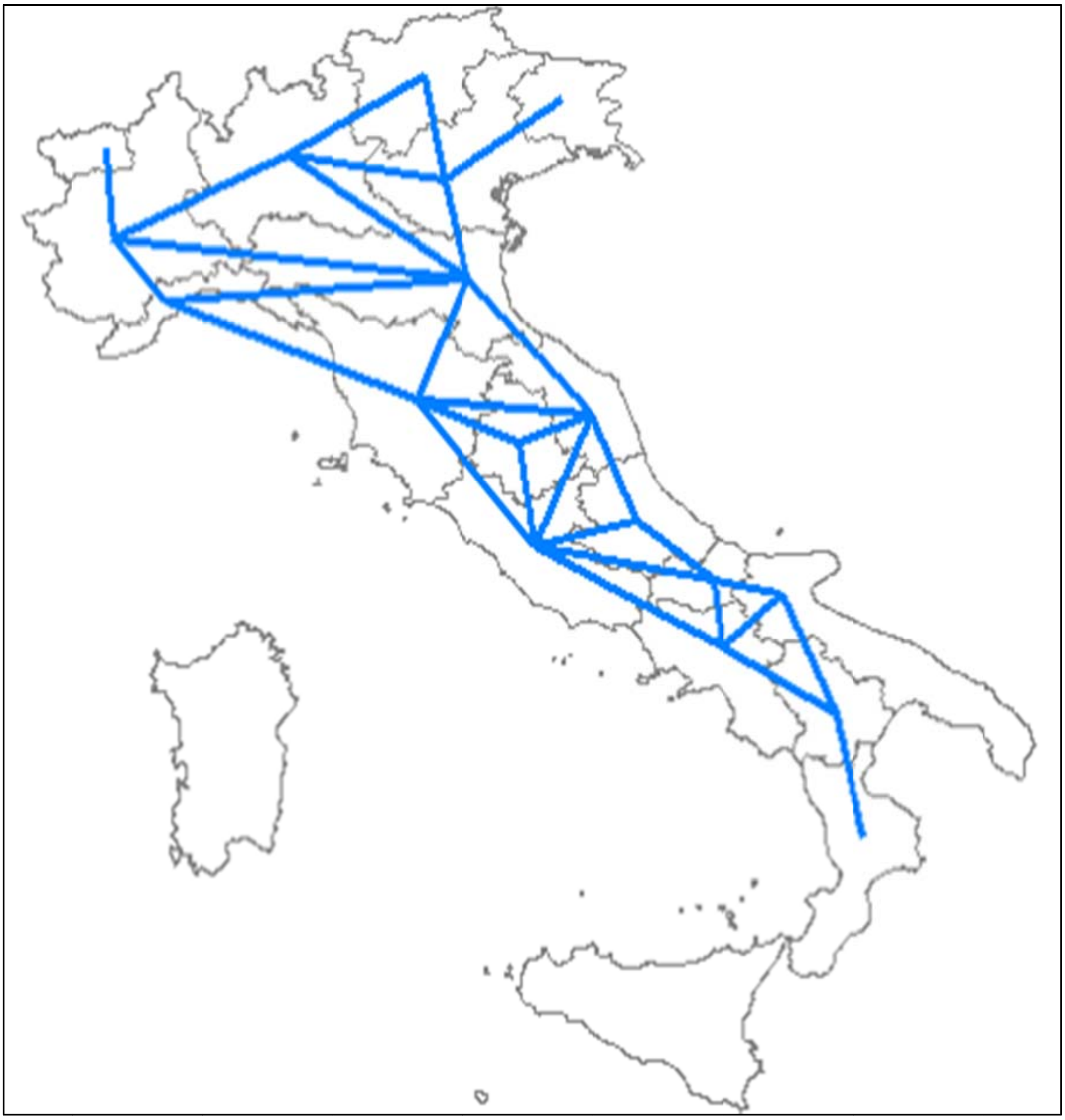


Figure 2: The schematic representation of the Italian common borders in STEM.

4. RESULTS & DISCUSSION

In this work, the authors aimed to demonstrate, considering a hypothetical scenario in Italy, the consequences of a bioterrorist attack carried out using smallpox virus in three different situations. In the first case, the consequences of the epidemic were studied and evaluated at 25 days after the attack and before the application of any intervention, applying to STEM the estimated parameters shown in Table 2. In this case, 25 days after the start of the simulated outbreak, 1,851 affected people were identified (Table 3). The STEM analysis of the simulated outbreak (Figure 3A) shows the evolution of the population in different epidemiological states during the time. Moreover, the geographical distribution of the involved population (Figure 4A) in different epidemiological states (*E*, *I* and *R*) after 25 days of simulation was obtained using STEM. The intensity of the colors in the different Italian regions is proportional to the number of people involved in one of the four epidemiological states.

In the second case, the consequences of the epidemic were studied and evaluated at 180 days after the attack and without the application of any intervention. As expected, a considerable number of affected people was identified, reaching the value of 14,943 (Table 3). As in the previous case, the STEM analysis of the simulated outbreak (Figure 3B) shows the evolution of the population in different epidemiological states during the time and geographical distribution of the involved population (Figure 4B).

In the third and last case studied the consequences of the outbreak 180 days after its start and considering the application of physical countermeasures (starting the 26th day after the attack). The obtained data were compared with the second case and, as expected, a strong difference between the second and third simulation results was identified (Table 3): the involved people were reduced by four times when physical countermeasures were applied. The STEM analysis and related geographic distribution of the involved population are showed in Figures 3C and 4C respectively.

All the simulations were adjusted using a logarithmic intensity scale and a specific gain factor (x 10^5), and the fourth-order Runge-Kutta algorithm to solve it numerically, written in JavaScript™.

Table 3: Results of the simulated scenario for 25 days and 180 days after the outbreak started, as well as 180 days after the outbreak started with the application of physical countermeasures.

	After 25 days	**After 180 days**	**After 180 days with physical interventions applied**
Index cases	1,000	1,000	1,000
Exposed (*E*)	666$^{(*)}$	1,476$^{(*)}$	23$^{(*)}$
Infectious (*I*)	489$^{(*)}$	1,049$^{(*)}$	22$^{(*)}$
Recovered (*R*)	953$^{(*)}$	8,695$^{(*)}$	2,557$^{(*)}$
Disease deaths (*D*)	409$^{(*)}$	3,723$^{(*)}$	1,093$^{(*)}$
People involved (*C*)	1,851	14,943	3,695

* Rounded off to the nearest whole number

5. CONCLUSION

Although smallpox has been declared eradicated, the possibility of releasing smallpox or smallpox like organisms brings the potential for a catastrophic scenario as it is for deadly emerging pathogens (Olson & Shchelkunov, 2017). In order to predict and prevent, or at least to reduce this kind of threat, or an epidemic outbreak or infection from spreading, decision makers and policymakers can benefit from simulation tools such as STEM.

In this paper, STEM was applied to a hypothetical bioterrorist scenario using historical epidemiological data from the literature, considering three different cases. The obtained data, starting from 1,000 index cases, were achieved by applying the deterministic SEIR compartmental model to a STEM using standard population and integrating the input data with real smallpox outbreak data from the literature (e.g., transmission, incubation, and recovery rates). The project was designed as the starting point for scenario evolution simulations. The final results obtained from STEM analysis allowed, with some preliminary assumptions, an easy and complete assessment of how the population size changes in the three cases considered as a whole during the simulations as well as in the different spatial disease compartments. The STEM simulations analyze the effects of epidemic behavior change alone and in combination with specific control measures. As a result, the provided information can suggest to decision makers, with a high level of accuracy, how the outbreak would spread and develop in space and time in different phases: in the early, during, and in the last phases of the epidemic, as well as in combination with control measures.

Thus, STEM can significantly improve preparedness and response in the field of bioterrorism. In fact, being able to estimate the spatial distribution and spread of an agent and the temporal disease outbreak patterns reflects a more effective emergency planning and response. As a result, this tool could help to develop (and test) control strategies based on computer simulations. Additionally, it would address the most important information gaps for the creation of faster and more specific exposure assessments and risk characterization.

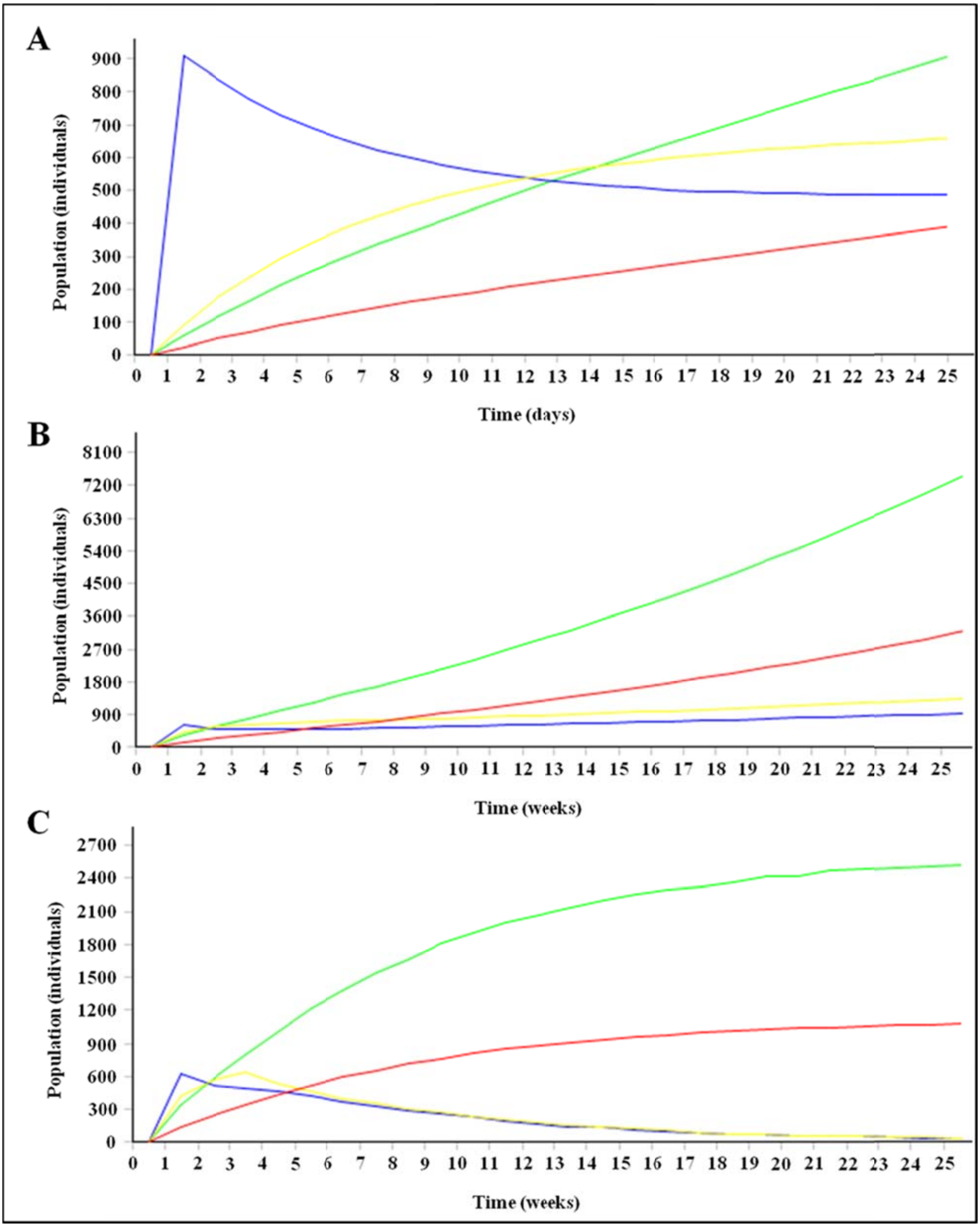


Figure 3. STEM analysis of the scenario in the different considered situations. The evolution of people compartmented in different states with respect to time is represented. Each state is reported in a specific color; exposed (yellow line), infected (blue line), recovered (green line), and dead (red line) individuals. A) First case: 25 days after the outbreak start. B) Second case: 180 days after the outbreak start. C) Third case: 180 days after the outbreak start with the application of physical countermeasures after the 26th day from the outbreak start.

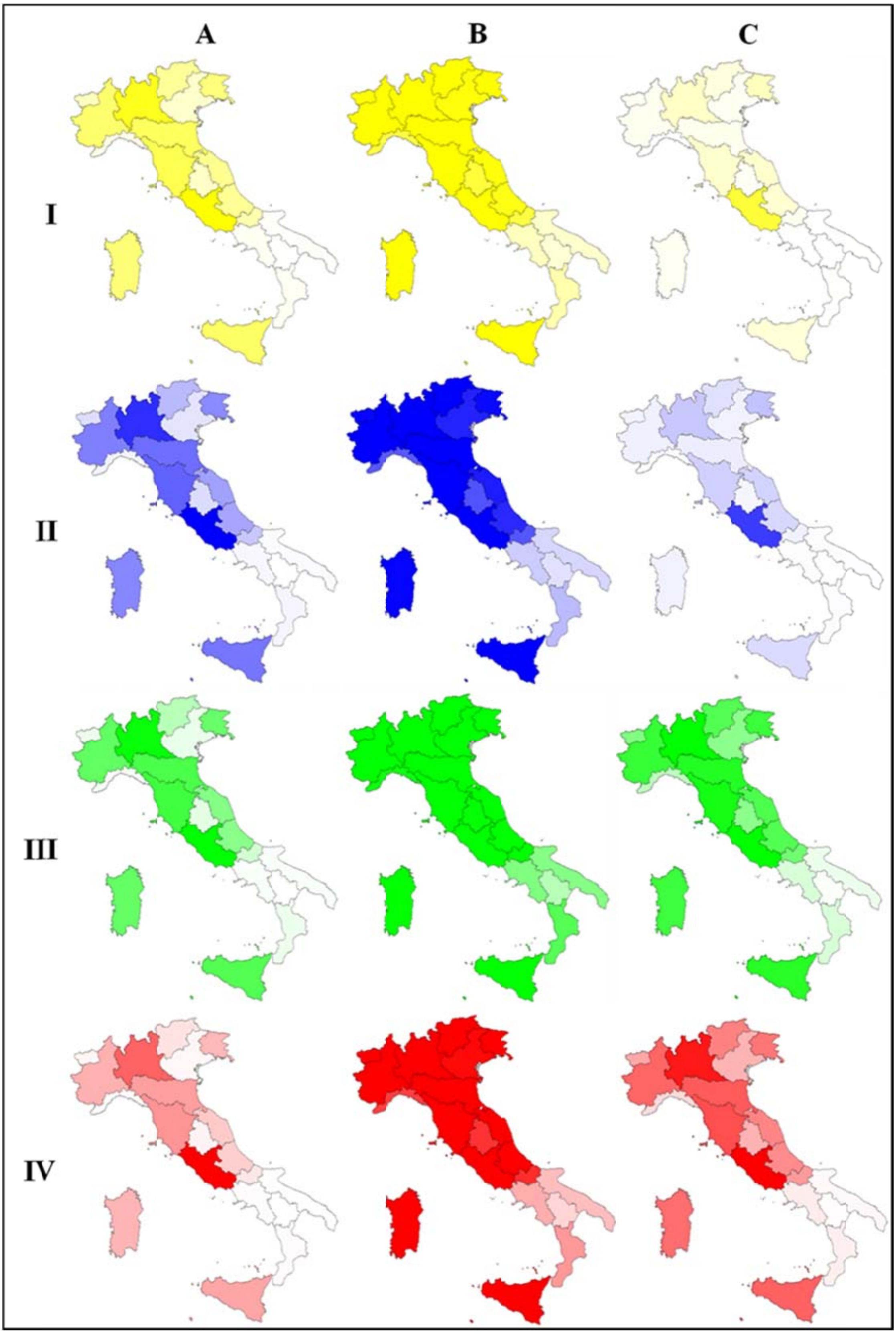


Figure 4. STEM analysis of the scenario in the different considered situations. The geographical distribution of people compartmented in the different states is represented. Each state is reported in a specific color; I) Exposed (*E*), yellow; II) Infected (*I*), blue; III) Recovered (*R*), green; IV) Disease Deaths (*D*), red. The color intensity is proportional to the number of involved people in the four different epidemiological states. A) First case: 25 days after the outbreak start; B) Second case: 180 days after the outbreak start. C) Third case: 180 days after the outbreak start with the application of physical countermeasures from the 26th day from the outbreak start.

REFERENCES

Adivar, B., & Selen, E. S. (2011). Compartmental disease transmission models for smallpox. *Discrete Cont. Dyn. Syst.*, **Vol. 2011**: 13-21.

Ahmed, F., Zviedrite, N. & Uzicanin, A. (2018). Effectiveness of workplace social distancing measures in reducing influenza transmission: a systematic review. *BMC Public Health*, **18**:518.

Bachinsky, A.G. & Nizolenko, L.P. (2013). A universal model for predicting dynamics of the epidemics caused by special pathogens. *BioMed Res. Int.*, **Vol. 2013**: 467078.

Balcan, D., Hu, H., Goncalves, B., Bajardi, P., Poletto, C., Ramasco, J.J., Paolotti, D., Perra, N., Tizzoni, M., Van den Broeck, W., Colizza, V. & Vespignani, A.(2009). Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC Med.*, **7**: 45.

Baldassi, F., D'Amico, F., Carestia, M., Cenciarelli, O., Mancinelli, S., Gilardi, F., Malizia, A., Di Giovanni, D., Soave, P. M., Bellecci, C., Gaudio, P. & Palombi, L. (2015). Testing the accuracy ratio of the Spatio-Temporal Epidemiological Modeler (STEM) through Ebola haemorrhagic fever outbreaks. *Epidemiol. Infect.*, **144**:1463-1472

Bhalla, D.K. & Warheit, D.B. (2004). Biological agents with potential for misuse: a historical perspective and defensive measures. *Toxicol. Appl. Pharm.*, **199**: 71-84.

Bozzette, S.A., Boer, R., Bhatnagar, V., Brower, J.L., Keeler, E.B., Morton, S.C. & Stoto, M.A. (2003). A model for a smallpox-vaccination policy. *New Eng. J. Med.*, **348**: 416-425.

Breman, J.G., & Henderson, D.A. (2002). Diagnosis and management of smallpox. *New Eng. J. Med.*, **346**: 1300-1308.

Cenciarelli, O., Pietropaoli, S., Gabbarini, V., Carestia, M. & D’Amico, F. (2014). Use of non-pathogenic biological agents as biological warfare simulants for the development of a stand-off detection system. *J. Microb. Biochem. Technol.*, **6**: 375-380.

Cenciarelli, O., Rea, S., Carestia, M., D’Amico, F., Malizia, A., Bellecci, C., Gaudio, P., Gucciardino, A., & Fiorito, R. (2013). Bioweapons and bioterrorism: a review of history and biological agents. *Defence S&T Tech. Bull.*, **6**: 111-129.

Centers for Diseases Control and Prevention (CDC) (2015). *Bioterrorism Agents / Diseases*. Available online at: http://www.bt.cdc.gov/agent/agentlistcategory.asp (Last access date: 30 May 2020).

Chowell, G., Hengartner, N. W., Castillo-Chavez, C., Fenimore, P. W., & Hyman, J. M. (2004). The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *J Theor Biol*, *229*, 119-126.

Del Valle, S., Hethcote, H., Hyman, J. M., & Castillo-Chavez, C. (2005). Effects of behavioral changes in a smallpox attack model. *Math. Biosci.*, **195**: 228-251.

Department of Health and Human Services (DHHS) (2009). *Biosafety in Microbiological and Biomedical Laboratories, 5th Ed*. U.S. Department of Health and Human Services (DHHS), Washington, D.C.

Dudley, J.P. & Woodford, M.H. (2002a). Bioweapons, biodiversity, and ecocide: Potential effects of biological weapons on biological diversity. *BioScience*, **52**: 583-592.

Dudley, J.P. & Woodford, M.H. (2002b). Bioweapons, bioterrorism and biodiversity: Potential impacts of biological weapons attacks on agricultural and biological diversity. *Rev. Sci. Tech.*, **21**: 125-138.

Eclipse Foundation (2020). *About the Eclipse Foundation*. Available online at: https://www.eclipse.org/org (Last access date: 30 May 2020).

Esposito, J.J. & Fenner, F. (2001). Poxviruses. *In* Knipe, D.M. & Howley P.M. (Eds.), *Fields Virology*. Lippincott Williams & Wilkins, Philadelphia, Pennsylvania, pp. 2885–2921.

Gani, R. & Leach S. (2001). Transmission potential of smallpox in contemporary populations. *Nature*, **414**: 748–751.

Gubser, C. & Smith, G.L. (2002). The sequence of camelpox virus shows it is most closely related to variola virus, the cause of smallpox. *J. Gen. Virol.*, **83**: 855-872.

Henderson, D.A. (1999). The looming threat of bioterrorism. *Sci.*, **283**: 1279-1282.

Henderson, D.A. (2009). *Smallpox - The Death of a Disease: The Inside Story of Eradicating a Worldwide Killer*. Prometheus Books, Buffalo, New York.

Hethcote, H.W. (2000). The mathematics of infectious diseases. *SIAM Rev.*, **42**: 599-653.

Kaiser, J. (2014). *Science Insider*. Available online at: http://news.sciencemag.org/health/2014/07/six-vials-smallpox-discovered-u-s-lab (Last access date: 30 May 2020).

Kaplan, E.H. (2003). Emergency response to a smallpox attack: The case for mass vaccination. *Math Biosci*, **185**: 33-72.

Lavine, J.S., Poss, M. & Grenfell, B.T. (2008). Directly transmitted viral diseases: modeling the dynamics of transmission. *Trends Microbiol.*, **16**: 165-172.

Legrand, J., Grais, R.F., Boelle, P.Y., Valleron, A.J. & Flahault, A. (2007). Understanding the dynamics of Ebola epidemics. *Epidemiol. Infect.*, **135**: 610-621.

Legrand, J., Viboud, C., Boelle, P.Y., Valleron, A. J. & Flahault, A. (2004). Modelling responses to a smallpox epidemic taking into account uncertainty. *Epidemiol. Infect.*, **132**: 19-25.

Lekone, P. & Finkenstädt, B. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, **62**: 1170-1177.

Madeley, C.R. (2003). Diagnosing smallpox in possible bioterrorist attack. *Lancet*, **361**: 97-98.

Mahy, B.W.J. (2003). An overview on the use of a viral pathogen as a bioterrorism agent: why smallpox?. *Antiviral Res.*, **57**: 1-5.

Meyer, H., Ehmann, R. & Smith G.L. (2020). Smallpox in the post-eradication era. *Viruses*, **12**: 138.

Moss, B. (1996). Genetically engineered poxviruses for recombinant gene expression, vaccination, and safety. *Proc. Natl. Acad. Sci.*, **93**: 11341-11348.

Moss, B. (2007). Poxviridae: The viruses and their replication. *In* Knipe D.M. & Howley P.M. (Eds.), *Fields Virology*. Lippincott Williams & Wilkins, Philadelphia, Pennsylvania, pp. 2905–2946.

Ndanguza, D., Tchuenche, J. M. & Haario, H. (2013). Statistical data analysis of the 1995 Ebola outbreak in the Democratic Republic of Congo. *Afrika Matematika*, **24**: 55-64.

O'Toole, T. (1999). Smallpox: An attack scenario. *Emerg. Infect. Dis.*, **5**: 540.

Olson, V. A. and Shchelkunov, S. N. (2017). Are we prepared in case of a possible smallpox-like disease emergence? *Viruses*, **9**: 242.

Reardon, S. (2014). *Vials of Smallpox Virus Found Unsecured at NIH*. Available online at: www.scientificamerican.com/article/vials-of-smallpox-virus-found-unsecured-at-nih/ (Last access date: 30 May 2020).

Riedel, S. (2005). Smallpox and biological warfare: a disease revisited. *Proc. Bayl. Univ. Med. Cent.*, **18**: 13-20.

Sulaiman, I. M., Tang, K., Osborne, J., Sammons, S. & Wohlhueter, R. M. (2007). GeneChip resequencing of the smallpox virus genome can identify novel strains: a biodefense application. *J. Clin. Microbiol.,*, **45**: 358-363.

Utgoff, V.A. (1993). The biotechnology revolution and its potential military implications. In Roberts, B. (Ed.), *Biological Weapons: Weapons of the Future?* Center for Strategic and International Studies, Washington D.C., pp. 28-31.

Whitley, R.J. (2003). Smallpox: a potential agent of bioterrorism. *Antiviral Res.*, **57**: 7-12.

Wolfe, N.D., Dunavan, C.P. & Diamond, J. (2007). Origins of major human infectious diseases. *Nature*, **447**: 279-283.

World Health Organization (WHO) (2008). WHO recommendations concerning the distribution, handling and synthesis of variola virus DNA. *Wkly. Epidemiol. Rec.*, **83**: 393-400.

Worldometers (2020). *Italy Demographics*. Available online at https://www.worldometers.info/demographics/italy-demographics (Last access date: 30 May 2020).