Handling Uncertainties and Unreliability in Data from Smart Sensors

Pitu B. Mirchandani

School of Computing, Informatics and Decision System Engineering

Arizona State University

Monica Gentili

Department of Mathematics and Computer Science

University of Salerno (Italy)

Introduction

Current sensor technologies allow one to identify particular vehicles in a subnetwork and re-identify them later in the subnetwork. These technologies include license plate readers, electronic toll tags, blue tooth signatures, etc. When time stamps are included in the data, one obtains travel times for each vehicle on the route segments traversed by the vehicles. These travel times can be used to study the performance of the network, including predictability of travel times for traveller information systems. However there are complexities in the use of this data for travel time estimations due to (1) correlations, (2) unreliability of sensors and (3) uncertainty of the underlying scenario. This paper attempts to explain these complexities and develops a Bayesian model for using such data, where the travel times are updated from prior estimates to new estimates in real time.

Given the framework of the data processing approach, the Bayesian model may be combined with an optimal location model to place additional data collection points to improve the predictability of travel times. Such an optimization model may be modified to include the following stochastic considerations to make the data collection more “robust”:

1. Reliability of the sensors themselves, and
2. Possible scenarios of major network disruptions (e.g., link failures).

Bayesian Travel time estimation

When a vehicle, which may be referred to as a *probe* vehicle, passes a sensor and then another sensor downstream, a travel time *t* is measured. Given reliable error free sensors, *t* depends on (a) the current mean flow on the link (or route segment), (b) the fluctuations of the flow, and (c) the aggressiveness of the driver. Therefore, the underlying network may be modelled as stochastic. Each arc (or route segment) will have a mean travel time that changes slowly over time within the day, assuming that flows do not change fast unless a major accident takes place on the arc (or the route segment). The actual travel time measured fluctuates due to two factors: (1) natural fluctuations of flow that occur in traffic flows (and would occur even if the population of driver was homogeneous with identical aggressiveness’), and (2) the aggressiveness of the particular driver of the probe that passed the two sensors. The first fluctuation can be treated through a distribution describing the mean of the travel time, and latter fluctuation as some sort of additive noise. In the model presented here, a priori knowledge of the means and variances of mean travel times on route segments are assumed while sensors (e.g., blue tooth signature readers) that are located at given points on the network provide sample data on these travel times.

These data can be used to update the a priori information on the mean route travel time to improve our knowledge about the current travel times on the route. Granted that each update is “slightly behind the curve” since the next vehicle travelling may not experience the same travel time, it is still better than off-line batch processing of the data collected over a large time period.

We model the network as a directed graph *G = (V, A)* with a given set of route segments *R = {R₁, R₂, …, Rₚ}* on which most traffic flows. Let us denote by τ_{R_i} the travel time on route *Rᵢ*. τ_{R_i} is a stochastic variable with mean μ_{R_i} and variance $\mathrm{var}(\tau_{R_i})$. We assume that $\mathrm{var}(\tau_{R_i})$ is constant and known from past data, while the mean μ_{R_i} is a stochastic variable with some known a priori distribution.

Let τ_i denote the travel time on arc *aᵢ*, and let τ_i be a random variable with mean μ_i and variance σ_i^2. We assume that the a priori variance σ_i^2 is known for all arcs *aᵢ* in the network from historical data. Our model assumes that μ_i changes dynamically, but slowly; that is, for our Bayesian update it is assumed constant for short periods of time (say, for few minutes). Suppose for the current short time period, instead of its true value, we have a priori distribution of μ_i, which is Normal with mean η_i and variance γ_i. Therefore, the variance of travel time on arc *aᵢ* has two components, that is, $\mathrm{var}(\tau_i) = \sigma_i^2 + \gamma_i$. For the short period of time when travel time data is collected, the expected travel time on arc *aᵢ* is η_i; then together with $\mathrm{var}(\tau_i)$, a confidence interval of τ_i can be obtained.

Suppose that we have *n* observations of τ_i, then by appealing to the Central Limit Theorem, $\overline{\tau_i}$, which is the average value of the observed travel times, can be approximated by a Normal distribution with mean μ_i and variance σ_i^2 / n. With an observation of $\overline{\tau_i}$, we can update the a priori distribution of μ_i by appealing to Bayesian statistical theory. The updated mean and variance of μ_i are

$$\eta_i^* = \frac{\dfrac{\eta_i \sigma_i^2}{n} + \overline{\tau}_i \gamma_i}{\dfrac{\sigma_i^2}{n} + \gamma_i}, \qquad (1)$$

$$\gamma_i^* = \frac{\gamma_i \sigma_i^2 / n}{\sigma_i^2 / n + \gamma_i}. \qquad (2)$$

Now the expected value of τ_i is η_i^* and $\mathrm{var}(\tau_i) = \sigma_i^2 + \gamma_i^*$. Therefore, an updated confidence interval of τ_i may be computed.

We can see from equation (2) that the variance of μ_i is reduced. The difference between γ_i and γ_i^* is

$$\gamma_i - \gamma_i^* = \frac{\gamma_i^2}{\sigma_i^2 / n + \gamma_i} \qquad (3)$$

The above discussion of travel time prediction on a single arc can be extended to a route *R* in the traffic network consisting of a sequence of arcs. Assuming that the travel times on different arcs are independent, the route travel time $\tau_R = \sum_{i \in R} \tau_i$ is a random variable with the mean and variance of

$$\mu_R = \sum_{i \in R} \mu_i , \qquad (4)$$

$$\mathrm{var}(\tau_R) = \sum_{i \in R} \sigma_i^2 + \sum_{i \in R} \gamma_i . \qquad (5)$$

The expected value of τ_R is $\sum_{i \in R} \eta_i$. Therefore, a confidence interval of τ_R can be obtained with $\sum_{i \in R} \eta_i$ and $\mathrm{var}(\tau_R)$.

Assume that we have *n* observations of travel times on a segment *s* of the route. Let $\overline{\tau}_s$ be the average travel time on the segment. $\overline{\tau}_s$ has an approximate Normal distribution with $\mu_s = \sum_{i \in s} \mu_i$ and $\sigma_s^2 = \sum_{i \in s} \frac{\sigma_i^2}{n}$

The updated distribution of μ_s is then

$$\eta_s^* = \frac{\frac{\sum_{i \in s} \eta_i \sum_{i \in s} \sigma_i^2}{n} + \overline{\tau}_s \sum_{i \in s} \gamma_i}{\frac{\sum_{i \in s} \sigma_i^2}{n} + \sum_{i \in s} \gamma_i}, \qquad (6)$$

$$\gamma_s^* = \frac{(\sum_{i \in s} \gamma_i \sum_{i \in s} \sigma_i^2)/n}{\sum_{i \in s} \sigma_i^2 / n + \sum_{i \in s} \gamma_i}. \qquad (7)$$

With the updated value of η_s^* and γ_s^*, we have more information on the travel time of that route. Now, the predicted expected value of route travel time τ_R is $\eta_s^* + \sum_{i\in R\mathrm{I}\ i\notin s} \eta_i$, and the variance of τ_R is $\mathrm{var}(\tau_R) = \gamma_s^* + \sum_{i\in R\mathrm{I}\ i\notin s} \gamma_i + \sum_{i\in R} \sigma_i^2$. The updated confidence interval of τ_R is tighter than the original confidence interval because γ_s^* is smaller than $\gamma_s = \sum_{i\in s} \gamma_i$.

The drawbacks of the above model are

1. Assumption of independence in the randomness of travel times on arcs. Historical data on each arc will give means η_i and variances γ_i . Although one also obtain cross correlations among the travel times, the update model will be computationally burdensome and may introduce unnecessary delays in computing updates. On the other hand, since travel times are not assumed to vary much from update to update, it is felt that the approach will track travel times well.
2. Assumption that route segments assumed carry most of the flows in the subnetwork. This is not a strong assumption as long as the objective is to simply update travel time distributions on the route segments monitored. However, if the estimation criterion includes some sort of objective that includes the total flow monitored, then additional complementary assumptions would be necessary. For example, if the density of sensors is sufficient high, there would be very few alternative routes between sensors, and if equilibrium exists on these routes then the travel times on the alternative route segments (and links) could be easily imputed.

Optimal location models

From equations (3) and (7), larger sample size *n*, results in larger decrease in the variance of the predicted arc and route travel time, respectively. Different sets of new sensor locations provide different travel time samples in the network. By optimally locating the set can result maximizing predictive knowledge, in the sense of decreased confidence interval.

A first attempt in this direction was presented in [3] where two mathematical models to optimally locate sensors on the arcs of a network are presented. The first model is a deterministic model that locates sensors to maximize total vehicle-miles monitored. The second model takes into account the stochastic nature of travel times and seeks the optimal location of *q* sensors on the network so that the posterior variance of the mean travel times is minimized using an objective based on the model presented through (1)-(7).

Let *q* denote the maximum number of sensors that can be installed on the network. Three sets of binary variables are defined: x_j whose value is equal to 1 if a sensor is located on arc a_j and 0 otherwise; y_{ij} whose value is equal to 1 if arc a_j has the most downstream sensor located on route R_i and 0 otherwise; z_{ij} whose value is equal to 1 if arc a_j has the most upstream sensor located on route R_i and 0 otherwise. Parameters $\Delta\gamma_i^{jk}$ are used to denote the reduction in variance of the mean travel time on route R_i if travel times can be measured from arc a_j to arc a_k on that route. Data need to be preprocessed to obtain $\Delta\gamma_i^{jk}$ for each route and each route segment on that route. When *s* is the route segment between arc a_j and arc a_k on route R_i, then $\Delta\gamma_i^{jk}$ can be calculated by $\Delta\gamma_i^{jk} = \gamma_s - \gamma_s^*$, where $\gamma_s = \sum_{m\in s}\gamma_m$, and $\gamma_s^* = \dfrac{(\sum_{m\in s}\gamma_m \sum_{m\in s}\sigma_m^{\ 2})/n_s}{\sum_{m\in s}\sigma_m^{\ 2}/n_s + \sum_{m\in s}\gamma_m}$.

The mathematical formulation of the problem is then:

$$\max \tfrac{1}{2}\textstyle\sum_{1=1}^{s}\sum_{j\in R_i}\sum_{k\in R_i}(y_{ij} - z_{ik})\Delta\gamma_i^{jk} \tag{8}$$

$$\sum_{j=1}^{m} x_j \le q \tag{9}$$

$$y_{ij} \le x_j \quad \forall i = 1,2,\ldots,p \quad \forall j = 1,2,\ldots,m \tag{10}$$

$$z_{ij} \le x_j \quad \forall i = 1,2,\ldots,p \quad \forall j = 1,2,\ldots,m \tag{11}$$

$$\sum_{j\in R_i} y_{ij} \le 1 \quad \forall i = 1,2,\ldots,p \tag{12}$$

$$\sum_{j\in R_i} z_{ij} \le 1 \quad \forall i = 1,2,\ldots,p \tag{13}$$

$$\sum_{j\in R_i} y_{ij} - \sum_{j\in R_i} z_{ij} = 0 \quad \forall i = 1,2,\ldots,p\,. \tag{14}$$

Objective function (8) maximizes the reduction in variance by maximizing the difference between y_{ij} and z_{ij} weighted by the reduction of variance that is due to updating travel times on the route segments monitored. Constraint (9) requires locating no more that *q* sensors on the network. Constraints (12)-(13) ensure that there is no more than one upstream sensor and one downstream sensor on each route. Constraints (10) and (11) are logical constraints linking the variables and ensuring that there cannot be an upstream (or downstream) sensor on a route if no sensor is located on it. Finally, constraint (14) states that if there is a most upstream sensor on route R_i, there must also be a most downstream sensor on the route and vice versa. Similarly, if there is no most upstream sensor on route R_i, then there cannot be a most downstream sensor on the same route and vice versa. Note that if on a route segment there is only one sensor installed, then the installed sensor is both the most upstream and downstream sensor, and the vehicle miles monitored on that route is zero.

The above model is extended in this paper considering both the reliability of the sensors themselves, and possible scenarios of major network disruptions (e.g., link failures). Based of literature of unreliable facility locations and back-up assignment of customers of failed facilities [1,4], recently, Li and Ouyang [2] modified the first deterministic model proposed in [3] to take into account failed most upstream/downstream sensors and re-assignment of located sensors as the most upstream/downstream. Indeed, in the long period, sensor failure over time can cause loss of data and loss of reliability in the estimates of predicted travel times. The authors in [2] consider the probability of failure of each sensor to be uniform and independent and develop an integer mathematical formulation that locates a given number of sensors to maximize the expected total weighted vehicle-miles monitored.

In this paper, starting from the proposed Bayesian scheme, we study:

1. A modification of the stochastic model presented in [3] that takes into account sensors failure: We assume the probability of sensor failure to be site specific. The resulting model minimizes the expected posterior variance in the prediction of travel time when *q* sensors are to be located.
2. A modification of the model that takes into account link failure: each link in the network has a given known probability of disruption due to external events. When one or more link fails, the connectivity of the network changes. This induces re-routing in the network and a consequent change in the travel times on the network. The model seeks for the optimal location of *q* sensors to minimize the expected posterior variance over all possible link failure scenarios.

In the presentation we discuss the Bayesian approach and some experimental results based on simulations, which provide insight of the developed models.

REFERENCES

[1]. G. Chen, M.S. Daskin, Z.J. Shen and S. Uryasev (2006). The α-Reliable Mean-Excess Regret Model for Stochastic Facility Location Modeling. *Naval Research Logistics, 53*,617-626.

[2]. X. Li and Y. Ouyang, Reliable Sensor Deployment for Network Traffic Surveillance, working paper (submitted for publication in *Transportation Research Part B*.) (2009).

[3]. P. Mirchandani, M. Gentili and Y. He, Location of vehicle identification sensors to monitor travel-time performance, *IET Intelligent Transportation Systems 3*, 3, p.289–303 (2009)

[4]. L. Snyder and M Daskin, Reliability Models for facility location: The expected failure cost case. *Transportation Science 39, 3,* 400-419 (2005).