Analyzing the Impact of COVID-19 on Demand for Subway Systems

Ilana Zane, David Tian, Srinivas Rajagopalan, Daniel Pattathil

May 12, 2020

1 Abstract

Subway systems are usually complicated, far reaching networks that are constantly in use. However, during an epidemic, people are encouraged to stay home and practice social distancing in order to reduce their chance of being infected and infecting others. It is common to see that when these measures are encouraged, there is a reduced demand for public transportation because most people are no longer traveling or going to work. It is important that a subway system can be adjusted for epidemics to prevent the spread of a virus and while still operating with reduced demand in order to maintain essential services.

In this paper we explore the demand of subway systems in large metropolitan areas during epidemics and determine how much a virus contributes to the demand for the subway system, how to respond to a changing demand, and predict the best way to alter the subway in order to prevent transmission. We are primarily focused on studying the effects of COVID-19 on the New York City subway system.

Previous work has been done that models the changing demand for subway systems in past epidemics. However, none of these models address the current COVID-19 pandemic. Our main goal is to analyze how the demand for the New York City subway system is changing not only through regular weekly fluctuations, but also through the pandemic. We take into account a number of external factors such as economic status, modeled using the number of unemployment returns, and the number of confirmed COVID-19 cases. We utilize several different data sources and modeling techniques in order to model the changing demand for the NYC subway as a result of these external factors.

2 Introduction

Public transportation systems are critical to the smooth, efficient functioning of a city. Subway systems in particular move vast amounts of people to and from work every day and play a major role in their city's economic growth and development. However, in a large-scale epidemic setting, such as the recent outbreak of COVID-19, people are encouraged to practice social distancing and remain home from work in order to contain the virus and prevent further spreading. During this time, travel is usually highly discouraged, which will likely decrease what is usually a high demand for public transportation. As usage of the subway system decreases, the operator cannot afford to run the system in the same way that it used to with full capacity. However, the system will still need to be kept running in order to provide essential services, such as groceries and healthcare, to residents of the city.

As one of the largest subway systems in the world, the New York City subway is currently faced with this dilemma. It is also one of the few systems to provide overnight services. On average, the NYC subway carries about 5 million people everyday, so it is important to predict how to keep such an important part of the city's infrastructure functioning efficiently. The system will have to reduce services in order to discourage travel and to follow guidelines set by the state. In these conditions, certain stations will be closed, others will not run as often, and the Metropolitan Transit Authority (MTA) will stop running trains after a certain time at night. Demand and the number of running cars also need to be reduced so that the cars can be disinfected.

All of these changes are implemented in order to restrict travel and the spread of the virus

amongst people, but essential workers are heavily reliant on the subway system. It is important to model the usage of the subway system not only to make sure that essential workers have access to transportation, but to determine which employees of MTA can be sent home, how often trains need to run and how often train cars need to be disinfected. In order to make an informed decision about which lines need to have reduced service or which stations need to be closed, we need to analyze which subway lines travel through which areas, the percentage of people who are essential workers in that area, overall socioeconomic status of that area, and the number of confirmed COVID-19 cases in that area. With this information we can draw conclusions about which stations or lines are most likely to contribute to the spread of COVID-19.

Past research has modeled the spread of viruses in cities by analyzing human mobility, but not through public transportation. Railway cars and train stations are much more crowded and confined than regular city blocks and other modes of transportation, and many passengers will come into contact with the same surfaces, such as handrails and seats. These factors lead to a high rate of contamination amongst passengers. Other research has focused on how the viruses spread throughout subways using historical data from the 1957 influenza pandemic in NYC. Other viruses such as the coronavirus might behave differently in terms of transmission rates and methods of contamination, so the data used may not accurately reflect the current situation.

Our goal in this paper is to evaluate the current usage of subway systems in cities and monitor how it changes over time when that geographical area is confronted by an epidemic. We begin by monitoring the ridership of the NYC transit system before the epidemic affected the area in order to get a baseline and to predict which lines are most likely to contribute to sickness. We use the results to predict the subway demand in the near future, which stations to close or limit, and ultimately predict the best way to modify the system in order to reduce the spread of the virus amongst travelers. We consider different models that can help us visualize the spread of viruses throughout a subway system. In New York City for example, we can use publicly available data to take into the account the ridership on each line and their cleanliness. To track busiest stations we analyze turnstile data from the MTA. We also analyze which neighborhoods of the city are most at risk of containing the virus based on density, socioeconomic status, and proximity to healthcare. In the following sections, our paper will discuss related work in the field of epidemiology and public transportation, the motivation behind our design, our main design and implementation, and an evaluation of our results.

3 Related Work

Plenty of previous work has been done in the realm of infectious diseases, considering that it is such an important topic for human well-being. Researchers have modeled the spread of various infectious diseases spatially using statistical methods [4, 5]. Unsurprisingly, public transportation use in the five days prior to the onset of a virus corresponds to an almost sixfold increased risk in consulting a doctor for acute respiratory infection [10].

Contact tracing is a method that has been explored to mitigate this. It identifies people who have been in close contact with a known infected person, so that proper action can be taken to reduce the chance of the disease spreading further [1,2]. This works well for flights, where all passengers are asked for identification and seats are usually assigned. If the disease in question has a low risk of transmission, such as tuberculosis or MERS, it is not difficult to alert the passengers who sat near the infected individual. However, contact tracing is an extreme measure that becomes impractical and inefficient at large scales and is usually only effective for diseases with a long latency and low reproductive ratio, such as sexually transmitted diseases. It is not a viable solution in massive ground-based public transportation systems where many millions of people come into contact every day [9].

There has also been work done on predicting the volume of commuters in a public transportation system. For example, machine learning has been applied to infer inflow and outflow in metros based on spatio-temporal factors [6,7]. Other research also takes into account the clustering of similar stations and anomalies such as special events [8]. We found only a few works that focused on the intersection of public transportation and infectious disease spread. Most are rather broad airline based models for a variety of disease types [3]. We were only able to

find one article that focused specifically on the role of the New York subway system in the influenza epidemic of 1957-1958 [11]. This work is helpful in establishing a model to demonstrate how a virus can travel through a subway, but it does not incorporate data from the recent outbreak of COVID-19.

4 Motivation

The main motivation for this proposal comes from the sudden decline in ridership that struck the New York City subway system after quarantining measures were put in place during the COVID-19 pandemic. The MTA was unprepared and now faces a large loss in revenue, which could negatively impact the system for years to come. These happenings revealed a need to accurately infer the effect that an epidemic has on the demand for a mass transit system, particularly the New York City subway. This form of research would help administrative bodies be able to clearly assess the need for closure of certain parts of the system and would be more widely applicable to different urban areas.

This issue is important to explore because there is a lack of data and research that analyzes this specific situation. As the density of urban areas around the globe increases, making pandemics more and more common, it is important that a versatile model is created in order to predict how a large-scale disease will affect millions of people.

Many epidemiologists have done previous work on the spread of contagious disease through various forms of transportation. In addition, Researchers studying smart cities have recognized that subway systems are popular forms of transportation in urban areas, so there is plenty of published literature on the subject. However, there is a lack in literature concerning the crossover between viruses and subway systems. We hope that our work will be able to help bridge that gap and inspire future work on the subject.

5 Main Design

In the next section we will go over our main design for modeling ridership on the New York City subway during the COVID-19 epidemic.

5.1 Modeling Ridership

Before creating our model, we had already assumed that ridership would be much lower for recent weeks compared to the beginning of the 2020 year. In order to validate this assumption, we made visualizations of subway ridership from January 2020 through April 2020. Data was retrieved from the MTA, which publishes turnstile entry and exit counts at every station, every four hours. Unfortunately, the data is not well formatted and required extensive pre-processing before we were able to gather any insights with it.

First, the data appeared to be incomplete for stations outside of the standard numbered and lettered lines. Therefore, we decided to leave out the Staten Island Railway, the Roosevelt Island Tramway, and the Path lines connecting New Jersey to the city. This left us with about 454 subway stations in Manhattan, Brooklyn, Queens, and the Bronx. In addition, since turnstile counts are cumulative, we needed to compute the difference between counts at every time interval. Turnstiles sometimes rolled over, so we filtered out any records with a difference greater than 10,000. Occasionally, some turnstiles would also count backward; to account for this, we took the absolute value of any difference greater than -5,000 and filtered out the rest. Since the actual number of entries and exits could not be recovered for that time interval, some counts are under counted. Fortunately, very few records were discarded. The final data set was aggregated into entry and exit counts for each station on every day.

Figure 1 shows ridership on Friday, February 7, 2020, before the pandemic had caused any widespread shutdowns. Each red circle represents a station; its radius is proportional to the number of entries at that station. We saw that ridership fluctuated consistently between weekdays and weekends, but suddenly dropped toward the end of March, when social distancing and work-at-home practices were put into place to prevent the spread of the virus. Figure 2 shows the reduced ridership on Friday, April 24, 2020. From the visualizations, we can see that ridership drastically decreased in almost all stations across each borough of the city.

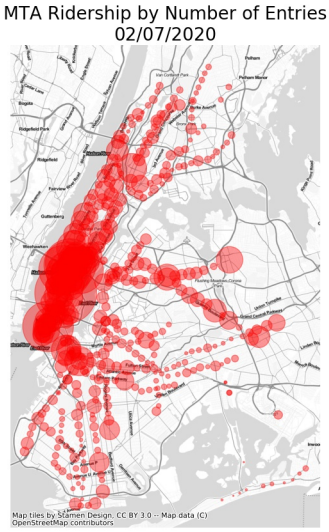

Figure 1: High volume of ridership before the epidemic

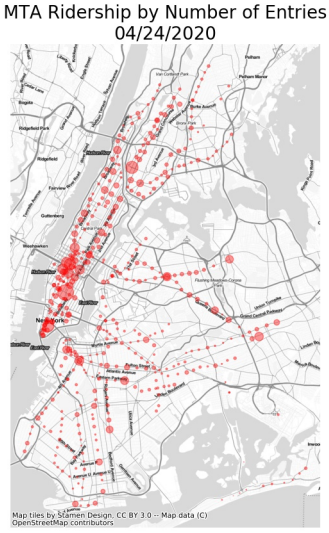

Figure 2: Low volume of ridership during the epidemic

5.2 Pre-epidemic Ridership

In order to get a better idea of the change in ridership due to the pandemic, we decided to first try to create a model to predict the demand in a normal situation, i.e. if the pandemic had not occurred. To do so we gathered and cleaned data from 2019 and used a single-layer neural network to try to predict ridership at each station based on a one-hot encoded value for the day of the week, and external factors such as weather. Unfortunately, this did not produce accurate results as we hoped, since we had assumed that the weekday-weekend cycle was the main contributor to ridership fluctuations. This was the case for large transfer hubs like Grand Central station, as shown in Figure 3.

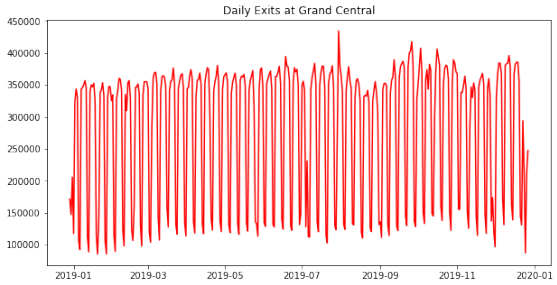

Figure 3: 2019 Ridership at Grand Central Station

However, when we investigated closer, we found that a good number of stations did not behave in such a regular pattern, especially those located near specific attractions. For example, the Mets-Willets Point station, located inside Flushing Meadows Corona Park, is the closest station to Citi Field Stadium and the Billie Jean King National Tennis Center. Exits at this station peak sharply on weekends during the warmer months, as shown in Figure Y.

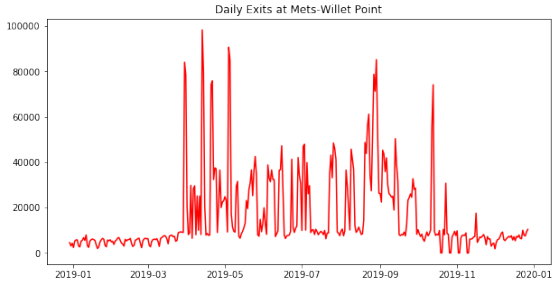

Figure 4: 2019 Ridership at Mets-Willets Point

It became evident that we would be unable to predict ridership for these situations, without knowing the expected attendance details for every nearby event taking place. Moreover, since none of these events would occur due to social distancing measures, we reasoned it would be safe to use median weekday and weekend ridership counts as our baseline. In essence, our model looks at the decline in ridership as if these events never happened.

5.3 Socioeconomic Factors

When looking at data regarding ridership rates and levels throughout the New York City metro

system, it is also crucial to consider external factors relating the changing rates and ways to isolate variables with heavier impact on ridership persistence through the pandemic. In order to gain a deeper understanding of the demographics, US Census Tract data was utilized in conjunction with the aforementioned MTA turnstile data to gain a general understanding of the socioeconomic status at various subway stations.

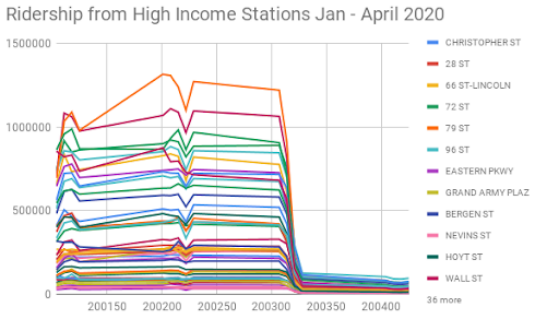

Figure 5: High-income ridership dropped sharply

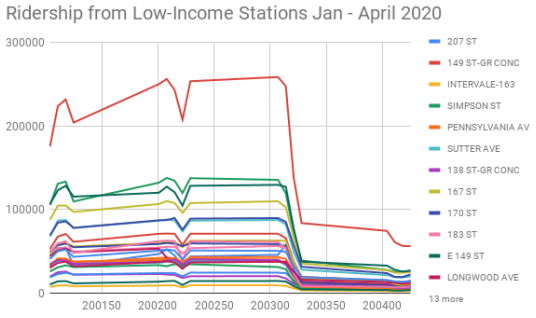

Figure 6: Low-income ridership dropped less

Figures 5 and 6 above denote the change in net ridership throughout different subway lines categorized by income level. The first figure is related to stations with the median income level over $60,000 and the new fraction remaining in subway usage was at most one tenth of what it previously had been when looking at the most frequented stations. This is a drastic difference compared to the ridership among low income stations where the economic level is estimated to be at less than $32,000 for the baseline poverty level established by New York City guidelines. Many of the changes in ridership were much closer to one third of what they had been, with some stations maintaining as many as half of their previous turnstile counts coming close to one third, a noticeable difference when compared to their higher income peers. As a result, there was a need to investigate further on the aspects of these stations that may require further mobility for the nearby residents. When looking at the data for ridership shifts over a three month period, this trend was much more visible.

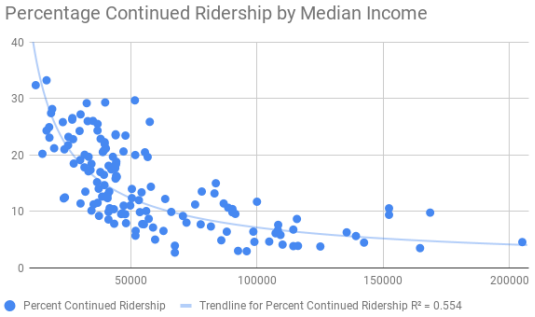

Figure 7: Ridership decline is correlated with median income

Figure 7 describes the relationship between subway station, median income and the percentage of remaining ridership between January 18th and April 18th. The choice of remaining ridership was made to show the relative difference in usage between stops. In addition, the data points that had above 20% continued ridership were entirely in the Bronx, Queens, and Brooklyn, with the exception of one borderline Manhattan stop.

This may be connected to the level of essential workers that are residents or employed near these subway stations and further research into employment data linked to socioeconomic status could clarify the sustained subway utilization in these areas.

5.4 Modeling Ridership

Our primary modeling problem was to investigate the number of exits made at New York City train stations, after the pandemic. On March 25th, certain New York City subway lines were cut as part of the city's "NY Essential Service Plan". We were interested in modeling the number of exits from a train station based on the number of coronavirus cases, and other factors that could relate to the overall pandemic. For example, other researchers have made claims that coronavirus cases are linked to poorer areas. Therefore, using data from the 2017 US government taxation report on New York City,

we gathered data on the number of unemployment returns in a given zip code area.

Our final data set is as follows: every row consists of a data, a station name, its zip code, the number of coronavirus cases in that zip code from a specific date, the number of unemployment returns filed in that zip code, and finally a 0 or 1 encoding to describe whether or not the train station lies on a subway line that has been closed due to the pandemic. We gathered the subway station data from the MTA, the unemployment returns from the US government, published in 2017, and the line factor data from news articles. Due to the statistics of our model, our paper does not focus on developing predictive strategies, but instead looks at why this problem is difficult and potentially significant factors. We seek to make claims based on the relationship of public transit systems and factors during a pandemic, such as the COVID-19 pandemic.

We are interested in modeling the subway system as being affected by a virus that encompasses a large metropolitan area. As of now there is not an abundance of research that models this. We found that the best way to model our data is through a linear regression model and create visualizations to see how subway demand is impacted. The MTA's historical and current data on entrances and exits from turnstile data help us determine how popular a station is. We also consider data on socioeconomic status, percentage of essential workers and number of confirmed COVID-19 cases in each borough of NYC.

6 Constraints

There were certain limitations and constraints that we considered while undertaking this project. Our model is made specifically for this subway system and for this virus, which means that our model may only be able to serve as a basis or outline for other cities and viruses. Another limitation we have encountered is access to data. We initially wanted to create a human mobility model that accurately shows the flow of commuters throughout the system, but the MTA only provides data that has low granularity. The data consists only of turnstile entrance and exit counts, which tells us the volume of commuters, but not their exact movements. Another constraint was the access to information on COVID-19. Research is constantly changing on a day to day basis and researchers/scientists are not completely aware of how the disease is transmitted (i.e. if it lingers in the air or how long it can survive on surfaces). The methods used for counting infected patients by zip code are also inconsistent.

We encountered several constraints when modeling the relationship between subway exits, confirmed cases, and unemployment. Even though we were able to create a model, we were unable to make any predictive conclusions. Before doing any model corrections, we had to remove 955 upper outliers from the dataframe, which we defined as an exits count that is greater than Q3 + 1.5*IQR, where Q3 is defined as the 75% quantile and IQR is defined as the interquartile range. For example, we see in Figure 8 that the histogram of the distribution of exits from train stations is distinctly right-skewed.

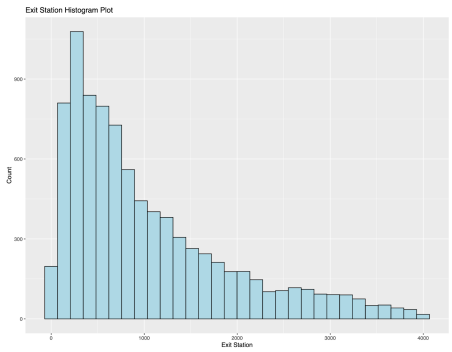

Figure 8: Exit Station Histogram Plot

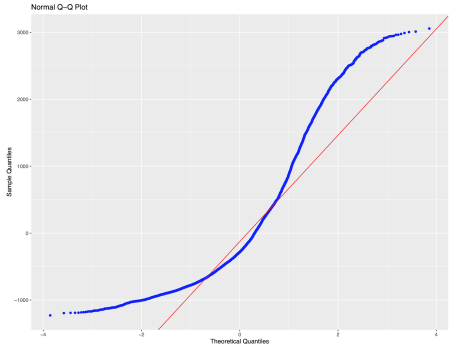

Figure 9: Normal Q-Q Plot

Because of this, we were presented with some potential conflicts when using linear regression. For example, we generated plots for the residuals of the linear regression model, as well as Q-Q plots for the linear regression model, which clearly show that the residuals are not distributed normally and that the Q-Q plot does

not show normality as well. These plots can be seen in Figures 9 and 10.

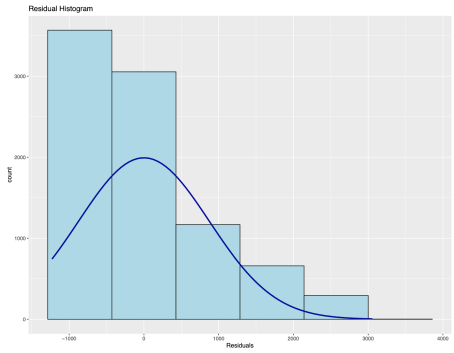

Figure 10: Residual Histogram

We see that the residual plot and Q-Q plot indicate that linear regression does not work suitably in this regard. We additionally applied log transform and square root transform, as well as more complicated models such as random forest. None of these models made improvements when comparing the R^2 values. We additionally tried Poisson regression, as we thought that the output variable counts data. Unfortunately, this also resulted in no improvement in R^2.

7 Performance Evaluation

While the R^2 value for our linear regression model was low, we do see several positives with our model. None of the 95% confidence intervals for the predictor coefficients cross 0. Additionally, the overall model significance for our linear regression model is highly significant (p-value ¡¡¡ 0.001), with coronavirus and the closed subway line binary predictor also having statistically significant p-values (p-value ¡¡¡ 0.001), while unemployment did not. While we cannot use this as a predictive approach due to its poor R^2, we are able to generate inferences based on this model. Additionally, we computed the Variance Inflation Factor for all of the predictors, and none of them were greater than 10. Additionally, none of them were different than the others, validating that our model did not suffer from multicollinearity. We additionally saw moderate correlation between Unemployment and coronavirus cases (R = 0.51).

8 Conclusion

While our model's R^2 value is not great, we believe that this is due to the underlying nature of the data we observed, and due to the relatively few dates that have gone through in this pandemic. While March 25th represented a change in transportation policy, many other companies had already begun shifting their work policies. Additionally, because of early problems with COVID-19 testing, the case numbers might not be accurate, which impacts our model. However, due to the highly significant p-value of the predictors and of the model, we deduce that the predictors of coronavirus cases and closed subway lines do have significant impact on the number of people using the train station.

We also found that between January and April, ridership dropped for two different economic classes- one which made more than $60,000 and the other that made less than $32,000. However, amongst subway riders from the first income group, only one tenth of them continued to ride the subway, while almost one half to one third of subway riders from the poorer group continued to ride the subway. Stations that maintained more than 20% of their original ridership were located in boroughs that have high rates of poverty. This means that there is a correlation between sustained ridership during a pandemic and groups of people from lower economic status, which pertains to most essential workers.

Through our observations we can conclude that there is a direct correlation between levels of ridership, economic status, and the number of confirmed cases. Given more accurate and descriptive data, we could further expand our model to definitively predict which stations need to remain open in order to accommodate essential workers. When the NYC economy slowly reopens in the future, it will also be important to predict which stations will be most used by civilians in order to determine station capacity, and further determine which guidelines need to be implemented in stations to prevent another outbreak.

References

[1] Eames, K. T. D., & Keeling, M. J. (2003). Contact tracing and disease control. Proceedings of the Royal Society of London. Series B: Biological Sciences, 270(1533), 2565–2571. https://doi.org/10.1098/rspb.2003.2554

[2] Kiss, I. Z., Green, D. M., & Kao, R. R. (2005). Disease contact tracing in random and clustered networks. Proceedings of the Royal Society B: Biological Sciences, 272(1570), 1407–1414. https://doi.org/10.1098/rspb.2005.3092

[3] Tatem, A. J., Rogers, D. J., & Hay, S. I. (2006). Global Transport Networks and Infectious Disease Spread. In Advances in Parasitology (pp. 293–343). Elsevier. https://doi.org/10.1016/s0065-308x(05)62009-x

[4] Riley, S. (2007). Large-Scale Spatial-Transmission Models of Infectious Disease. Science, 316(5829), 1298–1301. https://doi.org/10.1126/science.1134695

[5] Meyer, S., & Held, L. (2014). Power-law models for infectious disease spread. The Annals of Applied Statistics, 8(3), 1612–1639. https://doi.org/10.1214/14-aoas743

[6] Xie, P., Li, T., Liu, J., Du, S., Yang, X., & Zhang, J. (2020). Urban flow prediction from spatiotemporal data using machine learning: A survey. Information Fusion, 59, 1–12. https://doi.org/10.1016/j.inffus.2020.01.002

[7] Liu, Y., Liu, Z., & Jia, R. (2019). DeepPF: A deep learning based architecture for metro passenger flow prediction. Transportation Research Part C: Emerging Technologies, 101, 18–34. https://doi.org/10.1016/j.trc.2019.01.027

[8] Noursalehi, P., Koutsopoulos, H. N., & Zhao, J. (2018). Real time transit demand prediction capturing station interactions and impact of special events. Transportation Research Part C: Emerging Technologies, 97, 277–300. https://doi.org/10.1016/j.trc.2018.10.023

[9] Mohr O, Askar M, Schink S, Eckmanns T, Krause G, Poggensee G. Evidence for airborne infectious disease transmission in public ground transport – a literature review. Euro Surveill. 2012;17(35):pii=20255. https://doi.org/10.2807/ese.17.35.20255-en

[10] Troko, J., Myles, P., Gibson, J., Hashim, A., Enstone, J., Kingdon, S., Packham, C., Amin, S., Hayward, A., & Van-Tam, J. N. (2011). Is public transport a risk factor for acute respiratory infection? BMC Infectious Diseases, 11(1). https://doi.org/10.1186/1471-2334-11-16

[11] Cooley, P., Brown, S., Cajka, J., Chasteen, B., Ganapathi, L., Grefenstette, J., Hollingsworth, C. R., Lee, B. Y., Levine, B., Wheaton, W. D., & Wagener, D. K. (2011). The Role of Subway Travel in an Influenza Epidemic: A New York City Simulation. Journal of Urban Health, 88(5), 982–995. https://doi.org/10.1007/s11524-011-9603-4

[12] Goscé, L., & Johansson, A. (2018). Analysing the link between public transport use and airborne transmission: mobility and contagion in the London underground. Environmental Health, 17(1).