Estimating Public Transport Demand Information Using Crowdsourced Data

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

[1]Mobilab Transport Research Group, Faculty of Sciences, Technology and Communication, University of Luxembourg, Luxembourg

SHORT SUMMARY

The analysis of transit demand and its complex dynamics has typically relied on survey-based data that captures only a small fraction of the total demand. Recently, emerging data-driven approaches have been applied to transportation issues and these typically rely on sensing data gathered by mobile devices under the so-called mobile crowdsensing (MCS) paradigm. This type of data can be a powerful source of information especially in areas where transit data is not available. This work aims to investigate the possibility of using Google Popular Times (GPT), a widely available crowdsensed data, to estimate the passenger flows of individual subway stations. Our results show that we can estimate precisely both entrances and exits profiles, which is particularly challenging because GPT only provide popularity trends. Our analysis is carried out on more than 105 subway stations of Manhattan and it is validated using turnstile count data from the stations.

Keywords: Crowdsourced Data, Public Transport Demand, Google Popular Times, Big Data Analytics

1. INTRODUCTION

In the era of Advanced Public Transport Systems, new technologies have been introduced and deployed providing multiple sources of data that can be utilized for demand estimation and analysis. Operators can rely on smart card data and automated passenger counts that can be used to estimate transit demand and its variation under different operational conditions (Pelletier, Trépanier, & Morency, 2011). Simultaneously, the pervasiveness of mobile devices permits to use Information and Communications Technology (ICT) by unleashing unprecedented possibilities in urban environments to improve citizens' quality of life and services. Citizens carrying smart devices are a potential data source according to the mobile crowdsensing (MCS) paradigm (Capponi et al., 2019). The market for such data is projected to raise USD 5460.4 Million by 2028 at a compound annual growth rate of 24.73[1]. Crowdsourced data have been introduced in transportation, through different applications to identify special events and disruptions or to monitor travel behavior and provide complementary information (Nandan, Pursche, & Zhe, 2014). This work aims to bring one step further the research on Public Transport demand estimation by overcoming the limitations of traditional techniques with the use of crowdsensed data. Specifically, in this paper we leverage Google Popular Times (GPT) a crowdsourced dataset provided by Google; we want to investigate its potential as a source of information for public transport demand. The worldwide availability of GPT then opens up the possiblity to estimate public transport demand in areas where such information is not typically collected. Other types of crowdsourced data have been exploited for public transport analysis. An example is given by (Lau & Sabri Ismail, 2015), which propose a framework that provides realtime public transport data using crowdsourced information provided

[1]https://www.databridgemarketresearch.com/reports/global-crowd-analytics-market

by passengers smartphones, the shortcomings of this approach are that passengers have to actively contribute to the framework and that developing a crowdsourced campaign from scratch requires a big effort. By comparison, GPT are already available and do not require any data collection campaign, the main limitation of GPT is that since the data is directly provided by Google the processing made on the raw data are unknown. This study aims at overcoming this shortcoming by combining the GPT with real public transport data. Only few studies started to explore the importance of GPT for the transportation field, the authors of (Bandeira et al., 2020) investigate the possibility of using GPT to predict traffic volumes in a specific area, with their work they found a clear relationship between GPT, traffic, and environmental performances. Another interesting analysis of GPT is the one from (Timokhin, Sadrani, & Antoniou, 2020), this research 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. According to the best knowledge of the authors, no research has focused specifically on analysing the potential of GPT to estimate public transport information. This study intends to fill this research gap. In the remainder of the manuscript, Section 2 illustrates the data employed in the study, the methodology is described in Section 3. Finally, Section 4 present the results and Section 5 concludes the work and highlights the final remarks.

2. Dataset

In this section we present the types of data we use in our analysis of transit stations demand together with the preliminary observations

Google Popular Times

GPT is a feature of Google maps that visualise the standard temporal profile of 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 is generated from data sent anonymously by smartphones' users who have the google history location enabled, the locations of these devices is tracked in the background and sent to Google through WiFi or mobile networks. Together with the standard week profile, GPT provide a live value for the current hour, this value assesses the actual level of crowdedness at the place. 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). However, this hides the absolute quantity of the demand, i.e. the real number of customers. In this work we are not interested in the GPT of all activities, but only on the data regarding the category subway stations. Our dataset includes the GPT for 105 subway stations from the Manhattan region.

Turnstile Data

For analyzing the public transport demand we consider the dataset of the Metropolitan Transportation Authority (MTA) which provides information for boarding and alighting passengers for all the subway stations of New York[2]. The data consists of the number of turnstile entries and exits for subway stations of New York aggregated in four hour intervals. The data we considered includes 1,135 unique turnstile positions that are associated with 732 station entrances or exits and 105 stations. We collected three months of the MTA turnstile dataset. In order to compare the turnstile data with GPT we needed a dataset of the same length. To this end we exploited the first month of turnstile to create a typical weekly profile made by averaging the turnstile data of the same hours and days of the week. Since our GPT dataset includes only Manhattan, we consider during

[2]Source: http://web.mta.info/developers/turnstile.html

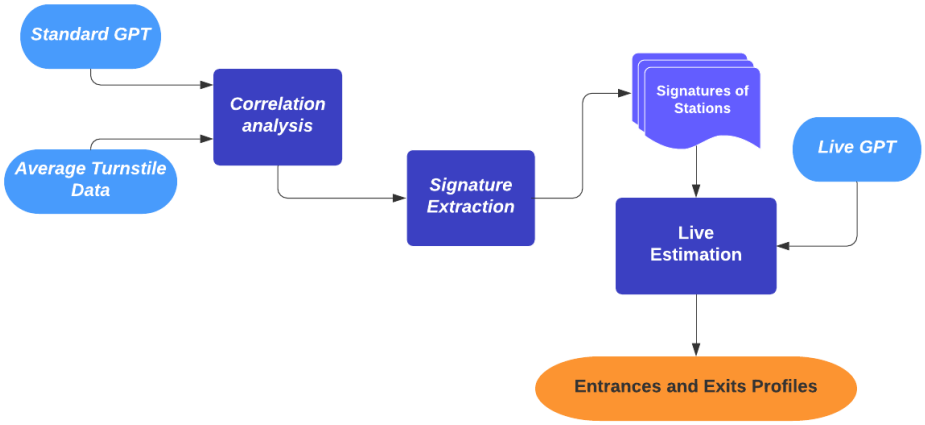


Figure 1: General estimation framework

analysis and evaluation only the subway stations from that area of the city.

3. THE ESTIMATION FRAMEWORK

Fig.1 shows the methodology framework exploited by this study. The methodology aims at estimating the exit and entrance profiles of a specific week for every subway station in Manhattan. The Inputs exploited are the standard GPT, the Turnstile data averaged, and the Live GPT. The framework is composed of three interconnected phases: Correlation analysis, Signature extraction, and Live Estimation.

Correlation analysis

In this first phase, we want to detect which information from the Turnstile dataset of a station is the most similar to the GPT profile. The scope is to understand how the increasing or decreasing of the GPT percentage is correlated with the real amount of passengers entering or exiting from the stations. To analyze the turnstile usage data and its correlation with the GPT we use the following linear regression model:

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

where G is the GPT value for station s and hour of the week h, β represents the regression coefficient, and ε is the residual error. T is the turnstile data, we tested the regression model for both information of turnstile information (Entrances and Exits), and the sum of the two. The performances of the regression models are evaluated using the coefficient of determination or R^2 score which is the proportion of variation explained by independent variables.

Signature Extraction

From the results of the correlation phase (described in Sec.4) it transpires that the GPT of each station has a different correlation with the turnstile data: in some stations the GPT is more correlated with the entrances while in others with the exits. This phase aims at extracting the signature that characterizes the relationship between the GPT of a single station and corresponding entrances and exits profiles. Similarly to the previous phase, we exploit as input the standard GPT and the averaged entrances and exits. First, we need to transform the entrances and the exits data in order to replicate the GPT scale (0-100). 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:

$$T_{scaled} = \frac{T - min(T)}{max(T) - min(T)} \cdot 100 \tag{2}$$

where T represents the exits or the entrances dataset for a single station, *min* and *max* are the corresponding minimum and maximum values, these two values are stored for each station and will be used on the Live estimation phase. Once scaled the turnstile data, we compute the signature of the stations, the signature represents the margin between the standard GPT and the scaled exits and entrances. For each station we compute two signatures, one for the entrances and one for the exits. The signature calculation is the following:

$$S_{en,s} = En_{scaled} - G_s \tag{3}$$

$$S_{ex,s} = Ex_{scaled} - G_s \tag{4}$$

where S is the signature for the station s corresponding to the turnstile data of entrances En or exits Ex.

Live estimation

In the third step, we try to estimate the real values of users exiting and entering the subway stations for a specific week leveraging the corresponding GPT Live data. Specifically, we exploit as input the signatures S_en and S_ex 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 Exits profile of a week w is the following:

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

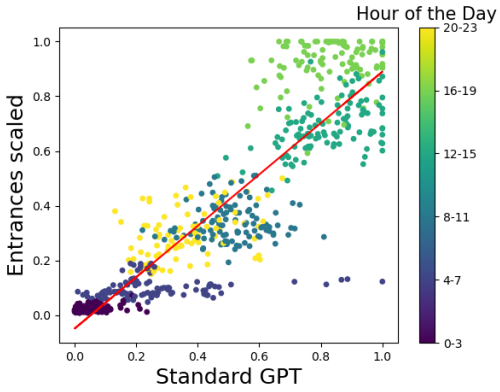
where *max* and *min* are the same stored from eq.2, $S_{ex,s}$ is the signature of exits for station s, and GL is the GPT Live data extracted for station s during week w. The same function applies also to the estimation of the entrances profile and it is repeated for every station in the dataset for 12 different weeks after the signature extraction. Finally, we calculate the estimation error using the Normalized Root Mean Square Error (NRMSE):

$$\text{NRMSE} = \frac{\sqrt{\sum_{1=1}^{n} \frac{(\overline{y_i} - y_i)^2}{n}}}{M}, \tag{6}$$

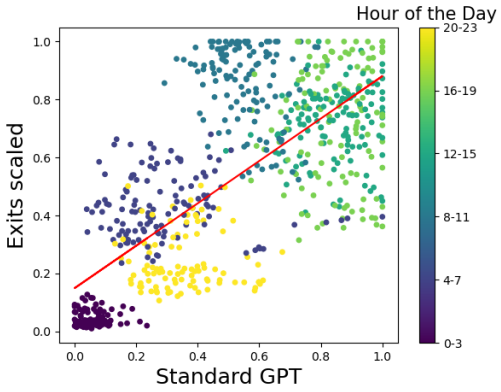
where $\overline{y_i}$ are the estimated values, y_i the observed values, or ground truth, M is the mean of the observed values, and n is the length of these two series.

4. RESULTS

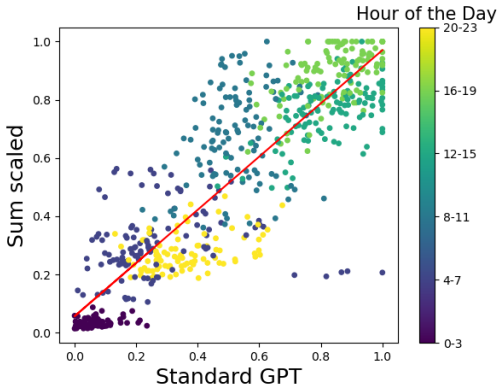
We first start by analyzing the results obtained in the correlation phase. We applied the linear regression described in eq.1 to the standard GPT of all stations, together with the data from entrances, exits, and the sum entrances+exits. Fig.2 shows the scatter plots of the dataset together with the results of the regression. The colorbar represents the hour of the day. Looking at the R^2 it is clear that the entrances have a better correlation ($R^2$0.79) than the exits ($R^2$0.42), and the sum entrances+exits does not improve on the entrances result. This outcome could be explained by the fact that passengers entering a station have to wait for the subway to arrive, 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



(a) Entrances
$R^2 = 0.79$
$G = 0.93T - 0.04$



(b) Exits
$R^2 = 0.52$
$G = 0.73T + 0.14$



(c) Sum
$R^2 = 0.77$
$G = 0.91T + 0.05$

Figure 2: Correlation between turnstile data and GPT, each station contributes 168 points (total hours of a week) to each plot

and GPT is more correlated with the exit flows. Fig.2 shows us another important aspect of the GPT-Turnstile relationship, the noise in the plots tend to intensify in specific intervals of the day, for the entrances the hours with more noise are $16h - 19h$, while for the exits are the $8h - 11h$ intervals. This characteristic of time dependency on the correlation GPT-Turnstile, together with the peculiar similarity to the exits of some stations lead us to develop a specific profile for each station able to identify the interconnection between the GPT and the Turnstile data for a generic week. Fig.3 presents the signatures of entrances and exits for the subway station "50th", 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 signature for the entrances and the one for the exits obtained by applying eq.3 and eq.4. Looking at the signatures for this station, it is interesting to note that are almost always negative for the full period, above all it is clear that the biggest differences between the GPT and the turnstile arise during morning peaks. GPT seems not to display the same high percentages of exits and entrances during mornings, this information contained in the signatures will be crucial for the estimation process. 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.4 shows the result of the estimation process in a single station (50th) for the 2 weeks 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

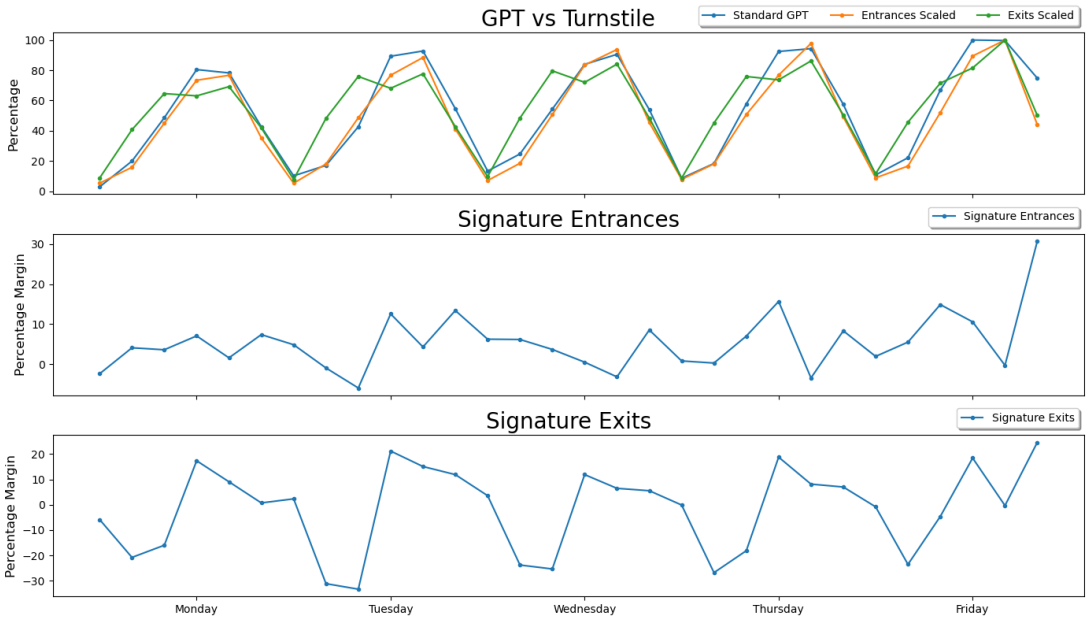


Figure 3: Extraction of signatures profiles for 50th subway station

of the peaks reached by the ground-truth are replicated by the estimated flows, it is interesting to notice that the NRMSE for this station is greater for the first week $0,25$ than for the second 0.24. Having illustrated the estimation results for a single station, tab.1 contains the performance of this estimation process on the entire dataset, it includes the averaged NRMSE of the estimations in all stations for the entrances and the exits for all the weeks in the data collection interval. Tab.1 shows that the estimation process is stable along the weeks, the NRMSE is always contained in the interval $0.3 - 0.4$, it is notable that the error does not appear to systematically increase along the different weeks, this means that the signatures extracted before week 1 are still valid also after the 3 months of the data collection.

Table 1: Estimation error for all stations

Week after extraction	**Error Exits** (Avg NRMSE)	**Error Entrances** (Avg NRMSE)
1	0.36	0.36
2	0.34	0.35
3	0.36	0.40
4	0.38	0.36
5	0.38	0.33
6	0.39	0.40

5. CONCLUSIONS

In this work, we investigated the potential to leverage GPT to estimate public transport demand flows. By exploring this crowdsourced data, we identified that GPT can be correlated with the entrance pattern of the majority of subway stations, while the crowdedness of a subset of stations is linked with the exits flows. We observed that the relationship between GPT and public transport demand depends on the day of the week and the hour of the day. Therefore we extracted from each station 2 signatures revealing the temporal profile of the correlation between GPT and entrances/exits. Finally, we estimated two months of entrance/exit flows by applying the extracted signatures to GPT Live data from each station. The estimation process produced promising results

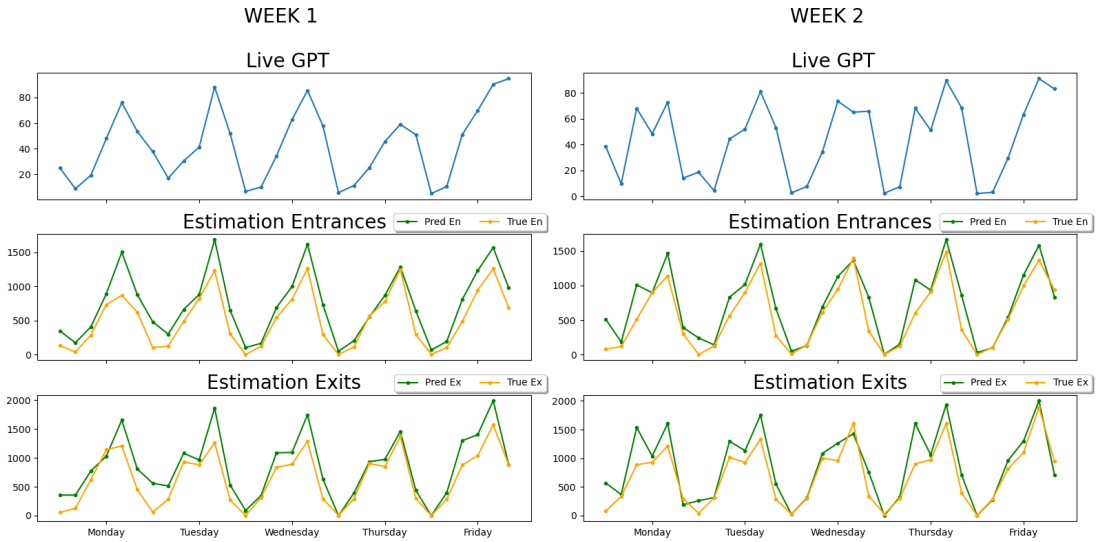


Figure 4: The profiles of the predicted and true values of turnstile data for week 1 and 2 after the signature extraction, for station 8th

whose accuracy appears to be stable over the different weeks. 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

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

Capponi, A., Fiandrino, C., Kantarci, B., Foschini, L., Kliazovich, D., & Bouvry, P. (2019, May). A survey on mobile crowdsensing systems: Challenges, solutions and opportunities. *IEEE Communications Surveys Tutorials*, 1-49. doi: 10.1109/COMST.2019.2914030

Lau, S. L., & Sabri Ismail, S. M. (2015). Towards a real-time public transport data framework using crowd-sourced passenger contributed data. In *2015 ieee 82nd vehicular technology conference (vtc2015-fall)* (p. 1-6). doi: 10.1109/VTCFall.2015.7391180

Nandan, N., Pursche, A., & Zhe, X. (2014). Challenges in crowdsourcing real-time information for public transportation. In *2014 ieee 15th international conference on mobile data management* (Vol. 2, pp. 67–72).

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.

Timokhin, S., Sadrani, M., & Antoniou, C. (2020). Predicting venue popularity using crowd-sourced and passive sensor data. *Smart Cities*, *3*(3), 818–841. Retrieved from https://www.mdpi.com/2624-6511/3/3/42 doi: 10.3390/smartcities3030042