ESTIMATING ROUTE TRAVEL TIME VARIABILITY FROM LINK DATA BY MEANS OF CLUSTERING

Wouter R. J. Charle

Francesco Viti

Chris M.J. Tampère

Centre for Industrial Management

Section Traffic and Infrastructure

Faculty of Mechanical Engineering

Katholieke Universiteit Leuven

wouter.charle@cib.kuleuven.be

ABSTRACT

Accurate route travel time estimation is today one of the most challenging problems in traffic theory. This research proposes a novel method for the estimation of route travel time distributions, based on historical link travel time observations. Central in the development of this framework is the distinction between (cheap) off-line storage and computations and (expensive) on-line computations. For that it is important to minimize the on-line computational effort of calculating a route travel time histogram. The key elements in the method are correlations in link travel time fluctuations and a clustering algorithm. Tests on the Belgian road network show that 1) the clustering method is on-line computationally efficient while keeping the off-line storage under control, 2) that it can accurately estimate route travel time distributions and 3) that it can be applied successfully for route reliability estimation.

Keywords: travel time variability, clustering, data mining, route information

INTRODUCTION

Transportation networks are facing more and more congestion issues all over the world. Under these conditions, the adoption of traffic management measures aimed at resolving traffic issues and at using the available network capacity with greater efficiency is a necessary (short-term) policy to guarantee a sustainable transportation system. To achieve this goal, it is necessary to estimate and predict how congestion affects the performance of a network in order to provide the road users with optimal advices and guidance, and to identify which management strategies to apply. However, the continuously growing demand for mobility and the increasing complexity of transportation networks (e.g., more travel alternatives, more information sources) make the estimation and prediction of congestion effects one of the most challenging problems in traffic theory.

One of the key issues is that congestion effects are typically local, i.e. they can be observed and confined within certain parts of the network and affect a limited number of links. However, their consequences can propagate through a much larger part of the network, since a local link travel time variation can be experienced by different users using different routes, whose reaction can also cause changes in the traffic patterns elsewhere. For that, even if congestion effects are observable and quantifiable at the link level, it is important to extrapolate their effects onto the route level.

Another key aspect which motivates this study is that congestion causes two undesirable effects to the transportation users: on one hand it causes an increase of expected travel time due to a reduction of the mean speed and queuing, and on the other hand increases the variability of travel times. Since the decisions of both traffic managers and road users rely on accurate and reliable travel time estimation and prediction, knowing the variability of travel time might be as important as knowing its expectation value. However, the mainstream research on travel time estimation concentrates on the estimation of mean route travel time or some measures of travel time reliability (i.e. 10th and 90th percentiles). As a consequence, route travel time variability is rarely considered in traffic applications, and route choice models often favour alternatives on the basis of shortest travelling distance or expected travel time. The research described in this paper aims to show that, given the growing availability and detail of travel time measurements, it is also possible to estimate route travel time distributions. This serves a broader spectrum of applications and provides more useful information.

Although new data sources, able to provide route travel time information directly, are becoming more and more popular (e.g., GPS data, Automatic Vehicle Identification technologies), they are often insufficient in size to derive statistically significant estimates for all system users. On the other hand, transversal aggregate data is more easily available in practice, through e.g. inductive loop detectors, it is a cheaper way to collect data and it can provide full statistics of flows at specific road sections. Although the calculation of link travel time distributions is possible through these data sources, the relationship between these distributions and travel time distributions at the route level is not straightforward. Because link flows and travel times are often highly mutually correlated, the calculation of route travel time variability can hardly be approximated by simply summing up the link travel time variability.

This research presents a new method to calculate route travel time histograms, based on historical link travel time observations. A computationally efficient procedure is proposed to relate link travel time variability to route travel time variability. Clustering techniques are used to combine successive links into clusters of links (hereafter called clusterlinks) such that the link travel times correlations between subsequent clusterlinks are minimized. This enables one to aggregate link travel time distributions into route travel time distributions ignoring only arbitrarily small travel time correlations between adjacent links.

It is important to stress that we aim to quantify link travel time correlations that are not caused by within-day and day-to-day fluctuations. The correlations we derive from the clustering approach will provide an indication of how congestion effects are expected to propagate onto the traffic system, for instance if an accident has occurred. In this sense link correlations are strongly related to the structure of the network and the connectivity between links.

Many applications could employ the results of the method presented in this paper. They can be used for instance in any route selection criteria, e.g., to quantify route travel time uncertainty in the travellers' decision making, to provide route advice and guidance with some form of confidence bounds, to consider trip chain strategies etc. Moreover, this information can help to improve travel time estimation and prediction models, or to provide performance indicators, such as reliability or uncertainty measures. It is also important to stress that the method proposed in this study can also be deployed beyond road traffic engineering applications. Other applications can be for instance the location optimization of logistic hubs (i.e. airports, packaging services ...) or public services (i.e. hospitals, fire stations ...) and the evaluation of any network performance (e.g., electric, social and internet networks).

In this paper the new clustering algorithm is incorporated in a robust route guidance framework. The aim is to allow the development of an advanced route planner that is able to optimize route choice reliability. The notion of reliability is defined by the end-user of the route planner and highly depends on the properties of the route travel time distribution.

Outline The remainder of this article is organized as follows. The second section gives a literature overview of the methods proposed to derive travel time variability and their application to route planning. A description of the data is given in the third section, followed by a description of the proposed method. The results on a real traffic network are discussed in the fourth section. Finally, the last section gives the conclusions.

LITERATURE REVIEW

Travel time variability is widely acknowledged as one of the main factors in determining the route preferences of road travellers. Abdel-Aty et al. (Abdel-Aty, Kitamura, & Jovanis, 1996) reported in a survey that 54% of daily commuters indicated travel time variability as the most important or the second most important factor for choosing their route. Small et al. reported in (Small, Noland, Chu, & Lewis, 1999) that the value of one minute of travel time saving is comparable to a minute reduction in travel time variability in route choice. From the perspective of the individual drivers, route travel time variability has much more importance than link travel time variability, since drivers are interested in knowing primarily how long a trip will last, or at what time they will arrive at destination. It is therefore important, when modelling the choice process of drivers, to provide a measure of route travel time variability.

The interest in methods for quantifying route travel time variability has also been growing since the increasing development of Intelligent Transportation Systems (ITS) as short-term measures to manage congested traffic networks (e.g. route guidance, Advanced Traveller

Information Systems, etc). However, most of the research has focused on determining long-term variability, such as peak-hour, daily, weekly and seasonal variability. Many studies have assumed route travel times to be well approximated by Normal or Log-Normal distributions, under the condition that link travel times are (sufficiently) independent variables on a day-to-day time scale (Noland & Polak, 2002), or that the network is under user equilibrium constraints (e.g. (Clark & Watling, 2005), (Davidson & Teye-Ali, 2008)). This assumption is acceptable in slow-varying and uncongested networks, while there is strong evidence of significant correlations between link travel times in a route otherwise (He, Liu, Kornhauser, & Ran, 2002).

Although in uncongested sections route travel time variability is a simple accumulation of each single link travel time variability, this assumption inevitably underestimates route travel time variability in general since it would neglect their correlation1. Despite this fundamental shortcoming most of the available methods neglect this dependency for the sake of computational tractability (e.g., (Lecluyse, Van Woensel, & Peremans, 2009), (Lu, 2001)) or calculate the correlations using approximate analytic functions (e.g., (Fu & Rilett, 1998)), or using a traffic simulation model (e.g., (He, Liu, Kornhauser, & Ran, 2002)), which still remains a simplified representation of real traffic patterns. Finding these correlations using real data is much more complex, as correlations are intermixed with other stochastic factors (within-day and day-to-day variations, external factors such as weather, incidents etc.).

In many applications a single measure of travel time variability may not be sufficient. He et al. (He, Liu, Kornhauser, & Ran, 2002) indicate that although mean and variance contain the most important information about path travel time, finding the single route with expected shortest travel time is not appropriate for routing when planners are not risk neutral. The entire travel time distribution contributes to the routing choice (Lecluyse, Van Woensel, & Peremans, 2009). However, the difficulty to derive some measure of route travel time variability from link-based data grows considerably if one aims to derive the complete distribution of travel times. In this respect there is very limited research, often bound to some specific application. For instance, Hellinga and Fu (Hellinga & Fu, 1999) and Key (Key, 2005) proposed methods to derive minimum travel time distributions (i.e. using only data from shortest paths) for application to respectively route choice in stochastic networks and to route guidance systems. Krishnamoorthy (Krishnamoorthy, 2008) proposed instead a method to derive route travel time variability from link travel time distributions, which is applicable only to signalized arterial corridors.

In this paper we propose an important step forward into the calculation of route travel time distributions from link travel time data. The method proposed in this paper is not bound to any application domain. The remainder of this paper will describe the data used and the proposed method, together with some preliminary results on a real case study.

METHODOLOGY

The focal point of this research is the on-line calculation of a route travel time distribution based on historical link travel time observations. This will improve the stochastic shortest

[1] It is well known in statistics that the variance of the sum of two or more random variables is equal to the sum of the variance of each variable plus twice the covariance of each couple of variables.

path algorithms as developed by Hellinga and Fu (Hellinga & Fu, 1999) and Key (Key, 2005) since more detailed route information is available. Moreover, the travel time distribution can be calculated for any route (i.e. not only shortest paths), and routes can be compared based on their travel time distribution. Consequently an optimal route is not necessarily a shortest route, neither is it always the fastest route. This is illustrated in Section 4.3
For the on-line calculation of a route travel time distribution based on historical link travel time observations, two cases can be distinguished:

- ad hoc: the route travel time distribution is calculated while searching a route
- ex post: the route is already known when calculating the histogram

The ad hoc calculation of the route travel time distribution has the advantage that the properties of the distribution can be optimized while calculating the route. This is because the statistics of several sections of the network can be compared in order to construct a route for which the travel time distribution is optimized (e.g. the variance is minimized). The result of the ad hoc calculation is 1 optimal route. This advantage is not present in the ex post calculation of a distribution of route travel times because a set of routes has to be provided a priori. This set of routes can for example be computed with a k-shortest-paths algorithm using simple weights (e.g., mean travel time, distance, etc.). The ex post calculation consists of calculating the travel time distribution for each of the routes. Based on these probability distributions the optimal route can be selected. The drawback of an ex post calculation is that the selected route will be 'an' optimal route, but not 'the' optimal route. The advantage is that the optimization criteria do not have to be know in advance. The ex post calculation of the distribution is in this way a simpler problem, but is non-trivial when on-line computational efficiency is an issue. In this paper we focus on the procedure to derive these distributions, not on the routing procedure. For this reason we will refer mainly to the ex post case.
In what follows, both a naïve approach as adopted by Lecluyse in (Lecluyse, Van Woensel, & Peremans, 2009), and an approximative approach to the ex post calculation of travel time distributions are discussed. It is shown that the approximative method improves the computational efficiency of the naïve method by pre-processing the link travel time data off-line. The approximative method can be applied to networks of any size.

Notation

In this section the notation used to describe the methodology is defined. A network contains L links λ indexed by i. For each of the links λ_i there is a collection of historical travel time observations $TT_{i,j}^d$ on T times τ indexed by j for D different days δ indexed by d. A route R, indexed by r, is a set of H_r links λ_i. The travel time of route R_r on day δ_d and time τ_j is denoted by $TT_{r,j}^d$. The empirical travel time distributions are approximated by histograms of B bins. The notation is summarized in Table 1.

Table 1: Summary of the notation

L	The number of links in the network
λ_i	A specific link, indexed by i

D	The number of observation days
δ_d	A specific day, indexed by d
T	The number of observations on 1 day
τ_j	A specific observation time, indexed by j
R_r	A route in the network, indexed by r
H_r	The number of links in route R_r
$TT_{i/r,j}^{\mathrm{d}}$	The travel time for link λ_{i} or route R_r, on day δ_d and time τ_j
G_r $(\leq H_r)$	The number of (cluster)links in route R_r
B $(\leq T \times D)$	The number of bins in a travel time histogram

Approaches

A first straightforward but naïve approach makes direct use of the available historical travel time data to calculate a route travel time distribution. The setup of a system that implements such an approach is illustrated in Figure 1. In this setup a route-guidance framework sends a query to the statistics system in order to obtain the travel time distribution for a specific route R_r. The statistics system calculates the distribution based on the historical data contained in a database and returns the distribution to the calling system.

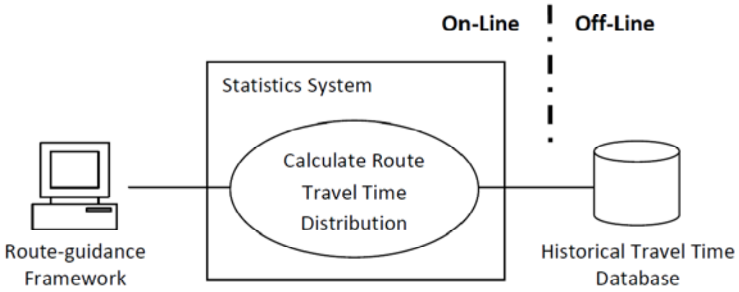
On-Line
Off-Line
Statistics System
Calculate Route Travel Time Distribution
Route-guidance Framework
Historical Travel Time Database

Figure 1: Architecture of the naïve approach. A route-guidance system sends a query to the statistics system in order to obtain the travel time distribution for a specific route. The statistics system calculates the distribution based on the historical data contained in a database. All calculations are done on-line.

Not considering temporal propagation of travel time fluctuations explicitly, the route travel time histogram is calculated from the sum of instantaneous travel times $TT_{i,j}^{d}$ over all links λ_i contained in the route R_r for each observation time:

$$TT_{r,j}^{d} = \sum_{i:i \in R_r} TT_{i,j}^{d} \tag{1}$$

The computational complexity of this approach is dominated by the summations in (\ref{eq:naive}) and are of order $H_r \times T \times D$. Considering that in practical applications this has to be done for each route in a set of relevant routes, it is easy to understand that this is not a feasible approach for an on-line application. The off-line pre-computation of all route statistics is neither an option because the total number of routes is at least a quadratic function of the number of links in the network.[2] Consequently shifting the naïve approach to an off-line environment results in a storage capacity increase by at least power 2.

With the second approach a method is developed to shift computational effort partially off-line while keeping the data storage size under control. This is done on one hand by pre-processing the travel time data off-line and on the other hand by using clustering techniques to group links based on the network topology and correlations between link travel time data. In the remainder this approach is also referred to as the clustering approach.

The setup of a system implementing the clustering approach is illustrated in Figure 2. In the on-line window of this setup, similar to the implementation of the naïve approach, a robust route-guidance framework sends a query to the statistics system to obtain the travel time distribution for a specific route. The statistics system calculates the distribution based on the statistical data contained in a statistics database. The statistics database is created and updated off-line by the clustering system. The clustering system uses the data contained in the historical travel time database to calculate the travel time histograms, using correlations and a clustering algorithm, and stores them in the statistics database.

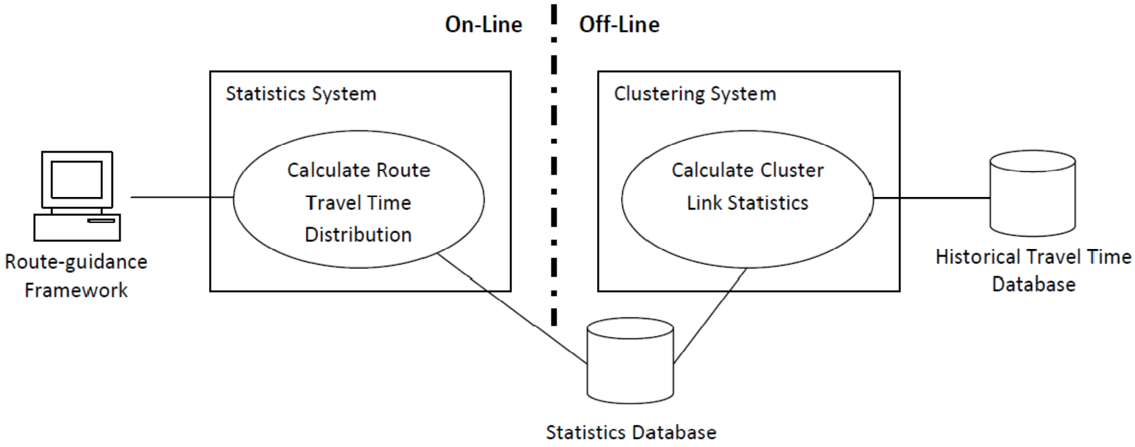
On-Line
Off-Line
Statistics System
Calculate Route Travel Time Distribution
Clustering System
Calculate Cluster Link Statistics
Route-guidance Framework
Historical Travel Time Database
Statistics Database

Figure 2: Architecture of the clustering approach. On-line (left): A robust route-guidance framework sends a query to the statistics system to obtain the travel time distribution for a specific route. The statistics system calculates the distribution based on the statistical data (i.e. averages, standard deviations, histograms) contained in a statistics database. Off-line (right): The statistics database is created and updated off-line by the clustering system. The clustering system uses the data contained in the historical travel time database to calculate the travel time histograms, using correlations and a clustering algorithm, and stores them in the statistics database.

By pre-processing the travel time data to travel time histograms, the amount of data required during the on-line computation of a route travel-time distribution is reduced from $T \times D$ to B. However, because the statistical properties of the sum of observables is not equal to the sum of these properties if the observables are statistically dependent, statistical dependencies between travel time observations on different links have to be taken into account. While this has no consequences for expected value calculations, neglecting these dependencies will result in an incorrect estimation of higher moments.

Most often the observed travel times of two subsequent links in a network are not statistically independent. These statistical dependencies can be due to, for instance, congestion spillback over links, similar road conditions (e.g. due to the weather, incidents, etc.), or other structural similarities. The dependencies between link travel time statistics are in this

[2] The number of routes considering a simple sequence of L links in series is $\frac{L^2+L}{2}$.

approach removed by redefining the network. This is done off-line by means of a clustering algorithm which defines new (artificial) links as a combination of links that have correlated travel time fluctuations. In this way the correlations between each cluster of links are minimized.

The travel time histogram of each cluster of links can be therefore calculated off-line by making use of the naïve approach. The clustering method is used because the route histogram is calculated as the convolution of the histograms of the clusterlinks, and the calculation of a density function of the sum of two stochastic variables by convolving their density functions requires that these variables are statistically independent. The clustering approach is thus an approximative method in which the quality of the approximation is determined by the extent to which the correlations between clusterlinks can be minimized. As a result the route travel time histogram is an approximation to the real route travel time distribution.

Because a clusterlink is the combination of 1 or more links, the length of a route R_r is reduced to $G_r \leq H_r$ clusterlinks. To calculate the convolutions efficiently, the histograms are stored in the database as Fourier Transformations (FT). The advantage of using the FT of the histograms is that, by the convolution theorem, the route travel time histogram is obtained as the inverse FT of the multiplication of the FT-ed clusterlink histograms instead to their convolution. For that the complexity of the on-line computation is of order $B \times (G_r + \log(B))$, with the $B \log(B)$ term due to the inverse FT.

The computational complexity of both approaches is compared in the following numerical example. Suppose the route contains $H_r = G_r = 100$ links, the histogram contains $B = 25$ bins and that there are $T \times D = 1440$ travel times observations for each link. It is found that the second approach (order $\sim 3.10^3$) outperforms the first approach (order $\sim 1.10^5$). This is because in this test case $B - T \times D < 0$. The performance of the second approach improves further as $G_r < H_r$. The naïve approach performs equal to or better than the clustering approach only if

$$\frac{B}{H_r}(G_r + \log(B)) \geq T \times D \qquad (2)$$

In general, even if $G_r = H_r$, this condition is not satisfied since $B - T \times D < 0$ must always be valid to calculate the empirical travel time histogram correctly.

The clustering algorithm

The clustering algorithm is used to combine successive links into clusterlinks such that the correlation between subsequent clusterlinks is minimized. As such, the travel time statistics of any 2 subsequent clusterlinks in the resulting network are approximately linearly independent. To identify clusters in the network the following heuristic is proposed. The clustering algorithm considers all links λ_i in the network and, assuming instantaneous travel times, calculates the correlation[3] C_{ik} with each successive link λ_k.

[3] Temporal propagation of travel time fluctuations is not considered explicitly and is subject of future research.

Calculating the correlations, it is important to capture the temporal travel time fluctuations around the stationary travel time pattern to identify dependencies between any 2 successive links λ_i and λ_k. Else the correlation will be large between most of the links because morning and evening peak effects in travel time data are universal accross the majority of links. To ensure unbiased correlations, it is necessary to transform the link travel time data into a symmetric distribution. For that the following assumption is made about the distribution of link travel times. Knowing that travel time observations on a link λ_i are bound from below by some minimal travel time θ_i and do virtually not have an upper bound, it is assumed that they are distributed as

$$TT_{i,j}^{d} \sim LogN\left(\mu_i, \sigma_i^2\right) + \theta_i, \tag{3}$$

where μ_i is the mean and σ_i^2 is the variance of the travel time distribution of link λ_i. This is in line with Noland et al. (Noland & Polak, 2002). By assumption (3) and defining

$$TT^{*}{}_{i,j}^{d} = \log\left(TT_{i,j}^{d} - \theta_i\right), \tag{4}$$

a vector of transformed travel times $\widetilde{TT}_{i,j}^{d}$ can be constructed:

$$\widetilde{TT}_{ij}^{d} = \frac{TT^{*}{}_{ij}^{d} - \frac{1}{D}\sum_{d=1}^{D} TT^{*}{}_{ij}^{d}}{\sum_{j=1}^{T}\sum_{d=1}^{D}\left(TT^{*}{}_{ij}^{d} - \frac{1}{D}\sum_{d=1}^{D} TT^{*}{}_{ij}^{d}\right)}. \tag{5}$$

In this transformation, the travel times are explicitly centralized around the historical mean (assumed stationary) travel time for each observation time τ_j. This is necessary to capture travel time fluctuations instead of travel time patterns caused by day to day fluctuations. The correlation of travel time fluctuations C_{ik} between any 2 subsequent links λ_i and λ_k is now readily calculated as the inner product of $\widetilde{TT}_{i,j}^{d}$ and $\widetilde{TT}_{k,j}^{d}$

$$C_{ik} = C_{ki} = \sum_{j=1}^{T}\sum_{d=1}^{D} \widetilde{TT}_{ij}^{d}\widetilde{TT}_{kj}^{d}. \tag{6}$$

The correlation can be interepreted geometrically as the angle between 2 vectors and lies in the range $[-1,1]$ because of the normalization in equation (5).

Having calculated the correlation C_{ik}, it is tested against the hypothesis H$_0$: $C_{ik} = 0$ to minimize the probability of a Type II error. This means that Pr(H$_0$ is accepted | $C_{ik} \neq 0$) is reduced to some specified value $0 \leq p \leq 1$. Link λ_i and λ_k are combined into a cluster of links if H$_0$ is rejected (i.e. it is possible that $C_{ik} \neq 0$). The statistics of this new clusterlink are obtained from the sum of the instantaneous travel times of both links for each τ_j and δ_d. Each such combination of links is added to the network and expanded until H$_0$ cannot be rejected (i.e. no links can be added to the cluster of links). Remark that assumption (3) has no influence on the functional form of the resulting clusterlink travel time distribution as it is only used to calculate correlations correctly.

Based on the function of the node between 2 links (i.e. traffic passage, origin or destination node), some links become redundant. Links that have become redundant due to the clustering are removed from the network. In this way the network is restructured such that the statistics of each 2 subsequent clusterlinks in the network are approximately linearly independent. The heuristic of the clustering algorithm is illustrated on a fundamental clustering diagram[4] in Figure 3.

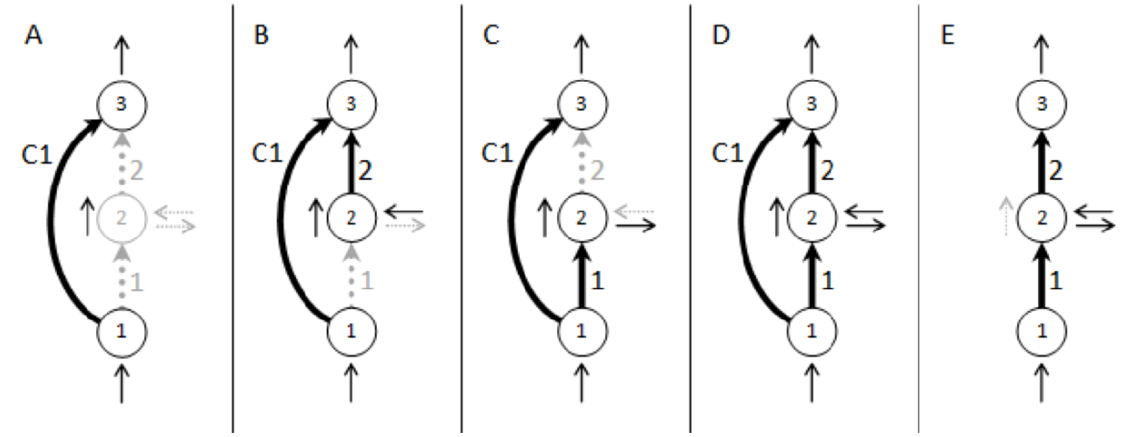

Figure 3: Fundamental clustering diagram of 3 nodes and 2 correlated links. Depending on the properties of node 2, the clustering of links results in diagram A, B, C, D or E with cluster C1 of links 1 and 2. A: Traffic can only pass by node 2. B: Traffic can enter the network at node 2. C: Traffic can exit the network at node 2. D: Traffic can both enter/exit the network at node 2. E: Traffic can enter/exit the network at node 2 but cannot pass over it, the links are not clustered.

RESULTS

Historical data is used to calculate empirical route travel time distributions.

The traffic data is generated on the Belgian motorway network using the floating car system of Be-Mobile (BeMobile, 2010). The system uses vehicle probes (e.g., taxi's, commercial vehicles, private cars, etc.) that communicate their position frequently to a central system. The individual data samples are processed to generate a traffic state for each individual road segment. The system covers up to 60000 km of Belgian roads and is fully operational since October 2007.

The methodology is tested on the Belgian road network for 6 different routes between Bruges and Leuven. The results in this section have been obtained by a comparison of the results for the routes. The routes are displayed in Figure 4 and all have a length of ≈ 130km (≈ 100 links). The historical travel time dataset has the structure as described in Section 3.1. It consists of observations every 15 minutes between 6:00 and 22:45 ($T = 68$) on 10 Tuesdays ($D = 100$). For certain links and observation times no observations have been made. The gaps in the data occur especially on times where there is low density traffic. Because of this the lacking data will be mostly free flow travel times. The frequency of data gaps varies per road type. Roughly 10% of the data is missing on motorways and up to 50% on rural roads. Using equation (4), a gap in the data on link λ_i and observation time τ_j on day δ_k has been replaced by the historical mean $\sum_{d \neq k} TT^{*d}_{i,j}$ to calculate the correlations correctly. In this way

[4] Any other structure can be reduced to this diagram.

the correlations are not altered by the gaps because $\widetilde{TT}_{i,j}^{d} = 0$. Observation times on which there is no data are not used in the calculation of the link travel time histograms.

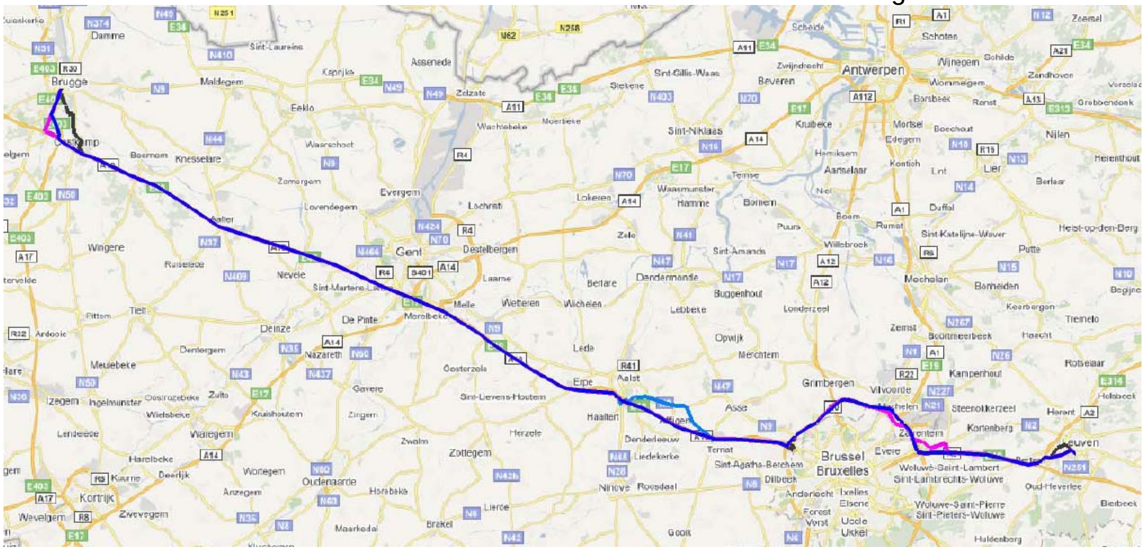

Figure 4: Illustration of the 6 routes between Bruges and Leuven. The routes only differ slightly in regions where congestion frequently arises. These are the exit of Bruges, the E40 section around Affligem, the entrance of the Brussels ringroad, the exit of the Brussels ringroad and the entrance of Leuven.

From an evaluation point of view it is more interesting to examine the impact of clustering on the results as a function of a correlation threshold ρ instead of using the hypothesis test. Using the threshold, links λ_i and λ_k are clustered if $|C_{ik}| > \rho$. The value of ρ ranges from $\rho = 1$ to $\rho = 0.02$.

In what follows, first the impact of clustering on the number of links in a route is examined. After this the variance and the histogram of route travel times is estimated using both approaches and in function of the correlation threshold ρ. Finally the clustering approach is used to compare the route travel time percentiles of the 6 routes.

Route size

The number of links plus clusterlinks in a route is examined for each route in function of the correlation threshold ρ. The results are averaged over all routes and are illustrated relative to the initial number of links in the route in Figure 5. It is found that the total number of links in a route is linearly related to ρ. This is because for lower correlation thresholds, more links are clustered into clusterlinks. As ρ decreases from 1 to approximately 0.5, the links are especially clustered into new clusterlinks. As ρ decreases below 0.5, the size of the clusterlinks begins to increase. This causes the number of clusterlinks in the route to remain stable for $0.3 \lesssim \rho \lesssim 0.5$ and to decrease for $\rho \lesssim 0.3$. The number of clusterlinks never coincides with the total number of links in a route. Moreover, while the total number of links in a route is linearly related to ρ, the route is never replaced by 1 clusterlink. This indicates that the clustering algorithm is not oversensitive to correlation estimation errors since also very small correlations can be obtained from the data, and that the correlations capture meaningful dependencies between adjacent links.

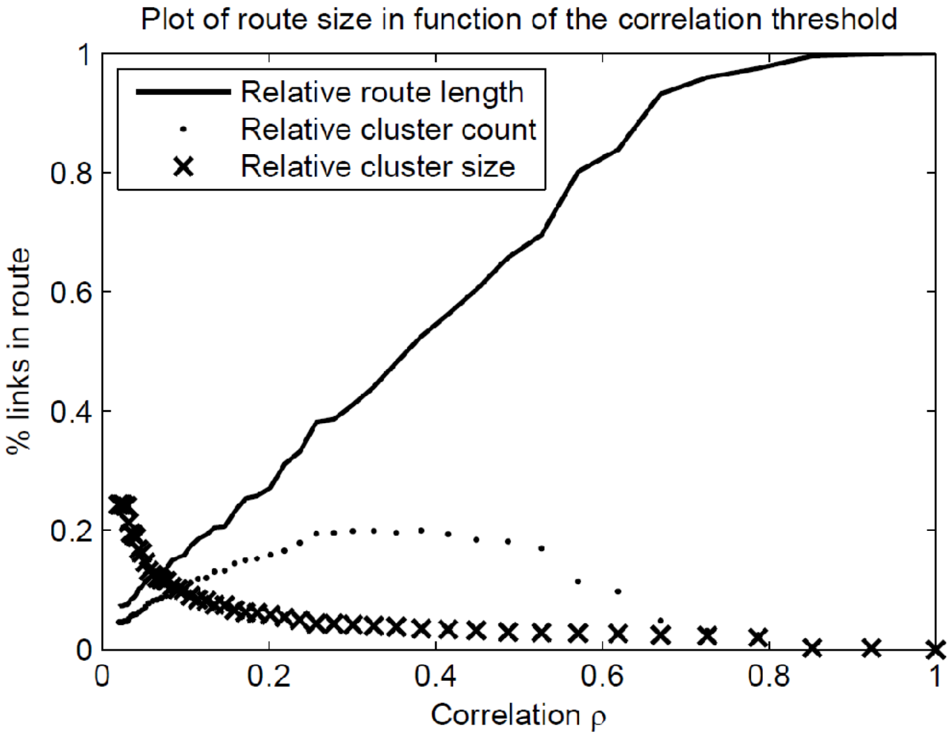

Figure 5: Plot of size effects relative to the initial number of links in a route, in function of the correlation threshold. Blue: Fractional number of links in a route after clustering. Red: Percentage of clusterlinks in the route. Green: Fractional cluster size. From right to left: As the correlation threshold ρ decreases from $\rho = 1$ to $\rho \approx 0.5$ the number of links in the route decreases as links become more clustered. From $\rho \approx 0.5$ to 0.3 the cluster size increases and the number of clusterlinks in the route remains stable. From $\rho \approx 0.3$ to 0 the relative cluster size further increases and the number of clusterlinks in the route decreases. Note that not all links are clustered as ρ decreases and that for $\rho \rightarrow 0$ the route is not completely clustered.

Estimation accuracy

To see the effect of clustering on the estimation of statistical parameters, the variance of route travel times is calculated using both approaches and in function of ρ. Let σ_n^2 denote the variance calculated using the naïve approach and $\sigma_c^2(\rho)$ the variance calculated using the clustering approach for a certain correlation threshold ρ. The fraction $\sigma_c^2(\rho)/\sigma_n^2$ has been calculated for $\rho \in [0.02,1]$ for each route. The results are averaged over all 6 routes and are displayed in Figure 6. Because data gaps are excluded in the naïve approach, σ_n^2 overestimates the real variance. This is seen in the results since the fraction $\sigma_c^2(\rho)/\sigma_n^2$ is always smaller than 1. For $\rho = 1$, no links are clustered. Consequently all statistical dependencies in the travel time data of subsequent links are ignored. This results in an underestimation of the route travel time variance.

The variance σ_c^2 varies almost monotonically in function of ρ between these 2 extremes. Because of this, the clustering method always gives better estimates for $\rho < 1$ as when correlations are ignored (i.e. $\rho = 1$). However for $\rho \approx 0$, the fraction $\sigma_c^2(\rho)/\sigma_n^2$ does not converge to 1 because not all links are clustered in 1 clusterlink. The clustering approach copes more efficiently with data gaps than the naïve approach and does not overestimate the variance in the limit $\rho \rightarrow 0$ as much as the naïve approach.

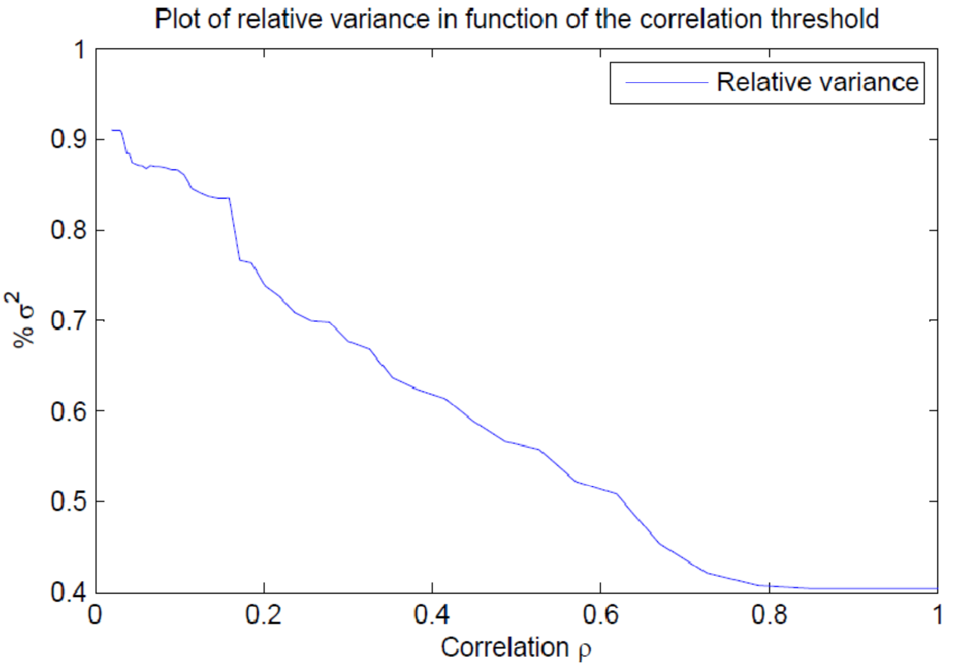

Figure 6: Plot of relative variance.

In Figure 7, a comparison of the histograms calculated under different settings is made. The histograms are calculated using the naïve approach and the clustering approach with $\rho = 0.02$ and $\rho = 1$. Since this comparison is similar for all 6 routes, 1 route was selected.
The histogram bins are wider for larger ρ. This is because as ρ increases, more correlations are ignored, and thus a wider range of travel times is possible. All histograms contain 100 bins, but the range of travel times has been cropped to $[4.10^3, 10^4]$ seconds since this region contains all the large bins.
Because in the naïve approach it is implicitly assumed that all links are correlated, the resulting histogram is noisier. Ignoring the smallest correlations (i.e. $\rho = 0.2$), the obtained histogram is much smoother, but still has approximately the same shape as the noisier naïve histogram. This is not the case if all correlations are ignored. The histogram for $\rho = 1$ gives more weight to small travel time values and less weight to large travel time values. Moreover, the centre of the distribution is shifted to larger travel time values. Because of this, the expected travel time and route reliability will always be overestimated. These errors are reduced by minimizing the correlations between subsequent links, as this results in a better estimation of the histogram shape and thus a better estimation of percentiles.

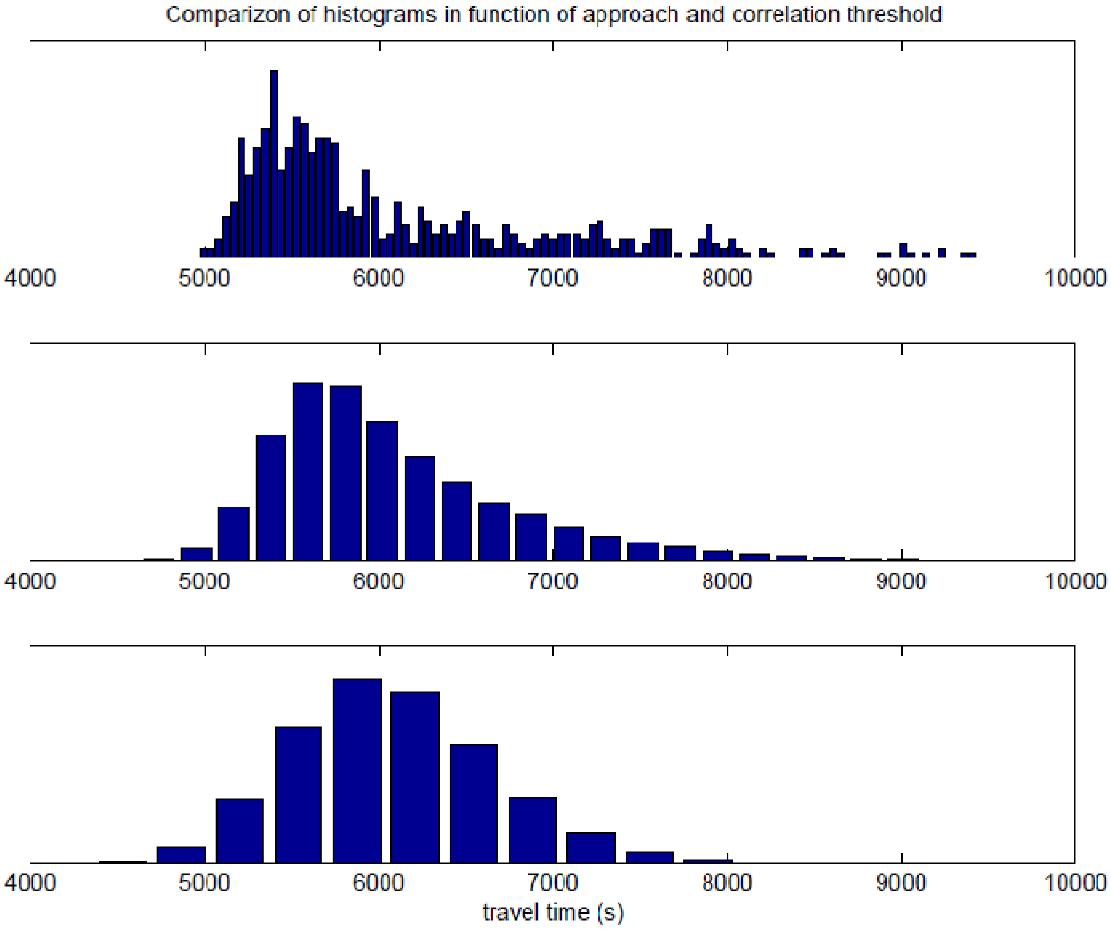

Figure 7: Plot of histograms: for the naïve approach (Top) and for the clustering approach with $\rho = 0.2$ (Middle) and with $\rho = 1$ (Bottom).

Route reliability

The results obtained from the clustering approach with $\rho = 0.2$ are used to calculate the 5%, 10%, 25%, 50%, 75%, 90% and 95% percentiles for each of the six routes. The routes are compared in Table 2, based on their percentile travel time values. This illustrates the applicability of the clustering methodology to robust route guidance, in which route reliability is important. A route with a low 95th travel time percentile is for example more reliable for risk averse drivers than a route with a higher 95th travel time percentile but maybe a lower expected travel time.

Route 4 has the best 50% travel time estimate. However, while route 4 would be chosen by risk neutral drivers, it is not at all a reliable route for risk averse drivers, since 4 of the 6 routes have a better 95% travel time estimate. This means that over all route 4 performs well, but that it is sensitive to delays. The most risk-averse route is route 2, which will be the best option in case of extreme events. In all other situations faster alternatives exist. Following this reasoning, route 1 and 5 are never a good option, and the best choice is a trade-off between route reliability and overall route travel time. In this view, route 3 and route 6 give the best results. For practical applications, reliability should be cast into a reliability measure depending on a user's preferences.

Table 2: Comparison of route performance based on the travel time estimates per percentile. The values in the table are route numbers. The first column indicates the percentile. Each row is ordered (left-to-right) from small to large percentile travel time value. For risk averse drivers, route 2 is the most reliable route. Contrary, route 4 is the most reliable route for risk neutral drivers since it has the lowest median travel time. In a trade-off situation, route 3 and 6 are both reliable and fast. Routes 5 and 1 are never a good option. For practical applications, reliability should be cast into a reliability measure depending on a user's preferences.

Prctl.	Route number					
95%	**2**	3	6	5	**4**	1
90%	3	**2**	6	**4**	5	1
75%	3	6	**4**	**2**	5	1
50%	**4**	6	3	**2**	5	1
25%	**4**	6	3	**2**	5	1
10%	**4**	6	3	**2**	5	1
5%	**4**	6	3	**2**	5	1

CONCLUSIONS

The clustering approach presented in this paper is able to efficiently calculate route travel time statistics on-line. This is done by shifting part of the calculations off-line. Not only it is at least an order of 10^2 more efficient as a naïve approach, it also improves the estimation accuracy by taking statistical dependencies into account. It is shown that the clustering with low correlation thresholds does not result in a completely clustered route. Rather, the route is divided in several almost uncorrelated clusterlinks. Hence, it deals more efficiently with incomplete data, which resulted in the naïve approach in an overestimation of the route travel time variance. The calculation of the route travel time histogram using the clustering approach is a coarse-grained approximation of the histogram calculated using the naïve approach. The shape is well reproduced, incorporating a long tail of higher travel time events with non-zero probability. It was shown that neglecting correlations results in a histogram in which long travel time events are underestimated. Testing the clustering approach on the Belgian road network shows that the method is sensitive enough to discriminate between routes, even if these are only slightly different. Moreover it shows that routes can be very different in terms of speed and reliability.

This research concentrated on the development of a novel method to calculate stationary route travel time statistics based on historical travel time data. The method assumes instantaneous travel times. The next step in the research is the development of different clustering heuristics and the study of their effect on the total network size and the size of the statistics database. Future research will also concentrate on the calculation of dynamic route travel time statistics making use of time-integrated travel time data. This will make it possible to calculate a histogram of route travel times for a certain time of the day, with the possibility to incorporate real-time travel time observations.

ACKNOWLEDGEMENTS

This research is part of the IBBT project MobiRoute in collaboration with Be-Mobile, Duo and Ghent University. We would like to thank IBBT for funding and our partners for the cooperation on this research. Explicit acknowledgements go to Steven Logghe and Be-Mobile (http://www.be-mobile.be) for providing the travel time data and description of their data collection method.

REFERENCES

Abdel-Aty, M. A., Kitamura, R., & Jovanis, P. P. (1996). Investigating Effect of Advanced Traveler Information on Commuter Tendency To Use Transit. *Transportation Research Record: Journal of the Transportation ResearchBoard , 1550*, 65 - 72.

BeMobile. (2010, February). *Be-Mobile: Floating Vehicle Data*. Opgeroepen op February 2010, van Be-Mobile: The leading provider of traffic & mobility information: http://www.be-mobile.be

Clark, S. D., & Watling, D. P. (2005). Modelling network travel time reliability under stochastic demand. *Transportation Research Part B: Methodological , 39* (2), 119 - 140.

Davidson, P., & Teye-Ali, C. (2008). Incorporating travel time reliability in traveller information systems.

Fu, L., & Rilett, L. R. (1998). Expected shortest paths in dynamic and stochastic traffic networks. *Transportation Research Part B: Methodological , 32* (7), 499-516.

He, R. R., Liu, H. X., Kornhauser, A. L., & Ran, B. (2002). *Temporal and Spatial Variability of Travel Time.* UC Irvine: Center for Traffic Simulation Studies.

Hellinga, B., & Fu, L. (1999). Route Selection Considering Travel Time Variability. *ITS World Conference.*

Key, R. (2005). Routing in stochastic networks. 861-866.

Krishnamoorthy, R. K. (2008). *Travel time estimation and forecasting on urban roads.* London: University of London.

Lecluyse, C., Van Woensel, T., & Peremans, H. (2009). Vehicle routing with stochastic time-dependent travel times. *4OR: A Quarterly Journal of Operations Research , 7* (1), 363 - 377.

Lu, X. (2001). *Dynamic and Stochastic Routing Optimization: Algorithm Developmentand.* Irvine: University of California Irvine.

Noland, R. B., & Polak, J. W. (2002). Travel time variability: a review of theoretical and empirical issues. *Transport Reviews , 22* (1), 39 - 54.

Small, K. A., Noland, R., Chu, X., & Lewis, D. (1999). *Valuation of Travel-Time Savings and Predictability in CongestedConditions.* NCHRP Report 431, Transportation Research Board, National Research.