Commutes and Contagions: Simulating Disease Propagation on Urban Transportation Networks

Ho Lum Cheung and Dimas Muñoz-Montesinos
HSE University, Moscow 101000, Russia
cheung.ho.lum@gmail.com, dmunosmontesinos@edu.hse.ru

Abstract. Public transportation plays a vital role in bringing people together. International trade and tourism is reliant on commercial aviation, and the most successful cities have buses, trams, or subways connecting workers to their workplaces. However, the efficiency and connectivity of modern transportation networks leaves us vulnerable to the spread of diseases.

In this paper, we introduce a general agent-based framework for modeling disease propagation on networks and use it to build models of major transportation networks. Using our models, we take a look at how disease spreads throughout a city. We pay particular attention to predicting the regions of a city which are likely to become disease hot-spots. This is of particular interest to city planners who want to allocate resources appropriately to combat an outbreak.

Using information about subway routes, ridership, and city-wide countermeasures, we were able to predict areas with high caseloads during the onset of the 2020 COVID-19 pandemic in New York City. Our findings suggest a correlation between these factors and the spread of COVID-19, Influenza-Like Illnesses, and other contagious diseases.

Keywords: Communicable · Disease · Transportation · Network

1 Introduction

When studying and discussing epidemics, news is often focused on national and international levels. However, there is significant empirical evidence that different regions and cities react differently to epidemics. And within individual cities, the spread of a disease such as the recent COVID-19 is uneven. Certain neighborhoods are much more likely to be affected than others.

While factors such as income and residential density are often studied for their correlation with this spread, in this paper we look at mobility data with a focus on subway systems. First, we built an agent-based framework to simulate infections in locations tied together by a transportation network. Localities (neighborhoods, postal codes, boroughs, etc.) are represented by an individual agent following a SEIR model, but modified to include an outside infection...
chance. We base this outside chance primarily on the commute time from the neighborhood and the primary subway routes serving it. We then use this agent-based framework to build models for COVID-19 in cities such as New York and London.

To build the transportation networks for our model, we pulled from publicly available data stores from various city government websites. We were also able to find satisfactorily granular geographical and demographic data from publicly available national census data.

We found that with minimal tuning based on the overall case rate, we can forecast the case rate of COVID-19 across all localities with a median MAPE less than 0.25. With better tuning, or location-specific hyper-parameters, we can further reduce this general result.

We also uncovered many interesting paths for future study such as creating a fuller mobility picture for individual cities, and leave some thoughts in the discussion. But the main conclusion is that commute times and subway routes are correlated with the propagation of disease within a city, and should be taken into consideration when planning for future outbreaks.

The paper is organized as follows. Section 2 contains a background on mathematical modeling of epidemics including SEIR model. Section 3 describes the used data sources. The proposed methodology is described in Sect. 4. The model fitting and obtained results are summarized in Sect. 5, while the discussion and future prospects are given in Sect. 6. Finally, Sect. 7 concludes the paper.

2 Background and Prior Research

2.1 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 [9], 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 the data and tools we have now.

2.2 Mathematical Modeling of Epidemics

SEIR Model. The SEIR compartmental model [17] 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). While statistical modeling has typically given more accurate forecasts for a well-known situation, mechanistic models such as SEIR do a better job with exploring general phenomena and explaining the impact of policy decisions. Below, we provide an explanation of each SEIR compartment and give the corresponding system of equations.
− **Susceptible** \((S)\) - These people are susceptible to infection from infected people.
− **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** \((\beta)\) - Rate at which infected people infect susceptible people.
− **Latent Rate** \((\alpha)\) - Rate at which exposed people become infected.
− **Removal Rate** \((\gamma)\) - Rate at which infected people become removed.

\[S(t) + E(t) + I(t) + R(t) = N\]

\[s(t) = \frac{S(t)}{N}, e(t) = \frac{E(t)}{N}, i(t) = \frac{I(t)}{N}, r(t) = \frac{R(t)}{N}\]

\[\frac{ds}{dt} = -\beta \cdot s(t) \cdot i(t)\]
\[\frac{de}{dt} = \beta \cdot s(t) \cdot i(t) - \alpha \cdot e(t)\]
\[\frac{di}{dt} = \alpha \cdot e(t) - \gamma \cdot i(t)\]
\[\frac{dr}{dt} = \gamma \cdot i(t)\]

The system of equations described above can be numerically solved given \(\beta, \alpha, \gamma\) and initial values \(S(0), E(0), I(0),\) and \(R(0).\) And if we have values for \(S, E, I,\) and \(R\) at certain times, we can fit \(\beta, \alpha, \gamma\) 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 \(R_0 = \beta/\gamma.\)

**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.

### 2.3 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 sub-area of research is exploring countermeasure policies. We refer the reader to the following papers [3,10,11], and discuss two of the most relevant ones below.

**Flu on the London Underground.** In [14], 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.

**A Simulation of New York City.** In [12], 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.

### 2.4 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 [22]. 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 [19].

### 2.5 Subway Nomenclature

Below we define some common terms used in subway systems.

- **Station** - Passengers enter subway stations in order to ride the subway to an exit station.
- **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.
3 Data Sources

3.1 London

Demographics. The finest granularity for which we found London COVID-19 data was by borough. Since March 1, the city government of London has been combining data from Public Health England and the Office of National Statistics (United Kingdom) to publish case, death, and recovery data for COVID-19 by borough [1].

For visualization, we used a shapefile for London boroughs also available from the city government of London [5].

Subway System. For the subway system, we used a parsed list of London Underground stations from Wikimedia [6], and verified the results against an official map [18].

3.2 New York City

Demographics. The finest granularity for which we found New York City COVID-19 data was by MODZCTA (Modified Zip Code Tabulation Area). Since March 26, the NYC Health Department has been releasing and updating this data on Github [21]. Some data is incomplete or unavailable due to technical or privacy issues. 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. The dataset also contains the estimated population of each MODZCTA.

For visualization, we used a shapefile for New York City from NYC Open Data [16].

Subway System. The New York City MTA (Metropolitan Transportation Authority), in charge of the New York City subway system, has a dataset containing basic subway station information [8]. It has stations, longitude, latitude, line, route, borough, and other information. While it does not have MODZCTA, we found it by using longitude and latitude.

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, and did the same for the branching ‘5’ line.

4 Methodology

4.1 MESA

To implement our ideas, we chose to use MESA [7], an ABM framework written in Python. MESA provides a simple framework with basic Agent, Model, and Schedule classes from which we can build more complex behavior.
4.2 NetworkX

NetworkX is a standard Python library used to model and visualize networks [15]. Its documentation can be found online.

4.3 Modeling Framework

First, 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 Fig. 1, the UML diagram of our framework is provided.

![UML Diagram of Framework](image)

**Fig. 1.** The UML diagram of our framework

4.4 Special Considerations for Subways

We next create the main components of the ABM using the following ideas:

- Infection on a subway network spreads along routes.
- Infection on a subway network depends on the average commute time of commuters using the station.
4.5 SubwayModel

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

SubwayAgent. This class inherits from the base SEIRAgent and represents the population living around a specific subway station. The base agent simulates a disease spreading through a closed population. The main additional functionality is adjustment of infection (and removal) rates based on outside exposure to the disease.

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 serves as a wrapper around a NetworkX Graph with additional functionality.

4.6 Hyper-parameters and Values

Basic Epidemic Characteristics

- DEFAULT_BETA - The default $\beta$ of the modeled disease in the SEIR model. It can be affected by countermeasures.
- DEFAULT_GAMMA - The default $\gamma$ of the modeled disease in the SEIR model. It can be affected by countermeasures.
- DEFAULT_ALPHA - The default $\alpha$ of the modeled disease in the SEIR model. It can be affected by countermeasures.

Countermeasures

- ISOLATION_COUNTERMEASURE_START - Models civic and government countermeasures to combat a disease. The countermeasures start after this number of people are infected.
- INITIAL_REDUCTION_TIME - Models the amount of time it takes to implement an initial set of countermeasures.
- INITIAL_REDUCTION_TGT - Models the effectiveness of the initial countermeasures. The $R_0$ of the virus will decrease to this number linearly over the INITIAL_REDUCTION_TIME.
- FULL_REDUCTION_TIME - Models the amount of time it takes to implement a full set of countermeasures.
- FULL_REDUCTION_TGT - Models the effectiveness of full countermeasures. The $R_0$ of the virus will decrease to this number linearly over the FULL_REDUCTION_TIME.
- STARTING_PERCENTAGE - Models the initial percentage of the population infected by the virus. Initial infected are evenly distributed.
STARTING_RATIO - Models the initial percentage of the population exposed to the virus. Initial exposed are evenly distributed.

CONTACT_MODIFIER - A parameter used to linearly modify the exposure from contacts by subway commuters. Note that comparing this modifier between different cities is meaningless unless all other exposure factors are normalized.

4.7 Algorithmic Description

Below is an algorithmic description of our agent-based model.

Algorithm 1. Simulation of Disease Spread on Subways

1: for $i = 1; i < \text{TIMESPAN}; i++$ do
2:     Check conditions ($i$, number of infected) to see if we should deploy COUNTERMEASURES
3:     for Station in SubwayModel.Environment.Nodes do
4:         Add ‘Exposure’ from infected at this station to routes stopping at this station.
5:     end for
6:     for Agent in SubwayNetwork.Agents do
7:         Get ‘Exposure’ from routes at this station.
8:         Get City-wide COUNTERMEASURES
9:         Get Percentage of commuters
10:        Update actual $\beta$ and $\gamma$ based on conditions
11:        Update (self) SEIR numbers
12:     end for
13: end for

Alternatively, the model can be described as changing SEIR to work as follows:

$$\Delta E = \frac{\beta SI}{N} + S\rho\mu,$$

where $\rho$ is the contact ‘density’ of the specific population and is directly linearly correlated to the factors listed in the algorithm (exposure, percentage of commuters), and $\mu$ is a global normalization hyperparameter. Note, of course, that $\Delta S$ is respectively altered.

5 Fitting and Results

Overall, in Table 1, we chose the following values for the hyper-parameters.
### Table 1. The choice of hyper-parameters

| Parameter                                      | London | NYC  |
|-----------------------------------------------|--------|------|
| DEFAULT_BETA                                  | 1.00   | 1.00 |
| DEFAULT_ALPHA                                 | 0.24   | 0.24 |
| DEFAULT_GAMMA                                 | 0.16   | 0.16 |
| ISOLATION_COUNTERMEASURE_START                | 200    | 500  |
| INITIAL_REDUCTION_TIME                        | 7      | 7    |
| INITIAL_REDUCTION_TGT                         | 1.00   | 1.50 |
| FULL_REDUCTION_TIME                           | 60     | 60   |
| FULL_REDUCTION_TGT                            | 0.60   | 0.60 |
| STARTING_PERCENTAGE                           | 1e$^{-6}$ | 1e$^{-6}$ |
| STARTING_RATIO                                | 2.5    | 2.5  |
| CONTACT_MODIFIER                              | 0.10   | 0.22 |
| COMMUTER_RATIO                                | 0.50   | 0.50 |

5.1 Choosing Epidemiological Characteristics

Default values for $\alpha$, $\beta$, and $\gamma$ were chosen to approximate the epidemiological characteristics of COVID-19. Other sources [2,13,20,23] have investigated the best numbers more thoroughly, but for our modeling it is only necessary that the values are reasonable.

5.2 Fitting to SEIR

Next, we fit overall SEIR numbers to case, death, and recovery numbers. We approximated the start of the epidemic in both cities to March 1 and picked STARTING_PERCENTAGE and ISOLATION_COUNTERMEASURE_START to match this. We chose INITIAL_REDUCTION_TIME and FULL_REDUCTION_TIME also based on these numbers. We chose a STARTING_RATIO of 2.5 based on default $R_0$ numbers. We chose a COMMUTER_RATIO of 0.50 based on ridership numbers compared to population [4].

We next sought to fit the total cases to $I(t) + R(t)$ and adjusted the remaining hyper-parameters (CONTACT_MODIFIER, INITIAL_REDUCTION_TGT, FULL_REDUCTION_TGT) to do so. The results are provided in Fig. 2.
5.3 Comparison by Localities

Next we compare the case numbers of our agents against their respective localities. In New York City, we compare the case numbers for a MODZCTA against the infected and recovered numbers for agents representing the MODZCTA. Due to a lack of data, we are only able to compare starting on April 1st ($t = 32$). In Fig. 3 we show a visual comparison of data on April 14.

The value of MAPE ($t \geq 32, Stations \geq 1$) is 0.233 and the scale is from 0.00 to 0.20 cumulative cases/person.
A MAPE of 0.233 is much worse than the overall error, but it still suggests that there is some correlation. It is also a great improvement over simply fitting the default value everywhere. 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.

5.4 Results Summary

The results summary in terms of MAPE is given in Table 2.

|                      | London (by Borough) | New York City (by MODZCTA) |
|----------------------|---------------------|----------------------------|
| Overall MAPE         | 0.056               | 0.071                      |
| MAPE by region (no model) | 0.571             | 0.623                      |
| Median MAPE by region| 0.310               | 0.233                      |
| Mean MAPE by region  | 0.457               | 0.310                      |
| Highest MAPE by region| 1.40               | 0.946                      |

6 Discussion and Further Research

6.1 Additional Parameters

While creating our model, we proposed and experimented with numerous other hyper-parameters which would increase the accuracy of our predictions. However, we excluded them to reduce the specificity and complexity of the model or omitted them due to a lack of reliable data. We list some of the most useful below:

- Awareness Increase Rate - An additional countermeasure modeling increasing public awareness of an outbreak.
- Defiance - The local population will defy government orders if very few people are infected in their area.
- Locality-specific Commuting Modifier - This modifier approximates the percent of the population which commutes in each locality.
- 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.
6.2 Other Modeling Improvements

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.

Next, our research into New York City suggests that there is a significant amount of commuting from outside the 5 boroughs [4], and we haven’t even included Staten Island as there is no subway there. This phenomenon is undoubtedly true for London and all other major cities as well. We have not tried to model these commuters.

Lastly, although we have done similar modeling for Madrid’s commuter rails and the world airline network, we have not formally tried to fit our model to other cities.

7 Conclusion

In this paper, we outlined how to create an agent-based framework to model the spread of diseases through transportation networks. We then built specific models for the spread of diseases through subway networks. Taking into account various factors such as commute time, subway routes, and city-wide countermeasures, we found that our model could fairly accurately predict New York City and London case data.

Our results suggest that city regions with longer commute times and commute routes linked to highly infected areas have a higher prevalence of disease than more central areas. This should be of interest to public health officials looking to allocate resources appropriately during the early days of a new disease outbreak.

References

1. Coronavirus (Covid-19) cases. https://data.london.gov.uk/dataset/coronavirus-covid-19-cases
2. Covid-19: Data. https://www1.nyc.gov/site/doh/covid/covid-19-data.page
3. Infection scenarios for Russian cities. https://story.tutu.ru/scenarii-zarazhenija-gorodov-rossii/
4. 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
5. London shapefile by ward. https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london
6. London underground geographic maps. https://commons.wikimedia.org/wiki/London_Underground_geographic_maps/CSV
7. Mesa documentation. https://mesa.readthedocs.io/en/master/overview.html
8. MTA station data. http://web.mta.info/developers/data/nyct/subway/Stations.csv
9. Who coronavirus information page. https://www.who.int/emergencies/diseases/novel-coronavirus-2019
10. World airport traffic rankings. https://aci.aero/news/2019/03/13/preliminary-world-airport-traffic-rankings-released/

11. Mo, B., et al.: 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

12. Cooley, P., et al.: The role of subway travel in an influenza epidemic: a New York city simulation. J. Urban Health 88(5), 982–995 (2011). https://doi.org/10.1007/s11524-011-9603-4

13. Gonzalez-Reiche, A.S., et al.: Introductions and early spread of SARS-CoV-2 in the New York City area. Science 369, 297–301 (2020). https://doi.org/10.1101/2020.04.08.20056929

14. Goscé, L., Johansson, A.: Analysing the link between public transport use and airborne transmission: mobility and contagion in the London underground. Environ. Health 17(1), 1–11 (2018). https://doi.org/10.1186/s12940-018-0427-5

15. Hagberg, A.A., Schult, D.A., Swart, P.J.: Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th Python in Science Conference (SciPy2008), pp. 11–15. Pasadena, CA USA, August 2008

16. Department of Health (DOHMH): Modified zip code tabulation areas (MOD-ZCTA): NYC open data, May 2020. https://data.cityofnewyork.us/Health/Modified-Zip-Code-Tabulation-Areas-MODZCTA-/pri4-ifjk

17. Hethcote, H.W.: The mathematics of infectious diseases. SIAM Rev. 42(4), 599–653 (2000). https://doi.org/10.1137/s0036144500371907

18. Medicine, I.O., Geller, A.B., Ogawa, V.A., Wallace, R.B.: Assessing the Use of Agent-based Models for Tobacco Regulation. National Academies Press, appendix A: Considerations and Best Practices in Agent-Based Modeling to Inform Policy (2015)

19. Ndaïrou, F., Area, I., Nieto, J.J., Torres, D.F.M.: Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan, April 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7184012/

20. NYCHealth: NYC coronavirus-data repository. https://github.com/nychealth/coronavirus-data

21. Schelling, T.C.: Dynamic models of segregation†. J. Math. Sociol. 1(2), 143–186 (1971). https://doi.org/10.1080/0022250x.1971.9989794

22. Xu, F., Mccluskey, C.C., Cressman, R.: Spatial spread of an epidemic through public transportation systems with a hub. Math. Biosci. 246(1), 164–175 (2013). https://doi.org/10.1016/j.mbs.2013.08.014