Dynamic Visualization Platform for Travel-Related Data Integration to Support Sustainability-Based Decision-Making for Smart Cities

Yufei Xu
School of Civil and Environmental Engineering
Georgia Institute of Technology
Atlanta, GA
yxu403@gatech.edu

Gulam Kibria
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA
mkibria7@gatech.edu

Chaojie Wang
School of Civil and Environmental Engineering
Georgia Institute of Technology
Atlanta, GA
chaojie.wang@gatech.edu

Einat Tenenboim
School of Civil and Environmental Engineering
Georgia Institute of Technology
Atlanta, GA
einat.tenenboim@gatech.edu

Viswa Sri Rupa Anne
School of Civil and Environmental Engineering
Georgia Institute of Technology
Atlanta, GA
vanne3@gatech.edu

Zhiwei Chen
Department of Civil, Architectural and Environmental Engineering
Drexel University
Philadelphia, PA
zc392@drexel.edu

Srinivas Peeta
School of Civil and Environmental Engineering
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA
peeta@gatech.edu

***Abstract*—Smart cities seek to leverage data from advanced information, communication, and sensor technologies (ICSTs) for achieving their transportation-related sustainability goals. However, the multi-source, multi-timescale nature of these disparate data sets introduces many challenges to community decision-makers, hindering the use of these technologies in an efficient, effective, and holistic manner. Here, using statistical and machine learning methods, we present a visualization platform developed for the City of Peachtree Corners, GA, comprising nine integrated data sets. This platform can capture dynamic interactions between data from different sources and has the potential to support decision-makers in developing different solution options for contemporary transportation-related problems in a smart city environment.**

Keywords—smart city, data integration, data visualization

I. INTRODUCTION

Smart cities seek to leverage the deployment of advanced information, communication, and sensor technologies (ICSTs) for improving community-level transportation-related decision-making, such as decisions related to attaining certain sustainability goals (e.g., improved mobility, enhanced transit accessibility, etc.). While these technologies are well-established and organically introduced into communities, they are typically not utilized adequately due to the lack of systematic methods and frameworks to leverage them. Hence, smart cities are faced with various challenges in realizing the benefits of ICSTs, the main one being their disparate nature. Specifically, the large-scale, multi-source, multi-timescale nature of these data precludes community decision-makers from using them in a deliberate, holistic manner. This study aims to utilize data mining, data fusion, and data analysis methods to develop a comprehensive platform to integrate and dynamically visualize data from ICSTs to support decision-makers (e.g., city planners) in attaining different transportation sustainability goals (in particular, mobility, safety, accessibility, and equity). The platform can aid decision-makers to develop innovative solutions for various transportation-related unique and contemporary problems, such as transit deserts and enabling societal benefits to be realized at their highest potential for various community stakeholders (e.g., residents, employees, city governance, travelers, etc.). In this context, the City of Peachtree Corners (PTC), GA, serves as a living lab for this study. PTC is part of the Atlanta metropolitan area and is the largest city in Gwinnett County, with a population of around 42,000. The study uses data obtained through our partnerships with PTC and Gwinnett County Transit (GCT), along with open-source data, to demonstrate the development of the visualization platform.

The remainder of this paper is organized as follows: Section II summarizes related work and gaps in literature. Section III provides methodological details. Section IV discusses potential applications along with a case study. Section V presents the conclusion.

II. RELATED WORK

Existing studies related to visualization of spatiotemporal transportation data mostly consider single-source data. Visualization platforms were usually developed to explore the spatiotemporal relationship of single-source data and analyze potential impact [1,2].

Recently, several studies have integrated transportation-related data from different sources to provide more consistent and accurate information than that provided by a single source. A key challenge is that transportation data is usually owned and operated by different organizations that may not share it with others. Hence, the data integrated in these studies are either open-source or managed by the same organization [3,4,5]. The effect of route choice, day of the week, and weather conditions on travel time variability was studied by using publicly available traffic data and presented using visualization methods [3]. With the focus on private and public transport, multiple-source data were integrated to create tools for mobility data analysis in [6]. Moreover, multiple attributes of transportation-related data, including space, time, and modes, have been utilized to investigate traffic pattern [4]. However, the data considered in these studies did not entail multi-timescale-related complexity.

In the context of smart cities, several studies have explored data integration and visualization methods to tackle transportation-related problems. Machine learning-based traffic visualization platform has given promising performance in processing and analyzing traffic data to solve congestion problems in smart cities [7]. The relationship between air quality parameters and traffic density was demonstrated using a GIS-based visualization method in [8].

The summary of the existing literature points to four key gaps in the context of smart cities. First, existing data visualization platforms have been developed to address specific problems, and hence cannot be effectively used by community decision-makers when faced with other transportation-related problems. For example, a platform developed for traffic surveillance cannot be directly used to address transit desert-related problems. Second, socio-demographic characteristics (e.g., age, income, households, etc.) influence the various transportation-related decisions (e.g., trip origins/destinations, mode choice, etc.). However, existing studies that developed data integration and visualization platforms did not consider socio-demographic data. Therefore, they cannot adequately identify disadvantaged population groups (e.g., people living in transit deserts, lower-income areas, etc.) and the associated transportation equity-related issues. Third, the few studies that consider multi-source data do not address multi-timescale data issues. Finally, most of these platforms are web-based where data from different sources are deposited and can be viewed on different web pages without the capability to visualize the data together. They cannot efficiently capture the interactions and relationships between different attributes from different data sets located on different web pages.

The next section describes the proposed methodology to address these gaps. It also illustrates how the challenge associated with integrating multi-timescale data can be addressed.

III. METHODOLOGY

This section describes the methodology followed to develop the dynamic visualization platform using 3 steps: data collection, data preprocessing and analysis, and data integration and visualization.

Data Collection. To address the first gap, a platform integrating a comprehensive list of transportation-related data sets is needed so that it can assist in solving virtually any transportation-related contemporary problems. To achieve this capability, we decided to collect a comprehensive list of transportation-related data sets associated with PTC. By leveraging our partnerships with PTC and GCT, we collected different private data sets. Open-source data portals, specifically the Georgia Department of Transportation (GDOT), OpenWeather, Foursquare, and HERE, were used to collect other open and paid public data sets.

To address the second gap, we collected socio-demographic data sets from open portals of the Census Bureau and Statistical Atlas. To provide decision-makers with a holistic perspective of PTC's transportation system, we collected nine city-level data sets. Figure 1 shows the data sets collected and used as input to the visualization platform, including source, key attributes, timescale, and spatial and temporal resolution characteristics. These data sets were selected based on PTC's transportation sustainability goals, which include mobility, safety, accessibility, and equity. Traffic speed, traffic count, and weather data sets are used to assess mobility. Accident data is used to assess safety. To measure accessibility, facility, transit, and autonomous vehicle (AV) shuttle data sets are used. Socio-demographic data is used to identify transportation equity issues within the city.

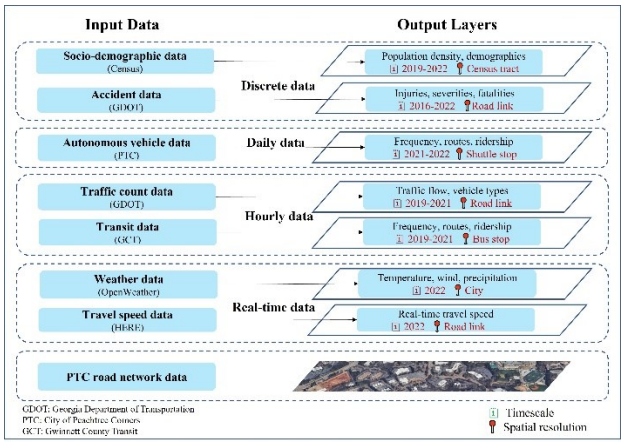


Figure 1. Input data (and sources) and key attributes in the dynamic visualization platform

Data Preprocessing and Analysis. From Figure. 1, it is evident that different data sets have different timescales, implying they are not compatible with each other. To address the challenge of integrating these multi-timescale data sets (i.e., the third gap mentioned in section II) and providing high-quality inputs to the visualization platform, data processing, and analysis was

done which consisted of three steps: (i) data exploration and characterization, (ii) data preprocessing, and (iii) data analysis.
Data exploration and characterization: The collected data sets have disparate formats (which is natural since these data sets came from different sources), including XLS (Excel Spreadsheet), CSV (Comma-Separated Value), and shapefiles (e.g., .shp, .dbf, etc.). Hence, it is critical to exhaustively explore each data set using different software that can handle these varying data formats to examine data quality. We examined these data sets using Python libraries (e.g., Pandas, CSV) and QGIS software. Table 1 summarizes the data formats and the software used to examine different data sets.

Table 1. Datasets and corresponding format

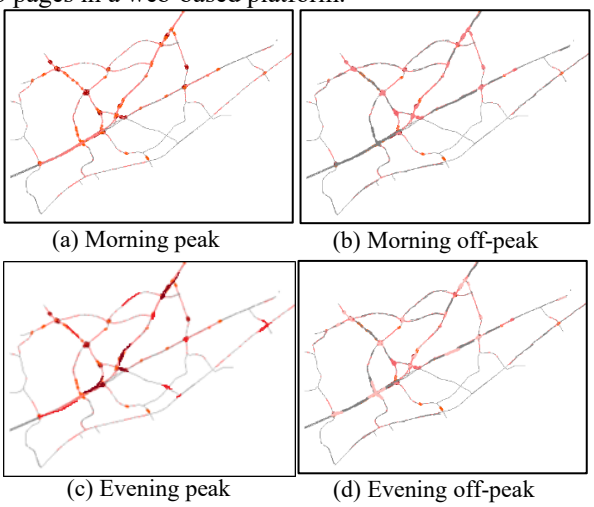
Data set	Format	Software	Data set	Format	Software
Socio-demographic	Shapefile	QGIS	Traffic count	CSV	Python
Accident	CSV	Python	Transit	XLS	Python
Facility	CSV	Python	Weather	Text	Python
AV	XLS	Python	Speed	CSV	Python

We inspected the characteristics of each data set by examining its features such as data attributes and temporal and spatial resolutions in each file, identifying and summarizing key characteristics (Figure 1) to be integrated and visualized on the platform.
Data preprocessing: Since multi-source, multi-timescale data sets typically have varying data units and uncertain quality (e.g., missing values, repeated values, etc.), data preprocessing was employed to ensure and enhance data quality. Data units across all data sets were modified as needed so that they become consistent with each other. Cells with repeated and missing values were either eliminated or imputed based on other available information. For example, the facility data set collected through Foursquare contains locations (longitude-latitude) and types (e.g., apartment complexes, schools, grocery stores, etc.) of the facilities throughout the city. However, there were instances where locations were missing for some facilities and facility types were missing for some locations. To handle such cases, a manual effort was made to validate the types and locations of those specific facilities.
Data analysis: Since different data sets have different timescales and temporal resolutions (e.g., daily, hourly, real-time), they reflect conditions at mismatched timescales and hence do not readily sync with each other. We handled this issue by selecting the same timescale for socio-demographic, AV, and transit data sets. Moreover, real-time data and historical data cannot be directly synced. For example, real-time traffic speed data and historical accident data cannot be synced without making some modifications. To address this challenge, data analytics tools such as distribution analysis and *k*-means clustering algorithm were employed to extract insights from historical data that are free from timescale-related influence. Considering the frequency and locations of traffic accidents can significantly vary during peak hours and off-peak hours, to dynamically visualize how accident occurrences may vary during a typical day, we decided to divide a day into peak and off-peak periods based on traffic speed data (higher speed is associated with off-peak hours and lower speed is associated with peak-hours). To determine the peak and off-peak periods for PTC, *k*-means clustering algorithm was applied to analyze the collected traffic speed data by assigning traffic conditions at different times to four clusters: morning peak, morning off-peak, evening peak, and evening off-peak. Within each cluster, traffic conditions are similar. The *k*-means clustering algorithm is as follows:

1: specify the number *k* of clusters to assign
2: Randomly initialize *k* centroids
3: **repeat**
4: **expectation**: Assign each point to its closest centroid.
5: **maximization**: Compute the new centroid (mean) of each cluster.
6: **until** the centroid positions do not change.

By applying spatial data mining, we analyzed the locations and counts of past traffic accidents to identify accident-prone zones along different road segments corresponding to each of the four clusters. Figure 2 shows the accident-prone zones using red color (the darker the color the higher the likelihood of accidents) and how these zones change during a typical day.

Data Integration and Visualization. It is vital to integrate processed data in a flexible yet unified manner and to interactively display them to provide decision-makers (e.g., city planners) with efficient and effective support for addressing various smart city problems. Therefore, the visualization platform should enable decision-makers to load the necessary data sets based on the problem context and relevant objectives, produce integrated data layers, and generate interactive spatial and temporal visualizations for the periods of interest. These capabilities, however, are largely missing in the existing web-based platforms as explained in the fourth and final gap in section II. To fill this gap, instead of developing a web-based platform, we developed a Python-based platform in which all the data sets are integrated at the same location thus eliminating the limitations associated with data sets located on different web pages in a web-based platform.



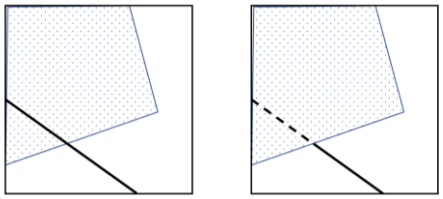
(a) Morning peak (b) Morning off-peak
(c) Evening peak (d) Evening off-peak

Figure 2. Accident-prone zones during the four different periods on a typical day

A python script was developed that can automatically reformat the real-time data (for example, HERE API data), and then

analyze and visualize it on a distinct layer. When selected data is loaded into the platform, each data set is represented as distinct layers using points (transit stops, AV shuttle stops, traffic counts), lines (accident-prone zones, AV shuttle route, transit routes, and road segments), and polygons (socio-demographic data) based on their characteristics. Python libraries HoloViews and Bokeh were used for data integration and visualization using distinct layers. Standard overlay operations (i.e., the placement of two or more distinct data layers on top of one another to create a more complex layer) were performed to generate cohesive data layers. While most of the overlay operations were straightforward (e.g., point-on-point, line-on-line), overlay operations for generating demand links (defined in the next subsection) were more involved as discussed below.

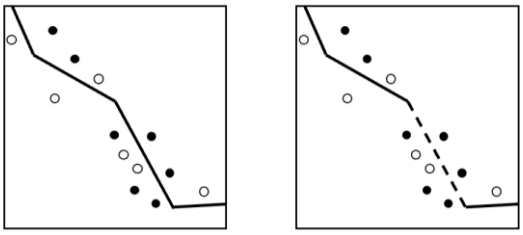
To visualize trips made within the city, trip origins and destinations as well as trip-starting links and ending links are needed. These links are referred to as demand links because these links create demands in the road network. Even though trip origin-destination data is available, data for demand links is absent in the original data source. Hence, we had to generate and visualize them using the available data sets. Since trips are associated with people moving between different facilities, we used the facility layer and the road network layer to generate the demand links. Line-on-polygon and point-on-line overlay operations were used in this regard. Figure 3(a) shows an apartment (the dot-patterned polygon) on the facility layer and a neighborhood road on the road network layer. The original road is divided into two segments by a line-on-polygon overlay procedure that preserves the polygon features on the overlapping portions of lines in the output layer. Both segments inherit the properties of the traffic link, while the segment overlaid by the apartment (shown using a dashed line in Figure 3(b) inherits demand-related properties from the polygon in the facility layer. Thus, the dashed segment corresponds to a demand link.



(a) (b)

Figure 3. Generating demand links using line-on-polygon overlay operations

Given the inevitability of misaligned coordinates on different layers, overlay operations are not as simple as combining them. Very often, aggregation and clustering were used to align adjacent lines or polygon edges within a predetermined tolerance. Figure 4 illustrates another instance of demand link generation based on aggregation, clustering, and point-on-line overlay operation. In this scenario, most of the facilities are single storefronts. Hence, they are depicted as points rather than polygons. Specifically, the black dots represent facilities that induce a high number of trips (e.g., restaurants, grocery stores, etc.), whereas the facilities associated with the white dots do not induce a significant number of trips (e.g., print services, storage services, etc.). Despite the proximity of these facilities to the road, the dots do not align with the lines in the traffic network layer. Consequently, we are unable to determine the number of demand-related facilities for each link. To address this issue, only higher demand-inducing facilities (i.e., the black dots) are selected and grouped with their nearest links. For a link, if the number of nearby demand-inducing facilities exceeds a threshold, the output layer labels the corresponding link as a demand link (shown by the dashed line in Figure 4(b). Other overlay operations, such as polygon-on-polygon overlay and polygon-on-point overlay, are also applied in practice depending on how data sets are represented in corresponding distinct data layers.



(a) (b)

Figure 4. Generating demand links using point-on-line overlay operation

Additionally, a feature was added to the visualization platform that allows users to play animations at a chosen speed and drag a slider to efficiently examine the spatiotemporal interactions across different data layers. Figure 5 depicts how the proposed platform may assist in visualizing the effects of weather on traffic speeds and how the accident-prone zones are correlated with traffic speeds.

Figure 5. An instance from the dynamic visualization platform, capturing traffic and weather states, and illustrating the accident-prone zones at a specific time

IV. APPLICATIONS

A city's transportation system consists of its transportation demand (which depends on the socio-demographic characteristics of the population) and transportation supply (e.g., available travel modes, transportation infrastructure, etc.). Therefore, to effectively address different transportation-related contemporary problems, the decision-makers need to consider data associated with both demand and supply. In the context of smart cities, these data are generally available from the ICSTs. However, the multi-source, multi-timescale nature of these disparate data sets makes it challenging for the

decision-makers to consider different data sets relevant to a specific problem together to efficiently determine the interactions and relationships between them. Determining these interactions is of crucial importance while identifying solution options for a specific problem. By developing the dynamic visualization platform, we have generated capabilities to support PTC in efficiently determining the aforementioned interactions by combining multiple layers on the visualization platform. This capability has the potential to aid the city in achieving its transportation sustainability goals.

A Case Study. A case study on how the platform can aid in addressing an emerging transportation problem in PTC is discussed hereafter. PTC has a 3-mile AV test track (Figure 6(a)), through which two AV shuttles (Figure 6(b)) serve the city residents, workers, and visitors by providing access to shops, office buildings, etc. This shuttle service is being used as a proof of concept, meaning the shuttles are serving a small number of people within a limited catchment area. The city planners want to commercialize the AV shuttle service by extending the current route and deploying more shuttles. One of the associated planning questions is: in which regions of the city should the AV shuttle service be extended to enhance the city's transportation accessibility and equity? In the absence of the visualization platform, the city planners' preliminary plan was to connect the existing route to the downtown of the city by extending the route to the north. After developing the platform, we decided to take a systematic approach to find the answer to the aforesaid question and check whether the preliminary plan is the most effective one. For this, we first need to understand the current accessibility level of different regions of the city and the socio-demographic characteristics of the residents living in those regions. The developed visualization platform can adequately aid in this regard.

Figure 6. The current AV shuttle route and the AV shuttles

To assess PTC's current accessibility and equity states, we can combine the following layers: transit, AV shuttle, facility, and median income. Figure 7(a) shows the visualization platform with three activated layers: median income, AV shuttle, and transit. Figure 7(b) shows the facility sub-layers combined with the income layer. From these figures, we make the following observations:

- High-income population groups mostly live in the north of the city, whereas lower-income groups live in the south.
- Almost all the apartment complexes are situated in the south-east and south-west of the city. There are no apartment complexes in the north of the city.
- Three major attractions can be identified: the two purple rectangles in Figure 7(b) (people living in the south of the city shop there) and the downtown (people living in the north of the city shop there).
- The current AV shuttle route is well-connected to transit route 35.

In Figure 7(c), we have combined all the layers shown in Figure 7(a) and 7(b) together. We identified two regions of interest that are shown using a blue oval and a red oval in the figure.

Blue oval: It covers areas in the south of the city. Around 45% of the total population lives there. Most of the apartment complexes are situated in this region. One major attraction is situated there and another one is nearby (in the north-west of this region). This region mostly contains lower-income population groups. A significant amount of area inside this region does not have a transit service indicating the presence of transit deserts.

Red oval: It covers areas in the north-west and north-east parts of the city. Around 35% of the total population lives there. The residents there fall into the higher-income groups. There are no apartment complexes, and this region is largely disconnected from existing transit routes, implying that the whole region is a transit desert. It is logical to assume that residents in this region use private vehicles for daily mobility needs, implying higher congestion. The downtown is situated in this region.

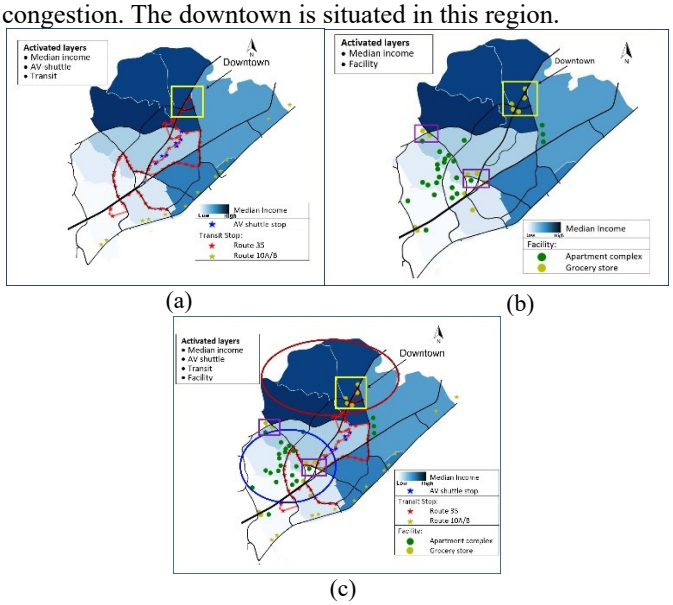


Figure 7. Screenshots from the visualization platform

Based on the current AV route location and the above observations, we have made the following recommendations to PTC:

- By extending the AV shuttle service in the region corresponding to the blue oval, the city can enhance access to the major attractions (especially to the one situated in the north-west of this region) as well as can enhance mobility and accessibility for the lower-income groups who may not own private vehicles, thus addressing equity issues.
- By extending the AV shuttle service in the region corresponding to the blue oval, the city can aim at enhancing transit accessibility (note that the AV shuttle is a transit service) for the people living in that region. Moreover, the city may extend the AV shuttle route in this region in such a way that the route also connects the

existing transit route. By doing so, transit desert-related issues can be mitigated. Even though the residents in this region have higher income and own private vehicles, providing them with transit options to move within the city has a high potential to promote travel sustainability (e.g., people switching modes from private vehicles to AV shuttle and transit).

Based on these recommendations, the city is currently working towards introducing the AV shuttle service in the aforementioned regions. It is important to compare these recommendations with the city's preliminary AV shuttle route extension plan (i.e., extending the route to the north to connect the downtown). Clearly, by leveraging the visualization platform, well-informed decisions can be made, as illustrated above, that have a higher likelihood of meeting the city's transportation sustainability goals, which may not have been possible in the absence of the visualization platform.

V. CONCLUSION

In this study, we have developed a multi-source, multi-timescale transportation-related data integration and visualization platform to support city planners in achieving different transportation sustainability objectives. The platform is dynamic in the sense that it can capture and visualize the spatiotemporal interactions and relations between different data sets. Although the dynamic visualization platform has been developed for a small city, the methodology for the platform can be scaled to larger-sized cities using appropriate computing resources. Hence, it is customizable and can be transferred to other smart cities characterized by disparate data sources and data owners.

The main contributions of this work are threefold. First, by integrating a comprehensive list of transportation data sets, we have made the visualization platform capable of aiding smart city decision-makers in addressing virtually any transportation-related contemporary problems. This capability fills the gap in existing similar visualization platforms which were developed to address specific problems. Second, the platform is capable of effectively identifying transportation equity issues within the cities. By integrating the socio-demographic data set, a data set that has been largely ignored in existing visualization platforms, we have generated capabilities to support city decision-makers identify different disadvantaged population groups. Hence, the platform has the potential to help smart cities move towards their transportation equity-related goals by identifying different equity issues and taking necessary steps to benefit disadvantaged population groups. Third, as discussed in section II, most of the existing visualization platforms are web-based which inhibits analyzing interactions and relations between different attributes from different data sets located on different web pages. This largely limits the capability of taking well-informed decisions by city decision-makers. Since the developed visualization platform can show any subset of data sets together, it does not suffer from this aforementioned limitation.

In term of future work, since the platform is built using Python and Jupyter notebook, it can be challenging for some users who are not familiar with the language. To improve the platform's accessibility, future work will be focused on creating a more user-friendly interface.

ACKNOWLEDGMENT

This research is funded through the National Science Foundation's Smart and Connected Communities (S&CC) program, award number 2125390. The authors would like to thank personnel of community partners PTC and GCT, and especially PTC's Assistant City Manager Brandon Branham, for their assistance throughout the data collection process. Any errors or omissions remain the sole responsibility of the authors.

REFERENCES

[1] Ferreira, N., Poco, J., Vo, H. T., Freire, J., & Silva, C. T. Visual exploration of big spatio-temporal urban data: A study of New York City taxi trips. *IEEE transactions on visualization and computer graphics*, *19(12),* 2149-2158, 2013.

[2] Xin, R., Ai, T., Ding, L., Zhu, R., & Meng, L. Impact of the COVID-19 pandemic on urban human mobility-A multiscale geospatial network analysis using New York bike-sharing data. *Cities*, *126*, 103677, 2022.

[3] Verovšek, Š., Juvančič, M., Petrovčič, S., Zupančič, T., Svetina, M., Janež, M., ... & Moškon, M. An integrative approach to neighbourhood sustainability assessments using publicly available traffic data. *Computers, Environment and Urban Systems*, *95*, 101805, 2022.

[4] Ding, C., Wang, Y., Yang, J., Liu, C., & Lin, Y. Spatial heterogeneous impact of built environment on household auto ownership levels: Evidence from analysis at traffic analysis zone scales. *Transportation Letters*, *8*(1), 26-34, 2016.

[5] He, S., Zhang, J., Cheng, Y., Wan, X., & Ran, B. Freeway multisensor data fusion approach integrating data from cellphone probes and fixed sensors. *Journal of Sensors*, 2016, 7269382, 2016.

[6] Fortini, P. M., & Davis Jr, C. A. Analysis, integration and visualization of urban data from multiple heterogeneous sources. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Advances on Resilient and Intelligent Cities (pp. 17-26)*, 2018.

[7] Zhang, Y., Wang, H., & Wang, X. (2022). Research on the improvement of transportation efficiency of smart city by traffic visualization based on pattern recognition. *Neural Computing and Applications*, 1-14, 2022.

[8] Bovkir, R., & Aydinoglu, A. C. Big urban data visualization approaches within the smart city: gis-based open-source dashboard example. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences,* 2021