

Mathematical Modelling of Epidemiology

Shreya Nagunuri¹ and William Tavernetti[#]

¹Granite Bay High School, Granite Bay, CA, USA

[#]Advisor

ABSTRACT

A mathematical model is a method to represent real world problems using mathematical terms like equations and formulas. In the medical sphere, these models are used to prevent, predict, or avoid these diseases. Devastating epidemics are inevitable due to the increase of global travel and urbanization which both contribute increasing points of contact for diseases to spread. In order to combat the spread of infectious diseases, scientists from a combination of fields have started to develop mathematical models to represent these diseases. With computer models and applied mathematics, one can predict and control various diseases. We present the evolution and implementation of different models used in mathematical modelling as well as a demonstration of the SIR model in MATLAB.

Introduction

The WHO estimates that 2030-2050 an additional 250,000 people will die due to infectious diseases specifically because of climate change. This is because of increased flooding created by climate change that breeds diseases like cholera and other diarrhoeal diseases.¹ Some other reasons that explain why epidemics are unpreventable are because of increased contact between animals and humans, similar to the cause of COVID-19 but is also applicable in the case of factory farming in the US, and a health worker shortage caused by a major exodus of doctors and nurses in third world countries like Africa and Asia.¹ It is clear that governments and corporations will continue innovating technology to combat these diseases similar to the international vaccine effort of 2021. Yet, reality warns humanity that the next pandemic will be more deadly than the previous one, so the question is not how can we prevent the next epidemic but instead how to analyze the disease so scientists understand how the disease will spread, its mortality rate, how many waves to expect, and what governmental policy is necessary in order to protect citizens. So, it is already clearly established that epidemics are not only costly in terms of human life but also a threat to the international economy. The effects of such epidemics include the closure of business, schools, markets, preventing traveling for leisure, intermingling, and implementing curfews and lockdown.¹ A reduction of these harms is only possible through understanding the disease and implementing countermeasures. Mathematical modelling of diseases and developing stronger global surveillance methods is clearly the way to ensure the smallest loss of lives due to future inevitable epidemics.

Background

History

History has been seriously affected by outbreaks and pandemics from the beginning of time. In fact, the first recorded major pandemic was in 430 BCE Athens during the Peloponnesian War.² But the first paper discussing mathematical modelling of diseases was in 1766 by a scientist of the name Daniel Bernoulli. In his groundbreaking paper *Essai d'une nouvelle analyse de la mortalité causée par la petite vérole*, he talked about the outbreak of smallpox in England and created the first model based on math that calculated the mortality rate of smallpox.³ Despite various mathematicians like Lambert and Laplace referencing similar methods to Bernoulli, mathematical modelling of diseases only

started again in 1911 by Ross to study “discrete-time dynamics of malaria through the mosquito-borne pathogen transmission”.⁹ He was the first to address mass-action incidents (“rate at which an infection passes in a population is proportional jointly to the product of the number of persons I who are infectious and the number of persons S who are susceptible to the infection $C = rIS$ ”⁴) and this was later created into the Law of Mass Action coined by Guldberg and Waage in 1864 which was later christened the SIR model since it would mean that a person’s contact would randomly spread throughout.⁹ Although separate from the SIR model, another model that came about in 1928 by W. H Frost was the Reed- Frost chain binomial model that rather than assuming homogeneous mixing instead believed that infection spread between people through discrete time Markov chain events (non-coincidental chronological events)⁵ and this became mainstream known as stochastic epidemic modeling.⁹

Common Terms in Mathematical Modelling

Before statistical models are created for particular diseases, first transmission models are made to help understand the pattern and evolution of transmission between various individuals. So for most infections (ie. influenza, smallpox, malaria) people can be put into three categories: “susceptible,” “infected”, or “recovered and immune.”⁶ There are still various nuances in these four categories since an exposed individual (infected) may not be capable of spreading the diseases, so we call these latent cases. Also, there are three main periods of time between the exposure of the disease, the transmission, and the later infectiousness.⁶ In the latent period, the individual doesn’t pass on the disease. The incubation period is the time between exposure to the disease and then the biological response to the infection. There are special cases where a person can transmit the disease before symptoms are felt and in those cases mathematical models must be adapted.⁶ Finally, the infection period is when the infected person can transmit the disease.⁶ Next, we have primary cases which are the first people in an outbreak and the secondary cases and the average number of people a single primary case can infect is known as the reproduction number, denoted with the letter R . So, in order for the pandemic to end, the last known case must not reach out to any susceptible individuals, meaning $R < 1$. The infection attack rate (A) is the random number of individuals that come in contact with the disease. Finally, the exponential growth rate of infected individuals is defined as r and it is determined with the timeline for infection which is different for each disease.⁶ In order to shape public policy of disease control like border control using mathematical models it is key that these variables be as close as possible to their true value.

Methods to Modelling

Statistical-Based Model

The first method, which is also one of the first methods created, includes mostly statistical methods for surveillance of outbreaks. An example of this would be a regression method, which focuses on the statistics and timeline of reported infected cases from when diseases weren’t there. A basic regression model is shown below:

$$y(t) = a + bt + \sum_{i=1}^m c_i \cos \theta + \sum_{i=1}^m d_i \sin \theta + e(t)$$

Where θ is a linear function of time (t), the coefficients are determined by parameter identification techniques, and $e(t)$ is the “noise” which is estimated from the time-series. Another example of a statistical based model, are the Hidden Markov models (HMM). This particular model is very unique for it allows scientists to model for diseases that don’t have specific parameters. For example, Johne’s disease passes its infection through animals’ feces but the amount of MAP (Mycobacterium avium subsp. paratuberculosis)- an enteritis that leads to diarrhea- in each varies and the disease progression has not one but two distinct patterns.⁷ Specifically, what makes Johne’s disease hard to model is that it is a slowly progressing disease so the parameters on the rate of infection or recovery are hard to process. So,

in order to fix this problem HMM's eliminate the parameters using machine learning to estimate them like a disease state transition rate.⁷

Empirical/Machine Learning-Based Model

The second model comes from machine learning for predicting how an ongoing epidemic will spread and this would include online data mining for information and statistics and surveillance networks. One common technique that comes under this type of model is data that comes from communication platforms and search engines. This means “surveys, administrative data (including vital registration data, census data, epidemiological, and/or demographic surveillance data), hospital data, insurance claims data, disease registries, and other related sources” are all used in order to decide if an epidemic is spreading. BioBERT is a “contextualized word representation model” that can understand what a text is saying and the context behind it.⁸ BioBERT can specifically understand articles related to the biomedical science (ie. NCBI, PubMed, and more).⁸ An example of where this might be applicable is that BioBERT can understand the circumstances needed for a disease to happen like overcrowding or warfare, and then can search for where circumstances like that can be happening.⁸ BioBERT has been particularly useful when predicting for previously unseen countries and is consistent across all age groups. However one common issue with these types of techniques is that they require lots of data set training in order to get proper results.⁸ Another application of a machine learning based model is that Hulth et al. (2009) model processed information coming from a Swedish website related to influenza between 2005-7. By comparing the amount of information from the website and the number of verified influenza cases at the time, the model was able to calculate the spread of the outbreak and predict where it would spread next.¹⁰

Mathematical/Mechanistic State-Space Models (SIR Model)

The third mathematical model and the most commonly used is the state-space model that predicts how “hypothetical” epidemics spread. The most simple of these methods is the SIR model without considering demographics that might make certain individuals more susceptible to a disease. The SIR method focuses on the transition of states that individuals will go through modeled by differential equations. The S in SIR stands for the group of individuals vulnerable to the disease, I stands for those infected, and the R represents the people who leave the process through recovery. N, which is not in the acronym, is the total number of individuals in the population; this means that $N = S + I + R$.¹⁴ It is important to note that an individual may leave this process at any time due to death. The differential equations that model the relationship between these variables are listed below.

$$\begin{aligned} dS/dt &= -\text{incidence} = -\beta SI/N \\ dI/dt &= \text{incidence} - \text{recovery} = \beta SI/N - \gamma I \\ dR/dt &= \text{recovery} = \gamma I \end{aligned}$$

The differential ($d(\cdot)/dt$) is supposed to represent the change in that particular variable over a certain period of time and β is the transmission rate and γ is the recovery rate. In the simplified model presented here, the inverse of the recovery rate would equal the average length of the period of infectiousness. Also, since we know that when $t \approx 0$, $S/N \approx 1$ we can make the simplification into this equation.

$$dI/dt = \beta I - \gamma I$$

Dividing the transmission rate (β) and the recovery rate (γ), defines a new variable as (R_0) for the reproduction number. With all of these equations and established relationships, we can solve for the final prevalence equation: $I(t) = I_0 e^{\gamma(R_0 - 1)t}$ where I_0 is the initial number of infections at the beginning of the epidemic.⁹ We can rewrite this equation in the form $I' = \gamma R_0 I - \gamma I$ and when this is added to the recovery equation one gets $I' + R' = \gamma R_0 I$ which leads us to

the final equation of incidence = γR_0 (prevalence). In this simplified explanation of the SIR model, it is important to note that we have omitted other facts like how the endemic state is dependent on vaccination coverage (p) and therefore herd immunity. More complex compartmental models include incorporating heterogeneity (the differences of a population i.e. race) or including a continuous age structure.⁹

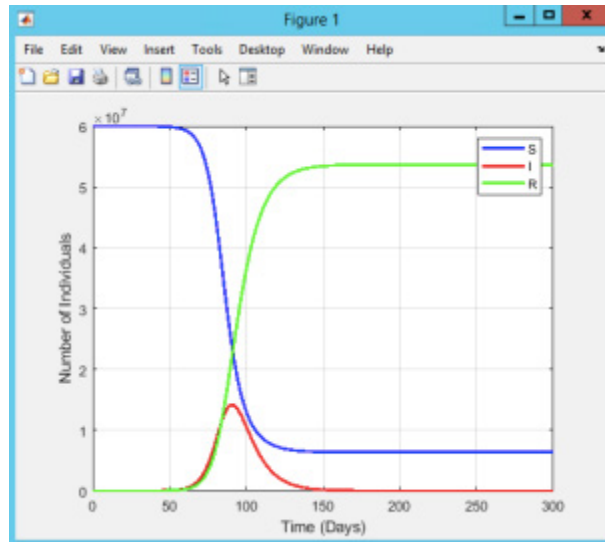
Example of Mathematical Modelling

In order to model a hypothetical disease, I coded a simple model using MATLAB. The model is relatively simple, with only seven parameters: beta, gamma, delta, N, I0, T, and dt. The beta in the equation represents the rate of infection, the gamma represents the rate of recovery, and delta is the rate of immunity loss. N, I0, and T are all the same as stated in the beginning of the paper and dt is the time interval used. It is important to notice that the assumptions in the for loop are different from the ones in the SIR model because this model includes a new variable (delta) that accounts for the immunity loss. In addition, the original equations under the SIR section divide $S(t)$ by N however in this case, the units inputted already reflect this, but other than that, the equations are relatively the same. The code after the for loop creates the plot.¹³

```
function [S,I,R] = sirs_model(beta,gamma,delta,N,I0,T,dt) 13
    S=zeros(1,T/dt);
    S(1)=N;
    I=zeros(1,T/dt);
    I(1)=I0;
    R=zeros(1,T/dt);
    for tt= 1:(T/dt)-1
        % Equations of the model-
        dS=(-beta*I(tt)*S(tt) + delta*R(tt)) * dt;
        dI=(beta*I(tt)*S(tt) - gamma*I(tt)) * dt;
        dR=(gamma*I(tt) - delta*R(tt)) * dt;
        S(tt+1)=(S(tt) + dS);
        I(tt+1)=I(tt) + dI;
        R(tt+1)=R(tt) + dR;
    end
    tt = 0:dt:T-dt;
% Curve
    plot(tt,S,'b',tt,I,'r',tt,R,'g','LineWidth',2); grid on;
    xlabel('Time (Days)'); ylabel('Number of Individuals');
    legend('S','I','R');
end
sirs_model(5*10^-9, 0.12, 0.0, 6*10^7,10, 300, 1/4)
```

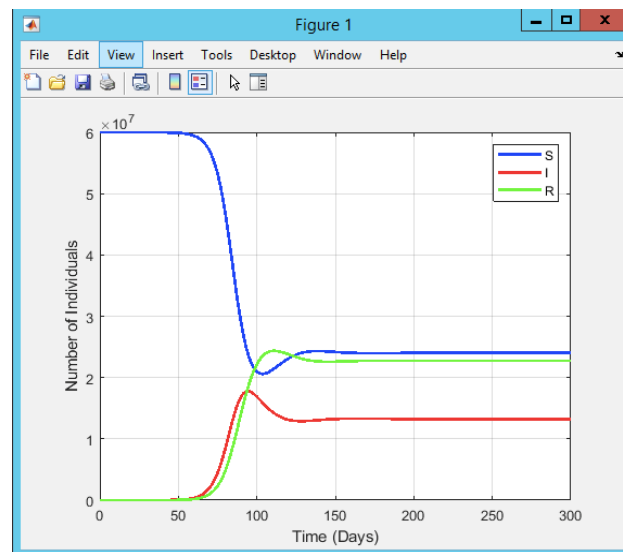
By inputting 5×10^{-9} for beta, .12 for gamma, 0.0 for delta, 6×10^7 as the total population, 10 as the number of initial infections, 300 days for time, and $\frac{1}{4}$ for T, the following graph is created.

In the graph, we see that the susceptible population significantly decreased over time while the recovered population increased over time. There are no fluctuations in this graph because someone who has recovered from the hypothetical disease can not get it again because there is no immunity loss.



This condition is not very common in the real world, for example after getting vaccinations many individuals have been infected with COVID-19 again and although unlikely, might pass away because of it. Let's introduce a new variable, delta(0.07), to add some nuance.

In this new graph, we can see some fluctuation and the graphs for the susceptible population and



the recovered population are very different from their previous one. It is important to note that this code and model is very simple and can by no means predict something as complex as COVID-19 because it doesn't take into account various factors like age, socioeconomic status, location, and more. However, the code does give an insight on how a coded mathematical model can help us predict the future of an epidemic.

Modelling Use in Public Policy

In order to properly fight pandemics, global leaders need to enact domestic policy to curb their citizens' "risky" behavior during pandemics. But to decide what policy to propose is based on mathematical modelling and extensive monitoring.⁹ The World Health Organization (WHO) and the Centers for Disease Control (CDC) in the US have

started creating a global surveillance network so sudden epidemics don't turn into pandemics. This was useful in the case of AIDS, SARS (2002), the 2009 H1N1 swine flu, and COVID-19 pandemic since constant surveillance and contact tracing were able to ensure variables in the mathematical model were as accurate as possible.⁹ In fact, for the COVID-19 pandemic the CDC started their own forecasting for deaths, cases, and hospitalizations due to COVID. With this information, the CDC has "published to help people in different risk groups (like healthcare workers or older people) stay safe in different settings (like grocery stores, home, or school)."¹¹ In addition, it was Giordano et al. new SIDARTHE model that recommended and later influenced Italian policy for comprehensive testing, contact tracing, and minimal contact with other individuals during the COVID-19 pandemic.¹² This model was expanded by Memon et al. as they provided a more comprehensive SIR model which included quarantine and isolation periods and proceeded that quarantining was an effective method to prevent the spread of the disease.¹²

Discussion

In order to properly fight pandemics, global leaders need to enact domestic policy to curb their citizens' "risky" behavior during pandemics. But to decide what policy to propose is based on mathematical modelling and extensive monitoring.⁹ The World Health Organization (WHO) and the Centers for Disease Control (CDC) in the US have started creating a global surveillance network so sudden epidemics don't turn into pandemics. This was useful in the case of AIDS, SARS (2002), the 2009 H1N1 swine flu, and COVID-19 pandemic since constant surveillance and contact tracing were able to ensure variables in the mathematical model were as accurate as possible.⁹ In fact, for the COVID-19 pandemic the CDC started their own forecasting for deaths, cases, and hospitalizations due to COVID. With this information, the CDC has "published to help people in different risk groups (like healthcare workers or older people) stay safe in different settings (like grocery stores, home, or school)."¹¹ In addition, it was Giordano et al. new SIDARTHE model that recommended and later influenced Italian policy for comprehensive testing, contact tracing, and minimal contact with other individuals during the COVID-19 pandemic.¹² This model was expanded by Memon et al. as they provided a more comprehensive SIR model which included quarantine and isolation periods and proceeded that quarantining was an effective method to prevent the spread of the disease.¹²

Conclusion

From my research, I was able to examine the background of mathematical modelling of epidemiology and explore different models of it. In addition, I was able to create my own hypothetical model using MATLAB to model a fictitious infectious disease. Finally, I was able to study the impact of such modelling in public policy in other countries as well as discussing further implications for newer technologies like increased surveillance.

Limitations

Since this is a research article, my research is limited by the data available on the web. Other than what I may not access, there are few limitations to my research. In addition, other limitations are created by using MATLAB since my research is limited to one software.

Acknowledgments

I would like to thank Dr. Tavernetti and Dr. Gravner for mentoring me on mathematics and for assisting me whenever I needed help on my paper. In addition, I would like to thank Professor Lea for providing me with a strong background on how to create a proper research paper. This paper would not be possible without their contributions.

References

1. 5 Reasons Why Pandemics like COVID-19 Are Becoming More Likely, GAVI, 28 Apr. 2020, www.gavi.org/vaccineswork/5-reasons-why-pandemics-like-covid-19-are-becoming-more-likely.
2. History.com Editors. "Pandemics That Changed History." History.com, A&E Television Networks, 27 Feb. 2019, www.history.com/topics/middle-ages/pandemics-timeline.
3. Blower S, Bernoulli D. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. 1766. *Rev Med Virol*. 2004;14:275–88. doi: 10.1002/rmv.443. [PubMed] [CrossRef] [Google Scholar]
4. Wilson, Edwin B, and Jane Worcester. The Law of Mass Action in Epidemiology, NCBI, 13 Dec. 1944, www.ncbi.nlm.nih.gov/pmc/articles/PMC1078744/pdf/pnas01670-0031.pdf.
5. Kumar, Ridhima. Marketing Analytics through Markov Chain, Towards Data Science, 28 Aug. 2019, towardsdatascience.com/marketing-analytics-through-markov-chain-a9c7357da2e8.
6. Milwid, Rachael, et al. Toward Standardizing a Lexicon of Infectious Disease Modeling Terms, *Frontiers*, 28 Sept. 2016, www.frontiersin.org/articles/10.3389/fpubh.2016.00213/full.
7. Ceres, Kristina M., et al. "Characterizing Infectious Disease Progression through Discrete States Using Hidden Markov Models." *PLOS ONE*, Public Library of Science, 20 Nov. 2020, journals.plos.org/plosone/article?id=10.1371/journal.pone.0242683.
8. Zhang, Yuanzhao, et al. "Applying Artificial Intelligence Methods for the Estimation of Disease Incidence: The Utility of Language Models." *Frontiers in Digital Health*, *Frontiers*, 15 Dec. 2020, www.frontiersin.org/articles/10.3389/fdgth.2020.569261/full.
9. Siettos, Constantinos I, and Lucia Russo. "Mathematical modeling of infectious disease dynamics." *Virulence* vol. 4,4 (2013): 295-306. doi:10.4161/viru.24041
10. Hulth, Anette et al. "Web queries as a source for syndromic surveillance." *PloS one* vol. 4,2 (2009): e4378. doi:10.1371/journal.pone.0004378
11. "About Covid-19 Epidemiology." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 1 July 2020, www.cdc.gov/coronavirus/2019-ncov/science/about-epidemiology/index.html.
12. Akyildiz, F Talay, and Fehaid Salem Alshammari. "Complex mathematical SIR model for spreading of COVID-19 virus with Mittag-Leffler kernel." *Advances in difference equations* vol. 2021,1 (2021): 319. doi:10.1186/s13662-021-03470-1
13. Giovanni Valentini (2021). `sirs_model` (https://www.mathworks.com/matlabcentral/fileexchange/82993-sirs_model), MATLAB Central File Exchange. Retrieved July 29, 2021.
14. Smith, David, and Lang Moore. "The Sir Model for Spread of Disease - the Differential Equation Model." *The SIR Model for Spread of Disease - The Differential Equation Model*, Mathematical Association of America, Dec. 2004, www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model.