Predicting Public Transportation Load to Estimate the Probability of Social Distancing Violations

Juan Martinez[1], Ayan Mukhopadhyay[1], Afiya Ayman[2], Michael Wilbur[1], Philip Pugliese[3], Dan Freudberg[4], Jonathan Gilligan[1], Aron Laszka[2], and Abhishek Dubey[1].

[1]Vanderbilt University, Department of Electrical Engineering and Computer Science
[2]University of Houston, Department of Electrical Engineering and Computer Science
[3]Chattanooga Area Regional Transportation Authority (CARTA)
[4]Nashville Metropolitan Transit Authority

Abstract

Public transit agencies struggle to maintain transit accessibility with reduced resources, unreliable ridership data, reduced vehicle capacities due to social distancing, and reduced services due to driver unavailability. In collaboration with transit agencies from two large metropolitan areas in the USA, we are designing novel approaches for addressing the aforementioned challenges by collecting accurate real-time ridership data, providing guidance to commuters, and performing operational optimization for public transit. We estimate ridership data using historical automated passenger counting data, conditional on a set of relevant determinants. Accurate ridership forecasting is essential to optimize the public transit schedule, which is necessary to improve current fixed lines with on-demand transit. Also, passenger crowding has been a problem for public transportation since it deteriorates passengers' wellbeing and satisfaction. During the COVID-19 pandemic, passenger crowding has gained importance since it represents a risk for social distancing violations. Therefore, we are creating optimization models to ensure that social distancing norms can be adequately followed while ensuring that the total demand for transit is met. We will then use accurate forecasts for operational optimization that includes *(a)* proactive fixed-line schedule optimization based on predicted demand, *(b)* dispatch of on-demand micro-transit, prioritizing at-risk populations, and *(c)* allocation of vehicles to transit and cargo trips, considering exigent vehicle maintenance requirements (*i.e.*, disinfection). Finally, this paper presents some initial results from our project regarding the estimation of ridership in public transit

Introduction

Public transit provides important services that enable residents of a city, especially those without personal cars, to commute to work and access essential services (Lao et al. 2016). Cities strive to maximize the coverage of transportation services under budgetary constraints. The novel coronavirus disease (COVID-19) pandemic and subsequent mitigation efforts have not only disrupted the lives of millions but also created powerful new challenges to public transit systems (Tirachini 2020).

Higher population density and connectivity can also aid the spread of pandemics (Olmo and Sanso-Navarro 2020; Medo 2020). One of the most effective measures for slowing down or stopping the spread of a contagious disease is *social distancing*, that is, reducing the number of times that people come into close contact with each other. To reduce contact between passengers and drivers, agencies are switching to fare-free operation on buses and forbidding front-door boarding; and to reduce contact between passengers, they are limiting the passenger capacity of their vehicles. In some cases, the agencies are also reducing the frequency of the fixed route service as ridership has declined. While these ad-hoc changes are reducing the number of close contacts between passengers, they are also affecting the people that have higher socio-economic vulnerabilities (Brough, Freedman, and Phillips 2020). Indeed, the effects of COVID-19 on public transit ridership are not uniform across demographic, spatial and temporal variations (Hu and Chen 2021).

Wilbur et al. (2020) found that public transport ridership in Nashville and Chattanooga, TN, dropped by 66.9% in the spring of 2020 and despite a partial recovery, ridership in July remained 48.4% below normal levels. This severely reduced revenues, and transit agencies have struggled to provide adequate service. Wilbur et al. found that the drop in ridership was higher on weekdays (57%) compared to weekends (44%), which was likely due to a combination of rising unemployment and a transition to remote work by many employers. This analysis also observed a positive correlation between the change in ridership in different neighborhoods and the economic characteristics of those neighborhoods (income and home value): ridership dropped by 77% in the highest-income neighborhoods, compared to 58% in the lowest-income neighborhoods.

Currently, most transit agencies make their planning decisions *ad hoc* and myopic. Some agencies have started partnering with software developers to provide passengers with crowd-sourced occupancy information via smart phone applications, enabling the commuters to take informed decisions (Darsena et al. 2020; Couture et al. 2020). Even then, the decision on service changes and frequency and capacity controls remain *ad hoc*, awhich leaves passengers to plan their travel under considerable uncertainty.

Considerable effort has gone into modeling the spread of COVID-19 and these models are widely used to inform public-health decision-making (Whitelaw et al. 2020; Wang, Ng, and Brook 2020; Bouffanais and Lim 2020), but most of these models have not been applied to planning by transit agencies, nor to helping transit users to plan safe travel during the pandemic (Lee and Lee 2020).

We are working with transit agencies in two large metropolitan areas in the USA to address these shortcomings by analyzing *(a)* how residents perceive public transit during the pandemic, *(b)* how transit agencies can ensure an adequate level of service with a combination of fixed-line and on-demand transit, *(c)* how high-fidelity statistical models can be developed using ridership data and applied to optimizing transit, and *(d)* how allocation of transit vehicles can be optimized for passenger and cargo trips, accounting for exigent vehicle maintenance requirements (e.g., disinfection).

Prior work has explored several different aspects of forecasting transit ridership. For example, Karnberger and Antoniou (Karnberger and Antoniou 2020) provide an insight into the relationship between public transit ridership and spatio-temporal influences from exogenous events. Zhou et al. (Zhou et al. 2017) explore the influence of daily weather condition changes on the usage of public transit. There has also been work that attempts to predict passenger occupancy on public transportation in the near future by using real-time information from smart cards (Van Oort, Brands, and de Romph 2015; Nuzzolo et al. 2013). In this paper, we create statistical models to forecast ridership based on automated passenger counting (APC) and transit data from the General Transit Feed Specification (GTFS). Further, we consider and assess the uncertainty of the input data by exploring different probability distributions.

Fixed Line Transit Model

Our model describes a transit agency that operates a set of buses $\mathcal{V}$. Each bus $\nu \in \mathcal{V}$ follows a fixed line route $h \in \mathcal{H}$, where $\mathcal{H}$ is the set of all possible routes in operation. If $\mathcal{S}$ represents the set of all bus stops and garages, then any directed graph connecting a subset of $\mathcal{S}$ is a possible bus route, and $\mathcal{H}$ represents the set of all possible directed graphs over $\mathcal{S}$. Note that two graphs traversing the same nodes in different directions represent two different possible bus routes. We represent the set of all daily trips in the agency's master schedule by $\mathcal{T}$.

Each trip $t \in \mathcal{T}$ follows a particular route $h \in \mathcal{H}$. For each trip, the arrival time $\mathbf{t}_s^{arrival}$ of the bus at a stop $s \in \mathcal{S}$ serving the trip is retrieved from the GTFS feed. We define the number of passengers boarding the bus in trip $t \in \mathcal{T}$ at stop s_i as $\gamma_t(s_i)$. As an example, Figure 1 shows the registered board counts for the busiest route in one of the cities under consideration from January to June.

The number of passengers getting off at s_i during trip $t \in \mathcal{T}$ is denoted by $\alpha_t(s_i)$. Automated passenger-counting

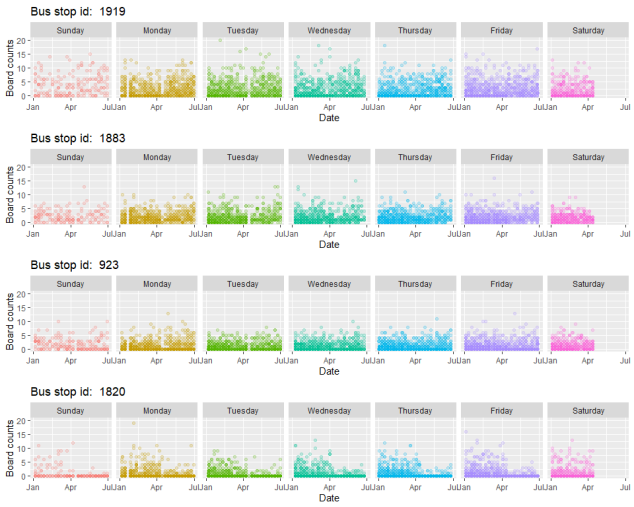

Figure 1: Overall board counts in one of the cities of interest for the 4 busiest bus stops on the busiest route.

devices attempt to count the number of passengers boarding and leaving the bus at each stop. These devices are known to be imperfect, so the data on passenger boarding and exiting that our partner agencies provide has errors that our model must account for.

We model the occupancy of a bus as a random integer L, which denotes the total number of passengers inside that bus at a given point in time. The occupancy on trip t at location s_i, denoted by $L_t(s_i)$, can be calculated from models of the boarding and alighting processes. We model occupancy as an autoregressive process:

$$L_t(s_i) = L_t(s_{i-1}) + \gamma_t(s_i) - \alpha_t(s_i) \tag{1}$$

It is important to note that our data does not directly reveal the ground truth demand. A way to estimate demand based on occupancy data is to treat it as a latent variable. We can then estimate demand based on the observed board counts. For example, we can model demand by considering order statistics of the board counts, such as the maximum or 90*th* percentile. For this paper, we focus on predicting the average board and alight counts per hour of the day for each trip at each stop, given ridership information of all the previous trips.

Formally, we denote probability distributions F_b and F_a over board and alight counts respectively. Our goal is to learn $F_b(\gamma_t(s_i) \mid w)$ and $F_a(\alpha_t(s_i) \mid w)$, where w is a set of features that characterize conditions at location s_i at time t. For example, w can include weather, time of day, and ridership data from previous trips. Our model starts from a predicted demand at the origin of the trip (we use the predicted board count at the origin as a proxy for this demand). As the bus follows its route, we sample from the distributions F_b and F_a to generate board and alight counts at every stop, thereby simulating an entire trip made by a vehicle. Note that F_b and F_a are not constant, but vary with

time of day, bus stop, weather, occupancy of the bus, etc. In particular, F_a is constrained not to allow more passengers to leave the bus than are currently aboard.

We use four statistical models, namely the Poisson, negative binomial, zero-inflated Poisson, and the zero-inflated negative binomial models to estimate the distribution of board and alight counts. For the sake of brevity, we refrain from presenting the probability density functions of all the models in this paper. The Poisson distribution is one of the most widely used approaches for modeling count data (Menon and Lee 2017). Each event (a single passenger boarding a bus given a trip and a stop) is considered a result of an independent Bernoulli trial. As the number of trials increases and the probability of success decreases, the count of the number of successes (total passengers boarding or alighting) takes the form of a Poisson distribution. A shortcoming of the Poisson model is that it assumes that the variance and mean of the distribution is the same. The negative binomial model (essentially a hierarchical Poisson model where the mean parameter follows a gamma distribution) has been shown to be more flexible (Mukhopadhyay et al. 2020). As with many real world datasets that model counts of events, ridership data consists of many zeros, i.e. during many trips and at many stops, no passengers board or get down from the vehicles. Zero-inflated models can easily handle excess zeros that cannot be explained by standard count-based models by considering a separate *state* in the underlying statistical process (Mukhopadhyay et al. 2020).

Experiments

We use APC data for board counts and alight counts from two large metropolitan areas of USA. Information about arrival times, routes, and stops were retrieved from the GTFS feed generated by the transit agencies of the cities. We consider route direction, total board counts, and occupancy as selection criteria to apply our models. Then, we focus on the bus stops with the highest board counts.

We used data from January 2020 to May 2020 for training the models, and data from June 2020 as the test set. Our data consists of a total number of 71,765 boardings and 71,971 alight counts. Naturally, the total number of passengers getting down from buses cannot be greater than the number of people that board, but this reflects that the data is noisy.

As a preliminary analysis, we tested the dependence of the board and alight counts on time of day. We considered each hour as our unit of discretization to compute the hourly average for board and alight counts. While prior work has shown that clustering spatial locations together can balance spatial heterogeneity and model variance (Mukhopadhyay and Vorobeychik 2017), we learn a separate regression model for each combination of trip and bus stop in order to capture the significant variation in ridership across trips.

Moreover, building regression models for these parameters allow us to capture the temporal heterogeneity of the board and alight count rates.

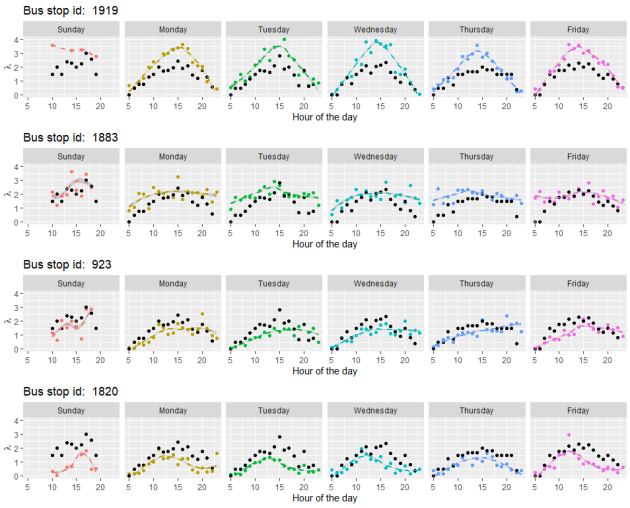

Figure 2: Actual and predicted average board counts per hour. The line in each plot represents a local polynomial regression. Black dots represent the predicted hourly averages (data from our partners show an overwhelmingly large number of missing values on Saturdays. As a result, we learned models only for the other six days of the week.)

Table 1: Model Performances for Board Counts

Model	AIC Score	$-log(\mathcal{L})$ (train)	RMSE (test)
Poisson	2071.62	-1018.56	2.1625
Neg-Bin	**1785.58**	**-879.15**	2.161
Zero-Inflated Poisson	1836.42	-883.71	**2.1564**
Zero-Inflated Neg-Bin	1883.49	-904.49	2.174

Table 2: Model Performances for Alight Counts

Model	AIC Score	$-log(\mathcal{L})$ (train)	RMSE (test)
Poisson	867.71	-416.60	1.261
Neg-Bin	1016.19	-492.45	1.289
Zero-Inflated Poisson	797.62	-364.31	**0.951**
Zero-Inflated Neg-Bin	**789.73**	**-359.36**	0.960

We see that all the statistical models can predict board and alight count averages fairly accurately with only hour of the day as a regressor. Also, zero-inflated models have the best performance (in terms of predicted error). This behavior is expected due to the prevalence of zero counts in our dataset. Moreover, the variability of the alight counts is better captured by the zero-inflated models in both the training and testing sets. However, the AIC Score and log-likelihood indicate that the Negative Binomial regression model has a better performance on the training set. This suggests that high overdispertion is also present for low board count rates. We also show plots of the actual and predicted board

and alight counts in Figure 2 and Figure 3 respectively.

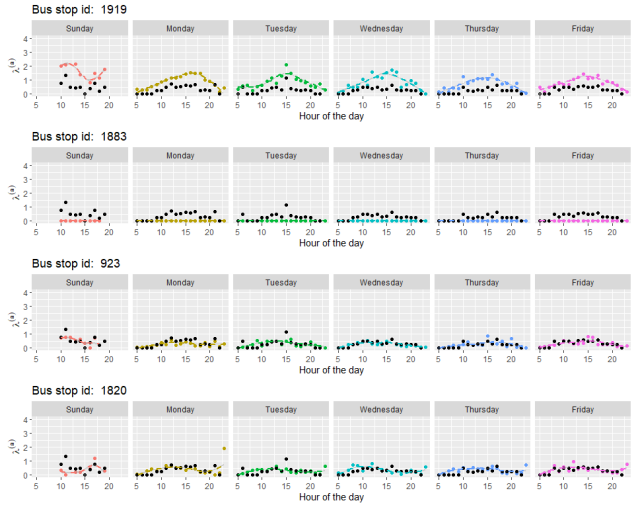

Figure 3: Actual and predicted average alight counts per hour. The line in each plot represents a local polynomial regression. Black dots represent the predicted hourly averages.

We observe in the Figures 2 and 3 that the predictive models can capture the overall trend of the observed data. However, there are certainly patterns that the models do not identify. As an example, with fewer trips and routes being active during the weekends (particularly Sundays), the models struggle to generalize to future predictions. We also observe that in the busiest stop, the model under-predicts board counts. As a data-driven research, we are currently incorporating features like weather, broadening our hypothesis class by using richer representations, and performing spatio-temporal clustering to improve the predictive performance of our models.

Conclusion

We are working with transit agencies of two metropolitan areas in the USA to estimate ridership patterns in order to better meet the demands of the residents and maintain safety during the pandemic. Our approach combines APC data and GTFS feed from the transit agencies to build ridership models. Our initial results show that simple statistical models can be useful in estimating average board and alight counts. These results suggest that these simple but efficient models can provide sensible information to predict board and alight counts defining the hour of the day as the time horizon. Moreover, we are exploring the effects of changing the time unit discretization on the accuracy of the models. This analysis would allow us to provide relevant information to the users. That is, we could estimate the probability of social distancing violations of a particular trip between any pair of consecutive bus stops for different time horizons.

Also, we are currently exploring how artificial neural networks and additional features can improve the performance of our models. Our predictive models can be used to optimize the schedule of public transit, and possibly replace (some) fixed lines with on-demand transit. Further, we are creating optimization models to ensure that social distancing norms can be adequately followed while ensuring that the total demand for transit is met.

Acknowledgement

This material is based upon work supported by the Department of Energy, Office of Energy Efficiency and Renewable Energy (EERE), under Award Number DE-EE0008467 and National Science Foundation through award numbers 1818901, 1952011, 2029950 and 2029952. The authors will also like to acknowledge the computation resources provided by the Research Computing Data Core at the University of Houston and through cloud research credits provided by Google.

References

Bouffanais, R.; and Lim, S. S. 2020. Cities — try to predict superspreading hotspots for COVID-19. *Nature* 583(7816): 352–355. Number: 7816 Publisher: Nature Publishing Group.

Brough, R.; Freedman, M.; and Phillips, D. 2020. Understanding Socioeconomic Disparities in Travel Behavior during the COVID-19 Pandemic. *SSRN Electronic Journal* ISSN 1556-5068.

Couture, V.; Dingel, J. I.; Green, A.; Handbury, J.; and Williams, K. 2020. Measuring movement and social contact with smartphone data: a real-time application to COVID-19. *NBER Working Paper* (w27560).

Darsena, D.; Gelli, G.; Iudice, I.; and Verde, F. 2020. Safe and Reliable Public Transportation Systems (SALUTARY) in the COVID-19 pandemic. *arXiv preprint arXiv:2009.12619* .

Hu, S.; and Chen, P. 2021. Who left riding transit? Examining socioeconomic disparities in the impact of COVID-19 on ridership. *Transportation Research Part D: Transport and Environment* 90: 102654. ISSN 13619209.

Karnberger, S.; and Antoniou, C. 2020. Network–wide prediction of public transportation ridership using spatio–temporal link–level information. *Journal of Transport Geography* 82: 102549.

Lao, X.; Zhang, X.; Shen, T.; and Skitmore, M. 2016. Comparing China's city transportation and economic networks. *Cities* 53: 43–50.

Lee, D.; and Lee, J. 2020. Testing on the Move South Korea's rapid response to the COVID-19 pandemic. *Transportation Research Interdisciplinary Perspectives* 100111.

Medo, M. 2020. Epidemic spreading on spatial networks with distance-dependent connectivity. *arXiv preprint arXiv:2003.13160* .

Menon, A. K.; and Lee, Y. 2017. Predicting short-term public transport demand via inhomogeneous Poisson processes. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2207–2210.

Mukhopadhyay, A.; Pettet, G.; Vazirizade, S.; Vorobeychik, Y.; Kochenderfer, M.; and Dubey, A. 2020. A Review of Emergency Incident Prediction, Resource Allocation and Dispatch Models. *arXiv preprint arXiv:2006.04200* .

Mukhopadhyay, A.; and Vorobeychik, Y. 2017. Prioritized allocation of emergency responders based on a continuous-time incident prediction model. In *International Conference on Autonomous Agents and MultiAgent Systems*.

Nuzzolo, A.; Crisalli, U.; Rosati, L.; and Ibeas, A. 2013. Stop: a short term transit occupancy prediction tool for aptis and real time transit management systems. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, 1894–1899. IEEE.

Olmo, J.; and Sanso-Navarro, M. 2020. Modelling and forecasting the spread of COVID-19 in New York City 30.

Tirachini, A. 2020. COVID-19 and Public Transportation: Current Assessment, Prospects, and Research Needs 21.

Van Oort, N.; Brands, T.; and de Romph, E. 2015. Short term ridership prediction in public transport by processing smart card data. *Transportation Research Record* (2015).

Wang, C. J.; Ng, C. Y.; and Brook, R. H. 2020. Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing. *JAMA* 323(14): 1341–1342. ISSN 0098-7484.

Whitelaw, S.; Mamas, M. A.; Topol, E.; and Van Spall, H. G. C. 2020. Applications of digital technology in COVID-19 pandemic planning and response. *The Lancet Digital Health* 2(8): e435–e440. ISSN 25897500.

Wilbur, M.; Ayman, A.; Ouyang, A.; Poon, V.; Kabir, R.; Vadali, A.; Pugliese, P.; Freudberg, D.; Laszka, A.; and Dubey, A. 2020. Impact of COVID-19 on Public Transit Accessibility and Ridership. *Proceedings of the Annual Conference of Transportation Research Board* .

Zhou, M.; Wang, D.; Li, Q.; Yue, Y.; Tu, W.; and Cao, R. 2017. Impacts of weather on public transport ridership: Results from mining data from different sources. *Transportation research part C: emerging technologies* 75: 17–29.