Epidemics on Transportation Networks

By Ho Lum Cheung, Dimas Muntesinos, Frank Acquaye, Elie Wanko

27 JUN 2020

Abstract

Public transportation plays a vital role connecting people with each other. International trade and tourism is reliant on commercial aviation. And the most successful cities all have buses, trams, or subways connecting workers to their workplaces. However, the speed and connectivity of these transportation networks leave us vulnerable to the spread of diseases.

In this paper, we introduce a general agent-based framework for modeling disease propagation on transportation networks. We use it to build models of major transportation networks, and take a special look at the New York City subway to see how well the model can predict infection rate and infection hotspots during the 2020 COVID-19 epidemic.

Predicating contagion on subway lines, ridership at stations, commute time, and city-wide countermeasures, we were able to fit our model closely to empirical data. This suggests a correlation between these factors and the spread of COVID-19 and other ILI.

1 Background

1.1 Epidemics and COVID-19

Coronavirus disease 2019 (COVID-19) is a disease caused by the SARS-CoV-2 coronavirus. Since being identified in December 2019, it has been labelled by the WHO as a pandemic, and spread around the world. Epidemics such as the coronavirus have been a subject of research for centuries, and is of special interest to those working in public health. Recent waves of new research came in 2002 (SARS), 2009 (H1N1), and 2014 (Ebola). However, in these prior epidemics, researchers did not have access to as much data as we have currently. In recent years, the state of data science research tools has also greatly improved, allowing researchers to answer questions in novel ways, but still following the scientific method.

1.1.1 Mathematical Modeling of Epidemics

There are two basic types of modeling for epidemics; statistical and mechanistic. While statistical modeling has typically given more accurate forecasts for a well-known situation, mechanistic models such as the SIR and SEIR compartmental models help to explain why phenomena such as epidemics spread the way they do, and what impact various policy decisions will have on the result.

1.1.2 SEIR Model

As we will be using the SEIR model[1] in our research, we will cover it briefly here. The SEIR compartmental model is a mathematical modeling of infectious diseases where

a closed population of people move successively from compartment to compartment (from susceptible to exposed, to infected, and to removed). We briefly provide an explanation of each compartment and leave the corresponding system of equations below.

- **Susceptible (S)** - These people are susceptible to getting the disease from someone infected.
- **Exposed (E)** - These people are no longer susceptible to the disease, and do not infect others. After a latent period, they become infectious.
- **Infected (I)** - These people will spread the disease to susceptible people. After a period of time they are removed by isolation, recovery, hospitalization, or death.
- **Removed (R)** - Sometimes known as resistant or recovered. We will do our modeling with the term 'removed'. These people no longer spread the disease.
- **Contact Rate (β)** - Rate at which infected people infect susceptible people.
- **Latent Rate (α)** - Rate at which exposed people become infected.
- **Removal Rate (γ)** - Rate at which infected people become removed.

$$S(t) + E(t) + I(t) + R(t) = N$$

$$s(t) = S(t)/N, e(t) = E(t)/N, i(t) = I(t)/N, r(t) = R(t)/N$$

$$ds/dt = -\beta * s(t) * i(t)$$

$$de/dt = \beta * s(t) * i(t) - \alpha * e(t)$$

$$di/dt = \alpha * e(t) - \gamma * i(t)$$

$$dr/dt = \gamma * i(t)$$

The system of equations described above can be numerically solved given β, α, γ and initial values $S(0), E(0), I(0), R(0)$. And if we have values for S,E,I,R at certain times, we can fit β, α, γ to better define the disease's epidemiological characteristics and predict its future course. Lastly, we note that R_0, an important characteristic known as the basic reproductive rate, can be calculated for the SEIR model as $R_0 = \beta/\gamma$.

1.1.3 Other Compartmental Models

While we have done research into simpler and more advanced models and are interested in cases such as super-spreaders, we believe the SEIR model to be sufficient for our needs.

Basic SIR is insufficient because public health officials often make policy decisions based on positive case numbers. For example, an official may decide to impose strict isolation only after 100 positive cases. But by the time there are 100 cases of 'infected' people, there may be 1000 exposed people who will meaningfully impact epidemic statistics.

1.2 Epidemics on Transportation Networks

Epidemics on transportation networks have been modeled in many different ways depending on the needs of the researcher. The most important differences are usually the type of transportation network, the duration of interest, the granularity of population data, and passenger flow characteristics. A popular motivation is exploring countermeasure policies. We highlight some prior research in the sections below.

1.2.1 Flu on the London Underground [2]

In this paper, the researchers modeled the "contact rate" of riders of the London Underground (subway system) by breaking down the stages of subway travel (entering, waiting, riding, exiting), and concluded that riders of some boroughs were at higher risk than riders from other boroughs. Their analysis was consistent with PHE data for influenza-like illnesses (ILI). The researchers also made additional observations such as that rush hour contributes to infection due to higher passenger density and longer waiting times.

1.2.2 Time-dependent links in Singapore[3]

In this paper, which is a pre-publication as of June 1, 2020, the researchers modeled the Singapore bus system while limiting the spread of disease to passengers in the same location. They investigated various countermeasures such as reducing bus service and requiring personal protective equipment (PPE).

1.2.3 Passenger Flow on the World Aviation Network[4]

In this paper, not directly related to epidemiology, the researchers proposed an open-source statistical modeling of passenger flow in the World Airline Network(WAN). This would be useful for modeling the spread of disease worldwide via aviation. It seems that while they were initially successful, their website and data are no longer available.

1.2.4 Domestic Passenger Flow in Russia[5]

In this blog post, Tutu, a Russian travel website, gathers, shares, and analyzes holistic passenger flow data (buses, trains, airplanes) across Russia. Treating Russia as a closed system due to closed borders, they estimate the start and end of the outbreak in Russian cities using an SIR epidemiological model.

1.2.5 A Simulation of New York City[6]

In this paper, the researchers make a very thorough model of New York City. While the title suggests they focused on subways, they modeled hospitals, schools, and major hubs with the subways delivering people to their places of work and education. They

fit their work to historical flu data as well as infection numbers from prior work and also investigated the effect of countermeasures (interventions). They estimated that the subway was directly responsible for 12.5% of all infections in NYC.

1.2.6 Our Own Research

In this paper, we focus on the NYC Subway. However, in the appendix we use our framework to model some other transportation networks to demonstrate the robustness of the framework, show similarities and differences between different types of transportation, and highlight key general facts about epidemics on transportation networks.

1.3 Agent-Based Models

An agent-based model(ABM) is a computational model used to simulate the overall effects of individual agents on a system. Some famous prior uses include Conway's Game of Life and Schelling's Segregation Model[7]. ABMs consist of one or several types of agents interacting with an environment. For example, in Schelling's Segregation Model, the agents have a race and a tolerance level (of other races). If the agent finds the surrounding environment intolerable, they will independently move away. Given certain hyper-parameters, segregated communities eventually form.

ABMs offer a number of benefits over traditional mathematical models. Complex systems which cannot be easily solved mathematically can be simulated. These simulations help policy makers with decisions when mathematical results are not available and real world experiments are impractical. [8].

For example, one popular use case of ABMs is in urban planning to simulate traffic flow. Given empirical traffic flow data, it is virtually impossible to calculate the optimal traffic light configuration for a city and extremely expensive to run multiple configuration experiments. However, ABM software with various vehicles as agents and the city traffic grid as the environment can easily find a satisfactory configuration [9].

1.4 COVID-19 in NYC

In addition to general knowledge about COVID-19, it would behoove the reader to know about the early spread of the disease in New York. Analysis of viral RNA in patients at the Mount Sinai Health System has lead researchers to conclude that the virus first came into the community through "multiple, independent but isolated introductions" from Europe and elsewhere in the USA[10]. Below we have also given an approximate timeline of some of the most relevant events as of June 1st, 2020. Note that as forensic researchers begin examining the data, significant new dates or corrections may arise:

- February 25 - First positive test in NYC from a 39 year-old female healthcare worker flying back from Iran.
- March 3 - First confirmed P2P spread in NYC.

- March 9 - Mayor holds press conference and notes that there have been 16 confirmed cases.
- March 12 - Mayor declares a local state of emergency.
- March 15 - Schools officially close.
- March 22 - State order (PAUSE) to shelter in place comes into effect.

1.5 COVID-19 and the NYC Subway

On April 24th, 2020, a researcher at MIT released a working paper finding that "The Subways Seeded the Massive Coronavirus Epidemic in New York City"[11]. The paper received a lot of attention from the media. There does not seem to be enough evidence to prove causation, and correlation could be related to a third factor. For example, high infection rate could be rated to dense housing, and housing is denser around subway stations.

From this paper, and other prior research into subways, we bring into our own work two main ideas to test about disease spread on subways:

- Infection on a subway network depends much more strongly on the line than the geographical distance or shortest path between stations.
- Infection on a subway network depends on the average commute time of commuters using the station.

1.6 Subway Nomenclature and Other Definitions

- Station - Passengers enter subway stations in order to ride the subway to an exit station.
- Complex - Multiple stations can reside in one station complex.
- Turnstiles - Barriers at the entrance and exit of stations which count people entering and exiting.
- Line - The train tracks on which services and routes run.
- Service/Route - Trains follow specific routes between stations based on a timetable.
- Borough - A geographical region. NYC has 5 boroughs.
- MODZCTA - Modified Zip Code Tabulation Areas. They are very similar to zip (postal) codes, but are used over postal codes to better identify regions and their populations.

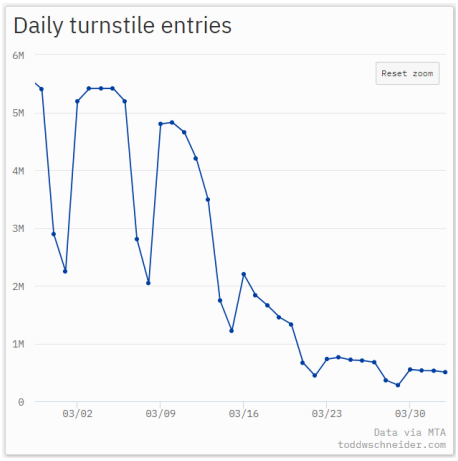


Figure 1: NYC Subway daily turnstile entries for March 2020 [14]

2 Data Sources and Preprocessing

2.1 MTA Station Data [12]

This dataset has basic subway station information. It has stations, their longitude, latitude, associated complex (if any), line, route, and borough. We can find the MODZCTA from the longtitude and latitude data.

In this dataset, station 167 is listed twice, so we took care to combine the route data. We also split service 'S' into 3 different services as it represents 3 different shuttle services. We also split services on the 'A' line, into 3 different services depending on the destination.

2.2 MTA Turnstile Data [13]

The MTA also publishes turnstile data with turnstiles reporting every 4 hours. However, we only need a basic picture of the daily in and out flow of each station, so we only process the aggregate passenger flow of each station between March 1 and March 21 (inclusive).

2.3 NYC COVID-19 Data

Since March 26, the NYC Health Department has been releasing and updating COVID-19 data on Github[15]. Some data is incomplete or unavailable due to technical or privacy issues. For example, detailed case, death, and recovery numbers by MODZCTA only became available on May 18, 2020. However, a rudimentary record of positive tests for COVID-19 by MODZCTA has been available since April 1, 2020.

3 Methodology

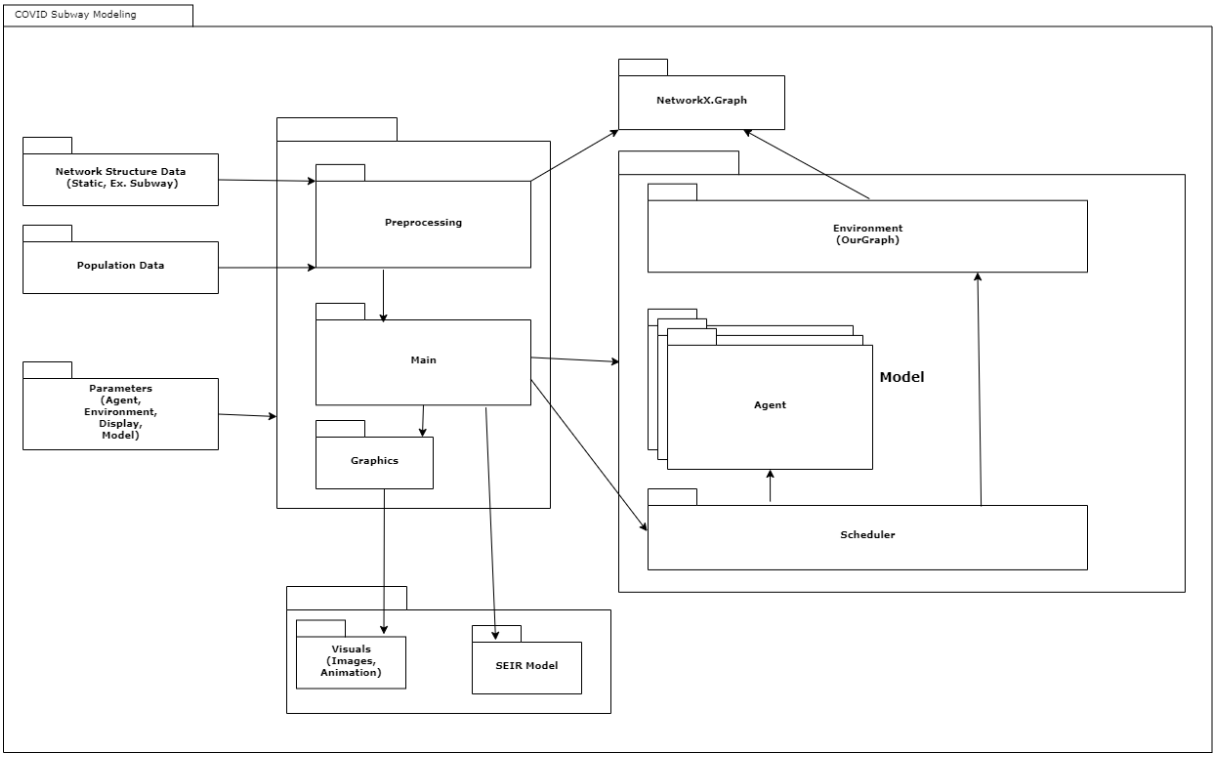
3.1 MESA[16]

To implement our ideas, we chose to use MESA, an ABM framework written in Python. There are many other ABM frameworks, but many of them are oriented towards begin-

ners and lacking features to organize the code necessary for complex behavior. Others are not suitable for modeling a complex environment (network). MESA provides a simple framework with basic Agent, Model, and Schedule classes from which we can build more complex behavior. In addition, our group already has a working knowledge of modeling networks using Python's NetworkX[17], which will be necessary for specifying the environments of our models.

3.2 Modeling Framework

Before focusing on the NYC subways, we built a general framework for modeling various transportation problems. This allowed us to see how subways (and subway commuters) differ from other types of transportation networks. It also allows us to extend our initial research into other networks and environments. In the appendix, the reader will find some other models we have built and the key point behind the modeling. Below is a UML diagram of our framework.



3.3 SubwayModel

This class inherits from the base TransportationModel. The main additional functionality is based on the model using SubwayAgents and a SubwayGraph to model the environment. It also has functionality related to deploying countermeasures given a certain threshold of infection.

3.3.1 SubwayAgent

This class inherits from the base SEIRAgent and represents the population living around a specific subway station. The main additional property is the number of commuters, and

the main additional functionality is adjustment of infection (and removal) rates based on exposure to the disease.

3.3.2 SubwayGraph

This class represents the environment of our agents. Each node represents a subway station. Each edge represents a connecting subway line or passage between stations. It is a wrapper around a NetworkX Graph with additional functionality.

3.4 NetworkX[17]

This is a Python library used to model networks.

3.5 Important Parameters, Hyper-parameters, and Values

3.5.1 Basic Epidemic Characteristics

- DEFAULT_BETA - The default β of the modeled disease in the SEIR model. It can be affected by countermeasures.
- DEFAULT_GAMMA - The default γ of the modeled disease in the SEIR model. It can be affected by countermeasures.
- DEFAULT_ALPHA - The default α of the modeled disease in the SEIR model. It can be affected by countermeasures.

3.5.2 Countermeasures

- ISOLATION_COUNTERMEASURE - Models a government order to stay isolated after a certain number of people are infected.
- RECOMMENDATION_COUNTERMEASURE - Models a government recommendation to be safe after a certain number of people are infected. The isolation countermeasure supercedes this countermeasure.
- AWARENESS_COUNTERMEASURE - Models increasing public awareness after a certain number of people are infected. As time passes since the public first becomes aware of a problem, the infection and removal rates decrease due to self-initiative.

3.5.3 Values

These values were chosen to approximate the epidemiological characteristics of COVID-19. Other sources[18][10][19][20] have investigated the best numbers more thoroughly, but for our modeling it is only necessary that the values are not preposterous (more than 200% difference).

Values	Contact Rate (β)	Latent Rate (α)	Removal Rate (γ)	Start Trigger
Default Rates	**1.75**	**0.2**	**0.5**	t = 0
Isolation Modifier	0.25	1	2	I >5000
Recommendation Modifier	0.67	1	1.5	I >500
Awareness Modifier	1 to 0.25	1	1	I >500

3.5.4 Other Parameters

We also used a number of other parameters.

- Global Exposure Rate - This adds an additional infection chance based on globally infected. The value of 0.7 indicates that we think approximately 30% of viral propagation is subway-based and the other 70% is not subway based.
- Defiance - The local population will defy government orders if very few people are infected in their area.
- Awareness Increase Rate - Awareness increases by a negative exponential up to 0.75. The rate is calculated so that most of this gain is achieved in the first month.
- Borough Commuting Modifier - This modifier approximates the percent of the population which commutes by borough. This is in lieu of better research which estimates the percentage of commuters by station.

3.6 Algorithm

Below is a simplified algorithm for our model.

Algorithm 1 Simulation of Disease Spread on Subways

for $i = 1; i < TIMESPAN; i++$ **do**
    Check conditions (i, number of infected) to see if we should deploy COUNTERMEASURES
    for Station in SubwayModel.Environment.Nodes **do**
        Calculate 'Local Exposure' from locally infected and commute time.
        Calculate 'Route Exposure' from infected on the same route.
        Calculate 'General Exposure' due to city-wide infected.
        Update 'Exposure' At Station based on above conditions
    end for
    for Agent in SubwayNetwork.Agents **do**
        Get 'Exposure' At Location
        Get City-wide COUNTERMEASURES
        Get Percentage of commuters
        Calculate SEIR beta and gamma based on conditions
        Update SEIR numbers
    end for
end for

4 Fitting and Results

4.1 Fitting to SEIR

We first fit overall SEIR numbers to NYC case, death, and recovery numbers. We only sought to fit the total cases to $I(t) + R(t)$ and adjusted the countermeasure properties to do so. Since we were at complete liberty to adjust these numbers, it is unsurprising that we created a near-perfect fit.

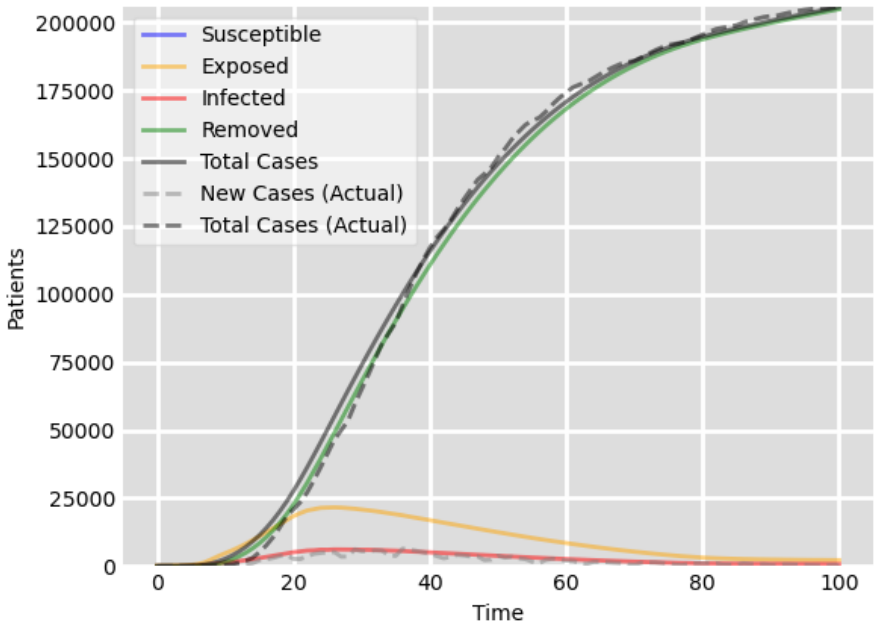
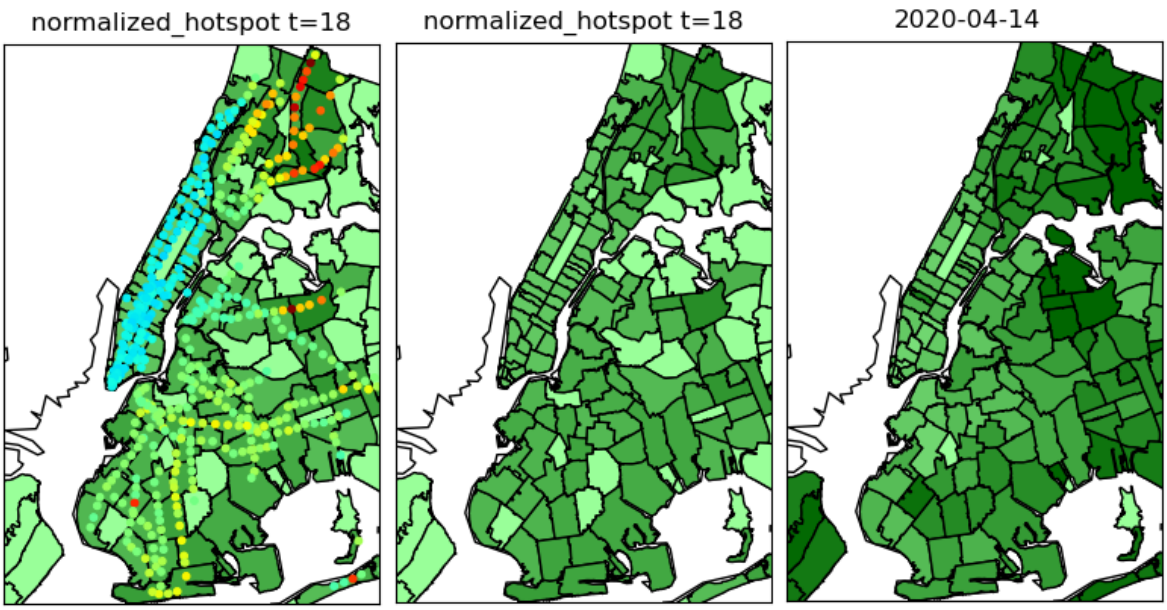


Figure 2: SEIR Fitting to NYC Case Total. *MAPE ($t \geq 30$): 0.0145*

4.2 Fitting to MODZCTA

Next we fit our data to MODZCTA. Due to the lack of numbers, we are only able to fit starting on April 1st (t=32). When calculating fitting error using MAPE, we exclude MODZTAs with no subway stations.



Scale(0 - 0.11 infections/person, 0 - 0.11 infections/person, 0 - 0.20 cumulative cases/person)
$MAPE(t \geq 32, Stations \geq 1 = 0.311)$

A MAPE Of 0.311 is much worse than the overall error. But it still suggests that there is some correlation. We would expect that even with additional tuning and more granular subway data, there would be a large amount of error due to infection spread which cannot be modeled with currently available data.

5 Further Research

5.1 Controlling for demographics

While there is good correlation between empirical infection rates and our modeling, there is also likely to be good correlation between these rates and demographic data like population density and income. We have not tried to control for this.

5.2 NYC: An open system

Our research into New York City suggests that there is a significant amount of commuting from outside the 5 boroughs [21]. In addition, we have not even considered Staten Island.

5.3 Fitting to a different system

We have not fit our research to a different subway system for a different city.

5.4 Unnecessary hyper-parameters

While we initially started with a minimal number of hyper-parameters, the model has become bloated with additional parameters which accommodate missing data and provide for a better fit. We should investigate ways to obtain or interpolate the data or question whether it is necessary to model it.

6 Conclusion

In this paper, we used a general agent-based framework to create a model for the spread of COVID-19 through New York City via the subway system. We took into account various important factors such as commute time, subway routes, and city-wide countermeasures in the model hyper-parameters and found a good fit for overall New York City case data.

Predicating contagion on these parameters, we were also able to closely match the empirical data by MODZCTA. This suggests a correlation between these factors and the spread of COVID-19 and other ILI.

References

[1] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

[2] Lara Goscé and Anders Johansson. Analysing the link between public transport use and airborne transmission: mobility and contagion in the london underground. *Environmental Health*, 17(1), 2018.

[3] Yu Shen Clarence Tam Daqing Li Yafeng Yin Jinhua Zhao Baichuan Mo, Kairui Feng. Modeling epidemic spreading through public transit using time-varying encounter network. *Preprint, ResearchGate*. https://www.researchgate.net/publication/340541290_Modeling_Epidemic_Spreading_through_Public_Transit_using_Time-Varying_Encounter_Network.

[4] Liang Mao, Xiao Wu, Zhuojie Huang, and Andrew J. Tatem. Modeling monthly flows of global air travel passengers: An open-access data resource, Sep 2015.

[5] Infection scenarios for russian cities. https://story.tutu.ru/scenarii-zarazhenija-gorodov-rossii/.

[6] Philip Cooley, Shawn Brown, James Cajka, Bernadette Chasteen, Laxminarayana Ganapathi, John Grefenstette, Craig R. Hollingsworth, Bruce Y. Lee, Burton Levine, William D. Wheaton, and et al. The role of subway travel in an influenza epidemic: A new york city simulation. *Journal of Urban Health*, 88(5):982–995, Sep 2011.

[7] Thomas C. Schelling. Dynamic models of segregation†. *The Journal of Mathematical Sociology*, 1(2):143–186, 1971.

[8] Institute of Medicine, Amy B. Geller, V. Ayano. Ogawa, and Robert B. Wallace. *Assessing the Use of Agent-based Models for Tobacco Regulation*. National Academies Press, 2015. Appendix A: Considerations and Best Practices in Agent-Based Modeling to Inform Policy.

[9] Davy Janssens, Ansar-Ul-Haque Yasar, and Luk Knapen. *Data science and simulation in transportation research*. Information Science Reference, an imprint of IGI Global, 2014.

[10] Ana S Gonzalez-Reiche, Matthew M Hernandez, Mitchell Sullivan, Brianne Ciferri, Hala Alshammary, Ajay Obla, Shelcie Fabre, Giulio Kleiner, Jose Polanco, Zenab Khan, and et al. Introductions and early spread of sars-cov-2 in the new york city area. Nov 2020.

[11] Jeffrey E. Harris. The subways seeded the massive coronavirus epidemic in new york city. http://web.mit.edu/jeffrey/harris/HarrisJE_WP2_COVID19_NYC_24-Apr-2020.pdf.

[12] Mta station data. http://web.mta.info/developers/data/nyct/subway/Stations.csv.

[13] Mta turnstile data. http://web.mta.info/developers/turnstile.html.

[14] New york city subway usage. https://toddwschneider.com/dashboards/nyc-subway-turnstiles/.

[15] NYCHealth. Nyc coronavirus-data repository. https://github.com/nychealth/coronavirus-data.

[16] Mesa docuementation. https://mesa.readthedocs.io/en/master/overview.html.

[17] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, August 2008.

[18] Covid-19: Data. https://www1.nyc.gov/site/doh/covid/covid-19-data.page.

[19] Faïçal Ndaïrou, Iván Area, Juan J Nieto, and Delfim F M Torres. Mathematical modeling of covid-19 transmission dynamics with a case study of wuhan, Apr 2020.

[20] Fei Xu, C. Connell Mccluskey, and Ross Cressman. Spatial spread of an epidemic through public transportation systems with a hub. *Mathematical Biosciences*, 246(1):164–175, 2013.

[21] The ins and outs of nyc commuting. https://www1.nyc.gov/assets/planning/download/pdf/planning-level/housing-economy/nyc-ins-and-out-of-commuting.pdf.

[22] Albert-László Barabási and Márton Pósfai. *Network science*. Cambridge University Press, 2017.

[23] M. Laskowski, B. C. P. Demianyk, J. Witt, S. N. Mukhi, M. R. Friesen, and R. D. Mcleod. Agent-based modeling of the spread of influenza-like illness in an emergency department: A simulation study. *IEEE Transactions on Information Technology in Biomedicine*, 15(6):877–889, 2011.

[24] Plan of metro station novokuznetskaya. http://www.karta-metro.ru/stations/96/428/.

[25] https://openflights.org/.

[26] https://aci.aero/news/2019/03/13/preliminary-world-airport-traffic-rankings-released

[27] https://www.routesonline.com/news/29/breaking-news/286313/busiest-routes-in-the-world-the-top-100/.

[28] Madrid commuter rail routing data. https://crtm.maps.arcgis.com/home/item.html?id=1a25440bf66f499bae2657ec7fb40144.

[29] Madrid commuter rail turnstile data. https://data.renfe.com/dataset/volumen-de-viajeros-por-franja-horaria-madrid.

7 Appendix

7.1 Additional Models

7.1.1 Madrid Commuter Rail

Data

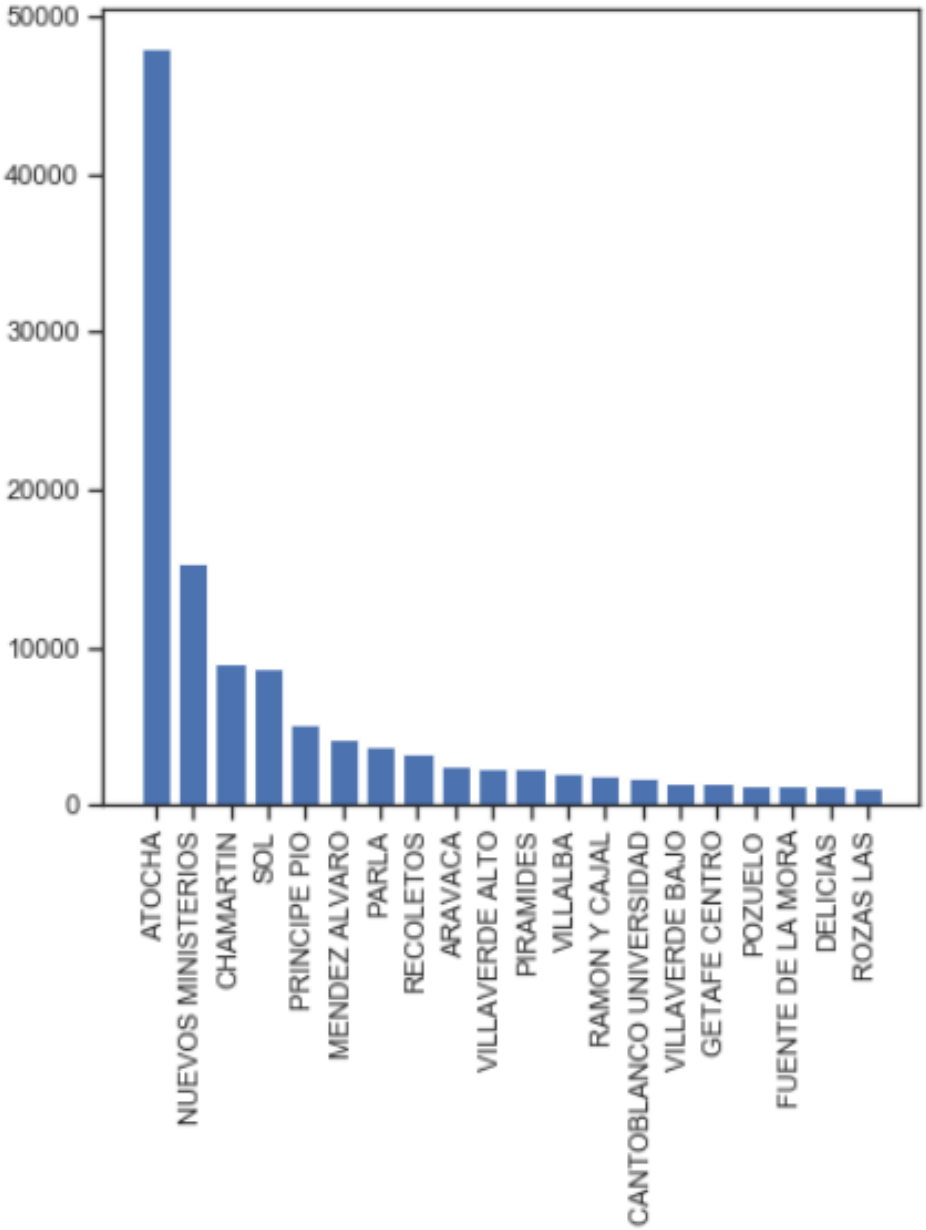
- Madrid Cercanias Routing Data [28]
- Madrid Cercanias Turnstile Data [29]

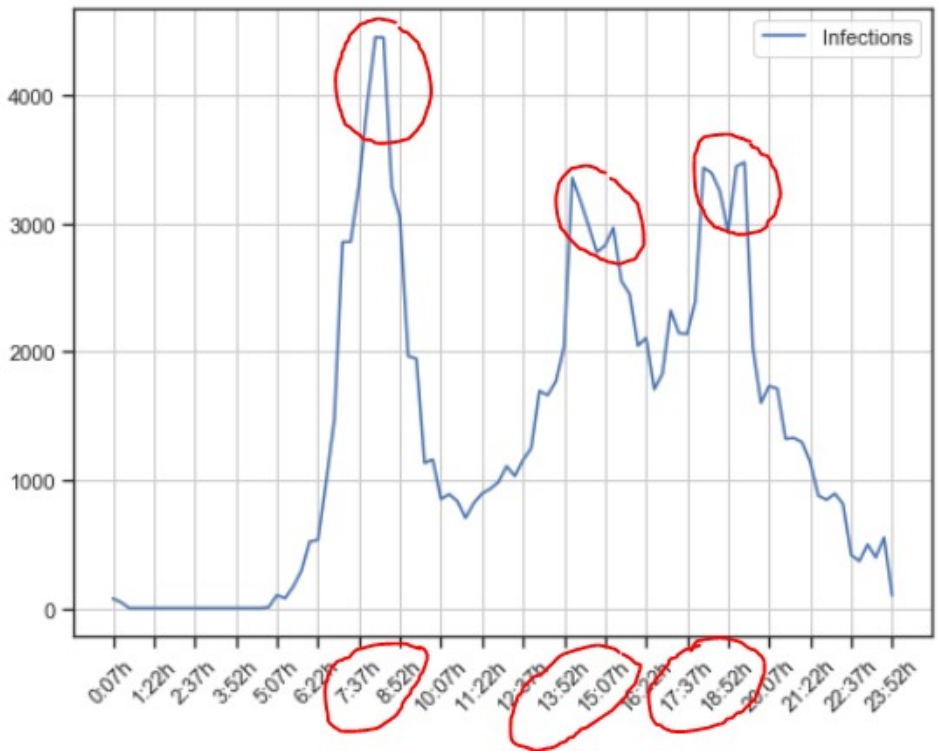
Methodology

1. Determine in and out flow weights for each station
2. Model trains running on the network
3. Model passengers entering, waiting, riding, exiting
4. Model passengers infecting passengers at the same location
5. Analyze infection rates and locations

Result

Our results show that the central hub at Madrid is the most dangerous station, and that most infections happen during Madrid's rush hours.





7.1.2 World Airline Network

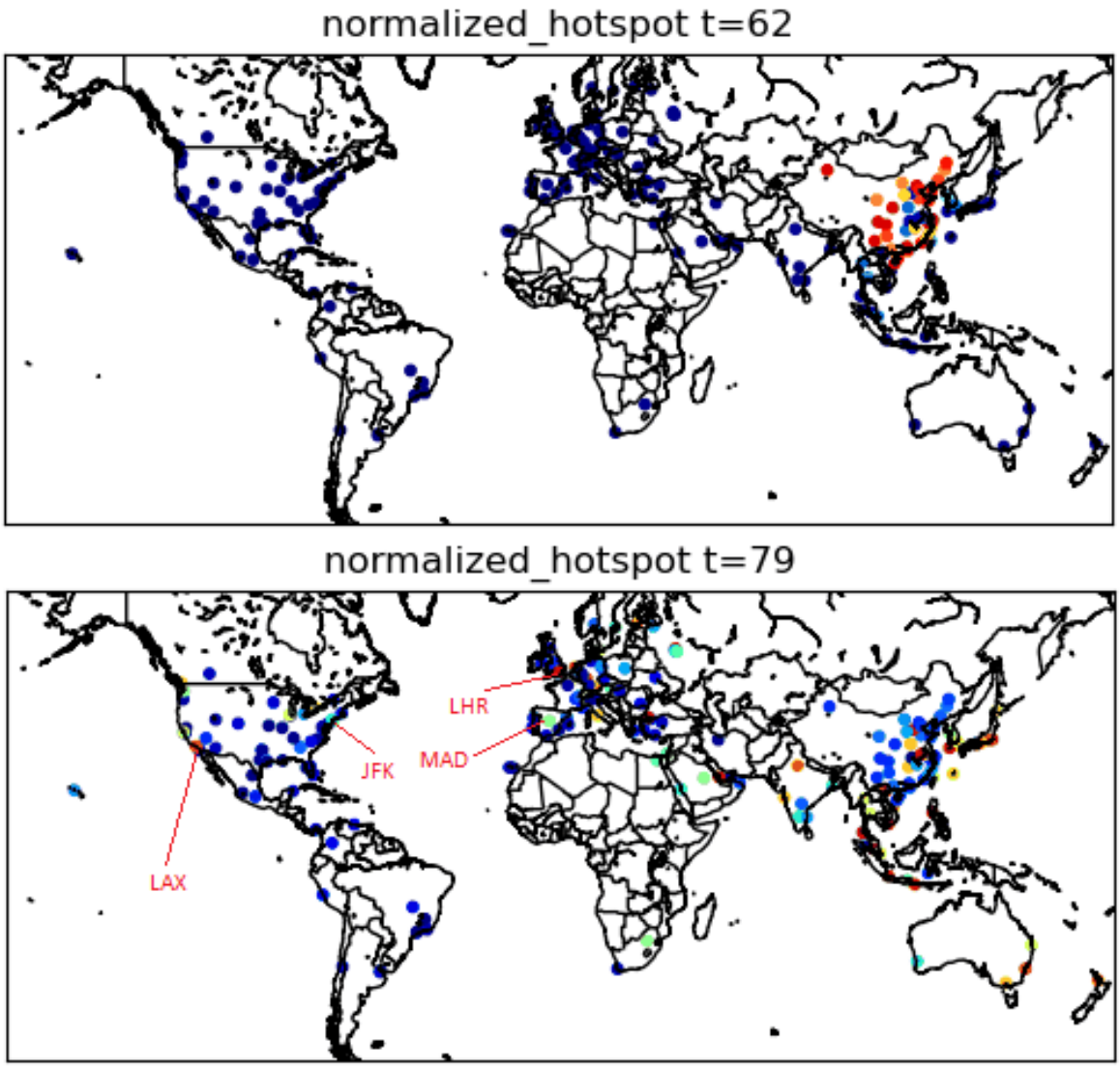
Data

- List of Airports - OpenFlights [25]
- List of Air Routes - OpenFlights [25]
- Airport Passenger Flow Numbers - [26]
- Top Air Routes [27]
- Other Air Routes

Methodology

1. Process airports, air routes
2. Add all known passenger flow numbers
3. Add all known airport and air route labels
4. Classify remaining airports and air routes
5. Interpolate missing air route data

Result Our results show that with accurate passenger flow data and labeling, it's possible to show the idea that diseases have a domestic and international wave.


normalized_hotspot t=62
normalized_hotspot t=79
LHR
MAD
JFK
LAX