Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities

Unsupervised origin-destination flow estimation for analyzing COVID-19 impact on public transport mobility

Lan Zhang, Kaijian Liu *

Dept. of Civil, Environmental and Ocean Engineering, Stevens Institute of Technology, 1 Castle Point Terrace, Hoboken, NJ 07030, United States of America

ARTICLE INFO

Keywords:
COVID-19
Urban mobility
Origin-destination flow estimation
Unsupervised machine learning

ABSTRACT

The outbreak of COVID-19 caused unprecedented disruptions to public transport services. As such, this paper proposes a methodology for analyzing COVID-19 impact on public transport mobility. The proposed methodology includes: (1) a new unsupervised machine learning (UML) method, which utilizes a decoder-encoder architecture and a flow property-based learning objective function, to estimate the origin-destination (OD) flows of public transport systems from boarding-alighting data; and (2) a temporal-spatial analysis method to analyze OD flow change before and during COVID-19 to unveil its impact on mobility across time and space. The validation of the UML method showed that it achieved a coefficient of determination of 0.836 when estimating OD flows using boarding-alighting data. Upon the successful validation, the proposed methodology was implemented to analyze the impact of COVID-19 on the mobility of the New York City subway system. The implementation results indicate that (1) the rise in the number of weekly new COVID-19 cases intensified the impact on the public transport mobility, but not as strongly as public health interventions; and (2) the inflows to and outflows from the center of the city were more sensitive to the impact of COVID-19.

1. Introduction

Public transport is an indispensable means for people to access essential resources such as goods, services, and opportunities. For example, in 2019 alone, the New York City subway system carried over 1.6 billion passengers to their desired destinations (MTA, 2019). However, normal public transport use in many urban areas was disrupted by the global outbreak and rapid spread of the coronavirus disease (COVID-19). On one hand, in the fight against this unprecedented pandemic, many public health interventions were undertaken by public transport authorities to control the transmission of COVID-19 and slow down its spread (Chakraborty & Maity, 2020). Although they were effective, some of these interventions inevitably disrupted public transport services, such as "stay-at-home" orders, temporal shutdown of public transport services, and social distancing on public transport vehicles (NYC, 2020; Cuomo, 2020; CDC, 2022c). On the other hand, the pandemic changed the psychology of people (e.g., injecting the fear of being infected by COVID-19) and, subsequently, their motives and decision making to move around the city (e.g., reducing unnecessary trips and changing transport modes). All these changes disrupted the normal patterns of public transport use (Humagain & Singleton, 2021; Teixeira et al., 2021), imposing restrictions on the mobility of people and limiting their ability to access essential resources to maintain life stability and well-being. There is, therefore, an evident need to analyze and understand how COVID-19 affected public transport mobility across time and space. Such analysis and understanding would allow decision makers at local and federal agencies to make effective recovery plans and long-term policies to mitigate the impact of the pandemic and build public transport resilience to combat potential pandemics of similar scale in the future.

To address this need, this paper proposes a new methodology that allows for estimating the origin-destination (OD) flows of public transport systems and analyzing OD flow change before and during COVID-19, to unveil the impact of the pandemic on public transport mobility. The proposed methodology includes two main components: OD flow estimation, and temporal-spatial analysis. However, OD flow estimation, as the first step toward the impact analysis, is particularly challenging. Estimating the OD flows of public transport systems requires capturing the number of passengers traveling from an origin station to a destination station over a given period of time (Barbosa-Filho et al., 2018). The challenges in estimating OD flows arise from two main reasons (further discussed in Section 2). First, existing methods for

* Corresponding author.
E-mail addresses: lzhang29@stevens.edu (L. Zhang), Kaijian.Liu@stevens.edu (K. Liu).

https://doi.org/10.1016/j.cities.2024.105086
Received 26 January 2023; Received in revised form 10 April 2024; Accepted 30 April 2024

collecting OD flow information are limited in covering diverse ridership groups of a public transport system, making the use of such biased information to estimate and provide a full view of OD flows challenging. Second, although boarding-alighting data provide comprehensive coverage of all public transport riders, they only capture the total number of riders entering and exiting a transport station and do not offer OD flow information in their original form (e.g., information about the number of riders, out of the total number of exits at a destination station, that come from a specific origin station).

To address the aforementioned challenges, the methodology in this paper offers a novel unsupervised machine learning (UML) method to learn from boarding-alighting data to effectively estimate OD flows. The proposed UML method does not require historical OD flows (which are typically not available for most public transport systems) as labels to supervise the learning. Instead, it leverages a proposed decoder-encoder architecture to achieve unsupervised learning: (1) a decoder that leverages a multilayer perceptron (MLP) mixer to decode input boarding-alighting data into estimated OD flow matrices, and (2) an encoder to encode each estimated OD flow matrix to reconstruct its corresponding input boarding-alighting data. To ensure estimation accuracy, the proposed UML method uses a flow property-based learning objective function to minimize reconstruction errors between input and reconstructed boarding-alighting data. This paper presents the proposed methodology for analyzing COVID-19 impact on public transport mobility. Furthermore, it presents and discusses the experiments and experimental results for validating the OD flow estimation method, and implementing the methodology in analyzing the pandemic impact in the New York City.

2. State of the art and knowledge gaps in OD flow estimation

There is an evident need to acquire public transport system OD flows to support the analysis of COVID-19 impact on public transport mobility across time and space. Despite the importance of research efforts that have been undertaken in the area of OD flow analysis, existing methods are still limited in estimating OD flows of public transport systems. Two primary knowledge gaps are identified.

First, there is a lack of methods that can effectively estimate public transport system OD flows. On one hand, a branch of research efforts have focused on leveraging the following methods to directly collect information about the origin and destination locations of each trip from public transport riders for OD flow estimation, including travel surveys (e.g., Katranji et al., 2019; Ali, 2022; Jin et al., 2022), call detail records (e.g., Zagatti et al., 2018; Mamei et al., 2019; Fekih et al., 2020), social media (e.g., Ebrahimpour et al., 2020; Osorio-Arjona & García-Palomares, 2019; Pourebrahim et al., 2018), and smart cards (e.g., Hussain et al., 2021; Li et al., 2018; Pelletier et al., 2011). However, such methods are limited in OD flow estimation: (1) methods that rely on surveys, call detail records, and social media are often associated with sampling bias, i.e., riders of certain ridership groups are not sufficiently sampled or may not even get sampled. For example, riders making short-distance trips are less likely to complete a travel survey before alighting and hence are not sampled (FHWA, 2017), not all riders make phone calls when using public transport services (Jiang & Luo, 2022), and only around 2 % of tweets on social media are tagged with location information (Twitter, 2022); and (2) smart card-based methods appear to be promising to mitigate such sampling bias because all riders need to use a smart card to swipe in and out of a transport system for fare collection, which allows collecting information about boarding and alighting stations from all riders. But, not all public transport systems currently use distance-based fare collection. Some are flat-fare or entry-only systems and thus do not collect destination information for fare calculation, limiting the applicability of smart card-based methods for certain public transport systems such as the Boston subway system, the New York City subway system, among others (Egu & Bonnel, 2020; Hussain et al., 2021; Zaragozí et al., 2021). On the other hand, another track of research efforts have focused on using optimization methods for OD flow estimation, including the maximum entropy method (e.g., Bell, 1983; López-Ospina et al., 2021), the maximum likelihood estimation method (e.g., Spiess, 1987; Yang & Rakha, 2017), the Bayesian inference method (Maher, 1983; Pitombeira-Neto et al., 2018), and the generalized least squares method (Cascetta, 1984; Guo et al., 2019). These optimization methods are typically used to account for the sampling bias by minimizing the variances between the target OD flow matrix and the estimated matrix, based on observed link counts (Bera & Rao, 2011; Dey et al., 2020; Wang et al., 2016). Although using these optimization methods can theoretically improve the quality of estimated OD flows, they require link counts as prior information for the optimization (Jeong & Park, 2021; Wang et al., 2016). However, link counts are generally not available in public transport systems, making such optimization methods limited in effectively estimating OD flows (Dey et al., 2020; Papageorgiou & Varaiya, 2009).

Second, there is a lack of methods that can leverage boarding-alighting data to enable effective OD flow estimation. Boarding-alighting data capture the total number of riders entering and exiting a transport station (Blum et al., 2010; Sharic et al., 2021; Sun et al., 2021). Compared to the aforementioned sources of data, boarding-alighting data provide comprehensive coverage of all riders (Ge et al., 2021), in addition to being publicly available and free of privacy concerns. For example, subway systems across U.S. cities rely on turnstiles to administrate subway entrances and exits, thereby collecting boarding-alighting/turnstile data for all the riders (MTA, 2022). Despite the advantage of boarding-alighting data, there is currently a lack of methods that can analyze boarding-alighting data for effective OD flow estimation. Boarding-alighting data capture the total number of entrances and the total number of exits at a station (*station A*), but do not capture the information about the origin and destination flows. For example, among the total exits at *station A*, which is the number of riders entering a specific station (*station B*) and exiting from *station A*? There is, therefore, a need to estimate OD flows from boarding-alighting data. However, existing methods in OD flow studies mainly focus on OD flow prediction, rather than estimation. OD flow prediction learns from previous OD flows to predict future OD flows of a given transport system, using supervised ML methods such as support vector machines (Wang et al., 2018; Luo et al., 2019; Tang et al., 2019; Toan & Truong, 2020), autoencoders (Moussavi-Khalkhali & Jamshidi, 2019; Zhao et al., 2019; Essien et al., 2020; Hou et al., 2021), graph convolutional networks (Hu et al., 2020; Ke et al., 2021; Shi et al., 2020; Wang et al., 2019), and convolution-based long short-term memory networks (Gu & Duan, 2022; Khalil et al., 2021; Selvarajah et al., 2020; Zhao et al., 2020). But, for public transport systems, there are typically no previous OD flows to learn how to capture the relationships between boarding-alighting data and OD flows (Mo et al., 2020; Ros-Roca et al., 2022; Yang et al., 2018) – limiting the applicability of these supervised ML-based methods in estimating OD flows for public transport systems. The lack of available OD flows to learn from demands an unsupervised learning method for public transport system OD flow estimation using boarding-alighting data, which is, however, missing from the current literature (Liu et al., 2019; Zargari et al., 2021).

3. Proposed methodology

This paper proposes a new methodology to enable the analysis of COVID-19 impact on public transport mobility. As shown in Fig. 1, the proposed methodology uses a new UML-based method to effectively estimate OD flows of public transport systems from boarding-alighting data, and utilizes a temporal-spatial analysis method to compare estimated OD flows before and during COVID-19 to derive insights on how the pandemic affected public transport mobility across time and space.

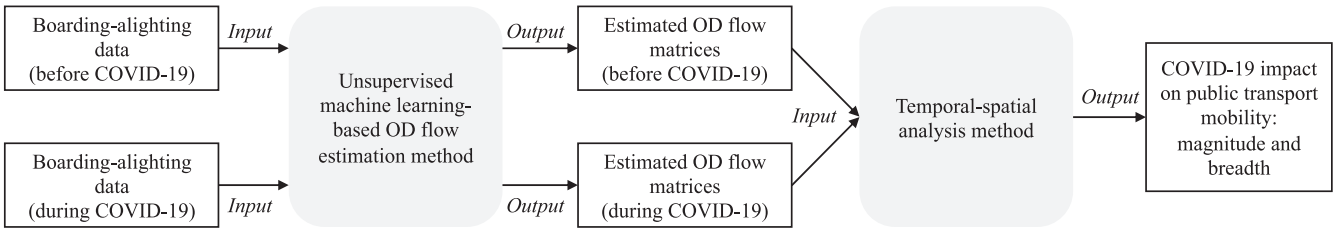


Fig. 1. Proposed methodology for analyzing COVID-19 impact on public transport mobility.

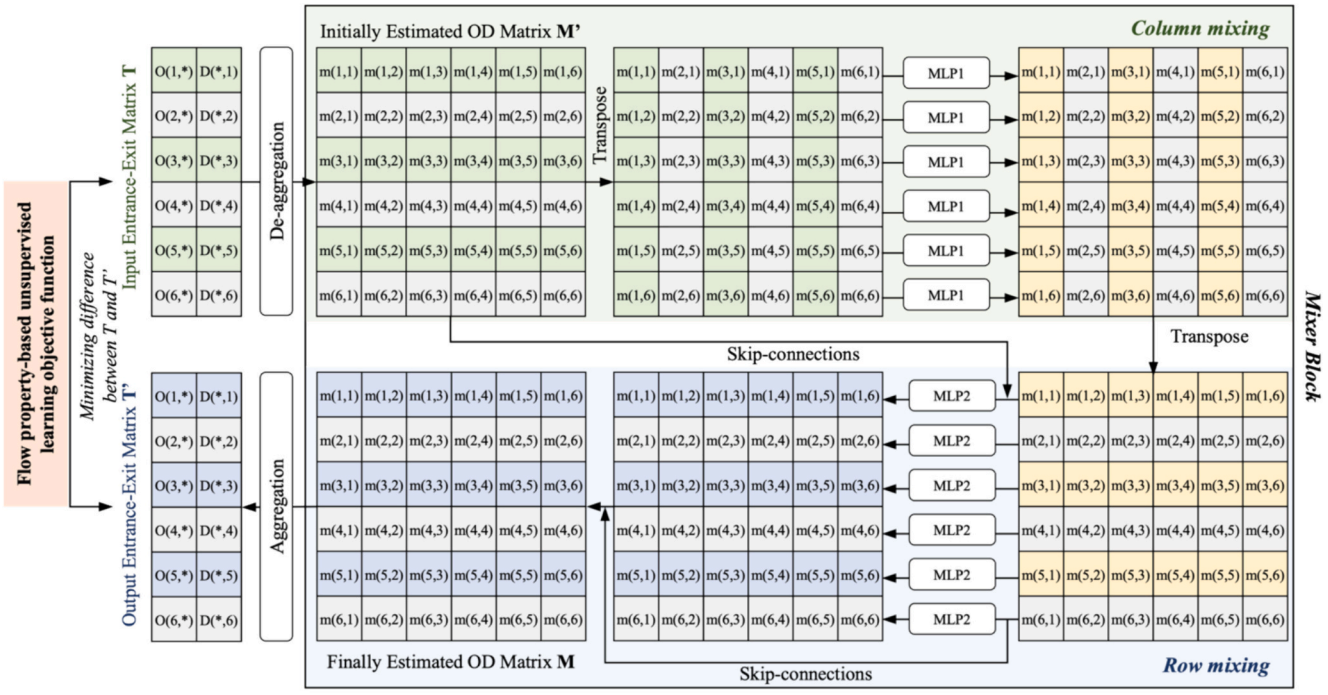
3.1. UML-based OD flow estimation

As shown in Fig. 2, the input to the proposed estimation method is daily boarding-alighting (i.e., turnstile) data, and the output is the estimated OD flow matrix. Daily turnstile data capture the total daily entrances and exists at each station in a public transport system, in form of an $N \times 2$ matrix where N is the number of unique stations, each cell in the first column of the matrix captures the total entrances at a station, and each cell of the second column captures the total exits at a station. Daily turnstile data do not capture OD flow information because the total number of exits of a station (*station A*) is only an aggregated count of all riders that enter from all the other stations and exit from *station A*, without capturing the numbers of riders traveling between each of the other stations and *station A*, i.e., the origin-destination flows. An OD flow matrix shows the number of riders between each station pair, in form of an $N \times N$ matrix with each cell capturing the number of riders between an entering station and an exiting station. The proposed estimation method aims to learn from turnstile data to estimate public transport system OD flows in an unsupervised way: it does not require historical OD flow matrices as labels to supervise the learning from turnstile data for the estimation. Instead, it leverages (1) an unsupervised decoder-encoder architecture to decode input turnstile data into estimated OD flow matrices by learning to distribute the total number of entrances and exits at each station across all the other stations, and encode each estimated OD matrix to reconstruct its corresponding input; and (2) a flow property-based learning objective function to minimize the errors between the input and reconstructed turnstile data to ensure the accuracy of the estimated OD flow matrices, thereby realizing the unsupervised learning/estimation.

3.1.1. Decoder-encoder learning architecture

A decoder-encoder learning architecture is proposed to estimate public transport system OD flows from turnstile data. The decoder architecture takes an $N \times 2$ entrance-exit turnstile matrix (**T**) that stores turnstile data as input, and aims to learn to transform the input matrix **T** into an $N \times N$ OD flow matrix (**M**).

The proposed decoder architecture includes an input layer, a de-aggregation layer, and a set of multilayer perceptron (MLP) mixer blocks. The input layer takes matrix **T** as input and passes it to a de-aggregation layer, which transforms **T** into an initially estimated $N \times N$ OD flow matrix (**M′**), where each element $m'(i, j)$ in **M′** represents an initially estimated number of riders traveling from stations i to j, $\forall i = 1, \ldots, N$, and $\forall j = 1, \ldots, N$. The de-aggregation layer does not involve ML operations but uses the maximum entropy method (Anstreicher et al., 1999; Bertsimas & Yan, 2018; Dubiner & Singer, 2011) to conduct the transformation because of two reasons. First, it allows for generating an initial, good-enough OD matrix to avoid random OD matrix initialization during the learning process, which can help prevent the learning



Note: *O(i, *) = Total number of entrances at origin station i; D(*, i) = Total number of exits at destination station i; m(i, j) = origin-destination flow between stations i and j; MLP = multilayer perceptron.*

Fig. 2. Proposed unsupervised machine learning-based method for public transport system origin-destination (OD) flow estimation.

from being stuck in local minima and facilitate learning convergence. Second, it can mathematically estimate $m'(i,j)$ in a closed form, without the need to train additional ML model parameters, to reduce the complexity of the decoder. According to the maximum entropy method, $m'(i,j)$ can be estimated as per Eq. (1), where $t(i,1)$ is the i^{th} cell in the first column of matrix **T** and represents the total number of entrances to station i, and $t(j,2)$ is the j^{th} cell in the second column of matrix **T** and represents the total number of exits from station j.

$$m'(ij) = t(i1)\frac{t(j2)}{\sum_{j=1}^{N} t(j2)}, \forall i = 1 \text{ to } N \text{and} \forall j = 1 \text{ to } N \tag{1}$$

The initially estimated $\mathbf{M}'$ is connected to a set of MLP mixer blocks to learn a final OD flow matrix **M**. MLP mixers are needed to refine the initially estimated matrix because the entropy-based estimation tends to lead to identical OD flow distributions, which does not reflect the OD flow distributions of real-world public transport systems and would lead to incorrect flow estimation (Bertsimas & Yan, 2018). MLP mixer was utilized as the learning backbone of the decoder because of three main reasons. First, as shown in Fig. 2, an MLP mixer block includes a column-mixing layer, which allows for capturing the interactions between a destination station and all of the origin stations to better estimate each column in **M**. Second, it also includes a row-mixing layer, which allows for capturing the interactions between an origin station and all of the destination stations to better estimate each row in **M**. Third, it uses skip-connections to avoid gradient vanishing in the learning process, without adding extra parameters or computational complexity (He et al., 2016; Liu et al., 2020). In the proposed estimation method, the column-mixing layer uses a set of MLPs – each learns from a column of $\mathbf{M}'$ – to capture the interactions between a destination station and all of the origin stations. Each MLP includes two fully-connected linear layers and a Gaussian error linear unit (GELU) nonlinear layer (Tolstikhin et al., 2021). The row-mixing layer is similar to the column-mixing layer, but uses a set of MLPs to learn from the rows of the matrix (created by adding $\mathbf{M}'$ and the output matrix of the column mixing using a skip-connection) to capture the interactions between an origin station and all of the destination stations, resulting in a finally estimated **M**. Because the number of mixer blocks affects the performance of OD flow estimation, different numbers of MLP mixer blocks were tested, and the optimal number was selected based on the testing results presented in Section 4.1.2.

The finally estimated OD matrix **M** is connected to the encoder, which reconstructs the output entrance-exit matrix $\mathbf{T}'$ by aggregating **M**. Unlike the decoder, the encoder does not include any learning process, because $\mathbf{T}'$ can be reconstructed by aggregating corresponding columns and rows in **M** as per Eqs. (2.1) and (2.2), where $t'(i,1)$ is the i^{th} cell in the first column of reconstructed $\mathbf{T}'$ and represents the total number of estimated entrances to station i; $t'(j,2)$ is the j^{th} cell in the second column of $\mathbf{T}'$ and represents the total number of estimated exits from station j, and $m(i,j)$ is an element in **M** and represents the estimated number of riders traveling from stations i to j.

$$t'(i1) = \sum_{j=1}^{N} m(ij), \forall i = 1 \text{ to } N \tag{2.1}$$

$$t'(j2) = \sum_{i=1}^{N} m(ij), \forall j = 1 \text{ to } N \tag{2.2}$$

3.1.2. Flow property-based learning objective function

The learning objective of the proposed decoder-encoder architecture is to minimize the error between **T** and its corresponding reconstructed $\mathbf{T}'$ to ensure the accuracy of the estimated **M**. The proposed learning objective function for assessing the reconstruction error is defined in Eq. (3), where $D1$ is the total of the entrance-normalized absolute difference between each element in **T** and the corresponding element in $\mathbf{T}'$, $D2$ is the total of the exit-normalized absolute difference between each element in **T** and $\mathbf{T}'$, $S1$ represents the symmetry property, $S2$ represents the smoothness property, λ is a weight that controls the contribution of the flow properties in regulating the estimation of OD flows, and other notations follow those defined in Eqs. (1)–(2.2). Furthermore, the learning objective function constrains that no riders enter and exit at the same station to follow the established convention (Kumar et al., 2019).

$$\begin{aligned} &min\left((1-\lambda)(D1+D2)+\lambda(S1+S2)\right) \\ &= min\left((1-\lambda)\left(\sum_{i=1}^{N}\left(\frac{|t(i,1)-t'(i,1)|}{t(i,1)}\right)+\sum_{j=1}^{N}\left(\frac{|t(j,2)-t'(j,2)|}{t(j,2)}\right)\right)+\lambda(S1+S2)\right) \end{aligned} \tag{3}$$

Subject to

$$m(ij) = 0, \quad when\ i = j, \forall i = 1 \text{ to } N, and \forall j = 1 \text{ to } N \tag{3.1}$$

$$m(i,i) = 0, \forall i = 1 \text{ to } N \tag{3.2}$$

Only minimizing $D1$ and $D2$ cannot provide a unique estimation of the **M** from **T** because several OD flow matrices can be aggregated into the same **T**, as long as $\sum_{j=1}^{N} m(i,j) = t'(i,1) = t(i,1)$ and $\sum_{i=1}^{N} m(i,j) = t'(j,2) = t(j,2)$. Therefore, $S1$ and $S2$ are included in Eq. (3), as origin-destination flow properties to be satisfied, to achieve a unique optimal estimation of **M**. The symmetry property $S1$ assumes that riders of public transport systems tend to make round trips (Lin et al., 2019; Hwang et al., 2020; Abbaas & Ventura, 2021; Zhang et al., 2022): for any pair of stations i and j, the number of riders traveling from i to j should be approximately equal to the number of riders traveling from j to i. Consequently, an OD flow matrix for a public transport system is approximately symmetric, i.e., $m(i,j) \cong m(j,i)$. $S1$ is defined in Eq. (4), where $AVG = (1/N) \times \sum_{i=1}^{N} t(i,1)$.

$$S1 = \sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{m(i,j)}{AVG} - \frac{m(j,i)}{AVG}\right| \tag{4}$$

The smoothness property $S2$ assumes that neighboring stations in a public transport system tend to have similar OD patterns (Lu et al., 2020; Miao et al., 2022; Tian et al., 2018; Zhang et al., 2020). This assumption indicates that (1) the OD flow patterns for station pair i and j are similar to that for pair k and j, when i and k are neighboring stations, i.e., $m(i,j)/t(i,1) \cong m(k,j)/t(k,1)$; and (2) the OD flow patterns for station pair i and j are similar to that for station pair i and l, when j and l are neighboring stations, i.e., $m(i,j)/t(j,2) \cong m(i,l)/t(l,2)$. $S2$ is defined in Eq. (5), where $\gamma(i)$ is the nearest neighboring station of i, $Dist(i,\gamma(i))$ is the distance between i and $\gamma(i)$ measured by geographical distance, $\gamma(j)$ is the nearest neighboring station of j, and $Dist(j,\gamma(j))$ is the geographical distance between j and $\gamma(j)$. The reciprocal of the distance between the stations and their nearest neighboring stations indicates that the neighboring stations with longer distances are less similar than the

neighboring stations with shorter distances.

$$S2 = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\left|\frac{m(i,j)}{t(i,1)} - \frac{m(\gamma(i),j)}{t(\gamma(i),1)}\right|}{Dist(i,\gamma(i))} + \frac{\left|\frac{m(i,j)}{t(j,2)} - \frac{m(i,\gamma(j))}{t(\gamma(j),2)}\right|}{Dist(j,\gamma(j))}\right) \quad (5)$$

3.2. Temporal-spatial analysis

The input to the temporal-spatial analysis method is an OD flow matrix during the pandemic and its corresponding matrix before the pandemic (both matrices were estimated in Section 3.1), and the output is the COVID-19 impact on public transport mobility across time and space. First, from the temporal perspective, the analysis method compares the estimated OD flow for each station pair during the pandemic to its corresponding OD flow before the pandemic, based on the rate of OD flow change. As defined in Eq. (6.1), the rate of OD flow change $C^t(i,j)$ is calculated as one minus the ratio of OD flow during the pandemic to that before the pandemic, where $M_{i,j}^{t,1}$ is the estimated daily OD flow for station pair i and j on date t of the year before the pandemic, and $M_{i,j}^{t,2}$ is the estimated daily OD flow for the same station pair on the same day of the year during the pandemic. For example, if a pair of stations had a daily OD flow equal to 1000 ridership for March 25th, 2019, and a daily OD flow equal to 90 ridership for the same day of 2020 (during the pandemic), the change rate is then calculated as 1 – (90/1000) = 91 %. Second, from the spatial perspective, the analysis method uses the ratio of station pairs between an origin and a destination district to analyze the impact of COVID-19 across different spatial areas. As per Eq. (6.2), the ratio is calculated as the number of origin-destination station pairs that have a certain OD flow change rate over the total number of station pairs, where $P_C(a,b)$ is the change of OD flows between an origin community district a and a destination district b, $Z(a,b)$ is the total number of station pairs in districts a and b, and $Z_C(a,b)$ is the number of station pairs that have a rate of OD flow change satisfying a certain threshold C (e.g., the rate of flow change is over 90 %). In the experiments of this study, each station in the OD flow matrices was mapped to a community district based on the geospatial location of the station and the geospatial boundaries of community districts.

$$C^t(i,j) = 1 - \frac{M_{i,j}^{t,2}}{M_{i,j}^{t,1}} \quad (6.1)$$

$$P_C(a,b) = \frac{Z_C(a,b)}{Z(a,b)} \quad (6.2)$$

4. Experiments, experimental results, and discussion

4.1. Validation of proposed origin-destination flow estimation method

4.1.1. Validation experiments

Although the proposed methodology includes both OD flow estimation and temporal-spatial analysis, its performance is only affected by the performance of the OD flow estimation. As such, a set of validation experiments were conducted to evaluate the performance of the proposed UML method in estimating public transport system OD flows using turnstile data. The validation included three steps: (1) dataset preparation, (2) OD flow estimation model training, and (3) performance evaluation.

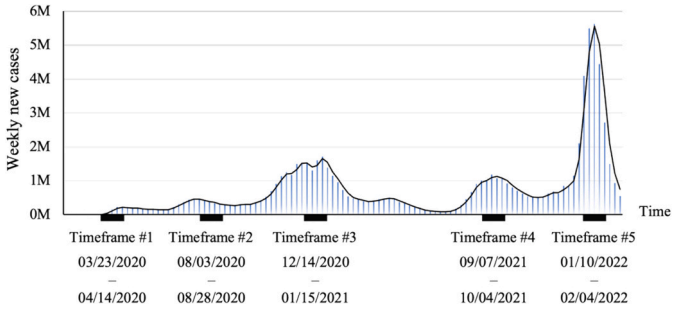


Fig. 3. Weekly trends of COVID-19 new cases in the United States. (Data source: CDC (2022b).)

A dataset, which contains daily OD flow matrices for the Bay Area Rapid Transit (BART) subway system of San Francisco (BART, 2022), was created. The BART subway system was used in the validation because it is one of a few systems in the U.S. that track both origins and destinations of subway trips to collect OD flow information, offering gold-standard OD flow matrices against which the OD flow matrices estimated by the proposed method can be compared for performance evaluation. In creating the dataset, five critical timeframes during the COVID-19 period were identified, as per Fig. 3: (1) the first timeframe is from March 23rd, 2020, to April 14th, 2020, which is around the first peak of weekly new COVID-19 cases and the declarations of "stay-at-home" order in most U.S. cities (CDC, 2022c); (2) the second timeframe is from August 3rd, 2020 to August 28th, 2020, which is around the first peak of weekly new cases after most U.S. cities have rescinded the "stay-at-home" orders, lifted the state of emergencies, and reopened to normal lives; and (3) the other three timeframes are from December 14th, 2020 to January 15th, 2021, September 7th, 2021 to October 4th, 2021, and January 10th, 2022 to February 4th, 2022. These timeframes correspond to three peaks of weekly new cases after the reopening. The daily OD matrices of the corresponding weeks in the year of 2019 (before the COVID-19 pandemic) were also included in the dataset to allow for analyzing the temporal and spatial changes in OD flow patterns before and during the COVID-19 pandemic.

The proposed estimation method was coded in Python by the authors to train OD flow estimation models. Before the model training, each training OD flow matrix $\overline{\mathbf{M}}$ in the created dataset was aggregated into a daily entrance-exit turnstile matrix $\mathbf{T}$ such that $t(i,1) = \sum_{j=1}^{N}\overline{m(i,j)}$ and $t(j,2) = \sum_{i=1}^{N}\overline{m(i,j)}$, where $\overline{m(i,j)}$ is a matrix element from $\overline{\mathbf{M}}$ that represents the gold-standard number of riders traveling from stations i to j. The entrance-exit turnstile matrices were used as the inputs for the algorithm training. Because the proposed algorithm is unsupervised, only entrance-exit turnstile matrices are needed as input data, and no previous OD flow matrices are needed as labels. During the training, model selection was conducted to select the optimal learning architecture and the optimal weight in the learning objective function [i.e., λ in Eq. (3)]. Learning architectures with different numbers of MLP mixer blocks, ranging from 1 to 4, were used to train estimation models and were compared to identify the optimal number. Learning objective functions with different weights, λs ranging from 0.1 to 1 with a step size of 0.1, were also compared to select the optimal weight.

Coefficient of determination (R^2) was selected as the metric to evaluate the performance of the trained estimation models. R^2, as per Eq. (7) (Ozer, 1985), is within the range between 0 and 1, with R^2 closer to 1 indicating a better estimation performance. In Eq. (7), $\bar{y}$ is the arithmetic mean of OD flows of all station pairs in the gold-standard matrix $\overline{\mathbf{M}}$, $\overline{m(i,j)}$ is a matrix element in $\overline{\mathbf{M}}$ that represents the gold-standard number of riders traveling from stations i to j, and other notations follow those defined in Eqs. (1)–(5).

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\overline{m(ij)} - m(ij)\right)^2}{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\overline{m(ij)} - \bar{y}\right)^2}, \text{where } \bar{y} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\overline{m(ij)} \quad (7)$$

4.1.2. Validation results and discussion

4.1.2.1. Model selection results and discussion. In terms of selecting the optimal learning architecture, as per Fig. 4(a), the architecture with a

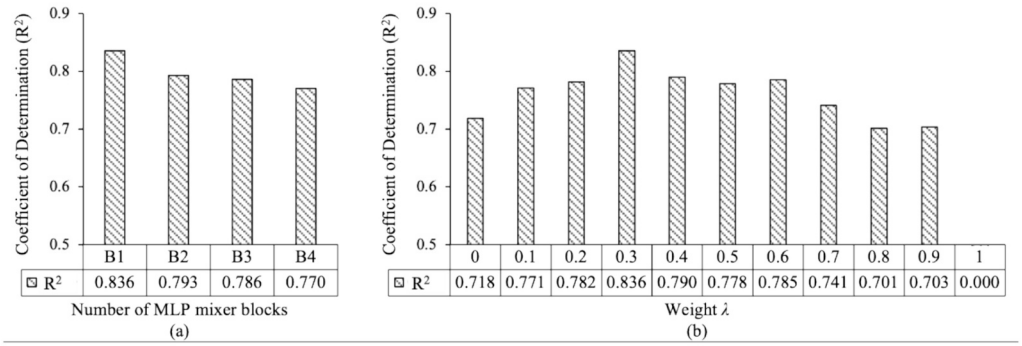


Fig. 4. Performance results for origin-destination (OD) flow estimation model selection: (a) selecting optimal number of multilayer perceptron mixer blocks; (b) selecting optimal weight.

single MLP mixer block outperformed those with more blocks. The single-block architecture achieved the R^2 of 0.836, which is 0.043, 0.050, and 0.066 higher than those achieved using the architectures with two, three, and four blocks, respectively. This might be because of two main reasons. First, the joint use of column mixing and row mixing in a single mixer block already allowed the estimation model to sufficiently capture the interaction patterns between each station pair. Even with a single mixer block, each initially estimated OD matrix was processed and estimated two times, each by a mixing layer, which provided sufficient learning for the estimation model to learn how to effectively estimate OD matrices. Second, the increase in the number of mixer blocks did not necessarily improve the estimation performance because additional mixer blocks also increased the complexity of the models, requiring more parameters to be learned and increasing the probability of model overfitting during the learning process.

In terms of selecting the optimal weight for the learning objective function, $\lambda = 0.3$ achieved the highest R^2 of 0.836, Fig. 4(b). When $\lambda = 0$, the training of the model only aimed to minimize the difference between input matrices **T**s and the corresponding reconstructed **T**′s. But there could be multiple OD matrices, both accurate and inaccurate ones, that can lead to the same **T**′ (as discussed in Section 3.1.2). As λ increased from 0.1 to 0.3, the symmetry and smoothness properties played a more important role in constraining the OD flow estimation to eliminate OD flow matrices that satisfy the difference terms in Eq. (3) but are not consistent with gold-standard OD matrices, leading to a better estimation performance. However, as λ kept increasing from 0.3 to 0.9, the flow-based properties became too stringent to the extent of requiring the estimated OD matrices to strictly follow these properties, even at the expense of increasing the difference. Hence, the large values of λ resulted in the estimated matrices inconsistent with the gold standard. When $\lambda = 1$, the training of the estimation model only meant to satisfy the symmetry and smoothness properties, without being constrained by the difference terms in Eq. (3). Based on the aforementioned discussion, the optimal model for public transport system OD flow estimation was identified, which uses a single MLP mixer block for the learning architecture and a weight of 0.3 for the learning objective function.

4.1.2.2. OD estimation performance results and discussion. Fig. 5 shows the performance results for the best-performing estimation model (with a single MLP mixer block and $\lambda = 0.3$). Each point in Fig. 5 represents the average of the estimated OD flows between a pair of stations across all the days in the selected analysis timeframes (Fig. 3), and the average of the gold-standard OD flows. Overall, the results show that the estimation model performed well: the majority of the points (out of a total of 2034) are distributed around the line of equality, especially for the points representing the station pairs with a smaller number of riders (< 800). In some cases, the proposed estimation model tended to inaccurately estimate the OD flows of the stations with a large number of riders (>

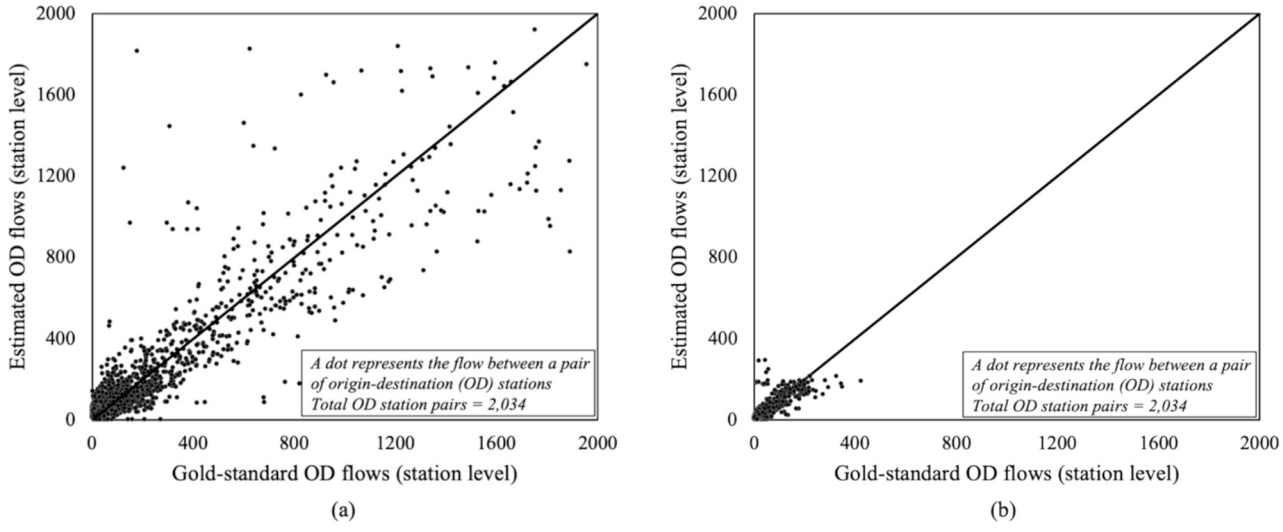


Fig. 5. Performance results for proposed OD flow estimation method: (a) comparison of estimated and gold-standard flows before COVID-19; (b) comparison of estimated and gold-standard flows during COVID-19.

800). Two important observations can be drawn from the results. First, the skewed OD flow data across station pairs had a negative impact on the accuracy of the estimation model. For example, as seen in Fig. 5(a), the numbers of riders varied significantly across station pairs, i.e., ranging from <100 to >2000 per station pair. Such flow data led to a skewed right distribution, where a large number of station pairs have a low ridership and a small number of station pairs have a high ridership. Despite the high ridership, the small number of station pairs with a high ridership made the error in estimating OD flows for these station pairs much smaller than the total error for all the stations. Consequently, the learning process focuses on reducing the estimation error for low-ridership stations in order to minimize the total error, which might have come at the expense of reducing the estimation error for high-ridership stations. Such data skewness is a common challenge to

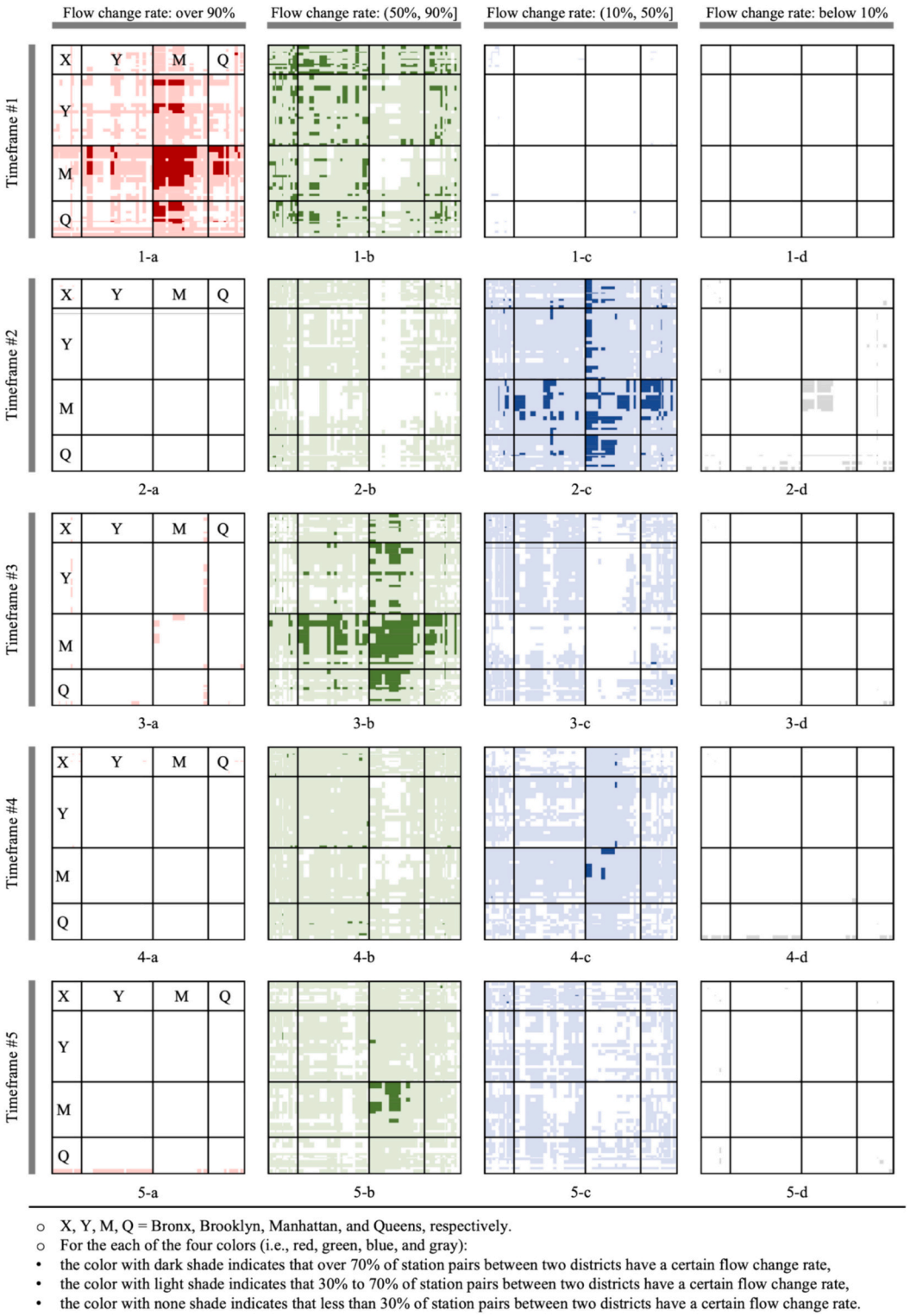


Fig. 6. Impact of COVID-19 on public transport mobility: Temporal-spatial changes in origin-destination (OD) flows of New York City subway system before and during COVID-19.

machine learning models, and has also shown negative impacts on the performance of the proposed UML-based OD flow estimation model. Second, although the proposed estimation model caused errors to both low-ridership and high-ridership stations, the errors were more significant for high-ridership stations. This might be attributed to the fact that most high-ridership stations are transfer stations that serve two or more subway lines. As a result, it was challenging for the estimation model to ensure the accuracy of the estimated OD flows due to high flow fluctuations caused by uncertain origins or destinations at transfer stations and more travel choices, i.e., riders at transfer stations can come from multiple lines and go to multiple lines (Ma et al., 2019; Shuai et al., 2022).

When comparing Fig. 5(a) and (b), it can be noted that the model achieved a better performance when estimating OD flows during the pandemic than before it (i.e., R^2 was 0.0447 higher). As per Fig. 5, the OD flows during the pandemic were more uniform than those before the pandemic, largely reducing data skewness. For example, as seen in Fig. 5 (b), the gold-standard OD flows during the pandemic were mostly centered within the range between 0 and 400. In addition to less skewed flow data, the reduction in the number of ridership and the simplification of movement trajectories of the passengers during the pandemic could also make the OD flow patterns less complex than the patterns before the pandemic (e.g., less trips requiring transfers). These observations indicate that the proposed OD flow estimation model, as expected, performed better when the OD flow patterns are more uniform and less complex.

4.2. Implementation of proposed methodology

Upon the successful validation of the estimation method, the proposed methodology was implemented in estimating the OD flows for the New York City (NYC) subway system to analyze the impact of the COVID-19 pandemic on the mobility. The NYC subway system was chosen in this study because of three main reasons. First, according to the Centers for Disease Control and Prevention, NYC is among the top cities that have the most COVID-19 cases (CDC, 2022a). Second, the NYC subway system, with a ridership of around 1.6 billion riders per year, is one of the largest subway systems in the U.S. and serves a wide range of populations (MTA, 2019). Third, the NYC subway system, like many public transport systems in the U.S., only collects turnstile data and does not capture the OD flows. All these make the NYC subway system suitable to demonstrate how the proposed methodology can be applied for estimating public transport system OD flows and supporting the subsequent temporal-spatial analysis to attain insights into how the pandemic affects the mobility.

4.2.1. Implementation setup

The implementation included two main steps. First, the proposed UML method was used to estimate OD flows of the NYC subway system by learning from turnstile data collected from the Metropolitan Transportation Authority (MTA, 2022). The NYC subway turnstile data for the five timeframes during the pandemic and their corresponding timeframes before it were collected, as per Fig. 3. The collected turnstile data were analyzed by following the OD estimation model training process discussed in Section 4.1, resulting in estimated daily OD flow matrices for both before and during the pandemic. Second, the temporal-spatial analysis method defined in Section 3.2 was followed to analyze how the pandemic affects the OD flows to understand its impact on the mobility.

To facilitate the interpretation of the analysis results, the estimated daily OD matrices for an analysis timeframe were averaged over time. The averaged matrix was marked using four colors, as per Fig. 6, to show the magnitude in the rate of OD flow change (i.e., the severity of flow change per station pair): red for rate over 90 %, green for rate between 50 % and 90 %, blue for rate between 10 % and 50 %, and gray for rate <10 %. Each color is also associated with three shade levels to show the breadth of flow changes, i.e., the number of station pairs with a specific flow change rate, which is calculated as the number of station pairs associated with the change rate out of the total number of pairs between two community districts. For example, dark red indicates that >70 % of the station pairs between two districts have a change rate over 90 %, light red indicates that 30 % to 70 % of the pairs have such a rate, and no shade indicates that <30 % of the pairs have the change rate.

4.2.2. Implementation results and discussion: COVID-19 impact on public transport mobility

Fig. 6 shows the implementation results for analyzing the COVID-19 impact on the public transport mobility in NYC. From the temporal perspective, two main conclusions were drawn from the analysis results.

- First, although the number of weekly new COVID-19 cases increased as the pandemic kept spreading across the five timeframes, the impact of the pandemic reduced its magnitude and breadth. For the first timeframe, the OD flows between most community districts were reduced by over 90 % in magnitude: more than half of the OD district pairs are marked by either dark or light red color in Fig. 6(1-a). However, after the first timeframe, only a minimum number of district pairs experienced a change rate over 90 %. On the other hand, the breadth of the impact also shrunk. As seen, the dark blue color for the second timeframe and the dark green color for the third timeframe all became light in the following timeframes, respectively, indicating that the number of the pandemic-affected OD station pairs (i.e., the impact breadth) was decreasing. The discussion above further indicates that: (1) although the normal OD flows have not yet been fully recovered, the magnitude and breadth of the pandemic impact were becoming contained as time went by; and (2) public health interventions such as “stay-at-home” orders during the first timeframe appeared to have more impact than the rise in the number of weekly new cases. Although the weekly new case number got to the climax in the fifth timeframe, the impact on OD flows in this timeframe was milder than that in the first period when the “stay-at-home” order was in place, in terms of both the magnitude and breadth.
- Second, starting at the third timeframe, as the number of new cases increased, the impact of the pandemic in the current timeframe was relatively more intensified than the previous timeframe. For example, the number of weekly new cases substantially increased when moving from the second timeframe to the third. Although the overall trend of the impact reduced across the five timeframes, the third timeframe experienced an impact that was severer in both magnitude and breadth compared to the second: as per Fig. 6(2-c) and (3-b), the dark colors shifted from blue to green when entering the third timeframe, and the number of district pairs with a dark color in the third timeframe was more than the number of district pairs with a dark color in the second. The analysis thus further indicates that after the lift of the “stay-at-home” order and with the reopening of the city, the magnitude and breadth of the pandemic impact increased as the rise of weekly new case numbers. Although the magnitude and breadth of the impact were not as strong as those from the “stay-at-home” order, the rise in new cases indeed limited the public transport mobility. This could suggest that the rise in new cases placed amplified fear of being infected by COVID-19 into the psychology and decision making of people, which affected their motives to move around the city and thus reduced the origin-destination flows of public transport.

From the spatial perspective, two main conclusions were drawn from the analysis results.

- First, the pandemic impact showed a radial pattern: the impact on the inflows to and outflows from Manhattan showed a pattern that was different from the others. For example, when moving from the

first timeframe to the second, the magnitude of the impact on the OD flows of most districts in the Manhattan neighborhood was reduced at a greater degree. As shown in Fig. 6(2-b), while most district pairs still experienced a change rate between 50 % to 90 %, the inflows to and outflows from Manhattan did not suffer from such a rate, but a rate below 50 %. When moving from the second timeframe to the third, the overall impact breadth increased, but the breadth of the impact for Manhattan was severer. As seen in Fig. 6(3-b), the increase in the coverage of the dark blue is for flows around Manhattan, including the OD flows between Manhattan and Brooklyn, Manhattan and Queens, and within Manhattan. For these neighborhoods, 9 out of the 11 districts with severe impact in Manhattan are commercial areas and have an average poverty measure of 11.8 % [i.e., 11.8 % of the residents are below the NYC poverty threshold (NYC Planning, 2022), 11 out of the 17 districts with severe impact in Brooklyn are residential and manufacturing areas and have a poverty measure of 20.6 %, and 6 out of the 10 districts with severe impact in Queens are residential and manufacturing areas and have a poverty measure of 18.6 %. Based on these community characteristics, what can be further revealed by the radial impact pattern is that: (1) when the overall impact was reduced compared to the previous timeframe, the impact on the flows from communities with relatively low income to commercial areas and the flows from manufacturing areas to commercial areas got more reduced; and (2) when the overall impact became severe, the impact on the aforementioned flows got even more severely affected.

- Second, short-range OD flows in Manhattan were more sensitive to the impact of COVID-19: the impact on the short-range flows became more reduced when the overall impact was milder, and they became more increased when the overall impact was severer. As shown in Fig. 7, OD flow distances are marked by three colors: dark purple for long-range trips (i.e., the top 10 % longest distances across all trip distances), light purple for middle-range trips (i.e., 10 % to 90 %), and white color for short-range trips (i.e., the top 10 % shortest distances across all trip distances). For example, while the overall magnitude of the impact was reduced when moving from the first timeframe to the second, the magnitude of short-range OD flows in Manhattan became more reduced. In reference to Fig. 6(2-b) and (2-c), most short-range OD flows in Manhattan had a change rate between 10 % to 50 % in the second timeframe, but most other district pairs still experienced a change rate between 50 % to 90 %. When moving from the second to the third, the overall breadth of the impact increased, and the breadth of the impact on the short-range OD flows in Manhattan was severer, as per Fig. 6(3-b). On the contrary, long-range OD flows in Manhattan were less sensitive to COVID-19 and showed a much steadier pattern: most of the long-range flows are in light green colors across all the timeframes. The spatial analysis, thus, indicates that (1) short-range flows in the Manhattan area, the center of the city, are more sensitive to the impact of COVID-19 than the long-range OD flows; and (2) public transport services are essential for people making long-range trips that have fewer alternative transport modes. Despite that the pandemic indeed reduced long-range public transport system flows, the remaining long-range flows remained relatively steady throughout the analysis timeframes, indicating the essentiality of long-range trips by public transport even during the pandemic.

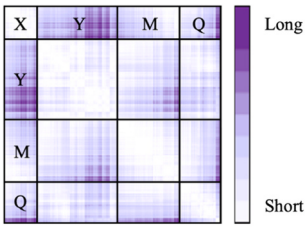


Fig. 7. Geographical distance between origin-destination (OD) station pairs of New York City subway system.

5. Limitations

Three main limitations of this study are acknowledged. First, the OD flow estimation model utilized in this paper was developed using turnstile data before and during COVID-19, and the data cover normal days and days with extreme weather and special events (e.g., snowstorms and large-scale gatherings). As a result, because the number of 'abnormal' days is much smaller than that of normal days, the UML-based estimation model may not be able to fully capture the flow patterns for 'abnormal' days (due to the imbalance in the number of data instances) and would thus result in less accurate OD flows. In the future, the authors will develop two separate estimation models, one for each type of days, to investigate if such a separation and balancing would result in improved flow estimation. Second, the proposed OD flow estimation method is limited in accounting for the impact of transfer stations (which usually have a high ridership) on estimating OD flows for public transport systems, as discussed in Section 4.1.2.2. As a result, the method achieved a relatively lower performance when estimating OD flows for such stations. To address this limitation, the authors plan to study how to account for such stations as an additional property in the proposed objective function to improve the estimation performance. Third, the proposed methodology was implemented in analyzing the pandemic impact on the public transport mobility of New York City. As a result, the analysis results presented in this study may not be directly applicable to other cities. This implementation was intended to demonstrate how the methodology can be used to support the analysis of the pandemic impact, and does not intend to imply that the method can only be used for the chosen city. Instead, the proposed methodology can be used to estimate public transport system OD flows for other cities, different temporal granularities, and/or different spatial granularities. In future work, the methodology can be performed to support various analysis needs (e.g., analyzing the impact of the pandemic on rush-hour OD flows by comparing the flows before, during, and after the pandemic).

6. Contributions to the body of knowledge

This research contributes to the body of knowledge in two primary ways. First, this research offers a new UML-based method that allows for using boarding-alighting data (e.g., turnstile data for subway systems) to estimate OD flows for public transport systems. On one hand, while other data sources that have been explored for OD flow estimation are either biased in terms of sampling coverage (e.g., travel surveys) or challenging to acquire (e.g., link data), turnstile data provide comprehensive coverage of all riders of a public transport system, in addition to being easy to collect, publicly available, and free of privacy concerns. Thus, by enabling the use of turnstile data, the proposed method opens doors to improve the ability to better estimate OD flows for public transport systems. On the other hand, although a plethora of studies have developed ML-based methods for OD flow analysis, these studies have focused on developing supervised ML approaches for OD flow prediction, which usually require learning from previous OD flows and are thus limited in estimating OD flows using turnstile data (due to the lack of historical public transport system OD flows). By using a decoder-encoder learning architecture with flow property-based unsupervised learning objective function, the proposed estimation approach addresses this limitation in the current state of the art to enable the effective use of turnstile data for OD flow estimation. Second, the application of the proposed methodology, which includes OD flow estimation and temporal-spatial analysis, allows for analyzing and understanding the impact of extreme events on public transport mobility. While this study

focused on applying the methodology to analyze the pandemic impact, the methodology itself can be generally applied to various scenarios, such as analyzing the impact of extreme events (e.g., flooding and hurricanes) on mobility in urban areas. In future applications, estimation models specific to the intended analysis scenario need to be developed by following the methodology to re-train models using turnstile data relevant to the scenario at hand. By offering a methodology to analyze public transport mobility, this research has the potential to create new knowledge about the impact of extreme events on urban mobility to support the making of effective recovery plans and long-term policies toward building resilience in our public transport systems and communities.

7. Conclusions and future work

This paper proposed a novel methodology to analyze the impact of COVID-19 on public transport mobility. The methodology includes a new unsupervised machine learning (UML) method to estimate public transport systems OD flows from boarding-alighting data, and a temporal-spatial analysis method to compare OD flows before and during COVID-19 to explore its impact on public transport mobility across time and space. The performance of the proposed UML-based flow estimation method was validated by comparing the OD flow matrices estimated by the method to the gold-standard matrices in the Bay Area Rapid Transit dataset. The validation results showed that the proposed estimation method was effective: it achieved an R^2 of 0.836. Upon the successful validation, the proposed methodology was implemented in estimating the OD flows of the NYC subway system before and during the pandemic to analyze and understand how the pandemic affected the mobility across time and space. The implementation results showed that (1) although the overall impact of the pandemic was becoming contained as time went by, the rise of weekly new COVID-19 cases in an analysis timeframe led to an impact relatively more intensified than the previous timeframe, in terms of both impact magnitude and breadth. Yet, the intensified impact was still much more contained than the impact for the initial phase of the pandemic, when public health interventions such as “stay-at-home” orders were in place; and (2) the impact showed a radial pattern: the impact on the inflows to and outflows from the Manhattan neighborhood, the center of the city, became more contained when the overall impact got milder, and the impact became more intensified when the overall impact got severer.

In the future, the authors will focus their research efforts on two main directions. First, they will further improve the capability of OD flow estimation by jointly learning from both turnstile data and OD flow context data, such as weather data, disaster data, and public transport operation data (e.g., closure and maintenance data). Such joint learning would allow for developing flow estimation models that are capable of accurately estimating OD flows for normal days and ‘abnormal’ days with extreme events such as flooding and large-scale gathering. Second, the authors will develop a computational framework that integrates OD flow estimation and data-driven socioeconomic analysis to enable deeper flow-based urban mobility analytics, in terms of the demographic and socioeconomic characteristics of people traveling between a certain pair of origin and destination locations, the purposes of the trips, and how the trips affect the well-being of the travels. These efforts could further advance our knowledge of urban mobility patterns – i.e., how mobility patterns change and evolve across time, space, ridership groups, and travel purposes – to enhance urban mobility analysis and understanding, thereby offering more opportunities for decision makers to stimulate effective and long-term urban development strategies toward enhanced urban resilience and sustainability.

CRediT authorship contribution statement

Lan Zhang: Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Kaijian Liu:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

None.

Data availability

Data will be made available on request.

Acknowledgements

None.

References

Abbaas, O., & Ventura, J. A. (2021). An edge scanning method for the continuous deviation-flow refueling station location problem on a general network. *Networks, 79*(3), 264–291.

Ali, A. (2022). *AI-based mode of transportation and destination classification and prediction in origin-destination surveys*. M.S. thesis. Concordia Univ.

Anstreicher, K. M., Fampa, M., Lee, J., & Williams, J. (1999). Using continuous nonlinear relaxations to solve constrained maximum-entropy sampling problems. *Mathematical Programming, 85*(2), 221–240.

Barbosa-Filho, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., … Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports, 734*, 1–74.

BART (Bay Area Rapid Transit). (2022). Ridership reports. https://www.bart.gov/about/reports/ridership. (Accessed 28 November 2022).

Bell, M. G. (1983). The estimation of an origin-destination matrix from traffic counts. *Transportation Science, 17*(2), 198–217.

Bera, S., & Rao, K. V. K. (2011). Estimation of origin-destination matrix from traffic counts: The state of the art. *European Transport, Institute for the Study of Transport within the European Economic Integration., 49*, 2–23.

Bertsimas, D., & Yan, J. (2018). From physical properties of transportation flows to demand estimation: An optimization approach. *Transportation Science, 52*(4), 1002–1011.

Blum, J. J., Sridhar, A., & Mathew, T. V. (2010). Origin-destination matrix generation from boarding-alighting and household survey data. *Journal of the Transportation Research Board, 2183*(1), 1–8.

Cascetta, E. (1984). Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator. *Transportation Research Part B: Methodological., 18*(4–5), 289–299.

CDC (Centers for Disease Control and Prevention). (2022a). COVID data tracker. https://covid.cdc.gov/covid-data-tracker/#county-view?list_select_state=New+York&data-type=CommunityLevels&list_select_county=36061. (Accessed 28 November 2022).

CDC (Centers for Disease Control and Prevention). (2022b). Weekly trends in number of COVID-19 cases in the United States reported to CDC. https://covid.cdc.gov/covid-data-tracker/#trends_weeklycases_select_00.

CDC (Centers for Disease Control and Prevention). (2022c). *COVID-19*. https://www.cdc.gov/coronavirus/2019-ncov/. (Accessed 20 November 2022).

Chakraborty, I., & Maity, P. (2020). Covid-19 outbreak: Migration, effects on society, global environment and prevention. *Science of The Total Environment., 728*, Article 138882.

Cuomo, M. A. (2020, March 20). Governor Cuomo signs the ‘New York State on PAUSE’ Executive Order. https://www.governor.ny.gov/news/governor-cuomo-signs-new-york-state-pause-executive-order. (Accessed 28 November 2022).

Dey, S., Winter, S., & Tomko, M. (2020). Origin-destination flow estimation from link count data only. *Sensors, 20*(18), 5226.

Dubiner, M., and Y. Singer. 2011. “Entire relaxation path for maximum entropy problems.” In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 941–948. Edinburgh, Scotland, UK. Association for Computational Linguistics.

Ebrahimpour, Z., Wan, W., Velázquez García, J. L., Cervantes, O., & Hou, L. (2020). Analyzing social-geographic human mobility patterns using large-scale social media data. *ISPRS International Journal of Geo-Information, 9*(2), 125.

Egu, O., & Bonnel, P. (2020). How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination surveys and household travel survey? An empirical investigation in Lyon. *Transportation Research Part A: Policy and Practice., 138*, 267–282.

Essien, A., Petrounias, I., Sampaio, P., & Sampaio, S. (2020). A deep-learning model for urban traffic flow prediction with traffic events mined from Twitter. *World Wide Web, 24*(4), 1345–1368.

Fekih, M., Bellemans, T., Smoreda, Z., Bonnel, P., Furno, A., & Galland, S. (2020). A data-driven approach for origin-destination matrix construction from cellular network signaling data: A case study of Lyon region (France). *Transportation, 48*(4), 1671–1702.

FHWA (Federal Highway Administration). (2017). Understanding traditional origin-destination data: A survey. https://rosap.ntl.bts.gov/view/dot/55804.
Ge, L., Sarhani, M., Voß, S., & Xie, L. (2021). Review of transit data sources: Potentials, challenges and complementarity. *Sustainability, 13*(20), Article 11450.
Gu, M., and Z. Duan. 2022. "Daily OD demand prediction in urban metro transit system: A convolutional LSTM neural network with multi-factor fusion channel-wise attention." 2022 IEEE 25th International Conference on Intelligent Transportation Systems.
Guo, J., Liu, Y., Li, X., Huang, W., Cao, J., & Wei, Y. (2019). Enhanced least square based dynamic OD matrix estimation using radio frequency identification data. *Mathematics and Computers in Simulation, 155*, 27–40.
He, K., X. Zhang., S. Ren., and J. Sun. 2016. "Deep residual learning for image recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition.
Hou, Y., Deng, Z., & Cui, H. (2021). Short-term traffic flow prediction with weather conditions: Based on deep learning algorithms and data fusion. *Complexity, 2021*, 1–14.
Hu, J., B. Yang., C. Guo., C. S. Jensen., and H. Xiong. 2020. "Stochastic origin-destination matrix forecasting using dual-stage graph convolutional, recurrent neural networks." 2020 IEEE 36th International Conference on Data Engineering.
Humagain, P., & Singleton, P. A. (2021). Exploring tourists' motivations, constraints, and negotiations regarding outdoor recreation trips during COVID-19 through a focus group study. *Journal of Outdoor Recreation and Tourism, 36*, Article 100447.
Hussain, E., Bhaskar, A., & Chung, E. (2021). Transit OD matrix estimation using smartcard data: Recent developments and future research challenges. *Transportation Research Part C: Emerging Technologies., 125*, Article 103044.
Hwang, S. W., Kweon, S. J., & Ventura, J. A. (2020). An alternative fuel refueling station location model considering detour traffic flows on a highway road system. *Journal of Advanced Transportation*, 1–27.
Jeong, I.-J., & Park, D. (2021). Stochastic programming approach for static origin-destination matrix reconstruction problem. *Computers and Industrial Engineering, 157*, Article 107373.
Jiang, W., & Luo, J. (2022). Big data for traffic estimation and prediction: A survey of data and tools. *Applied System Innovation., 5*(1), 23.
Jin, Z., Chen, Y., Li, C., & Jin, Z. (2022). Trip destination prediction based on hidden Markov model for multi-day global positioning system travel surveys. *Transportation Research Record: Journal of the Transportation Research Board., 036119812211079*.
Katranji, M., Kraiem, S., Moalic, L., Sanmarty, G., Khodabandelou, G., Caminada, A., & Selem, F. H. (2019). Deep multi-task learning for individuals origin-destination matrices estimation from census data. *Data Mining and Knowledge Discovery, 34*(1), 201–230.
Ke, J., Qin, X., Yang, H., Zheng, Z., Zhu, Z., & Ye, J. (2021). Predicting origin-destination ride-sourcing demand with a spatio-temporal encoder-decoder residual multi-graph convolutional network. *Transportation Research Part C: Emerging Technologies., 122*, Article 102858.
Khalil, S., Amrit, C., Koch, T., & Dugundji, E. (2021). Forecasting public transport ridership: Management of information systems using CNN and LSTM architectures. *Procedia Computer Science, 184*, 283–290.
Kumar, P., Khani, A., & Davis, G. A. (2019). Transit route origin-destination matrix estimation using compressed sensing. *Journal of the Transportation Research Board, 2673*(10), 164–174.
Li, T., Sun, D., Jing, P., & Yang, K. (2018). Smart card data mining of public transport destination: A literature review. *Information, 9*(1), 18.
Lin, Q., Gong, Z., Wang, Q., & Li, J. (2019). RILNET: A reinforcement learning based load balancing approach for datacenter networks. *Machine Learning for Networking.*, 44–55.
Liu, F., X. Ren., Z. Zhang., X. Sun., and Y. Zou. 2020. "Rethinking skip connection with layer normalization." Proceedings of the 28th International Conference on Computational Linguistics.
Liu, Z., Z. Wang., X. Yin., X. Shi., Y. Guo., and Y. Tian. 2019. "Traffic matrix prediction based on deep learning for dynamic traffic engineering." 2019 IEEE Symposium on Computers and Communications.
López-Ospina, H., Cortés, C. E., Pérez, J., Peña, R., Figueroa-García, J. C., & Urrutia-Mosquera, J. (2021). A maximum entropy optimization model for origin-destination trip matrix estimation with fuzzy entropic parameters. *Transportmetrica A: Transport Science, 18*(3), 963–1000.
Lu, B., X. Gan., H. Jin., L. Fu., and H. Zhang. 2020. "Spatiotemporal adaptive gated graph convolution network for urban traffic flow forecasting." Proceedings of the 29th ACM International Conference on Information and Knowledge Management.
Luo, C., Huang, C., Cao, J., Lu, J., Huang, W., Guo, J., & Wei, Y. (2019). Short-term traffic flow prediction based on least square support vector machine with hybrid optimization algorithm. *Neural Processing Letters, 50*(3), 2305–2322.
Ma, X., Zhang, J., Du, B., Ding, C., & Sun, I. (2019). Parallel architecture of convolutional bi-directional LSTM neural networks for network-wide metro ridership prediction. *IEEE Transactions on Intelligent Transportation Systems, 20*(6), 2278–2288.
Maher, M. J. (1983). Inferences on trip matrices from observations on link volumes: A Bayesian statistical approach. *Transportation Research Part B: Methodological., 17*(6), 435–447.
Mamei, M., Bicocchi, N., Lippi, M., Mariani, S., & Zambonelli, F. (2019). Evaluating origin-destination matrices obtained from CDR data. *Sensors, 19*(20), 4470.
Miao, Y., Xu, Y., & Mandic, D. (2022). Hyper-GST: Predict metro passenger flow incorporating GraphSAGE, hypergraph, social-meaningful edge weights and temporal exploitation. *Machine Learning*. arXiv:2211.04988.
Mo, B., Li, R., & Dai, J. (2020). Estimating dynamic origin-destination demand: A hybrid framework using license plate recognition data. *Computer-Aided Civil and Infrastructure Engineering, 35*(7), 734–752.
Moussavi-Khalkhali, A., & Jamshidi, M. (2019). Feature fusion models for deep autoencoders: Application to traffic flow prediction. *Applied Artificial Intelligence, 33*(13), 1179–1198.
MTA (Metropolitan Transit Authority). (2019). Subway and bus ridership for 2019. https://new.mta.info/agency/new-york-city-transit/subway-bus-ridership-2019. (Accessed 28 November 2022).
MTA (Metropolitan Transit Authority). (2022). Turnstile data. http://web.mta.info/developers/turnstile.html. (Accessed 28 November 2022).
MTA (Metropolitan Transportation Authority). (2022). The MTA network. https://new.mta.info/about-us/the-mta-network. (Accessed 28 November 2022).
NYC (The official website of the City of New York). (2020). Mayor de Blasio issues state of emergency. https://www.nyc.gov/office-of-the-mayor/news/138-20/mayor-de-blasio-issues-state-emergency. (Accessed 28 November 2022).
NYC Planning (City of New York - Department of City Planning). (2022). *Community district profiles*. https://communityprofiles.planning.nyc.gov/. (Accessed 28 November 2022).
Osorio-Arjona, J., & García-Palomares, J. C. (2019). Social media and urban mobility: Using Twitter to calculate home-work travel matrices. *Cities, 89*, 268–280.
Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97*(2), 307–315.
Papageorgiou, M., & Varaiya, P. (2009). Link vehicle-count – The missing measurement for traffic control. *IFAC Proceedings Volumes., 42*(15), 224–229.
Pelletier, M.-P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies., 19*(4), 557–568.
Pitombeira-Neto, A., Loureiro, C., & Carvalho, L. (2018). Bayesian inference on dynamic linear models of day-to-day origin-destination flows in transportation networks. *Urban Science, 2*(4), 117.
Pourebrahim, N., S. Sultana., J. C. Thill., and S. Mohanty. 2018. "Enhancing trip distribution prediction with Twitter data." Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery.
Ros-Roca, X., Montero, L., Barceló, J., Nökel, K., & Gentile, G. (2022). A practical approach to assignment-free dynamic origin-destination matrix estimation problem. *Transportation Research Part C: Emerging Technologies., 134*, Article 103477.
Selvarajah, K., K. Ragunathan., Z. Kobti., and M. Kargar. 2020. "Dynamic network link prediction by learning effective subgraphs using CNN-LSTM." 2020 International Joint Conference on Neural Networks.
Sharic, S., S. Bandara., and S. Fernando. 2021. "Methods to estimate bus revenue from passenger boarding and alighting data: case study for Sri Lanka." 2021 Moratuwa Engineering Research Conference.
Shi, H., Q. Yao., Q. Guo., Y. Li., L. Zhang., J. Ye., Y. Li., and Y. Liu. 2020. "Predicting origin-destination flow via multi-perspective graph Convolutional Network." 2020 IEEE 36th International Conference on Data Engineering.
Shuai, C., Shan, J., Bai, J., Lee, J., He, M., & Ouyang, X. (2022). Relationship analysis of short-term origin-destination prediction performance and spatiotemporal characteristics in urban rail transit. *Transportation Research Part A: Policy and Practice., 164*, 206–223.
Spiess, H. (1987). A maximum likelihood model for estimating origin-destination matrices. *Transportation Research Part B: Methodological., 21*(5), 395–412.
Sun, W., Schmöcker, J.-D., & Fukuda, K. (2021). Estimating the route-level passenger demand profile from bus dwell times. *Transportation Research Part C: Emerging Technologies., 130*, Article 103273.
Tang, J., Chen, X., Hu, Z., Zong, F., Han, C., & Li, L. (2019). Traffic flow prediction based on combination of support vector machine and data denoising schemes. *Physica A: Statistical Mechanics and its Applications, 534*, Article 120642.
Teixeira, J. F., Silva, C., & Moura e Sá, F. (2021). The motivations for using bike sharing during the COVID-19 pandemic: Insights from Lisbon. *Transportation Research Part F: Traffic Psychology and Behaviour., 82*, 378–399.
Tian, Y., Zhang, K., Li, J., Lin, X., & Yang, B. (2018). LSTM-based traffic flow prediction with missing data. *Neurocomputing, 318*, 297–305.
Toan, T. D., & Truong, V.-H. (2020). Support vector machine for short-term traffic flow prediction and improvement of its model training using nearest neighbor approach. *Transportation Research Record: Journal of the Transportation Research Board., 2675*(4), 362–373.
Tolstikhin, L., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., … Dosovitskiy, A. (2021). MLP-Mixer: An all-MLP architecture for vision. *Computer Vision and Pattern Recognition.*. https://doi.org/10.48550/arXiv.2105.01601
Twitter. (2022). Advanced filtering with geo data. https://developer.twitter.com/en/docs/tutorials/advanced-filtering-for-geo-data. (Accessed 28 November 2022).
Wang, X., Zhang, N., Zhang, Y., & Shi, Z. (2018). Forecasting of short-term metro ridership with support vector machine online model. *Journal of Advanced Transportation, 2018*, 1–13.
Wang, Y., Ma, X., Liu, Y., Gong, K., Henrickson, K. C., Xu, M., & Wang, Y. (2016). Correction: A two-stage algorithm for origin-destination matrices estimation considering dynamic dispersion parameter for route choice. *PLoS One, 11*(2).
Wang, Y., H. Yin., H. Chen., T. Wo., J. Xu., and K. Zheng. 2019. "Origin-destination matrix prediction via graph convolution." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
Yang, H., & Rakha, H. (2017). A novel approach for estimation of dynamic from static origin-destination matrices. *Transportation Letters, 11*(4), 219–228.
Yang, H., Y. Wang., and D. Wang. 2018. "Dynamic origin-destination estimation without historical origin-destination matrices for microscopic simulation platform in urban network." 2018 21st International Conference on Intelligent Transportation Systems.
Zagatti, G. A., Gonzalez, M., Avner, P., Lozano-Gracia, N., Brooks, C. J., Albert, M., … Bengtsson, L. (2018). A trip to work: Estimation of origin and destination of

commuting patterns in the main metropolitan regions of Haiti using CDR. *Development Engineering, 3*, 133–165.
Zaragozí, B., Trilles, S., Gutiérrez, A., & Miravet, D. (2021). Development of a common framework for analysing public transport smart card data. *Energies, 14*(19), 6083.
Zargari, S. A., Memarnejad, A., & Mirzahossein, H. (2021). Hourly origin-destination matrix estimation using intelligent transportation systems data and deep learning. *Sensors, 21*(21), 7080.
Zhang, Q., Li, C., Yin, C., Zhang, H., & Su, F. (2022). A hybrid framework model based on wavelet neural network with improved fruit fly optimization algorithm for traffic flow prediction. *Symmetry, 14*(7), 1333.
Zhang, Z., Li, M., Lin, X., & Wang, Y. (2020). Network-wide traffic flow estimation with insufficient volume detection and crowdsourcing data. *Transportation Research Part C: Emerging Technologies., 121*, Article 102870.
Zhao, J., Qu, H., Zhao, J., Dai, H., & Jiang, D. (2020). Spatiotemporal graph convolutional recurrent networks for traffic matrix prediction. *Transactions on Emerging Telecommunications Technologies, 31*(11).
Zhao, X., Y. Gu., L. Chen., and Z. Shao. 2019. "Urban short-term traffic flow prediction based on stacked autoencoder." CICTP 2019.