Azimian, Amin, and Junfeng Jiao. 2021. "Prediction of 'L' Train's Daily Ridership in Downtown Chicago During the COVID-19 Pandemic." *Findings*, December. https://doi.org/10.32866/001c.30181.

TRANSPORT FINDINGS

Prediction of "L" Train's Daily Ridership in Downtown Chicago During the COVID-19 Pandemic

Amin Azimian[1], Junfeng Jiao[1]

[1] Urban Information Lab, University of Texas at Austin

Keywords: rail, ridership, random forest, prediction, covid-19

https://doi.org/10.32866/001c.30181

Findings

In this study, we utilized a random forest model to predict the "L" train's daily ridership in the Chicago downtown area during the pandemic based on environmental, transportation, and COVID-19-related factors. The results indicated that the model accurately predicts ridership one month in advance. However, its accuracy degraded over time. Moreover, average temperature, stay-at-home order status, and percentage of home renters were found to be the most important factors contributing to ridership.

1. Questions

Over the last year, many studies (da Silva et al. 2021; Hu and Chen 2021; Jiao and Azimian 2021; Liu, Miller, and Scheff 2020; Schneider and Schinkowsky 2021; Tokey 2021) have investigated the change in travel patterns for different transportation modes because of the COVID-19 outbreak. Nevertheless, most of them have not focused on rail transit systems, and they have failed to develop short- and long-term predictions of daily transit ridership during the pandemic, which is critical for transit authorities to manage and optimize their services. Consequently, the aim of this paper is two-fold. First, we performed a data-driven analysis of the impacts of the environmental, transportation, and COVID-19-related factors on the "L" train system (Metro) in the Chicago downtown area. Second, we attempted to use a random forest model to predict the "L" train's daily ridership during the pandemic based on the aforementioned predictors. Our findings contribute to filling the knowledge gap regarding the pandemic's impact on daily rail transportation demand patterns in large cities, specific to each transport mode, and propose short- and long-term predictors of future public transport demand.

2. Methods

To perform the analysis, we collected various data from different sources, COVID-19 cases, and "L" train total ridership for all stations (entries at all turnstiles) in the Chicago downtown area, including zip codes 60601 through 60605, and 60607, obtained from the Chicago's open data portal (CTA 2021). Timeline for COVID-19 stay-at-home was derived from the city of Chicago's official website (https://www.chicago.gov) Percentage of people renting a home/an apartment, and job rating (job availability score) were collected from Niche's website (https://www.niche.com/), and weather-related data were derived from the Meteostat portal (https://meteostat.net). Figure 1 shows the data collection procedure and the variables used in the model. Additionally, Table 1 represents the summary statistics of the variables used in our model.

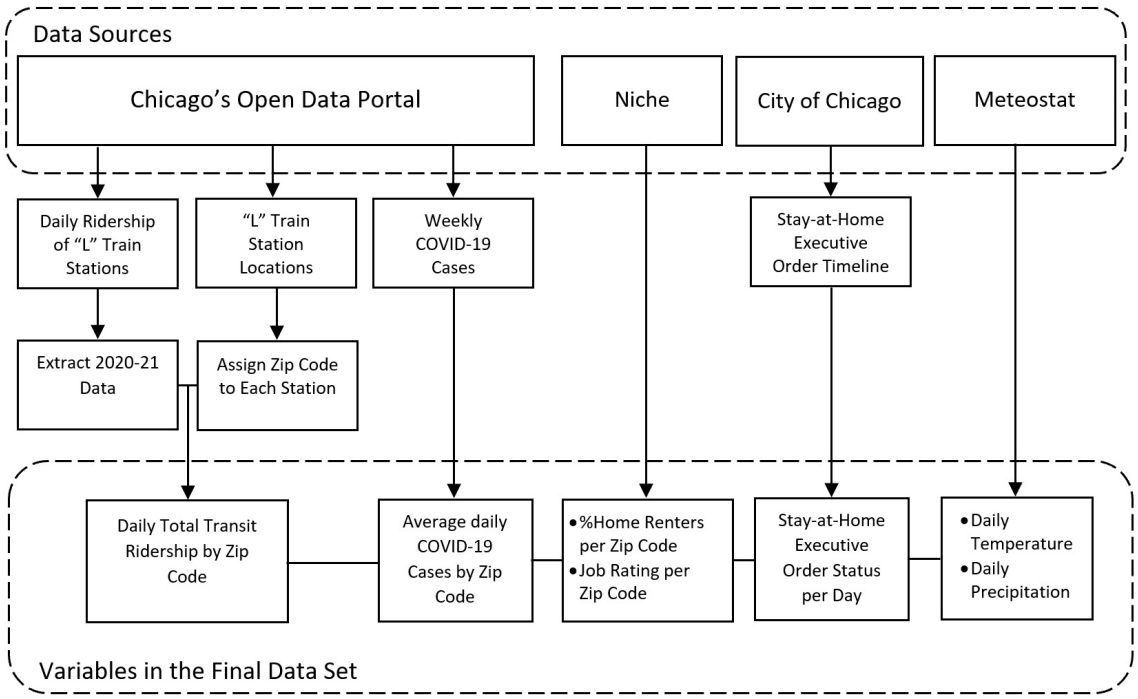

Figure 1. Data collection and aggregation framework

Table 1. Summary statistics of the variables used in the random forest model

Variable	Mean	Std. Dev.	Min	Max
Rides	4162.19	5217.74	3	53340
No. COVID-19 Cases	2.39	4.27	0	29
Stay-at-Home Executive Order (1 if the order was in effect, 0 otherwise)	0.23	0.42	0	1
% Renters	58.00	9.35	43	71
Job Rating	3.22	0.55	2.3	4
Avg. Temperature (°C)	13.63	8.96	-9.4	28.5
Precipitation (mm)	3.23	9.25	0	96.3
Type of Day (1= If weekday, 0 otherwise)	0.70	0.46	0	1

As for methodology, we proposed a random forest model to predict the log transform of daily ridership. The random forest approach was first proposed by Ho (1995). Breiman (1996) developed an extension of the approach, which is a machine learning algorithm that combines Breiman's "bagging" idea and the random selection of features, introduced first by Ho (1995). This method has a reputation for its simplicity and diversity, as it can be used for both classification and regression tasks. In the random forest model, first, the data set is split into training and validation data sets. As for the training data set, we utilized all data in zip codes 60601 through 60605, and 60607, for March 1, 2020, to December 31, 2020, whereas we used data for zip code 60601 for January 1, 2021, to May 31, 2021, for validation. In the next step, many decision trees are randomly created with "boot-strap samples" from the data set. The branching of each tree is determined by randomly selected predictors

Table 2. Performance measures of the random forest model

Measure	Training Data	Test Data
MAE	0.04	0.12
MSE	0.01	0.03

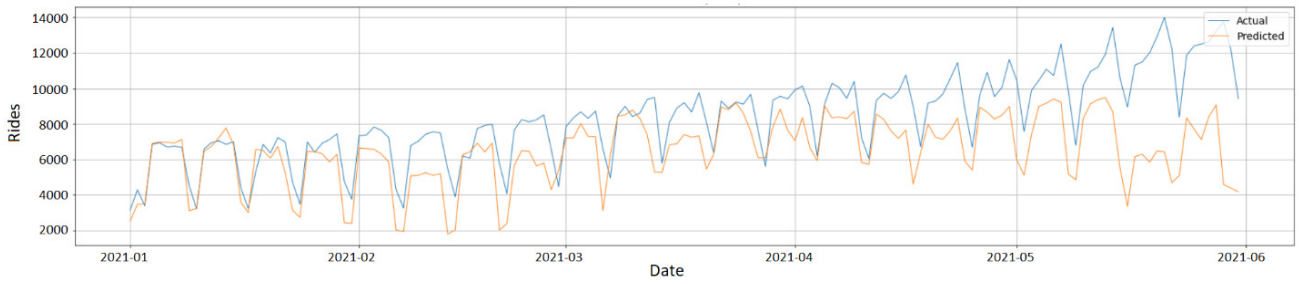

Figure 2. Actual versus predicted log transform of ridership in zip code 60601

at node points. The random forest final estimate is the average of all results from each tree. Therefore, each individual tree affects its estimation at certain weights (Yeşilkanat 2020). The random forest algorithm is superior to other machine learning algorithms, as it can randomly receive training data from subsets and form trees with a random algorithm (Panov and Džeroski 2007).

3. Findings

We fitted a random forest to the training dataset (1,836 observations) and generated forecasts of the log transform of transit ridership five months in advance based on test data (151 observations). To assess the model's performance, we utilized the mean squared error (MSE) and mean absolute error (MAE) as study measures. As shown in Table 2, the estimated MAE and MSE values for the test dataset were slightly higher than those in the training dataset, suggesting that the model's performance was slightly degraded in the test data. From Figure 2, the model accurately predicted transit ridership in January 2021. Furthermore, the forecast accuracy performed moderately well from early February 2021 until the end of April 2021. However, the predicted values tended to diverge from the actual values in May 2021 and beyond.

To explore the impact of explanatory variables on metro ridership, the relative contributions of those variables were calculated using the permutation method. Relative variable importance values range from 0% to 100%. The most important variable always has a relative importance of 100% and is the basis for measuring less important variables' relative importance.

As shown in Figure 3, the average temperature contributed the most to predicting ridership, with a relative importance of 100%. This finding is consistent with the findings of Guo, Wilson, and Rahbee (2007), who reported that temperature significantly affects daily transit ridership in the Chicago

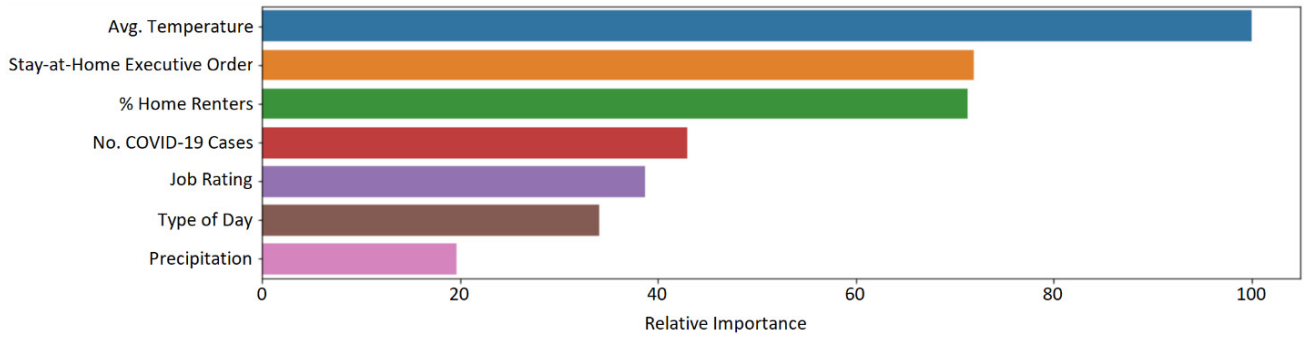

Figure 3. Relative importance of predictors as determined by the random forest model

area. The stay-at-home order, which was 72% as important as the average temperature, ranked second. As discussed in previous works (Jiao and Azimian 2021; Roy 2021), stay-at-home orders mandated all residents to avoid outings except to travel to an essential job. Moreover, many employers switched to remote work, which significantly affected transit ridership.

Percentage of home renters, average daily number of COVID-19 cases, and job rating accounted for 71%, 43%, and 38% of relative contributions, respectively. Regarding the percentage of home renters, its high relative importance implied that home renters were more likely to use transit than cars, which is in line with the findings of Pongprasert and Kubota (2017). In terms of the average daily number of COVID-19 cases, we can interpret our findings from two different perspectives. First, the number of COVID-19 cases could be a proxy for risk perception (Liu, Miller, and Scheff 2020). That is, the higher the number of COVID-19 cases, the lower the transit ridership. Second, an increase in transit ridership would pose a risk of COVID-19 transmission to riders and increase the number of COVID-19 cases (Hu and Chen 2021). As for the relative importance of job rating, it is reasonable to believe that areas with higher job rating/opportunities (i.e., downtown) likely attracted a surge of commuters from surrounding areas (Dave et al. 2020), and many favored public transportation because of downtown traffic congestion and high parking fees (Florida 2019). Type of day (weekends/holidays versus weekdays) had a relative importance of 34%, suggesting that ridership would change based on weekdays and weekends (Palaio et al. 2021).This finding is consistent with the fact that passengers mostly commute to workplaces and school during weekdays rather than weekends. Last, precipitation had a relative importance of 19%, and rain and snowfall were likely to have a different influence level on ridership as it varied over a year (Kashfi, Lee, and Bunker 2013).

Regarding research implications, the model developed in this research can be used by transit agencies to accurately predict demand for the "L" train system one month in advance or more, while accounting for the three most important factors (temperature, stay-at-home orders, and percent of home renters). Such a model would assist planners in adjusting their level of service

(e.g., frequency) to avoid overcrowding and maintain social distancing during the pandemic. These strategies are even more critical during severe weather when more customers are likely to use rail systems, which increases boarding time and causes trains to fall behind schedule. Last, people with financial uncertainty (e.g., home renters) require more attention as they highly contribute to transit ridership. For example, important changes may include implementing alternate schedules when serious service disruptions do not allow for regularly scheduled service and providing reduced transit fare based on monthly income and rent.

As for future work, this study can be extended by developing a comprehensive model to predict transit ridership in rural or low-density urban areas. In addition, the inclusion of a new temporal variable/measure that accounts for the severity of the various governmental COVID-19 restrictions (e.g., executive orders, advisories, etc.) that periodically limit public space access is needed.

Submitted: November 14, 2021 AEDT, Accepted: December 06, 2021 AEDT

REFERENCES

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24 (2): 123–40.

CTA. 2021. "Open Data." CTA. Chicago Transit Authority. 2021.

da Silva, Denise Capasso, Sara Khoeini, Deborah Salon, Matthew W Conway, Rishabh S Chauhan, Ram M Pendyala, Ali Shamshiripour, Ehsan Rahimi, Tassio Magassy, and Abolfazl Kouros Mohammadian. 2021. "How Are Attitudes Toward COVID-19 Associated with Traveler Behavior During the Pandemic?" *Findings*, 24389.

Dave, Dhaval M, Andrew I Friedson, Kyutaro Matsuzawa, Drew McNichols, Connor Redpath, and Joseph J Sabia. 2020. "Risk Aversion, Offsetting Community Effects, and Covid-19: Evidence from an Indoor Political Rally." National Bureau of Economic Research.

Florida, Richard. 2019. "The Great Divide in How Americans Commute to Work." 2019.

Guo, Zhan, Nigel HM Wilson, and Adam Rahbee. 2007. "Impact of Weather on Transit Ridership in Chicago, Illinois." *Transportation Research Record* 2034 (1): 3–10.

Ho, Tin Kam. 1995. "Random Decision Forests." In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–82. IEEE.

Hu, Songhua, and Peng Chen. 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.

Jiao, Junfeng, and Amin Azimian. 2021. "Exploring the Factors Affecting Travel Behaviors during the Second Phase of the COVID-19 Pandemic in the United States." *Transportation Letters* 13 (5–6): 331–43.

Kashfi, Syeed Anta, Brian Lee, and Jonathan Bunker. 2013. "Impact of Rain on Daily Bus Ridership: A Brisbane Case Study." In *Australasian Transport Research Forum 2013 Proceedings*, 1–18. Australasian Transport Research Forum.

Liu, Luyu, Harvey J Miller, and Jonathan Scheff. 2020. "The Impacts of COVID-19 Pandemic on Public Transit Demand in the United States." *Plos One* 15 (11): e0242476.

Palaio, Lori, Tung Vo, Michael Maness, Robert L Bertini, and Nikhil Menon. 2021. "Multicity Investigation of the Effect of Holidays on Bikeshare System Ridership." *Transportation Research Record* 2675 (7): 404–23.

Panov, Pančе, and Sašo Džeroski. 2007. "Combining Bagging and Random Subspaces to Create Better Ensembles." In *International Symposium on Intelligent Data Analysis*, 118–29. Springer.

Pongprasert, Pornraht, and Hisashi Kubota. 2017. "Why TOD Residents Still Use Car? A Study on Factors Affecting the Automobile Ownership and Use of Residents Living near Transit Stations of Bangkok." In *55th Civil Engineering Planning Research Presentation / Spring Meeting*. Japan: Japan Society of Civil Engineers.

Roy, Shuvangkor Shusmoy. 2021. "The Impacts of COVID-19 Pandemic on" L" Rail Transit Ridership in Chicago." *MUP Capstone*.

Schneider, Robert J, and Hayley Schinkowsky. 2021. "Reactions to University Campus Commute Mode Shifts During COVID-19." *Findings*, 29446.

Tokey, Ahmad Ilderim. 2021. "Spatial Association of Mobility and COVID-19 Infection Rate in the USA: A County-Level Study Using Mobile Phone Location Data." *Journal of Transport & Health* 22: 101135.

Yeşilkanat, Cafer Mert. 2020. "Spatio-Temporal Estimation of the Daily Cases of COVID-19 in Worldwide Using Random Forest Machine Learning Algorithm." *Chaos, Solitons & Fractals* 140: 110210.