Non-Parametric Estimation of Route Travel Time Distributions from Low-Frequency Floating Car Data

Mahmood Rahmani\textsuperscript{a,\#}, Erik Jenelius\textsuperscript{a}, Haris N. Koutsopoulos\textsuperscript{a,b}

\textsuperscript{a}Department of Transport Science, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden
\textsuperscript{b}Department of Civil and Environmental Engineering, Northeastern University, Boston, MA 02115

Abstract

The paper develops a non-parametric method for route travel time distribution estimation using low-frequency floating car data (FCD). While most previous work has focused on link travel time estimation, the method uses FCD observations for estimating the travel time distribution on a route. Potential biases associated with the use of sparse FCD are identified. The method involves a number of steps to reduce the impact of these biases. For evaluation purposes, a case study is used to estimate route travel times from taxi FCD in Stockholm, Sweden. Estimates are compared to observed travel times for routes equipped with Automatic Number Plate Recognition (ANPR) cameras with promising results. As vehicles collecting FCD (in this case, taxis) may not be a representative sample of the overall vehicle fleet and driver population, the ANPR data along several routes are also used to assess and correct for this bias. The method is computationally efficient, scalable, and supports real time applications with large data sets through a proposed distributed implementation.

Keywords: floating car data, automatic number plate recognition, kernel estimation, route travel time distribution, sampling bias, non-parametric.

1. Introduction

Monitoring traffic conditions in light of increasing congestion in urban areas is critical for traffic management and effective transport policy. Provision of travel time information is also important as a means of dealing with congestion. There are a number of well-established technologies for travel time data collection, including loop detectors and automatic vehicle identification (AVI) sensors (Antoniou et al., 2011). AVI systems (Automatic Number Plate Recognition (ANPR) cameras, Bluetooth devices, etc.) provide direct measurements of route travel times. However, the spatial coverage is typically small, and may not be representative of the network as a whole.

Floating car data (FCD) collected from GPS devices installed in vehicle fleets or smart phones are becoming increasingly available. FCD can complement stationary sensors by providing information from the entire network. Methods for highway and arterial traffic state estimation based on speed measurements from FCD and loop detectors have been developed (Treiber et al., 2011; Cipriani et al., 2012; Yuan et al., 2014). Other studies use the concept of a virtual probe vehicle inferred from detectors and signal status data along a corridor helps improve the allocation of observed travel time of FCD to different sections of the road (Liu and Ma, 2009). Travel time estimation from FCD is often challenging because of low polling frequency (less than once or twice per minute due to bandwidth limitations and data transmission costs), which means that vehicles may traverse multiple links between consecutive probes (Jenelius and Koutsopoulos, 2015).

The research on FCD-based travel time estimation has largely focused on network links (Sanaullah et al., 2013; Hofleitner et al., 2012; Zheng and van Zuylen, 2012; Hellinga et al., 2008). In general, these methods allocate the travel time between two consecutive probes to the traversed links. Average route travel times can be estimated from the average link travel times. However, a drawback of the link-based
approach is that statistics of the route travel time distribution (apart from the mean value) are not straightforward to derive from the travel time distributions of the constituent links. For many applications, for example, monitoring of path travel time reliability, estimation of the variance and percentiles is as important as the calculation of the mean. While several models have been proposed (Hofleitner et al., 2012; Westgate et al., 2013; Ramezani and Geroliminis, 2012; Jenelius and Koutsopoulos, 2013), they typically rely on strong assumptions about the functional form of the link travel time distributions and the correlation structure. For real-time applications, there is also a trade-off between the complexity of the model and the computational efficiency of route travel time calculations.

A somewhat different approach is taken in Westgate (2013), with trip travel time distributions estimated from origin-destination FCD travel time observations. The proposed parametric method avoids the problems of aggregating link travel times into route travel times, but does not utilize the information from intermediate FCD reports.

The objective of the paper is to propose a computationally efficient, non-parametric method for estimating the distribution of route travel times using low-frequency FCD, incorporating all available information. No assumptions are made regarding the form of the distribution. Given the complexity of urban traffic, it is likely that the form of the travel time distribution varies by route and time of day. Thus, the flexibility of the proposed method is highly valuable whenever the variability of travel times is of interest.

The method incorporates all available observations (fully or partially overlapping with the route of interest). The reason that partially overlapping observations are utilized is the limited availability of fully overlapping ones. In the case study presented in Section 4.2, even with FCD from 1500 taxis over the period of one year, the number of direct observations is rather low. Therefore, it is important to develop a method that incorporates both fully and partially overlapping observations.

The methodology addresses several challenges due to the nature of FCD (which to a different degree also exist when estimating link travel time distribution). Examples include partial overlap, oversampling of sections of the route, etc. These factors can bias the estimation unless they are accounted for. The paper identifies and categorizes the most important sources of bias. The proposed methodology uses ideas from kernel-based estimation (e.g. Hastie et al., 2009) and takes into consideration the particular features of network routes and FCD observations to correct for these biases.

Another potential bias when using FCD from vehicle fleets to estimate general traffic conditions is the representativeness of the sample with respect to driver behavior and traffic regulations applicable to the vehicles in the fleet.

A case study using FCD from 1500 taxis in Stockholm, over a period of one year, is used to illustrate the proposed method. Data from ANPR on selected routes is also available. The results from the FCD compare well with the ANPR travel data (which provide direct observations of route travel times, but at a very limited spatial coverage). The results also illustrate that the proposed approach deals effectively with the identified biases. Furthermore, the case study is used to evaluate the sampling bias (not being representative) in the Stockholm region. A significant amount of this bias is explained by the use of bus lanes by taxis (while regular vehicles are not allowed to use them).

The fact that estimation is route-specific places high demand on computational efficiency for on-demand (real-time) applications where the route is provided as part of a request. The computational complexity of the proposed method is evaluated, and a framework for distributed computation is presented. It is demonstrated that, in addition to off-line applications, the method is suitable for real time applications where short response times are required, even for large data sets.

The main contributions of the paper are:

- Identifies important biases introduced when FCD are used for the estimation of path travel times.
- Develops efficient methods for their correction and demonstrates their effectiveness.
- Uses ANPR data for comparison purposes (which is probably one of very few papers that have done so). It also uses ANPR data to correct for sampling biases (i.e. the sample of taxis is not representative of the population of drivers).
- Proposes a computationally efficient implementation that is scalable and can support real time applications with large data sets.

The paper is organized as follows. Section 2 introduces the basic concepts and discusses the nature of FCD observations. Section 3 presents the travel time estimation methodology. Section 4 discusses the
application of the methodology in a case study with taxi FCD from Stockholm, with results in Section 5. The computational performance of the method and approaches to its efficient implementation are presented in Section 6. Section 7 concludes the paper.

2. Preliminaries

2.1. Network route travel time

A network route is defined as a path \( \pi = (k_s, k', \ldots, k'', k_e) \) connecting the beginning and end links \( k_s \) and \( k_e \), and two offsets \( o_s \) and \( o_e \), where \( o_s (o_e) \) is the distance from the start of link \( k_s (k_e) \) to the start (end) location of the route. A route may thus begin and end at points in the interior of links and not necessarily at the beginning or end of links. The path \( \pi \) is assumed to be acyclic. The length of overlap between the network route and link \( k \), denoted by \( a_k \), is

\[
a_k = \begin{cases} 
  o_e - o_s & k = k_s = k_e \\
  t_k - o_s & k = k_s \neq k_e, \\
  o_e & k = k_e \neq k_s, \\
  t_k & k \in \pi \setminus \{k_s, k_e\}, \\
  0 & \text{otherwise.} 
\end{cases}
\]

(1)

The ratio of the length of overlap to the total length of link \( k \) is denoted by \( \alpha_k = a_k/t_k \).

The route travel time is denoted by \( T \) and varies stochastically between trips. Furthermore, the route travel time varies as a function of the route entry time \( T = T(s) \), where \( s \) is the time that a vehicle passes the beginning of the route.

2.2. Travel time measurements from FCD

FCD consist of sequences of probes from vehicles traveling on the network. Each probe is a triplet \((q, s, < x, y >)\), where \( q \) is a unique vehicle identifier, \( s \) is a timestamp, and \(< x, y >\) are the GPS coordinates of the vehicle location at that time. In some systems other information such as instantaneous speed, heading and odometer are also reported; such information is not considered in this paper.

Low-frequency FCD require preprocessing to be useful for travel time estimation. Most importantly, reported positions \(< x, y >\) must be matched to the model of the road network and the paths taken by the vehicles between probes must be inferred. A number of different approaches for map-matching and path inference have been presented in the literature, e.g., Miwa et al. (2012); Bierlaire et al. (2013); Rahmani and Koutsopoulos (2013). In the remainder of this paper, it is assumed that map-matching and path inference have been performed.

For a pair of consecutive probes \( i \) and \( i + 1 \) from the same vehicle, a travel time measurement \( t_i = s_{i+1} - s_i \) is obtained. Furthermore, map-matching and path inference yield a path \( p_i = (k_i, k', \ldots, k'', k_{i+1}) \) and two offsets \( o_i \) and \( o_{i+1} \), referred to as the FCD route. Probe \( i \) is thus matched to link \( k_i \) and probe \( i + 1 \) to \( k_{i+1} \), and the offsets \( o_i \) and \( o_{i+1} \) are defined as the distances from the beginning of the links to the location of the first and second probes, respectively. The distance traversed on link \( k \) by observation \( i \), denoted \( r_{ik} \), is

\[
r_{ik} = \begin{cases} 
  o_{i+1} - o_i & k = k_i = k_{i+1} \\
  t_k - o_i & k = k_i \neq k_{i+1}, \\
  o_{i+1} & k = k_{i+1} \neq k_i, \\
  t_k & k \in p_i \setminus \{k_i, k_{i+1}\}, \\
  0 & \text{otherwise.} 
\end{cases}
\]

(2)

The ratio of the traversed distance to the total length of link \( k \) is denoted by \( \rho_{ik} = r_{ik}/t_k \).

The part of the FCD route overlapping with the network route is referred to as the overlap route for short. The fraction of link \( k \) covered by both FCD route \( i \) and the network route is denoted by \( \beta_{ik} \); note that \( \beta_{ik} \leq \alpha_k \) and \( \beta_{ik} \leq \rho_{ik} \).

The notation used in the paper is summarized below; an illustrative example is given in Figure 1.
\( \ell_k \) length of link \( k \)
\( \alpha_k \) fraction of link \( k \) included in the definition of the network route
\( \tau_i \) travel time observation \( i \)
\( s_i \) timestamp of the first probe in observation \( i \)
\( \rho_{ik} \) fraction of link \( k \) covered by observation \( i \)
\( \beta_{ik} \) fraction of link \( k \) covered by both observation \( i \) and network route

\[ \begin{align*}
\emptyset & = r_1 \quad \alpha_1 = 0 \quad \beta_1 = 0 \\
\emptyset & = r_2 \quad \alpha_2 = \frac{a_2}{l_2} \quad \beta_2 = \frac{a_2}{l_2} \\
\emptyset & = r_3 \quad \alpha_3 = \frac{a_3}{l_3} \quad \beta_3 = 1 \\
\emptyset & = r_4 \quad \alpha_4 = \frac{a_4}{l_4} \quad \beta_4 = \frac{a_4}{l_4}
\end{align*} \]

Figure 1: Notation example. Since there is only one observation in this example the observation index \( i \) is omitted.

2.3. Sources of bias

As illustrated in Figure 2, FCD probes are not in general generated at the start and end points of routes or links but wherever the vehicles were polled. There are typically very few trips with reports located at the beginning and end of a route, which means that estimating a distribution of travel times using only fully overlapping observations may not be feasible. Therefore, all observations, with full or partial overlap with the route, are considered for the route travel time distribution estimation. However, incorporating partial observations also introduces a number of potential sources of bias in estimating route travel time distributions from FCD observations.

Incomplete traversal of route

An FCD observation may cover only a fraction of the network route; this is the case for all observations illustrated in Figure 2. The travel time of the vehicle on other parts of the route is thus unknown and needs to be inferred based on the partial observation and a priori information about the route. If the speed characteristics of the traversed and the non-traversed parts of the route are different and assumptions used for the extrapolation are not correct, the estimated route travel time will be biased.

Influence of adjacent network

An FCD observation may partially traverse parts of the network not included in the route, as illustrated by observations 1, 4–6 in Figure 2. The allocation of the observed travel time between the route and the adjacent links is unknown and needs to be inferred based on appropriate assumptions. Incorrect allocation means that the travel speed on the adjacent links will influence and bias the estimated route travel time.

Figure 2: Examples of FCD observations after map-matching.
Non-uniform coverage of route

Vehicles may enter and leave the route via side streets, as illustrated by observations 5 and 6 in Figure 2. This means that the number of FCD observations may vary along the route so that some sections are under-represented relative to other sections. If the speed characteristics are not homogeneous along the route, the estimated route travel time will be biased without appropriate correction.

Non-representative vehicle sample

FCD from opportunistic sensors may be generated by vehicle fleets that are not representative samples of the whole population. For example, the data may be generated from commercial vehicles (taxis, distribution vehicles, etc.). As a result, the vehicles may be driven differently (in terms of aggressiveness, en-route stops, etc.), or may be bound by different traffic regulations (speed limits, access to bus lanes, etc.). In contrast to the previously discussed issues, the nature of this bias and suitable approaches to correct it strongly depend on the particular application.

3. Methodology

The methodology for estimating the travel time distribution on a network route using low-frequency FCD is a non-parametric kernel-based approach. The output of the estimation includes statistics of the travel time distribution of interest (moments, percentiles, probability density function, etc.). No assumption is made about the functional form of the distribution; rather, this is an output of the estimation. The method is developed to address the challenges of partial route coverage, traversal of the adjacent network, and non-uniform coverage of the route discussed in Section 2.3. Correcting for the potential bias due to the non-representativeness of the sample of probe vehicles depends on the characteristics of the source of FCD and the network. A correction method developed specifically for FCD from taxis in Stockholm, Sweden is presented in Section 5.3.

The estimation method consists of a sequence of steps: transformation, weighting, and aggregation. The first step transforms FCD observations that only partially overlap with the network route to observations of the route travel time. Each observation is then weighted according to its relevance as a route travel time observation. The final step is to aggregate all weighted observations and calculate the sought statistics.

It is assumed that a set of FCD observations \(i = 1, \ldots, M\) is available, with the corresponding FCD route having some overlap with the network route of interest. Furthermore, it is assumed that for each network link \(k\) a time-dependent prior travel time \(t_{ik}\) is known. The index \(i\) indicates that \(t_{ik}\) is the prior link travel time corresponding to the time stamp of observation \(i\). The prior travel times may be estimated from FCD as well, obtained from some other sources, or approximated from the free-flow speed of each link. They are link and time period dependent and may even be clustered by weather, day of the week, etc, if such information is available.

The method is flexible with respect to the network representation. For example, links between intersections may be divided into multiple segments, which may increase the precision of the estimation. Furthermore, turn movements at intersections can be represented with specific virtual links to capture the fact that certain movements (e.g., left turns) may experience longer delays than others (see for example Jenelius and Koutsopoulos, 2013).

3.1. Transformation

The first step of the methodology is to transform each FCD observation partially covering the route into an observation of the actual route travel time. The step is performed independently for each observation and consists of four processes: concatenation, allocation, scaling, and route entry time estimation.

Concatenation

Depending on the length of the network route in relation to the sampling frequency of the FCD, a vehicle may report multiple probes along the route (e.g., observations 1–4 in Figure 2). It is reasonable, however, to consider one passage of a vehicle on the route as one travel time observation. Consecutive observations from the same vehicle are thus concatenated into a single travel time observation. The result is a new, smaller set of FCD observations to be used in the subsequent steps of the methodology.
Furthermore, concatenation yields observations that better capture the correlation between the links. Concatenating observations decreases the effects of several sources of bias and improves the coverage of the route.

There is a trade-off between having a larger fraction of the route traversed on one hand and a larger influence of the adjacent network on the other hand. If a vehicle is polled multiple times along the route, there is therefore a choice whether to include the observations during which the vehicle enters and leaves the route, respectively, in the concatenated observation (e.g., observations 1 and 4 in Figure 2). In general, there are four possibilities of concatenating the trace (1-3, 1-4, 2-3, and 2-4 for this example). The solution proposed here is to use the concatenation that results in an observation with the highest weight according to the kernel function described in Section 3.2.

Concatenation is done in the following steps:

1. Group observations of the same trip together as a trace.
2. Select a subset of the trace that overlaps the route of interest, namely sub-trace $j$, which corresponds to a sequence of $n$ observations $(i_1, \ldots, i_n)$.
3. Merge the trace $j$ into a single observation so that:
   - The first timestamp of the trace becomes the first timestamp of the merged observation, i.e., $s_j = s_{i_1}$.
   - The total travel time of the trace is the travel time of the merged observations, i.e., $\tau_j = \sum_{m=1}^{n} \tau_{i_m}$.
   - The concatenation of the paths of the trace becomes the path of the merged observation, i.e., $p_j = (p_{i_1}, \ldots, p_{i_n})$.
   - The fraction of the trace traversed on link $k$ is the fraction traversed on link $k$ for the merged observation, i.e., $p_{jk} = \sum_{m=1}^{n} p_{i_m,k}$.
   - The total overlap of the trace and the route on link $k$ becomes the overlap of the merged observation and the route on link $k$, i.e., $\beta_{jk} = \sum_{m=1}^{n} \beta_{i_m,k}$.

In uncommon cases, a vehicle may pass along a route multiple times within the same trace, or bypass part of the route, as illustrated in Figure 3. In such cases, concatenation applies only on sub-parts of the trace as opposed to the entire trace, excluding the intermediate off-route parts. In the example of Figure 3, observation 3 is excluded and two new observations are created by merging 1 with 2, and 4 with 5.

**Allocation**

The next step considers the FCD observations that partially traverse the network adjacent to the route. For each observation $i$, the observed travel time $\tau_i$ is allocated between the network route and the adjacent network. The allocation is based on the prior link travel times $t_{0jk}$ and the distance traversed on each link. Using information about prior link travel times rather than only the distance traversed on each link is expected to reduce the bias introduced by the influence of side streets.

The prior travel time estimate of the observation $i$ is $\sum_k \beta_{jk} t_{0jk}$, while the prior travel time estimate on the overlap route is $\sum_k \beta_{jk} t_{0jk}$. The main assumption in the allocation is that the fraction of time spent on the overlap route in relation to the whole FCD route by observation $i$, $\phi_i$, is the same as for the prior travel times on the same sections. The travel time allocated to the overlap route is then

$$\tau_i' = \phi_i \tau_i,$$

(3)
where the allocation factor $\phi_i$ is calculated from the prior link travel times,
\[
\phi_i = \frac{\sum_k \beta_{ik} t_{0ik}}{\sum_k \rho_{ik} t_{0ik}}.
\] (4)

The allocation factor is 1 for observations traversing only the network route and approaches 0 as the distance traversed on adjacent links increases.

**Scaling**

While the allocation step estimates the time spent on the network route for each FCD observation, the observations do not, in general, traverse the entire route. The next step, therefore, is to scale up the travel time observations to the entire route. Similar to the allocation, the scaling is based on the assumption that the ratio between the travel time on the overlap route and the travel time on the entire network route is the same as for the prior travel time estimates on the same sections. The scaled route travel time observation is then
\[
T_i = \frac{1}{\eta_i} \tau'_i,
\] (5)

where $\eta_i$ is the scaling factor
\[
\eta_i = \frac{\sum_k \beta_{ik} t_{0ik}}{\sum_k \alpha_k t_{0ik}}.
\] (6)

The scaling factor is 1 for observations covering the entire route and approaches 0 as the overlap with the route decreases. In addition to scaling, $\eta_i$ can be used for filtering out observations with route overlap below a predetermined threshold. For example, if observations with full overlap with the route are of interest, then only observations with $\eta_i$ close to 1 are used.

**Route entry time extrapolation**

The time that each probe vehicle passes the beginning of the network route (the route entry time) is in general not observed or may not even exist if the vehicle joins the route at some point further along the route (such an example is shown in Figure 4). However, the route entry time is the basis for clustering observations and aggregating statistics. For each observation the route entry time, real or hypothetical, is estimated based on the prior travel time estimates along the same lines as the allocation and the scaling.

The travel time between the first probe (with timestamp $s_i$) and the first node in the interior of the network route that is reached is estimated as
\[
\tau'_i = \frac{\sum_{k \in \pi'} \alpha_k t_{0ik}}{\sum_k \rho_{ik} t_{0ik}} \tau_i,
\] (7)

where $\pi'$ is the path of the network route between the beginning of the route and the first reached node within the route.

The route entry time $s'_i$ is estimated as
\[
s'_i = s_i + \tau'_i - \tau^{(2)}_i.
\] (9)
3.2. Weighting

After transformation, each route travel time observation \( T_i \) is assigned a weight \( \omega_i \) that determines the influence of the observation in the estimation of route travel time statistics. Observations are weighted for two reasons: to reflect the level of representativeness of the observation in relation to the network route; and to correct for sampling bias due to uneven coverage of the route. The final weight is the product of the representativeness weight \( \nu_i \) and the sampling bias weight \( \lambda_i \), i.e., \( \omega_i = \nu_i \lambda_i \).

The range of \( \omega_i \) depends on route (e.g. number of links, type and location in the urban area) and time of day. Examples of the range of \( \omega_i \) is presented in the routes of the application (see Section 5.1).

**Route and adjacent network traversal weighting**

The confidence in each observation \( T_i \) as drawn from the true distribution of route travel times depends on the potential amount of error introduced in the transformation step. The potential for error in the allocation is greater if the overlap between the FCD route and the adjacent network is large, introducing more bias from the conditions outside the route. Similarly, the potential for error in the scaling is greater if the overlap between the FCD route and the network route is small. Lower overlap with the network route means that the relevance of the FCD observation as a network route observation is lower. An FCD observation is perfectly relevant if the start and end positions coincide with the start and end positions of the network route.

Standard kernel-based methods assign weights to measurements according to a distance metric between the measurement points and the target point in Euclidean space. In the present setting, the measurements and the target are not points in Euclidean space but routes in a network. The negative logarithms of the allocation factor \( \phi_i \) and the scaling factor \( \eta_i \),

\[
\delta_{i,\phi} = -\log \phi_i = \log \sum_k \rho_{ik} t_{ik}^0 - \log \sum_k \beta_{ik} t_{ik}^0, \tag{10}
\]

\[
\delta_{i,\eta} = -\log \eta_i = \log \sum_k \alpha_{ik} t_{ik}^0 - \log \sum_k \beta_{ik} t_{ik}^0, \tag{11}
\]

can be interpreted as one-dimensional measures of the distance between the FCD route and the overlap route, and between the network route and the overlap route, respectively. Note that \( \beta_{ik} \leq \alpha_{ik} \) and \( \beta_{ik} \leq \rho_{ik} \) guarantees that \( \delta_{i,\phi} \) and \( \delta_{i,\eta} \) are positive. The proposed method applies the kernel weighting on these distances. In the most general formulation, a two-dimensional kernel is defined considering the distance between the points \( p_1 = (\log \sum_k \rho_{ik} t_{ik}^0, \log \sum_k \alpha_{ik} t_{ik}^0) \) and \( p_2 = (\log \sum_k \beta_{ik} t_{ik}^0, \log \sum_k \beta_{ik} t_{ik}^0) \) according to some distance metric in \( R^2 \). In the following, we use separable kernel functions,

\[
\nu_i = \nu_{\phi} \left( \frac{\delta_{i,\phi}}{\theta_1} \right) \nu_{\eta} \left( \frac{\delta_{i,\eta}}{\theta_2} \right), \tag{12}
\]

where \( \theta_1 > 0 \) and \( \theta_2 > 0 \) are two bandwidth parameters.

In the machine learning literature, common choices for the kernel functional form include the Epanechnikov and tri-cube kernels (with compact support) and the Gaussian density function (Hastie et al., 2009). Another common kernel function is the negative exponential of the distance metric, which has been demonstrated (Treiber and Helbing, 2002; van Lint and Hoogendoorn, 2010) to perform well in traffic state estimation when used for smoothing in space and time. Based on this evidence from the literature, the negative exponential kernel is used for the analysis in this paper, given by:

\[
\nu_i = \exp \left( -\frac{\delta_{i,\phi}}{\theta_1} \right) \exp \left( -\frac{\delta_{i,\eta}}{\theta_2} \right), \tag{13}
\]

or equivalently,

\[
\nu_i = \phi_i^{1/\theta_1} \eta_i^{1/\theta_2}. \tag{14}
\]

Observations are thus weighted directly based on the allocation and scaling factors. The parameters \( \theta_1 \) and \( \theta_2 \) control how fast the weight function decays as the overlap with the adjacent networks increases, and the overlap with the network route decreases, respectively.

If ground-truth travel time data were available, the kernel function and the associated bandwidth parameters can be selected using cross-validation techniques.
Non-uniform route coverage weighting

As discussed, vehicles may enter and leave the route on side streets, meaning that the number of observations on different parts of the route may differ. Non-uniform coverage of the route means that the FCD do not constitute a representative sample of the route travel time. To reduce this bias in the absence of traffic counts, observations are weighted according to the relative number of observations covering the same parts of the route. Route coverage is evaluated at the link level. Let $N_k$ be the number of observations covering (partially or fully) link $k$:

$$N_k = \sum_i \lceil \beta_{ik} \rceil,$$

where $\lceil x \rceil \in \{0, 1\}$ is the ceiling function. A weighted average is calculated for the traversed part of the route using link lengths as weights. The weight $\lambda_i$ given to each observation $i$ is the inverse of this weighted average coverage,

$$\lambda_i = \sum_k \beta_{ik} \ell_k / \sum_k \beta_{ik} \ell_k N_k.$$

3.3. Aggregation

The last step of the estimation process calculates statistics of the route travel time distribution from the travel time observations $T_i$ and the associated weights $\omega_i$. The observations are aggregated based on the corresponding route entry time according to pre-specified clusters (i.e. time-of-day intervals, weekday, season, etc.) Let $C$ denote the set of observations belonging to such a cluster and let $N_C$ denote the number of observations in the set.

Moments

The most common statistic of the travel time distribution, also available from link-based approaches, is the mean value $\hat{\mu}_{T(C)}$.

$$\hat{\mu}_{T(C)} = \frac{\sum_{i \in C} \omega_i T_i}{\sum_{i \in C} \omega_i}. \tag{17}$$

The estimator of the travel time variance is given by:

$$\hat{\sigma}^2_{T(C)} = \frac{\sum_{i \in C} \omega_i (T_i - \hat{\mu}_{T(C)})^2}{\sum_{i \in C} \omega_i}. \tag{18}$$

Higher order moments (skewness, kurtosis, etc.) are estimated similarly.

Percentiles

Percentiles of the travel time distribution are often used as indicators of reliability and traffic safety. The $p$-th percentile of the travel time distribution can be estimated as follows. Let $T_{(i)}$ be the $i$-th travel time observation in ascending order and $\omega_{(i)}$ the associated weight, $i = 1, \ldots, N_C$. Let $W_i = \sum_{j=1}^{i} \omega_{(j)}$ denote the $i$-th partial sum of the weights, and $p_i = 100/W_i \cdot (W_i - \omega_{(i)}/2)$ denote the weighted percent rank of the $i$-th smallest observation. The $p$-th percentile, $\hat{T}(C)_p$, is given by:

$$\hat{T}(C)_p = \begin{cases} T_{(1)} & p < p_1, \\ T_{(N_C)} & p \geq p_{N_C}, \\ T_{(i)} + \frac{p - p_i}{p_{i+1} - p_i} (T_{(i+1)} - T_{(i)}) & p_i \leq p < p_{i+1}, i \in \{1, N_C - 1\} \end{cases} \tag{19}$$

Probability density function

The probability density function of the travel time distribution may be estimated by kernel density estimation. Let $g_h(\cdot)$ be a kernel function with bandwidth parameter $h$. The kernel density estimator is

$$\hat{f}_{T(C)}(t) = \frac{\sum_{i \in C} \omega_i g_h(t - T_i)}{\sum_{i \in C} \omega_i}. \tag{20}$$

Common choices for the kernel function include uniform, triangular and Gaussian density functions (Silverman, 1986).
4. Application

This section describes an application of the proposed method for the estimation of travel time distributions of several routes in the arterial street network of Stockholm, Sweden. The routes are routes used by the city for monitoring purposes. ANPR cameras installed on each route collect travel time data. The estimated travel time distributions from FCD are compared to the corresponding empirical distributions from the ANPR system. Furthermore, a statistical regression model is used to explain the differences between FCD and ANPR mean travel times in terms of the characteristics of the routes. The model can be used to adjust the estimates from FCD to be representative of the overall driver population as captured by the ANPR cameras.

4.1. Data

FCD are collected by a GPS-based taxi fleet management system covering about 1500 taxis. Each taxi broadcasts its location (latitude, longitude), timestamp, id, and status (free/hired) once every two minutes on average (for more detail about the data see Rahmani et al., 2010).

Based on criteria such as reliability of ANPR data and availability of FCD (e.g. avoiding routes with long tunnels where collecting FCD is infeasible), 27 routes are selected for the analysis (see Figure 5). For map-matching and path inference of the raw FCD, the method developed by Rahmani and Koutsopoulos (2013) is used. For the travel time estimation problem only probes related to a hired taxi are used. Excluding probes with status free reduces the impact of reports from stand still vehicles at taxi stations, cruising taxis searching for passengers, and taxis changing passengers. Also, consecutive probes from the same vehicle with gaps longer than 3 minutes are ignored.

ANPR data typically have a lot of noise, including vehicles that stop or take detours along the route; unless the data are filtered, travel time measures will thus tend to be biased upwards. Several studies have developed filtering methods for AVI-based travel time data (see for example Dion and Rakha, 2006; Kazagli and Koutsopoulos, 2013). This study uses the method proposed in Kazagli and Koutsopoulos (2013) which assumes that ANPR observations are drawn from a mixture of two populations: one representing normal movement through the network and one representing vehicles that diverge or stop for whatever reason along the route. A mixture model is estimated to separate the noise from the valid observations. The method was tested with ANPR from the same network used in this study and the results from the analysis of a number of routes (also used in this paper) show better performance than standard outlier removal methods.

4.2. Experimental design

For both sources (FCD and ANPR), data from Mondays through Thursdays between 6 a.m. and 10 p.m. are used, collected from September 15, 2012 to September 15, 2013.

The proposed route travel time estimation method requires prior link travel times $t_{0k}$ as input. The prior link travel times are estimated from the same set of FCD by applying the proposed method on each individual link. The estimation approach may thus be interpreted as a two-stage method based on a single data set. Values of $t_{0k}$ depend on time of day, calculated for each 15-minute interval. $t_{0k}$ as such is link and time period specific and incorporates a lot of prior information related to each individual link. Furthermore, as more data become available, $t_{0k}$ can be clustered by day of week, weather conditions, etc. This is a step that is performed off-line and updated periodically. For the weighting step of the method, the negative exponential kernel is used with bandwidth parameters $\theta_1 = \theta_2 = 1$ (refer to Equation 14).

The route travel time estimated from FCD is compared to ANPR data in terms of the mean, 25th, 50th (median), and 75th percentiles of the distribution, and the probability density function. For comparison, the mean route travel time is also estimated from the same set of FCD as the sum of the link average travel times (link-based approach), $\hat{\mu}_{T(n)} = \sum_k \alpha_k t_{0nk}$, where $n$ refers to the time interval the observation belongs to.

Travel time observations from both ANPR and FCD are grouped by route entry time into 15-minute intervals, and the travel time statistics are calculated for each time interval $n = 1, \ldots, N$ from 6 a.m. to 10 p.m. The similarity between each travel time statistic of FCD, $z$, and ANPR, $z'$, for a given route is evaluated using the following performance measures:
Figure 5: The study area indicated by the shadowed polygon (about 34 km$^2$ and 7278 links).
1. Root mean squared error (RMSE):

\[ \text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (z_n' - z_n)^2} \]

(21)

2. Normalized RMSE:

\[ \text{RMSNE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left( \frac{z_n' - z_n}{z_n'} \right)^2} \]

(22)


\[ U = \frac{\sqrt{\frac{1}{N} \sum_{n=1}^{N} (z_n' - z_n)^2}}{\sqrt{\frac{1}{N} \sum_{n=1}^{N} z_n'^2} + \sqrt{\frac{1}{N} \sum_{n=1}^{N} z_n^2}} \]

(23)

U takes values between 0 and 1 with values closer to 0 indicating better performance. U can be decomposed into proportions of inequality: bias \( U^M \), variance \( U^S \), and covariance \( U^C \):

\[ U^M = \frac{(\mu_{z'} - \mu_z)^2}{\frac{1}{N} \sum_{n=1}^{N} (z_n' - z_n)^2}, \]

(24)

\[ U^S = \frac{(\sigma_{z'} - \sigma_z)^2}{\frac{1}{N} \sum_{n=1}^{N} (z_n' - z_n)^2}, \]

(25)

\[ U^C = \frac{2(1 - \rho)\sigma_{z'} \sigma_z}{\frac{1}{N} \sum_{n=1}^{N} (z_n' - z_n)^2}, \]

(26)

\[ U^M + U^S + U^C = 1, \]

(27)

where \( \mu_z = \frac{1}{N} \sum_{n=1}^{N} z_n \) is the average of \( z \) statistic of a route over all time intervals, \( \sigma_z \) the standard deviation, and \( \rho \) the correlation coefficient. For any value of \( U > 0 \), low values of \( U^M \) and \( U^S \) indicate better performance.

5. Results

5.1. Comparison with ANPR

Travel times of the selected routes are estimated using the proposed method and compared with corresponding ANPR statistics. Table 1 summarizes the overall errors over all 27 routes and 64 time intervals. The link-based route travel time method can only estimate mean travel times (not distributions). The results suggest that in terms of mean travel times, link and route-based methods have practically the same overall performance. \( U^M \) is small indicating low systematic bias. Figure 6 depicts the mean travel time of twelve of the routes as a function of time of day according to the route-based and link-based estimation approaches and the ANPR observations. The difference between FCD-based estimates and ANPR is usually smaller during non-congested hours. In congested hours, especially where ANPR shows
Figure 6: Mean travel times for 12 routes in Stockholm estimated using the FCD, route and link-based methods, compared with ANPR travel times, in 15-minute intervals.
sharp peaks, the FCD-based method underestimates the travel time. In general, the method captures the trend by time of day similar to ANPR except for some routes in peak hours. The differences are mainly due to the fleet population bias, as taxis are allowed to use bus lanes, while the general cars are not. The issue is discussed in detail in Section 5.3 where the corresponding bias is quantified. While the route-based travel time estimation approach performs similarly to the link-based approach for the mean travel time, the main benefit of the route-based approach is the capability of estimating other statistics. Figure 7 shows the median and 25th and 75th percentiles of the travel time distributions according to ANPR and FCD (Equation 19). The medians are shown by markers and the vertical lines are drawn from the first to the third quartiles. In general, FCD percentiles are close to ANPR except for some routes during congested hours. For some routes (e.g. 67, 68, and 116) the travel time variability differs significantly by time of day and this is captured by both FCD and ANPR. Other routes (e.g. 19, 34, and 55) display stable travel times along the day and less intra-period variability. As expected, longer routes have greater travel time variability. Figure 8 depicts the travel time distributions of the same routes.
routes estimated by a kernel density smoothing (Equation 20 with normal kernel function) based on FCD and ANPR data for the time interval 07:00-07:15 a.m. The ANPR distributions confirm that the travel times are not drawn from a consistent distribution across routes. Overall, the shapes of the distributions (location, concentration, skewness, etc.) are estimated reasonably well from the FCD. As mentioned

As mentioned

![Comparison of the estimated travel time distributions using FCD and ANPR data for the period of 07:00-07:15 a.m.](image)

A two-sample Kolmogorov-Smirnov test is conducted for each route and time interval to test whether the ANPR and FCD distributions are equal. In total, the null hypothesis that the two distributions are equal cannot be rejected for 53 (119) out of 1728 (27 × 64) route-interval combinations at significance level $\alpha = 0.05$ (0.01). This is consistent with expectations since indeed the two samples are from different distributions (although, such tests with large number of observations tend to reject the null hypothesis often).

---

1A two-sample Kolmogorov-Smirnov test is conducted for each route and time interval to test whether the ANPR and FCD distributions are equal. In total, the null hypothesis that the two distributions are equal cannot be rejected for 53 (119) out of 1728 (27 × 64) route-interval combinations at significance level $\alpha = 0.05$ (0.01). This is consistent with expectations since indeed the two samples are from different distributions (although, such tests with large number of observations tend to reject the null hypothesis often).
of links, length, location in the urban area, type of route such as through route or not, etc) and time of day. As shown in Figure 9, some routes, such as 55, may have relatively high weights during the day (indicating vehicles traversing large sections of the route), while some other routes, for example 64, have low values of weights. There are also some routes, for example 68, with mixture distributions of weights caused by, among other things, multiple entries/exits along the route.

Figure 9: $\omega_i$ distribution for routes 55, 64, 68 (top). 25th, 50th, 75th percentiles of $\omega_i$ as a function of time of day (bottom).

5.2. Analysis of FCD bias corrections

As discussed in Section 2.3, three potential sources of bias in FCD are: incomplete traversal of the route, influence of the adjacent network, and non-uniform coverage of the route. For each FCD observation, these aspects are captured by the factors $\eta$, $\phi$, and $\lambda$, respectively. A question of interest is how much of the difference between FCD and ANPR travel times is attributed to these factors. To analyze this, the mean and the standard deviation of the three factors across observations ($\hat{\mu}_\eta, \hat{\mu}_\phi, \hat{\mu}_\lambda, \hat{\sigma}_\eta, \hat{\sigma}_\phi, \hat{\sigma}_\lambda$) are used as explanatory variables in a regression model for the difference between the FCD and ANPR travel times of the 27 routes. Since the estimation methodology takes these factors into account through the weighting, the expectation is that their contribution is not significant.

Table 2 summarizes the distribution of the variables over the 27 routes, separately for congested and non-congested periods. Congested periods are different from route to route and defined individually for each route based on the ANPR data. The regression model used is:

$$\log\left(\frac{\mu_T^*}{\hat{\mu}_T}\right) = \beta_0 + \beta_1 \hat{\mu}_\eta + \beta_2 \hat{\mu}_\phi + \beta_3 \hat{\mu}_\lambda + \beta_4 \hat{\sigma}_\eta + \beta_5 \hat{\sigma}_\phi + \beta_6 \hat{\sigma}_\lambda + \epsilon_r$$

where $\hat{\mu}_T^*$ and $\hat{\mu}_T$ are mean travel times from ANPR and FCD respectively, for route $r \in \{1, ..., 27\}$.

The results of testing the hypothesis $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ are shown in Table 3, indicating that the null hypothesis cannot be rejected (for both, congested and non-congested periods). Hence, the effect of $\eta$, $\phi$, and $\lambda$ on the deviation of FCD (which has already been corrected for corresponding biases using the weights) from ANPR is insignificant. The remaining difference between the FCD estimates and ANPR observations may be due to:

1. Difference in accounting for intersection delays. Since ANPR cameras are installed after intersections, the delay at the downstream intersection is included in the observed travel time by ANPR system. On the other hand, in the case of the FCD-based estimates, intersection delays are spread over intermediate links before and after the intersection. Therefore, in the FCD-based method, in
Table 2: Descriptive statistics for $\hat{\mu}_\eta$, $\hat{\mu}_\phi$, $\hat{\mu}_\lambda$, $\hat{\sigma}_\eta$, $\hat{\sigma}_\phi$, and $\hat{\sigma}_\lambda$ (over 27 routes).

<table>
<thead>
<tr>
<th></th>
<th>Congested period</th>
<th>Non-congested period</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\mu}_\eta$</td>
<td>0.33</td>
<td>0.35</td>
</tr>
<tr>
<td>$\hat{\mu}_\phi$</td>
<td>0.50</td>
<td>0.48</td>
</tr>
<tr>
<td>$\hat{\mu}_\lambda$</td>
<td>0.24</td>
<td>0.63</td>
</tr>
<tr>
<td>$\hat{\sigma}_\eta$</td>
<td>0.31</td>
<td>0.25</td>
</tr>
<tr>
<td>$\hat{\sigma}_\phi$</td>
<td>0.12</td>
<td>0.30</td>
</tr>
<tr>
<td>$\hat{\sigma}_\lambda$</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>stddev</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td>min</td>
<td>0.16</td>
<td>0.14</td>
</tr>
<tr>
<td>max</td>
<td>0.64</td>
<td>0.69</td>
</tr>
<tr>
<td>mean</td>
<td>0.35</td>
<td>0.30</td>
</tr>
<tr>
<td>stddev</td>
<td>0.17</td>
<td>0.04</td>
</tr>
<tr>
<td>min</td>
<td>0.22</td>
<td>0.57</td>
</tr>
<tr>
<td>max</td>
<td>0.59</td>
<td>0.38</td>
</tr>
<tr>
<td>mean</td>
<td>0.31</td>
<td>0.23</td>
</tr>
<tr>
<td>stddev</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>min</td>
<td>0.11</td>
<td>0.15</td>
</tr>
<tr>
<td>max</td>
<td>0.23</td>
<td>0.22</td>
</tr>
<tr>
<td>mean</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td>stddev</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>min</td>
<td>0.08</td>
<td>0.22</td>
</tr>
<tr>
<td>max</td>
<td>0.19</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table 3: Regression summary for the impact of FCD bias corrections.

<table>
<thead>
<tr>
<th></th>
<th>Congested period</th>
<th>Non-congested period</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>1.42</td>
<td>0.37</td>
</tr>
<tr>
<td>Std. Error</td>
<td>0.61</td>
<td>0.20</td>
</tr>
<tr>
<td>t-value</td>
<td>2.33</td>
<td>1.82</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.86</td>
<td>-0.08</td>
</tr>
<tr>
<td>Std. Error</td>
<td>0.48</td>
<td>0.14</td>
</tr>
<tr>
<td>t-value</td>
<td>-1.80</td>
<td>-0.59</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.04</td>
<td>-0.02</td>
</tr>
<tr>
<td>Std. Error</td>
<td>0.23</td>
<td>0.10</td>
</tr>
<tr>
<td>t-value</td>
<td>0.19</td>
<td>-0.20</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-0.81</td>
<td>-0.48</td>
</tr>
<tr>
<td>Std. Error</td>
<td>1.00</td>
<td>0.31</td>
</tr>
<tr>
<td>t-value</td>
<td>-0.82</td>
<td>-1.56</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>1.10</td>
<td>-0.17</td>
</tr>
<tr>
<td>Std. Error</td>
<td>0.96</td>
<td>0.31</td>
</tr>
<tr>
<td>t-value</td>
<td>1.15</td>
<td>-0.56</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-1.38</td>
<td>0.20</td>
</tr>
<tr>
<td>Std. Error</td>
<td>0.96</td>
<td>0.34</td>
</tr>
<tr>
<td>t-value</td>
<td>-1.44</td>
<td>0.59</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>-2.82</td>
<td>-0.20</td>
</tr>
<tr>
<td>Std. Error</td>
<td>1.68</td>
<td>0.59</td>
</tr>
<tr>
<td>t-value</td>
<td>-1.69</td>
<td>-0.33</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.32</td>
<td>0.30</td>
</tr>
<tr>
<td>$F$</td>
<td>1.68</td>
<td>1.46</td>
</tr>
</tbody>
</table>

some cases, part of the delay that belongs to the section before the camera location is assigned to the section after that. This may result in minor underestimation of actual travel times by the FCD-based approach.

2. **Data from multiple cameras for the same route.** The current implementation of the ANPR system in Stockholm defines a route as a path between two sites as opposed to a path between two cameras. As illustrated by the example in Figure 10, a route is defined between Site 1 and Site 2. The reports from Camera 2 and Camera 3 are grouped together and considered as reports of Site 2. This causes the problem of mixing the travel times of through and left turning traffic, which can be different in cases such as Figure 10, where left turns are prohibited and vehicles have to detour around a block instead. Such a mixture of travel times can result in overestimation of actual route travel times by ANPR (although extreme observations have been removed by the ANPR noise removal method).

![Figure 10: An example of a mixture of travel times distributions.](image)

3. **Measurement errors.** Errors in GPS measurements, map-matching/path inference processes, and ANPR measurements.

4. **Variation in drivers’ behavior and vehicle mix.** The FCD used in this application are generated by taxis, which is a subset of the population of vehicles. Taxis are not necessarily representative of the total population of vehicles in a city. For example, taxis are often allowed to take advantage of facilities with public transport priorities (e.g. dedicated bus lanes). Besides, taxi drivers normally know the city better than average drivers resulting in more aggressive driving and lane selection, especially during congested periods. Thus, taxis are expected to be faster than the general population, especially in congested periods. The next section examines this issue in greater detail.
5.3. Correction for non-representative vehicle sample

A regression analysis is carried out to investigate the relation between route attributes and the deviation of the taxi mean travel times from those of the entire population observed by the ANPR system. Theoretical considerations and regression results suggest that a multiplicative model structure based on the relative deviation between the mean travel times from ANPR and FCD is suitable. A number of route attributes were examined as explanatory variables (Table 4). Products of the attributes are also used to test whether interactions are present. The same set of 27 routes is used for the analysis and 15-minute intervals are grouped to form the two congested and non-congested classes. The regression

\[
\log \left( \frac{\hat{\mu}_T}{\mu_T} \right) = \beta_0 + \beta_1 (\kappa_r + 1) b_r + \varepsilon_r \quad r = 1, \ldots, 27
\]

The model specification is consistent with traffic regulations in Stockholm which allow bus lane use by taxis. This is expected to be advantageous during congested hours since taxis travel faster than normal traffic, but of little value during non-congested hours. The interaction between the bus lane ratio and the number of traffic signals is used as a proxy for the number of signals with bus priority along the bus lane. To allow a bus lane to impact travel time even in the absence of traffic signals, the variable \((\kappa + 1) b\) is used.

Table 5 summarizes the regression analysis for congested and non-congested periods. During the congested period FCD travel times are about 93\% (=1/\(e^{0.068}\)) of the corresponding ANPR travel times for routes with no signals and bus lanes and as low as 60\% for routes with signals and bus lanes. During the off-peak periods it ranges between 93\% and 99\% of the corresponding ANPR travel times. In congested periods, the bus lane-signal interaction term is highly significant in explaining the difference between ANPR and FCD travel times. Under non-congested conditions, the model reduces to a simple shifting of FCD travel times. This does not mean that traffic signals are not important for travel time estimation, but rather that they impact FCD and ANPR data similarly when the network is not congested as Table 5 shows. Under non-congested conditions the advantage of bus lanes is reduced and the opportunities for signal priority are infrequent. The overall goodness-of-fit of the regression, shown in Table 6, is also assessed with cross-validation of the model using the leave-one-out method. 27 regressions are performed, each time leaving out one route, performing regression on other routes.
and using the model to estimate the travel time for the route excluded from the estimation. As shown, the average errors are reduced after regression for both congested and non-congested periods, with more significant reduction of error in congested periods. $U^M$ for the congested period is reduced from 0.34 to almost zero indicating a drastic reduction of systematic bias. The models are stable when the leave-one-

<table>
<thead>
<tr>
<th></th>
<th>RMSE (min)</th>
<th>RMSNE</th>
<th>$U$</th>
<th>$U^M$</th>
<th>$U^S$</th>
<th>$U^L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.43 (0.59)</td>
<td>0.08 (0.10)</td>
<td>0.05 (0.07)</td>
<td>0.00 (0.15)</td>
<td>0.02 (0.19)</td>
<td>0.97 (0.66)</td>
</tr>
<tr>
<td>Non-congested</td>
<td>0.21 (0.23)</td>
<td>0.06 (0.06)</td>
<td>0.03 (0.03)</td>
<td>0.01 (0.10)</td>
<td>0.16 (0.23)</td>
<td>0.83 (0.67)</td>
</tr>
<tr>
<td>Congested</td>
<td>0.72 (1.03)</td>
<td>0.12 (0.16)</td>
<td>0.07 (0.11)</td>
<td>0.00 (0.34)</td>
<td>0.01 (0.10)</td>
<td>0.99 (0.56)</td>
</tr>
</tbody>
</table>

out method is used. For the non-congested class, $\beta_0$ is in the range $[0.00, 0.01]$ and $\beta_1$ is stable around 0.01. For the congested class, $\beta_0$ is in the range $[0.06, 0.07]$ while $\beta_1$ is in the range $[0.06, 0.09]$.

### 5.3.1. Hourly correction for non-representativeness bias

As an alternative to congested/non-congested periods (which were defined individually for each route), the regression analysis can be performed in a more granular way, for each hour of the day based on the formulation in Equation 29. The results from 16 regression models (one per hour from 6 a.m. to 10 p.m.) show that the importance of the bus lane ratio variable varies during the day. Consistent with a-priori expectations, the bus lane ratio becomes more significant in peak hours and less significant in off-peak hours.

Both the two-period (congested/non-congested) and the hourly approaches can be used to correct the FCD-based estimates. The corrections for the example routes are shown in Figure 11. In the hourly approach, travel time gradually increases (decreases) before (after) the two peaks, while the two-period model generates relatively higher gradient at the boundaries of the peaks (see, e.g. route 64, 95, and 116). In general, the correction for the non-representativeness bias reduces the error significantly during congested hours and has less effect during the non-congested hours.

Each of the two approaches, congested/non-congested periods and hourly periods, has its own advantages and disadvantages. The former needs the definition of congested periods specifically for each route. This requires knowledge of traffic conditions on that route during the day which may be difficult to acquire with no alternative data sources (such as ANPR) available. On the other hand, the latter assumes that the hour of the day captures the major characteristics and hence can be used for arbitrary routes. As such, the model does not take into account the fact that congestion may occur at different times for different routes.

The regression-based correction of FCD-based travel times can also be applied to correct other statistics such as percentiles and higher order moments.

### 6. Computational performance

Route travel time estimation can be performed off-line if routes are predefined. This may not be the case, however, in interactive/real-time applications. For example, a request arrives at a travel planner for alternative paths for a specific OD pair. In such requests, assuming that a set of reasonable paths has been predefined for each OD pair, the travel time (and its attributes) may be determined on the fly. Similarly, a request may require the travel time on a specific route (e.g. for monitoring purposes). Traditional link-based estimation methods can have good performance in on-line applications because link travel time profiles can be computed in advance, and the mean travel time of any arbitrary route can be computed on demand from link mean travel times. In contrast, the proposed route-based method computes the mean and other statistics of a given route directly from the map-matched FCD. Hence, it is important for the method to be efficient in order to be used in low-latency applications such as journey planners.

The computational performance of the proposed method for one route (in this case 3.4 km long) is estimated by processing several data sets with different sizes. The method was implemented in Java and tested on a machine with a 2.2 GHz Intel Core i7 CPU and 8 GB memory. The road network from Section 4.1 and data with different sizes representing the equivalent to 3 months up to 5 years of FCD observations were used. Figure 12 illustrates the computational time of estimating travel time of a route, 3.4 km long, using various FCD sizes.
Figure 11: Corrected (hourly and two-period) mean travel times and comparison to corresponding ANPR travel times.
The performance is measured for two parts of the process separately: a) concatenation, allocation, scaling, and weighting together (i.e. steps before aggregation); and b) aggregation (including calculation of the mean, percentiles, and other statistics). The reason for evaluating the performance of these two parts separately is that the pre-aggregation processes can be decomposed and performed in parallel if needed, and it is thus beneficial to know how long it takes on its own.

As the size of the input (FCD) $n$ increases the computation time increases quadratically, or $O(n^2)$, but with small coefficients for the second order term. As illustrated in Figure 12-a, a data set of ~1 million trajectories takes about 5 seconds for transformation and weighting. As mentioned before, concatenation decreases the number of effective observations used in the calculation. For example, in the case study of 27 routes about 35% reduction is observed after concatenation (averaged over all routes). In addition, only trajectories that overlap with the route of interest are used in the aggregation step, resulting in a smaller input size for aggregation compared to the initial set (1 million). Allocation and weighting of one million trajectories result in 50k records. It takes about 4 seconds to aggregate them (see Figure 12-b). The time for calculation of the correction factors $\lambda_k$ (Equation 16) is included in the aggregation time as opposed to the weighting time. Although $\lambda_k$ is part of the weighting, it has to be calculated in the aggregation phase because it requires $N_k$ (the total number of observations covering link $k$) which is known only in this phase.

\[ f(x) = a x^2 + b x + c \]
\[ a = 5.16e-12 \]
\[ b = 1.97e-06 \]
\[ c = 0.232 \]

\[ f(x) = a x^2 + b x + c \]
\[ a = 1.85e-09 \]
\[ b = 1.23e-05 \]
\[ c = 0.024 \]

Figure 12: Computation time as a function of size of the input for (a) transformation and weighting; and (b) aggregation for one route.

In conclusion, the computational performance of the method is sufficient for real-time applications such as on-demand information services which require estimation of travel time distribution of any arbitrary route.

Although a single instance of the process can handle millions of trajectories in a matter of seconds, even more throughput can be achieved in more demanding applications where there may exist billions of trajectories through parallel processing. As illustrated in Figure 13, the pre-aggregation steps can be executed independently for each individual observation. Thus, a large data set can be decomposed and processed by multiple instances of the program in parallel. The output of these processes can then be aggregated using one aggregator per time interval.

7. Conclusion

The paper presented a non-parametric method for the estimation of the distribution of travel times along routes from low-frequency FCD. The main objective of the method is to provide estimates not only of the mean but also any statistics of the route travel time distribution. FCD have several sources of bias, including incomplete and uneven coverage of the route, and partial coverage of the adjacent network. The method involves a number of steps designed to reduce the impacts of these factors. It is designed to be efficient for real-time, on-demand applications such as trip planning services. Furthermore, many steps in the calculation procedure can be performed in parallel which reduces computing time further.
The paper also discusses correction of the bias due to non-representativeness of the FCD sample, when other data sources, such as ANPR, are available.

The results of the case study demonstrate that the proposed method is able to accurately estimate the mean and other statistics of the travel time distribution, such as the median and various percentiles. Also, the proposed method performs equally well or better than the link-based approach when the mean route travel time is compared to corresponding results from ANPR data. In conclusion, FCD can complement stationary sensors, such as AVI systems, to enable large scale estimation of travel time and travel time variability.

A major issue in the area of big data and opportunistic sensors is the lack for ground truth, especially in the field of transportation. Hence, comparisons of alternative data sources, despite their limitations, can be very instructive. A good example, in the case of this study, is how the differences between ANPR and FCD data led to the quantification of the impact of bus lanes. An important direction of future research is the fusion of alternative data sources to improve the estimation accuracy. Understanding the comparative advantages of each source can be very useful in that regard. In this paper, the data is clustered by time of day (using 15-minute intervals). Another direction of future research is to apply higher resolution of clustering by considering, for example, within week variations, weather conditions, etc.

In summary, the paper makes a number of contributions: it identifies the main biases introduced when FCD are used for the estimation of route travel times and proposes methods for their correction. It compares the travel time distribution based on FCD to corresponding estimates from ANPR data and provides insights into their differences. Finally, the method is computationally efficient. A distributed implementation is also proposed that can support real time applications.

Acknowledgments

The authors would like to thank the Swedish National Transport Authority for funding part of this work through the Mobile Millennium Stockholm project and Trafik Stockholm for their support and provision of data. Many useful comments and suggestions from the reviewers and colleagues at the Swedish ITS Graduate School are also acknowledged.

References


