IEEE Open Journal of
Intelligent Transportation
Systems

Received 11 January 2022; revised 3 June 2022 and 17 August 2022; accepted 21 September 2022. Date of publication 28 September 2022;
date of current version 11 October 2022.

Digital Object Identifier 10.1109/OJITS.2022.3210301

Link Travel Time Estimation for Arterial Networks Based on Sparse GPS Data and Considering Progressive Correlations

ZAHRA GHANDEHARIOUN (Student Member, IEEE), AND ANASTASIOS KOUVELAS (Senior Member, IEEE)

Institute for Transport Planning and Systems (IVT), Department of Civil, Environmental and Geomatic Engineering,
Eidgenössische Technische Hochschule Zürich (ETH), 8093 Zürich, Switzerland

CORRESPONDING AUTHOR: Z. GHANDEHARIOUN (e-mail: zahra.ghandeharioun@ivt.baug.ethz.ch)

ABSTRACT Understanding complicated city traffic patterns has been recognized as a critical goal by twenty-first-century urban planners and traffic management systems, resulting in a significant rise in the quantity and variety of traffic data gathered. For example, in a growing number of large cities, taxi firms have begun collecting metadata for each vehicle trip, such as origin, destination, and travel duration. Taxi data offer information on traffic patterns, allowing the study of urban flow—what will traffic look like between two sites on a particular day and time in the future? This paper proposes a method based on sparse GPS probe data, that focuses on allocating travel time data to the different links traveled between GPS observations. This model incorporates the progressive spatial correlations between the links in a network. The main goal of this work is to show how we can consider progressive spatial correlations and improve our results more realistically with a simple adjustment in the previously known parametric methods. For estimating arterial travel time, the methodology is applied to a case study for the partial network of New York City-based on the data collected from the taxicabs in New York City, providing the locations of origins, destinations, and travel times. The model estimates quarter-hourly averages of urban link travel times using OD trip data. This study proposes a more accurate approach for estimating link travel times, that fully utilizes the partial information received from taxi data in cities.

INDEX TERMS Correlation coefficient, iterative methods, iterative algorithms, maximum likelihood estimation, parametric statistics.

I. INTRODUCTION

FOR THE purpose of optimizing urban traffic operations and identifying major bottlenecks in the traffic network, accurate estimates, and forecasts of urban link travel times are critical. User advantage may also be gained by giving precise travel time information, allowing for improved path selection within the network, and reducing total trip traveling time. The use of real-time information from either in-road sensors such as loop detectors, microwave sensors, or road-side cameras, or mobile sensors (e.g., floating vehicles), or global positioning system (GPS) devices is required in order to estimate link travel times (e.g., cell phones) correctly. While there is little information available about the speed or the location of the connection in most of these instances, it is necessary to establish suitable methods for correctly estimating the performance measure of interest at the link, path, or network level.

There has been an increasing trend for GPS-equipped taxicabs in metropolitan regions in recent years. While GPS-equipped cabs have many benefits, they also act as valuable real-time probes for the traffic network. Taxis equipped with a GPS device collect a large quantity of data over days and months, offering a rich data supply for calculating network-wide performance indicators. Within this context, we present a technique based on sparse GPS probe data, and that is

The review of this article was arranged by Associate Editor Emmanouil Chaniotakis.

concerned with how to assign trip time data to the various links traversed between GPS observations to improve accuracy. The spatial correlations between the connections in a network are taken into account by this approach. Ultimately, the purpose of this study is to demonstrate that by modifying the previously established techniques, we may include spatial correlations in our calculations and enhance our findings more realistically.

The present paper is organized as follows: first, the problem at hand is described briefly. In Section II, we review related publications and briefly discuss several approaches to similar problems. The subsequent section introduces the detailed methodology of the current work. In Section IV proposed modification is explained in details. The result of applied methods on a case study is presented in Section V, with more details on the initial assumption and estimation results. The paper is then closed by conclusions and final remarks in Section VI.

A. PROBLEM STATEMENT

Urban travel time estimation based on GPS probe data has attracted many researchers recently [1], [2], [3], [4], [5]. The goal is to determine the urban link travel time based on the large amount of reported trip data for a network. Taxi trip data consist of the following information: exact coordinates of origin and destination with the trip distance and travel time. In most of the available data sets, the precise trajectory of the taxi trip is unknown, and different assumptions are made to discover the most probable path for a given origin and destination of trip data. In order to estimate the link travel time, the following problems should be solved:

1) Represent the network in a digital form.
2) Match the recorded geographic coordinates of the trip origin and destination on the produced digital network.
3) Discover the most probable path for the given trip.
4) Allocate the travel time to the links belonging to the discovered path.
5) Estimate the travel time of the link based on the observed travel times.

The first two steps are usually solved in similar ways by different researchers [4], [6]. For the third step, most researchers benefit from applying the k-shortest path algorithms to minimize the difference between the observed path and the assumed one [4], [6], [7], [8]. The methodologies used in the fourth and fifth steps can be classified into three categories:

a) Parametric approaches rely on statistical models and, based on mathematical assumptions, estimate the travel times. The majority of the parametric approaches assume that the link travel time is spatially and temporally independent of the rest of the network [7]. However, in reality, travel time on different road segments and at other times of day are spatially and temporally associated with one another [9]. Incorporating information on the spatio-temporal correlations of trip times may improve the estimation performance.
b) Non-parametric approaches are based on data-driven methods such as machine learning and neural networks. These approaches are free from assumptions and highly dependent on the amount of input data. The fusion of parametric and non-parametric approaches is classified as a third category called:
c) Hybrid approaches, which utilize a combination of both statistical models and data-driven methodologies. The details of the relevant works regarding this classification are presented in Section II.

Based on the classification mentioned above, in the current work, we focus on extending a parametric approach, introducing static and progressive spatial correlations between the links on the network, and modifying a statistically proven method to have more realistic travel time estimations.

B. CONTRIBUTIONS

In the current work, the main contribution is as follows:

- We propose a method based on sparse GPS probe data that focuses on allocating travel time data to the different links traveled between GPS observations. This model incorporates the spatial correlations between the links in a network.
- The main goal of this work is to show how we can consider progressive spatial correlations and improve our results more realistically with a simple adjustment in the previously known parametric methods.
- The methodology is applied to a case study for the partial network of New York City; based on the data collected from the taxicabs in New York City. By estimating link travel times with the proposed method, we show that travel time estimation accuracy is improved compared to the previously known parametric approaches.

II. RELATED WORKS

In this section, we investigate the related works focusing on two topics. First, we review the works contributing to different travel time estimation methods. Second, we explore the literature considering the travel time correlation between the links.

A. TRAVEL TIME ESTIMATION

Urban travel time estimation methods depend on the technologies deployed. The majority of the studies are based on data from technologies requiring extensive investment in sensor installation and maintenance, such as loop detectors in the following works: [10], [11], [12]; Automated Vehicle Identification (AVI) in [13], [14], [15]; video cameras in [16]. Therefore, travel time estimation becomes expensive depending on the network coverage and the accuracy of the sensors.

An alternative approach is to develop methods of estimation based on emerging large-scale data sources, such as GPS devices in either a dedicated fleet of vehicles, available from taxis, transit, commercial vehicles, and service vehicles or

even users' mobile phones. Reference [7] used GPS trace data from a fleet of around 500 taxis in San Francisco, USA, to estimate and predict traffic conditions. References [1], [8] and [17] utilize methods that are based on OD data, such as the New York City data set. Reference [4] proposed a statistical approach for path and travel time inference using GPS probe vehicle trajectory data. Furthermore, [18] states that reliable traffic estimation based on taxi data is provided when an adequate historical traffic database is available and the data covers long road segments sufficiently. Nevertheless, more complex approaches are needed to generate valuable output compared to the methods for traditional sensors stated in [19].

The methodologies based on GPS data introduced in different approaches can be categorized as follows:

Parametric approaches rely on mathematical and statistical equations. These approaches are limited by the assumptions made in the analytical and statistical models. However, they are proven mathematically correct and less computationally expensive [4]. Yeon *et al.* in [16] developed a model that can estimate travel time on a freeway using Discrete Time Markov Chains (DTMC), where the states correspond to whether or not the link is congested. Ramezani and Geroliminis in [20] also used a Markov chain approach to estimate arterial trip travel time distributions by capturing the spatial correlations using a Transition Probability Matrix (TPM) calibrated from historical data.

Most parametric estimations assume the spatially or temporally independent link travel time [4], [7], [16]. Bertsimas *et al.* in [1] introduce the general approach for travel time estimation based on OD data that can recover interpretable city traffic and routing information from potentially noisy and incomplete data. Zhan *et al.* in [17] combine the statistical model with MNL for path selection and minimize the least square error between the observed and expected path travel times.

Among the parametric approaches, only a few consider spatial correlation; the model presented in [5] separates trip travel times into link travel times and intersection delays and allows the correlation between travel times on different network links based on a spatial moving average (SMA) structure. Tang *et al.* in [21] develop a tensor-based Bayesian probabilistic model for citywide and personalized travel time estimation, using the large-scale and sparse GPS trajectories generated by taxicabs in Beijing. His model incorporates both the spatial and temporal correlation between different road segments and the person-specific variation between different drivers. Ma *et al.* in [22] propose a generalized Markov chain approach for estimating the probability distribution of trip travel times from link travel time distributions and take into consideration correlations in time and space.

Non-parametric approaches rely on data-driven methods such as machine learning and neural network [23]. These methods are free of the assumptions but highly dependent on the amount of input data and, therefore, computationally expensive. Reference [8] introduces a neighbor-based approach and considers a dynamic traffic condition using temporal speed references. Furthermore, [11] developed a method based on artificial neural networks to estimate the complete link travel time for an individual probe vehicle traversing the link, using the low frequency data collected by probe vehicles.

Hybrid approaches utilize a combination of data driven methods and statistical models. The fusion of parametric and non-parametric methods is generally more precise than the methods mentioned earlier. Reference [24] combines parametric and non-parametric traffic state prediction techniques through assimilation in an ensemble Kalman filter. For a non-parametric prediction, a neural network method is adopted; the parametric prediction is carried out with a cell transmission model with velocity as the state. In [25] similarly, benefit from a hybrid approach and develop a model on traffic flow through signalized intersections and combines it with a machine learning framework to both learn static parameters of the roadways as well as to estimate and predict travel times through the arterial network.

B. TRAVEL TIME CORRELATION

Correlation between travel times of links in a network or a path is empirically and theoretically discussed in many previous studies [2], [9], [26], [27], [28], [29], [30]. The problem of how to estimate the travel time correlation between links on a corridor was also introduced by Sen *et al.* in [9]. The theoretical analysis of this correlation is presented in [26] and [31]. Rilett and Park in [27] developed a one-step approach using artificial neural networks (ANN) to predict corridor travel times directly and consider inter-correlation between link travel times. The authors suggested that using a separate model to predict the travel time on each link without considering the covariance with other links can lead to significant errors. Zeng *et al.* in [32] extended the Lagrangian relaxation algorithm by representing travel time correlations based on the Cholesky decomposition. Chen *et al.* [33] further extended the multi-criteria A* algorithm to consider travel time correlations among adjacent K links. In addition, they show that adjacent link travel times are strongly correlated. For example, traffic accidents on a link may also lead to serious travel delays on its upstream links.

In his work [29], Gajewski and Rillet also estimate the link travel time correlation in the range of -1 to $+1$ by using a nonparametric regression technique based on Bayesian natural cubic splines. Rachtan *et al.* [34] developed three regression models to describe the correlation variation by considering various combinations of variables such as spatial distance, temporal distance, traffic state, and the number of lanes. They found that the primary factor in the correlation is spatial distance.

Based on the literature above and the logic presented as Tobler's first law of Geography, that 'all things are related, but nearby things are more related than distant things' (Tobler, 1970 [35]), we introduce the spatial correlation formulation to incorporate it with the previously proven historic

model of traffic introduced by Herring [36]. Furthermore, El Esawey and Sayed in [3], show that the correlation is usually very low for links that are spatially distant, even on the same street. Also, they show, for the determination of the correlation coefficient between the links, using the exponential model form outperformed the linear and power model forms under the chosen acceptance limits for the goodness of fit criteria.

III. METHODOLOGY – ESTIMATION WITHOUT SPATIAL CORRELATIONS

In this section, we present the methodology based on the steps introduced in Section I-A, and explain how we have approached each problem. It is worth mentioning that, the core of our methodology is built on the work presented in [36]. However, our approach addresses the gap of considering spatial correlations between network links and modifies the aforementioned work.

A. NETWORK MODEL

Basis of this work is a digital representation of a physical network. A directed graph $G(L, N)$ is generated utilizing Open Street Map, where links (L) and nodes (N) represent roads and intersections, respectively. For example, if a road is a two-way street, two links will be defined for that segment. The weight of the links in this graph is the length of the link in the real network.

B. MAP MATCHING AND PATH INFERENCE

In this work, we benefit from the origin destination of trips reported by a reliable source in [37], which has been used in many previous works [1], [17], [38]. This type of data is usually reported in GPS format. We know the exact geographical coordinates of the origin and destination of each trip. If the origin or destination location of a trip is in the middle of a link, it is projected to the nearest node/intersection. This step is a source of error at two levels. On the one hand, GPS data are unavoidably inaccurate, and on the other hand, it is neglected that trips generally do not start and end at intersections. However, the consequences of the latter are not significant, if the trips reported are sufficiently long.

Since we are not aware of the exact path that the taxi has taken in this type of data, we apply the k-shortest path algorithm based on Yen's algorithm explained in [39] to determine the inferred path as the one that minimizes the difference between the inferred and observed path distance. Since the k-shortest path is a computationally expensive task, defining k depends on the available resources for each study. After this step, the observations that violate the following inequality are removed.

$$0.5 \times \text{observed distance} < k \text{ shortest path distance} < 1.5 \times \text{observed distance} \quad (1)$$

After this step, the data are in the form of path observations. The set of all available path observations for time interval t, is denoted as P_t and a single path as p.

C. TRAVEL TIME ESTIMATION MODEL

The proposed travel time estimation methodology is built on [36] methodology, and requires path observations as input data. This work is based on the following assumptions:

- The travel time distribution for each network link is independent of all other network links. Therefore, the set of all network links, that we have observations for is denoted as L.
- Any given moment in time belongs to exactly one historical time period, during which, traffic conditions are assumed to be constant.
- All travel time observations from a specific link l are independent and identically distributed within a given time period t.
- Sparse probe measurements are the only data available to the model.

Admittedly, the first and second assumptions are very strong and proven incorrect. Spatial correlations exist at both the local and non-local levels. Temporal dependencies exist in a short-term neighboring and long-term periodic timescale [11]. While that might hold true, capturing these spatial-temporal dependencies is challenging, independent of whether you try to estimate them or incorporate literature values into the model, given that they even exist. In this approach, we explain the solution with independent variables and try to consider the dependencies of the link and improve the Herring [36] approach to a more realistic one.

1) PROBABILISTIC SETTING

The random variable capturing the link travel time for link l in time period t is denoted as $X_{l,t}$, where l can be any element of L. The set of links lying on path p is denoted as L_p, so let $Y_{p,t}$ be the random variable representing the path travel time for path p in time period t. Then, the path travel time $Y_{p,t}$ can be represented as follows

$$Y_{p,t} = \sum_{l \in L_p} X_{l,t}. \quad (2)$$

It is assumed that all link travel times in the network follow some probabilistic distribution. This generally can be any probability distribution function for any link l. In the current work, we assume that all link travel times follow Gaussian distributions, and we define $\mu_{l,t}$ as mean value and $\sigma^2_{l,t}$ as variance, thus: $X_{l,t} \sim N(\mu_{l,t}, \sigma^2_{l,t}),\ \forall l \in L_p$.

The parameters describing the distribution for link l and time period t are denoted as $Q_{l,t}$. The link travel time probability density function for link l during time period t is denoted as $G_{Q_{l,t}}(X_{l,t})$. Path travel time probability density function is denoted as $G_{Q_{L_{p,t}}}(Y_{p,t})$, where the indices $Q_{L_{p,t}}$ denote the parameters of the links along the path p in time period t. The probability distribution of the sum of two or more independent random variables is the convolution of their individual distributions. Therefore, $G_{Q_{L_{p,t}}}(Y_{p,t})$ is the convolution of the link travel time distributions along the path p. In this case, all link travel times are

assumed to be independent from one another and to follow Gaussian distributions. Hence, for a path observation, it holds, $Y_{p,t} \sim N(\Sigma_{l \in L_p} \mu_{l,t}, \Sigma_{l \in L_p} \sigma^2_{l,t})$.

The goal is to find the parameter values $Q_{l,t}$ for each link and time period, which make the observed data most probable. This is achieved by maximizing the likelihood function, which can be written in a general case as follows:

$$\underset{Q_t}{\operatorname{argmax}} \prod_{p \in P_t} G_{Q_{L_{p,t}}}(Y_{p,t}). \tag{3}$$

To transfer the product into a sum, the logarithm of the function is calculated. The maximum still occurs at the same parameter values since the logarithm is a monotonic function.

$$\underset{Q_t}{\operatorname{argmax}} \sum_{p \in P_t} \ln\Big(G_{Q_{L_{p,t}}}(Y_{p,t})\Big). \tag{4}$$

Given the assumption that all link travel times follow Gaussian distributions, problem (4) can be reformulated with optimization problem (5).

$$\underset{Q_t}{\operatorname{argmax}} \sum \ln\left(f\left(\sum_{l \in L_p} \mu_{l,t}, \sum_{l \in L_p} \sigma^2_{l,t}\right)\right), \quad p \in P_t, \tag{5}$$

where $f(\sum_{l \in L_p} \mu_{l,t}, \sum_{l \in L_p} \sigma^2_{l,t})$ denotes the Gaussian probability density function as a function of $\mu_{l,t}$ and $\sigma^2_{l,t}$ for a given $Y_{p,t}$.

This optimization problem is challenging on two levels. On the one hand, it simultaneously solves for the mean and variance. On the other hand, the number of variables is large, particularly in a network-wide study. The number of variables can be calculated as the number of links multiplied by the number of parameters per link.

Herring in [7] explained that the methodology can be extended to cases beyond the Gaussian distribution but leads to more complex optimization problems because it simultaneously solves for the mean and variance of every link in the network. It is possible to solve this problem directly if using a commercial-grade non-linear optimization engine with a lot of computational power. However, it is assumed that such resources may not be available, and an alternative solution strategy is proposed. The Gaussian case is presented here to show an example of the algorithm from start to finish in complete detail.

Since we extend the Herring methodology to a correlated version, we present the work by considering the Gaussian distribution. In general, the choice of a Gaussian distribution restricts the model's flexibility to capture unique traffic characteristics, but it is also far more tractable to solve in practice [36]. When using this model with certain classes of link travel time distributions, the travel time allocation problem is efficient, even for large amounts of data (such distributions include the standard distributions like Gaussian, Log-Normal, Gamma [40], and others). The parameter estimation problem is also efficient for the same set of distributions listed above [40].

Furthermore, recent empirical studies based on field observations show that the use of normal distributions appears to reflect observed path travel time distributions [41]. In addition, Chen *et al.* [33] found that the normal distribution can reasonably approximate the path travel time distribution. The normal distribution approximation can achieve 98.3% and 94.9% accuracy at the 10th and 90th percentiles. Also, Zeng *et al.*, in their work [32], used the empirical link and path travel time data from probe vehicles to characterize travel time distributions at the link and path level. Several typical distributions are tested, such as normal, log-normal, truncated normal, and truncated lognormal. Further, he explains the observed data distribution is approximated by a normal distribution, which is more computationally tractable and has an acceptable compromise on accuracy.

Herring in [36], suggests an intuitive decomposition scheme reaches near-optimal solutions efficiently. Also, note that for each time interval t, the problem can be solved separately, given the assumption, each time interval is independent.

2) DECOMPOSITION SCHEME

The core concept is to decouple the optimization problem into two more manageable sub-problems and iterate between these two until converging to an optimal solution. These two sub-problems are travel time allocation and parameter optimization. Herring's explanation [36] of why his decomposition scheme makes sense, though it cannot be derived mathematically, goes as follows. It would be straightforward to estimate the link parameters if it was known how much time each probe vehicle spent on each link on its path. However, in the case of sparsely sampled and OD data, this information is not available. Instead, one could try to determine the most likely link travel times, which depend on the link travel time parameters that in turn need to be estimated with the most likely link travel times. This is a chicken-and-egg type of problem. It is solved by assuming some initial link parameters, which are then used to determine the most likely link travel times. Following, the most likely link travel times are used to update the link parameters, which then are utilized to determine the most likely travel times again. This iterative process is repeated until convergence is reached. By reaching the convergence, the algorithm's output is $X_{l,t}$ variable that contains all the individual travel times allocated in an optimal manner to the links $l \in L$ for time period t. This $X_{l,t}$ can be used to compute our final set of parameters Q_t.

3) TRAVEL TIME ALLOCATION

The travel time allocation determines the most likely link travel times corresponding to a path p. To solve this problem, estimates of the link parameters must be available for all links in time period t, $l \in L_t$. This means that all link parameters are fixed for this part of the algorithm. Furthermore, it is essential to define lower bounds for the link travel times; otherwise, the most likely travel time is smaller than the free-flow travel time, or in extreme cases, even negative.

The free-flow travel time is denoted as b_l and is the time needed to travel link l with the maximum allowed speed. It is calculated by dividing the link length by the maximum allowed speed. For example, despite the existence of some highways and areas with narrower streets, the speed limit in Manhattan is 25mph [37]. It is suggested by [7] to assume that the taxi drivers will travel at 40 to 50 mph to compute the minimum link travel time. However, in the case study presented in Section V, we use 25mph as the free flow speed to calculate the free flow travel time. This constraint implies that path observations with an average speed greater than 25mph do not have a solution, and thus are removed from the path set.

The goal of finding the most likely travel times is also achieved by formulating a maximum likelihood function and finding its maximum. Still assuming that, all link travel time distributions are Gaussian, the problem can be formulated as in problem (6), where $f\ (X_{l,t}|\mu_{l,t},\ \sigma^2_{l,t})$ denotes Gaussian probability density function for a given mean μ and a given variance σ^2 as a function of the link travel time $X_{l,t}$

$$\underset{X}{\operatorname{argmax}} \prod_{l\in L_p} f\left(X_{l,t}|\mu_{l,t},\ \sigma^2_{l,t}\right). \tag{6}$$

Again, to convert the product to a sum, the logarithm of the function is computed. Moreover, two constraints are added. The sum of the link travel times lying on a path must be equal to the observed path travel time $Y_{p,t}$, and the link travel times $X_{l,t}$ must be larger than the free-flow travel time.

$$\begin{aligned}\underset{X}{\operatorname{argmax}}\ & \sum_{l\in L_p} \ln\left(f\left(X_{l,t}\mid \mu_{l,t},\ \sigma^2_{l,t}\right)\right)\\ &\text{s.t.}\\ &\sum_{l\in L_p} X_{l,t} = Y_{p,t}\\ &X_{l,t} \geq b_l,\ \ \forall l \in L_p.\end{aligned} \tag{7}$$

This problem needs to be solved for every observation $p \in P_t$, and this is done by the following method. First, the total expected path variance V and the difference between expected and observed path travel time Z need to be calculated

$$V = \sum_{l\in L_p} \sigma^2_{l,t}, \tag{8}$$

$$Z = Y_{p,t} - \sum_{l\in L_p} X_{l,t}. \tag{9}$$

As the next step, the expected travel time, adjusted by some proportion of Z, is allocated to each link. This proportion is computed by dividing the link variance by the total path variance

$$X_{l,t} = \mu_{l,t} + \frac{\sigma^2_{l,t}}{V} Z. \tag{10}$$

Links with high variance are the most likely source of discrepancies between observed and expected path travel time. The links with high variance get attributed to the largest part of Z. After this attribution, some links may violate the free flow constraint. These links are saved in the set J. After identifying the violating links and saving them in the set J, we calculate V and Z again. At this step, all the identified violating links saved in J ($l \in J$) have an expected travel time equal to the free-flow travel time, and these links do not contribute to the calculation of the total path variance V

$$V = \sum_{l\in L_p/J} \sigma^2_{l,t}, \tag{11}$$

$$Z = Y_{p,t} - \sum_{l\in L_p/J} X_{l,t} - \sum_{l\in J} b_l. \tag{12}$$

Then, the updated difference between the expected and observed travel time Z is attributed again with Equation (10). After this step, some links may still violate the constraint. Thus, J is updated, V and Z are recalculated, and Z is attributed to the links again. This procedure is repeated until the free-flow travel time constraint is met. On average, 1 to 5 iterations were necessary to meet the constraint in the use case at hand. Having solved the travel time allocation for all path observations $p \in P_t$, the output of the algorithm $X_{l,t}$ contains all the individual travel times allocated to the links $l \in L$ and time period t.

IV. INTRODUCING SPATIAL CORRELATIONS

Considering the aforementioned theoretical backgrounds in Section II and the criteria of spatial correlation they all show in their works, we introduce our heuristic for both progressive and static correlations as follows:

The Travel time allocation method presented in III can be extended for correlated links if we assume that the travel time on these links is jointly normally distributed. Based on the multivariate central limit theorem [42], the summation of all links' travel times is still normally distributed; therefore, this does not affect the maximum likelihood function formulation in the historic traffic model.

For each link in the set of L_p, we define the correlation between link $l_i \in L_p$ and $l_j \in L_p$ in path p by ρ^p_{ij} the Equations (8) and (10) will be updated as follows:

$$V = \sum_{l_i\in L_p} \sigma^2_{l_i,t} + 2 \sum_{l_i,l_j\in L_p, i\neq j} \sigma_{l_i,t}\sigma_{l_j,t}\rho^p_{ij,t} \tag{13}$$

$$X_{l_i,t} = \mu_{l_i,t} + \frac{\sigma^2_{l_i,t} + \sum_{l_i,l_j\in L_p, i\neq j} \sigma_{l_i,t}\sigma_{l_j,t}\rho^p_{ij,t}}{V} Z. \tag{14}$$

The correlation between the links can be considered both static and progressive. In the static version, we allocate the travel time in each iteration based on the same correlation coefficient defined at the beginning. In the progressive version, we update the correlation coefficient in each iteration based on the changes in parameters (in here, the mean value) in the last two iterations.

It is worth mentioning that the correlation coefficient here focuses on spatial correlation, and the temporal correlation is neglected in this study, and we assume that the travel time estimation is independent between different time periods.

In the current work, the main contribution is to show the effect of considering spatial correlations to understand the model's performance regardless of considering temporal correlations. Also, since for every 15-minute time interval, we have an extensive amount of taxi trip data, it can provide us with enough input for that time interval reflecting the conditions propagated from the previous time interval (e.g., spillback). However, one can include the temporal correlation by incorporating the parameters about the travel time of each link from the previous interval to the next interval. If we include both correlations simultaneously, it is hard to understand the effects separately.

In the following, we explain each version in more detail:

- *Static Correlation*
  Defining a realistic spatial correlation matrix is a challenging task, and it is highly dependent on network characteristics [9]. A basic rational approach for spatial correlation coefficient can follow the logic of the further you get from a link; the correlation coefficient will decrease accordingly [35]. Following this logic and the aforementioned background, the mathematical formulation of the spatial correlation should meet the following criterion: a) The correlation function should be descending by increasing the spatial distance b) The correlation coefficients should be near zero for very distant links. In our approach the static spatial correlation is calculated as follows: In a path with k links, the path p is a set of links: $L_p = \{l_1, l_2, l_3, \ldots, l_k\}$, $\rho^p_{ij,t}$ is the correlation coefficient between link l_i and l_j in time interval t in path p, where $i, j \in \{1, k\}$ and $|i-j|$ is the rank order distance of l_i to l_j in the set of L_p

$$\rho^p_{ij,t} = \frac{1}{\alpha \cdot |i-j| + 1}, \forall l_i, l_j \in L_p, \quad \text{where } 0.1 \leq \alpha \leq 0.9. \tag{15}$$

  The α value defines how quickly the correlation between the links in a path can decrease by increasing the distance. The higher α value corresponds to the quicker reduction in correlation coefficient between the links in the path by increasing distance. The correlation coefficient is calculated only on the basis of the paths, as the path observations are the only input in the proposed model. If two paths have mutual links, the spatial correlation is calculated for each path separately, and the correlation coefficient for the mutual link is calculated in each path towards the other links in the path.
  Remark 1: We note that the function in Equation (15) is only a candidate function and does not necessarily provide the best result among all the possible functions. One can find a near optimal correlation function through hybrid approaches [24]. However, the main focus of our work is to show how we can consider static spatial correlations and improve our results more realistically with a simple adjustment in the previously known parametric methods.

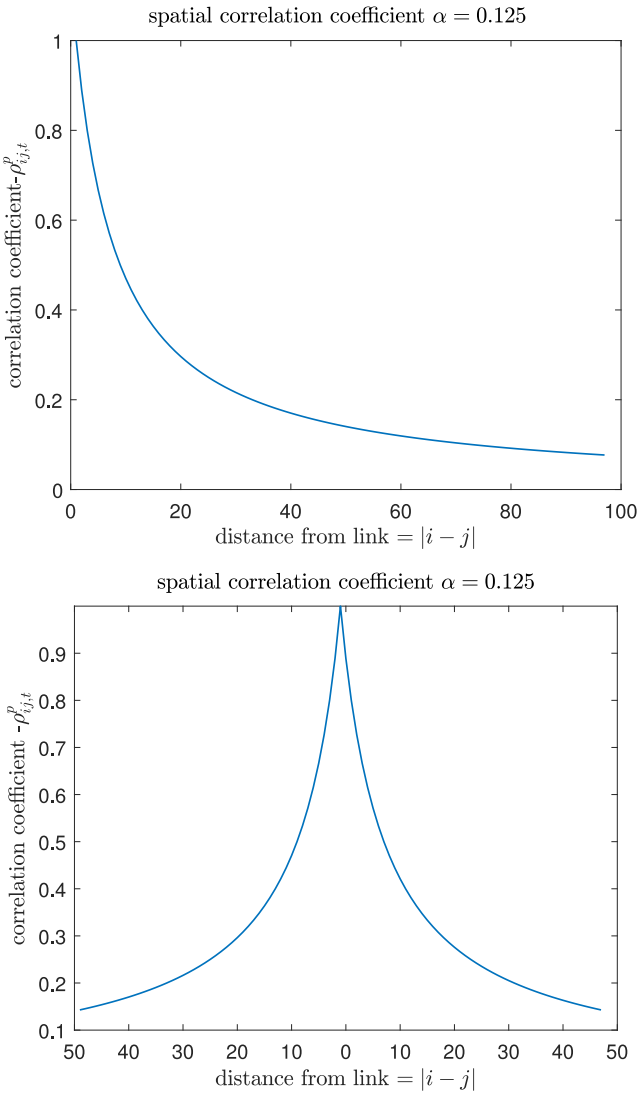

FIGURE 1. Static correlation coefficient example of a path with 98 links with $\alpha = 0.125$ (top: from the first link, bottom: from the middle link).

For example, the static correlation coefficient for the first and the middle link in a path is depicted in Figure 1. The profile definition in both diagrams in Figure 1 follows the same Function (15); the only difference is the starting link. We calculate the correlation coefficient between the first link and all other links in the path in the top figure. The bottom figure shows the correlation coefficient between the link in the middle of the path and all other links in the path, the links before the middle link and after the middle link. In the static version, the value of $\rho^p_{ij,t}$ remains the same through iterations for the calculation of Equation (13) and Equation (14).

- *Progressive Correlation*
  In the progressive version, we start by defining the correlation of the links in a path similar to Equation (15) in the first iteration ($n = 1$) and increase or decrease it based on the changes in the $\mu_{l_{i,t}}$, and $\mu_{l_{j,t}}$ in

previous iterations. Iteration number is defined by n. Suppose $\Delta\mu_{l_{i,t,n}}$ and $\Delta\mu_{l_{j,t,n}}$ both are positive or negative ($\lambda_n > 0$), meaning that both link trends are following the same direction. Then, we increase the correlation coefficient $\rho^p_{ij,t}$ in the next iteration. If one is positive and the other negative ($\lambda_n < 0$), we decrease the correlation coefficient. We assume that the trend in the changes in the mean travel time of a link through iterations can reflect the correlation between the two links. This can be seen in the travel time distribution of the links and thus in the mean travel time changes in the iterative approach. The amount that the correlation coefficient is increased or decreased in iteration n follows the function introduced in (16). The mathematical formulation of progressive approach needs to meet the following criterion: a) the function should gradually increase to an upper bound or gradually decrease to a lower bound, b) The changing increment should be adjustable by defining a parameter. For example, in Equation (16), we gradually increase the correlation coefficient up to the upper bound of $+0.8$, and similarly, we decrease it down to -0.8, that is the lower bound [29].

$$\rho^p_{ij,t,n} = \rho^p_{ij,t,n-1} + C_{ij,t,n} \; \forall l_i, l_j \in L_p$$

$$\lambda_n = \frac{\Delta\mu_{l_{i,t,n}}}{\Delta\mu_{l_{i,t,n}}} = \frac{\mu_{l_{i,t,n}} - \mu_{l_{i,t,n-1}}}{\mu_{l_{j,t,n}} - \mu_{l_{i,t,n-1}}}$$

$$C_{ij,t,n} = \begin{cases} -a^\beta + a, a = |0.8 - \rho^p_{ij,t,n-1}|, & \text{if } \lambda_n > 0 \\ b^\beta - b, b = |-0.8 - \rho^p_{ij,t,n-1}|, & \text{if } \lambda_n < 0 \\ 0, & \text{if } \rho^p_{ij,t,n-1} > 0.8 \\ & \text{or } \rho^p_{ij,t,n-1} < -0.8 \end{cases}$$

$$\text{where } 0.01 \le \beta \le 0.09. \tag{16}$$

The β value corresponds to the increment that we increase or decrease the correlation coefficient between two links. The higher the β value, the faster we reach the upper/lower bounds. As an example, the progressive correlations for a link at the beginning of the path and a link in the middle of the path are depicted in Figure 2. In this figure, we present the changes in the correlation coefficient through iterations. Each line in Figure 2 is the correlation coefficient of the chosen link i to all the other links j in the path. For instance, in Figure 2 on top, we have the correlation coefficient of the first link ($i = 1$) of a path with 98 links to all the other $j = 1{:}98$ links. The X axis is $|i-j|$ and the Y axis is the correlation coefficient ρ^p_{ij} for each iteration. Here we presented only 20 iterations, each with a distinct color and line pattern, with the number of iterations and the line pattern in the graph's legend. The graph at the bottom presents the correlation coefficient ρ^p_{ij} of the link in the middle $i = 50$ to all other links $j = 1{:}98$ in the path p. As we see, the first iteration starts with the same values calculated for the static version and changes through iterations based on Equation (16). Negative correlations between the links can occur, for instance, due to having traffic signals in the path. If one link is highly congested due to a red signal, having a longer travel time, the others are empty and have free flow travel time. A negative correlation in our study can explain this situation. It means that an increase in travel time in the link i can strongly reduce the travel time in link j. We note that (16) may not provide us with the best mathematical formulation for the optimal performance indicator using in the progressive approach. However, we show that the results improve by taking into account the changes in distribution function parameters through iterations for defining the correlation coefficient (see Table 3).

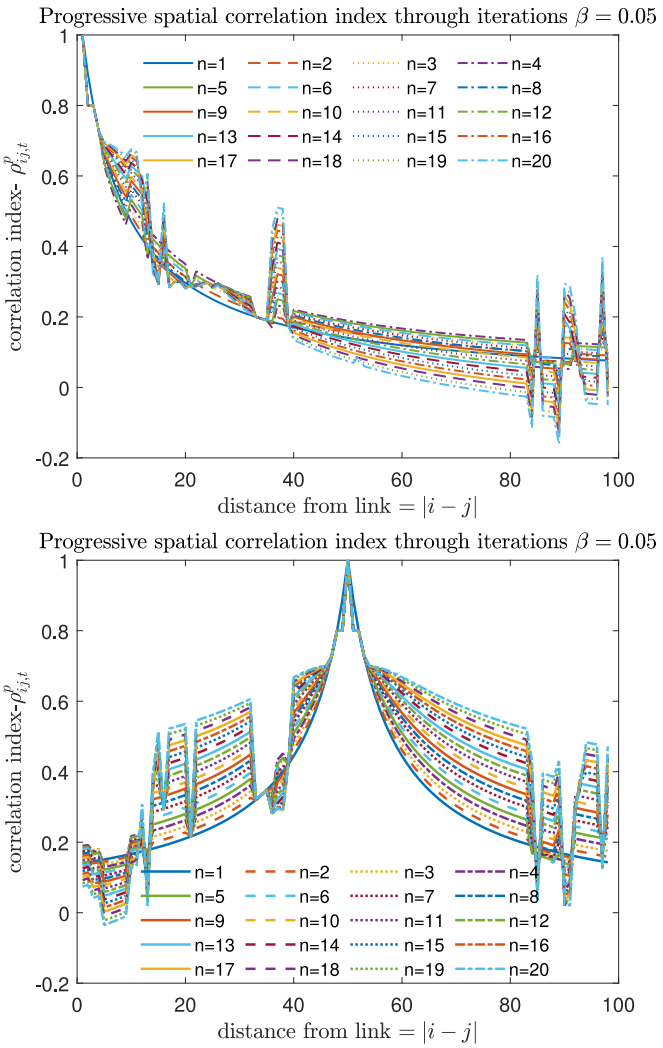

FIGURE 2. Progressive correlation coefficient example of a path with 98 links through 20 iterations $\beta = 0.05$ (top: from the first link, bottom:from the middle link, number on each line shows the iteration number).

1) PARAMETER OPTIMIZATION

Receiving $X_{l,t}$ from the travel time allocation step, optimizing the parameters is straightforward. Mean and variance are updated based on Equations (17) and (18), respectively. Note that $X_{l,t}(m)$ denotes the m_{th} observation of $X_{l,t}$. Reliable

estimates are not possible for links with less than ten observations available. Thus, the parameters are not updated, but the initial ones are kept.

$$\mu_{l,t} = \frac{1}{|X_{l,t}|} \sum_{m=1}^{|X_{l,t}|} X_{l,t}(m), \tag{17}$$

$$\sigma_{l,t}^2 = \frac{1}{|X_{l,t}|} \sum_{m=1}^{|X_{l,t}|} \left(X_{l,t}(m) - \mu_{l,t}\right)^2. \tag{18}$$

To solve the chicken-and-egg problem entirely, initial parameters for all links $l \in L$ are still required. Herring [36] suggests that these should be chosen according to literature values, which are in keeping with the link characteristics (number of lanes, traffic lights, etc.). For this work, the initial parameters are based on assuming that all cabs had a constant velocity along their path. This allows allocating the travel times based on the length of the links (see Equation (19) below). D_l denotes the length of link l and D_{L_p} the sum of all link lengths lying on path p.

$$X_{l,t} = \frac{D_l}{D_{L_p}} Y_{p,t}. \tag{19}$$

The output of this initial travel time allocation is of the same type as $X_{l,t}$. The initial parameters are therefore calculated with Equations (17) and (18), having $X_{l,t}$ based on the constant velocity assumption as the input argument.

2) CONVERGENCE

With each iteration (going back and forth between travel time allocation and parameter optimization), the parameter values should become smaller until the parameter values no longer change significantly. This is called convergence.. The parameters are the near-optimal solution Q_t for optimization problem (4) by reaching convergence. Herring [36] suggests that a global parameter $n_{\max}$ can define the criterion for convergence that stipulates the number of maximum iterations. In this work $n_{\max} = 100$ is set, which led to a reasonable convergence. In Table 2, the mean relative differences for the mean travel time values for all links in all time intervals in different models for the case study is presented in Section V.

Alternatively, after each iteration, one could compute the absolute difference between the individual link parameters of the previous and the current iteration. These differences are then divided by the parameter values of the previous iteration, revealing the relative differences as well. We denote this difference as Δ_Q. The convergence criterion itself is defined as a maximum allowed relative difference of the parameters between two iterations that we call $\Delta_{Q,\max}$. For instance, an appropriate value for $\Delta_{Q,\max}$ is 0.01, meaning that convergence is reached as soon as none of the parameters change by more than one percent between two subsequent iterations. The downside of this type of convergence criterion is that a single iteration needs more computing time. However, this criterion is more general, and one can also avoid unnecessary iterations and therefore may save total computing time for the algorithm as a whole. For the second proposed convergence method, if we consider the relative difference in mean values for all links to be less than 0.01, which means 1% on average. With the presented values in Table 2 for the case study in Section V, it is obvious that we need less than 100 iterations.

V. TRAVEL TIME ESTIMATION IN MANHATTAN: A CASE STUDY

In this section, the previously mentioned methodology is applied to the NYC taxi trip data set provided by the Taxi and Limousine Commission (TLC), available online at [37]. In this case, the time periods of interest are quarter-hourly intervals from 7 am until 9 am on Tuesday the 1st of February 2011. According to [43], traffic in Manhattan intensifies significantly between 7 am and 9 am and then remains relatively constant until 7 pm. The area of interest is limited to Manhattan; since it particularly suffers from congestion and has a high number of taxi trip observations available relative to its size (165737 on Tuesday the 1st of February 2011 [37]).

The Manhattan network includes a grid road network consisting of 228 numbered streets running in the East-West direction and 11 avenues running in the South-North direction. The network presented as a directed graph is generated using Open Street Map [44], with the nodes representing intersections, and edges representing links. The weight of an edge represents the road distance between two intersections, and the direction of an edge represents the allowed driving direction. Also, the geographical coordinates of the nodes are known. However, other network information, such as the number of lanes, bus stops, and traffic lights are not considered.

The observed GPS coordinates of the starting and end points need to be assigned to a specific point in the graph. This can either be the points lying on an edge or a node. For simplicity, we chose the starting and ending point as the node that is closest to the observed GPS coordinate based on Euclidean distance. After this step, the GPS coordinates are no longer used. Instead, all the starting and ending points are now represented by node IDs corresponding to the graph. As explained in Section III, this step is a source of error on two levels. On the one hand, GPS data is unavoidably noisy, and on the other hand, it is neglected that trips generally do not start and end at intersections. However, the consequences of the latter are not grave, since the average taxi trip observation from the NYC data set covered roughly 40 links (this number is based on the applied path inference method).

The straightforward method explained in Section III is used for the path inference problem. By applying Yen's algorithm [39], up to 20-shortest paths are calculated to find the path with the least difference between the reported trip length and generated trip length. In addition, all trips violating the Inequality (1) are removed. The next step is to find out if the

TABLE 1. Number of observations for different time intervals.

Time interval	Number of Observations	Number of links with more than 10 data points
7:00 – 7:15	1822	1781
7:15 – 7:30	1858	1987
7:30 – 7:45	2008	2011
7:45 – 8:00	2175	2222
8:00 – 8:15	2380	2321
8:15 – 8:30	2477	2390
8:30 – 8:45	2474	2497
8:45 – 9:00	2045	1976

shortest path assumption suffices. For this, we calculate the difference between the individual observed trip length and the shortest distance relative to the observed distance. The mean of this relative difference is 0.088, and the median is 0.052. Judging on behalf of this, the accuracy of the shortest path assumption suffices. One could argue that multiple paths corresponding to an OD pair can have a very similar length but differ widely regarding the links they travel. A large number of path observations compensates for this.

After this step, the data are in the form of path observations. The number of path observations and the number of links with more than 10 data points are presented in Table 1. The time interval a path observation belongs to is defined by the pickup time.

In order to observe the effect of progressive spatial correlation modification, we present the results by comparing the outcomes of both static and dynamic correlated algorithms. Moreover, we present the results of the historic traffic model of Herring [36] in which the links' travel time are assumed to be independent and labeled as an uncorrelated model. The comparison of the mean travel times of individual links is understandable when they are normalized. This is achieved by dividing the individual mean link travel times $\mu_{l,t}$ by the link length. Hereby, we receive the travel time rates, which can be considered as the inverse of the mean velocity. Here, we use the unit seconds per meter. The travel time rate corresponding to the maximum allowed speed suggested by [36] (25mph) is 0.0894 s/m. In Figure 3 and Figure 4, the normalized mean travel time rates are depicted relative to the free-flow travel time rate, where 1 is equal to the free-flow travel time rate, and 5 is five times the free flow travel time rate. In Figure 3, we show the distribution of the link travel times in each time interval by box plots. The top of the rectangle in the box plot indicates the third quartile (75%), the horizontal line near the middle of the rectangle indicates the median (50%), and the bottom of the rectangle indicates the first quartile (25%). In Figure 4, we present the normalized mean value of link travel time on each link on Manhattan network. To highlight the links with particularly high travel time rates, the line widths are adjusted according to the mean link travel time rates.

Figure 3 supports the indication that the traffic overall becomes slower from 7 am to 9 am, which is in line with the earlier work conducted on the New York City taxi data set [43]. It also shows that the difference in travel time

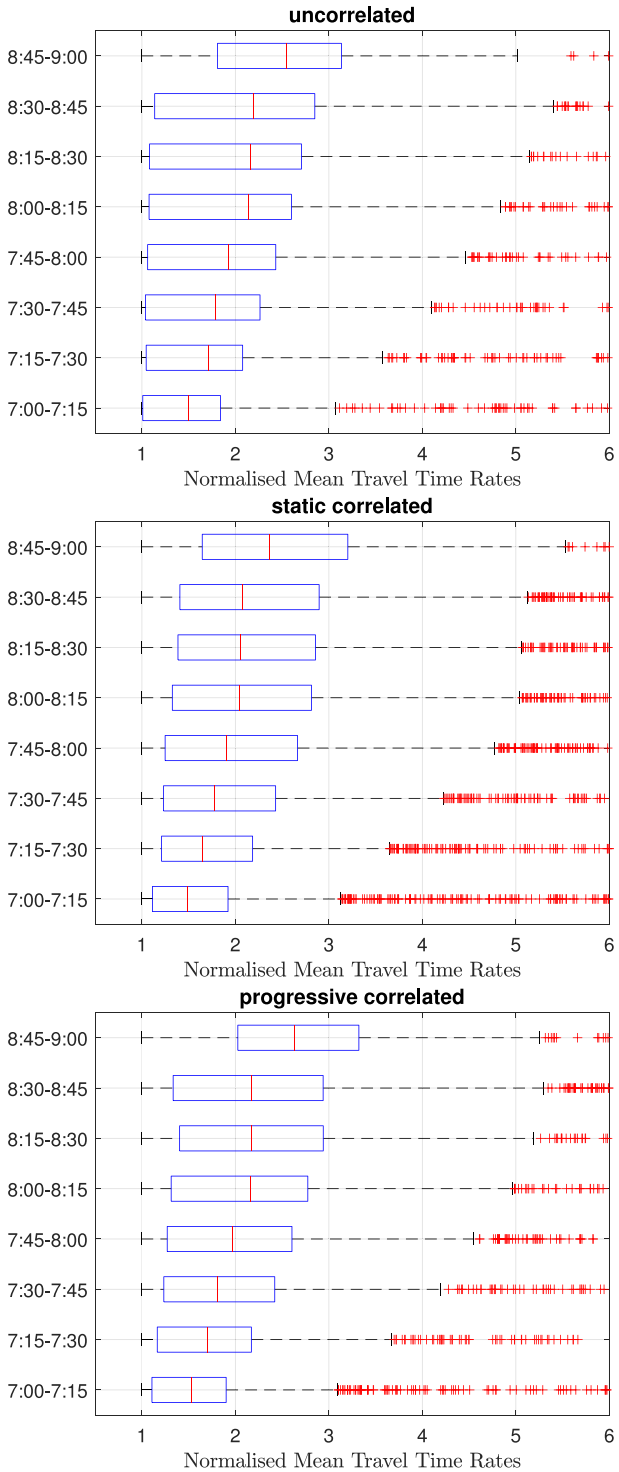

FIGURE 3. Normalized Mean Travel Time Rates (top: uncorrelated middle: static correlated, bottom: progressive correlated).

rates increases among the links; this can be judged from the widening of interquartile boxes over time. Comparing the results of static correlated and progressive correlated, we can observe that the median value rates in progressive correlated box plots are slightly higher than the static correlated ones.

In Figure 4, there is a clear tendency in the depicted time interval that streets converge toward much higher travel time

TABLE 2. Convergence criteria results.

Model	mean relative difference % with 100 iterations
Uncorrelated (Herring's baseline model)	-0.01
Static Correlated	-0.02
progressive Correlated	-0.02

rates than avenues. This confirms the empirically known fact that traffic on streets is slower than traffic on avenues [1]. These values are consistent with previous studies, which have found that the average traffic speed during the day in eastern Midtown is 6.3 mph [17]. This corresponds to the values 3 to 4 in Figure 4. Similar to Figure 3, the mean value rates depicted in the progressive correlated version are slightly higher than the static correlated one.

Moreover, in Figure 5, we show the results in the form of normalized relative differences. The relative differences are calculated based on the order of the models written in the title of each diagram. For instance, the relative difference progressive - static is calculated as follows:

$$\frac{\text{Normalized } \mu_{\text{progressive model}} - \text{Normalized } \mu_{\text{static model}}}{\text{Normalized } \mu_{\text{static model}}} \cdot 100\%$$

Figure 5 gives an instant overview of the changes in mean travel time for each link; however, the best comparison between the performance of the models is presented in Table 3, which is discussed later.

A. CONVERGENCE ANALYSIS

As explained in Section IV--2, the change in the parameter (mean and variance) values should become smaller up to a point where the parameter values will no longer change significantly through iterations. This is called convergence. Table 2 presents the mean relative differences for the mean travel time values associated with all links and all time intervals in different models for the case study. The result shows that all three models, after 100 iterations, have converged to an acceptable mean relative difference.

B. COMPARING OUR RESULTS AGAINST OTHER BENCHMARKS

In this section, we present the results of our exploration through available benchmark data and compare our results against them. For one of the benchmarks, we decided on travel time data provided by the Google direction API [45]. Google historical data is used among other researchers as a comparison benchmark [46]. Google historical travel time data is fetched through third party website Outscraper [47] in which we could extract the instantaneous travel time from an origin to a destination exactly for the study time and date. The complete manual of how to extract historical data from google is explained in [47] for an interested reader. First, we tried to fetch all the travel times for all links in our network and produce travel times of the traveled paths by taxis reported by TLC [37] by adding the travel time of the links. Since TLC does not report the exact path, we used the 20-shortest path calculated based on Yen's algorithm [39] for each observation and chose the path with the lowest length difference from TLC's reported path's length.

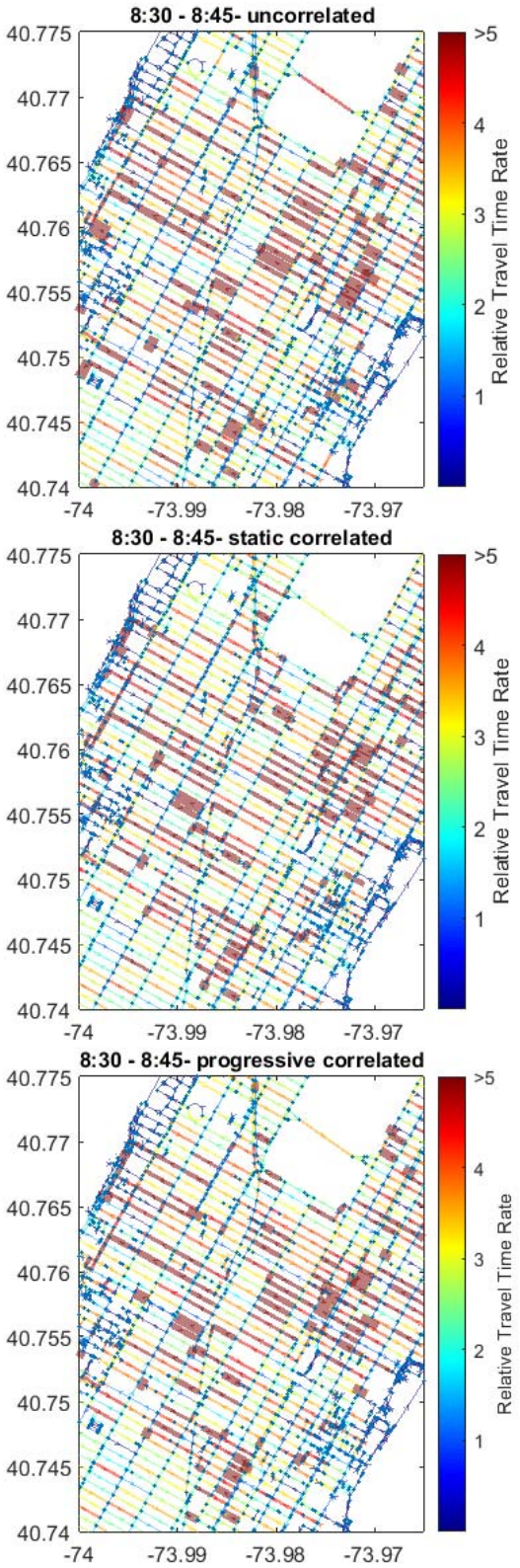

FIGURE 4. Normalized Mean Travel Time Rates on Manhattan network(top: uncorrelated, middle: static correlated, bottom: progressive correlated).

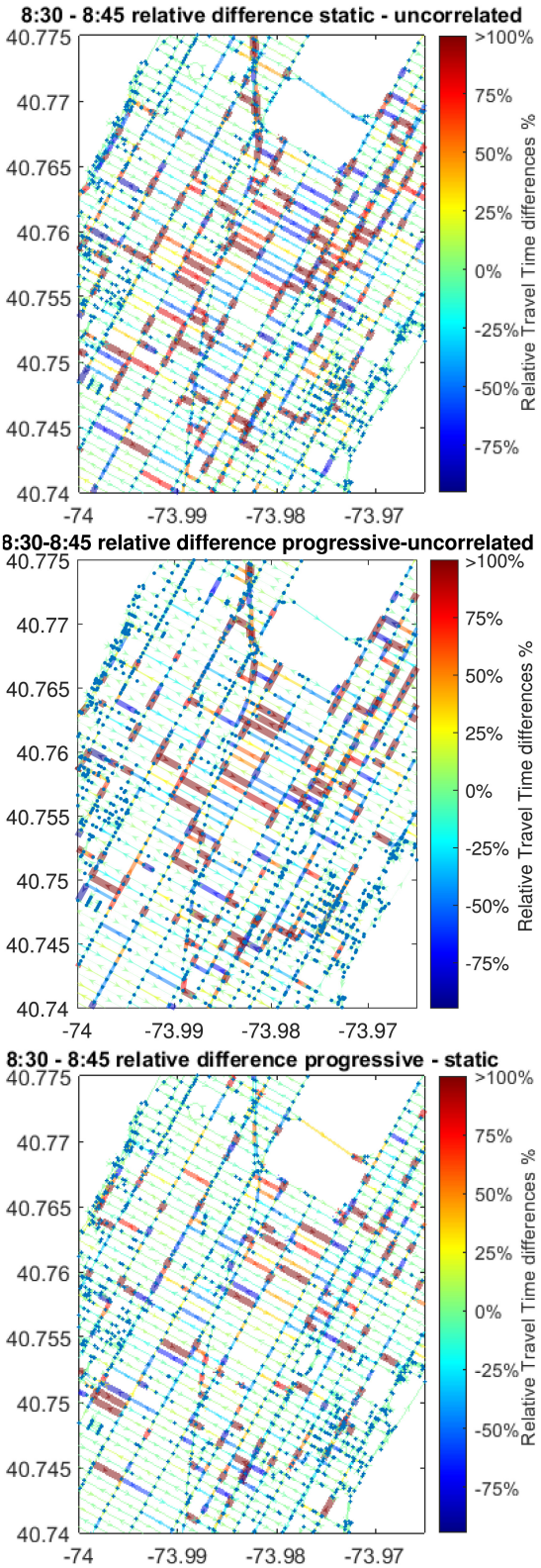

FIGURE 5. Normalized relative differences of mean travel times in different models on Manhattan network.

In this approach, we realized there is a large discrepancy between the path travel time reported by TLC and the one we calculated by adding up the google links travel times.

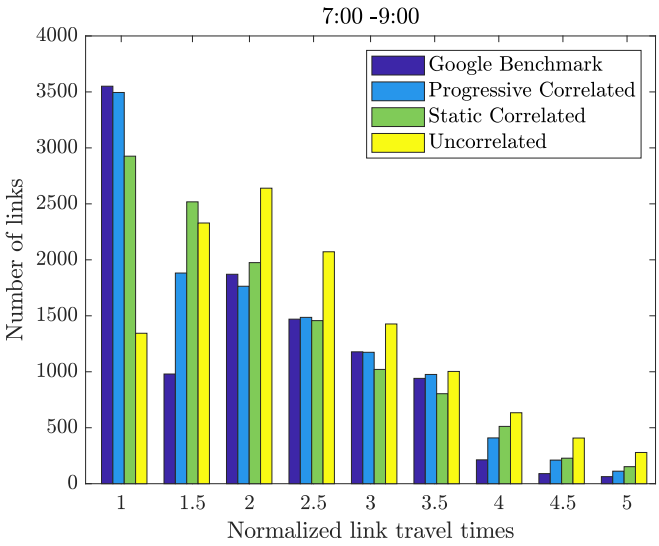

FIGURE 6. Normalized Mean Travel Time Rate Comparisons.

Therefore, we extracted the exact path travel times from google data with the same origin and destination reported by TLC. By this step, we tried to understand if the problem was raised by summing up the link travel times or not. Unfortunately, the same pattern was observed in the path travel time difference. Considering this problem, we could not directly consider the google data as a benchmark and tried to use their data in the following way.

We assume that the ratio of a link travel time to the path travel time is the only valuable data from google we can benefit from. Since both summation of link travel time and path travel time from google are very different from the travel times reported by TLC, the only useful information is the proportion of the link travel time over the path travel time reported by Google. By obtaining all the link travel data and path travel time data from google, we calculated the ratios for each link and path. By multiplying this ratio by the path travel time reported by TLC based on the following equation:

$$X_{l,\text{google benchmark}} = \frac{X_{l,\text{google}}}{Y_{p,\text{google}}} \times Y_{p,\text{TLC}}, \forall l \in L_p$$
$$X_{l,\text{google benchmark}} \sim N\left(\mu_{l,\text{google benchmark}}, \sigma^2_{l,\text{google benchmark}}\right), \quad \forall l \in L_p \tag{20}$$

we get the distribution of google benchmark instantaneous travel times for each link. The mean of this distribution is considered as google benchmark data for each link in our analysis.

Furthermore, to have another data set to compare our results, we use the baseline model proposed by Herring [36] and show the result against this benchmark. In Figure 6, the histograms of normalized travel time rates are depicted for progressive correlated, static correlated, uncorrelated, and calculated google benchmark as explained previously. Moreover, the comparison of RMSE is presented for all

TABLE 3. Experimental results comparison between the proposed models and the baseline model.

Model	RMSE (sec)	MPE	MAPE
Uncorrelated(Herring's baseline model)	143.85	-5.47%	19.67%
Static Correlated	134.39	-4.41%	18.17%
progressive Correlated	127.73	-3.15%	16.71%

three methods in the following table. The metrics in Table 3 are calculated based on the following equations:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(\hat{x}_i - x_i)^2},$$

$$MPE = \frac{100\%}{n}\sum_{i=1}^{n}(\frac{x_i - \hat{x}_i}{x_i}),$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}|\frac{x_i - \hat{x}_i}{x_i}|,$$

where x_i is the i^{th} path observed value for travel time reported by TLC [37] and $\hat{x}_i$ is the estimated path observation achieved by summation of the link travel times in that path. Negative values of MPE mean that the estimated value is larger than the observed value.

The trend in Figure 6 shows that, in all normalized travel time rates, the dynamic correlated model is closer to the google benchmark data compared to the static correlated model results. However, the result in Figure 6 is very aggregated, and the comparison between the three models is best achieved by comparing the metrics in Table 3. In Table 3, we observe that the progressive model values are showing the best result. Therefore, the progressive model can estimate the links' travel time more accurately than the other models.

VI. CONCLUSION

This work proposes a methodology to estimate historical link travel times based on GPS OD data; historical means that the parameters uniquely belong to a past time period. Of course, such a process could be applied in a real-time setting or a hybrid model by combining historical estimates and real-time measurements. The proposed model infers the unknown path by the cabs with the simple assumption that the cabs always travel the shortest path based on the distance, and the difference between the observed and calculated path is reduced by calculating up to the 20-shortest path utilizing Yen's algorithm [39]. The link travel times and their corresponding variances can then be estimated by formulating a maximum likelihood function. This optimization problem is computationally challenging but can be tackled by an iterative decomposition scheme suggested by [36]. In order to consider the spatial correlation, we have proposed a spatial correlation matrix for each sub-network and adopted the methodology for correlated links.

The model was applied to the Manhattan network for quarter-hourly time intervals from 7 am to 9 am on Tuesday, 1st of February 2011. The data used in this study were collected by the yellow New York City taxi cabs and are provided by the New York City Taxi and Limousine Commission [37]. The time of day had a significant effect on the means and variability of the travel times, with travel times gradually increasing on many links from 7 am to 9 am. The algorithm correctly detected a spatial pattern of streets having higher relative travel times than avenues in all time intervals. Furthermore, by comparing our results against other benchmarks, we show that the consideration of progressive correlation can improve the results, thus leading to a more accurate parametric travel time estimation approach.

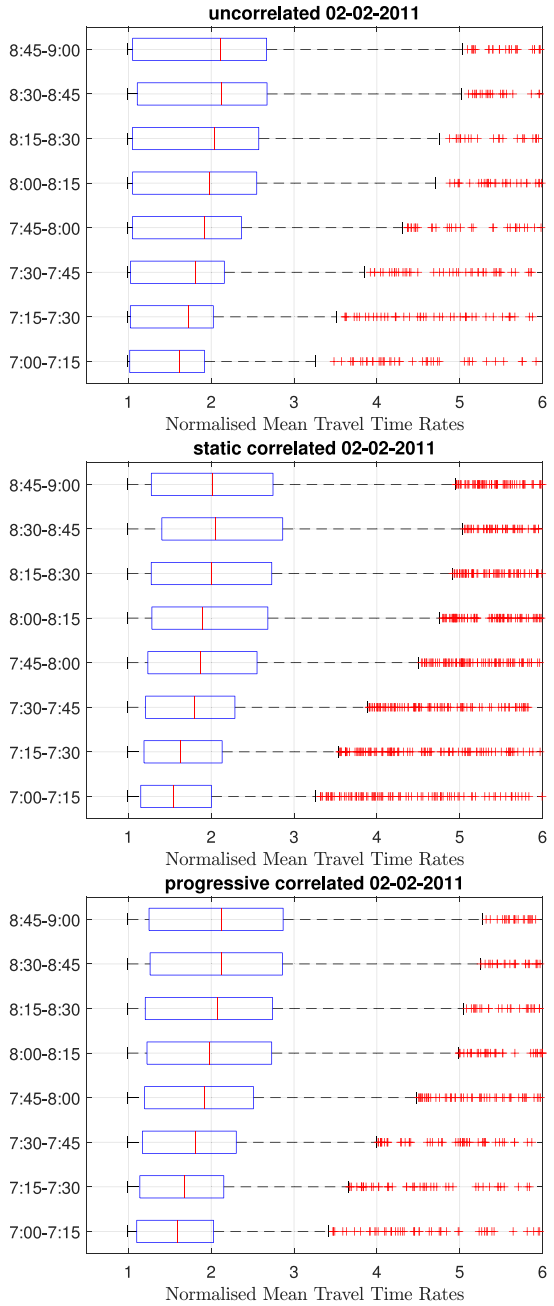

FIGURE 7. Normalized Mean Travel Time Rates for all models for Wednesday 02-02-2011.

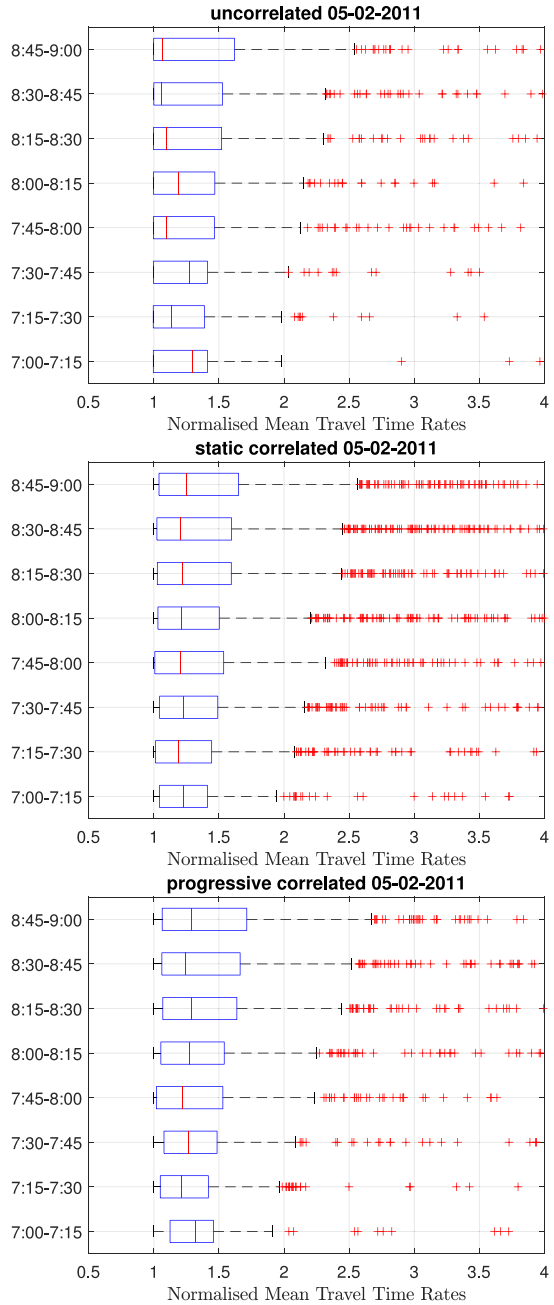

FIGURE 8. Normalized Mean Travel Time Rates for all the models for Saturday 05-02-2011.

The proposed methodology can be applied to any GPS probe vehicle data set, for instance, synthetic data provided by [48], [49] or real data set [37], [50], given that the data provide the origin, destination, and path travel time. Furthermore, the higher number of observations for a link travel time can increase the accuracy of the proposed methodology [7].

This study proposes a more accurate approach for estimating travel times that fully utilizes the partial information received from taxi data in cities as well as known or constructed (static or progressive) spatial correlations.

TABLE 4. Experimental results comparison between the proposed models and the baseline model for Wednesday 02-02-2011.

Model	RMSE (sec)	MPE	MAPE
Uncorrelated(Herring's baseline model)	144.72	-5.83%	20.89%
Static Correlated	137.98	-5.42%	19.85%
Progressive Correlated	127.73	-4.15%	18.08%

TABLE 5. Experimental results comparison between the proposed models and the baseline model for Saturday 05-02-2011.

Model	RMSE (sec)	MPE	MAPE
Uncorrelated(Herring's baseline model)	76.25	-5.20 %	15.77%
Static Correlated	72.00	-4.49 %	15.30%
Progressive Correlated	70.17	-3.68%	14.32%

APPENDIX

In this Appendix, we present the results of all the proposed models for another day of the week (Wednesday 02-02-2011) in Figure 7 and a Weekend day (Saturday 05-02-2011) in Figure 8. In addition, the experimental result comparison between the proposed models is presented in Table 4 for Wednesday 02-02-2011 and in Table 5 for Saturday 05-02-2011. We can conclude that, the progressive model has the best performance comparing to the other models.

ACKNOWLEDGMENT

The authors would like to thank Moritz Rau for his contribution to this work as part of his B.Sc. thesis at ETH Zurich.

REFERENCES

[1] D. Bertsimas, A. Delarue, P. Jaillet, and S. Martin, "Travel time estimation in the age of big data," *Oper. Res.*, vol. 67, no. 2, pp. 498–515, Mar. 2019. [Online]. Available: https://ideas.repec.org/a/inm/oropre/v67y2019i2p498-515.html

[2] M. Chen and S. I. J. Chien, "Dynamic freeway travel-time prediction with probe vehicle data: Link based versus path based," *Transp. Res. Rec.*, vol. 1768, no. 1, pp. 157–161, 2001. [Online]. Available: https://doi.org/10.3141/1768-19

[3] M. El Esawey and T. Sayed, "Travel time estimation in urban networks using limited probes data," *Can. J. Civil Eng.*, vol. 38, no. 3, pp. 305–318, 2011. [Online]. Available: https://doi.org/10.1139/L11-001

[4] T. Hunter, R. Herring, P. Abbeel, and A. Bayen, "Path and travel time inference from GPS probe vehicle data," *Anal. Netw. Learn. Graphs*, vol. 12, no. 1, p. 2, 2009.

[5] E. Jenelius and H. N. Koutsopoulos, "Travel time estimation for urban road networks using low frequency probe vehicle data," *Transp. Res. B, Methodol.*, vol. 53, pp. 64–81, Jul. 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191261513000489

[6] J. Yuan *et al.*, "T-drive: Driving directions based on taxi trajectories," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2010, pp. 99–108. [Online]. Available: https://doi.org/10.1145/1869790.1869807

[7] R. Herring, A. Hofleitner, P. Abbeel, and A. Bayen, "Estimating arterial traffic conditions using sparse probe data," in *Proc. 13th Int. IEEE Conf. Intell. Transp. Syst.*, 2010, pp. 929–936.

[8] H. Wang, X. Tang, Y.-H. Kuo, D. Kifer, and Z. Li, "A simple baseline for travel time estimation using large-scale trip data," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, p. 19, Jan. 2019. [Online]. Available: https://doi.org/10.1145/3293317

[9] A. Sen, P. Thakuriah, X. Zhu, and A. F. Karr, "Frequency of probe vehicle reports and variances of link travel time estimates," *J. Transp. Eng.*, vol. 123, no. 4, pp. 290–297, 1997.

[10] B. Coifman, "Estimating travel times and vehicle trajectories on freeways using dual loop detectors," *Transp. Res. A, Policy Pract.*, vol. 36, no. 4, pp. 351–364, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0965856401000076

[11] F. Zheng and H. Van Zuylen, "Urban link travel time estimation based on sparse probe vehicle data," *Transp. Res. C, Emerg. Technol.*, vol. 31, pp. 145–157, Jun. 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X12000575

[12] C.-H. Wu, J.-M. Ho, and D. T. Lee, "Travel-time prediction with support vector regression," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 4, pp. 276–281, Dec. 2004.

[13] D. Park and L. R. Rilett, "Forecasting multiple-period freeway link travel times using modular neural networks," *Transp. Res. Rec.*, vol. 1617, no. 1, pp. 163–170, 1998. [Online]. Available: https://doi.org/10.3141/1617-23

[14] R. Li and G. Rose, "Incorporating uncertainty into short-term travel time predictions," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 6, pp. 1006–1018, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X11000842

[15] H. D. Sherali, J. Desai, and H. Rakha, "A discrete optimization approach for locating automatic vehicle identification readers for the provision of roadway travel times," *Transp. Res. B, Methodol.*, vol. 40, no. 10, pp. 857–871, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191261506000038

[16] J. Yeon, L. Elefteriadou, and S. Lawphongpanich, "Travel time estimation on a freeway using discrete time Markov chains," *Transp. Res. B, Methodol.*, vol. 42, no. 4, pp. 325–338, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191261507000768

[17] X. Zhan, S. Hasan, S. V. Ukkusuri, and C. Kamga, "Urban link travel time estimation using large-scale taxi data with partial information," *Transp. Res. C, Emerg. Technol.*, vol. 33, pp. 37–49, Aug. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X13000740

[18] H. X. Liu and W. Ma, "A virtual vehicle probe model for time-dependent travel time estimation on signalized arterials," *Transp. Res. C, Emerg. Technol.*, vol. 17, no. 1, pp. 11–26, 2009.

[19] G. Leduc *et al.*, "Road traffic data: Collection methods and applications," Working Papers Energy Transport Climate Change, Inst. Prospect. Technol. Stud., Seville, Spain, 2008.

[20] M. Ramezani and N. Geroliminis, "On the estimation of arterial route travel time distribution with Markov chains," *Transp. Res. B, Methodol.*, vol. 46, no. 10, pp. 1576–1590, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191261512001087

[21] K. Tang, S. Chen, Z. Liu, and A. J. Khattak, "A tensor-based Bayesian probabilistic model for citywide personalized travel time estimation," *Transp. Res. C, Emerg. Technol.*, vol. 90, pp. 260–280, May 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X18303103

[22] Z. Ma, H. N. Koutsopoulos, L. Ferreira, and M. Mesbah, "Estimation of trip travel time distribution using a generalized Markov chain approach," *Transp. Res. C, Emerg. Technol.*, vol. 74, pp. 1–21, Jan. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X16302248

[23] M. Rahmani, E. Jenelius, and H. Koutsopoulos, "Non-parametric estimation of route travel time distributions from low-frequency floating car data," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 343–362, Sep. 2015.

[24] A. Allström *et al.*, "Hybrid approach for short-term traffic state and travel time prediction on highways," *Transp. Res. Rec.*, vol. 2554, no. 1, pp. 60–68, 2016. [Online]. Available: https://doi.org/10.3141/2554-07

[25] A. Hofleitner, R. Herring, and A. Bayen, "Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning," *Transp. Res. B, Methodol.*, vol. 46, no. 9, pp. 1097–1122, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191261512000513

[26] R. W. Hall, "The fastest path through a network with random time-dependent travel times," *Transp. Sci.*, vol. 20, no. 3, pp. 182–188, 1986. [Online]. Available: https://doi.org/10.1287/trsc.20.3.182

[27] L. Rilett and D. Park, "Direct forecasting of freeway corridor travel times using spectral basis neural networks," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1752, no. 1, pp. 140–147, 2001.

[28] W. Eisele and L. Rilett, "Estimating corridor TravelTime mean, variance, correlation with intelligent Transporta-tion systems link travel time data," in *Proc. 81st Annu. Meeting Transp. Res. Board*, 2002, p. 20.

[29] B. Gajewski and L. Rilett, "Estimating link travel time correlation: An application of Bayesian smoothing splines," *J. Transp. Stat.*, vol. 7, nos. 2–3, pp. 53–70, 2005.

[30] K. Chan, W. Lam, and M. Tam, "Real-time estimation of arterial travel times with spatial travel time covariance relationships," *Transp. Res. Rec.*, vol. 2121, no. 1, pp. 102–109, Dec. 2009.

[31] L. Fu and L. Rilett, "Expected shortest paths in dynamic and stochastic traffic networks," *Transp. Res. B, Methodol.*, vol. 32, no. 7, pp. 499–516, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191261598000162

[32] W. Zeng, T. Miwa, Y. Wakita, and T. Morikawa, "Application of lagrangian relaxation approach to α-reliable path finding in stochastic networks with correlated link travel times," *Transp. Res. C, Emerg. Technol.*, vol. 56, pp. 309–334, Jul. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X1500159X

[33] B. Y. Chen, W. Lam, and Q. Li, "Efficient solution algorithm for finding spatially-dependent reliable shortest path in road networks," *J. Adv. Transp.*, vol. 50, pp. 1413–1431, Nov. 2016.

[34] P. Rachtan, H. Huang, and S. Gao, "Spatiotemporal link speed correlations: Empirical study," *Transp. Res. Rec.*, vol. 2390, no. 1, pp. 34–43, 2013. [Online]. Available: https://doi.org/10.3141/2390-04

[35] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Econ. Geography*, vol. 46, pp. 234–240, Jun. 1970. [Online]. Available: http://www.jstor.org/stable/143141

[36] R. J. Herring, "Real-time traffic modeling and estimation with streaming probe data using machine learning," Ph.D. dissertation, Ind. Eng. Oper. Res., Univ. California, Berkeley, CA, USA, 2010.

[37] "NYC TLC trip record data." NYC Taxi and Limousine Commission. 2019. Accessed: Jul. 1, 2019. [Online]. Available: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[38] J. Alonso-Mora, S. Samaranayake, A. Wallar, E. Frazzoli, and D. Rus, "On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment," *Proc. Nat. Acad. Sci.*, vol. 114, no. 3, pp. 462–467, 2017. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1611675114

[39] J. Y. Yen, "An algorithm for finding shortest routes from all source nodes to a given destination in general networks," *Quart. Appl. Math.*, vol. 27, pp. 526–530, Jan. 1970.

[40] O. A. Nielsen, "On the distributions of the stochastic components in sue (stochastic user equilibrium) traffic assignment models," in *Proc. Seminar PTRC Eur. Transp. Forum*, Sep. 1997, pp. 77–93.

[41] H. Rakha, I. EL-Shawarby, M. Arafeh, and F. Dion, "Estimating path travel-time reliability," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2006, pp. 236–241.

[42] *The Multivariate Normal Distribution*. Hoboken, NJ, USA: Wiley, 2002, ch. 4, pp. 82–111. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471271357.ch4

[43] M. M. Grynbaum, *Gridlock May Not Be Constant, but Slow Going is Here to Stay*, New York Times, New York, NY, USA, Mar. 2010.

[44] "Open street map." Accessed: Jul. 1, 2021. [Online]. Available: https://www.openstreetmap.org/

[45] "The directions API overview." Google Developers. Accessed: Oct. 2021. [Online]. Available: https://developers.google.com/maps/documentation/directions/overview

[46] A. Genser, N. Hautle, M. Makridis, and A. Kouvelas, "An experimental urban case study with various data sources and a model for traffic estimation," *Sensors*, vol. 22, no. 1, p. 144, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/1/144

[47] "Google maps traffic extractor." Outscraper. Accessed: Mar. 2022. [Online]. Available: https://outscraper.com/google-maps-traffic-extractor/

[48] S. F. A. Batista, G. Cantelmo, M. Menéndez, and C. Antoniou, "A Gaussian sampling heuristic estimation model for developing synthetic trip sets," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 37, no. 1, pp. 93–109, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12697

[49] G. Flötteröd and M. Bierlaire, "Metropolis-hastings sampling of paths," *Transp. Res. B, Methodol.*, vol. 48, pp. 53–66, Feb. 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S091261512001 52X

[50] "Taxi trips reported to the city of chicago." Chicago Open Data. 2020. Accessed: Jul. 11, 2022. [Online]. Available: https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew#column-menu

ZAHRA GHANDEHARIOUN (Student Member, IEEE) received the B.Sc. degree in civil engineering and the M.Sc. degree in transportation systems from the Technical University of Munich, Munich, Germany, in 2010 and 2014, respectively. She is currently pursuing the Ph.D. degree in traffic engineering and control with the Institute for Transport Planning and Systems, Department of Civil, Environmental and Geomatic Engineering, ETH Zurich, Switzerland. Her research interests include developing algorithm for travel time estimation based on probe data and optimization of on demand transportation systems in particular ride-sharing systems.

ANASTASIOS KOUVELAS (Senior Member, IEEE) received the Diploma, M.Sc., and Ph.D. degrees in modeling, control, and optimization of large-scale transport systems from the Department of Production and Management Engineering (Operations Research), Technical University of Crete, Greece, in 2004, 2006, and 2011, respectively. He has been the Director of the Research Group Traffic Engineering and Control with the Institute for Transport Planning and Systems (IVT), Department of Civil, Environmental and Geomatic Engineering, ETH Zurich, since August 2018. Prior to joining IVT, he was a Research Scientist with Urban Transport Systems Laboratory, EPFL, from 2014 to 2018, and a Postdoctoral Fellow with Partners for Advanced Transportation Technology with the University of California at Berkeley, Berkeley, from 2012 to 2014. Before this, he was appointed as an Adjunct Professor with the Technical University of Crete in 2011, and a Research Associate with the Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece, from 2011 to 2012. In 2009, he was a Doctoral Visiting Scholar with the Center for Advanced Transportation Technologies of the Viterbi School of Engineering, Department of Electrical Engineering, University of Southern California, Los Angeles. He has been awarded with the 2012 Best IEEE ITS Ph.D. Dissertation Award from IEEE Intelligent Transportation Systems Society. He has been serving as an Associate Editor for *IET Intelligent Transport Systems* since 2017 and has guest-edited several special issues in transportation journals. He has been a member of the Transportation Research Board Standing Committee (AHB15) on Intelligent Transportation Systems, since 2019.