Correlative Analysis on Transportation Ecosystems

Joshua Cabral, Nikhil Chawla, Edmund Chen, Ankit Kaul, Yunbo Zhang
Georgia Institute of Technology

1 INTRODUCTION

Many visualization systems for public transit do not offer predictive analysis. Some advanced GPS systems for cars can give estimated times for a given trip but not give overall predictions congestion in various areas. This is needed to help riders and travelers to better distribute themselves across days and route to make travel more efficient.

2 PROBLEM DEFINITION

Our project will focus on the impact of confounding factors between various modes of transportation. We will use several analytical (clustering, regression analysis, correlative analysis) and visualization techniques (heatmaps, network graphs, temporal graphs, chords) to accomplish this goal. We hope to find any connection between the various features that are described below in order to help give guidance to future city planners.

3 LITERATURE SURVEY

Prommaharaj et al., Zhang et al., and *Kunama et al.* build pre-processing and visualization tools for General Transit Feed Specification (GTFS) data and provide easy to use interface for users to visualize raw GTFS data. They provide insights on understanding, processing and visualizing the raw data from GTFS. However, they do not provide any non-trivial analyses using this dataset. On the other hand, *Sobral et al.* provides visualization techniques for analyzing people dynamics through granular data. It provides useful methods of heat-mapping to explore congestion but lacks a uniform method to compare different modes of transit.

Sewall et al. and *Setiawan et al.* leverages the density and conservation properties of traffic flow to establish a single-road Lighthill-Whitman-Richards traffic model. It provides novel ways to optimize the placement of stop lights to optimize throughput but fails to consider complicated road networks.

Fortin et al., Krause et al. both introduce graph-oriented frameworks to develop analysis tools characterizing the different congestion and distances of transit networks over time. They provide insight on how to effectively visualize characteristics of network structure, though both lack the flexibility to compare multiple or non-network transit modes.

Kalamaras et al. and *Kurkcu et al.* present an interactive visual analytics platforms to explore historical road traffic data, to predict future traffic through user interaction, and to compare speeds of different bus routes. Their main contributions include utilization of visual analytics techniques on multiple features, including user-defined ones, to identify more informative patterns. These papers includes a study on evaluating traffic prediction accuracy of different prediction models using a NRMSE metric.

Barbieri et al., Liu et al., Munawar et al. analyse the impact of COVID-2019 on different aspects of transit. They give a good reference to our work on how analyzing the impact of the pandemic can be done, but they only consider one form of transit, whereas we are proposing to analyze multiple forms of transit systems.

Cheng et al., Jiang et al., Miaoyi et al. explore the transit demand correlations between taxi and subway. They are useful in exploring the inter-dependencies and interactions across multiple transport system. However, other modes of transportation such as buses and bikes in a city are not studied.

4 PROPOSED METHOD

4.1 Intuition

There are currently lots of ways to visualize transit routes on map. What these methods lack is a way to translate past data into predictions that can be intuitively visualized to communicate expected travel time information to the rider. Most public transit system have a way to communicate daily issues but these are typically in the form of message dashboards and only show messages after something goes wrong. We hope to give predictions on how busy a given transit system is. This can impact a rider's decision on which form and when that want to travel for a specific reason. Integrating our prediction model analysis with these visualization will allow travelers to predict busy times and save them from the arduous task of navigating busy

crowds while reducing load on transportation systems do to more informed traveler actions.

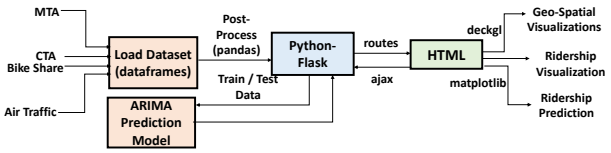

Figure 1: Integrated Methodology to visualize and predict ridership/traffic for multiple transit systems

4.2 Description

In order to build a better transit visualization system, we will analyze traffic patterns and make predictions based on the data. We will then use displayed gradients on the routes to show areas that may become busy based on the features in the data. We show different visualizations and analyses of a variety of transit systems - including metros, airlines, bike shares, taxis, and car traffic trips as shown in Figure 1. As each mode of transit provides us with a different form of data, we have to take into account various limits to be able to effectively mesh these different metrics together. For instance, for some subway networks where they have a tap-in and tap-out system we are able to obtain the specific origin-destination numbers, whereas other subway networks with a single-fare system only yield entry and exit numbers.

For sake of example, we elaborate on the following visualizations considering the several modes of transit system under mentioned dataset constraints:
1. GTFS Network: Transit data from GTFS includes a list of stops associated with different trips, a list of trips corresponding to different routes, and a list of routes with unique route IDs. The goal is a nodal visualization of distinct routes for which a list of all route-specific stations is required. We noticed that stops and routes do not have a direct relationship in the available datasets because stops are associated with trips, which can be different. To determine a list of stops served by a route, we cross-referenced the available datasets. As a starting point, we visualized all the current routes and stops in the Metropolitan Transportation Authority (MTA) and the Chicago Transit Authority (CTA) networks, as illustrated in section 5. The next steps include visualizing temporal variation of ridership due to COVID-19.
2. Ridership Load: Ridership load is classified in many different ways between varying agencies. We define three subclasses of ridership data in increasing granularity: 1) general ridership by week or less specific 2) ridership by station by day 3) directed origin-destination by station by hour. For the classes 2) and 3) we can derive heatmap visualizations plotting the relative load on stations over a time period. In order to process such data, we first run this through pre-processing stages to calculate the throughput of each station per day, which can be then used to plot the ridership visualizations mentioned in the latter section.
3. Directed Trips: Some of our more granular datasets such as airline transit, bikeshare, taxi, and selected metros consist of collections of directed trips between stations. Since we may not know the exact route between two endpoint of the trip, we visualize each trip using an arc connecting the endpoints of the trip. In the latter section we cover examples of this for bikeshare and air travel during 2019-2020

4.3 List of Innovations

- Show predictive congestion expectations of various modes of travel over geospatial map
- Show historical data that can be played forward to help users see busy time of year and busy areas

4.4 Analysis

To further analyze the traffic data, we use machine learning techniques to train prediction model that predict the future traffic within user selected station within a selected year. Specifically, we train an Auto-regressive Integrated moving average (ARIMA) model to predict daily number of trips that start from a user selected station for the transit mode. Specifics of this algorithm are further elaborated in 5.3, where we have fleshed out the rest of our process.

5 EXPERIMENTS

Our test beds are aimed to answer the following questions:

(1) What are the steady state and temporal impacts of COVID-19 on public transport ridership? For e.g. how does transit ridership compare in April 2019 vs April 2020?

(2) How does the traffic/ridership/air-traffic prediction accuracy compare for years 2019 and 2020 due to the pandemic?

In this section, we describe our testbeds and experiments to address these questions.

5.1 Geospatial Visualizations

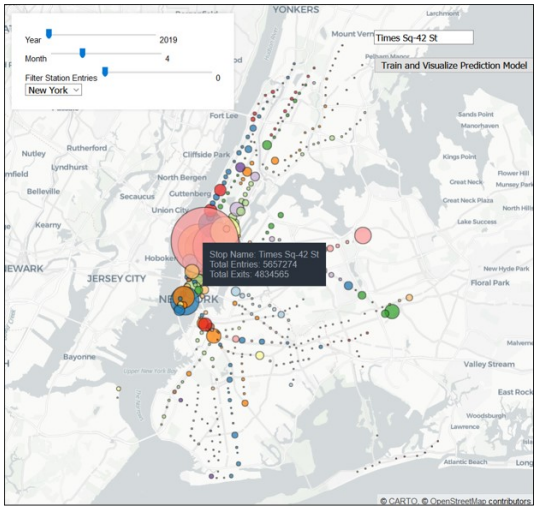

Figure 2: MTA (New York) Transit Routes visualized by ridership (total entries) for April 2019

5.1.1 Nodal Visualization of Routes. Figure 2 illustrates the MTA (New York) transit network routes for April 2019 via a nodal visualization. All routes and stops in the MTA network from January 2019 to March 2021 have been visualized on a map using the DeckGL API. The temporal variation of ridership can be observed and number of entries adjusted using the slider inputs. Ridership prediction based on historical time series data is described in Section 5.3.

Figure 3 is a layout of Chicago's bus routes. We are showing how busy each route is relative to its historical ridership. This will help inform potential passengers about days when they need to leave early or to prepare for large crowds. If a user selects a route, they can then see an ARIMA model prediction of the number of passengers on that route. Users can select any year to run the ARIMA prediction on it it will compare it to its true value data. There is also a tool tip that gives users information about the specific route that they are focusing on. This can also be used by city planners to give insight on the best times to give extra support to that specific rail system.

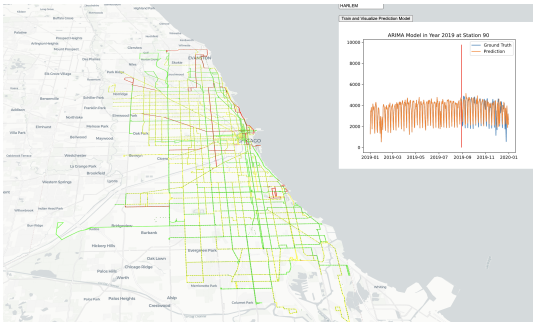

Figure 3: CTA Bus

5.1.2 Arc Visualization of Trips. To visualize the trips without fixed routes, we use colored arcs to represent each trip. To make the visualization less messy, we group the trips by their start and end stations and filter out the less important trips. As shown in Figure 4, the bike share trip data in January 2019 is shown. Each arc in the visualization represent all the trips from one station to the other within that time of the year, and the thickness represent the number of trips between the stations. In addition, we allow the user to filter out the arcs if the number of trips is fewer than certain threshold. This option enables use to have a clear view on which stations are the major nodes for the entire transit system.

Figure 4: LA Bike Share Trips in January 2020

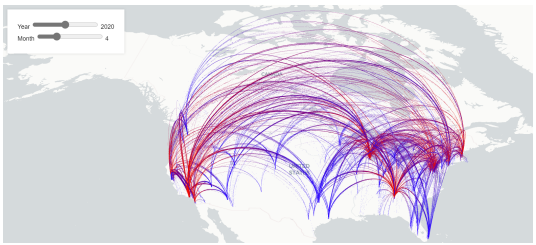

Figure 5: #Flights from Major Airport Locations in Mar 2020

To visualize traffic at given airport location, we grouped trips between two locations and displayed trends in trips every month. A trip is denoted using colored arc between origin and destination airport locations. The thickness of arc denotes frequency of trips between two locations. Figure 5 shows the flights departures for major airports in Chicago, LA , Atlanta, Boston and NY for Mar 2020

5.2 Ridership Visualizations

5.2.1 Heatmap Visualization of Ridership. A question we want to answer falls upon the relative ridership of various stations throughout a certain time frame. Viewed across a large timeframe, this may identify which stations have seen the most change in relative congestion throughout the system? As such, we develop a framework for a time-series heatmap, which can visualize any change in metric across various physical datapoints (station, route, county) over time. The most straightforward data we can apply this on would be a ridership exits for each station. Figure 6 displays this, with it already apparent that ridership drops down on the weekends and holidays such as Jan. 1st (New Years) and Jan. 17th (MLK Day). To view trends beyond the weekly shifts, we introduce normalization to view relative station load. Figures are cropped for purposes of this report. As we can see, this is fairly cyclic with the

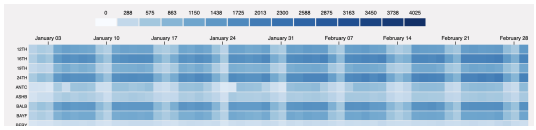

Figure 6: Bart Ridership Heatmap 2021

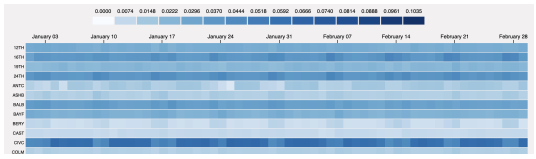

Figure 7: Bart Normalized Ridership Heatmap 2021

weekly cycle, so it will be helpful to further analyze the percentage change of ridership over extended periods of time. Doing this process for Chicago's Metro, we can in fact see that though March saw a steady decrease of percent ridership, the most severe dropoff came on June 1st, see 7. Providing both raw count and normalized views across different time series in a couple cities are shown in our visualization. Using the normalized percentage change for instance, makes it visible which stations are especially reliant on weekday traffic.

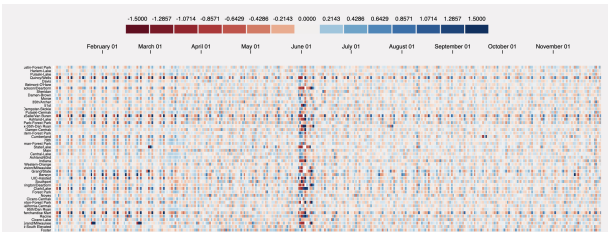

Figure 8: CTA Metro Ridership Percentage Change

5.2.2 Network Visualization of Ridership. Another useful tool we may use for further analyses are force-directed graphs of directed trips between two given stations. For instance, we may introduce a temporal feature where we can visualize which station to station combination receive greater or less traffic across a period of time.

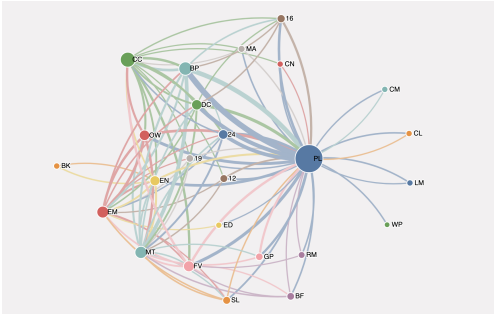

Figure 9: Bart Network Ridership 2020

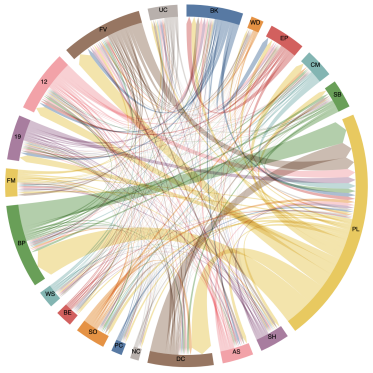

Figure 10: Bart Chord Ridership 2020

5.2.3 Chord Visualization of Ridership. Another method of visualizing directed trips would be a chord graph, where the arrows and their respective thicknesses represent the throughput of passengers from the two endpoints. However, this turns quite convoluted past a relatively small network, so it's important to put an adjustable number of stations parameter. It gives a better view of station to station specifics, as the links are much clearer than those in a force-directed graph.

5.3 ARIMA Prediction Model Visualizations

For each of the transit types, we implement the functionality for user to pick a station and train an ARIMA model that predicts the daily number of trips that will start from that station. To format the dataset to train the model, we sum the number of trips groupbing by each day of the selected year to get a list of numbers representing the number of trips on each day of the year. We partition the first two-thrid of the dataset as the training set and the rest as testing set for our ARIMA model. To visualize our trained model, for the number of trips in each day in the testing set, we make a prediction on that day and append the ground truth result into the training set to update the model. We rollout model's prediction over the entire dataset and compare the prediction and ground truth in a plot. Shown in figure 11, an ARIMA predictive model is trained to predict the trip counts at station Time Sq-42 St over the year 2019. In the figure, we use a vertical red line to separate the plot into two part. The left side of the plot shows the ground truth of training data on the trip count, and the right side shows the comparison between the predicted trips counts and the ground truth. Results indicate that the more number of travels in a station, the higher accuracy we can get in the prediction model.

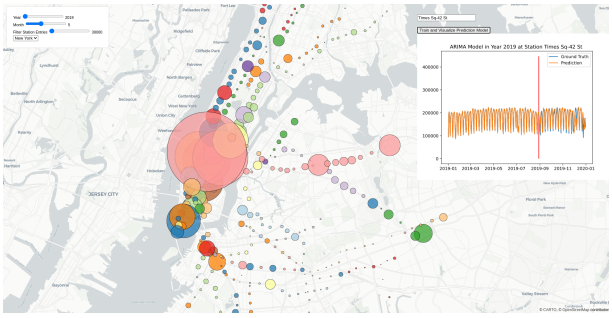

Figure 11: MTA subway total entries (Apr. 2019) and predicted entries (2019)

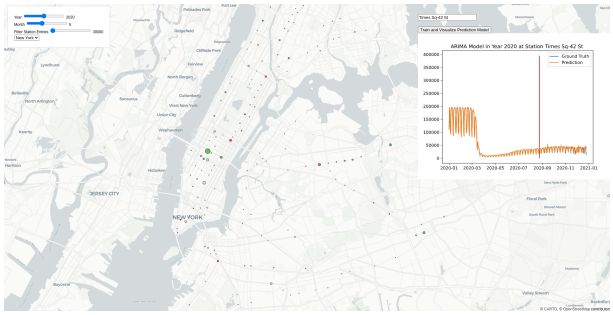

Figure 12: MTA subway total entries (Apr. 2020) and predicted entries (2020)

5.4 Impact of COVID on Transit Systems

From all of our visualizations, we are able to clearly see the impact of COVID-19 on different transit modes across the country. For all the transit mode's visualization, we compare the number of trips for the entire transit systems on record during May in 2019 and 2020. In addition, we compare in detail the number of trips at one of the busiest stop in individual transit systems between two entire years.

5.4.1 Air Traffic Across the United States. For the air traffic data, as shown in Figure 13 and 14, we can see the number of flight in May 2020 is significantly fewer the same month in 2019. On the plot, we visualize the number of flight departs from Hartsfield - Jackson Atlanta International Airport in Atlanta over the entire year and see a clear dent during May, June, and July 2020.

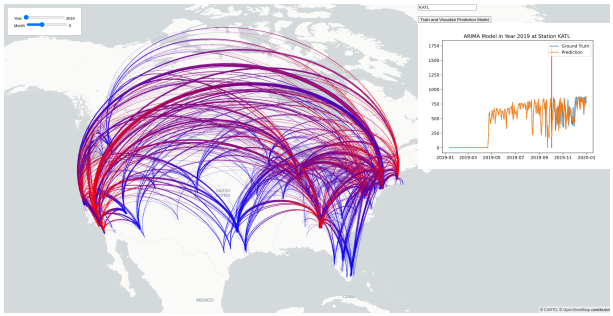

Figure 13: Air Traffic In 2019

5.4.2 Bus Traffic In Chicago. In Chicago, we compare the trips count across all bus routes on May 1st, 2019 and 2020, and also compare the yearly trip count over the busiest route Inner Drive/Michigan Express. We can

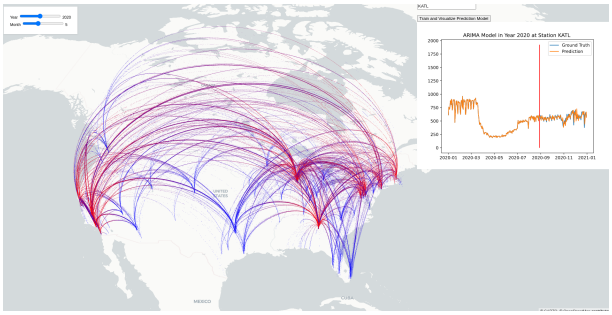

Figure 14: Air Traffic In 2020

see in Figure 15 and 16, much fewer bus route usages are recorded in 2020 than 2019.

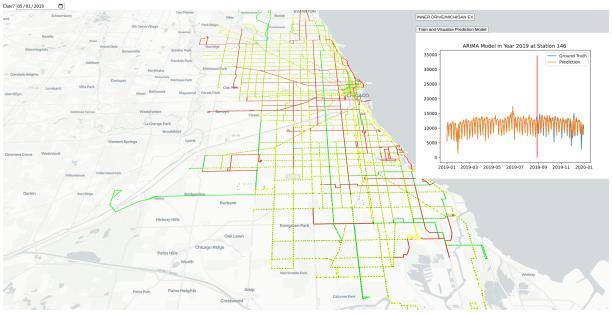

Figure 15: CTA Bus traffic in 2019

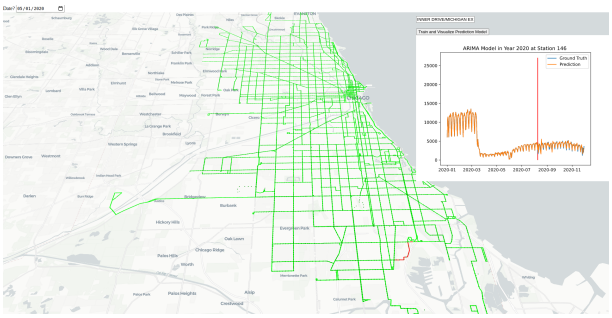

Figure 16: CTA Bus traffic in 2020

5.4.3 Metro System in New York. For the metro system in New York, we see the trips across all stations with daily trip number more than 30000 (Fig. 11 and 12). For the month of May, ridership in 2020 was ≈6.3% of that in 2019 for the busiest station Times Square 42 St. A significant valley in ridership reduction was observed across all stations during summer 2020, including the busiest station (Times Square 42 St).

5.4.4 Shared Bike System in Los Angeles. For bike share trips in Los Angeles, we see the similar trend as shown in Figure 17 and 18. More severely even at the busiest bike station, no trip is recorded at all through April, May, and June in 2020 as shown in the plot.

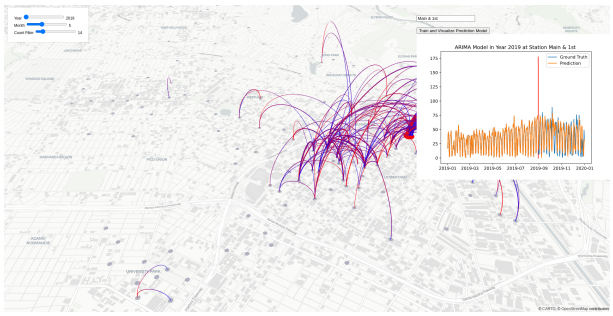

Figure 17: LA Bike Share in 2019

Figure 18: LA Bike Share in 2020

6 CHALLENGES

This project uses several unique data sets each of which provides a unique set of challenges when analyzing and visualizing it. Some data sets only have a daily resolution whereas others have exact timestamps for each passenger. The features are not always the same and different modes of transit have different features of varying importance.

7 CONCLUSION

The approach we have taken could improve the ability of travelers to decide which routes to take and what are the best days to travel on. It can also help city planners to be able to find the utilization of the various modes of transit in their city or of other city to help them decide where to invest scarce resources. All team members have contributed a similar amount of effort to this project.

REFERENCES

[1] Diego Maria Barbieri, Baowen Lou, Marco Passavanti, Cang Hui, Inge Hoff, Daniela Antunes Lessa, Gaurav Sikka, Kevin Chang, Akshay Gupta, Kevin Fang, et al. 2021. Impact of COVID-19 pandemic on mobility in ten countries and associated perceived risk for all transport modes. *PLoS One* 16, 2 (2021), e0245886.

[2] X. Cheng, K. Huang, L. Qu, and L. Li. 2020. A Cooperative Data Mining Approach for Potential Urban Rail Transit Demand Using Probe Vehicle Trajectories. *IEEE Access* 8 (2020), 24847–24861. https://doi.org/10.1109/ACCESS.2020.2970863

[3] Philippe Fortin, Catherine Morency, and Martin Trépanier. 2016. Innovative GTFS data application for transit network analysis using a graph-oriented method. *Journal of Public Transportation* 19, 4 (2016), 2.

[4] Shixiong Jiang, Guan Wei, Zhengbing He, and Liu Yang. 2018. Exploring the Intermodal Relationship between Taxi and Subway in Beijing, China. *Journal of Advanced Transportation* 2018 (06 2018), 1–14. https://doi.org/10.1155/2018/3981845

[5] Ilias Kalamaras, Alexandros Zamichos, Athanasios Salamanis, Anastasios Drosou, Dionysios D Kehagias, Georgios Margaritis, Stavros Papadopoulos, and Dimitrios Tzovaras. 2017. An interactive visual analytics platform for smart intelligent transportation systems management. *IEEE Transactions on Intelligent Transportation Systems* 19, 2 (2017), 487–496.

[6] Josua Krause, Marc Spicker, Leonard Wörteler, Matthias Schäfer, Leishi Zhang, and Hendrik Strobelt. 2012. Interactive visualization for real-time public transport journey planning. In *Proceedings of SIGRAD 2012; Interactive Visual Analysis of Data; November 29-30; 2012; Växjö; Sweden*. Citeseer, 95–98.

[7] Narumon Kunama, Mudtana Worapan, Santi Phithakkitnukoon, and Merkebe Demissie. 2017. GTFS-VIZ: Tool for preprocessing and visualizing GTFS data. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 388–396.

[8] Abdullah Kurkcu, Fabio Miranda, Kaan Ozbay, and Claudio T Silva. 2017. Data visualization tool for monitoring transit operation and performance. In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. IEEE, 598–603.

[9] Luyu Liu, 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 (2020), e0242476.

[10] Li Miaoyi, Lei Dong, Zhenjiang Shen, Wei Lang, and Xinyue ye. 2017. Examining the Interaction of Taxi and Subway Ridership for Sustainable Urbanization. *Sustainability* 9 (02 2017), 242. https://doi.org/10.3390/su9020242

[11] Hafiz Suliman Munawar, Sara Imran Khan, Zakria Qadir, Abbas Z Kouzani, and MA Mahmud. 2021. Insight into the Impact of COVID-19 on Australian Transportation Sector: An Economic and Community-Based Perspective. *Sustainability* 13, 3 (2021), 1276.

[12] Postsavee Prommaharaj, Santi Phithakkitnukoon, Merkebe Getachew Demissie, Lina Kattan, and Carlo Ratti. 2020. Visualizing public transit system operation with GTFS data: A case study of Calgary, Canada. *Heliyon* 6, 4 (2020), e03729.

[13] Erwin B Setiawan, D Tarwidi, and Rian F Umbara. 2015. Numerical simulation of traffic flow via fluid dynamics approach. *International Journal of Computing and Optimization* 3 (2015), 93–104.

[14] Jason Sewall, David Wilkie, Paul Merrell, and Ming C Lin. 2010. Continuum traffic simulation. In *Computer Graphics Forum*, Vol. 29. Wiley Online Library, 439–448.

[15] Thiago Sobral, Teresa Galvão, and José Borges. 2019. Visualization of urban mobility data from intelligent transportation systems. *Sensors* 19, 2 (2019), 332.

[16] Tianchi Zhang, Mei-Hwa Chen, and Catherine Lawson. 2014. General transit feed specification data visualization. In *2014 22nd International Conference on Geoinformatics*. IEEE, 1–6.