DEFENCE S&T TECHNICAL BULLETIN

VOL. 13 NUM. 1 YEAR 2020 ISSN 1985-6571

CONTENTS

Ministry of Defence
Malaysia

SCIENCE & TECHNOLOGY RESEARCH
INSTITUTE FOR DEFENCE (STRIDE)

AIMS AND SCOPE

The Defence S&T Technical Bulletin is the official technical bulletin of the Science & Technology Research Institute for Defence (STRIDE). The bulletin, 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 bulletin 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 bulletin 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 bulletin (Vol. 13, Num. 2) is expected to be published in November 2020. The due date for submissions is 13 August 2020. **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:

Books

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

Book Chapters

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.

Journals / Serials

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

Online Sources

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).

Unpublished Materials (e.g. theses, reports and documents)

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

A REVIEW OF HARDWARE TROJAN DETECTION: AN OVERVIEW OF DIFFERENT PRE-SILICON TECHNIQUES

Hau Sim Choo[1], Chia Yee Ooi[1*], Michiko Inoue[2], Nordinah Ismail[1] & Chee Hoo Kok[1]

[1]Embedded System Research Laboratory, Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia (UTM), Malaysia
[2]Dependable System Laboratory, Graduate School of Science and Technology, Nara Institute of Science and Technology (NAIST), Japan

[*]Email: ooichiayee@utm.my

ABSTRACT

As integrated circuit (IC) design gets more complicated, outsourcing parts of the IC design and fabrication is commonly applied to simplify the production and reduce the cost. This leads to the threat of malicious manipulation to the design by the third parties involved. Such threat is considered as hardware Trojan attack, which could pose adverse impacts to a system or network. Recently, hardware Trojan is gaining more interest as a research subject, especially pre-silicon detection. Various kinds of hardware Trojan detection approaches have been proposed to detect different Trojan types in different circuits. This study summarises the existing hardware Trojan detection methods and discusses the attributes of the methods to better distinguish them between each other. Existing pre-silicon detection methods are reviewed, which includes verification-based, threshold-based and machine learning-based feature analysis techniques. The objective of this study is to ease the future hardware-Trojan-related research by providing an organised summary of detection techniques, especially pre-silicon detection.

Keywords: *Hardware Trojan; hardware security; integrated circuit (IC) authentication; trusted integrated circuit.*

1. BACKGROUND OF HARDWARE TROJAN

In the information technology (IT) realm, a Trojan refers to a malicious computer software program that may appear legitimate but schemingly modifies the true intents of the original software or system. In the semiconductor industry, almost similarly in context, hardware Trojan describes the malicious modification of the circuitry of an integrated circuit (IC) chip. The presence of hardware Trojan implies that the compromise of hardware security on IC chips has become a major concern in the semiconductor industry. Due to the rapid growth and advancement of technology, the demand for sophisticated and dedicated system boards and ICs has risen similarly. Technology companies whose fab-less business and research focus are in sophisticated and complex system development are inclined to find third-party solutions for IC chips or system-on-chips (SoCs) fabrication. This widespread trend poses greater threat to hardware security with the continuous increase in complexity and cost of electronics design. In order to simplify and reduce the cost of the supply chain, fab-less technology companies sometimes turn to some untrusted third parties that may be involved in inserting malicious modification to the ICs during the design and manufacturing process. Such an attack is defined as hardware Trojan insertion (Tehranipoor & Koushanfar, 2010). The possible attacks that can be launched by hardware Trojan include information leaking, denial-of-service attack and degradation of the system performance (Shakya *et al.*, 2017).

Although hardware Trojan attack is not a prevalent issue today, a few cases have been reported. Adee (2008) reported that in 2007, a Syrian radar failed to warn of an incoming air strike from Israel. It was suspected that a backdoor was built into the system's circuitry, which might be the cause of the radar

system's failure. This issue showed the effect of hardware Trojan in endangering a nation's security. Governments around the world are aware of this threat and has started focusing on hardware Trojan research. For example, a French government funded project named HOMERE researches on hardware Trojan (Francq *et al.*, 2015). A hardware Trojan attack could cause adverse impact to a system or network. When a device connected to a network has been attacked by a hardware Trojan, other devices in the same network will be vulnerable as well (Koley & Ghosal, 2015). As the Internet-of-Things (IoT) starts to be pervasive and integrates into our daily life, we are also exposed to hardware Trojan attacks. Besides military electronic systems, hardware Trojan can be inserted into any electronic systems, such as transportation, household appliances and healthcare systems. Hardware Trojans could already now exist in electronic systems around us without us realising it. Such worry of hardware Trojan is not without reason since a Trojan is usually stealthy and hidden.

Thus, hardware Trojan is now increasingly gaining interest by the research communities. The research topics related to hardware Trojan include design of Trojan, Trojan prevention and Trojan detection. Design of Trojan research aims to obtain a more practical Trojan to be used in experiments to evaluate new detection methods' effectiveness, as a real Trojan is currently unavailable. Trojan prevention research aims to analyse electronic IC design susceptibility to Trojan and develop techniques to prevent Trojan insertion. Trojan detection research aims to obtain methods for detecting any inserted Trojan in a design.

This study focusses on Trojan detection and consists of several sections of discussions. First, an overview of existing Trojan detection methods is discussed. Second, the existing pre-silicon Trojan detection methods are reviewed. Third, the discussion on reviewed studies is presented. The final section contains the conclusions of this study.

2. OVERVIEW OF HARDWARE TROJAN DETECTION

Detecting Hardware Trojan is not a trivial task. Currently, there is no standard detection method available to detect all hardware Trojans. Adversaries will always try to develop their Trojan to make it stealthier and able to evade detection during verification and testing process. In addition, the diversity of Trojan attacks and Trojan effects make the detection difficult and challenging. In recent years, many detection methods have been proposed. Each method aims for different purposes and has different motivations. As shown in Figure 1, the differences between each method can be seen from several general attributes of the Trojan detection method (Tehranipoor & Koushanfar, 2010; Bhunia *et al.*, 2014; Francq *et al.*, 2015; Moein *et al.*, 2016; Xiao *et al.*, 2016), such as manipulation to circuit, the detection level and the analysed parameter. This section discusses on each general attribute and the specific attributes of Trojan detection method. This can help researchers to classify and review different existing Trojan detection techniques and ease the future hardware Trojan related research works.

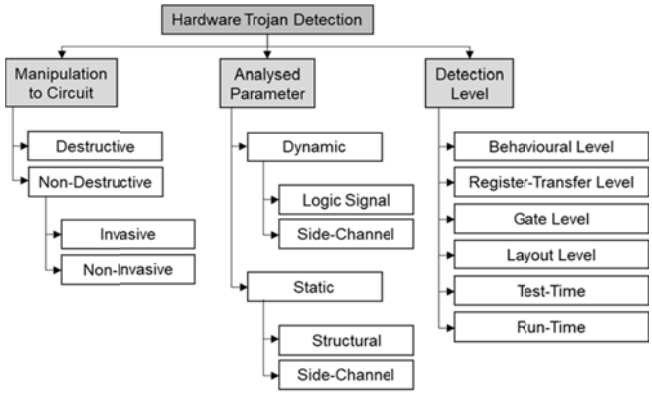


Figure 1: Hardware Trojan detection method attributes.

2.1 Types of Hardware Trojan

Hardware Trojans are classified based on their attributes. A Trojan taxonomy system was proposed by Shakya *et al.* (2017), as shown in Figure 2, in which six general Trojan attributes, namely insertion phase, abstraction level, activation mechanism, effect, location and physical characteristics, are used to further classify attributes of Trojan. For instance, a Trojan that is inserted at design phase can have an attribute of abstraction level of gate level, user input triggered mechanism, and so on. With such classification, several Trojan detection techniques were proposed to address the challenges in specific Trojan attributes. For example, Chen & Liu (2017) proposed a methodology to detect Trojan with single-trigger-pattern activation mechanism, while Wang *et al.* (2017) proposed a methodology to detect information-leaking Trojan.

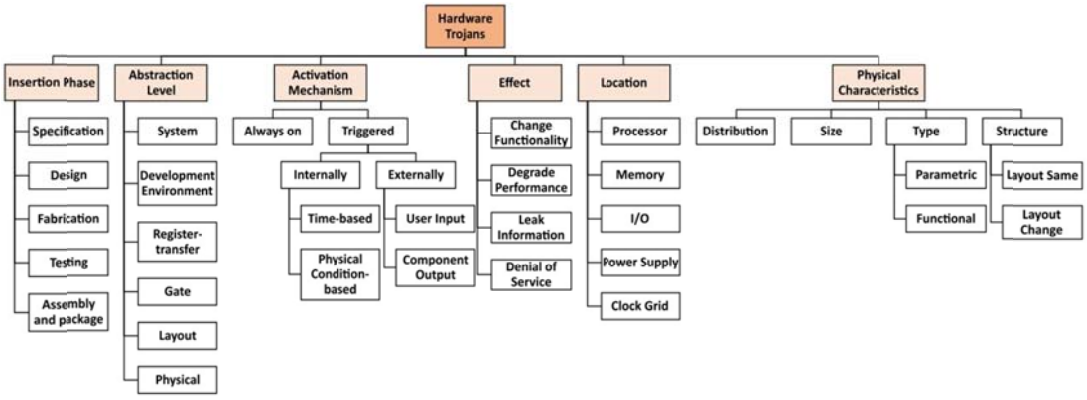


Figure 2: Hardware Trojan taxonomy (Source: Shakya *et al.*, 2017).

2.2 Types of Detection Techniques

Hardware Trojan detection methods have three general attributes which are manipulation to circuit, analysed parameter, and detection level, as shown in Figure 1. These general attributes can help to address the differences between the existing detection methods.

2.2.1 Manipulation to Circuit

There are three types of circuit manipulation in performing Trojan detection: destructive, non-destructive but invasive, and non-destructive and non-invasive.

- ***Destructive***
  Destructive hardware Trojan detection techniques refer to reverse engineering techniques or physical inspection techniques such as optical scanning (Jin & Makris, 2008; Bao *et al.*, 2014, 2016). These techniques involve backside thinning and de-processing operations, which will destroy the circuit after the analysis.

- ***Non-Destructive***
  Non-destructive approaches refer to techniques that do not destroy the circuits after the analysis, but instead allow the circuit to be used after authentication. It can be classified further again under two categories: non-invasive and invasive. The difference of the two techniques is non-invasive techniques leave the original design without modification while the invasive techniques make changes to the original design to assist in hardware Trojan detection.

  - ***Invasive***
    For invasive detection methods, some circuit modifications are made to the circuit to facilitate the Trojan detection, especially post-silicon detection (Bilzor *et al.*, 2011, 2012, 2017; Forte *et al.*, 2013; Ngo *et al.*, 2015; Alsaiari *et al.*, 2017). These approaches relate

to design for hardware trust techniques, which eases the detection of Trojans and provide effective circuit authentication by making some beneficial alterations to the original design.

- ***Non-Invasive***
  Non-invasive hardware Trojan detection methods detect the existence of Trojans directly from the circuit design or circuit itself without any augmentation on the design (Hasegawa *et al.*, 2016; Salmani, 2016). This kind of detection methodologies can be applied in either the pre-silicon or post-silicon stage. Unlike the invasive technique, this technique does not alter the original design. Thus, it does not affect circuit performance, such as power consumption.

2.2.2 Analysed Parameter

For each hardware Trojan detection technique based on different manipulation to circuit, either static or dynamic analysis can be applied, based on the parameters being analysed. In machine learning and statistical approaches, the parameters in analysis are also called features. For dynamic analysis, or methods that analyse dynamic parameter, logic simulation or functional testing is required. Examples of extractable dynamic parameters are power, path delay, voltage, logic signal and current. A Trojan detection method by analysing power was introduced by Wang *et al.* (2013). Power traces of genuine circuits are extracted and used to train a machine learning classifier. By giving the power trace of a circuit-under-test generated by an identical test bench, the classifier can recognise whether it is a Trojan circuit or a Trojan-free circuit. Static analysis, on the other hand, analyses static parameters, which remain unchanged during circuit operation. This kind of information can be obtained directly from the circuit or its design without using the test vectors. Examples of static parameters used to identify Trojans are interconnection of gates, testability of nets, and leakage current. Salmani *et al.* (2016) analysed the testability of nets to identify Trojan nets by using machine learning approach. The information can be extracted directly from netlist design without logic simulation. Based on the analysed parameters, the analysis method used in Hardware Trojan detection approaches can be known as:

- ***Logic Signal Analysis***
  Logic testing is a dynamic analysis that observes the circuit primary output to identify any malicious functionality introduced by Trojan (Banga & Hsiao, 2010; Zhang & Tehranipoor, 2011a ; Bilzor *et al.*, 2011, 2012; Ngo *et al.*, 2015; Hu *et al.*, 2016; Wang *et al.*, 2017; Alsaiari *et al.*, 2017). Circuit primary output data, or logic signals can be obtained from either simulation of circuit design (pre-silicon) or testing on fabricated circuit (post-silicon). Trojan detection by analysing circuit primary output would require activation of Trojan to show its effect to be detected. Compared to other analysis methods, process variations have less impact on logic signal analysis, and usually need not to be considered, which is an advantage when using logic signal analysis to detect Trojan.

- ***Side-Channel Analysis***
  Side-channel analysis can be used to detect Trojan by investigating any malicious impacts caused by Trojan to the circuit's side-channel parameters (Agrawal *et al.*, 2007; Potkonjak *et al.*, 2009; Zhang & Tehranipoor, 2011b; Forte *et al.*, 2013; Kumar & Srinivasan, 2014). Dynamic side-channel parameters, including path delay, current and electromagnetic field, can be obtained from either simulation of circuit design or testing on fabricated circuit. In addition, side-channel information can be also a static parameter such as current leakage. As Trojan is usually infused by modifying a circuit, some side-channel parameters would be affected. The process variation complicates post-silicon Trojan detection that relies on analysis of side-channel parameters due to the fact that Trojan's response might hide within the process variation and escapes from the detection. In addition, Trojan detection using side-channel

analysis often requires a golden circuit or fingerprint that is Trojan free, which is difficult to obtain.

- ***Structural Analysis***
  Structural analysis refers to techniques used to analyse static structural characteristics of a circuit (Yao *et al.*, 2015; Oya *et al.*, 2015; Salmani, 2016; Chen & Liu, 2017; Shen & Zhao, 2017; Piccolboni *et al.*, 2017; Hasegawa *et al.*, 2017b; Chen *et al.*, 2018). The structural information can be obtained from either the circuit design or the physical layout. The extractable information includes structure of control-data flow graph, size, density, geometric information, interconnection of gates, as well as routing and placing of logics. Structural analysis identifies the existence of hardware Trojan through any suspicious structural pattern that shows Trojan characteristics.

2.2.3 Detection Level

Another important attribute of detection method that is useful in categorising Trojan is the detection level. In this study, detection level, as shown in Figure 3, refers to the stage of the IC development process where the Trojan detection is applied. In typical IC design, the levels involved are behavioural level, register-transfer level, gate level and layout level. Post-silicon Trojan detection can be performed during post-silicon testing or during the system operation during which the circuit is installed.

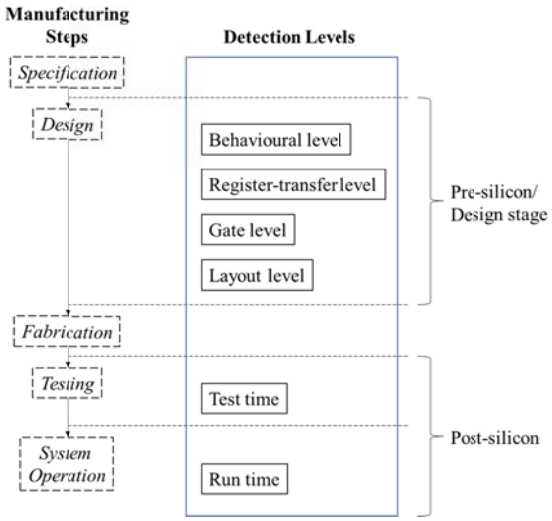


Figure 3: Detection levels corresponding to manufacturing steps.

- ***Behavioural Level***
  Behavioural level description defines the functionality of a design although the design is not necessarily synthesisable. At this abstraction level, the algorithm is described by a set of instructions running in sequence to perform some tasks. The possible effect of Trojan attack is modification of functionality.

- ***Register-Transfer Level (RTL)***
  At the register-transfer-level abstraction, hardware description languages, such as Verilog and VHDL, are used to create high-level representations of a circuit. The RTL design is developed by a RTL designer or generated using high-level synthesis from algorithm description. The flow of digital signals between hardware registers, and the logical operations performed on those signals are described in this level. The possible effect of Trojan attack is alteration of circuit operations and data transfers.

- ***Gate Level***
  A gate-level netlist is a description of the interconnections of gates in a circuit. The netlist can be obtained from synthesising a RTL design through logic synthesis. A list of electronic components in the circuit, and their connectivity is presented in a gate-level netlist. The possible effect of Trojan attack is alteration of logical links and timing properties.

- ***Layout Level***
  IC layout is the geometric representation of an integrated circuit, which corresponds to the pattern of metal, oxide or semiconductor layers that form the components of the IC. Placement and routing of the components are to be done at this level. The aim is to meet the criteria of performance, size, density and manufacturability. The possible effect of Trojan attack includes alteration of thickness or length of the layers and wires as well as position and interconnection of the components.

- ***Post-Silicon, Test Time***
  After circuit fabrication, the produced circuits will undergo manufacturing test. Post-silicon Trojan detection is suggested to be performed after the testing. This is because the conventional testing may recognise the Trojan effect as a design fault or manufacturing fault. This can avoid unnecessary detection process when the Trojan effect can be observed during testing. Destructive detection techniques are usually applied at this stage, before the circuit is installed into the system.

- ***Post-Silicon, Run Time***
  Run time detection analyses the circuit-under-test operating behaviour in real-time to determine any Trojan effect and warns the users. This requires the circuit to be installed into a system. Run-time detection techniques require users' effort on observing any changes or alert reported to the users. This can be challenging since users do not usually own the technical knowledge of the circuit. This technique becomes the last resort if all other relevant techniques fail to detect the Trojan in previous stages.

2.3 EXISTING HARDWARE TROJAN DETECTION TECHNIQUES

In this section, existing hardware Trojan detection techniques are discussed based on their detection method attributes defined in Section 2.2.

- ***Destructive***
  - *Structural Analysis*
    Bao *et al.* (2014, 2016) introduced reverse engineering-based detection methods using machine learning. The introduced methods are test-time detection, which analyse the scanning electron microscope image of circuit layout to identify Trojan by using machine learning classifier (support vector machine and *k*-means clustering). The features used are related to the area and centroid difference between scanned image and golden layout.

  - *Side-Channel Analysis*
    Jin *et al.* (2008) proposed to use reverse engineering to generate a path delay fingerprint that can be used for Trojan detection. This approach only de-processes a few circuits instead of every single circuit. However, which circuits are qualified to be used to generate fingerprint is still a challenge.

Reverse engineering-based methods destroy the fabricated circuit while extracting the information. In addition, it usually requires a lot of time and cost although these methods are effective for authentication purposes. Thus, it is not suited for applications where every circuit needs to be authenticated. Furthermore, due to the nanometre semiconductor device fabrication technology, this approach is becoming increasingly ineffective with the increase in transistor integration density.

- ***Non-Destructive but Invasive***

  - *Logic Signal Analysis*
    Ngo *et al.* (2015) introduced an assertion-based Trojan detection method at run-time. Assertions are 'active comments', which are created to check the defined critical properties based on the circuit specification. Violation of property can be determined by observing signal behaviours. Ngo *et al.* (2015) converted the assertions into synthesisable hardware description language that was inserted into the target circuit. Assertions report to user during run time if any property is violated, which could be caused by Trojan. Bazzazi *et al.* (2017) proposed a Trojan detection method based on run-time logical testing. Some signatures are defined and implemented in the circuit. The user may verify the circuit operation to ensure Trojan attack is absent by checking the signature. Logic signal analysis in post-silicon stage is not effective to detect Trojan, as activation of Trojan under rare conditions is usually required (Chakraborty *et al.*, 2009). Therefore, design-for-trust techniques can be useful to improve the detection ability.

  - *Side-Channel Analysis*
    Zhang & Tehranipoor (2011b) suggested a design-for-trust technique to implement a ring oscillator network into circuit to detect hardware Trojan at test time. The ring oscillator network generates a power supply fingerprint that can be used to determine malicious modification.

Invasive detection techniques add some overheads into the circuit to assist the detection. These techniques are popular in post-silicon detection because it may enhance and extend the conventional testing for Trojan detection. Instead of trying to activate the Trojan during test time, the additional circuit may act as a sensor that helps to monitor the circuit operation and report to the user when malicious behaviour is observed. This gradually reduces the detection effort, since activation of Trojan is a challenge.

- ***Non-Destructive and Non-Invasive***
  - *Logic Signal Analysis*
    Wang *et al.* (2017) proposed a Trojan detection method by checking assertion at register transfer level. Assertion is defined based on the suspicious connection patterns of nodes found in the proposed coarse-grained control-data flow graph. The suspicious connection patterns correspond to the triggering mechanism of Trojan.

  - *Side-Channel Analysis*
    Kumar *et al.* (2014) used the path delay parameter to detect Trojan at the design stage. However, this method requires fingerprint generated from simulation of a golden circuit design. Thus, it is impractical because there is usually no Trojan issue at the design stage when we have the golden design. Wang *et al.* (2013) proposed to evaluate the power traces of circuit at test time to determine Trojan circuit by using one-class support vector machine. This method is robust, with no specific design being targeted for the classifier.

  - *Structural Analysis*
    Yao *et al.* (2015) introduced a threshold-based feature analysis method to analyse flip-flop level control-data flow graph generated from gate-level netlist for Trojan detection. The features analysed correspond to the triggering mechanism of Trojan. This method was then extended by Chen *et al.* (2018). Features at combinational logic level control-data flow graph were also analysed to reduce false positive issue. Bang *et al.* (2010) and Zhang *et al.* (2011a) extended conventional verification techniques to integrate Trojan detection ability at the design stage. Hasegawa *et al.* (2016) and Inoue *et al.* (2018) proposed to analyse gate-level netlist to identify Trojan gates based on five gate-level structural features of Trojan using support vector machine.

The non-invasive method is relatively popular in detection at the design stage, especially structural analysis at gate level. As detection at design may obtain the parameters or circuit information from simulation, no additional circuit is required. Even though some detection methods require additional hardware description language for detection purpose (Wang *et al.*, 2017), it can be removed before the fabrication, when its purpose is achieved.

With the suggested hardware Trojan detection method categorisation based on the detection attributes, the similarities and differences between each existing method can be observed clearer. In next section, the existing pre-silicon detection methods, with detection level at layout level or earlier, are reviewed.

3. REVIEW ON PRE-SILICON HARDWARE TROJAN DETECTION

It is highly desirable if a hardware Trojan can be detected as early as possible in the manufacturing process. Thus, compared to post-silicon detection, pre-silicon Trojan detection is more preferable in the semiconductor industry in the view of cost effectiveness to avoid further financial implication due to infected design. However, the lower the abstraction level is, the more challenging it takes to detect the Trojan since the Trojan attributes become less observable. For example, when Trojan is inserted at register-transfer level with the aim to amend the circuit functionality, it is difficult to be identified from a gate-level netlist where only the interconnection of gates is described.

One advantage of pre-silicon detection has over post detection is that the former does not suffer from process variation. Process variation is the slight variation that occurs to the attributes of transistors during circuit fabrication, in which every fabricated circuit will show slightly different parameter value and the process is unavoidable. Since pre-silicon detection uses information generated from simulation that only depends on the design and technology library provided, the process variation does not exist and will not deter the Trojan detection.

In this section, existing pre-silicon hardware Trojan detection methods related to the following categories are reviewed:

(I) Verification-based detection
- **(a)** Detection through improved IC design methodology
- **(b)** Assertion-based hardware Trojan detection
- **(c)** Information-flow tracking
- **(d)** Subgraph isomorphism

(II) Threshold-based feature analysis
- **(a)** Feature analysis based on fixed threshold
- **(b)** Scored-based classification

(III) Machine-learning-based feature analysis

3.1 Verification-Based Feature Analysis

Unlike hardware Trojan detection, verification is the process of ensuring that the circuit design meets the specifications or requirements without any defects or design faults. Errors or bugs are usually placed not on purpose, detectable, and its effects can be observed when the intended purposes of the design are not fulfilled. Unlike design bugs, Trojans are more complicated. Trojans are referred to as intended modification by the adversary, which are always designed to be highly stealthy as to ensure they cannot be detected during the conventional verification process. Some improvements have been made to current verification techniques in order to accommodate for Trojan detection. In the semiconductor industry, verification is vital in manufacturing process. Using improved verification techniques to detect Trojans could be beneficial to manufacturers as these methods can be cost-

effective since they are an extension of conventional verification techniques that are commonly implemented in the semiconductor industry today.

3.1.1 Detection Through Improved IC Design Methodology

Hardware Trojans are usually unpredictable as the adversary will try to modify and enhance the Trojan to attack specific circuit designs. It causes the Trojan detection to be an intensive process and the complexity scales with the circuit size. In order to simplify the detection, Banga & Hsiao (2010) and Zhang & Tehranipoor (2011a) introduced redundant net filtering methods based on verification processes such as automatic test pattern generation (ATPG), code coverage analysis and equivalence analysis. The first objective is to remove any untestable net or stuck-at-fault net from the list of Trojan candidates, because they are assumed not to have any ability to modify the original functionality since, they are not controllable. Even if the nets are Trojan nets, it will be easily recognised as a design fault during conventional verification process. The second objective of the filtering process is to remove high switching frequency nets, as Trojan nets are expected to be stealthy. A list of Trojan candidate nets is identified after the filtering process.

Both Banga & Hsiao (2010) and Zhang & Tehranipoor (2011a) leverage on Trojan detection methodology at gate level, presented in Figures 4 and 5 respectively. A suspect-signal-guided equivalence checking method was introduced by Banga & Hsiao (2010) to detect hardware Trojan. Trojan candidate nets are cross-checked against a circuit generated based on the circuit specification. If a net is unable to give the same behaviour as the specification requested, it is highly suspected as a Trojan net. The suspicious nets are then analysed by using an infected region isolation algorithm to conclude whether they are Trojan nets. For the experiment, 11 ISCAS'89 benchmarks are used. A 10-bit counter Trojan was designed by the authors and inserted into the 11 benchmarks. The experimental results are excellent where 10 out of 11 benchmarks have 100% of isolated gates belonging to Trojan.

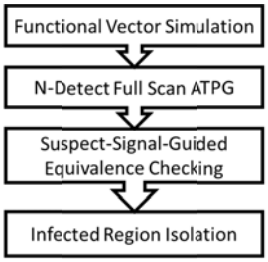


Figure 4: The hardware Trojan detection methodology proposed by Banga & Hsiao (2010).

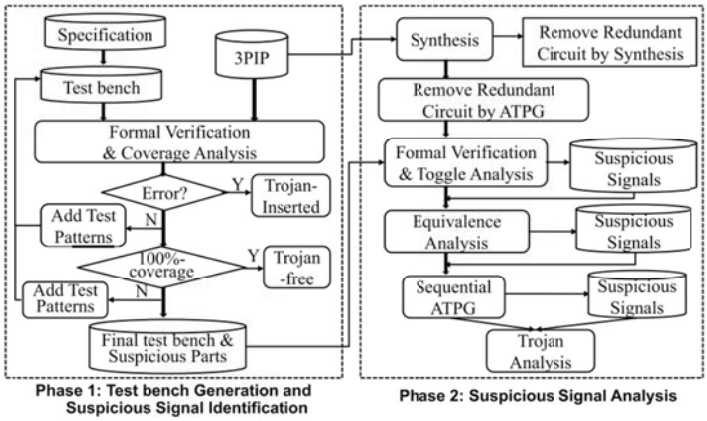


Figure 5: Flow of the Trojan detection methodology proposed by Zhang & Tehranipoor (2011a).

ATPG was also proposed to be used to confirm the existence of Trojan nets by Zhang & Tehranipoor (2011a). After the redundant net filtering, sequential ATPG is applied to trigger the suspicious nets to observe any violation of circuit functionality. For the performance evaluation, 19 RS232 benchmark circuits were used, with different Trojans inserted in each circuit. The authors designed nine of the Trojans, while 10 Trojans are adopted from Trust-Hub. However, only 10 benchmark results were shown. The results indicated that the method is unable to detect Trojans that have no impact on circuit functionality.

Hardware Trojan detection using improved IC design methodology could be beneficial as it does not involve other new and complicated testing or analysis. However, the downside of these methods is that their accuracy is highly dependent on the efficiency of the filtering process. In addition, methods with logic simulation is effective to detect functionality-changing Trojans because these methods directly observe the functionality of the design, but the detection capability and its accuracy is associated with the code coverage of the test pattern, which scales with the required time to derive the test pattern. In order to obtain a higher accuracy detection result, it would require huge amount of test pattern generation time especially when the design is complex.

3.1.2 Assertion-Based Hardware Trojan Detection

Assertion is an expression that indicates whether an event is true or false. When an assertion reports that an event is false during model checking or logic simulation, an unexpected or undesired behaviour is identified. Assertion has been generally used in both computer software and hardware development to ease the verification especially during the debugging process. In computer hardware development, assertion is commonly used for IC verification to identify any behaviour violating the circuit specification or property through logic simulation or model checking. Assertion directly observes the signal behaviour at the point of its insertion, instead of the primary outputs. It improves the observability of a circuit. Hardware Trojans are usually stealthy to bypass the testing process as the testability of Trojans circuitry is very poor. Assertion can tackle such kinds of Trojans, especially in the scenario where the Trojan payload does not propagate to the circuit primary output. In hardware Trojan detection, during logic simulation or model checking, assertion will fail if any violation of property is caused by the Trojan.

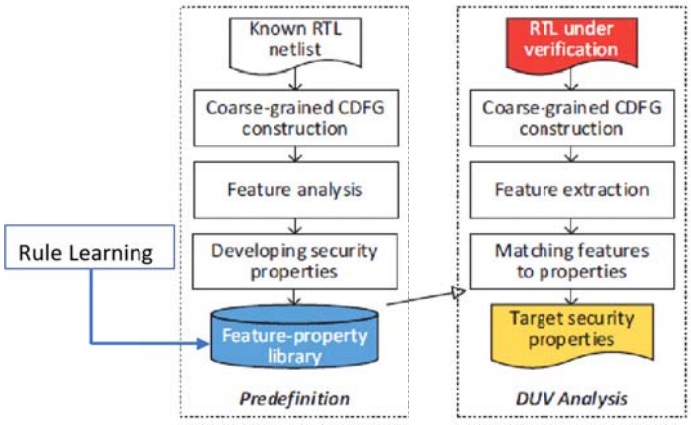


Figure 6: Flow chart of ASPG with rule learning improvement (Source: Wang *et al.*, 2017).

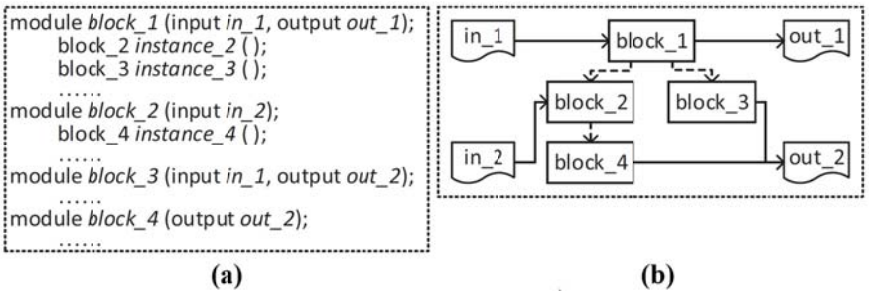


Figure 7: Example of: (a) RTL code (b) Coarse-grained CDFG (Source: Wang *et al.*, 2017).

An assertion-based method for information-leaking hardware Trojan detection at register-transfer level (RTL) was proposed by Wang *et al.* (2017). Figure 6 shows the flowchart of the method, namely automatic security property generation (ASPG). A feature-property library is constructed for supporting the detection. The Trojan features based on the triggering mechanism are first studied and extracted from the existing known Trojan design at coarse-grained control-data flow graph (CDFG). A set of properties are defined based on each feature. A list of the Trojan features and its corresponding properties form the library. To detect if any Trojan exists in an RTL design, the RTL design is first modelled into coarse-grained CDFG (Figure 7), then the Trojan features are extracted from the graph, and the corresponding properties are generated for each found Trojan feature by the library. After the properties are translated into assertion statements and added into the design, the detection will be carried out through model checking. If any assertion is unsatisfied, the Trojan is detected. The library is further improved using a rule learning algorithm to obtain a more reliable detection result that can overcome the future unknown Trojan with its self-learning ability. For benchmarking, 15 Trojan benchmarks from Trust-Hub and Zhang *et al.* (2014) are selected. The authors claimed that the method can detect all Trojan benchmarks and has 0% false positive.

Rajendran *et al*. (2015) proposed a data-modifying hardware Trojan detection method at register-transfer level. The proposed method detects Trojans that modify the critical data stored in registers. "No-data-corruption property" is defined for every critical register, which will be verified by assertion during bounded model checking or ATPG. If a Trojan is detected, the checker reports violation of property and a set of sequence inputs that caused the violation. The authors modified the Trojan structures of nine benchmarks from Trust-Hub, based on Zhang *et al*. (2014). The experimental results showed that eight out of them can be detected. The only undetected Trojan remained inactive within given simulation clock cycles.

The detection performance and accuracy of assertion-based detection can be very high, if a Trojan is successfully activated and causing the assertion to fail. However, for simulation-based detection, its detection performance relies highly on the test bench. If the code coverage of the test bench is too low, the Trojan could not be activated, and thus cannot be detected. In addition, if the test bench is too large, it would be highly time-consuming. Therefore, design of the test bench for Trojan detection is still a challenge. Besides, defining assertion at the right place that can observe Trojan behaviours is another challenge. The property to be checked and location of assertion will determine the detection coverage.

3.1.3 Information-Flow Tracking

Information-flow tracking is a method that is widely used to ensure the security properties related to confidentiality and data integrity. A hardware Trojan detection method using gate-level information flow tracking (GLIFT) was introduced by Hu *et al*. (2016), as illustrated in Figure 8. The proposed method assigns a label to each signal in the gate-level netlist, either as HIGH or LOW. Two logic signal analyses are performed in the proposed method, which are confidentiality analysis that checks for confidential data leaking, while integrity analysis checks for illegally critical data overwriting. The signal classification and labelling examples are shown in Table 1. The GLIFT logics are generated

automatically and mapped to the design together with security properties defined by the designers. The property is based on the rule of "HIGH data should never flow to LOW data". Then formal verification is carried out. If it passes the verification with all properties satisfied, there is no Trojan. In contrary, a counterexample will be generated, which enables functional testing on the GLIFT logics to identify the suspicious behaviour caused by the Trojan. In the experiment, 11 Trojan benchmarks from Trust-Hub were successfully identified. The result showed that the method is effective to detect Trojans that cause undesirable logic information flow.

The limitation of this technique is that it only targets Trojans that violate the security properties by leaking information through logical payload. It also requires effective logic signal analysis or formal verification to detect any suspicious behaviours.

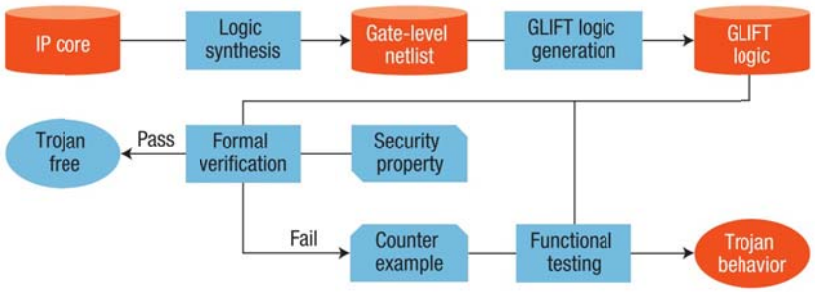


Figure 8: Detection flow of GLIFT (Source: Hu *et al.*, 2016).

Table 1: Examples of signal classification and labelling (Source: Hu *et al.*, 2016).

Confidential Analysis			**Integrity Analysis**		
Data type	Example	Label	Data type	Example	Label
Secret	Plaintext and key	HIGH	Critical	Program counter	LOW
Not Secret	Clock, reset, and start of encryption signal	LOW	Noncritical	Input from open network or keyboard	HIGH

3.1.4 Subgraph Isomorphism

Graph isomorphism has been implemented in layout versus schematic checking, which is one of the traditional verification processes. The purpose of the checking is to ensure that the circuit netlist is equivalent to the design schematic. Some similar approaches based on subgraph isomorphism for hardware Trojan detection methods were also introduced by Shen & Zhao (2017) and Piccolboni *et al.* (2017). These methods search for the gate-level structural features of Trojan within a circuit design to detect the existence of Trojan.

A gate-level hardware Trojan detection method using improved subgraph isomorphism algorithm was introduced by Shen & Zhao (2017). The detection involves graph matching between graphs of reference circuits and subgraphs of circuit-under-verification. The authors designed the Trojan circuits that were used as reference circuits. During the graph matching, if any graph of reference circuits is found within the subgraphs of circuit-under-verification, Trojan is detected. This study focused on the improvement over the traditional subgraph matching techniques and showed its performance improvement in Trojan detection. The detection method uses subgraph isomorphism technique based on directed graphs of gates, as shown in Figure 9. The introduced algorithm shows that it outperforms the traditional algorithm with smaller runtime overhead, and it is claimed to be efficient even with an actual VLSI size circuit. Trojans were successfully located in all seven processor designs used in the experiment. However, this method requires the information of inserted Trojans to detect them, which

is not practical. This method is also unable to detect distributed Trojan whose circuitry is spread throughout the design.

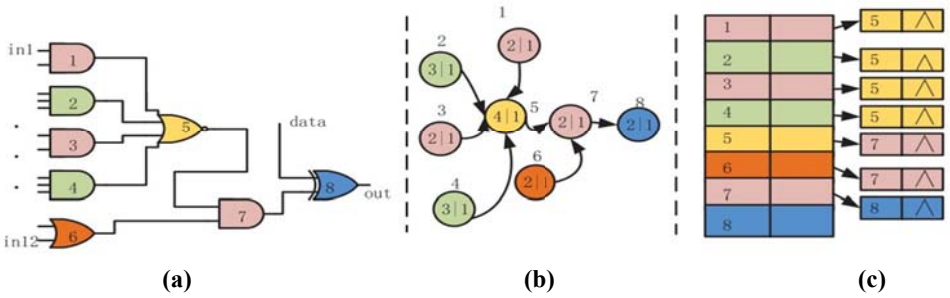


Figure 9: Gate-level hardware Trojan detection method (Source: Shen & Zhao, 2017): (a) Example circuit (b) Directed graph (c) Adjacency list.

Another RTL hardware Trojan detection method using control-flow subgraph matching technique was proposed by Piccolboni *et al*. (2017). The example of RTL to control-flow subgraph modelling is presented in Figure 10. In this proposed method, a Trojan library is constructed, which includes the control-flow subgraphs of wide categories of known Trojans and its variants. By including the variants of the Trojans, the capability of the detection can be enhanced to overcome the alterations of Trojan by different adversaries. The benchmarking showed that all Trojans included in the library can be detected. The weakness is that this method is unable to detect Trojan that is not included in the library. A false positive problem is also found for those Trojans with similar structure to the genuine circuit.

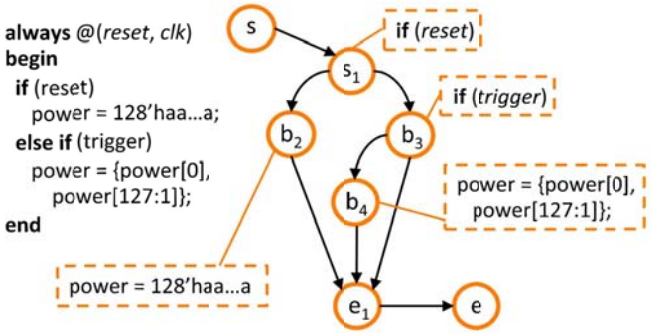


Figure 10: Example of RTL to control-flow graph modelling (Source: Piccolboni *et al.*, 2017).

The limitation of the subgraph isomorphism method is the requirement of reference circuits containing Trojan. This leads to the difficulty of detection when facing future unknown Trojans.

3.2 Threshold-Based Feature Analysis

Feature analysis has been commonly used in software virus detection and it is popular in pre-silicon hardware Trojan detection. Threshold-based feature analysis gives a threshold for each targeted feature. When a feature exceeds the set threshold, the detection is positive.

3.2.1 Feature Analysis Based on Fixed Threshold

A feature-analysis for gate-level hardware Trojan detection at flip-flop level control-data flow graph (CDFG) level was introduced by Yao *et al*. (2015). An example of a flip-flop level CDFG is presented in Figure 11. The features used in the method are gate-level structural features based on triggering mechanism of both combinational and sequential Trojans. The features are extracted from flip-flop level CDFG. A threshold is manually set for each feature, based on the pattern of interconnection of

CDFG nodes. Trojan candidates are determined if any of the features exceeds the set threshold. Although all 16 gate-level Trojans benchmarks from Trust-Hub and Zhang *et al*. (2014) were successfully identified, a high false positive rate problem was found.

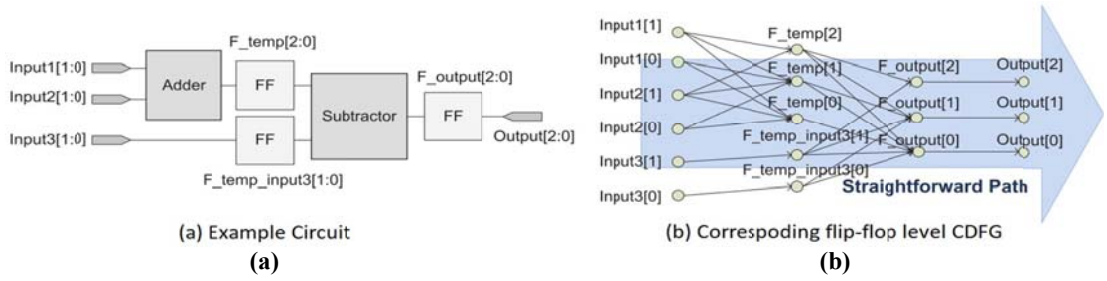


Figure 11: Example of a flip-flop level CDFG (Source: Yao *et al.*, 2015): (a) Example circuit (b) Corresponding flip-flop level CDFG.

A complementary method to Yao *et al*. (2015) was proposed by Chen *et al*. (2018), which used multilevel feature analysis instead of single abstraction level features. Besides using flip-flop level CDFG features, Chen *et al*. (2018) also proposed to analyse the features at combinational logic level information flow graph (IFG). The example of a combinational logic level IFG is shown in Figure 12. The reason to implement multi-level features is because some Trojan information might not be observable in specific abstraction levels. Thus, the detection result is not reliable. With such more refined information at combinational logic level for the analysis, a more accurate result can be obtained. The proposed multi-level analysis flow is presented in Figure 13. The benchmarking showed a reduced false positive rate result as compared to Yao *et al*. (2015). Thus, Trojan detection accuracy can be improved by analysing different level features.

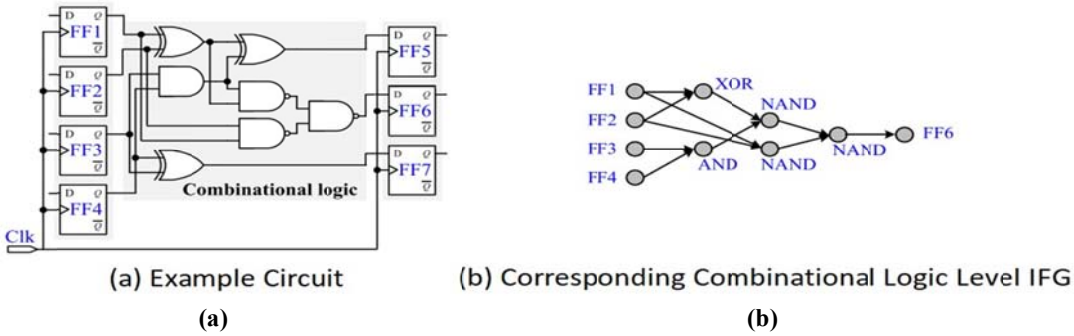


Figure 12: Illustration of a combinational logic level IFG (Source: Chen *et al.*, 2018): (a) Example circuit (b) Corresponding combinational logic level IFG.

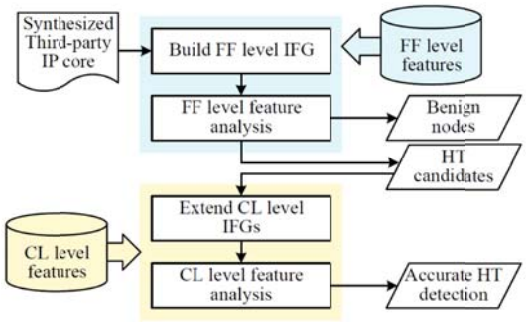


Figure 13: Flowchart of two-level feature analysis (Source: Chen *et al.*, 2018).

3.2.2 Score-Based Classification

A score-based classification is a statistical classification based on a scoring metric. A score is assigned to each feature associated with its weight. After feature analysis, the cumulative score is determined and if it is above the threshold, the result is positive.

A three-phase score-based classification for hardware Trojan detection was proposed by Oya *et al.* (2015), as illustrated in Figure 14. The first phase of detection is weak classification, which calculates the score of a net based on nine gate-level Trojan structural features. The features are extracted directly from the gate-level netlist. The score of a net is increased according to the introduced scoring metric for each Trojan feature matched. The higher the score, the net is more likely a Trojan net. The second phase is strong classification, where another three Trojan factors (max score, max constant cycles and max score net count) are determined for each net with maximum score in the first phase. The last phase is Trojan identification, where the three Trojan factors are compared to a manually set threshold. If all three factors surpass the thresholds, the result is positive, whereby a Trojan is detected. Although the experimental results showed that all 15 Trojan benchmarks from Trust-Hub could be detected, this method is not robust enough to overcome other Trojans, because the features and corresponding thresholds are fixed regardless of the circuits-under-verification. Thus, it might not be suited for other circuits and Trojans. More factors and dynamic thresholds are required.

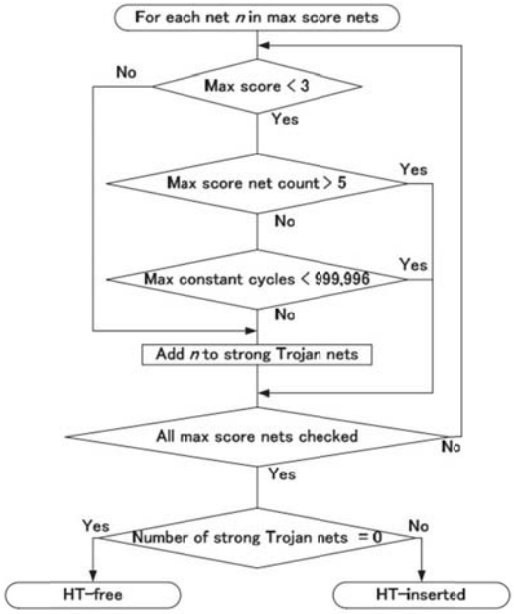


Figure 14: Flowchart of the score-based Trojan detection method proposed by Oya *et al.* (2015).

Another score-based single-triggered hardware Trojan detection method based on triggering mechanism was proposed by Chen & Liu (2017). The method first extracts combinational logic, including AND, OR, NAND and NOR gates, to reduce the complexity of the detection. A scoring algorithm is introduced to analyse the extracted combinational logic gates. A score is determined for each gate. In this method, a proposed score outlier determination algorithm is introduced to dynamically set a threshold to distinguish Trojan gates and non-Trojan gates, based on the score distribution. Any gate with score higher than the threshold will be identified as Trojan gate. In the experiment, 39 out of 40 Trojan benchmarks from Trust-Hub were successfully identified. The authors also claimed that the method can detect most of the Trojans with low false positive rate and small runtime overhead.

The threshold should be dynamic and robust to be applied for different circuits. This is because the feature's correlation with Trojan is design dependent. One feature could be a strong feature for a circuit but a weak feature in different circuits. Additional analysis is required to determine the correlation of feature and Trojan, for a better detection performance.

3.3 Machine Learning-Based Feature Analysis

Machine learning is becoming popular and important in for both industry and our daily lives. It helps in building an analytical model automatically by learning the data without explicit programming. It can further improve and develop by themselves by learning from new data. Due to these advantages, machine learning is also introduced to be implemented in hardware Trojan detection.

Hasegawa *et al.* (2016) introduced a hardware Trojan classification method using machine learning at gate-level netlist. The authors proposed five Trojan net features at gate level, as listed in Table 2. The five Trojan net feature values of every net are considered as the five-dimensional input vectors to support vector machine classifier. The classifier is first trained with the vectors extracted from existing known Trojan netlists (Inoue *et al.*, 2018). Each vector is labelled as Trojan net or normal net. The trained classifier can classify a set of nets of a netlist into Trojan nets and normal nets, without logic simulation required. As the number of Trojan nets are usually relatively small as compared to te total number of nets in a netlist, the authors introduced a dynamic weighting method to balance the training data between Trojan nets and normal nets. The experimental result showed that a true positive rate of 80% was achieved in most cases. However, the true negative rates were quite low, where many normal nets were classified as Trojan nets mistakenly.

Table 2: Trojan net features proposed by Hasegawa *et al.* (2016).

#	Feature	Description
1	*LGFi*	The number of inputs of the logic gates two-level away from the net *n*.
2	*FFi*	The number of logic levels to the nearest flip-flop input from the net *n*.
3	*FFo*	The number of logic levels to the nearest flip-flop output from the net *n*.
4	*PI*	The minimum logic level from any primary input to the net *n*.
5	*PO*	The minimum logic level to any primary output from the net *n*.

Inoue *et al.* (2018) extended the study of Hasegawa *et al.* (2016). The authors designed three types of hardware Trojan based on the triggering mechanism: combinational-triggered Trojan, sequential-triggered Trojan and always-on Trojan. The Trojans were inserted into a RS232 transceiver circuit's netlist. The Trojan classification method of Hasegawa *et al.* (2016) was applied to the netlist. The training of classifier used seven Trojan benchmarks from Trust-Hub. The experimental result showed 100% true positive rate for combinational-triggered Trojan and sequential-triggered Trojan, but always-on Trojan was unable to be detected. The results showed a low true negative rate problem.

Input variables with high correlation are important to build a high accuracy classification model. Hasegawa *et al.* (2017b) proposed a feature extraction at gate-level netlist to look for a set of high correlation features that can better represent Trojans. As shown in Table 3, 51 gate-level features were introduced and analysed to determine the best features within them. Random forest classifier was used to select the best features whose F-measure value was maximised, based on the Trojan benchmarks from Trust-Hub. F-measure is the harmonic mean of precision and recall, which quantifies a test's accuracy of a binary classification. Out of the 51 features, 11 best features were selected as listed in Table 4. To show the performance of the selected features, the 11 best features were then used as input to random forest classifier to detect Trojan. The benchmarking result showed it had high true negative rate. However, some results showed very high true positive rate while others had very low true positive rate.

Table 3: The analysed gate-level features ($1 \leq x \leq 5$) (Source: Hasegawa *et al.*, 2017b).

Trojan Feature	Description
fan_in_*x*	The number of logic-gate fanins up to *x*-level away from the net *n*.
in_flipflop_*x*	The number of flip-flops up to *x*-level away from the input side of the net *n*.
out_flipflop_*x*	The number of flip-flops up to *x*-level away from the output side of the net *n*.
in_multiplexer_*x*	The number of multiplexers up to *x*-level away from the input side of the net *n*.
out_multiplexer_*x*	The number of multiplexers up to *x*-level away from the output side of the net *n*.
in_loop_*x*	The number of up to *x*-level loops.
out_loop_*x*	The number of up to *x*-level loops.
in_const_*x*	The number of constants up to *x*-level away from the input side of the net *n*.
out_const_*x*	The number of constants up to *x*-level away from the output side of the net *n*.
in_nearest_pin	The minimum level to the primary input from the net *n*.
out_nearest_pout	The minimum level to the primary output from the net *n*.
{in,out}_nearest_flipflop	The minimum level to any flip-flop from the input or output side of the net *n*.
{in,out}_nearest_multiplexer	The minimum level to any multiplexer from the input or output side of the net *n*.

Table 4: The proposed best set of 11 gate-level Trojan features (Source: Hasegawa *et al.*, 2017b).

#	Trojan Feature
1	fan_in_4
2	fan_in_4
3	in_flipflop_4
4	out_flipflop_3
5	out_flipflop_4
6	in_loop_4
7	out_loop_5
8	in_nearest_pin
9	out_nearest_pout
10	out_nearest_flipflop
11	out_nearest_multiplexer

Hasegawa *et al.* (2017a) adopted the 11 features previously introduced (Hasegawa *et al.*, 2017b), and fed them as inputs to multi-layer neural networks to classify Trojan nets and normal nets. The experiment used 17 benchmarks from Trust-Hub, and the result showed that three layer neural networks can achieve the best result of 84.8% of average true positive rate and 70.1% of average true negative rate.

A hardware Trojan detection based on controllability and observability value at gate-level netlist was introduced by Salmani (2016). An unsupervised learning algorithm, namely *k*-mean clustering is used. Considering Trojans are always stealthy and hard to be triggered, nets with poor controllability and observability are very likely to be Trojans. The benchmarking results showed high performance with no false positives and false negatives for the benchmarks with obvious difference of combinational controllability (CC) and combinational observability (CO) values between genuine and Trojan nets. However, there is a possibility of Trojan nets whose testability is similar the genuine nets. In this case, the Trojan is not detectable.

4. DISCUSSION & CONCLUSION

Table 5 shows the summary of pre-silicon detection methods reviewed in the previous section. In this section, discussions are made based on the reviews.

Table 5: Summary of reviewed pre-silicon trojan detection methods.

Authors	Method	Features	Detection Level	Target Trojan	Result
Banga *et al*. (2010)	Verification	Switching Frequency	Gate	(unspecified)	(10/11) Isolated region is 100% HT gates
Zhang *et al*. (2011a)	Verification	Switching Frequency	Gate	(unspecified)	(6/10) Suspicious signals are 100% Trojan signals, other > 67.7%
Rajendran *et al*. (2015)	Assertion	Corruption of data	RTL	Critical register corruption	(8/9) HT circuits can be identified, except AES-T1200
				Bypass register	
				Pseudo-critical register	
Wang *et al*. (2017)	Assertion	Connectivity at coarse-grained CDFG	RTL	Information leaking Trojan	(15/15) All Trojan circuits are identified, 0% FPR
Hu *et al*. (2016)	Information flow tracking, Assertion	Switching Frequency	Gate	Information leaking Trojan	11 Trojan benchmarks from Trust-Hub are correctly identified
Shen *et al*. (2017)	Subgraph isomorphism	Connectivity at gate level	Gate	(unspecified)	(7/7) Trojans are successfully located within different processor designs.
Piccolboni *et al*. (2017)	Subgraph isomorphism	Connectivity at control-flow level, Triggers and payloads of Trojan	RTL	(unspecified)	All Trojan benchmarks are correctly identified
Yao *et al*. (2015)	Feature analysis	Connectivity at flip-flop level CDFG, Triggering mechanism of Trojan	Gate	Data-triggered	(16/16) All Trojan circuit can be identified, but FPR found
				Time-triggered	
				Implicit-triggered	
Chen *et al*. (2018)	Feature analysis	Connectivity at flip-flop level CDFG, Connectivity at combinational level CDFG, Triggering mechanism of Trojan	Gate	Data-triggered	Reduced false positive compared to Yao *et al.* (2015)
				Time-triggered	
				Implicit-triggered	
Oya *et al*. (2015)	Score-based classification	Connectivity at gate level, Switching Frequency	Gate	(unspecified)	All (15/15) Trojan and (9/9) clean benchmarks are correctly identified
Chen *et al*. (2017)	Score-based classification	Connectivity at gate level	Gate	Single-triggered	(39/40) Trojan benchmarks are correctly identified
Salmani (2016)	Machine Learning	Testability	Gate	(unspecified)	91.3% TPR
Hasegawa *et al*. (2016)	Machine learning	Connectivity at gate level	Gate	(unspecified)	80% TPR in most cases, Low TNR
Hasegawa *et al*. (2017a)	Machine learning	Connectivity at gate level	Gate	(unspecified)	Average TPR 84.8%, Average TNR 70.1%
Hasegawa *et al*. (2017b)	Machine learning	Connectivity at gate level	Gate	(unspecified)	>98% TNR, but not consistent TPR
Inoue *et al*. (2018)	Machine learning	Connectivity at gate level	Gate	Sequential-triggered	100% TPR for sequential-triggered & combinational-triggered Trojan, Unable to detect always-on Trojan, Low TNR
				Combinational-triggered	
				Always-on	
Kumar *et al*. (2014)	Side-channel analysis	Path delay	Gate	Comparator	(no accuracy is discussed)
				Bypass Trojan (MUX)	
				Trigger-Payload	

Assertion-based methods (Rajendran *et al*., 2015; Hu *et al*., 2016; Wang *et al*., 2017) are claimed to have high detection performance and accuracy. However, the detection outcome depends on the efficiency of assertion or property definition. Assertions must be defined corresponding to Trojan behaviours. This could be a problem, especially when information of inserted Trojan is unknown.

Subgraph-matching methods (Shen & Zhao, 2017; Piccolboni *et al*., 2017) are claimed to have high efficiency in Trojan detection, contributed by its small time overhead and high accuracy. However, these methods require the information of inserted Trojan to match with the suspicious circuit. This may not be practical since in most cases, the inserted Trojan is unknown.

Machine learning for Trojan detection has become popular in recent years. Most existing methods using machine learning are at gate level (Salmani, 2016; Hasegawa *et al*., 2016; Hasegawa et al., 2017a, b; Inoue *et al*., 2018). The performance evaluation of most methods focuses on achieving high sensitivity, where high number of Trojan nets are correctly detected as Trojan, in other words, high true positives. Although many proposed methods have proven to achieve high sensitivity, the problem lies in the false positives and it remains unresolved.

One of the advantages of using machine learning is that it can dynamically identify Trojans based on the learning result. With more test data provided, the classifier can detect more types of Trojan and achieve better detection performance. However, the currently available Trojan benchmark circuits are limited. The lack of training data set could result in high false positive problem. In addition, Trojans that are not included in the training data set are unlikely to be detected by the classifier.
Side-channel analysis is not popular in pre-silicon detection because the lack of golden circuit availability in the pre-silicon stage. Availability of golden circuit is a problem, unless all the previous manufacturing processes are authenticated. In addition, pre-silicon detection targets suspicious design stage, which is a relatively early level in manufacturing steps, so it would be very difficult to obtain a golden circuit before or at the detection level. Recently, most of the Trojan researches are focusing on developing detection method without requiring a golden circuit. On the other hand, without using a golden circuit, the detection accuracy becomes a concern.

Trojan detection at register-transfer level (RTL) is less popular. There is a research gap where we may explore more into Trojan detection by analysing features at RTL. Trojan detection at RTL is a challenge, as RTL is relatively high in the abstraction level and only functionality and signal flow information are described. The information that can be used to identify Trojan are limited.

In conclusion, hardware Trojan detection is gaining more attention in recent years, especially the pre-silicon detection. This is a sign that the community is aware of the threat of hardware Trojan. However, the existing countermeasures of hardware Trojan are not yet extensive enough. Research in hardware Trojan is necessary before hardware Trojan issues becomes widespread.

REFERENCES

Adee, S. (2008). The hunt for: The kill switch. *IEEE Spectrum*, **45**: 34-39.

Agrawal, D., Baktir, S., Karakoyunlu, D., Rohatgi, P. & Sunar, B. (2007). Trojan detection using IC fingerprinting. *IEEE Symp. Secur. Priv.*, 20-23 May 2007.

Alsaiari, U., Gebali, F. & Abd-El-Barr, M. (2017). Programmable assertion checkers for hardware Trojan detection. *Conf. PhD Res. in Microelectron. Electron. Latin America (PRIME-LA)*, 2023 February 2017.

Banga, M. & Hsiao, M.S. (2010). Trusted RTL: Trojan detection methodology in pre-silicon designs. *IEEE Int. Symp. Hardware-Oriented Secur Trust (HOST)*, 13-14 June 2010.

Bao, C., Forte, D. & Srivastava, A. (2014). On application of one-class SVM to reverse engineeringbased hardware Trojan detection. *Int. Symp. Qual. Electron. Design*, 3-5 March 2014.

Bao, C., Forte, D. & Srivastava, A. (2016). On reverse engineering-based hardware Trojan detection. *IEEE Trans. Comput.-Aided Design of I.C. Syst.*, **35**: 49-57.

Bazzazi, A., Manzuri Shalmani, M.T. & Hemmatyar, A.M.A. (2017). Hardware Trojan detection based on logical testing. *J. Electron. Testing*, **33**: 381-395.

Bhunia, S., Hsiao, M.S., Banga, M. & Narasimhan, S. (2014). Hardware Trojan attacks: Threat analysis and countermeasures. *Proc. the IEEE*, **102**: 1229-1247.

Bilzor, M., Huffmire, T., Irvine, C. & Levin, T. (2011). Security checkers: Detecting processor malicious inclusions at runtime. *IEEE Int. Symp. Hardware-Oriented Secur. Trust (HOST)*, 56 June 2011.

Bilzor, M., Huffmire, T., Irvine, C. & Levin, T. (2012). Evaluating security requirements in a generalpurpose processor by combining assertion checkers with code coverage. *IEEE Int. Symp. Hardware-Oriented Secur. Trust (HOST)*, 3-4 June 2012.

Chakraborty, R.S., Narasimhan, S. & Bhunia, S. (2009). Hardware Trojan: Threats and emerging solutions. *IEEE Int. High Level Design Valid. Test Workshop*, 4-6 November 2009.

Chen, F. & Liu, Q. (2017). Single-triggered hardware Trojan identification based on gate-level circuit structural characteristics. *IEEE Int. Symp. Circuits Syst. (ISCAS)*, 28-31 May 2017.

Chen, X., Liu, Q., Yao, S., Wang, J., Xu, Q., Wang, Y., Liu, Y. & Yang, H. (2018). Hardware Trojan Detection in Third-Party Digital Intellectual Property Cores by Multi-Level Feature Analysis. *IEEE Trans. on Comput.-Aided Design of I.C. Syst.*, **37**: 1370-1383.

Forte, D., Bao, C. & Srivastava, A. (2013). Temperature tracking: An innovative run-time approach for hardware Trojan detection. *IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, 18-21 November 2013.

Francq, J. & Frick, F. (2015). Introduction to hardware Trojan detection methods. *Proc. Design, Automation Test in Europe (DATE)*, 9-13 March 2015.

Hasegawa, K., Oya, M., Yanagisawa, M. & Togawa, N. (2016). Hardware Trojans classification for gate-level netlists based on machine learning. *IEEE Int. Symp. On-Line Testing Robust Syst. Design (IOLTS)*, 4-6 July 2016.

Hasegawa, K., Yanagisawa, M. & Togawa, N. (2017a). Hardware Trojans classification for gate-level netlists using multi-layer neural networks. *IEEE Int. Symp. on On-Line Testing Robust Syst. Design (IOLTS)*, 3-5 July 2017.

Hasegawa, K., Yanagisawa, M. & Togawa, N. (2017b). Trojan-feature extraction at gate-level netlists and its application to hardware-Trojan detection using random forest classifier. *IEEE Int. Symp. Circuits Syst. (ISCAS)*, 28-31 May 2017.

Hu, W., Mao, B., Oberg, J. & Kastner, R. (2016). Detecting hardware Trojans with gate-level information-flow tracking. *Comput.*, **49**: 44-52.

Inoue, T., Hasegawa, K., Yanagisawa, M. & Togawa, N. (2018). Designing hardware trojans and their detection based on a SVM-based approach. *IEEE Int. Conf. ASIC (ASICON)*, 25-28 October 2018.

Jin, Y. & Makris, Y. (2008). Hardware Trojan detection using path delay fingerprint. *IEEE Int. Workshop Hardware-Oriented Secur. Trust*, 9 June 2008.

Koley, S. & Ghosal, P. (2015). Addressing Hardware Security Challenges in Internet of Things: Recent Trends and Possible Solutions. *IEEE Int. Conf. UIC-ATC-ScalCom*, 10-14 August 2015.

Kumar, P. & Srinivasan, R. (2014). Detection of hardware Trojan in SEA using path delay. *IEEE Students' Conf. Elect., Electron. Comput. Sci.*, 1-2 March 2014.

Moein, S., Subramnian, J., Gulliver, T.A., Gebali, F. & El-Kharashi, M.W. (2016). Classification of hardware Trojan detection techniques. *Int. Conf. Comput. Eng. & Syst. (ICCES)*, 23-24 December 2015.

Ngo, X.T., Danger, J.L., Guilley, S., Najm, Z. & Emery, O. (2015). Hardware property checker for run-time hardware Trojan detection. *European Conf. Circuit Theory Design (ECCTD)*, 24-26 August 2015.

Oya, M., Shi, Y., Yanagisawa, M. & Togawa, N. (2015). A Score-Based Classification Method for Identifying Hardware-Trojans at Gate-Level Netlists. *Design, Automation & Test in Europe Conf. & Exhibition (DATE)*, 9-13 March 2015.

Piccolboni, L., Menon, A. & Pravadelli, G. (2017). Efficient Control-Flow Subgraph Matching for Detecting Hardware Trojans in RTL Models. *ACM Trans. Emb. C. Syst. (TECS)*, **15**: 137.

Potkonjak, M., Nahapetian, A., Nelson, M. & Massey, T. (2009). Hardware Trojan horse detection using gate-level characterization. *ACM/IEEE Design Automation Conf.*, 26-31 July 2009.

Rajendran, J., Vedula, V. & Karri, R. (2015). Detecting malicious modifications of data in third-party intellectual property cores. *ACM/EDAC/IEEE Design Automation Conf. (DAC)*, 8-12 June 2015.

Salmani, H. (2016). COTD: Reference-free hardware Trojandetection and recovery based on controllability and observability in gate-level netlist. *IEEE Trans. Inform. Forensics Secur.*, **12**: 338-350.

Shakya, B., He, T., Salmani, H., Forte, D., Bhunia, S. & Tehranipoor, M. (2017). Benchmarking of hardware Trojans and maliciously affected circuits. *J. Hardware Syst. Secur.*, **1**: 85-102.

Shen, H. & Zhao, Y. (2017). HTChecker: Detecting hardware trojans based on static characteristics. *IEEE Int. Symp. Circuits Syst. (ISCAS)*, 28-31 May 2017.

Tehranipoor, M. & Koushanfar, F. (2010). A survey of hardware Trojan taxonomy and detection. *IEEE Design & Test of Comput.*, **27**: 10-25.

Trust-Hub. Available online at: http://trust-hub.org/ (Last access date: 2 March 2019).

Wang, C., Cai, Y. & Zhou, Q. (2017). Automatic security property generation for detecting information-leaking hardware Trojans. *IEEE Int. Conf. Comput. Design (ICCD)*, 5-8 November 2017.

Wang, C., Li, J., Yu, M. & Wang, J. (2013). An intelligent classification method for Trojan detection based on side-channel analysis. *IEICE Electron. Express*, **10**: 20130602.

Xiao, K., Forte, D., Jin, Y., Karri, R., Bhunia, S. & Tehranipoor, M. (2016). Hardware Trojans: Lessons learned after one decade of research. *ACM Trans. Design Automation of Electron. Syst. (TODAES)*, **22**: 6.

Yao, S., Chen, X., Zhang, J., Liu, Q., Wang, J., Xu, Q., Wang, Y. & Yang, H. (2015). FASTrust: Feature analysis for third-party IP trust verification. *IEEE Int. Test Conf. (ITC)*, 6-8 October 2015.

Zhang, J., Yuan, F. & Xu, Q. (2014). DeTrust: Defeating hardware trust verification with stealthy implicitly-triggered hardware Trojans. *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 3-7 November 2014.

Zhang, X. & Tehranipoor, M. (2011a). Case study: Detecting hardware Trojans in third-party digital IP cores. *IEEE Int. Symp. Hardware-Oriented Secur. Trust (HOST)*, 5-6 June 2011.

Zhang, X. & Tehranipoor, M. (2011b). RON: An on-chip ring oscillator network for hardware Trojan detection. *Design, Automation Test Europe*, 14-18 March 2011.

DEVELOPMENT OF SPACE COMMUNICATION, NAVIGATION AND SURVEILLANCE (CNS) SOLUTIONS FOR MILITARY APPLICATIONS

Dimov Stojce Ilcev

Space Science Center (SSC), Durban University of Technology (DUT), Durban, South Africa

E-mail: ilcev@dut.ac.za

ABSTRACT

This paper introduces modern projects and solutions of space systems for military and civilian Communication, Navigation and Surveillance (CNS) at sea, on the ground and in the air. Deploying new CNS technologies and techniques, military mobile assets, such as ships, land vehicles and aircrafts, including their personnel, will be connected, controlled and managed in more tactical, reliable and safe ways. The new satellite communication and navigation transponders onboard Geostationary Earth Orbit (GEO) satellites will able to provide connection and control of all military vehicles and personnel, enhance traffic control and management, improve safety and security of movements, and augment collision avoidance, especially for navy and air forces assets. The integrated CNS components for military applications described in this paper are solutions that provide reliable Mobile Satellite Communications (MSC), Digital Video Broadcasting-Return Channel via Satellite (DVB-RCS) and augmentation of Global Navigation Satellite Systems (GNSS). Separately, modern satellite CNS applications for the navy, ground and air forces with their technical concepts and advantages are discussed.

Keywords: *Communication, Navigation and Surveillance (CNS); Geostationary Earth Orbit (GEO); Mobile Satellite Communications (MSC); Digital Video Broadcasting-Return Channel via Satellite (DVB-RCS); Global Navigation Satellite Systems (GNSS).*

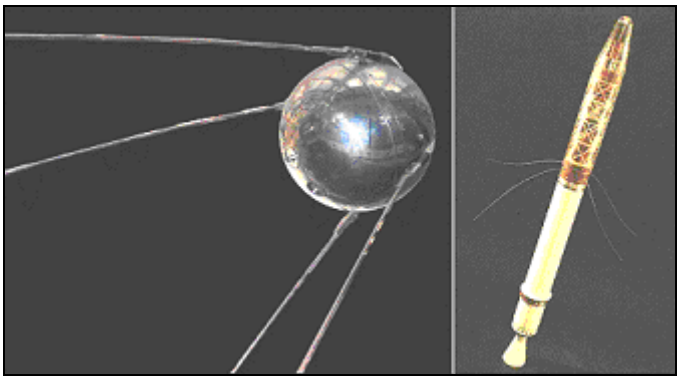
1. INTRODUCTION

The main objective of this paper is to carry out a general overview of the current space Communication, Navigation and Surveillance (CNS) systems via Geostationary Earth Orbit (GEO) satellite constellations for all mobile civilian and military applications. In addition, this paper shortly introduces non-GEO satellite networks with the objective of upgrading some civilian and military applications with more cost effective solutions for ships, land vehicles (road and rail) and aircrafts. For instance, space CNS is very important for improvement of awareness, safety and security for ships navigating in open oceans, sea passages, cramped channel strips, coastal waters, approaching anchorages and inside seaports. For instance, a more critical scenario is when ships are sailing across oceans and coastal areas in very bad weather conditions, where visibility and audibility are zero, where sometimes is not possible to use surveillance radar, because of bad signal propagation, and in situations where it is very difficult to detect surrounding ships for collision avoidance. On the other hand, space CNS systems are improving safety and security of aircrafts in flights over oceans, continents and in all other phases, including takeoff, precision landing capabilities and collision avoidance in the air and on the ground in airports. Deploying CNS systems at seaports and airports can provide more secure and safe traffic control for all vehicles movements. New CNS systems will also improve traffic control and management of land vehicles and especially for signalization (Ilcev, 2012).

Therefore, the development of modern reliable and precise systems for communication, positioning, surveillance and time measurement is crucial for sea, land and air military operations, where the

effects from diverse sources must be coordinated in any real time and space. New CNS techniques integrated with GNSS receivers and satellite transceivers will be able to provide users with the ability to detect, locate and track civilian and military assets and personnel via GEO satellites and new projected small Low Earth Orbit Satellites (LEO) satellite constellations (Ilcev, 2013).

The satellite communication and navigation era began when the Soviet Union shocked the globe with the launch of the ever first artificial satellite, Sputnik-1, on 4 October 1957, which is illustrated in Figure 1 (a). This initiative was followed up on 31 January 1958 with the launch of the US satellite Explorer-1, illustrated in Figure 1 (b). Explorer contained a cosmic ray detector, radio transmitter and temperature and micrometeoroid sensors. In such a way, the development of launchers and satellite systems for future cosmic explorations started, and the space race began (GEOLSOC, 2018).



(a) **(b)**

Figure 1: (a) Sputnik-1 and (b) Explorer-1 satellites (Source: GEOLSOC, 2018; Breitman, 2018).

After the launch of the Sputnik-1 and Explorer-1 satellites, a sustained effort by the US to catch up with the USSR space program was started. This was reflected with the first active communications satellite, named SCORE, which was launched on 18 December 1958 by the US Air Force. The second satellite, Courier, was launched on 4 October 1960 in the High Elliptical Orbit (HEO) with its perigee at about 900 km and its apogee at about 1,350 km using solar cells and frequency of 2 GHz. The next challenge was the third US satellite, Telstar-1, designed by Bell Telephone Laboratories experts and launched by the National Aeronautics and Space Administration (NASA) on the 10 July 1962 in HEO configuration with its perigee at about 100 km and apogee at about 6,000 km. The world's first GEO satellite, Syncom-1, was launched by NASA on 14 February 1963. While this satellite failed during launch, Syncom-2 and 3 were successfully placed in orbit on 26 July 1963 and 19 July 1964, respectively (Breitman *at al*., 2018).

The first commercial GEO satellite was Early Bird (renamed as Intelsat-1), which was developed by Comsat for Intelsat, while the first formal military satellite program in the US, Weapon System 117L, was developed in the mid-1950s. Within this program, a number of sub-programs were developed, including the Corona satellite program, which carried different code names and continued until 25 May 1972. There have also been a number of subsequent programs, including Canyon (seven launches between 1968 and 1977) and other programs. The USSR (today Russia) began the Almaz (Russian: Алмаз) satellite program in the early 1960s. This program involved placing space stations in Earth orbit as an alternative to satellite constellations. In fact, three stations were launched between 1973 and 1976: Salyut-2, Salyut-3 and Salyut-5 respectively (Ilcev, 2019).

2. MARISAT MILITARY MARITIME MSC SYSTEM

The world's first maritime Mobile Satellite Communication (MSC) Marisat system developed for military applications was unveiled in 1976 with only three GEO satellites and ocean networks that provide MSC services in the Atlantic, Pacific and Indian Oceans. The Hughes Aircraft Company,

known today as Boeing Satellite Systems Inc., under contract to the US Comsat General Corp, built three multi-frequency communications spacecraft, known as Marisat (Maritime Satellite), for the space segment of the world's first MSC operator. In 1971, frequency bands around 1.6 GHz were allocated for satellite communications for military ships (navy) and aircrafts (air force). The Marisat satellites were designed initially for US Navy vessels and they had an Ultra High Frequency (UHF) transponder onboard in the band of 240 - 400 MHz. As there was sufficient margin for additional payload, L- and C-band transponders were installed on the Marisat satellite to provide commercial MSC traffic for maritime applications (Seedhouse, 2012).

These satellites had a dual role at that time: to provide space segment facilities that were leased to the US Navy for military communications with naval ships and it also enabled the use of transponders for Comsat General itself to operate MSC for traffic with merchant ships virtually worldwide. All three Marisat satellites were launched with the same type of US rocket, McDonnell Douglas 2914 Delta, in 1976, on 19 February, 9 June and 14 October, for the needs of Comsat General. The Marisat-F1, F2 and F3 satellites were placed in GEO satellite planes at 15 ˚W, 72.5 ˚E and 176.5 ˚E longitudes respectively.

All the satellites have been leased from Comsat in effect; the Marisat-F1 spacecraft served as an in-orbit spare for the maritime European Communications Satellite (ECS) or Marecs-A spacecraft in the Marisat Atlantic Ocean Region (AOR) constellation at position of 15 ˚W. Then, this satellite was leased as a spare in the Inmarsat Atlantic Ocean Region – East (AOR-E) and moved to 106 ˚W. In addition to the Marisat network, Marecs (1973) and Inmarsat (1979) were developed at first for maritime applications only. However, in the next stage, Inmarsat also developed mobile satellite services for land, aeronautical and semi-fixed civilian, and later military applications.

The Marisat-F2 spacecraft served as a spare in Indian Ocean Region (IOR) at 73 ˚E and later, this satellite was leased for the Inmarsat space segment. The Marisat-F3 spacecraft served at 176.5 ˚E in the Pacific Ocean Region (POR), and afterwards reprogrammed as a spare for the Marecs-B2 satellite. This satellite was finally relocated as an in-orbit spare at 182.5 ˚E (Comsat, 2018).

The Marisat service at that time was welcomed by merchant shipping companies and by 1982, around 1,000 vessels were equipped to use the Marisat system. All three Marisat satellites also served as an emergency backup, one in each of the three ocean regions: AOR, IOR and POR. After many years of service, these satellites are no longer in use either by the Comsat or Inmarsat systems (Ilcev, 2013; Launius, 2015; Comsat; 2018; Brown, 2019).

The navy and air force fleets used P-band frequencies (P) for transmitter (Tx) at 248 - 260 MH and for receiver (Rx) at 130 - 312 MHz for the military links. The merchant oceangoing shipping element of the L-band Marisat payload used the newly allocated L-band frequencies (L) for Tx at 1,537 – 1,541 MHz and Rx at 1,638.5 – 1,642.5 MHz) as the service links. C-band (C) for Tx at 6174.5 – 6424 MHz and Rx at 3945.5 – 4199 MHz were used for both military and civilian applications as the feeder links (Comsat, 2018).

Fixed Ground Earth Station (GES) infrastructures for mobile services were located at Santa Paola, US, for POR (GES 1), at Southbury, US for AOR (GES – 2) and at Fucino and Yamaguchi, Italy, for IOR (GES – 3), which is shown in Figure 2. The system provided access to the satellites, linking ships at sea through the Public Switched Telephone Network (PSTN) with the Terrestrial Telecommunication Network (TTN) subscribers ashore for telephone, telex, fax, data and High Speed Data (HSD) transmissions. The Marisat system was controlled by the Network Control Center (NCC) located at Washington. Satellite Tracking, Telemetry and Command (TT&C) was also conducted over C-band frequencies (Ilcev, 2012; Launius, 2015).

However, governments in many other countries were not quite content that some foreign commercial or military corporations controlled the MSC service of merchant vessels managed by their shipping companies. Owing to this problem, in 1976 under the aegis of the International Maritime Organization

(IMO) an agreement was drawn up for the establishment of an Inmarsat organization, initially for maritime service only (Ilcev, 2013).

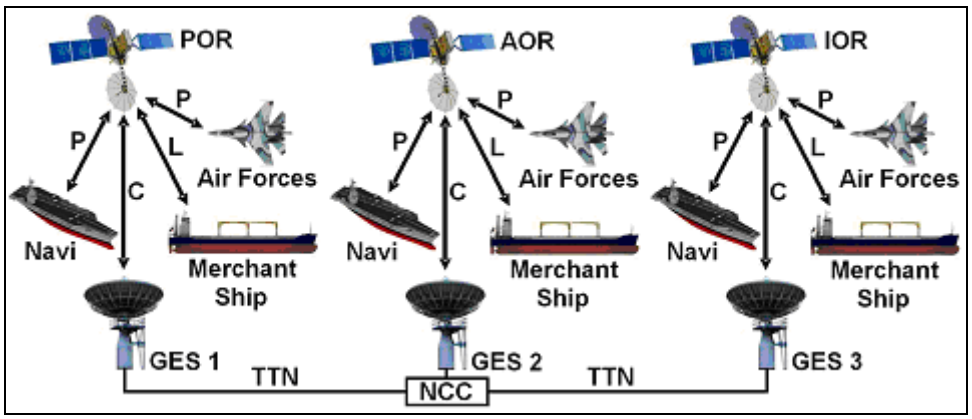


Figure 2: Marisat space and ground segments (Source: Ilcev, 2013).

3. MILITARY SPACE COMMUNICATION NETWORKS

All modern civilian and military MSC networks for GNSS augmentation and CNS systems can utilize GEO satellite constellations only, via additional communication and navigation transponders, such as Wide Area Augmentation System American (WAAS), European Geostationary Navigation Overlay Service (EGNOS), Japanese Multi-Functional Transport Satellite (MTSAT) serving MTSAT Satellite-based Augmentation System (MSAS), and other augmentation systems of the US Global Positioning System (GPS) and Russian Global Navigation Satellite System (GLONASS) networks. However, Medium Earth Orbit (MEO) MSC networks (multinational O3b) and LEO MSC networks, such as Big LEO (the US Iridium and Globalstar), Little LEO (the US Orbcomm and Russian Gonets), and HEO (the Russian Molniya) can be used for Communication, Positioning and Tracking (CPT) of all mobile assets. In fact, Big and Little LEO satellites are applicable for CPT only, but not for augmentation of GNSS, because they have not enough room in satellite payloads to include GNSS and special communication transponders for CNS and augmentation of GNSS networks.

In the meantime, GEO fixed and mobile satellite systems were developed, such as Inmarsat, Intelsat, Eutelsat and other satellite networks, including a combination of these constellations in Hybrid Satellite Orbits (HSO). These MSC networks are providing civilian and military applications as well.

The Inmarsat MSC system has been supporting navy sailors, ground troops including ground vehicles with personnel, and air force staff to carry out their duties via GEO satellite constellations for military operations for many years. In such a way, morale is recognized as essential to the success of military deployments, personnel and crew retention. The range of MSC services gives individuals a choice of ways to stay in touch with their families and friends. In addition, correct storage and timely distribution of different material to meet the mission commander's requirements is a critical component of any military campaign. Reliable communication of the latest information underpins operational readiness.

Multi-military forces today can operate in a complex, multinational environment with responsibility for sea, land and air military assets deployed to any region of the world with reliable MSC networks and equipment. For instance, effective and reliable global satellite communications are recognized as a key enabler for modern military MSC, with an example of hypothetical space and ground segments shown in Figure 3. Here, it is possible to connect through interactive (two-way) MSC systems all types of military forces at sea, land and air through GES terminals and GEO satellites. In fact, seagoing vessels of navies, all vehicle types and troops of ground forces, as well as all kinds of aircrafts serving air forces are carrying specialized MSC transceivers that provide mutual satellite

Voice, Data and Video (VDV) communications via GES terminals in the footprint of GEO satellite constellations (Ilcev, 2018).

For this purpose, Inmarsat network provides service with various types of MSC devices, such as Broadband Global Area Network (BGAN), which can be used as portable, transportable or installed onboard vehicles; Global Xpress, which can be installed onboard Navy and Air Force units; other Inmarsat MSC standards, such as maritime FleetBroadband and Fleet One; land mobile, such as BGAN and mini-M; and aeronautical, such as SwiftBroadband. In addition, the Inmarsat, Iridium and Globalstar satellite networks are providing global telephone services via handheld or portable telephones, which can also be installed onboard commercial and military ships, land vehicles and aircrafts (Kaplan *at al*., 2017; Ilcev, 2018; Inmarsat, 2018).

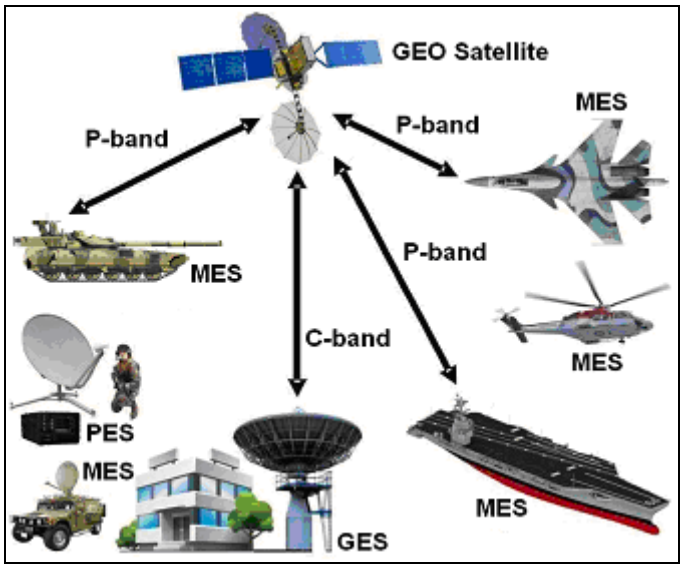


Figure 3: Military MSC space and ground segments (Source: Ilcev, 2018).

A new mobile and fixed satellite service was developed more than two decades ago using only GEO satellite constellations, commonly known as Digital Video Broadcasting-Return Channel via Satellite (DVB-RCS) standards. Despite its unusual name, DVB-RCS is actually all about implementing fast, efficient and reliable Internet Protocol (IP) networks using Very Small Aperture Terminals (VSAT) or larger ones, if necessary. It also provides excellent support for VSAT mobility for civilian and military applications. In fact, high-throughput Communications-on-the-Move (COTM) is delivered with roaming across beams and countermeasures against interference for ships, off-road vehicles and aircraft. The DVB-RCS equipment works with any type of electronic array or stabilized antenna system in any satellite band or beam configuration (Ilcev, 2013).

4. MILITARY SPACE NAVIGATION NETWORKS

The US GPS and Russian GLONASS are the first generation of Global Navigation Satellite Systems (GNSS), known as the GNSS-1 network, while the new upcoming Chinese BeiDou (Compass) and European Galileo are the GNSS-2 network. Both GNSS networks are providing civilian or military transport requirements of high operating Integrity, Continuity, Accuracy and Availability (ICAA) for space navigation systems utilizing precise range measurements and determination of Position, Velocity and Time (PVT) to onboard mobile GNSS receivers, which for ground controllers are able to determine Identification Number (ID), call sign, altitude and another positioning data anywhere in the world, whereby the space and ground segments are illustrated in Figure 4.

The current GNSS-1 networks are providing military and commercial maritime, land vehicle, aeronautical, portable and personal solutions via Intermediate Circular Orbit (ICO) or MEO satellite constellation, with highly accurate worldwide three-dimensional, common-grid, positioning and location data, velocity and precise timing with accuracies that have not previously been easily attainable. In the meantime, China is developing its own BeiDou system, while Europe is developing Galileo system, which will be fully operational in the near future. On the other hand, regional satellite navigation systems, such as Quasi-Zenith Satellite System (QZSS) and Indian Regional Navigation Satellite System (IRNSS) are not designed to provide providing global coverage. The Japanese QZSS is a four-satellite regional time transfer system and a satellite-based augmentation system developed to enhance the GPS coverage in the Asia-Oceania region, with a focus on Japan, while the IRNSS is an autonomous regional satellite navigation system that provides accurate real-time positioning and timing services extension of about 1,500 km around India (Del Re & Ruggieri, 2008).

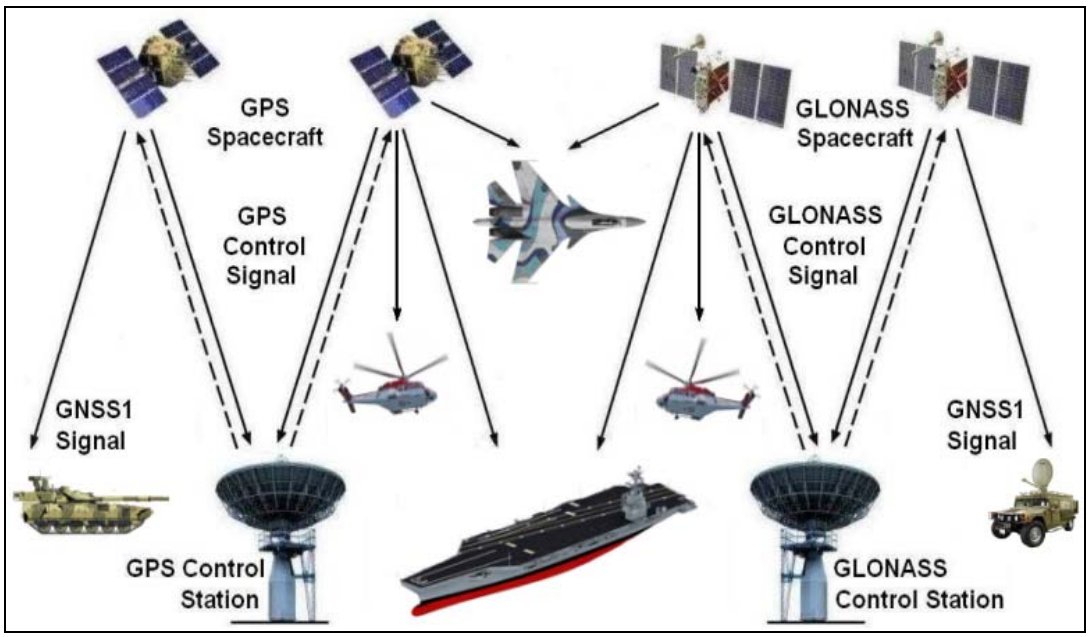


Figure 4: GNSS space and ground segments (Source: Ilcev, 2012).

The GNSS network service is based on the concept of triangulation from known points similar to the technique of "resection" used with a map and compass, except that it is done with radio signals transmitted by satellites. The GNSS receiver must determine when a signal is sent and the time it is received. Nothing except onboard GNSS receivers is needed to use the system, which does not transmit any signals and thus, they are not electronically detectable (Ilcev, 2013; Kaplan *at al*., 2017).

At the beginning of this millennium, two GNSS augmentation networks of Regional Satellite Augmentation Systems (RSAS) were developed, which are the US WAAS and Japanese MSAS networks. The next similar project, the European EGNOS is the first European real venture into satellite navigation systems, which augment the two GNSS-1 network of GPS and GLONASS, and make them suitable for CNS and safety critical applications of mobile movements. The RSAS network will in the future be able to support GNSS-2 networks as well, when Galileo and BeiDou became fully operational in 2020. On the other hand, the Wide Area GPS Enhancement (WAGE) is a GNSS augmentation system operated by the US Department of Defense (DOD) for use by military and authorized receivers (Nejat, 1992; Ilcev, 2013).

The Global Satellite Augmentation Satellite (GSAS) network will integrate three operational RSAS networks, namely WAAS, EGNOS and MSAS, as shown in Figure 5. Newly developed RSAS

networks include the Russian System of Differential Correction and Monitoring (SDCM), Chinese Sino Navigation Augmentation System (SNAS), and Indian GPS/GLONASS and GEOS Augmented Navigation (GAGAN). The new African Satellite Augmentation System (ASAS) project will be a RSAS network for the Africa and Middle East (Ilcev, 2012).

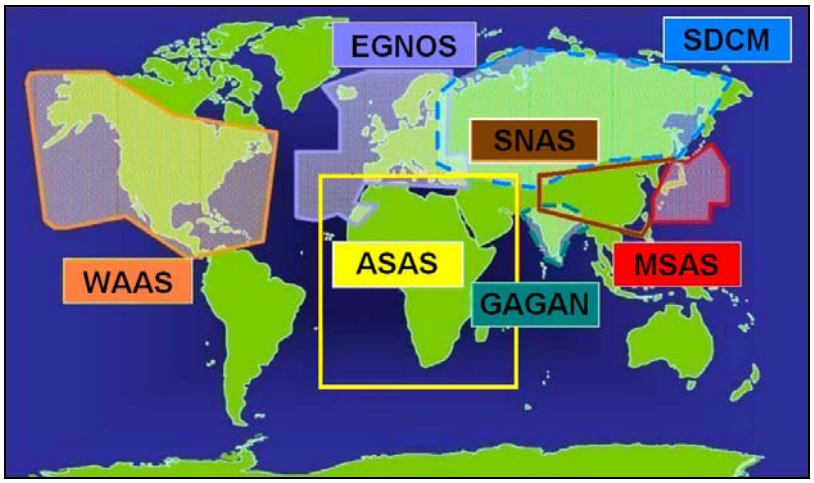


Figure 5: Integration of RSAS networks (Source: Ilcev, 2012).

4.1 Main GNSS Error Sources

The augmentation of an GNSS networks is a method of improving the navigation system's attributes, such as ICAA facilities, through the integration of external information into the calculation process. There are many systems based on how the GNSS sensor receives the external information. First systems transmit additional information about sources of error, such as clock drift, ephemeris or ionospheric delay, and second systems provide direct measurements of how much the signal was off in the past, while a third systems provide additional vehicle information to be integrated in the calculation process.

1. Satellite Clocks – The atomic clocks in the GNSS satellites are very accurate, but they do drift a small amount, so this inaccuracy in the satellite clock results a significant error in the position calculated by the GNSS receiver. For example, 10 ns of clock error results in 3 m of position error. The clock in the satellite is monitored by the GNSS ground control system and compared to the even more accurate clock used in the ground control system. In the downlink data, the satellite provides the user with an estimate of its clock offset. Typically, the estimate has an accuracy of about ±2 m, although the accuracy can vary between different GNSS systems. To obtain a more accurate position, the GNSS receiver needs to compensate for the clock error. The precise satellite clock information contains corrections for the clock errors that are calculated by the RSAS or Precise Point Positioning (PPP) system.

2. Orbit Errors – The GNSS satellites travel in very precise and well known orbits, but they do vary a small amount, with a small variation in the orbit resulting in a significant error in the position calculated. The GNSS ground control system continuously monitors the satellite orbits, so when they change, the ground control system sends a correction to the satellites and the satellite ephemeris is updated. Even with the corrections from the GNSS ground control system, there are still small errors in the orbit that can result in up to ±2.5 m of position error. One way of compensating for satellite orbit errors is to download precise ephemeris information from an RSAS system. Another way of compensating for satellite orbit errors is to use a Differential GNSS (DGNSS) or Real-Time Kinematics (RTK) receiver configuration (Kaplan *at al.*, 2017).

3. Ionospheric Delay - The ionosphere is the layer of atmosphere between 80 and 600 km above the Earth that contains electrically charged particles called ions. These ions delay the satellite signals and can cause a significant amount of satellite position error, typically ±5 m, but can be more during periods of high ionospheric activity. The onboard mobile GNSS receivers use L1 and L2 frequency bands as their advantage, whereby by comparing the measurements for L1 to the measurements for L2, the receiver can determine the amount of ionospheric delay and remove this error from the calculated position. In addition, the base station and rover receivers experience very similar delay, so this allows DGNSS and RTK systems to compensate for ionospheric delay (Kaplan *at al*., 2017; NovAtel, 2018).

4.2 Additional GNSS Error Sources

Other GNSS vulnerabilities also cause errors, including:

1. Tropospheric Delay – The troposphere is the layer of atmosphere closest to the surface of the Earth. Variations in tropospheric delay are caused by the changing humidity, temperature and atmospheric pressure in the troposphere. Since tropospheric conditions are very similar within a local area, the base station and rover receivers experience very similar tropospheric delay. This allows DGNSS and RTK systems to compensate for tropospheric delay. GNSS receivers can also use tropospheric models to estimate the amount of error caused by tropospheric delay.

2. Receiver Noise – Receiver noise refers to the position error caused by the GNSS receiver hardware and software. High end GNSS receivers tend to have less receiver noise than lower cost GNSS receivers.

3. Jamming - Jamming is intentional transmission in the GNSS frequency bands aiming to block the satellite signals at the user antenna. Thus, GNSS jamming over short distances can easily be done by using low-cost jammers that are hard to detect, while GNSS jamming over larger distances beyond Line-of-Sight (LOS) requires more effort and is easier to detect. Using Controlled Radiated Pattern Antennas (CRPA) will reduces the risk. On the other hand, the unintentional transmission of signals is distrupting the reception of the GNSS signals, which can be in-band or out-band interference.

4. Spoofing – GNSS spoofing is done by transmitting a fake GNSS signal to fool the reference system to believe it is in a different position. To spoof a receiver, an adversary needs to faithfully recreate the signals from multiple satellites and then transmit that spoofing signal to capture a local GNSS receiver. If the targeted GNSS receiver is unable to tell the difference between the real satellite signals and the spoofed signals, the spoofing will fool the target receiver into appearing to be at a different location. Mitigation of spoofing is computationally complicated or limited to a specific spoofing scenario. A new approach uses a two-antenna array to steer a null toward the direction of the spoofing signals, taking advantage of spatial filtering, and the periodicity of the authentic and spoofing signals. It requires neither antenna-array calibration nor a spoofing detection block, and can be employed as an inline anti-spoofing module at the input of conventional GNSS receivers (Kaplan *at al*., 2017).

5. Multipath – Multipath occurs when a GNSS signal is reflected off an object, such as the wall of a building or some effects construction elements onboard mobiles, to the GNSS antenna. As the reflected signal travels farther to reach the antenna, the reflected signal arrives at the receiver slightly delayed. This delayed signal can cause the receiver to calculate an incorrect position. The simplest way to reduce multipath errors is to place the GNSS antenna in a location that is away from the reflective surface objects. When this is not possible, the GNSS receiver and antenna must deal with the multipath signals. Long delay multipath errors are typically handled by the GNSS receiver, while short delay multipath errors are handled by the GNSS antenna. Due to the additional technology required to deal with multipath signals, high end GNSS receivers and antennas tend to be better at rejecting multipath errors (Kaplan *at al*., 2017; Kongsberg, 2019).

5. MILITARY CNS SYSTEMS

CNS systems and equipment are the main functions that form the infrastructure for traffic control and management at sea, on the ground and in the air. The CNS system has to ensure that the traffic of navy, ground and air force vehicles is safe, efficient and easy manageable. Military mobile assets will experience the benefits of having quality GNSS navigation information available to support their movements and tactical missions in real time and space. In such a way, another mission for improvements in situational awareness is the new Communication, Navigation and Surveillance facilities integrated in CNS systems.

5.1 Satellite Communication Subsystem (SCS)

Most current communications between mobiles and traffic controllers are conducted via Very High Frequency (VHF), Ultra High Frequency (UHF) and High Frequency (HF) Radio Frequency (RF) bands, which in some busy portions of the world are reaching their limits and which propagation is sometimes problematic because of different atmospheric effects. In fact, the conventional RF-bands are congested and additional frequencies are not available. However, to improve the communication and traffic control facilities of all mobile assets, civilian MSC networks were implemented more than 35 years ago, whereby transmission takes shorter time, less propagation problems and is able to handle more information than conventional radio systems alone. Before that, the military maritime MSC system Marisat was unveiled in 1976 by the US Comsat General with only three satellites and networks in the Atlantic, Pacific and Indian oceans.

Figure 6 depicts a modern military MSC network for navy, ground and air forces using L/C-band. Military MSC can also use UHF, S, X, Ka and Ka bands between Ship Earth Station (SES), Vehicle Earth Station (VES), Aircraft Earth Station (AES) or Transportable Earth Station (TES), and Military Control Centers. As illustrated, all military vehicles and troops will be able to communicate via voice, fax, low / medium / high speed data, telex and video facilities (Ilcev, 2013).

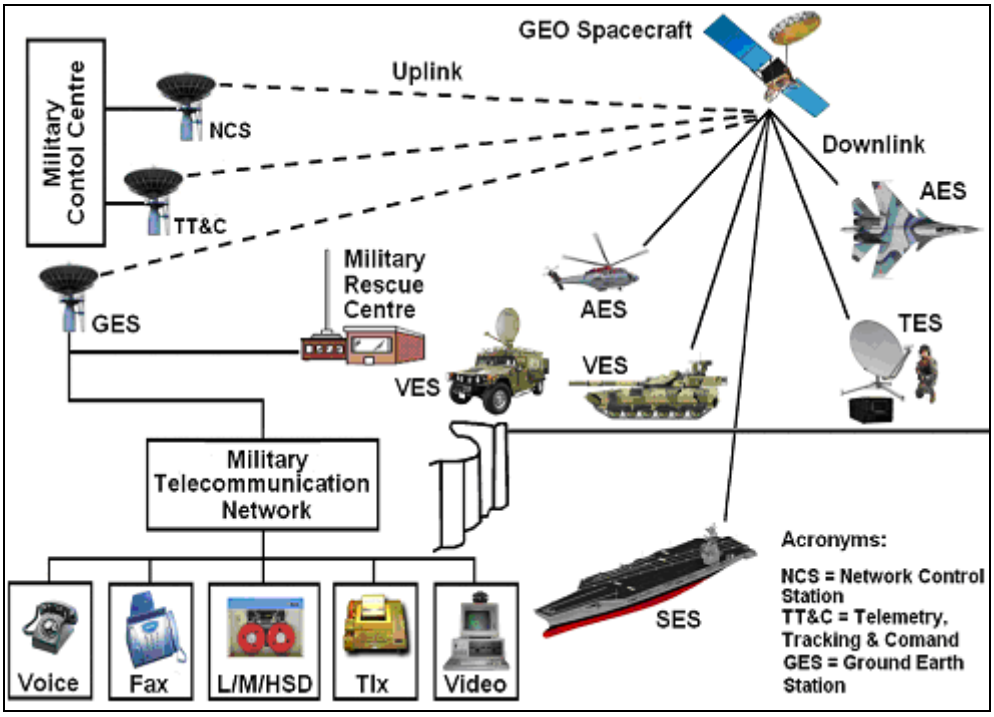


Figure 6: Military SCS (Source: Ilcev, 2012).

The MSC systems are designed not only to provide more cost effective, reliable, redundant and fast communication links between sea, ground and air mobile assets, but also to integrate GNSS data for implementing new services for enhanced navigation and surveillance solutions. At this point, the convergence of MSC and Internet techniques has opened many opportunities to deliver new multimedia services over hybrid satellite systems to Mobile Earth Terminals (MES) stations. Thus, with the need for increased bandwidth capability, the deployed numbers and facilities of MSC satellites is increasing dramatically every year (Nejat, 1992; Seedhouse 2012; Ilcev, 2013).

The size of the Earth requires multiple satellites to be placed in orbit in a constellation to cover uncovered areas of interest. Typically, a minimum of three to four satellites are needed to provide adequate communications coverage, so each of three satellites are covering 1/3 area of Earth surface between 80° North and South, excluding Polar Areas. Secondly, for existing users, upgrading satellites is not feasible, which means new capabilities are required and new satellites means new launches. Thirdly, more developed countries are recognizing the huge advantages of Military Satellite Communication (MILSATCOM) capabilities and are looking to implement or expand their networks. In such a way, some countries need to implement MILSATCOM, and some to upgrade it with modern CNS solutions

New mobile DVB-RCS system illustrated in Figure 7 is derived from current fixed DVB-RCS sometimes in 2000/01, which is proposed in 2000 by the South African Company CNS Systems [www.cnssystems.co.za]. The DVB-RCS mobile network includes a special ground receiving and transmitting equipment know as a HUB or GES (Gateway) terminals with C, Ku or Ka-band antenna system. The HAB terminals interface the TTN terminals or DVB-Terrestrial (DVB-T) cells via corresponding GEO satellite connections at C, Ku or Ka-band with MES and Transportable Earth Station (TES) terminals or remote DVB-Satellite (DVB-S) cells onboard navy, ground and air forces assets. The new DVB-RCS standards are the best solutions for establishment a network for connection all military assets at sea, on the ground and in the air (Ilcev, 2013).

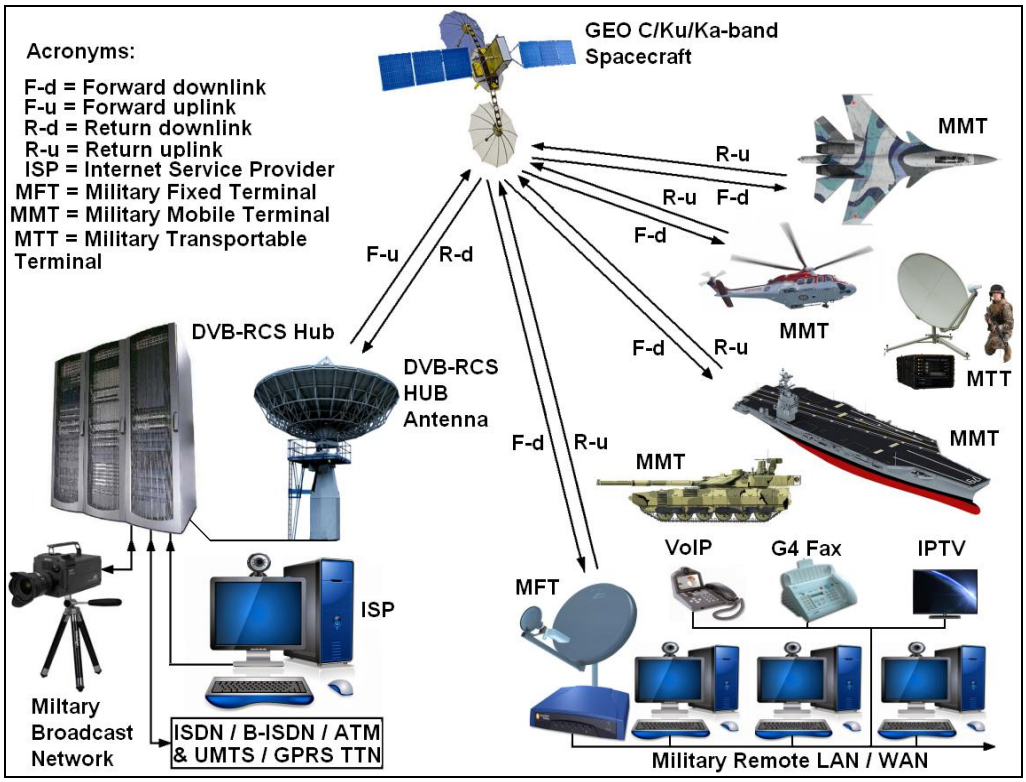


Figure 7: Military DVB-RCS communication network (Source: Ilcev, 2012).

New mobile DVB-RCS networks are very suitable for mobile-to-ground and ground-to-mobile solutions. Both systems are providing sophisticated Voice, Data and Video over IP (VDVoIP) for corporate, private, military, traffic control and management, meteorological and navigation information, training and medical service, technical and maintenance data, Search and Rescue (SAR), and Satellite Augmentation Data (SAD) of GNSS signals for surveillance and Inter Mobile Links (IML). The problem of current satellite fixed and mobile operators is that they are providing service via GEO satellite constellations and in this case are not able to cover both polar areas, such as Inmarsat, Eutelsat and Intelsat. In fact, because GEO constellation cannot cover both poles areas, to build a real global satellite coverage, it will be necessary to implement hybrid GEO and HEO or MEO satellite constellations.

The commercial and military SCS networks are very important in missions for the following reasons:

- To provide commercial or defense satellite communication links between mobiles and ground infrastructures, and between mobiles alone;
- To provide alert, distress and Search and Rescue (SAR) satellite communication links between mobiles and ground infrastructures, and between mobiles alone;
- To transfer augmented and not-augmented navigation PVT data from mobiles to traffic control centers via GEO satellite communication transponders; and
- To transfer augmented surveillance PVT data from traffic control centers to all mobiles via GEO satellite GNSS transponders, which will be used for enhanced navigation data and collision avoidance (Nejat, 1992; Del Re *at al*., 2008; Seedhouse, 2012; Ilcev, 2013).

5.2 Satellite Navigation Subsystem (SNS)

The GPS and GLONASS space segment consist of 24 GNSS-1 spacecrafts each as well as ground segment, which contains Ground Control Station (GCS) and Users Segment, shown in Figure 8. The GNSS-1 network is providing services for ships, land vehicles and aircrafts, which are receiving PVT signals by onboard installed mobile GPS or GLONASS receivers.

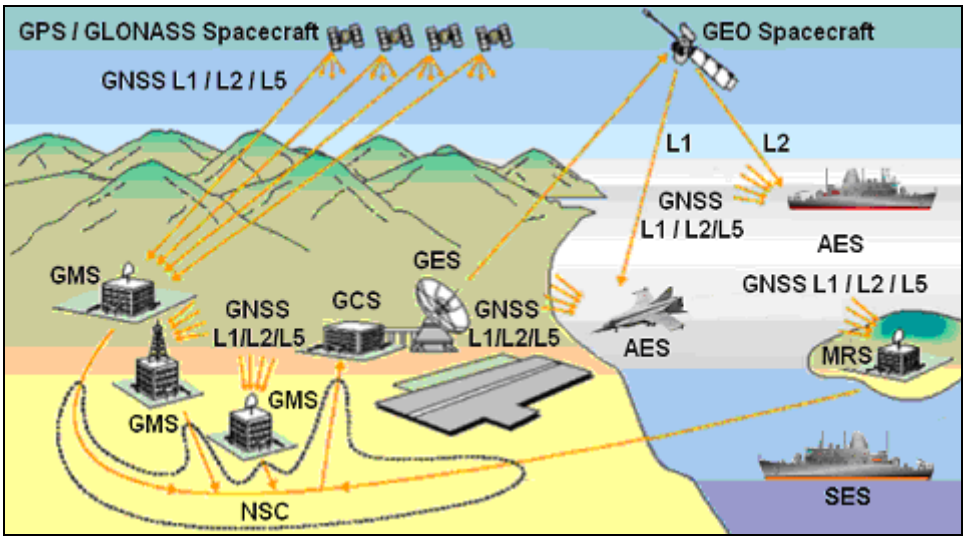


Figure 8: Military Satellite Navigation Network (Source: Ilcev, 2012).

As stated before, the US has its own Navstar GPS and Russians have the GLONASS system as parts of the GNSS-1 network. Europeans will eventually have Galileo and China is implementing its global BeiDou system, whereby both are part of the new GNSS-2 network. The current GPS constellation is consisted of the United States Air Force (USAF) GPS II satellites that provides around-the-clock, ultra-precise navigation and timing services for military and civilian users. The first satellite of the new GPS III next generation constellation SV-1 was launched on 23 December 2018, while the

second GPS III satellite, SV-2 USAF was launched on 22 August 2019, with entry into the operational constellation planned for the spring of 2020. The future GPS III constellation is designed to modernize the GPS constellation and to provide three times better accuracy and up to eight times improved anti-jamming capabilities.

In addition, the new generation of GPS III satellite constellation, when it becomes fully operational, will provide better accuracy through advanced atomic clocks, a more jam-resistant military signal and longer design life than earlier satellites. This GPS constellation will also provide increased precision navigation and timing to combat forces, as well as increased signal power, precision and capacity of the system and form the core of the GPS constellation for years to come. The GNSS-1 systems and their accuracy are upgraded by VHF or satellite augmentation of GPS and GLONASS solutions. In such a way, there is Differential GPS (DGPS) developed by the US Coast Guard, whereby the modern name is Local VHF Augmentation System (LVAS).

On the other hand, almost the entire Northern Hemisphere is already covered with modern Regional Satellite Augmentation System (RSAS), such as the US WAAS, European EGNOS, Chinese SNAS, Russian SCDN and Indian GAGAN. The new project African Satellite Augmentation System (ASAS) network is proposed by the South African CNS Systems Company and Space Science Center (SSC) at the Durban University of Technology (DUT) as the first RSAS network in the Southern Hemisphere. The RSAS network, also known as Satellite-Based Augmentation System (SBAS), was proposed by the International Civil Aviation Organization (ICAO). All the above stated RSAS systems will be integrated in the Global Satellite Augmentation System (GSAS) to serve civilian and military applications (Del Re *at al*., 2008; Ilcev, 2012; Ilcev, 2013; Kaplan *at al*., 2017).

5.3 Satellite Surveillance Subsystem (SSS)

The new SSS network will integrate both SCS and SNS solutions with Wide Area Multilateration (WAM) military system. The traditional surveillance radar with Automatic Dependent Surveillance (ADS) facilities can be also included in this integration or can be even used as backups. The ADS system is used on civilian aircrafts for communication and tracking solutions, and can be old Radio ADS (R-ADS) and new Satellite ADS (S-ADS), which can be used onboard ships as well.

The current surveillance is achieved through the use of long-range terminal surveillance radars, which sometimes cannot work properly because of very bad weather condition or other natural influences, such as heavy fog or dust coming from volcanoes. The new SSS infrastructures are set up for the traffic control systems to know where the mobile is and where it is heading, whose network is shown in Figure 9**.** The SSS solution known as Satellite Automatic Dependent Surveillance - Broadcast (SADS-B) is working in the way that all mobiles will be able to derive their GNSS PVT data from non-augmented or augmented onboard mobile GNSS receivers and send PVT surveillance data via GEO GNSS satellite transponder to the traffic control centers for computer processing and displaying of surveillance information to the ground controllers on screens.

The already developed RSAS networks will be the core of the future development of military GNSS augmentation infrastructures of the US GPS, Russian GLONASS, Chinese BeiDou and European Galileo GNSS networks for implementation of SSC/SSN and SSS technique via communication and navigation payload onboard multipurpose GEO spacecrafts. The GNSS augmented satellite network will serve navy, ground and air forces via Military Geospatial Augmentation System (MGAS), whereby the space and ground segments are illustrated in Figure 10 (Ilcev, 2012).

The MGAS network is working in the same way as civilian RSAS infrastructure. Wide Reference Station (WRS) sites and military mobiles are getting GNSS signals at the same time. While WRS is determining the differences of GNSS signals, Wide Monitoring Station (WMS) provides augmentation, which is sent via GEO Uplink Subsystem (GUS) or GES to military mobiles at GNSS frequency bands. The main part of MGAS is that Military Command Center (MCC), which is receiving GNSS augmentation navigation data from all military mobiles via GEO and GUS,

processing PVT and shown onto screens such as radar display. For instance, the MCC station sends via GUS and GEO satellites to certain aircrafts position data of all aircraft in its vicinity for collision avoidance and awareness.

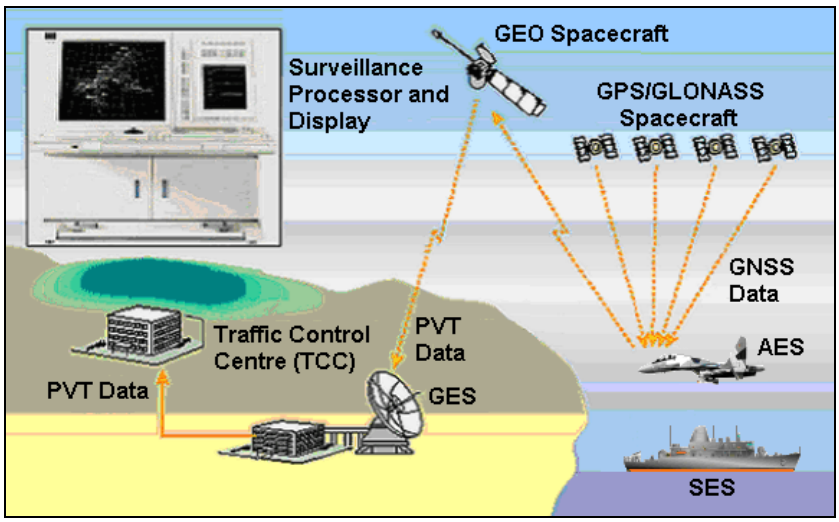


Figure 9: Military satellite surveillance network (Source: Ilcev, 2012).

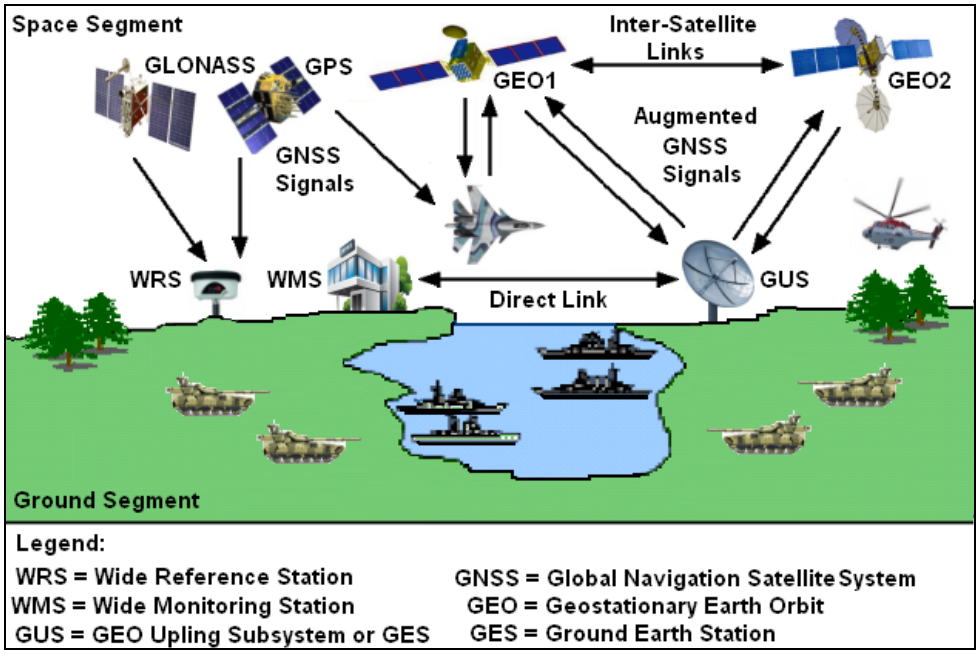


Figure 10: Military Geospatial Augmentation System (MGAS) (Source: Ilcev, 2012).

Many satellite research centers in many countries worldwide are working with the aeronautical agencies, research institutions and other stakeholders on the design of more sophisticated and reliable Satellite Automatic Dependent Surveillance - Broadcast (SADS-B) capability that periodically broadcasts an aircraft's position. The SADS-B network is supporting information, heading, altitude including aircraft identification (ID or name) and short-term intent, more accurately and reliable than the current radar capabilities. At the same time, other centers are providing research and solutions for development of the new maritime SADS-B, which has to upgrade Satellite - Automatic Identification System (S-AIS) and Long Range Identification and Tracking (LRIT) systems.

The Airborne SADS-B system can be used to increase a pilot's onboard situational awareness, particularly important in unfamiliar places, such as where aviation is vital with minimum ground

infrastructures, because of extreme harsh conditions and weather changes in unpredictable fashions. Shipborne S-ADSB will help ship captains to manage their vessels in a more safe way, especially during extremely bad weather conditions. Satellite ADS-B can be implemented for land vehicles and for all military applications as well (Del Re *at al*., 2008; Ilcev, 2013, 2018).

6. NON-GEO SATELLITE SYSTEMS FOR MILITARY SOLUTIONS

The non-GEO satellite systems started with operations at the end of 1990s, which as new global MSC structures are space solutions that can provide different services. such as VDV, tracking and positioning in integration with GNSS networks, as well as surveillance, monitoring, observation, remote sensing and imaging for both civilian and military applications. The major problems of LEO satellites are enormous satellite cost, complex network and short satellite visibility and lifetime. However, the recent developments of small satellites in LEO altitude up to 500 km will provide new applications with more cost effective possibilities designed by many countries, companies and even individuals. Thus, this year sees the launch of more than 200 nanosatellites to create a global network of reliable and affordable Internet and telecommunications services, helping to reach the unreachable.

However, it is important to highlight that small satellites are not able to provide CNS systems and DVB-RCS standards for seamen and airmen. At this point, only GEO satellites can carry MSC, GNSS and DVB-RCS transponders and provide CNS Systems for ship-, vehicle- and air-borne commercial and military applications. In the spite of that, Table 1 presents all main types of satellites and their masses. The mass of the smallest Fenito is less than 100 g, while the mass of a GEO satellite is higher than 1,000 kg and up to 7,000 kg (Ilcev, 2013).

Table 1: Masses of different types of satellites (https://www.nanosats.eu/cubesat).

Type of Satellite	Mass
Extra Large GEO	> 1000 kg
Medium MEO	500 to 1000 kg
Small LEO	100 to 500 kg
Micro LEO	10 to 100 kg
Nano LEO	1 - 10 kg
Pico LEO	0.1 to 1 kg
Fenito LEO	< 0.1 kg

6.1 Iridium Big LEO Satellite Network

The Iridium GMSC system was proposed in 1989 by Motorola and after the research phase, the Iridium LLC system was founded in 1991, with an investment of about USD 7 billion and became operational on 1 November 1998. With complete coverage of the Earth, including polar regions, the Iridium GMSC system delivers access to remote or rural areas, where no other forms of commercial and military communication are available. The company main office is situated in Leesburg, Virginia, where the Satellite Network Operations Centre is located, while the gateway facilities are located in Tempe, Arizona and Oahu, Hawaii. Through its own gateway in Hawaii, the US Department of Defense relies on Iridium for global communications capabilities. This system comprises three principal components: satellite network, ground network and Iridium subscriber products including phones and pagers, as shown in Figure 11.

The first generation of Iridium Big LEO satellites is situated in a near-polar orbit at an altitude of 780 km. They circle the Earth once every 100 minutes traveling at a rate of about 26,856 km/h. Each Iridium satellite is cross-linked (inter satellite link) to four other satellites; two satellites in the same orbital plane and two in an adjacent plane. In such a way, the Iridium constellation consists in 66

operational satellites and 14 spares orbiting in a constellation of six polar planes. Each plane has 11 mission satellites performing as nodes in the telephony network. The 14 additional satellites orbit as spares ready to replace any unserviceable satellite. This constellation ensures that every region on the globe is covered by at least one satellite at all times, providing voice, facsimile, paging and data solutions, as well as imaging, video and tracking, which also include the GPS capability already developed, for commercial and military solutions (Iridium, 2006; Ilcev, 2013, 2018).

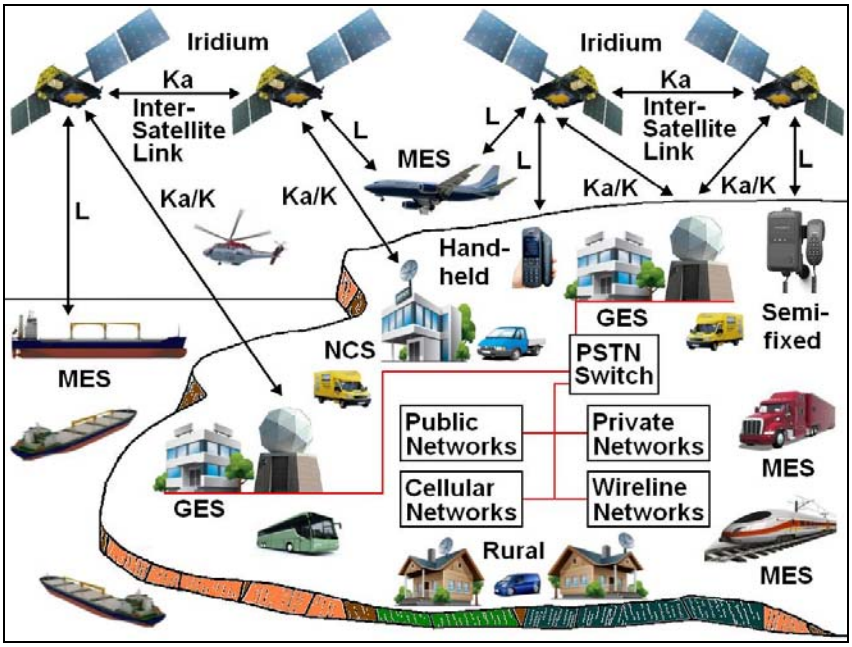


Figure 11: Iridium GMSC network (Source: Ilcev, 2018).

6.2 Globalstar Big LEO Satellite Network

Loral Space & Communications, with Qualcomm Incorporation developed the concept of Globalstar LEO system at a similar time to Iridium. Globalstar offers telephone, data, tracking and other services worldwide, starting from the end of the last century. Globalstar phones are similar to that of cellular systems, but with one main advantage; while a cellular phone works only with its compatible system in its coverage areas, the Globalstar system will offer worldwide coverage and interoperability with current and future public switched telephone and land mobile networks. The Globalstar network is also providing services for military solutions.

The Globalstar system consists of three major segments, which are the Space, Ground and User segments, including a Terrestrial Telecommunication Network (TTN), as well as public, private, wireless and cellular systems, as is illustrated in Figure 12. The Globalstar satellites are receiving signals from mobiles through the S-band forward link and sending signals to mobiles through the L-band return link. Globalstar has not intersatellite links, so it cannot provide coverage of both poles. The link between satellites and Ground Earth Stations (GES) is at C-band and the entire Globalstar network is controlled by Operations Control Centre (OCC) (Globalstar, 2000; Ilcev, 2013, 2018).

6.3 Orbcomm Small LEO Satellite Network

The Orbcomm system is a wide area packet switch and two-way data transfer network that is providing satellite messaging, tracking and monitoring services between mobile, remote, semi-fixed and fixed users via GES terminals and Little LEO satellites. It also provides Satellite - Automatic Identification System (S-AIS) integrated with Radio R-AIS, with the space and ground segments shown in Figure 13.

In the S-AIS network, all ships are receiving GNSS PVT signals from the US GPS (1) or Russian GLONASS (2), then ships out of R-AIS coverage are sending via service link (3) PVT data to the AIS satellite, which transmits this data via feeder link to the GES (Gateway) terminal (4). On the other hand, all ships sailing inside of R-AIS coverage are sending GNSS PVT data to R-AIS Base Station (BS) via radio link (5), while all these ships have AIS data communication via inter-ship links (6). Received AIS data GES and AIS BS are forwarding via terrestrial links (7) to the SCS terminal for processing. In such a way, AIS data with positions of all ships in certain sailing regions can be displayed on radar screens and used for collision avoidance (Orbcomm, 2000; Ilcev, 2018).

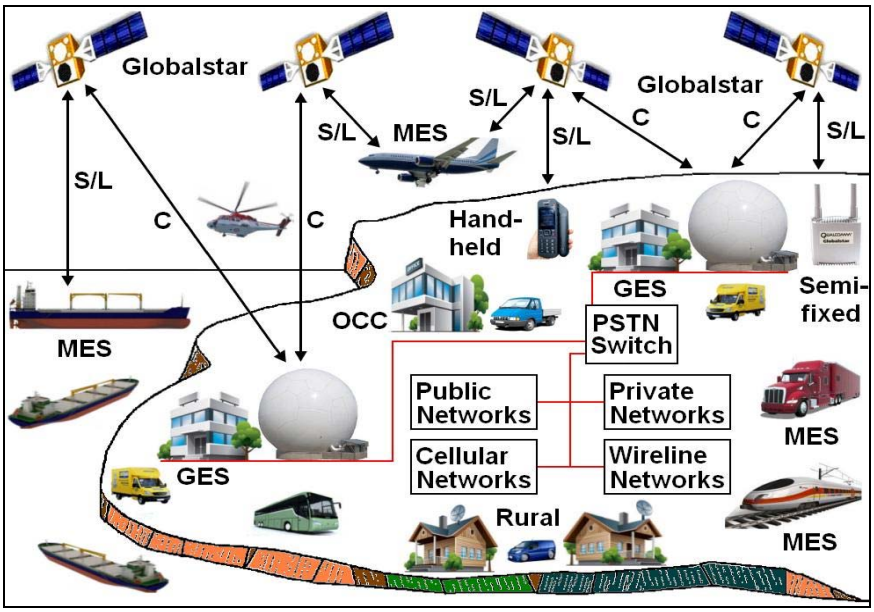


Figure 12: Globalstar GMSC network (Source: Ilcev, 2018).

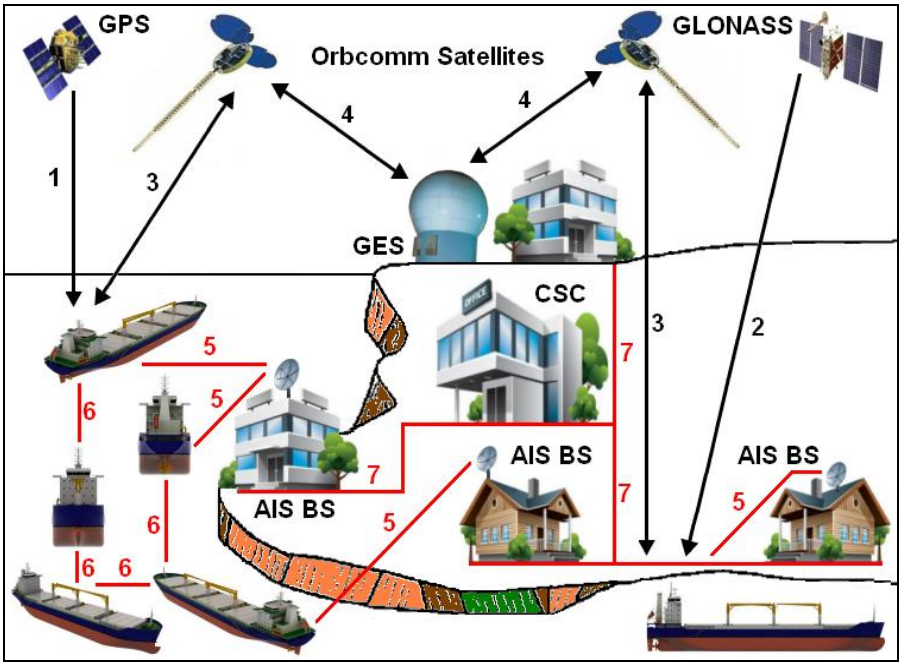


Figure 13: Orbcomm Satellite AIS (S-AIS) (Source: Ilcev, 2018).

6.4 Pico, Nano and Micro (CubeSat) Satellite Networks

Recently many countries and some space centers started to provide research, design and development of cost effective small satellites for LEO constellations. Another name for some of the smallest satellites is CubeSat, a term that is not used in the ITU resolution but often is applied to nanosatellites with a volume of 1 L or 10 cm^3. The ITU regulations are not geared for these smaller satellites and until now, there has been no consideration of how the rules might be adopted for them. In an effort to foster their use, a group of European countries submitted a proposal to WRC-12 suggesting that the next ITU conference should consider the frequency bands and regulatory requirements for small types of satellites. The proposal was dated on 30 January 2012, which maintains that more than 500 of these smaller satellites are under development, which is putting pressure on currently used frequency bands, usually within the range of 137 MHz to 2,450 MHz.

For instance, the University of Toronto Institute for Aerospace Studies/Space Flight Laboratory (UTIAS/SFL) has been developed a prototype of Generic Nanosatellite Bus (GNB) to fly a variety of payloads, ranging from S-AIS tracking solutions to precision formation flying. With the successful launch of the CanX-2 mission, technological validation is paving the way for the next generation of GNB derived CanX missions. The current GNB satellites in development include AISSat-1, CanX-3 (BRITE) and CanX-4&5, which the new design of the AISSat spacecraft offering end-to-end capability from mission design and spacecraft manufacturing to launch services and on-orbit operations. It is currently the only laboratory in Canada that has built and retains the capability to build very low-cost spacecraft, such as microsatellites under 100 kg and nanosatellites under 10 kg for S-AIS maritime services, of which interior details and components are shown in Figure 14. The spacecraft structure comprised of two trays that accommodate all of the electronics. The AIS system is a ship-to-ship and ship-to-shore system that is used as an aid for collision avoidance and vessel traffic management.

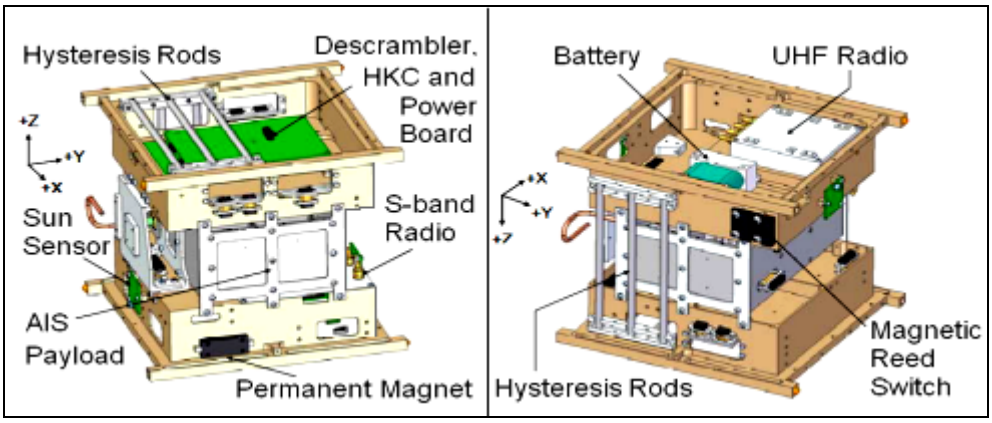


Figure 14: Nanosatellite Tracking Ships (NTS) interior details (Source: Ilcev, 2018).

The primary problem of small satellites is the short lifetime of batteries. The lifetime of a battery is depending on the amount of charge cycles , robustness for low voltage and operating temperature range. In addition, the weight and size of the battery have to be minimized with the focus being the space usability of the battery. In order to develop the best battery for CubeSat satellites, these factors have to be solved with new battery types (Newland *at a*l., 2009; Ilcev, 2018; Aerospace, 2019).

6. CONCLUSION

The development of modern civilian and military CNS solutions depends on the research, design and improvement of contemporary satellite systems, networks and ground infrastructures. While early satellite communication systems had a life span of days or weeks, today's systems have design lives extending to 20 years and beyond, with typical mean mission duration of 15 years. Thus, it is necessary to justify system effort and cost of development and operations.

Another change over time is that modern satellite CNS onboard terminals for mobile civilian and military applications have become smaller and more numerous. These terminals have evolved from a few large fixed terminals to thousands of small mobile terminals. On the other hand, from the beginning, the weight of GEO communication satellites was about 50 kg and today we have modern 10-ton structures with solar panels spanning several tenths of meters. However, a nanosatellite's mass range is between 1 and 10 kg, while a picosatellite weight is between 100 g and 1 kg.

Modern accurate, reliable and precise systems for communication, positioning, surveillance and time measurements are crucial for sea, land air military operations, where the effects from diverse sources must be coordinated in space and time. New GNSS augmented and CNS techniques and technologies address the needs of new Navigation Warfare (NAVWAR) defense systems, including Electronic Protection, Electronic Support and Electronic Attack. This equipment ensures continuous positioning even in the face of interference and jamming. New techniques provide users with the ability to detect, locate and characterize interference sources, such as the Global Mobile Tracking (GMT) network of military assets via GEO or LEO satellite constellations. New satellite multi-constellation capabilities ensure maximum signal availability deploying software, such as NovAtel CORRECT with Real Time Kinematics (RTK) provides centimeter-level accuracy for precise operations. Augmented systems, such as Synchronous Position, Attitude and Navigation (SPAN) of GNSS integrated with inertial technology ensure continual navigation in difficult environments.

REFERENCES

Aerospace (2019). *Small Satellites*. Available online at: https://aerospace.org/small-satellites (Last access date: 20 December 2019).

Breitman, D. & Byrd, D. (2018). *Today in Science: Launch of Explorer 1 Satellite.* Available online at: https://earthsky.org/space/launch-of-explorer-1-jan-31-1958 (Last access date: 20 December 2019).

Brown, M. (2019). *First Commercial Communications Satellite Activates*. Available online at: https://thisdayintechhistory.com/06/28/first-commercial-communications-satellite-activates (Last access date: 20 December 2019).

Comsat (2018). *Comsat*. Available online at: https://www.comsat.com (Last access date: 20 December 2019).

Del Re, E & Ruggieri, M (2008). *Satellite Communications and Navigation Systems*, Springer, New York.

GEOLSOC (2018). *Sputnik and the Shape of the Earth.* Geological Society, Piccadilly, London.

Globalstar, (2000) *Globalstar System – Description*. Globalstar, San Jose, CA.

Ilcev, D.S. (2012). *African Satellite Augmentation System (ASAS)*. Durban University of Technology (DUT), Durban South Africa.

Ilcev, D.S (2013). *Global Aeronautical Communications, Navigation and Surveillance (CNS) – Theory and Applications*, AIAA, Reston, Virginia.

Ilcev, D.S. (2018). *Global Communication, Navigation and Surveillance (CNS)*, Durban University of Technology (DUT), Durban, South Africa.

Inmarsat (2018). *Inmarsat*. Available online at: https://www.inmarsat.com (Last access date: 20 December 2019)

Iridium (2006). *Implementation Manual for Iridium Satellite Communications Service*, McLean, VA, USA.

Launius, R. (2015). *Beginning the Age of Satellite Communications, Commenting on Spaceflight History*. Available online at: https://launiusr.wordpress.com/2015/08/03/beginning-the-age-of-satellite-communication-echo-1-august-1960 (Last access date: 20 December 2019).

Nejat, A. (1992). *Digital Satellite Communications Systems and Technologies – Military and Civil Applications*. Kluwer Academic Publishers (Springer), Dordrecht, Holland.

Newland, F., Coleshill, E., DSouza, I. & Cain, J. (2009). *Nanosatellite Tracking of Ships Review of the First Year of Operations,* AIAA 7th Responsive Space Conference, Los Angeles.

Orbcomm (2001). *Orbcomm System Overview*. Orbcomm, NJ, USA.

Seedhouse, E. (2012). *Military Satellites - Current Status and Future Prospects*. Available online at: http://www.spaceref.com/news/viewnews.html?id=1622 (Last access date: 20 December 2019).

ADAPTIVE WINDOW SIZE AND STEPPED FREQUENCY SCAN SPECTROGRAM ANALYSIS FOR DRONE SIGNAL DETECTION IN MULTI-SIGNAL ENVIRONMENT

Chia Chun Choon & Ahmad Zuri Sha'ameri*

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

*Email: ahmadzuri@utm.my

ABSTRACT

In this paper, a spectrogram based on stepped frequency scanning and adaptive window size algorithm is proposed to detect drone signals that operate at the 2.4 and 5.8 GHz Industrial, Scientific and Medical (ISM) bands in a multi-signal environment. In this algorithm, the received signal is divided into multiple sub-bands and scanned through a large analysis bandwidth. The window size is automatically adjusted by balancing the time and frequency resolution. The adaptive stepped frequency scan spectrogram (ASFSS) is then implemented to obtain the time-frequency representation (TFR). From the TFR, signal parameters, such as the hop duration, bandwidth, and instantaneous frequency (IF), are estimated. Three possible drone signal types are used in the study: fast frequency hopping spread spectrum (FHSS), slow FHSS, and hybrid spread spectrum (HSS). The performance of ASFSS is verified using Monte-Carlo simulation with 20 realisations at signal-to-noise ratio (SNR) range from -16 to 12 dB. In the presence of additive white Gaussian noise (AWGN), the detection cut-off point is -12 dB for fast and slow FHSS and -5 dB for HSS. Additional environment signals, such as direct sequence spread spectrum (DSSS) and WiFi, increase the cut-off point to 5 dB for fast FHSS, 7 dB for slow FHSS and 8 dB for HSS.

Keywords: *Adaptive stepped frequency scan spectrogram (ASFSS); adaptive window size estimation; instantaneous frequency (IF); large analysis bandwidth; time-frequency representation (TFR).*

1. INTRODUCTION

The applications of drones for civilian and commercial purpose have grown over the past decade. This is as drones, commonly referred to as radio controlled unmanned aircraft systems (UASs), provide beneficial visual services in the sky. Examples of legal and beneficial uses of drones include industrial and building safety inspection, geographical mapping, aerial monitoring, photography and videography, delivery, and agricultural research (Dobie, 2016). However, the abuse of drone applications has caused security issues, crime, terrorism and privacy problems, such as cross-border drug or weapon smuggling, spying on critical infrastructure or government facilities, flying a drone by carrying gun or bomb, and endangering commercial or military air traffic especially at airports. A recent incident at a restricted airspace over Gatwick airport in the UK resulted in disruption of services over the Christmas holiday week in 2018 (BBC, 2018). Thus, an approach for countering the threats caused by drones is described, which includes monitoring the radio frequency (RF) link between the remote control and drone, and drone signal detection.

Drone detection technologies refer to technologies used to detect the presence of drones within a given coverage area. The most commonly used techniques in drone detection are RF signal detection, radar, acoustic signal detection, computer vision and thermal sensing (Musa *et al.*, 2019). Each detection

technology has its strength and weaknesses. Radar is a possible solution for detecting drones, However, the small radar cross section of drones, due to small size and low altitude makes drone detection challenging (Shi *et al.*, 2018) as well as reduces the probability of detection (Ochodnický *et al.*, 2017). Acoustics detection is not affected by drone size since it uses low cost and simple audio sensors such as microphones to sense the buzzing noise come from drone propellers. However, the detection range is very short, which is easily affected by ambient noise (Mezei *et al.*, 2015; Shi *et al.*, 2018) and highly relies on a database of drone sounds (Mezei *et al.*, 2015). For computer vision detection, the visibility could be impacted by fog, trees and buildings (Güvenç *et al.*, 2018). Furthermore, it requires a database for pattern recognition due to diverse shapes and different angles of view of drones (Rozantsev *et al.*, 2015). For thermal sensing, the detection is based on the emitted heat signature from motor and battery, which is useful for night time detection (Andraši *et al.*, 2017). However, the detection range is short (Fu *et al.*, 2017) and affected by ambient noise especially in urban areas, such as heat sources from lightings. Some of these challenges can be addressed using RF-based detection techniques by monitoring the RF link between the remote control and drone. Unlike the other detection techniques mentioned above, RF-based detection relies on electromagnetic emissions from the drone, which in most cases contain telemetry and video signals. The only disadvantage is that when the drone goes into autonomous mode, the transmission is cut-off. Furthermore, many other wireless technologies, such as WiFi and Bluetooth, share the same frequency bands as drones (Shi *et al.*, 2018), which could reduce the effectiveness of detecting drone signals. The wireless technologies that are commonly used in drones are frequency hopping spread spectrum (FHSS), direct sequence spread spectrum (DSSS), hybrid spread spectrum (HSS), WiFi and Bluetooth (Rohde & Schwarz, 2016).

In RF-based detection, the characteristics of drone signals are represented in time and frequency domains, and used to detect the presence of drones (Ezuma *et al.*, 2019). One of the most commonly used methods to represent signals in time-frequency domain is short-time Fourier transform (STFT). In Luo *et al.* (2009), Luan *et al.* (2010) and Zhang *et al.* (2016), the time-frequency representation (TFR) plot generated from STFT is converted into image and signal parameters that are extracted from the image. The short-time Goertzel algorithm proposed in Sha'ameri *et al.* (2016) is a method similar to STFT. Instead of using fast-Fourier transform (FFT), the Goertzel algorithm is used, whereby the received signal goes through a bandpass filtering operation at every channel using a second order infinite impulse response (IIR) filter. In addition, the Wigner-Ville distribution method was used in Javed *et al.* (2010) and Guo *et al.* (2011) to obtain better time and frequency resolution in TFR. Better signal detection is obtained in terms of probability of false detection, probability of lost detection, buffer size or sensing time, signal-to-noise ratio (SNR) and receiver operating curve.

In a spectrogram, there is a trade-off between time and frequency resolution due to the uncertainty principle. In order to obtain an optimal window size adaptively, Zhong *et al.* (2010) proposed that the analysis window size be measured using the instantaneous frequency (IF) gradient of the signal, where the IF of the signal is obtained by detecting the ridge of wavelet transform. This implies that a wide window should be employed as the IF of the signal varies smoothly, while a narrow window should be employed as the IF varies sharply. Pei *et al.* (2012) achieved the same objective by using a concentration measure to quantitatively evaluate the TFR energy concentration.

Once the signal representation and window size optimisation are completed, the next step would be detecting the presence of a drone. In Deng *et al.* (2011) and Meng *et al.* (2013), the autocorrelation method is used in signal detection for searching autocorrelation peaks and judging the cumulative peak-to-average ratio. The autocorrelation function has the advantage of discriminating between noise and actual signal coming from the drone. Noise which is random appears as an impulse signal while the drone signal exhibits periodicity whose feature allows distinction in the autocorrelation function. Furthermore, Han *et al.* (2013) proposed an energy detection method. The detection threshold is based on the mean

energy of the background signal, which is estimated by averaging. Input that is bigger than the threshold is considered as a signal. This method is simple, but has poor detection for low SNR signals.

The objective of this paper is to describe an adaptive stepped frequency scan spectrogram (ASFSS) approach that allows representation of signals and large bandwidth analysis at lower sampling rate with better time and frequency resolution. The input signal is divided into multiple sub-bands and scan through a large analysis bandwidth. The window size is automatically adjusted by balancing the time and frequency resolution. Most of the drones are operated at 2.4 and 5.8 GHz Industrial, Scientific and Medical (ISM) bands, which occupy a large bandwidth of 100 and 150 MHz respectively (Rohde & Schwarz, 2016). Normally, a large bandwidth analysis involves high sampling rate as stated in the Nyquist sampling theorem. Furthermore, the time-varying nature of drone signals, such as FHSS and HSS, requires analysis in both time and frequency domains. For drone signal detection, frequency channel and hop duration are estimated from TFR, which is then use to derive the IF. In addition to drone signals, the presence of additive white Gaussian noise (AWGN) and other wireless technologies in the environment, such as WiFi and Bluetooth, that also operate in the same ISM band would cause signal interference and complicate the drone signal detection. The performance of ASFSS is then verified using Monte-Carlo simulation for different cases of multi-signal environment at various SNRs.

This paper is organised as follows: Section 2 defines the signal model and problem statement, followed by Chapter 3, which discusses time-frequency analysis (TFA), signal detection method and IF estimation. The results are presented in Section 4, and the paper is concluded in Section 5.

2. SIGNAL MODEL AND PROBLEM STATEMENT

The signals model that used in ASFSS verification is first described in this section followed by the problem statement.

2.1 Signal Model

In this study, three types of drone signals and two types of background signal would be used to verify the accuracy of TFR produced by ASFSS. The drone signals are fast-FHSS, slow-FHSS and HSS, which is a combination of FHSS and DSSS. For background signals, there are orthogonal frequency division multiplexing (OFDM) and DSSS. The OFDM signal represents WiFi, which is commonly used by many wireless devices. The signals parameters for the drone and background signals are described in Tables 1 and 2. These signals are assumed to operate at 2.4 GHz ISM band, and are down-converted and sampled at the Nyquist rate with sampling rate $f_s = 50$ MHz. The received drone signals can be expressed as (Deng *et al.*, 2011; Liu *et al.*, 2016):

$$\text{Fast-FHSS} \quad : s_1(t) = a(t)e^{j(2\pi f_c(t)t)} \quad , \ 0 \le t \le T \tag{1}$$

$$\text{Slow-FHSS} \quad : s_2(t) = a(t)e^{j(2\pi f_c(t)t)} \quad , \ 0 \le t \le T \tag{2}$$

$$\text{HSS} \quad : s_3(t) = a(t)p(t)e^{j(2\pi f_c(t)t)} \quad , \ 0 \le t \le T \tag{3}$$

where $a(t)$ represents the information bearing signal, $f_c(t)$ is the time-varying channel frequency, T is the total signal length, and $p(t)$ is the pseudorandom sequence code. The channel frequency f_c of the FHSS and HSS signals will keep changing from time to time, and the time spent for each f_c on each hop is the hop duration T_{hop}, with the bandwidth occupied by each hop is f_{BW}. For binary frequency shift

keying (BFSK) signals, such as fast-FHSS and slow-FHSS, the expected f_{BW} to be seen on the TFR plot is calculated as:

$$f_{BW} = 2f_{\Delta} + R_s \tag{4}$$

The received background signals, OFDM and DSSS, can be described as (Schulze *et al.*, 2005; Deng *et al.*, 2011):

$$\text{Strong-OFDM} \quad : v_1(t) = A \sum_{k=1}^{64} a(t) e^{j2\pi(f_c + f_k)t} \ , \ 0 \le t \le T \tag{5}$$

$$\text{Weak-OFDM} \quad : v_2(t) = (0.707A) \sum_{k=1}^{64} a(t) e^{j2\pi(f_c + f_k)t} \ , \ 0 \le t \le T \tag{6}$$

$$\text{DSSS} \quad : v_3(t) = a(t)p(t)e^{j2\pi f_c t} \ , \ 0 \le t \le T \tag{7}$$

where $a(t)$ represents the information bearing signal, f_c is the constant channel frequency, f_k is the subcarrier frequency, $p(t)$ is the PN pseudorandom sequence code, k is the subcarrier frequency index in the range of $1 \le k \le 64$, and A is the signal gain of the OFDM signal. The signal gain of a weak-OFDM is multiplied by $1/\sqrt{2}$ to produce an OFDM signal that has half the power of a strong-OFDM signal. According to the IEEE802.11 b/g/n standard for WiFi, subcarrier spacing $\Delta f_k = 312.5$ kHz and number of subcarriers is 64, which results in OFDM signal bandwidth per channel of 20 MHz.

Table 1: Drone signal parameters: hop duration (T_{hop}), channel frequency (f_c), channel index (k_c), bandwidth (f_{BW}), frequency different (f_{Δ}), symbol rate (R_s) and chirp rate (R_c).

Signal	Modulation	Time Parameter (Hop duration)	Channel frequency	Other parameters
Fast-FHSS	BFSK	200 μs	$f_c = k * 6$ MHz $k = 1, 2, 3, \ldots\ldots, 16$	$R_s = 50$ kbit/s $f_{\Delta} = 0.5$ MHz
Slow-FHSS	BFSK	500 μs	$f_c = k * 15$ MHz $k = 1, 2, 3, \ldots\ldots, 6$	$R_s = 20$ kbit/s $f_{\Delta} = 0.25$ MHz
HSS	Binary phase shift keying (BPSK)	800 μs	$f_c = k * 15$ MHz $k = 1, 2, 3, 4$	$R_s = 500$ kbit/s $R_c = 750$ kbit/s $f_{BW} = 1.5$ MHz

Table 2: Background signal parameters: signal length (T), channel frequency (f_c), bandwidth (f_{BW}), symbol rate (R_s) and chirp rate (R_c).

Signal	Modulation	Time Parameter (Signal length)	Channel frequency	Other Parameters
OFDM	Quadrature Phase Shift Keying (QPSK)	3200 μs	$f_c = 12, 37, 84$ MHz	$R_s = 16$ Mbit/s $f_{BW} = 20$ MHz 64 subcarriers
DSSS	BPSK	3200 μs	$f_c = 12$ MHz	$R_s = 500$ kbit/s $R_c = 1.5$ Mbit/s $f_{BW} = 2$ MHz

In practice, signals received from drones are interfered by background signals that can be represented as:

$$y(t) = s(t) + v(t) + n(t) \quad (8)$$

where $s(t)$ is the actual drone signal, $v(t)$ is the background signal, and $n(t)$ is the complex-valued additive white Gaussian noise (AWGN) process of independent and identically distributed real and imaginary parts with zero mean and total variance ${\sigma_v}^2$ (Chee *et al.*, 2014). AWGN is included because it is a basic noise source that appears in almost every practical environment. Furthermore, the OFDM and DSSS interference signals are included. Thus, five cases of environment are considered as follows:

1) Case 1: Fast-FHSS + Slow-FHSS + HSS
2) Case 2: Fast-FHSS + Slow-FHSS + HSS + AWGN
3) Case 3: Fast-FHSS + Slow-FHSS + HSS + AWGN + Strong-OFDM
4) Case 4: Fast-FHSS + Slow-FHSS + HSS + AWGN + Weak-OFDM
5) Case 5: Fast-FHSS + Slow-FHSS + HSS + AWGN + Strong-OFDM + DSSS

where Case 1 is an ideal environment, Cases 3 and 4 contain three strong-OFDMs and weak-OFDMs at channels 12, 37 and 84 MHz respectively. In Case 5, there are two strong-OFDMs at 37 and 84 MHz, with a DSSS at channel 12 MHz.

2.2 Problem Statement

Drone signal detection refers to RF detection of the communication link between the remote control and drone. Drones are frequently flown in public and noisy environments. In order to protect against interference, drones adopt spread spectrum technologies, such as FHSS, DSSS, HSS and OFDM (such as WiFi in the 2.4 GHz ISM band) (Shin *et al.*, 2016). For these time-varying signals, signal parameters such as hop duration and channel frequency are important to detect and identify the presence of a drone. Therefore, representing signal information in time and frequency domains is important to drone signal analysis.

Most of drones, especially the recreational types, operate at the 2.4 and 5.8 GHz ISM bands, where large bandwidths of 100 and 150 MHz are utilised respectively. For Malaysia, the Malaysian Communications and Multimedia Commission (MCMC) has specified the frequency bands and transmit power for operating drones (MCMC, 2015). Typically, the signal is first down-converted to the intermediate frequency before the analysis is performed. The Shannon-Nyquist theorem states that to perfectly represent an analog signal, it must be sampled at a frequency higher than twice the signal's bandwidth (Fyhn *et al.*, 2013). For a recreational drone that uses FHSS, the signal can hop over the very large 100 MHz bandwidth and thus, capturing the full FHSS bandwidth requires very high sampling rate (Liu *et al.*, 2016). The higher the sampling rate, the bigger the signal samples that contribute to higher computational complexity. Due to this frequency hopping over the large bandwidth, the signal is difficult to detect while maintaining a low sampling rate to reduce computational complexity in the real-time. In order to reduce the cost induced by high sampling rate, Joo *et al.* (2007) proposed a scanning receiver divides the frequency range into sub-bands and scans through every sub-band. Similar to Lehtom¨aki *et al.* (2004), the channelisation and sweeping method are introduced, where the signal is channelised into sub-channels, which are swept by changing the centre frequency of the receiver in order to cover entire bandwidth. Unlike channelised receivers (Tang *et al.*, 2012), the large analysis bandwidth of the receiver is divided into multiple sub-channels where each input of sub-channel is filtered and down-converted in parallel before reconstructing the signal from each sub-channel output. Furthermore, there is a method discussed

by Shin *et al.* (2016) to make use of power threshold to select the sub-channels and record whenever signal power exceeded a threshold.

In TFR, there is a trade-off between resolution of time and frequency due to the uncertainty principle (Zhong *et al.*, 2010). It is not possible to get high resolution TFR in both time and frequency simultaneously. However, this is crucial if precision is required to estimate the time and frequency parameters of a signal. This is best illustrated by a modulated Gaussian pulse (Boualem, 2015). It is shown that this is the only signal where the product of effective time t_{eff} and frequency f_{eff} is constant at 1/4π, and other signals should conform to this equality $t_{eff}f_{eff} \geq 1/4\pi$. The result of this inequality is also known as the uncertainty principle, similar to the Heisenberg uncertainty principle in quantum physics. Since the product is a constant, a short duration pulse will have a broad bandwidth and vice versa. Thus, increase in time resolution will cause decrease in frequency resolution, and vice versa. In addition, different drone uses different signal bandwidths and hop durations that require an analysis window size adapting to the signal characteristics by balancing between time and frequency resolutions.

The 2.4 GHz ISM bands that are used by recreational drones are also used by other wireless technologies, such as WiFi and Bluetooth. In this multi-signal environment, these wireless technologies and white noise are mixed together with drone signals making it difficult to distinguish each of them (Luan *et al.*, 2010). Unlike white noise, interference signals such as WiFi have dedicated transmission channels and only affect drone signals at certain frequency ranges. Therefore, the threshold for drone signals detection has to be adapted according to the change of frequency spectrum.

3. TFA, SIGNAL DETECTION, AND IF ESTIMATION

In this section, the proposed spectrogram with adaptive window size method is presented, followed by the signal frequency channel and hop duration estimation and finally, the IF estimation algorithm is explained.

3.1 ASFSS

In ASFSS, the RF signal at 2.4 GHz is demodulated and down-converted to the intermediate frequency signal with frequency range from 0 to 25 MHz, the input signal is divided into seven sub-bands, and scanned through every sub-band. From one sub-band to next sub-band, there is a constant frequency step of 12.5 MHz. The analysis bandwidth is equal to the sub-band bandwidth, which is 25 MHz as shown in Figure 1. The sub-bands are 50% overlapped with each other to prevent amplitude loss in between two sub-band boundaries due to the nature of frequency response. To describe how the ASFSS is applied, the choice for bandwidth and number of sub-bands is limited to 25 MHz and seven sub-bands, and the bandwidth depends on the sampling rate of 50 MHz, which is based on the Nyquist sampling theorem. The time and frequency parameters can be varied according to the application requirement. Consequently, the higher number of sub-bands, the time taken to complete a full spectrum scan is longer, resulting in lower scanning speed. In order not to miss any part of the drone signal, the scanning speed has to be made higher than the signal hop rate. The scanning speed also depends on the window size and overlap percentage of one window to the next window. In this paper, ASFSS using the Hamming window function is overlapped continuously, where the window function is advanced by one sample from current window to next window, which is around 99% of window overlapping. By considering the maximum window size of 4096, 99% window overlapping for seven sub-bands would have a total window length of 4102, and the shortest hop duration that can be analysed at 50 MHz sampling rate is 82 μs. Thus, the proposed ASFSS settings described earlier can represent the signal with the shortest hop duration shown in Table 1.

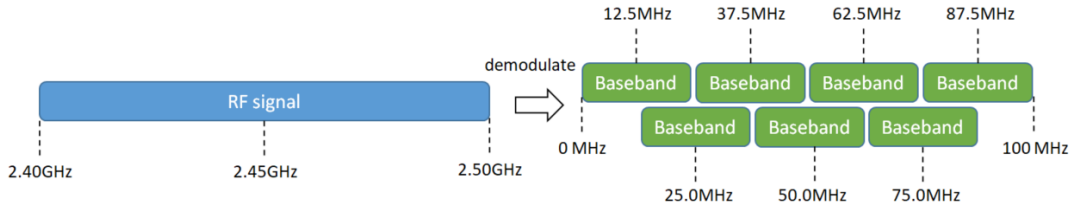


Figure 1: ASSFS analysis bandwidth.

The block diagram for ASFSS illustrated in Figure 2 shows that the input RF signal is demodulated through a mixer with local oscillator at frequency $f_{LO} = 2.4$ GHz. After that, the demodulated signal is split into seven sub-bands using variable prefilter, mixer and variable local oscillator. The parameters of the variable prefilter and variable local oscillator can be found in Table 3. All the components of the system and the necessary processing before the analog to digital converter (ADC) should be implemented based on analog technology. The variable prefilter band is defined based on the start and stop frequency of each sub-band, while the variable local oscillator frequency is based on the start frequency of each sub-band. Next, the signal now with smaller frequency bandwidth is filtered using a low pass filter (LPF), where the filter bandwidth $BW = 25$ MHz. Then, the filtered signal is digitised via the ADC at sampling rate $f_s = 50$ MHz instead of 100 MHz. Subsequently, a Hamming window function and FFT is applied on the digital signal to obtain a sub-spectrum for the particular sub-band. This sub-spectrum will be stored in memory temporary until all the seven sub-spectrums are obtained from every sub-band. These seven sub-spectrums would be merged into a complete spectrum that covers the analysis bandwidth of 100 MHz. The last part would be accumulating the spectrum to obtain a TFR of the signal.

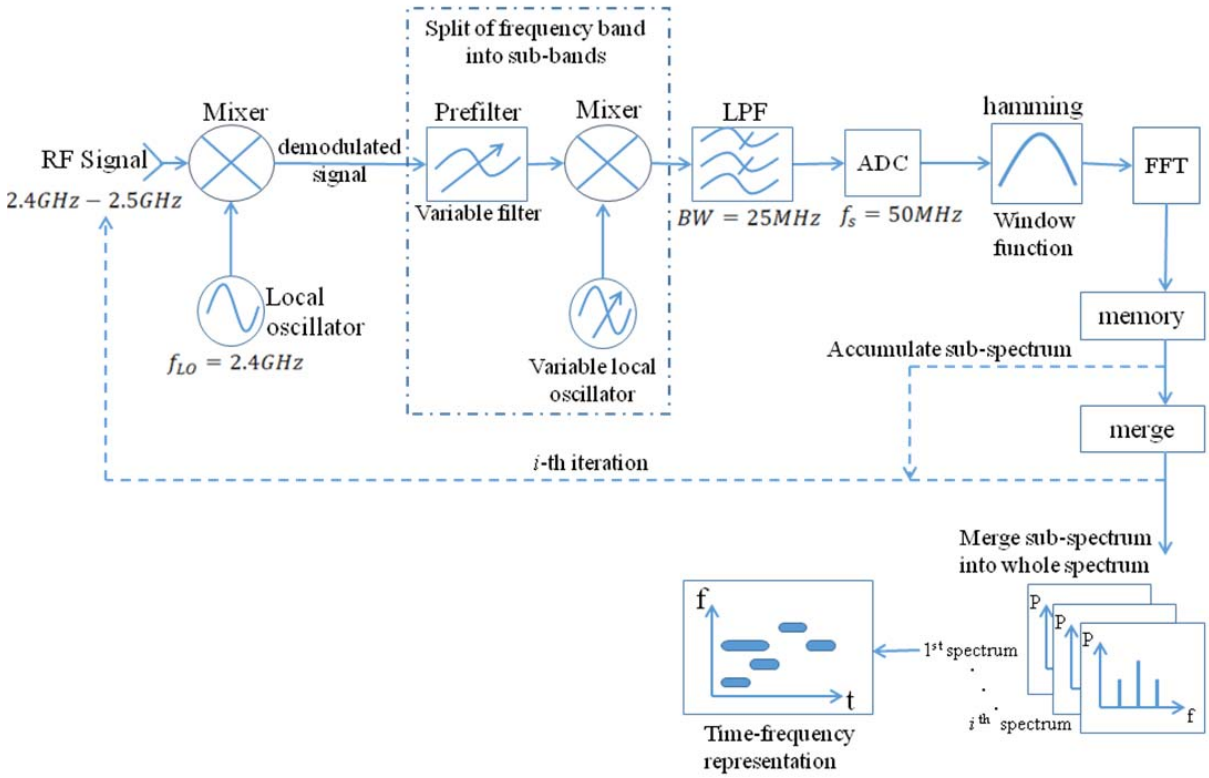


Figure 2: ASFSS block diagram.

The ASFSS approach can be modelled as:

a) The input RF signal down-converted to demodulated signal can be derived as:

$$x_d(t) = x_r(t)e^{-j(\omega_{LO}t+\emptyset_{LO})} \quad (9)$$

where $x_r(t)$ is the input RF signal, $x_d(t)$ is the demodulated signal, $\omega_{LO} = 2\pi f_{LO}$ is the local oscillator's frequency in radian/sec, and $\emptyset_{LO}$ is local oscillator's phase angle in radian/sec.

Table 3: Variable prefilter and variable local oscillator parameters corresponding to sub-bands.

Sub-band index	Variable prefilter band (MHz)	Variable local oscillator frequency (MHz)
1	0.0 – 25.0	0.0
2	12.5 – 37.5	12.5
3	25.0 – 50.0	25.0
4	37.5 – 62.5	37.5
5	50.0 – 75.0	50.0
6	62.5 – 87.5	62.5
7	75.0 – 100.0	75.0

b) $x_d(t)$ is split into sub-bands using the following equation:

$$x_i(t) = \left[\int_{-\infty}^{\infty} g_a(\tau)x_b(t-\tau)d\tau\right] e^{-j(\omega_a t+\emptyset_a)} \quad (10)$$

where $x_i(t)$ is the intermediate frequency signal and $g_a(t)$ is a variable prefilter kernel with variable filter band, $\omega_a = 2\pi f_a$ is the variable local oscillator's frequency in radian/sec, and $\emptyset_a$ is the variable local oscillator's phase angle in radian/sec. The variable filter band and variable local frequency are changing based on the sub-band index $1 \leq a \leq 7$ as per in Table 3.

c) The filtering of intermediate frequency signal $x_i(t)$:

$$x_f(t) = h(t) * x_i(t)$$

$$= \int_{-\infty}^{\infty} h(\tau)x_i(t-\tau)d\tau \quad (11)$$

where $x_f(t)$ is filtered signal, and □(t) is a LPF kernel with cut-off frequency at 25 MHz.

d) ADC conversion from continuous signal to discrete signal:

$$x_f(t) \rightarrow x(n) \quad (12)$$

e) The transformation of the digital signal to frequency domain and obtaining the TFR can be modelled as:

$$X_s(n, k_s) = \sum_{m=0}^{N-1} x(m)w(m-n)e^{-j\omega m} \quad (13)$$

$$\omega = \frac{2\pi k_s}{N_w} \quad (14)$$

$$X(n,k) = \sum_{l=0}^{6} X_s(7n+l,k_s) \tag{15}$$

$$S_x(n,k) = \left|X(n,k)\right|^2 \tag{16}$$

where $X_s(n,k_s)$ is the sub-spectrum, N is the signal length, $x(m)$ is the signal and $w(n)$ is the window function with length N_w. Equation 15 is an algorithm to merge all the seven sub-spectrums $X_s(n,k_s)$ into a complete spectrum $X(n,k)$ and $S_x(n,k)$ is the TFR.

3.2 ADAPTIVE WINDOW SIZE

Window size affects the time and frequency resolution on a given signal in the TFR. It is proportional to the time resolution but inversely proportional to the frequency resolution. Thus, a balance between the time and frequency resolution is desired. This can be achieved by having 1:1 ratio for the number of time bins N_T to number of frequency bins N_F (Hz/bin). The time resolution T_r is the distance in s between two adjacent data points in TFR, which is measured in time-per-bin (s/bin). For frequency resolution F_r, it is the distance in Hz between two adjacent data points in TFR, which is measured in frequency-per-bin (Hz/bin). Mathematically, the relationship between T_r and F_r can be represented as:

$$T_r = N_w / f_s \tag{17}$$

$$F_r = f_s / N_w \tag{18}$$

where f_s is the sampling frequency and N_w is the window size. In the TFR, N_T and N_F is the total number of bins that is required to represent T_{hop} and f_{BW} of a signal in terms of time and frequency. The N_T and N_F can be derived as:

$$N_T = \hat{T}_{hop}/T_r \tag{19}$$

$$N_F = \hat{f}_{BW}/F_r \tag{20}$$

where $\hat{T}_{hop}$ is signal hop duration estimated using the smallest window size (highest time resolution), and $\hat{f}_{BW}$ is signal bandwidth estimated using the largest window size (highest frequency resolution). The relationship between N_T and T_r can be derived by substituting Equation 17 into Equation 19 to obtain:

$$N_T = \hat{T}_{hop} f_s \left(\frac{1}{N_w}\right) \tag{21}$$

where N_T is inversely proportional to N_w. The relationship between N_F and F_r is derived by substituting Equation 18 into Equation 20:

$$N_F = \left(\frac{\hat{f}_{BW}}{f_s}\right) N_w \tag{22}$$

where N_F is directly proportional to N_w. The estimated window size $\widehat{N}_W$ can be obtained by achieving the time and frequency resolution $N_T = N_F$, or allowing Equation 21 to equal Equation 22:

$$\widehat{N}_w = \sqrt{\frac{(f_s)^2\left(\hat{T}_{hop}\right)}{\hat{f}_{BW}}} \tag{23}$$

The flowchart for estimating the adaptive window size is presented in Figure 3. Both the drone signal and background signal is captured, and then inputted to the spectrogram to obtain the TFR. First, spectrogram is performed using the minimum window size to estimate the minimum hop duration of signal. After that, the process is repeated using the maximum window size to estimate the minimum bandwidth of signal. The estimated minimum bandwidth and minimum hop durations is then used in Equation 23 to calculate $\widehat{N}_W$.

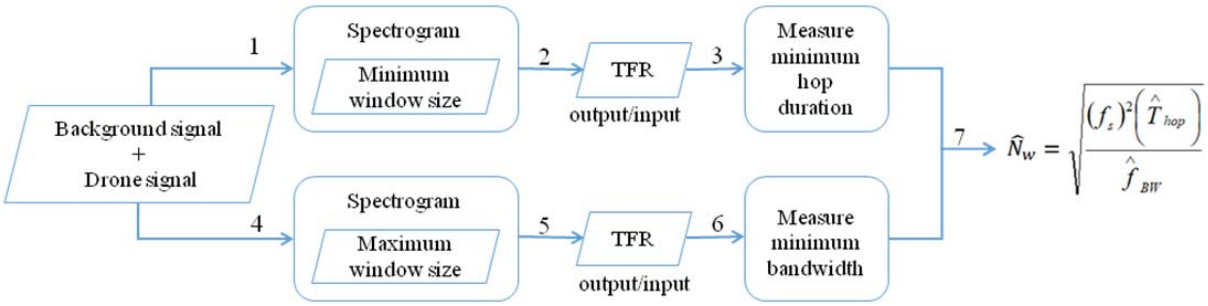


Figure 3: Flowchart of adaptive window size estimation.

3.3 Threshold Setting

Once the signal is captured, a threshold value is needed to separate the drone signals from background signals, such as AWGN, strong-OFDM, weak-OFDM and DSSS. The threshold is determined from a baseline signal, which is without the presence of drone signals, and it contains at least AWGN or any possible combination of the background signals. The baseline signal would be captured at the very beginning and the power spectrum of the baseline signal is derived from the frequency marginal of TFR:

$$P_x(k) = \sum_{n=0}^{\infty} S_x(n,k) \tag{24}$$

Since AWGN exists across the frequency range, a default threshold P_{T_def} is determined by taking the peak value of the AWGN power spectrum. For OFDM and DSSS, they appear at certain frequency ranges and have higher power as compared to AWGN. A higher threshold is needed at frequency where the OFDM and DSSS signals appear and can be derived as:

$$P_T(k) = \max(P_{T\_def}, P_B(k)) \tag{25}$$

where $P_T(k)$ is the threshold at k frequency, and $P_B(k)$ is the maximum power level of background signals other than AWGN.

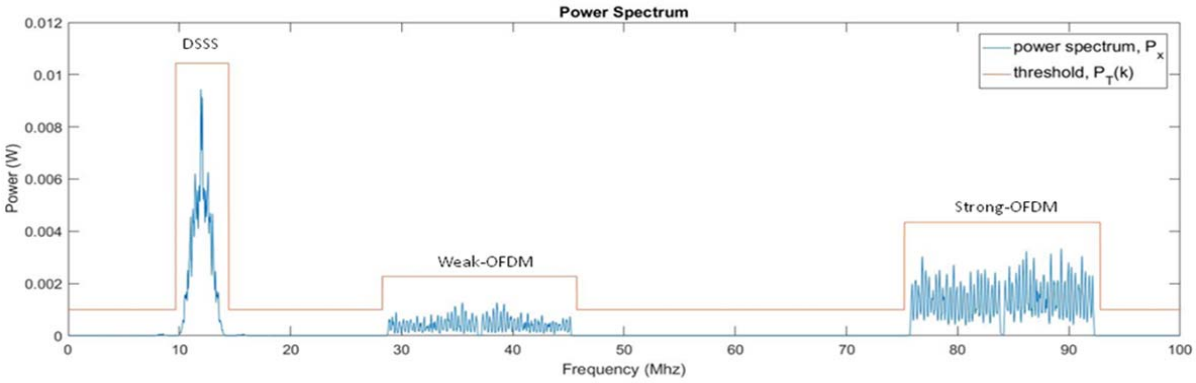


Figure 4: Threshold setting from baseline signal.

The Figure 4 shows the threshold setting $P_T(k)$ across the spectrum where the threshold setting at frequency bands occupied by DSSS and OFDM are adaptively adjusted to its peak value. For this example, the threshold values selected for certain frequency bands are:

(a) For 0 – 10 MHz, 14 – 28 MHz, 45 – 75 MHz, and 93 – 100 MHz, the threshold is 0.001 W
(b) For 10 – 14 MHz, the threshold is 0.012 W
(c) For 28 – 45 MHz, the threshold is 0.002 W
(d) For 75 – 93 MHz, the threshold is 0.004 W

3.4 Signal Detection

After window size and threshold estimation, the next step is to acquire the drone and background signals to obtain the TFR using the estimated window width in Equation 23. The purpose of signal detection is to extract the channel frequency f_c and hop duration T_{hop} of the drone signals from TFR. The channel frequency may vary within an observed time instant but the total number of hop frequencies is the same for a given signal type. Similarly, this is true for the hop duration which for a given signal type is the same. Thus, both hop frequencies and hop duration could be used to identify drone signals.

To estimate the frequency parameters from the TFR, the background signal is separated by applying the estimated threshold on the TFR, and then the power spectrum is derived from the TFR using the frequency marginal in Equation 24. From the power spectrum shown in Figure 5, the frequency channel and signal bandwidth for BFSK can be estimated as:

$$\text{Estimated frequency channel} \quad : \hat{f}_c = (f_{p2} + f_{p1})/\,2 \tag{26}$$

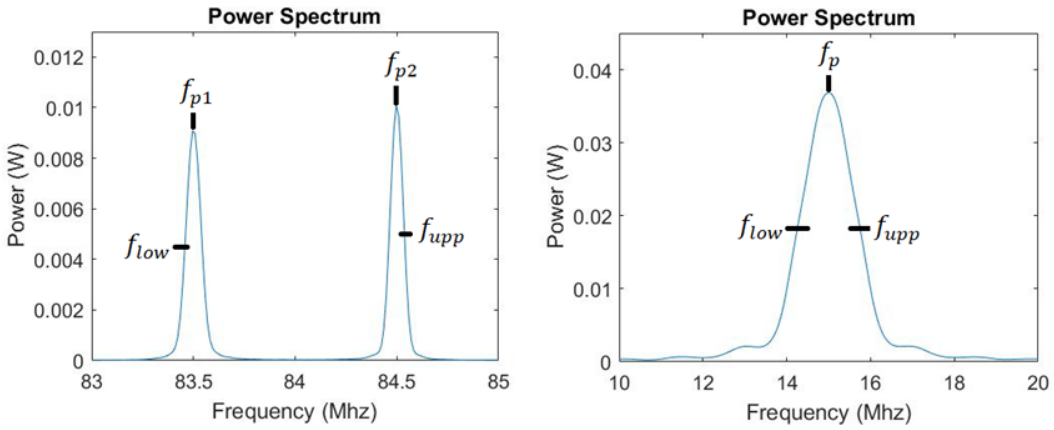
$$\text{Estimated signal bandwidth} \quad : \hat{f}_{BW} = f_{upp} - f_{low} \tag{27}$$

where f_{p1} and f_{p2} are the frequencies at first and second peak, and f_{low} and f_{upp} are the lower and upper frequencies measured from 50% power of the first and second peaks respectively. The BPSK signal can be estimated as:

$$\text{Estimated frequency channel} \quad : \hat{f}_c = f_p \tag{28}$$

$$\text{Estimated signal bandwidth} \quad : \hat{f}_{BW} = f_{upp} - f_{low} \tag{29}$$

where f_p is the frequency at the only peak, f_{low} and f_{upp} are the lower and upper frequencies measured from 50% power of the peak, which is equivalent to the 3 dB or half power point.



(a) Power spectrum for BFSK **(b) Power spectrum for BPSK**

Figure 5: Power spectrum from TFR.

To determine the signal duration at a particular channel frequency, the TFR is evaluated within a frequency range close to the channel frequency and the time characteristics is given by

$$P_t(n) = \sum_{k=f_{low}}^{f_{upp}} S_x(n,k) \tag{30}$$

where f_{low} and f_{upp} are the lower and upper frequency limit respectively within the channel frequency estimated in Equation 28. For each estimated channel frequency, the time characteristics is derived and then used to estimate the hop duration. Figure 6 describes how the hop duration is estimated from time characteristics. The reference level used to estimate hop duration is 50% of waveform peak value, which is consistent with the practice in filter design, where the cut-off point of 50 % power or -3 dB is used to determine the transition between pass band and stop band.

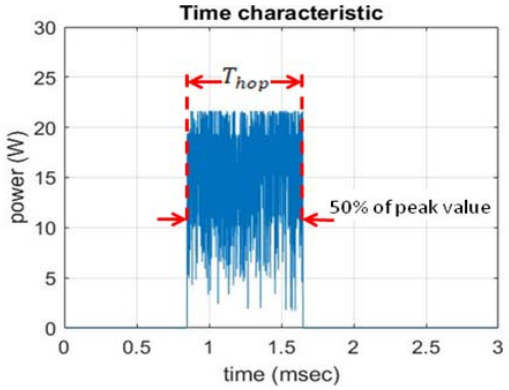


Figure 6: Time characteristic from TFR.

3.5 Instantaneous Frequency (IF) Estimation

The IF provides all the details about the hop pattern as defined by the hop frequency f_c and hop duration T_{hop} throughout the frequency band. This is important as it would provide the input parameters for the signal classifier. In the IF estimation, the first step is to categorise the drone signals based on the hop duration estimate shown in Figure 6 since it remains unchanged throughout the transmission for FHSS

and HSS. This can be done by using the binning method with three predefined ranges of hop duration as per stated in Table 4 where all the estimated hop durations are grouped into the bins that fit the range.

Table 4: Predefined hop duration ranges for binning.

Bin	Range of hop duration (μs)
Bin 1	150 – 250
Bin 2	450 – 550
Bin 3	750 – 850

Each estimated hop duration also carries the information on the hop frequency since the time characteristics is derived from the TFR described in Equation 30. Furthermore, the hop duration estimation also estimates the start time t_s and end time t_e. In the binning method, the bin count for the estimated hop duration is i and the estimated IF can be represented as:

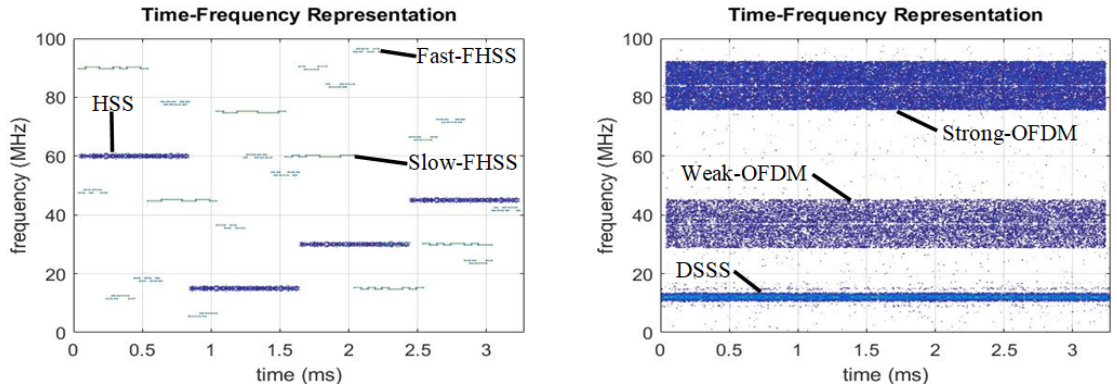
$$\hat{f}_i(t_s(i): t_e(i)) = \hat{f}_c(i) \tag{31}$$

4. RESULT & DISCUSSION

In this section, we demonstrate the estimation of drone signals parameters in a multi-signal environment TFR for ASFSS. The adaptive window size method is verified using the same spectrogram method but without adaptive window size functionality, which is referred to as the stepped frequency scan spectrogram (SFSS). The performance of ASFSS in different multi-signal environments for various SNR is verified by Monte-Carlo simulation.

4.1 TFR

Figure 7 shows the TFR for the captured signal that contains both the drone and background signals. In Figure 7(a), the TFR contains three types of drone signals, which are fast-FHSS, slow-FHSS, and HSS. The background signal TFR in Figure 7(b) contains AWGN and interference signal, such as strong-OFDM, weak-OFDM, and DSSS. The window size N_w used in ASFSS is 978 and this value will be explained in Section 4.2. In general, the TFR describes the signals as defined in Table 1.



(a) Drone signals **(b) Background signals**

Figure 7: TFR of drone and background signals

The fast-FHSS has the shortest hop duration among three drone signals, thus more hops are captured within the signal length of 3.2 ms. Since the fast-FHSS and slow-FHSS are BFSK signal, this result in

two squared power per hop in TFR. For HSS, which is a BPSK signal, it would show one squared power per hop in TFR. For the background signal, AWGN is scattered across all frequencies. The DSSS and OFDM have a constant frequency channel all the time, but DSSS has smaller bandwidth as compared to OFDM. In addition, strong-OFDM has darker colour because it has higher power as compare to the weak-OFDM.

4.2 Adaptive Window Size in Spectrogram

Based on the uncertainty principle for TFR, Equations 21 and 22 can be presented in two hypotheses to ensure that the adaptive window size algorithm functions correctly:

a) The shorter the hop duration, the smaller the smaller the window width. This is because shorter hop duration requires higher time resolution or smaller time per bin T_r to resolve the interval.
b) The smaller the bandwidth, the larger the window width. This is because larger bandwidth requires higher frequency resolution or smaller frequency per bin F_r to resolve frequency.

For concept proving, the results in Table 5(a) shows that when T_{hop} becomes shorter, $\widehat{N}_w$ will become smaller. This satisfies the first hypothesis, where shorter T_{hop} requires higher time resolution or smaller N_w. The experiment is conducted by manipulating the T_{hop} parameter with f_{BW} as a constant and using Equation 23 to get $\widehat{N}_w$.

For the second hypothesis, it is proven by the results of Table 5(b) that shows when f_{BW} becomes smaller, $\widehat{N}_w$ will become smaller. This also matches the seconds hypothesis where larger f_{BW} requires higher frequency resolution or bigger N_w. The experiment is conducted by manipulating the f_{BW} parameter with T_{hop} as a constant and using Equation 23 to get $\widehat{N}_w$.

Table 5: Window size estimation results:

(a) Estimated T_{hop}

Input		output
f_{BW} (kHz)	T_{hop} (us)	$\widehat{N}_w$ (sample)
500	30	387
500	20	316
500	10	224

(b) Estimated f_{BW}

Input		output
T_{hop} (us)	f_{BW} (kHz)	$\widehat{N}_w$ (sample)
50	700	423
50	600	456
50	500	500

The accuracy of the spectrogram is defined by the accuracy of the T_{hop} and f_{BW} of a signal in the TFR plot. Indeed, higher time and frequency resolution would produce a TFR plot with higher accuracy or lower error. The errors of T_{hop} and f_{BW} are represented in absolute percentage error (APE) that can be calculated as:

$$\text{APE for } T_{hop} : APE_T = (\hat{T}_{hop}/T_{hop}) * 100 \quad (32)$$

$$\text{APE for } f_{BW} : APE_F = (\hat{f}_{BW}/f_{BW}) * 100 \quad (33)$$

where $\hat{T}_{hop}$ is the estimated signal hop duration, T_{hop} is actual signal hop duration, $\hat{f}_{BW}$ is the estimated signal bandwidth, and f_{BW} is the actual signal bandwidth. After that, both APE is combined as Average APE (AAPE), which is calculated by:

$$\text{AAPE} : AAPE = (APE_T + APE_F) \,/\, 2 \quad (34)$$

There are many hops that would be appear on the TFR plot, but only the shortest $\hat{T}_{hop}$ and lowest $\hat{f}_{BW}$ are taken account in the spectrogram accuracy evaluation. As long as the spectrogram is able to resolve the shortest T_{hop} and smallest f_{BW}, it can also resolve the longer T_{hop} and larger f_{BW}. Thus, the shortest $\hat{T}_{hop}$ and smallest $\hat{f}_{BW}$ would be recorded and used for AAPE calculation. Based on Table 1, the Fast-FHSS has the shortest T_{hop} of 200 μs and the Slow-FHSS has the smallest f_{BW} of 520 kHz based on Equation 4. The SFSS without the help of adaptive window size is used to compare with ASFSS by performing the TFA using all the commonly used N_w, in the range from 64 to 4096 with progression power of two (64, 128, 256, 512, 1024, 2048, 4096). Therefore, the minimum window size is $N_{w\_min} = 64$ and maximum window size is $N_{w\_max} = 4096$.

Table 6 shows the results of AAPE for different N_w on TFR. Case 1 signals that are defined in Section 2.1 is used since the objective is adapting to drone signal characteristics instead of background signals. For SFSS, the AAPE decreases from $N_w = 64$ until reaching the minimum of 3.76%, when $N_w = 1024$. Then, AAPE increases when $N_w = 2048$ onwards. When $N_w = 64$, the $\hat{T}_{hop}$ is the most accurate estimation while the $\hat{f}_{BW}$ estimation has the highest error, and vice versa when $N_w = 4{,}096$, which contributes to higher AAPE. $N_w = 1{,}024$ has better compromise between time and frequency resolutions, improving the AAPE result.

For ASFSS, since $\hat{T}_h = 200.62\ us$ when $N_w = 64$ and $\hat{f}_{BW} = 524.90\ kHz$ when $N_w = 4096$, the $\widehat{N}_w$ is 978 based on Equation 23. Table 6 shows that ASFSS has the best AAPE at 3.66 % when $N_w = 978$, where the balance between time and frequency resolutions is achieved. Apart from that, it is worth to mention that the TFR accuracy evaluation in (Chee *et al.*, 2014) is based on the symbol duration (SD) and main lobe width (MLW), while ASFSS is based on T_{hop} and f_{BW}, since these are the signal parameters that would help to identify a drone. The SD and MLW estimation requires higher TFR time and frequency resolutions because there are the fine parameters that build up the T_{hop} and f_{BW}. Provided that the cross terms are effectively attenuated, quadratic time-frequency distribution (TFD), such as Wigner-Ville Distribution will always give better measurement accuracy than a spectrogram. However, the spectrogram has the advantage of lower complexity as compared to quadratic TFD.

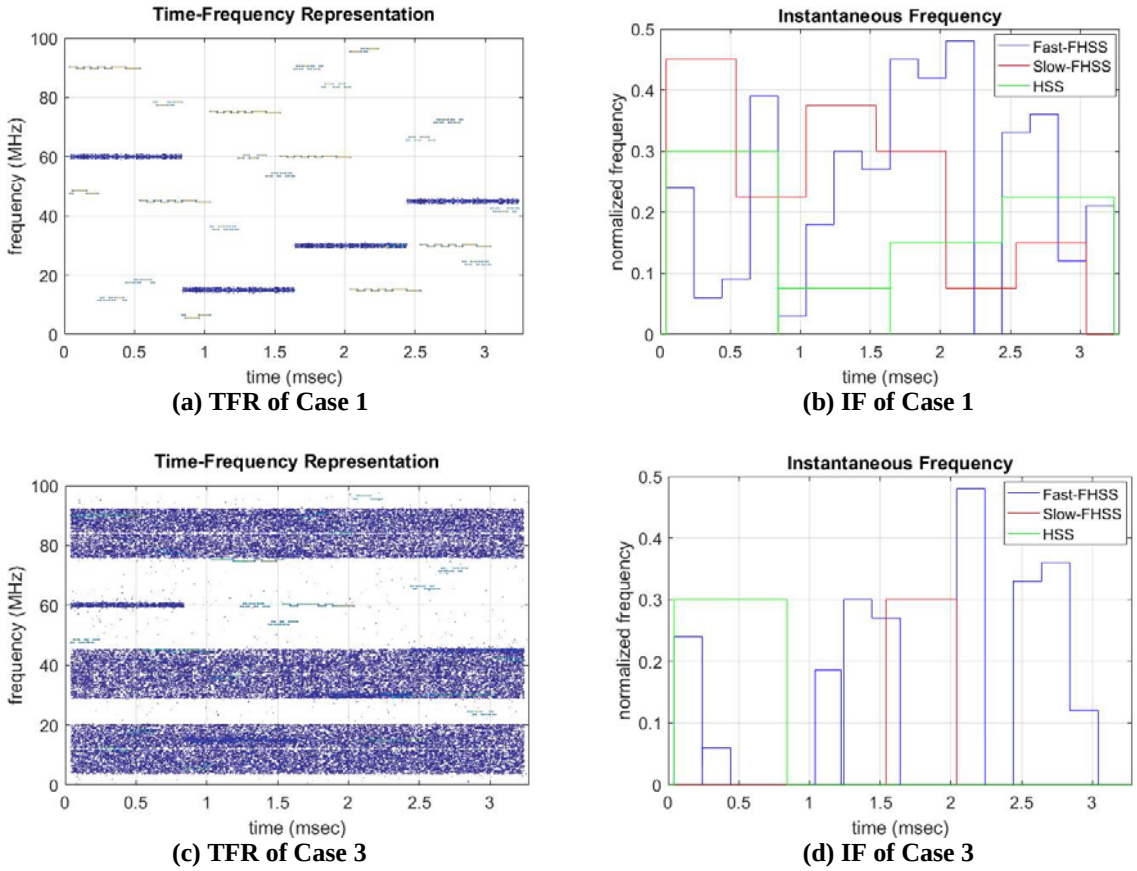
Table 6: AAPE results of TFR for different window sizes N_w.

Method	N_w (sample)	$\hat{T}_h$(μs)	$\hat{f}_{BW}$ (kHz)	APE_T (%)	APE_F (%)	$AAPE$ (%)
Adaptive-SFSS	978	208.04	537.11	4.02	3.29	3.66
SFSS	4096	233.52	524.90	16.76	0.94	8.85
SFSS	2048	216.86	524.90	8.43	0.94	4.69
SFSS	1024	208.46	537.11	4.23	3.29	3.76
SFSS	512	204.12	585.94	2.06	12.68	7.37
SFSS	256	202.16	671.39	1.08	29.11	15.10
SFSS	128	201.04	805.66	0.52	54.93	27.73
SFSS	64	200.62	878.91	0.31	69.02	34.67

4.3 IF Estimation Results

In this section, the IF of drone signals is estimated as explained in Section 3.5. In Figure 8, the IF estimation is evaluated at 7 dB SNR, corresponding to its TFR in multi-signal environment Case 1 (ideal case) and Case 3 (worst case). The 7 dB SNR is used here because the drone signals transmitted power are low enough to show all the three drone signals being detected in the IF plot, especially for Case 3. From the IF estimation in Figure 8(b), all the hop durations of the drone signals are detected because there is no interference from background signals in Case 1. For Figure 8(d), many of the hop durations of the

drone signals could not be detected because of the strong-OFDM higher power and the larger affected frequency band. This causes part of the drone signals to be buried inside the strong-OFDM, resulting in the overall weak drone signals or low SNR. A similar effect is shown for other multi-signal environment cases. In general, the presence of background interference signals, such as strong-OFDM, weak-OFDM and DSSS, would degrade signal detection where the frequency band is interfered. Other than that, the SNR could impact the detection quality, which will be discussed in the next section, where a Monte-Carlo simulation is conducted on the IF at various SNRs in these five multi-signal environment cases.



(a) TFR of Case 1 **(b) IF of Case 1**

(c) TFR of Case 3 **(d) IF of Case 3**

Figure 8: IF estimation from TFR in multi-signal environment Cases 1 and 3.

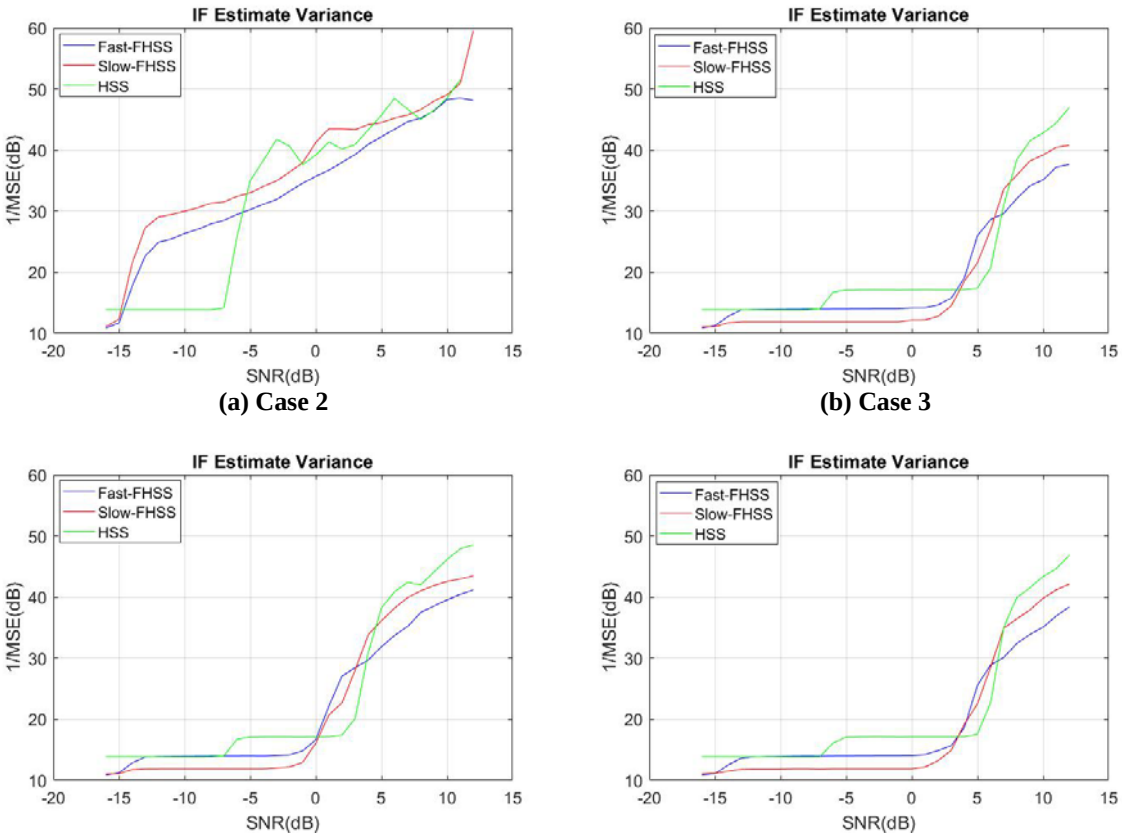
4.4 Monte-Carlo Simulation Results

In this section, IF estimate variance is generated using Monte-Carlo simulation with 20 realisations at SNR range from -16 to 12 dB. The IF estimation for various SNR would be verified in the five multi-signal environment cases defined in Section 2.1, by assuming the IF of Case 1 is the reference IF f_i. The mean-squared-error (MSE) for the IF estimate is calculated by (Chee *et al.*, 2014):

$$var(\hat{f}_i) = \frac{1}{N}\sum_{n=0}^{N-1}(\hat{f}_i(n) - f_i(n))^2 \tag{35}$$

where N is the number of samples. Figure 9 shows the IF estimate variance graph for the drone signals at various SNRs in different multi-signal environments. From the graph, the accuracy of IF decreases gradually as SNR decreases, because the background signal power is increasing, until the cut-off point

where the IF estimate variance starts to fall-off sharply and remain flat. The lower the cut-off point, the better because this is indicated that the drone signals are able to be detected at lower SNR. Based on Table 7, the cut-off point for Case 2 is the lowest because there is no interference signal, such as OFDM and DSSS. Case 4 is the second lowest due to lower power of the weak-OFDM. However, for Cases 3 and 5 are badly interfered by the strong-OFDM and DSSS.



(a) Case 2 **(b) Case 3**

(c) Case 4 **(d) Case 5**

Figure 9: IF estimate variance of drone signals in different multi-signal environments.

Table 7: IF estimate variance cut-off point.

	Drone signals cut-off point (dB)		
Multi-signal environment	Fast-FHSS	Slow-FHSS	HSS
Case 2	-12	-12	-5
Case 3	5	7	8
Case 4	2	4	6
Case 5	5	7	8

The plots below the cut-off point in the graphs are flat, indicating that the drone signal has fallen under the detection threshold. Lowering the threshold could lower cut-off point, but there is always a trade-off between the detection threshold and the background signal separation. Some parts of the signal cannot be detected if the threshold is too high. On the contrary, a low threshold results in false detection, where the background signal is treated as drone signal.

4.5 Relevance to Practical Drone Detection Application

Since Table 7 describes the performance of the ASFSS for estimating the parameters of a drone signal, it is important to relate the relevance of the method in a practical drone detection scenario. There are two possible ways to detect the drone signal: the ground controller and the down link signal. For this scenario, it is assumed that the signal received is from the downlink signal, which may contain the flight and sensor parameters, and the video captured by the onboard camera.

The performance is described by making the following assumptions on first the transmitted drone signal and the receiving station that employs the ASFSS. For the transmitted signal, it is assumed that the frequencies used are 2.4 and 5.8 GHz, transmit power is 100 mW, and an omnidirectional antenna used has 3 dBi gain (Rohde & Schwarz, 2016). The transmitter parameters selected comply with the requirements of MCMC (2015). For the receiver, the parameters selected are described in Table 8. A high antenna gain of 12 dBi can be obtained either using a co-linear antenna or directional antenna, such as Yagi in a sectorised arrangement (SELEX, 2014). The sensitivity is obtained based on a noise figure of 10 dB and receiver bandwidth of 25 MHz. Since line of sight is available up to 45 km at an altitude of 400 ft, the free space loss propagation model is applied and the performance based on the capability to estimate the drone signal parameters according to the range is presented in Table 9 (Sklar *et al.*, 2020).

Table 8: Receiver parameters.

No.	Parameter	Values
1	Received antenna gain	12 dBi
2	Carrier frequency	2.4 and 5.8 GHz
3	Receiver sensitivity	-90 dBm
4	Other losses	3 dB

Table 9: Maximum range for detecting drones for various cases and carrier frequencies.

Case	Cut-off SNR	2.4 GHz	5.8 GHz
2	-5	25 km	10 km
5	8	5 km	3 km

From the results shown in Table 7, higher cut-off SNR is required at 8 dB for Case 5 where the background signal is present as compared to the lower cut-off SNR of -5 dB for Case 2 where the source of interference is only AWGN. When translated to distance based on the receiver parameters defined in Table 8 and free space propagation loss propagation model, the range in general is shorter when a higher frequency is used and in the presence of background signals, as shown in Table 9. Background signals, such as WiFi, have to be considered since there is a possibility that drones operate close to industrial areas, airports or commercial areas, as shown in recent incidents. Further reduction in range is expected if obstacles such as buildings, towers or vehicles are present within the area due to multipath which further downgrades the quality of reception.

5. CONCLUSION

In this paper, The ASFSS is described for use to generate the TFR and detect drone signals in various multi-signal environments. In the 2.4 GHz ISM band with analysis bandwidth of 100 MHz, a smaller sampling rate of 50 MHz is possible, which is four times smaller than the Nyquist rate of 200 MHz. Thus, the method allows the use of lower sampling rate for a given bandwidth. Furthermore, the adaptive window size algorithm introduced with the spectrogram has been proven to have the lowest error based on the AAPE. From the TFR, signal separation and detection are done based on the threshold setting from

baseline signal, estimate f_c and f_{Bw} from power spectrum, estimate $T_{\square op}$ from time characteristic on each f_c, and signal categorisation based on T_{hop} binning. Based on these parameters, the IF estimated from the TFR clearly defines the signal parameters, such as f_c, T_{hop} and hopping pattern. Finally, Monte Carlo simulation conducted to access the robustness of the ASFSS to background signals and AWGN for SNR range from -16 to 12 dB shows that the detection cut-off point is higher by about 12 dB in the presence of background signals such as DSSS or OFDM. Therefore, the detection range from a practical perspective in the presence of background signals is expected to be lower.

ACKNOWLEDGEMENT

The authors would like to thank Universiti Teknologi Malaysia (UTM) for providing the resources for this research.

REFERENCES

Andraši, P., Radišić, T., Muštra, M. & Ivošević, J. (2017). Night-time detection of UAVs using thermal infrared camera. *Transp. Res. Proc.*, **28**: 183-190.

BBC (2018). *Gatwick chaos: Police 'could shoot down drone'*. Available online at: https://www.bbc.com/news/uk-england-sussex-46640033 (Last access date: 1 Jan 2020).

Boualem, B. (2015). The time-frequency approach. *In* Boualem, B., *Time-Frequency Signal Analysis and Processing 2nd Edition*. Elsevier, Academic Press, pp. 20 - 21.

Chee, Y.M. & Sha'ameri, A.Z. (2014). IF estimation of FSK signal using adaptive smoothed windowed cross Wigner-Ville distribution. *Signal Process.*, **100**: 71-84.

Deng, Z., Shen, L., Bao, N., Su, B., Lin, J. & Wang, D. (2011). Autocorrelation based detection of DSSS signal for cognitive radio system. *Int. Conf. Wire. Commun.*, pp. 1-5.

Dobie, G. (2016). *Rise of the Drones*. Reports, Allianz Global Corporate & Specialty.

Ezuma, M., Erden, F., Anjinappa, C. K., Ozdemir, O. & Guvenc, I. (2019). Micro-UAV detection and classification from RF fingerprints using machine learning techniques. *Aerosp. Conf. Proc.*, pp. 1-13.

Fu, H., Abeywickrama, S., Zhang, L. & Yuen, C. (2018). Low-complexity portable passive drone surveillance via SDR-based signal processing. *IEEE Commun. Mag.*, **56**: 112-118.

Fyhn, K., Jensen, T.L., Larsen, T. & Jensen, S.H. (2013). Compressive sensing for spread spectrum receivers. . *IEEE T. Wirel. Commun.*, **12**: 2334-2343.

Guo, J., Liu, L. & Wang, L. (2011). Time frequency representation of frequency hopping signals based on cyclic spectral correlation. *2011 Int. Conf. on Elect. Commun. Cont. (ICECC)*, pp. 1178-1181.

Güvenç, I., Koohifar, F., Singh, S., Sichitiu, L.M. & Matolak, D. (2018). Detection, tracking, and interdiction for amateur drones. *IEEE Commun. Mag.*, **56**: 75-81.

Han, Y. & Jia, G. (2013). Adaptive acquisition threshold algorithm based on mean energy. *Int. Conf. Instr. Meas,*, pp. 634-637.

Javed, F. & Mahmood, A. (2010). The use of time frequency analysis for spectrum sensing in cognitive radios. *2010 4th Int. Conf. on Signal Proces. and Commun. Sys.*, pp. 1-7.

Joo, J., Won, J., Lee, C., Park, S. & Lee, K. (2007). Detection of an unknown FH signal using scanning receiver and DF receiver in practical environments. *IEEE Wcnc.,* pp. 1226-1230.

Lehtom¨aki, J.J., Juntti, M. & Saarnisaari, H. (2004). Detection of frequency hopping signals with a sweeping channelized radiometer. *Conf. Rec. Asilomar C.*, **2**: 2178-2182.

Liu, F., Marcellin, W. M., Goodman, N.A. & Bilgin, A. (2016). Compressive sampling for detection of frequency-hopping spread spectrum signals. *IEEE T. Signal Proces.*, **64**: 5513-5524.

Luan, H. & Jiang, H. (2010). Blind detection of frequency hopping signal using time-frequency analysis. *I. C. Wirel. Comm. Netw.*, pp. 1-4.

MCMC (Malaysian Communications and Multimedia Commission) (2015). *Notice: Use Of Frequencies For Unmanned Aircraft Systems.* Available online at: https://www.skmm.gov.my/en/media/announcements/notice-use-of-frequencies-for-unmanned-aircraft-s (Last access date: 3 Jan 2020).

Meng, F., Zhang, L. & Wang, Y. (2013). Detection of DS &FH hybrid spread spectrum signal in TT & C communication. *Int. Conf. Info. Sci.*, pp. 1242-1245.

Mezei, J., Flaska, V. & Molnár, A. (2015). Drone sound detection. *2015 16th IEEE Int. Symp. Comp. Intelli. Inform.*, pp. 333-338.

Musa, S. A., Abdullah, R. S. A. R., Sali, A., Ismail, A., Rashid, N. E. A., Ibrahim, I. P. & Salah, A. A. (2019). A review of copter drone detection using radar systems. *Defence S&T Tech. Bull.*, **12**: 16-38.

Ochodnický, J., Matoušek, Z., Babjak, M. & Kurty, J. (2017). Drone detection by Ku-Band battlefield radar. *Icmt-Int.Conf. Milit.*, pp. 613-616.

Pei, S. C., & Huang, S. G. (2012). STFT with adaptive window width based on the chirp rate. *IEEE T. Signal Proces.*, **60**: 4065-4080.

Rohde & Schwarz (2016). *Protecting The Sky Whitepaper.* Available online at: http://www.rohde-schwarz-usa.com/rs/324-UVH-477/images/Drone_Monitoring_Whitepaper.pdf (Last access date: 12 Jan 2020).

Rozantsev, A., Lepetit, V. & Fua, P. (2015). Flying objects detection from a single moving camera. *Proc. Cvpr. IEEE*, pp. 4128-4136.

Schulze, H. & Luders, C. (2005). *Theory and Applications of OFDM and CDMA*. John Wiley & Sons Ltd, pp147.

Sha'ameri, A. Z. & Kanaa, A. (2016). Robust multiple channel scanning and detection of low probability of intercept (LPI) communication signals. *Defence S&T Tech. Bull.*, **9**: 1-17.

SELEX Sistemi Integrati (2014). *ADS-B Subsystem: Standard E5010015201SDD*.

Shi, X., Yang, C., Xie, W., Liang, C., Shi, Z. & Chen, J. (2018). Anti-drone system with multiple surveillance technologies: architecture, implementation, and challenges. *IEEE Commun. Mag.*, **56**: 68-74.

Shin, H., Choi, K., Park, Y., Choi, J. & Kim, Y. (2016). Security analysis of FHSS-type drone controller. *Lect. Notes. Comput. Sc.*, **9503**: 240-253.

Sklar, B. & Harris F. J. (2020). *Digital Communications 3rd Edition*. Prentice Hall.

Tang, P., Lin, Q., Yuan, B. & Chen, Z. (2012). Efficient digital channelized receiver based on subband decomposition and DFT filter banks. *Int. Conf. Sign. Proces.*, **1**: 417-420.

Zhang, Q., Zhang, X. & Liu, Y. (2016). Parameter estimation of non-modulated or modulated frequency-hopping signals. *IEEE Int. Conf. on Signal Proces. Commun. and Comput. (ICSPCC)*, pp. 1-4.

Zhong, J. & Huang, Y. (2010). Time frequency representation based on an adaptive short-time Fourier transform.*Int. Conf. Sign. Proces.*, **58**: 5118-5128.

A METHOD FOR SYNTHESIZING THE STRUCTURE OF ACTIVE LOADED BLOCK RESERVATION OF SUBSYSTEMS USING THE GRAPH-ANALYTICAL MODEL AND A WAVE OPTIMIZATION ALGORITHM

Vyacheslav M. Grishin

Department of Systems Analysis and Management, Moscow Aviation Institute (National Research University), Russian Federation

Email: vyacheslav.grishin@gmail.com

ABSTRACT

One of the most important problems facing developers of subsystems of technical systems (TS) for various purposes is the task of increasing their reliability. The purpose of this article is to develop a method for synthesis of structures for active loaded block reservation of subsystems. The main method of study of this problem was the analysis method, which allowed us to develop a way for optimizing the structure of block reservation, which can significantly increase the likelihood of uptime as compared to non-optimal options. Among the many ways to increase the reliability of subsystems associated with the introduction of reservation to the system, the study discusses a method of increasing the reliability of subsystems with active loaded reservation by choosing the optimal block size. The task of selecting the optimal block size of reservation subsystems was set in the work. A graphical and analytical model of the reservation structure was developed using a wave algorithm for optimizing the reliability of subsystems with sudden independent element failures. A method for synthesizing the structure of block reservation was described. Examples of calculations were presented. The proposed method is applicable to any problems that can be reduced to optimization on both oriented and non-oriented weighted graphs, with the complexity not exceeding polynomial of third-order.

Keywords: *Reliability scheme; reservation ratio; optimal block size; structural reliability scheme; edge graph.*

1. INTRODUCTION

With passive (permanent) reservation, as well as with active reservation in the case of perfect operation of the switches, the smaller the reservation scale, the higher is the reliability of the technical system (TS) (Gliniak *et al.*, 2018). The level (scale) at which the reservation is made may be different. We can reserve individual elements, units, subsystems, and finally the whole system. This research considers block active loaded reservation of TS subsystems. This is due to the fact that active reservation, in contrast to passive, does not lead to a change in the parameters of considered class of subsystems in case of failures. Moreover, active loaded reservation, although it requires a resource consumption of reservation elements, significantly reduces the time delay when replacing failed elements due to their continuous readiness for work (Rabinskiy & Vakhneev, 2018; Unaspekov *et al.*, 2019).

As it is known, active loaded reservation, as well as passive reservation (Epifanov, 1975; Grishin & Trong Tuan, 2018), when comparing the failure-free operation indices of reserved and non-reserved subsystems, have critical probabilities of elements of non-reserved subsystems and the corresponding areas in which the redundancy of reserved subsystems is worse than non-reserved ones. Nonetheless, the reasons for the presence of these areas, as well as their properties for these types of reservations, differ significantly. With active reservation, the cause of critical probabilities lies in imperfect

operation of the switches. As the reliability of the switches approaches one, the critical probabilities tend to zero (Grishin & Ko, 2009).

One of the properties of active loaded reservation of subsystems is that separate (element-wise) connection of reserve elements is not the best from the point of view of their failure-free operation. Thus, with active loaded reservation and imperfect operation of the switches, the smallest reservation scale (element-by-element reservation) may be far from optimal. The problem arises when optimizing the structure of active loaded reservation associated with finding the optimal block size (Redko, 2002).

The solution to this problem is enhanced by the possibility of dividing technical systems into subsystems of various levels of decomposition with a small number of elements. Then, any vehicle can be divided into subsystems of the first level. Each subsystem of the first level can be divided into subsystems of second level. The division process can continue until indivisible components are obtained. Then, at any level of division, each subsystem consists of a small number (from 3 to 20) of subsystems of the underlying level, which can be considered as its indivisible elements. The TS subsystems, depending on type and purpose, are distinguished by a wide variety of their design features and reliability indicators (Redko, 2002; Lugovtsova, 2015; Tretyakov, 2016).

Nonetheless, despite the wide variety of TS subsystems, all of them, in absence of reservation, have a structural diagram for calculating reliability indicators, often called a reliability structural diagram (RSD), in the form of a basic (serial) connection of elements. This is because of the fact that in the absence of structural reservation associated with the introduction of reservation elements into the system, the failure of any element leads to a failure of the subsystem. In this case, the problem arises of how to draw up a block diagram of sequences of reservation blocks (in the range of resizing blocks from general to separate reservation) so that it provides the maximum failure-free rate, if there are restrictions on the technical implementation of reservation of some blocks (Reinhold *et al.*, 1980).

In the course of the study, a method for synthesizing the structure of the active loaded block reservation of subsystems using a graph-analytical model and a wave optimization algorithm was developed. This method contributes to solving the problem of increasing the reliability of systems in the aerospace field, in particular, the search for new methods for synthesizing the reservation structures of systems containing a large number of elements.

2. STATEMENT OF THE PROBLEM OF OPTIMIZING THE STRUCTURE OF THE ACTIVE LOADED BLOCK RESERVATION SUBSYSTEM

Let us consider the problem of optimizing the structure of the active loaded block reservation of a subsystem in case of sudden independent failures, when the failure-free performance of its elements and switches can vary significantly. Let us assume that the switches are combined with parameter indicators and have a generalized reliability indicator. As an indicator of uptime, we will use a very important and popular indicator in the form of probability of uptime (Tolstov & Pantyukhov, 2016; Boykov *et al.*, 2018). Let us accept that the following notations: $P_1(t_з) = P_1$, $P_2(t_з) = P_2$, $P_3(t_з) = P_3$, ... $P_m(t_з) = P_m$ are the probability of failure-free operation of non-reserved elements during the execution of the task $t_з$; while m is the number of basic elements in the subsystem.

The probability of uptime of the unreserved subsystem P_c is obviously determined by multiplication of probabilities of the failure-free operation of the elements. As mentioned in Polovko (1964), there are two ways to connect this type of reserve:

1) All blocks of main and backup subsystems are equipped with switches;

2) The switches are located only in the blocks of all reserve subsystems and are not contained in the blocks of the main subsystem.

Without loss of generality, let us consider the first method of connecting a reserve. Figure 1 shows a block diagram of a block active reservation of a subsystem with connecting reserve according to the first method.

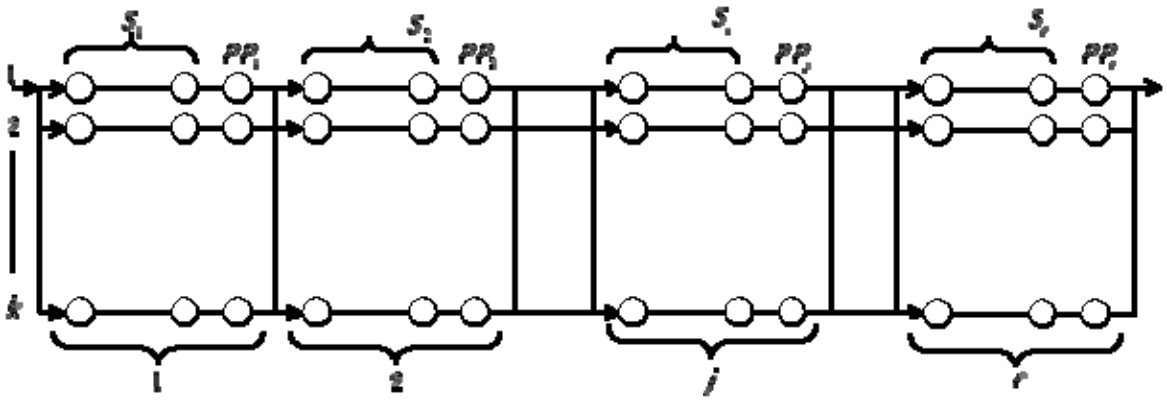


Figure 1: Structure diagram of subsystem block active reservation.

The following notations are used in the diagram:

S_j is the size of j^{th} reservation block;
r is the number of backup units;
k is reservation ratio;
$P_c(t_з) = P_c$ is the probability of failure-free operation of the reservation subsystem during the execution of the task $t_з$;
$PP_j(t_з) = PP_j$ is probability of failure-free operation of switches of j reservation block in time $t_з$.

All reserve elements of the unit, as well as the switches serving them, are identical with the main ones respectively. By the multiplicity of reservation in work, we mean the ratio of the total number of all elements of the system (primary and reserve) to the number of basic elements. The probability of failure-free operation of the reserved subsystem with the same reservation ratio of individual blocks is determined using the following equation:

$$P_s = \prod_{j=1}^{r} \{1 - [1 - PP_j \cdot \prod_{i=1}^{S_j} P_{i,j}]^k\}, \quad (1)$$

where j is the backup block number, and i is the element number (primary or backup) in j^{th} block.
From Equation 1, we can see that at $r = 1$, block reservation is transformed into general reservation, and at $r = m$, into separate reservation.

The task of optimizing the reservation structure is as follows: for given system parameters of $m, k, P_{i,j}$, and PP_j we need to find such block sizes $S_1, S_2, ..., S_j, ..., S_r$ that provide the maximum reliability P_c^* of the reserved subsystem, i.e.:

$$P_S^* = \max_{\{S_j\}} \prod_{j=1}^{r} \{1 - [1 - PP_j \cdot \prod_{i=1}^{S_j} P_{i,j}]^k\} \quad (2)$$

with restrictions on technical implementation of some blocks S_j:

$$- \sum_{j=1}^{r} S_j = m,\ 1 \le r \le m, \tag{3}$$

where r, S_j – integers.

The solution of Equation 2 under the constraints of Equation 3 results in a complex combinatorial problem. In order to resolve the computational difficulties arising from this, an optimization method was developed on non-classical (weighted) graphs. Many studies have been devoted to such methods (Soltani & Shafiei, 2011; Mahajan & Tatikonda, 2015; Chattamvelli & Shanmugam, 2019).The method is based on: 1) graph-analytical model using oriented weighted edge graph containing analytical formulas for calculating reliability indicators of all edge blocks with all kinds of subsystem block reservation methods; 2) wave algorithm for finding the optimal route on the graph.

3. GRAPHICAL ANALYTICAL MODEL OF BLOCK RESERVATION

The model is based on oriented weighted edge graph. In the absence of restrictions on technical implementation of some blocks, a fully connected oriented edge graph can be used to solve the problem. In the presence of these restrictions, the edges of the graph corresponding to the forbidden blocks should be deleted. In the beginning, we construct an edge graph for the non-reserved subsystem. It contains $(m+1)$ top and m ribs. Figure 2 shows the edge graph for the non-reserved subsystem.

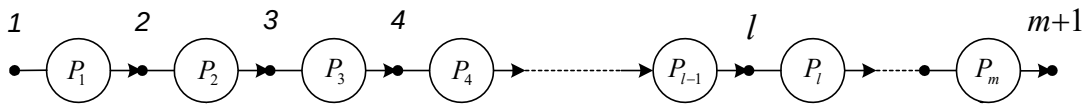


Figure 2: Edge graph for non-reserved subsystem.

Then, each edge of the graph is associated with a "value" in the form of the probability of failure-free operation of the corresponding unreserved element of the subsystem. The edge graph of the reserved subsystem is constructed based on the edge graph of non-reserved subsystem. In order to do this, at first, on each oriented edge of the graph of the unreserved subsystem, a new value is affixed, corresponding to the probability of failure-free operation of the length unit (separate reservation) with a given reservation ratio k. For example, the edges between the vertices 1-2,2-3, ..., m - $(m+1)$ – should have values in the form of probability of failure-free operation (Equations 4 - 6):

$$P^{r}_{1,2} = 1 - (P_{1,2} \cdot PP_{1,2})^{k}; \tag{4}$$

$$P^{r}_{2,3} = 1 - (P_{2,3} \cdot PP_{2,3})^{k}; \tag{5}$$

$$P^{r}_{m,m+1} = 1 - (P_{m,m+1} \cdot PP_{m,m+1})^{k}, \tag{6}$$

where: $P^{r}_{1,2}, = P_1, P^{r}_{2,3} = P_2, \ldots, P^{r}_{m,m+1} = P_m$, and $PP_{1,2}, PP_{2,3}, \ldots, PP_{m,m+1}$ are probabilities of uptime of the switches serving the respective units. The initial graph with new values takes the form presented in Figure 3.

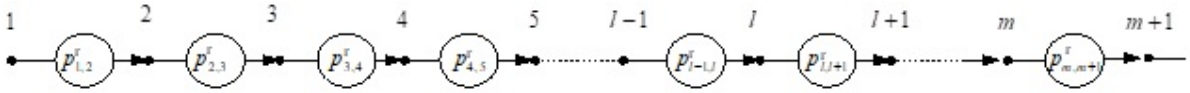


Figure 3: The edge graph of the original subsystem with new separate reservation values.

Figure 4 shows a structural diagram of reliability of the subsystem with separate reservation of elements corresponding to the edge graph shown in Figure 3.

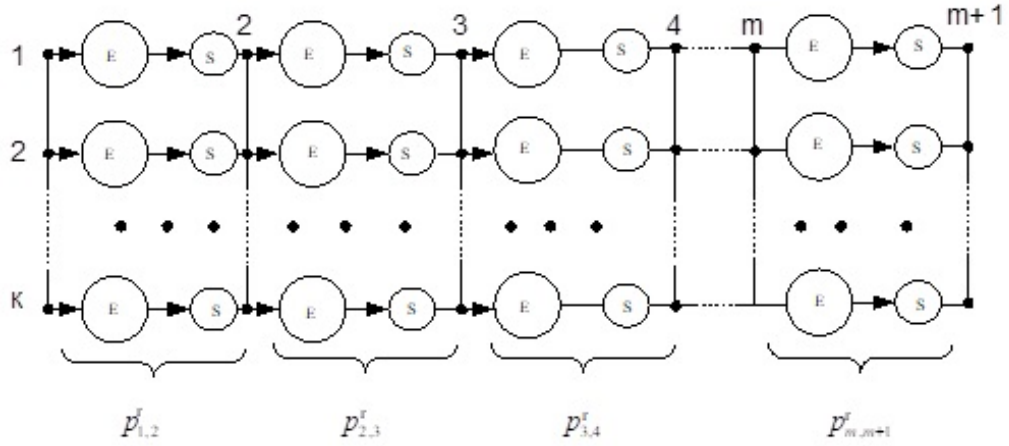


Figure 4: The structure diagram of reliability of separate block reservation (block size is 1, where *E* are subsystem elements and *S* are switches serving shared redundant units.

A comparison of Figures 3 and 4 shows that the probability of failure-free operation of this kind of the reserved subsystem is determined by the multiplication of the values of *m* for the consecutively-connected edge blocks. This block reservation kind (Figures 3 and 4) is the first step in converting the original edge graph (Figure 2). Further transformations of the edge graph of the original subsystem goes as follows. In Figure 2, sequentially oriented edges are drawn from each vertex (starting from the first) to all the others (from left to right), avoiding the image of two or more oriented edges between any pair of vertices. The resulting edge graph is shown in Figure 5.

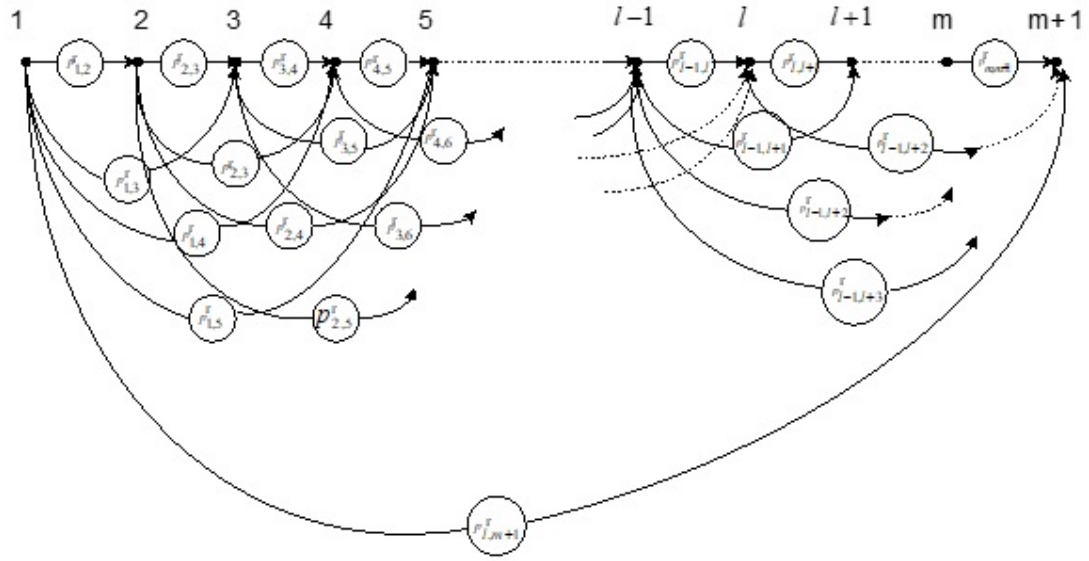


Figure 5: Formation of the oriented edge graph of the reserved subsystem.

The block reservation structure corresponding to the graph shown in Figure 5 is impossible to depict, since this structure reflects all possible ways of reserving the original system. Only separate block reservation options can be depicted. For example, the oriented edge of 1 – (m + 1) in Figure 5 with value of $P^{r}_{1,m+1}$ corresponds to the structural scheme of one block with a common reservation, with the block reservation structure presented in Figure 6.

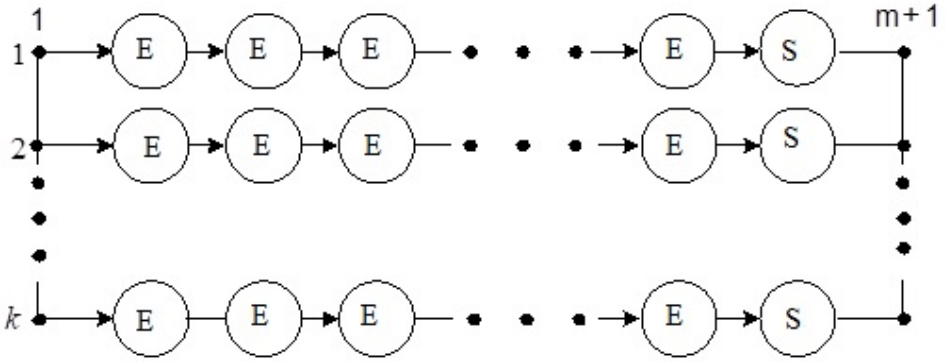


Figure 6: Block reservation diagram for 1 – (m + 1) corresponding to the graph, where *E* are subsystem elements and *S* are switches serving shared redundant units.

The generated model of the edge graph demonstrated in Figure 6 is convenient for presentation in the form indicated in Figure 7.

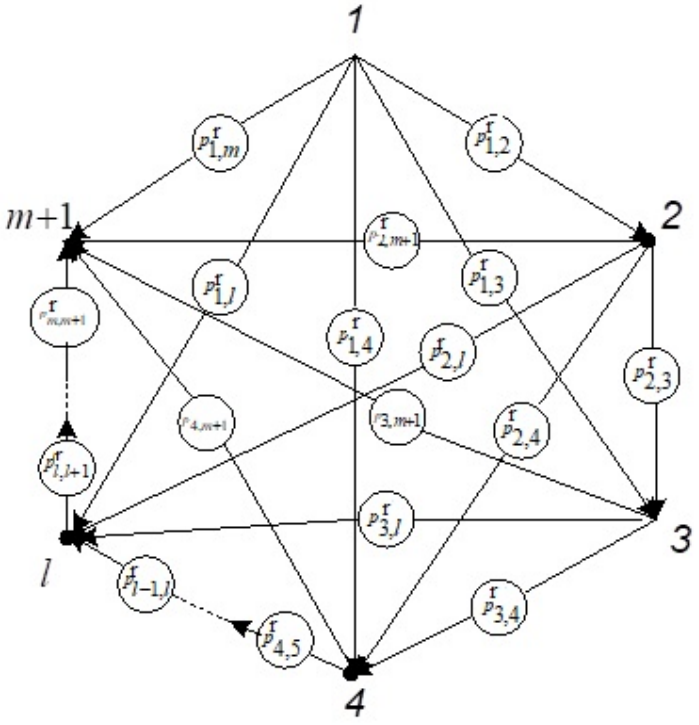


Figure 7: Edge graph of reserved subsystem, where $P^{r}_{1,2},\ldots,P^{r}_{1,m+1}$ are uptime probabilities of various redundant units.

This graph contains $(m + 1) \cdot m / 2$ oriented edge blocks in the absence of technical restrictions on implementation of some blocks. The entire set of edges of the graph reflects all ways to form reservation blocks. Therefore, any block has a start vertex and an end vertex. Each oriented edge, as in the previous case, is assigned a value equal to the probability of failure-free operation of the corresponding reservation block. For example, a block including a sequence of vertices g, $g + 1$, $g + 2$, ..., $h + 1$ of the graph shown in Figure 7 has the following value:

$$P^{r}_{g,} = 1 - (1 - P_g \cdot P_{g+1} \cdot \ldots \cdot P_h \cdot PP_{g,h})^k \cdot \quad (7)$$

Figure 7 shows that on the formed graph, there are many routes from vertex 1 to vertex $m + 1$ in the form of sequences (chains) of oriented edge blocks. Each route defines one of the methods for block reservation of the system in the range from separate (route 1-2-3– ... -$m + 1$) to general (route 1– ($m +$ 1)). We define the value of the route as multiplication of the values included in its edge blocks. Then, these values are equal to the probability of system uptime with this reservation option. Thus, the solution to Equation 2 with the restrictions from Equations 3-5 is reduced to finding a route with the maximum value.

4. DEVELOPMENT OF WAVE ALGORITHM FOR OPTIMIZING BLOCK RESERVATION

In order to find the optimal route (in the sense of its maximum value), it is proposed to use wave algorithm. Its essence can be reduced down to the following. A wave is launched from the initial vertex, which propagates in steps in all possible directions of the graph. In one step, the wave moves from the current state to a new state by one edge. Conventionally, we can separate the leading wave front and the trailing wave front, between which optimization takes place. At the first step, the wave moves from the initial vertex to adjacent vertices along all edges emanating from it. Vertices to which the first wave has reached define its leading edge. Each vertex to which the wave has reached in the first step is assigned the value of the edge along which the wave has come into it (input value), and also the number of the vertex from which the wave has arrived (input address) is stored (Polovko, 1964).

Then the peaks shall be fixed, to which the wave reaches the second step by enumerating all the edges coming from the vertices to which the wave reached in the first step. The value of any vertex in the second step is determined by the multiplication of the values of the edges entering into it and the value of the vertex coming from this edge (outgoing). If any – top of the second step (and subsequent steps) includes several of oriented edges, the value of this peak is defined as the maximum of possible values. This ensures that the vertices of the graph are assigned a relative maximum (assignment to the vertices of the maximum value of several possible values in the current step). The step-by-step process of wave propagation continues until the trailing edge of the wave passing along the longest paths (in terms of number of edges included in them) reaches all the vertices (Soltani & Shafiei, 2011).

Then, behind the leading front, the wave can repeatedly come to the same peak along different routes at different steps. Each time in these situations, the algorithm updates the input value and the input address of the vertex if the new value is higher than previously recorded. This ensures that each vertex is assigned a conditional maximum value. Upon reaching the end of the wave propagation, each vertex contains an absolute unconditional maximum, i.e., the maximum value with which the wave came to each vertex along the optimal route from the initial first vertex. Optimal routes are easily traced by the input addresses assigned to each vertex (Kozlov & Ushakov, 1975).

The developed method is programmed in MATLAB. For using the wave algorithm, the following data is used. The input for the algorithm are a mathematical model of connectedness of directed graph of reserved system in the form of list of connectedness; a mathematical model of the values of the edges

of the directed graph of the redundant system in the form of a two-dimensional array of probabilities of failure-free operation; and the values of all oriented edges. Each line in the list corresponds to a specific vertex of the graph. The i^{th} line of the list indicates the numbers of the vertices of the graph with which the i^{th} vertex is connected. The elements of the array of values are described by Equations 8 – 16 (Reinhold *et al.*, 1980):

$$MV(1,2) = 1 - (1 - p_1 \cdot p_{p1,2})^k; MV(1,3) = 1 - (1 - p_1 \cdot p_2 \cdot p_{p1,3})^k \quad (8)$$

$$MV(1, m+1) = 1 - (1 - p_1 \times p_2 \times \ldots p_m \times p_{p1,m+1})^k; \quad (9)$$

$$MV(2,3) = 1 - (1 - p_2 \times p_{p2,3})^k; \quad (10)$$

$$MV(2,4) = 1 - (1 - p_2 \times p_3 \times p_{p2,4})^k; \quad (11)$$

$$MV(2, m+1) = 1 - (1 - p_2 \times p_3 \times \ldots p_m \times p_{p2,m+1})^k; \quad (12)$$

$$MV(3,4) = 1 - (1 - p_3 \times p_{p3,4})^k; \quad (13)$$

$$MV(3,5) = 1 - (1 - p_3 \cdot p_4 \cdot p_{p3,5})^k; \quad (14)$$

$$MV(3, m+1) = 1 - (1 - p_3 \times p_4 \times \ldots p_m \times p_{p3,m+1})^k; \quad (15)$$

$$MV(m, m+1) = 1 - (1 - p_m \times p_{pm,m+1})^k. \quad (16)$$

The output data are a vector of optimal routes (1, m + 1) and a vector of optimal route values (1, m + 1). The first element is considered as a source of the wave. Each i^{th} element of the vector of optimal routes (i > 1) contains the number of the vertices from which the wave came to the i^{th} vertex along the edge of vector's length (equals one) in the optimal route. According to the vector of optimal routes, they are easily formed from the first peak to any other. Each i^{th} element of the value vector of optimal routes (i > 1) contains the maximum value with which the wave came to the i^{th} vertex from the initial first vertex. Therefore, the two output vectors contain the solution to the problem posed, not only containing subsystem m elements, but for all smaller subsystems. Elements of the output vectors (with the exception of the first one) are regularly updated during the step-by-step operation of the wave algorithm, receiving the final values after termination of the wave motion.

The intermediate data are a vector of the old state of the wave (a vector that determines the position of the wave at the current step) and a vector of the new state of the wave (a vector that determines the position of the wave at the new step). The elements of the vector of the old state of the wave contain the numbers of the vertices that the wave have reached by the current step. Elements of the vector of new state of the wave has the numbers of vertices reached by the wave in the next step. At the end of each step, the vector of the new wave state is rewritten, and the content of the vector of the old wave state and old vector is zeroed. The work of the algorithm ends (wave motion ends) when the two output vectors (vectors of optimal values and routes) cease to change, and the vector of the new wave state becomes zero.

It will be interesting to compare the complexity of the considered method with the enumeration method for solving the problem. In order to solve the problem by enumeration, we need to determine the number of oriented routes leading from the initial vertex of a fully connected edge graph to the final vertex. Since each route determines one block reservation method, it is necessary to calculate the

probability of the subsystem's failure-free operation for each reservation method as the product of the probability of failure-free operation of the edge blocks included in each route with a given multiplicity. The number of oriented routes is easy to calculate by summing the number of routes of different lengths.

By the length of the route, we understand the number of intermediate vertices included in the route between the initial vertex and the final vertex. For the number of vertices of the graph of the non-reserved subsystem (m + 1), the lengths of the routes take the values: 0, 1, 2, ..., (m - 1). Therefore, it is necessary to count the number of routes: without intermediate vertices between the initial and final nodes; with one intermediate vertex and then two, three, intermediate vertices and so on; and finally to (m - 1)th intermediate vertex. It is clear that the number of routes with i intermediate vertices is determined by the number of combinations of (m - 1) elements with i elements. The total number of oriented routes M from the initial vertex to the final one is determined by their sum, which comes from the Newton binomial expansion formula in the following form:

$$M = C^0{}_{m-1} + C^1{}_{m-1} + \ldots + C^{v-1}{}_{m-1} = 2^{m-1}. \quad (17)$$

As we can see from Equation 17, the problem posed in the work for the enumeration method is exponentially complicated, that is, it is difficult to solve; while for the developed method it is polynomially solvable. Actually, a wave moving along edges of unit length maximally makes m steps (along the longest route). During each step (except the first), the wave can propagate maximally from r ($0 < r < m + 1$) vertices when forming the vector of the old state of the wave. Each vertex of this vector can be connected with a maximum of m vertices during the formation of the vector of the new wave state. Therefore, the complexity of solving the problem using the developed method is not higher than a third-order polynomial from the size of the problem m. In comparison, the Dijkstra method allows us to solve the problem in time set by a second-order polynomial. While the developed method is less efficient than the Dijkstra method, it is more universal, since, unlike the latter, it allows solving optimization problems on both oriented and non-oriented graphs with the same complexity, defined by a third-order polynomial. With the closed loops in non-oriented graphs, the wave motion never loops due to the fact that when the waves re-enter them, the values of the vertices of these loops cease to be updated.

5. THE RESULTS OF CALCULATIONS PERFORMED BY THE DEVELOPED METHOD

Here are some results illustrating the possibilities of the proposed approach for various options of the source data. In our calculations, we will assume that there are no technical restrictions on the options for the formation of blocks of reservation.

Option 1. The unreserved subsystem has five elements (m = 5). The edge graph of original system has six vertices and it is shown in Figure 8.

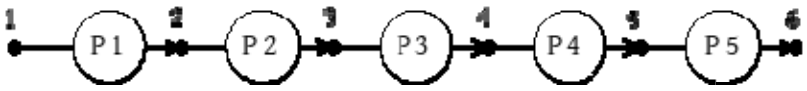


Figure 8: Edge graph of the unreserved subsystem.

The probability of failure free operation of the main and reserved elements have the following values: P_1 = 0.978; P_2 = 0.975; P_3 = 0.970; P_4 = 0.972; and P_5 = 0.964. The reliability of the switches for all combinations of block reservation is assumed PP = 0.95, with reservation ratio k = 2.

As a result of the calculations, we get the optimal route and the corresponding reliability indicator. The optimal route (OR) from vertex 1 to vertex 6 has the form: OR = 1-3-5-6, where 1, 3, 5 and 6 are the numbers of the vertices of the edge graph of the unreserved system. The subsystem failure free index P_c is 0.9734. Figure 9 shows the optimal reservation structure corresponding to the calculated optimal route, whereby Elements 1 and 2 form the first block, Elements 3 and 4 form the second block, and Element 5 forms the third block.

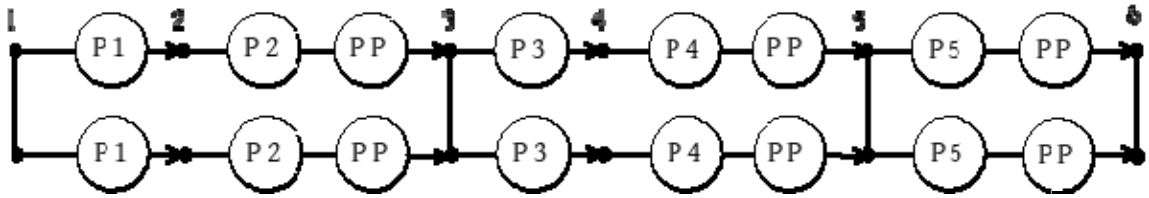


Figure 9: The optimal structure of block reservation for Option 1.

Option 2. Let us change the probability values P_1 and P_5, leaving the rest of the original data the same, i.e., P_1 = 0.964 (instead of 0.978), and P_5 = 0.978 (instead of 0.964). As a result of the calculations, we obtain OR = 1-2-4-6 and P_c = 0.9734. We can see that the option of block reservation has changed with a constant indicator P_c.

Option 3. The source data of this option corresponds to Option 1. Let us increase the reservation ratio by 1, i.e., k = 3. As a result of the calculations, the optimal route (OR) from vertices 1 to 6 has the form of 1-2-3-4-6, where 1, 2, 3, 4 and 6 are the numbers of the vertices of the edge graph of the unreserved system, while the subsystem failure free index P_c is 0.9992. The optimal reservation structure corresponding to this option is presented in Figure 10, whereit can be seen that elements 1 and 2 form the first block, elements 3 and 4 form the second block, and elements 5 and 6 form the third block with a reservation ratio of 3.

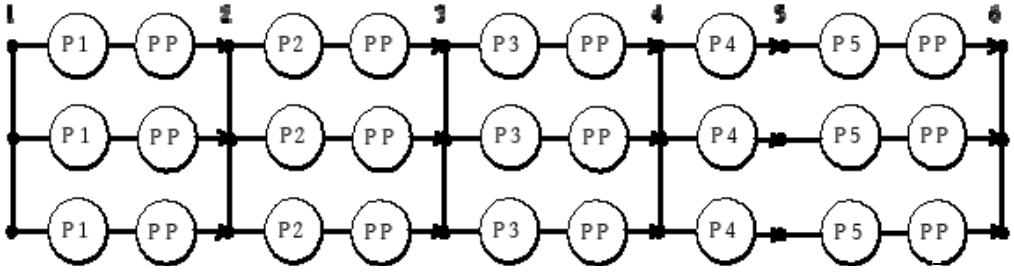


Figure 10: Optimal block reservation structure for Option 3.

Option 4. The source data of this option corresponds to Option 3. Let us increase the multiplicity once more by 1, i.e., k = 4. As a result of the calculations, we obtain OR = 1-2-3-4-5-6 and P_c = 0.9999.The optimal reservation structure corresponding to this option is presented in Figure 11. It is observed that with increasing reservation multiplicity, the optimal block reservation structure tends to be separated reservation.

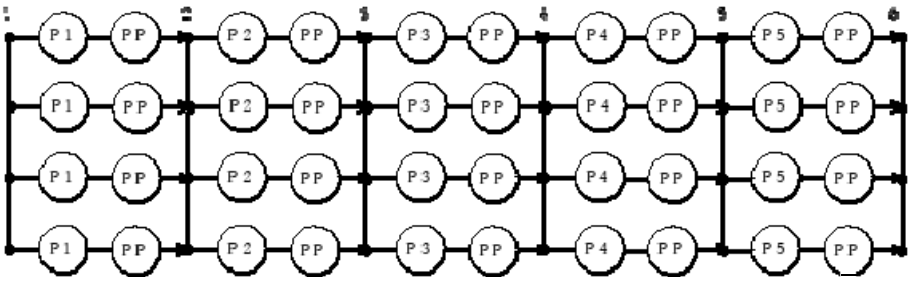


Figure 11: Optimal block reservation structure for Option 4.

Option 5. Let us accept $m = 5,\ k = 2$. We will set various values of the reliability indicators not only of elements P_i, but also switches of various blocks $PP_{l,h}$, where l is the beginning of the block on the edge graph, and h is the end of block *l*, whereby $h = 1,2, ..., m,\ h > l$ (Table 1).

Table 1: Various values of the reliability indicators.

Indicator in the form of probability of uptime, *P*	Probability of failure-free operation of switches, *PP*
1 = 0.978	1,2 = 0.950
2 = 0.975	1,3 = 0.989
3 = 0.970	1,4 = 0.973
4 = 0.972	1,5 = 0.987
5 = 0.964	1,6 = 0.985
	2,3 = 0.995
	2,4 = 0.980
	2,5= 0.982
	2,6 = 0.985
	3,4 = 0.940
	3,5 = 0.968
	3,6 = 0.971
	4,5= 0.945
	4,6 = 0.977
	5,6 = 0.952

As a result of the calculations, we get OR = 1-3-6 and P_c = 0.9892. In this option, the optimal reservation structure contains two blocks. The first block contains Elements 1 and 2 with a switch $PP_{1,3}$, while the second block contains Elements 3 - 5 with a switch $PP_{3,6}$.

Option 6. Let us accept $m = 20,\ k = 2$. The switches of all blocks shall be the same - $PP = 0.95$. The reliability of the main and reserved elements is set by the probabilities shown in Table 2.

As a result of the calculations, we obtain OR = 1-5-6-13-18-18-21 and P_c = 0.8243. The optimal reservation structure contains five blocks with elements: 1 - 4; 5; 6 - 12; 13 - 17; and 18 - 20. Let us calculate the probability of failure of the subsystem for the source data of this option in addition to the case of general and separate reservation, whereby we obtain: P_{gen} = 0.7673 and P_{sep} = 0.7249.

The calculated indicators are much lower than the optimal value of P_c = 0.8243. Therefore, optimization of the structure of block reservation can significantly increase the likelihood of failure free operation as compared to non-optimal options.

Table 2: the probabilities that set the reliability of the main and reserved elements.

Probability No.	Value
1	0.985
2	0.955
3	0.957
4	0.959
5	0.960
6	0.973
7	0.985
8	0.981
9	0.976
10	0.964
11	0.967
12	0.958
13	0.955
14	0.962
15	0.959
16	0.973
17	0.979
18	0.961
19	0.987
20	0.980

6. CONCLUSION

We have developed a method for synthesizing the structures of active loaded block reservation of subsystems based on the use of a graph-analytical model and a wave algorithm for optimizing reliability. This method allows us to build the optimal block reservation structure, which significantly increases the probability of failure-free operation of the subsystem both with the same and with different probabilities of failure-free operation of elements and switches under the conditions of independence of the occurring failures. It also makes it easier to take into account technical limitations imposed by design and technical documentation on the feasibility of individual reservation units. The method is applicable to all kinds of problems that can be reduced to optimization on both oriented and non-oriented graphs with assigned value and with the complexity not exceeding a third-order polynomial.

REFERENCES

Boykov, I.R., Kitaev, S.V. & Smorodova, O.V. (2018). A set of indicators for assessing the reliability of gas pumping units. *Reliab.*, **18**: 16-21.

Chattamvelli, R. & Shanmugam, R. (2019). Generating functions in engineering and the applied sciences. *Synthesis Lect. Eng.*, **13**: 1-111.

Epifanov, A.D. (1975). *Reliability of Control Systems*. Mashinostroyeniye, Moscow.

Gliniak, M., Oziembłowski, M., Dróżdż, M., Dróżdż, T., Nawara, P., Kiełbasa, P. & Ostafin, M. (2018). The possibilities of automation of the manual line for dismantling waste electrical and electronic equipment. *Przeglad Elektrotechniczny*, **94(6)**: 136-139.

Grishin, V.M. & Ko, P.M. (2009). Optimization of the reliability of one class of components of aircraft control systems with active loaded redundancy. *Bull. Mosc. Aviat. Instit.*, **1**: 116-123.

Grishin, V.M. & Trong Tuan, V. (2018). Specific features of research and development of the passive redundant subsystems of the aircraft with due consideration of tolerances. *J. Mechan. Engin. Res.* Devel., **41**: 43-48.

Res. Devel., **41**: 43-48.
Kozlov, B.A. & Ushakov, I.A. (1975). *Handbook for the Calculation of the Reliability of Electronic Equipment and Automation*. Sovetskoye Radio, Moscow.
Lugovtsova, N.Yu. (2015). *Reliability of Technical Systems and Technological Risk*. Mediasfera, Moscow.
Mahajan, A. & Tatikonda, S. (2015). An algorithmic approach to identify irrelevant information in sequential teams. *Autom.*, **61**: 178-191.
Polovko, A.M. (1964). *Fundamentals of Reliability Theory*. Nauka, Moscow.
Rabinskiy, L.N. & Vakhneev, S.N. (2018). Study of the features of gas flow behind the vortex burner of the combustion chamber of an aircraft gas turbine engine. *Period. Tche Quim*., **15**: 359-367.
Redko, P.G. (2002). *Increased Uptime and Improved Performance of Electro-Hydraulic Servo Drives*. Yanus-K, Moscow.
Reinhold, E., Nivergelt, J. & Deo, N. (1980). *Combinatorial Algorithms Theory and Practice*. Mir, Moscow.
Soltani, H. & Shafiei, S. (2011). Heat exchanger networks retrofit with considering pressure drop by coupling genetic algorithm with LP (linear programming) and ILP (integer linear programming) methods. *Energy*, **36**: 2381-2391.
Tolstov, A.S. & Pantyukhov, D.V. (2016). An approach to ensuring the reliability of complex systems based on parametric optimization of reliability schemes. *Reliab.*, **16**: 26-30.
Tretyakov, A.M. (2016). *Fundamentals of Reliability Theory*. Publishing house of Altai State Technical University, Biysk.
Unaspekov, B.A., Zhumadilova, Z.O., Auelbekov, S.S., Taubaldieva, A.S. & Aldabergenova, G.B. (2019). Investigation of heat distribution processes on the inner surface of the enclosing structure, taking into account the movement of air in the boundary area between the device and fencing. *Period. Tche Quim*., **16**: 345-361.

MAPPING CRIME HOTSPOTS USING KERNEL DENSITY ESTIMATION (KDE) FOR DEFENSIBLE SPACE

Hasranizam Hashim[1,2*], Wan Mohd Naim Wan Mohd[1], Eran Sadek Said Md Sadek[1], Sahabudin Abd Manan[2] & Mohd Kamal Kordi[2]

[1]Centre of Studies for Surveying Science & Geomatics, Faculty of Architecture, Planning & Surveying, Universiti Teknologi MARA (UiTM), Malaysia
[2]Logistic & Technology Department, Royal Malaysia Police (RMP), Malaysia

[*]Email: 2014994239@isiswa.uitm.edu.my

ABSTRACT

This paper aims to map crime hotspots, and to determine high and low crime areas in Petaling Jaya, Selangor. The data is analysed using kernel density estimation (KDE) in the ArcGIS 10.5 spatial analysis tool software to identify hotspot areas and buffer analysis for hotspot streets, which is displayed in raster format. The study uses index crimes, which has 26,739 points (x, y coordinates) from 2010 to 2016. This study involves ten boundaries of police stations in Petaling Jaya. The results indicate that there is a clustering of crime hotspots found in seven areas in the study area, namely Kelana Jaya, Sg. Way, Sea Park, Kota Damansara, Seksyen 17, Damansara and the city centre of Petaling Jaya. The crime hot streets are found to be of very high priority are in seven roads; 23 roads are of high priority; 83 roads are medium; low priority in 694 roads; and 13,902 roads have very low crime. The outcome of this paper allows law enforcement agencies and policymakers to consider the hotspot areas and streets in the implementation of strategic planning for reducing crime.

Keywords: *Geographical information systems* (*GIS); crime hotspot; crime mapping; kernel density estimation (KDE); defensible space.*

1. INTRODUCTION

Crime hotspots typically represent the identified areas where crime occurs most frequently, and geographical information system (GIS) has become widely used to support policing and crime prevention with spatial location information (United Nation, 2010). Chainey (2013) explained that hotspots could exist at the city level where crime is highest, at the local residential housing level, or specific streets and clusters of buildings where crime is highly concentrated. The GTP Annual Report (2010) stated that 50 crime hotspot areas were identified as areas of high crime frequency.

A hotspot, as defined by Chainey and Ratcliffe (2005), is a geographical area of concentration higher than average crime with cluster incidents location. Identifying crime hotspots is the first step in spatial analysis when conducting crime mapping. The traditional method of displaying crime hotspot is placing pins for crime locations onto a wallpaper map board. Crime mapping studies have been going on for almost 190 years, and it is based on the earliest study about mapping crime by André-Michel Guerry in early 19th century in France (Eck &Weisburd, 1995; Paulsen & Robinson, 2004; Chainey & Ratcliffe, 2005; Weisburd *et al.,* 2012; Wortley & Mazerolle, 2017). An analysis of crime hotspots with mapping provides a distribution of geographic patterns of crime to identify problem areas. Jefferis (2009) found that the most common method for displaying geographic patterns of crime is by point mapping. The mapping of crime hotspots has become a common practice among police forces across the world (United Nation, 2010; Gorr & Kurland, 2012; Chainey, 2013; Gill *et al.,* 2015; Leigh *et al.,* 2016; Kalinic & Krisp, 2018).

Mapping crime hotspots aims to identify spatial clusters for proper strategic policing policy. The most common method of policing policy depends on the pattern of crime. Cluster crime patterns enable policing policies to focus on locations for defensible preventing crime. (Chainey & Ratcliffe, 2005; Braga, 2007; United Nation, 2010; Ratcliffe *et al.,* 2011; Gorr & Kurland, 2012; Chainey, 2013).

In Malaysia, high crime areas (hotspots) are in states with significant urban hierarchy status and big cities, such as Selangor, Kuala Lumpur, Penang and Johor, which are recognised by the Government Transformation Programme (GTP) Annual Reports. High-risk urban areas had experienced a significant reduction in crime from 486 crimes that occurred daily in 2010 to 272 crimes each day in 2017 when the government launched the Transformation Plan 2010-2020 to reduce the national crime rate (GTP Annual Report, 2017). The omnipresent patrol initiative programme is one of the policing policies introduced by the government in 2011. A total of 21,624 additional police personnel have been patrolling in 50 crime hotspots to combat crime.

The standard spatial analysis methods and techniques to produce hotspot maps of crime include point density mapping, spatial ellipses, thematic mapping, kernel density estimation (KDE) and G_i^* statistics. However, KDE is the most popular of these standard hotspot mapping techniques, and widely used among crime analysts who use GIS to analyse and display geographic concentrations of crime events (Chainey & Ratcliffe, 2005; Gorr & Kurland, 2012; Chainey, 2013; Boba, 2016; Kalinic & Krisp, 2018). Kernel estimation aims to estimate how the density of events changes spatially based on any given crime point patterns. It creates a continuous surface of density values whereby any location reflects the concentration of points in the area surrounding that location. One of the main advantages of kernel estimation is that it can quickly turn an intricate crime pattern of points into a smooth and understandable image. According to Chainey & Ratcliffe (2005), there is no universal standard of numerical number in an area to be the crime hotspot.

The objective of this study is to generate crime street hotspot maps for Petaling Jaya, with the aim of mapping crime hotspots, and determining high and low crime areas and streets using KDE. This study provides spatial information for law enforcement on crime hotspots for creating defensible spaces and safe cities.

2. STUDY AREA, DATASETS AND BACKGROUND STUDY

This study focuses on Petaling Jaya, Selangor, as this area is considered as one of the hottest spots in Selangor. The Petaling Jaya district is covered by ten police station boundaries as shown in Figure 1. Dataset of the crime data consists set of *x*, *y* coordinates using the WGS 84 Web Mercator projection, and was obtained from the police district departments and the iSelamat.my web portal for the period of 2010 to 2016.

3. METHODOLOGY

KDE and buffer analysis are used to understand the crime problem. The methodology adopted in this study is based on the crime pattern theory proposed by Brantingham & Brantingham (1984), and KDE technique proposed by Chainey *et al*. (2002) and Eck *et al*. (2005). The GIS software used is ArcGIS 10.5 with the kernel density tool extension.

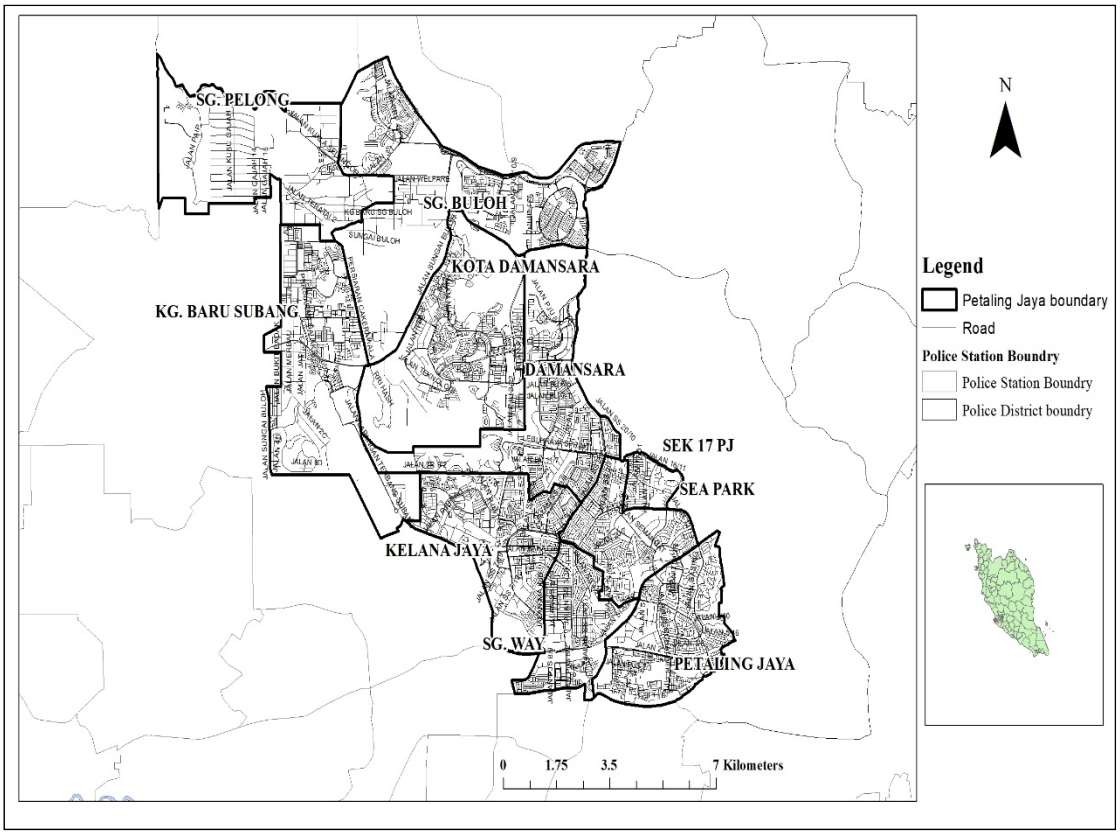


Figure 1: The police station boundaries for the study area.

3.1 Kernel Density Estimation (KDE) technique

There are four stages in the process of creating hotspot maps (Chainey *et al*., 2002; Eck *et al*., 2005) as shown in Figure 2, which are geocoding crime incidents, creating the grid, defining the bandwidth (search radius), and assigning values to squares. Most GIS software, such as ArcGIS, MapInfo and CrimeStat, provide quartic KDE, which requires two important parameters - the grid cell and bandwidth. The lead parameter is bandwidth. Therefore, bandwidth selection will determine the density or volume of crime pattern distribution with a smooth continuous surface to represent hotspots across the study area.

Stage 1 is the processing of the geocoded or *x, y* coordinate crime incidents before the hotspot exists. In Stage 2, a grid is created to overlay the area of the crime points. The size of the grid is referred to as the output of grid size. This grid size is measured in metres and the output cell size is set for one size: 50 m, which represents a square of 50 m x 50 m on the ground. This grid size will determine how detailed the resulting map will be and how much processing time it will take. If the cell set for the grid size is too large, the result will be a map of blocks. If the cell set for the grid size is too small, the map will have high resolution but will take a much longer time to process. Both grids will technically be correct but with different data resolutions.

Stage 3 is defining bandwidth of the grid. The process looks at each grid cell, in turn, counting all the crimes that have occurred within this bandwidth distance. This study provides the search radius for two distances: i) 100 m and ii) 500 m. The points closer to the centre of the cell have a higher influence value than those further away. These values accumulate to provide the density value for that cell. Stage 4 involves processing each grid square, then assigning a value relating to the number of points in or near it. When these density values are applied to each cell, they provide a continuous surface density map, and these values will then form the density or hotspot map. Therefore, the classification is done using standard deviations $n = 2.5$.

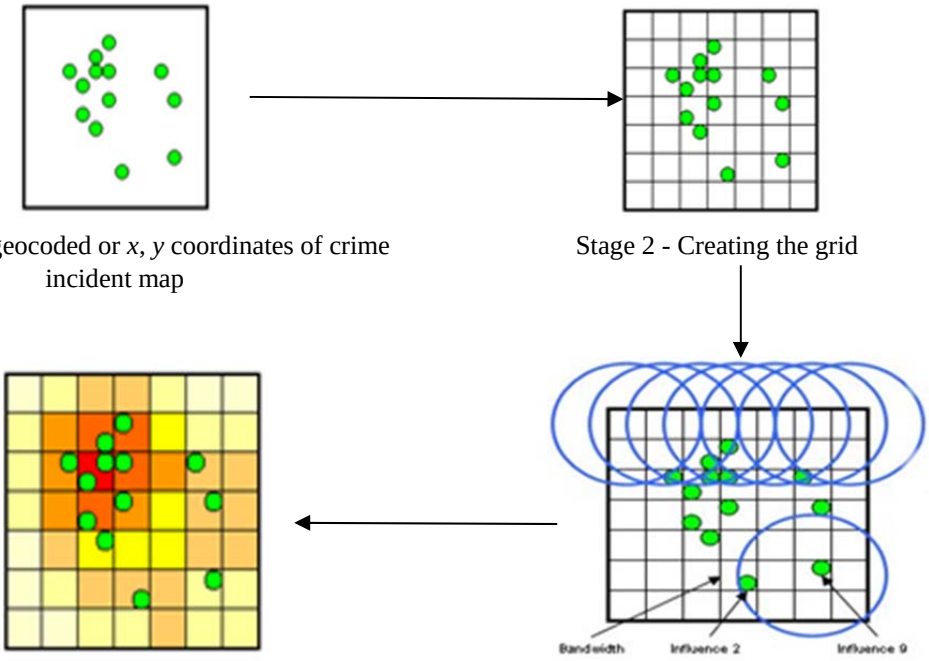


Figure 2: Process of the KDE technique.

The kernel density for the points calculated is computed using the ArcGIS kernel density tool (Esri, 2018), which uses a quartic kernel (Silverman, 1986) that is expressed as follows:

$$Density = \frac{1}{(radius)^2}\sum_{i=1}^{n}\left[\frac{3}{\pi}\cdot pop_i\left(1-\left(\frac{dist_i}{radius}\right)^2\right)^2\right] \quad (1)$$

For $dist_i < radius$

where:

$i = 1\ldots, n$ are the input points. Points are only included in the sum if they are within the radius distance of the x, y location.

pop_i is the population field value of point i, which is an optional parameter.

$dist_i$ is the distance between point i and the x, y location.

The calculated density is then multiplied by the number of points where the output raster is applied to the centre of every cell.

3.2 Workflow and Buffer Tool Technique for Street Level Density Hotspot

An analysis with buffer has two benefits. First, the streets are lines and buffers with polygons created around them. When the crime points are joined to the polygons, the number of points within the polygon is summed. The joint buffer allows ranked crime statistics to be displayed on each street level (Gill *et al.*, 2015). The crime events are then aggregated to street level, and crime events on the street level can be further aggregated and placed into defined classes. Weisburd *et al*. (2012) and Curman *et al*. (2015) found that crime events on the street level can be labelled in various permutations of low,

medium, and high crimes as well as stable, increasing and decreasing in the analysis segment density for hotspot maps. The workflow process used in this study that combines buffer analysis with KDE analysis is shown in Figure 3.

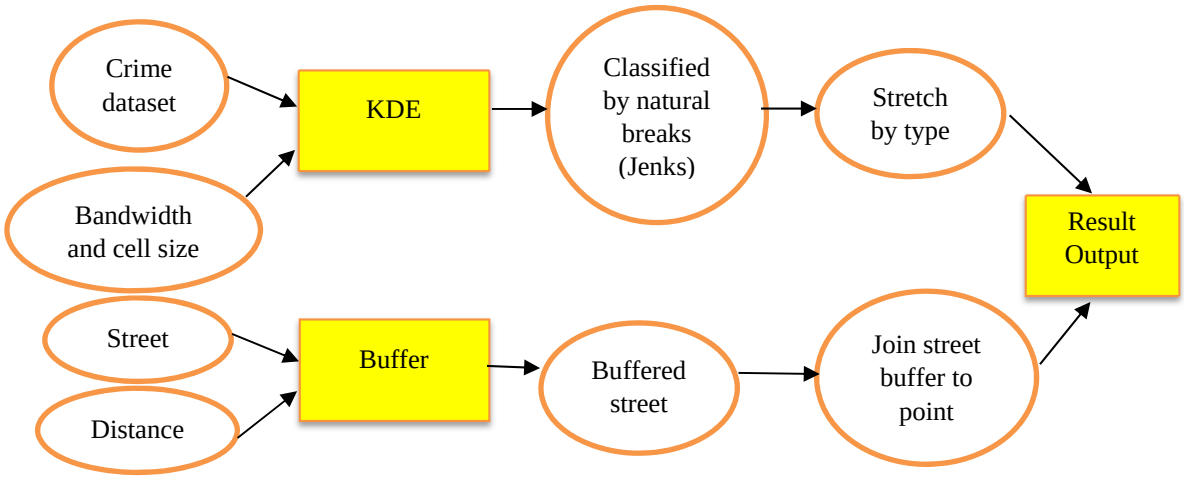


Figure 3: Workflow of KDE and buffer analysis technique.

3.3 Crime Data

The crime data used is between 2010 to 2016, with 26,739 points (*x*, *y* coordinates) obtained from the district police stations. The type of crime data for the analysis consists of ten types of crime index, whereby seven of them are property crime, namely theft, snatch theft, motorcycle theft, car theft, van / lorry / heavy machine theft, night-time house break-in and daytime house break-in; while the other three are violent crimes, which are gang robbery, robbery and assault. KDE is used to show the macro-level results of hotspot crime locations over the study area, while buffer analysis is used to show the micro-unit results of hotspot crime street locations. The study of macro and micro-levels will determine high and low crime densities in the study area.

4. RESULTS AND DISCUSSION

4.1 Crime Distribution by Classification of Police Station Boundary

As shown in Figure 4, from five classifications determined using the natural breaks (Jenks) tool use in ArcGIS, the crime distribution for the ten police stations in the study area shows that Sg. Way has a very high crime distribution (total count ranges from 4,694 to 6,095 crime incidents). Kelana Jaya and the city centre of Petaling Jaya have high crime rates (total count ranges from 2,943 to 4,694 crime incidents). Kota Damansara has a medium crime rate (total count ranges from 2,649 to 2,943), while low crime areas are Damansara, Sek.17, and Sea Park (total count range from 730 to 2,649 crime incidents). The very low crime areas are Kg. Baru Subang, Sg. Buloh and Sg. Pelong. The results show that the high crime areas are close to the city centre of Petaling Jaya, which is a developed urban area. The very low and low crime areas are within the villages and intra-urban areas with moderate urbanisation. Kota Damansara, which is classified as a medium crime area, is also moderately urbanised. Damansara, Sek.17 and Sea Park have the characteristics of an old urbanisation development in Petaling Jaya. The results of the total crime show that the indication of high and low crime is affected by the urbanisation of the areas.

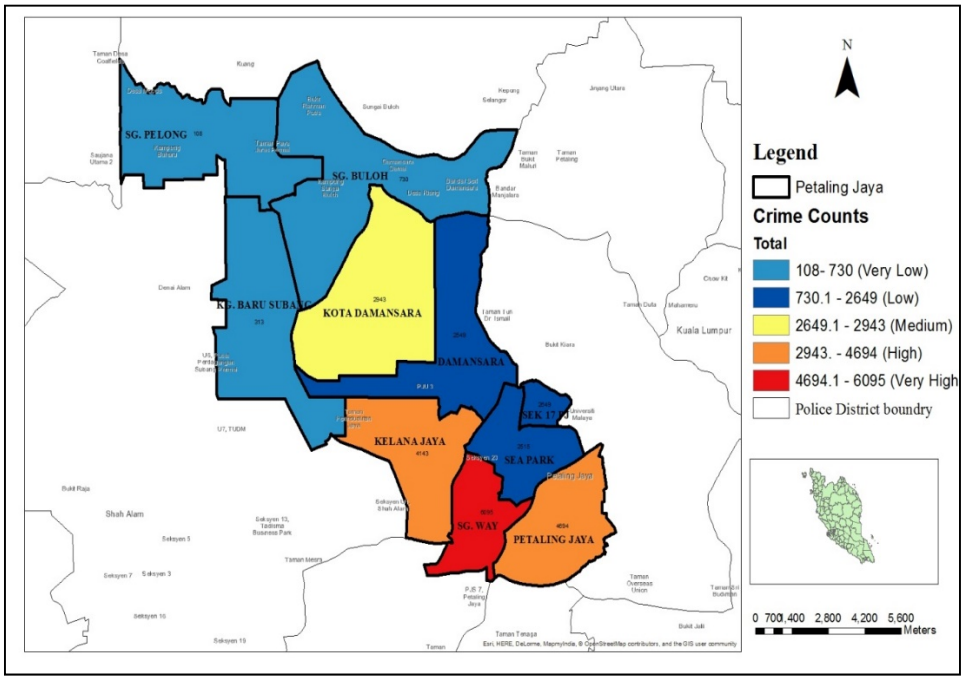


Figure 4: High and low crime areas in Petaling Jaya according to police station boundaries (2010 to 2016).

4.2 Crime Pattern Hotspots Generated Using KDE

As discussed in Section 3.1, there are two hotspot maps produced to visualise the data in this study. The first search radius bandwidth is 500 m with 50 m grid cell size, while the second is 100 m with 50 m grid cell size. As the search radius is increased from 100 to 500 m, as shown in Figure 5, the hotspots become more generalised. The kernel density function is best used to locate high crime areas and provide a broad idea of crime pattern areas based on raster format. A 50-m raster size is selected because this distance covers approximately half a block, and this raster size shows variation at the block level as described by Spicer *et al.* (2016). KDE analysis for bandwidth of 500 m shows that there are seven high crime areas. Kota Damansara, Kelana Jaya, Sg Way, Sea Park, Sek. 17, Damansara and the city centre of Petaling Jaya are the highest crime areas, while Sg Pelong, Sg Buloh and Kg Baru Subang are low crime areas based on the raster format. The second map for of 100 m with 50-meter grid cell size is produced to show the different densities and is more specific, as shown in Figure 6.

4.3 Crime Pattern Within Street Buffer for High and Low Crime Hotspot Map

The street-level study was conducted to determine the streets with the hottest crime spots. From Table 1 and Figure 7, the street-level map shows that areas in Sg Way, the city centre of Petaling Jaya and Kelana Jaya with Jalan 51A/224, Jalan PJS 8/9, Jalan PJS 2C/2, Jalan PJS 2/1, Jalan PJS 20/19, Jalan SS 7/26 and Lorong SS 7/19 have high crime concentrations in the study area. The study found that 23 streets are categorised into high priority crime, 83 streets in medium crime, and low priority crime in 694 streets, while 13,902 streets in very low crime.

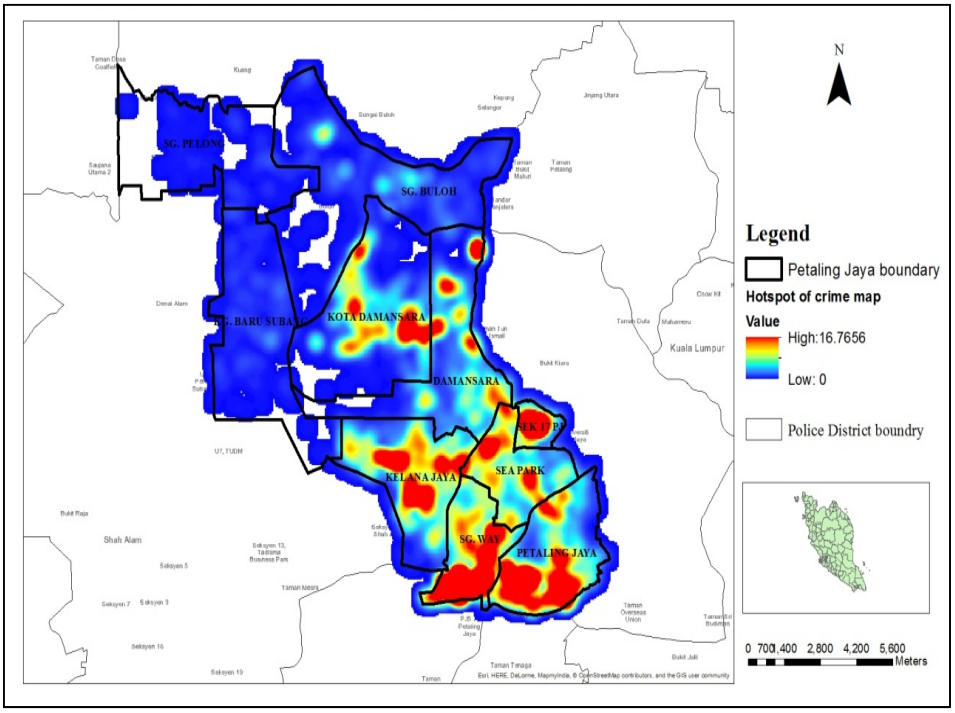


Figure 5: High and low crime areas map generated using KDE (Bandwidth: 500m and cell size: 50m).

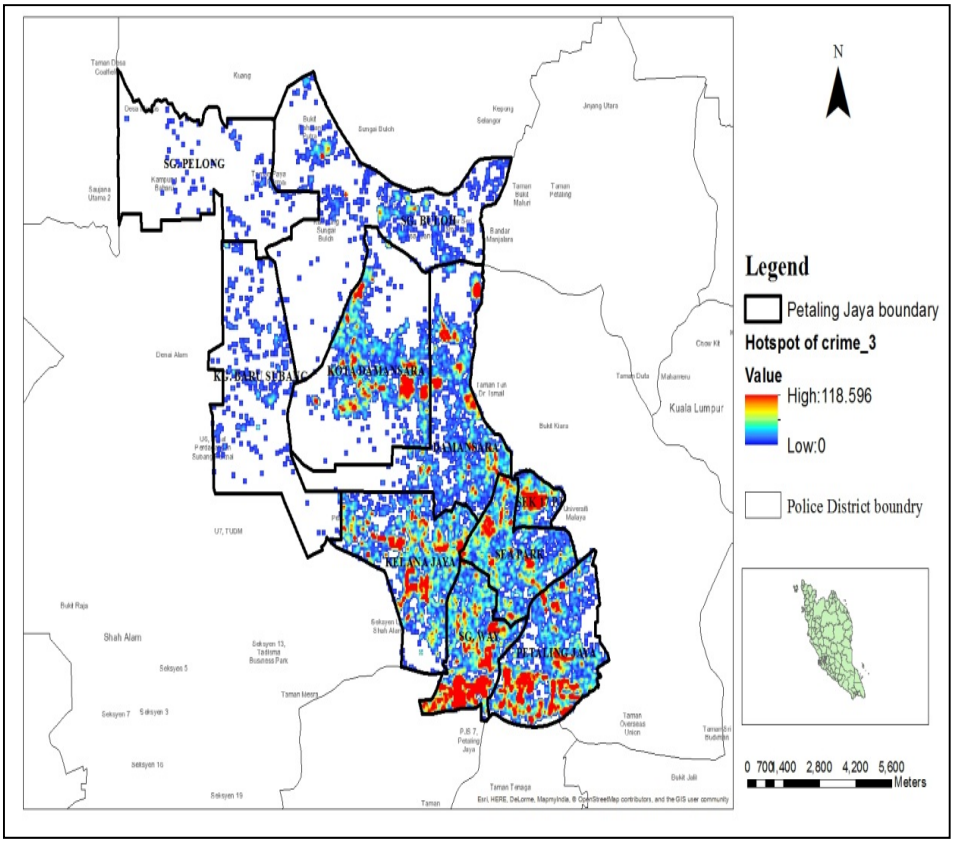
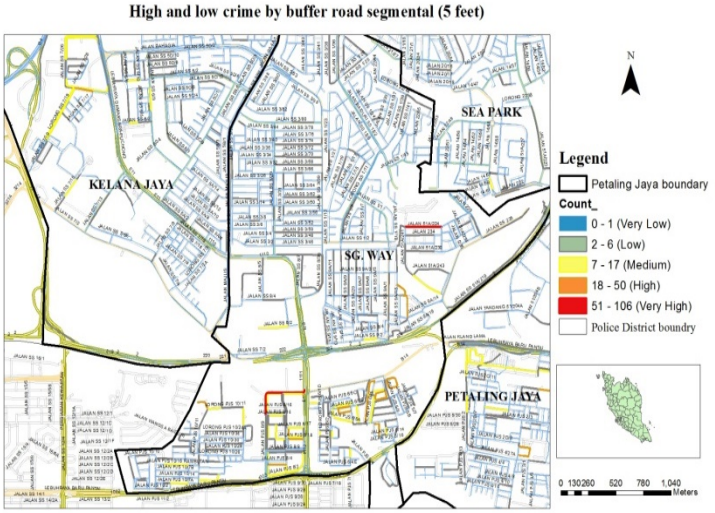
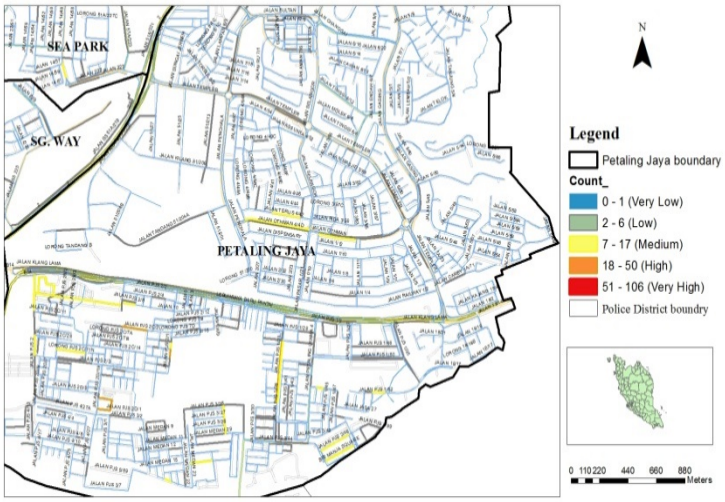
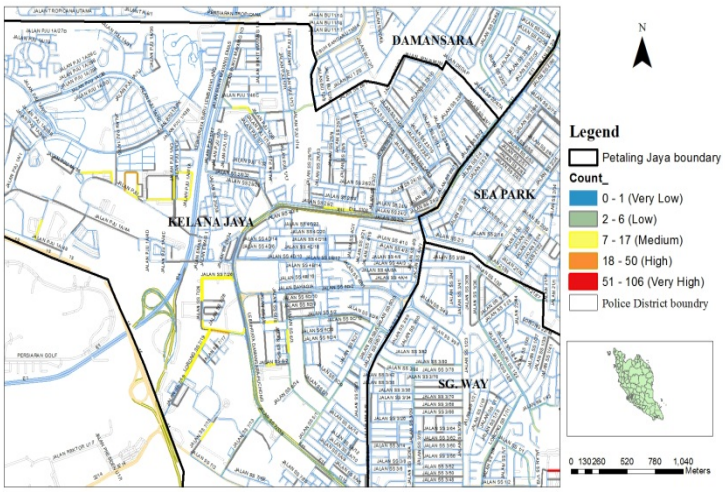


Figure 6: High and low crime map area by KDE analysis (Bandwidth: 100m and cell size: 50m)

Table 1: Crime hotspots on the street (2010-2016).

Rank	Areas	Buffer zone map	Road name
1	Sg. Way		1. Jalan 51A/224 2. Jalan PJS 8/9
2	City centre of Petaling Jaya		1. Jalan PJS 2C/2 2. Jalan PJS 2/1 3. Jalan PJS 20/19
3	Kelana Jaya		1. Jalan SS 7/26 2. Lorong SS 7/19

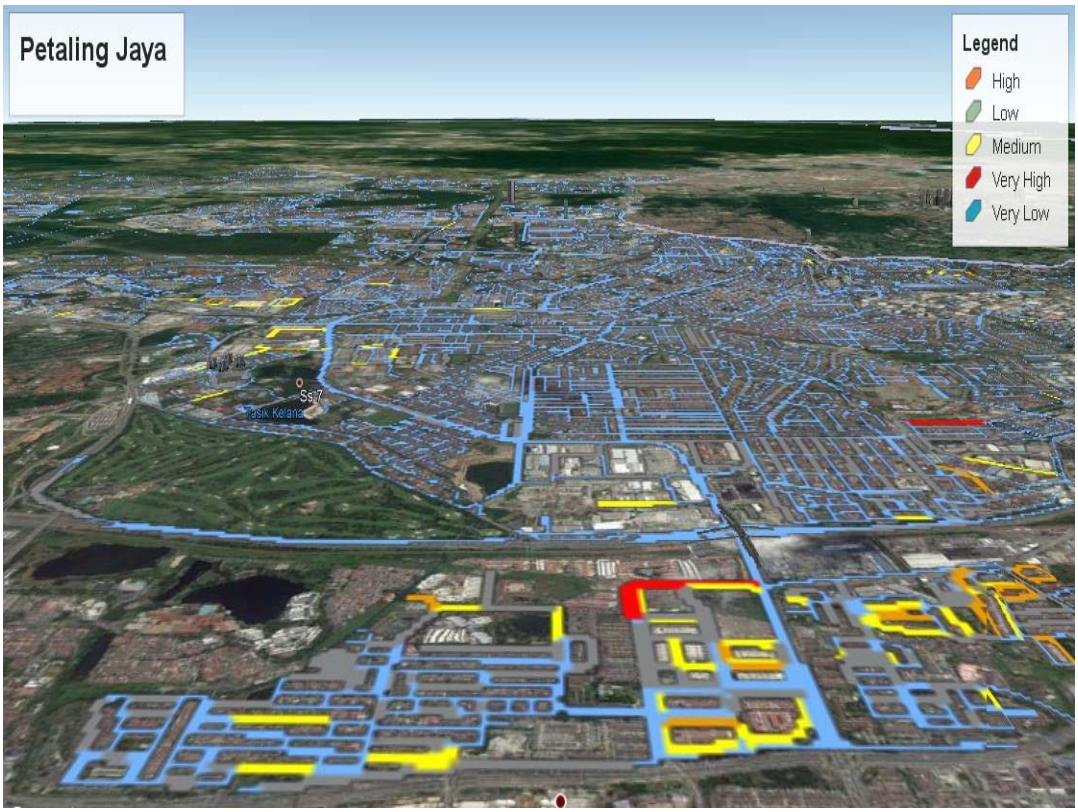


Figure 7: Crime hotspots in 3D view.

4.4 Discussion

Based on Figures 4 -7 the results obtained vary with different types of analysis. By using the natural breaks (Jenks) method, the map in Figure 4 shows that the areas with very high crime are only Kelana Jaya, Sg Way and the city centre of Petaling Jaya. By using KDE analysis, as shown in Figures 5 and 6, when the search radius bandwidth increases, the hotspots becomes more generalised. By using buffer tool analysis, as shown in Table 1, the hotspots on the streets have been identified in specific locations, namely Jalan 51A/224 and Jalan PJS 8/9 in Sg Way; Jalan PJS 2C/2, Jalan PJS 2/1 and Jalan PJS 20/19 in the city centre of Petaling Jaya; and Jalan SS 7/26 and Lorong SS 7/19 in Kelana Jaya. The analysis also shows Kelana Jaya, Sg. Way and the city centre of Petaling Jaya as the high crime areas in the street level. From the results obtained, the study found that crime hotspots depend on urbanisation density and business areas.

5. CONCLUSION

This study has identified seven areas in Petaling Jaya as crime hotspots. By using classifications with the natural breaks (Jenks) method, KDE, and buffer analysis, the study found that Kelana Jaya, Sg. Way, Sea Park, Kota Damansara, Seksyen 17, Damansara and the city centre of Petaling Jaya as crime hotspots. The analysis of crime on the streets level found that Jalan 51A/224 and Jalan PJS 8/9 in Sg Way; Jalan PJS 2C/2, Jalan PJS 2/1 and Jalan PJS 20/19 in the city centre of Petaling Jaya; and Jalan SS 7/26 and Lorong SS 7/19 in Kelana Jaya as hotspots street. It can be concluded that mapping crime hotspots in neighbourhood areas with street-level analysis could identify the specific locations for defensible crime prevention.

REFERENCES

Braga A. A. (2007). Effects of hot spots policing on crime. *Campbell Systemat. Rev.* **3**: 1-36.

Brantingham, P.J. & Brantingham, P.L. (1984). *Patterns in Crime*. Macmillan Press, New York.

Boba, R.S. (2016). *Crime Analysis with Crime Mapping, 4th Ed.* Sage Publications, Inc, Thousand Oaks, California.

Chainey, S.P. Reid, S. & Stuart, N. (2002). When is a hotspot a hotspot? A procedure for creating statistically roboust hotspot maps of crime. *In* Kidner, D. Higgs, G. & White, S. (Eds.), *Innovations in GIS 9*. Taylor & Francis Press, London, pp. 21-36.

Chainey, S. (2013). Examining the influence of cell size and bandwidth size on kernel density estimation crime hotspot maps for predicting spatial patterns of crime. *Bull. Geogr. Soc. Liege.*, **60**: 7-19.

Chainey, S. & Ratcliffe, J. (2005). *GIS and Crime Mapping*. Mastering GIS: Technology, Applications and Management. John Wiley & Sons Press, England.

Eck, J. E. & Weisburd, D. (1995). Crime Places in Crime Theory. *In* Eck, J. E. & Weisburd, D. (Eds.). *Crime and Place, Crime Prevention Studies.* Willow Tree Press, Monsey, NY, pp. 1–33.

Eck, J. Chainey, S.P. Cameron, J. & Wilson, R. (2005). *Mapping Crime: Understanding Hotspots*. National Institute of Justice Press, Washington.

Esri (2018). *How Kernel Density Works*. Available online at: https://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/how-kernel-density-works.htm (Last access date: 21 January 2019).

Gill, C. Vitter, Z. & Weisburd, D. (2015). *Identifying Hot Spots of Juvenile Offending: A Guide for Crime Analysts.* Available online at: https://ric-zai-inc.com/Publications/cops-p298-pub.pdf (Last access date: 10 April 2019).

GTP Annual Report (2010). *Reducing Crime Chapter*. Available online at: http://ntp.epu.gov.my/images/ntp/ pastreports/ 2010/ GTP_2010_ENG.pdf (Last access date: 2 Mac 2018).

GTP Annual Report (2011). *Reducing Crime Chapter*. Available online at: http://ntp.epu.gov.my/images/ntp/ pastreports/ 2011/ GTP_2011_ENG.pdf (Last access date: 2 Mac 2018).

GTP Annual Report (2012). *Reducing Crime Chapter*. Available online at: http://ntp.epu.gov.my/images/ntp/ pastreports/ 2012/ GTP_2012_ENG.pdf (Last access date: 2 Mac 2018).

Jefferis, E. (1999). *A Multi-Method Exploration of Crime Hot-Spots: A Summary of Findings*. National Institute of Justice Press, Washington.

Kalinic, M. & Krisp, J. (2018). *Kernel density estimation (KDE) vs. hot-spot analysis - Detecting criminal hot spots in the city of San Francisco*. *21st AGILE Conf. Geogr. Inform. Sci.*, 12-15 June 2018, Lund, Sweden.

Leigh, J.M. Dunnett, S.J. & Jackson, L.M. (2016). Predictive policing using hotspot analysis. *Int. Multi-Conf. Eng. Comput. Sci. (IMECS 2016)*, 16-18 March 2016, Hong Kong.

NTP Annual Report. (2015). *Reducing Crime Chapter*. Available online at: http://ntp.epu.gov.my/images/ntp/ pastreports/ 2015/ NTP_AR2015_ENG.pdf (Last access date: 2 Mac 2018).

NTP Annual Report (2016). *Reducing Crime Chapter*. Available online at: http://ntp.epu.gov.my/images/ntp/ pastreports/ 2016/ NTP_AR2016_ENG.pdf (Last access date: 2 Mac 2018).

NTP Annual Report (2017). *Reducing Crime Chapter*. Available online at: http://ntp.epu.gov.my/images/ntp/NKRA/ pengurangan_jenayah.pdf (Last access date: 2 Mac 2018).

Paulsen, D. J. & Robinson, M. B. (2004). *Spatial Aspects of Crime*: *Theory and Practice*. Pearson Education Inc. Press, USA.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall Press, London.

United Nations (2017). *World Crime Trends and Emerging Issues: Responses in The Field of Crime Prevention and Criminal Justice*. UN Press, Vienna.

United Nation (2010). *Handbook on the United Nations Crime Prevention Guidelines - Making Them Work. A Guide for Implementing the UN Crime Prevention Guidelines*. UN Press, Vienna.

Wortley, R. & Mazerolle, L. (2017). *Environmental Criminology and Crime Analysis: Situating the Theory, Analytic Approach and Application*. Routledge Press, New York.

Weisburd, D. L. Groff, E.R. & Yang, S.M. (2012). *The Criminology of Place: Street Segments and Our Understanding of The Crime Problem*. Oxford University Press, New York.

VALIDATION OF HAND ARM VIBRATION (HAV) MONITORING USING INTEGRATED KURTOSIS-BASED ALGORITHM FOR *Z*-NOTCH FILTER (I-KAZ) VIBRO VIA INDEPENDENT COMPONENT ANALYSIS (ICA)

Shamsul Akmar Ab Aziz[1*], Mohd Zaki Nuawi[2], Nakamura Hiroki[3] & Yamazaki Toru[3]

[1]Science and Technology Research Institute for Defence (STRIDE), Ministry of Defence, Malaysia
[2]Department of Mechanical and Material Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia (UKM), Malaysia
[3]Department of Mechanical Engineering, Kanagawa University, Japan

[*]Email: shamsulakmar.abaziz@stride.gov.my

ABSTRACT

The objective of this paper is to validate the integrated kurtosis-based algorithm for Z-notch filter (I-kaz) Vibro method for monitoring hand arm vibration (HAV) in three-tonne truck steering wheels using independent component analysis (ICA). From the recorded HAV signals for acceleration (Acc), velocity (Vel) and displacement (Dis), ICA is used to obtain the independent components, IC_1 and IC_2. Comparison between the I-kaz display and I-kaz Vibro coefficient (Z_v^{∞}) with the proposed ICA display and independent component kurtosis coefficient (K_{IC}) shows that HAV acceleration (a), Z_v^{∞} and K_{IC} values increase when vehicle speed is increased. In addition, these results are also supported by the observation of the scattering of the I-kaz Vibro and ICA displays, whereby larger space of scattering is found for increasing vehicle speed, indicating that Z_v^{∞} and K_{IC} values are comparatively increasing. Comparison of the fast Fourier transform (FFT) plots of the original HAV acceleration signal with the two independent components shows that the vibration value for each frequency in the FFT graph did not change between the recorded HAV data with each of the two independent components, and the relative error between the maximum level is considered to be within practically acceptable relative error limits. Thus, it can be concluded that the I-kaz Vibro method for HAV monitoring has been validated accurately using ICA.

Keywords: *Integrated kurtosis-based algorithm for Z-notch filter (I-kaz) Vibro; independent component analysis (ICA); three-tonne truck steering wheel; hand arm vibration (HAV); fast Fourier transform (FFT) plots.*

Nomenclature

a	HAV acceleration
K_{IC}	Independent component kurtosis coefficient
x_i^A	Data for the i^{th}-sample of time for acceleration value
x_i^D	Data for the i^{th}-sample of time for displacement value
x_i^V	Data for the i^{th}-sample of time for velocity value
Z_v^{∞}	I-kaz vibro coefficient
σ^2	Variance
$\sigma_A{}^2$	Variance for acceleration value
$\sigma_D{}^2$	Variance for displacement value

$\sigma_V{}^2$ Variance for velocity value
μ_A Mean for acceleration value
μ_D Mean for displacement value
μ_V Mean for velocity value

1. INTRODUCTION

The steering wheel is the part of the steering system in a vehicle that is manipulated by the driver to control the direction of the vehicle. When the engine is started, vibrations from the engine are transmitted to the steering wheel through the steering rod and can be felt by the driver's hands and arms. In Aziz *et al.* (2015, 2017), we proposed that Z_v^{∞} be used to represent the hand arm vibration (HAV) experienced by the driver from the steering wheel by calculating kurtosis and standard deviation values for each axis. Therefore, HAV levels acting on the driver's hands and arms can be predicted by determining Z_v^{∞}. It was found that HAV exposure to drivers based on vehicle speed changes affects the value of the measured vibration signal. I-kaz Vibro was used to measure the spread of data distribution by calculating the distance of each centroid signal from the raw data. The concept of data dispersion and diffusion space can be observed in the I-kaz Vibro display with higher values corresponding to a wider dispersion of data. This process produces a 3D graphic representation of time discrete continuous signals measuring the displacement, velocity and acceleration on the frequency distributions of the *x*-, *y*- and *z*-axes respectively.

The I-kaz Vibro method was developed based on the concept of data scattering about its centroid. Z_v^{∞} can be computed in terms of variance (σ^2) as follows:

$$Z_v{}^{\infty} = \sqrt{(\sigma_A{}^2)^2 + (\sigma_V{}^2)^2 + (\sigma_D{}^2)^2} \tag{1}$$

$$Z_v{}^{\infty} = \sqrt{\frac{\sum_{i=1}^{N}(x_i{}^A - \mu_A)^4}{N^2} + \frac{\sum_{i=1}^{N}(x_i{}^V - \mu_V)^4}{N^2} + \frac{\sum_{i=1}^{N}(x_i{}^D - \mu_D)^4}{N^2}} \tag{2}$$

where $\sigma_A{}^2$, $\sigma_V{}^2$ and $\sigma_D{}^2$, and x_i^A, x_i^V and x_i^D, and μ_A, μ_V and μ_D are the variance for each axis, data for the i^{th}-sample of time and means of each axis respectively (Aziz *et al.,* 2015, 2017).

The objective of this paper is to validate the I-kaz Vibro method for monitoring HAV in three-tonne truck steering wheels using independent component analysis (ICA). The verification tests are necessary to ensure that this statistical analysis method is suitable and any follow-up studies that use this method can produce reliable results.

ICA is a signal processing method that extracts statistically independent components from a set of measured signals (Wang *et al.*, 2011; Naik & Kumar, 2011; Ahmad *et al.*, 2013). In ICA, basic original signals are recovered from linearly mixed signals (Kang *et al.*, 2009). A generative model is defined for an observed multivariate data, which is typically given in a large amount of data. In the model, the data variables are assumed to be linear or nonlinear mixtures of some latent variables. The mixing system is also unknown. The latent variables are supposed to be mutually independent and non-Gaussian. They are called the independent components of the observed data. The ICA technique estimates the unmixing

matrix iteratively by using the maximisation of independence of the unmixed signals as the cost function (Hyvarinen *et al.*, 2001). The underlying processes are assumed to be independent of each other, which is realistic if they correspond to distinct physical processes. Its power resides in the physical assumptions that the different physical processes generate unrelated signals (Hyvarinen *et al.*, 2013). We choose the FastICA algorithm mainly because of its ease of implementation and speed of operation. FastICA attempts to separate underlying components from a given set of mixed measurement channels based on their 'non-Gaussianity'. Furthermore, FastICA provides a fast iterative algorithm that undertakes to find projections that maximize the non-Gaussianity of components by their kurtosis (the 4^{th} order cumulant given to a random variable) or negentropy (Wang *et al.*, 2011). The simple and generic nature of this assumption has allowed ICA to be successfully applied in a diverse range of research fields, including biomedical signal processing, audio signal separation, telecommunications, fault diagnosis, feature extraction and data mining (Hyvarinen *et al.*, 2001; Naik & Palaniswami, 2010; Wang *et al.*, 2011; Naik & Kumar, 2011; Ahmad & Ghanbari, 2011; Guerrero-Mosquera & Navia-Vázquez, 2012).

Previous studies have shown that ICA is a suitable tool for vibration analysis. For example, Razaghi *et al.* (2013) used vibration analysis and ICA to assess the density of multiple materials making up a single structure. A relationship was found between the densities of the liquids and the vibration frequency of the ICA extracted frequency components. Nakamura *et al.* (2014) used a preprocessing algorithm using ICA to improve the model construction for statistical energy analysis (SEA). Through an experiment targeted on subsystems structure and bended steel plate, the feasibility of separation of inconvenient vibration using ICA was determined. The results suggested that ICA-SEA can predict subsystem energy more accurately than normal SEA, especially for low frequencies, where normal SEA is not suitable. Furthermore, the energy of remote subsystems is well estimated with ICA-SEA. Zang *et al.* (2002) presented a novel approach for the decomposition of time domain vibration signals using ICA. ICA has the potential to reduce the size of measured response data and filter unwanted measurement noise. Therefore, the approach may be used as a tool to pre-process measured data for further vibration analysis.

The findings of these studies indicate that ICA is an appropriate method to validate the I-kaz Vibro method for HAV monitoring. I-kaz Vibro consists of three types of time discrete continuous HAV signals, measuring the displacement, velocity and acceleration of the *x*-, *y*- and *z*-axes respectively. ICA will statistically extract the independent components of the three HAV signals.

2. METHODOLOGY

Based on the recommendation of ISO 5349 (Mechanical Vibration - Measurement and Evaluation of Human Exposure to Hand-Transmitted Vibration), the most important quantity used to describe the magnitude of the HAV transmitted by the steering wheel to the driver's hands is root mean square (RMS) frequency-weighted acceleration. Dynamic movement, such as vibration, as in the case of this study, can be studied using various parameters, such as acceleration, velocity and distance.

In Aziz *et al.* (2015, 2017), HAV acceleration (*Acc*) was measured in the tangential direction using a single axis piezoelectric accelerometer on the top left side of the steering wheel. The acceleration time history raw data was sampled at 800 samples/s for 10 s segments, and integrated to produce velocity and distance time history data. The HAV signal recorded at speed of 80 kmh^{-1} was used for sampling in this analysis. The recorded signals shown in Figure 1 look as if they are completely noisy. However, there are actually some quite structured underlying source signals hidden in the observed signals (Hyvarinen *et al.*, 2013). The objective here is to recover the source signals, consisting of the independent components, from the recorded signals.

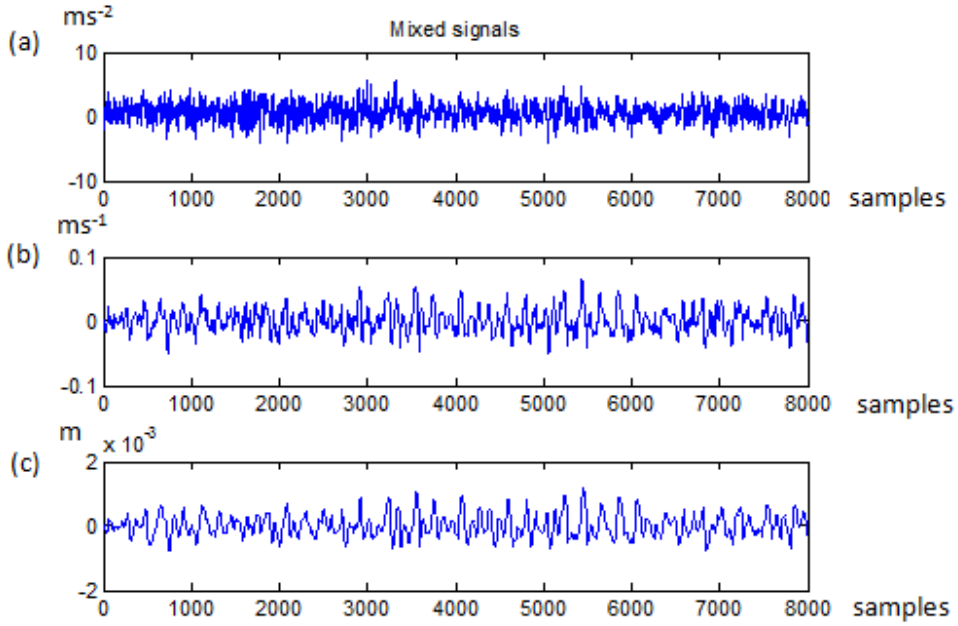


Figure 1: The recorded HAV signals - (a) acceleration (b) velocity (c) displacement - in the tangential direction of the steering wheel while truck is moving at speed of 80kmh^{-1}.

Figure 2 shows both the mixing and unmixing processes involved in ICA. The source signals are mixed by the mixing matrix (*A*). If the estimate of the unmixing matrix (*W*) is accurate, a good approximation of the sources can be obtained. The ICA model shown is a simple model as it ignores all noise components and any time delays in the recorded signals (Naik & Kumar, 2011).

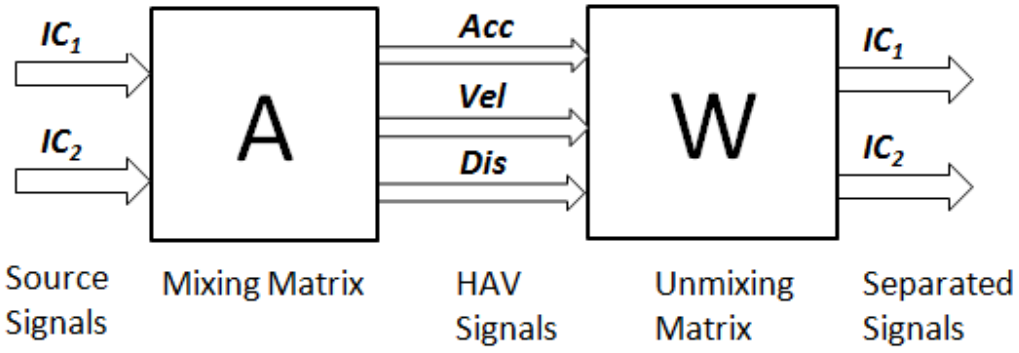


Figure 2: The ICA block diagram for this study. *IC$_1$* and *IC$_2$* are the source signals; *Acc, Vel* and *Dis* are the recorded HAV signals; *A* is the mixing matrix; and *W* is unmixing matrix.

The general model of ICA is as follows:

$$X = AS \tag{3}$$

$$\begin{bmatrix} Acc \\ Vel \\ Dis \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} IC_1 \\ IC_2 \end{bmatrix} \tag{4}$$

where X and S are vectors representing the recorded and source signals respectively; and a_{ij} are the mixing weights, which are unknown as all the properties of the physical mixing system are also unknown. The source signals are unknown as well, since the very problem is that they cannot be recorded directly (Hyvarinen *et al.*, 2001). For the matrix above, each component for *Acc*, *Vel* and *Dis* can be separated using the following equations:

$$Acc = a_{11}IC_1 + a_{12}IC_2 \tag{5}$$
$$Vel = a_{21}IC_1 + a_{22}IC_2 \tag{6}$$
$$Dis = a_{31}IC_1 + a_{32}IC_2 \tag{7}$$

In the ideal condition, W is the inverse of A:

$$W = A^{-1} \tag{8}$$

Thus, S can be computed as follows:

$$S = WX \tag{9}$$

$$\begin{bmatrix} IC_1 \\ IC_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} Acc \\ Vel \\ Dis \end{bmatrix} \tag{10}$$

$$IC_1 = w_{11}Acc + w_{12}Vel + w_{13}Dis \tag{11}$$

$$IC_2 = w_{21}Acc + w_{22}Vel + w_{23}Dis \tag{12}$$

where a_{ij} are the unmixing weights.

Using the above equations, an arbitrarily chosen mixing matrix A, as follows, was used to compute the independent components IC_1 and IC_2 (Figure 3):

$$A = \begin{bmatrix} -0.8823 & 0.9824 \\ 0.0061 & -0.017 \\ 1.63 \times 10^{-4} & -1.8714 \times 10^{-4} \end{bmatrix} \tag{13}$$

Similar to I-kaz Vibro, an ICA display method was developed based on the concept of data scattering about its centroid. This method provides a 2D graphical representation of the frequency distribution and we developed a new parameter to represent the ICA display, known as independent component kurtosis coefficient (K_{IC}), which calculates the distance of each data point from signal centroid. K_{IC} can be written in terms of kurtosis (K) as follows:

$$K_{IC} = (K_{IC1} + K_{IC2})/2 \tag{14}$$

where K_{IC1} and K_{IC2} are the kurtosis of IC_1 and IC_2 respectively.

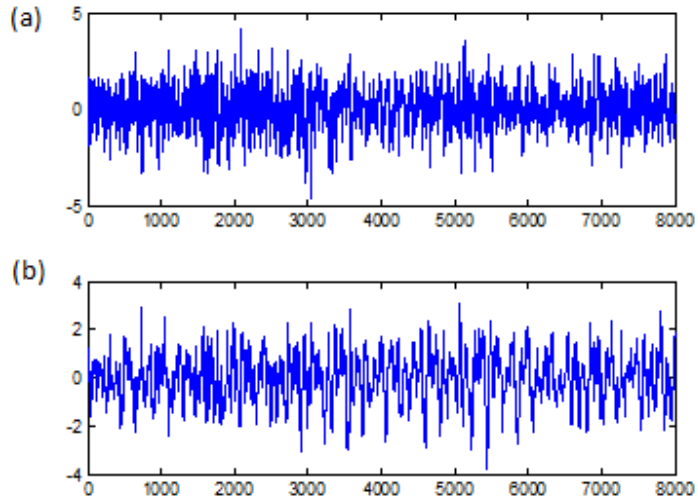


Figure 3: Independent components of the HAV signals - (a) IC_1 (b) IC_2- while the truck is moving at speed of 80kmh^{-1}.

3. RESULTS AND DISCUSSION

The HAV acceleration (a), $\not{Z}_v^{\infty}$ and K_{IC} values for various vehicle speeds are tabulated in Table 1. It is observed that a increases with increasing vehicle speed, indicating that the truck experienced higher vibration. This resulted in a higher amount of vibration being transmitted through the steering wheel and felt by the driver. The highest value of a (3.41 ms^{-2}) was obtained at maximum speed of 80 kmh^{-1}. This was followed by medium speed (40 kmh^{-1}), with a value of 1.56 ms^{-2}. The idle condition (0 km/h) had the lowest a value of 0.62 ms^{-2}.

Table 1: Values of a, $\not{Z}_v^{\infty}$ and K_{IC} values for the various vehicle speeds.

Truck Speed (kmh^{-1})	a (ms^{-2})	$\not{Z}_v^{\infty}$	K_{IC}
Idle (0)	0.62	6.09×10^{-5}	2.31
Medium speed (40)	1.56	8.45×10^{-5}	2.92
Highest speed (80)	3.41	3.82×10^{-4}	3.22

Comparisons were made between the I-kaz Vibro (Figure 4) and ICA (Figure 5) displays, and the corresponding $\not{Z}_v^{\infty}$ and K_{IC} values for the varying vehicle speeds. It is found that increasing values of a with increasing vehicle speed resulted in increasing dispersions, and values of $\not{Z}_v^{\infty}$ and K_{IC}. The highest values of $\not{Z}_v^{\infty}$ and K_{IC} are 3.82×10^{-4} and 3.22 respectively, which was obtained at a speed of 80 kmh^{-1}. This is followed by speed of 40 kmh^{-1} with $\not{Z}_v^{\infty}$ and K_{IC} values of 8.45×10^{-5} and 2.92 respectively. The idle condition gave the lowest $\not{Z}_v^{\infty}$ and K_{IC} values of 6.09×10^{-5} and 2.31 respectively.

The different dispersions based on the different vehicle speeds show the levels of HAV vibration experienced by driver, with larger dispersions indicating higher levels of HAV and vice-versa. The data for higher a values distributed further from the mean and show higher scattering. On the other hand, the data for lower a values distributed closer to the mean. These graphical representations show that larger

space of scatterings indicate that the values of $\not{Z}_v^{\infty}$ and K_{IC} values are increasing. The reason for this is that the I-kaz Vibro and ICA displays were developed based on standard deviation and kurtosis, which are used for the detection of changing of HAV because of their sensitivity to changing of amplitude. I-kaz Vibro and ICA displays for the highest speed (80 kmh^{-1}) distribute far off from the mean and show the largest scattering of data distribution. On the other hand, the I-kaz Vibro and ICA displays for idle condition distribute closely to the mean, with low scattering of data distribution. This approach provides better understanding about the data analysis through the visualisation of 3D display.

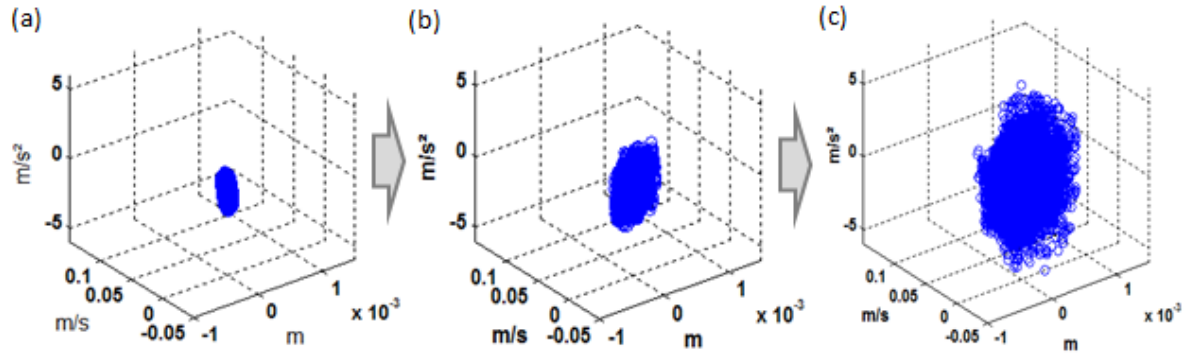


Figure 4: I-kaz Vibro displays for the different vehicle speeds: (a) Idle, $\not{Z}_v^{\infty}$= 6.09 × 10^{-5}(b) 40kmh^{-1}, $\not{Z}_v^{\infty}$= 8.45 × 10^{-5}(c) 80 kmh^{-1}, $\not{Z}_v^{\infty}$= 3.82 × 10^{-4}.

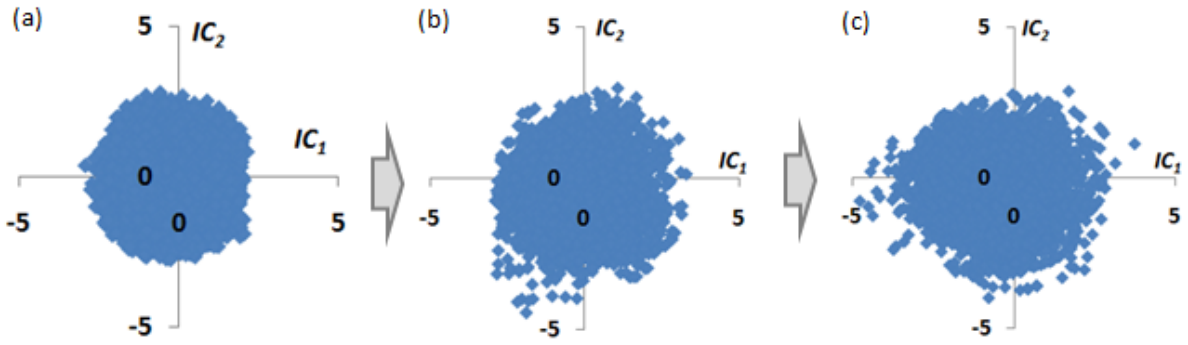


Figure 5: ICA displays for the different vehicle speeds: (a) Idle, K_{IC} = 2.31(b) 40kmh^{-1}, K_{IC} = 2.92 (c) 80 kmh^{-1}, K_{IC} = 3.22.

Comparison of the fast Fourier transform (FFT) plots of the original HAV acceleration signal with the two independent components are shown in Figure 6. The graph shows that the vibration value for each frequency is not changed between the recorded original HAV data and each of the two independent components. The FFT graph shows the level of the maximum vibration occurring at frequency of 3.9 Hz. The vibration was 2.81 ms^{-2}, with one of the independent components generating 2.89 ms^{-2}. The relative error between the two readings is 3%, which is considered to be within the practically acceptable relative error limit of 10% (sen *et al.*, 2006; Aziz *et al.*, 2014). Therefore, the I-kaz Vibro method for HAV monitoring has been validated using ICA.

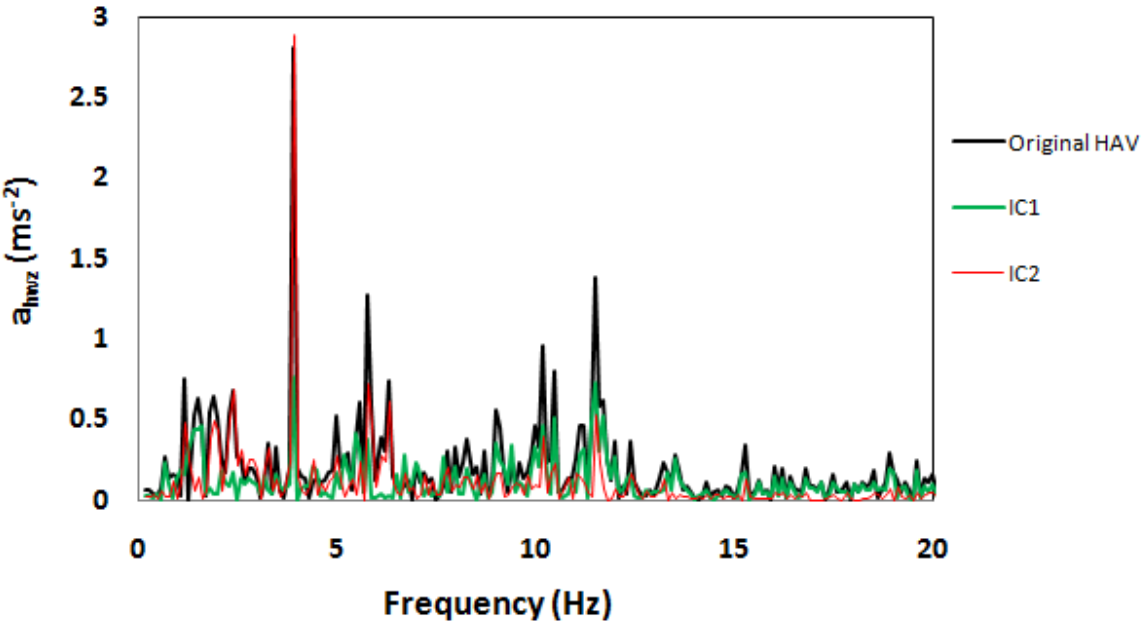


Figure 6: Comparison FFT graph between the original HAV with the IC components.

5. CONCLUSION

This paper validate the I-kaz Vibro method for HAV monitoring using ICA. From the results obtained for the I-kaz Vibro and ICA analyses, the a, Z_v^∞ and K_{IC} values increased when vehicle speed was increased. In addition, the results were also supported by the observation of the scattering of the I-kaz Vibro and ICA displays, which illustrate larger space of scattering for increasing vehicle speed and a, indicating that the Z_v^∞ and K_{IC} values are comparatively increasing. Comparison of the FFT plots of the original HAV acceleration signal with the two independent components show that the vibration value for each frequency is not changed between the original HAV data and each of two independent components. It was also found that the maximum vibration in the FFT graph was close to one of the independent components with the low relative error of 3%. Thus, it can be concluded that the I-kaz Vibro method for HAV monitoring has been validated accurately using ICA.

REFERENCES

Ahmad, T., Alias, N., Ghanbari, M., and Askaripour, M. (2013). Improved Fast ICA algorithm using eighth-order Newton's method, *Res. J. Appl. Sci. Eng. Tech.*, **6**: 1794–1798.

Ahmad, T. & Ghanbari, M. (2011). A review of independent component analysis (ICA) based on kurtosis contrast function. *Aust. J. Basic Appl. Sci.*, **5**: 1747–1755.

Aziz, S.A.A., Nuawi, M.Z., Nor, M.J.M. & Daruis, D.D.I. (2014). New regression models for predicting noise exposure in the driver's compartment of Malaysian Army three-tonne trucks. *Adv. Mech. Eng.,* **Vol. 2014:** 616093.

Aziz, S.A.A., Nuawi, M.Z. & Nor, M.J.M. (2015). New regression model for predicting hand-arm vibration (HAV) of Malaysian Army (MA) three-tonne truck steering wheels. *J Occup Health.* **57**: 513–520

Aziz, S.A.A., Nuawi, M.Z. & Nor, M.J.M. (2017). Monitoring of hand-arm vibration. *Int. J. Acoust. Vib.*, **22**: 34-43.

Guerrero-Mosquera, C. & Navia-Vázquez, A. (2012). Automatic removal of ocular artefacts using adaptive filtering and independent component analysis for electroencephalogram data. *IET Signal Process.*, **6**: 99–106.

Hyvarinen, A., Karhunen, J. & Oja E. (2001). *Independent Component Analysis*. John Willey & Sons, New Jersey.

Hyvarinen, A. (2013). Independent component analysis : Recent advances. *Phil. Trans. R. Soc. A*, **371**: 20110534.

ISO 5349-1 (2001). *Mechanical Vibration—Measurement and Evaluation of Human Exposure to Hand-Transmitted Vibration—Part 1: General Requirements*. International Organization for Standardization (ISO), Geneva.

Kang, H., Qi, Y., Yan, W. & Zheng, H. (2009). Application of independent component analysis in denoising the instantaneous signals. *1st Int. Conf. Inform. Sci. Eng.*, pp. 501–504.

Naik, G.R., Kumar, D.K. & Palaniswami, M. (2010). Classification of low level surface electromyogram using independent component analysis. *IET Signal Process,* **4**: 479–487.

Naik, G.R., & Kumar, D. K. (2011). An overview of independent component analysis and its applications. *Informatica*, **35**: 63–81.

Nakamura, H., Chida, S., & Yamazaki, T. (2014). Improvement of experimental sea model accuracy using independent component analysis, in inter noise. *Inter.Noise 2014*, 16-19 November 2019, Melbourne, Australia

Razaghi, H., Saatchi, R., Offiah, A., Burke, D., Bishop, N. & Gautam, S. (2013). Assessing Material Densities by Vibration Analysis and Independent Component Analysis, *Malaysian J. Fund. Appl. Sci.,* **9**: 123–128.

Sen, Z., Altunkaynak, A. & Ozger, M. (2006). Space-time interpolation by combining air pollution and meteorologic variables. *Pure Appl Geophys*, **163**, 1435–1451.

Wang, Z., Chen, J., Dong, G. & Zhou, Y. (2011). Constrained independent component analysis and its application to machine fault diagnosis, *Mech. Syst. Signal Proc.,* **25**: 2501–2512.

Zang, C., Friswell, M.I. & Imregun, M. (2002). Decomposition of time domain vibration signals using the independent component analysis technique. *3rd Int. Conf. Identification Eng. Syst.*, Swansea, April 2002, pp. 434-445.

AUTOMOTIVE DRIVE-SHAFT HEALTH CONDITION MONITORING AND RELAYING USING INTERNET OF THINGS (IOT)

Sivakumar Subburaj[1], Siva Irulappasamy[2*] & Ramalakshmi Ramar[1]

[1]Department of Computer Science and Engineering,
[2]Department of Mechanical Engineering
Kalasalingam Academy of Research and Education, India

[*]Email: isiva@klu.ac.in

ABSTRACT

Internet of Things (IoT) is fast growing and now becoming a fundamental need in our practical life. Predictive maintenance is one of the top IoT-use cases today, specifically in automotive applications. Among the several major components, power transmission shaft is critical in military wheeled vehicles and failure of the shaft leads to risk in the product's service life. In military ground vehicles, the fatigue cracks in drive shaft were induced by the cyclic stresses. Prediction and indication of such fatigue cracks would save the fault diagnosis time during the vehicle maintenance. Reading and recognising vibrational pattern of the drive shaft would help to continuously warrant the drive shaft health and ease of reporting the faulty condition when it occurs. Military wheeled vehicles have an enormous potential in using sensor-based diagnose system to track the engine and prime mover faults, whereas, now it is indeed to develop an electronic system with data transmission in order to monitor the health of drive shaft of military vehicle to promise safety riding. In this study, a laboratory-scale unit is developed to simulate the typical fatigue response of military ground vehicle drive-shaft with the necessary IoT units. Both online and IoT based test rigs with control systems developed for condition monitoring in a complete automation approach.

Keywords: *Drive shaft; fatigue failure; health condition monitoring; Internet of Things (IoT); data acquisition.*

1. INRODUCTION

In the history of general automotive failures from ancient eras, power-shafts have played a vital role since the produced power from the engine reaches the differentials directly through the power shafts. Disconnection is not only interrupting power transmission; perhaps terrible life lost may also occur due to the severe structural damages (Souflas *et al*., 2018). Online condition monitoring is installed in most of the modern vehicles mainly to check the engine conditions. Perhaps, the transmission line may demand indispensably. Among several types of research on condition monitoring of automotive components, Vicuña (2014), has developed a condition monitoring unit for planetary gear-box through acoustic sensing unit to diagnose the faulty conditions in gears. Acoustic signals for varying load and rotation are acquired to set the healthy signal database. Based on the experimental results, the health condition of gears is determined. The author has concluded with the potential importance of condition monitoring in power transmission units to enable a hassle-free running of automotive. In the recent year (Coppola & Morisio, 2016), the number of vehicles connected to the internet is increasing; which in turn, some new methodologies are required in storage/handling the data when connecting multiple vehicles at different locations. Present condition monitoring technology involves many advanced instrumentations but faces

challenges in storage and data analysis. The vibration threshold values are useful for failure identification, perhaps lacks to provide enough information to identify the fault on the real-time situation.

In power transmission shaft, failure occurs when a material subjected to cyclic loading in addition to the chemical environment (corrosion), production and maintenance strains. Failure has treated as either traverse or vertical cracks on the shaft (Loutas *et al*. 2011) and too challenging to identify at the earlier stage, namely during crack initiation period. This failure detection problems have to be controlled by monitoring and diagnosed at the earlier stages (Sinou & Lees, 2005).

In rotating equipment, vibration and sound signals are directly related to the structural dynamics of the machine and contain sufficient information about the condition of components (Abu-mahfouz & Banerjee, 2017). Vibration signals are mainly characterised using time domain, frequency domain and time-frequency domains analysis, which shows the primary failure information (Dalpiaz *et al.* 2000). Each domain has a different pattern of vibration signal analysis. The time-domain technique uses statistical parameters, frequency domains uses fast Fourier transform (FFT), and time-frequency domains use non-stationary waveform signals (Zhang & Nandi 2007; Lei *et al.* 2013; Feng *et al*. 2013). Also wavelet transform is the tool for fault diagnosis of rotating machinery (Yan *et al*., 2014).

Many researchers have conducted failure detection in rotating members of an automotive vehicle using signals processing methods, and prediction has changed numerical and simulation data to support the experimental purity. In principle, crack size and location have been identified using the change of frequency and amplitude vibration signals in a rotating shaft (Adewusi & Al-Bedoor, 2002; Mohamed *et al*. 2011). Analytical approaches have demonstrated that vibration monitoring has tremendous potential in detecting localised defects in the machines. When mounted in proximity of bearing housing (a general case), these modules can collect the stationery as well as non-stationary signature data of vibration of bearing housing reliably. Since most vibration sensors mounted in proximity of bearing housings (based on mechanical impedance considerations); bearing fault detection techniques can be implemented for online bearing condition monitoring. This module can easily be deployed for different rotating machines for vibration monitoring purposes (Lei *et al.* 2013).

Lu & Chu (2011) have studied a novel smart sensing unit for vibration measurement and machinery condition monitoring. Here the author has embedded accelerometer with the Arduino microcontroller board, where microprocessor-based smart sensor collects 3-D vibrations; and these stored data in the microcontroller can be sent to third party devices (Laptop, PC, etc.) for further signal analysis. Jaber and Bicker (2015) have studied online condition monitoring based fault detection of an unbalanced rotor induction motor (IM). An open-source quick and secure protocol MQTT (Message Queue Telemetry Transport) is implemented for reliable communication in connected cars (Dhall & Solanki, 2017).

The main contributions of this paper are:

i. An effective IoT based condition monitoring system developed for automotive power transmission shaft.
ii. Development of experimental test rig using IoT hardware, software and integration with the cloud using IoT communication protocol discussed.
iii. Successful implementation of control methodology using experimentations.

2. LABORATORY SIMULATOR AND EXPERIMENTATION

In order to simulate the mechanical response of beam member, a laboratory-scale rotary bending test rig was modified and AA7475 samples are taken as the representative to the military vehicle drive shaft.

Figure 1 shows the experimental set-up, along with a data acquisition system and virgin specimen adopted in this study. Experiments conducted for different loading and rotational speed and frequency generated at the moment of fracture has taken as a response.

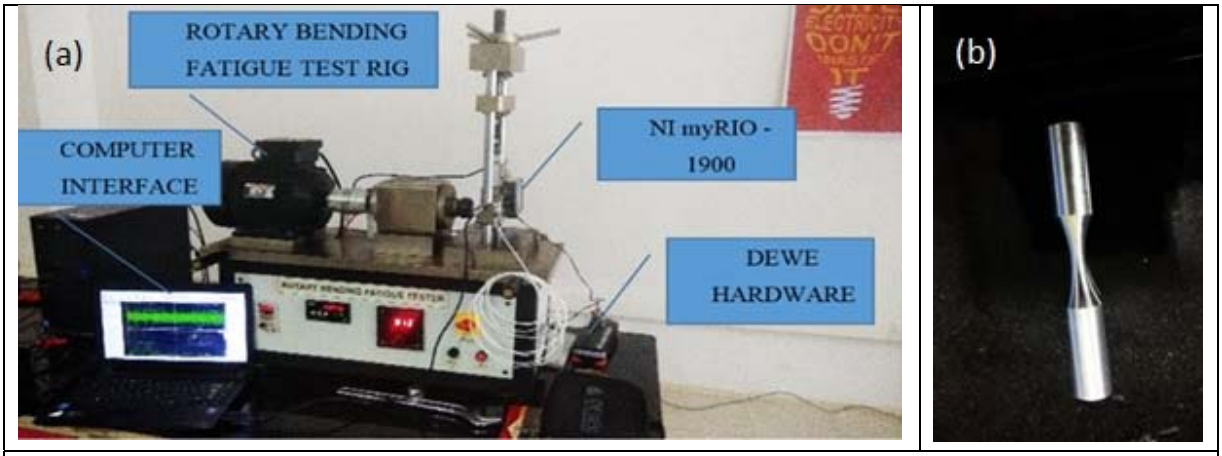


Figure 1: (a) Automotive drive shaft simulator. (b) AA7475 specimen.

2.1 Online Condition Monitoring Using LabVIEW

Aluminum 7475 (AA7475), which is mostly used in automotive application material test sample was modelled as a representative to the actual driveshaft and loaded into the test rig. Various sensors and Arduino microcontroller was used for implementing data acquisition system. An adxl345 accelerometer was mounted on the top of the bearing which connects the propeller shaft for measuring the vibration signal and connected to the digital input pin of Arduino. An adxl345 accelerometer was mounted on the top of the bearing which connects the propeller shaft for acquiring the vibration signal and was connected to the digital input pin of Arduino.

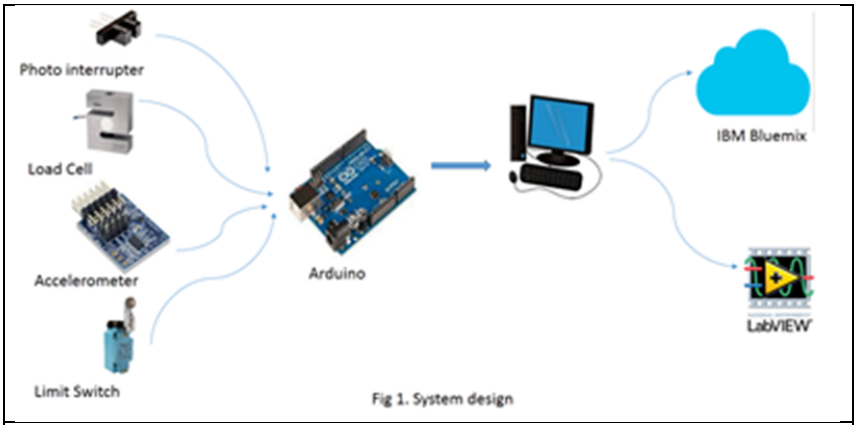


Figure 2: System design.

2.2 IoT-based Condition Monitoring System

Real-time condition monitoring achieved by integrating IBM Bluemix cloud platform with the help of Node-RED. Node-RED is an open-source visual tool which has a lot of support for IoT. It helps IoT developer to integrate hardware, API and cloud. The cloud receives the real-time sensor data from sensors using the MQTT protocol. MQTT is a lightweight communication protocol, and it receives data in the form of JSON.

3. METHODOLOGY

The flow diagram shown in Figure 3 represents the methodology implemented in an online condition monitoring system of fatigue testing unit. In this case, LabVIEW acts as master and Arduino acts as a slave device. The sensors (photo interrupter and load cell) were used to acquire machine cycle counts data, and applied data interfaced with Arduino microcontroller. The limit switch was also interfaced with Arduino.

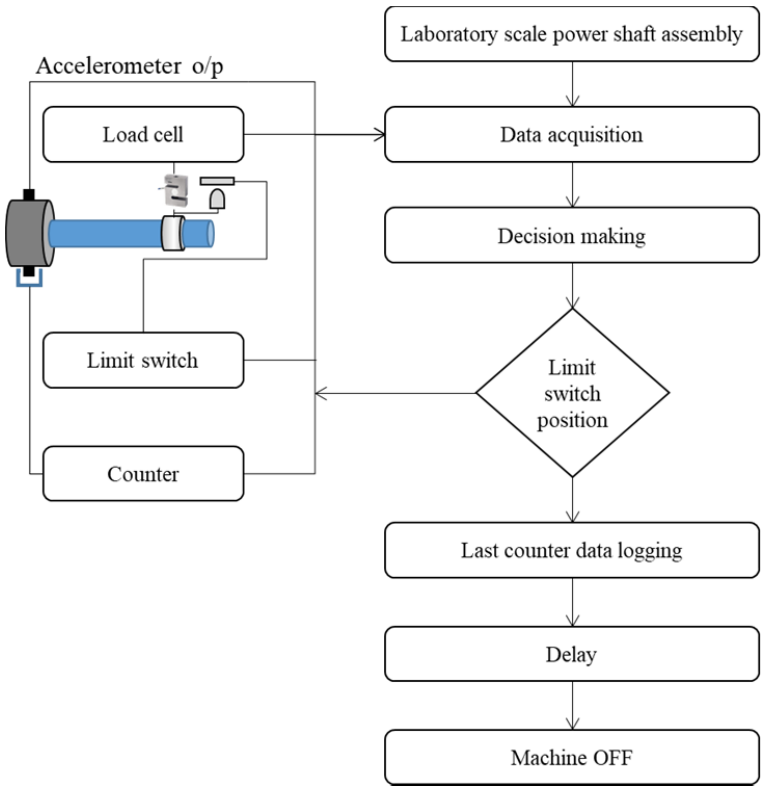


Figure 3: The methodology employed in this study.

The purpose of Arduino is to acquire data from sensors and sends data to LabVIEW for monitoring purpose. During the actual condition, the limit switch will be in NO logic and fracture condition; the limit

switch is forced to be in NC logic. NO to NC change can be captured by Arduino and LabVIEW for identifying the specimen's fracture to log the machine cycle count and load data at the instant of fracture.

3.1 IoT Based Condition Monitoring

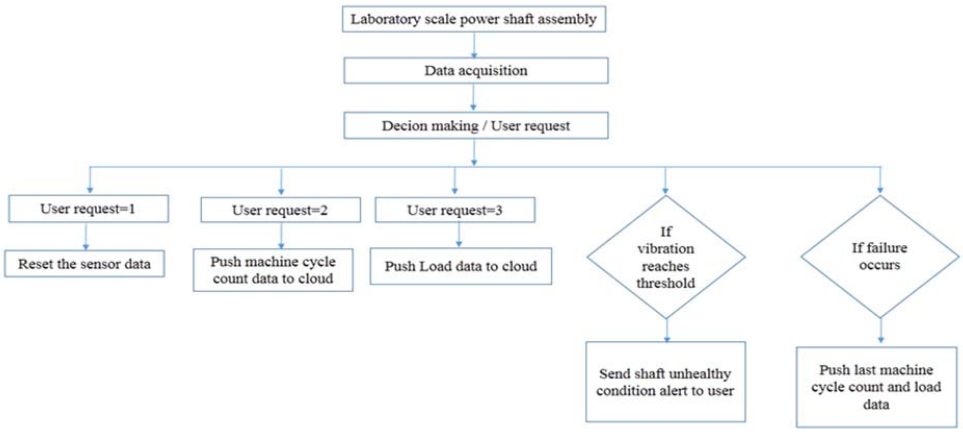


Figure 4: IoT based condition monitoring.

The rest of the paper deals with global monitoring and health prediction of power shaft test unit via IoT and the flow diagram as shown in Figure 4, which explains the methodology implemented in this drive-shaft condition monitoring system through IoT. In this system, IoT Bluemix platform at end user acts as master and Arduino acts as slave based on the user request or fatigue fracture status, Arduino pushes sensor data to end user (IoT Bluemix Node-RED platform). The detailed working of a system was divided into 5 cases and is explained below.

Case 1: Data '1'
If the user sends Data '1' from IoT Bluemix Node-RED platform to Arduino, Data '1' is the command to reset the system parameters of fatigue tester unit.

Case 2: Data '2'
If the user sends Data '2' from IoT Bluemix Node-RED platform to Arduino, Data '2' is the command to push current machine cycle count data of fatigue testing unit.

Case 3: Data '3'
If the user sends Data '3' from IoT Bluemix Node-RED platform to Arduino, Data '3' is the command to push load data of fatigue tester unit.

Case 4: vibration level reaches the threshold
If the vibration level of power transmission shaft reaches its threshold level, Arduino pushes unhealthy condition of power shaft alert message to IoT user end. The threshold value will vary depending upon the material.

Case 5: Power transmission shaft fracture
If power transmission shaft fracture occurs, Arduino pushes instantaneous data of machine cycle count and load at fracture point to IoT.

4. RESULTS & DISCUSSION

4.1 Online monitoring using LabVIEW

ATMEGA 328P controller acquires machine cycle count, load and fatigue failure data from respective sensors in the test rig, and it uploads the acquired data to LabVIEW via serial communication. For Fatigue Test rig monitoring and data logging, a customised VI was developed, and its logging result is shown in below Figure 5.

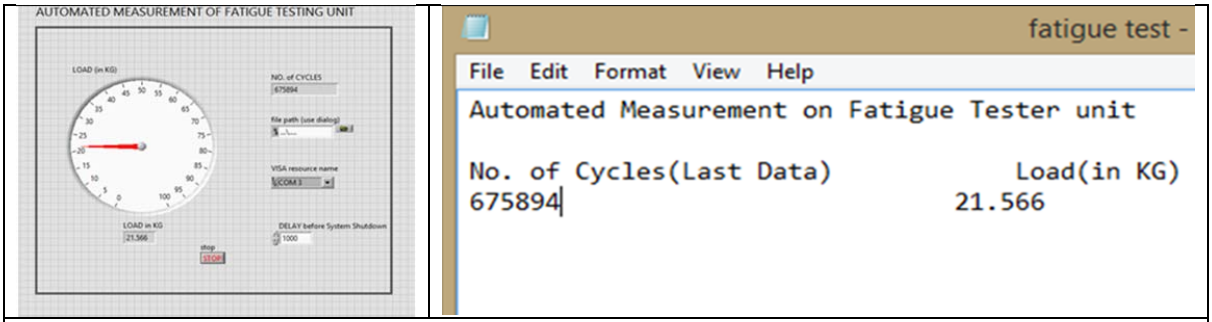


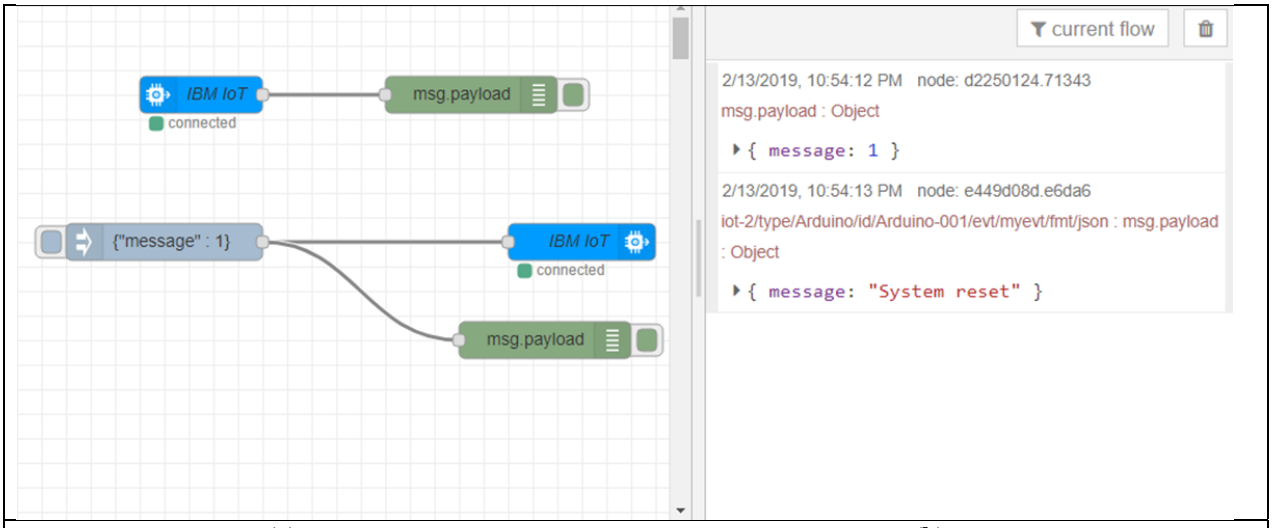
Figure 5: (a) LabVIEW VI front Panel for monitoring and data logging. (b) The result obtained

4.2 IoT Based Monitoring

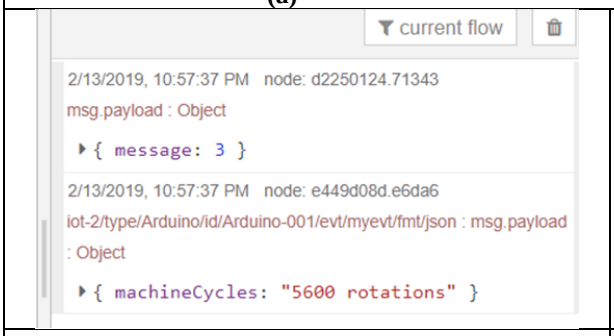
ATMEGA 328P controller collects data from sensors and uploads the acquired data to the PC via serial communication. From the serial port, Node.js is the software which pushes the serial port data to IoT server. Node.js acts as a bridge between the serial port and internet cloud server. The below listed several actions will happen in the system according to the client request and in fracture conditions.

Table 1: List of actions that occurred during catastrophic failures.

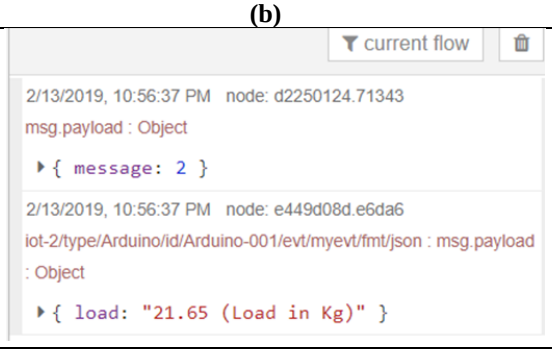
S.No	User Request / Fatigue Status	Corresponding Action
1	User sends data 1	Arduino Resets the machine
2	User sends data 2	Arduino pushes Machine cycle count data to IoT Bluemix Page
3	User sends data 3	Arduino pushes Load data to IoT Bluemix Page
4	Shaft vibration reaches continuous threshold	Arduino pushes unhealthy condition alert message to IoT Bluemix Page
5	Fatigue failure occurs	Arduino pushes last machine cycle count and load data to IoT Bluemix Page



(a) (b)

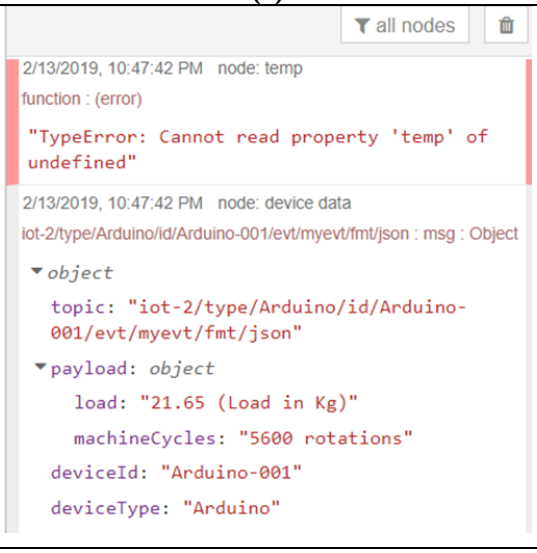


(c)



(d)

(e)



(f)

Figure 6: (a) Node-RED platform. (b) User Request to reset the sensor measurement in Arduino. (c) User request to push cycle count value from Arduino to cloud. (d) User request to push load data from Arduino to cloud. (e) Push shaft health status. (f) Last machine cycle count and load data.

4.3 Vibration signal analysis

The time domain and frequency domain analysis results are plotted in Figure 8 for different loads and fault conditions.

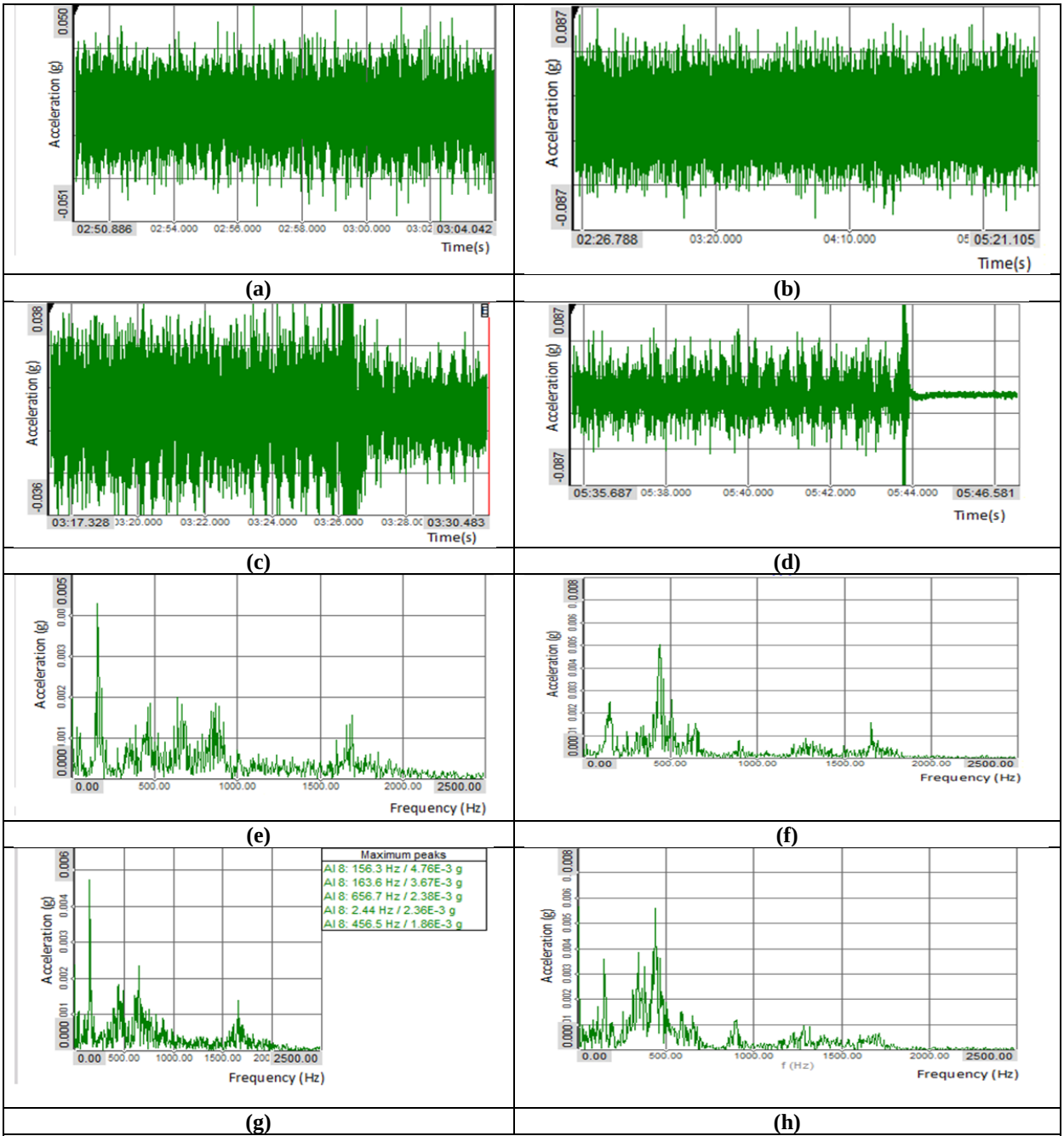


Figure 7: The time-domain plot of vibration data for healthy shaft for (a) Load 10 Kg, Speed 800 rpm and (b) Load 20 Kg, Speed 800 rpm. The time-domain plot of vibration data for faulty shaft for (c) Load 10 Kg, Speed 800 rpm and (d) Load 20 Kg, Speed 800 rpm. FFT plot of vibration data for healthy shaft for (e) Load 10 Kg, Speed 800 rpm and (f) Load 20 Kg, Speed 800 rpm. FFT plot of vibration data for faulty shaft for (g) Load 10 Kg, Speed 800 rpm and (h) Load 20 Kg, Speed 800 rpm.

From the time domain plot, it is observed that the amplitude of vibration for healthy shaft and faulty shaft increases as the load increased. It is found that the amplitude of vibration of a faulty shaft is greater than that of normal shaft. Since, the faulty or shaft with significant crack would be in imbalanced rotation, hence the vibration amplitude would be higher than the healthy shaft; the same was also reported by Prasad & Sekhar (2018).

5. CONCLUSION

The importance of safe riding with automotive used in military purpose would be a challenging research since the vehicle operation is rugged. Perhaps developing a system to do online monitoring of those vehicles also create lot of challenges. In this report, a LabVIEW based online monitoring and control set up was developed which logs the cycle counts and load data at fracture condition. To ensure the reduced power conception in online mode, the whole system is programmed to shut down after fatigue failure with a delay of 60 seconds. The effective IoT-based condition monitoring and health prediction system were also developed for power transmission shaft in rotary bending fatigue tester. The IoT-based model integrates sensors and Arduino with cloud in real-time for an effective condition monitoring system. The load, machine cycle count data and crack status acquired from test rig are sent to IBM Bluemix cloud platform by using Arduino and node.js.

REFERENCES

Abu-Mahfouz, I. & Banerjee, A. (2017). Crack detection and identification using vibration signals and fuzzy clustering. *Procedia. Comput. Sci.*, **114**: 266–274.

Adewusi, S.A. & Al-Bedoor B. (2002). Experimental study on the vibration of an overhung rotor with a propagating transverse crack. *Shock. Vib*, **9**: 91–104.

Coppola, R. & Morisio, M. (2016). Connected car: technologies, issues, future trends. *ACM. Comput. Surv. (CSUR)*, **49**: 46(1-36)

Dalpiaz, G., Rivola, A. & Rubini, R. (2000). Effectiveness and sensitivity of vibration processing techniques for local fault detection in gears. *Mech. Syst. Signal. Pr.*, **14**: 387-412.

Dhall, R. & Solanki, V.K. (2017). An IoT Based Predictive Connected Car Maintenance Approach. *Int. J. Interact. Multimed. Artif. Intell.*, **4**: 16-22

Feng, Z., Liang, M. & Chu, F. (2013). Recent advances in time – frequency analysis methods for machinery fault diagnosis : A review with application examples. *Mech. Syst. Signal. Pr.*, **38**: 165–205.

Jaber, A.A. & Bicker, R. (2015). Real-Time Wavelet Analysis of a Vibration Signal Based on Arduino-UNO and LabVIEW. *Int. J. Mat. Sci.*, **3**: 66–70.

Lei, Y., Lin, J., He, Z. & Zuo, M. J. (2013). A review on empirical mode decomposition in fault diagnosis of rotating machinery. *Mech. Syst. Signal. Pr.*, **35**: 108–126.

Loutas, T.H., Roulias, D., Pauly, E. & Kostopoulos, V. (2011). The combined use of vibration , acoustic emission and oil debris on-line monitoring towards a more effective condition monitoring of rotating machinery. *Mech. Syst. Signal. Pr.*, **25**: 1339–1352.

Lu, W. & Chu, F. (2011). Shaft crack identification based on vibration and AE signals. *Shock. Vib.*,**18**: 115–126.

Mohamed, A.A., Neilson, R., MacConnell, P., Renton, N. C. & Deans, W. (2011). Monitoring of fatigue crack stages in a high carbon steel rotating shaft using vibration. *Procedia. Engineer.*, **10**: 130–135.

Prasad, S.R. & Sekhar, A. S. (2018). Life estimation of shafts using vibration based fatigue analysis. *J Mech. Sci. Tech.*, **32**: 4071-4078.

Sinou, J.J. & Lees, A. W. (2005). The influence of cracks in rotating shafts. *J. Sound. Vib.*, **285**: 1015–1037.

Souflas, I., Pezouvanis, A. & Ebrahimi, K. M. (2018). Health monitoring system for transmission shafts based on adaptive parameter identification. *Mech. Syst. Signal. Pr.*, **104**: 673–687.

Vicuña, C.M. (2014). Effects of operating conditions on the Acoustic Emissions (AE) from planetary gearboxes. *Appl. Acoust.*, **77**: 150–158.

Yan, R., Gao, R. X. & Chen, X. (2014). Wavelets for fault diagnosis of rotary machines: A review with applications. *Signal. Process.*, **96**: 1–15.

Zhang, L. & Nandi, A. K. (2007). Fault classification using genetic programming. *Mech. Syst. Signal. Pr.*, **21**: 1273–1284.

INVESTIGATION OF THE MECHANICAL PROPERTIES OF STANDARD MALAYSIAN RUBBER WITH CONSTANT VISCOSITY AND EPOXIDISED NATURAL RUBBER USING NANO-INDENTATION TEST

Mohd Azli Salim[1,2,*], Adzni Md. Saad[1], Azmi Naroh[3], Mohd Nizam Sudin[1], Crtomir Donik[4], Norbazlan Mohd Yusof[5] & Intan Raihan Asni Rosszainily[1]

[1]Fakulti Kejuruteraan Mekanikal, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia
[2]Advanced Manufacturing Centre, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia
[3]Jabatan Kejuruteraan Mekanikal, Politeknik Ungku Omar, Malaysia
[4]Institute of Metals and Technology, Slovenia
[5]Centre of Excellence, Projek Lebuhraya Usahasama Berhad (PLUS), Malaysia

[*]Email: azli@utem.edu.my

ABSTRACT

The usage of natural rubber (NR) in the development of laminated rubber-metal spring (LR-MS) for automotive applications has led to the investigation on its mechanical properties. This paper describes the use of nano-indentation test to investigate the mechanical properties of standard Malaysian rubber (SMR) with constant viscosity and epoxidised natural rubber (ENR). The mechanical properties of SMR Constant Viscosity 60 (SMR CV-60) and 25 mole % ENR (ENR 25) reinforced with 0, 20, 40 and 60 phr of carbon black (CB) was investigated. The nano-indentation test was carried out using Berkovich tips at constant load of 2 mN with holding times of 0, 5, 10, 15, and 20 s. It was found that the SMR CV-60 and ENR 25 compounds with 60 phr CB loading recorded the highest hardness and elastic modulus values, and also had the lowest penetration depth. The test also revealed that the hardness and penetration depth were independent to the holding times. In contrast, the indentation elastic modulus was found to be highly affected by the holding time.

Keywords: *Standard Malaysian rubber (SMR); natural rubber (NR); epoxidised natural rubber (ENR); nano-indentation.*

1. INTRODUCTION

Rubbers are available in two types, which are synthetic and natural. Synthetic rubber is an artificial rubber that is produced through the polymerisation process of petroleum by-product. It was invented to reduce dependencies on the natural rubber (NR) during the World War II. As for today, there are more than 20 types of synthetic rubber that are available for various applications. On the other hand, NR is naturally extracted from rubber trees through the tapping process, which is a process of latex secretion by shearing a thin layer of rubber tree's bark by using a special tapping knife. There are more than 2,500 known species of plants that can produce latex. In spite of that, only latex produced from the *Hevea Brasiliensis* tree is mostly used for rubbery products due to its quality (Lindley, 1964; Aravind *et al*., 2015).

Although synthetic rubbers are well known to offer numerous advantages in terms of mechanical and chemical properties, NR also possesses advantages over synthetic rubbers. NR can be considered as green material since it can be continuously supplied by nature as compared to the synthetic rubber, which is derived from the non-renewable sources (Boonkerd, 2017). Besides that, the price of raw NR is more economical than synthetic rubber. The price of synthetic rubber normally fluctuates depending on the current crude oil prices. Other than that, the properties of NR compounds can be derived from better formulation and processing. NR also allows low heat build-up in dynamic force to

make it able to withstand large deformations and grants it with the ability to instantly recover when distorted at room temperature (Chandrasekaran, 2010). The application of NR as the main material for vibration isolator in automotive applications does not require extensive maintenance due to its capability of maintaining itself for a long period of time. Additionally, NR also exhibits inherent damping and spring-like performance in order to reduce the resonance effect. NR can be bonded with other materials such as metal in order to increase its strength (Heide *et al*., 2018).

As one of the largest NR manufacturers and exporters, Malaysia has taken an initiative to promote local NR in various fields, including construction, automotive and sports. The Malaysian government has established the Malaysian Rubber Board (MRB) and Tun Abdul Razak Research Centre (TARRC) to provide the best infrastructure for research and development of local NR. The quality and competitiveness of Malaysian NR products, especially in vibration isolation, has been proven as one of the best in the world. One of the examples of Malaysian NR applications in vibration isolation can be found in the structure of the Sultan Abdul Halim Muadzam Shah Bridge, Penang that is equipped with high damping NR to withstand from possible large earthquakes (Picken *et al*., 2012). Besides that, most of the Malaysian ports also use the Malaysian-made dock fenders as a bumper to absorb impact from collisions between boats and jetties (Yu *et al*., 2001; Kim *et al*., 2004).

Based on previous research, it was found that the development of the laminated rubber-metal spring (LR-MS) for automotive mounting applications is currently in trend (Salim *et al*., 2013, 2014, 2016). The study on the transmissibility in the axial direction and parameter assessment on LR-MS was conducted. This study looks into the application of Malaysian NR, which is the vulcanised standard Malaysian rubber constant viscosity 60 (SMR CV-60) reinforced with carbon black (CB) as the potential main material, combined with a metal plate for the development of LR-MS. In order to determine the compatibility of local NR grade as the main substance in the development of LR-MS, the mechanical properties of rubber reinforced with CB need to be investigated.

This study focuses on the investigation of the mechanical properties of vulcanised SMR CV-60, reinforced with different CB loadings ranging from 0, 20, 40 and 60 part per hundred of rubber (phr) using nano-indentation test. Besides that, the commercial 25% mol ENR (ENR 25) with similar compounding formula and CB loading was also used as a comparison to SMR CV-60. Nano-indentation test is used in this study as it is more cost-effective and non-destructive, as well as requiring smaller test pieces as compared to conventional techniques.

2. METHODOLOGY

This section describes in detail the methodologies for material preparation and nano-indentation test.

2.1 Material Preparation

Two types of NR compound were used in this study, namely vulcanised SMR CV-60 and vulcanised ENR 25. As a collaboration with the Malaysian Rubber Board (MRB), the material preparation of milling and mixing, as well as the vulcanising processes were conducted in MRB's accreditation laboratory in order to maintain the quality of the compounded rubber.

The composition ingredients for the rubber compound were prepared based on phr. The complete formulation for both rubbers is tabulated in Table 1. The addition of sulphur helped to increase the crosslinking in the rubbers, which resulted in the improvement of rubber texture from soft to hard. Meanwhile, the other ingredients, such as zinc oxide, stearic acid and N-cyclohexyl benzothiazyl sulphenamide (CBS), worked as the accelerator to boost up the sulphur crosslink efficiency. Santoflex 13 and paraffin wax were also added into the compound as ozone protective agents. N330 CB was used as the filler due to its good surface activity and chemical properties, which influenced good

interface interaction between the rubber chain and CB particles (Sangwichien *et al*., 2008). In this study, the CB loading in the NR compounds was set as the changing parameter, which varied at 0, 20, 40 and 60 phr.

Table 1: Formulation for the SMR CV-60 and ENR 25 compounds.

Ingredient	Amount (phr)
SMR CV-60/ ENR 25	100
Zinc Oxide	5
Stearic Acid	2
CBS	0.8
Sulphur	3.25
Black HAF, N330	0, 20, 40, 60
Santoflex 13	3
Parafin Wax	2

The equipment specifications and procedures for the mixing and vulcanisation processes were based on ASTM D3182: Standard Practice for Rubber-Materials, Equipment, and Procedures for Mixing Standard Compounds and Preparing Standard Vulcanized Sheets, which is the guideline for preparing standard vulcanised rubber sheets. The compounding process was started by mixing the dried NR with other ingredients in the rolling machine for about 10 to 15 min. Then, the compound was left to rest for at least 4 h to reach its steady state before moving to the next process. The rheometer test was conducted on the compound in order to obtain the parameter for the vulcanisation process. The compounds were vulcanised at temperature of 150 ˚C for 8 to 12 min, depending on the CB compositions in the NR compound. The vulcanising process refers to the treating of rubber materials with sulphur in the presence of great heat to improve its elasticity and hardness.

2.2 Test Load Determination

This study provides the findings of conducting the nanoindentation test on vulcanised natural rubber compounds of SMR CV-6 and ENR 25. An indentation-creep test was conducted at various holding times starting from 0 to 20 s. A fixed 2 mN test force was used as the maximum peak load. As there was no specific peak load stated in the ASTM standard, the peak load value was determined by referring to Oyen (2007) as a precaution in order to avoid excessive applied force or the indenter tip to indent over the specimen.

Oyen (2007) determined the peak load level (P_{max}) range, depending on the material's elastic modulus ($E_{elastic}$) and Poisson's ratio (υ). This researcher applied P_{max} ranging from 1 mN and 10 mN to indent materials with $E_{elastic}$ = 210 MPa and 4 MPa with υ = 0.42 and 0.50 respectively. The researcher used a spherical tip to indent up to 3.05 mm from the thickness of the material. By referring to the range of P_{max} as stated in the previous research as the limit for maximum applied force and the Young's Modulus value obtained from the tensile and compression tests, 2 mN test force was used in this study.

2.3 Nano-Indentation Setup

The indentation test was used to determine the strength properties of material through its surface. In this study, the nano-indentation test was conducted to measure the specimens Young's Modulus and other properties. The effects of holding time on the material properties was also investigated. The experiment was conducted by referring to the ASTM standard E2546-15: Standard Practice for Instrumented Indentation Testing. This standard only covers the practice requirement of indentation testing without specifying any test force, indentation depth range or any specific indenter types.

All the tests were performed at ambient temperature of 23 ˚C. The indentation test was conducted on eight samples with four different CB loadings using a nano-hardness tester (Shimadzu Dynamic Ultra Micro Hardness Tester Model DUH-211). The nano-indenter device was equipped with an imaging device that is capable to switch back and forth from the optical microscope to the indenter tips. The mounted imaging device helps to accurately identify the desirable indentation point. A standard Berkovich type 115˚ triangular pyramidal indenter was used in this study. The machine was operated with the aid of the DUH211 software that comes together with the equipment.

Both vulcanised rubber sheets with 2 mm thickness were cut into 10 mm x 10 mm squares for the tests. The test piece size was determined by referring E2546-15. The test piece thickness was set to be at least 10 times of the maximum indentation depth (if conducting displacement assessment) or six times of the indentation radius. Meanwhile, the test piece size should not be too small or too big, depending on the size of indenter holder. All the test pieces' surfaces were wiped off with a damp cloth to remove dirt that might affect the indentation results. The test pieces were let to completely dry before testing.

Each test piece was mounted on a platform to keep it stationary while conducting the indentation test. The test piece was placed under an optical microscope to determine the indentation point before the platform was moved to the indenter side. The load-hold-unload test was conducted with a fixed maximum test force of 2 mN and constant indenter loading speed of 0.2926 mN/sec. The maximum test force was determined by considering certain conditions, such as the test piece thickness, types of materials, probability of materials to deform during indentation, surface finish, and machine competency. The indentation tests were also conducted at different holding times, which varied from 0, 5, 10, 15, and 20 s. All the tests were conducted with three repetitions, where the mean and standard deviation were calculated using Equations 1 and 2 respectively.

$$\bar{x} = \frac{\sum fx}{n} \tag{1}$$

and

$$S = \sqrt{\frac{\sum \left(x - \bar{x} \right)^2}{n-1}} \tag{2}$$

where $\bar{x}$ is the mean values of the data set, n is the number of data set, and $\sum fx$ is the summation of the all the data set. S represents the standard deviation of the sample, $\sum$ means "sum of", and x represent each value in the dataset.

This study was carried out with two assumptions: no sink-in and pile-up effect were accounted during the experiment. The nano-indentation results were recorded and presented in terms of maximum force (F_{max}), maximum indentation depth (h_{max}), hardness (H_u), and Elastic Modulus (E_{it}).

3. RESULTS & DISCUSSION

3.1 Effect of CB Loading

The experimental results for the nano-indentation test on SMR CV-60 and ENR 25 filled with CB loadings of 0, 20, 40 and 60 phr are shown in Tables 2 and 3. Based on the obtained results, the recorded maximum force ranged from 2.00 to 2.01 MPa in response to the test constant parameter, which was the maximum indentation load at 2 mN.

Table 2: The nano-indentation properties for the SMR CV-60 compounds.

Nano-indentation properties	Carbon loading (phr)			
	0	20	40	60
Maximum Force (F_{max}, mN)	2.01±0.01	2.01±0.01	2.00±0.01	2.01±0.01
Maximum Indentation Depth ($h_{max,}$ μm)	15.36±1.00	13.13±0.25	9.25±1.00	7.20±0.87
Hardness (H_u , MPa)	0.33±0.04	0.44±0.02	0.91±0.21	1.80±0.41
Elastic Modulus (E_{it} , MPa)	5.01±0.79	7.21±0.23	18.29±4.27	42.51±4.27

Table 3: The nano-indentation properties for the ENR 25 compounds.

Nano-indentation properties	Carbon loading (phr)			
	0	20	40	60
Maximum Force (F_{max}, mN)	2.01±0.01	2.00±0.01	2.02±0.01	2.01±0.01
Maximum Indentation Depth ($h_{max,}$ μm)	12.89±0.45	12.47±1.09	8.69±0.87	6.65±1.54
Hardness (H_u , MPa)	0.44±0.03	0.50±0.09	1.03±0.22	1.42±1.04
Elastic Modulus (E_{it} , MPa)	7.21±1.61	8.57±2.12	21.17±2.71	51.26±18.84

The relationship between h_{max} and the CB loading of SMR CV-60 and ENR 25 is presented in Figure 1. It shows that the h_{max} values decreased with the increase of CB loadings for both NR compounds. Besides that, the h_{max} value for SMR CV-60 was slightly higher as compared to ENR 25. The highest h_{max} value recorded by SMR CV-60 was 15.36 μm, while the lowest was 7.20 μm. Meanwhile, the highest h_{max} value for the ENR 25 compounds was 12.89 μm, while the lowest was 6.65 μm.

The graph of H_u versus CB loadings is presented in Figure 2. Both NR compounds exhibited similar graph trend, where H_u was found to increase as the CB loadings was increased. Additionally, the graph also showed that the H_u value of SMR CV-60 was slightly lower than ENR 25 with the CB loadings of 0 phr to 40 phr and conversely higher at 60 phr. The H_u values at 0 phr were 0.33 MPa for SMR CV-60 and 0.46 MPa for ENR 25. H_u increased slightly at 20 phr with the value of 0.44 MPa for SMR CV-60 and 0.50 MPa for ENR 25. Then, at 40 phr, the H_u value was found to gradually increase at the value of 0.91 MPa for SMR CV-60 and 1.03 MPa for ENR 25. The H_u values for SMR CV-60 and ENR 25 at 60 phr were 1.80 and 1.42 MPa respectively.

The correlation between h_{max} and H_u can be found in Figures 1 and 2, where h_{max} was found to decrease while H_u increased with the increase of CB loadings. This is as the increase of CB loadings increased the hardness of NR, thus enhancing the NR stiffness to become less elastic. This restrained the indenter tips to penetrate further into the materials, resulting in lower h_{max} as the compound contains high CB loadings (Callister & Rethwisch, 2011).

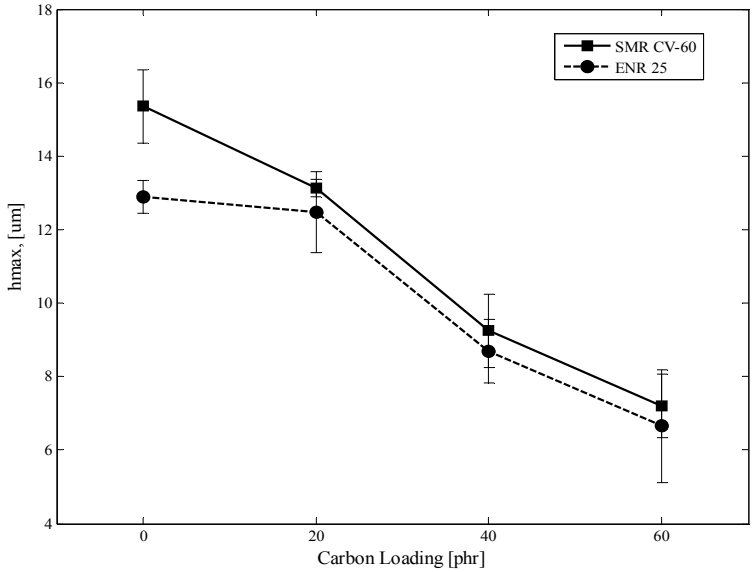


Figure 1: The graph of maximum indentation depth (h_{max}) versus CB loadings for SMR CV-60 and ENR 25.

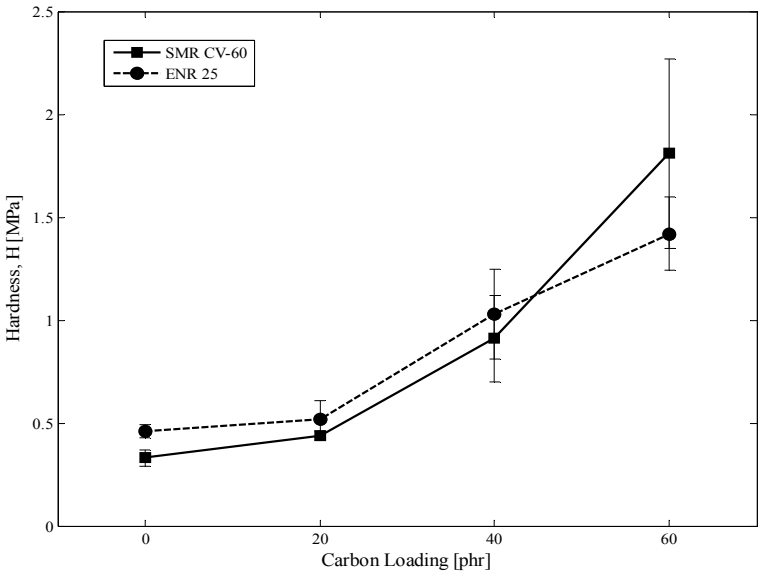


Figure 2: The graph of hardness (H_u) versus the CB loadings for SMR CV-60 and the ENR 25.

The graph of E_{it} for SMR CV-60 and ENR 25 at different CB loadings is depicted in Figure 3. It is found that E_{it} for SMR CV-60 was slightly increased from 0 to 20 phr of CB loadings. Contrarily, E_{it} for ENR 25 was found to slightly decrease from 0 to 20 phr. Subsequently, the E_{it} values for both NR compounds were found to exhibit a significant increment as the CB loadings was increased up to 60 phr. The recorded E_{it} values for SMR CV-60 were 5.01, 7.21, 18.29, and 42.51 MPa at 0, 20, 40 and 60 phr of CB loadings respectively. On the other hand, the recorded values for ENR 25 were 8.58, 8.57, 21.17 and 51.26 MPa at 0, 20, 40 and 60 phr of CB loadings respectively. The graph also showed that the E_{it} values of the ENR 25 compounds were higher as compared to the SMR CV-60 compounds. This shows that the CB particles interact better with the ENR 25 particles to produce stronger NR chains as compared to the SMR CV-60 particles.

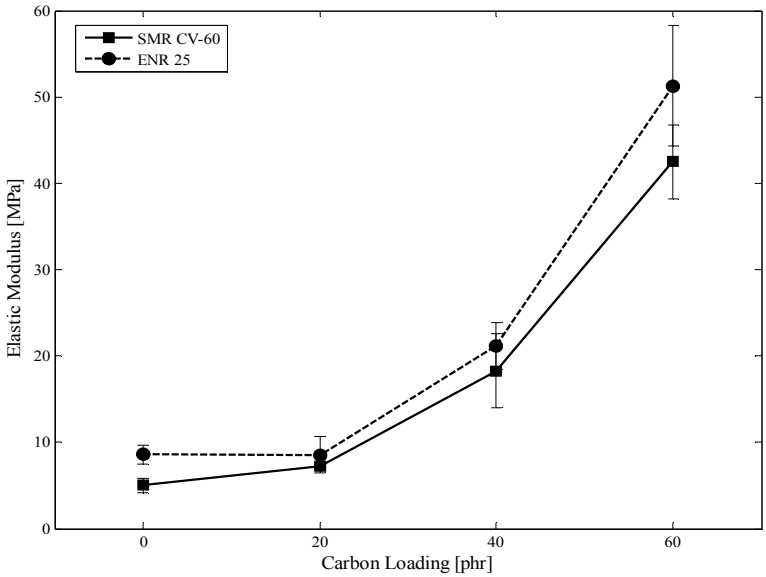


Figure 3: The graph of elastic modulus (E_{it}) versus the CB loadings for the SMR CV-60 and the ENR 25.

3.2 Effect of Different Holding Times

In this study, both compounds with different CB loading was indented at holding times of 5, 10, 15, and 20 s, with the results presented accordingly in Table 4 and Table 5. Based on the table, it was found that h_{max} varied at different holding time. This is due to the adhesion in rubbery materials. In compared to metals, the NR possesses a soft and tacky surface that will be attracted to the indenter tips during preloading, resulting in negative load and displacement values at the beginning of test (Chandrashekar *et al*., 2015).

Table 4: The nano-indentation properties of SMR CV-60 at different holding times.

Time (s)	Nano-indentation properties	Carbon loading (phr)			
		0	20	40	60
5	Maximum Force (F_{max}, mN)	1.97±0.03	1.97±0.03	2.00±0.03	2.01±0.02
	Maximum Indentation Depth ($h_{max,}$ μm)	16.55±0.19	13.65±0.56	10.08±1.51	6.22±0.56
	Hardness (H_u , MPa)	0.28±0.01	0.41±0.03	0.78±0.21	2.36±0.94
	Elastic Modulus (E_{it} , MPa)	4.15±0.10	6.87±0.43	14.55±2.93	45.47±3.06
10	Maximum Force (F_{max}, mN)	1.99±0.02	1.99±0.00	1.99±0.00	2.01±0.02
	Maximum Indentation Depth ($h_{max,}$ μm)	18.14±1.18	13.30±0.27	9.85±2.21	6.68±0.28
	Hardness (H_u , MPa)	0.23±0.03	0.43±0.02	0.88±0.21	1.71±0.94
	Elastic Modulus (E_{it} , MPa)	3.61±0.38	7.85±0.22	16.15±6.43	48.93±1.34
15	Maximum Force (F_{max}, mN)	1.98±0.01	1.99±0.00	1.99±0.00	2.00±0.01
	Maximum Indentation Depth ($h_{max,}$ μm)	19.28±1.10	13.50±0.55	10.10±1.20	5.34±0.79
	Hardness (H_u , MPa)	0.20±0.02	0.42±0.03	0.80±0.20	2.23±0.65
	Elastic Modulus (E_{it} , MPa)	3.09±0.31	7.90±0.38	15.12±3.87	56.16±2.04
20	Maximum Force (F_{max}, mN)	1.98±0.02	1.99±0.02	1.99±0.02	2.00±0.01
	Maximum Indentation Depth ($h_{max,}$ μm)	18.90±0.57	12.47±1.10	9.15±1.28	6.13±0.57
	Hardness (H_u , MPa)	0.21±0.01	0.49±0.09	1.02±0.29	2.06±0.42
	Elastic Modulus (E_{it} , MPa)	3.21±0.17	9.08±1.16	18.17±3.98	55.21±3.81

Table 5: The nano-indentation properties of ENR 25 at different holding times.

Time (s)	Nano-indentation properties	Carbon loading (phr)			
		0	20	40	60
5	Maximum Force (F_{max}, mN)	1.99±0.03	1.99±0.02	1.99±0.02	1.99±0.02
	Maximum Indentation Depth ($h_{max,}$ μm)	15.71±0.91	11.51±0.65	8.86±1.14	5.39±0.77
	Hardness (H_u , MPa)	0.31±0.04	0.57±0.07	1.00±0.27	2.71±0.70
	Elastic Modulus (E_{it} , MPa)	5.48±0.71	9.64±0.29	19.18±3.07	54.00±3.68
10	Maximum Force (F_{max}, mN)	1.99±0.03	1.99±0.02	1.98±0.03	1.99±0.03
	Maximum Indentation Depth ($h_{max,}$ μm)	13.89±1.13	11.02±1.06	8.75±0.64	7.38±0.95
	Hardness (H_u , MPa)	0.40±0.06	0.63±0.13	1.00±0.14	1.43±0.36
	Elastic Modulus (E_{it} , MPa)	7.22±1.05	11.67±2.56	21.56±2.59	46.68±7.29
15	Maximum Force (F_{max}, mN)	1.98±0.04	1.98±0.03	1.98±0.03	1.99±0.02
	Maximum Indentation Depth ($h_{max,}$ μm)	15.35±1.12	12.56±1.50	8.89±1.24	5.80±0.66
	Hardness (H_u , MPa)	0.32±0.05	0.47±0.12	0.98±0.25	2.30±0.54
	Elastic Modulus (E_{it} , MPa)	5.88±0.89	8.51±2.57	21.97±4.50	66.74±4.56
20	Maximum Force (F_{max}, mN)	2.00±0.01	2.00±0.01	1.99±0.01	2.00±0.00
	Maximum Indentation Depth ($h_{max,}$ μm)	15.53±0.57	12.92±1.32	8.27±0.42	5.50±0.69
	Hardness (H_u , MPa)	0.34±0.08	0.46±0.09	1.11±0.11	2.58±0.64
	Elastic Modulus (E_{it} , MPa)	6.67±1.07	8.34±1.36	26.17±1.95	84.56±5.29

The relationship between H_u and the different holding times are presented in Figures 4 and 5 for the SMR CV 60 and ENR 25 compounds respectively. Based on the figures, it is found that the H_u values increased as the CB content was increased. However, it is also observed that the H_u values varies with the increase of holding time for both compounds at different CB loadings.

Figure 4 shows that the H_u values of SMR CV-60 with 0 phr of CB loading, was slightly decreased with the increase of holding time. The values ranged from 0.28 MPa to 0.21 MPa. For SMR CV-60 with 20 phr of CB loading, the H_u values fluctuated slightly from 0.41 to 0.49 MPa. The fluctuation differences increased as the CB loading increased. For SMR CV-60 with 40 phr of CB loading, the H_u values ranged from 0.78 to 1.02 MPa, while for SMR CV-60 with 60 phr of CB loading, the H_u values ranged from 1.71 to 2.36 MPa.

As shown in Figure 5, fluctuating trends for H_u were also observed for the ENR 25 compounds with different CB loadings. For ENR 25 with 0 phr of CB loading, the H_u values ranged for 0.31 to 0.40 MPa, while for ENR 25 with 20, 40 and 60 phr of CB loadings, the H_u values ranged from 0.47 to 0.63 MPa, 0.98 to 1.11 MPa, and 1.43 to 2.71 MPa respectively.

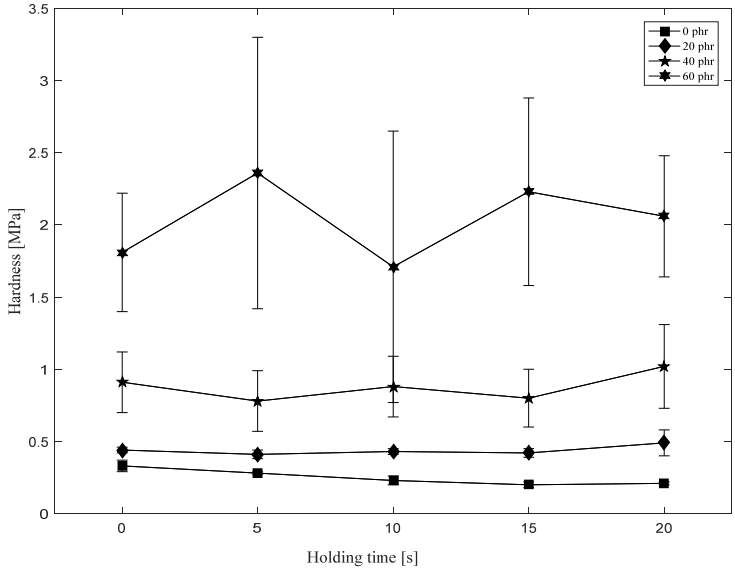


Figure 4: The graph of hardness (H_u) versus the holding time for SMR CV-60.

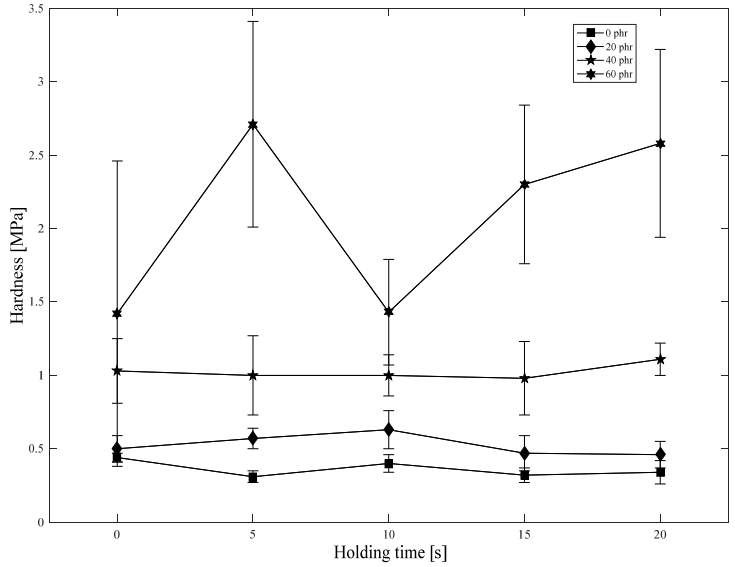


Figure 5: The graph of hardness (H_u)versus the holding time for ENR 25.

The recorded E_{it} values at different holding times were plotted separately in Figures 6 and 7 for SMR CV-60 and ENR 25 at different CB loadings respectively. The E_{it} values for SMR CV-60 at 0 phr CB loading were found to decrease slightly from 5.01 to 3.21 MPa as the holding time increased. In contrast, for SMR CV-60 with the other CB loadings, the E_{it} values were found to increase with the increasing holding time. The E_{it} values for SMR CV-60 with 20 phr CB loading increased slightly from 7.21 to 9.09 MPa. On the other hand, SMR CV-60 with 40 phr and 60 phr CB loadings showed a fluctuating trend for E_{it} as the holding time increased. The E_{it} values for SMR CV-60 with 40 phr CB loading ranged from 14.55 to 18.29MPa, for SMR CV-60 with 60 phr CB loading, it ranged from 42.51 to 56.16 MPa.

The fluctuating trend was also observed for the recorded E_{it} values at different holding times for ENR 25 at different CB loadings. The recorded E_{it} values for ENR 25 showed slight change at various holding times with the range of 5.48 to 8.58 MPa for 0 phr CB loading and 8.33 MPa to 11.67 MPa for 20 phr CB loading. The graph also showed that the E_{it} values for ENR 25 with 40 phr CB loading was significantly higher than ENR 25 with 20 phr CB loading, with the values ranging from 19.18 to 26.17 MPa. As compared to SMR CV-60, ENR 25 with 60 phr CB loading showed a sudden increase in the E_{it} values from 46.68 MPa at 10 s to 66.74 and 84.86 MPa at 15 and 20 s respectively.

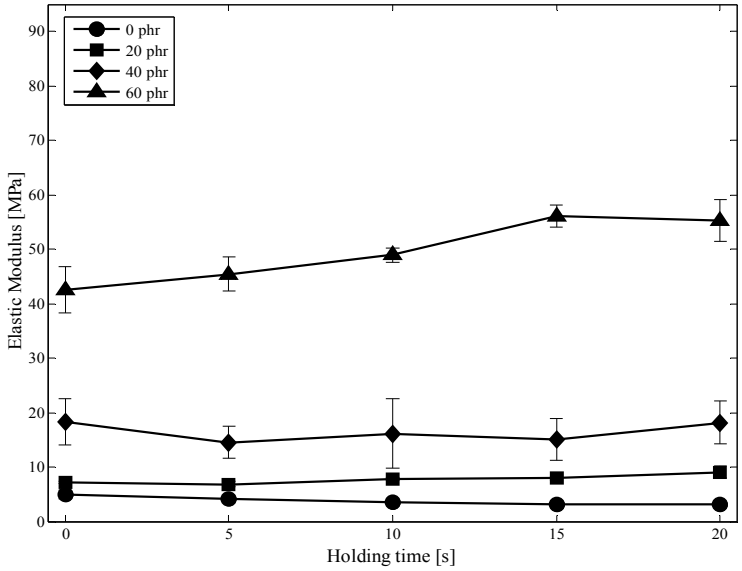


Figure 6: The graph of elastic modulus versus the holding time for SMR CV-60 at different CB loadings.

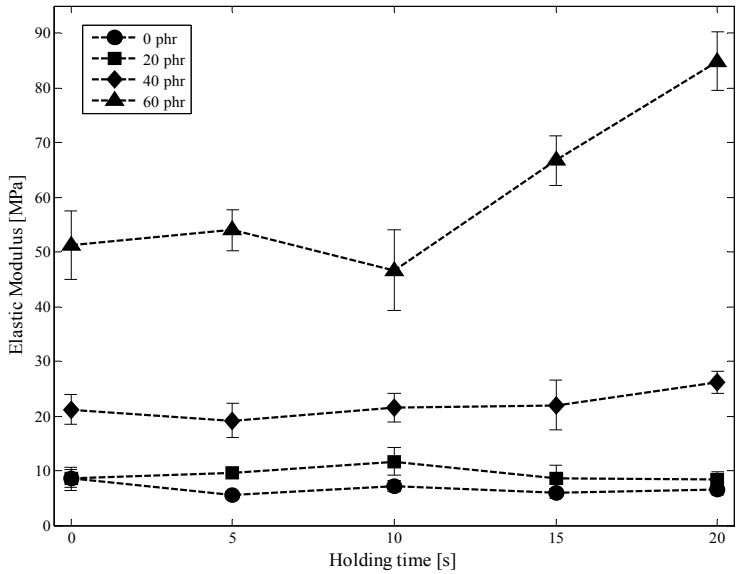


Figure 7: The graph of elastic modulus versus the holding time for ENR 25 at different CB loadings.

Figures 6 and Figure 7 show that the E_{it} values are slightly dependent on the holding time for both NR compounds at 0, 20 and 40 phr CB loadings. In contrast, significant changes were observed the compounds with high CB content (60 phr), where the E_{it} values were found to increase with increasing holding time. As in nano-indentation, the holding time was applied to minimise the creep effect on the unloading curve, which may affect the hardness and elastic modulus reading. The holding time also highly affects the E_{it} for some materials that are indented by the Berkovich tip (Han *et al*., 2016).

Additionally, the fluctuating E_{it} values observed for SMR CV-60 and ENR 25 with lower CB content were caused by experimental uncertainties that affecting the nano-indentation test results, such as machine compliance, thermal drift and NR surface roughness. The standard deviation bar showed in each graph represents the variation of values in the collected data, which was caused by the experimental uncertainties that could directly affect the instrument performance and indentation results. The instrument or machine compliance for calibration of the initial contact between the sample and mounted tips, as well as imprecise machine stiffness value could significantly affect the measured indentation force and penetration depth data. Besides that, the effect of thermal drift, which occurs due to the machine thermal expansion and heat generated by continuously running the

electronic device is quite significant on soft polymer, especially when indented at lower maximum depth and time relaxation (Lin *et al*., 2007; Chandrashekar *et al*., 2015).

3.3 Indentation Work Dissipation

The results for the work dissipation coefficient (μ_{in}) in the indentation of SMR CV-60 and ENR 25 at different CB loadings under different holding times are presented in Tables 6 and 7. μ_{in} can be described as the ratio of indentation work value for the elastic component to the total indentation work (W_{total}). W_{total} is the total of elastic and inelastic work components, which is expressed as in Equation 3.

$$W_{total} = W_{in}^{elastic} + W_{in}^{inelastic} \tag{3}$$

The dissipated work depends not only on H_u and E_{it}, but also on the elastic recovery of material during the unloading process (Recco *et al*., 2009). Based on the tables, at 0 s holding time, the μ_{in} values decreased as the CB loading increased. Besides that, both compounds showed similar behaviour, where fluctuating μ_{in} values were found, except for SMR CV6 with 20 phr CB loading, whereby μ_{in} decreased with increasing holding time. It was also found that the μ_{in} values for SMR CV60 were higher than ENR 25 with similar CB loadings. This behaviour shows the irreversibility of plastic deformation of both rubber compounds and its capability to store elastic energy at different CB loadings and holding times.

Table 6: Work dissipation coefficient (μ_{in}) for SMR CV-60 at different holding times.

Work dissipation coefficient (μ_{in} ,%)						
Carbon loading (phr)		Holding time (s)				
		0	5	10	15	20
SMR CV-60	0	66.77	66.74	59.51	61.26	60.90
	20	58.48	54.78	49.32	47.35	47.29
	40	47.60	47.86	49.93	47.51	45.32
	60	34.98	31.99	28.27	27.64	34.32

Table 7: Work dissipation coefficient (μ_{in}) of ENR 25 at different holding times.

Work dissipation coefficient (μ_{in} ,%)						
Carbon loading (phr)		Holding time (s)				
		0	5	10	15	20
ENR 25	0	52.02	50.17	41.26	45.24	48.56
	20	47.12	38.35	36.20	44.94	43.64
	40	43.67	39.28	36.19	30.69	31.72
	60	29.60	28.93	24.72	24.00	24.43

4. CONCLUSION

The results of the nano-indentation test conducted in this study shows that the properties of the SMR CV60 and ENR 25 compounds were highly affected by the CB loadings. The indentation depth decreased as the CB loadings in the SMR CV-60 and ENR 25 compounds were increased. This shows that the increasing CB loading increased the stiffness of the NR compounds, thus increasing the resistance of the NR compounds from indenter tip penetration. In addition, the hardness and elastic modulus values also increased with the increasing CB loading for both NR compounds. Meanwhile, in the investigation of the effect of holding time on the indentation of the Berkovich tips, it showed that the indentation properties were slightly dependent on the holding time, whereby the penetration depth, hardness and elastic modulus values fluctuated as the holding time increased. The dependency

of holding time was clearest for the SMR CV-60 and ENR 25 compounds with 60 phr CB loading. The work dissipation coefficient study also showed that the capability of both compounds to store elastic energy is highly depended on the CB loadings and holding times.

ACKNOWLEDGEMENTS

The authors are grateful to the Advanced Manufacturing Centre and Fakulti Kejuruteraan Mekanikal, Universiti Teknikal Malaysia Melaka (UTeM) for providing the laboratory facilities used in this study.

REFERENCES

Ab-Malek, K., Ahmadi, H.R., Muhr, A.H., Stephens, I.J., Gough, J., Picken, J.K. & Taib, I.M. (2012). Seismic protection of 2nd Penang crossing using high damping natural rubber isolators. *15th World Conf. Earthquake Eng.,* Lisbon.

Aravind, A., Joy, M. L., & Nair, K. P. (2015). Lubricant properties of biodegradable rubber tree seed (Hevea brasiliensis Muell. Arg) oil. *Ind. Crop. Prod.*, **74**: 14-19. ASTM, D. 3182. (2016). *Standard Practice for Rubber—Materials, Equipment, and Procedures for Mixing Standard Compounds and Preparing Standard Vulcanized Sheets*. Annual Book of ASTM standards. ASTM International, West Conshohocken, Pennsylvania.

ASTM, E. 2546. (2015). *Standard Practice for Instrumented Indentation Testing*. Annual Book of ASTM Standards. ASTM International, West Conshohocken, Pennsylvania.

Atrian, A., Majzoobi, G. H., Nourbakhsh, S. H., Galehdari, S. A. & Nejad, R. M. (2016). Evaluation of tensile strength of Al7075-SiC nanocomposite compacted by gas gun using spherical indentation test and neural networks. *Adv. Powder Technol*, **27**: 1821-1827.

Boonkerd, K. (2017). Development and modification of natural rubber for advanced application. In *Applied Environmental Materials Science for Sustainability*. IGI Global, Pennsylvania, pp. 44-76.

Callister, W.D. & Rethwisch, D.G. (2011). *Materials Science and Engineering: An Introduction*. Wiley, New York.

Chandrashekar, G., Alisafaei, F. & Han, C.S. (2015). Length scale dependent deformation in natural rubber. *J Poly. Sci.*, **132**: 42683

Han, C. S., Sanei, S. H., & Alisafaei, F. (2016). On the origin of indentation size effects and depth dependent mechanical properties of elastic polymers. *J. Polym. Eng.*, **36**: 103-111.

Heide-Jørgensen, S., Møller, R.K., Buhl, K.B., Pedersen, S.U., Daasbjerg, K., Hinge, M. & Budzik, M.K., (2018). Efficient bonding of ethylene-propylene-diene M-class rubber to stainless steel using polymer brushes as a nanoscale adhesive. *Int. J. Adhes. Adhes*, **87**: 31-41.

Kim, W.D., Lee, H.J., Kim, J.Y., & Koh, S.K. (2004). Fatigue life estimation of an engine rubber mount. *Int. J. Fatigue*, **26**: 553-560.

Lin, D.C., Dimitriadis, E.K. & Horkay, F. (2007). Elasticity of rubber-like materials measured by AFM nano-indentation. *Express Polym. Lett.*, **1**: 576-584.

Lindley, P.B. (1964). *Engineering Design with Natural Rubber*. Malayan Rubber Board, Malaysia.

Oliver, W.C. & Pharr, G.M. (1992). An improved technique for determining hardness and elastic modulus using load and displacement sensing indentation experiments. *J. Mater. Res.*, **7**: 1564-1583.

Oyen, M.L. (2007). Sensitivity of polymer nanoindentation creep measurements to experimental variables. *Acta Mater.*, **55**: 3633-3639.

Recco, A.A.C., Viáfara, C.C., Sinatora, A. & Tschiptschin, A.P. (2009). Energy dissipation in depth-sensing indentation as a characteristic of the nanoscratch behavior of coatings. *Wear*, **267**: 1146-1152.

Salim, M. A., Putra, A., & Abdullah, M. A. (2014). Analysis of axial vibration in the laminated rubber-metal spring. *Adv Mater Res.,* **845**: 46-50.

Salim, M. A., Putra, A., Mansor, M.R., Musthafah, M.T., Akop, M.Z. & Abdullah, M.A. (2016). Analysis of parameters assessment on laminated rubber-metal spring for structural vibration. *IOP Conf Ser. Mater. Sci. Eng.*, **(Vol. 114**: 012014.

Salim, M. A., Putra, A., Thompson, D., Ahmad, N. & Abdullah, M.A. (2013). Transmissibility of a laminated rubber-metal spring: A preliminary study. *Appl. Mech. Mater.*, **393**: 661-665.

Sangwichien, C., Sumanatrakool, P. & Patarapaiboolchai, O. (2008). Effect of filler loading on curing characteristics and mechanical properties of thermoplastic vulcanizate. *Chiang Mai J.Sci.*, **35**:141–149.

Thomas, S. & Stephen, R. (Eds.). (2010). *Rubber Nanocomposites: Preparation, Properties, and Applications*. John Wiley & Sons, Singapore.

Xu, F., Ding, Y.H., Deng, X.H., Zhang, P. & Long, Z L. (2014). Indentation size effects in the nano- and micro-hardness of a Fe-based bulk metallic glass. *Physica B: Condensed Matter*, **450**: 84-89.

Yu, Y., Naganathan, N.G., & Dukkipati, R.V. (2001). A literature review of automotive vehicle engine mounting systems. *Mech Mach Theory*, **36**: 123-142.

THE INVESTIGATION OF THE TENSILE AND QUASI-STATIC INDENTATION PROPERTIES OF PINEAPPLE LEAF / KEVLAR FIBRE REINFORCED HYBRID COMPOSITES

Ng Lin Feng[1*], Sivakumar Dhar Malingam[1], Kathiravan Subramaniam[1], Mohd Zulkefli Selamat[1] & Woo Xiu Juan[2]

[1]Centre for Advanced Research on Energy, Fakulti Kejuruteraan Mekanikal
[2]Fakulti Kejuruteraan Elektrik
Universiti Teknikal Malaysia Melaka (UTeM), Malaysia

[*]Email: nglinfeng@yahoo.com

ABSTRACT

Fibre reinforced polymers are the contemporary advanced materials that possess high specific properties when compared to metallic alloys. They have been widely explored since the past few decades. This research study intends to investigate the quasi-static indentation behaviour of non-hybrid and hybrid pineapple leaf / Kevlar fibre reinforced composites with various stacking configurations. Four fibre stacking configurations were fixed in the composite laminates. The composite panels were fabricated through hot moulding compression process. Quasi-static tensile and indentation test were then conducted to measure the tensile properties, energy absorption, maximum indentation force and specific energy absorption of the non-hybrid and hybrid composites. The front and rear fracture surface of the non-hybrid and hybrid composites were analysed after the indentation test. The findings evidenced that the hybrid pineapple leaf / Kevlar fibre reinforced composites showed a positive hybrid effect where the tensile properties, maximum indentation force and energy absorption were drastically improved after partial incorporation of Kevlar fibre in the composites. However, the hybrid composites with the outermost Kevlar fibre and middle pineapple leaf fibre showed comparable tensile and indentation properties to the non-hybrid Kevlar fibre reinforced composites. The results demonstrated the high potential of hybrid pineapple leaf / Kevlar fibre reinforced composites in the replacement of non-hybrid Kevlar fibre reinforced composites.

Keyword: *Hybrid composites; pineapple leaf fibre; Kevlar fibre; indentation properties; energy absorption*

1. INTRODUCTION

A continuous effort has been given to the exploration of the environmental friendly materials with high specific properties. Fibre reinforced polymers (FRPs) are regarded as lightweight materials that possess high specific properties when compared to metallic alloys (Feng *et al*., 2018). However, current FRPs applied in the industrial applications are mainly dominated by synthetic fibres and thermoset matrix. Kevlar fibres are among the high strength synthetic fibres which encompass high impact resistance, stiffness and toughness (Johnson & Venkatesan, 2018). Due to their high impact resistance, Kevlar fibres are commonly employed for military and defence applications such as bulletproof vests, body armours and combat helmets. Despite the contemporary FRPs have high specific properties but such materials have led to the negative impacts on the environment and human health (Mostafa, 2019). One of the possible techniques to resolve the problem is by using natural fibre based composites instead of synthetic fibre based composites in industrial applications.

To date, natural fibre based composites have been employed in aerospace, automotive, electrical and household applications (Sanjay *et al*., 2018). Due to the increasing demand for lightweight composite

materials, it is indispensable to continuously explore the potential of natural fibre based composites. The shift of synthetic fibre towards natural fibre based composites is expected as they can improve renewability, biodegradability, environmental friendliness, and reducing the cost and density of the materials (Afzaluddin *et al*., 2019). Indeed, the natural fibres have offered innumerable virtues such as carbon dioxide neutral, less abrasive, low energy consumption, low cost, biodegradability and lightweight properties (Arthanarieswaran *et al*., 2014; Paul *et al*., 2015; Karahan & Karahan, 2015; Vieira *et al*., 2017; Feng & Malingam, 2019). On this token, pineapple leaf fibre (PALF) has shown a great potential in which PALF possesses a relatively high mechanical strength compared to other natural fibres due to its high cellulose content. Furthermore, PALF is considered as an agricultural waste as their plant is mainly grown for the fruit rather than the fibres. As a result, the incorporation of PALF in the composite materials is beneficial to the environment and the mechanical properties of the composites.

In spite of several attractive features in natural fibres, they also exhibit various shortcomings such as batch-to-batch variation, lack of thermal stability, weak impact resistance, susceptible to strength degradation and high moisture sensitivity (Santulli, 2007; Sgriccia *et al*., 2008; Adekunle *et al*., 2011; Asgarinia *et al*., 2015). The high moisture uptake of the natural fibres eventually leads to poor interfacial bonding with hydrophobic polymer matrices (Kushwaha & Kumar, 2010). Such impediments of natural fibres have retarded their use in structural applications. In fact, the demerits of natural fibres can be resolved using several techniques. Chemical treatments such as silane and alkali treatments are among the most commonly applied techniques to resolve the demerits of natural fibres. Silane treatment provides a siloxane bridge across the fibre-matrix interface, thereby improving the compatibility between fibre and matrix, resulting in the improvement in the mechanical properties. In contrast, alkali treatment alters the fibre surface structure, which removes a certain amount of hemicellulose, lignin and pectin, increasing the number of reactive hydroxyl groups on the fibre surface (Suardana *et al*., 2011; Sullins *et al*., 2017). Thus, the alkali treatment results in the improvement in the aspect ratio of the fibres, leading to the enhancement in the mechanical properties of natural fibre based composites. However, a more efficient and direct method to remedy the shortcomings of natural fibres is via the hybridisation with synthetic fibres (Feng *et al*., 2019a).

Hybrid composites are regarded as the materials that are formed through blending of more than one type of fibres within the polymer matrix. The hybrid composites allow the freedom of tailoring the mechanical properties in accordance with certain industrial applications. Hybrid composites can be grouped into three major classes which are synthetic / synthetic, synthetic / cellulosic and cellulosic / cellulosic fibre based hybrid composites. However, the synthetic / cellulosic fibre reinforced hybrid composites are the most widely explored composite materials. The mechanical properties of synthetic / cellulosic fibre reinforced hybrid composites have been reported in the literature and the findings demonstrated that the partial incorporation of synthetic fibre in the natural fibre based composites improved the mechanical properties (Shahzad, 2011; Ng *et al*., 2017; Kureemun *et al*., 2018; Feng *et al*., 2019b). Through hybridisation of high strength synthetic fibre with low strength natural fibre, it is believed that the energy absorption as well as the indentation properties can be enhanced as well.

Generally, composite materials are susceptible to the indentation loading during their service life (Zhou *et al*., 2017). Therefore, it is particularly important to investigate the indentation properties of the composite materials. Liu *et al*. (2015) studied the temperature effects on the indentation behaviour of carbon fibre reinforced composites with pyramidal truss cores. The findings revealed that the increase of temperature reduced the maximum indentation load and energy absorption of the composite materials due to the degradation of the polymer matrix and the interfacial bonding. Azwan *et al*. (2014) investigated the quasi-static indentation behaviour of composite sandwich structures with glass fibre reinforced polyester as top and bottom layers and polyurethane foam as the core. The indentation test was conducted with several strain rates on the composite materials. They concluded that the increase in strain rate eventually led to an increase in energy absorption of composite materials. Wagih *et al*. (2016) analysed the quasi-static indentation behaviour of carbon fibre reinforced epoxy composites. They found that the indentation damage can be divided into four stages, elastic deformation with no damage, indentation load drop due to matrix cracking, progressive growth

of delamination and finally drastic load drop. A recent research study has been conducted by Bulut & Erkliğ (2018) to investigate the quasi-static indentation behaviour of glass / Kevlar / carbon fibre reinforced epoxy hybrid composites. The results demonstrated that the hybrid composites exhibited the highest indentation force and energy absorption compared to non-hybrid composites. Another recent research study was carried out by Salman *et al*. (2018) to compare the indentation properties of non-hybrid and hybrid kenaf / Kevlar fibre reinforced polyvinyl butyral composite laminates. It was observed that the energy absorption of hybrid composites was superior to those of non-hybrid kenaf fibre reinforced composites. Overall, non-hybrid Kevlar fibre reinforced composites achieved the highest energy absorption. However, it should be emphasised that the energy absorption of hybrid kenaf / Kevlar fibre reinforced composites was comparable to the non-hybrid Kevlar fibre reinforced composites.

Since composite materials are susceptible to the localised impact loading which results in the damage of such materials and thus it is necessary to study the indentation behaviours of the composite materials. The literature studies have shown composite materials exhibited excellent indentation properties and energy absorption. In this case, it is vital to continuously explore the potential of composite materials particularly cellulosic fibre based composites. Thus far, the tensile and indentation properties of thermoplastic based PALF / Kevlar reinforced hybrid composites still remain unexplored. Therefore, this study intends to evaluate the tensile and indentation behaviours of PALF / Kevlar fibre reinforced polypropylene hybrid composites with various fibre stacking configurations. Moreover, the damage properties of the penetrated non-hybrid and hybrid composite laminates are also analysed.

2. METHODOLOGY

2.1 Materials

Woven PALF and Kevlar fabrics as shown in Figure 1(a) and Figure 1(b) were used as reinforcement in the non-hybrid and hybrid composites. PALF fabric with an areal density of 315 g/m^2 was purchased from Mecha Solve Engineering, Malaysia. Kevlar fabric with an areal density of 200 g/m^2 was provided by DuPont Knowledge Center, India. Homopolymer polypropylene (PP) granules were supplied by the Al Waha Petrochemical Company, Saudi Arabia. The properties of PALF and Kevlar fibre are summarised in Table 1.

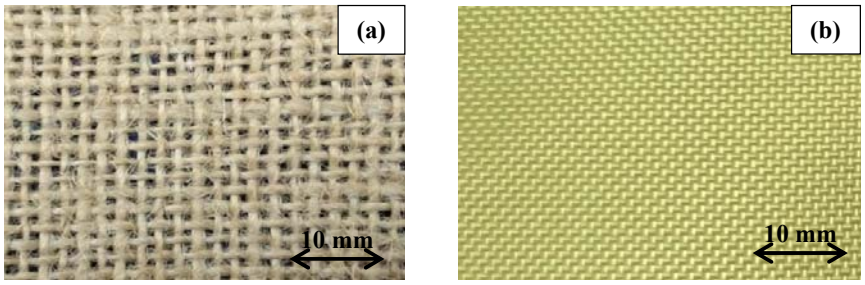


Figure 1: Plain weave woven: (a) PALF (b) Kevlar.

Table 1: Properties of PALF and Kevlar fibre (Ahmad *et al*., 2014; Gurunathan *et al*., 2015).

Properties	PALF	Kevlar fibre
Tensile strength (MPa)	170 – 1627	3000
Tensile modulus (GPa)	60 – 82	124
Strain at break (%)	1 – 3	2.5
Density (g/cm^3)	1.5	1.44

2.2 Fabrication of Composite Laminates

Non-hybrid and hybrid PALF / Kevlar reinforced composite laminates were prepared via hot moulding compression method using Gotech hydraulic hot press machine. Furthermore, film stacking technique was applied to arrange the PALF and Kevlar fabrics as well as PP films before they were compressed in the hot press machine. PP granules were firstly compressed into the form of films prior to the composite fabrication process. Meanwhile, PALF fabric was dried in an oven at a temperature of 80 ºC for 24 hours to eliminate the excessive moisture content to avoid the formation of voids during the composite fabrication. Thereafter, PALF and Kevlar fabrics were stacked alternatively with the PP films in a 3 mm-thick picture frame mould to allow the optimum fibre impregnation. The stack was then subjected to a compression moulding process with a temperature of 175 ºC and pressure of 3.5 MPa. Preheating was applied to the composite laminates before it was fully compressed in order to ensure the heat was evenly distributed throughout the composite laminate. Subsequently, the composite laminate was allowed to cool down until the ambient temperature. Finally, the composite laminate with a nominal thickness of 3 mm was taken out from the hot press machine for visual inspection of any defects.

Non-hybrid and hybrid composites were prepared to study the effects of fibre stacking configurations on the quasi-static tensile and indentation behaviours, including the tensile strength, tensile modulus, maximum indentation load, energy absorbing capacity and specific energy absorption of such materials. Figure 2 depicts the different fibre stacking configurations in the PALF / Kevlar based composite laminates. Each of the non-hybrid and hybrid composite laminates consists of three layers of woven fabric. The non-hybrid PALF and Kevlar fibre reinforced composites are denoted as [PF] and [KV]. When one layer of the middle PALF fabric was replaced with Kevlar fabric in the composite laminate, the hybrid composite is referred to as [H1]. Moreover, the hybrid composite laminate is represented by [H2] when the outermost PALF fabrics were superseded with Kevlar fabrics. The fibre weight and volume fractions along with their respective standard deviations are shown in Table 2. The fibre volume fraction of each composite laminate was calculated with reference to Equation 1.

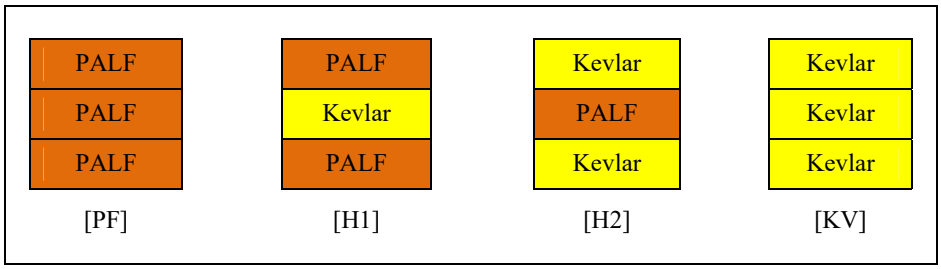


Figure 2: Fibre stacking configurations in composite laminates.

Table 2: Fibre weight and volume fraction in composite laminates.

Fibre stacking configuration	Fibre weight fraction (%)	Fibre volume fraction (%)		
		PALF	Kevlar	Total
[PF]	31.25 (0.85)	21.62 (0.67)	–	21.62 (0.67)
[H1]	28.21 (1.35)	14.62 (0.79)	4.84 (0.26)	19.46 (1.05)
[H2]	26.94 (1.23)	8.02 (0.41)	10.61 (0.54)	18.63 (0.95)
[KV]	22.58 (1.03)	–	14.81 (0.76)	14.81 (0.76)

$$V_{fibre} = \frac{\frac{w_{PALF}}{\rho_{PALF}} + \frac{w_{Kevlar}}{\rho_{Kevlar}}}{\frac{w_{PALF}}{\rho_{PALF}} + \frac{w_{Kevlar}}{\rho_{Kevlar}} + \frac{w_{pp}}{\rho_{pp}}} \quad (1)$$

where w_{PALF} is the weight of PALF, w_{Kevlar} is the weight of Kevlar fibre, w_{pp} is the weight of PP, ρ_{PALF} is the density of PALF, ρ_{Kevlar} is the density of Kevlar fibre and ρ_{pp} is the density of PP.

2.3 Experimental Works

The effects of fibre stacking configurations on the tensile properties of non-hybrid and hybrid PALF / Kevlar based composites were investigated through the quasi-static tensile test. The tensile test was performed with reference to ASTM D3039 at ambient temperature using the Instron model 8872 servo-hydraulic universal testing machine (UTM) with a load cell capacity of 25 kN. A quasi-static cross-head displacement rate of 2 mm/min was fixed throughout the tensile test. Extensometer was equipped on the specimens to monitor the tensile strain. The average tensile strength and modulus were then measured and recorded for further analysis and evaluation.

Quasi-static indentation test was conducted on non-hybrid and hybrid PALF / Kevlar based composites with reference to ASTM D6264 using Instron model 5585 Universal Testing Machine. Maximum indentation force, energy absorption and specific energy absorption were measured during the indentation test. The composite laminates were cut into the dimension of 100 mm x 100 mm and arranged in an edge support configuration as demonstrated in Figure 3. A quasi-static cross-head displacement rate of 1.27 mm/min was fixed throughout the indentation test. Prior to the indentation test, the composite laminate was clamped and tightened between the top and bottom support plates using four screws at the corners to avoid any slippage that affects the accuracy and reliability of the results during the indentation test. An indenter with a 12.7 mm diameter hemispherical tip was used to perform the indentation test. The indentation test was repeated three times for each of the fibre stacking configurations and the average findings were recorded for analysis and evaluation. The results were then represented by force-displacement curves to evaluate the energy absorption and maximum indentation load of the composite laminates. Finally, the damage behaviours of the composite laminates resulted from the indentation force were studied.

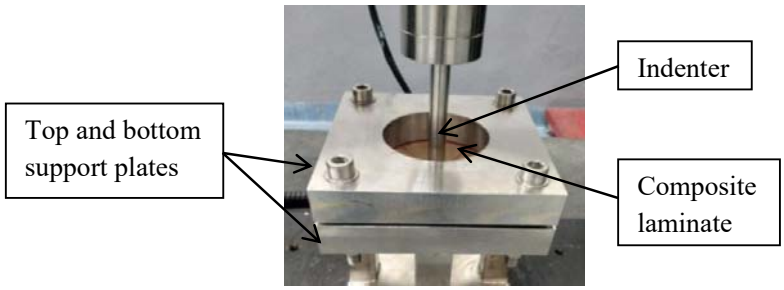


Figure 3: Setup of the quasi-static indentation test.

3. RESULTS AND DISCUSSION

3.1 Tensile Properties

The tensile properties of PALF / Kevlar reinforced composites with different fibre stacking configurations were investigated. The findings obtained from the tensile test were then summarised and recorded as shown in Figure 4. It is undoubtedly that the hybrid PALF / Kevlar composites

improved the tensile properties when compared to non-hybrid PALF based composites. On average, the highest tensile properties were noticed in the non-hybrid Kevlar reinforced composites whereas the lowest tensile properties were obtained in non-hybrid PALF based composites. It was shown that the [KV] composite laminates had tensile strength and modulus of 111.73 MPa and 3.56 GPa, which are respectively 281.20 % and 42.97 % higher than [PF] composite laminates. However, improvements of 33.50 % and 9.64 % were observed in the tensile strength and modulus when one middle PALF fabric was superseded with Kevlar fabric in the [H1] hybrid composite laminates. A further improvement was obtained in [H2] composite laminates when two outermost PALF fabrics were replaced with Kevlar fabrics. It can be seen that those [H2] composite laminates exhibited tensile strength and modulus of 100.85 MPa and 3.19 GPa which are 244.08 % and 28.11 % higher than [PF] composite laminates. From Figure 4, the overall findings demonstrated the incorporation of Kevlar fibre in the hybrid composite laminates indeed attested positive hybrid effect on the tensile properties.

Furthermore, it is worth noting that the [H2] composite laminates possessed comparable tensile properties to the [KV] composite laminates. The tensile strength and modulus of [H2] composite laminates are merely 9.74 % and 10.39 % lower than [KV] composite laminates. This trend showed the high potential of employing hybrid PALF / Kevlar composite laminates, [H2], to substitute those of non-hybrid Kevlar based composite laminates [KV]. Apart from the comparable tensile properties of [H2] composite laminates to those of [KV] composite laminates, it can be noticed that [H1] composites had comparable tensile properties to the [PF] composite laminates as well. These behaviours evidenced that the tensile properties of the materials are highly dependent on the outermost layer in the composite laminates. The outermost layers in the composite laminates act as the main load carriers which attract and sustain more loads during the tensile test. Thus, the skin layers have a decisive effect on the tensile properties of composite laminates. Feng *et al*. (2017) obtained similar results in which the hybrid composite laminates with high strength fabrics as the outermost layers exhibited higher tensile properties.

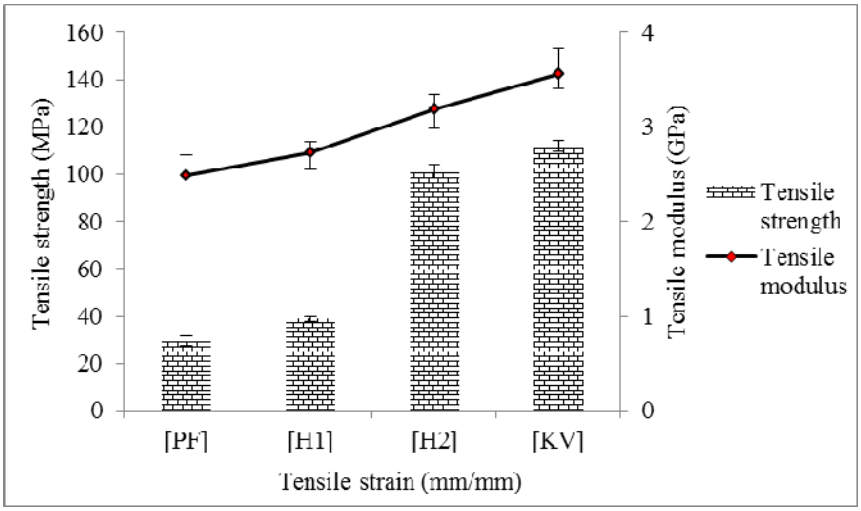


Figure 4: Tensile properties of non-hybrid and hybrid PALF / Kevlar based composites.

3.2 Quasi-Static Indentation Properties

The indentation behaviours of non-hybrid and hybrid PALF / Kevlar fibre reinforced composite laminates were investigated through the quasi-static indentation test. Table 3 records the indentation properties including maximum indentation force, energy absorption and specific energy absorption. The standard deviations of each of the indentation properties are included in the parentheses as well.

Table 3: Indentation properties of non-hybrid and hybrid PALF / Kevlar based composites.

Fibre stacking configurations	Maximum indentation force (N)	Energy absorption (J)	Specific energy absorption ($J.m^2/kg$)
[PF]	866.13 (40.66)	8.59 (0.86)	2.85 (0.08)
[H1]	2654.95 (140.76)	23.40 (3.37)	7.97 (0.38)
[H2]	4280.06 (206.87)	48.31 (2.47)	18.20 (0.83)
[KV]	5819.00 (135.99)	55.75 (4.19)	20.04 (0.96)

Findings obtained from the quasi-static indentation test are represented by the force-displacement curves as depicted in Figure 5. As can be seen in Figure 5, the force-displacement curves of the non-hybrid and hybrid composite laminates showed a very similar trend irrespective of fibre stacking configurations. The trend demonstrated the increase of indentation force along with the increase of displacement up to a maximum indentation force was reached. In fact, the indentation process of the composite laminates can be divided into three major regions. The load increased at the initial stage up to the peak point where the matrix cracking and initial delamination were noticed. After that, fibre breakage together with the higher extent of delamination occurred that resulted in the reduction in the load-carrying capacity of the composite laminates. Finally, the composite laminates were penetrated, that indicates the complete fracture, leading to the friction between the indenter and the composite laminates. The observation is in agreement with the results obtained by Bulut & Erkliğ (2018) in which the non-hybrid and hybrid Kevlar / glass / carbon fibre reinforced composites evidenced the similar trend during the indentation process.

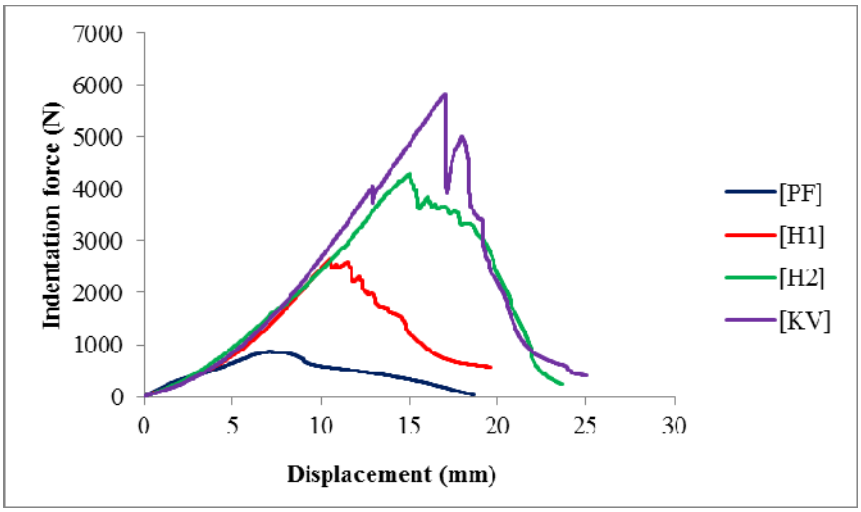


Figure 5: Force-displacement curves of non-hybrid and hybrid PALF / Kevlar based composites.

The energy absorption of the composite laminates was measured by calculating the area under the force-displacement curves. Figure 6 elucidates the maximum indentation force and energy absorption of the non-hybrid and hybrid PALF / Kevlar reinforced composites. Overall, the maximum indentation force and energy absorption were noticed in the non-hybrid Kevlar fibre reinforced composites. On the contrary, the non-hybrid PALF reinforced composites showed the lowest indentation force and energy absorption. The maximum indentation force and energy absorption of non-hybrid Kevlar fibre reinforced composites are 571.84 % and 549.01 % higher than non-hybrid PALF based composites. Nonetheless, the results demonstrated a positive hybrid effect where the maximum indentation force and energy absorption of the composite laminates were significantly

improved when Kevlar fibre was partially incorporated. In comparison with the non-hybrid PALF based composites, the improvements in the maximum indentation force and energy absorption are 206.53 % and 172.41 % when one middle layer of PALF was substituted with Kevlar fibre. Moreover, it was shown that the [H2] composite laminates exhibited maximum indentation force and energy absorption of 4280.06 N and 48.31 J which are 394.16 % and 462.40 % higher than [PF] composite laminates. These trends demonstrated the potential of Kevlar fibre in enhancing the indentation resistance and energy absorption of the hybrid composite laminates.

Nevertheless, it is interesting to emphasise that the [KV] composite laminates evidenced the maximum indentation force and energy absorption which are only 35.96 % and 15.40 % higher than [H2] composite laminates, implying that [H2] composite laminates had comparable indentation properties to the [KV] composite laminates. Since the indentation resistance and energy absorption of the composite laminates are highly dependent on the outermost fabric layers, hence the [H2] and [KV] composite laminates had attested superior indentation properties particularly the energy absorption over those of [PF] and [H1] composite laminates. In fact, the indentation resistance of the composite laminates is governed by the bending stiffness of each fabric layer and the outermost fabric layers are the main constituents that have the decisive effect on the indentation properties. Therefore, the placement of the high strength Kevlar fibre as the skin layers is considered as an alternative fibre stacking configuration which contributes to higher indentation properties.

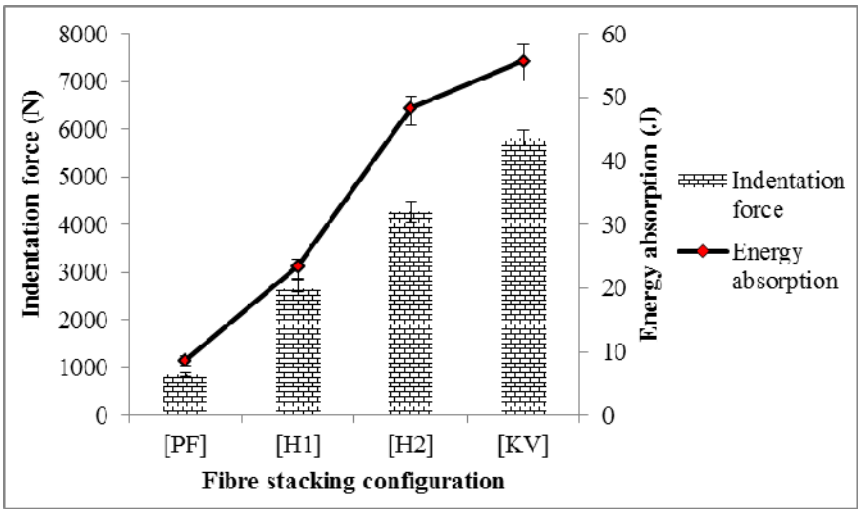


Figure 6: Maximum indentation force and energy absorption of composite laminates under quasi-static indentation.

Lightweight and environmental friendly characteristics are those of criteria which are taken into consideration when searching for an alternative material to supersede synthetic fibre based composites. In this context, it is pivotal to consider the specific properties of non-hybrid and hybrid PALF / Kevlar reinforced composites. Figure 7 shows the specific energy absorption of the composite laminates with different fibre stacking configurations. The specific energy absorption was measured according to the Equation (2) by dividing the total energy absorption by the areal density of the composite laminates.

$$E_{\text{specific}} = \frac{E_{absolute}}{Areal\ density} \tag{2}$$

where $E_{absolute}$ is the total energy absorption.

On average, the trend of the specific energy absorption of the composite laminates as shown in Figure 7 is very similar to the absolute energy absorption. The highest specific energy absorption was noticed in [KV] composite laminates while [PF] composite laminates evidenced the lowest specific energy absorption. The excellent specific energy absorption was still observed in those of composite laminates with the incorporation of high strength Kevlar fabric as the outermost layers. When the areal density of the composite laminates was taken into consideration, [H2] composite laminates still attested the comparable specific energy absorption of 18.20 J.m^2/kg which is only 9.18 % lower than [KV] composite laminates. It is worth noting that the difference in the specific energy absorption of [H2] and [KV] composites was even diminished when compared to their absolute energy absorption. Due to the similar density of PALF and Kevlar fibre, the overall trend of specific energy absorption did not show any significant difference to the absolute energy absorption of PALF / Kevlar based composites.

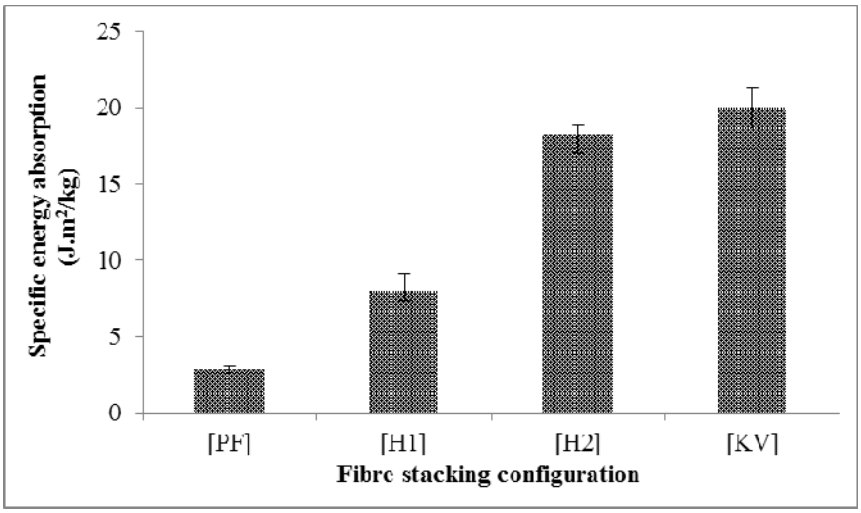


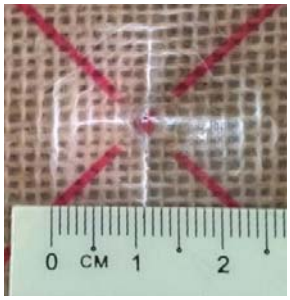
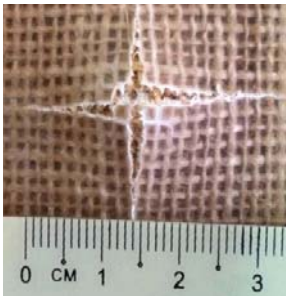
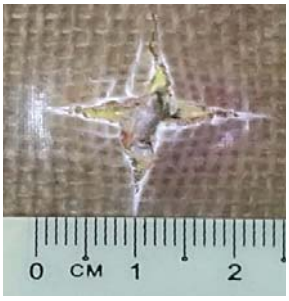
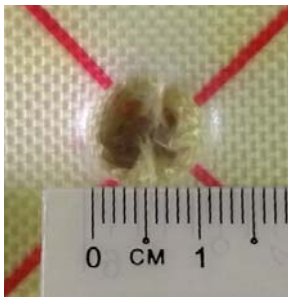
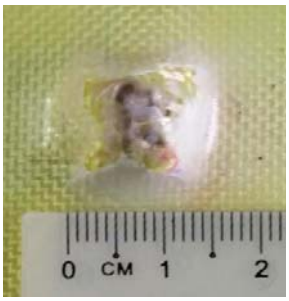
Figure 7: Specific energy absorption of composite laminates subjected to indentation.

3.3 Damage Assessment

The damage assessment of the non-hybrid and hybrid PALF / Kevlar reinforced composites was conducted on the fracture composite laminates to justify the findings obtained from the quasi-static indentation test. The damage behaviours of the non-hybrid and hybrid PALF / Kevlar reinforced composites were evaluated on the front and rear surface of the fracture composite laminates. Table 4 depicts the fracture surface of the front and rear sides of PALF / Kevlar reinforced composites. From Table 4, it was noticed that the damage behaviours of the composite laminates were highly influenced by the fibre types and fibre stacking configurations.

As can be seen in Table 4, the damage was more severe in the PALF based composites in comparison with Kevlar based composites. The crack length of the fracture composite laminate was diminished in both the indented and rear sides when the Kevlar fibre was partially incorporated in the composite laminates. When comparing the damage behaviours of the composite laminates with different fibre stacking configurations, [PF] and [H1] composite laminates exhibited similar damage behaviours while the damage behaviours of [H2] composite laminates were similar to those of [KV] composite laminates. On average, the crack propagation of [H2] and [KV] composite laminates was apparently smaller than the [PF] and [H1] composite laminates. This is attributed to the placement of high strength Kevlar fibre in the outermost layers of the composite laminates, leading to the reduction in the crack propagation of the composite laminates.

Table 4: Fracture surface of PALF / Kevlar reinforced composite laminates.

Composites	Front	Rear
[PF]		
[H1]		
[H2]		
[KV]		

In addition, the composite laminates had evidenced more severe damage on the rear surface instead of the indented surface, implying that the rear surface of the composite laminates was susceptible to higher damage and deformation level during the indentation process. The damage due to the indentation force can be due to the tension-shear and compression-shear. During the indentation test, the damage was initiated with a dent on the indented side, which was then followed by crack initiation

and propagation on the rear surface. The crack propagated continuously along with the increase of the indentation displacement until the complete fracture of the composite laminates. Fibre pull-out, fibre-matrix delamination and fibre breakage were observed in the composite laminates as well, which were shown in Table 4. However, the fibre-matrix delamination was more significant in Kevlar based composites compared to PALF based composites, indicating that PALF based composites encompassed better fibre-matrix adhesion. Nonetheless, it should be emphasised that the incorporation of Kevlar fibre indeed resulted in the global deformation of the composite laminates after subjected to indentation force, resulting in an increase in the energy absorbing capacity.

4. CONCLUSION

This research study intends to investigate the quasi-static tensile and indentation properties of non-hybrid and hybrid PALF / Kevlar reinforced composites with various fibre stacking configurations. According to the findings obtained from the quasi-static indentation test, several conclusions have been drawn:

1. The average findings evidenced the high potential of hybrid composites especially [H2] composite laminates in terms of tensile properties. The results showed the partial replacement of PALF with Kevlar fabrics had undoubtedly improved the tensile properties of the composite laminates, indicating the positive hybrid effect towards the enhanced tensile properties. Moreover, it should be noted that the [H2] composite laminates had shown comparable tensile properties to the [KV] composite laminates. The tensile strength and modulus of [KV] composite laminates are only 10.79 % and 11.60 % higher than [H2] composite laminates. Therefore, [H2] composite laminates had evidenced a high potential to substitute the non-hybrid [KV] laminates.

2. The partial addition of Kevlar fibre in the composite laminates had led to the positive hybrid effect in the indentation properties. The improvement in the indentation properties was noticed when the Kevlar fibre was partially incorporated in the composite laminates. However, the highest indentation force and energy absorption were observed in the [KV] composite laminates whereas [PF] composite laminates showed the lowest indentation properties. Moreover, [H2] hybrid composite laminates had demonstrated a comparable indentation resistance and energy absorption to the [KV] composite laminates. The [H2] composite laminates evidenced the indentation force and energy absorption of 4280.06 N and 48.31 J which are only 26.45 % and 13.35 % lower than [KV] composite laminates.

3. In the context of energy absorption, a similar trend to the indentation force was observed. The partial substitution of PALF with Kevlar had improved the energy absorbing capacity of the composite laminates. When one middle layer of PALF was replaced with Kevlar fibre, the energy absorption was enhanced by 172.41 % while the improvement of 462.40 % was noticed when the outermost layers of PALF were substituted with Kevlar fibre. When comparing the energy absorption of composite laminates with different fibre stacking configurations, composite laminates with Kevlar in the outermost layers attested higher potential in the energy absorbing capacity in comparison with the composite laminates with PALF as the skin layers.

4. Since the specific properties and the environmental friendliness are the main concerns in searching for alternative materials to supersede the synthetic based composites, the specific energy absorption of PALF / Kevlar reinforced composites was measured. In accordance with the results obtained, hybrid composites [H2] still demonstrated a high potential in the energy absorbing capacity to replace those of non-hybrid [KV] composite laminates. [H2] composite laminates still showed comparable energy absorption of 18.20 $J.m^2/kg$ which is merely 9.18 % lower than [KV] composite laminates. The trend of the specific energy absorption is very similar to the absolute energy absorption since the PALF and Kevlar have a similar density. The overall results have concluded the high potential of employing PALF in the hybrid composite laminates in order to achieve the balance in environmental friendliness and the indentation properties.

ACKNOWLEDGEMENT

The authors would like to thank Universiti Teknikal Malaysia Melaka (UTeM) for the continuous support. The authors would also like to express their gratitude to the Skim Zamalah UTeM provided by UTeM.

REFERENCES

Adekunle, K., Cho, S.W., Patzelt, C., Blomfeldt, T. & Skrifvars, M. (2011). Impact and flexural properties of flax fabrics and Lyocell fiber-reinforced bio-based thermoset. *J. Reinf. Plast. Compos*., **30**: 685–697.

Afzaluddin, A., Jawaid, M., Salit, M.S. & Ishak, M.R. (2019). Physical and mechanical properties of sugar palm/glass fiber reinforced thermoplastic polyurethane hybrid composites. *J. Mater. Res. Technol*., **8**: 950–959.

Ahmad, F., Choi, H.S. & Park, M.K. (2014). A review: natural fiber composites selection in view of mechanical, light weight, and economic properties. *Macromol. Mater. Eng*., **300**: 10–24.

Arthanarieswaran, V.P., Kumaravel, A. & Kathirselvam, M. (2014). Evaluation of mechanical properties of banana and sisal fiber reinforced epoxy composites: Influence of glass fiber hybridization. *Mater. Des*., **64**: 194–202.

Asgarinia, S., Viriyasuthee, C., Phillips, S., Dube´, M., Baets, J., Vuure, A.V., Verpoest, I. & Lessard, L. (2015). Tension–tension fatigue behaviour of woven flax/epoxy composites. *J. Reinf. Plast. Compos*., **34**: 857–867.

Azwan, S., Abdi, B., Yahya, M.Y. & Ayob, A. (2014). Quasi-static flexural and indentation behavior of glass fiber reinforced polymer composite sandwich panel. *Adv. Mater. Res*., **845**: 320–323.

Bulut, M. & Erkliğ, A. (2018). The investigation of quasi-static indentation effect on laminated hybrid composite plates. *Mech. Mater*., **117**: 225–234.

Feng, N.L. & Malingam, S.D. (2019). Monotonic and fatigue responses of fiber-reinforced metal laminates. *In* Jawaid, M., Thariq, M. and Saba, N. (Eds.), *Mechanical and Physical Testing of Biocomposites, Fibre-Reinforced Composites and Hybrid Composites*. Woodhead Publishing, Duxford, United Kingdom, PP. 307–323.

Feng, N.L., Malingam, S.D., Jenal, R., Mustafa, Z. & Subramonian, S. (2018). A review of the tensile and fatigue responses of cellulosic fibre-reinforced polymer composites. *Mech. Adv. Mater. Struct*., 1–16. doi:10.1080/15376494.2018.1489086

Feng, N.L., Malingam, S.D., Subramaniam, K., Selamat, M.Z., Ali, M.B. & Bapokutty, O. (2019a). The influence of fibre stacking configurations on the indentation behaviour of pineapple leaf / glass fibre reinforced hybrid composites. *Def. S&T Tech. Bull*., **12**: 103–113.

Feng, N.L., DharMalingam, S., Zakaria, K.A. & Selamat, M.Z. (2019b). Investigation on the fatigue life characteristic of kenaf/glass woven-ply reinforced metal sandwich materials. *J. Sandw. Struct. Mater*., **21**: 2440–2455.

Gurunathan, T., Mohanty, S. & Nayak, S.K. (2015). A review of the recent developments in biocomposites based on natural fibres and their application perspectives. *Compos. Part A Appl. Sci. Manuf*., **77**: 1–25.

Johnson, H.M. & Venkatesan, S.P. (2018). Impact resistance analysis of hybrid composite material for aircraft fuselage. *Int. J. Eng. Sci. Comput*., **8**: 16712–16715.

Karahan, M. & Karahan, N. (2015). Investigation of the tensile properties ofnatural and natural/synthetic hybrid fiber woven fabric composites. *J. Reinf. Plast. Compos*., **34**: 795–806.

Kureemun, U., Ravandi, M., Tran, L.Q.N., Teo, W.S., Tay, T.E. & Lee, H.P. (2018). Effects of hybridization and hybrid fibre dispersion on the mechanical properties of woven flax-carbon epoxy at low carbon fibre volume fractions. *Compos. Part B Eng*., **134**: 28–38.

Kushwaha, P.K. & Kumar, R. (2010). Effect of silanes on mechanical properties of bamboo fiber-epoxy composites. *J. Reinf. Plast. Compos*., **29**: 718–724.

Liu, J.,Qiao, W., Liu, J., Xie, D., Zhou, Z., Wu, L. & Ma, L. (2015). High temperature indentation behaviors of carbon fiber composite pyramidal truss structures. *Compos. Struct.*, **131**: 266–272.

Mostafa, N.H. (2019). Tensile and fatigue properties of jute-glass hybrid fibre reinforced epoxy composites. *Mater. Res. Express*., **6**: 085102.

Ng, L.F., Sivakumar, D., Zakaria, K.A., Bapokutty, O. & Sivaraos. (2017). Influence of kenaf fibre orientation effect on the mechanical properties of hybrid structure of fibre metal laminate. *Pertanika J. Sci. Tech*., **25**: 1–8.

Paul, V., Kanny, K. & Redhi, G.G. (2015). Mechanical, thermal and morphological properties of a bio-based composite derived from banana plant source. *Compos. Part A Appl. Sci. Manuf*., **68**: 90–100.

Salman, S.D., Leman, Z., Ishak, M., Sultan, M. & Cardona, F. (2018). Quasi-static penetration behavior of plain woven kenaf/aramid reinforced polyvinyl butyral hybrid laminates. *J. Ind. Text*., **47**: 1427–1446.

Sanjay, M.R., Madhu, P., Jawaid, M., Pradeep, S., Senthamaraikannan, P. & Senthil, S. (2018). Characterization and properties of natural fiber polymer composites: a comprehensive review. *J. Clean Prod*., **172**: 566–581.

Santulli, C. (2007). Impact properties of glass/plant fibre hybrid laminates. *J. Mater. Sci*., **42**: 3699–3707.

Sgriccia, N., Hawley, M.C. & Misra, M. (2008). Characterization of natural fiber surfaces and natural fiber composites. *Compos. Part A Appl. Sci. Manuf*., **39**: 1632–1637.

Shahzad, A. (2011). Impact and fatigue properties of hemp–glass fiber hybrid biocomposites. *J. Reinf. Plast. Compos*., **30**: 1389–1398.

Sullins, T., Pillay, S., Komus, A. & Ning, H. (2017). Hemp fiber reinforced polypropylene composites: The effects of material treatments. *Compos. Part B Eng*., **114**: 15–22.

Suardana, N.P.G., Piao, Y. & Lim, J.K. (2011). Mechanical properties of hemp fibers and hemp/pp composites: Effects of chemical surface treatment. *Mater. Phys. Mech*., **11**: 1–8.

Vieira, L.M.G., Santos, J.C., Panzera, T.H., Rubio, J.C.C. & Scarpa, F. (2017). Novel fibre metal laminate sandwich composite structure with sisal woven core. *Ind. Crops Prod*., **99**: 189–195.

Wagih, A., Maimí, P., Blanco, N. & Costa, J. (2016). A quasi-static indentation test to elucidate the sequence of damage events in low velocity impacts on composite laminates. *Compos. Part A Appl. Sci. Manuf*., **82**: 180–189.

Zhou, G., Zhang, B. & Pasricha, A. (2017). A study of indentation behaviour of sandwich panels supported rigidly. *Int. J. Struct. Integr*., **8**: 439–451.

INDOOR AND OUTDOOR BIOAEROSOL SAMPLING AND BACTERIAL COUNTING ANALYSIS

Nik Nur Ilyani Mohamed Nazri[1*], Ann Nurrizka Abd. Hamid[2], Nur Amira Aminuddin[2], Asmariah Jusoh[1] & Noor Hafifi Zuraini Abdul Rahim[1]

[1]Biosurveillance and Biological Defence Branch, Protection and Biophysical Technology Division (BTPB), Science and Technology Research Institute for Defence (STRIDE), Ministry of Defence, Malaysia
[2]School of Biology, Faculty of Applied Sciences, Universiti Teknologi MARA (UiTM), Malaysia

[*]Email: nikilyani.nazri@stride.gov.my

ABSTRACT

Bioaerosol concentrations are significantly affected by environmental conditions. Therefore, bioaerosol sampling is important to monitor and control air quality, in particular to control airborne diseases. This study was conducted in August 2019 to investigate the air quality in a government premise in Selangor, Malaysia. The sampling sites were selected randomly from the same building. One of the samples was collected from a garden as outdoors sample, while others were collected indoors, which included the pantry, toilet, staff rooms, store rooms, meeting room, prayer room, exhibition hall and training room. The eleven sampling sites differ from each other in terms of the presence of personnel, windows, air conditioning system, water source, organic substrates and flooring materials. For bioaerosol sampling, the Andersen impaction method was applied using Nutrient Agar (NA) non-selective media plates. Three units of biological air samplers (SKC Quick Take 30, USA) with setting of 2 min sampling time and volume flow rate of 28.3 L/min were used in each sampling location. The plates were incubated at 35 ºC for 24 h and underwent the plate count process using an automatic colony counter (Interscience Scan 4000, USA). The total number of bacterial colonies for 24 and 48 h were counted and recorded before the bacterial concentrations were calculated using the concentration of biological contaminants calculation formula. The results showed that all the indoor and outdoor locations were still at healthy levels and do not exceed the maximum limit of 1,000 CFU/m^3 for the total number of bioaerosol particles that has been recommended by the National Institute of Occupational Safety and Health (NIOSH) and American Conference of Governmental Industrial Hygienists (ACGIH). However, the staff room exceeded the culture count for total bacteria of 500 CFU/m^3 that was recommended by ACGIH. The higher bacterial concentrations in the staff room and prayer room than the garden could be influenced by several factors including presence of personnel, air ventilation, flooring material, water sources and organic substrates. In conclusion, the airborne bacteria presence could possibly originate from outdoor sources, with textile and humans as the main vectors of the contamination in both locations, especially during the haze period.

Keywords: Bioaerosol; indoor and outdoor environments; biological air sampler; automatic colony counter; haze period.

1. INTRODUCTION

Bioaerosols are airborne particles that originate from living organisms, such as viruses, intact bacterial cells and spores, protozoa and their cysts, fungal cells and spores, and plant pollen grains and spores. Bioaerosols have been shown to cause health problems related to indoor air quality (Schwab *et al.,* 2003). Approximately 5 to 34% of air contamination is caused by bioaerosol particles (Gizaw *et al.,* 2016), which could also lead to the declination of indoor air quality. Air sampling is a popular method of conducting microbial examinations, as it allows for direct toxicological evaluation with results that

can be related to a concentration that is expressed in terms of colony forming units per cubic meter (CFU/m^3) (Yassin & Almouqatea, 2010). Furthermore, monitoring of environmental factors, such as temperature and relative humidity, can be a useful tool to explain possible bioaerosol sources (Stetzenbach *et al*., 2004). Since bacteria require specific environmental conditions to grow and propagate, their concentrations are significantly affected by these factors (Mouli *et al*., 2005). Therefore, bioaerosol sampling is important to monitor and control air quality, in particular to control airborne diseases.

Many studies have been conducted to investigate the relationship between bioaerosols as one of the air components and human infectious diseases (Douwes *et al.,* 2003; WHO, 2009). A number of airborne bioaerosol assessments have been conducted by researchers over these past two decades, whereby it has been reported that the concentration of pollutants vary between indoor and outdoor environments (Pastuszka *et al.,* 2000, 2005; Bernstein *et al.,* 2008; Aydogdu *et al.,* 2010; Nasir & Colbeck, 2010; Pegas *et al.,* 2010; Salleh *et al.,* 2011; Dumała & Dudzińska, 2013; Moon *et al.,* 2014; Karottki *et al.,* 2015). It is not clear which specific components primarily account for the presumed health effects. Dose-response relationships have often not been described and knowledge about threshold values is also not available. This relative lack of knowledge is mainly due to the lack of valid quantitative exposure assessment methods (Douwes *et al.,* 2003). The information of the dispersion size of particles is also limited (Latif *et al.,* 2014). Bioaerosols differ from aerosols based on their biological characteristics and a detailed study on controlling the dispersion of the particles is required (Lee, 2011).

Air contamination can occur if the air components are disturbed and imbalance. Hence, solid, fluid or gas air contaminants would not just affect human health and quality of life, but also would affect animals and plants too. Its effects could be either direct or indirect because the air flow in the Earth's atmosphere is very turbulent and dynamic. The direct effect can be seen in declination of human health level, while the indirect effect could pollute, spoil and damage infrastructure, especially building structure (Rao *et al.,* 1996).

Many would have thought that indoor air is safer and cleaner than outdoor air. Outdoors is assumed to be more polluted because of its wide, open and unlimited space. Besides that, there are a lot of possible sources of air contamination, such as from vehicles, industrial areas, open burning, cigarette smoke, as well as from natural phenomenon such as drought, acid rain and volcano eruption. However, according to the United State Environmental Protection Agency (EPA) (1987), indoor levels of pollutants may be up to 100 times higher than outdoor pollutant levels. Indoor contamination has been ranked as among five top public health risks (Kotzias, 2005; Gawrońska & Bakera, 2015). Poor air quality is often related to the factors of heart and lung diseases (Shin *et al*., 2015). Indoor air quality is particularly important to humans since we spend up to 90% of time indoors (Klepeis, 2001).

Common sampling techniques such as impactors, filters, impingers and cyclones are used to separate and collect bioaerosols (Macher *et al*., 1995; Willeke & Macher, 1999; Haig *et al*., 2016). In microbiological research, after air sampling, it is essential to accurately enumerate microbial cells. For example, as a key step in developing new antimicrobial agents, the determination of minimum inhibitory concentration (MIC) requires inoculation of a precise number of viable microorganisms (Balouiri *et al*., 2016). The microbial concentration calculation can be done by direct microscopic count (Brock & Bohlool, 1974), flow cytometry (Muirhead *et al.,* 1985; Porter *et al.,* 1997) and most probable number counting assay (MPN) methods (Makkar & Casida, 1987). The agar plate method is inexpensive and the procedures are simple. However, it takes a long time for cells to be cultured and to obtain the laborious manual counting results, which could take up to a few days (Hazan *et al.,* 2012; Davis, 2014). Therefore, to reduce manual work and its related human error, commercial automated colony counters, such as ProtoCOLTM (Synbiosis, UK), EC2TM (BioMérieux, Marcy-l0Étoile, France) and Scan TM 500 (Interscience, Woburn, MA, USA), have been developed through applications of image processing algorithms (Song *et al*., 2018).

Several advantages of using an automated colony counter system are: it increases sample throughput; it can count colonies of different colours simultaneously so that counts can be less time-consuming; it can improve data accuracy as count data are automatically transferred onto digital files; and plate count data can be checked at a later date since automated counting systems use electronic data files that allow storage of plate images and numerical results. Moreover, its performance is superior to routine manual counting of plates especially in the presence of higher numbers of colonies (Brugger *et al.*, 2012).

After the total number of bacteria is obtained and recorded, the bacterial concentration can be determined using the concentration of biological contaminants calculation formula:

$$\frac{\text{Colony Forming Unit (CFU)} \times 1000}{\text{Flow Rate (L/min)} \times \text{Duration (min)}} \quad (1)$$

For indoor air quality assessments, concentrations measured in test environments are typically compared to baseline data from reference areas or data reported in the literature (Rao *et. al*, 1996). Reponen *et al.* (1992) recommended using the extreme of the data distributions culturable bacteria and fungi (5,000 and 500 CFU/m^3 respectively) as indicators of the presence of abnormal indoor sources or insufficient ventilation in urban and suburban residences in a subarctic climate (e.g., Finland). However, the general consensus is that it is not possible to set numeric concentration limits for bacteria or fungi in indoor environments (Heikkinen *et al.*, 2005).

Although indoor environments are considered to be protective, they can become contaminated with particles that present different and sometimes more serious risks than those related to outdoor exposures, when their concentrations exceed recommended maximum limits. These are 1,000 CFUs/m^3 for total number of bioaerosol particles set by the National Institute of Occupational Safety and Health (NIOSH) and American Conference of Governmental Industrial Hygienists (ACGIH) with the culturable count for total bacteria not to exceed 500 CFUs/m^3 (Cox & Wathes, 1995; Jensen & Schafer, 1998).

The objective of this study is to determine indoor and outdoor bioaerosol concentration levels in selected locations using the impaction air sampling method and automatic colony counter for bacterial concentration plate count. In addition, the relation of the bioaerosol concentration levels with the sources of airborne microorganisms in the built environment will also be studied.

2. MATERIALS AND METHODS

2.1 Preparation of Nutrient Agar (NA) Medium

A total of 10 g of NA powder was weighed and dissolved in 50 ml distilled water in a conical flask. The solution was stirred using magnetic stirrer (700 rpm) and heated up on a hotplate (130 °C) until the murk was almost cleared. The solution was poured into an autoclave bottle and autoclaved at 121 °C for 20 min. A volume of 15 ml of the autoclaved medium solution was poured into each sterile petri dish until the bottle was empty. The agar medium was left to dry at room temperature in laminar air flow before placing the stock in a refrigerator at 4 °C. The medium was pre-warmed to room temperature before being used in the bioaerosol sampling.

2.2 Air Sampling

The sampling sites were selected randomly from the same building of a government premise in Selangor, Malaysia. One of the samples was collected outdoors (garden), while the others were collected indoors, which included the pantry, toilet, staff rooms, store rooms, meeting room, prayer room, exhibition hall and training room. The 11 sampling sites differ from each other in terms of the

presence of personnel, windows, air conditioning system, water source, organic substrates and flooring materials. The required equipment were prepared and sterilised to avoid sample contamination. The Andersen impaction method (Andersen, 1958) was applied using NA non-selective media plates. All measurements were performed in triplicate. Three units of SKC biological air sampler (SKC Quick Take 30, USA) with setting of 2 min sampling time and volume flow rate of 28.3 L/min were used in each sampling site. All the sampling plates were incubated immediately at 35 °C for 24 h.

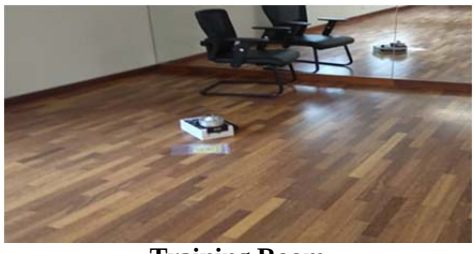
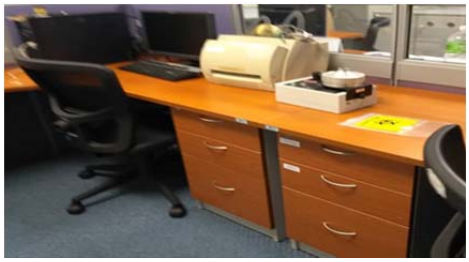
2.3 Bacterial Enumeration

The sampling plates were taken out from the incubator every 24 and 48 h for the plate count process using an automatic colony counter (Interscience Scan 4000, USA). The total number of colonies for 24 and 48 h incubation for each sampling plate was counted and recorded. After the plate count process for 48 h, all the sampling plates were sealed with parafilm and stored back in the refrigerator at 4 °C. The bacterial concentration for each location was determined using the concentration biological contaminants calculation formula.

3. RESULTS AND DISCUSSION

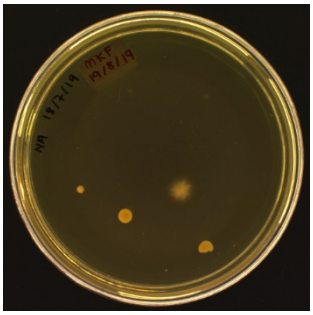
The characteristics for each sampling sites are described in Table 1, which could possibly contribute to the presence of airborne bacteria at the sampling sites. The images of the sampling plates captured using the automatic colony counter for both indoor and outdoor environments are shown in Figure 1. Figure 2 shows the bacterial concentration at every sampling site after 24 and 48 h of incubation.

Table 1: Description of characteristics of each sampling site.

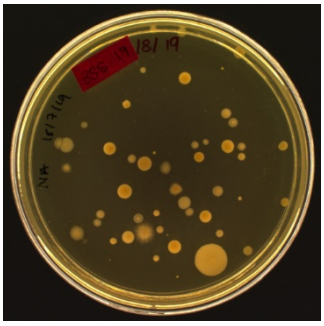
Sampling Site	Characteristics of Sampling Site
 **Training Room**	• Air conditioning • Wood flooring material • Water source • Low human presence • Windows
 **Staff Room**	• Air conditioning • Carpet flooring material • High human presence • Organic substrate source (e.g., food) • Windows

 **Administration Room**	• Air conditioning • Carpet flooring material • High human presence • Windows
 **Pantry**	• Air conditioning • Tile flooring material • Water source • High human presence • Organic substrate source (e.g., food) • Windows
 **Prayer Room**	• Air conditioning • Wood flooring material • High human presence • Textiles • Windows
 **Meeting Room**	• Air conditioning • Carpet flooring material • Moderate human presence
 **Store Room**	• Tile flooring material • Low human presence • Textiles

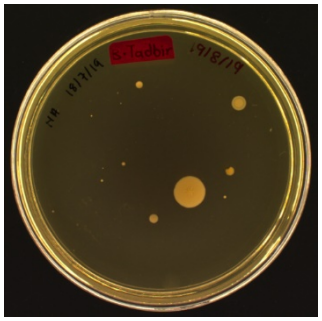
 **File Room**	• Air conditioning • Tile flooring material • Low rate of human occupancy
 **Toilet**	• Tile flooring material • Plants • Moderate human presence • Water sources • Windows
 **Exhibition Hall**	• Air conditioning • Tile flooring material • Low human presence • Organic substrate source (e.g., food) • Textiles • Windows
 **Garden**	• Open space • Plants • Moderate human presence • Organic substrate source (e.g., soil) • Water source



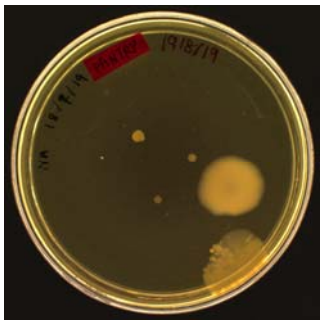
Training Room



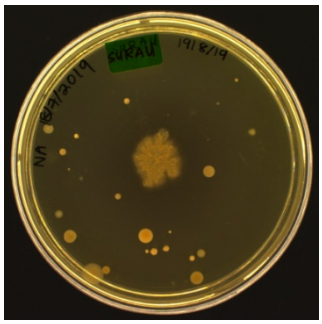
Staff Room



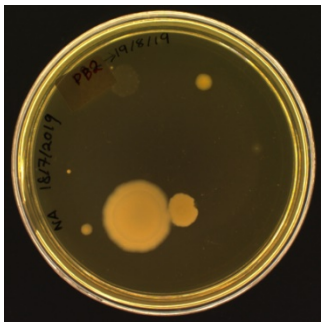
Administration Room



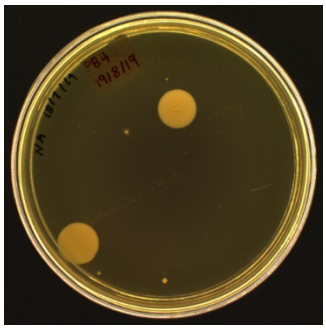
Pantry



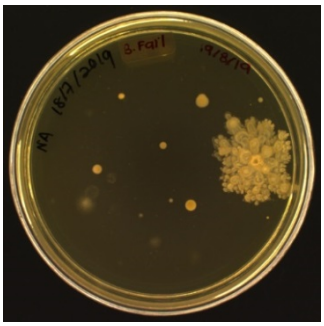
Prayer Room



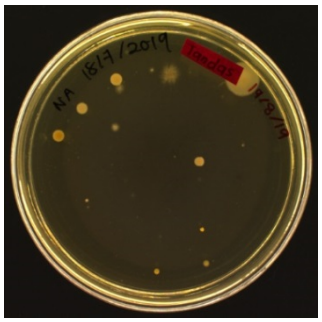
Meeting Room



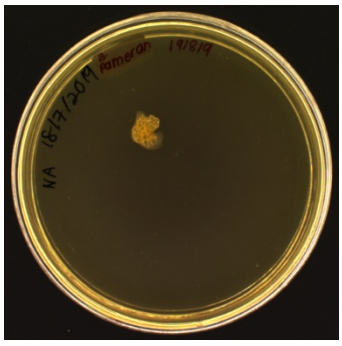
Staff Room



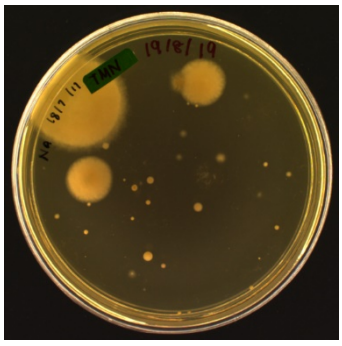
File Room



Toilet



Exhibition Hall



Garden

Figure 1: Images of sampling plates for the various locations.

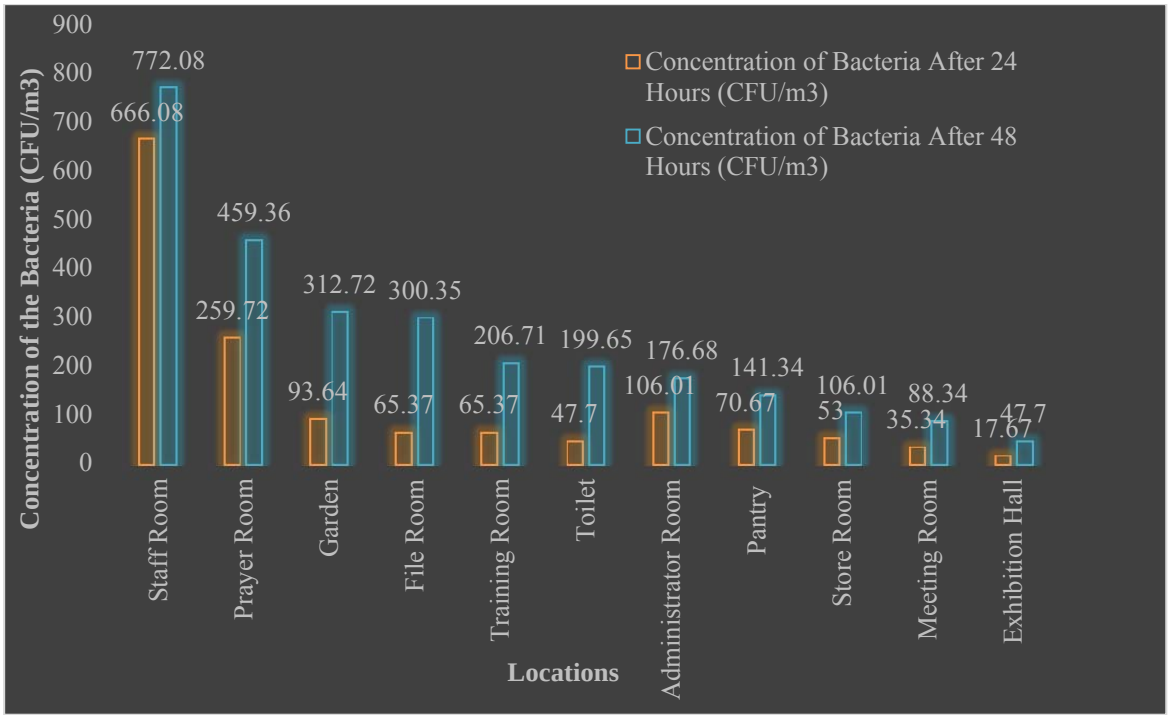


Figure 2: Bacterial Concentration at each sampling site after 24 and 48 h of incubation.

The results show that all the indoor and outdoor locations had under 1,000 CFU/m^3 of total number of bioaerosol particles that has been recommended by NIOSH and ACGIH. However, the staff room exceeded the culture count for total bacteria of 500 CFU/m^3, which is recommended by ACGIH.
It was also found that the staff room and prayer room had higher bacterial concentrations than the garden (the outdoor location), which could be explained by several factors.

Air contamination in a building is normally caused by the building itself or outdoor sources (EPA, 1987). In indoor environments, humidity, temperature and availability of organic substrate could promote the growth of microbial (Lamoth & Greub, 2010). Air conditioning (AC) systems is one of the technical solutions to improve air quality and provide a comfortable workplace environment indoors. AC systems could reduce up to 80% of the pollutant rate from the air. However, without regular maintenance and cleaning, it could promote the growth of many bioaerosol agents (Law *et al.,* 2001). Therefore, building users have some degree of control over AC systems as a source of airborne microorganisms (Prussin *et al.,* 2015).

The water source in the prayer room contributes to high bacterial concentration since bacteria could also be transmitted through water, in addition to transmission of bacteria through the air, which comes from the opened windows. Bacteria could also be transferred indoors by dust particles as the main vector of the contamination (Di *et al.,* 2010; Hewitt *et al.,* 2012; Nazaroff, 2016). In addition, during the air sampling process, there was a haze crisis that occurred in Malaysia from mid-July to late September 2019, which could have contributed to the high bacterial concentration in the indoor and outdoor environments. It was reported that the haze was caused by intense forest fires that had raged across the Indonesian regions. Haze contains dust and smoke particles that could affect human health. Some studies have found that haze periods are associated with high bioaerosol concentrations (Li *et al.,* 2015; Wei *et al.,* 2016; Li *et al.,* 2017; Xie *et al*., 2018). Bioaerosols can live well in the air with nutrients supported by cloud, rainwater and particles (Womack *et al.,* 2010). Even in harsh atmospheric hard conditions, such as high solar radiation, low moisture and nutrients, and large dispersing capability, the atmosphere can gradually evolve into one of the habitats for microbes and thus, there is a large amount of microbes in the air (Henderson & Salem, 2016).

The sampling sites with high human presence had high numbers of bacterial concentrations, including the staff room and prayer room. Humans are one of the primary sources of bioaerosols in built environments (Qian *et al*., 2012; Adams *et al*., 2013), whereby the main source of airborne contamination from outdoors into indoors is through transfer from human activities (Koistinen *et el*., 2004; Kopperud, 2004). Airborne bacteria, especially during haze periods, could attach to human clothes and be brought into indoor locations.

In addition, skin transient flora also plays an important role as a source of bioaerosol contamination. Transient flora could replicate and contaminate the air by the shedding of skin or cleansing routine such as washing hands (Noble *et al.,* 1976; Mackintosh *et al.,* 1978; Fox *et al.,* 2010). Bacteria in human bodies could also contaminate the environment by mucous and respiratory fluid secreted orally, such as through sneezing, coughing, talking and breathing (Klepeis *et. al.,* 2001; Brasche & Bischof, 2005). These bacteria might be trapped on the carpet in the staff room and wood surfaces in the prayer room since people keep coming into the room and stay in the room for long periods. Moreover, sweat and respiration from human bodies could be substrate sources for the microorganisms (Lamoth & Greub, 2010). Human skin, oral cavity and nasal passage are the common habitats of bacteria in indoor environments (Bouillard *et al.,* 2005; Prussin *et al*., 2015; Hayleeyesus & Manaye, 2014; Brągoszewska *et al.*, 2018).

While the floor in the staff room was fully covered by carpet, textiles such as prayer clothes and mats were provided in the prayer room, which could act as potential surfaces for microbes to move from one place to another and even reproduce on it. Almost all natural textile materials and clothing made of cotton, wool, silk, etc., are known to be susceptible to microbial attacks, as these fabrics provide and offer large surface areas and absorb moisture, thus providing a suitable atmosphere for microbial growth and reproduction (Cardamone, 2002). Natural fibres have protein (keratin) and cellulose, which provide basic requirements such as moisture, oxygen, nutrients and temperature for bacterial growth and multiplication. When in contact with human bodies, they provide an ideal environment for the growth and multiplication of pathogenic microbes, giving rise to objectionable odour, dermal infection, allergic responses and other related diseases (Thiry, 2001).

4. CONCLUSION

From the results given by the automatic colony counter, the staff room and prayer room have higher bacterial concentration as compared to the garden. These high bacterial concentrations could be affected by various factors that caused the bacterial growth or transmission into these rooms, such as presence of personnel, air ventilation, flooring material, water sources and organic substrates. The airborne bacteria presence could possibly originate from outdoor sources, with textile and humans as the main vectors of the contamination in both locations, especially during the haze period.

Even though the total number of bioaerosol particles at all the indoor and outdoor locations were still at healthy levels and did not exceed the maximum level of 1,000 CFU/m^3 that has been recommended by NIOSH and ACGIH, attention should be given to control the environmental factors that favour the growth and multiplication of microbes in indoor environments to safeguard the health of personnel, especially in the staff room which has exceeded the culture count for total bacteria of 500 CFU/m^3 that has been recommended by ACGIH. It is highly recommended to conduct further investigation in the future on why there were high plate count results at certain locations in the building, and also to identify the genus and species of airborne bacteria from the bioaerosol samples in order to determine whether the bacteria present in the air at the locations are pathogenic to humans. This would allow for appropriate further action to be taken after the sample identification is done.

REFERENCES

ACGIH (1989). *Guidelines for the Assessment of Bioaerosols in the Indoor Environment*. American Conference of Governmental Industrial Hygienists (ACGIH), Cincinnati, Ohio.

Adams, R.I., Miletto, M., Taylor, J.W. & Bruns, T.D. (2013). Dispersal in microbes: fungi in indoor air are dominated by outdoor air and show dispersal limitation at short distances. *ISME J.,* **7**: 1262-1273.

Andersen, A.A. (1958). New sampler for the collection, sizing, and enumeration of viable airborne particles. *J. Bacteriol.*, **76**: 471.

Aydogdu, H., Asan, A. & Tatman, O.M. (2010). Indoor and outdoor airborne bacteria in child day-care centers in Edirne City (Turkey), seasonal distribution and influence of meteorological factors. *Environ. Monit. Assess*., **164**: 53-66.

Balouiri, M., Sadiki, M. & Ibnsouda, S.K. (2016). Methods for in vitro evaluating antimicrobial activity: a review. *J. Pharmaceut. Biomed.,* **6**: 71-79.

Bernstein, J.A., Alexis, N., Bacchus, H., Bernstein, I.L., Fritz, P., Horner, E. & Reijula, K. (2008). The health effects of nonindustrial indoor air pollution. *J. Allergy. Clin. Immun.*, **121**: 585-591.

Bouillard, L., Michel, O., Dramaix, M., & Devleeschouwer, M. (2005). Bacterial contamination of indoor air, surfaces, and settled dust, and related dust endotoxin concentrations in healthy office buildings. *Ann. Agr. Env. Med.,* **12**: 187-192.

Brągoszewska, E., Biedroń, I., Kozielska, B. & Pastuszka, J.S. (2018). Microbiological indoor air quality in an office building in gliwice, poland: analysis of the case study. *Air Qual. Atmos. Hlth.,* **11**: 729-740.

Brasche, S. & Bischof, W. (2005). Daily time spent indoors in German homes baseline data for the assessment of indoor exposure of German occupants. *Int. J. Hyg. Envir. Heal.*, **208**: 247-253.

Brock, T.D. & Bohlool, B.B. (1974). Immunofluorescence approach to the study of the ecology of thermoplasma acidophilum in coal refuse material. *Appl. Microbiol*., **1**: 11-16.

Brugger, S.D., Baumberger, C., Jost, M., Jenni, W. & Brugger U. (2012). Automated counting of bacterial colony forming units on agar plates. *PLOS One,* **7**: e33695.

Cardamone, J.M. (2002). *Proteolytic Activity of Aspergillus Flavus on Wool*. AATCC Rev.

Colbeck, I., Nasir, Z.A. & Ali, Z. (2010). Characteristics of indoor/outdoor particulate pollution in urban and rural residential environment of Pakistan. *Indoor Air*, **20**: 40-51.

Cox, C.S. & Wathes, C.M. (1995). Bioaerosols in the environment. *In Bioaerosols Handbook*, Lewis Publishers, New York, pp. 11-14.

Davis, C. (2014). Enumeration of probiotic strains: review of culture-dependent and alternative techniques to quantify viable bacteria. *J. Microbiol. Meth.,* **103**: 9-17.

Di, G.M., Grande, R., Di, C.E., Di, B.S. & Cellini, L. (2010). Indoor air quality in university environments. *Environ. Monit. Assess.,* **170**: 509-517.

Douwes, J., Thorne, P., Pearce, N. & Heederik, D. (2003). Bioaerosol health effects and exposure assessment: progress and prospects. *Ann. Occup. Hyg.,* **47**: 187-200.

Dumała, S.M. & Dudzińska, M.R. (2013). Microbiological indoor air quality in Polish schools. *Ann. Set. Envir. Prot. (Rocznik Ochrona Środowiska)*, **15**: 231-244.

EPA (1987). *The Total Exposure Assessment Methodology* (*TEAM) Study: Summary and Analysis*. EPA/600/6-87/002a, Washington, DC.

Fox, K., Fox, A., Elßner, T., Feigley, C. & Salzberg, D. (2010). Maldi-tof mass spectrometry speciation of staphylococci and their discrimination from micrococci isolated from indoor air of schoolrooms. *J. Environ. Monitor.*, **12**: 917-923.

Gawrońska, H. & Bakera, B. (2015). Phytoremediation of particulate matter from indoor air by Chlorophytum comosum L. plants. *Air Qual. Atmos. Health.,* **8**: 265-272.

Gizaw, Z., Gebrehiwot, M. & Yenew, C. (2016). High bacterial load of indoor air in hospital wards: the case of University of Gondar teaching hospital, Northwest Ethiopia. *Multidiscip. Respir. Med.,* **11**: 24.

Haig, C.W., Mackay, W.G., Walker, J.T. & Williams, C. (2016). Bioaerosol sampling: sampling mechanisms, bioefficiency and field studies. *J. Hosp. Infect.,* **95**: 242-245.

Hayleeyesus, S.F. & Manaye, A.M. (2014). Microbiological quality of indoor air in university libraries. *Asian Pac. J. Trop. Biomed.,* **4**: S312-S317.

Hazan, R., Que, Y.A., Maura, D., & Rahme, L.G. (2012). A method for high throughput determination of viable bacteria cell counts in 96-well plates. *BMC Microbiol.*, **12**: 259.

Heikkinen M.S.A., Hjelmroos M.K., Haggblom, M.M. & Macher, J.M. (2005). Bioaerosols. *In* Ruzer, L. & Harley N.H. (Eds.), *Aerosols Handbook: Measurement, Dosimetry, and Health Effects.* CRC Press, Boca Raton, pp. 291-342.

Henderson, T.J. & Salem, H. (2016). The atmosphere: its developmental history and contributions to microbial evolution and habitat. *In* Salem, H. & Katz, S.A. (Eds.), *Aerobiology: The Toxicology of Airborne Pathogens and Toxins.* The Royal Society of Chemistry, The State University of New Jersey, New Jersey, pp. 1-41.

Hewitt, K.M., Gerba, C.P., Maxwell, S.L. & Kelley, S.T. (2012). Office space bacterial abundance and diversity in three metropolitan areas. *PLOS One,* **7**: e37849.

Jensen, P.A. & Schafer, M.P. (1998). *Sampling and Characterization of Bioaerosols*. NIOSH/DPSE NIOSH Manual of Analytical Methods, New York.

Karottki, D.G., Spilak, M., Frederiksen, M., Jovanovic, A.Z., Madsen, A.M., Ketzel, M. & Loft, S. (2015). Indoor and outdoor exposure to ultrafine, fine and microbiologically derived particulate matter related to cardiovascular and respiratory effects in a panel of elderly urban citizens. *Int. J. Env. Res. Pub. He.,* **12**: 1667-1686.

Klepeis, N.E., Nelson, W.C., Ott, W.R., Robinson, J.P. & Tsang, A.M. (2001). The national human activity pattern survey: a resource for assessing exposure to environmental pollutants. *J. Expo. Anal. Env. Epid.,* **11**: 321-352.

Koistinen, K.J., Edwards, R.D., Mathys, P., Ruuskanen, J., Künzli, N. & Jantunen, M.J. (2004). Sources of fine particulate matter in personal exposures and residential indoor, residential outdoor and workplace microenvironments in the helsinki phase of the EXPOLIS study. *Scand. J. Work. Env. Hea.,* **30**: 36-46.

Kopperud, R.J., Ferr, A.R. & Hildemann, L.M. (2004). Outdoor versus indoor contributions to indoor particulate matter determined by mass balance methods. *J. Air. Waste. Manage.,* **54**: 1188-1196.

Kotzias, D. (2005). Indoor air and human exposure reassessment - needs and approaches. *Exp. Toxicol. Pathol*., **57**: 5-7.

Lamoth, F. & Greub, G. (2010). Amoebal pathogens as emerging causal agents of pneumonia. *FEMS Microbiol. Rev.*, **34**: 260-280.

Latif, M.T., Yong, S.M., Saad, A., Mohamad, N., Baharudin, N.H., Mokhtar, M.B. & Tahir, N.M. (2014). Composition of heavy metals in indoor dust and their possible exposure: a case study of preschool children in Malaysia. *Air Qual. Atmos. Health,* **7**: 181-193.

Law, A.K.Y, Chau, C.K. & Chan, G.Y.S (2001). Characteristics of bioaerosol profile in office buildings in Hong Kong. *Build. Environ.,* **36**: 527-541.

Lee, B.U. (2011). Life comes from the air: a short review on bioaerosol control. *Aerosol Air Qual. Res.*, **11**: 921-927.

Li, Y.P., Fu, H.L., Wang, W., Liu, J., Meng, Q.L. & Wang, W.K. (2015). Characteristics of bacterial and fungal aerosols during the autumn haze days in Xi'an, China. *Atmos. Environ*., **122**: 439-447.

Li, Y.P., Lu, R., Li, W.X., Xie, Z.S. & Song, Y. (2017). Concentrations and size distributions of viable bioaerosols under various weather conditions in a typical semi-arid city of Northwest China. *J. Aerosol. Sci.*, **106**: 83-92.

Macher, J.M., Chatigny, M.A. & Burge, H.A. (1995). Sampling airborne microorganisms and aeroallergens. *In* Cohen, BS. & Hering, SV. (Eds.), *Air Sampling Instruments for Evaluation of Atmospheric Contaminants.* 8th ed. American Conference of Governmental Industrial Hygienists, Inc., Cincinnati, Ohio, pp. 589-617.

Mackintosh, C.A., Lidwell, O.M., Towers, A.G., & Marples, R.R. (1978). The dimensions of skin fragments dispersed into the air during activity. *Epidemiol. Infect.,* **81**: 471-480.

Makkar, N.S. & Casida, L.E. (1987). Technique for estimating low numbers of a bacterial strain(s) in soil. *Appl. Environ. Microbiol.,* **53**: 887-888.

Moon, K.W., Huh, E.H. & Jeong, H.C. (2014). Seasonal evaluation of bioaerosols from indoor air of residential apartments within the metropolitan area in South Korea. *Environ. Monit. Assess*., **186**: 2111-2120.

Mouli, P.C., Mohan, S.V. & Reddy, S.J. (2005). Rainwater chemistry at a regional representative urban site: influence of terrestrial sources on ionic composition. *Atmos. Environ.*, **39**: 999-1008.

Muirhead, K.A., Horan, P.K. & Poste, G. (1985). Flow cytometry: present and future. *Biotechnol.,* **3**: 337-356.

Nasir, Z.A. & Colbeck, I. (2010). Assessment of bacterial and fungal aerosol in different residential settings. *Water Air Soil Pollut*., **211** :367-377.
Nazaroff, W.W. (2016). Indoor bioaerosol dynamics. *Indoor Air,* **26**: 61-78.
Noble, W.C., Habbema, J.D.F., Van, F.R., Smith, I. & De, R.C. (1976). Quantitative studies on the dispersal of skin bacteria into the air. *J. Med. Microbiol.*, **9**: 53-61.
Pastuszka, J.S., Marchwinska, W.E. & Wlazlo, A. (2005). Bacterial aerosol in Silesian hospitals: preliminary results. *Pol. J. Environ. Stud.*, **14**: 883.
Pastuszka, J.S., Paw, U.K.T., Lis, D.O., Wlazło, A. & Ulfig, K. (2000). Bacterial and fungal aerosol in indoor environment in Upper Silesia, Poland. *Atmos. Environt.*, **34**: 3833-3842.
Pegas, P.N., Evtyugina, M.G., Alves, C.A., Nunes, T., Cerqueira, M., Franchi, M. & Freitas, M.D.C. (2010). Outdoor/indoor air quality in primary schools in Lisbon: a preliminary study. *Quim. Nova.,* **33**: 1145-1149.
Porter, J., Deere, D., Hardman, M., Edwards, C. & Pickup, R. (1997). Go with the flow - use of flow cytometry in environmental microbiology. *FEMS Microbiol. Ecol.,***24**: 93-101.
Prussin, A.J., Garcia, E.B. & Marr, L.C. (2015). Total concentrations of virus and bacteria in indoor and outdoor air. *Environ. Sci. Tech. Let.,* **2**: 84-88.
Qian, J., Hospodsky, D., Yamamoto, N., Nazaroff, W.W. & Peccia, J. (2012). Size-resolved emission rates of airborne bacteria and fungi in an occupied classroom. *Indoor Air,* **22**: 339-351.
Rao, C.Y., Burge, H.A. & Chang. J.C.S. (1996). Review of quantitative standards and guidelines for fungi in indoor air. *J. Air Waste Manage.*, **46**: 899-908.
Reponen, T., Nevalainen, A., Jantunen, M., Pellikka, M., & Kalliokoski, P. (1992). Normal range criteria for indoor air bacterial and fungal spores in a subarctic climate. *Indoor Air*, **2**: 6.
Salleh, N.M., Kamaruzzaman, S.N., Sulaiman, R. & Mahbob, N.S. (2011). Indoor air quality at school: ventilation rates and it impacts towards children - a review. *2nd Int. Conf. Envir. Sci. Tech.,* **6**: 418-422.
Schwab, C.J., Cooley, J.D., Brasel, T., Jumper, C.A., Graham, S.C. & Straus, D.C. (2003). Characterization of exposure to low levels of viable penicillium chrysogenum conidia and allergic sensitization induced by a protease allergen extract from viable p. chrysogenum conidia in mice. *Int. Arch. Allergy. Imm*., **130**: 200-208.
Shin, S.K., Kim, J., Ha, S., Oh, H.S., Chun, J., Sohn, J. & Yi, H. (2015). Metagenomic insights into the bioaerosols in the indoor and outdoor environments of childcare facilities. *PLOS One*, **10**: 1-17.
Song, D., Liu, H., Dong, Q., Bian, Z., Wu, H., & Lei, Y. (2018). Digital, rapid, accurate and label free enumeration of viable microorganisms enabled by custom built on glass slide culturing device and microscopic scanning. *Sensors.*, **18**: 3700.
Stetzenbach, L.D., Buttner, M.P. & Cruz, P. (2004). Detection and enumeration of airborne biocontaminants. *Curr. Opin. Biotech.*, **15**:170-174.
Thiry, M.C. (2001). Small game hunting: Anti microbials take the field. *AATCC Rev.,* **1**:11-17.
Wei, K., Zou, Z., Zheng, Y., Li, J., Shen, F., Wu, C.Y. & Yao, M. (2016). Ambient bioaerosol particle dynamics observed during haze and sunny days in Beijing. *Sci. Total. Environ.*, **550:** 751-759.
WHO (2009). *WHO Guidelines for Indoor Air Quality: Dampness and Mould.* World Health Organization 2009, WHO Regional Office for Europe, Europe.
Willeke, K. & Macher, J.M. (1999). Air sampling. *In* Macher, J.M. (Ed.), *Bioaerosols: Assessment and Control.* American Conference of Governmental Industrial Hygienists, Inc., Cincinnati, Ohio, pp. 11-25.
Womack, A.M., Bohannan, B.J. & Green, J.L. (2010). Biodiversity and biogeography of the atmosphere. *Philos. Trans. R. Soc. Lond.,* **365**: 3645-3653.
Xie, Z., Fan, C., Lu, R., Liu, P., Wang, B., Du, S. & Li, Y. (2018). Characteristics of ambient bioaerosols during haze episodes in China: a review. *Environ. Pollut.*, **243**: 1930-1942.
Yassin, M.F. & Almouqatea, S. (2010). Assessment of airborne bacteria and fungi in an indoor and outdoor environment. *Int. J. Environ. Sci. Tech*., **7**: 535-544.

BIOAEROSOL SAMPLING AND IDENTIFICATION OF AIRBORNE BACTERIA IN INDOOR AND OUTDOOR ENVIRONMENTS

Nik Nur Ilyani Mohamed Nazri[1*], Ann Nurrizka Abd. Hamid[2] & Nur Amira Aminuddin[2]

[1]Biosurveillance and Biological Defence Branch, Protection and Biophysical Technology Division(BTPB), Science and Technology Research Institute for Defence (STRIDE), Ministry of Defence, Malaysia
[2]School of Biology, Faculty of Applied Sciences, Universiti Teknologi MARA (UiTM), Malaysia

[*]Email: nikilyani.nazri@stride.gov.my

ABSTRACT

People are exposed every day to a variety of bioaerosols, including airborne bacteria, which can lead to both beneficial and detrimental effects to the environment and human beings. This study was conducted in July 2019 to investigate the air quality at a government premise in Selangor, Malaysia. The sampling sites were selected randomly at the same building. One of the samples was collected outdoors (garden), while the others were collected indoors which include the pantry, toilet, staff rooms, store rooms, meeting room, prayer room, exhibition hall and training room. The 11 sampling sites differ from each other in terms of the presence of personnel, windows, air conditioning system, water source, organic substrates and flooring material. For bioaerosol sampling, the Andersen impaction method was applied using a Tryptic Soy Agar (TSA) non-selective media plate. Three units of SKC biological air sampler (SKC Quick Take 30, USA) with setting of 2 min sampling time and volume flow rate of 28.3 L/min were used in each sampling location. The plates were incubated at 35 ºC for 24 h, before the bacterial colonies were observed and segregated into different groups based on their morphology, such as colours, shapes and sizes. The selected bacterial colonies were then serially subcultured using the streaking method in order to obtain the pure colonies for identification purposes using a Biolog microbial identification system (Biolog Gen III Technology, USA). The isolates were from four genera, which were Bacillus, Staphylococcus, Micrococcus and Pseudomonas. Thirteen bacteria were identified at species level. The most dominant species was Bacillus marisflavi, which was found in all the indoor sampling sites. Meanwhile, for the outdoor sampling site, only Bacillus pumilus was found. In conclusion, airborne bacteria presence could possibly originate from outdoor sources, with human activities as the main vector of the contamination.

Keywords: *Bioaerosol; airborne bacteria; indoor and outdoor environments; biological air sampler; microbial identification system.*

1. INTRODUCTION

Bioaerosol is one of the airborne pollutants that is usually associated with compounds of biological origin. Bioaerosols include all pathogenic or non-pathogenic, live or dead fungi and bacteria, bacterial endotoxins, mycotoxins, peptidoglycans, β (1, 3)-glucans, viruses, high molecular weight allergens, pollens, and biofilm (Bloomfield *et al*., 1998; Douwes *et al*., 2003; Ariya & Amyot, 2004; Bigg & Leck, 2008). Bioaerosols, as important components of atmospheric aerosol, account for 30 to 80% of the particulate matter (PM) in the atmosphere (Jaenicke, 2005; Huffman *et al*., 2012; Fröhlich *et al*., 2016). Airborne microorganisms could be pathogenic and cause various respiratory tract disease and allergies (Douwes *et al.,* 2003). They also play important functions in atmospheric chemistry and nucleation processes, such as biotransformation of organic matter, carbon

cycling, photochemical reactions and cloud formation to influence global climate (Morris *et al*., 2014; Fröhlich *et al*., 2016).

Bioaerosol monitoring in occupational environments is a rapidly emerging area of industrial hygiene. It is one of the many tools industrial hygienist uses in the assessment of indoor environmental quality, infectious disease outbreaks, agricultural health and clean rooms. This assessment includes the measurement of viable (culturable and nonculturable) and nonviable microorganisms in both indoor (e.g., industrial, office or residential) and outdoor (e.g., agricultural and general air quality) environments. In general, indoor bioaerosol sampling need not be performed if visible growth is observed. In addition, contamination, such as microbial growth on the floor, wall or ceiling, should be remedied. If personnel remain symptomatic after remediation, air sampling may be appropriate. However, industrial hygienists should bear in mind that false negative results are quite possible and should be interpreted with caution (Jensen *et al.,* 1994; Jensen & Schafer, 1998).

In general, indoor microflora concentrations of a healthy work environment are lower than outdoor concentrations at the same location (ACGIH, 1989; Macher *et al*., 1995). If one or more genera are found indoors, in concentrations greater than outdoor concentrations, then the source of amplification must be found and remedied. Bioaerosol sampling is often performed outdoors for pollen and fungi to assist allergists in their treatment of patients by identifying taxa distribution and concentration in air over time. On occasion, outdoor bioaerosol sampling is conducted in an occupational environment (e.g. agricultural investigations and sewage treatment plants). Indoor bioaerosol sampling is often conducted in both occupational (e.g., industrial and office environments) and non-occupational (e.g., residential and educational buildings) settings (Jensen *et al.,* 1994; Jensen & Schafer, 1998).

Most bioaerosol sampling devices involve techniques that separate particles from the airstream and collect them in or on a preselected medium. Impactors, filters, impingers and cyclones are four common sampling techniques used to separate and collect bioaerosols (Macher *et al*., 1995; Willeke & Macher, 1999; Reponen *et al*., 2009; Haig *et al*., 2016). After air sampling, microbial identification methods are required to distinguish the taxa of the microbes. There are many microbial identification methods, such as microscopic examination (Burge, 1995), polymerase chain reaction (PCR) (Saiki *et al*., 1985), laser induced fluorescence (LIF) (Rösch *et al*., 2005) and cellular fatty acid analysis (Sasser, 1990a, b).

2. MATERIALS AND METHODS

2.1 Preparation of Tryptic Soy Agar (TSA) Medium

A total of 3 g of Tryptic soy broth (TSB) powder was weighed and dissolved in 100 ml distilled water in a conical flask. The solution was stirred using a magnetic stirrer (700 rpm) and heated up on a hotplate (130 °C) until the murk was almost cleared. The solution was poured into an autoclave bottle and autoclaved at 121 °C for 20 min. A volume of 15 ml of the autoclaved medium solution was poured into each sterile petri dish until the bottle was empty. The agar medium was left to dry at room temperature in laminar air flow before placing the stock in a refrigerator at 4 °C. The medium was pre-warmed to room temperature before being used in the bioaerosol sampling.

2.2 Air Sampling

The sampling sites were selected randomly from the same building of the government premise. One of the samples was collected outdoors (garden), while others are collected indoors, which include the pantry, toilet, staff rooms, store rooms, meeting room, prayer room, exhibition hall and training room. The 11 sampling sites differ from each other in terms of the presence of personnel, windows, air

conditioning system, water source, organic substrates and flooring materials. The required equipment was prepared and sterilised to avoid sample contamination. The Andersen impaction method was applied using a TSA non-selective media plate. All measurements were performed in triplicate. Three units of SKC biological air sampler (SKC Quick Take 30, USA) with setting of 2 min sampling time and volume flow rate of 28.3 L/min were used in each sampling site. All the sampling plates were incubated immediately at 35 °C for 24 h.

2.3 Colony Selection from Sampling Plates and Subculture

After 24 h of incubation, the microbial colonies at the sampling plates were observed and segregated into different groups based on their morphology, such as colours, shapes and sizes. Selected microbial colonies were then serially subcultured using the streaking method in order to obtain the pure colonies for identification purposes. The pure colonies were incubated at 35 °C for 24 h before being used in the pre-identification process.

2.4 Pre-Identification Process: Inoculum Preparation

Before proceeding into the microbial identification process, the inoculating fluid should be prepared in a biosafety cabinet. The turbidity of the inoculating fluid that was specialised for bacteria identification (IF-AN) was examined with a turbidimeter with 100% transmittance set as control.

2.5 Pre-Identification Process: Inoculation of Microplates

Each cell from the pure colony plates was collected and transferred into IF-AN solution using a sterile cotton swab. The cotton swab was stirred and swirled slowly until it mixed with the fluid. The turbidity of the suspension fluid was tested using a turbidimeter to be at an acceptable range of 97 – 98 %. Each microplate was labelled with the date and sample name. Each suspension fluid was poured into a multichannel pipette reservoir before a volume of 100 ml was pipetted into each microplate using a multichannel pipettor until all 96 microplate wells were filled with the suspension fluid. Each microplate was covered with its lid and incubated at 35 °C for 24 h. All the pure colony plates were sealed with parafilm and stored in the refrigerator at 4 °C.

2.6 Identification Process Using a Microbial Identification System

After 24 h of incubation, all the microplates were observed a using a Biolog microbial identification system (Biolog Gen III Technology, USA) with the Biolog GEN III Database for the identification process, and was continuously incubated and identified for up to 5 days. For the remaining unidentified colonies, resubculturing was done from the pure colony plates using the streaking method and incubated at 35 ºC for 24 h. The steps for inoculum preparation and microplate inoculation were repeated for the identification process. Microplates were discarded after all the bacterial colonies were identified.

2.7 Subculture, Incubation and Observation of Fungi

After the identification process, there were four remaining unidentified colonies that were suspected to be fungi. Resubculturing was done from the fungi pure colony plates using cork borer and potato dextrose agar (PDA) plates. The subculture plates were incubated at 26 ºC and the morphology of fungi growth on the subculture plates were observed daily for three days before being used in the DNA sequencing identification process.

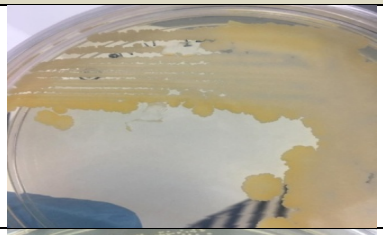
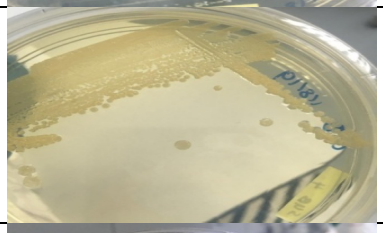
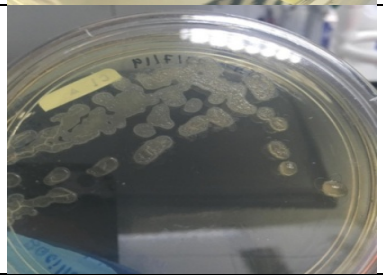
2.8 Identification Process Using DNA Sequencing

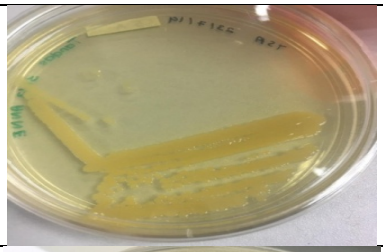
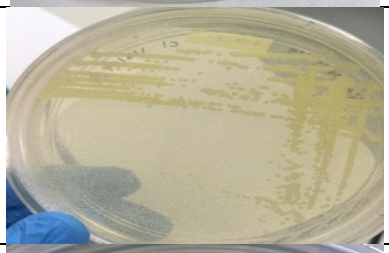
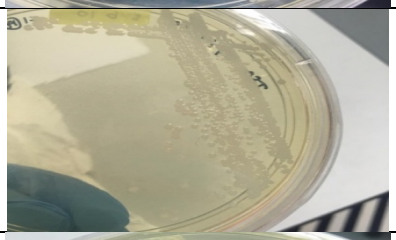
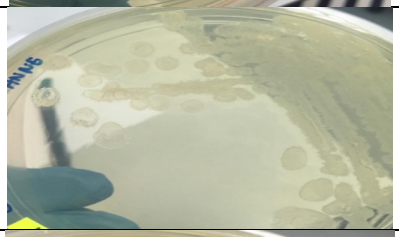
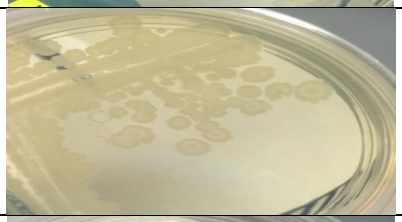
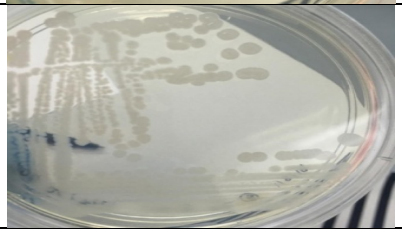
The fungal ITS genes were amplified using universal primers ITS1 and ITS4. The total reaction volume of 25 µl containing gDNA were purified using a commercial kit extraction method, 0.5 pmol of each primer, 200 µM of each deoxynucleotides triphosphates (dNTPs), 0.5 U of DNA polymerase, a standard PCR buffer and water. The PCR was performed as follows: one cycle for initial denaturation (98 ºC for 2 min); 25 cycles for annealing and extension (98 ºC for 15 sec, 60 ºC for 30 sec and 72 ºC for 30 sec), and one cycle for the final extension of the amplified DNA (72 ºC for 10 min). The PCR products were then purified using a standard method and directly sequenced using a BigDye® Terminator v3.1 cycle sequencing kit (Applied Biosystems).

3. RESULTS AND DISCUSSION

The species identification of bacteria and fungi for both indoor and outdoor environments are shown in Tables 1 and 2 respectively. Figures 1 and 2 indicate the number of bacteria and fungi found from each sampling site, while Figure 3 shows the frequency of locations found for each bacteria species. *Bacillus marisflavi*, *Micrococcus luteus*, *Bacillus pumilus* and *Bacillus idriensis* were found in ten, six and three sampling sites respectively. *Bacillus megaterium*, *Staphylococcus aureus* and *Staphylococcus epidermidis* were all found in two sampling sites, while the rest were found in only one sampling site.

Table 1: Streaking plates of bacteria species identified with probability value (P) and locations found.

Streaking Plates	Species Identification	Locations Found
	*Staphylococcus xylosus* (*P* = 0.638)	Staff Room 2
	*Kytococcus sedentarius* (*P* = 0.726)	Pantry
	*Pseudomonas stutzeri* (*P* = 0.516)	Pantry Prayer Room
	*Bacillus marisflavi* (*P* = 0.653)	Pantry Training Room Staff Room 1 & 2 Store Room 1& 2 Meeting Room Prayer Room Toilet Exhibition Room

	*Bacillus idriensis* (*P* = 0.596)	Toilet Meeting Room Store Room 1
	*Micrococcus luteus A* (*P* = 0.779)	Prayer Room Toilet Staff Room 1 & 2 Meeting Room 2 Store Room 2
	*Bacillus megaterium* (*P* = 0.734)	Toilet Staff Room 1
	*Staphylococcus epidermidis* (*P* = 0.596)	Training Room Prayer Room
	*Staphylococcus aureus ss aureus* (*P* = 0.569)	Store Room 1 & 2
	*Bacillus simplex* (*P* = 0.624)	Training Room
	*Staphylococcus sciuri ss sciuri* (*P* = 0.582)	Staff Room 2
	*Bacillus pumilus* (*P* = 0.582)	Garden Toilet Staff Room 2

	*Bacillus endophyticus* ($P = 0.610$)	Staff Room

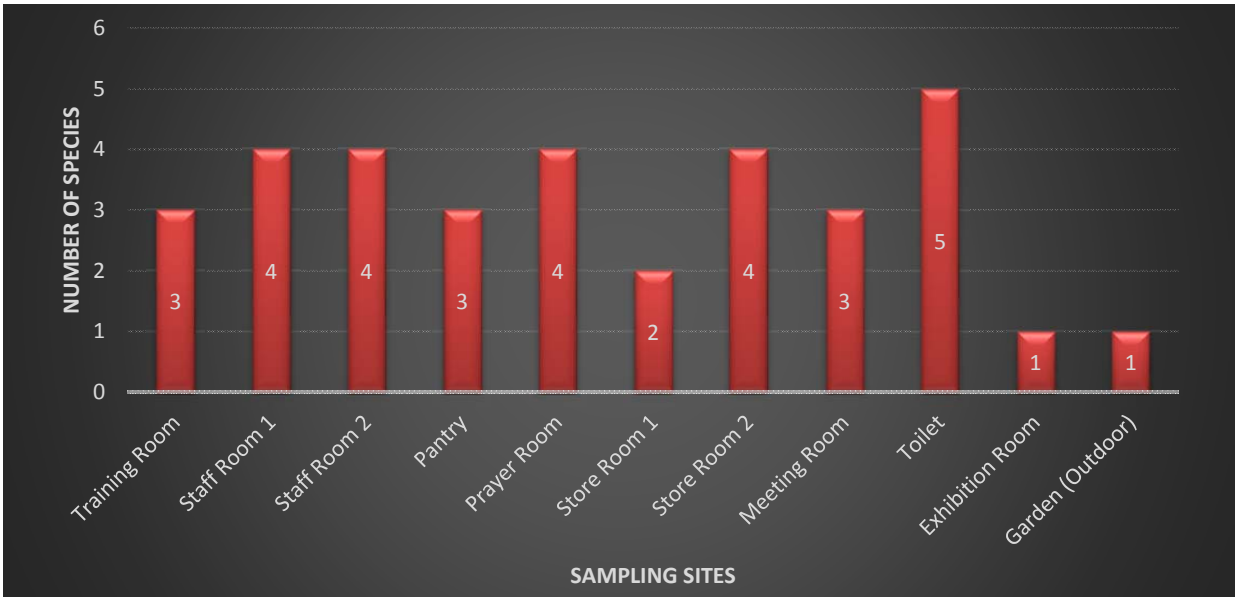
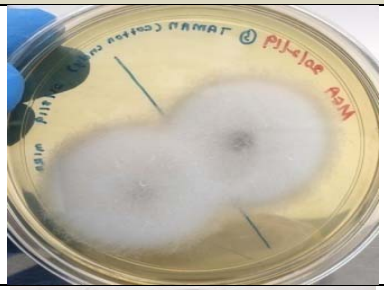
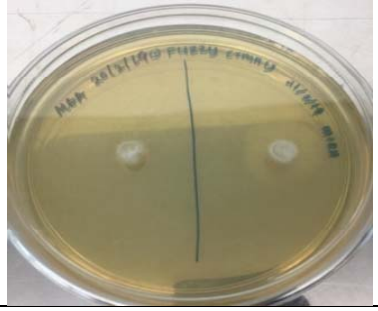
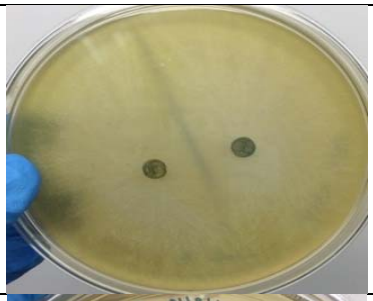


Figure 1: Number of bacteria species identified in the sampling sites.

Table 2: Subculture plates of fungi species identified and locations found

Subculture Plates	Species Identification	Locations Found
	*Cunninghamella echinulata*	Garden Training Room
	*Coprinellus* sp.	Garden

	*Trichoderma reesei*	Garden Toilet Meeting Room Exhibition Room Store Room 1 & 2 Staff Room 1 & 2 Prayer Room
	*Aspergillus carbonarius*	Staff Room 1 & 2 Toilet Prayer Room

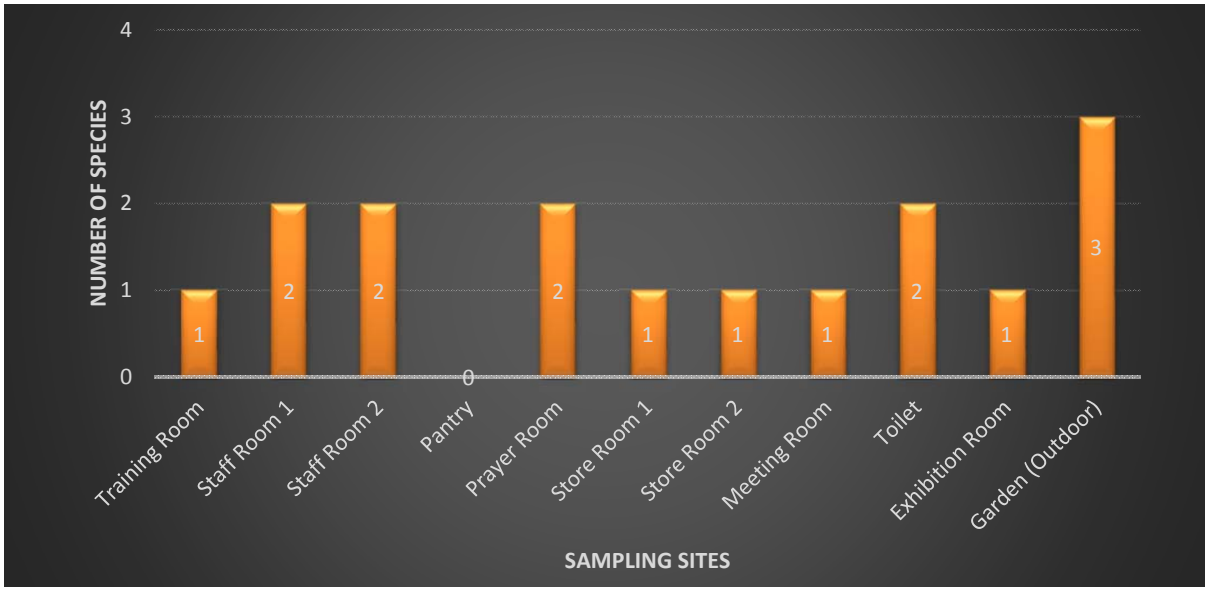


Figure 2: Number of fungi species in identified in the sampling sites.

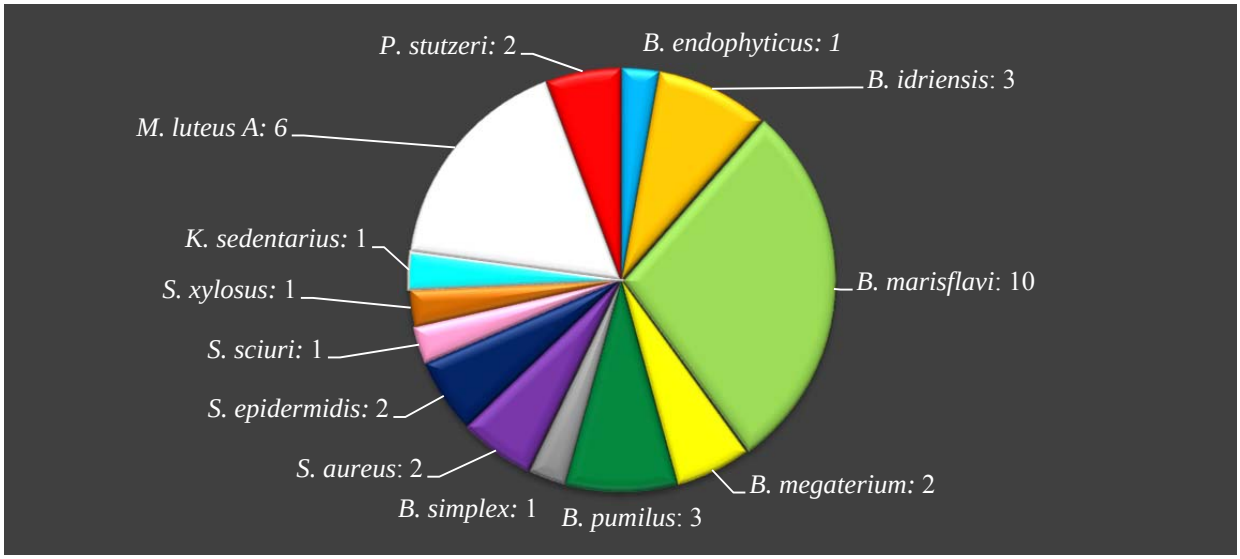


Figure 3: Number of sampling sites where each bacteria species was found.

For the bacteria found in all the sampling sites, basically, all the isolated genera were common bacteria found in the air and non-pathogenic (Hayleeyesus & Manaye, 2014). The sources of temporal and spatial distribution of bioaerosols, such as sampling sites, seasons, meteorological factors (air temperature, relative humidity (RH), wind speed and wind direction) and anthropogenic emissions ($PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3) may increase the chances of airborne bacteria being found at the sampling sites (Bowers *et al.,* 2012; Boreson *et al.*, 2004). In this study, several factors were taken into consideration, such as presence of personnel, air ventilation, carpet, wood and tile flooring materials, water source, organic substrates from food, and the haze season.

The various bacteria species isolated might be influenced by the haze season that occurred in Malaysia from mid-July to the end of September 2019. Bioaerosols can live well in the air with nutrients supported by clouds, rainwater and particles (Womack *et al.,* 2010). Although the atmosphere provides survival conditions for airborne microorganism due to microbial intrinsic properties, different microbes can better adapt themselves to certain environment and develop into preponderant species in disparate atmospheric settings (Zhai *et al.*, 2018).

Bacillus is one of the common genera in the air (Hayleeyesus & Manaye, 2014), in residences (Adams *et al.,* 2014) and retail stores (Hoisington *et al.,* 2016). *B. marisflavi* has been identified in marine waters (Yoon *et al.,* 2003; Khaneja *et al.,* 2010). It has also been detected in marine water samples, whereby Akayli *et al*. (2016) found its presence in intestinal microflora of a marine fish species. However, like all members of the genus *Bacillus, B. marisflavi* is capable to form highly resistant dormant endospores in response to nutrient deprivation and other environmental stresses (Sonenshein *et al.* 2002). These spores can be easily made airborne and disperse by wind (Jaenicke, 2005; Merrill *et al*., 2006). Thus, spores might migrate long distances, land in a given environment but never germinate there. Considering that the traditional method for isolating *B. marisflavi* requires that the organism be in its spore form, there is no guarantee that when a strain is isolated from a particular environment, it is actually growing at that location. Thus, to date, the question of where *B. marisflavi* grows remains largely unanswered. *B. pumilus* was only found in the garden because it is naturally found in soil, marine and terrestrial areas (Satomi *et al*., 2006; Liu *et al*., 2013; Lai *et al*., 2014), in the air at high altitudes (Shivaji *et al*., 2006), and in the roots of some plants used as antimicrobial agents (Kaushal *et al*., 2017).

Both genera of *Micrococcus* and *Staphylococcus* are dominant in office buildings (Bonetta *et al*., 2010). They are also among the resident microorganisms on human skins (Kloos & Musselwhite, 1975; Peters *et al*., 1990; Bouillard *et al*., 2005) and are also available in the air via the shedding of skin to the surrounding (Hayleeyesus & Manaye, 2014; Prussin & Marr, 2015) that might be trapped on the carpet or wood surfaces. This explains why *Staphylococcus* sp. was found in sites such as the staff room, prayer room and training room, which have higher presence of personnel, in addition to the flooring material in the rooms being made of wood and carpet. The *Staphylococcus* sp. and *Micrococcus* sp. found are non-pathogenic but could lead to opportunistic infections (Albertson *et al*., 1978; Gozalo *et al*., 2010; Nemeghaire *et al*., 2014; Zhou *et al.*, 2016).

Stagnant water is known to be a breeding ground for *Pseudomonas* sp. and this could be a cause for the persistence of this bacterium in the sampling sites. In addition, the presence of personnel infected with *Pseudomonas* sp. in the sampling sites might have led to the occurrence of *Pseudomonas* sp. in the study. It is also found in food courts where there is high organic substrate from food (Rajasekar & Balasubramanian, 2011). Among cultivable bacteria, the Gram-positive bacteria, which is generally the dominant bacteria isolated from bioaerosols, take up a large part of all detected genera in different studies (Amato *et al*., 2007; Fröhlich *et al*., 2016; Raisi *et al*., 2012). These proportions may change with other parameters, but the overall trends are basically constant (Kowalski *et al*., 2017). *P. stutzeri* is found everywhere, including soil, water and eye makeup (Özbek *et al*., 2010), which explains the existence of the bacteria in prayer room and pantry. However, *P. stutzeri* infections are rare.

The toilet had the highest number of bacteria species found as plumbing system is a major contributor to bioaerosols in built environments (Prussin *et al*., 2015). Some uncommon species (such as *B.*

marisflavi and *P. stutzeri*) that originate from the soil, marine plants and animal microflora are also found contaminating indoor environments (Koistinen *et al.*, 2004; Kopperud *et al*., 2004). The bacteria could be transferred into the indoors by dust particles as the main vector of the contamination (Di *et al*., 2010; Nazaroff, 2016). The sampling sites with high human presence have high number of bacteria species found, such as the toilet, prayer room, staff room and pantry. This is as humans are one of the greatest sources of bioaerosol in built environment (Qian *et al*., 2012; Adams *et al*., 2013) .

Fungi samples were found due to their high survivability and longevity. *Cladosporium*, *Aspergillus* and *Penicillium* have the highest frequencies in relevant reports, which shows that they are the dominant genera of airborne fungi (Sharma, 2011; Pyrri & Kapsanaki, 2011; Sepahvand *et al*., 2013; Bezerra *et al*., 2014). In addition, fungi could also grow in TSA medium as it is a non-selective medium. Thus, *Cunninghamella echinulata*, *Coprinellus* sp., *Trichoderma reesei* and *Aspergillus carbonarius* were also found in the bioaerosol samples. All the isolated fungi were found in the outdoor environment (garden). The fungi species were also found in every indoor sampling sites, except in the pantry. For future research, it is beneficial to conduct sampling on airborne fungi at the same sampling sites using specific medium agar for fungal growth such, as PDA. Before conducting the DNA sequencing identification process, the fungi samples were first identified using the microbial identification system. However, this can be considered as a poor identification method because it detected the fungi samples as bacteria with low probability value (less than 0.5), even though using inoculating fluid that is specifically used for bacterial identification only.

4. CONCLUSION

From the results obtained using the microbial identification system, four genera and thirteen bacteria species were identified. In addition, four fungi species have been identified using DNA sequencing. *Bacillus* sp., *Micrococcus* sp., *Staphylococcus* sp. and *Pseudomonas* sp. were the four main genera of airborne bacteria found in the indoor and outdoor environments. Several factors, such as presence of personnel, air ventilation, flooring material, water source and organic substrate, are considered as the main sources of bacterial contaminations in both environments. As *B. marisflavi* was found to be the most dominant species, a better understanding of its ecology could be useful for uncovering the still largely unknown determinants enabling the dispersal and fate of such populations in environmental conditions. Even though the bacteria found were mostly non-pathogenic and not harmful to human beings, attention should be given to control environmental factors that favour the growth and multiplication of microbes in indoor environments of the building to safeguard health of personnel.

While the information provided from the microbial identification system was time saving and very helpful for bacteria identification, it was not effective enough to distinguish between fungi and bacteria species at the initial stage. Therefore, it is important to develop a smarter, and more time saving and accurate technology in the future to obtain more reliable data for the identification analysis of the bioaerosol samples. Observation through microscope can also be done in future research to study the morphological details of each species and to compare with the results given by the Biolog GEN III Database. In addition, to study precisely on bacteria, anti-fungal agents should be applied on the agar medium to avoid the unwanted fungal growth.

REFERENCES

ACGIH (1989). *Guidelines for the Assessment of Bioaerosols in the Indoor Environment*. American Conference of Governmental Industrial Hygienists (ACGIH), Cincinnati, Ohio.

Adams, R.I., Miletto, M., Lindow, S.E., Taylor, J.W., Bruns, T.D. (2014). Airborne bacterial communities in residences: similarities and differences with fungi. *PLOS One*, **9**: e91283.

Adams, R.I., Miletto, M., Taylor, J.W. & Bruns, T.D. (2013). Dispersal in microbes: fungi in indoor air are dominated by outdoor air and show dispersal limitation at short distances. *ISME J.,* **7**: 1262-1273.

Akayli, T., Albayrak, G., Ürkü, C., Canak, Ö & Yörük, E. (2016). Characterization of micrococcus luteus and bacillus marisflavi recovered from common dentex (dentex dentex) larviculture system. *Medit. Mar. Sci.,* **17**: 163-169.

Albertson, D., Natsios, G.A. & Gleckman, R. (1978). Septic shock with Micrococcus luteus. *Arch. Intern. Med.,* **138**: 487-488.

Amato, P., Parazols, M., Sancelme, M., Laj, P., Mailhot, G. & Delort, A.M. (2007). Microorganisms isolated from the water phase of tropospheric clouds at the Puy de Dome: major groups and growth abilities at low temperatures. *FEMS Microbiol. Ecol.*, **59**: 242-254.

Ariya, P.A. & Amyot, M. (2004). New Directions: The role of bioaerosols in atmospheric chemistry and physics. *Atmos. Environ.,* **38**: 1231-1232.

Bezerra, G.F., Gomes, S.M., Neto Silva, M.A., Santos, R.M., Filho Muniz, W.E., Viana, G.M. & Nascimento, M.D. (2014). Diversity and dynamics of airborne fungi in Sao Luis, State of Maranhao, Brazil. *Rev. Soc. Bras. Med. Trop*., **47**: 69-73.

Bigg, E.K. & Leck, C. (2008). The composition of fragments of bubbles bursting at the ocean surface. *J. Geophys. Res. Atmos.,* **113**: D11.

Bloomfield, S.F., Stewart, G.S., Dodd, C.E., Boot, I.R. & Power, E. (1998). The viable but non-culturable phenomenon explained? *J. Microbiol.,* **144**: 1-3.

Bouillard, L., Michel, O., Dramaix, M., & Devleeschouwer, M. (2005). Bacterial contamination of indoor air, surfaces, and settled dust, and related dust endotoxin concentrations in healthy office buildings. *Ann. Agr. Env. Med.,* **12**: 187-192.

Bonetta, S., Mosso, S., Sampo, S. & Carraro, E. (2010). Assessment of microbiological indoor air quality in an Italian o□ce building equipped with an HVAC system. *Environ. Monit. Assess.,* **161**: 473-483.

Boreson, J., Dillner, A.M. & Peccia, J. (2004). Correlating bioaerosol load with PM2.5 and PM10 of concentrations: a comparison between natural desert and urban-fringe aerosols. *Atmos. Environ.,* **38**: 6029-6041.

Bowers, R.M., McCubbin, I.B., Hallar, A.G. & Fierer, N. (2012). Seasonal variability in airborne bacterial communities at a high-elevation site. *Atmos. Environ.*, **50**: 41-49.

Burge, H.A. (1995). Bioaerosols in the residential environment. *In* Cox, C.S. & Wathes, C.M. (Eds.), *Bioaerosols Handbook*. Lewis Publishers, Boca Raton, Florida, pp. 579-597.

Di, G.M., Grande, R., Di, C.E., Di, B.S. & Cellini, L. (2010). Indoor air quality in university environments. *Environ. Monit. Assess.,* **170**: 509-517.

Douwes, J., Thorne, P., Pearce, N. & Heederik, D. (2003). Bioaerosol health effects and exposure assessment: progress and prospects. *Ann. Occup. Hyg.,* **47**: 187-200.

Fröhlich, N.J., Kampf, C.J., Weber, B., Huffman, J.A., Pöhlker, C., Andreae, M.O. & Elbert, W. (2016). Bioaerosols in the earth system: climate, health, and ecosystem interactions. *Atmos. Res.,* **182**: 346-376.

Gozalo, A.S., Hoffmann, V.J., Brinster, L.R., Elkins, W.R., Ding, L. & Holland, S.M. (2010). Spontaneous *staphylococcus xylosus* infection in mice deficient in NADPH oxidase and comparison with other laboratory mouse strains. *J. Am. Assoc. Lab. Anim. Sci.,* **49**: 480-486.

Haig, C.W., Mackay, W.G., Walker, J.T. & Williams, C. (2016). Bioaerosol sampling: sampling mechanisms, bioefficiency and field studies. *J. Hosp. Infect.,* **95**: 242-245.

Hayleeyesus, S.F. & Manaye, A.M. (2014). Microbiological quality of indoor air in university libraries. *Asian Pac. J. Trop. Biomed.,* **4**: S312-S317.

Hoisington, A., Maestre, J.P., Kinney, K.A. & Siegel, J.A. (2016). Characterizing the bacterial communities in retail stores in the United States. *Indoor Air*, **26**: 857-868.

Huffman, J., Sinha, B., Garland, R., Snee, P.A., Gunthe, S., Artaxo, P. & Pöschl, U. (2012). Size distributions and temporal variations of biological aerosol particles in the Amazon rainforest characterized by microscopy and real-time UV-APS fluorescence techniques during AMAZE-08. *Atmos. Chem. Phys.,* **12**: 11997-12019.

Jaenicke, R. (2005). Abundance of cellular material and proteins in the atmosphere. *Sci.,* **308**: 73.

Jensen, P.A., Lighthart, B., Mohr, A.J. & Shaffer, B.T. (1994). Instrumentation used with microbial bioaerosol. *In* Lighthart, B. & Mohr, A.J. (Eds.), *Atmospheric Microbial Aerosols: Theory and Applications.* Chapman and Hall, New York, pp. 226-284.

Jensen, P.A. & Schafer, M.P. (1998). *Sampling and Characterization of Bioaerosols*. NIOSH/DPSE NIOSH Manual of Analytical Methods, New York.

Kaushal, M., Kumar, A. & Kaushal, R. (2017). Bacillus pumilus strain YSPMK11 as plant growth promoter and biocontrol agent against sclerotinia sclerotiorum. *Biotech*., **7**: 90.

Khaneja, R., Perez, F.L., Fakhry, S., Baccigalupi, L. & Steiger, S. (2010). Carotenoids found in Bacillus. *J. Appl. Microbiol.*, **108**: 1889-1902.

Kloos, W.E. & Musselwhite, M.S. (1975). Distribution and persistence of staphylococcus and micrococcus species and other aerobic bacteria on human skin. *Appl. Environ. Microbiol.,* **30**: 381-395.

Koistinen, K.J., Edwards, R.D., Mathys, P., Ruuskanen, J., Künzli, N. & Jantunen, M.J. (2004). Sources of fine particulate matter in personal exposures and residential indoor, residential outdoor and workplace microenvironments in the helsinki phase of the EXPOLIS study. *Scand. J. Work. Env. Hea.,* **30**: 36-46.

Kopperud, R.J., Ferr, A.R. & Hildemann, L.M. (2004). Outdoor versus indoor contributions to indoor particulate matter determined by mass balance methods. *J. Air Waste Manage.,* **54**: 1188-1196.

Kowalski, M., Wolany, J., Pastuszka, J.S., Płaza, G., Wlazło, A., Ulfig, K. & Malin, A. (2017). Characteristics of airborne bacteria and fungi in some polish wastewater treatment plants. *Int. J. Environ. Sci. Technol.*, **14**: 2181-2192.

Lai, Q., Liu, Y. & Shao, Z. (2014). Bacillus xiamenensis sp. nov., isolated from intestinal tract contents of a flathead mullet (mugil cephalus). *Anton. Van. Lee. J. M. S.,* **105**: 99-107.

Liu, Y., Lai, Q., Dong, C., Sun, F., Wang, L., Li, G. & Shao, Z. (2013). Phylogenetic diversity of the Bacillus pumilus group and the marine ecotype revealed by multilocus sequence analysis. *PLOS One,* **8**: e80097.

Macher, J.M., Chatigny, M.A. & Burge, H.A. (1995). Sampling airborne microorganisms and aeroallergens. *In* Cohen, BS. & Hering, SV. (Eds.), *Air Sampling Instruments for Evaluation of Atmospheric Contaminants.* 8th ed. American Conference of Governmental Industrial Hygienists, Inc., Cincinnati, Ohio, pp. 589-617.

Merrill, L., Dunbar, J., Richardson, J. & Kuske, C.R. (2006). Composition of Bacillus species in aerosols from 11 U.S. cities. *J. Forensic Sci.*, **51**: 559-565.

Morris, C.E., Conen, F., Alex, H.J., Phillips, V., Pöschl, U. & Sands, D.C. (2014). Bioprecipitation: a feedback cycle linking earth history, ecosystem dynamics and land use through biological ice nucleators in the atmosphere. *Glob. Change Biol.,* **20**: 341-351.

Nazaroff, W.W. (2016). Indoor bioaerosol dynamics. *Indoor Air,* **26**: 61-78.

Nemeghaire, S., Argudín, M.A., Haesebrouck, F. & Butaye, P. (2014). Epidemiology and molecular characterization of methicillin-resistant staphylococcus aureus nasal carriage isolates from bovines. *BMC Vet. Res.,* **10**: 153.

Özbek, A., Aktas, O., Uyanik, M.H., Bİlİcİ,, D. & Yildirim Z.K. (2010). A case of pseudomonas stutzeri bacteremia in a patient with hematologic malignancy. *FLORA,* **15**: 34-36.

Peters, G., Schumacher, P.F. & Jansen, B. (1990). Staphylococcus epidermidis-a versatile pathogen. *Curr. Top. Microbiol.,* Springer, pp. 309-315.

Prussin, A.J., Garcia, E.B. & Marr, L.C. (2015). Total concentrations of virus and bacteria in indoor and outdoor air. *Environ. Sci. Tech. Let.,* **2**: 84-88.

Prussin, A.J. & Marr, L.C. (2015). Sources of airborne microorganisms in the built environment. *Microbiome,* **3**: 78.

Pyrri, I. & Kapsanaki, G.E. (2011). Diversity and annual fluctuations of culturable airborne fungi in Athens, Greece: a 4 year study. *Aerobiologia,* **28**: 249-262.

Qian, J., Hospodsky, D., Yamamoto, N., Nazaroff, W.W. & Peccia, J. (2012). Size□resolved emission rates of airborne bacteria and fungi in an occupied classroom. *Indoor Air,* **22**: 339-351.

Raisi, L., Aleksandropoulou, V., Lazaridis, M. & Katsivela, E. (2012). Size distribution of viable, cultivable, airborne microbes and their relationship to particulate matter concentrations and meteorological conditions in a Mediterranean site. *Aerobiologia*, **29**: 233-248.

Rajasekar, A. & Balasubramanian, R. (2011). Assessment of airborne bacteria and fungi in food courts. *Build. Environ.*, **46**: 2081-2087.

Reponen, T., Nevalainen, A., Willeke, K. & Grinshpun S.A. (2009). Biological particle sampling. *In* Baron, P.A, Willeke, K. & Kulkarni P. (Eds.), *Aerosol Measurement: Principles, Techniques and Applications.* 3rd ed., John Wiley and Sons, Inc., New York, Chapter 24.

Rösch, P., Harz, M. & Schmitt, M. (2005). Chemotaxonomic identification of single bacteria by micro-Raman spectroscopy: application to clean-room-relevant biological contaminations. *Appl. Environ. Microbiol.,* **71**: 1626–1637.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. & Arnheim, N. (1985) Enzymatic amplication of B-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Sci.*, **230**:1350-1354.

Sasser, M. (1990a). Identification of bacteria through fatty acid analysis. *In* Klement, Z., Rudolph, K. & Sands, D.C. (Eds.), *Methods in Phytobacteriology*. AkadémiaiKiadó, Budapest, Hungary, pp. 199-204.

Sasser, M. (1990b). *Identification of Bacteria by Gas Chromatography of Cellular Fatty Acids*. Newark, DE, Microbial Identification, Inc. (MIDI), Technical Note #101.

Satomi, M., La, D. & Venkateswaran, K. (2006). Bacillus safensis sp. nov., isolated from spacecraft and assembly-facility surfaces. *Int. J. Syst. Evol. Micr.,* **56**: 1735-1740.

Sepahvand, A., Shams, G.M., Allameh, A. & Razzaghi, A.M. (2013). Diversity and distribution patterns of airborne microfungi in indoor and outdoor hospital environments in Khorramabad Southwest Iran. *Jundishapur J. Microbiol*., **6**: 168-192.

Sharma, K. (2011). Concentration and species diversity of airborne fungi of Dongargarh. *Int. Multidiscip. Res. J.,* **1**: 34-36.

Shivaji, S., Chaturvedi, P., Suresh, K., Reddy, G.S.N., Dutt, C.B.S., Wainwright, M., Narlikar, J.V. & Bhargava, P.M. (2006). Bacillus aerius sp., bacillus aerophilus sp. nov., bacillus stratosphericus sp. nov. and bacillus altitudinis sp. nov., isolated from cryogenic tubes used for collecting air samples from high altitudes. *Int. J. Syst. Evol. Micr.,* **56**: 1465-1473.

Sonenshein, A.L., Losick, R. & Hoch, J.A. (2002). *Bacillus subtilis and Its Closest Relatives: from Genes to Cells.* ASM Press, Boston, United States.

Willeke, K. & Macher, J.M. (1999). Air sampling. *In* Macher, J.M. (Ed.), *Bioaerosols: Assessment and Control.* American Conference of Governmental Industrial Hygienists, Inc., Cincinnati, Ohio, pp. 11-25.

Womack, A.M., Bohannan, B.J. & Green, J.L. (2010). Biodiversity and biogeography of the atmosphere. *Philos. Trans. R. Soc. Lond.,* **365**: 3645-3653.

Yoon, J.H., Kim, I.G., Kang, K.H., Oh, T.K. & Park, Y.H. (2003). Bacillus marisflavi sp. nov. and bacillus aquimaris sp. nov., isolated from sea water of a tidal flat of the yellow sea in Korea. *Int. J. Syst. Evol. Micr.*, **53**: 1297-1303.

Zhai, Y., Li, X., Wang, T., Wang, B., Li, C. & Zeng, G. (2018). A review on airborne microorganisms in particulate matters: composition, characteristics and influence factors. *Environ. Int.*, **113**: 74-90.

Zhou, Y., Yang, J., Zhang, L., Zhou, X., Cisar, J.O. & Palmer, R.J.J. (2016). Differential utilization of basic proline-rich glycoproteins during growth of oral bacteria in saliva. *Appl. Environ. Microbiol.,* **82**: 5249-5258.

ASSESSMENT OF IMMUNE FUNCTION IN WELL TRAINED MILITARY PERSONNEL AFTER STRENUOUS PHYSICAL ACTIVITY IN A TROPICAL ENVIRONMENT

Raja Zarith Fatiah[1], Victor Feizal Knight[1], Brinnell Caszo[2], Justin Gnanou[2*] & Ananthan Subramaniam[3]

[1]Faculty of Medicine and Defence Health, National Defence University of Malaysia (UPNM), Malaysia
[2]School of Medicine, International Medical University, Malaysia
[3]Centre of Tropicalisation, National Defence University of Malaysia (UPNM), Malaysia

[*]Email: justingnanou@gmail.com

ABSTRACT

Military training involves activities that are strenuous and requires high endurance. Strenuous activity is associated with depression of immune functions and can lead to immune function defects. This can compromise the performance of the military personnel in the field. Thus, we aimed to study the pattern of urinary cytokine levels in military personnel after a 5 km jungle trek with full battle gear for five consecutive days. Six male army physical instructors of similar age (32.7 ± 2.2 years) were recruited for the study. After collection of a baseline urine sample, the subjects completed a 5 km jungle trek with full battle gear in a tropical environment. At the end of the jungle trek, another urine sample was collected. The same protocol was repeated for five days and at the end of the jungle trek on the Day 5, another urine sample was collected. The urine samples were analysed for the cytokines (IL-1 β, IL-6, TNF-α, IL-1ra and IL-10). The Wilcoxon-Mann-Whitney test was used to compare the urine cytokines between the baseline, and the end of Days 1 and 5. Our study showed that pro-inflammatory cytokines 1L-1β and IL-6 showed a steady increase from baseline to Day 5 and remained persistently high even at the end of the five days' of the jungle trek. Anti-inflammatory cytokine IL-1ra showed an initial increase, but was decreased by Day 5, while no changes were seen in IL-10 levels. On the whole, studying cytokine secretion pattern would help us in understanding the body's response to strenuous activities conducted by military personnel.

Keywords: *Interleukins; urinary cytokines; strenuous exercise; military training; tropical environment.*

1. INTRODUCTION

Cytokines are hormone-like molecules that help regulate and mediate our immune response to injury at the cellular level. They are classified based on their function as either pro-inflammatory or anti-inflammatory interleukins (IL). Pro-inflammatory cytokines include IL-1, IL-6, IL-1α and IL-1 β. Of these cytokines, IL-6 is considered the most important of the pro-inflammatory cytokines, while IL-1 beta is an acute phase cytokine. Anti-inflammatory cytokines include interleukin 1 receptor antagonist (IL-1ra), IL-4 and IL-10. These cytokines are released in response to pro-inflammatory cytokines and their action is to inhibit the release of pro-inflammatory cytokines and thus stop the inflammatory process. A classic example is IL-1ra, which is the receptor antagonist to the IL-1 family of cytokines (Turner *et al.*, 2014)

Studies have found that prolonged and strenuous exercise leads to an increase in the activities of pro-inflammatory cytokines, such as tumour necrosis factor-α (TNF-α), 1L-1β and IL6. This is followed by increase of anti-inflammatory cytokines, such as IL10 and IL-1ra in order to balance the pro-inflammatory cytokines (Markovitch *et al*., 2008; Nielsen *et al*., 2016). Strenuous exercise is associated with muscle soreness and injury, which provokes the release of pro-inflammatory cytokines. This pattern of cytokine release is responsible for the rapid migration of neutrophils to the site of muscle injury and followed later by monocytes to aid in the healing process (Prame Kumar *et al*., 2018).

Strenuous physical activity and exertion is an inherent part of military training. Thus, one would expect a milieu of cytokine activity in military personnel undergoing training. Whilst cytokine production and inhibition is a normal process of the cellular and tissue protection and healing, the pattern and production of cytokines can vary due to many external factors, such as environmental conditions in which the training is conducted (Izquierdo *et al*., 2009). A tropical environment with high humidity and temperature can influence the type and amount of cytokines produced by the body (Cosio-Lima *et al*., 2011). Characterisation of the cytokine pattern can be useful to optimise training protocols. Thus, in this study, we aimed to quantify urinary cytokine levels in military personnel after a 5 km jungle trek with full battle gear in a tropical jungle environment for five consecutive days.

2. METHODOLOGY

2.1 Baseline Measurements

Six male army physical instructors of similar age were recruited for the study. Table 1 provides the subject characteristics. The subjects were briefed about the study protocol and an informed consent was obtained from the subjects. The study protocol was reviewed by the University Research Committee and the study was conducted in accordance with the Declaration of Helsinki and the guidelines set by Resolution 198/96 of the National Health Council (World Medical Association, 2001).

Table 1: Subject characteristics.

Characteristics	Average	Standard Deviation
Age (years)	32.7	2.2
Height (cm)	166.0	5.0
Weight (kg)	69.0	5.0

2.2 Experimental Design

The study protocol consisted of a 5 km jungle trek within the complex of the National Defence University of Malaysia (UPNM). It is built in a dense tropical jungle forest with a typical tropical climate and environment. At the start of the experiment, the subjects were allowed to be familiarised with the jungle trek course. The experiment protocol consisted of trekking the jungle in full battle field gear for five consecutive days. On the first day of the jungle trek, baseline urine samples were collected. Then, the subjects were asked to don full battle gear and undertake the 5 km jungle trek. At the end of the course, urine samples were collected again. A similar jungle trek and urine protocol were repeated for Days 2 to 5. The final urine samples were collected at the end of Day 5 of the jungle trek.

2.3 Biochemical Analysis

The urine samples collected were stored at -80°C until analysis. On the day of the analysis, the urine samples were thawed and centrifuged, and the supernatant was used for analysis. The urine samples were used for the analysis of five different types of cytokines, namely, IL-1β, IL-6, TNF-α, IL-1ra, and IL-10. The cytokines were measured using commercially available enzyme linked immunosorbent assay kits (Cloud-Clone Corp., Texas, USA). The protocols of the respective kits were followed and absorbance was measured using a SpectraMax 5M (Molecular Devices, CA, USA) analyser. Calibration curves were generated using a four-parameter logistic regression analysis via the SpectraMax software (Molecular Devices, CA, USA).

2.4 Statistical Analysis

Tests of normality were conducted and a non-parametric test (Wilcoxon-Mann-Whitney test) was chosen to compare the urine cytokines between baseline, and end of Days 1 and 5. A p-value of <0.05 was considered significant.

3. RESULTS AND DISCUSSION

3.1 Interleukin Assay Results

The results of our study showed a 40 and 12% increase in pro-inflammatory cytokines IL-1β and IL-6 at the end of the five days of the jungle trek respectively. Both these cytokines also showed a gradual increase from baseline, and end of Days 1 and 5. However, TNF-α, another pro-inflammatory cytokine did not show any increase. Anti-inflammatory cytokine IL-1ra showed a 65% decrease, while IL-10 showed a 13% increase. Similar to cytokines IL-1β and IL-6, the gradual increase and decrease of IL-1ra and increase of IL-10 was also noted (Figures 1 and 2). However, these changes at the end of the fifth day were not statistically significant for all the cytokines.

3.2 Discussion on Pro-Inflammatory Cytokines - TNFα, IL-6, and IL-1beta (Figure 1)

Pro-inflammatory cytokines are secreted by activated macrophages and play a central role in mediating and enhancing the inflammatory process. These cytokines facilitate the influx of lymphocytes, monocytes and neutrophils to region of tissue injury and thus promote tissue healing (Dinarello, 1992). There is also evidence that shows pro-inflammatory cytokines, such as TNFα, IL-6 and IL-1β, are involved in the regulation of pain (Zhang & An, 2007). The results from our study show that the concentration of IL-1β, a pro-inflammatory cytokine increased gradually over the five consecutive days of strenuous exercise among the military personnel. This gradual increase can be due to the effect of strenuous exercise on the inflammatory pathway. Strenuous exercise is known to induce skeletal muscle damage as well as DNA damage (Zainuddin *et al*., 2019). This may lead to activation of the inflammatory pathway and secretion of IL-1β by monocytes and macrophages as a defensive mechanism by the injured cells (Peake, 2002). Existing literature also suggests non-immune cells, such as fibroblast and endothelial cells, also have the ability to secrete IL-1β (Teijaro *et al*., 2011; Peake *et al*., 2015). Once secreted by these cells, IL-1β acts as a priming agent for the initiation of a downstream cascade of events leading to tissue fibrosis. It is also important to note that normal healthy cells do not normally express IL-1β, and cell injury acts as a trigger for its secretion (Lopez-Castejon & Brough, 2011). Cytokine profiling following strenuous exercise has been extensively studied, but the findings are not consistent. Ostrowski *et al*. (1999) found a two-fold increase in IL-1β after a marathon race, but Nielson *et al*. found no changes in IL-1β after a marathon run. Our study involved five days of consecutive strenuous exercise. We found at the end of day 1 that there

was no difference in IL-1β from the baseline. However, at the end of the fifth day, a 40% increase was noted. This shows that IL-1β does not increase immediately, but over a period of time and with constant injury elicited on the muscle tissue, IL-1β secretion increases and initiates inflammation. This is seen in a study on ranger-training course, where IL-1β showed a small increase during the course, however, IL-1β remained increased during the post-exercise period (Nielsen *et al*., 2016).

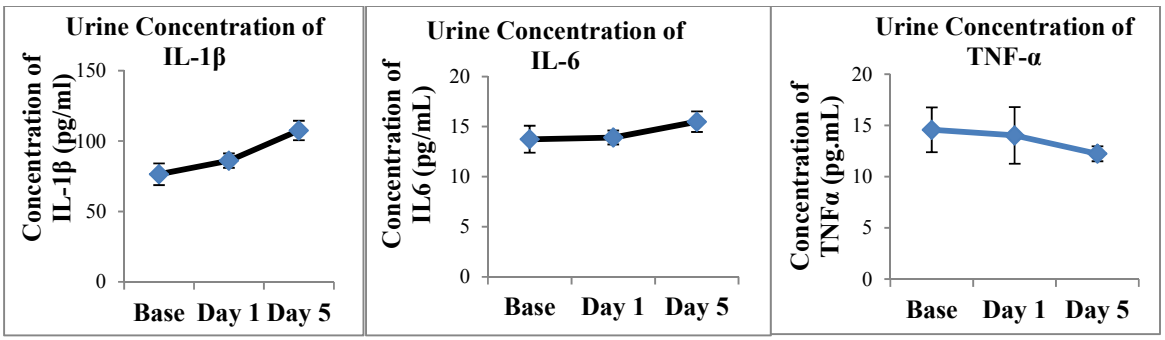


Figure 1: Pro-inflammatory cytokine response for the baseline, and Days 1 and 5 of the experiment protocol.

IL-6 is another important pro-inflammatory cytokine apart from IL-1β in the inflammatory pathway. It is considered both as a pro-inflammatory cytokine and an anti-inflammatory myokine (Petersen & Pedersen, 2005). Increased levels of IL-6 have been associated with muscle damage following strenuous exercise (Kim *et al*., 2015). Its increase following strenuous exercise is considered beneficial as it promotes the healing process (Pedersen *et al*., 2001). However, increased IL-6 following sustained strenuous exercise can lead to impairment of immune function through decrease in mucosal immunity. In a study conducted on military personnel after a five day military training regime, high circulating levels of IL-6 were found and linked to decrease in mucosal immunity (Nielsen *et al*., 2016). This finding is similar to our study, which showed a 12% increase at the end of five days of strenuous jungle trek.

Unlike IL-1β and IL-6, the concentration of TNF-α was reduced, though not statistically significant by the end of Day 5 as compared to the baseline. This finding was against our preliminary hypothesis where we anticipated increase in all the pro-inflammatory cytokines. A number of studies have found increase in TNF-α levels following marathon runs, strenuous exercise as well as acute exercise (Ostrowski *et al*., 1999; Moldoveanu *et al*., 2000). However, there are studies that have shown no change in TNF-α levels after marathon races (Drenth *et al*., 1995). This finding is explained by the fact that high levels of anti-inflammatory cytokines, such as IL-10, could have prevented the secretion of pro-inflammatory cytokines including IL-6 (Nielsen *et al*., 2016). However, this was not the case in our study. In our study, we found increase in IL-1β and IL-6, but almost no increase in IL-10. We feel that exercise induced increase in catecholamines, acting through beta- adrenergic receptors, that caused this decrease in TNF-α secretion. Yano *et al*. (2010) in a study on TNF-α production following pathogen stimulation by strenuous exercise found that the suppression of TNF-α depends on the translation of TNF-α mRNA in the tissues.

3.3 Discussion on Anti-Inflammatory Cytokines - IL-1ra and IL-10 (Figure 2)

Anti-inflammatory cytokines, such as IL-1ra and IL-10, have been shown to reduce the cytokine mediated inflammatory responses. Following cell injury, IL-6 is known to stimulate the secretion of IL-1ra, which in turn binds to and blocks IL-1 receptors, and thus exerts anti-inflammatory effects. Thus, IL-1ra levels usually rise much later after IL-6 secretion (Duzova *et al*., 2009). Mild to moderate exercise causes IL-1ra to increase about 2 to 3 h after the peak of IL-6 (Duzova *et al*., 2009). In our study, we found a 65% decrease in levels of IL-1ra at the end of five days of strenuous exercise. Initially at the end of Day 1,

there was a gradual increase, but by the fifth day, the levels were markedly reduced. This could be explained by the fact that increase in IL-6 over the five day period was only 12% and this increase would not have been sufficient to sustain the stimulation to secrete IL-1ra. IL-10 is also known to cause inhibition of IL-1ra and TNF-α secretion, but in our study we did not find an increase in IL-10. The main function of IL-10 is to inhibit the activation of T cells, NK cells and macrophages. Thus, IL-10 by orchestrating the inflammatory responses acts as a negative feedback to limit the activation of these immune cells. However, in our study, this function of IL-10 was absent and thus caused pro-inflammatory cytokines to rise constantly.

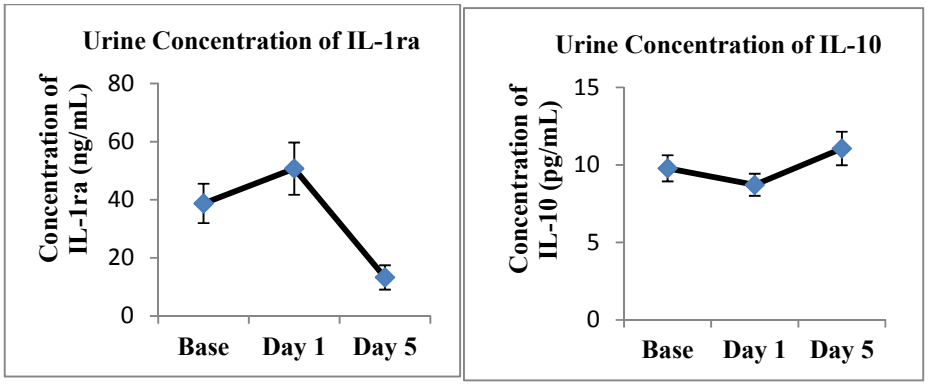


Figure 2: Anti-inflammatory cytokine response for the baseline, and Days 1 and 5 of the experiment protocol.

4. CONCLUSION

In conclusion, the present study provides new data on cytokine profiling of a strenuous military activity in a tropical jungle environment. We found there was a steady increase in pro-inflammatory cytokines mainly mediated by IL-1β. Even though IL-6 is considered as a main pro-inflammatory cytokine responsible for inducing anti-inflammatory cytokine secretion, our results showed that increase in IL-6 was not sufficient to induce secretion of anti-inflammatory cytokines. Instead, we found that anti-inflammatory cytokines decreased as the strenuous physical activity continued. Thus, further studies are required to look into the role of IL-1β in long term strenuous physical activity and strategies to inhibit IL-1β might prove beneficial in accelerating the healing process during the recovery period.

REFERENCES

Cosio-Lima, L.M., Desai, B.V., Schuler, P.B., Keck, L. & Scheeler, L. (2011). A comparison of cytokine responses during prolonged cycling in normal and hot environmental conditions. *Open. Access. J. Sports. Med*., **11**:7-11.

Dinarello, C.A. (1992). Role of interleukin□1 in infectious diseases. *Immunol. Rev*., **127**, 119–146.

Drenth, J.P., van Uum, S.H., van Deuren, M., Pesman, G.J., van der Ven-Jongekrijg, J. & van der Meer, J.W. (1995). Endurance run increases circulating IL-6 and IL-1ra but downregulates ex vivo TNF-alpha and IL-1 beta productions. *J. Appl. Physiol*., **79**:1497-1503.

Duzova, H., Karakoc, Y., Emre, M.H., Dogan, Z.Y., Kilinc, E. (2009). Effects of acute moderate and strenuous exercise bouts on IL-17 production and inflammatory response in trained rats. *J. Sports Sci. Med*., **8**:219–224.

Izquierdo, M., Ibanez, J., Calbet, J.A., Navarro-Amezqueta, I., Gonzalez-Izal, M., Idoate, F., Hakkinen, K., Kraemer, W.J., Palacios-Sarrasqueta, M., Almar, M. & Gorostiaga, E.M. (2009). Cytokine and hormone responses to resistance training. *Eur. J. Appl. Physiol*., **107**: 397-409.

Kim, H.K., Konishi, M., Takahashi, M., Tabata, H., Endo, N., Numao, S., Lee, S.K., Kim, Y.H., Suzuki, K. & Sakamoto, S. (2015). Effects of acute endurance exercise performed in the morning and evening on inflammatory cytokine and metabolic hormone responses. *PLoS One*., **10**:e0137567.

Lopez-Castejon, G. & Brough, D. (2011). Understanding the mechanism of IL-1β secretion. *Cytokine Growth Factor Rev*., **22**:189–195.

Markovitch, D., Tyrrell, R.M. & Thompson, D. (2008). Acute moderate-intensity exercise in middle-aged men has neither an anti nor pro-inflammatory effect. *J. Appl. Physiol*., **105**: 260-5.

Moldoveanu, A.I., Shephard, R.J. & Shek, P.N. (2000). Exercise elevates plasma levels but not gene expression of IL-1beta, IL-6, and TNF-alpha in blood mononuclear cells. *J. Appl. Physiol*., **89**: 1499-1504.

Nielsen, H.G., Øktedalen, O., Opstad, P.K. & Lyberg, T. (2016). Plasma cytokine profiles in long-term strenuous exercise. *J. Sports Med*., **2016**:7186137.

Ostrowski, K., Rohde, T., Asp, S., Schjerling, P. & Pedersen, B.K. (1999). Pro and anti-inflammatory cytokine balance in strenuous exercise in humans. *J. Physiol*., **515**: 287-91.

Peake, J. M. (2002). Exercise-induced alterations in neutrophil degranulation and respiratory burst activity: possible mechanisms of action. *Exerc. Immunol. Rev*., **8**:49–100.

Peake, J.M., Gatta, P.D., Suzuki, K. & Nieman, D.C. (2015). Cytokine expression and secretion by skeletal muscle cells: regulatory mechanisms and exercise effects. *Exerc. Immunol. Rev*., **21**: 8–25.

Pedersen, B.K., Steensberg, A. & Schjerling, P. (2001). Muscle-derived interleukin-6: possible biological effects. *J. Physiol*. 536:329–337.

Petersen, A.M. & Pedersen, B.K. (2005). The anti-inflammatory effect of exercise. *J. Appl. Physiol*., **98**: 1154-1162.

Prame Kumar, K., Nicholls, A.J. & Wong, C.H.Y. (2018). Partners in crime: neutrophils and monocytes/macrophages in inflammation and disease. *Cell. Tissue. Res*., **371**: 551-565.

Teijaro, J. R., Walsh, K. B., Cahalan, S., Fremgen, D. M., Roberts, E., Scott, F., Martinborough, E., Peach, R., Oldstone, M.B. & Rosen, H. (2011). Endothelial cells are central orchestrators of cytokine amplification during influenza virus infection. *Cell*., **146:** 980–991.

Turner, M.D., Nedjai, B., Hurst, T. & Pennington, D.J. (2014). Cytokines and chemokines: At the crossroads of cell signalling and inflammatory disease. *Biochim. Biophys. Acta*., **1843**:2563-2582.

World Medical Association. (2001). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *Bull. World Health Organ*., **79**: 373-374.

Yano, H., Uchida, M., Nakai, R., Ishida, K., Kato, Y., Kawanishi, N. & Shiva, D. (2010). Exhaustive exercise reduces TNF-α and IFN-α production in response to R-848 via toll-like receptor 7 in mice. *Eur. J. Appl. Physiol*., **110**:797-803.

Zainuddin, H., Caszo, B., Knight, V.F. & Gnanou, J. (2019). Training induced oxidative stress-derived DNA and muscle damage in triathletes. *Eurasian J. Med*., **51**: 116–120.

Zhang, J.M. & An, J. (2007). Cytokines, inflammation, and pain. *Int. Anesthesiol. Clin*., **45**: 27–37.

THE MULTIDIMENSIONAL IMPACT OF CBRNe EVENTS ON HEALTH CARE IN THE MIDDLE EAST: THE ROLE OF EPIDEMIOLOGICAL SURVEILLANCE IN THE LONG-TERM RECOVERY OF PUBLIC HEALTH SYSTEMS

Stefania Moramarco[1*], Leonardo Palombi[1], Faiq B. Basa[1,2] & Leonardo Emberti Gialloreti[1]

[1]Department of Biomedicine and Prevention, University of Rome Tor Vergata, Italy
[2]Rizgary Hospital, Erbil, Kurdistan Region, Iraq

*Email: stefania.moramarco@gmail.com

ABSTRACT

Chemical, biological, radiological, nuclear and explosive (CBRNe) hazards used as agents during conflicts and terrorism are a threat to any civil society, negatively impacting economical, human, political and environmental aspects. When a country is stricken by CBRNe, several related events can significantly debilitate the public health system's assets, placing additional demands on the health-care organization of the country and neighbors. The consequences for the health of a population might be both direct, such as deaths and injuries, and indirect with long-term aftereffects such as disruption of basic health services, damages to infrastructures and lack of medical personnel. This article provides an overview on the multidimensional impact of CBRNe events on the healthcare sector in the Middle East (Afghanistan, Iraq, Lebanon and Syria) and the countries' efforts in setting up the basis for health system recovery. Our main aim is to emphasize that the rebuilding of an effective healthcare system is a long-term process that requires multiple actions and actors, with epidemiological surveillance being the cornerstone. We maintain that reinvesting in effective health monitoring systems is essential to support countries in meeting the health needs of their populations in post-CBRNe events. This is the first necessary step for appropriately allocating resources, driving investment, setting up preventive strategies and decision-making responses. Furthermore, especially when referring to fragile states, acquiring knowledge on the health needs of the threatened population holds a paramount importance for providing preparedness planning and enhancing resilience of the public health system in case of future CBRNe events.

Keywords: *CBRNe events; epidemiological surveillance; health monitoring system; health system recovery; Middle East*.

1. THE IMPACT OF CBRNe EVENTS ON HEALTHCARE

The use of chemical, biological, radiological, nuclear, and explosive (CBRNe) agents during armed conflicts, as an act of terrorism or when accidents happen, can have complex and multidimensional impact on a country's structure, affecting from economic, political, social and health sectors to food security (Moramarco, 2018). Specifically, when referring to the health sector, all facets of a health system are likely to be stricken, especially during protracted long-lasting events: basic health delivery services are jeopardized through general damage and specific destruction of health structures - as recently happened in Iraq (The Guardian, 2016) and Syria (The Washington Post, 2018) - with health facilities sometimes becoming specific targets of strikes. Consequently, human capital and medical staff are suppressed, with medical personnel directly killed or forced to flee, looking for better life conditions and a more comfortable environment to work in. The crippling of healthcare infrastructure compromises routine health delivery systems - at times vaccines and other lifesaving drugs have been intentionally blocked from reaching civilians (The National, 2018) - creating a population susceptible to potential disease outbreaks (Nnadi *et al.,* 2017). When outbreaks occur in these settings, they rapidly spread, often with a high probability of prolonged transmission, owing to unavailable treatment facilities, and poorly trained or lack of health personnel, coupled with population displacements. These are all events that significantly heighten the risk of disease transmission, as well as

increase the morbidity and mortality rates of the affected population (Murray *et al*., 2002). Indeed, populations that have experienced such events present the worst indicators of infant, child (Black *et al*., 2003) and maternal mortality (Johnson, 2017). In addition, even neighboring countries are threatened not only because of the direct consequences of CBRNe, but also for their indirect impact: people forced to flee increase the number of displaced people and refugees, putting more pressure on the health services of the hosting countries; and outbreaks of infectious diseases can easily spread, representing a global public health hazard (UN News, 2018). Refugee camps experience the highest rates of infection and case fatality due to scarce resources (Brown *et al*., 2002).

The countries' first priority becomes limiting injuries and deaths, with public investments inevitably drawn away from sustainable development goals and redirected to emergency response activities, but also to military and defensive pursuits. In such settings of security and time limitations, assessing healthcare (both health needs of the population and availability of health services) is not easy. Lack of systematic information gathering hampers effective decision-making, mobilization, resource allocation and advocacy for health. This complex scenario can last even long after the emergency cessation.

2. LONG-TERM RECOVERY OF HEALTH SYSTEMS IN POST-CBRNe EVENTS

The recovery of health systems after direct CBRNe events or their indirect consequences is a complicated endeavor and a long-term process that requires multiple human and financial investments (Rutherford & Saleh, 2019). Recovery is defined as the process of rebuilding, restoring and rehabilitating the public health system following an emergency; it includes remediation techniques that may be applicable to facilitate the return to normality, as well as moving to self-sufficiency, sustainability and resilience. While the immediate response to an emergency can be relatively short, recovery is a long-running process that continues until the disruption is rectified, the needs of the affected are met, and the demands of services are satisfied (HM Government, 2013). This long-term phase can last months or even years, as its aim is to revitalize, rebuild and repopulate / re-use affected areas, favoring the progress to self-sufficiency, sustainability and resilience.

Recovery should be considered not only as a response to a specific disaster, but since the latter might be due to lack of preparedness, it has to be considered also as the basis for future readiness. Preparedness includes plans to be prepared for, and respond to, a wide range of hazards and threats, essential to save lives and to facilitate early response and timely recovery operations in case of future emergencies. It is, indeed, important to provide continuity plans integrated into wider emergency plans (Boyd *et al.,* 2013). Therefore, monitoring the needs of populations in terms of health and availability of health services - tracking their evolution over time - requires tools for information gathering and data analysis in order to drive effective decision-making, as well as mobilization and resource allocation.

2.1. Epidemiological Surveillance as the Basis for Health System Recovery

Especially in fragile states, the reconstruction of a functional public health system – which includes stabilization initiatives and investments in reestablishing or in some cases establishing for the first time a system of health services for the population - is a key element to address both the acute and ongoing health needs, to guide decisions, and to optimize future activities (OECD, 2011). The priorities and resources devoted to planning, preparedness, and response are of paramount importance, and the only way to appropriately allocate them is conducting a reliable initial situation analysis and a routinely epidemiological surveillance[1]. As applied to public health, the term "surveillance" means the close monitoring of the occurrence of selected health conditions in the population (Berkelmann *et al*., 1997). In 1963, disease surveillance was defined as "*the continued watchfulness over the distribution and trends of incidence through systematic collection, consolidation, and evaluation of morbidity and mortality reports and other relevant data*" (Langmuir, 1963), which implies information for action (World Health Assembly 21, 1968).

[1] "*Epidemiological Surveillance is the ongoing, systematic collection, analysis, interpretation, and dissemination of data about health-related events for use in public health action to reduce morbidity and mortality and to improve health*" (Centers for Disease Control and Prevention, 2001).

Surveillance data can be in fact also used to detect changes in health practices, evaluate control measures, and describe the natural history of a health event in a community. Therefore, since it is crucial to identify the important contextual variables that influence the country's assets and development, surveillance requires commitment to data collection. Its ultimate goal is the formulation of public health policies to promote health and prevent diseases (Bonita *et al*., 2003).

After CBRNe events, as well as during conflicts (especially when protracted), affected local governments are often deeply weakened, having usually limited capacity and resources to manage on their own all the necessary reconstruction tasks. Therefore, public health recovery is generally also supported by private or international donors (Rubenstein, 2009). These principles go along with building the capacity of the local health authorities to engage in the essential tasks of leadership, planning and oversight.

3. EVIDENCE FROM MIDDLE EAST COUNTRIES

Examples of the devastating effects of CBRNe events on countries' asset come from the Middle East. The location of the region's confines, generally on the eastern side of Mediterranean Sea, has always led to some confusion over changing definitions (Hazbun, 2012). In the most common definition, usually the region comprises of Afghanistan, Bahrain, Iran, Iraq, Jordan, Kuwait, Lebanon, Oman, Pakistan, Palestine, Israel, Qatar, Saudi Arabia, Syria, Turkey, the United Arab Emirates, and Yemen. Historically, the Middle East has been one of the most critical regions of the world. Crises in the region resulted from various reasons, and led to long-lasting conflicts and terrorism attacks, most of which have involved CBRNe agents, or have been triggered by their real or perceived threats. Such events had a major impact on health systems and, consequently, on populations' health (Coutts *et al.,* 2013). A recent Global Burden of Disease Study (GBD 2013) analyzed the burden of disease and injuries in the eastern Mediterranean region as of 2013, when the region faced unrest as a result of revolutions and wars during the so-called Arab uprisings (Mokdad *et al.*, 2016). The study showed that the eastern Mediterranean region was going through a critical period of the health sector, with a call for increasing health investments in the region, in addition – of course - to reducing conflicts.

In the Thirteenth General Program of Work for 2019–2023 (GPW 13), the World Health Organization (WHO) emphasized the need for reliable and timely health information in order to ensure healthcare delivery, health policy development and implementation, with the strengthening of health information systems being a priority in the region (WHO, 2018).

3.1 The Case of Afghanistan

Afghanistan had a reasonably functioning health sector before the Soviet invasion in 1979, when the country entered into a prolonged war. There have been reports that both Soviet and Mujahedeen forces used various types of chemical agents (US Department of State, 1982). Such reports were mainly based on testimonies and on some symptoms of the victims; thus, they have never been confirmed. Details and official documentation on the use of these agents are indeed still lacking (Schwartzstein, 1982). Following the Soviet withdrawal of 1988-1989, Afghanistan plunged into a civil war, resulting in the rise of the Taliban regime, which exacerbated the already compromised situation. After the September, 11 terrorist attacks in 2001, the US-led intervention resulted in the fall of the Taliban regime. In 2002, when the transitional government, with the support of the international community, tried to re-establish systems and services in all sectors, it became clear that physical and social assets were severely damaged, although little information was available about their extent (Fujita *et al.,* 2011).

Today, the country is still facing challenges and fragility. Several chemical attacks have been reported (55 between 2008 and 2013, with 55 people killed and 2,683 injured), but not all of them were confirmed (Johnston's Archive, 2015). For example, in 2012, 16 poison attacks on high schoolgirl in the Northern Region, Takhar province, were reported. Several – may be all - of these cases seemed to involve poisoning of the school's water supply; overall, 1,355 children and 28 teachers and staff were reported injured in these

attacks. However, WHO inspectors did not confirm the presence of a poison, alleging that the cases might be the result of mass hysteria (Johnson, 2015).

As a consequence of all these events, many health infrastructures were damaged and the health system almost stopped working (Cook, 2003). Afghanistan's health system has become one of the worst in the world, with destroyed public structures, reduced technical and health staff (killed or forced to live in exile), and no investments made in the health sector. One of the major challenges has been the unavailability of medicines and the low quality of the few available drugs. In 2002, health indicators were very poor, including a maternal mortality ratio of 1,600 per 100,000 live births, an infant mortality rate of 165 per 1,000, and an under-five mortality of 257 per 1,000 (Waldman & Hanif, 2002; Bartlett *et al.,* 2002).

The post-conflict situation provided a unique opportunity to redesign and rebuild the health system, almost from scratch, with the Ministry of Public Health (MoPH) promising on various occasions that there would be extra attention (The Daily Outlook Afghanistan, 2017). One of the earliest activities in the health sector was the national health resources assessment, in 2002, when available resources in the country (human, material, financial) were mapped (Ministry of Health Transitional Islamic Government of Afghanistan, 2002)[2].

To reverse this decline, the MoPH began working with the international community and civil society, first to provide emergency health services to a long-suffering population, and then to reconstruct the health system. The establishment of a "Basic Package of Health Services", with a major expansion of the availability of primary health services and the reach of vaccination programs, started to bring important improvements in key health indicators (Ministry of Health, Transitional Islamic Government of Afghanistan, 2003). To give an example, in 2010 infant mortality fell to 77 per 1,000, while under-five mortality decreased to 97 per 1,000 (USAID, 2014). Nowadays, the WHO is still providing technical assistance to the MoPH for the implementation of the National Disease Surveillance and Response system (NDSR), including the Early Warning component for responding to epidemics outbreaks.

3.2 The Case of Iraq

In the 1970s, the Iraqi health system was one of the best and most advanced in the Middle East (Middle East Health Magazine, 2012). However, its capacity and performance started to deteriorate during the 1980s, then further worsening in the last few decades as a result of continuous wars, sanctions, loss of health workers, looting, political interference, and economic sanctions. Some of those events have been a direct consequence of CBRNe events or their perceived threats. During the Iran-Iraq War (1980-1988), the use of *samarium*, *tabun*, and *mustard gas* on a large scale was reported against both Iran and the Kurdish populations in northern Iraq. For a period of about five years from 1983 to 1988, the Iraqi army used chemical weapons several times (Razavi *et al*., 2014). The worst attack was conducted in 1988 on the Iraqi Kurdish town of Halabja. The event was also known as the Halabja Massacre or Bloody Friday,[3] since it was one of the deadliest chemical attacks in history after the First World War (Hiltermann, 2007). The chemical gasses lasted some 45 minutes, causing 6,000 civilians deaths, 5,000 of them within less than 10 minutes (eight times more than the daily victims of one of the big epidemics of the middle ages) (Mohamed-Ali, 1992; The New York Times, 2003). Severe damage was reported in people exposed - between 7,000 to 10,000 - most of them civilians (The Times, 2010). Before and after the attack on Halabja, poison gas was used in various other towns and villages in the border regions with Iran and Turkey during the so-called Anfal Campaign (Mlodoch, 2017). Those events triggered the involvement of multinational military operations, from the US-led, which dragged the country into recurring and long-lasting conflicts, up to the recent ISIS (the so-called Islamic State) occupations. Nowadays, Iraq is still one of the world's most landmine-affected countries: years of conflicts have left landmines, cluster munitions, and unexploded bombs all over, preventing

[2] One of the main findings was the limited number of female health professionals (approximately one female doctor, one female nurse and one midwife for every 50,000 individuals). This was a major constraint, which resulted from the post-Taliban society.

[3] The attack was conducted by Saddam Hussein's forces in the final months of the eight-year Iraq-Iran war and was also part of Iraqi efforts to counter-attack Kurdish militias forces, supported by Iran.

communities from using their land and displaced populations from returning home (Mine Action Review, 2018).

One of the results of this complex scenario was that health indicators fell over the years to levels comparable to some of the least developed countries. Most updated health-care data, when available, confirms that the Iraqi population is still paying a crippling price for the continuing violence, political infighting and widely acknowledged rampant corruption within the health system (Webster, 2011). Over the years, the number of hospital beds in Iraq declined from a high of 1.95 per 1,000 people in 1970 to a low of 1.30 per 1,000 people in 2012 (Trading Economics, 2019). As a consequence of lack of proper sanitation system and inadequate timely response, Iraq experienced recurring cholera outbreaks in 2007, 2009, 2012, and 2015. In October 2007, the outbreak spread to 9 out of the 18 provinces across Iraq, with more than 3,315 blood samples testing positive for *Vibrio cholerae*, the bacterium causing the disease. Fourteen people died from the disease. The most affected areas were Kirkuk (2,309 cases) and Sulaymaniyah (870 cases). In 2015, Iraq faced an outbreak of cholera that started in September along the Euphrates valley. A total of 2,810 laboratory-confirmed cases of cholera, from 17 governorates, were reported by WHO from mid-September to November, with two related deaths (WHO, 2015). The health system received a shock after the ISIS took over about one third of the country in 2014. Nowadays, in spite of the defeat of ISIS and much rebuilding, health infrastructure is still not fully restored (Devi, 2018). In terms of areas of displacement, at the end of 2018, the Kurdistan Region of Iraq was the area hosting the largest number of internally displaced people - IDPs: 1.8 million people, a nearly 30% increase in the population of the region (IOM, 2018). The increasing number of displaced people and, consequently, their disease caseload poses new challenges to the local government and humanitarian agencies in the provision of health care, diagnostics and medications. In particular, concern has been raised on the access to care among minority groups displaced by ISIS, such as the Yazidis (Cetorelli *et al.,* 2017). Moreover, the latest offensive of the Turkish army against the Kurdish population in Syria is a rising concern for new refugee flowing to the Iraqi Kurdistan Region (KRI) (north of Iraq), with the reopening of refugee camps.

Field researches conducted in the KRI before the refugee crisis found that if the population continued to grow at current rates and physician utilization rates were similar to nearby countries, the KRI would have needed an additional 2,097 physicians by 2020 (an increase of more than 33%) (Moore *et al.,* 2014). However, contrary to the national needs, many skilled health workers and young graduates have continued to leave.

National development plans call for a realignment of the health system with primary health care as the basis (Al Hilfi *et al.,* 2013). Several problems in the epidemiological monitoring system have been identified. Specifically, existing data collection and use is fragmented, inconsistent and often of poor quality, with routine data collection not standardized across the country. The result of it is inadequate knowledge of the health status of the population and insufficient understanding of its needs, limited public health vision and organization, as well as lack of prevention activities and preparedness response plan (Ross *et al.,* 2017). For example, despite 30 years having passed since the chemical attack in Halabja, the Kurdistan population continues to suffer from its long-term psychological, health and environmental consequences (Mlodoch, 2017). After many years, survivors are still reporting chronic symptoms, such as respiratory, digestive and neurological disorders; increased miscarriages and infertility; blindness; leukemia, lymphoma and other cancers; post-traumatic stress disorder, depression and feelings of guilt, aggression, and alienation; as well as congenital malformations and other birth defects (Jiyan Foundation for Human Right, 2014). These deteriorated medical conditions put survivors in need of continuous medical care and support. Additionally, mental health with declared psychiatric and social dysfunction, have been observed even decades after the initial incident. Consequences are found not only among survivors but also the second generation seems to be affected. These observations call for field-researches and for further investigations. Recent articles have suggested that the long-term impact of war on psychosocial, physical health and emotional life is often overlooked by policy makers, despite the evidence that traumatized populations remain disadvantaged on a number of economic and social fronts long after fighting has ceased (Galea *et al.,* 2003). Further researches are needed to determine the societal costs of human rights abuses and to identify groups of persons at increased risk of psychological dysfunction years after the fighting stops (Dworkin *et al.,* 2008).

In order to respond to these health priorities, the University of Rome Tor Vergata (Italy) and Ministry of Health of KRI have been implementing, since 2015, an electronic system for epidemiological monitoring and health surveillance in the Kurdistan Region, with the aim of extending it to the whole of Iraq. The system - launched by means of the project "Development and Implementation of a Health Monitoring and Epidemiological Surveillance System in Iraqi Kurdistan" - has been designed to manage healthcare data, by collecting, storing, managing, and transmitting patients' electronic medical records, including diagnoses, vaccinations, births and deaths. At the end of October 2019, 59 centers were active, with more than 600,000 disease events recorded in the four governorates of Dohuk, Erbil, Halabja and Sulaymaniyah (Emberti Gialloreti *et al.*, 2020a). The software is run by the health and administrative personnel of the local health centers, who have already been trained in data collection, analysis and disease coding (Moramarco *et al.,* 2019).

3.3 The Case of Lebanon

Since the 70s Lebanon has been threatened by years of social and political instability, deeply influenced by the neighboring Syria, which controlled Lebanon from 1976 to 2005. After Syria's withdrawal, the Israeli and Hezbollah militias continued to attack and counterattack against each other, culminating in a war in 2006 that left much of south Lebanon devastated. The Israeli government admitted to have used phosphorus weapons in Lebanon during this conflict. These are not forbidden by international law but several experts believe they should be re-classified as chemical weapons. Despite the Israeli government claiming their use according to the rules of international law, there have been numerous reports that phosphorus munitions injured and killed civilians, with unexploded cluster bombs threatening the Lebanese population even after the end of the war (The Guardian, 2006).

Since the start of the Syrian crisis in 2011, in light of the volatility and complexity of the conflict, concerns exist over the potential for the Syrian regime to transfer chemical weapons to non-state actors, such as Hezbollah in Lebanon (Brooks *et al*., 2018), that might also threaten a fragile peace in the country (The Globe and Mail, 2018).

Nowadays, Lebanon is indirectly experiencing the consequences of neighboring conflicts and CBRNe related events. With some 1.5 million Syrians refugees - about a quarter of the Lebanese population - in addition to a large community of Palestinian refugees, the country hosts the largest concentration of refugees per capita and the fourth largest refugee population in the world (Government of Lebanon & the United Nations, 2019). A robust response has been timely mounted by the local government in partnership with the international community, helping to avert dire consequences. Over the years, the situation for refugees has been stabilized and even slightly improved in many sectors, but over two thirds of Syrian refugees still remain in poverty and 90% are experiencing food insecurity. The large presence of displaced populations has increased demand on infrastructure and health services, which lack the capacity to fully meet the needs. Currently, nearly one third of refugee households remains unaware of where to access medical services in case of an emergency (UNHCR/UNICEF/WFP, 2018), suggesting the need for strengthening communication and preparedness. These impacts are likely to be significant, and to continue into the long-term, with particular concern for health security and the likely increase of infectious diseases risk (Republic of Lebanon MoPH, 2016). Overall vaccination coverage rates remain sub-optimal, with recently experienced outbreaks of vaccine preventable diseases (867 cases of measles in 2018) and water-borne diarrhea (Government of Lebanon & the United Nations, 2019). The pressure on healthcare institutions caused by the increased demand for services could be an additional potential source of conflict and CBRNe-related events in this already fragile state.

It is impossible to completely define the epidemiological and health services impacts of such human flows even in the future. The vulnerable situation of Syrian refugees could be only indirectly deducted from annual household surveys: for example, in 2018, 4,446 Syrian refugee households from 26 districts across the country were randomly visited, feeding the data on the Vulnerability Assessment of Syrian Refugees in Lebanon (VASyR) (UNHCR/UNICEF/WFP, 2018). This data may underestimate the overall load of the situation, or not being anymore representative of the situation due to the relighting of the tensions in Syria in October 2019.

In order to address this, efforts have been geared towards strengthening the local MoPH both centrally and peripherally, as well as the overall primary healthcare system. Recently in 2017, the MoPH has established a Health Information System to monitor health indicators and related outcomes necessary for future health planning and decision-making. The online system connects 75 Public Health Centres to the Primary Health Care Unit in the MoPH, ensuring the centralization of information and rapid transferring of data to the Ministry of Health (WHO, 2017).

3.4 The Case of Syria

The Syrian conflict started in March 2011, involving the government and a spectrum of anti-government factions including the so-called Islamic State - ISIS. Warring parties have proliferated, from additional jihadist groups, to Russian and US-led coalition forces. Since its beginning, the Syrian crisis claimed hundreds of thousands of lives and countless injuries among civilians. Aerial bombings and shelling rapidly became primary causes of direct deaths of civilians, calling into question the use of wide-area explosive weapons in urban areas (Guha-Sapir *et al.,* 2018). According to United States intelligence, Syria has had a stockpile of chemical weapons since 2012, and – according to the same source - over the past years both the Syrian government and ISIS have been responsible for chemical weapons attacks (Arms Control Association, 2019). In 2013, the United Nations called for an immediate investigation, since the government was suspected of using chemical weapons in an attack on civilians outside Damascus, where more than 1,300 people were killed, many of them women and children (CNN, 2013). The latest attack was reported in Douma, in April 2018, where the Organization for the Prohibition of Chemical Weapons (OPCW) concluded in its final report that a toxic chemical, likely chlorine, had been used (UN Security Council, 2019).

Just recently, in October 2019, Turkey launched an offensive into north-eastern Syria against the Kurdish forces who controlled the region. Images showed smoke rising with civilians fleeing towns. During the airstrike on the border town of Ras-al-Ayn, some civilians were reporting suspected signs of being exposed to chemical substances (phosphorus gas). The OPCW claimed to be aware of the situation in Syria and to be investigating the alleged use of chemical weapons (Independent, 2019). At the time we concluded this article, the chemical attack was not yet confirmed but the situation might have evolved.

The protracted conflict had a cumulative destructive effect on the economy, infrastructure, agricultural production, food systems, social institutions, as well as human resources. Syria has become one of the most dangerous places for healthcare providers, where hundreds of healthcare workers have been killed and / or tortured, and several health facilities deliberately destroyed (Fouad *et al*., 2017). As a main result, the health system has been directly and indirectly impacted catastrophically, with supply lines interrupted and a general degradation of key services (The Syrian Centre for Policy Research, 2015). The large scale breakdown of health services led to a decrease in life expectancy and an increase in childhood mortality (Guha-Sapir *et al*., 2018), that has obliterated the public health gains made in the past (Abbara *et al*., 2015). Despite moderately high vaccination coverage rates in pre-conflict Syria (UNICER & WHO, 2012), recent reports of infectious disease outbreaks have become increasingly common. Today's reduced coverage rates are striking, if compared to 2010, a time in which more than 80% of the target age group was vaccinated (de Lima Pereira *et al*., 2018). After having being considered for 15 years as a polio-free country, Syria reported a polio outbreak in 2013-2014, which also spread to neighboring countries (Ozaras *et al*., 2016). The formal recognition of the outbreak prompted a multi-country regional response, with a variety of actors involved in immunization campaigns (WHO Regional Office for the Eastern Mediterranean, 2016).

The armed conflict has caused massive and continuous exoduses of Syrians: more than 5 million people are still displaced inside the country, with another five million living in neighboring countries, mainly Egypt, Iraq, Jordan, Lebanon, and Turkey (Operational Portal Refugee Situation, 2019). The majority of the refugees rely on humanitarian assistance to meet their basic needs. The health systems of the hosting countries have been challenged to respond to the diverse health needs of the refugees, while trying to preserve services for their own citizens. The need for emergency and basic health services, such as reproductive and maternal / child health, draws a disease burden profile consistent with that of middle-income countries (Akik *et al*., 2019). High risk of epidemics in neighboring countries were registered (WHO

Regional Office for the Eastern Mediterranean, 2013), with increased rates of tuberculosis reported among Syrian refugees in Lebanon and Jordan.

While there has been some attention on the challenges of meeting health needs of Syrian refugees in neighboring countries, very little has been documented about the humanitarian challenges within Syria. It is estimated that over seven million people in the country are without access to basic healthcare (UNOCHA, 2016), but insufficient information is available about the effects of the conflict on the health of the population inside Syria (Ismail *et al.,* 2016). A recent World Bank study found – even if the results are not yet confirmed - that in Syria, more people may have died because of the breakdown of the health system than because of direct fatalities due to fighting (World Bank Group, 2017).

In order to monitor the health care in the country, in early 2013 the Syrian Ministry of Health and sector partners adapted a tool called HeRAMS (Health Resources and Services Availability Mapping System). The key information assessed the functionality status of the public health system, including health infrastructure, human resources, availability of health services, equipment, medicines at primary and secondary care level. HeRAMS has been developed to provide timely information, in order to support decision-making and coordination of health sector actors in emergencies, but can also be applied to post-emergencies, recovery and development contexts (WHO Regional Office for the Eastern Mediterranean, 2017).

4. DISCUSSION

After CBRNe events, especially during long-lasting conflicts, a country's situation is often characterized by a weakening of the health system, complicated also by the limited quantity and quality of human resources (Lanjouw *et al*., 1999). The field-reports from Afghanistan, Iraq, Lebanon and Syria provide evidence of CBRNe conflict-related threats on the public health asset, exposing the existing systems to fragility, and difficulties in maintaining and strengthening them (Sharara & Kanj, 2014). Health structures collapsed, while many health professionals have been killed or forced to flee, causing a shortage of personnel. The loss of human resources and talents has been huge, while investments have been diverted from civilian to military facilities (Mokdad *et al*., 2016). The breakdown of health services, and water and sanitation treatment plants, increased exposure of people to vulnerabilities, lack of investments in public health, as well as huge flux of refugees and displaced people have opened the door to disease outbreaks, even beyond borders. As an example of the impact of Syrian war to neighboring countries, Syria and Iraq experienced the reemergence of polio in areas declared polio-free before the conflicts (Arie, 2014), due to deep decline in health infrastructures and delivering devices (i.e., immunizations). The caseload of diseases among refugees and displaced people (especially in Iraq and Lebanon), posed new challenges for hosting government and humanitarian agencies in the provision of early assessment, health care, diagnostics, and medications.

Alongside the direct and relatively short-term causalities, long-term and potentially health devastating consequences stemmed also from the erosion of the states' ability to conduct health monitoring and surveillance, thus being not only unable to identify, prevent and treat adverse health conditions, but also to set up effective strategic plans. In these fragile states, health data are still mostly estimated by modeling techniques using other available variables and figures, often collected from neighboring countries or from countries with a similar health profile. Occasionally, data are deduced from surveys, censuses, household recalls for death, or United Nations estimates that account for migration. Frequently, data trends show discrepancies in case report numbers between government and non-government controlled areas, and interpretation is hampered by uncertainties over sentinel surveillance coverage and base population numbers (Ismail *et al.,* 2016). Uncertainty about population denominators (due to unknown numbers of killed people, limited capacity to monitor in- and out-flows, collapse of prewar statistical services, and outdated census figures) impedes meaningful analysis of health data (e.g., coverage indicators) and accurate service delivery planning (e.g., vaccination) (Diggle *et al*., 2017). This condition can last even long after emergencies and CBRNe events, especially in war-torn countries, so that targeted health interventions and preparedness plans do not exist, or are not fully and timely implemented.

The rebuilding process includes a wide array of actions and actors (Rutherford & Saleh, 2019). Collecting, analyzing and interpreting data on the health of the population (epidemiological surveillance) is the

fundamental starting point to know its status and the hazards threatening it, in order to address its specific needs, thus providing timely and useful evidence, which is essential for effective decision making (Murray, 2009; Lopez & Setel, 2015). Indeed when talking about public health systems, one of the primary responsibilities of any government is to protect health throughout three key elements: prevention, early detection, and timely and effective response. Public health surveillance is the epidemiological foundation for modern public health (Bonita *et al*., 2003), which has proven to have a central role in health policies planning and evaluation, thus being termed "the foundation of all public health practice" (Henderson, 2016). Investment and proper research towards post-events reconstruction is imperative for addressing healthcare problems and establishing an effective resilience (Rutherford & Saleh, 2019). A key requirement for effective health system development is the availability of reliable health indicators (Kruk *et al.,* 2010). Therefore, health monitoring systems providing useful data for epidemiological surveillance are essential to support local governments during the effective rebuilding of an efficient public health system. In addition, the epidemiological surveillance system will also act as an early warning system, which is extremely important especially for fragile states (Emberti Gialloreti *et al.*, 2020b). Only with rapid alerting of suspected cases of diseases can countries implement appropriate response measures to mitigate their negative impact. After CBRNe emergencies, as well as after each complex emergency, guidance for health interventions consistently highlights in fact the need for simple but effective health monitoring systems covering basic health information data of the population (i.e., mortality and morbidity data), health services delivery and response planning (UNHCR, 2007). Innovative strategies are needed to provide long-term strengthening of public health services and provision of health care access for the whole population, including refugees, displaced persons and minorities.

Nevertheless, a critical role is played by the strengthening of leadership and management capacities at all levels. Supporting capacity building and technical leadership should not be neglected and always be considered a key priority integrated into the humanitarian health response in emergency areas (Diggle *et al*., 2017).

5. CONCLUSION

CBRNe events often leave a damaged health system behind. Building or re-building a public health system has epidemiological surveillance as a paramount basis for producing relevant statistics and access the health needs of the population in post-CBRNe events. This allows governments to support day-to-day health management, population's health protection and promotion, while providing useful data for long-term planning and health policy development. Therefore, the recovery phase after CBRNe events, especially in war-torn countries, is a long process which requires inputs at all levels from multiple stakeholders. Coordinated actions offer an opportunity to set in place the foundations for nationwide governance and are the backbone for a post-conflict empowerment of the health system. Hitting those targets, especially in fragile states, serves as guide in planning future interventions and in strengthening the country's health institutions, preventing deep impact consequences in case of further CBRNe events.

REFERENCES

Abbara, A., Blanchet, K., Sahloul, Z., Fouad, F., Coutts, A. & Wasim, M. (2015). The effect of the conflict on Syria's health system and human resources for health. *World Health Population,* **16**: 87-95.

Akik, C., Ghattas, H., Mesmar, S., Rabkin, M., El-Sadr, W.M. & Fouad, F.M. (2019). Host country responses to non-communicable diseases amongst Syrian refugees: a review. *Confl. Health*, **13**: Art. 8.

Al Hilfi, T,K., Lafta, R. & Burnham, G. (2013). Health services in Iraq. *Lancet*, **381** (9870): 939-948.

Arie, S. (2014). Polio virus spreads from Syria to Iraq. *BMJ*, **348**.

Arms Control Association (2019). *Timeline of Syrian Chemical Weapons Activity, 2012-2019.* Available online at: https://www.armscontrol.org/factsheets/Timeline-of-Syrian-Chemical-Weapons-Activity (Last access date: 15 August 2019).

Bagcchi, S. (2016). Cholera in Iraq strains the fragile state. *Lancet Infect. Dis*., **16**: 24-25.

Bartlett, L., Whitehead, S., Crouse, C., Bowens, S., Mawji, S., Ionete, D. & Salama, P. (2002). *Maternal Mortality in Afghanistan: Magnitude, Causes, Risk Factors and Preventability.* Afghanistan, Ministry of Public Health, CDC, UNICEF.

Berkelmann, R.L., Stroup, D.F. & Buehler, JW. (1997). Public health surveillance. In: Detels R, Holland WW, McEwen J. & Omenn G.S. (Eds.) *Oxford Textbook of Public Health, 3rd Ed.* Oxford University Press, New York, pp. 735-750.

Black, R., Morris, S. & Bryce, J. (2003). Where and why are 10 million children dying every year? *Lancet*, **361**: 2226-2234.

Bonita, R., Winkelmann, R., Douglas, K.A. & de Courten, M. (2003) The WHO Stepwise Approach to Surveillance (Steps) of Non-Communicable Disease Risk Factors. In: McQueen D.V., Puska P. (Eds.) *Global Behavioral Risk Factor Surveillance*. Springer, Boston MA, pp. 9-22.

Boyd, A., Chambers, N., French, S., Shaw, D., King, R. & Whitehead, A. (2013). Emergency planning and management in health care: priority research topics. *Health Syst.*, **3**: 83–92.

Brooks, J., Erickson, T.B., Kayden, S., Ruiz, R., Wilkinson, S. & Burkle, F.M. Jr. (2018). Responding to chemical weapons violations in Syria: legal, health, and humanitarian recommendations. *Confl. Health,* **12**: Art. 12.

Brown, V., Jacquier, G., Bachy, C., Bitar, D. & Legros, D. (2002). Prise en charge des épidémies de choléra dans un camp de réfugiés. *Bull. Soc. Pathol. Exot.*, **95**: 351-354.

Centers for Disease Control and Prevention (2001). *Updated Guidelines for Evaluating Public Health Surveillance Systems*. July 27, 2001. Available online at: https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm (Last access date: 26 November 2019).

Cetorelli, V., Burnham, G. & Shabila, N. (2017). Prevalence of non-communicable diseases and access to health care and medications among Yazidis and other minority groups displaced by ISIS into the Kurdistan Region of Iraq. *Conf. Health,* **11**: Art. 4.

CNN (2013). *UN, US Call for urgent probe of Syria Chemical Attack Claim.* August 23, 2013. Available online at: https://edition.cnn.com/2013/08/22/world/meast/syria-civil-war/ (Last access date: 5 September 2019).

Cook, J. (2003). Post-conflict reconstruction of the health system of Afghanistan: assisting in the rehabilitation of a provincial hospital–context and experience. *Med. Confl. Surviv.,* **19**: 128–141.

Coutts, A., Stuckler, D., Batniji, R., Ismail, S., Maziak, W. & McKee, M. (2013). The Arab Spring and health: two years on. *Int. J. Health Serv.*, **43**: 49-60.

De Lima Pereira, A., Southgate, R., Ahmed, H., O'Connor, P., Cramond, V. & Lenglet A. (2018). Infectious disease risk and vaccination in Northern Syria after 5 years of civil war: The MSF experience. *PLoS Curr*., **10**: 29511602.

Devi, S. (2018). Reconstructing Iraq. *Lancet*, **392**: 541-542.

Diggle, E., Welsch, W., Sullivan, R., Alkema, G., Warsame, A., Wafai, M., Jasem, M., Ekzayez, A., Cummings, R. & Patel, P. (2017). The role of public health information in assistance to populations living in opposition and contested areas of Syria, 2012–2014. *Conf. Health,* **11**: Art. 33.

Dworkin, J., Prescott, M., Jamal, R., Hardawan, S.A., Abdullah, A. & Galea, S. (2008). The long-term psychosocial impact of a surprise chemical weapons attack on civilians in Halabja, Iraqi Kurdistan. *J. Nerv. Ment. Dis.,* **196**: 772-775.

Emberti Gialloreti, L., Basa, F.B., Moramarco, S., Salih, A.O., Alsilefanee HH, Qadir, S.A., et al. (2020a) Supporting Iraqi Kurdistan health authorities in post-conflict recovery: The development of a health monitoring system. *Front. Public Health*, **8**: Art. 7.

Emberti Gialloreti, L., Moramarco, S. & Palombi, L. (2020b). Investing in epidemiological surveillance for recovering health systems in war-torn countries. *Perspect. Public Health,* **140**: 25-26.

Fouad, F.M., Sparrow, A., Tarakji, A., Alameddine, M., El-Jardali, F., Coutts, A.P., El Arnaout, N., Karroum L.B., Jawad, M., Roborgh, S., Abbara, A., Alhalabi, F., AlMasri, I. & Jabbour, S. (2017). Health workers and the weaponisation of health care in Syria: a preliminary inquiry for the lancet-American University of Beirut Commission on Syria. *Lancet*, **390**: 2516 – 2526.

Frieden, T.R., Tappero, J.W., Dowell, S.F., Hien, N.T., Guillaume, F.D. & Aceng, J.R. (2014). Safer countries through global health security. *Lancet,* **383**: 764 – 766.

Fujita, N., Zwi, A.B., Nagai, M. & Akashi, H. (2011). A comprehensive framework for human resources for health system development in fragile and post-conflict states. *PLoS Med*., **8**: e1001146

Galea, S., Vlahov, D., Resnick, H., Ahern, J., Susser, E., Gold, J., Bucuvalas, M. & Kilpatrick, D. (2003) Trends of probable post-traumatic stress disorder in New York City after the September 11 terrorist attacks. *Am. J. Epidemiol.,* **158**: 514–524.

Government of Lebanon & the United Nations (2019). *Lebanon Crisis Response Plan 2017-2019 (update 2019).* Available online at: https://reliefweb.int/report/lebanon/lebanon-crisis-response-plan-2017-2020-2019-update. (Last access date: 29 October 2019).

Guha-Sapir, D., Schlüter, B., Rodriguez-Llanes, J.M., Lillywhite, L. & Hsiao-Rei Hicks, M. (2018). Patterns of civilian and child deaths due to war-related violence in Syria: a comparative analysis from the Violation Documentation Center dataset, 2011–16. *The Lancet Glob. Health.,* **6**: 103–110.

Hazbun, W. (2012). The Middle East through the lens of critical geopolitics: Globalization, terrorism and the Iraq War. In Bonine, A. (Ed.), *Is there a Middle East?: The Evolution of a Geopolitical Concept*. Stanford University Press, Stanford.

Henderson, D.A. (2016). The Development of Surveillance Systems, *Am. J. Epidemiol.,* **183**: 381–386.

Hiltermann, J.R. (2007). *A Poisonous Affair: America, Iraq, and the Gassing of Halabja.* Cambridge University Press, Cambridge, p. 195.

HM Government (2013). *Emergency Response and Recovery*. Cabinet Office Civil Contingencies Secretariat, London. Available online at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/253488/Emergency_Response_and_Recovery_5th_edition_October_2013.pdf (Last access date: 29 October 2019).

Independent (2019). *Turkey faces scrutiny over alleged use of white phosphorus on children in northern Syria.* October 18, 2019. Available online at: https://www.independent.co.uk/news/world/middle-east/syria-turkey-ceasefire-war-crimes-middle-east-a9161586.html (Last access date: 20 November 2019).

IOM (2018). *Iraq Reasons to remain: Categorizing Protracted displacement in Iraq*. November 2018. Available online at: https://displacement.iom.int/reports/iraq-%E2%80%93-reasons-remain-categorizing-protracted-displacement-iraq-november-2018 (Last access date: 15 August 2019).

Ismail, S.A., Abbara, A., Collin, S.M., Orcutt, M., Coutts, A.P., Maziak, W., Sahloul, Z., Dar, O., Corrah, T. & Fouad, F.M. (2016). Communicable disease surveillance and control in the context of conflict and mass displacement in Syria. *Int. J. Infect. Dis.,* **47**: 15–22.

Jiyan Foundation for Human Right (2014). *Program for survivors of Genocide*. October, 13. Available online at: https://www.jiyan-foundation.org/programs/genocide (Last access date: 20 November 2019).

Johnson, S.A. (2017). The cost of war on Public Health: An Exploratory Method for Understanding the Impact of Conflict on Public Health in Sri Lanka. *PLoS One*, **12**: e0166674.

Johnston's Archive (2015). *Chemical weapon terrorism in Iraq and Afghanistan*. April 15, 2015. Available online at: http://www.johnstonsarchive.net/terrorism/wmdterrorism-1.html (Last access date: 20 November 2019).

Kruk, M.E., Freedman, L.P., Anglin, G.A. & Waldman, R.J. (2010). Rebuilding health systems to improve health and promote statebuilding in post-conflict countries: a theoretical framework and research agenda. *Soc. Sci. Med.,* **70**: 89–97.

Langmuir, A.D. (1963). The surveillance of communicable diseases of national importance. *N. Engl. J. Med.,* **268:** 182-192.

Lanjouw, S., Macrae, J. & Zwi, A.B. (1999). Rehabilitating health services in Cambodia: the challenge of coordination in chronic political emergencies. *Health Policy Plan.*, **14**: 229–242.

Lopez, A.D. & Setel, P.W. (2015). Better health intelligence: a new era for civil registration and vital statistics? *BMC Med*., **13**: Art 73.

Middle East Health Magazine (2012). *Iraq report. On the road to recovery*. January 21, 2012. Available online at: http://www.middleeasthealthmag.com/cgi-bin/index.cgi?http://www.middleeasthealthmag.com/jan2012/feature1.htm (Last access date: 20 November 2019).

Mine Action Review (2018). *Iraq: Clearing the Mines 2018*. http://www.mineactionreview.org/assets/downloads/NPA_Clearing_the_Mines_2018_Web.pdf (Last access date: 20 November 2019).

Ministry of Health, Transitional Islamic Government of Afghanistan (2002). *A basic package of health services for Afghanistan.* Available online at:

http://unpan1.un.org/intradoc/groups/public/documents/apcity/unpan018852.pdf (Last access date: 29 October 2019).

Mlodoch, K. (2017). The indelible smell of apples: poison gas survivors in Halabja, Kurdistan-Iraq, and their struggle for recognition. In: Friedrich B., Hoffmann D., Renn J., Schmaltz F. & Wolf M. (Eds.) *One Hundred Years of Chemical Warfare: Research, Deployment, Consequences*. Springer, Cham, pp. 349-362.

Mohamed-Ali, H. (1992). Late lesions due to poison gas in survivors of the Iraqi poison warfare against the Kurdish people. *Wien Med. Wochenschr*., **142**: 8-15.

Mokdad, A.H., Forouzanfar, M.H., Daoud, F., El Bcheraoui, C., Moradi-Lakeh, M., Khalil, I., et al. (2016). Health in times of uncertainty in the eastern Mediterranean region, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet Glob. Health.*,**4:** e704-13.

Moore, M., Ross, A.C., Yee-Wei Lim, Jones, S.S., Overton A. & Yoong, J.K. (2014). *The Future of Health Care in the Kurdistan Region—Iraq: Toward an Effective, High-Quality System with an emphasis on Primary Care*. Santa Monica, Calif.: RAND Corporation and Ministry of Planning of the Kurdistan Regional Government, MG-1148-1-KRG. Available online at: https://www.rand.org/pubs/monographs/MG1148-1.html (Last access date: 29 October 2019).

Moramarco, S. (2018). Food security and proper nutrition: A public health and humanitarian priority in pre- and post-CBRN events. *Defence S&T Tech. Bull*; **11**: 299-309.

Moramarco, S., Basa, F.B., Alsilefanee, H.H., Qadir, A.S. & Emberti Gialloreti, L. (2019). Developing a public health monitoring system in a war-torn region: A field report from Iraqi Kurdistan. *Disaster Med. Public Health Prep;* 1-3.

Murray, C.J., King, G., Lopez, A.D., Tomijima, N. & Krug, E.G. (2002). Armed conflict as a public health problem. *BMJ,* **324**: 346–349.

Murray, C.J.L. (2009). *Assessing Health Systems Performance Using Information on Effective Coverage of Interventions*. Working paper series No.3. University of Queensland School of Population Health, Health Information Systems Knowledge Hub, Brisbane.

Nnadi, C., Etsano, A., Uba, B., Ohuabunwo, C., Melton, M., Wa Nganda, G., Esapa, L., Bolu, O., Mahoney, F., Vertefeuille, J., Wiesen, E. & Durry, E. (2017). Approaches to vaccination among populations in areas of conflict. *J. Infect. Dis*., **216**: S368–S372.

OECD. (2011). *International Engagement in Fragile States: Can't We Do Better?*. Conflict and Fragility. OECD Publishing.

Operational Portal Refugee Situation (2019). *Syria Regional Refugee Response*. November 14, 2019. Available online at: http://data.unhcr.org/syrianrefugees/regional.php (Last access date: 20 November 2019).

Ozaras, R., Leblebicioglu, H., Sunbul, M., Tabak, F., Balkan, I.I., Yemisen, M., Sencan, I. & Ozturk R. (2016). The Syrian conflict and infectious diseases. *Expert Rev. Anti Infec. Ther*., **14**: 547-555.

Razavi, S.M., Razavi, M.S., Pirhosseinloo, M. & Salamati, P. (2014). Iraq-Iran chemical war: calendar, mortality and morbidity. *Chin. J. Traumatol*., **17**: 165-169.

Republic of Lebanon MoPH (2016). *Health Strategic Plan. Strategic Plan for the Medium Term (2016 to 2020).* Ministry of Public Health, Beirut. Available online at: https://www.moph.gov.lb/en/Pages/0/11665/strategic-plan-2016-2020- (Last access date: 29 October 2019).

Rodier, G., Greenspan, A.L., Hughes, J.M. & Heymann, D.L. (2007). Global public health security. *Emerg. Infect. Dis.,* **13**: 1447–1452.

Ross, A.C., Moore, M., Hilborne, L.H., Rooney, A., Hickey, S., Ryu, Y. & Botwinick, L. (2017). *Strengthening Health Care in the Kurdistan Region of Iraq.* Santa Monica, Calif.: RAND Corporation, RB-9990-KRG. Available online at: https://www.rand.org/pubs/research_briefs/RB9990.html (Last access date: 10 October 2019).

Rubenstein, L. (2009). *Post-Conflict Health Reconstruction*. United States Institute of Peace, Washington, DC. Available online at: https://www.usip.org/publications/2009/09/post-conflict-health-reconstruction (Last access date: 10 October 2019).

Rutherford, S. & Saleh, S. (2019). Rebuilding health post-conflict: case studies, reflections and a revised framework. *Health Policy Plan*., **34**: 230-245.

Schwartzstein, S.J.D. (1982). Chemical Warfare in Afghanistan: An Independent Assessment. *World Affairs,* **145**: 267–272.

Sharara, S.L. & Kanj, S.S. (2014).War and infectious diseases: challenges of the Syrian civil war. *PLoS Pathog.,* **10**: e1004438.

The Daily Outlook Afghanistan (2017). *Afghanistan's Poor Health Care System.* October 07, 2017. Available online at: http://www.outlookafghanistan.net/editorialdetail.php?post_id=19087 (Last access date: 20 November 2019).

The Globe and Mail (2018). *Strikes on Syria threaten a fragile peace in Lebanon.* April 13, 2018. Available online at: https://www.theglobeandmail.com/world/article-in-lebanon-renewed-conflict-in-syria-threatens-a-fragile-peace/ (Last access date: 20 November 2019).

The Guardian (2006). *Israel admits it used phosphorus weapons.* October 23, 2006. Available online at: https://www.theguardian.com/world/2006/oct/23/israel (Last access date: 20 November 2019).

The Guardian (2016). *US launches airstrike on Mosul hospital used by Isis, military says.* December 7, 2016. Available online at: https://www.theguardian.com/us-news/2016/dec/07/islamic-state-iraq-mosul-hospital-airstrike-us-military (Last access date: 20 November 2019).

The National (2018). *Yemen setback as Houthi rebels block aid deliveries*. May 3, 2018. Available online at: https://www.thenational.ae/world/yemen-setback-as-houthi-rebels-block-aid-deliveries-1.727166 (Last access date: 20 November 2019).

The New York Times (2003). *Halabja: America didn't seem to mind poison gas*. January 17, 2003. Available online at: https://www.nytimes.com/2003/01/17/opinion/IHT-halabja-america-didnt-seem-to-mind-poison-gas.html (Last access date: 20 November 2019).

The Syrian Centre for Policy Research (2015). *Syria: alienation and violence, impact of Syria crisis report 2014.* Damascus – Syria. Available online at : https://reliefweb.int/sites/reliefweb.int/files/resources/alienation_and_violence_impact_of_the_syria_crisis_in_2014_eng.pdf

The Times (2010). *Halabja, the massacre the West tried to ignore.* January 18, 2010. Available online at: https://www.thetimes.co.uk/article/halabja-the-massacre-the-west-tried-to-ignore-qsl8n6nspc7 (Last access date: 20 November 2019).

The Washington Post (2018). Middle east. *Warplanes bomb 3 hospitals in southern Syria as Assad's army presses offensive*. June 27, 2018. Available online at: https://www.washingtonpost.com/world/middle_east/warplanes-bomb-3-hospitals-in-southern-syria-as-assads-army-presses-offensive/2018/06/27/c3850054-798b-11e8-ac4e-421ef7165923_story.html (Last access date: 20 November 2019).

Trading Economics (2019). *Iraq - Hospital beds.* Available online at: https://tradingeconomics.com/iraq/hospital-beds-per-1-000-people-wb-data.html (Last access date: 20 November 2019).

UN News (2018). *Fresh Yemen hospital attack raises risk of new cholera epidemic*. August 3, 2018. Available online at: https://news.un.org/en/story/2018/08/1016272 (Last access date: 20 November 2019).

UN Security Council (2019). *Report of the fact-finding mission regarding the incident of alleged use of toxic chemicals as a weapon in Douma, Syrian Arab Republic, on 7 April 2018*. April 18, 2019 (S/2019/208). Available online at: https://reliefweb.int/report/syrian-arab-republic/report-fact-finding-mission-regarding-incident-alleged-use-toxic (Last access date: 29 October 2019).

UNHCR (2007). *Handbook for emergencies*. 3rd Ed. UNHCR, Geneva.

UNHCR/UNICEF/WF (2018). *VASyR 2018: Vulnerability Assessment for Syrian Refugees in Lebanon.* Available online at: https://data2.unhcr.org/en/documents/download/67380 (Last access date: 20 November 2019).

UNOCHA (2016). *Middle East and North Africa: crises in focus—an overview of humanitarian needs in the region.* Available online at: https://reliefweb.int/report/world/middle-east-and-north-africa-crises-focus-overview-humanitarian-needs-region-enar (Last access date: 20 November 2019).

US Department of States (1982). *Chemical Warfare in Southeast Asia and Afghanistan*. *Report to the Congress for Secretary of State Alexander M Haig Jr. March 22, 1982.* Available online at: https://www.cia.gov/library/readingroom/docs/CIA-RDP97M00248R000500010018-6.pdf (Last access date: 20 November 2019).

USAID (2014). *Rebuilding the Health Sector in Afghanistan professionalizing leadership and management as a pillar of the health system.* Available online at: https://www.lmgforhealth.org/sites/default/files/LMG%20Afghanistan%20Program%20Brief.pdf (Last access date: 10 October 2019).

Verrecchia, R., Thompson, R. & Yates, R. (2019). Universal Health Coverage and public health: a truly sustainable approach. *Lancet Pub. Health*; **4**: e10 - e11.

Waldman, R. & Hanif, H. (2002). *The Public Health System in Afghanistan: Current Issues*. Afghanistan Research and Evaluation Unit. Available online at: http://unpan1.un.org/intradoc/groups/public/documents/APCITY/UNPAN015804.pdf (Last access date: 10 October 2019).

Webster, P.C. (2011). lraq's health system yet to heal from ravages of war: Special Report. *Lancet,* **378**: 863 – 866.

World Bank Group (2017). *The Toll of War: the Economic and Social consequences of the conflict in Syria*. The World Bank, Washington DC. Available online at: http://www.worldbank.org/en/country/syria/publication/the-toll-of-war-the-economic-and-social-consequences-of-the-conflict-in-syria (Last access date: 10 October 2019).

WHO (2015). *Emergencies, preparedness, response. Cholera Iraq.* November 26, 2015. Available online at: https://www.who.int/csr/don/26-november-2015-iraq-cholera/en/ (Last access date: 10 October 2019).

WHO (2017). *Primary health care systems (PRIMASYS): case study from Lebanon, abridged version*. World Health Organization, Geneva. Available online at: https://www.who.int/alliance-hpsr/projects/alliancehpsr_lebanonabridgedprimasys.pdf?ua=1 (Last access date: 25 November 2019).

WHO (2018). *Regional Office for the Eastern Mediterranean Eastern Mediterranean Region: framework for health information systems and core indicators for monitoring health situation and health system performance 2018/WHO Regional Office for the Eastern Mediterranean.* Available online at: http://applications.emro.who.int/docs/EMROPUB_2018_EN_20620.pdf?ua=1 (Last access date: 10 October 2019).

WHO (2019). *What is health financing for universal coverage?* World Health Organization, Geneva. Available online at: http://www.who.int/health_financing/universal_ coverage_definition (Last access date: 25 November 2019).

WHO Regional Office for the Eastern Mediterranean (2017). *Developing health centres and hospitals indices for Syria. Based on HeRAMS dataset 2014*. World Health Organization. Available online at: http://applications.emro.who.int/dsaf/EMROPUB_2017_EN_19363.pdf?ua=1 (Last access date: 25 November 2019).

WHO Regional Office for the Eastern Mediterranean (2013). *Regional Office for the Eastern Mediterranean Eastern Mediterranean Region. WHO warns of increased risk of disease epidemics in Syria and in neighboring countries as summer approaches, 3 June 2013*. June 3, 2013. Available online at: http://www.emro.who.int/press-releases/2013/disease-epidemics-syria.html (Last access date: 10 October 2019).

WHO Regional Office for the Eastern Mediterranean (2016). *Syria achieves polio milestone; 2 years without a reported case.* January 23, 2016. Available online at: http://www.emro.who.int/syr/syria-events/syria-achieves-polio-milestone-2-years-without-a-reported-case.html (Last access date: 25 November 2019).

World Health Assembly 21 (1968). *Report of the technical discussions on "national and global surveillance of communicable diseases"*. World Health Organization, Geneva. Available online at: https://apps.who.int/iris/handle/10665/143808 (Last access date: 25 November 2019).