Modeling the Spread of the Zika Virus

CS224W Final Paper

Aashna Garg, Anusha Balakrishnan, Sheema Usmani, Vinaya Polamreddi

Dec 11th, 2016

1 Abstract

In this paper, we use various epidemic models to model the spread of Zika virus through Brazil. We use a dataset that reports the number of infected cases across each of the states in Brazil over an 8 week timespan as our ground truth. We also use open source flight data and geographic information to help understand connections between the locations and cases to help us model the network. We test models in which the network nodes are locations and models in which the network nodes are people and the edges are modeled using geographic information and flight information. We also model the internal connections of people with a location using a random graph model where the edge probability was determined by population density of the location which we mined from available data sources. We explored the ability of standard epidemic models such as SIR and SEIR to capture the spread of Zika virus. We found that the SIR models and SEIR models fall short of modeling the Zika virus well due to a few reasons including that our ground truth data doesn't include recovered cases and that Zika uses human contact of bodily fluids to spread which our network does not model those connections. Our findings indicate that the SIR and SEIR models might not be that effective at representing the Zika epidemic. Further, we need to obtain more fine-grained data regarding the spread of the virus, such as the number of recoveries and detailed social network relationships that cause the virus to spread.

2 Related/Previous Work

In Nah et al. (2016)[1], the authors proposed a model to predict the risk of importation (introduction) of the Zika virus over time to different countries, based on several factors, including flight network information, the country-specific presence of Zika-transmitting mosquitoes, and country-specific data regarding the chikungunya and dengue epidemics (since they often tend to be transmitted by the same mosquito species locally). The probabilistic model that the authors used for prediction was able to somewhat roughly capture the spread of the virus considering Brazil as the only source.

Amit et al.(2010)[3] label edges with probabilities of influence (between nodes) from a log of past propagations. They described both static and time-dependent models for capturing influence. Additionally they also developed techniques for predicting the time by which a user may be expected to perform an action.

3 Dataset and Collection

The dataset that we used to construct the network and model the spread of the Zika virus through that network were derived from three sources: a Zika epidemic dataset released by the CDC, a manually collected dataset of land border information and population statistics for every state in Brazil, and two datasets from OpenFlights ontaining airport and airline route data.

3.1 CDC Zika Virus Epidemic Dataset

For our project, we use the Zika Virus Epidemic dataset released by the CDC on Kaggle[1]. This dataset contains statistics on the number of Zika-related cases per province/state, per country, and per week, for a number of countries in North and South America, including some countries in the Caribbean. For this project, we only make use of the per-state statistics provided in this dataset for Brazil. There are 8 weeks of data available in this dataset, from April to June 2016.

[1] https://github.com/cdcepi/zika/

3.2 Airports and flight route data

The OpenFlights website [2] provides several datasets related to airlines and airports, two of which we use for this project. First, we used the airports dataset to collect the names and cities of all of the airports in Brazil. Since our network (described in section 4) is organized into clusters representing states, we then manually mapped each of these airports to a state, based on the city that the airport is located in. Finally, we used the flight routes dataset to determine the number of flight connections between every pair of states in our dataset.

3.3 Geographic and demographic data

For each state, we manually compiled a dataset containing geographic and population-based statistics that we used when creating the network:

- **Land border information**. For each state, we collected the names of the states that it shares a land border with from Wikipedia. This information is used to determine which states are connected to each other in our network.
- **Population statistics**. We gathered information on the total population and population density of each state. This information is used to determine how dense or sparse the clusters representing the state in our network should be.

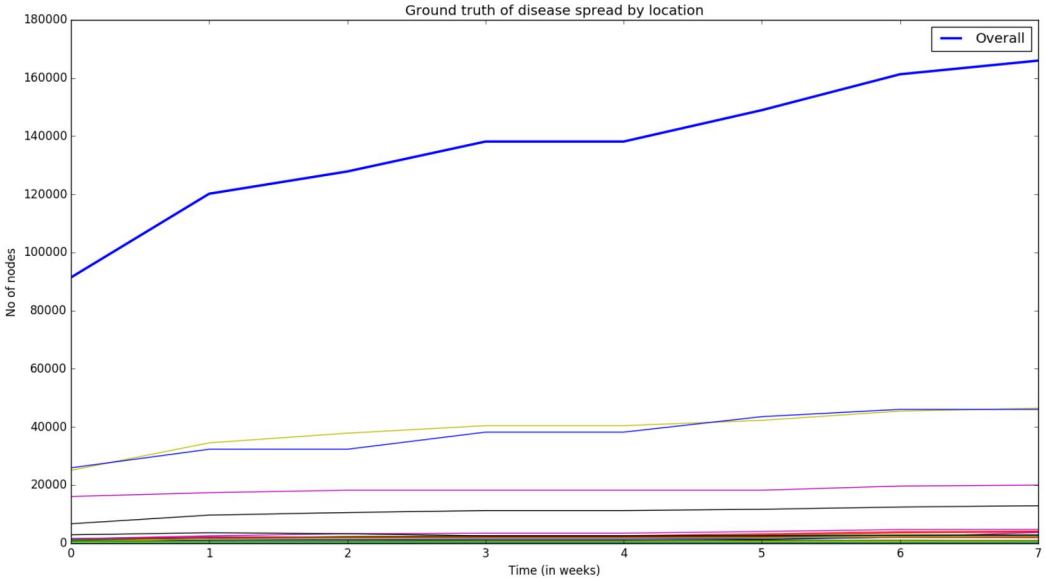
3.4 Dataset collection

To construct the network that we describe in Section 4 below, we manually collected geographic proximity information: for each state, we collected information regarding the states that border it, and used this information to construct our graph. To get the connectivity between every state of Brazil, we collected flight connectivity information by getting the number of flight routes between each pair of states and assuming a constant number of passengers and a constant flight frequency. We also collected information regarding the population of each state (to compute the fraction of infected individuals in that state), as well as the population density (to estimate how connected people within a given state are to others in the state). We then combined this information with the publicly available CDC Zika virus epidemic dataset for each state. Thus, the final dataset that we composed consists of two major parts: a connectivity dataset that defines the graph structure (i.e. defines how people in different states are connected to other people intra- and inter-state), and a timestamped dataset containing metadata relevant to each state. In this case, the metadata that we store is the number of reported cases of Zika in that state (from the Brazil Epidemiological Bulletin data provided by the CDC), the total population of the state at that timestep [3], and the population density of the state. Since the CDC dataset only provides granularity of up to a week, each timestep in this case is a week (and the number of infected individuals is cumulative). There are 8 weeks of data available in this dataset, from April to June 2016.

[2] http://openflights.org/data.html

[3] Since the population of the state is several orders of magnitude higher than the number of infected individuals, small changes in population don't affect the fraction of infected individuals to a great degree. Thus, we simply use the population estimate from the most recent publicly available census of that state.

Figure 1: Spread of the Zika virus, in terms of number of infected individuals, over time. The blue line represents the total number of cases across all states, and the other lines represent the number of cases per state.



4 Network

In this project, we attempt to model both the more global spread of the Zika virus across different states in Brazil, as well as the local spread of the virus within each state. Thus, our network has a global structure to represent high-level interactions between states and a local structure to represent local interactions within each state. At the global level, our network consists of $K = 27$ "clusters" of nodes, where each cluster represents one of the states in Brazil, and edges are created between nodes in different clusters to capture inter-state connectivity and model the spread of the virus across states.

4.1 Local connectivity

At the local level, in order to allow our model to accurately capture the number of individual cases in each state, each state in our network is actually represented by a cluster of nodes, where each cluster is an Erdös-Renyï (ER) random graph with n nodes and m edges. The number of nodes n_i in the i-th state or cluster is $n_i = P_i \cdot c$, where P_i is the population of the i-th state, and $0 < c \leq 1$ is a scaling factor to make the number of nodes in each cluster feasible, since the populations of each state are very high (the lowest population is 400K and the highest is 16M). Thus, the nodes in each state represent groups of c people in that state, and edges between nodes represent social interactions between We use the population density d_i to calculate the number of edges in each cluster, by assuming that the cluster has the same number of edges as a k-regular graph with n nodes and $k = d$. Thus, the number of edges m_i in the i-th cluster is $m_i = n_i \cdot d_i/2$. To create the ER graph, we first create the cluster with n_i nodes, and then add m_i random edges, ensuring that a different edge (u, v) is picked every time. In practice, we set $c = 1/10000$.

4.2 Global connectivity

We used two different types of data to create inter-state connectivity (i.e. connect nodes that belong in different states/clusters). First, we used land border information to create l edges between states that share a land border. Second, we used flight route information from the OpenFlights dataset to create edges between states that are connected by a flight route. In order to understand the effectiveness of each of these data types, we created three different networks using this data:

- A network where global edges were created using only the flight data. In this network, the number of edges $e_{A,B}$ between two states A and B is $e_{A,B} = r \cdot F \cdot f_{A,B}$, where $f_{A,B}$ is the number of flight routes between A and B, and F is the average frequency of flights along each route per week.
- A network where global edges were created using only land border information. In this network, the number of edges $e_{A,B}$ between two states A and B is $e_{A,B} = l \cdot Iborder(A, B)$, where $Iborder(A, B)$ is an indicator variable that indicates whether A and B share a land border.
- A network where global edges were created using both land border information and flight data. In this network, the number of edges $e_{A,B}$ between two states A and B is $e_{A,B} = l \cdot Iborder(A, B) + r \cdot F \cdot f$.

  In practice, we set $r = 200$, $F = 1$, and $f = 200$ (the average capacity of a passenger flight). Since each state is itself a cluster of nodes, we created $e_{A,B}$ global edges by randomly choosing one node each from A and B respectively, and connecting those nodes.

  Thus, the total number of edges in our graph is $E = \left(\sum_{i=1}^{i=K} P_i \cdot c \cdot d_i/2\right) + \sum_{i=1}^{i=K} \sum_{j=i+1}^{j=S} e_{i,j}$, where $e_{i,j}$ is determined by the network type as above.

5 Methods

5.1 Assumptions

We make the following assumptions for the following models [6].

- The population is fixed with no births or deaths.
- The only way a person can leave the susceptible group is to become infected. The only way a person can leave the infected group is to recover from the disease. Once a person has recovered, the person received immunity.
- All people are treated equally.

5.2 SIR Model

The SIR Model is used in epidemiology to compute the amount of susceptible, infected, recovered people in a population.

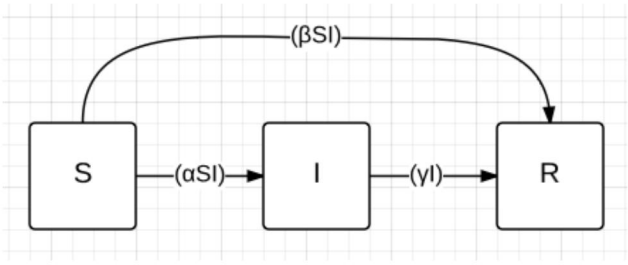


Figure 2: SIR flow [8]

5.2.1 SIR Formulas

The SIR models the flows of people between three states: susceptible (S), infected (I), and resistant (R). This model is represented by the following:

$S(t)$ is the number of susceptible individuals at time t
$I(t)$ is the number of infected individuals at time t
$R(t)$ is the number of recovered individuals at time t
N is the total population size

$$\frac{dS}{dt} = -\beta S(t)I(t) \tag{1}$$

$$\frac{dI}{dt} = \beta(S(t) - k)I(t) \tag{2}$$

$$\frac{dR}{dt} = k(I(t)) \tag{3}$$

where k is the recovery rate (with greater or equal to zero), α is the probability of becoming infected,γ is the number of people infected person comes in contact with in each period of time on average, β is the average number of transmissions from an infected person in a time period (with β greater or equal to zero), and

$$S(t) + I(t) + R(t) = N \tag{4}$$

5.3 SEIR Model

The SEIR models the flows of people between four states: susceptible (S), exposed (E), infected (I), and resistant (R). Each of those variables represents the number of people in those groups. The parameters α and β control how fast people move from being susceptible to exposed, from exposed to infected (σ), and from infected to resistant (γ) [5].

The SEIR differs from the SIR model in the addition of a latency period. Individuals who are exposed (E) have had contact with an infected person, but are not themselves infectious. Figure 3 shows the patterns that emerge in the number of infected, susceptible, exposed, and recovered individuals in a typical SEIR model.

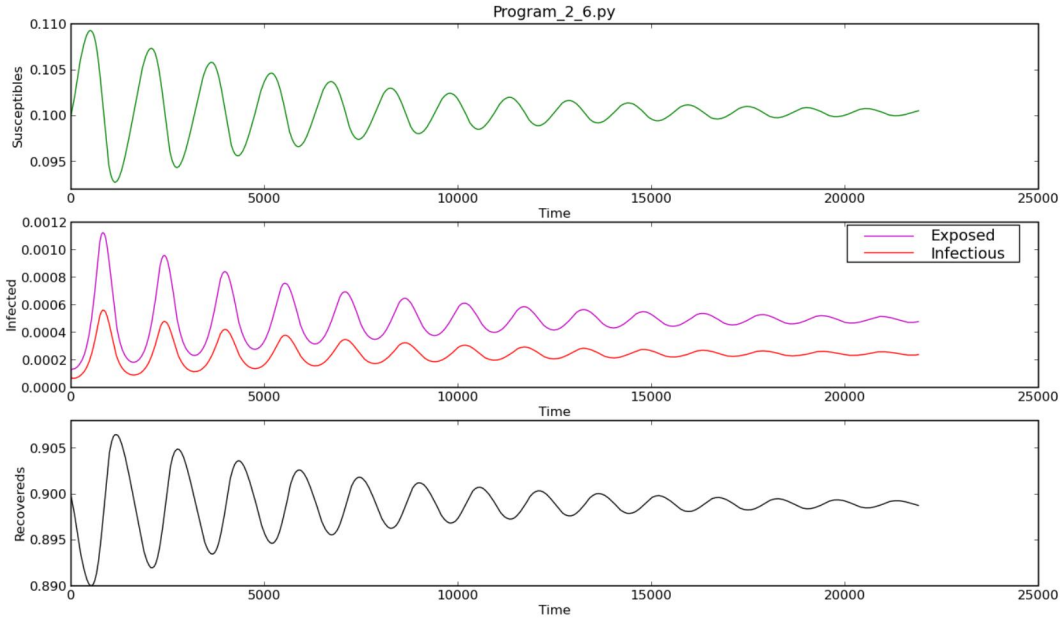


Figure 3: Number of infected, susceptible, exposed, and recovered individuals in a typical SEIR model [9].

5.3.1 Parameters

- S is the fraction of susceptible individuals (those able to contract the disease)
- E is the fraction of exposed individuals (those who have been infected but are not yet infectious)
- I is the fraction of infective individuals (those capable of transmitting the disease)
- R is the fraction of recovered individuals (those who have become immune)
- β The parameter controlling how often a susceptible-infected contact results in a new exposure.
- γ The rate an infected recovers and moves into the resistant phase.
- σ The rate at which an exposed person becomes infective.

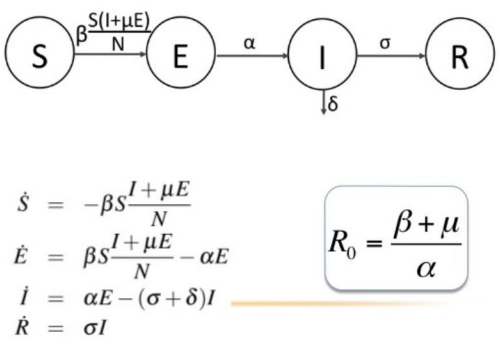


Figure 4: SEIR equations [7]

6 Results

For both SIR and the SEIR models, we tried using different values of the model parameters (for SIR, the transmission probability α and the recovery probability ; for SEIR, the exposure probability α, the infection probability σ, and the recovery probability γ) in increments of 0.01 to try to determine the best range of values. To evaluate the performance of the model, we calculated the L1 loss function

$$\sum t = 1^{t=T} \sum_{i=1}^{i=K} |\hat{y_i^{(t)}} - y_i^{(t)}|$$

,

where K is the total number of states (in this case, 27), $y_i^{(t)}$ is the true number of infected individuals that are infected in state i at time t, and $\hat{y_i^{(t)}}$ is the predicted number of infected individuals in state i at time t. We also evaluated a baseline model that assumes a very simple network structure.

6.1 Baseline

For our baseline model, we constructed a simple network consisting of $N = 27$ nodes. In this network, each node represents a state, and the number of infections spread from node A to node B is $\frac{k_i}{p_i} \cdot \alpha$, where k_i is the number of infected individuals, p_i is the population of the state, and α is the transmission probability. We used the SIR model to model the spread of the virus through this network, and tried different values of the parameter α in increments of 0.01 to find the value that minimized the L2 loss. Figure 5 shows the change in the L2 loss as α is varied. We see that this naive model is extremely noisy, and varying the α values has no real effect in predicting the spread of the virus through the network.

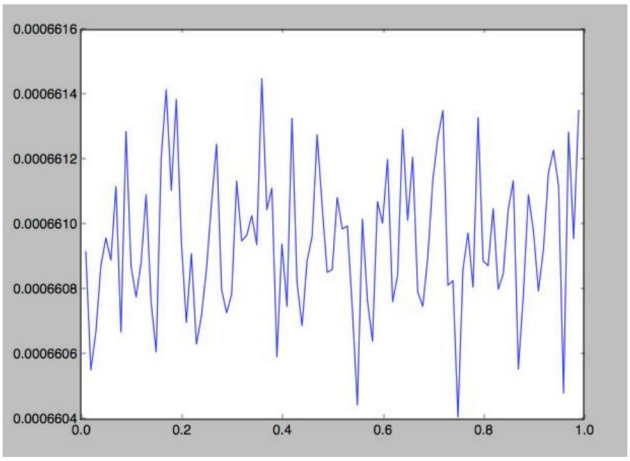


Figure 5: L2 loss vs. α for our baseline model.

6.2 SIR model

As shown in Figure 6, the number of infected individuals grows extremely fast over time in the SIR model, and thus does not in any way capture real trends. Further, we see that the patterns predicted by this model don't change significantly when inter-state edges are created using more information (i.e. when using flight and land border information, as opposed to just using flight or land information).

6.3 SEIR model

Figure 7 shows the predictions of the SIR and SEIR models (respectively) with the lowest loss on each network type. We see that the number of infected individuals predicted by the SEIR model grows slowly and thus bears some resemblance to the actual change found in the ground truth data. Further, the change in number of infected individuals resembles the change observed in the typical SEIR model in Figure 3. However, the patterns are still significantly different overall. Further, this model is also largely unaffected by any changes in the global structure of the network. These results suggest that the SIR and SEIR models are not sophisticated enough to capture the spread of the Zika virus.

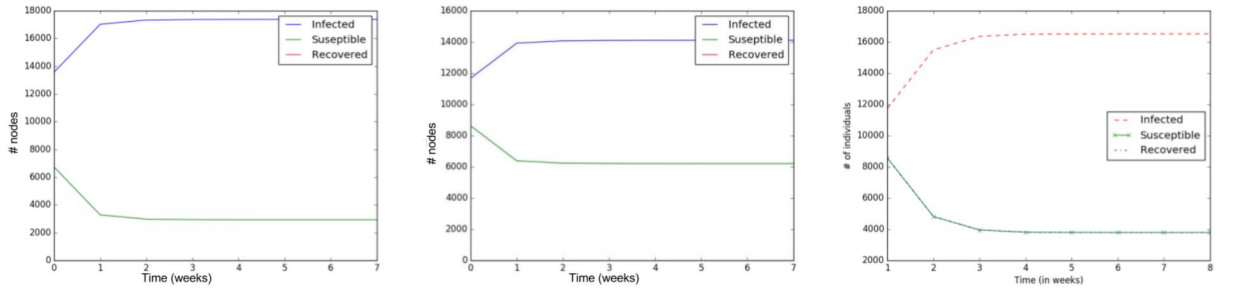


Figure 6: SIR model predictions on a network created using (from left to right): a) flights and land border information; b) flight information only; c) land border information information. Blue lines indicate number of infected individuals, and green lines indicate number of susceptible individuals.

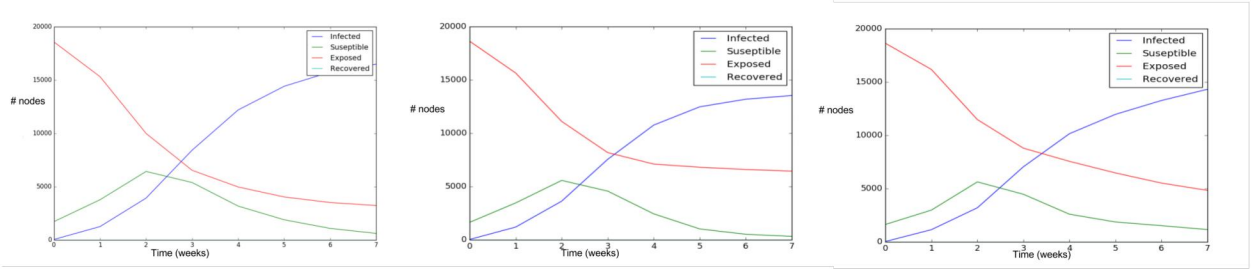


Figure 7: SEIR model predictions on a network created using (from left to right): a) flights and land border information; b) flight information only; c) land border information information. Blue lines indicate number of infected individuals, green lines indicate number of susceptible individuals, and red lines indicate the number of exposed individuals.

7 Conclusion and Further Work

Since the SIR and SEIR models with the network architecture as built are unable to represent the spread of Zika accurately, there are a few directions that can be explored in the future. First, the network can be built from more relevant and representative data. The Zika virus travels from human contact rather than just location proximity or other factors, so the graph connections, for example, could be built from social graphs rather than travel or location information. Second, the SIR and SEIR models are also represent fundamentally different models of propagation than the Zika virus, as shown from the ground truth data over time and how it compares to the SIR and SEIR model simulations. Third, further work can include modifying the SIR/SEIR models to include techniques such as representing the the recovery rate as a Gaussian random variable rather than a constant, and trying different probabilistic techniques such as $\alpha = \beta \cdot Y/N$ (where Y = number of reported cases, N = total population size, β = product of contact rates and transmission probability), maximum likelihood estimate and basic reproductive ratio. We can also include features regarding Zika-transmitting mosquitoes, such as the presence and spread of mosquitoes, rate of contact between mosquitoes and humans and the number of times a female mosquito bites in her lifetime.

This analysis in exploring the representative power of SIR/SEIR models for Zika on a network based on location, travel and geographic information showed us that there is more complexity to the spread of the Zika virus than was captured by these models. We need to understand and represent hidden and latent factors that play a role in the spread of the Zika virus are important to factor into the network and model.

References

[1] Nah, Kyeongah, et al. "Estimating risks of importation and local transmission of Zika virus infection." PeerJ 4 (2016): e1904. APA

[2] Myers, Seth A., and Jure Leskovec. "Clash of the contagions: Cooperation and competition in information diffusion." 2012 IEEE 12th International Conference on Data Mining. IEEE, 2012. APA

[3] Amit Goyal , Francesco Bonchi , Laks V.S. Lakshmanan, Learning influence probabilities in social networks, Proceedings of the third ACM international conference on Web search and data mining, February 04-06, 2010, New York, New York, USA [doi¿10.1145/1718487.1718518]

[4] Rothenberg, Richard B., et al. "Social network dynamics and HIV transmission." Aids 12.12 (1998): 1529-1536. APA

[5] "SEIR Model." SEIR Model. N.p., n.d. Web. 11 Dec. 2016.

[6] Johnson, Teri. "Mathematical Modeling of Diseases: Susceptible-Infected-Recovered (SIR) Model." Math 4901 Senior Seminar (2009 Mathematical Modeling of Diseases: Susceptible-Infected-Recovered (SIR) Model): n. pag. Mathematical Modeling of Diseases: Susceptible-Infected-Recovered (SIR) Model. Web.

[7] SEIR Model. N.p., n.d. Web.

[8] SIR Model. N.p., n.d. Web.

[9] SEIR Model Wikipedia. N.p., n.d. Web.

Individual Contributions

- **Aashna**: Processed airports and flight connectivity data, worked on paper
- **Anusha**: Collected land border information and total population numbers; wrote code to create Erdos-Renyi based states network using collected data; worked on paper
- **Sheema**: Collected population density data, implemented visualizations for the results; worked on paper
- **Vinaya**: Implemented the SIR and SEIR models for networks based on nodes representing people and locations; worked on paper