EPIDEMIC SURVEILLANCE OF NOVEL CORONAVIRUS 2019 THROUGH PROBABILISTIC MODELS

SHERILYNN NGERNG SIEW FONG

A project report submitted in partial fulfilment of the requirements for the award of Master of Mathematics

Lee Kong Chian Faculty of Engineering and Science

Universiti Tunku Abdul Rahman

April 2021

DECLARATION

I hereby declare that the dissertation is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTAR or other institutions.

Signature:

Name: Sherilynn Ngerng Siew Fong

ID No.: 2000740

Date: 03/04/2021

APPROVAL FOR SUBMISSION

This dissertation/thesis entitled "**EPIDEMIC SURVEILLANCE OF NOVEL CORONAVIRUS 2019 THROUGH PROBABILISTIC MODELS**" was prepared by SHERILYNN NGERNG SIEW FONG and submitted as partial fulfilment of the requirements for the degree of Master of Mathematics at Universiti Tunku Abdul Rahman.

Approved by:

_____

(Dr. Denis Wong Chee Keong)

Date:é é é é é ..

Professor/Supervisor

Department of Mathematics and Actuarial Sciences (DMAS)

Faculty of Lee Kong Chian Faculty of Engineering and Science (LKC FES)

Universiti Tunku Abdul Rahman

The copyright of this report belongs to the author under the terms of the copyright Act 1987 as qualified by Intellectual Property Policy of Universiti Tunku Abdul Rahman. Due to acknowledgements shall always be made of the use of any material contained in, or derived from, this report.

© 2021, Sherilynn Ngerng Siew Fong, All rights reserved.

ACKNOWLEDGEMENTS

I would like to thank everyone who had contributed to the successful completion of this project. I would like to express my immense gratitude to my research supervisor, Dr Denis Wong Chee Keong for his invaluable advice, guidance and his enormous patience throughout the development of the research. Over the span of my research project, Dr Denis has provided his upmost support by providing me directions towards my research area and resources when I encountered technical issues, such as suggesting Ms Wong Kuan Wai to guide me when I encountered data transformation issues with Excel. Dr Denis has also been nothing but kind and patient towards my research writing process by readjusting his schedule when reviewing late draft submission due to the event that my mother was exposed to covid-19. I am also most grateful for my family's support over the pandemic while I work on my research paper.

EPIDEMIC SURVEILLANCE OF NOVEL CORONAVIRUS 2019 THROUGH PROBABILISTIC MODELS

ABSTRACT

This research paper explores the fundamentals behind epidemic surveillance models and the characteristics that stand out towards the construction of modern made drop-in surveillance systems developed during the height of covid-19 pandemic. It also allows a glimpse on creating a surveillance system for personal monitoring of disease outbreaks, whilst properties of the surveillance systems studied can be applied in personal monitoring of similar unique events such as economic crisis. This research paper is typical in the application of ECDC's publicly sourced surveillance data on covid-19 and this disease had little to no historical data prior to the pandemic. Amongst the epidemic surveillance models discussed, Farrington's Quasi-Poisson model predominantly works well with the aid of historical data to study previous trends and better predict incoming outbreaks while handling over-dispersed data. The Early Aberration Reporting System (EARS) model was developed by CDC after the 9/11 incident to predict sudden terrorist attacks and sudden outbreaks. Whereas the Spatio-Temporal Endemic-Epidemic model monitors the disease outbreak in transitioning stages of Suspected-Infected-Removed/Recovered which allows the observation of covid-19's infection rate. These models could help us obtain essential information on the pandemic to possibly brace the next wave of disease outbreaks.

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF SYMBOLS/ABBREVIATIONS

Covid-19	Novel Coronavirus 19
SARS	Severe Acute Respiratory System
WHO	World Health Organization
ECDC	European Centre for Disease Prevention and Control
CDC	Centres for Disease Control and Prevention
CDSC	Communicable Disease Surveillance Centre
MC-Health	Munich Centre of Health Sciences
EARS	Early Aberration Reporting System
CUSUM	Cumulative Sum
EARS-C1	Early Aberration Reporting System ï Cumulative Sum Variant 1
EARS-C2	Early Aberration Reporting System ï Cumulative Sum Variant 2
EARS-C3	Early Aberration Reporting System ï Cumulative Sum Variant 3
SIR	Suspected-Infected-Removed
FPR	False-positive rates
TPR	True-positive rates

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.0 Introduction

Bill Gates, a technologist, philanthropist and business leader presenting in a security conference, uttered, "Whether it occurs by a quirk of nature or at the hand of a terrorist, epidemiologists say a fast-moving airborne pathogen could kill more than 30 million people in less than a year. And they say there is a reasonable probability the world will experience such an outbreak in the next 10 to 15 years." Despite nuclear weapons and climate change, Bill Gates has often stressed that humanity's downfall would not be World War 3 but the outbreak of diseases. We are always ill-equipped to prevent the attacks, much less contain and overcome them without proper resources. Peering into his Ted Talk, we relive the incident whereby the Ebola virus outbreak occurred in third world countries. First world countries, for instance, the United States, should provide for these third world countries with proper healthcare while doubling as a station to monitor and possibly contain the spread of diseases before it spreads globally. Humanity may have been spared from the wrath of influenza disease outbreaks occurring over the previous years. However, the ghost of the deadly coronavirus, identified as the Novel Coronavirus 2019 (covid-2019), haunts us as a fatal reminder of our once again fragile humanity. In less than ten years, the lethality of covid-2019 has surpassed its predecessor, Severe Acute Respiratory Syndrome (SARS) that took place in early 2003.

1.1 Background of Study

Considering WHO labelling the novel coronavirus as a global epidemic stresses the threat of biological warfare. One of the precautions would entail constructing a universal epidemic surveillance model, which is essential. There is a delay in relaying information on disease spread from doctors in hospitals to governing boards. This incident can be observed from the outbreak of novel coronavirus as a doctor, Li WenLiang, had identified this new virus and began to inform other medical professionals when first encountering such a disease. However, he was withheld from information sharing as it was yet to be verified by governing boards and was labelled as 'igniting unnecessary concerns by spreading false information'. It is lamented by many that the disease could have been better well contained if encountering such a deadly disease is communicated efficiently. Hence, this case illustrates the need for an automated epidemic surveillance system to facilitate communication on conditions and automatic disease monitoring.

Public health surveillance faces the challenge of the early identification of potentially high mortality and high-risk infectious diseases. As the incoming data volume from public health agencies is large, surveillance analysts cannot manually analyse them. Digitalised data collection methods then facilitate automating the outbreak detection process (Lombardo & Buckeridge, 2007).

As the epidemic spreads, the care-seeking infected population will provide clues to the background statistics. Identifying discrepancies in the number of infections is fundamental to detecting an outbreak. The task of processing outbreaks is to analyse large amounts of data, and issue alerts to epidemiologist on anomalies (Khameneh, J., & Nastaran., 2014). Automated surveillance systems can screen clinical data from various origins for quick and accurate detection of disease's possible outbreaks.

Mechanical health monitoring systems use statistical algorithms to discover anomalies for the investigation and design of preventive measures.

The advancement of technologies has empowered humanity with great tools in the scientific research process. Based on the seriousness of the epidemic, precautionary measures should be taken steps in advance to control and prevent the next outbreak. This entails the importance of epidemic surveillance to study the spread pattern of diseases to track their source and find a way to halt the spread process.

Spatio-temporal endemic epidemic model is the primary approach adopted by MC-Health; this model can track the Susceptible, Infected and Removed (SIR) developmental stages of the disease and provide geospatial variance for the authority to comprehend the disease spreads' characteristics and movement. This approach's major weakness is its inability to interpolate past missing information or extrapolate to project development trend. It is merely a tool for reporting current status based on whatever input data obtained. This system may lack sensitivity for controlling covid-19 which can be crucial for life-saving purposes.

As such, we look forward to the Early Aberration Reporting System (EARS), a drop-in surveillance system with its adaptive detection approach in signalling alarms by projecting the next disease outbreak wave. Whereas Farrington's model may patch up missing history and project trends with a minimum amount of data. This property could complement the shortcoming of the Spatio-temporal model.

As the EARS model is typically built for handling localised data to combine the time series at source from different locations, we need a system that is capable of handling overdispersion data or clusters of data. Hence, we consider the model based on Farrington's algorithm.

A more detailed description and evaluation of the three methods mentioned above are covered in this research study's literature review section. The discussion above highlights that these three main routes of developing an abbreviation detection system may have its strength and weakness. In this research study, we will build a version of implementation for each of the three approaches to investigating their properties as a covid-19 surveillance system to develop suggestions to integrate different methods' strength to compensate for the individual system's weakness.

1.2 Importance of the Study

In this research study, each algorithm's performances will be studied using the R surveillance package's algorithms consolidated by Höhle. The model's adaptability is analysed with the spread of covid-19 that is sudden and unsuspecting. Consecutively, insights regarding the distance of the covid-19 can also be reviewed based on the given dataset provided by European Centre for Disease Prevention and Control time-series datasets. This can be achieved by studying the correlation of infected cases as the days goes on using Farrington's model, also known as the quasi-Poisson regression model. Farrington's model can detect inconsistently reported data and project these missing data to create a whole picture of how the disease spread over time. The spread of outbreaks can be better studied through data visualisation using geospatial data obtained from the geotagged patient's medical dataset, incorporated into the Spatio-Temporal Endemic-Epidemic modelling coded on Python using the geopandas package. The movement of covid-19 can be studied to visualise better how the spread of this disease is affected and possibly controlled. Learning the disease spread solely based on time series and geographical distribution is not enough. The incorporation of alarming systems is crucial for an epidemic study. Thus, the Early Aberration

Reporting System (EARS) model, a surveillance model created by CDC, is chosen to study its performance and understand its pattern recognition in a pandemic event. This study is significant as nobody has anticipated covid-19 to outlast its predecessor (the SARS outbreak) while affecting our livelihood for two years and counting. Most studies of the models discussed were conducted with seasonal diseases or diseases rich in historical data. Hence, this research study could enable insight into the models' shortcomings towards this non-historical but long-lasting pandemic and demonstrate how they complement each other for monitoring covid-19.

1.3 Problem Statements

Epidemic surveillance is imperative in the covid-19 pandemic era. This virus outbreak requires attention in building a surveillance system that enables us to know the current status of covid-19 and be quick to detect any situational changes. Based on the literature review, the epidemic surveillance models discussed can better foresee and detect the occurrence of an epidemic as it allows real-time monitoring of disease spread based on data input. The three distinct epidemic surveillance models each have unique approaches in filtering surveillance data to monitor disease spread, which helps supplement researchers with quantitative information extracted through the model's given perspective. However, each model was described in its positive elements without comparing its virtues and shortcomings side by side.

These models were often studied using diseases that were familiar or reoccurring during typical epidemic seasons. Therefore, there is sufficient background data to compare the disease spread scale and check the epidemic surveillance models' reliability. On the other hand, rare diseases such as the coronavirus provide little to no background data for research studies. The models reviewed in the research studies

often rely on such background data to slowly study and identify an epidemic's event but rarely tested under pressure with diseases that spread rapidly.

Moreover, information regarding the coronavirus is not reflected on a large scale to observe and identify the possible disease hotspot or origins in real-time. Hence, there is a delay in relaying information on logging diseases between hospitals and governing boards. This problem heavily impacts the reliability of the epidemic surveillance models by human error.

1.4 Objectives

To incorporate machine learning on R programming for three essential epidemic surveillance models, including the Early Aberration Reporting System (EARS), quasi-Poisson regression method and Spatio-Temporal Endemic-Epidemic modelling. This research study implements a version of the three algorithms' sustainability with distinct approaches towards monitoring and determining an epidemic's event using the present data from the recent covid-19 disease outbreak. The main objective is to investigate three models' properties as a surveillance system for covid-19 and identify each model's best properties to integrate them to form a more holistic surveillance system that can wholesomely combat a sudden pandemic outbreak.

Throughout this research, the recommendation will be formed by comparing their performances in adapting to the rapid spread of diseases with little to no background data to monitor the area of study closely. This issue is raised as incomplete surveillance data collected as infected individuals are undetermined or unreported. Sparse cases of background data are recorded. The algorithms' performance can be evaluated through the probability of false alarms raised if the disease is not declared as an epidemic, along with the likelihood of accurately identified epidemic events.

Per this research study, a network visualisation of the disease spread will be mapped through the R Studios client to illustrate the studied epidemic impact. Moreover, it acts as a visual aid in comparing the performance of each machine learning algorithm.

1.5 Project Scope

This project aims to investigate the performance of three extensively used multi-purpose outbreak detection algorithms in monitoring daily surveillance data gathered through tracking the spread of covid-19 disease. Current covid-19 outbreaks present a challenging yet pressing needed situation to search for the best aberrant event detection algorithm. Therefore, we propose to evaluate three practical approaches in developing such an algorithm, namely:

- CUSUM based EARS are developed and represented the standard system at the United States Centres for Disease Control and Prevention (CDC).
- The improved quasi-Poisson regression-based method developed and currently used in Communicable Disease Surveillance Centre (CDSC), the United Kingdom, for weekly detection of infectious disease outbreaks.
- Spatio-Temporal Endemic-Epidemic method is a statistical surveillance system and multivariate counting process-based SIR model developed by the Munich Centre of Health Sciences (MC-Health), Germany. It raises potential alarms by screening daily counts, which is suitable for the emergency of covid-19 disease outbreak.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview of Epidemic Surveillance Models

Method	Year	Paper	Used Cases
Early Aberration Reporting System (EARS)	2005	Multifaceted Syndromic Surveillance in a Public Health Department Using the Early Aberration Reporting System	The EARS method pioneered during the dawn of terrorism in the United States. This method heavily relies on past trends and medical logs, which is outdated over the decade.
	2009	CDC releases first report on terrorism preparedness, and emergency response funded activities	
	2010	Assessing the Effectiveness of the Early Aberration Reporting System (EARS) for Early Event Detection of the H1N1 ("Swine Flu") Virus	
	2013	Implementing a Weather-Based Early Warning System to Prevent Traffic Accidents Fatalities	
	2011	Assessing the Early Aberration Reporting System's Ability to Locally Detect the 2009 Influenza Pandemic	
Quasi-Poisson Regression	2014	A Quasi-Poisson Approach on Modelling Accident Hazard Index for Urban Road Segments	The Quasi-Poisson Regression method is a generalised

	2015	Time series regression model for infectious disease and weather	linear model that discusses the factors involved in developing an epidemic. It's very versatile and accountable for cases where there is an over-dispersion of data.
	2016	Quasi-Poisson versus negative binomial regression models in identifying factors affecting initial CD4 cell count change due to antiretroviral therapy administered to HIV-positive adults in Northï West Ethiopia (Amhara region)	
	2020	Investigating the Significant Individual Historical Factors of Driving Risk Using Hierarchical Clustering Analysis and Quasi-Poisson Regression Model	
	2020	Modelling Burglar Incidents Data Using Generalized and Quasi Poisson Regression Models: A Case Study of Nairobi City County, Kenya	
Spatio-Temporal Endemic-Epidemic	2017	Applying Spatio-temporal models to assess variations across health care areas and regions: Lessons from the decentralised Spanish National Health System	The Spatio-Temporal Endemic-Epidemic method is modern and very recently used for related covid-19 pandemic studies.
	2018	Joint Spatio-Temporal Shared Component Model with an Application in Iran Cancer Data	

	2019	Spatio-temporal pattern of two common cancers among Iranian women: An adaptive smoothing model	It allows for the investigation of geospatial mapping of diseases.
	2020	Right and Yet Wrong: A Spatio-Temporal Evaluation of Germany's COVID-19 Containment Policy	
	2020	A Spatio-temporal approach to short-term prediction of visceral leishmaniasis diagnoses in India	

Table 1: Summary of Cases for each surveillance model

2.2 Implementation of Epidemic Surveillance Models on the Healthcare Industry

The Early Aberration Reporting System (EARS) is a biosurveillance system dynamically developed by the Centres for Disease Control and Prevention (CDC). The EARS written in SAS acts as drop-in surveillance for wide-ranging incidents that can operate with minimal to no baseline data (Hagen *et al*., 2011). It is currently adapted for routine biosurveillance in the United States to equip state and local health departments to apply aberration detection methods of significant sensitivity degrees. The EARS encases quality-control charts, moving average (MA), and cumulative sum variations (CUSUM). This algorithm is designed to circumvent the challenges opposed by modelling the syndrome's baseline trend. The challenge is complicated by the individuality, serial correlation, seasonality, and daily fluctuation of the syndromic data (Henning, 2004). It conducts early event detection (EED) by surveilling for syndromic growth derived from chief complaints.

Syndromic outbreaks are monitored based on the presence of keywords in chief complaint records. This procedure is achieved in two fundamental steps. First, the occurrence of specific terms, word variant, common typos ad associated medical abbreviations and jargons are imperative in monitoring the syndrome's development. Following that, the syndromes are then analysed to deduce the presence of a syndrome. Consequently, every individual included in the data for each syndrome is concluded in contracting the syndrome or not. An example of the EARS application includes locally detecting the seasonally occurring H1N1 pandemic. The data was collected on a series of significant chief complaints full of jargon, acronyms, and medical personnel's abbreviations (Hagen *et al*., 2011). Syndrome indicators are siphoned from chief complaints by parsing keywords, including variations of misspelled vital words.

The results derived from the EARS can be adjusted at the local level by modifying a syndrome's definition. This incident was observed in the emergence of the 2009 H1N1 virus, whereby the syndrome's definition was expanded to increase the probability the system would signal during an outbreak. However, there was no consideration for individuals in the study that have taken a flu shot and were incorrectly detected under the syndromic surveillance system in the presence of the word "flu" in their chief complaint text. Hence, this resulted in a syndrome that substantially increased in the daily counts, and the estimated rate of such syndrome significantly exceeded the actual reported rate. It is evident that the EARS' detection algorithms heavily rely on the choice of syndrome definition, which impacts the daily syndrome counts (Hagen *et al*., 2011). The EARS C1, C2 and C3 variants are implemented to detect public health interest outbreaks, including the influenza season's start. Contrary to epidemiologists, little is known about how these algorithms'

signalling patterns correspond to identifying public health interest events (Watkins *et al*., 2008).

Following that, we discuss the detection system implemented at the Communicable Disease Surveillance Centre (CDSC) described in Farrington *et al.* (1996), which is designed to overcome a wide array of organism frequencies and temporal patterns. The improved quasi-Poisson regression-based method, also known as Farrington Flexible, was developed and currently used in the Communicable Disease Surveillance Centre (CDSC), the United Kingdom, for weekly detection of infectious disease outbreaks. It is an ideal generalised linear model when there is an over-dispersion of data anticipated. Compared to the Poisson distribution, where the mean is strictly equal to the variance, the quasi-Poisson regression's variance is a mean linear function. This mean and variance structure is unique to the quasi-Poisson regression as it allows the estimation of an epidemic risk in varying risk environments. Based on a study comparing the EARS method and Farrington Flexible, it is observed that the Farrington Flexible is best suited for a multi-purpose daily surveillance system as it has higher sensitivity and specificity. This justifies the stance that a quasi-Poisson regression is preferred in the event of accurately predicting daily alarms (Noufaily *et al*., 2019). The quasi-Poisson regression is extensively used in epidemic studies, including Hermansen *et al.*, to uncover factors correlating to Nontuberculous mycobacteria in Denmark and accommodate for data overdispersion (Hermansen *et al*., 2017).

Spatio-Temporal Endemic-Epidemic model is a statistical surveillance system and multivariate counting process-based SIR model developed by Munich Centre of Health Sciences (MC-Health), Germany. Monitoring daily counts for potential alarms is a concept that has not been extended in the literature. The focus has been on daily

surveillance due to a covid-19 disease outbreak (Meyer *et al*., 2017). Traditionally, epidemic models are used to characterise the spread of a transmittable disease in a population. Occasionally, a partitioned population sample is recorded by categorising samples into one of the three states: (S)usceptible, (I)nfectious, or (R)emoved. Depending on the types of Spatio-temporal data, there are three variants of regression-oriented modelling frameworks based on spatial and temporal resolution (Meyer *et al*., 2017). Its leading feature is the additive decomposition of disease risk based on endemic and epidemic characteristics.

The endemic component studies new event risks due to external factors that do not depend on the chronicle of the epidemic process. In the context of infectious diseases, these determinants can include seasonality, population density, socio-demographic variables, and vaccination rates, relative variables that are influenced by time and location. Then, clear accreditations of events are verified by the epidemic component, as guided by past observations (Höhle, M., 2010). First, the Spatio-temporal point patterns are displayed when the entire region is continuously monitored for infection, transmitted on time, georeferenced and potentially enhanced by specific data of an event. Such continuous time flow data is an exciting realisation of space-time processes. The second type of data we consider is the history of the individual units traced from time to time as they transition between the SIR states. These data are consistent with the SIR spatial model structure represented as a multivariate time point process. The third data type is event counts by region and period, collecting the individual event data mentioned first. This data type is employed when there are complications with data due to privacy protection or filtered data due to reporting intervention. Such regional count time series matches the multivariate negative binomial time-series model. The three model classes are all constructed upon the

Poisson branching process's fundamentals with the immigration approach proposed by Held *et al*. (2005).

CHAPTER 3

METHODOLOGY

3.1 Research Flowchart

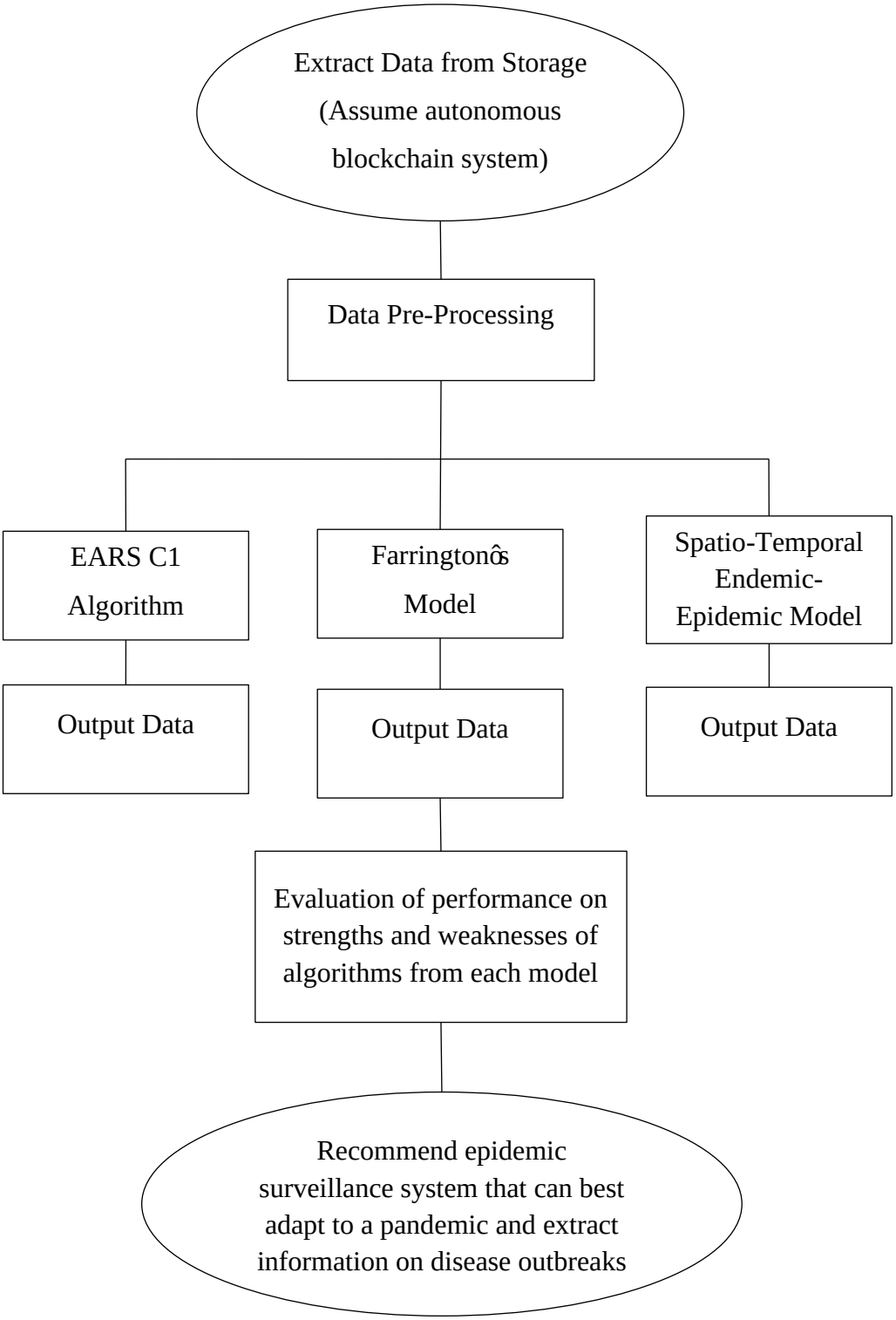


Figure 1: Flow chart of research study

From Figure 1, this research study has assumed a hypothetical environment that allows the reporting of surveillance data to be instantaneous. Hence, it is hypothetically acknowledged that, in this research study, surveillance data is automatically extracted via a blockchain system. This blockchain system involves the real-time logging of chief complaints from related medical staffs. The chief complaint is then interpreted into covid-19 symptoms and diagnosis, which will then be translated into the surveillance data. However, in reality, only reports of diagnosed cases are published and then extracted from ECDC. This hypothetical setting is assumed for the EARS model that facilitates epidemic surveillance modelling through chief complaints.

Nevertheless, this research study would bypass the diagnosing of covid-19 through chief reports and directly model through a confirmed number of cases or deaths. This setting also makes sense in reality as data is only publicly shared through WHO's daily updated health reports. Hence, there could only be data of verified cases.

After the surveillance data has been extracted, the data is transformed into an $n \times m$ matrix format to study the development of covid-19 over the days. The transformed data is now fitted into the three distinct epidemic surveillance models for this research study. In turn, the output data for each model is obtained. The model's best features will be determined through its data modelling and standard comparison parameters such as sensitivity, specificity and false alarm rate. The model will also be compared to adapt to rapid disease outbreaks with almost little to no historical surveillance data provided, which is often the most alarming situation that humankind least anticipated. Finally, the best surveillance model will be studied to select for its capability to predict disease outbreaks over the pandemic or best model the overwhelming spike of infected cases due to disease outbreaks.

3.2 Type of Data to be Stored

The surveillance data and related information are encoded in multivariate time series of counts. The counts are denoted as $y_{it}; i = 1,2, \dots, m,\ t = 1, \dots, n$, where n entails the length of time series (n days, weeks, years) and m is the number of entities such as geographical regions being monitored (m geographical units). The critical slots for each surveillance time series class describe the observed counts and the corresponding aggregation periods. The observed counts,y_{it} are stored in the $n \times m$ matrix observed while the remaining slots characterise time (Paul, M. *et al.*, 2008).

The surveillance data cluster will be updated weekly by ECDC through globally published health reports, including WHO and CDC. The models detailed as studied through literature review will be programmed through R Studios to learn the response and rate of alerting an epidemic.

Once all surveillance data has been collected, it will be fitted into the three epidemic surveillance models selected for this research. Each model's efficiency and accuracy will be determined through the ratio of false-positive rates (FPR) and true favourable rates (TPR). Aside from analysing each algorithm's statistical accuracy, the models will also be compared for their contribution in producing informative output data in observing an epidemic outbreak.

3.3 Qualitative Analysis Parameters

The objective of outbreak detection algorithms is instantaneous detection of outbreaks with minimal false alerts. The following three quantities used to assess detection algorithms' accuracy are specificity, sensitivity, and timeliness (Wagner *et al*., 2006).

Specificity is the probability of no alarm when there is no outbreak:

$$specificity = P(alarm = 0 \,|\, outbreak = 0) = \frac{n(alarm=0, outbreak=0)}{n\,(outbreak=0)} \quad (1)$$

where $n(alarm = 0, outbreak = 0)$ is the number of ordinary days. The ordinary days in this context refers to days where disease outbreaks do not exist, in which the algorithm does not raise the alarm and $n\,(outbreak = 0)$ is the number of standard days in an analysis interval.

Sensitivity is the probability of raising the alarm given that an outbreak has taken place:

$$sensitivity = P(alarm = 1 | outbreak = 1) = \frac{n(alarm=1, outbreak=1)}{n\,(outbreak=1)} \quad (2)$$

where $n(alarm = 1, outbreak = 1)$ is the number of alerts in outbreak days, and $n\,(outbreak = 1)$ is the number of outbreak days. There exist two translations of the sensitivity of a detection algorithm.

One might tally the outbreak days. The algorithm signals an alarm with sensitivity in the equation's numerator and the total number of outbreak days in the denominator. This equation is a probability called sensitivity per day works by accurately identifying outbreak days. Otherwise, the number of detected outbreaks can be measured in the numerator, regardless if multiple alarms been signalled within an episode. The number of outbreak intervals is then summed in the denominator, which varies from the number of outbreak days if an outbreak persists. In an epidemiological context, the second interpretation of sensitivity, called sensitivity per outbreak, makes more sense because excessive alerts within a single episode are trivial.

The quantity of ***false alarm rate*** is the probability that an algorithm classifies a typical day incorrectly and raises the alarm in the absence of an outbreak:

$$falseAlarmRate = P(alarm = 1|outbreak = 0) = \frac{n(alarm=1, outbreak=0)}{n\,(outbreak=0)} \quad (3)$$

False alarm rate and specificity would both sum up to one, so it is used in evaluations interchangeably (Wagner *et al*., 2006).

CHAPTER 4

SOFTWARE IMPLEMENTATION

4.1 Overview of the Software

The software covered in this research study includes R and Python programming. The epidemic mentioned above surveillance models is derived from the surveillance package maintained by Sebastian Meyer. The surveillance package comprises monitoring application specialising in aberration detection using count data time series, particularly from public health surveillance of infectious diseases. Hence, this open-sourced surveillance package on R best fits this research study involving ECDC's publicly available surveillance data on covid-19. Hitherto, the surveillance data for this research has only been studied from the lenses of a statistician. It would be best to mention that diseases develop outbreaks as it is allowed to spread from person to person across the region. Therefore, a geospatial analysis of disease spread using Python's geopandas package will be implemented to understand the disease spread better and assist in real-time detection of outbreak points geographically.

4.2 Program Setup and Display

The surveillance package (Version 1.19.0) including EARS, Farrington (Quasi-Poisson) and Spatio-Temporal models, are executed on RStudio. The geospatial demonstration of the Spatio-Temporal model is visualised with the aid of the geopandas package from Python programming.

4.2.1 Data Pre-Processing

The worldwide covid-19 dataset implemented for this research study was extracted from the European Centre for Disease Prevention and Control (ECDC). The covid-19 number of cases and death reports collected spanned from January 22 2020 to February 4 2021.

The data in this research study was collected from January to adopt the adjustments made by ECDC on switching from daily reporting to a weekly reporting schedule, thus discontinuing all daily updates from December 14 2020. The reporting channel achieves this by aggregating the number of cases and deaths reported worldwide and publishing it every Thursday. A snapshot of the raw ECDC's globally reported data on covid-19 cases and deaths is shown below.

K26 | Europe

	A	B	C	D	E	F	G	H	I	J	K	L
1	dateRep	day	month	year	cases	deaths	countries	geoId	countryte	popData2	continentExp	
2	4/4/2021	4	4	2021	3693	23	Austria	AT	AUT	8901064	Europe	
3	3/4/2021	3	4	2021	3077	33	Austria	AT	AUT	8901064	Europe	
4	2/4/2021	2	4	2021	3058	23	Austria	AT	AUT	8901064	Europe	
5	1/4/2021	1	4	2021	3107	28	Austria	AT	AUT	8901064	Europe	
6	31/3/2021	31	3	2021	2810	28	Austria	AT	AUT	8901064	Europe	
7	30/3/2021	30	3	2021	3240	24	Austria	AT	AUT	8901064	Europe	
8	29/3/2021	29	3	2021	2667	11	Austria	AT	AUT	8901064	Europe	
9	28/3/2021	28	3	2021	3896	27	Austria	AT	AUT	8901064	Europe	
10	27/3/2021	27	3	2021	3487	24	Austria	AT	AUT	8901064	Europe	
11	26/3/2021	26	3	2021	3516	16	Austria	AT	AUT	8901064	Europe	
12	25/3/2021	25	3	2021	3262	29	Austria	AT	AUT	8901064	Europe	
13	24/3/2021	24	3	2021	2715	40	Austria	AT	AUT	8901064	Europe	
14	23/3/2021	23	3	2021	2306	23	Austria	AT	AUT	8901064	Europe	
15	22/3/2021	22	3	2021	2918	19	Austria	AT	AUT	8901064	Europe	
16	21/3/2021	21	3	2021	4051	22	Austria	AT	AUT	8901064	Europe	
17	20/3/2021	20	3	2021	3304	15	Austria	AT	AUT	8901064	Europe	
18	19/3/2021	19	3	2021	3507	33	Austria	AT	AUT	8901064	Europe	
19	18/3/2021	18	3	2021	2814	33	Austria	AT	AUT	8901064	Europe	
20	17/3/2021	17	3	2021	2487	23	Austria	AT	AUT	8901064	Europe	
21	16/3/2021	16	3	2021	1910	22	Austria	AT	AUT	8901064	Europe	
22	15/3/2021	15	3	2021	2664	17	Austria	AT	AUT	8901064	Europe	
23	14/3/2021	14	3	2021	3325	17	Austria	AT	AUT	8901064	Europe	

Figure 2: Snapshot of raw data obtained from ECDC's global covid-19 report

Hence it is acknowledged that the data collection process differs from the traditional EARS method of deriving cases through chief complaints as this study uses already reported the number of cases of individuals infected by covid-19. This factor is accounted for as data was not readily available for research studies until the disease

was identified. Therefore, this study is focused on the model's adaptability towards rapid incoming disease data on the pandemic with little to no historical data supplied. Data transformation was applied to fit the data into the models in R's surveillance package. To achieve an $n \times m$ matrix format for the surveillance data, Excel's pivot table function was used to transpose the dates into n rows and allocate the number of cases occurring each day to its respective occurring days. The m column, as shown in the snapshot below, is the pivot of countries where data is collected. I retained some valuable information for several countries identified by their province by adding a province column. However, the models will read the data by matching the number of cases occurring in the day with its country as the models will study the data from column B and onwards. Moreover, to supplement information for the Spatio-Temporal Model, each country's longitude and latitude were obtained from Google and allocated to each country using the "VLOOKUP" function for precise information match. A demonstration of the transformed data is detailed in the snapshot attached below, whereby the data is now converted to an $n \times m$ matrix format.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Province/	Country/R	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	1/30/20	1/31/20	2/1/2020	2/2/2020	2/3/2020	2/4/2020	2/5/2020	2/6/2020	2/7/2
2		Afghanist	33.93911	67.70995	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3		Albania	41.1533	20.1683	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4		Algeria	28.0339	1.6596	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5		Andorra	42.5063	1.5218	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6		Angola	-11.2027	17.8739	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7		Antigua ar	17.0608	-61.7964	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8		Argentina	-38.4161	-63.6167	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9		Armenia	40.0691	45.0382	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	Australian	Australia	-35.4735	149.0124	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	New Sout	Australia	-33.8688	151.2093	0	0	0	0	3	4	4	4	4	4	4	4	4	4	4	4	
12	Northern	Australia	-12.4634	130.8456	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	Queensla	Australia	-27.4698	153.0251	0	0	0	0	0	0	0	1	3	2	3	2	2	3	3	4	
14	South Aus	Australia	-34.9285	138.6007	0	0	0	0	0	0	0	0	0	0	1	2	2	2	2	2	
15	Tasmania	Australia	-42.8821	147.3272	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	Victoria	Australia	-37.8136	144.9631	0	0	0	0	1	1	1	1	2	3	4	4	4	4	4	4	
17	Western A	Australia	-31.9505	115.8605	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18		Austria	47.5162	14.5501	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19		Azerbaijar	40.1431	47.5769	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20		Bahamas	25.02589	-78.0359	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21		Bahrain	26.0275	50.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Figure 3: Snapshot of transformed surveillance dataset in $n \times m$ matrix format

4.2.2 Explanatory Data Analysis of Preliminary Data and Findings

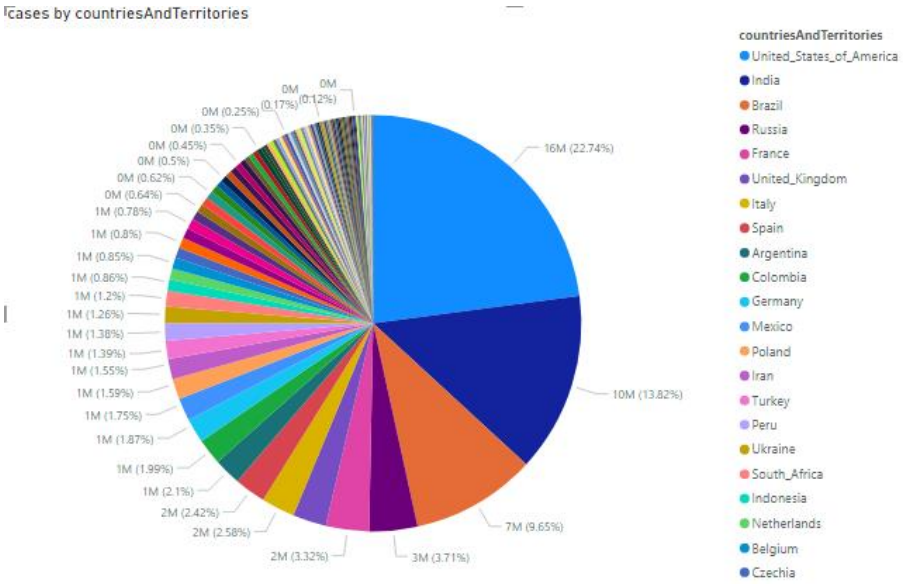


Figure 4: Pie Chart Summary on Number of Infected Cases by Countries and Territories

It is observed from the pie chart summary that the United States of America has the highest amount of infected cases, followed by India and Brazil. The rest of the countries are in smaller proportions than the three largest countries from the dataset.

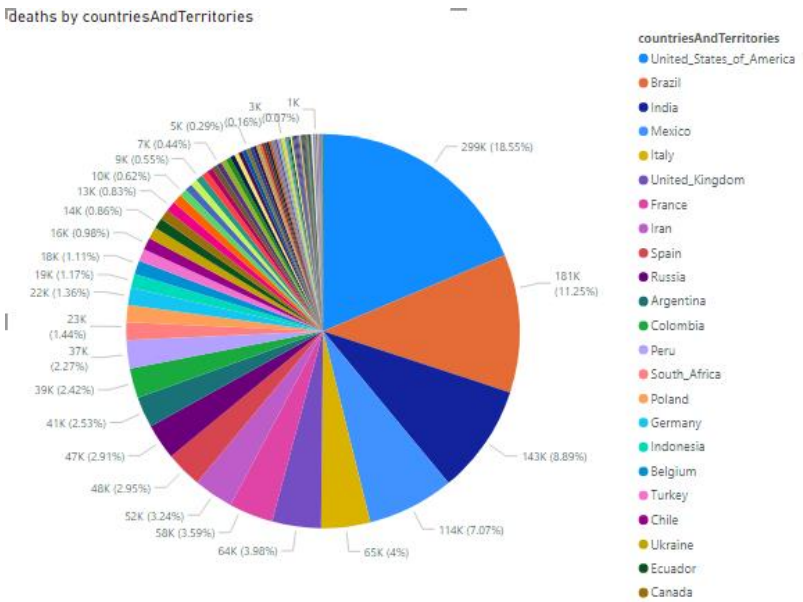


Figure 5: Pie Chart Summary on Number of Deaths by Country and Territories

Like Figure 4, the United States of America has a relatively high death rate caused by covid-19. However, it can be observed that Brazil has higher mortality rates than India as it was the third leading number of infected cases but has the following highest number of deaths. The proportion of death cases for India is closely proportionate to the number of death cases from Mexico.

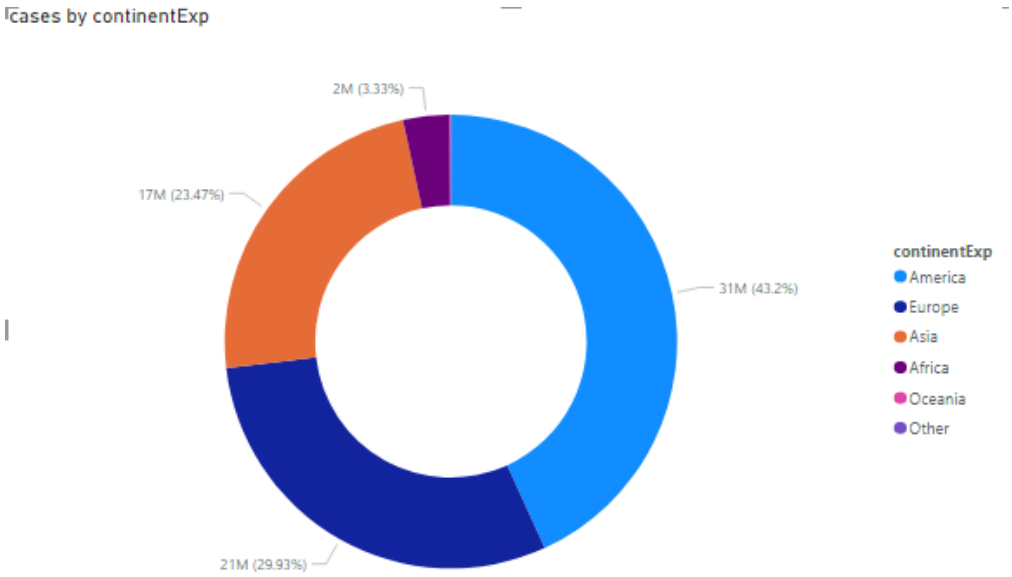


Figure 6: Donut Chart of Number of Infected Cases by Continent

Based on the donut chart from Figure 6, it is evident that America is leading by the number of infected cases because it is leading in the number of infected patients and deaths by countries. Due to the developments of disease spread in Europe that is recently rampant, it is also reflected in this dataset. It has the second-highest number of infected cases amongst other continents. On the other hand, Asia is third leading in the number of infected patients ranked by continents.

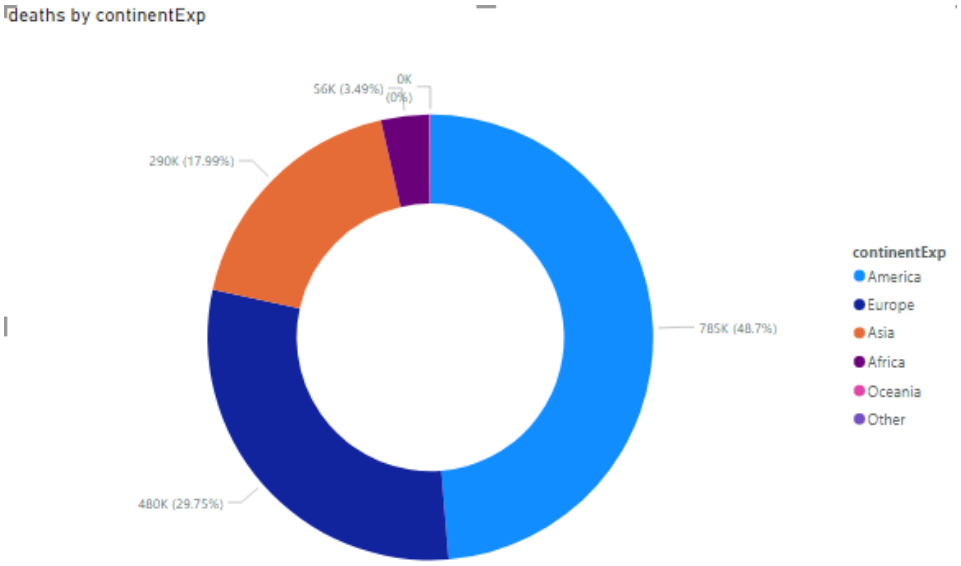


Figure 7: Donut Chart of Number of Deaths by Continent

By comparing Figure 6 with Figure 7, the ranking of infected cases by continents is reflected in continents' scale of deaths. However, it can be deduced from Figure 7 that covid-19 has relatively high mortality rates from a rapidly uncontrolled disease spread. This phenomenon can be observed because there are many infected cases for America, while the resulting death cases are significantly high compared to other countries.

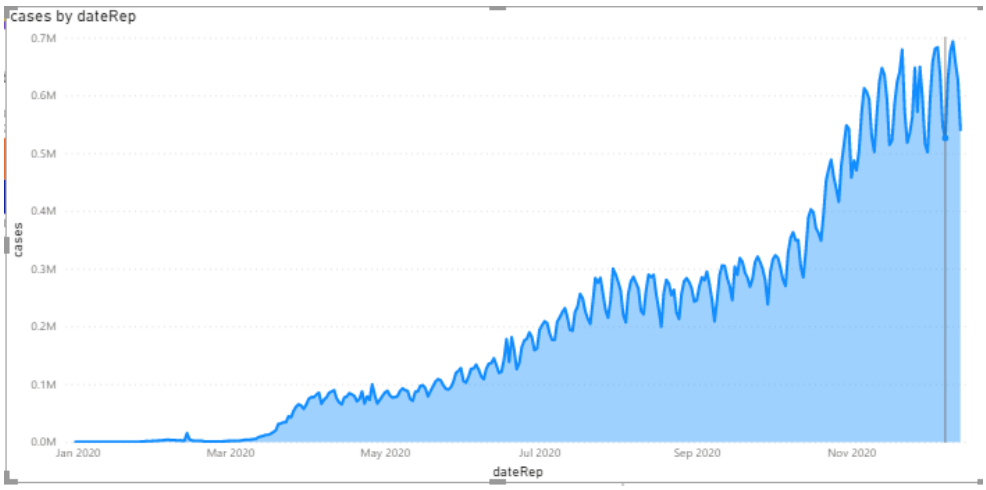


Figure 8: Area Chart of Number of Infected Cases diagnosed from January 22, 2020, to February 4, 2021

As seen in Figure 8, the number of infected cases for covid-19 shows an overall exponential increase from April 2020. It seems to plateau around November 2020 as the disease spread has become relatively potent.

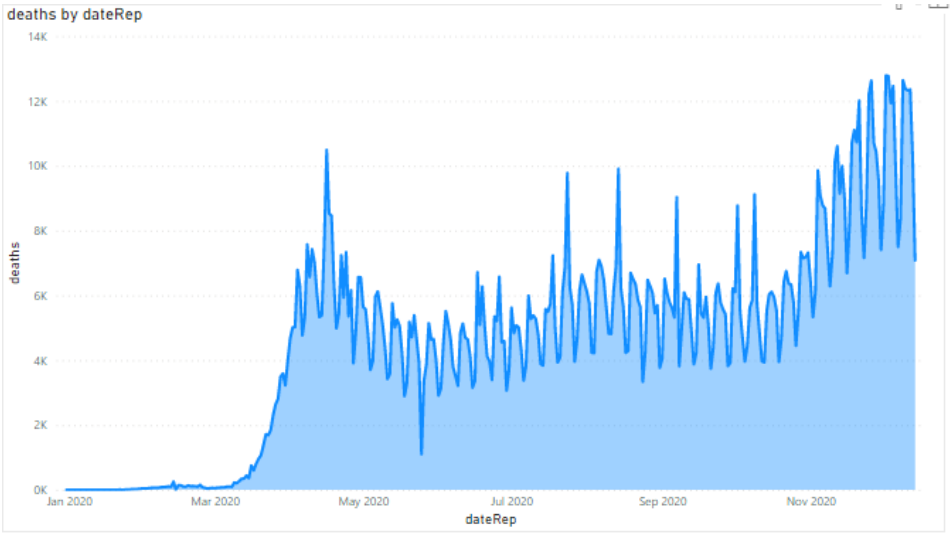


Figure 9: Area Chart of Number of Deaths occurring from January 22, 2020, to February 4, 2021

From Figure 9, the number of death cases spiked around April 2020-time period. It displayed an erratic stochastic pattern of sharp increases in death cases, followed by a steep decreased number of death cases.

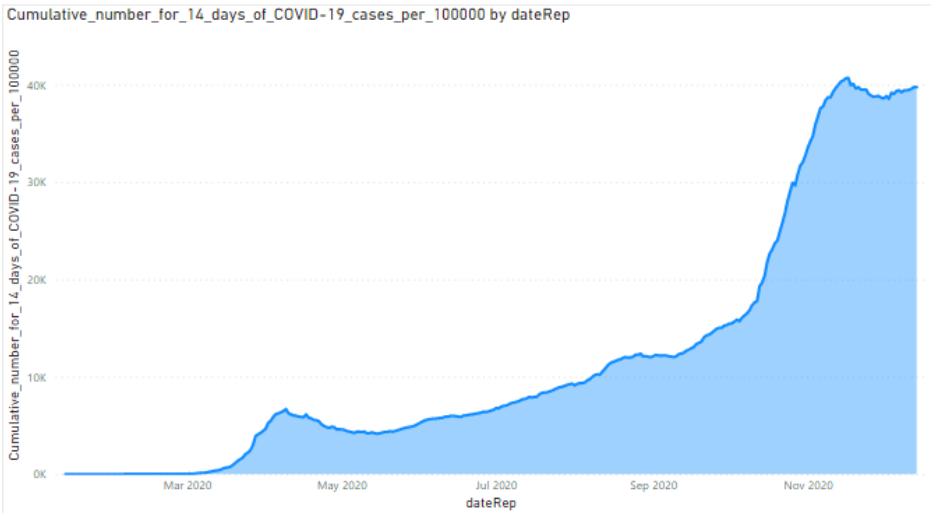


Figure 10: Area Chart of Cumulative Infected Cases per 100,000 individuals accumulated over 14-days interval from January 22, 2020, to February 4, 2021

The number of infected cases for the covid-19 disease is observed to be spreading exponentially from April 2020 to November 2020, as observed from the cumulative number of infected cases per 100,000 individuals per 14 days' interval. The spread of disease hit a plateau after November 2020 whereby the following cases then proceed to decrease slightly but continue to increase steadily.

4.2.3 Construction of Epidemic Surveillance Models

A syndrome refers to "a set of symptoms or conditions occurring together and suggests the presence of a certain disease or an increased chance of developing the disease". A syndrome may indicate a bioterrorism agent's release or outbreak of natural diseases (Henning, 2004).

Data for syndromic surveillance can be commonly sourced through records of death-on-death certificates, patient diagnosis through emergency and clinical visits, school or work absenteeism and over the counter medication sales in pharmacies. In essence, the data for syndromic surveillance is referred to as surveillance data gathered in time series. During the absence of an outbreak event, these surveillance data can also be identified as background data for epidemic surveillance studies. The care-seeking infected population adds an outbreak signal to the surveillance data when an outbreak occurs (Khameneh, J., & Nastaran., 2014). Hence, the automated biosurveillance system analyses the surveillance data to find patterns of a disease outbreak.

Background data accumulated over the years of monitoring a disease can help observe long-term trends, such as surveying and identifying the seasonal outbreaks of bird flu in the event of bird migration season (Khameneh, J., & Nastaran., 2014). Many algorithms were developed to extract helpful information about disease control or track an epidemic's source over a timeline.

The three following distinct outbreak detective algorithms are reviewed for their features and approach to screening surveillance data. The Early Aberration Reporting System (EARS) was developed by the CDC, consisting of a class of quality-control charts, moving average (MA) and variations of the cumulative sum (CUSUM). This algorithm is designed to circumvent the challenges opposed by modelling the

syndrome's baseline trend, which is complicated by the discreteness, serial correlation, seasonality, and daily fluctuation of the syndromic data (Henning, 2004).

CUSUM (Cumulative Sum) charts improve the ability to detect small shifts by charting a statistic that incorporates current and previous data values from the process. Specifically, the CUSUM chart plots the cumulative sums of the sample values' deviations from a target value. The inclusion of several samples in the cumulative sum results in greater sensitivity for detecting shifts or trends over the traditional Shewhart charts (Li, Y. *et al.*, 2016).

The QC charts in EARS use daily syndromic counts or incidences (daily counts of a specific syndrome divided by total ED volume for the day) y_t between day $t - K + 1$ and day t (current day) to derive a monitoring statistic, m_t. The monitoring statistic is

$$m_t = y_t \ ,$$

$$m_t = \frac{1}{K}\sum_{k=1}^{K} y_{t-k+1} \text{ and } m_t = (1-\lambda)\sum_{k=1}^{\infty} \lambda^{k-1}\, y_{t-k+1} \qquad (4)$$

for P-chart, MA, and EWMA, respectively.

P-chart uses the selected day's data; MA is the average of K days before and including the day chosen; EWMA is a weighted average of all previous days with an exponentially decreasing weight given to days further away from the present day. The system generates a signal if mt exceeds a threshold $c\sigma m$ above the expected level μm. The PSD of a single alert (sensitivity) is the probability Pr $(m_t - \mu_m > c\sigma m)$ given an outbreak. The constant c determines the threshold in multiples of the standard deviation σm. Because the distribution of m_t is complex, EARS uses a sample estimate of μ_m and σm based on data in a baseline window of B days: $y_t - B -$

$g, y_t - B + 1 - g, \ldots\ldots, y_t - 1 - g$ with a gap of g days before the present day t. Specifically,

$$\mu_m = \frac{1}{B}\sum_{b=1}^{B} y_{t-b-g} \text{ and } \sigma_m^2 = \frac{1}{B-1}\sum_{b=1}^{B}(y_{t-b-g} - \mu_m)^2 \quad (5)$$

EARS also utilises three variations of the CUSUM method (*C1*, *C2*, and *C3*). *C1* uses data from the current day only and a baseline window of the preceding seven days: day $t - 7$ to -1 ($B = 7$ and $g = 0$). If *C1* generates a signal on day *t*, day *t* will become a part of the baseline for day $t + 1$, which might inflate the corresponding baseline mean μm, and reduce the PSD for that day. *C2* differs from *C1* by shifting the 7-day baseline window to the left with a gap of $g = 3$ days. As a result, its PSD of signalling on day 2 and 3 are not affected by a signal on day 1. Analytically, *C1* and *C2* are nearly equivalent in the absence of outbreaks, but *C2* is more sensitive than *C1* in signalling a continued attack past its onset. For this reason, *C1* was not evaluated. *C3* differs further from *C2* by using a partial sum of positive daily deviations for the current and two previous days ($t - 1$ and $t - 2$):

$$m_t = \frac{(y_t - \mu_m(t) - \sigma_m(t))^+}{\sigma_m(t)} + \frac{(y_{t-1} - \mu_m(t-1) - \sigma_m(t-1))^+}{\sigma_m(t-1)} I_{t-1} + \frac{(y_{t-1} - \mu_m(t-2) - \sigma_m(t-2))^+}{\sigma_m(t-2)} I_{t-2} \quad (6)$$

The EARS method is famous for its simplicity and relatively reliable source of development from the CDC. Hence, it is widely used in public health organisations as a basis for epidemic surveillance modelling. In this study, the *C2* algorithm will be selected to study the nature of the EARS algorithm whilst overcoming the limitation of the *C1* algorithm for the absence of a 2-day guard band interval. This argument is supported by the discussion earlier that the sensitivity of outbreak detection of the *C2* algorithm can be improved using a guard band because outbreaks spread over several days are not missed (Salmon, M. *et al.*, 2015).

Following that, we discuss the detection system implemented at the Communicable Disease Surveillance Centre (CDSC) described in Farrington *et al.* (1996), which is designed to overcome a wide array of organism frequencies and temporal patterns. There are five critical aspects of the statistical model for this quasi-Poisson regression-based algorithm.

Firstly, the seasonal variation is adjusted by basing the expected value in the current week on the counts observed in comparable weeks in the past. Let t be the current week of year h and b the number of years back to be considered. We take baseline counts only from weeks $t - w$ to $t + w$ (where w is the window half-width) of years $h - b$ to $h - 1$.

Following that, the log-linear model is used to take account of a linear trend if it is significant at the 5% level, $\log E(y_i) = \theta + \beta t_i$ (7)

where y_i is the count in week t . The model will be refitted without the log-linear model if the trend is not significant or results in an expected value outside the baselines' range.

Baseline weeks with outlying counts are then down-weighted to reduce their impact on current predictions. The weighting function on empirical grounds is assigned to very low weights to counts with large residuals. The weighting procedure is iterative, and the weights w_i at week t are defined by

$$w_i = \begin{cases} \gamma s_i^{-2} & \text{if } s_i > 1 \\ \gamma & \text{otherwise} \end{cases} \quad (8)$$

where γ is a constant such that $\sum_{i=1}^{n} w_i = n$ and s_i are scaled Anscombe residuals.

The statistical model is also a weighted quasi-Poisson model with mean μ_i and variance $\phi\mu_i$.

The dispersion parameter ϕ is estimated by $\hat{\phi} = \max\left\{\frac{1}{n-p}\sum_{i=1}^{n} w_i \frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i}, 1\right\}$ (9)

where $p = 1$ or $p = 2$ depending on whether a time trend is fitted. The scaled Anscombe residuals are $s_i = \frac{3}{2\hat{\phi}^{1/2}} \frac{y_i^{2/3} - \hat{\mu}_i^{2/3}}{\hat{\mu}_i^{1/6}(1-h_{ii})^{1/2}}$ (10)

where h_{ii} are the diagonal elements of the hat matrix. For Poisson data, for which $\phi = 1$, the s_i are the standardised Anscombe residuals.

Finally, the current expected value is calculated and the threshold value above which an observed count is declared unusual. The current expected count is estimated by $\mu_0 = \exp(\theta + \beta t_0)$ (11)

Traditionally, epidemic models are used to characterise the spread of an infectious disease in a population. Occasionally, a partitioned view of the population is recorded, thus categorising individuals into one of the three states: (S)usceptible, (I)nfectious, or (R)emoved.

Taking into consideration that a stochastic version of the modest homogenous SIR model in a closed population of size N, the hazard rate for a prone individual $i \epsilon S(t)$ to transition to the infectious state at time t, and thus the "force of infection" is

$\lambda_i(t) = \sum_{j \epsilon I(t)} \beta$ (12)

The index sets of presently susceptible and infectious individuals are denoted by $S(t), I(t)\{1, \dots \dots, N\}$ and likewise, the transmission rate refers to the parameter > 0.

The stochastic SIR model is supplemented by a distributional assumption about the duration when individuals are infective, which typically entails the exponential or the gamma distribution. The series of recovered individuals at time t is found as $R(t) = \{1, \dots \dots, N\} \backslash (S(t) \cup I(t))$.

The above homogeneous SIR model has since been extended in an assortment of ways, e.g., by additional states or population demographics.

Depending on Spatio-temporal data types, there are three variants of regression-oriented modelling frameworks based on spatial and temporal resolution.

First, the Spatio-temporal point patterns are displayed when the entire region is continuously monitored for infection, transmitted on time, georeferenced and potentially enhanced by specific data of a particular event. Such continuous time flow data can be seen as an exciting realisation of space-time processes. The second type of data we consider is the history of the individual units traced from time to time, as recorded when they become susceptible, infected, and potentially removed (neither at risk nor infectious). These data are consistent with the SIR spatial model structure represented as a multivariate time point process. The third data type is focused on event counts by region and period collected from individual event data mentioned first. This surveillance data is aggregated in such a way as to circumvent data privacy complications. These sectional count time series best fits the multivariate negative binomial time-series model. The three model mentioned above classes built upon the Poisson branching process's fundamentals with the immigration approach proposed by Held *et al*. (2005).

CHAPTER 5

RESULTS AND DISCUSSION

5.1 Selection of Epidemic Surveillance Models

Method	Year	Used by	Function	Limitation
Early Aberration Reporting System (EARS) method	2001	Centres for Disease Control and Prevention (CDC), United States	It is used to monitor syndromic outbreaks using keywords in chief complaint records.	Human errors easily influence this surveillance method as it relies on the choice of keywords selected from chief complaint records.
Quasi-Poisson Regression method	1996	Communicable Disease Surveillance Centre (CDSC), United Kingdom	Ideal GLM when there's an over-dispersion of data. This regression's variance is a linear function of the mean.	The false-positive rate (FPR) is too high, mainly due to excessive down-weighting of high baselines and reliance on too few baseline weeks.
Spatio-Temporal Endemic-	2017	Munich Centre of Health Sciences	• Studies the syndromic outbreaks in 3	The epidemic component with its additive covariates

Epidemic method		(MC-Health), Germany	states, suspected, infected and recovered. • It can be modelled on spatial and temporal resolution. • Address the challenge of monitoring daily counts for potential alarms.	does not provide the complete flexibility to model heterogeneity in infectivity and susceptibility.

Table 2: Summary of Function and Limitations of Surveillance Models studied

The EARS methods were initially designed for a drop-in surveillance system with little or no baseline data available (Fricker, 2010). One of CUSUM's potential disadvantages is that, if measurement continues, CUSUM will lose sensitivity over time because it will take longer to react to small changes in the process average as errors accumulate. However, the CUSUM can be reset to zero periodically to maintain sensitivity (Li, Y. *et al.*, 2016).

The quasi-Poisson algorithm evaluation's main conclusion is that the false positive rate (FPR) is too high. This incident owes to a combination of factors, notably excessive down-weighting of high baselines and reliance on too few baseline weeks

(Noufaily, A. *et al.*, 2013). Several alternatives have been revised to minimise the FPR without sacrificing the POD. An adaptive re-weighting scheme is one of the alternatives that broadly equivalent results to the scaled Anscombe residuals' re-weighting method. The latter approach was recommended as it has a higher threshold in minimising the FPR.

A unique feature of the newly adapted models applying baseline data allows for a better estimation of trend and variance. Besides, the movement should always be fitted even when non-significant or extreme. The discrepancies in the results will decrease when moving from one week to another. This dispersion could vary with the mean. Thorough empirical analyses of HPA data over 20 years suggest that this is not a severe problem (Noufaily, A. *et al.*, 2013). These investigations also support the credibility of the negative binomial model but only in suitable conditions. Therefore, the quasi-Poisson model is preferred from the negative binomial. It should be noted that the overall results from 2011 data displayed only moderate differences between these models. Hence, the quasi-Poisson model is still relevant to other modern epidemic surveillance models.

In the review of the Spatio-Temporal Endemic-Epidemic model, the additive covariate in the epidemic component is not flexible in terms of modelling infectivity and susceptibility. If infectivity or susceptibility depends solely on a single continuous covariate, this situation can be averted by scaling the essential functions with the covariate value. However, this results in a general measure of distance instead of Euclidean space. This factor becomes tedious when the additive combination of several covariates is exponential to the distance function (Höhle, M., 2009). So, the resulting model is no longer linear and would be time-consuming to generate. A log-

linear modification of the space-time distance could help circumvent this issue with further research.

Close cooperation between epidemiologists, microbiologists, computer scientists and statisticians is empirical towards a successful digitalised detection system. Outbreak detection should seek to combine different methods to obtain the best outcome. This system includes computer-based screening procedures incorporated with inter-disciplines to enhance existing strategies and improve quality data collection (Gierl, L., 1998).

5.2 Implementation of Surveillance Data for Models

In this research study, the surveillance data was loaded into the model, using the command in Figure 11, as a critical object. This key object is identified as a time series of stsObj class, which are "structured objects" of an $n \times m$ matrix for surveillance models.

```
stsObj <- read_csv("covid19_global.csv")
```

Figure 11: Importing data as a structured object for R surveillance package

5.2.1 EARS C1 Model

This method is beneficial for data without many historic values, since it only needs counts from the recent past.

The following commands were scripted in R for the EARS C1 model:

```
# Call earsC function and show result
covid19_ears<- earsC(stsObj, control = list(range =
10:70, method="C1"))
plot(covid19_ears, legend.opts=list(horiz=TRUE,
x="topright"))
```

Figure 12: Modelling the EARS C1 function

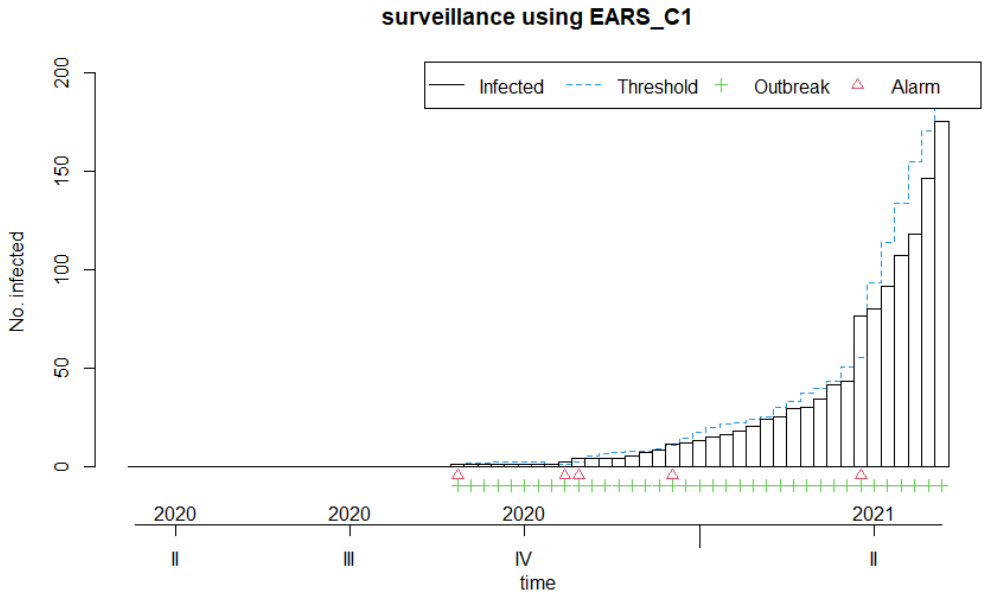


Figure 13: Performance of the EARS C1 model

Given that the EARS method acts as a drop-in surveillance system on studying diseases with minimal historical data, it can be observed from Figure 13 that it outshone its rivals over this pandemic by identifying and setting alarms on the subsequent waves of outbreaks. The model constantly rescaled itself to zero to maintain sensitivity towards the outbreak development, and the statistical warning is only produced at week t with observed count $Y(t)$ if the statistics of the C1 variation exceeds the threshold, which is the baseline count mean added by a multiple of the standard deviation. Therefore, the EARS model would assist in quality control upon reviewing outbreaks from each time interval and identifying the next potential surge of infected cases.

However, resetting back to zero may act as a dual-edged sword as it may result in over-scaling while plotting along with the moving average for each outbreak term. This reasoning can be observed from the low sensitivity rate of 0.1351351 for this model. The underreporting of cases can be followed by comparing the model's predictions with the Farrington model's performance (Figure 15) as the possibilities

are projected to be significantly lower than Farrington's. Since there is no historical data provided before the pandemic, the model could only study this disease outbreak if an outbreak has occurred, so it is justified that the specificity is 1as the alarms were raised at the beginning of this research study.

5.2.2 Farrington's (Quasi-Poisson) Model

The following commands were scripted in R for the Farringtonôs model:

```
#Set control parameters.
control <- list(b=1,w=3,range=56:100,reweight=TRUE,
verbose=FALSE,alpha=0.05)
farr <- algo.farrington(disProgObj =
stsObj,control=control)
#Plot the result.
plot(farr,disease="covid19",method="Farrington")
```

Figure 14: Setting control parameters and modelling the Farrington (Quasi-Poisson) function

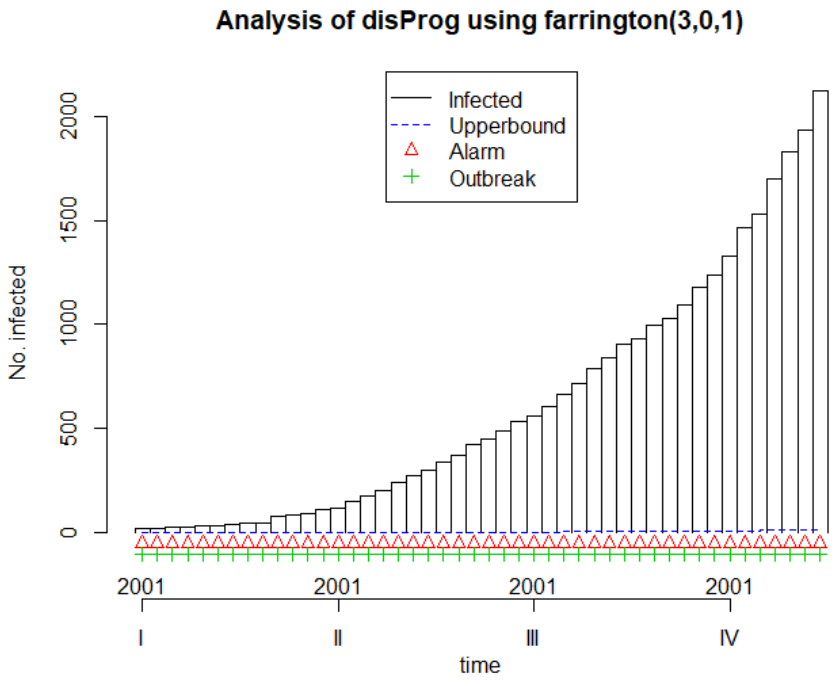


Figure 15: Performance of the Farrington (Quasi-Poisson) model

Farrington's model is a good match for the data since ECDC updated their reports weekly, and this model allows for weekly detection. The Farrington model's quasi-Poisson concept is popular in current epidemic studies due to its simplicity and familiarity with the Poisson model. However, no feature was implemented for the Farrington model to overcome the pandemic by monitoring sudden outbreaks in the surveillance package. There were many false positive alarms, as observed from Figure 15 and was unable to act as quality control for disease monitoring. It can be deduced from past studies that this model is better suited for seasonal disease outbreaks and only serves as a detector for an outbreak. It can control the overdispersion from extreme rises or drops in the data by plotting an exponential increase in infected cases. Since the alarms flared off in the initial stages of the outbreak, the Farrington model could not detect any further attacks, nor could it be studied for its sensitivity, specificity, and accuracy rate. The values displayed for this model were disregarded

as its sensitivity output was 1, signifying that there is no sensitivity for this model, and the specificity output was "NaN" or "Not Available". Moreover, the outbreak period was not correctly identified by the Farrington model as compared to the EARS C1 model. The EARS C1 model was able to locate and display surveillance results from 2020 to 2021, whilst the Farrington model demonstrated 2001 when the same dataset was used throughout this research study.

5.2.3 Spatio-Temporal Models (Spatio-Temporal Endemic-Epidemic Model and Spatio-Temporal SIR Model)

The following commands were scripted in R for the Spatio-Temporal Endemic-Epidemic Model:

```
## convert the original data frame to an "epidata" event
history
myEpi <- as.epidata(head(global, 1000), t0 = 0,
                    tI.col = "tI", tR.col = "tR", id.col
= "Id",
                    coords.cols = c("Long", "Lat"))
# Checking the data structure for Spatio-Temporal
Epidemic model
str(myEpi)
head(as.data.frame(myEpi))  # "epidata" has event history
format
summary(myEpi)              # see 'summary.epidata'
plot(myEpi)                 # see 'plot.epidata' and also
'animate.epidata'
```

Figure 16: Modelling of the Spatio-Temporal Endemic-Epidemic function

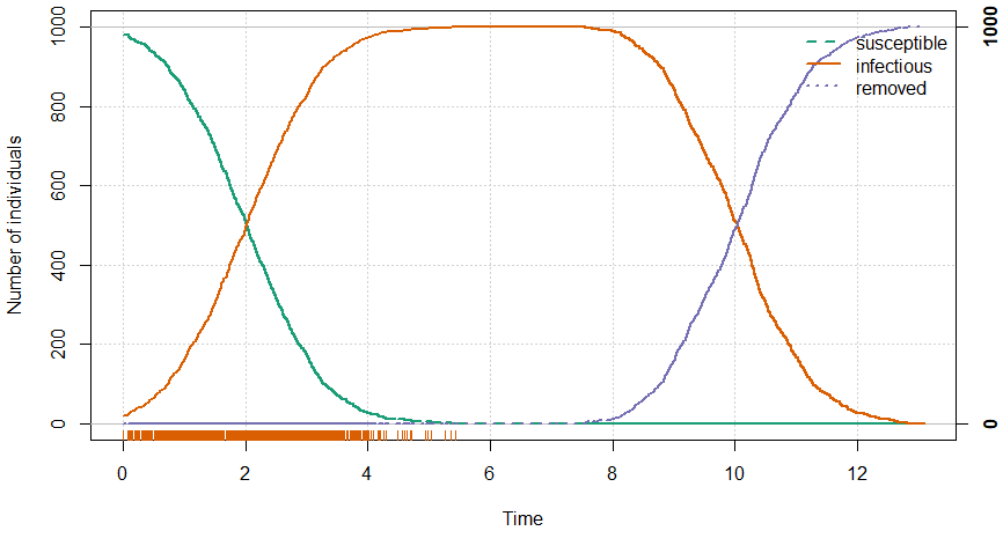


Figure 17: Study of the syndromic outbreaks in the SIR states using the Spatio-Temporal model

From the vast dataset provided by ECDC, 1000 individuals were sampled to create the plot summary using the Spatio-Temporal model below:

```
              time type  id nSusceptible nInfectious
nRemoved
1      0.000000000                   974          26
0
2      0.007452495    I 429          973          27
0
3      0.008473771    I 504          972          28
0
             [....]
1974 13.017109829    R 913            0           1
999
```

```
1975 13.601011493    R 439            0            0
1000
```

Figure 18: Output of ñsummary(myEpi)ò function

Based on Figure 18, a summary amongst the 1000 individuals sampled for this was made. In the early stages of the pandemic, 26 individuals among the 1000 individuals were infected, while the rest of the unreported cases are susceptible to the disease outbreak as the outbreak progresses. The research study ended with all the individuals being infected and then removed in an estimated time width of 14 days (rounded up from 13.601 days). This transition is interpreted in the model that most of the individuals studied are ultimately removed from the study. This interpretation makes sense because the dataset only logged in when there WHO reports on individuals upon contracting the illness and "remove" the individuals from the dataset after they are said to be deceased or have recovered.

Therefore, from Figure 17, it can be observed that the disease outbreak took place at an exponential rate as it becomes highly infectious. The disease outbreak broke off with its potential to infect such high numbers of susceptible individuals. The numbers exponentially decrease with the rapid spread of disease, and these susceptible individuals become infected cases. Eventually, after an infection period where it reaches a plateau around the first quarter of the year 2020, and this status persisted for approximately four months until the number of infected individuals goes on an exponential dive again as they are now "removed". With the aid of the Spatio-Temporal SIR model, we can understand the disease's capability to infect individuals and develop an outbreak in a mere two months. There is an average 4-month period until these cases either survive the illness or succumb to it.

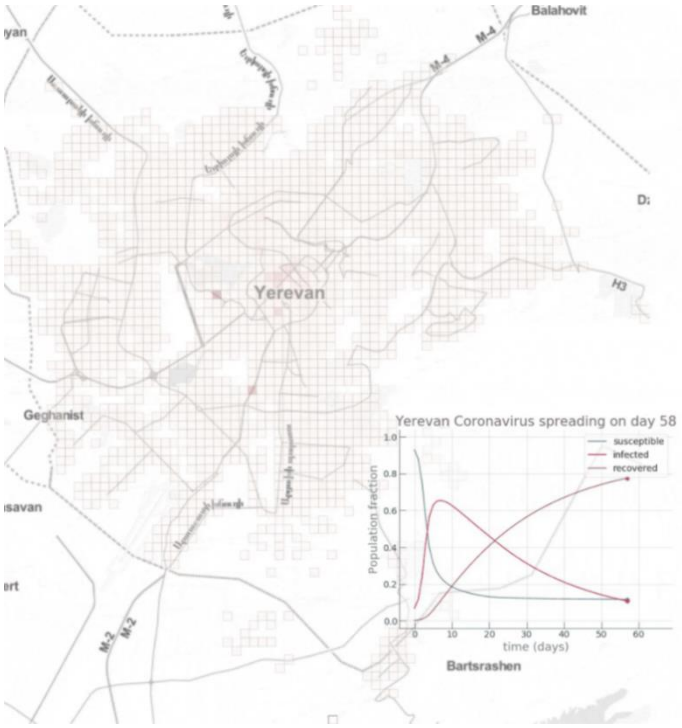


Figure 19: Visualisation of the Spatio Temporal model on Geospatial Modelling for Yerevan

The spread of covid-19 is not solely based on complex data and facts; it should be best studied through human movements to understand where the outbreak originated to devise plans on containing it. Hence, plotting the disease on a geospatial view would allow us to study the disease movement shown in the model adopted by Yeghikyan on urban epidemic modelling (Yeghikyan, 2020). Like the Spatio-Temporal model employed using the surveillance package. It studies the pandemic's development through three "S-I-R" states, Suspected, Infected and Removed. Based on the geospatial model's snapshot attached above, Yeghikyan's model further supported the results shown by the Spatio-Temporal model generated using the surveillance package. There are initially high numbers of individuals susceptible to the disease during the early stages of the outbreak. Slowly but surely, the numbers of susceptible cases exponentially decreased in this rapidly spreading disease as more patients have now become infected, and finally, they are "removed". This model's detailed code is

attached in the appendix, adapted for local surveillance of desired countries, such as Malaysia when population density data is available.

5.3 Challenges of the Proposed Model

The availability of surveillance data is an imperative influence on this research study. It has been discussed that the surveillance data sourced from ECDC has been reported weekly with data extracted from WHO's health reports related to covid-19's latest developments. This reporting system entails that there are several lags between the data reports at hand. The first lag exists where hospitals would have to report to the officials representing their country, translating to WHO's health reports. ECDC would finally consolidate the health reports weekly to publish these publicly available surveillance data. This lengthy and tedious process brings up several arguments over the reporting of surveillance data, such as the possibility of health officials potentially underreporting covid-19 cases from their representative countries. There might be several reasons to do so. One may argue that there might be unreported cases since some citizens could not afford to be clinically tested or receive medical help; There were reports on countries being underprepared financially with tight resources over the pandemic.

Hence, data availability primarily impacts the research study in two aspects, the development of a blockchain and the results from epidemic surveillance models studied. The concept of implementing a blockchain for this research study is theoretically proposed in the research framework to acknowledge real-time reports of surveillance data. However, in reality, building a blockchain for this research study would be unnecessary as there are no real-time reports from hospitals directly logging in chief reports upon diagnosing suspected or infected patients. Thus, this also affects

the study of the Spatio-Temporal Model. The results could only be derived from theoretically assuming the new incoming cases over the week were suspected in the previous period.

Moreover, there could be manipulation of data leading to disease outbreaks. It has been observed from the raw data that there are negative values of the number of cases reported between the following day, and there are occurrences where a large volume of patients was introduced back into the report. This manipulation could be caused by ECDC's attempt to correct over reported cases by subtracting the over-reported issues on the following day. The country faced a delay in reporting cases to WHO, which is consecutively reflected upon ECDC's surveillance data log. Therefore, the epidemic surveillance model's output would be inaccurate, with human error involved, as there are sudden spikes from the high report of cases and overdispersion of data.

The geospatial visualisation of covid-19's disease spread could not be privately applied to this research study as there was no readily available population density data provided. Yeghikyan noted that population density data are rarely available in the open-sourced environment, and he has suggested that this data could be provided to companies paying a reasonable sum. Likewise, this research is conducted through open-sourced resources with minimal costs as a research student. In such wise, a demonstration of the geospatial analysis was shown through Yeghikyan's model. He could hypothetically fit Yerevan's taxi movements as the country's population density to study covid-19's disease spread.

Categorically, the concept of the epidemic surveillance models discussed in the literature review is adapted into recent studies based on the pandemic. However, the results derived from the surveillance package models were not ideal as they were not

redesigned for pandemic events. Aside from the EARS C1 model, the Farrington model was only best fitted for seasonal disease outbreaks with historical data to understand disease outbreaks' progression. In some measure, the Farrington model could handle the overdispersion of data due to the occasional underreporting and delayed reporting of surveillance data and was selected for this study to compare its disease modelling and prediction performance against the EARS C1 model.

CHAPTER 6

CONCLUSION

Overall, the EARS model is a robust surveillance system that best served its purpose in epidemic surveilling the covid-19 outbreak without any historical data provided and monitored the disease outbreak's progression. It could help healthcare professionals prepare for rapid incoming numbers of infected cases by sounding alarms on the next more significant epidemics wave. Contrariwise, the model quickly lost its sensitivity and projected the number of infected cases with a higher threshold. Given that the data is collected from inconsistent weekly reports, there exists a higher variance in the dataset. On account of this, the upper point of infected cases provides no information by itself, so Farrington's model complements this shortcoming.

The globally published dataset is vulnerable to inconsistent reports, and it can be seen during data pre-processing that there is significant missing data from January 2020 to February 2020. Farrington's model overcame this issue by backtracking and projecting the underreported cases to provide a complete overview of covid-19's disease development. On the grounds of this, Farrington's model can be incorporated into the EARS model to handle overly dispersed and inconsistent data. In contrast, the EARS model is best principled in signalling the next wave of disease outbreaks over a pandemic.

The Spatio-Temporal model could best demonstrate its capabilities in the Spatio-Temporal Endemic-Epidemic variant. In this variant, the model exemplified the development of covid-19 outbreak by filling in on the global infection rate, incubation period and termination period. This crucial information gives us an insider view of a pandemic, given its full potential with the global dataset. Suppose the disease

is not contained in a population; It infects a population in 2 months. The minimum period needed to take full effect and terminate this population is six months. This disease spread and rapid development are exceptionally lethal to humanity if uncontained.

Open source of data would lead to more unrealised potentials from this research study. Sadly, this cannot be achieved currently in the real-world setting as there are complications from open sharing of data. Thus, civilians are not able to identify potential disease outbreaks with this significant limitation.

It should be noted that this research study employs data with confirmed cases of patients infected with covid-19. The model's predictions using solely chief complaints may skew the model's predictions. This instance can be taken note of through the case study of Dr Li WenLiang, who first reported on a new strain of virus affecting his patient, but his report was diffused until the outbreak occurred. Moreover, there are several symptoms for covid-19 that could be easily confused with other diseases. These symptoms include fever, sore throat or runny nose. There may still be underreported cases as asymptomatic patients do not exhibit specific symptoms logged for this model's detection.

The EARS model's feature on readjusting to maintain its sensitivity may have caused the shift in its moving average for the model. There are many extreme numbers of cases occurring between syndromic periods as this disease is spreading rapidly. Therefore, the EARS model's sensitivity is relatively low without historical data involvement to help the model develop a steady moving average by studying the disease spread before its outbreak.

However, resetting back to zero may act as a dual-edged sword as it may result in over-scaling while plotting along with the moving average for each outbreak term.

This deduction can be observed from the low sensitivity rate of 0.1351351 for this model. The underreporting of cases can be followed by comparing the model's predictions with the Farrington model's performance (Figure 15) as the possibilities are projected to be significantly lower than Farrington's. Since there is no historical data provided before the pandemic, the model could only study this disease outbreak if an outbreak has occurred, so it is justified that the specificity is 1 as the alarms were raised at the beginning of this research study.

Farrington's model could not adapt to monitoring disease outbreaks like covid-19 as it was unable to provide any features for quality control after an attack has occurred. This shortcoming is also crucial to note that disease information was only released after it was identified and an outbreak has already taken place. The Farrington model could not act as a surveillance model without the supply of historical data. It was efficient in providing a more accurate detail of the disease monitoring as it took into account the overdispersion of the rapidly changing number of infected cases.

This pandemic has exposed the world's digital vulnerability towards preparing for or even preventing a disease outbreak. This research study could be better improved if there were real-time updates of the data readily and openly available. There is also potential masked data reporting, resulting in the underreported number of cases or sudden spike cases from delay reporting. Besides, the geospatial mapping of covid-19's disease spread could not be easily achieved. For example, this research study could not focus on Malaysia not having an openly sourced geographical population mapping to study the population's rate transitioning from susceptible to infected and removed. Therefore, an adapted model solely for Malaysia could not be constructed as necessary information must be obtained from sources up to the Ministry level. The research with a geospatial plot of the Yerevan citizen's movement conducted by

Yeghikyan was based upon an approximation from the transportation data provided by the frequency of taxi drivers driving in Yerevan, and the frequencies are multiplied to match the estimated population.

Essentially, amongst the three standard epidemic surveillance models studied, the EARS model C1 variation stood out for its capability to adjust to the pandemic setting immediately. There were consistent alarms raised throughout the surveillance period from January 22, 2020, to February 4, 2021, as surveillance data was publicly shared after covid-19's rampant disease outbreak. However, the EARS model could help predict the time interval between the next disease outbreak wave that out-scaled the previous wave. This is an advantageous property to allow healthcare professionals to prepare and allocate resources for the next outbreak.

Even though the Farrington model consistently raised alarms upon encountering surveillance data amidst the pandemic, this model should not be overlooked. The Farrington model incorporating the Quasi-Poisson algorithm helped control the overdispersion of data and model predictions for the disease outbreak realistically. The EARS model has the most significant disadvantage of losing sensitivity over alarm intervals, as observed from its output that under-predicted the number of infected cases and attempted to circumvent this by setting a seemingly significant threshold. There were also erratic stochastic increases and decreases of infected points observed from ECDC's raw data analysis. The drastic changes in the number of instances influence the EARS model prediction to be not as specific or closely accurate to the actual value.

Perhaps the epidemic surveillance model would work soundly by incorporating the EARS model's drop-in surveillance and utilise Farrington's model to handle overly dispersed data when buffering between alarm interval property to predict disease

outbreaks. Meanwhile, the Spatio-Temporal Endemic-Epidemic model works well by monitoring the rate of disease spread. It also served its purpose in studying the infection rate where an individual becomes infected and then recovered/removed from the cycle. Therefore, healthcare professionals can inspect the covid-19's potency through a wide-eye view to quickly learn how this disease affects an individual.

REFERENCES

Farrington C, Andrews N, Beale A, Catchpole M (1996) A statistical algorithm for the early detection of outbreaks of infectious disease. J R Stat Soc Ser A 159:547ï 563

Fricker, R. D. (2010). Introduction to statistical methods for biosurveillance: With an emphasis on syndromic surveillance. In Introduction to Statistical Methods for Biosurveillance: With an Emphasis on Syndromic Surveillance. Cambridge University Press. https://doi.org/10.1017/CBO9781139047906

Gierl, L. (1998). Geomed 97 proceedings of the International Workshop on Geomedical Systems, Rostock, Germany, September 1997. Stuttgart: Teubner.

Hagen, K. S., Fricker, R. D., Hanni, K., Barnes, S., & Michie, K. (2011). Assessing the Early Aberration Reporting System's Ability to Locally Detect the 2009 Influenza Pandemic.

Held, L., Höhle, M. and Hofmann, M. (2005) A statistical framework for the analysis of multivariate infectious disease surveillance counts. Statistical Modelling, Vol. 5(3), pp. 187-199

Henning, K. J. (2004) Overview of Syndromic Surveillance What is Syndromic Surveillance? Retrieved 17, 2020, from https://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a3.htm

Höhle, M. (2009) Additive-multiplicative regression models for spatio-temporal epidemics. Biometrical Journal, Vol. 51(6), pp. 961-978

Höhle, M. (2010). Online Change-Point Detection in Categorical Time Series In: Statistical Modelling and Regression Structures (Kneib, T. & Tutz, G., eds.) Physica-Verlag HD, pp. 377-397

Khameneh, J., & Nastaran. (2014, December 1). Machine Learning for Disease Outbreak Detection Using Probabilistic Models. Retrieved from https://publications.polymtl.ca/1659/

Li, Y., Whelan, M., Hobbs, L., Fan, W. Q., Fung, C., Wong, K., é Johnson, I. (2016). Data-driven approach of CUSUM algorithm in temporal aberrant event detection using interactive web applications. Canadian Journal of Public Health, 107(1). doi: 10.17269/cjph.107.5228

Lombardo, J. S., & Buckeridge, D. L. (2007). Disease surveillance: a public health informatics approach. Hoboken, NJ: Wiley-Interscience.
M. M. Wagner et G. Wallstrom, "Methods for algorithm evaluation," Handbook of biosurveillance, pp. 301ï 310, 2006.

Meyer, S., Held, L., & Höhle, M. (2017). Spatio-temporal analysis of epidemic phenomena using the R package surveillance. Journal of Statistical Software, 77(1), 1ï 55. https://doi.org/10.18637/jss.v077.i11

Noufaily, A., Enki, D., Farrington, P., Garthwaite, P., Andrews, N., & Charlett, A. (2013). An Improved Algorithm for Outbreak Detection in Multiple

Surveillance Systems. Online Journal of Public Health Informatics, 5(1). doi: 10.5210/ojphi.v5i1.4497

Paul, M., Held, L. and Toschke, A. M. (2008) Multivariate modelling of infectious disease surveillance data. Statistics in Medicine, Vol. 27(29), pp. 6250-6267

Salmon, M., Schumacher, D., Stark, K. and Höhle, M. (2015) Bayesian outbreak detection in the presence of reporting delays. Biometrical Journal, Vol. 57(6), pp. 1051ð 1067

Yeghikyan, G. (2020). Modelling the coronavirus epidemic in a city with Python | by Gevorg Yeghikyan | Towards Data Science. Towards Data Science. https://towardsdatascience.com/modelling-the-coronavirus-epidemic-spreading-in-a-city-with-python-babd14d82fa2

APPENDICES

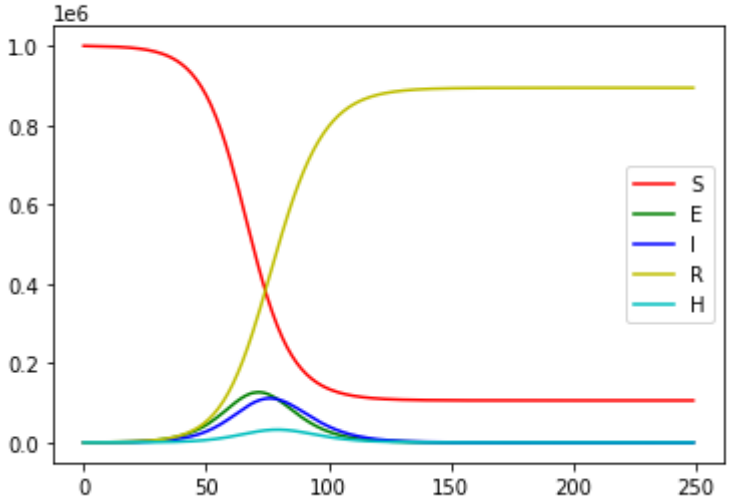
Creating a geospatial analysis of the covid-19 disease spread by simulating the epidemic in Yerevan as referenced from Yeghikyanôs Spatio-Temporal SIR geospatial epidemic surveillance model.

```
#import the basic libraries
import numpy as np
import pickle
import matplotlib.pyplot as plt
# define plot function
def seir_plot(res):
    plt.plot(res[::12, 0], color='r', label='S')
    plt.plot(res[::12, 1], color='g', label='E')
    plt.plot(res[::12, 2], color='b', label='I')
    plt.plot(res[::12, 3], color='y', label='R')
    plt.plot(res[::12, 4], color='c', label='H')
plt.legend()
# load OD matrices
pkl_file = open('Materials/Yerevan_OD_matrices.pkl',
'rb') # change to your desired directory
OD_matrices = pickle.load(pkl_file)
pkl_file.close()


np.set_printoptions(suppress=True, precision=3)


# load population densities
pkl_file = open('Materials/Yerevan_population.pkl', 'rb')
```

```
pop = pickle.load(pkl_file)
pkl_file.close()


pop[13] == pop[1]
# set up model
%run virus-sim.py


r = OD_matrices.shape[0]
n = pop.shape[1]
N = 1000000.0


initialInd = [334, 353, 196, 445, 162, 297]
initial = np.zeros(n)
initial[initialInd] = 50


model = Param(R0=2.4, DE= 5.6 * 12, DI= 5.2 * 12,
I0=initial, HospitalisationRate=0.1, HospitalIters=15*12)
# run simulation
%run virus-sim.py


alpha = np.ones(OD_matrices.shape)
iterations = 3000
res = {}
inf = 50
res['baseline'] = seir(model, pop, OD_matrices, alpha,
iterations, inf)
```

```
print(
"Max number of hospitalised people: ",
int(res["baseline"][0][:,4].max()),
"\n",
"Day with max hospitalised people: ",
int(res["baseline"][0][:,4].argmax()/12)
)
# plot result
seir_plot(res["baseline"][0])
```

```
Max number of hospitalised people:  32416
 Day with max hospitalised people:  79
```



Constructing the geospatial visualisation model,

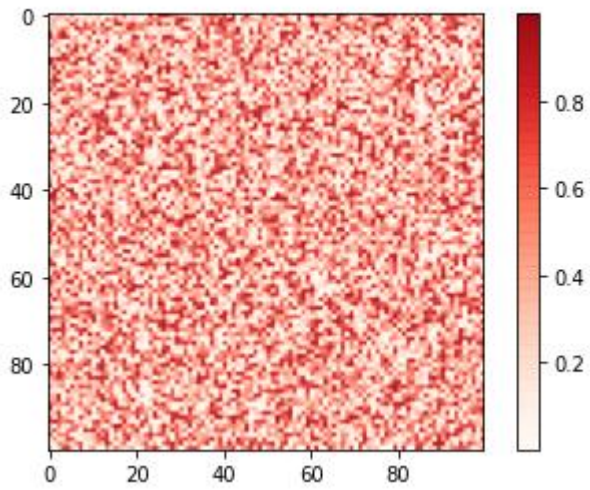
```
# import libraries
import geopandas as gpd, pandas as pd
import contextily as ctx


from pyproj import CRS
crs = CRS.from_epsg(4326)
```

```
# load Yerevan grid file
yerevan_gdf = gpd.read_file("Yerevan grid
shapefile/Yerevan.shp")
#yerevan_gdf.crs = {'init':'epsg:4326'}
yerevan_gdf.crs = crs

# convert to crs used by contextily
yerevan_gdf_3857 = yerevan_gdf.to_crs(epsg=3857)
west, south, east, north =
yerevan_gdf_3857.unary_union.bounds

# declare baseline array storing the dynamics of the
compartments
baseline = res['baseline'][1][::12, :, :]

# declare hopsitalisation array storing the dynamics of
the hospitalised
hosp = res['baseline'][0][::12, 4]

# find maximum hospitalisation value to make sure the
color intensities in the animation are anchored against
it
max_exp_ind = np.where(baseline[:, 1, :] == baseline[:,
1, :].max())[0].item()
max_exp_val = baseline[:, 1, :].max()
```

```
ncolors = 256
# get cmap
color_array = plt.get_cmap('Reds')(range(ncolors))

# change alpha values
color_array[:, -1] = np.linspace(0.3, 1, ncolors)

# create colormap object
import matplotlib.colors as colors
from matplotlib.colors import LinearSegmentedColormap

map_object =
LinearSegmentedColormap.from_list(name="Reds_transp",
colors=color_array)

# register the colormap object
plt.register_cmap(cmap=map_object)

# plot some example data
fig, ax = plt.subplots()
h = ax.imshow(np.random.rand(100,100),
cmap='Reds_transp')
plt.colorbar(mappable=h)

def trunc_colormap(cmap, minval=0.0, maxval=1.0, n=100):
```

```
    new_cmap =
LinearSegmentedColormap.from_list('trunc({n}, {a:.2f},
{b:.2f})'.format(n=cmap.name, a=minval, b=maxval),


cmap(np.linspace(minval, maxval, n)))
    return new_cmap


cmap = plt.get_cmap('Reds_transp')
new_cmap = trunc_colormap(cmap, 0.0, .9)


# plot some example data
fig, ax = plt.subplots()
h = ax.imshow(np.random.rand(100,100), cmap=new_cmap)
plt.colorbar(mappable=h)
```

```
<matplotlib.colorbar.Colorbar at 0x29b81361d08>
```




```
print("baseline dimensions: ", baseline.shape)
baseline dimensions:  (250, 5, 549)
```

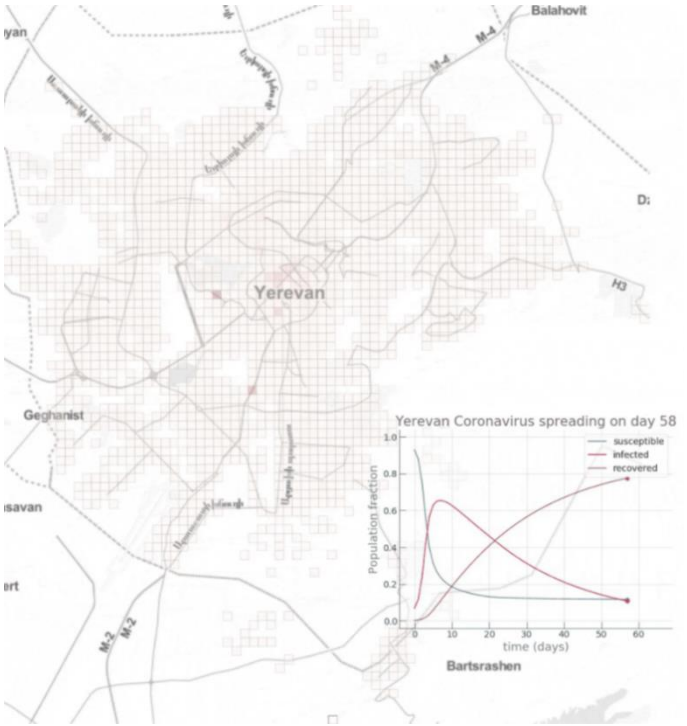
```
print("hosp dimensions: ", hosp.shape)
hosp dimensions:  (250,)


params = {"axes.labelcolor":"slategrey"}
plt.rcParams.update(params)
cmap = plt.cm.get_cmap("Blues")
blue = cmap(200)


from tqdm import tqdm_notebook


for time_step in tqdm_notebook(range(1,251)):


    yerevan_gdf_3857['exposed'] = baseline[time_step-1,
1, :]


    #plot
    fig, ax = plt.subplots(figsize=(14,14), dpi=72)
    yerevan_gdf_3857.loc[yerevan_gdf_3857.index==84,
'exposed'] = max_exp_val + 1
    yerevan_gdf_3857.plot(ax=ax, facecolor='none',
edgecolor='gray', alpha=0.5, linewidth=0.5, zorder=2)
    yerevan_gdf_3857.plot(ax=ax, column='exposed',
cmap=new_cmap, zorder=3)
    # add background
    ctx.add_basemap(ax, attribution="",
url=ctx.sources.ST_TONER_LITE, zoom='auto', alpha=0.6)
```

```
    ax.set_xlim(west, east)
    ax.set_ylim(south, north)
    ax.axis('off')
    plt.tight_layout()

    inset_ax = fig.add_axes([0.6, 0.14, 0.37, 0.27])
    inset_ax.patch.set_alpha(0.5)

    inset_ax.plot(baseline[:time_step, 0].sum(axis=1),
label="susceptible", color=blue, ls='-', lw=1.5,
alpha=0.8)
    inset_ax.plot(baseline[:time_step, 1].sum(axis=1),
label="exposed", color='g', ls='-', lw=1.5, alpha=0.8)
    inset_ax.plot(baseline[:time_step, 2].sum(axis=1),
label="infectious", color='r', ls='-', lw=1.5, alpha=0.8)
    inset_ax.plot(baseline[:time_step, 3].sum(axis=1),
label="recovered", color='y', ls='-', lw=1.5, alpha=0.8)
    inset_ax.plot(hosp[:time_step], label="hospitalised",
color='purple', ls='-', lw=1.5, alpha=0.8)

    inset_ax.scatter((time_step-1), baseline[(time_step-
1), 0].sum(), color=blue, s=50, alpha=0.2)
    inset_ax.scatter((time_step-1), baseline[(time_step-
1), 1].sum(), color='g', s=50, alpha=0.2)
```

```
    inset_ax.scatter((time_step-1), baseline[(time_step-
1), 2].sum(), color='r', s=50, alpha=0.2)
    inset_ax.scatter((time_step-1), baseline[(time_step-
1), 3].sum(), color='y', s=50, alpha=0.2)
    inset_ax.scatter((time_step-1), hosp[(time_step-1)],
color='purple', s=50, alpha=0.2)

    inset_ax.scatter((time_step-1), baseline[(time_step-
1), 0].sum(), color=blue, s=20, alpha=0.8)
    inset_ax.scatter((time_step-1), baseline[(time_step-
1), 1].sum(), color='g', s=20, alpha=0.8)
    inset_ax.scatter((time_step-1), baseline[(time_step-
1), 2].sum(), color='r', s=20, alpha=0.8)
    inset_ax.scatter((time_step-1), baseline[(time_step-
1), 3].sum(), color='y', s=20, alpha=0.8)
    inset_ax.scatter((time_step-1), hosp[(time_step-1)],
color='purple', s=20, alpha=0.8)

    inset_ax.fill_between(np.arange(0, time_step),
np.maximum(baseline[:time_step, 0].sum(axis=1), \

baseline[:time_step, 3].sum(axis=1)), alpha=0.035,
color='r')
    inset_ax.plot([time_step, time_step], [0,
max(baseline[(time_step-1), 0].sum(), \
```

```
baseline[(time_step-1), 3].sum())], ls='--', lw=0.7,
alpha=0.8, color='r')

    inset_ax.set_ylabel('Population', size=18, alpha=1,
rotation=90)
    inset_ax.set_xlabel('Days', size=18, alpha=1)
    inset_ax.yaxis.set_label_coords(-0.15, 0.55)
    inset_ax.tick_params(direction='in', size=10)
    inset_ax.set_xlim(-4, 254)
    inset_ax.set_ylim(-24000, 1024000)
    plt.xticks(fontsize=14)
    plt.yticks(fontsize=14)
    inset_ax.grid(alpha=0.4)

    inset_ax.spines['right'].set_visible(False)
    inset_ax.spines['top'].set_visible(False)

    inset_ax.spines['left'].set_color('darkslategrey')
    inset_ax.spines['bottom'].set_color('darkslategrey')
    inset_ax.tick_params(axis='x',
colors='darkslategrey')
    inset_ax.tick_params(axis='y',
colors='darkslategrey')
    plt.legend(prop={'size':14, 'weight':'light'},
framealpha=0.5)
```

```
    plt.title("Yerevan Covid-19 spreading on day:
{}".format(time_step), fontsize=18, color= 'dimgray')
    #plt.savefig("Plots/flows_{}.jpg".format(time_step),
dpi=fig.dpi)
    plt.show()
```



Appendix 1: Construction of geospatial visualisation of the Spatio-Temporal SIR Model

TENTATIVE PLAN AND GANTT CHART

Project I	
October 26 2020	Review epidemic surveillance models proposed in literature review to code the algorithm component
November 15 2020	Organise the data to be input into the algorithms for research study
December 1 2020	Collect quantitative data and evaluate data for its specificity, sensitivity and false alarm rates.
December 10 2020	Review data analysis and write down findings.
December 30 2020	Presentation of overall progress
Project II	
January 2 2021	Review the research project and update on data findings
February 15 2021	Rewrite the research project for better reporting flow of research analysis
April 30 2021	Finalisation of research project