TransitCrowd: Estimating Subway Stations Demand with Mobile Crowdsensing Data

Piergiorgio Vitello

piergiorgio.vitello@uni.lu

University of Luxembourg

Claudio Fiandrino

IMDEA Networks

Richard D. Connors

University of Luxembourg

Francesco Viti

University of Luxembourg

Research Article

Keywords: Public transport, Crowdsensing, Machine learning, Google Popular Times

Posted Date: December 5th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2555834/v2

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Data Science for Transportation on April 13th, 2024. See the published version at https://doi.org/10.1007/s42421-024-00091-4.

TransitCrowd: Estimating Subway Stations Demand with Mobile Crowdsensing Data

Piergiorgio Vitello[1*], Claudio Fiandrino[2], Richard D. Connors[1] and Francesco Viti[1]

[1*]Faculty of Sciences Technology and Communication, University of Luxembourg, Esch-sur-Alzette, 4365, Luxembourg.
[2]IMDEA Networks Institute, Leganés, Madrid, 28918, Spain.

*Corresponding author(s). E-mail(s): piergiorgio.vitello@uni.lu;
Contributing authors: claudio.fiandrino@imdea.org;
richard.connors@uni.lu; francesco.viti@uni.lu;

Abstract

Traditionally Public Transport (PT) demand estimation relies on manual survey-based or, where available, smartcard passenger data. However, transport service providers and authorities make it rarely available to researchers. An additional challenge is the variety of formats and the low granularity in which such data is available. Recently, first steps towards the use of advanced ICT-based data-driven approaches have started to emerge. These new data sources can provide new opportunities for generating more data and insights into transit demand patterns and behaviour. In this paper, we propose a novel data-driven transit demand estimation process, TransitCrowd, and apply it to subway stations. TransitCrowd estimates the passengers entering and exiting each station using as proxy the subway crowdness provided by Google Popular Times (GPT) crowdsensed information often available at sheer scale in any city. TransitCrowd's key component is its one-time calibration process, which creates temporal signatures of the stations based on historical GPT information, and regression-based machine learning and live GPT to predict passenger flows. We assess TransitCrowd's estimation accuracy for two cities across a two-months period, i.e., New York and Washington., showing very promising results for both estimation and real-time prediction of transit flows at subway stations.

Keywords: Public transport, Crowdsensing, Machine learning, Google Popular Times

1 Introduction

Transportation planners and researchers need accurate and precise data to monitor and manage Public Transport PT systems. Today, Public Transport Authorities (PTAs) rely on smartcard data and automated passenger counts that can be used to estimate transit demand and its variation under different operational conditions [1]. However, most PTAs are unwilling to share this data easily, and if they do, it is usually restricted to a short time period. Data are crucial to truly understand the complex dynamics of transit demand, especially in the last decades, where public transit demand has faced many challenges, such as the introduction of new on-demand services, the evolution of new modes of transport, and disruptive events like the Covid19 pandemic. According to the European Parliament's Committee on Transport and Tourism (TRAN), the usage of public transport will settle on a decrease of 10-15% compared to the pre-pandemic levels in the next 3 years [2]. Despite this variation, PT is widely used in metropolitan areas and medium sized cities today. Some examples are London (population of 9 million) where PT accounts for 5 million trips a day [3], and Amsterdam (population 900,000) where 300,000 passengers travel by PT on a normal weekday [4].

Recently, new technologies have been introduced and deployed, which provide multiple sources of data and information that can be utilized for demand estimation and analysis [5]. The widespread of mobile devices enables the use of Information and Communications Technology (ICT) by unleashing interesting opportunities to improve the quality of PT. This large amount of smart devices is a potential source of data according to the mobile crowdsensing (MCS) paradigm [6].

In the literature, many types of crowdsensed data have been exploited for PT analysis. An example is given by [7], which proposes a framework that provides real time public transport data using crowdsensed information provided by passengers' smartphones. The shortcomings of this approach are that passengers have to actively contribute to the framework and that developing a crowdsensed campaign from scratch requires a big effort. Most of the studies exploit WiFi and Bluetooth technologies as a measure of passengers [8][9]. Although these works can achieve an accurate and precise estimation of transit demand, they suffer severe drawbacks. First, they require the creation from scratch of a new crowdsensed campaign which usually requires a big effort. Second, they require user consent and cooperation.

Since 2015, Google has made available a new service called Google Popular Times (GPT) that is based on anonymized crowdsensed data passively collected from Google users. GPT is a feature of Google Maps and Search that visualises the temporal profile of the level of crowding of a certain location or point of interest (retail shops, restaurants, public places). Hence, this information can be potentially useful for assessing how busy a certain activity is. In [10] the authors used GPT to model the demand at local businesses level.

In this paper we investigate the potential of GPT information to for public transport demand estimation and prediction applications. This dataset has the

advantage of being already provided by Google and not requiring new data collection. Moreover, widespread availability of GPT opens up possibilities for estimating public transport demand in areas where such information is not typically collected. However, the main limitation of GPT is the lack of transparency and details in the data acquisition and processing, as Google only provides the information in an aggregated and normalised way, and at 1-hour intervals. This study aims at overcoming this shortcoming by combining GPT with real public transport data in order to leverage the GPT data for transit demand estimation and prediction at subway stations.

Hence, in this paper, we explore the potential of exploiting GPT data as substitute of transit data. Moreover, since GPT is worldwide available, and provides live and historical information each hour, it can enrich transit data where the transit data granularity is low. Finally, we show how GPT can be used to estimate the in- and outflow of transit users at subway stations.

To pursue these goals, we design and test TransitCrowd, a framework that is able to make live estimations of transit data exploiting only the GPT of stations.

In summary, the contributions we make with this paper are as follows:

Contribution 1. We analyse and quantify the correlation between transit entry and exit flow data in order to gain insight into the relationship between passenger flows at subway stations and GPT information.

Then, we introduce the TransitCrowd tool, which is able to estimate and predict transit demand data regardless of the granularity of the input transit data. Our tool comprises two different estimators, a signature-based estimation and a regression-based machine learning model.

Contribution 2. The first estimator (Reg estimator) is trained separately in each city, it requires an initial transit dataset and focuses on obtaining the maximum accuracy in the trained area. This tool is suited for areas where a transit dataset is available with low granularity or that is limited on time.

Contribution 3. The second estimator (Sig estimator) is more flexible. It gives a potentially transferable methodology without requiring starting transit data, but at the cost of lower accuracy. This estimator can be leveraged in case no transit data is available for the area under analysis.

The paper proceeds with an overview of the related works presented in the next section, followed by a description of the data exploited in the study, the methodology behind the framework, and evaluation of the results. Finally, the last section gives conclusions of the work, and some final remarks.

2 Related works

Understanding how to leverage large-scale datasets is fundamental to investigate the potential of big data for public transport analysis [11]. Typical approaches infer mobility of PT users' from smartcard data and automated data collection systems. These approaches have been used for instance to infer

bus passengers origin-destination [12], to extract information about passengers routines to predict transportation usage [13], or to measure the impact of individual characteristics on PT accessibility [14]. The issue with this type of data is that it is controlled by PTAs, and only a few share their datasets. Consequently, the studies based on smartcard data are usually characterized by a short data collection period or focus on a specific city.

During the past few years, several researches have exploited cellular network usage (i.e., LTE) as alternative to automatic data collection methods. In [15], the authors created a new framework able to exploit cellular data to measure passenger flows in subway stations in Paris, France. Mobile and wireless network data analysis can also be applied to classify subway users, distinguishing subway residents from commuters [16]. In [17] the authors created a methodology that leverages cell phone usage as a proxy to extract passengers' travel demand. Their findings help PTAs examine their public transportation options and effectively develop new transit routes or expand current routes to meet users' requirements.

Unfortunately, these approaches carry significant drawbacks due to technical constraints, such as lack of location accuracy, poor network coverage, and the unwillingness of network operators to share their datasets [18]. Similar to mobile phone data, WiFi and Bluetooth technology have emerged in the literature to capture mobility of PT users. In particular, WiFi sensors have been exploited to identify trajectories of metro passengers [19], to estimate real-time passengers' peak flow in order to avoid accidents [20], and to measure bus passengers' loads [21]. Although Bluetooth connections are explored more for proximity-based studies, in [22] the authors leverage this technology to detect bus passengers' origin and destination, while in [23] the authors analyze passenger dynamics and connectivity in Beijing subway. Another essential source of data for counting crowds is cameras in combination with machine vision approaches. Several works obtained interesting results studying the integration of video information with WiFi connection [24][25][26].

Although these approaches are shown to obtain promising results, they require every time new data collection campaigns for each specific city. This problem raises the issue of comparing a developed methodology in different cities since it would be challenging and expensive to carry multiple data collections.

As an alternative, mobile crowdsensing (MCS) allows to collect mobility data from users, e.g. identifying their usual habits and inferring special events [27]. The use of crowdsensed data has become a win-win solution in different domains of transportation, such as monitoring traffic dynamics and demand analysis on special events [28]. Crowdsensed-based approaches can be applied to better tackle transit demand and understand citizens' mobility. For example, crowdsensed data from the web can help to detect origin and destination of passengers in public transport [29].

In the context of MCS, this study focuses on a specific crowdsensed dataset, the Google Popular Times (GPT), due to its characteristics of wide availability

and ease to collect it. Despite this potential, few studies analyzed the possible importance of GPT for the transportation domain. In [30] the authors focused on venue popularity, they developed a WiFi microcontroller to measure the real number of people in a place, the comparison of their data with the corresponding GPT revealed promising results.

Another interesting analysis of GPT is presented in [31], where the temporal variation of electric vehicle charging demand using the GPT of the activities around the charging stations was analysed. In [32], the authors analyze the relationship between GPT, park use, and built-environment density.

GPT resulted a fundamental data source during Covid19 pandemic, many studies exploited this dataset to analyze citizens' mobility during lockdown [33] [34], since live GPT values can be an important source to make comparisons between different time periods. Unlike the above studies, this work aims to exploit GPT in order to extract public transport demand information. The closest work to our research is [35] that has the objective of investigating the possibility of using GPT to predict traffic volumes in a specific area. Their analysis indicated a clear relationship between GPT, traffic, and environmental performances. By contrast, no research analyzed the importance of GPT for the public transport domain. In this study, we investigate the possibility to provide an estimation of the passenger flows of transit stations relying only on GPT data.

3 Dataset and Preliminary Analysis

In this section, we describe the dataset we exploit in our analysis of transit stations demand.

3.1 Google Popular Times

GPT is a feature of Google maps that visualise the standard temporal profile of the number of people visiting a place (retail shops, restaurants, public places) as a vector of normalized per-hour weekly values in the range $[0 : 100]$ (0: closing hours, 1: lowest amount of visits per-hour in a week and 100: the highest).

The GPT information is generated from data sent anonymously by smartphones with the google history location enabled, the location of these devices is tracked in the background and sent to Google through WiFi or mobile networks.

Fig. 1 shows and instance of how GPT displays information for a specific place on Google Maps. We distinguish between two types of GPT information: live and standard. The blue bars show the standard GPT, which tells us how busy the place usually is at different times of the day. This standard view is made by looking at past data from many weeks and taking an average. On the other side, the red bar represents the live GPT value, reflecting the current crowd density at the location, with updates every hour. The live GPT offers

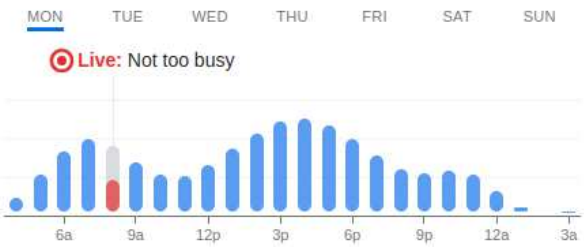
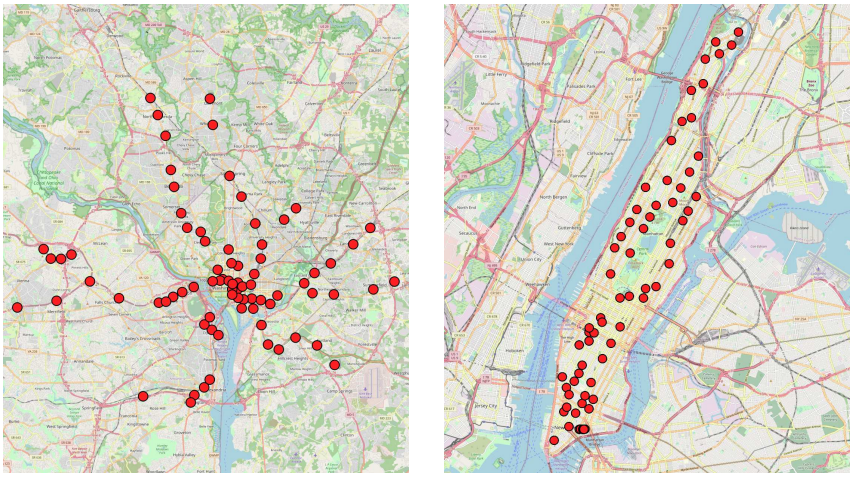


Fig. 1 An example of GPT record



(a) Washington stations (b) New York stations

Fig. 2 Maps of the cities considered in our study

real-time information, ensuring a thorough understanding of the location's current popularity dynamics.

The use of normalized values indicates the trend of an activity during a week and inherently the factors that influence such behaviour (e.g., a restaurant that has more success during weekends in touristic areas or at lunchtime in business districts). This data offers unexplored opportunities for linking human mobility with activity pattern dynamics, since the nature behind the GPT data is the individuals' traces collected by Google when entering and exiting the locations, and in between performing a the activity in the visited place. However, this information hides the absolute quantity of the demand, i.e. the real number of customers. In this work, we focus on leveraging GPT of the subway stations to investigate if such information can be exploited to determine the inflow and outflow of users at the same station.

3.2 Transit Data

Our dataset includes the GPT for 105 subway stations from the Manhattan region, NYC (fig. 2b), and 80 subway stations from Washington DC, USA (fig. 2a). We considered the data shared by the PTAs of the two cities. For New York the Metropolitan Transportation Authority (MTA) provides information for boarding and alighting passengers for all the subway stations[1], while for Washington we exploited entrances and exits data provided by the Washington Metropolitan Area Transit Authority (WMATA)[2]. The data of New York consists of the number of turnstile entries and exits for subway stations aggregated in four hour intervals. The information we considered includes 1.135 unique turnstile positions that are associated with 732 station entrances or exits of 105 subway stations within the island of Manhattan. The data of Washington include directly the information of entrances and exits per hour for every subway station in the city, our dataset contains the entrances/exits values for 80 subway stations in Washington area. We collected two months of transit data for both cities.

In order to compare the transit data with GPT we needed a dataset of the same length. To this end, we exploited the first month of transit to create a typical weekly profile made by averaging the transit data of the same hours and days of the week.

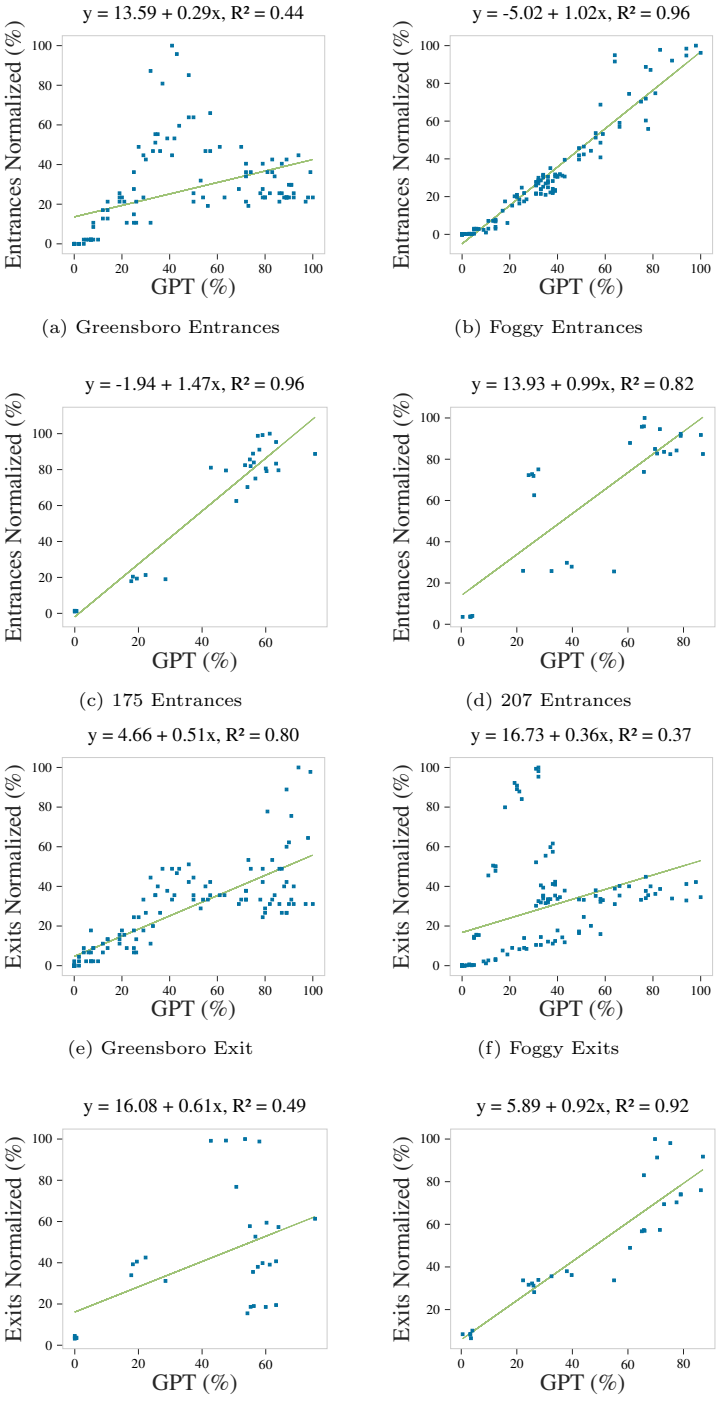
To align the transit data with the GPT of stations, we adopt a time-based aggregation approach for both the GPT and transit datasets. Washington transit data is provided on an hourly basis, allowing us to directly correlate each transit value with its respective GPT value. In contrast, New York transit data is segmented into 4-hour intervals. To match this with the GPT, we consolidate the GPT data by averaging values across the preceding four-hour span, ensuring a comparison with the corresponding transit data point. It's worth noting that while this method ensures alignment with the corresponding transit data point, it introduces potential risks. Averaging over a four-hour period can mask sudden pikes or drops in GPT within those intervals, potentially leading to inaccuracies in our estimations.

3.3 Correlation Analysis

In a preliminary analysis phase, we want to understand which information from the transit dataset of a station is the most similar to the GPT profile. The scope is to understand how the increase or decrease of the GPT percentage is correlated with the real amount of passengers entering and/or exiting from the stations. To analyze the transit usage data and its correlation with the GPT we use the following simple linear regression model:

$$G_{h,s} = \beta T_{h,s} + \epsilon, \tag{1}$$

[1]Source: http://web.mta.info/developers/turnstile.html
[2]Source: https://www.wmata.com/initiatives/ridership-portal/Rail-Data-Portal.cfm



(a) Greensboro Entrances (b) Foggy Entrances

(c) 175 Entrances (d) 207 Entrances

(e) Greensboro Exit (f) Foggy Exits

(g) 175 Exits (h) 207 Exits

Fig. 3 Correlation between GPT and transit data for 4 exemplifying station, 2 in New York and 2 in Washington

where $G_{h,s}$ is the GPT value for station s and hour of the week h, β represents the regression coefficient, ϵ is the residual error, and T is the corresponding transit data for the same station and hour. We tested this basic regression model for both transit information (entrances and exits), the sum, and the difference between the two. The performance of the regression models are evaluated using the coefficient of determination, i.e., R^2 score, which is the proportion of variation explained by independent variables.

We began by analyzing the aggregated results from the two cities in our dataset. Using the linear regression described in (1), we considered the standard GPT value across all stations in conjunction with data from entrances, exits, the sum of entrances and exits, and the difference between entrances and exits. The logic behind using these four variables is that GPT information contains human traces of both people entering and exiting each station. However, this information is expected to strongly depend on the volume of passengers transiting the station. Stations located in residential areas may reveal a much larger number of individuals entering the station in the morning than exiting, whereas the flow inverts in the afternoon peak when the commuters return home from their workplaces. Working activities are instead expected to show the opposite behavior. Therefore, we expect that GPT information contains both inflow and outflow information with different relative importance both from a spatial and a temporal dimension. The relation emerging from the computed correlation will be at the basis of the models developed within the TransitCrowd tool introduced and presented in the next section.

We compute the R^2 score derived from comparing the entire dataset of the standard GPT from all stations across both cities with the transit data. The goal is to see if the overall pattern of GPT matches any specific set of transit data more closely.

Analysing the R^2 for all stations it is clear that for both cities the entrances have generally a higher correlation ($R^2 = 0.91$ New York and $R^2 = 0.81$ Washington) than the exits ($R^2 = 0.70$ New York and $R^2 = 0.71$ Washington), and the sum of entrances and exits result ($R^2 = 0.89$ New York and $R^2 = 0.89$ Washington) shows a much higher correlation than the difference entrances-exits, which obtains the lowest scores of correlation ($R^2 = 0.29$ New York and $R^2 = 0.16$ Washington). This outcome could be explained by the fact that passengers entering a station have to wait for the subway to arrive, hence leaving a longer trace at the station as picked up by GPT, while the process of exiting is generally faster. Despite the general trend suggesting that GPT is mainly driven by the entrances profiles, at the single station level we notice the existence of a minority of the stations where the relationship is the opposite and GPT is more correlated with the exit flows. Fig. 3 shows this important aspect of the GPT-Transit relationship, we selected 2 stations per city, the R^2 values and the regression lines reveal that certain stations have high correlation with entrances (fig.3b,3c) and low correlation with exits (fig.3b, 3c) and at the same time some stations reveal an opposite behavior; examples are stations Greensboro (Washington) and 207st (New York). For both stations the GPT

is more correlated with exits (fig.3e,3h) than entrances (fig.3a,3d), but these remain a large minority of all analysed stations.

This characteristic of peculiar similarity to the exits of some stations leads us to develop a specific profile for each station able to identify the interconnection between the GPT and the transit data for a generic week.

4 The TransitCrowd Estimation Framework

The correlation analysis suggests that a linear combination of entrances and exits at each subway station was used to generate the data visualised by the Google Popular Times. Conversely, we can leverage the correlation between these two data sources to be able to infer from GPT data their input parameters. This is the idea behind a novel data-driven approach to estimate and predict transit demand, i.e. the TransitCrowd framework. However, the relative weights to assign to entrances and exit in this linear combination varies by station, and within a station by time of day, day of week and in relation to the volume of entries, exits and the time each individual dwells inside each station. We do not seek to derive a functional relationship to derive the weights, but instead we first develop a time series model to leverage the transit data and the standard GPT data for each station individually (signature of a station), which is used to define the weights in a data-drive approach and be used in combination with live GPT information to predict the specific entries and exits of each station. In order to rescale the demand to estimate and predict the actual entrances and exit flows, a suite of regression-based machine learning models is trained using the same standard GPT data and the average turnstile data.

Fig. 4 shows the process behind the TransitCrowd framework developed in this study. The methodology aims at estimating the exit and entrance profiles of every subway station in a city for a specific week. The inputs are the standard GPT, the averaged transit data, and the live GPT. The framework combines the two estimation tools, the *Sig Estimator* and the *Reg Estimator*. Both tools estimate the flows of entrances and exits at subway stations, but with different characteristics. The Reg estimator prioritizes the accuracy of the results in terms of rescaling. The Sig Estimator is based on simpler statistics methods, at the cost of a lower transferability compared to the Reg estimator.

For both estimators, we employ a station-specific model for each subway station rather than a singular overarching model. This decision was made to account for the unique characteristics and demand patterns of individual stations. Each station, depending on its location, surrounding amenities, and connectivity, can exhibit a distinct relationship of GPT and entrance and exit patterns. By focusing our models to individual stations, we aim to capture these characteristics more accurately. In the following, we describe the details of the two estimation tools.

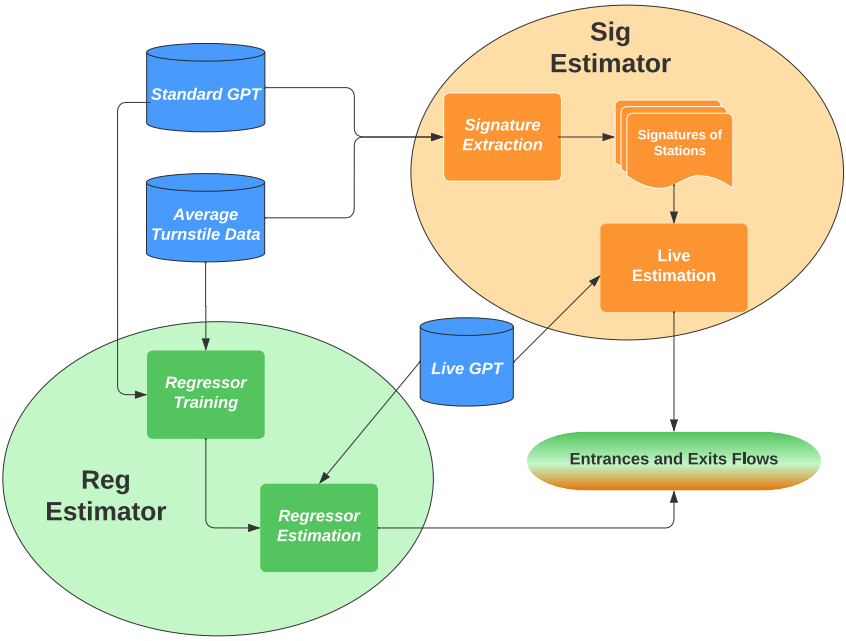


Fig. 4 TransitCrowd framework, blue symbols represent input data, orange blocks are the Sig estimator, and green ones the Reg estimator

4.1 Sig Estimator

The Sig estimator is composed by two interconnected modules: Signature extraction, and Live Estimation. The first module aims at extracting the signature that characterizes the relationship between the GPT of a single station and corresponding entrances and exits profiles. To extract this signature, we exploit the standard GPT and the averaged entrances and exits as inputs. First, we need to normalise the entrances and the exits data from the transit dataset in order to replicate the GPT scale (0-100). To do so, we apply to both entrances and exits a mix-man normalization scaling the dataset on the 0-1 interval and we then multiply by 100. The scaling procedure is the following:

$$t_{scaled} = \frac{t - \min(T)}{\max(T) - \min(T)} \cdot 100, \forall t \in T, \tag{2}$$

where T represents the exits or the entrances dataset for a single station for the selected time period, *min* and *max* are the corresponding minimum and maximum values within the same period; these two values are stored for each station and will be used in the live estimation phase. Once scaled the transit data, we compute the signature of the stations. The signature represents the additive error between the standard GPT and the scaled exits and entrances. This additive error captures the unique behavior of each station, emphasizing its distinct characteristics. For each station, we compute two signatures, one for

the entrances and one for the exits. It is important to note that this signature is not merely an error or residual but a representation of the station's unique behavior in relation to the normalized transit data. The term signature was chosen intentionally to underscore this specific station-level characteristic.

The signature calculation for entrances and exits, respectively, is the following:

$$S_{en,s} = t_{en,scaled} - GPT_s, \tag{3}$$

$$S_{ex,s} = t_{ex,scaled} - GPT_s, \tag{4}$$

where S is the signature for the station s corresponding to the transit data of entrances $t_{en,scaled}$ or exits $t_{ex,scaled}$.

In the second step, we try to estimate the real values of users exiting and entering the subway stations for a specific week by leveraging the corresponding GPT Live data. Specifically, we exploit as input the signatures $S_{ex,s}$ and $S_{en,s}$ extracted in the previous phase using past information and we combine them with the information of the current week from the Live GPT. The estimation function for the Entrances and Exits profiles of a week w is the following:

$$En_{w,s} = (S_{en,s} + GL_{s,w}) \cdot (\max_{en,s} - \min_{en,s}) + \min_{en,s}, \tag{5}$$

$$Ex_{w,s} = (S_{ex,s} + GL_{s,w}) \cdot (\max_{ex,s} - \min_{ex,s}) + \min_{ex,s}, \tag{6}$$

where max and min are the same stored from (2), $S_{ex,s}$ is the signature of exits for station s, and GL is the GPT Live data extracted for station s during week w. The same function applies also to the estimation of the entrances profile and it is repeated for every station in the dataset for 12 different weeks after the signature extraction.

4.2 Reg Estimator

With the aim of estimating the entrances/exit flows from each subway station, we selected as input the corresponding standard GPT and the averaged entrances and exits to train a suite of ML models, which were selected among the most widely and successfully used in the literature dealing with regression problems [36]. Our model is structured in two phases: regression training and regression estimation. During the regression training phase, the standard GPT, which encapsulates the historical crowdedness patterns of each station, serves as our primary input. To enhance the model's accuracy, we integrated a temporal dimension into our analysis by including information about the day of the week and the hour corresponding to the GPT value. The regression's target is shaped by the averaged data of entrances and exits, aggregated over a four-week period for each station. It is pivotal to highlight that this training approach is meticulously tailored to each station, ensuring a representation of both entrance and exit dynamics. Transitioning to the regression estimation phase, we leverage the model trained in the previous step. Instead of the

standard GPT, here we incorporate the Live GPT for each station. This real-time data, when combined with the specific day and hour corresponding to the Live GPT value, empowers our model to estimate a week of entrance and exit patterns, individually for each station in our dataset.

To validate the trained models, we employed the stratified k-fold cross-validation method, specifically opting for the standard ML choice of 10 folds [37]. This method is known for its robustness in evaluating model performance, particularly with imbalanced datasets. Stratified k-fold cross-validation maintains the same class distribution in each fold as the entire dataset, reducing bias and overfitting risks. It ensures accurate evaluation metrics and enhances the model's generalization to new data. In our study, each station is treated individually, providing a customized validation for each. While detailed errors for each fold aren't provided due to space limits, our validation aims to ensure model accuracy and generalizability across various stations and scenarios.

Moreover, each ML model has a set of hyperparameters that need to be tuned in order to improve its performance. This process, commonly known as "hyperparameter tuning", is carried out by implementing the random search method, which allows assessing the values of the hyperparameter with a larger impact on model performance.

Using R^2 as performance parameter, we assessed that the best-trained model for our approach is the Extra trees regressor [38]. It is a model of ensemble learning technique that aggregates the results of different de-correlated decision trees. Unlike traditional decision tree methods, the Extra Trees regressor selects random splits for each decision tree, which helps in adding an extra layer of randomness and reducing the variance. This characteristic makes it distinct from other ensemble methods, such as the Random Forest, where the best split among a random subset is chosen. Due to its inherent randomness and ability to handle large datasets with higher dimensionality, the Extra Trees regressor is particularly well-suited for complex regression tasks where capturing intricate patterns and relationships in the data is crucial. Once the training and the choice of the model are done, we move to the real estimation step. In this phase, we replace the GPT standard used for training with GPT live of a specific week that we want to estimate.

5 Performance Evaluation

We evaluate the performance of TransitCrowd calculating the estimation error at station level using the weighted Mean Absolute Percentage Error (wMAPE) [39]. We start analyzing the results provided by the Sig Estimator. As described in the previous section, the signature extraction is the first step of Sig estimator.

Fig. 5 gives an example by presenting the signatures of entrances and exits for the subway station "50th" in New York, the first row of the plot reveals the three datasets exploited for the signature extraction: standard GPT, entrances, and exits (scaled 0-100). The second and the third row of the plot show the

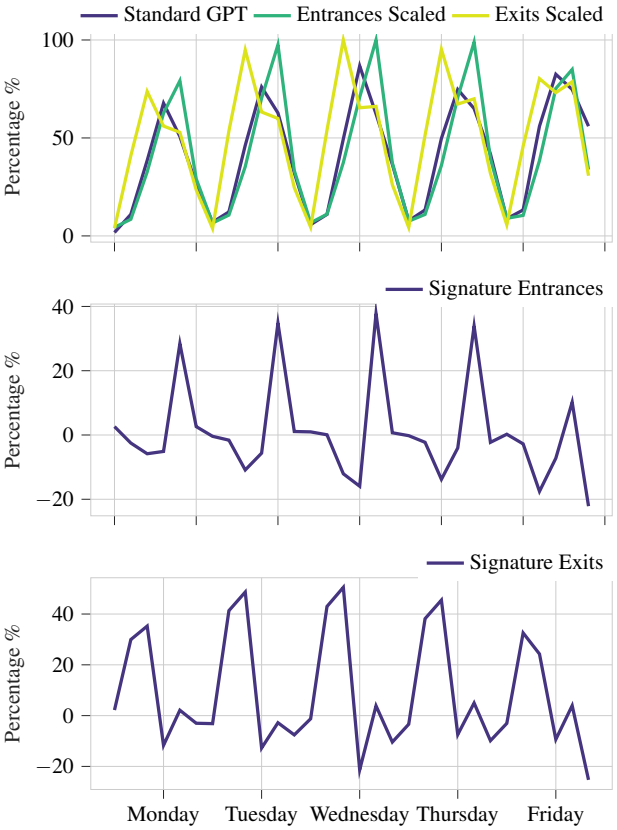


Fig. 5 Extraction of signatures profiles for subway station named "50th"

signature for the entrances and the one for the exits obtained by applying (3) and (4). It is interesting to observe that for this station, the most significant discrepancies between the GPT and the transit data occur during the morning peaks for exits signature and in the afternoon for entrances signature. These results indicate that the signature of this specific station assigns a higher weight to the passengers exiting the station in the morning, and a higher weight to the individuals exiting the station in the afternoon, suggesting that the station is located in a working area where there is a higher share of commuters arriving in the morning and leaving in the evening.

To further showcase the added value of the Sig estimation results for classifying stations' behavior and identifying similarities between stations Fig. 6 shows the results of a simple k-means clustering for all stations in NYC. The results showing the highest Silhouette score indicates 5 distinct clusters, with four (indicated with cyan, purple, violet and green dots) containing the largest part of the stations. Notably, the two major working areas, i.e. the areas around Downtown/Wall Street and the one around Times Square, are clearly identified. This analysis demonstrates the potential of the Sig estimator to

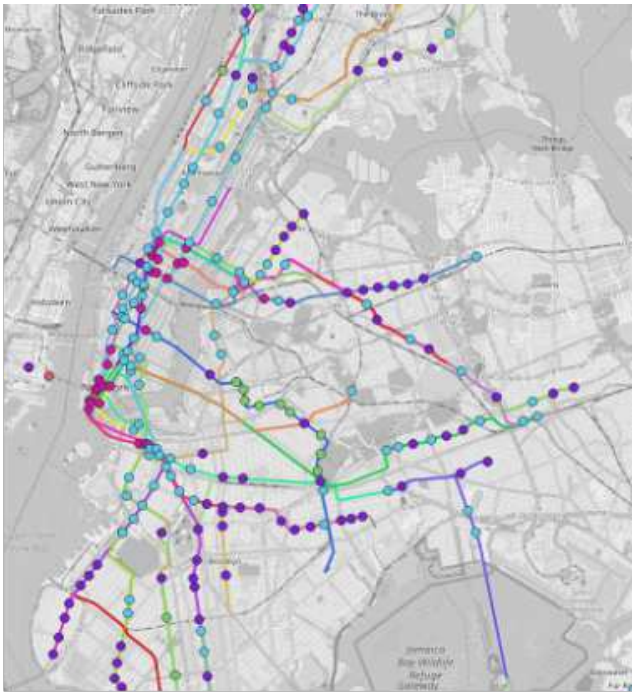


Fig. 6 Clustering of stations based on the signatures estimations

identify commonalities in mobility patterns and classify the stations using a data-driven approach. We will further explore this application opportunity in future research.

Once the signatures for all stations are extracted from the reference month, we are ready to leverage the GPT Live data for estimation of the real flows of entrances and exits. Fig. 7 shows an example of the result of the estimation process in a single station (Dupont Circle, Washington) for 1 week following the signature extraction month. The upper part of the figure reveals the profile of the GPT Live for the corresponding week, then the lower part presents the real estimation for entrances and exits produced by applying the matching signature. The figure depicts a good result for this single station, most of the peaks reached by the ground-truth are replicated by the estimated flows. It is interesting to notice that the estimation error for this station is stable throughout the week, this is a first signal that our prediction results are not deteriorating along different days.

We continue our analysis by looking at the results of Reg estimator. Fig. 8 presents the entrances estimation errors in terms of weighted Mean Absolute Percentage Error (wMAPE) of Reg for New York stations at different hours of the day. The wMAPE is expressed as:

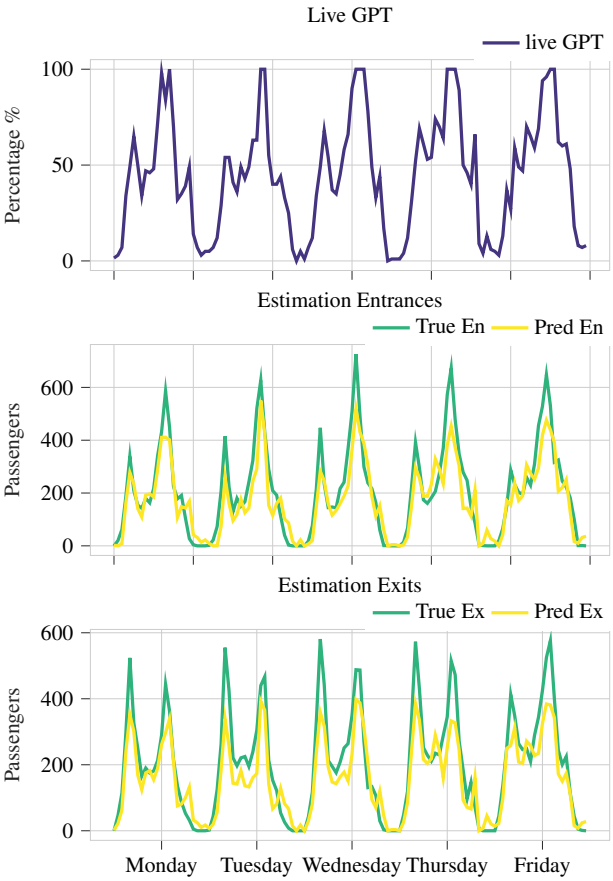


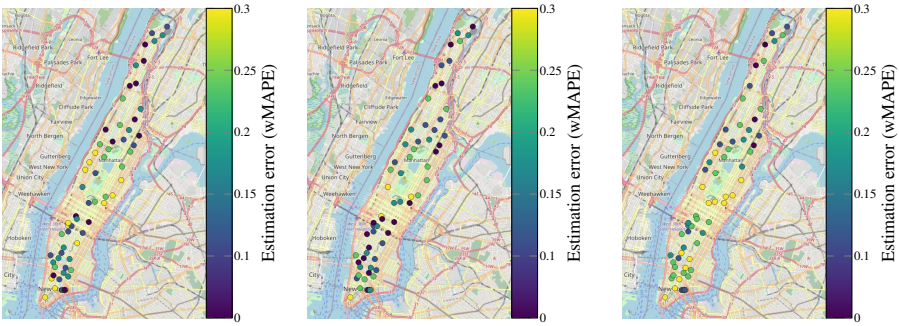
Fig. 7 Live Estimation results for sig estimator, profiles of the predicted and true values of turnstile data for week 1 after the signature extraction, for station Dupont Circle, Washington

$$\text{wMAPE} = \frac{\sum_{i=1}^{n} \|\overline{y_i} - y_i\|}{\sum_{j=1}^{n} y_j} \tag{7}$$

where $\overline{y_i}$ are the estimated values, y_i the observed values, or ground truth, and n is the length of these two series.

From the maps it is interesting to notice that the stations in the center of Manhattan are characterized by higher errors throughout the day. Moreover, Fig. 8c depicts how the errors in the evening are larger than in other day periods.

It is worth noting that other factors, such as station type, may also play a role in these errors. Overall, this figure highlights the need for further analysis and improvements to the Reg estimator to ensure more accurate entrance estimations for subway stations in New York City.



(a) H: 12:00 (b) H: 16:00 (c) H: 20:00

Fig. 8 Estimation error for stations in New York at different hours of a working day

Table 1 Estimation error for all stations of Manhattan, New York

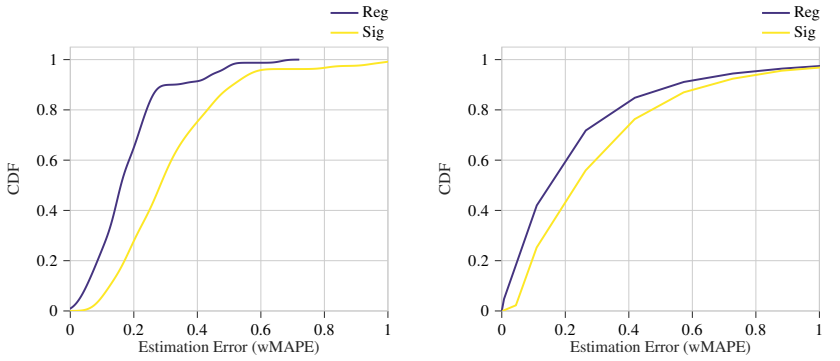
Week after training	**Error Entrances(wMAPE)**		**Error Exits(wMAPE)**	
	Sig	Reg	Sig	Reg
1	0.378	0.350	0.370	0.305
2	0.309	0.218	0.278	0.118
3	0.306	0.236	0.278	0.150
4	0.308	0.263	0.271	0.178
5	0.263	0.235	0.263	0.183
6	0.390	0.269	0.396	0.193

Having illustrated the estimation results for single stations for Sig and Reg estimator, Fig. 9 shows the performances of our framework on entrances for every station in our dataset in terms of wMAPE. The results are in form of a cumulative distribution function (CDF), every station contributes to the plot with a value of wMAPE that represents the estimation error made by the framework to estimate the entrance flow.

As expected, the Reg estimator produces lower errors, since for both plots the violet line representing Reg is always on the left of the Sig line. The Reg estimator obtains errors lower than 0.2 for the 60% of the estimations, while the errors of Sig estimator are less than 0.3 for the 60% of dataset in both cities.

The difference between the two cities is more evident in the interval $[0.6-1]$, here we can notice that New York CDF shows higher errors, both estimators reach values greater than 0.5 for a small portion of estimations (10%). The main outcome of the CDFs is that Reg estimator obtains better estimation results than Sig tool. Therefore, the proposed idea of an estimator prioritizing accuracy (Reg tool) is confirmed.

Once analyzed the estimation performances on the full dataset we want to analyze the evolution throughout the weeks, the scope is to recognize if our results are deteriorating along the weeks after the training.

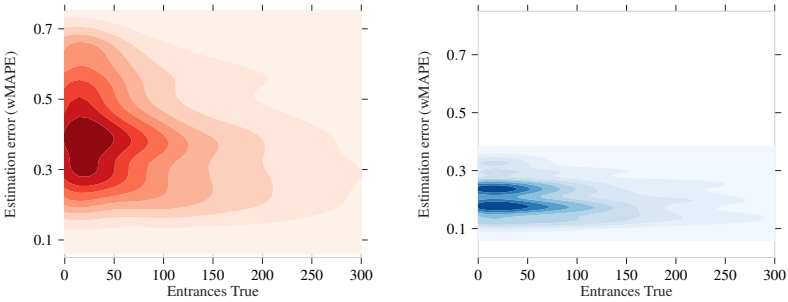


(a) Washington (b) New York

Fig. 9 Cumulative estimation error entrances for all stations using both Sig and Reg estimator

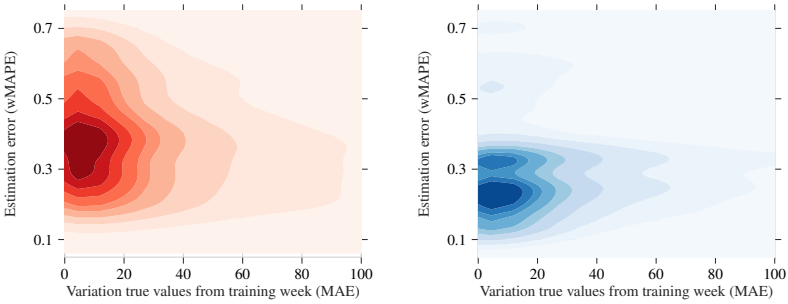
Tab. 1 shows the performance of this estimation process on the New York dataset period for the two estimators, it includes the average wMAPE of the estimations in all stations for the entrances and the exits for all the weeks in the data collection interval. The table shows that the estimation process is stable over the weeks, and the wMAPE is always contained in the interval $[0.2 - 0.3]$. it is notable that for both estimators the error does not appear to systematically increase along the different weeks, moreover the week with the lower errors is the 5th after training. This means that for Sig estimator the signatures extracted before week 1 are still valid also after the 2 months of the data collection, at the same time the Reg estimator does not require new training process after several weeks of estimations. Continuing with our analysis, We want to understand if the estimation errors of TransitCrowd are influenced by the amount of entrances/exits we are estimating. While Fig. 7, representing a single station, displayed higher errors for peak entrance and exit values, we present Fig. 10 to provide a broader perspective on how the model performs at high and low entrance and exit values across multiple stations.

Fig. 10 presents two density plots in order to visualize the distribution of the estimation errors over the values of entrances. The density plots are based on Washington results and concern both the Sig estimator(fig.10a) and the Reg estimator(fig.10b). Errors from Sig estimator are concentrated around wMAPE values of 0.35. In contrast, as for the previous results, the performances of Reg estimator are slightly better, and the errors focus on the interval $[0.25 - 0.3]$. The significant outcome of Fig. 10 is that for both estimators errors are not increasing with the rise of entrances values, it is remarkable that for entrances around $200 - 250$ the errors remain the same that for entrances $0 - 50$. This result indicates that, overall, our model performs well across the entire dataset of stations, even during peak entrance and exit periods. Finally, Fig. 11 presents the relationship between the prediction error of entrances and the variation of true values from the values of the reference week (i.e. the



(a) Distribution estimation error Sig (b) Distribution estimation error Reg

Fig. 10 Distribution estimation error entrances along true values



(a) Distribution estimation error Sig (b) Distribution estimation error Reg

Fig. 11 Distribution estimation error entrances along variation from training week

week of the signature extraction). The y-axis indicates the accuracy of the estimations (wMAPE between prediction and true values), while the x-axis depicts the deviation of the true values from the reference week (Mean absolute error between reference week and true values). These density plots reveal that the estimation error is not influenced by the variation of each week, this means that TransitCrowd is able to estimate with similar errors standard weeks and weeks different from the reference one.

6 Discussion

Despite the promising results of our framework, TransitCrowd, for estimating public transport demand flows, there are several limitations that must be discussed. The first limitation concerns the testing of the framework under normal conditions only. While the signature concept was able to accurately replicate the normal trends of transit demand, it is unclear how well it will perform in unusual or unexpected situations, such as extreme weather events, public holidays, or other events where the demand at subway stations significantly deviates from normal weekly patterns. The second limitation relates

to the transferability of the methodology. Currently, our approach has only been tested in two cities with similar characteristics, and it is uncertain how well the methodology will be transferable to other cities. Different cities have different structures, cultural behaviors, and daily routines, which can impact the accuracy of the results. The signature concept is based on the correlation between GPT data and entrance/exit flows at subway stations, and these correlations may not exist in the same way in another city. Additionally, the city's infrastructure, such as the size and layout of the subway network, may also impact the accuracy of the results. Further studies are required to fully understand the conditions under which our framework can be successfully transferred to other cities, as well as to determine the characteristics of cities that are suitable for transfer. Another limitation of TransitCrowd is tied to our station-specific modeling approach. While this approach provides a more precise representation of each station's transit demand patterns, it also introduces logistical challenges. Deploying and managing multiple models for each station can be resource-intensive. This is particularly true for the regressor estimator, which demands more resources compared to the signature method. As a consequence, these logistical challenges highlight an advantage of the signature method over the regression approach. The regression approach remains a viable solution when data is available for a specific period, and the necessary resources for model computation are available. These limitations of our study highlight the importance of further investigation in this field to fully understand the potential of using GPT data for estimating public transport demand flows.

6.1 Future Works

In our future research, we aim to extend the scope of this study, with a particular focus on enhancing the transferability of the Signature Estimator. Transferability is a key feature of the signature model, and we plan to develop it in the following manner: First, we will extract the signatures from a "training city" where transit data is readily available, exploiting the TransitCrowd tool for this purpose. Subsequently, we will analyze various characteristics of the station, including its type (whether it is a transfer station or not), the number of lines passing through it, and features of the surrounding catchment area. This will involve studying both static and dynamic data. Static elements include the quantity and types of activities around the station, while dynamic elements are represented by temporal demand data, sourced from GPT of the activities. Once we have identified the features that most significantly influence a station's signature, we will apply this knowledge to a "testing city" where transit data is unavailable. Our goal will be to identify the signature of stations in that city based on the characteristics defined in the previous steps. After estimating the signatures for the stations in the testing city, we will reintroduce the Transitcrowd tool, specifically in the "Live Estimation" section. By applying the corresponding GPT Live to the station signatures, we aim to

obtain an indication of the entrances and exits patterns at the stations in a city where no transit data is available.

7 Conclusion

In this work, we investigated the potential to leverage GPT to estimate public transport demand flows, specifically focusing on the subway. By exploring this crowdsensed data, we identified that GPT data can be correlated with entrance patterns of the majority of subway stations, while the crowdedness of a subset of stations is linked with their exits flows.

We developed TransitCrowd, a framework that exploits GPT to make live estimations of transit data at subway station level. Our framework is flexible, being composed of two distinct estimator tools. The first, Reg estimator, prioritizes the accuracy of results focusing on the city level. The second, Sig estimator, extracts signatures from stations revealing the temporal profile of correlations between GPT and entrances/exits. Through this fundamental information, it is possible to apply the presented methodology to other cities. Finally, we evaluated the performance of TransitCrowd, estimating two months of entrance/exit flows using as input the GPT Live data for each station.

The estimation process produced promising results whose accuracy appears to be stable over the different weeks considered. We observed that TransitCrowd is able to properly estimate weeks that are different from the training one, and that the errors are not influenced by the high or low values of entrance/exit flows.

Future works will focus on analyzing the signatures of different stations to identify influential factors, such as activities around the stations or sociodemographic data. Once such factors are detected, the final goal is to estimate signatures for stations in another city in order to test the transferability of our estimation process to a new environment.

ACKNOWLEDGMENT

Mr. Vitello is supported by the Luxembourg National Research Fund (PRIDE17/12252781/DRIVEN).

References

[1] Pelletier, M.-P., Trépanier, M., Morency, C.: Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies **19**(4), 557–568 (2011)

[2] Rodrigues, M., Teoh, T., Ramos, C., Knezevic, L., Marcucci, E., Lozzi, G., Gatta, V., Cré, I., for Internal Policies of the Union, E.P.D.-G., Panteia, *et al.*: Relaunching Transport and Tourism in the EU After COVID-19: Transport Workers vol. pt. 2. European Parliament, ??? (2021)

[3] London, T.: Travel in london report 14. Transport London: London, UK (2021)

[4] GVB: Jaarverslag 2021 (2021)

[5] Tao, Z., Tang, J., Hou, K.: Online estimation model for passenger flow state in urban rail transit using multi-source data. Computer-Aided Civil and Infrastructure Engineering **36**(6), 762–780 (2021). https://doi.org/10.1111/mice.12671

[6] Capponi, A., Fiandrino, C., Kantarci, B., Foschini, L., Kliazovich, D., Bouvry, P.: A survey on mobile crowdsensing systems: Challenges, solutions and opportunities. IEEE Communications Surveys Tutorials, 1–49 (2019). https://doi.org/10.1109/COMST.2019.2914030

[7] Lau, S.L., Sabri Ismail, S.M.: Towards a real-time public transport data framework using crowd-sourced passenger contributed data. In: Proc. of IEEE VTC-Fall), pp. 1–6 (2015). https://doi.org/10.1109/VTCFall.2015.7391180

[8] Myrvoll, T.A., Håkegård, J.E., Matsui, T., Septier, F.: Counting public transport passenger using WiFi signatures of mobile devices. In: Proc. of IEEE ITSC, pp. 1–6 (2017). https://doi.org/10.1109/ITSC.2017.8317687

[9] Shlayan, N., Kurkcu, A., Ozbay, K.: Exploring pedestrian bluetooth and WiFi detection at public transportation terminals. In: Proc. of IEEE ITSC, pp. 229–234 (2016). https://doi.org/10.1109/ITSC.2016.7795559

[10] Capponi, A., Vitello, P., Fiandrino, C., Cantelmo, G., Kliazovich, D., Sorger, U., Bouvry, P.: Crowdsensed data learning-driven prediction of local businesses attractiveness in smart cities. In: Proc. of IEEE ISCC, pp. 1–6 (2019). https://doi.org/10.1109/ISCC47284.2019.8969771

[11] Welch, T.F., Widita, A.: Big data in public transportation: a review of sources and methods. Transport Reviews **39**(6), 795–818 (2019). https://doi.org/10.1080/01441647.2019.1616849

[12] Wang, W., Attanucci, J.P., Wilson, N.H.M.: Bus passenger origin-destination estimation and related analyses. (2011)

[13] Foell, S., Kortuem, G., Rawassizadeh, R., Phithakkitnukoon, S., Veloso, M., Bento, C.: Mining temporal patterns of transport behaviour for predicting future transport usage. In: Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication. UbiComp '13 Adjunct, pp. 1239–1248. Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2494091.2497354

[14] Dixit, M., Sivakumar, A.: Capturing the impact of individual characteristics on transport accessibility and equity analysis. Transportation Research Part D: Transport and Environment **87**, 102473 (2020). https://doi.org/10.1016/j.trd.2020.102473

[15] Aguiléra, V., Allio, S., Benezech, V., Combes, F., Milion, C.: Using cell phone data to measure quality of service and passenger flows of paris transit system. Transportation Research Part C: Emerging Technologies **43**, 198–211 (2014). Special Issue with Selected Papers from Transport Research Arena

[16] Lou, X., Yan, M.: Classifying subway passengers based on mobile network data analysis. In: Proc. of IEEE/ACIS ICIS, pp. 92–96 (2021). https://doi.org/10.1109/ICIS51600.2021.9516871

[17] Demissie, M.G., Phithakkitnukoon, S., Sukhvibul, T., Antunes, F., Gomes, R., Bento, C.: Inferring passenger travel demand to improve urban mobility in developing countries using cell phone data: A case study of senegal. IEEE Transactions on Intelligent Transportation Systems **17**(9), 2466–2478 (2016). https://doi.org/10.1109/TITS.2016.2521830

[18] Wang, X., Zhou, Z., Yang, Z., Liu, Y., Peng, C.: Spatio-temporal analysis and prediction of cellular traffic in metropolis. In: Proc. of IEEE ICNP, pp. 1–10 (2017). https://doi.org/10.1109/ICNP.2017.8117559

[19] Zhao, J., Zhang, L., Ye, K., Ye, J., Zhang, J., Zhang, F., Xu, C.: Gltc: A metro passenger identification method across afc data and sparse wifi data. IEEE Transactions on Intelligent Transportation Systems, 1–15 (2022). https://doi.org/10.1109/TITS.2022.3171332

[20] Ding, X., Liu, Z., Xu, H.: The passenger flow status identification based on image and wifi detection for urban rail transit stations. Journal of Visual Communication and Image Representation **58**, 119–129 (2019). https://doi.org/10.1016/j.jvcir.2018.11.033

[21] Oransirikul, T., Nishide, R., Piumarta, I., Takada, H.: Measuring bus passenger load by monitoring Wi-Fi transmissions from mobile devices. Procedia Technology **18**, 120–125 (2014). https://doi.org/10.1016/j.protcy.2014.11.023. International workshop on Innovations in Information and Communication Science and Technology, IICST 2014, 3-5 September 2014, Warsaw, Poland

[22] Kostakos, V., Camacho, T., Mantero, C.: Towards proximity-based passenger sensing on public transport buses. Personal and Ubiquitous Computing **17**, 1807–1816 (2013). https://doi.org/10.1007/s00779-013-0652-4

[23] Wu, R., Cao, Y., Liu, C.H., Hui, P., Li, L., Liu, E.: Exploring passenger dynamics and connectivities in beijing underground via bluetooth networks. In: 2012 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), pp. 208–213 (2012). https://doi.org/10.1109/WCNCW.2012.6215492

[24] Hu, X., Zheng, H., Wang, W., Li, X.: A novel approach for crowd video monitoring of subway platforms. Optik **124**(22), 5301–5306 (2013). https://doi.org/10.1016/j.ijleo.2013.03.057

[25] Zhang, J., Liu, J., Wang, Z.: Convolutional neural network for crowd counting on metro platforms. Symmetry **13**(4) (2021). https://doi.org/10.3390/sym13040703

[26] Solmaz, G., Baranwal, P., Cirillo, F.: CountMeIn: Adaptive crowd estimation with Wi-Fi in smart cities. In: Proc. of IEEE PerCom, pp. 187–196 (2022). https://doi.org/10.1109/PerCom53586.2022.9762354

[27] Pender, B., Currie, G., Delbosc, A., Shiwakoti, N.: Social media use during unplanned transit network disruptions: A review of literature. Transport Reviews **34**(4), 501–521 (2014). https://doi.org/10.1080/01441647.2014.915442

[28] Pereira, F.C., Rodrigues, F., Ben-Akiva, M.: Using data from the web to predict public transport arrivals under special events scenarios. Journal of Intelligent Transportation Systems **19**(3), 273–288 (2015). https://doi.org/10.1080/15472450.2013.868284

[29] Moyo, T., Musakwa, W.: Using crowdsourced data (twitter & facebook) to delineate the origin and destination of commuters of the gautrain public transit system in south africa. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences **III-2**, 143–150 (2016). https://doi.org/10.5194/isprs-annals-III-2-143-2016

[30] Timokhin, S., Sadrani, M., Antoniou, C.: Predicting venue popularity using crowd-sourced and passive sensor data. Smart Cities **3**(3), 818–841 (2020). https://doi.org/10.3390/smartcities3030042

[31] Dixon, J., Elders, I., Bell, K.: Evaluating the likely temporal variation in electric vehicle charging demand at popular amenities using smartphone locational data. IET Intelligent Transport Systems **14**(6), 504–510 (2020). https://doi.org/10.1049/iet-its.2019.0351

[32] Fry, D., Aaron Hipp, J., Alberico, C., Huang, J.-H., Lovasi, G.S., Floyd, M.F.: Land use diversity and park use in new york city. Preventive Medicine Reports **22**, 101321 (2021). https://doi.org/10.1016/j.pmedr.2021.101321

[33] Mahajan, V., Cantelmo, G., Antoniou, C.: Explaining demand patterns during COVID-19 using opportunistic data: a case study of the city of munich. European Transport Research Review **13**(1), 1–14 (2021)

[34] Vitello, P., Capponi, A., Klopp, P., Connors, R.D., Viti, F., Fiandrino, C.: The CORONA Business in Modern Cities: Poster Abstract

[35] Bandeira, J.M., Tafidis, P., Macedo, E., Teixeira, J., Bahmankhah, B., Guarnaccia, C., Coelho, M.C.: Exploring the potential of web based information of business popularity for supporting sustainable traffic management. Transport and Telecommunication Journal **21**(1), 47–60 (2020). https://doi.org/10.2478/ttj-2020-0004

[36] Hagenauer, J., Helbich, M.: A comparative study of machine learning classifiers for modeling travel mode choice. Expert Systems with Applications **78**, 273–282 (2017). https://doi.org/10.1016/j.eswa.2017.01.057

[37] Nti, I., Nyarko-Boateng, O., Aning, J.: Performance of machine learning algorithms with different k values in k-fold cross-validation. International Journal of Information Technology and Computer Science **6**, 61–71 (2021). https://doi.org/10.5815/ijitcs.2021.06.05

[38] Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine learning **63**, 3–42 (2006)

[39] Kolassa, S., Schütz, W.: Advantages of the mad/mean ratio over the mape. Foresight: The International Journal of Applied Forecasting, 40–43 (2007)