Transportation Research Interdisciplinary Perspectives 37 (2026) 101973

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

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

Estimating traffic from crowdsourced location activity: A new data source perspective

Olivér Törő [a,*], Domokos Esztergár-Kiss [b,c], Jan-Dirk Schmöcker [d], Tamás Tettamanti [a,c]

[a] Department of Control for Transportation and Vehicle Systems, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, Budapest, H-1111, Hungary
[b] Department of Transport Technology and Economics, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, Budapest, H-1111, Hungary
[c] Systems and Control Lab, HUN-REN Institute For Computer Science and Control, Kende u. 13-17, Budapest, H-1111, Hungary
[d] Department of Urban Management, Faculty of Engineering, Kyoto University, Katsura Nishikyo-ku, Kyoto, 615-8540, Japan

ARTICLE INFO

Keywords:
Traffic estimation
Google Popular Times (GPT)
Point of Interest (POI)
Principal Component Analysis (PCA)

ABSTRACT

Accurate road traffic estimation is essential for efficient traffic management and strategic urban planning. However, traditional infrastructure-based traffic sensing systems face significant limitations in providing high-resolution, network-wide coverage due to prohibitive installation and operational costs. Consequently, there is a growing need to explore alternative data sources to complement conventional traffic monitoring. To address this gap, this paper investigates and justifies the applicability of crowdsourced social activity data as a novel data source for road traffic estimation. In the era of smartphones, social activity information is widely available, with representative data accessible at the majority of Points of Interest (POIs). In this study, a relationship is established between the social activity represented by Google Popular Times (GPT) data at POIs and conventional traffic information, such as speed (from Floating Car Data) and number of vehicles (from video sensors). Spatiotemporal correlation and prediction capabilities are presented and proven using real-world data from two different metropolitan areas: Budapest (Hungary) and Kyoto (Japan). A probabilistic model is introduced, enabling the regression of traffic data on GPT using a dimensionality reduction approach. To capture temporal features, forward- and backward-shifted social activity data is also applied in the estimation methodology. The analysis justifies that only a fraction of the social activity information is enough for efficient traffic prediction (both for speed or number of vehicles), resulting in a few percent relative errors.

1. Introduction

There has been a strong focus on road traffic speed and flow measurement, estimation, and prediction in recent decades. Accordingly, a plethora of papers, as well as surveys, have been published in this field (Shaygan et al., 2022; Gomes et al., 2023). Consequently, traffic detection has become a highlighted topic as this is the base of efficient traffic management (Papageorgiou, 2004), e.g., for applying traffic-responsive traffic light control or for calculating the eco-routing of vehicles. Furthermore, civil engineers also need traffic data for traffic infrastructure planning and development projects (Asmael et al., 2024; Farooq et al., 2026). Besides, reliable traffic data is also the preliminary input for traffic engineers' modeling activities, i.e., microscopic or macroscopic traffic simulation (Lopez et al., 2018; Thonhofer et al., 2018).

A known problem of current traffic detection is the lack of proper infrastructure based traffic sensing, i.e., maintaining a complete network-wide traffic sensing system with high resolution is unrealistic due to the high installation and operational costs of roadside infrastructure sensors (Tettamanti et al., 2014; Hruboš et al., 2021). Instead, this paper focuses on alternative data from which reliable traffic estimation can be deduced, i.e., social activity (Moro Zamprogno and Esztergár-Kiss, 2024) data at Points of Interest (PoI) can also be leveraged efficiently for traffic estimation purposes. Nowadays, we are facing an increasing amount of data from navigation (e.g., Waze, Google Maps), most of them provide traffic data in terms of travel time. That is, the traffic flow (in other words, how many vehicles are present in the networks) is not provided. Traffic flow data, however, is an essential data for both strategic planning and real-time traffic operation. Our solution, combined with cross-sectional traffic data (validated via real-world

* Corresponding author.
E-mail addresses: toro.oliver@kjk.bme.hu (O. Törő), esztergar-kiss.domokos@kjk.bme.hu (D. Esztergár-Kiss), schmoecker@trans.kuciv.kyoto-u.ac.jp (J.-D. Schmöcker), tettamanti.tamas@kjk.bme.hu (T. Tettamanti).

https://doi.org/10.1016/j.trip.2026.101973
Received 17 June 2025; Received in revised form 24 March 2026; Accepted 26 March 2026
Available online 28 March 2026

vehicle count data later in our case study), is capable of extending and predicting traffic flow data.

Modern traffic detection systems may include other devices in addition to traditional detection techniques (loop detector, camera, radar, and any other roadside sensors), i.e., alternative sensor technologies (which are not specifically developed for traffic sensing) can be used for traffic monitoring too. An emerging technology is traffic monitoring by means of distributed acoustic sensing using fiber optic cables already installed for communication purposes (Fakhruzi et al., 2025; van den Ende et al., 2022). Another example of alternative traffic detection is the data gathered from fleet management systems (Gunawan and Chandra, 2014; Tettamanti et al., 2024), or from mobile networks (Fekih et al., 2022).

Nowadays, we have access to an emerging amount of data. This is in contrast to the traffic detection techniques commonly used in traffic engineering (Alessandretti et al., 2020; Balsa-Barreiro et al., 2022). Pentland (2015) showed that it is possible to reveal the information flow patterns in a social network, regardless of the information actually exchanged. In this vein, the features of the traffic network under study can be accurately predicted. Furthermore, various social media-based studies are available to reveal urban travel behavior (Zhang et al., 2017) or to predict road traffic jams (Moyano et al., 2021). These studies are mainly based on Twitter data, thanks to the formerly freely available Twitter API. Twitter data were used, besides the traditional traffic detectors, for congestion prediction by Adetiloye and Awasthi (2019). Alomari et al. (2020) developed a traffic event detection tool and implemented an event classifier using Twitter data via machine learning. Osorio-Arjona and García-Palomares (2019) presented a Twitter data based method for origin–destination travel matrix estimation. Pourebrahim et al. (2019) utilized random forest models for Twitter data to calculate the trip distribution in an urban traffic network. Mariani et al. (2019) examined the history of the use of Facebook for travel decision-making. D'Andrea et al. (2015) invented a traffic event detector applying a text-mining algorithm for Twitter data with high efficiency.

Early research explored the use of social media and online mapping services to infer trip purposes and activity patterns. Cui et al. (2018) integrated data from Google Places and Twitter to predict individuals' trip purposes using a Bayesian neural network, demonstrating that combining POI-based contextual information with social activity data significantly improves prediction accuracy. Their work showed that the spatial distribution and popularity of POIs can serve as proxies for travel demand and activity intensity, laying the foundation for data-driven mobility modeling.

Building on this, recent studies have directly leveraged Google Popular Times (GPT) data, a real-time and anonymized measure of activity levels at POIs, as an indicator of human presence. Mahdi et al. (2023) modeled the time spent at POIs using GPT data across several activity categories, confirming that GPT closely reflects observed behavioral patterns such as visit duration and temporal activity cycles. Similarly, in Vongvanich et al. (2023) GPT data is applied to predict transit station demand, showing that the aggregated GPT activity of surrounding POIs explains a substantial portion of passenger flow variability. These findings establish GPT as a reliable proxy for local mobility dynamics and urban activity intensity.

Despite these advances, the use of GPT in road traffic estimation remains limited. Prior work has primarily focused on individual travel behavior or public transport demand, while the relationship between POI activity levels and vehicular traffic conditions has not been extensively explored. Our study fills this gap by integrating GPT data with camera-based and floating car traffic data to estimate traffic states via a spatiotemporal modeling framework. In particular, we investigate how temporal variations in aggregated POI activity correlate with road-level traffic metrics, providing new insights into the collective dynamics linking land-use activity and traffic flow.

All the studies mentioned above dealt with social-media data for traffic analysis or estimation. However, the activity of travelers, measured at POI locations as Google Popular Times (provided by Google Maps), has not yet been exploited for traffic estimation to the best of our knowledge. Although a few papers have been published, e.g., Kan et al. (2022), Gong et al. (2020), investigating the traffic dynamics with the consideration of POI data sets, they also considered fleet data as input for the estimation, and POI activity was therefore always measured in an indirect way from vehicular traffic in the vicinity of the POIs. Additionally, our previous paper (Kolat et al., 2023) dealt with POI activity data in the context of traffic dynamics by revealing the relationship between the POI activity and road traffic applying heuristics. Yet, it did not provide a model to describe this relationship or leverage its capability for traffic prediction. Accordingly, the present study provides solutions to these problems with the specific contributions listed below:

- Traffic data and GPT data are synchronized and resampled to allow a meaningful correlation analysis. Missing database entries are interpolated where needed.
- A probabilistic model, based on Principal Component Analysis (PCA), is proposed to capture the road traffic dynamics solely from GPT data.
- To capture temporal features, we created an estimation method based on time-shifted POI data, resulting in a generalized regression model capable of predicting both vehicle count and fleet traffic data.
- The proposed methods have been validated via real-world data: Floating Car Data (FCD) provided by TomTom in Budapest, as well as via vehicle count measurements from Kyoto.

The paper is structured as follows. After the introductory section, Section 2 introduces the applied data sources. Section 3 presents the methodology utilized in the research, including the probabilistic model approach, the Principal Component Analysis, and the regression method. Section 4 presents the case study locations and discusses the results of applying the proposed methods to real-world data. Finally, Section 5 summarizes the paper and gives an outlook for further research.

2. Data sources

2.1. Google popular times at points of interests

In Google Maps, a Point of Interest is defined as a specific location that may be useful or interesting to an individual. These locations can be restaurants, cafes, shops, tourist attractions, parks, museums, transportation hubs, public services, offices, banks, and other commercial establishments. There are several methodologies to determine the level of activity at a given location, collectively referred to as popular times. These methods involve collecting data from user devices, which is then used to generate a graphical representation of crowd levels throughout the day. The data collected is categorized into three main types: crowdsourced data, location history, and real-time data. The popular times feature provides insights into the typical crowd levels at a specific POI throughout the day and week.

The GPT data tries to represent how busy a particular POI is. The numeric values are between 0 and 100, so these are not absolute but relative to the POI's normal attendance.

The method to generate GPT data is proprietary; however, we know that it contains the following four kinds of information: (1) how busy a location usually is during a day; (2) real-time data showing how busy a location is right now; (3) average duration of visits; and (4) wait time estimates. For our analysis, we used the real-time activity data, which is generated at an hourly resolution; thus, it forms a time series (Google, 2025).

2.2. Vehicle flow data

Vehicle counting using cameras is a key method for estimating traffic, leveraging computer vision and Artificial Intelligence (AI). Traditional approaches use image processing techniques like background subtraction or edge detection to sense and count vehicles. However, deep learning-based methods provide higher accuracy by detecting and segmenting vehicles in real-time.

Camera-based traffic count is of key importance as traffic flow (also called traffic volume) is a base data for civil and traffic engineers for transportation planning or operation.

In this research, we utilized cross-sectional traffic flow data, i.e., vehicle count per time period, processed from video cameras.

2.3. Floating car data

The Floating Car Data (FCD) was collected by TOMTOM (2025). One advantage of FCD over vehicle count data is its ability to measure both historical and real-time data. Additionally, unlike vehicle counters, which provide information at specific locations (measured link cross sections), TomTom data covers the entire District XI of Budapest. FCD is based on Global Navigation Satellite System data, utilizing navigation devices and applications. FCD is a sample dataset relative to the total traffic data, as it is gathered solely from vehicles equipped with TomTom-based navigation systems. The core principle of the FCD technique involves the exchange of information between a fleet of floating cars and a central data system. The floating cars frequently transmit recently collected data on their positions and speeds. The central data system tracks this data along traveled routes by matching the related trajectories to the road network.

Therefore, the accuracy of FCD travel time estimations depends on the percentage of floating cars participating in the traffic flow. The average travel time and speed along road links are the most common and valuable information derived from FCD. Overall, evidence indicates that FCD-based travel time estimation is a reliable method, as demonstrated by the efficiency of common navigation systems (such as Waze, Google Maps, Here, or TomTom) that utilize their own FCD information.

In this work, we use the hourly average speed values at specific road sections.

3. Methodology

This section provides the theoretical background to establish a relationship between social activity represented by the Google Popular Times data at POIs and the conventional traffic information, such as speed (from Floating Car Data) and number of vehicles (video sensor measurement). The latter two will be referred to as traffic data when distinguishing between them is not necessary.

The following notations will be used. Lowercase bold letters are vectors, uppercase bold letters are matrices, calligraphic fonts represent random variables, and $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ stands for a Gaussian probability density function (pdf) in the variable $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

3.1. Probabilistic model

The GPT data originates from users' phones who chose to opt-in for this service; the data cannot perfectly represent the true activity and holds uncertainty. Only a fraction of the total visitors of a POI carry a phone enabled to use the GPT service, and this fraction cannot be known precisely. As a consequence, we will regard the GPT data as a random variable that is assembled as the sum of the true activity plus random noise. In particular, we model the GPT data of a POI as a truncated normal ($\mathcal{TN}$) random variable. By stacking all the POI's data into the random vector $\boldsymbol{\mathcal{X}}$, we get the probabilistic description with pdf $p_{\mathcal{X}}(\boldsymbol{x})$ as

$$p_{\mathcal{X}}(\boldsymbol{x})=\mathcal{TN}_{\mathcal{X}}\left(\boldsymbol{x};\boldsymbol{\mu}_{\mathcal{X}},\boldsymbol{\Sigma}_{\mathcal{X}},\mathcal{I}\right)=\frac{\exp\left(-1/2(\boldsymbol{x}-\boldsymbol{\mu}_{\mathcal{X}})^\top\boldsymbol{\Sigma}_{\mathcal{X}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\mathcal{X}})\right)}{\int_{\mathcal{I}}\exp\left(-1/2(\boldsymbol{x}-\boldsymbol{\mu}_{\mathcal{X}})^\top\boldsymbol{\Sigma}_{\mathcal{X}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\mathcal{X}})\right)\mathrm{d}\boldsymbol{x}}\,, \tag{1}$$

where $\boldsymbol{\mu}_{\mathcal{X}}$ is the true activity, $\boldsymbol{\Sigma}_{\mathcal{X}}$ represents the uncertainty, and $\mathcal{I}$ is the truncation interval. For our case $\mathcal{I}$ is taken to be the P-dimensional hypercube $[0,100]^P$.

Let $\mathcal{Y}$ represent a random scalar quantity describing the traffic at a specific road section. We are to establish a relationship between $\mathcal{Y}$ and $\boldsymbol{\mathcal{X}}$. Specifically, we are interested in the conditional pdf $p_{\mathcal{Y}|\mathcal{X}}(y|\boldsymbol{x})$ that will also serve as a likelihood function. Let us consider a linear Gaussian latent variable model (Jiang et al., 2012)

$$p_{\mathcal{Y}|\mathcal{X}}(y|\boldsymbol{x})=\mathcal{N}\left(y;f(\boldsymbol{x}),\Sigma\right)\,, \tag{2}$$

where $f(\boldsymbol{\mathcal{X}})$ is the latent scalar variable which is a linear function of $\boldsymbol{\mathcal{X}}$. The model has the form

$$f(\boldsymbol{\mathcal{X}})=\boldsymbol{\beta}^\top\boldsymbol{\mathcal{X}}+\mu\,, \tag{3}$$

where $\boldsymbol{\beta}^\top$ is a constant row vector and μ is a constant scalar. This model corresponds to the relation

$$\mathcal{Y}=\boldsymbol{\beta}^\top\boldsymbol{\mathcal{X}}+\mu+\epsilon\,, \tag{4}$$

where ϵ is a zero mean Gaussian noise with variance Σ. With these the likelihood (2) writes

$$p_{\mathcal{Y}|\mathcal{X}}(y|\boldsymbol{x})=\mathcal{N}\left(\mathcal{Y};\boldsymbol{\beta}^\top\boldsymbol{x}+\mu,\Sigma\right)\,. \tag{5}$$

If μ is taken in the form

$$\mu=\mu_{\mathcal{Y}}-\boldsymbol{\beta}^\top\boldsymbol{\mu}_{\mathcal{X}} \tag{6}$$

then we have

$$p_{\mathcal{Y}|\mathcal{X}}(y|\boldsymbol{x})=\mathcal{N}\left(y;\boldsymbol{\beta}^\top(\boldsymbol{x}-\boldsymbol{\mu}_{\mathcal{X}})+\mu_{\mathcal{Y}},\Sigma\right)\,. \tag{7}$$

By marginalizing out $\boldsymbol{\mathcal{X}}$ one gets

$$p_{\mathcal{Y}}(y)=\int p_{\mathcal{Y}|\mathcal{X}}(y|\boldsymbol{x})p_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{8}$$

$$=\int\mathcal{N}\left(y;\boldsymbol{\beta}^\top(\boldsymbol{x}-\boldsymbol{\mu}_{\mathcal{X}})+\mu_{\mathcal{Y}},\Sigma\right)\mathcal{TN}\left(\boldsymbol{x};\boldsymbol{\mu}_{\mathcal{X}},\boldsymbol{\Sigma}_{\mathcal{X}},\mathcal{I}\right)\mathrm{d}\boldsymbol{x}\,. \tag{9}$$

The integral in (9) is generally not tractable, although one can approximate the distribution using Monte Carlo methods. Alternatively, an analytic approximation could be applied. Since the conditional pdf (5) depends on $\boldsymbol{x}$ only through $\boldsymbol{\beta}^\top\boldsymbol{x}$ we introduce the random variable $\mathcal{Z}=\beta^\top\mathcal{X}$ with pdf

$$p_{\mathcal{Z}}(z)=\int\delta(z-\beta^\top\boldsymbol{x})p_{\mathcal{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\,. \tag{10}$$

Now we can write the marginal pdf (9) as

$$p_{\mathcal{Y}}(y)=\int p_{\mathcal{Y}|\mathcal{Z}}(y|z)p_{\mathcal{Z}}(z)\mathrm{d}z\,. \tag{11}$$

If the truncation is slight moment matching is a viable approximation. Using

$$\mu_{\mathcal{Z}}=\boldsymbol{\beta}^\top\mu_{\mathcal{X}} \tag{12}$$

$$\mathrm{Var}(\mathcal{Z})=\boldsymbol{\beta}^\top\mathrm{Cov}(\boldsymbol{\mathcal{X}})\boldsymbol{\beta} \tag{13}$$

we obtain

$$p_{\mathcal{Y}}(y)\approx\mathcal{N}\left(y;\boldsymbol{\beta}^\top\mu_{\mathcal{X}}+\mu,\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\mathcal{X}}\boldsymbol{\beta}+\Sigma\right)\,. \tag{14}$$

Formulas (9) and (14) can be interpreted as generative models that can be used to estimate missing $\mathcal{Y}$ values. Our uncertainty in the model (4) is represented by Σ. If we were to consider a deterministic relation or assume low uncertainty, then Σ can be neglected. However, the POI data has a low time resolution and, by its nature, cannot be regarded as precise.

3.2. Determining the β coefficient

To determine the value of the β coefficient, the traffic and POI data need to be analyzed. Let the $N \times 1$ column vector y store the time series data of a road segment and X contain the POI data as

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1P} \\ x_{21} & x_{22} & \dots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NP} \end{bmatrix} \tag{15}$$

with N measurements and P POI locations.

We are interested in a linear relation between y and X. The mean value of y will be denoted by μ_y so the centralized data is

$$y_0 = y - \mathbf{1}_N \mu_y , \tag{16}$$

where $\mathbf{1}_N$ stands for the N-element column vector containing all ones. Similarly, the column-wise centralized matrix X will be denoted by X_0.

Using the above notations, we are interested in the linear relation

$$y_0 = X_0 \beta_0 , \tag{17}$$

that helps determine the regression coefficient β_0 for the centralized data.

Since typically $P > N$, we need dimension reduction to solve for β_0. To this end, we apply Principal Component Regression (PCR) (Pearson, 1901; Jolliffe, 2002; Greenacre et al., 2022). PCR is a two-step procedure that first applies Principal Component Analysis (PCA) to the predictor matrix X_0 and then performs regression using the transformed variables. PCR is particularly suitable for our problem because the number of predictors greatly exceeds the number of observations, which makes direct regression ill-posed. Moreover, POI activity data are often highly correlated, and PCA provides an effective way to identify dominant activity patterns and mitigate multicollinearity. Thus, PCR offers a robust and interpretable approach to exploit the latent structure of the data while reducing overfitting.

PCA decomposes X_0 as

$$X_0 = SL , \tag{18}$$

where, according to the usual terminology, L is the loading matrix and S is the score matrix.

The regression coefficient is computed by

$$\beta_0 = (SS^\top)^{-1} S^\top y_0 . \tag{19}$$

We want to use the regression on the original POI data. Thus, we will use $L\beta_0$ instead of β_0. Furthermore, for the decentralized data, we need the extended regression coefficient

$$\beta = \begin{bmatrix} \frac{1}{N} \mathbf{1}_N^\top \left(y - XL\beta_0 \right) \\ L\beta_0 \end{bmatrix} \tag{20}$$

which enables to make predictions about y in the form

$$\hat{y} = \left[\mathbf{1}_N , X \right] \beta . \tag{21}$$

3.3. Regression with time-shifted data

To capture temporal features with the analysis, time-shifted values of the POI data are also used in the regression discussed above.

The time-shifted POI data matrix Z assembles as

$$Z = \left[\dots , X_{t-2}, X_{t-1}, X_t, X_{t+1}, X_{t+2} , \dots \right] , \tag{22}$$

where t is the time index. The amount of temporal difference, or timestep, and the number of components in Z can be chosen arbitrarily. The ordering also does not matter since the regression and the eigendecomposition are unaffected by the arrangement. Substituting Z in place of X, the formulas in the previous section apply for the time-shifted evaluation.

In the case studies presented below, we used a 15-minute shift both forward and backward in time. Hence, for a given time index t, the time-shifted data that goes into the model is

$$Z = [X_{t-1}, X_t, X_{t+1}] . \tag{23}$$

3.4. Note on Gaussian assumption

Assuming approximate normality provides a useful probabilistic interpretation. Under the Gaussian assumption, the covariance matrix fully characterizes the dependence structure of the data, and PCA corresponds to the maximum-likelihood solution of a linear Gaussian latent variable model. In our case, this assumption is applied as a simplification to the residual component of the POI activity data rather than to the entire time series, implying that deviations around the mean activity patterns are modeled as Gaussian noise. This formulation ensures that PCA captures the dominant linear modes of variation across POIs, providing a tractable approach to extract the main spatiotemporal structure of the dataset.

4. Results

This section presents the experimental evaluation and validation of the proposed traffic estimation framework. The objective is to assess how effectively the probabilistic regression model captures the relationship between social activity, represented by GPT data, and conventional traffic indicators, such as average vehicle speed and traffic volume. To ensure a comprehensive assessment, the analysis proceeds through multiple stages.

First, the preprocessing and synchronization of heterogeneous datasets are described, including procedures for temporal alignment, interpolation, and data completeness verification. Subsequently, the estimation performance is examined in two distinct urban contexts, Budapest (Hungary) and Kyoto (Japan), representing different data sources and traffic characteristics. To reveal how the input data and the performance scales, an ablation study is conducted, which in this setup involves varying how many days are included to compute the regression coefficient. The days not included are the ones for which prediction is made, and since the ground truth is known, error metrics such as Root Mean Squared Error (RMSE) and Mean Relative Error (MRE) can be computed.

Further experiments are conducted to investigate the influence of input parameters on estimation accuracy. These include a sensitivity analysis with respect to the number of POIs and observation days, as well as a large-scale statistical evaluation over the entire Kyoto dataset. Collectively, these tests aim to quantify both the estimation reliability and scalability of the proposed approach under real-world conditions.

4.1. Data synchronization and data imputation

To evaluate the proposed estimation method, we prepare the acquired real-world data as follows. The GPT data, which is a `.csv` file, is time-based and contains data for each hour for every POI in a separate file. This is restructured to be a single POI-based dataset that enables the extraction of social activity data for a selected location and a time slot.

The regression (21) is established between the traffic data and GPT data individually for every timeslot using multiple weekday data. We construct the matrix X such that it contains GPT data column-wise and the rows are for the individual observations. Thus, by selecting P POIs and N days for the regression, we create a N by P matrix.

The GPT time series data is approximately hourly with 10–15 min typical deviations. To be able to correlate meaningfully with the traffic data, the GPT data is resampled at a 15-minute rate using linear interpolation. The TomTom speed data is exactly hourly, and there are no missing entries.

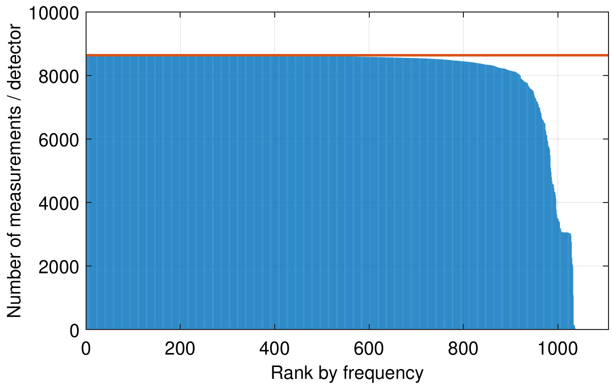


Fig. 1. Distribution of vehicle count data in the Kyoto dataset. Red line indicates the theoretical maximum value of $8640 = 12 \cdot 24\text{h} \cdot 30\text{days}$ measurements/detector throughout a month using 5 min sample time (12 observations per hour). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

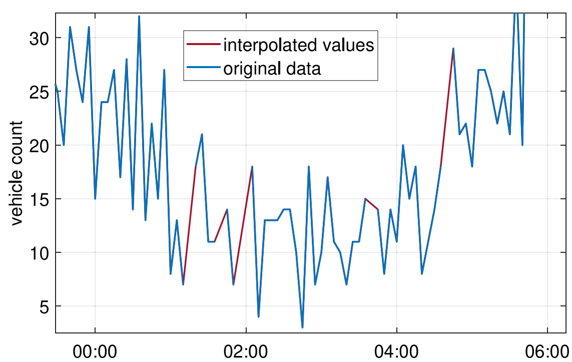


Fig. 2. Example of estimating missing vehicle count data entries with linear interpolation on 28.9.2021.

The sampling time for the vehicle count data is exactly 5 min. The distribution of the registered measurements per camera is shown in Fig. 1. We discarded the detectors that had less than 3000 measurements in total. For most of the cameras, the missing measurements are isolated, meaning the missing entries are flanked by existing data points on both sides. Consequently, the missing entries are filled with interpolated values, as can be seen in Fig. 2 as an example.

4.2. Case study locations

The case studies were conducted in two different locations, in Budapest, Hungary, and Kyoto, Japan. In Kyoto, a rectangular region is selected as the area of interest, which amounts to $131.66\,\text{km}^2$ (see Fig. 3(a)). In Budapest, District, XI. is selected, which is a $33.48\,\text{km}^2$ area (see Fig. 3(b)). POI locations for both locations are shown in Fig. 4.

The reason behind the choice of the case study locations is twofold. On the one hand, the traffic data sources were different in the two spots, still very common in practice, i.e., the Budapest area was analyzed using FCD (providing average speed information) and the Kyoto area via camera sensors measuring vehicle flow (Fig. 5). On the other hand, the two study areas reflect different traffic, i.e., typical European as well as Japanese travel behavior. This diversity makes the validation of the methodology more credible.

4.3. Case study in Budapest (Hungary)

We have collected data from 290 POIs in the region of interest. Data are distributed across POIs according to the histogram in Fig. 6(b). The theoretical maximum number of observations per POI is determined by the sampling resolution and the length of the observation period. In the Budapest dataset, the observation period covers 19 days, from 08.11.2021 to 26.11.2021, resulting in a maximum of $19 \times 24 = 456$ observations per POI. In practice, not all POIs reach these theoretical maxima because the availability of GPT data depends on user activity and data collection consistency. Some POIs have fewer valid entries due to temporary unavailability or missing hourly data, while a few Budapest POIs slightly exceed the theoretical maximum owing to occasional sampling intervals shorter than one hour.

From the 290 POIs, there were 140 that had data for the 12 evaluated weekdays in the given time interval. The POI data thus constitute a 12×140 matrix $\boldsymbol{X}$. The covariance matrix of this dataset is of rank 11 because there are 12 observations for each POI. Since the number of observations is much smaller than the number of POIs, we can regard the data in X as linearly independent, $\boldsymbol{X}$ having rank 12. The traffic data coming from the TomTom service is the average speed at a specific link. The speed values have a 60-minute resolution.

Results for Nagyszőlős road are presented in the following. Fig. 7(a) shows the true (blue) and estimated (orange) traffic data values for the afternoon 13–14 time slot. The regression used days $4 \dots 9$, and the estimation happened for the first and last 3 days. The vertical stems indicate the estimation errors.

Fig. 7(b) shows results for the 17–18h time slot. In this example, the regression used data from only the last 5 and the estimation happened for the first 7 days. In both figures, it can be seen that practically two principal components are enough to explain the variance in the data.

The RMSE and MRE values are computed only for the days on which the estimations happened (red stems). Table 1 summarizes the error metrics for selected, typical off-peak and peak periods, i.e., early and late afternoon hours in both cities.

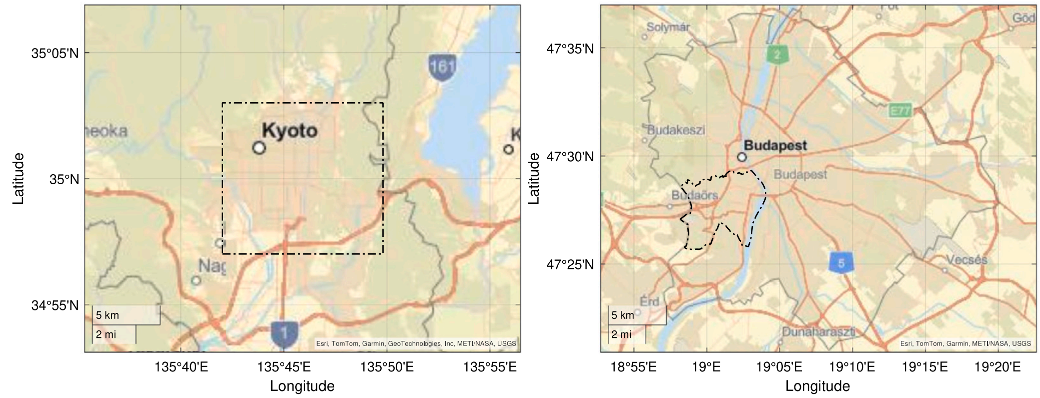
4.4. Case study in Kyoto (Japan)

In the Kyoto region, we have 1366 POIs in the area of interest, of which 624 have no missing data for the selected time intervals. The observation period spans 30 days, from 01.09.2021 to 30.09.2021, with a quite precise hourly resolution; the theoretical maximum equals $24 \times 30 = 720$ observations per POI. The traffic data comes from video sensors that have a 5-minute sampling rate.

Two locations are selected: Gojo-Dori St. and Marutamachi-Dori St. (see Fig. 5). Fig. 8(a) shows estimation results for Gojo-Dori St. in the 17–18h timeslot. The regression used the odd days from the first half of the dataset. The same setup with ± 15 minutes time-shift is presented in Fig. 8(b). The estimation error decreased significantly due to the inclusion of time-shifted data.

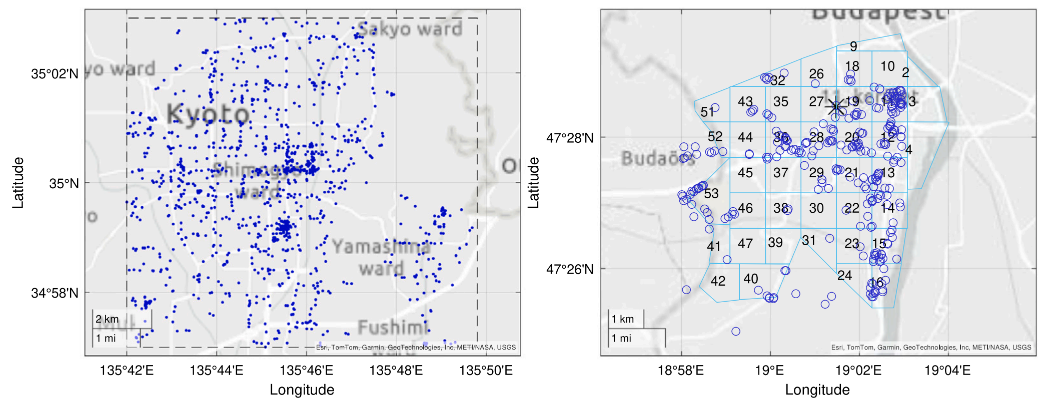
Fig. 9 shows an example for Marutamachi-Dori St. where the original estimator, using non-shifted data, had high errors (top figure). In this case, we can see blue error stems, indicating that the estimator did not manage to fit a model that reproduces the ground truth data correctly at these points. Again, including time-shifted POI data, the performance of the estimator increased. Expressed quantitatively, the RMSE and the MRE changed from 6.408 to 3.69 and from 10.2756 to 5.5503, respectively. The results of the Kyoto case study are also tabulated in Table 1, showing the RMSE and MRE values. As it is observable, the error values remain in a relatively low region, justifying the viability of the proposed solution. Moreover, the results also clearly reflect the superiority of the time-shifted solution.

Fig. 10 is a summary graph of how the number of observations used in the regression affects the estimation performance. Creating the regression with only one observation and predicting all the other results in 10.88% MRE. The right end of the graph indicates that if the regression is based on 21 observations, the estimation results in 4.77% MRE.



(a) Kyoto (b) Budapest - District XI.

Fig. 3. Case study locations.



(a) Kyoto (b) Budapest - District XI.

Fig. 4. POI locations.

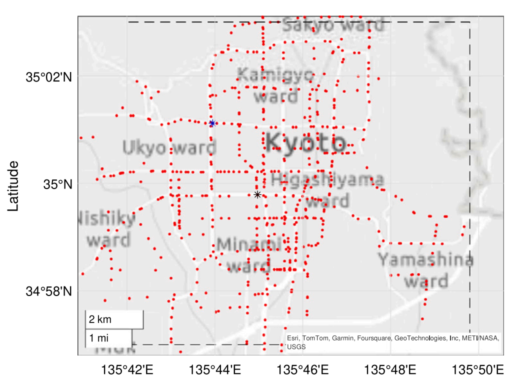


Fig. 5. Video sensor locations in the Kyoto are of interest. The two locations investigated in detail are marked with asterisks: Gojo-Dori St. (black), Marutamachi-Dori St. (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

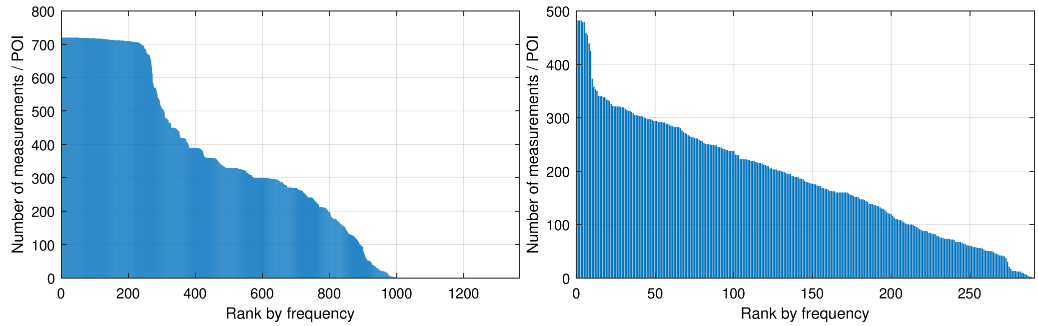
Table 1
RMSE and MRE values of the estimations, calculated for typical off-peak and peak periods, i.e., early and late afternoon hours in both cities. RMSE unit is km/h for the Budapest dataset and number of vehicles for the Kyoto dataset.

Case	RMSE	MRE (%)
Budapest 14 h	2.241	4.325
Budapest 17 h	0.713	3.704
Kyoto 17 h	7.789	3.704
Kyoto 17 h (with time-shift)	5.280	2.427
Kyoto 13 h	6.408	10.276
Kyoto 13 h (with time-shift)	3.690	5.550

4.5. Sensitivity analysis

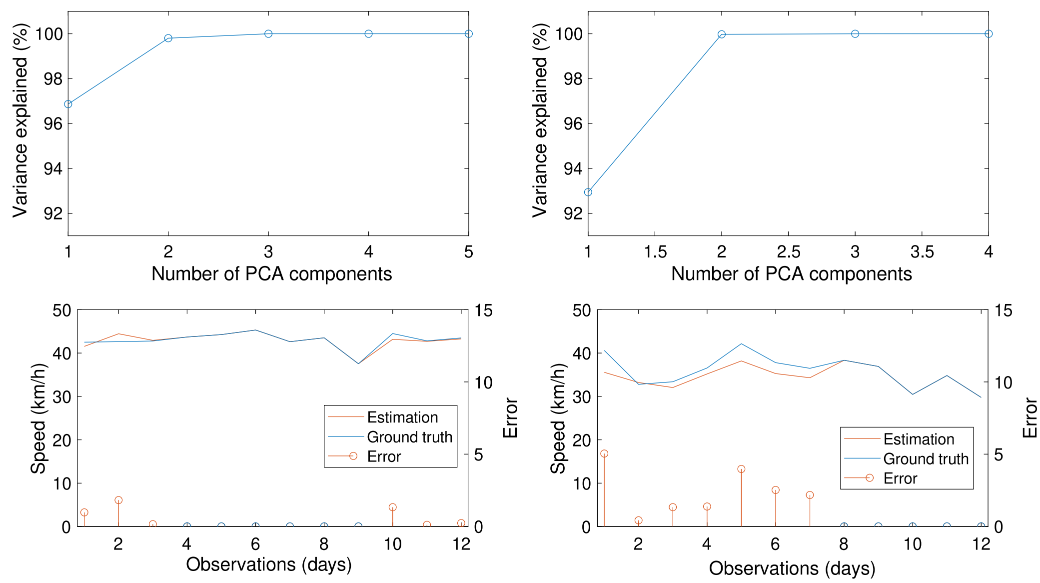
To evaluate the possibility of reducing the number of required POIs, we also examine how the POI number affects the estimation performance. First, we can compute the correlation coefficients between the vehicle count time series data and the POI data. POIs are then sorted according to the absolute correlation values. At this point, a truncation is performed, which discards POIs with less than 0.2 absolute correlation.

The evaluation involves two parameters to be changed, the number of POIs and the number of days (2, 5, 8, and 11 days) included in the regression, and considers the Marutamachi-Dori St, Kyoto. Having 221 POIs that have a correlation value of at least 0.2, Fig. 11 depicts how the RMSE values change along with the reduction of the POI number. The results show that the lowest error can be achieved in the case of the highest number of days included in the regression (11 days in the example). This, of course, seems natural since more data enables more effective estimations. The conclusion of the graphs is that two days are not enough, and the RMSE value is roughly constant regardless of how many POIs we include. With more days used, the RMSE actually scales with the POI number, and above 10, the estimation error settles in a reasonable range.



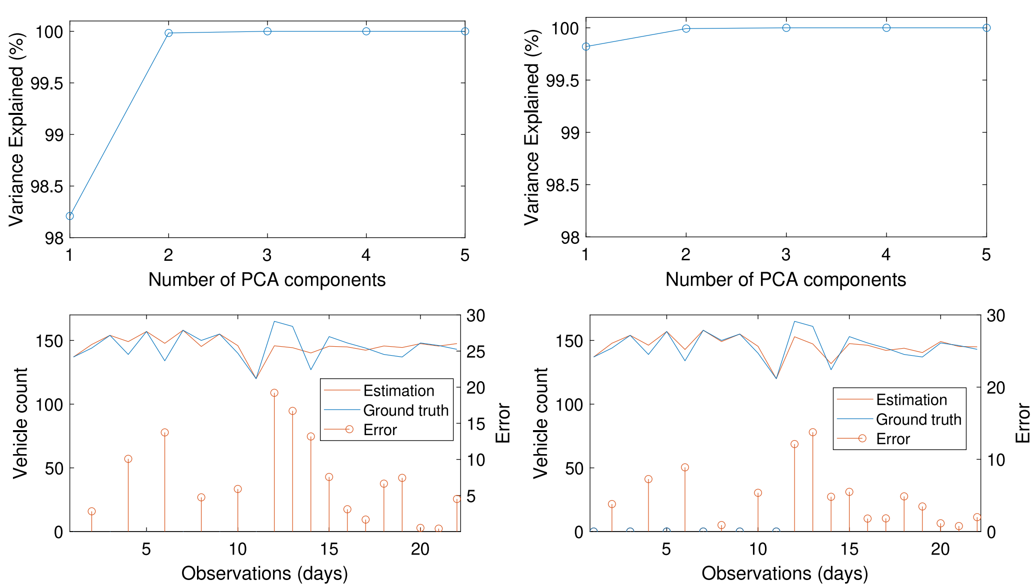
(a) Kyoto (b) Budapest - District XI.

Fig. 6. Distribution of collected data across the POIs sorted in descending order.



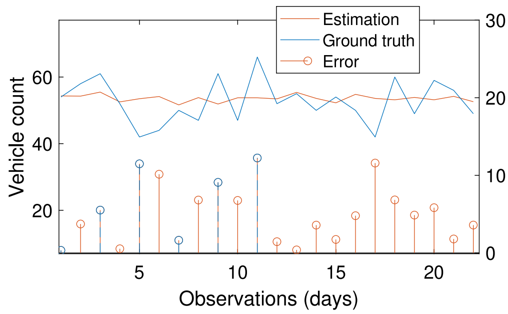
(a) 1-2 PM (b) 5-6 PM

Fig. 7. Results for traffic estimation on Nagyszőlős road, district XI. Budapest, between 08.11.2021 and 26.11.2021 for 1–2PM (a) and 5–7PM (b).

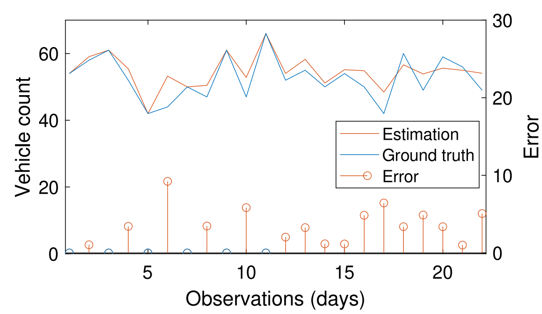


(a) 5-6 PM (b) 5-6 PM

Fig. 8. Results for traffic estimation on Gojo-Dori St. Kyoto, between 01.09.2021 and 30.09.2021 for 5–6 PM on weekdays without time-shift (a) and with time-shift (b).



(a)



(b)

Fig. 9. Comparison of traffic estimation on Marutamachi-Dori St. Kyoto, between 01.09.2021 and 30.09.2021 for 1–2 PM on weekdays; without time-shift (a) with time-shift (b).

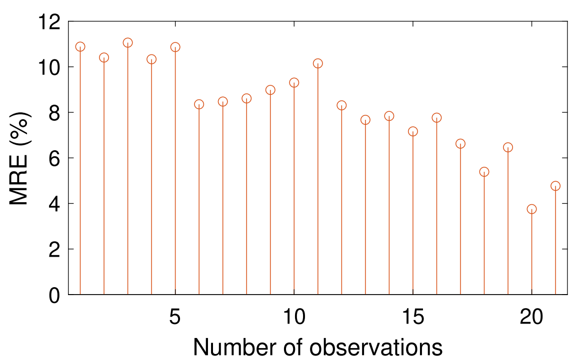


Fig. 10. Mean Relative Error (MRE) as a function of the number of observations used in the regression. The case study is for Marutamachi-Dori St. Kyoto, between 01.09.2021 and 30.09.2021 for 1–2 PM on weekdays, as in Fig. 9.

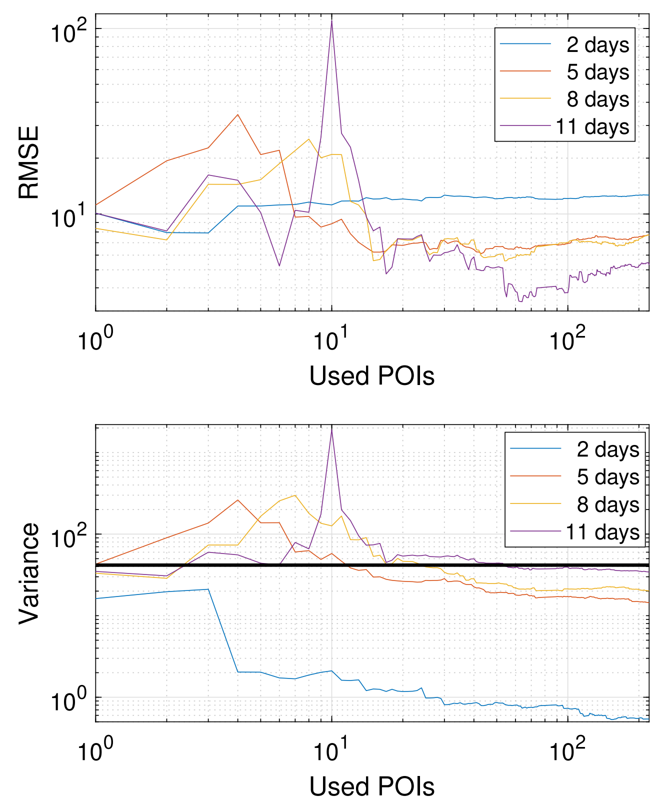


Fig. 11. Estimation RMSE (top) and variance (bottom) as a function of used POI data. Black straight line is the variance of the ground truth traffic data. Colored curves show different setups regarding how many days were used for the regression. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The lower graph in Fig. 11 shows how the variance of the estimations compares to the variance of the ground truth data. The variance value for the ground truth (thick horizontal line) is computed for the full 22-day time series. For the estimations, the variances are computed for the days the estimation is conducted. Using (2,5,8,11) days to create the regression, the estimation happens for (20,17,14,11) days, respectively. The interesting feature is that using only two days in the regression, the variance of the estimation is too low. This means that the output of the estimator is nearly constant, indicating that it cannot capture the variability of the data to be estimated. Using more days as input data, the variance of the estimated values starts to converge to the variance of the ground truth data. Again, it can be seen that more than 10 POIs are needed to obtain satisfactory results.

The results of a similar analysis with higher resolution are presented in Fig. 12. The left contour plot shows the logarithm of the RMSE values across various POI numbers and days used in the regression. From this plot, we can verify that at least 10 POIs are needed for a satisfactory performance. The right plot shows the logarithm of the relative variance of the estimation compared to the variance of the true data. This indicates that if we use data from at least 5 days and 20 POIs, the variance of the estimation will converge to the ground truth variance.

4.6. Large scale statistics

In the Kyoto region, we have performed a large-scale analysis involving all video sensors (1033) that had data for the selected time slots. The chosen POIs are the ones that have data throughout the 30 days, which means that the same POIs were used for every video sensor. The regression analysis was conducted as follows. From the 22 workdays, every third is used to create the regression, and two-thirds are used for evaluation.

Figs. 13(a)–13(d) show the histograms of the estimation RMSE and MRE values for all video sensor locations for 3 AM, 8 AM, 12 PM, and 5 PM. The shaded bars indicate where the mean values reside. For daytime, the histograms show similar characteristics. The typical RMSE and MRE values are around 10 and 20. However, for the night, the RMSE values are lower, and the MRE values are much higher. This is because at night, the traffic volume is much lower, resulting in higher relative and lower absolute errors.

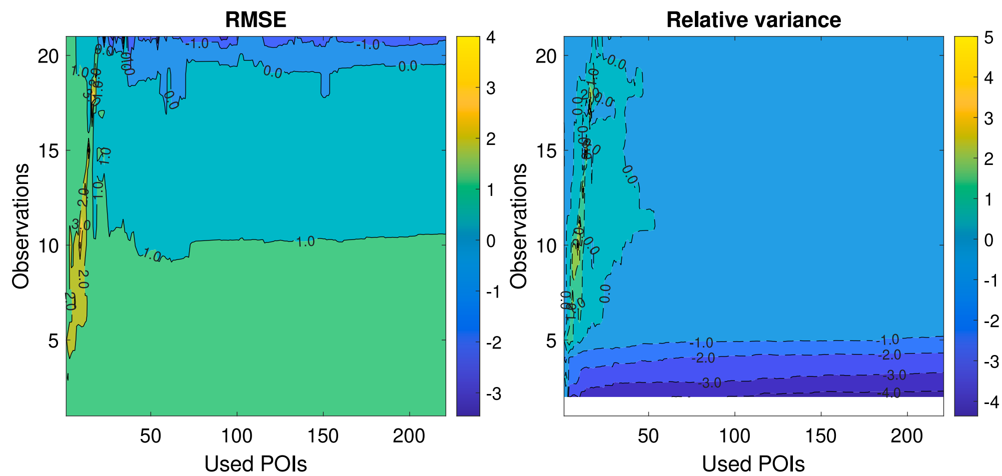


Fig. 12. Estimation RMSE (left) and relative variance (right) as a function of used POI data and observations. The values are in a logarithmic scale for better visibility.

4.7. Discussion

The research results demonstrate the usability of the proposed method. The crowdsourced social activity data is recognized as a new data source for traffic measurement or strategic urban planning, tackling the critical problem of incomplete network coverage due to the traffic sensor costs. Unlike navigation apps (such as Waze and Google Maps) that primarily provide readily available travel time data, our model specifically focuses on estimating the often missing traffic flow data. By combining a probabilistic model and time-shifted regression, the solution effectively predicts flow data, which is crucial for strategic urban planning and real-time traffic operations. The results, validated across diverse metropolitan areas like Budapest (speed) and Kyoto (flow), confirm that leveraging a fraction of social activity data leads to accurate and efficient traffic estimations with low relative errors. This hybrid approach, utilizing ubiquitous social data to augment sparse cross-sectional measurements, represents a significant, cost-effective advancement in network-wide traffic sensing.

4.8. Limitations

The results proved that using GPT data to estimate traffic-related quantities is feasible and promising. POIs offer several advantages for this purpose. They are not limited to commercial establishments but encompass a wide variety of locations that reflect human activity patterns, including schools, factories, bus stops, parking areas, petrol stations, government offices, and other public facilities. Essentially, any location represented in Google Maps can serve as a POI, provided it exhibits measurable temporal activity in Google Popular Times data.

The goal of our approach is not to rely on a specific type of POI (e.g., shopping malls or restaurants) but to identify and select those that show relevant spatiotemporal correlation with nearby traffic flow. This allows the method to be effective even outside commercial areas, such as residential or mixed-use zones, where certain categories of POIs (e.g., schools, local markets, or public transport stops) still exhibit activity patterns influencing local traffic.

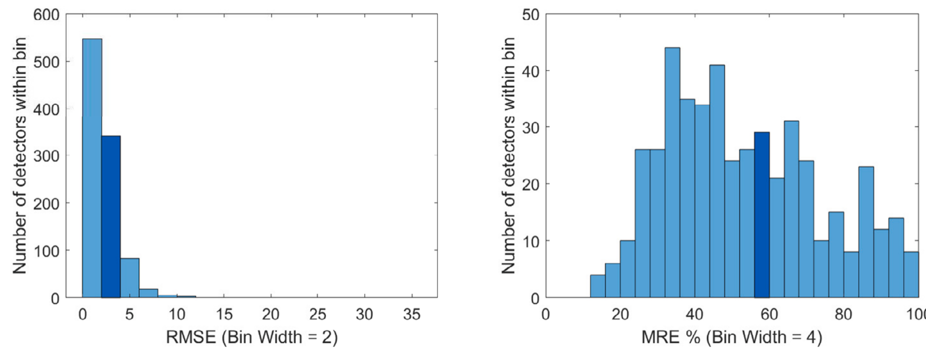
Nevertheless, certain limitations apply. First, the method by which GPT data are generated is proprietary and may be subject to future changes. Second, unique and irregular events — such as accidents, road closures, or special occasions — cannot be captured by the proposed approach, which is designed to model regular weekday traffic trends. Third, while POIs are widely distributed, they tend to be denser in commercial and central areas, potentially leading to lower spatial coverage in suburban or sparsely populated regions. This limitation can be mitigated by combining POI-based data with other complementary sources, which represents a potential direction for future research. Finally, the proposed method remains phenomenological, relying on statistical correlations rather than explicit physical modeling of traffic dynamics.

5. Conclusion

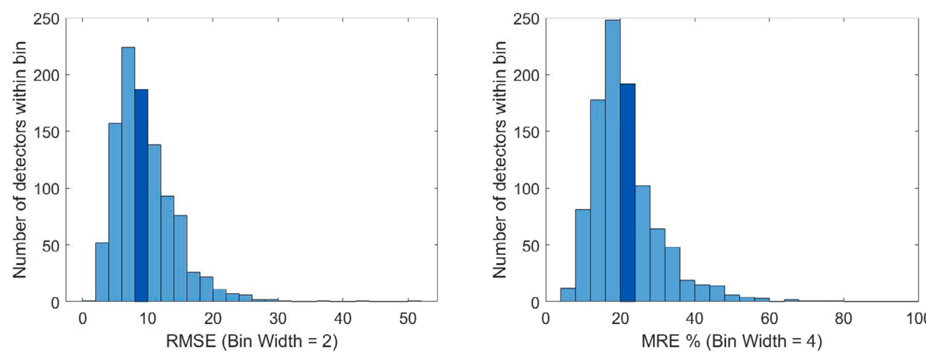
In this paper, we examined the spatiotemporal correlation of social activity and road traffic data. To be able to predict road traffic-related data, a general estimator is proposed. Our approach considered a linear relationship between Google Popular Times and specific descriptive traffic data. A probabilistic model was established that allows the construction of a regression of traffic data on GPT. To make the problem tractable, dimensionality was reduced using ideas adopted from the principal component analysis technique. The predictive power of the proposed method was validated on real-world data: the Budapest dataset containing GPT and FCD data, and the Kyoto dataset having GPT and vehicle count measurements. Analysis justified that by using only social activity information, traffic data can be predicted effectively. Moreover, by extending the GPT data with forward–backward temporal shifts, the performance increases. In all, our demonstration shows that a fraction of the available GPT data is enough to provide reliable traffic data estimations with a few percent relative errors.

The results achieved are promising in terms of policy recommendations. If the POI activity information is also considered together with the traditional traffic data, more credible traffic estimation and prediction can be achieved. Furthermore, in case of missing measurement data (e.g., typically due to technical failures of the sensors) social activity information can be effectively utilized to substitute conventional data.

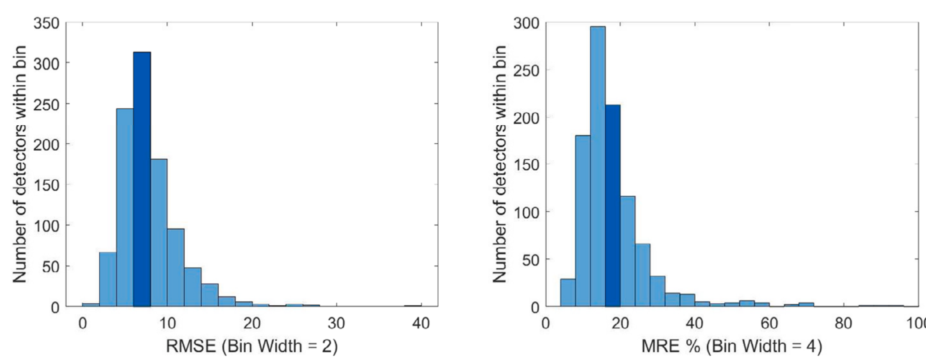
Future research can go in two directions. Still, regarding linear relations, more refined estimators of covariance matrices can be considered, which work robustly. Since there was data abundance in this work, the sample covariance matrix served well as an estimator. In a data-poor environment, perhaps with outlying data entries, the estimation needs more care as covariance matrices do not form a vector space but a Riemannian manifold (Smith, 2005). Another refined method would use copulas of spatiotemporal data to interpolate missing entries (Yang et al., 2025). If enough data is available, one might turn to a more detailed but still probabilistic phenomenological description of the traffic using the master equation (Mahnke et al., 2005). The master equation is typically written as a gain–loss equation that describes how the probability of a system being in a state evolves. In our context, it would describe how the number of vehicles changes on a link, and the POIs act as source and sink locations. Estimating the parameters or transition rates of a master equation is the task of fitting a high-dimensional stochastic and discrete spatiotemporal model to time series data.



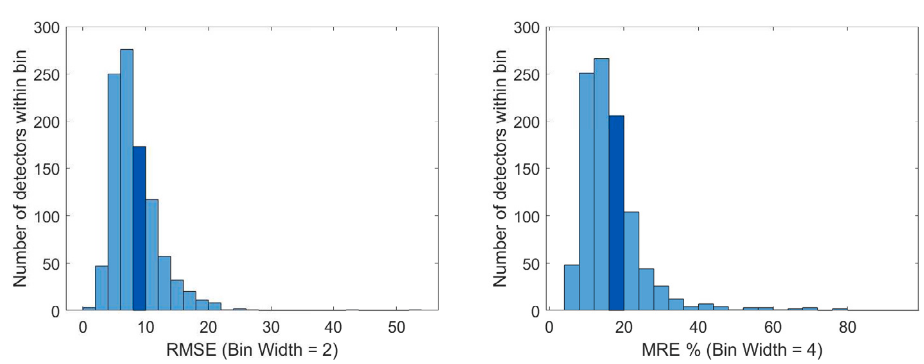
(a) 3 AM



(b) 8 AM



(c) 12 PM



(d) 5 PM

Fig. 13. Error metrics for all video sensor locations in Kyoto.

CRediT authorship contribution statement

Olivér Törő: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Domokos Esztergár-Kiss:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Jan-Dirk Schmöcker:** Writing – review & editing, Supervision, Project administration, Data curation, Conceptualization. **Tamás Tettamanti:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Data curation, Conceptualization.

Funding

The research was supported by the European Union within the framework of the National Laboratory for Autonomous Systems (RRF-2.3.1-21-2022-00002).

This project (2019-2.1.7-ERA-NET-2021-00019) has received funding from the NRDI Fund by the National Research, Development and Innovation Office Hungary and the ERA-NET COFUND /EJP COFUND Programme with co-funding from the European Union Horizon 2020 research and innovation programme.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors are unable or have chosen not to specify which data has been used.

References

Adetiloye, T., Awasthi, A., 2019. Multimodal big data fusion for traffic congestion prediction. In: Multimodal Analytics for Next-Generation Big Data Technologies and Applications. Springer, pp. 319–335. http://dx.doi.org/10.1007/978-3-319-97598-6_13.

Alessandretti, L., Aslak, U., Lehmann, S., 2020. The scales of human mobility. Nature 587, 402–407. http://dx.doi.org/10.1038/s41586-020-2909-1.

Alomari, E., Katib, I., Mehmood, R., 2020. Iktishaf: a big data road-traffic event detection tool using Twitter and spark machine learning. Mob. Netw. Appl. 1–16. http://dx.doi.org/10.1007/s11036-020-01635-y.

Asmael, N.M., Al-Taweel, H.M., Waheed, M., 2024. Exploring an interaction model for land used intensity-traffic congestion. Period. Polytech. Transp. Eng. 52, 270–275. http://dx.doi.org/10.3311/PPtr.23305.

Balsa-Barreiro, J., Menendez, M., Morales, A.J., 2022. Scale, context, and heterogeneity: The complexity of the social space. Sci. Rep. 12, 9037. http://dx.doi.org/10.1038/s41598-022-12871-5.

Cui, Y., Meng, C., He, Q., Gao, J., 2018. Forecasting current and next trip purpose with social media data and Google places. Transp. Res. Part C: Emerg. Technol. 97, 159–174. http://dx.doi.org/10.1016/j.trc.2018.10.017.

D'Andrea, E., Ducange, P., Lazzerini, B., Marcelloni, F., 2015. Real-time detection of traffic from Twitter stream analysis. IEEE Trans. Intell. Transp. Syst. 16, 2269–2283. http://dx.doi.org/10.1109/TITS.2015.2404431.

Fakhruzi, I., Titos, M., Benítez, C., García, L., 2025. Urban traffic monitoring through distributed acoustic sensing: Trial analysis of a potent monitoring tool. Measurement 117668. http://dx.doi.org/10.1016/j.measurement.2025.117668.

Farooq, D., Iqbal, H.W., Masoud, M., Haider, T., Abbas, H., Riaz, M.A., Asghar, M.A., 2026. Optimization of public transport route accessibility in high congestion areas using arc gis: A case study of Rawalpindi city. Period. Polytech. Transp. Eng. 54, 24–33.

Fekih, M., Bonnetain, L., Furno, A., Bonnel, P., Smoreda, Z., Galland, S., Bellemans, T., 2022. Potential of cellular signaling data for time-of-day estimation and spatial classification of travel demand: a large-scale comparative study with travel survey and land use data. Transp. Lett. 14, 787–805. http://dx.doi.org/10.1080/19427867.2021.1945854.

Gomes, B., Coelho, J., Aidos, H., 2023. A survey on traffic flow prediction and classification. Intell. Syst. Appl. 20, 200268. http://dx.doi.org/10.1016/j.iswa.2023.200268.

Gong, S., Cartlidge, J., Bai, R., Yue, Y., Li, Q., Qiu, G., 2020. Extracting activity patterns from taxi trajectory data: a two-layer framework using spatio-temporal clustering, Bayesian probability and Monte Carlo simulation. Int. J. Geogr. Inf. Sci. 34, 1210–1234. http://dx.doi.org/10.1080/13658816.2019.1641715.

Google, 2025. Popular times, wait times, and visit duration on Google business profile. URL https://support.google.com/business/community-guide/333628336/popular-times-wait-times-and-visit-duration-on-google-business-profile?hl=en.

Greenacre, M., Groenen, P.J., Hastie, T., d'Enza, A.I., Markos, A., Tuzhilina, E., 2022. Principal component analysis. Nat. Rev. Methods Prim. 2, 100. http://dx.doi.org/10.1038/s43586-022-00184-w.

Gunawan, F.E., Chandra, F.Y., 2014. Optimal averaging time for predicting traffic velocity using floating car data technique for advanced traveler information system. Procedia - Soc. Behav. Sci. 138, 566–575. http://dx.doi.org/10.1016/j.sbspro.2014.07.240, the 9th International Conference on Traffic and Transportation Studies (ICTTS 2014).

Hruboš, M., Nemec, D., Bubeníková, E., Holečko, P., Spalek, J., Mihálik, M., Bujňák, M., Andel, J., Tichý, T., 2021. Model-based predictive detector of a fire inside the road tunnel for intelligent vehicles. J. Adv. Transp. 2021, 6634944. http://dx.doi.org/10.1155/2021/6634944.

Jiang, X., Gao, J., Wang, T., Zheng, L., 2012. Supervised latent linear Gaussian process latent variable model for dimensionality reduction. IEEE Trans. Syst. Man Cybern. B 42, 1620–1632. http://dx.doi.org/10.1109/TSMCB.2012.2196995.

Jolliffe, I.T., 2002. Principal Component Analysis. Springer, http://dx.doi.org/10.1007/b98835.

Kan, Z., Kwan, M.P., Liu, D., Tang, L., Chen, Y., Fang, M., 2022. Assessing individual activity-related exposures to traffic congestion using GPS trajectory data. J. Transp. Geogr. 98, 103240. http://dx.doi.org/10.1016/j.jtrangeo.2021.103240.

Kolat, M., Tettamanti, T., Bécsi, T., Esztergár-Kiss, D., 2023. On the relationship between the activity at point of interests and road traffic. Commun. Transp. Res. 3, 100102. http://dx.doi.org/10.1016/j.commtr.2023.100102.

Lopez, P.A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., Wießner, E., 2018. Microscopic traffic simulation using SUMO. In: The 21st IEEE International Conference on Intelligent Transportation Systems. IEEE, pp. 2575–2582. http://dx.doi.org/10.1109/ITSC.2018.8569938.

Mahdi, A.J., Tettamanti, T., Esztergár-Kiss, D., 2023. Modeling the time spent at points of interest based on Google popular times. IEEE Access 11, 88946–88959. http://dx.doi.org/10.1109/ACCESS.2023.3305957.

Mahnke, R., Kaupužs, J., Lubashevsky, I., 2005. Probabilistic description of traffic flow. Phys. Rep. 408, 1–130. http://dx.doi.org/10.1016/j.physrep.2004.12.001.

Mariani, M., Styven, M.E., Ayeh, J.K., 2019. Using facebook for travel decision-making: an international study of antecedents. Int. J. Contemp. Hosp. Manag. 31, 1021–1044. http://dx.doi.org/10.1108/IJCHM-02-2018-0158.

Moro Zamprogno, M., Esztergár-Kiss, D., 2024. Applying cluster analysis for the investigation of travel behavior and user profiles. Period. Polytech. Transp. Eng. 52, 362–371. http://dx.doi.org/10.3311/PPtr.25663.

Moyano, A., Stepniak, M., Moya-Gómez, B., García-Palomares, J.C., 2021. Traffic congestion and economic context: changes of spatiotemporal patterns of traffic travel times during crisis and post-crisis periods. Transportation 48, 3301–3324. http://dx.doi.org/10.1007/s11116-021-10170-y.

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. http://dx.doi.org/10.1016/j.cities.2019.03.006.

Papageorgiou, M., 2004. Overview of road traffic control strategies. IFAC Proc. Vol. 37, 29–40. http://dx.doi.org/10.1016/S1474-6670(17)30657-2.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. 2, 559–572. http://dx.doi.org/10.1080/14786440109462720.

Pentland, A., 2015. Social Physics: How Social Networks Can Make Us Smarter. Penguin.

Pourebrahim, N., Sultana, S., Niakanlahiji, A., Thill, J.C., 2019. Trip distribution modeling with Twitter data. Comput. Environ. Urban Syst. 77, 101354. http://dx.doi.org/10.1016/j.compenvurbsys.2019.101354.

Shaygan, M., Meese, C., Li, W., Zhao, X.G., Nejad, M., 2022. Traffic prediction using artificial intelligence: Review of recent advances and emerging opportunities. Transp. Res. Part C: Emerg. Technol. 145, 103921. http://dx.doi.org/10.1016/j.trc.2022.103921.

Smith, S., 2005. Covariance, subspace, and intrinsic Cramér-Rao bounds. IEEE Trans. Signal Process. 53, 1610–1630. http://dx.doi.org/10.1109/TSP.2005.845428.

Tettamanti, T., Horváth, M., Varga, I., 2014. Road traffic measurement and related data fusion methodology for traffic estimation. Transp. Telecommun. 15, 269–279. http://dx.doi.org/10.2478/ttj-2014-0023.

Tettamanti, T., Tőkés, L., Varga, B., 2024. Fleet data based traffic modeling. Commun. Transp. Res. 4, 100138. http://dx.doi.org/10.1016/j.commtr.2024.100138.

Thonhofer, E., Palau, T., Kuhn, A., Jakubek, S., Kozek, M., 2018. Macroscopic traffic model for large scale urban traffic network design. Simul. Model. Pr. Theory 80, 32–49. http://dx.doi.org/10.1016/j.simpat.2017.09.007.

TOMTOM, 2025. Traffic data & traffic stats. URL https://www.tomtom.com/products/traffic-stats/.

van den Ende, M., Ferrari, A., Sladen, A., Richard, C., 2022. Deep deconvolution for traffic analysis with distributed acoustic sensing data. IEEE Trans. Intell. Transp. Syst. 24, 2947–2962. http://dx.doi.org/10.1109/TITS.2022.3223084.

Vongvanich, T., Sun, W., Schmöcker, J.D., 2023. Explaining and predicting station demand patterns using Google popular times data. Data Sci. Transp. 5, 10. http://dx.doi.org/10.1007/s42421-023-00072-z.

Yang, H., Zhao, X., Luan, S., Qi, J., 2025. A ST-copula method for highway traffic data interpolation in monitoring blind areas. Measurement 242, 116274. http://dx.doi.org/10.1016/j.measurement.2024.116274.

Zhang, Z., He, Q., Zhu, S., 2017. Potentials of using social media to infer the longitudinal travel behavior: A sequential model-based clustering method. Transp. Res. Part C: Emerg. Technol. 85, 396–414. http://dx.doi.org/10.1016/j.trc.2017.10.005.