

**REPORT**

**FINDING**

**STEPHEN VU**



**Table of Contents**

[**Abstract** 3](#_Toc51713271)

[**Problem Introduction** 3](#_Toc51713272)

[**An Thorough Explanation of Mathematical Algorithms in this project** 4](#_Toc51713273)

[Hidden Markov Modelling (HMM) 4](#_Toc51713274)

[INAR 5](#_Toc51713275)

[Viterbi 5](#_Toc51713276)

[**Approaches from popular literature and explanation of methodology selection** 5](#_Toc51713277)

[Why HMM? 6](#_Toc51713278)

[**Appendix** 7](#_Toc51713279)

# **Abstract**

As conventional practice of deciding the severity of a pandemic, experts normally take 2 aspects into consideration:

* Rate of Spread (ROS)
* Fatality rate of the pandemic

ROS will be discussed further with some peer-reviewed findings. To determine the fatality rate, we need to divide number of the dead from virus over the infected population. (And we assume that the deadh toll is fully reported by authorities). To ensure the correctness of statistics imputations, the population of the infected must be carefully rectified as it is classified into 3 categories:

* People get tested and publicly reported/recorded
* People in incubation stage
* **People have second attack from virus**
* **People get tested but receive wrong results**

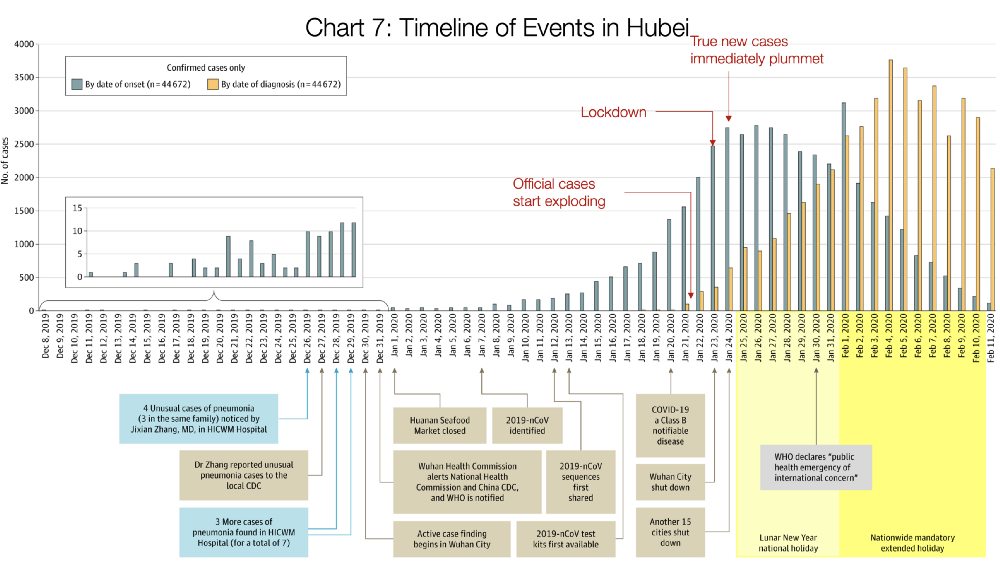
**The paper will use data from WHO on Covid outbreak in Australia from the start to 31st May**. *(Due to the need of quick execution from my computer)*

# **Problem Introduction**

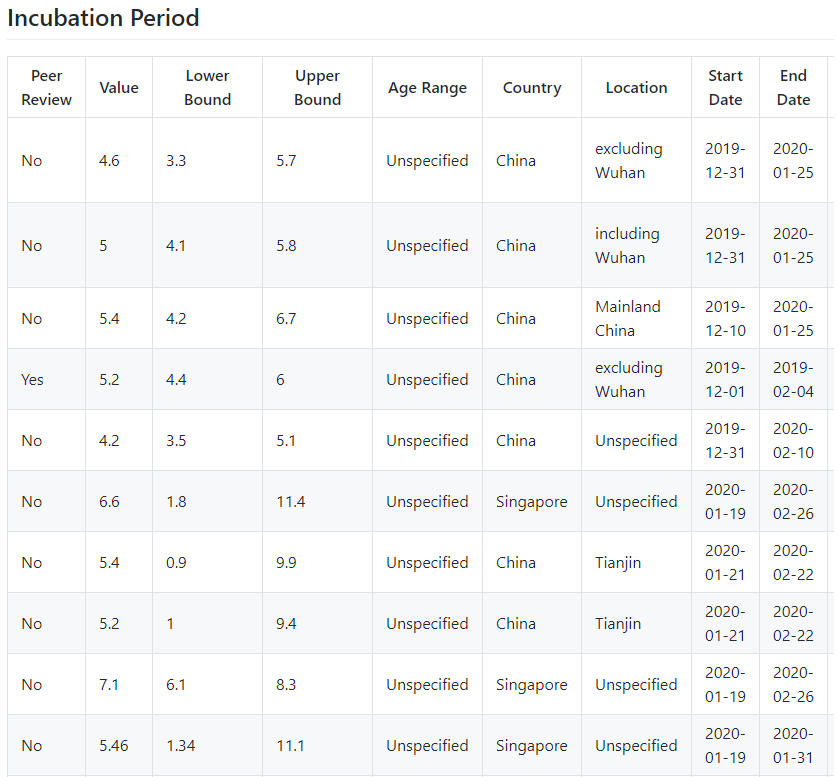
The under-reporting problem in current available dataset is confirmed by Science Magazine (page 489):

“…The researchers combined data from Tencent, one of the world's largest social media and technology companies, with a networked dynamic metapopulation model and Bayesian inference to analyze early spread within China. They estimate that ∼86% of cases were undocumented before travel restrictions were put in place. Before travel restriction and personal isolation were implemented, the transmission rate of undocumented infections was a little more than half that of the known cases. However, because of their greater numbers, undocumented infections were the source for ∼80% of the documented cases. Immediately after travel restrictions were imposed, ∼65% of cases were documented…”

To further understand the Australian problems as to later explain why this project choose Hidden Markov Model as the main method, we need to trace the root of this issue – HUBEI (a province of Wuhan) – You can see bigger size picture in appendix:



As witnessed from the chart, before the first case of 2019 nCoV is identified in Jan 7th 2020, it is believed people contracting the virus had been infected 11 days before that. Hence, the complexity of delayed symptoms perplexes the appropriate acts of resolution.



The average incubation period is 5 days. This is a good funding that can verify the correctness of incubating parameters later.

Within limited data resources, we must assume that all tests produce precise results with marginal rate of misconduct. (Because in Indian, most of test kits imported from China provided a low accuracy of diagnosis – according to Ander at. el, 2020).

# **An Thorough Explanation of Mathematical Algorithms in this project**

*#Note: This section only serves my personal need of future review of the paper. Thus, a well-documented clarification is required to fully embrace such advanced statistical concepts. Since this will be written in my own words, you can skip the section.*

## Hidden Markov Modelling (HMM)

***HMMs can be applied in many fields where the goal is to recover a data sequence that is not immediately observable (but other data that depend on the sequence are) (Wiki, 2020).***

In the purpose of the project, it will help to identify both reported cases and non-reported cases. For this section, to generalize my understanding of the concept, I will use examples of real-life objects for better illustration.

For instance, a day can be classified by only 3 main types of weathers: SUNNY, RAINY, SNOWY. For fundamental Markov chain, we are predicting tomorrow weather based on today’s weather.

Assume that we have a set of data based on empirical statistics:

|  |  |  |  |
| --- | --- | --- | --- |
|  | SUNNY | RAINY | SNOWY |
| SUNNY | 80% | 15% | 5% |
| RAINY | 38% | 60% | 2% |
| SNOWY | 75% | 5% | 20% |

Transition Probabilities of a state in a time series depending on the current state of an object is called First Markovian Order.

The set of states is named S.

S = {S1, S2, … Sn}

Transition probabilities of States is measured:

aij = P(qt+1=Si | qt=Sj)

Explain: Given that day t has the weather Sj, the aij calculates the possibility of day t + I has weather Si.

For example: from the above table, if today is sunny then 80% of sunny tomorrow, hence: asunny|sunny = 0.8

In this case we have 3 types of weather meaning n = 3.

* S = { S1, S2, S3 }

Π is the distribution of first day posiblilities. And π is measured:

Π = P[qi = Sj]

Suppose that first day has 70% sunny, 25% rainy, 5% snowy: π = (0.7 0.25 0.5)

The question: What is the probability of this sequence happening?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mon | Tue | Wed | Thu | Fri | Sat |
| Sunny | Rainy | Rainy | Rainy | Snowy | Snowy |

**P= 0.7x0.15x0.6x0.6x0.02x0.2 = 0.0001512**

As you see, only **0.015%** the above sequence can happen

With Q = {q1, q2,… qT} is the sequence of states.

In real life, some states might not be observable, we can make prediction of hidden observations based on its emitted behaviours. Emitted objects are events arised from the states. For example, if it rains, people might wear shorts, coats or umbrellas. Each of this is called an emitted object. And it must be completely agreed that emitted objects are clearly observable. Assume that we have this statistics:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Short | Coat | Umbrella |
| SUNNY | 60% | 30% | 10% |
| RAINY | 5% | 30% | 65% |
| SNOWY | 0% | 50% | 50% |

Within any states (sunny, rainy, snowy), Markov Process will generate/emit one or more visible events and it will be calculated as follows:

bj(k)=P(Ot = k | qt = Si)

The probability of k happening in day t when the state is I (Si)

To sum up, HMM comprises of three main inputs: Hidden sets, Observed Sets and Initial Sets:

**λ = (A,B,π) as:**

Martrix A = aij;

Martrix B=bj(k); Note: Emission probalitities is determined by ***multimomial distributions, Gaussian distributions, mix-ed Gaussian distributions.***

Martrix Π = P[qi = Sj];

HMMs is a perfect tool for performing supervised learning process. Hence, there are 3 problems HMMs can be implemented:

* Given the model parameters and observed data, estimate the optimal sequence of hidden states.
  + Find **λ that P(Ot | λ)max . People normally use Expectation – Maximization algorithm**
* Given the model parameters and observed data, calculate the likelihood of the data.
  + Given Oi => Find P(Q)
  + I tried backward and forward algorithm, does not appeal to be convincing on my dataset
* Given just the observed data, estimate the model parameters.
  + With this we can use Viterbi (which I will use in the project), Naïve Solution and Efficient Solution – This also used Viterbi.
  + Naïve Solution is good on big data. (long explanation so I attached my hand writing clarification in Apendix) . This is a very expensive model due to the large calculation samples. **It can work but my computer is too weak so I could not properly run this model.**
  + Efficient Solution (also in appendix): Using Recursion algorithm to extract all possible emission.

## INAR

[concerning to put in because it mainly for stationary random variables (like HIV, HPV or sth like that). For non-stationary, I am looking for a better estimating model.

## Viterbi

**The goal of Viterbi is to find the best Sequence of State Q ={q1,q2,…qT} corresponding with observable sets x1,x2,…xT that make LT maximal:**

As:



And with t = 1,2…T:



We can see the Viterbi can meet the recursion process:



Hence, the best sequence states Q is identified by:



# **Approaches from popular literature and explanation of methodology selection**

First there are 2 types of under-reporting that we need to take for consideration:

+ Death tolls under reporting – The number of fatality case in some countries.

+ Testing under reporting

Approaches: Influenza or Pneumonia

* Estimate the novelty of corona outbreak based on the comparison between the number of Severe Acute Respiratory Infection in 2020 and that of the period from 2016-2019 (Brazilian)
* It means that they assume people contracted with normal respiratory conditions as covid infected potential. They are detecting the anomolies in the novelty dataset to identify the possibility of under-reporting coronavirus data.
* We can apply this method for not only SARI but also for Pneumonia & Influenza. By having more data from other diseases, it helps to increase the confidence of our model
* According to WHO, from week 3 to week 36, there have been 194 cases suspected to be contracted with influenza, however, all patients are proven with negative testing results.
* Extrapolate expected deaths and hospitalizations with respect to observations in countries that passed the exponential growth curve
  + This is an unattainable approach since facts indicate that most death cases are correlated with pre-existed medical syndromes. Therefore, it will be even more difficult to cluster the ones died with covid infection
* Some use integer-valued autogressive hidden Markov model to identify the hidden process of recording covid cases in the world, then they applied for Viterbi estimation to predict the exact infection numbers in a specific location . For the moment I am trying to understand the math behind this algorithm to make it applicable for our research purpose.
* We have the spread rate of some nations, taking that as a model to compare with past viral pandemic to see our confidence
* Another popular methodology I found is that some applied susceptible-exposed-infected-recovered model with assuming the incubation period of 6 days to describe the dynamics of disease spread
  + However, their model did not take into account of covid importation which was our main target for this scientific research.
  + We assume that the time-rate of change of  **S(t)**,  the *number* of susceptibles,1 depends on the number already susceptible, the number of individuals already infected, and the amount of contact between susceptibles and infecteds. In particular, suppose that each infected individual has a fixed number  **b**  of contacts per day that are sufficient to spread the disease. Not all these contacts are with susceptible individuals. If we assume a homogeneous mixing of the population, the *fraction* of these contacts that are with susceptibles is  **s(t)**.  Thus, on average, each infected individual generates  **b s(t)**  new infected individuals per day. [With a large susceptible population and a relatively small infected population, we can ignore tricky counting situations such as a single susceptible encountering more than one infected in a given day.

=> My backup plan if HMM fails

Limitations:

* Most of official medical datasets from government’s health webpage are outdated (latest data I can find was in 2017 in New Zealand)
* Data in Taiwan is integrated with Chinese intelligence records, CCP hardly provide the authenticated data for the public audience

## Why HMM?

Hidden Markov Model is Hidden Markov Model has great promise in the surveillance of COVID-19 pandemic. This is because certain transitions happening in the progression of virus in the larger society have an association with time series data on confirmed/active cases, recoveries and deaths reported from Hospitals.

The project will classify states into: Healthy, infected, symptomatic, second attack. These states are invisible but from the dataset, it can be inferred that these states emit following events:

O = {confirmed,deaths, tests,positives,recovered,hosp,icu,vent}

[*need more time to make statistics for this one, I am quite confusing with choosing proper distribution methods on this data*]

**What’s bugging me is that should I use global statistics for the initial state distributions or I need to compute my own? (Appendix covers some of states I got from global data- and it is peer review)**

Following is my intention to perform my work:

* HMM can be used for estimating the whole underlying probabilities of under-reporting.
* From that we can either:
  + Trace back the full data in the early stage of pandemic
  + Predict when we the country reaches the stage of Healthy thereby determining the date of opening the border.
* Then ILAR(2) (Integer latent autogressive) will be applied to list out the possible counts of covids. This is a good algorithm that complements the uncertainty of HMM. The output will be in specific numbers. This one has not been done in precedented literature, and I still dig in the details in it to make the best estimation of the parameters. The reason I choose ILAR instead of INAR is because INAR mainly for stationary diseases, while ILAR can be used on a more advanced population.
* All data I gathered is from here: (with assumption that this is correctly collected)

<https://github.com/M3IT/COVID-19_Data>

For the dataset, we will divide the total counts of cases Zn in day n into Xn and Yn:

As Xn is underlying series for unobserved data (target variables) and Yn is observed series (dependent variables). We can have and ILAR structure:

*Xn* = *𝛼*◦*Xn*−1 + *Wn,*

# C:\Users\n10648771\OneDrive - Queensland University of Technology\DataScienceCovid\1_r-ddYhoUtP_se6x-NOEinA.png**Appendix**

