Transportation Research Part B 53 (2013) 64–81

Contents lists available at SciVerse ScienceDirect

Transportation Research Part B

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

Travel time estimation for urban road networks using low frequency probe vehicle data

Erik Jenelius *, Haris N. Koutsopoulos

KTH Royal Institute of Technology, Department of Transport Science, Division of Traffic and Logistics, Sweden

ARTICLE INFO

Article history:
Received 9 December 2012
Received in revised form 26 March 2013
Accepted 26 March 2013

Keywords:
Travel time
Network
Estimation
Probe vehicle
Low frequency sampling

ABSTRACT

The paper presents a statistical model for urban road network travel time estimation using vehicle trajectories obtained from low frequency GPS probes as observations, where the vehicles typically cover multiple network links between reports. The network model separates trip travel times into link travel times and intersection delays and allows correlation between travel times on different network links based on a spatial moving average (SMA) structure. The observation model presents a way to estimate the parameters of the network model, including the correlation structure, through low frequency sampling of vehicle traces. Link-specific effects are combined with link attributes (speed limit, functional class, etc.) and trip conditions (day of week, season, weather, etc.) as explanatory variables. The approach captures the underlying factors behind spatial and temporal variations in speeds, which is useful for traffic management, planning and forecasting. The model is estimated using maximum likelihood. The model is applied in a case study for the network of Stockholm, Sweden. Link attributes and trip conditions (including recent snowfall) have significant effects on travel times and there is significant positive correlation between segments. The case study highlights the potential of using sparse probe vehicle data for monitoring the performance of the urban transport system.

1. Introduction

Many urban road transport systems today experience increasing congestion that threatens the environment and the transport efficiency. To tackle these problems, knowledge about traffic conditions is critical at many levels of traffic management and transport policy. Through information and personalized advice, individuals and fleet management companies can plan their trips more accurately and increase the efficiency of the system. For traffic management, speed information at the segment level can reveal problematic locations where new or revised traffic control schemes may be introduced to increase performance. For transport policy, network-wide travel time information provides input for travel demand forecasting and impact assessments of policy instruments such as congestion charges.

There are a number of well-established technologies for collecting speed and travel time data, including loop detectors, automatic vehicle identification (AVI) sensors and, probe car data. Loop detectors and AVI sensors have the merit that they, once installed, continuously record every vehicle passing the monitored road section. However, the share of segments in the network equipped with these sensors is typically low and not representative of the urban network as a whole, which leaves the traffic conditions in most of the network unknown. Dedicated probe vehicles, meanwhile, are used to collect travel time and other data for designated routes in the network. However, due to cost considerations the number of traffic studies with

* Corresponding author. Address: Brinellvägen 32, SE-100 44 Stockholm, Sweden. Tel.: +46 8 790 8032; fax: +46 8 21 2899.
E-mail addresses: erik.jenelius@abe.kth.se (E. Jenelius), haris.koutsopoulos@abe.kth.se (H.N. Koutsopoulos).

0191-2615/$ - see front matter
http://dx.doi.org/10.1016/j.trb.2013.03.008

probe vehicles is typically small and the number of vehicles involved very low. Hence, they can only cover a limited number of routes for a limited duration of time.

Most recently, GPS devices, already installed for other purposes in vehicle fleets (e.g., taxis, commercial vehicles, service vehicles, etc.) or smartphones, provide a new type of traffic sensor. These opportunistic sensors have a great potential for provision of data for traffic management applications. Unlike stationary sensors, they can collect travel time and speed data for any part of the network where equipped vehicles move, without the need for traffic flow measurements. Unlike designated probe cars, they can continuously collect data for any time and day that equipped vehicles are active. Liu et al. (2009) investigated the feasibility of using data from taxi dispatching systems to collect reliable traffic information. They concluded that real-time detection of congestion on links is possible, provided that an adequate database of historical traffic conditions has been constructed and that sufficiently long road segments are considered.

However, a number of limitations mean that new sophisticated methods are needed to process the data and generate useful information, compared to traditional sensors (Leduc, 2008). The most significant challenge is that sampling (also called reporting, or polling) frequencies are often low (less than one per minute), so that vehicles may have traversed significant distances between reports. For arterial networks, low sampling frequency creates difficulties in inferring the true path of the vehicle between two position reports, which may involve a considerable number of network segments (Rahmani and Koutsopoulos, 2013; Miwa et al., 2008). Furthermore, the fraction of the reported travel time that is spent on each individual segment is not observed, which creates challenges for network travel time estimation.

The problem considered in this paper is to estimate the travel time for any route between any two points in the network under specified trip conditions; both the mean travel time and the variability are of interest, considering that link travel times along the route may be correlated. The observations used for the estimation come from probe vehicles that travel on the network, reporting the time and their positions at certain intervals. The sampling frequency is assumed to be low, in the sense that the distances between consecutive reports are typically longer than the scale desired to estimate travel times. The only information considered is the observed travel times and distances between reports; thus, data on, for example, instantaneous speeds, are not available. This may be a common case due to limitations in communication bandwidth.

The literature on travel time estimation and forecasting using probe vehicle data has grown in recent years as the technology has become more available. Most papers, however, have dealt with high frequency probe data, (e.g., Zou et al. 2005; Work et al., 2008), which eliminates many of the challenges of interest here. Further, much of the research has focused on highway traffic estimation, in particular on the use of probe vehicle data for online estimation of macroscopic flow models (e.g., Nanthawichit et al., 2003; Work et al., 2008; Yuan et al., 2012).

Most proposed methods for travel time model estimation using low-frequency data divide the process into two steps, performed either once or iteratively. First, each probe vehicle travel time observation is decomposed into a travel time for each traversed network segment. This is sometimes known as the travel time allocation problem (Hofleitner et al., 2012a; Zheng and van Zuylen, in press). Second, the travel time distributions for the network components are estimated or calculated empirically. Heuristic methods for the travel time allocation are proposed by Hellinga et al. (2008), and Zheng and van Zuylen (in press).

Hunter et al. (2009) present a probabilistic model of travel times in the arterial network, based on low frequency taxi GPS probes. The model takes into account that the path between two consecutive position reports may contain multiple segments, and the authors formulate a maximum likelihood problem to estimate the segment travel time distributions based on the set of observed route travel times. A simulation-based EM estimation algorithm is proposed, which iterates between travel time allocation and parameter estimation. The authors assume that the travel times on different segments are independent and report estimation results using normal and log-normal distributions.

A development of the approach in Hunter et al. (2009), more clearly aimed towards travel time forecasting, is presented in Hofleitner et al. (2012a). The model assumes that each segment can be in one of two possible states (congested or uncongested), each with its own conditional, independent normal travel time distribution. The transitions between states among neighboring segments are modeled as a dynamic Bayesian network model. The unobserved state and transition probabilities and the travel time distribution parameters are estimated in a simulation-based EM approach. A further development using travel time distributions inspired by traffic theory, state variables representing the number of queuing vehicles on each link, and turn fractions at intersections, is presented in Hofleitner et al. (2012b).

Another approach, using low frequency GPS data from ambulances, is presented in Westgate et al. (in press). In the paper, path inference and travel time estimation are performed simultaneously using a Bayesian approach. The framework makes use of instantaneous speed information reported by the vehicles. The travel times on the road links are assumed to be independent and log-normally distributed, and the parameters are estimated using Markov chain Monte Carlo methods.

Although the previous work clearly demonstrates the possibility to extract useful information from low-frequency probe vehicle data, the proposed methods have certain drawbacks. First, the formulation of the travel time allocation problem requires that each network segment has a travel time distribution with its own unique parameters. This means that significant amounts of data may be required for all parts of the network to obtain reliable and significant estimates of all model parameters. Second, the travel times on different segments are assumed to be independent conditional on the state of the system (which is typically held fixed in 15-min or longer intervals). Since travel times are often correlated across links (Bernard et al., 2006; de Fabritiis et al., 2008; Ramezani and Geroliminis, 2012), this may lead to incorrect estimates of the travel time variability on longer routes. Third, the models do not incorporate any structural information about factors having an effect on travel times, which means that they cannot be used to predict the impacts of policy measures or other variables influencing travel conditions.

The methodology proposed in this paper extends previous work on travel time estimation using probe vehicle data by considering the effect of explanatory variables on travel times. For the network components themselves this may include attributes such as speed limit, number of lanes, functional class, bus stops, traffic signals, stop signs, and left turns. The effects of the conditions for the trip, such as weather, time of day, weekday, time of year, and so on, are considered.

The use of explanatory variables is attractive for at least three reasons. First, it allows the identification of the underlying causes that contribute to the variability in speeds between links and points in time. This is important for system management and planning, where one needs to know the relationships between possible instruments and network travel times in order to improve the mobility and accessibility in the transport system. Second, the approach reduces the number of parameters to estimate and travel times can be estimated even with low frequency probe data and in areas with relatively few observations. This aspect has not been discussed much in previous work (it is handled implicitly in Bayesian approaches) but proved essential in practical applications of the probe vehicle data source used in this paper. Third, the integration of trip conditions in the model makes it possible to extend the historical estimation to prediction of future travel times based on forecasted conditions.

Furthermore, the methodology extends previous work by utilizing the sequences of position reports from the same vehicle to incorporate and estimate the correlation experienced by the vehicle traversing the segments sequentially on a trip. The network model assumes that segment travel times have a multivariate normal distribution according to a spatial moving average (SMA) structure. The observation model takes into account that the correlation between segments is incorporated not only in each probe vehicle travel time observation, but in the entire sequence of observations from the same vehicle trajectory.

The proposed methodology is based on a statistical model that consists of two layers: The network model specifies the joint distribution of travel times at the network segment level and expresses it in terms of a set of unknown parameters; the observation model specifies the information contained in sequences of probe vehicle reports and provides the link from the network model to the maximum likelihood estimation of the unknown parameters. The applicability of the method and the impacts of various factors on mean and variance of travel times are demonstrated empirically for a network in Stockholm, Sweden, using probe vehicle data from a taxi fleet. The case study highlights the potential in probe vehicle data for estimating traffic conditions even without the availability of classical traffic data such as flows.

The problem of determining the true positions and paths of the vehicles from noisy GPS data is not considered in this paper. It is assumed that the most likely network location (i.e., the network segment and position along segment) corresponding to each GPS measurement, as well as the path (i.e., the sequence of network segments) taken between each pair of consecutive measurements, have been determined by a map-matching and path inference process (Rahmani and Koutsopoulos, 2013; Miwa et al., 2012; Hunter et al., 2012).

The paper is organized as follows. The network model is presented in Section 2 and the observation model is presented in Section 3. Section 4 discusses some practical considerations regarding the specification and estimation of the model. This is followed by a description of a real-world application in Section 5 and a concluding discussion in Section 6.

2. Network model

The travel time of a trip is assumed to consist of two components:

1. Running travel time along links,
2. Delay at intersections and traffic signals (turns).

A *link* is defined to be the road section between two adjacent intersections or traffic signals (traffic signals are not always located at intersections). A link may be divided into one or more *segments*, with each segment part of one specific link.

N_S and N_L denote the total number of segments and links in the network, respectively. The relationship between segments and links is described by an $N_S \times N_L$ incidence matrix $\mathbf{S}$, so that element S_{sl} is 1 if segment s belongs to link l and 0 otherwise. Since a segment can only belong to a single link, $\sum_l S_{sl} = 1$ for all s. The link $l(s)$ of segment s is identified as the link such that $S_{s,l(s)} = 1$.

While the links are largely determined by the inherent network structure, the number of segments per link depends on the traffic characteristics of the link. Segments are designed to capture homogeneous traffic behavior. In this model the average speed of a vehicle can vary between segments but is assumed to be constant along each segment. The travel time on a segment s is presented as the length of the segment, ℓ_s, multiplied with the inverse speed or *travel time rate* X_s. As described in the following subsections, the travel time rate may depend on observed and unobserved properties of the segment and conditions for the trip.

The second travel time component of a trip is intersection and traffic signal delay. Let *turn* t define the movement from a link l_1 to the downstream link l_2, and N_T denote the total number of turns in the network. A turn can thus be defined by the pair (l_1, l_2). Turn t is assumed to impose a travel time penalty h_t. Factors that influence the magnitude of h_t may include the type of traffic control in the intersection, whether it involves a left or a right turn, time of day, etc.

Conceptually, turns can be seen as links having zero length, as illustrated in Fig. 1. This means that the probability of a vehicle reporting its position on a turn link is zero. While one can readily incorporate both types of components in a single set of variables, in this paper separate sets of variables for link running times and turn penalties are used for ease of exposition.

The segment travel time rates X_s, $s = 1, \ldots, N_S$, are modeled as stochastic variables, in general not independent. Both mean travel times and the variability around the mean values are of interest. Assuming that the mean value is finite, $X_s = g_s + v_s$, where g_s is the mean travel time rate and v_s; is a stochastic component with $E[v_s] = 0$ capturing the variability around the mean. For compact notation, it is convenient to introduce the $N_S \times 1$ vector $\mathbf{X}$ with elements X_s. Then

$$\mathbf{X} = \mathbf{g} + \boldsymbol{v}, \tag{1}$$

where $\mathbf{g}$ is the vector of mean segment travel time rates, and $\mathbf{v}$ is the vector of zero-mean stochastic terms. Conditional on trip conditions, the model assumes that the stochastic components of the segment travel time rates $\boldsymbol{v}$ follow a multivariate normal distribution according to a covariance structure defined in Section 2.2.

The turn penalties can be treated as deterministic or stochastic variables. In reality the delay at an intersection certainly varies according to unobserved changes in traffic flows, signal cycles, etc. For estimation purposes, however, it is difficult to separate the total travel time variability into variability in running travel times and in turn delays. Hence, in the application in this paper turn delays are treated as deterministic travel time penalties. The turn penalties h_t are represented by the $N_T \times 1$ vector $\mathbf{h}$.

2.1. Mean structure

Vectors $\mathbf{g}$ and $\mathbf{h}$ may be further expressed as functions of a number of factors with associated parameters to be estimated from data. The parametric structure should reflect the way different factors affect travel times, while also allowing for convenient and efficient estimation. The structure can also be used to partition the segments into larger groups to ensure that all parameters can be identified through the available observations. The explanatory variables for the mean segment travel time rates $\mathbf{g}$ capture *segment characteristics* that vary across the network, and *trip conditions* that vary in time.

The explanatory factors related to segment characteristics include regulatory factors, such as speed limit and classification, link length, nearby land use and fixed effects for specific segments. Segment flows are not specifically considered in the model structure but can be included if available. The N_B different attributes are represented by the $N_S \times N_B$ design matrix $\mathbf{B}$. The baseline segment travel time rates are then $\mathbf{B}\boldsymbol{\beta}_B$, where $\boldsymbol{\beta}_B$ is an $N_B \times 1$ parameter vector to be estimated. Note that the segments can be modeled as having fully distinct means, without any other explanatory variables, by setting $\mathbf{B}$ equal to the $N_S \times N_S$ identity matrix.

The observed trip conditions are assumed to act as a multiplier to the baseline segment travel time rates. Thus, a certain trip condition multiplies the baseline travel time rates $\mathbf{B}\boldsymbol{\beta}_B$ on all segments according to a certain percentage relative to some reference conditions. Relevant trip attributes could include temporal variations within the day, week and year, weather conditions, etc. For a given trip, the attributes are represented by an $1 \times N_o$ design vector $\mathbf{o}$. The multiplier for the trip conditions is then $(1 + \mathbf{o}\boldsymbol{\beta}_o)$, where β_o is an $N_o \times 1$ parameter vector to be estimated. In total, the mean segment travel time rates can be written as

$$\mathbf{g} = (1 + \mathbf{o}\boldsymbol{\beta}_o) \cdot \mathbf{B}\boldsymbol{\beta}_B. \tag{2}$$

The mean turn penalties $\mathbf{h}$ have a similar structure. The explanatory factors for the turn penalties, which may include indicators for traffic signals, left/right turns, etc., are collected in the $N_T \times N_E$ design matrix $\mathbf{E}$. The trip conditions are assumed to influence the turn penalties in the same way as the travel time rates. Thus, with the $N_E \times 1$ parameter vector $\boldsymbol{\beta}_E$,

$$\mathbf{h} = (1 + \mathbf{o}\boldsymbol{\beta}_o) \cdot \mathbf{E}\boldsymbol{\beta}_E. \tag{3}$$

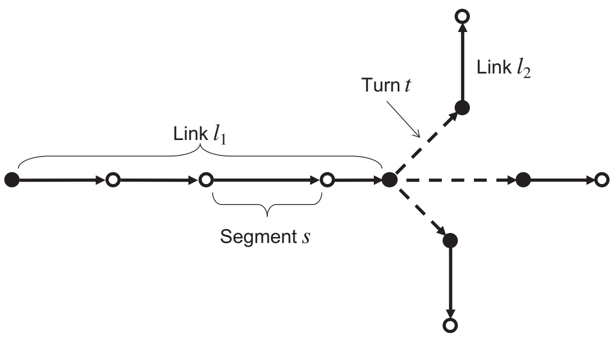

Fig. 1. Illustration of the three types of network components: Links, segments and turns.

The model can be extended to also allow the impact of trip conditions to differ between segments or turns. This would be appropriate, e.g., if network-coded information is used about traffic incidents or construction works that do not cover the whole period of observations, or if temporal variations are known to be more dominant in some parts of the network than others.

2.2. Variance structure

The stochastic components $\boldsymbol{v}$ in (1) represent the variability in segment travel time rates due to unobserved heterogeneity in traveler characteristics, traffic conditions, and local network characteristics. For the purposes of this paper, the error terms are modeled at the link level; travel time rates on different segments in a link are allowed to differ in means but are assumed to have the same variability around the mean, which implies that segment travel time rates are perfectly correlated within the link. This assumption can be relaxed if needed. The stochastic component of link l is denoted u_l. It follows that $v_s = u_{l(s)}$, or in vector notation,

$$\boldsymbol{v} = \mathbf{Su}. \tag{4}$$

In general, the segment-level covariance matrix can be obtained from the link-level stochastic components as

$$\boldsymbol{\Omega} = \mathrm{E}[\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}] = \mathbf{S}\mathrm{E}[\mathbf{u}\mathbf{u}^{\mathrm{T}}]\mathbf{S}^{\mathrm{T}}. \tag{5}$$

An approach from spatial econometrics is used to model the covariance between links (LeSage and Pace, 2009; Cheng et al., 2012). The general approach is to specify the structure for how independent stochastic components $\boldsymbol{\varepsilon}$ for all links interact to determine the total stochastic travel time rate components $\boldsymbol{v}$. In this paper, a *spatial moving average* (SMA) specification is used (Hepple, 2004). In the SMA model the stochastic component u_l of each link is expressed as an independent term ε_l with zero mean and variance σ_l^2, plus a linear combination of the independent components of the other links $\varepsilon_{l'}$, $l' \neq l$. The independent error term ε_l captures the variability in travel time rates that originates on the particular link. Independence implies that $\mathrm{E}[\varepsilon_l\varepsilon_{l'}] = 0$ for $l \neq l'$. The level of influence from link l' on link l, denoted $w_{ll'}$, is specified through the analysis, while the overall magnitude of the covariance is captured by a parameter ρ to be estimated. Assuming that $w_{ll} = 0$ for all l, the total stochastic component of link l is thus,

$$u_l = \varepsilon_l + \rho\sum_{l' \neq l} w_{ll'}\varepsilon_{l'}, \tag{6}$$

or in matrix notation, $\mathbf{u} = (\mathbf{I} + \rho\mathbf{W})\boldsymbol{\varepsilon}$. The $N_L \times N_L$ weight matrix $\mathbf{W}$ with elements $w_{ll'}$ captures the interaction among links. This is important since empirical evidence suggests that the structure of $\mathbf{W}$ must be defined with much care in order to represent the dependencies between link travel time rates properly and allow a meaningful estimation of the correlations between links.

The SMA model can be extended to include multiple weight matrices $\mathbf{W}_i$, $i = 1, \ldots, N_\rho$, with each matrix representing a separate dimension of spatial dependence (Hepple, 2004). This more general model is useful when there are multiple factors that contribute to correlations and are distributed differently in the network. The structure for the stochastic travel time components is then

$$\mathbf{u} = \left(\mathbf{I} + \sum_{i=1}^{N_\rho}\rho_i\mathbf{W}_i\right)\boldsymbol{\varepsilon}. \tag{7}$$

With the (extended) SMA model (7), the covariance between two links l_1 and l_2 is

$$\mathrm{E}[u_{l_1}u_{l_2}] = \sum_{i=1}^{N_\rho}\rho_i(w_{i,l_2,l_1}\sigma_{l_1}^2 + w_{i,l_1,l_2}\sigma_{l_2}^2) + \sum_{i=1}^{N_\rho}\sum_{j=1}^{N_\rho}\rho_i\rho_j\sum_{l' \neq \{l_1,l_2\}} w_{i,l_1,l'}w_{j,l_2,l'}\sigma_{l'}^2. \tag{8}$$

As can be seen, there is a first-order term that arises from the direct influences between the links, and a second-order term that arises from influences through common neighbors in the different dimensions.

Let $\boldsymbol{\sigma}^2$ denote the $N_L \times 1$ vector with elements σ_l^2. Similarly to the means, the travel time rate variances $\boldsymbol{\sigma}^2$ may be decomposed into a number of explanatory factors with associated parameters to be estimated from data. The explanatory variables are divided into two different categories: static link characteristics and dynamic trip conditions.

The link characteristics may include geometric properties and fixed effects for specific links or groups of links. The N_U variance components are represented by the $N_L \times N_U$ design matrix $\mathbf{U}$. The baseline link variances are then $\mathbf{U}\boldsymbol{\sigma}_U^2$, where $\boldsymbol{\sigma}_U^2$ is an $N_U \times 1$ parameter vector to be estimated. Note that the simplest model formulation would be that all links share a single variance parameter σ^2, in which case $\mathbf{U}$ is an $N_L \times 1$ vector of ones. In the other extreme, each link may have an individual variance parameter σ_l^2, in which case $\mathbf{U}$ is the $N_L \times N_L$ identity matrix.

Further, the observed travel conditions for the trip may impact the travel time variances as well as the means. Relevant trip attributes may be similar as for the mean travel time rates, but the impact of each attribute may be different. For a given trip, the attributes are represented by the $1 \times N_p$ design vector $\mathbf{p}$. The multiplier for the trip conditions is then $(1 + \mathbf{p}\boldsymbol{\beta}_p)^2$, where $\boldsymbol{\beta}_p$ is an $N_p \times 1$ parameter vector to be estimated. Note that the parameters actually capture the effect on the square root of the variance, i.e., the standard deviation, which is convenient since it has the same dimension as the mean. The variances of the independent stochastic components ε are thus in total

$$\boldsymbol{\sigma}^2 = (1 + \mathbf{p}\boldsymbol{\beta}_p)^2 \cdot \mathbf{U}\boldsymbol{\sigma}_U^2. \tag{9}$$

Let $\sigma_{U,n}^2$ be the nth link variance component and $\mathbf{U}_n$ the diagonal matrix with the nth column of $\mathbf{U}$ along its diagonal. The covariance matrix of ε is diagonal and given by

$$\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}] = (1 + \mathbf{p}\boldsymbol{\beta}_p)^2 \cdot \sum_{n=1}^{N_U} \sigma_{U,n}^2 \mathbf{U}_n. \tag{10}$$

Combining (5), (7), and (10), the covariance matrix for the network segments becomes

$$\boldsymbol{\Omega} = (1 + \mathbf{p}\boldsymbol{\beta}_p)^2 \cdot \sum_{n=1}^{N_U} \sigma_{U,n}^2 \left(\mathbf{S}\mathbf{U}_n\mathbf{S}^{\mathrm{T}} + \sum_{i=1}^{N_\rho} \rho_i \left(\mathbf{S}\mathbf{W}_i\mathbf{U}_n\mathbf{S}^{\mathrm{T}} + \mathbf{S}\mathbf{U}_n\mathbf{W}_i^{\mathrm{T}}\mathbf{S}^{\mathrm{T}} \right) + \sum_{i=1}^{N_\rho}\sum_{j=1}^{N_\rho} \rho_i\rho_j \mathbf{S}\mathbf{W}_i\mathbf{U}_n\mathbf{W}_j^{\mathrm{T}}\mathbf{S}^{\mathrm{T}} \right). \tag{11}$$

The SMA model (7) does not take into account that there may be dependencies in travel times due to unobserved variations between vehicles or drivers, or perhaps more importantly, between different days. However, it is straightforward to extend the model to capture such dependencies by including a stochastic component at the day or vehicle levels.

3. Observation model

The travel time measurements considered in this paper consist of sparsely sampled vehicle trajectories through the network obtained from GPS devices or similar sensor technologies. In general, GPS location measurements are associated with errors. Here it is assumed that the most likely network location corresponding to each GPS measurement, as well as the path (i.e., the sequence of network segments) taken between each pair of consecutive measurements, have been determined by a map-matching and path inference process (Rahmani and Koutsopoulos, 2013). A basic observation then consists of

1. a vehicle identification number,
2. a pair of time stamps τ_1, τ_2,
3. a path representing the trajectory of the vehicle between the two time stamps, involving a sequence of segments $(s_1, \ldots, s_n)$ and two offsets δ_1, δ_2 specifying the vehicle locations at times τ_1, τ_2 in relation to the upstream nodes of the first and last segments of the path, respectively.

The concepts are illustrated in Fig. 2. The sequence of segments defines the corresponding sequences of links $(l(s_1), \ldots, l(s_n))$ and turns $(t_1, \ldots, t_m)$, respectively, according to the network model.

Let d_s denote the distance traversed on segment s. With ℓ_s denoting the length of segment s,

$$d_s = \begin{cases} \delta_2 - \delta_1 & \text{if } n = 1, \quad s = s_1, \\ \ell_s - \delta_1 & \text{if } n > 1, \quad s = s_1, \\ \ell_s & \text{if } n > 1, \quad s = s_2, \ldots, s_{n-1}, \\ \delta_2 & \text{if } n > 1, \quad s = s_n, \\ 0 & \text{otherwise}. \end{cases} \tag{12}$$

Further, let a_t be equal to 1 if turn t was undertaken and 0 otherwise. The travel time observation $y = \tau_2 - \tau_1$ can then be written as

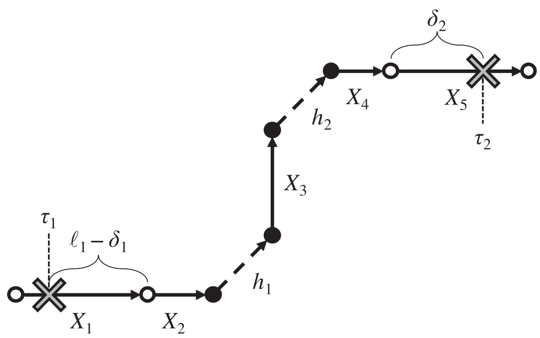

Fig. 2. Illustration of the components of a probe vehicle travel time observation.

$$y = \sum_{s=1}^{N_S} d_s X_s + \sum_{t=1}^{N_T} a_t h_t. \tag{13}$$

Based on this model, a probe vehicle travel time observation is a linear combination of segment travel time rates and turn delays. Furthermore, under the assumption of multivariate normal segment travel time rates and deterministic turn delays, the observed travel time follows a normal distribution. The closed-form distribution of the observed travel time allows the network model parameters to be identified directly through the probe vehicle observations. This attractive property is not preserved if other distributions than the multivariate normal distribution are used. The travel time is given by

$$y = \mu + \eta, \tag{14}$$

where μ is the mean travel time given by

$$\mu = \mathrm{E}[y] = \sum_{s=1}^{N_S} d_s g_s + \sum_{t=1}^{N_T} a_t h_t, \tag{15}$$

that is, the sum of the mean segment travel times plus the turn delays. The zero-mean stochastic term η is normally distributed and given by

$$\eta = \sum_{s=1}^{N_S} d_s v_s = \sum_{l=1}^{N_L} d_l u_l, \tag{16}$$

where $d_l = \sum_s S_{sl} d_s$ is the distance traveled on link l. The variance of η, and hence of observation y, is calculated from the variances and covariances of the traversed links as

$$\mathrm{Var}[\eta] = \sum_{l=1}^{N_L} d_l^2 \sigma_l^2 + 2\sum_{i=1}^{N_\rho} \rho_i \sum_{l_1=1}^{N_L} \sum_{l_2=l_1+1}^{N_L} d_{l_1} d_{l_2} (w_{i,l_2,l_1} \sigma_{l_1}^2 + w_{i,l_1,l_2} \sigma_{l_2}^2) + 2\sum_{i=1}^{N_\rho} \sum_{j=1}^{N_\rho} \rho_i \rho_j \sum_{l_1=1}^{N_L} \sum_{l_2=l_1+1}^{N_L} d_{l_1} d_{l_2} \sum_{l' \neq \{l_1, l_2\}} w_{i,l_1,l'} w_{j,l_2,l'} \sigma_{l'}^2. \tag{17}$$

The first term is the sum of the link travel time variances, which would be the only term if the links were independent. The second term contains the direct dependencies between the traversed links, and the third term contains the second-order dependencies through common neighbors, on or off the traversed path.

Consecutive observations from the same vehicle are correlated through their common links and whenever the links traversed in the different observations are correlated. Incorporating this in the model helps the consistent estimation of the link dependence parameters ρ_i. A *trace* is a contiguous sequence of observations from the same vehicle as it moves through the network. Observations from the same trace are correlated through the link correlation structures, while observations from different traces are considered independent. Fig. 3 illustrates the relationship between observations and traces.

Let index $k = 1, \ldots, N_K$ represent a certain vehicle trace, where N_K (upper-case K) is the number of traces in the data. Index $r = 1, \ldots, N_k$ represents a certain observation in trace k, where N_k is the number of observations in the trace. The total number of observations in the data is $N_R = \sum_{k=1}^{N_K} N_k$. $\mathbf{D}_k$ is an $N_k \times N_S$ matrix where element d_{rs} is the distance traversed on segment s for observation r. $\mathbf{A}_k$ is an $N_k \times N_T$ matrix with element a_{rt} equal to 1 if turn t is made in observation r and 0 otherwise. $\mathbf{y}_k$ is an

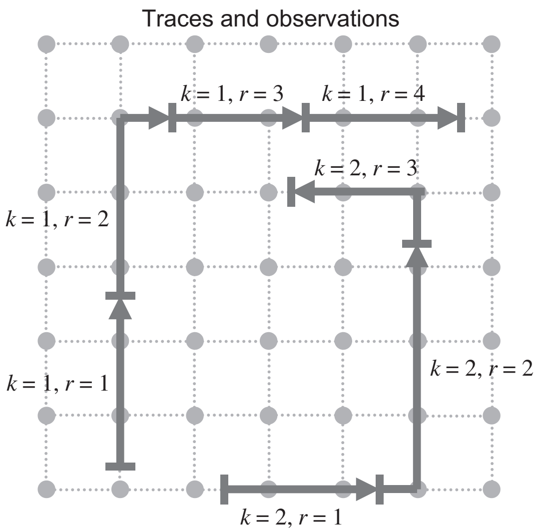

Fig. 3. Illustration of two traces ($k = 1, 2$) each containing four and three observations, respectively.

$N_k \times 1$ vector with element y_r the travel time of observation r (dependent variable). The vector version of (13) is then given by

$$\mathbf{y}_k = \mathbf{D}_k\mathbf{X} + \mathbf{A}_k\mathbf{h}. \tag{18}$$

The travel times $\mathbf{y}_k$ are a linear transformation of multivariate normal stochastic variables and are thus distributed multivariate normal. It follows from (18) that

$$\mathbf{y}_k = \boldsymbol{\mu}_k + \boldsymbol{\eta}_k, \tag{19}$$

where $\boldsymbol{\mu}_k$ is a vector of mean travel times and $\boldsymbol{\eta}_k$ is a vector of correlated zero-mean stochastic terms.

Trip conditions are assumed to be the same for all observations in a trace. Thus, each trace k is associated with two vectors of trip attributes $\mathbf{o}_k$ and $\mathbf{p}_k$ (weather conditions, weekday, season, etc.) affecting the means and variances, respectively. The vector of mean travel times for the trace is then simply the vector form of (15), where the mean segment travel time rates can be expressed in the model parameters using (2) and (3),

$$\boldsymbol{\mu}_k(\boldsymbol{\beta}_B, \boldsymbol{\beta}_E, \boldsymbol{\beta}_o) = \mathrm{E}[\mathbf{y}_k] = (1 + \mathbf{o}_k\boldsymbol{\beta}_o) \cdot (\mathbf{D}_k\mathbf{B}\boldsymbol{\beta}_B + \mathbf{A}_k\mathbf{E}\boldsymbol{\beta}_E). \tag{20}$$

Trip conditions affect all segments and observations uniformly and can be moved outside the other terms. The vector of stochastic terms for the observations in trace k is given by the vector form of (16),

$$\boldsymbol{\eta}_k = \mathbf{D}_k\boldsymbol{\nu} = \mathbf{D}_k\mathbf{S}\mathbf{u}. \tag{21}$$

For two different traces k and k', $\mathrm{E}[\boldsymbol{\eta}_k\boldsymbol{\eta}_{k'}^{\mathrm{T}}] = 0$ by assumption. Within the same trace k, meanwhile, (5), (11), and (21) give the $N_k \times N_k$ covariance matrix

$$\boldsymbol{\Sigma}_k(\boldsymbol{\beta}_p, \boldsymbol{\sigma}_U^2, \boldsymbol{\rho}) = \mathrm{E}[\boldsymbol{\eta}_k\boldsymbol{\eta}_k^{\mathrm{T}}] = (1 + \mathbf{p}_k\boldsymbol{\beta}_p)^2 \cdot \sum_{n=1}^{N_U}\sigma_{U,n}^2\left(\mathbf{D}_k\mathbf{S}\mathbf{U}_n\mathbf{S}^{\mathrm{T}}\mathbf{D}_k^{\mathrm{T}} + \sum_{i=1}^{N_\rho}\rho_i\mathbf{D}_k\mathbf{S}(\mathbf{W}_i\mathbf{U}_n + \mathbf{U}_n\mathbf{W}_i^{\mathrm{T}})\mathbf{S}^{\mathrm{T}}\mathbf{D}_k^{\mathrm{T}} + \sum_{i=1}^{N_\rho}\sum_{j=1}^{N_\rho}\rho_i\rho_j\mathbf{D}_k\mathbf{S}\mathbf{W}_i\mathbf{U}_n\mathbf{W}_j^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{D}_k^{\mathrm{T}}\right). \tag{22}$$

Note that the trip conditions for the trace enter the covariance as the scalar multiplier $(1 + \mathbf{p}\boldsymbol{\beta}_p)^1$ to a baseline covariance matrix. This means that the correlation between two observations depends on link characteristics but not on trip conditions, since the trip conditions factor enters both the covariance and the variances and cancels out.

4. Estimation

4.1. Maximum likelihood formulation

Eqs. (20) and (22) provide the basis for estimation of the network model parameters $\boldsymbol{\beta}_B$, $\boldsymbol{\beta}_E$, $\boldsymbol{\beta}_o$, $\boldsymbol{\beta}_p$, $\boldsymbol{\sigma}_U^2$ and $\boldsymbol{\rho}$ using the probe vehicle travel time observations. The observations are multivariate normal within each trace and assumed independent among traces. Hence, given all observed travel times $\mathbf{y}$, the log-likelihood function is analytical and given by

$$LL(\boldsymbol{\beta}_B, \boldsymbol{\beta}_E, \boldsymbol{\beta}_o, \boldsymbol{\beta}_p, \boldsymbol{\sigma}_U^2, \rho|\mathbf{y}) = -\frac{1}{2}N_R\log(2\pi) - \frac{1}{2}\sum_{k=1}^{N_K}(\mathbf{y}_k - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_k - \boldsymbol{\mu}_k) - \frac{1}{2}\sum_{k=1}^{N_K}\log|\boldsymbol{\Sigma}_k|. \tag{23}$$

The computation times for parameter estimation can be reduced considerably by pre-computing factors that do not depend on the parameter values outside of the optimization routine. For the mean structure, we note from (20) that the mean vector for trace k can be written as

$$\boldsymbol{\mu}_k = (1 + \mathbf{o}_k\boldsymbol{\beta}_o) \cdot (\mathbf{D}_k\mathbf{B}, \mathbf{A}_k\mathbf{E})(\boldsymbol{\beta}_B^{\mathrm{T}}, \boldsymbol{\beta}_E^{\mathrm{T}})^{\mathrm{T}}. \tag{24}$$

The parameters for the mean segment travel time rates and the turn delays are thus merged into a single parameter vector. Here $(\mathbf{D}_k\mathbf{B}, \mathbf{A}_k\mathbf{E})$ is an $N_k \times (N_B + N_E)$ matrix that is independent of the model parameters and may be computed and stored for all traces prior to the optimization. For the covariance structure, we note from (22) that the model parameters are multipliers to constant matrices that may be pre-computed and stored for every combination of parameters.

Matrix inversion is a costly operation, even if techniques such as LU or Cholesky factorization are used. The SMA model has the attractive property that it is not necessary to invert the full $N_L \times N_L$ link-level covariance matrix to compute the likelihood function, but only the $N_k \times N_k$ covariance matrices for the individual traces. Since the number of observations in a trace is typically much lower than the number of links in the network, this means that the SMA formulation requires significantly less computational effort to estimate. In addition to the theoretical considerations discussed in Section 2 this is another reason why the SMA formulation is chosen for the model.

The sampling of the vehicle trajectories can be interpreted as a linear projection from the space of segment travel time rates and turn penalties to the space of observed travel times. The dimension of the observation space depends on the data: the dimension increases with the sampling frequency and with the number of observations, assuming that the vehicle trajectories are sampled at random locations. The network model represents another projection from the network components to the parameter space. The identification of the model parameters depends on the relative dimensions of the parameter space and the observation space: a larger number of observations allows for the estimation of a richer model. This determines, for example, to what extent fixed effects for specific segments or groups of segments can be included in the model explicitly.

4.2. Network delimitations

One may often be interested in estimating travel times in a subnetwork, here referred to as the *primary* network, which is smaller than the network spanned by the available GPS probes. If the primary network is small, for example a single street or road, the number of observations that traverse only segments in the primary network may be insufficient to estimate the travel times reliably. There may also be a bias since short traversed distances may be over-represented. Rather, it is important to utilize all observations that to some extent traverse at least one of the segments in the primary network (referred to as *primary* observations).

With low frequency probe vehicle data, however, the primary observations will involve traversals of many segments and transitions also outside the primary network. These components, which depend on the used data, constitute the *secondary* network. If the primary network is small, the size of the secondary network can be many times greater. Once the secondary network has been identified, all observations that only traverse the secondary network, that is, do not extend the number of segments and transitions in the estimation any further, are used. These observations are referred to as *secondary* observations, and may be many times greater in number than the primary observations. The concepts are illustrated in Fig. 4.

4.3. Spatial clustering of network links

Sparse probe vehicle data may not have the resolution required to identify the travel time rate on all individual segments; indeed, this is the case in the application presented in Section 5. The use of explanatory variables can reduce the dimensionality of the problem. Still, it is very likely in practice that there are systematic differences in segment travel time rates between different parts of the network not captured by observed segment attributes. Spatial clustering of segments provides a compromise between the two extremes of fixed travel time rate effects for each segment (link) and a single baseline travel time rate for all segments.

As a first step, segments (links) can be grouped together into classes expected to have similar traffic performance characteristics. All links in the same class share the same functional form that characterizes their performance. Such a process can be used for grouping the links that belong to the primary network, since they are of the most interest. The secondary network only plays an auxiliary role supporting the estimation of the main links. For the links in the secondary network the following process is used that applies spatial criteria to automate the grouping process. The algorithm operates at the link level, so that all segments in the same link always belong to the same group.

1. *Initialization*
   - Assign each link to its own group
   - Select a min, max, and target value for the number of links and the number of observations in each group. The choice of these parameters should provide a good balance between robust estimation and model resolution.
2. *Iteration*
   - Select the group with the smallest number of observations (in case of ties, an arbitrary group is chosen).

     For every link in the current group
     - Check whether it is connected with some link in another group through a common node. All such identified adjacent groups are added to a list.
     - Rank the groups in the list in increasing number of observations associated with each group

     Merge the current group with the first adjacent group for which any of two conditions hold:
     - The total number of links and the total number of observations in the two groups do not exceed the corresponding target values,
     - The number of links or the number of observations in any of the groups is less than the minimum value, and the total number of links and the total number of observations in the two groups do not exceed the maximum values.
   - Repeat with next group
3. *Termination*
   - STOP if no groups can be merged

The algorithm aims at grouping links to balance the number of observations within each group (and therefore support estimation of the parameters of the corresponding distribution of travel time rates). Any constraint can be made non-binding by setting the corresponding threshold value sufficiently low or high.

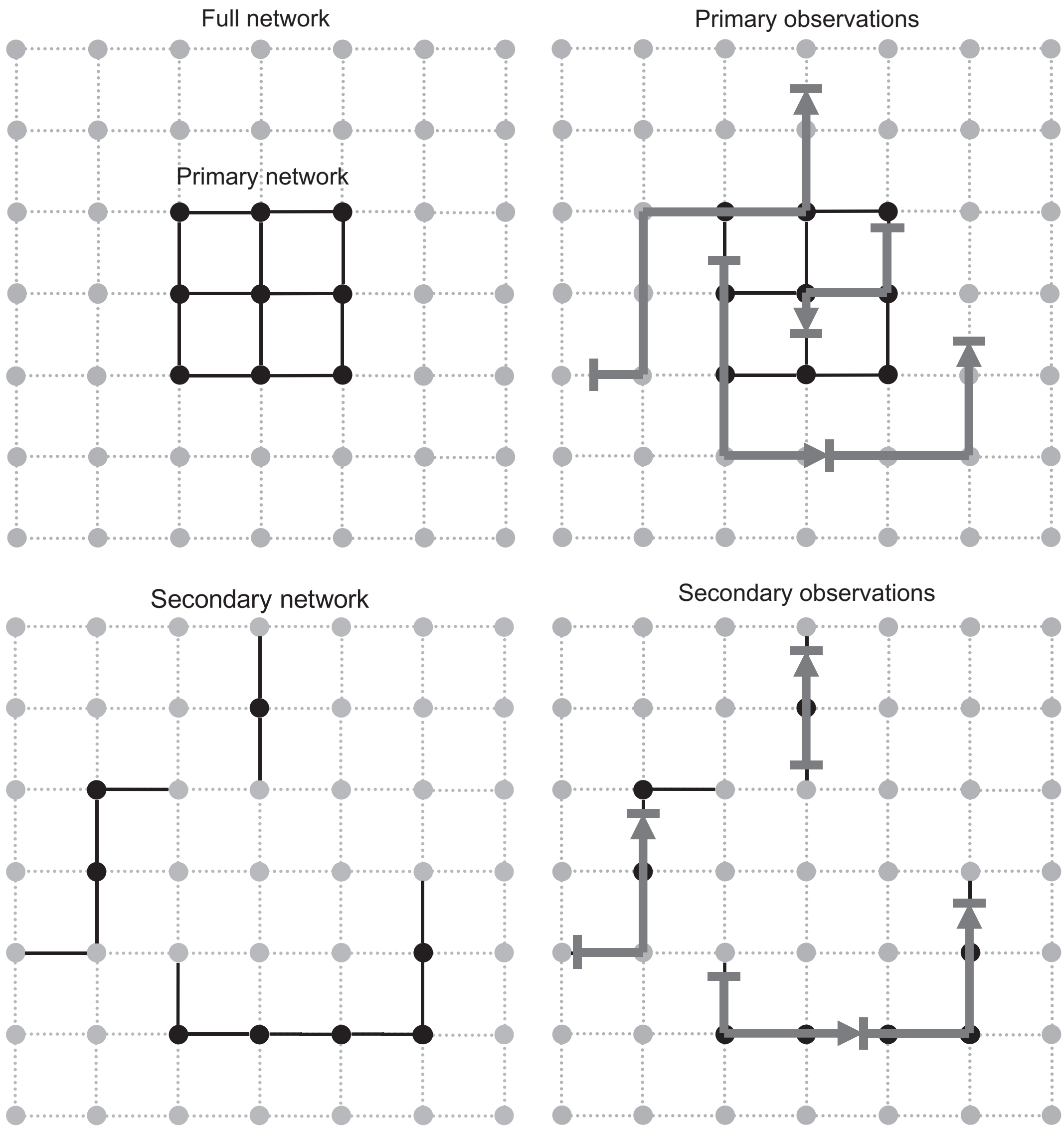

Fig. 4. Illustration of the concepts of the primary and the secondary network and observations.

The output of the algorithm can be summarized as a link-group incidence matrix **C**, where element C_{lc} is equal to 1 if link l belongs to group c and zero otherwise. Since each link can only belong to one group we must have $\sum_c C_{lc} = 1$ for all l. The mapping of segments to groups is then obtained as the composite projection **SC**.

5. Case study

The methodology proposed in Sections 3 and 4 is applied in the urban network of Stockholm, Sweden. The primary network consists of a route along one of the major inner city streets, the southern half of Birger Jarlsgatan, southbound direction, shown in Fig. 5. The shaded area represents the primary network. The secondary network extends, in some cases, even beyond the shown area in Fig. 5. The main route is chosen to coincide as closely as possible with a pair of automatic number plate recognition (ANPR) sensors mounted at each end of the route (see further Section 5.5 and Kazagli and Koutsopoulos, 2012).

The main route is about 1.4 km long and contains 28 links, divided into 36 segments, 26 intersections, and 10 traffic signals; the speed limit is constant at 50 km/h. There is a busy commercial and entertainment center in the middle of the route with a nearby taxi station, where it is expected that the mean travel time rates are higher than in adjacent parts, in particular

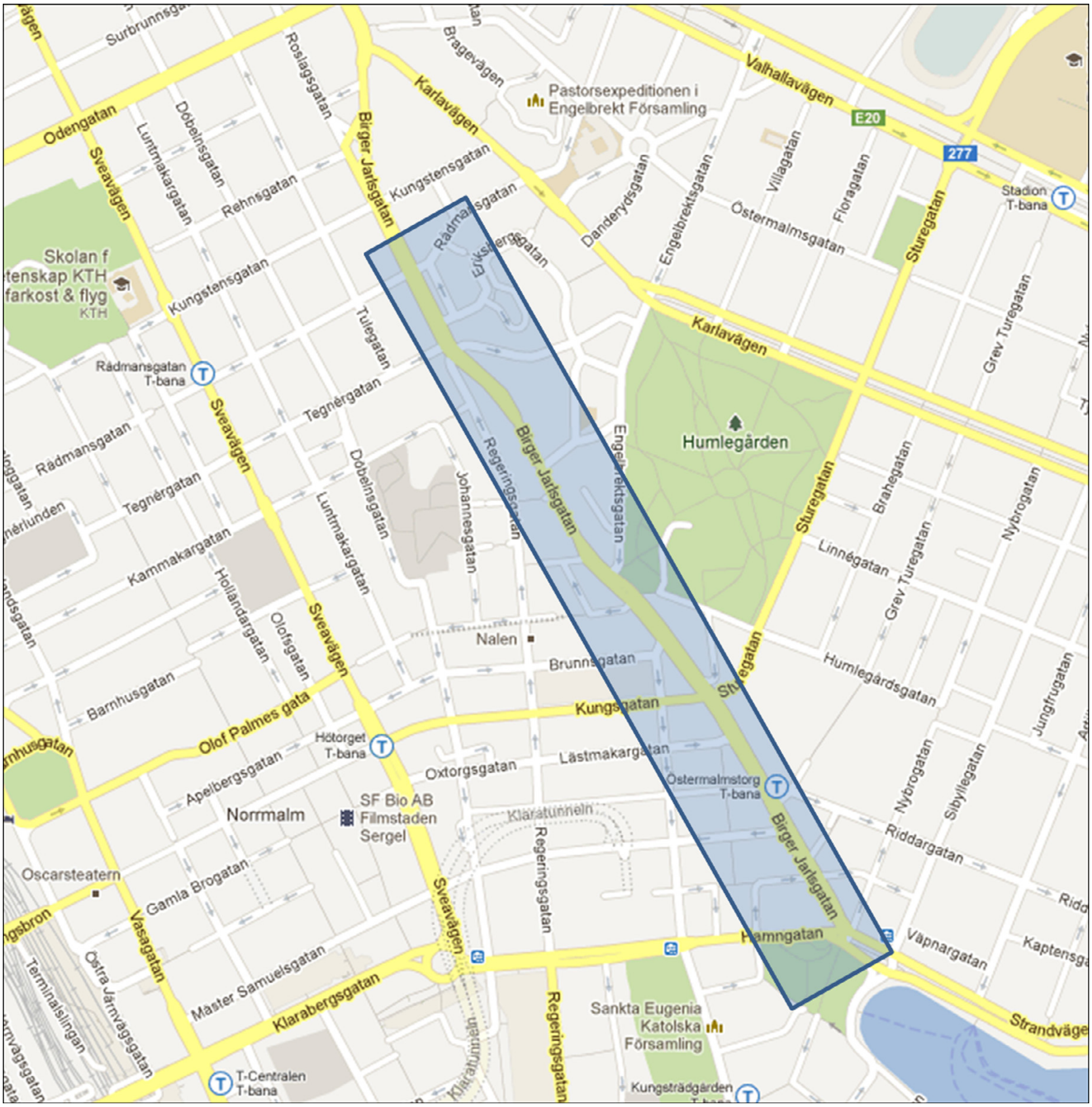

Fig. 5. The case study area in Stockholm, Sweden (Birger Jarlsgatan, southbound). Shaded area shows the primary network. Map data ©Google 2012.

for taxis. The route ends with a complicated signalized intersection in the south where delays can be significant. On the second half of the route there is a bus lane that taxis are allowed to use.

Initially, different specifications of the model structure are considered. The focus is not to derive the best model specification possible, but to demonstrate the structure of the model and the impact and significance of different explanatory factors on the observed travel times during a specific time interval (7:30–8:00 AM). The estimated travel time for the main route under different trip conditions is also evaluated. In the second part, the estimated route travel time is compared with observed travel times from the ANPR sensors.

5.1. Data

The GPS probe vehicle data are obtained from the fleet dispatching system of a taxi company operating in total about 1500 vehicles in the Stockholm network (Rahmani and Koutsopoulos, 2013). The average sampling frequency is about one report per 2 min and 780 m, which is considerably lower than in most previously reported studies (e.g., one per minute in Hunter et al., 2009; Hofleitner et al., 2012a; one per 200 m in Westgate et al., in press).

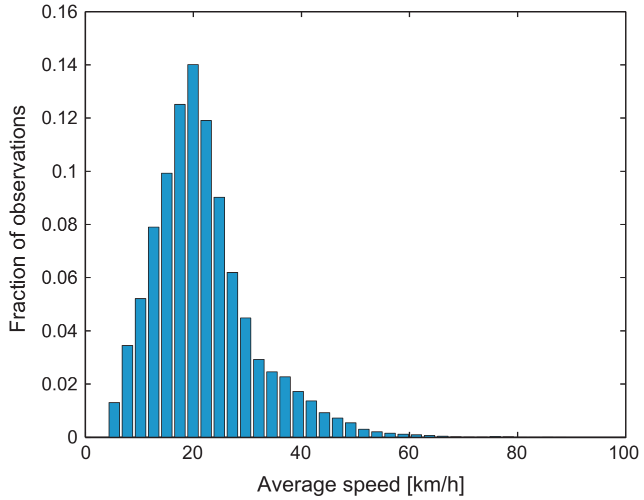

Fig. 6. Distribution of average speeds among the observations (7:30–8:00 AM).

The digital network representation contains information about various geometric attributes, including segment speed limit, functional class (a five-level hierarchical classification of the segments with 1 denoting highways and 5 denoting the most peripheral side streets), traffic signals, and facility type (ramps, tunnels, roundabouts, etc.). The network model is also used to identify intersections, left and right turns, and one-way streets.

Data for weekdays (Monday to Friday) and the time interval 7:30–8:00 AM between January 1, 2010 and December 31, 2011 were used. For this two-year period, 63,680 observations in 44,844 traces are available after filtering. Of these, 10,604 observations are primary (that is, they cover the main route to some extent) and 53,076 secondary (covering only the surrounding network). Across the primary and the secondary networks, the observations cover in total 1300 segments, 832 links and 1373 turns. On average, each observation covers 14.0 segments, 8.6 links and 7.7 turns. Fig. 6 shows the histogram of the average vehicle speed for each observation, calculated as the traversed (actual) distance divided by the time between the reports. The speed distribution has mean 21.8 km/h, median 20.4 km/h and standard deviation 9.4 km/h.

Further, historical weather data for the same period as the probe vehicle data were collected from a weather station in the area, reporting every 20–30 min. The data includes temperature, qualitative precipitation information (light/heavy rain/snow, etc.), and other related information.

Observations of link flows were not available for this study. Several network attributes that potentially have significant impact on travel times were also not available, including the number of lanes, locations of bus stops, pedestrian crossings, on-street parking, stop signs, traffic signal cycle times, nearby land use, etc. Information about the time and location for incidents and road works were also not available. If available, inclusion of this type of information in the model is straightforward.

5.2. Model specification

5.2.1. Specification of link groups

The spatial and temporal coverage of the available probe vehicle data is not high enough to identify and reliably estimate all segment specific parameters (for example, because of the sparse nature of the GPS probes). The number of observations covering each segment is generally lower in the secondary network than in the primary network. The network links are therefore partitioned into groups using the manual approach for the main route combined with the automated method described in Section 4.3 for the secondary network. The main route is divided into seven groups, each containing four links and 5.1 segments on average. The number of groups is selected so that fixed effects for the mean travel time rate of the segments in each group can be identified through the data, and the boundaries between groups are selected to capture the hypothesized heterogeneity of traffic conditions along the route.

The secondary network is partitioned with the algorithm in Section 4.3 using the following threshold values which were found to produce good results with the given dataset. For the number of observations covering each group, the minimum value is set to 300 and the target and maximum values are set to infinity. For the number of links in each group, the minimum, target and maximum values are set to 10, 15 and 38, respectively. This produces 47 link groups for the secondary network, each containing 17.1 links and 26.9 segments on average. Together with the seven link groups for the primary network, there are 54 link groups in total.

5.2.2. Specification of spatial weights

For the stochastic components of the travel time rates an SMA structure as described in Section 2.2 is used, with a single weight matrix $\mathbf{W}$ and associated parameter ρ. The correlation between the travel time rates of consecutive observations within vehicle traces in the dataset was about 6% and statistically significant, suggesting that, in general, there should be positive correlation between the network links. The structure of **W** was chosen after extensive experimentation with alternative specifications. Each link is assumed to be influenced by its nearest upstream neighbors in the network as well as the nearest upstream neighbors of those links (first and second-order neighbors, respectively). The second-order neighbors are given a weight of 0.25 relative to the first-order neighbors to represent the fact that influence should decay with distance. Finally, each row of **W** is normalized to add to 1. This normalization is the standard approach in spatial econometrics (LeSage and Pace, 2009), and was found to improve the model fit.

In order to investigate the effect and potential of different types of variables, four different specifications are considered.

5.2.3. Model 1: Link groups

In the first specification (Model 1), the mean segment travel time rates are explained only with fixed effects at the link group level (54 groups in total). The mapping of every segment to its assigned group thus forms the design matrix **B** = **SC**. Explanatory variables for segment attributes, trip conditions or turn penalties are not included. For the stochastic components, for the sake of simplicity, a formulation in which the variance parameter is common for all links is used.

5.2.4. Model 2: Segment and turn attributes

In the second specification (Model 2), explanatory variables for segment and turn attributes, but no variables for trip conditions, are added. This is a relevant level of specification for strategic applications, where typical travel times across different travel conditions are of interest in modeling and forecasting. The following variables are added, based on their causal effects and their significance in the estimation.

Segment attributes (matrix **B**):

- Dummy variables for every combination of segment speed limit (in km/h) and segment functional class (abbreviated FC) in the data. The combinations (50,1), (50,2) and (50,3) are grouped and used as reference level since the number of observations of links with combinations (50,1) or (50,2) are too few to estimate the effects separately. The other existing combinations of speed limit and functional class are (10,5), (30,3), (30,4), (30,5), (50,4) and (50,5).
- A dummy variable for the existence of a taxi station on the segment. The hypothesis is that the mean travel time rates for taxis on such segments are higher due to many stopping there.
- Dummy variables for the length of the link to which the segment belongs: length <50 m and length = 200 m, with the reference level length $\in$ [50, 200) m. The hypothesis is that vehicles have less chance to reach high speeds on short links, since the speed is typically lower close to intersections and traffic signals. Hence, short links should have higher mean travel time rates.

Turn attributes (matrix **D**):

- Three dummy variables for the cases that the turn involves a signalized left turn, right turn and straight through movement, respectively.
- Two dummy variables for the cases that the turn involves a non-signalized left turn and a right turn, respectively.

Through movements are used as reference level for all turn delays. The hypothesis is that straight-through movements in general have priority in intersections so that left or right turns lead to longer delays.

5.2.5. Model 3: Trip conditions

In the third specification (Model 3), explanatory variables for the trip conditions influencing travel time rates and turn delays are introduced. For the travel time means (vector **o**) the following variables are considered:

- A dummy variable for the late half of the time interval, i.e., 7:45–8:00 (not July or public holiday). This allows the estimation of travel time variations within the interval.
- One dummy variable for each weekday from Tuesday to Thursday (reference level is Monday and Friday).
- A dummy variable for public holidays on weekdays (Christmas, Easter, etc.)
- Dummy variables for the summer (June–August) and winter (January–February) seasons. The reference level is spring (March–May) and fall (September-December).
- A dummy variable for the month of July, the peak vacation period in Sweden. The total effect of July is obtained by adding the effect of the summer season (June–August).
- A dummy variable for the year 2011 (reference level is 2010). This accounts for changes in traffic levels between years.
- Recent snow: Number of hours of consecutive snowfall reports preceding the trip.
- A dummy variable for the taxi being free, as opposed to occupied or assigned to a customer.

For the travel time rate standard deviations (vector $\mathbf{p}$) the following variables are used:

- A dummy variable for the late half of the time interval, i.e., 7:45–8:00 (not July or public holiday). This allows the estimation of travel time variations within the time period.
- One dummy variable for each weekday from Tuesday to Friday (reference level is Monday).
- Dummy variables for the summer (June–August) season.
- A dummy variable for the month of July, the peak vacation period in Sweden. The total effect of July is obtained by adding the effect of the summer season (June–August).
- A dummy variable for the taxi being free, as opposed to occupied or assigned to a customer.

5.2.6. Model 4: Independent links

The fourth specification (Model 4) is identical to Model 3 except that all links are assumed to be independent; in other words, the dependence parameter ρ is omitted. By comparing the performance of Model 3 and Model 4, the contribution of the correlation structure in the model can be assessed.

Table 1
Travel time rate model parameter estimation results (standard errors in parenthesis).

Parameters	Model 1		Model 2		Model 3		Model 4	
Mean segment travel time rates ($\mathbf{B}$) [s/m]								
54 segment groups	Yes		Yes		Yes		Yes	
Speed limit 10, functional class 5	–		0.3869	(0.0243)	0.3668	(0.0233)	0.3684	(0.0235)
Speed limit 30, functional class 3	–		0.0285	(0.0025)	0.0283	(0.0024)	0.0285	(0.0024)
Speed limit 30, functional class 4	–		0.0579	(0.0031)	0.0550	(0.0030)	0.0544	(0.0030)
Speed limit 30, functional class 5	–		0.0975	(0.0025)	0.0936	(0.0024)	0.0930	(0.0024)
Speed limit 50, functional class 4	–		0.0777	(0.0026)	0.0748	(0.0025)	0.0747	(0.0025)
Speed limit 50, functional class 5	–		0.0507	(0.0063)	0.0496	(0.0060)	0.0490	(0.0060)
Taxi station	–		0.0830	(0.0087)	0.0794	(0.0084)	0.0789	(0.0084)
Link length < 50 m	–		0.0144	(0.0053)	0.0147	(0.0050)	0.0162	(0.0050)
Link length $\geqslant$ 200 m	–		−0.0500	(0.0029)	−0.0488	(0.0028)	-0.0504	(0.0028)
Turn penalties ($\mathbf{D}$) [s]								
Traffic signal, left turn	–		10.3225	(0.5118)	9.8008	(0.4931)	9.5615	(0.4902)
Traffic signal, right turn	–		7.2435	(0.4486)	6.9870	(0.4303)	6.8669	(0.4295)
Traffic signal, straight through	–		2.9030	(0.2359)	2.8125	(0.2259)	2.6814	(0.2238)
Non-signalized left turn	–		5.7913	(0.4215)	5.7915	(0.4026)	5.7113	(0.4016)
Non-signalized right turn	–		5.4097	(0.4481)	5.1150	(0.4280)	5.0733	(0.4243)
Mean trip conditions ($\mathbf{o}$) [%]								
7:45–8:00 (not July, not holiday)	–		–		5.0458	(0.3315)	5.0401	(0.3302)
Tuesday	–		–		2.0399	(0.4287)	2.0466	(0.4271)
Wednesday	–		–		1.8665	(0.4367)	1.9024	(0.4349)
Thursday	–		–		0.8671	(0.4127)	0.9065	(0.4109)
Public holiday	–		–		−14.2244	(1.4531)	−14.2385	(1.4474)
Winter (January–February)	–		–		2.1370	(0.5064)	2.1567	(0.5043)
Summer (June–August)	–		–		−4.1027	(0.3966)	−4.1210	(0.3948)
July (vacation month)	–		–		−5.9422	(0.6524)	−5.8928	(0.6499)
2011	–		–		1.3469	(0.3131)	1.3708	(0.3118)
Recent snow [% h^{-1}]	–		–		0.2390	(0.0477)	0.2396	(0.0475)
Free taxi	–		–		3.9911	(0.3863)	4.0177	(0.3847)
Standard deviation trip conditions ($\mathbf{p}$) [%]								
7:45–8:00 (not July, not holiday)	–		–		6.0935	(0.6902)	6.1497	(0.6914)
Tuesday	–		–		4.4533	(1.0073)	4.6261	(1.0097)
Wednesday	–		–		5.4067	(1.0316)	5.5028	(1.0336)
Thursday	–		–		4.8131	(0.9950)	4.9069	(0.9969)
Friday	–		–		5.7447	(0.9915)	5.8010	(0.9931)
Summer (June–August)	–		–		−3.9381	(0.7656)	−3.9298	(0.7671)
July (vacation month)	–		–		−4.5229	(1.2957)	−4.4054	(1.2981)
Free taxi	–		–		8.4385	(0.7184)	8.6944	(0.7192)
Travel time rate variance (σ^2) [$(s/m)^2$]	0.0290	(0.0006)	0.0282	(0.0006)	0.0245	(0.0006)	0.0265	(0.0004)
Travel time rate correlation (ρ)	0.2519	(0.0313)	0.1821	(0.0293)	0.1330	(0.0289)	–	
Log likelihood	−261,982		−259,670		−259,004		−259,016	
AIC	524,076		519,480		518,186		518,208	
Log likelihood(0)	−265,245		−265,245		−265,245		−265,245	
Observations	63,680		63,680		6,3680		63,680	
Variables	56		70		89		88	

5.3. Estimation results

The model specifications were estimated using the maximum likelihood estimation routine in the Statistics Toolbox for MATLAB and a trust-region reflective Newton optimization algorithm (MATLAB, 2009). Gradients of the log-likelihood function were evaluated analytically, while the Hessian used to calculate standard errors was inverted numerically.

Estimation results are shown in Table 1. The 54 segment group fixed effects, not shown in the table, are all significant with p-values less than 0.0001 in all four model specifications. This demonstrates that the automated grouping method is capable of handling the identification problems associated with the low sampling frequency, even in the outer parts of the secondary network with few observations per segment. The explanatory power of Model 1, captured by the log-likelihood and Akaike's information criterion (AIC), is considerably better than a model (''Model 0'') with a single mean travel time rate parameter and a single travel time rate variance parameter (the log-likelihood for this model is −265,745, and the AIC is 531,494). The logarithm of the likelihood ratio is −3763, and a likelihood ratio test with 54 degrees of freedom rejects Model 0 in favor of Model 1 with a p-value equal to 0 within machine precision.

The fit of the model increases when segment and turn attributes are added (Model 2 compared to Model 1). The logarithm of the likelihood ratio is −2312, and a likelihood ratio test with 14 degrees of freedom rejects Model 1 in favor of Model 2 with a p-value equal to 0 within machine precision. This result is encouraging as it suggests that the low sampling frequency and identification power of the observations can be counter-balanced, to some extent, by making use of attributes of the network components. As expected, speed limits and functional classes have a strong impact on travel time rates. The directions and magnitudes are also intuitive: all else equal, a lower speed limit increases the travel time rate, as does a higher functional class (i.e., a lower hierarchical level).

Travel time rates are significantly higher on segments with taxi stations, no doubt due to many taxis stopping there to wait for customers. Therefore, the corresponding variable can be used for estimating travel times for personal trips by controlling for the bias in the taxi data. Further, travel time rates are higher on shorter links and lower on longer links, reflecting the effect of acceleration and deceleration as hypothesized.

Traffic signals impose delays depending on the direction of movement: about ten seconds for left turns, seven seconds for right turns and three seconds for through movements. A non-signalized left turn or right turn gives a delay of about 5 s, shorter than for signalized turns. Recall that these are average delays for all vehicles passing the traffic signals and intersections, whether they need to break or not. However, the delays are probably somewhat underestimated (and segment running travel times correspondingly overestimated), since the delay caused by the traffic signal or intersection may be distributed along the preceding link(s) due to queues, etc., so that a vehicle sending a report just before reaching a turn as specified in the network model may already have experienced some of the associated delay.

The explanatory variables for the conditions of the trip improve the fit of the model further (Model 3 against Model 2). The logarithm of the likelihood ratio is −666; again, a likelihood ratio test with 19 degrees of freedom rejects Model 2 in favor of Model 3 with a p-value equal to 0 within machine precision. Mean travel time rates are significantly higher in the late half 7:45–8:00 of the period (+5%), suggesting a build-up of the morning peak compared to 7:30–7:45. There is also variation across the week: travel time rates are lowest on Mondays and Fridays and highest on Tuesdays (2% higher), after which they decay towards the weekend. Travel time rates are lower during public holidays, dropping by 14%. The summer season and in particular the peak vacation month of July are also associated with big reductions in travel time rates (−4% and −10%, respectively), whereas the winter season is associated with higher travel time rates (+2%). The latter result is more likely an effect of weather conditions (although recent snowfall is considered separately below) than increased traffic since demand levels in Stockholm are generally low in the winter. Interestingly, there is a small but significant increase in travel time rates from 2010 to 2011 of about 1.3%. This may be indicative of a long-term increase of congestion in the inner city of Stockholm.

Recent snowfall significantly increases mean travel time rates. The estimate suggests that every four hours of consecutive snowfall before the trip increases travel time rates by about 1%. This effect is expected since snow on the ground makes driving more difficult. Meanwhile, no significant impact of concurrent or recent rainfall on travel times was found. It is quite possible that further analysis of the weather data could lead to more refined insights into the impact of weather conditions (including for example combined effects of precipitation and temperature changes) on travel time rates.

Travel time rates are higher when the taxi is free (+4%), most likely due to the fact that they may be cruising for customers at low speeds. When travel times are predicted for personal cars it may be appropriate to set this variable to zero, or estimate the model only on data from hired taxis.

The standard deviation is higher (+6%) in the late half of the period. Thus, both the average and the variability of the travel time rates increase, which is the expected effect of increasing congestion. Interestingly, the variability on weekdays other than Mondays increases by roughly the same factor (+5%). Summer and vacation days have lower standard deviations as expected due to less congestion.

The dependence parameter ρ for the SMA structure is positive and statistically significant, indicating positive correlation between network link travel time rates. For the main route, the estimated parameter value in Model 3 translates to travel time correlations between links up to about 10%, which is quite low. Treating links as independent reduces the fit of the model only moderately (Model 4 vs. Model 3). This would seem to suggest that correlations between network components are not vital to consider when estimating travel times. However, more research is needed to determine the characteristics of inter-link correlation and appropriate structures for the spatial weight matrices.

5.4. Analysis of results

Using the estimation results of the previous section, the travel time on the main route was estimated as the sum of the travel times on the segments corresponding to the route. Fig. 7 illustrates how the estimated mean and variability vary under different combinations of trip conditions using the specification of Model 3. The reference, baseline conditions are those of the period 7:30–7:45 on a non-holiday Monday in spring/fall 2010 with no recent snow. As the figure shows, the estimated mean travel time may vary up to a minute depending on the travel conditions considered. Also, the variations in travel time may be very small but still statistically significant, such as the increase in travel time between 2010 and 2011.

Located at both ends of the main route are two ANPR cameras that record license plate numbers and time stamps when vehicles pass. The travel time of a vehicle on the route is calculated by taking the difference between the time stamps as the vehicle passes the two ANPR sensors. Note that the cameras capture all vehicles, including personal cars, taxis, trucks, buses, etc. Travel time data from the ANPR sensors at the individual vehicle level were available for the period August 15, 2011–April 12, 2012.

For a number of reasons, the ANPR data cannot be regarded as "ground truth" data for the taxi probe vehicle data. First, the ANPR data is noisy and a considerable number of observations has to be filtered out as outliers due to vehicles stopping along the route, taking detours, mismatched license plates, etc. (Kazagli and Koutsopoulos, 2012). Second, it is not known exactly at which points the ANPR sensors register the vehicles; it may even vary between vehicles depending on when the cameras are able to read the license plates. This means that there may be differences between the length of the taxi route and the ANPR route. Third and most important, it is generally acknowledged that taxis behave systematically different from overall traffic: they may drive faster than the average vehicle when occupied and far from the destination, whereas they stop more frequently and drive slower than average when picking up or dropping off customers. Hence, comparison of the estimated travel time against the ANPR data should not be considered as validation in any strict sense.

The distribution of travel times obtained from the ANPR data is shown in Fig. 8. As can be seen, the distribution is unimodal and slightly skewed to the right. The mean travel time is 3.52 min, the median 3.47 min, and the standard deviation 0.63 min.

The ANPR observations were compared to the main route travel time estimates using the specification and parameter values of Model 3 shown in Table 1. Using the estimated model, the main route travel time distribution, i.e., mean and standard deviation, was calculated under the trip conditions of each ANPR observation (weekday, season, year, etc.). Analysis showed occupied taxis to be more similar to the overall traffic than free taxis, and the "free taxi" dummy was set to 0 for all observations. A mixture distribution was then generated over all observation-specific distributions, with equal weights for all observations. The final distribution is thus a mixture of normal distributions and represents the travel time distribution under the varying trip conditions captured in the ANPR data. The main route travel time has an estimated mean of 4.04 min and

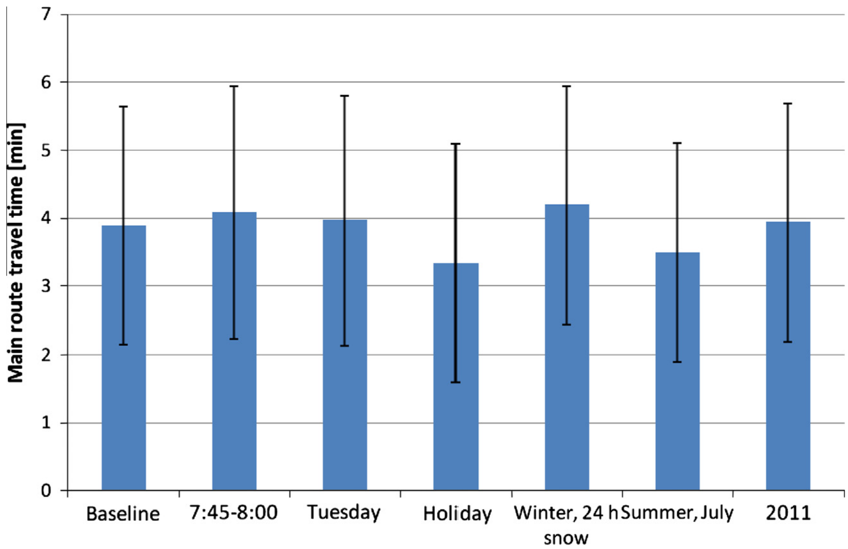

Fig. 7. Estimated impact of trip conditions on the travel time for the main route (Birger Jarlsgatan southbound, 7:30–8:00). The filled bars show the mean travel time, the error bars show 95% confidence intervals.

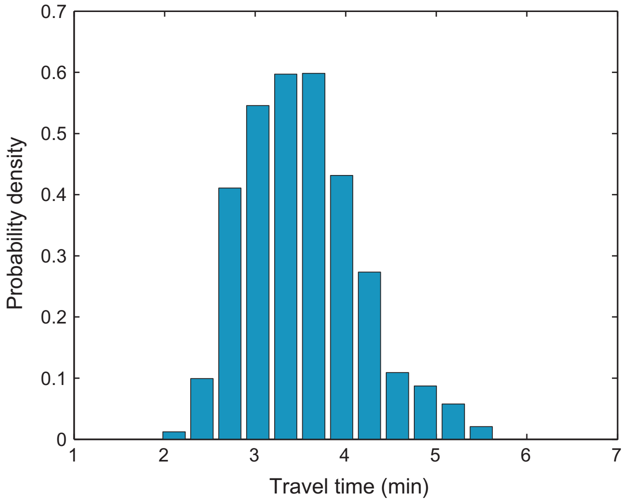

Fig. 8. Empirical distribution of travel times for the main route (Birger Jarlsgatan southbound) from ANPR data.

a standard deviation of 0.95 min. The traffic signals and intersections along the route constitute 23% of the mean travel time, while the remaining 77% is the running time on the links.

Hence, the model estimates of the mean and standard deviation using the probe vehicle data are higher compared to the ANPR data. The higher mean may be caused by a combination of the issues mentioned above: queuing delays partially included in link travel times, mismatching definitions of the start and end points of the main route, filtering of outliers, and systematic differences between taxis and the overall traffic (especially considering the land use characteristics of the corridor). The higher standard deviation may be further caused by uncertainty in GPS locations and path choices. Also, the specification of the variance used here is simple with only a single variance parameter for all links; it is possible that refining the estimate by adding link groups and explanatory variables for the variance may reduce the variance for the main route travel time.

6. Conclusion

The paper presented a statistical model for urban road network travel time estimation based on low frequency GPS probe vehicle data. Network travel times are modeled as running travel times on links and turn delay at traffic signals and intersections. To analyze the effects of network characteristics and trip conditions (time of day, season, weather conditions, etc.), the mean and variance of the link travel times and the turn delays are expressed as functions of explanatory variables in combination with fixed effects for groups of segments. This approach also allows the model to handle the low resolution of the probe data. The network model allows correlation between travel times on different network links according to a spatial moving average (SMA) structure. The observation model presents a way to estimate the parameters of the network model through low frequency samplings of vehicle traces, including the correlation as experienced by a driver traversing the links sequentially. The methodology can be used for the development of databases of historical travel times.

Conditional on trip conditions (weather, accidents, construction work, etc.), the network model assumes that link travel times follow a multivariate normal distribution. Some empirical studies have found the normal distribution to be a good approximation (Rakha et al., 2006), while other studies have found more skewed distributions to provide better fit (e.g., Uno et al., 2009; Fosgerau and Fukuda, 2012). Observations collected over an extended period of time (as is often the case with empirical travel time distributions shown in the literature), however, may encompass a range of different trip conditions. In the model, the marginal distribution across different conditions is a mixture of normal distributions and may have a shape very different from a normal distribution (Park et al., 2010). To some extent the model is thus capable of representing skewed travel times distributions.

The model was applied to a part of the arterial network in Stockholm, Sweden. Attributes such as speed limits, functional classes, one-way streets and signalized or non-signalized left turns and right turns have significant effects on travel times. Trip conditions such as time-of-day, weekday, public holiday, vacation, year, recent snowfall and recent rainfall also have significant effect on both the mean and in some cases the variability of travel times. Further, there is positive correlation between links.

The case study highlights the potential of using sparse probe vehicle data to monitor the urban road transport system and identify changes in travel times and speeds, without the availability of classical traffic data such as flows. Even small variations in travel times between days, seasons and years, can be identified with statistical precision, which suggests that opportunistic, sparse probe vehicle data can provide a cost-effective way of assessing the impacts of management actions, policy instruments and investments on traffic conditions. The increasing availability of such data also opens up the possibility for new types of control strategies; for example, congestion charges could be tied directly to a congestion index that is calculated from the estimated travel times and updated at suitable intervals.

The comparison with ANPR data recording travel times for all vehicles (personal cars, trucks, etc.), however, suggests that there may be systematic differences in traffic behavior between opportunistic probe vehicles, in this case taxis, and the overall vehicle fleet. These deviations may vary depending on the characteristics of the network and the source of data. In order to determine the best utilization of opportunistic probe vehicle data, the similarities and differences compared to the overall traffic need to be investigated carefully. Further research is needed to determine the extent to which such differences can be controlled for.

On the methodological side, the case study also reveals the importance of specifying the spatial dependence weight matrices properly in order for estimated correlations between link travel times to be meaningful. This is still an undeveloped area of research where more analysis of the characteristics of inter-link correlations using probe vehicle data is needed.

Acknowledgments

We would like to thank Tomas Julner from Trafik Stockholm for his support and provision of the data to the *iMobility* Lab at KTH, and Evanthia Kazagli for her assistance with processing the ANPR data. We are also grateful for the comments and suggestions from the Editor and the reviewers which helped improve the paper. The work was funded by the Swedish Transport Authority through the Mobile Millennium Stockholm project, and TRENoP Strategic Research Area.

References

Bernard, M., Hackney, J., Axhausen, K.W., 2006. Correlation of segment travel speeds. In: 6th Swiss Transport Research Conference, 2006.
de Fabritiis, C., Ragona, R., Valenti, G., 2008. Traffic estimation and prediction based on real time floating car data. In: 11th International IEEE Conference on Intelligent Transportation Systems, pp. 197–203.
Cheng, T., Haworth, J., Wang, J., 2012. Spatio-temporal autocorrelation of road network data. Journal of Geographical Systems 14 (4), 389–413.
Fosgerau, M., Fukuda, D., 2012. Valuing travel time variability: characteristics of the travel time distribution on an urban road. Transportation Research Part C 24, 83–101.
Hellinga, B., Izadpanah, P., Takada, H., Fu, L., 2008. Decomposing travel times measured by probe-based traffic monitoring systems to individual road links. Transportation Research Part C 16 (6), 768–782.
Hepple, L.W., 2004. Bayesian model choice in spatial econometrics. In: LeSage, J.P., Pace, R.K. (Eds.), Spatial and Spatiotemporal Econometrics, Advances in Econometrics. Elsevier Ltd., pp. 101–126.
Hofleitner, A., Herring, R., Abbeel, P., Bayen, A., 2012a. Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network. IEEE Transactions on Intelligent Transportation Systems 13 (4), 1679–1693.
Hofleitner, A., Herring, R., Bayen, A., 2012b. Arterial travel time forecast with streaming data: a hybrid approach of flow modeling and machine learning. Transportation Research Part B 46 (9), 1097–1122.
Hunter, T., Herring, R., Abbeel, P., Bayen, A., 2009. Path and travel time inference from GPS probe vehicle data. In: Neural Information Processing Systems Foundation (NIPS) Conference, Vancouver, Canada.
Hunter, T., Herring, R., Abbeel, P., 2012. The path inference filter: model-based low-latency map matching of probe vehicle data. arXiv preprint.
Kazagli, E., Koutsopoulos, H.N., 2012. Arterial travel time estimation from automatic number plate recognition data. In: Proceedings of the Annual Meeting of the Transportation Research Board, 2013.
Leduc, G., 2008. Road traffic data: collection methods and applications. JRC Technical Notes, Working Papers on Energy, Transport and, Climate Change, N.1.
LeSage, J., Pace, R.K., 2009. Introduction to Spatial Econometrics. Chapman & Hall/CRC, Boca Raton, Florida.
Liu, K., Yamamoto, T., Morikawa, T., 2009. Feasibility of using taxi dispatch system as probes for collecting traffic information. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations 13 (1), 16–27.
MATLAB, 2009. MATLAB version 7.8.0. Natick, Massachusetts: The MathWorks Inc., 2009.
Miwa, T., Sakai, T., Morikawa, T., 2008. Route identification and travel time prediction using probe-car data. International Journal of ITS Research 2 (1), 21–28.
Miwa, T., Kiuchi, D., Yamamoto, T., Morikawa, T., 2012. Development of map matching algorithm for low frequency probe data. Transportation Research Part C 22, 132–145.
Nanthawichit, C., Nakatsuji, T., Suzuki, H., 2003. Application of probe vehicle data for real-time traffic state estimation and short-term prediction on a freeway. In: Proceedings of the Annual Meeting of the Transportation Research Board, 2003.
Park, B.-J., Zhang, Y., Lord, D., 2010. Bayesian mixture modeling approach to account for heterogeneity in speed data. Transportation Research Part B 44 (5), 662–673.
Rahmani, M., Koutsopoulos, H.N., 2013. Path inference of low-frequency GPS probes for urban networks. Transportation Research Part C 30, 41–54.
Rakha, H., El-Shawarby, I., Arafeh, M., Dion, F., 2006. Estimating path travel-time reliability. In: Proceedings of the 2006 IEEE Intelligent Transportation Systems Conference, pp. 236–241.
Ramezani, M., Geroliminis, N., 2012. On the estimation of arterial route travel time distribution with Markov chains. Transportation Research Part B 46 (10), 1576–1590.
Uno, N., Kurauchi, F., Tamura, H., Iida, Y., 2009. Using bus probe data for analysis of travel time variability. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations 13 (1), 2–15.
Westgate, B.S., Woodard, D.B., Matteson, D.S., Henderson, S.G., in press. Travel time estimation for ambulances using Bayesian data augmentation. Annals of Applied Statistics.
Work, D.B., Tossavainen, O.-P., Blandin, S., Bayen, A.M., Iwuchukwu, T., Tracton, K., 2008. An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices. In: Proceedings of the 47th IEEE Conference on Decision and, Control, pp. 5062–5068.
Yuan, Y., van Lint, J.W.C., Wilson, R.E., van Wageningen-Kessels, F., Hoogendoorn, S.P., 2012. Real-time Lagrangian traffic estimator for freeways. IEEE Transactions on Intelligent Transportation Systems 13 (1), 59–70.
Zheng, F., van Zuylen, H., in press. Urban link travel time estimation based on sparse probe data. Transportation Research Part C.
Zou, L., Xu, J.-M., Zhu, L.-X., 2005. Arterial speed studies with taxi equipped with global positioning receivers as probe vehicle. In: Proceedings of the 2005 International Conference on Wireless Communications, Networking and Mobile, Computing, pp. 1343–1347.