Travel Time Estimation for Ambulances using Bayesian Data Augmentation

Bradford S. Westgate[*], Dawn B. Woodard, David S. Matteson, Shane G. Henderson

Cornell University

January 31, 2012

Abstract

Estimates of ambulance travel times on road networks are critical for effective ambulance base placement and real-time ambulance dispatching. We introduce new methods for estimating the distribution of travel times on each road segment in a city, using Global Positioning System (GPS) data recorded during ambulance trips. Our preferred method uses a Bayesian model of the ambulance trips and GPS data. Due to sparseness and error in the GPS data, the exact ambulance paths and travel times on each road segment are unknown. To estimate the travel time distributions using the GPS data, we must also estimate each ambulance path. This is known as the map-matching problem. We simultaneously estimate the unknown paths, travel times, and the parameters of each road segment travel time distribution using Bayesian data augmentation. We also introduce two alternative estimation methods based on GPS speed data that are simple to implement in practice.

We test the predictive accuracy of the three methods on a subregion of Toronto, using simulated data and data from Toronto EMS. In both cases, out-of-sample point and interval estimates of ambulance trip durations from the Bayesian method substantially outperform estimates from the alternative methods. We also construct probability-of-coverage figures using the Bayesian estimates, which are essential for emergency medical service (EMS) providers. Finally, map-matching estimates from the Bayesian method interpolate well between sparsely recorded GPS readings and are robust to GPS location errors.

Keywords: Reversible jump, Markov chain, map matching, Global Positioning System, emergency medical services

[*] Corresponding author. Email: bsw62@cornell.edu

1 Introduction

Emergency medical service (EMS) providers prefer to assign the closest available ambulance to respond to a new emergency [Dean 2008]. Thus, it is vital to have accurate estimates of the travel time of each ambulance to the emergency location. An ambulance is often assigned to a new emergency while away from its ambulance base [Dean], so the problem is more complicated than estimating response times from several fixed bases. Travel times also play a central role in locating bases and parking locations [Brotcorne et al. 2003, Goldberg 2004, Henderson 2010]. Travel times are variable, and recent EMS research has shown the importance of accounting for this uncertainty [Erkut et al. 2008, Ingolfsson et al. 2008]. In this paper, we estimate the distribution of ambulance travel times on each road segment (the section of road between neighboring intersections) in a city. This enables estimation of fastest paths in expectation between any two locations, using a shortest path algorithm, and also estimation of the probability an ambulance will reach its destination within a certain time threshold.

Available data are historic Global Positioning System (GPS) readings, stored during ambulance trips. Most EMS providers record this information; we use GPS data from Toronto EMS from 2007 and 2008. The GPS data contain locations, timestamps, speeds, vehicle and emergency incident IDs, and other information. The GPS data are sparse; readings are stored every 200 meters (m) or 240 seconds (s), whichever comes first. The true GPS sampling rate is much higher, but this scheme minimizes data transmission and storage. This is standard practice across EMS providers, though the frequencies vary [Mason 2005]. In related non-EMS applications the GPS readings can be even sparser. For example, Lou et al. [2009] analyzed data from taxis in Tokyo in which GPS readings are separated by 1-2 km or more.

The GPS location and speed data are also subject to error. For example, accuracy degrades in “urban canyons,” where GPS satellites may be obscured and signals reflected [Chen et al. 2005, Mason 2005, Syed 2005]. Chen et al. observed average location errors of 27 m in parts of Hong Kong with narrow streets and tall buildings, with some errors over 100 m. Location error is also present in the Toronto data; see Figure 1. Witte and Wilson [2004] found GPS speed errors of roughly 5% on average, with largest error at high speeds and when few GPS satellites were visible.

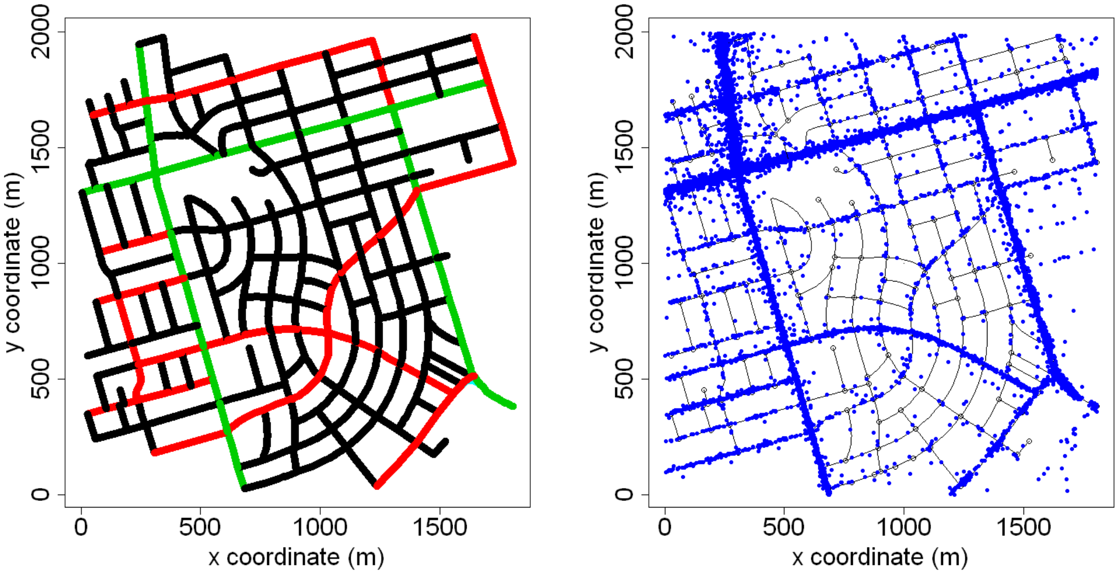

Figure 1: Left: A subregion of Toronto, with primary roads (green), secondary roads (red) and tertiary roads (black). Right: All GPS locations on this region from the Toronto EMS "standard travel" dataset.

We first introduce "local" methods using only the GPS locations and speeds (Section 4). Each GPS reading is mapped to the nearest road segment, and the mapped speeds are used to estimate the travel time for each segment. We call these methods "local" because they do estimation independently for each segment. We introduce two local methods. In the first, we use the harmonic mean of the mapped GPS speeds to create a point estimator of the travel time. We are the first to propose this estimator for mapped GPS data, though it is commonly used for estimating travel times using speed data recorded by loop detectors [Rakha and Zhang 2005, Soriguera and Robuste 2011, Wardrop 1952]. We give theoretical results supporting this approach in Appendix B. This method also yields natural interval and distribution estimates of the travel time. In our second local method, we assume a parametric distribution for the GPS speeds on each segment, and calculate maximum likelihood (MLE) estimates of the parameters of this distribution. These can be used to obtain point, interval, or distribution estimates of the travel time.

We also propose a more sophisticated method, modeling the GPS data at the trip level and combining them with additional trip information, including start and end locations (Sections

2 and 3). Using Bayesian data augmentation [Tanner and Wong 1987], we simultaneously estimate the path taken for each ambulance trip (solving the so-called map-matching problem [Mason]) and the distribution of travel times on each road segment. This method improves upon the local methods, because it incorporates GPS timestamps and other trip information. Also, GPS readings that are mapped to incorrect arcs by the local methods can be used on the correct arcs by the Bayesian method, if the map-matching is effective.

We compare the performance of the local methods and the Bayesian method on a subregion of Toronto (Figure 1), using simulated data and historical data from Toronto EMS (Sections 5 and 6). In both cases, the Bayesian method substantially outperforms the local methods in point and interval estimates of travel times for out-of-sample trips. In simulated data, point estimates from the Bayesian method outperform the local methods by over 60% (in root mean squared error), relative to an "Oracle" method with the lowest possible error. In the Toronto EMS data, interval estimates from the Bayesian method have dramatically better coverage.

Parameter estimation in the Bayesian approach is more computationally intensive than for the local methods. However, once the parameter estimation stage is complete, prediction for new routes and generation of the graphical displays in this paper can be done virtually instantaneously. Also, the parameter estimates are updated infrequently and offline, so the increased computation time in this stage is not an operational handicap.

We apply distribution estimates from the Bayesian method to produce probability-of-coverage figures [Budge et al. 2010] for the Toronto subregion (Section 6.4), showing the probability of traveling from a given start node to any other node within a time threshold. This is the performance standard in many EMS contracts; a typical EMS organization attempts to respond to, e.g., 90% of all emergencies within 9 minutes [Fitch 1995]. The intervals of the Bayesian method are wider than those from the local MLE method, leading to fewer nodes with extremely high or low probabilities.

We also assess the time-dependence of the travel times in the Toronto EMS dataset, by applying our methods to the rush-hour and non-rush-hour data separately. The Toronto EMS dataset is divided into "lights-and-sirens" (L-S) and "standard travel" (Std) ambulance trips.

Binning by rush-hour greatly improves predictive accuracy of the Bayesian method on the Std data, though it has little effect on performance for the L-S data. Ambulances at L-S speeds are less affected by traffic than at Std speeds, so the distribution of L-S travel times changes less throughout the day.

Finally, we assess the Bayesian method as a solution to the map-matching problem. Unlike most map-matching techniques that analyze each trip in isolation [Krumm et al., Lou et al., Marchal et al., Mason], the entire dataset of trips is used to produce the path estimate for each trip. Also, the posterior distribution over paths can capture multiple high-probability paths when the true path is unclear from the GPS data. We find that path estimates are interpolated accurately between widely-separated GPS points and are robust to GPS error.

Recent work on estimating ambulance travel time distributions has been done by Budge et al. [2010] and Aladdini [2010], using estimates based on total travel distance and duration, not GPS data. Neither of these papers considered travel times on individual road segments. For this reason they cannot capture some desired features, such as faster response times to locations near major roads (see Section 6.4).

2 Bayesian Formulation

2.1 Model

Consider a network of J directed road segments, called "arcs," and a set of I ambulance trips on this network. Assume that each trip i begins and ends on known nodes (intersections) d_i^1 and d_i^2 in the network, at known times. In practice, trips sometimes begin or end in the interior of a road segment; however, road segments are short enough that this is a minor issue (we show how to estimate start and end nodes in Section 6). The data associated with each trip i consist of the observed GPS readings, indexed by $\ell = 1, \ldots, r_i$ (where r_i is the number of GPS readings in trip i), and gathered at fixed times t_i^ℓ. GPS reading ℓ is the triplet $\left(X_i^\ell, Y_i^\ell, V_i^\ell\right)$, where X_i^ℓ and Y_i^ℓ are the measured geographic coordinates and V_i^ℓ is the measured speed. Denote $G_i = \left\{\left(X_i^\ell, Y_i^\ell, V_i^\ell\right)\right\}_{\ell=1}^{r_i}$.

The relevant unobserved variables for each trip i are the following:

1. The unknown path (sequence of arcs) $A_i = \{A_{i,1}, \ldots, A_{i,N_i}\}$ traveled by the ambulance from d_i^1 to d_i^2. The path length N_i is also unknown.
2. The unknown travel times $T_i = (T_{i,1}, \ldots, T_{i,N_i})$ on the arcs in the path. We also use the notation $T_i(j)$ to refer to the travel time in trip i on arc j.

We model the observed and unobserved variables $\{A_i, T_i, G_i\}_{i=1}^I$ as follows. Conditional on A_i, each element $T_{i,k}$ of the vector T_i follows a lognormal distribution with parameters $\mu_{A_{i,k}}, \sigma^2_{A_{i,k}}$, independently across i and k. Specifically,

$$T_{i,k}|A_i \sim \text{log-}N\left(\mu_{A_{i,k}}, \sigma^2_{A_{i,k}}\right) = \frac{1}{T_{i,k}\sqrt{2\pi\sigma^2_{A_{i,k}}}} \exp\left(-\frac{\left(\log T_{i,k} - \mu_{A_{i,k}}\right)^2}{2\sigma^2_{A_{i,k}}}\right). \tag{1}$$

In the literature, ambulance travel times have been observed and modeled to be lognormal [Aladdini 2010, Alanis et al. 2010]. Denote the expected travel time on each arc $j \in \{1, \ldots, J\}$ by $\theta(j) = \exp\left(\mu_j + \sigma_j^2/2\right)$. We use a multinomial logit choice model [McFadden 1973] for the path A_i, with likelihood

$$f(A_i) = \frac{\exp\left(-C\sum_{k=1}^{N_i} \theta\left(A_{i,k}\right)\right)}{\sum_{a_i \in \mathcal{P}_i} \exp\left(-C\sum_{k=1}^{n_i} \theta\left(a_{i,k}\right)\right)}, \tag{2}$$

where $C > 0$ is a fixed constant and $\mathcal{P}_i$ is the set of possible paths from d_i^1 to d_i^2 (with no repeated nodes) in the network. This model captures the fact that the fastest routes in expectation have the highest probability.

We assume that ambulances travel at constant speed on a single arc in a given trip. This is a necessary approximation since the data sparseness means that there is typically one or zero GPS readings on any arc in a given trip, and thus very little information in the data regarding changes in speed on individual arcs. Thus, the true location and speed of the ambulance at time t_i^ℓ are deterministic functions $\text{loc}\left(A_i, T_i, t_i^\ell\right)$ (short for "location") and $\text{speed}\left(A_i, T_i, t_i^\ell\right)$ of A_i and T_i. Conditional on A_i, T_i, the measured location $\left(X_i^\ell, Y_i^\ell\right)$ is assumed to have a bivariate normal distribution (a standard assumption; see [Krumm et al. 2007, Mason 2005]) centered at loc $\left(A_i, T_i, t_i^\ell\right)$, with known covariance matrix Σ. Similarly, the measured speed V_i^ℓ is assumed to have a lognormal distribution with expected value equal to speed $\left(A_i, T_i, t_i^\ell\right)$ and variance parameter ζ^2:

$$\left(X_i^\ell, Y_i^\ell\right)\Big| A_i, T_i \sim N_2\left(\mathrm{loc}\left(A_i, T_i, t_i^\ell\right), \Sigma\right), \tag{3}$$

$$\log V_i^\ell \Big| A_i, T_i \sim N\left(\log \mathrm{speed}\left(A_i, T_i, t_i^\ell\right) - \frac{\zeta^2}{2}, \zeta^2\right). \tag{4}$$

We assume independence between all the GPS speed and location errors. Combining Equations 1, 2, 3, and 4, we obtain the complete-data likelihood

$$f\left(\{A_i, T_i, G_i\}_{i=1}^I \Big| \{\mu_j, \sigma_j^2\}_{j=1}^J, \zeta^2\right) = \prod_{i=1}^I \left[f(A_i) \prod_{k=1}^{N_i} \log\text{-}N\left(T_{i,k}; \mu_{A_{i,k}}, \sigma^2_{A_{i,k}}\right)\right.$$
$$\left.\prod_{\ell=1}^{r_i}\left[N_2\left(\left(X_i^\ell, Y_i^\ell\right); \mathrm{loc}\left(A_i, T_i\right), \Sigma\right) \times \log\text{-}N\left(V_i^\ell; \log \mathrm{speed}\left(A_i, T_i, t_i^\ell\right) - \frac{\zeta^2}{2}, \zeta^2\right)\right]\right]. \tag{5}$$

In practice we use data-based choices for the constants Σ and C; see Appendix A. The unknown parameters in the model are the arc travel time parameters $\left\{\mu_j, \sigma_j^2\right\}_{j=1}^J$ and the GPS speed error parameter ζ^2. In Section 3, we give a computational method to estimate these parameters and the unobserved variables $\{A_i, T_i\}_{i=1}^I$ simultaneously.

2.2 Prior Distributions

To complete the Bayesian model, we must specify prior distributions for the unknown parameters ζ^2 and $\left\{\mu_j, \sigma_j^2\right\}_{j=1}^J$. We use

$$\mu_j \sim N\left(m_j, s^2\right), \qquad \sigma_j \sim \mathrm{Unif}\left(b_1, b_2\right), \qquad \zeta \sim \mathrm{Unif}\left(b_3, b_4\right), \tag{6}$$

independently, where m_j, s^2, b_1, b_2, b_3, b_4 are fixed hyperparameters. A normal prior is a standard choice for the location parameter of a lognormal distribution. We use uniform priors on the standard deviations σ_j and ζ so that the information in the data will dominate inference regarding these parameters; the prior ranges $[b_1, b_2]$ and $[b_3, b_4]$ are set to be wide enough to capture all remotely plausible parameter values. The prior mean for μ_j depends on j, while the other hyperparameters do not, because there are often existing road speed estimates that can be used in the specification of m_j. By contrast, prior information regarding the values s^2, b_1, b_2, b_3, b_4 is more limited. We use a combination of prior information and information in the data to specify all hyperparameters, as described in Appendix A.

3 Bayesian Computational Method

We use a Monte Carlo method to obtain samples $\left(\zeta^{2(\ell)}, \left\{\mu_j^{(\ell)}, \sigma_j^{(\ell)}\right\}_{j=1}^J, \left\{A_i^{(\ell)}, T_i^{(\ell)}\right\}_{i=1}^I\right)$ from the joint posterior distribution of all unknowns [Tanner and Wong 1987]. This is done by constructing a Markov chain with invariant distribution equal to the posterior distribution. There are many ways to do this and the precise construction does not affect the estimates or predictions resulting from the Bayesian method, provided that the chain is run for long enough, but it can affect the convergence time and thus the efficiency of the computational method. Here we give a simple but effective construction based on Metropolis-within-Gibbs [Robert and Casella 2004, Tierney 1994]. Each unknown quantity is updated in turn, conditional on the other unknowns. Each update is either a draw from the closed-form conditional posterior distribution or a Metropolis-Hastings move. Estimation of any desired function $g\left(\zeta^2, \left\{\mu_j, \sigma_j^2\right\}_{j=1}^J\right)$ of the unknown parameters is done via Monte Carlo, taking $\hat{g} = \frac{1}{M}\sum_{\ell=1}^M g\left(\zeta^{2(\ell)}, \left\{\mu_j^{(\ell)}, \sigma_j^{2(\ell)}\right\}_{j=1}^J\right)$.

3.1 Markov Chain Initial Conditions

First we describe the initial conditions used for the Markov chain. To initialize each path A_i, select the "middle" GPS reading, reading number $\lfloor r_i/2 \rfloor + 1$. Find the nearest node in the road network to this GPS location, and route the initial path A_i through this node, taking the shortest-distance path to and from the middle node. To initialize the travel time vector T_i, distribute the known trip duration across the arcs in the path A_i, weighted by arc length. Finally, to initialize ζ^2 and each μ_j and σ_j^2, draw from their priors.

3.2 Updating the Paths

Next we describe the updating of each path A_i in the Markov chain. We update A_i using a reversible-jump M-H proposal, since the number of arcs in the path may vary, changing the number of parameters in the model [Green 1995, Richardson and Green 1997]. We propose a small change to the current path A_i, giving proposed sample A_i^*. Because the path changes, the travel times T_i must be updated, giving proposed sample T_i^*. The new values (A_i^*, T_i^*) are accepted with the appropriate M-H acceptance probability, detailed below.

The proposal changes a contiguous subset of the path. The length (number of arcs) of this subpath is limited to some maximum value K; K is specified in Section 3.5. The proposal works as follows.

1. With equal probability, choose a node d' from the path A_i, excluding the final node.

2. Let a be the number of nodes that follow d' in the path. With equal probability, choose an integer $u \in \{1, \ldots, \min(a, K)\}$. Denote the uth node following d' as d''. The subpath from d' to d'' is the section to be updated (the "current update section").

3. Collect the alternative routes of length up to K from d' to d''. With equal probability, propose one of these routes as a change to the path (the "proposed update section"), obtaining the proposed path A_i^*.

Next, we propose new travel times T_i^* that are compatible with the new path A_i^*. Let $\{c_1, \ldots, c_m\} \subset A_i$ and $\{p_1, \ldots, p_n\} \subset A_i^*$ denote the arcs in the current and proposed update sections, noting that m and n will be different if the number of arcs has changed. As defined above, let $T_i(j)$ denote the travel time in path i on arc j. For each arc $j \in A_i^* \setminus \{p_1, \ldots, p_n\}$, set $T_i^*(j) = T_i(j)$. Let $S_i = \sum_{\ell=1}^{m} T_i(c_\ell)$ be the total travel time of the current update section. We must have $\sum_{\ell=1}^{n} T_i^*(p_\ell) = S_i$ also, because the total duration of the trip is known. The travel times $T_i^*(p_1), \ldots, T_i^*(p_n)$ are proposed as follows.

- Draw $(r_1, \ldots, r_n) \sim \text{Dirichlet}\,(\alpha\theta(p_1), \ldots, \alpha\theta(p_n))$, for a constant α (specified below). Set the proposed travel times $T_i^*(p_\ell) = r_\ell S_i$, for $\ell = 1, \ldots, n$.

This gives a proposal that is reasonable (and thus likely to be accepted), because the expected value of the new travel time on arc p_ℓ is (see Gelman et al. [2004])

$$E\left(T_i^*(p_\ell)\right) = S_i \frac{\theta(p_\ell)}{\sum_{k=1}^{n} \theta(p_k)},$$

so the total travel time on the current arcs is randomly distributed over the proposed arcs, weighted by the arc expected travel times. The constant α influences the variance of each component, but not the expected values. In our experience $\alpha = 1$ works well for our application; one can also tune α to obtain a desired acceptance rate for a particular dataset [Robert and Casella 2004, Roberts and Rosenthal 2001].

The M-H acceptance probability for this reversible-jump proposal [Green 1995, Richardson and Green 1997] is

$$p_{\mathrm{A}} = \min\left\{1, \frac{f\left(A_i^*, T_i^*, G_i \,\middle|\, \left\{\mu_j, \sigma_j^2\right\}_{j=1}^{J}, \zeta^2\right)}{f\left(A_i, T_i, G_i \,\middle|\, \left\{\mu_j, \sigma_j^2\right\}_{j=1}^{J}, \zeta^2\right)} \frac{q\left(A_i, T_i \,\middle|\, A_i^*, T_i^*, \left\{\mu_j, \sigma_j^2\right\}_{j=1}^{J}\right)}{q\left(A_i^*, T_i^* \,\middle|\, A_i, T_i, \left\{\mu_j, \sigma_j^2\right\}_{j=1}^{J}\right)} |\mathrm{Ja}|\right\},$$

where f denotes the complete data likelihood (Equation 5), q denotes the proposal density, and $|\mathrm{Ja}|$ denotes the Jacobian of the transformation between the parameter spaces corresponding to the current and proposed paths [Green]. We calculate q and $|\mathrm{Ja}|$ in Appendix C.

3.3 Updating the Trip Travel Times

The travel times $\{T_i\}_{i=1}^{I}$ are changed (by necessity) in the path proposal above, but we also include another M-H update of only the travel times [Tierney], to improve mixing of the Markov chain. The proposal works as follows.

1. With equal probability, choose arcs j_1 and j_2 in the path A_i. Let $S_i = T_i(j_1) + T_i(j_2)$.
2. Draw $(r_1, r_2) \sim \text{Dirichlet}(\alpha'\theta(j_1), \alpha'\theta(j_2))$. Set $T_i^*(j_1) = r_1 S_i$ and $T_i^*(j_2) = r_2 S_i$.

Similarly to the path proposal above, this proposal randomly distributes the travel time over the two arcs, weighted by the expected travel times $\theta(j_1)$ and $\theta(j_2)$, with variances controlled by the constant α' [Gelman et al. 2004]. In our experience $\alpha' = 0.5$ is effective for our application; the value of α' can also be tuned to improve the acceptance rate for a particular dataset [Robert and Casella, Roberts and Rosenthal]. The M-H acceptance probability may be calculated in a similar manner as in Appendix C.

3.4 Updating the Parameters μ_j, σ_j^2, and ζ^2

To update each μ_j, we sample from the full conditional posterior distribution, which is available in closed form. We have $\mu_j \left| \sigma_j^2, \{A_i\, T_i\}_{i=1}^I \sim N\left(\hat{\mu}_j, \hat{s}_j^2\right)\right.$, where

$$\hat{s}_j^2 = \left[\frac{1}{s^2} + \frac{n_j}{\sigma_j^2}\right]^{-1}, \qquad \hat{\mu}_j = \hat{s}_j^2 \left[\frac{m_j}{s^2} + \frac{1}{\sigma_j^2}\sum_{i\in I_j} \log T_i(j)\right],$$

the index set $I_j \subset \{1, \ldots, I\}$ indicates the subset of trips using arc j, and $n_j = |I_j|$.

To update each σ_j^2, we use a local M-H step [Tierney]. We propose $\sigma_j^{2*} \sim \text{Log-}N(\log \sigma_j^2, \eta^2)$, having fixed variance η^2. Using Equations 1 and 6, the M-H acceptance probability is

$$p_\sigma = \min\left\{1, \frac{\sigma_j}{\sigma_j^*}\mathbf{1}_{\{\sigma_j^* \in [b_1, b_2]\}} \left(\frac{\prod_{i\in I_j} \text{Log-}N\left(T_i(j); \mu_j, \sigma_j^{2*}\right)}{\prod_{i\in I_j} \text{Log-}N\left(T_i(j); \mu_j, \sigma_j^2\right)}\right) \frac{\text{Log-}N\left(\sigma_j^2; \log\left(\sigma_j^{2*}\right), \eta^2\right)}{\text{Log-}N\left(\sigma_j^{2*}; \log\left(\sigma_j^2\right), \eta^2\right)}\right\}.$$

To update ζ^2, we use another M-H step with a lognormal proposal, with different variance ν^2. The M-H acceptance probability may be calculated similarly. The proposal variances η^2, ν^2 are tuned to achieve an acceptance rate of approximately 25% [Roberts and Rosenthal].

3.5 Markov Chain Convergence

The Markov chain converges to the posterior distribution as long as the M-H transition kernels are reversible, aperiodic, and irreducible [Tierney]. The proposals for ζ^2, σ_j^2, and T_i satisfy these requirements. The A_i transition kernel is aperiodic, and reversible because the current path has update section length $m \leq K$ and the Dirichlet distribution has positive density everywhere on the simplex Δ^m, so it is possible to transition from state $\{A_i^*, T_i^*\}$ to state $\{A_i, T_i\}$. Rarely, a path is initialized with a repeat node (see Section 3.1), in which case the reverse transition is not allowed. However, this initial state is transient, so it can be neglected.

The A_i kernel is irreducible if for any path i, it is possible to move between any two paths in $\mathcal{P}_i$ in a finite number of iterations. For a given road network, the maximum update section length K can be set high enough to meet this criterion. On road networks with high connectivity, a low K is sufficient. For example, $K = 3$ is sufficient for a square grid, because it allows a path using one edge of a grid box to be replaced with a path using the other three edges. The value of K should be set as low as possible, because increasing K tends to lower the acceptance rate.

If there is a region of the city with sparse connectivity, the required value of K may be impractically large. For example, this could occur with a highway parallel to a small road. If the small road intersects other small roads, with each intersection beginning a new arc, there could be many arcs of the small road alongside a single arc of the highway. Then, a large K would be needed to allow transitions between the highway and the small road. If K is kept smaller, the Markov chain is not irreducible. In this case, the chain converges to the conditional posterior distribution for the closed communicating class in which the chain is absorbed. If this class contains much of the posterior mass, as might arise if the initial path follows the GPS data reasonably closely, then this should be a good approximation.

In Sections 5 and 6, we apply the Bayesian method to simulated data and data from Toronto EMS, on a subregion of Toronto with 623 arcs. Each Markov chain was run for 50,000 iterations (where each iteration updates all parameters), after a burn-in period of 25,000 iterations. We calculated Gelman-Rubin diagnostics [Gelman and Rubin 1992], using two chains, for the parameters ζ^2, μ_j, and σ_j^2. Results from a typical simulation study were: potential scale reduction factor (using the second half of both chains) of 1.06 for ζ^2, of less than 1.1 for μ_j for 549 arcs (88.1%), between 1.1-1.2 for 43 arcs (6.9%), between 1.2-1.5 for 30 arcs (4.8%), and less than 2 for the remaining 1 arc, with similar results for the parameters σ_j^2. These results indicate no lack of convergence.

Each Markov chain iteration for these experiments takes roughly 0.1 s on a personal workstation. The calculation of the local method estimates (Section 4) for all arcs also takes roughly

0.1 s. Therefore, parameter estimation for the Bayesian method is 4-5 orders of magnitude slower than for local methods. This does not include mapping each GPS reading to the nearest arc for the local methods (Section 4), which increases linearly with the number of nodes in the network, if implemented naively. If calculation of the Bayesian estimates for all data on an entire city is computationally impractical, the city can be divided into several regions and estimated in parallel. In future work, we will introduce approximate computational methods for this same model, allowing estimation for large spatial regions in a reasonable computation time.

4 Local Methods

Here we describe the two “local” methods outlined in Section 1. Each GPS reading is mapped to the nearest arc (both directions of travel are treated together). Let n_j be the number of GPS points mapped to arc j, L_j the length of arc j, and $\left\{V_j^k\right\}_{k=1}^{n_j}$ the mapped speed observations. We assume constant speed on each arc, as in the Bayesian method (Section 2.1). Thus, let $T_j^k = L_j/V_j^k$ be the travel time associated with observed speed V_j^k.

In the first local method, we calculate the harmonic mean of the speeds $\left\{V_j^k\right\}_{k=1}^{n_j}$, and convert to a travel time point estimate

$$\hat{T}_j^H = \frac{L_j}{n_j}\sum_{k=1}^{n_j}\frac{1}{V_j^k}.$$

This is equivalent to calculating the arithmetic mean of the associated travel times T_j^k. The empirical distribution of the associated times $\left\{T_j^k\right\}_{k=1}^{n_j}$ can be used as a distribution estimate. Because readings with speed 0 occur in the Toronto EMS dataset, we set any reading with speed below 5 miles per hour (mph) equal to 5 mph. Results are fairly sensitive to this correction. If the speed threshold is lower, some associated times are significantly higher, and the mean travel time estimates are higher. This harmonic mean estimator is well-known in the transportation research literature (where it is called the “space mean speed”) in the context of estimating travel times on a particular road using speed data recorded by loop detectors

[Rakha and Zhang 2005, Soriguera and Robuste 2011, Wardrop 1952].

In Appendix B, we consider this travel time estimator $\hat{T}_j^H$ and its relation to the GPS sampling scheme. We show that if GPS points are sampled by distance (for example, every 100 m), $\hat{T}_j^H$ is an unbiased and consistent estimator for the true mean travel time. However, if GPS points are sampled by time (for example, every 30 s), $\hat{T}_j^H$ overestimates the mean travel time. The Toronto EMS dataset uses a combination of sampling-by-distance and sampling-by-time. However, the distance constraint is usually satisfied first (see Figure 5, where the sampled GPS points are regularly spaced). Thus, the travel time estimator $\hat{T}_j^H$ is appropriate.

In the second local method, we assume $V_j^k \sim \text{Log-}N(m_j, s_j^2)$, independently across k, for unknown travel time parameters m_j and s_j^2. This distribution on the travel speed implies a related lognormal distribution on the travel time. Specifically, $T_j^k \sim \text{Log-}N\left(\log(L_j) - m_j, s_j^2\right)$. We use the maximum likelihood estimators

$$\hat{m}_j = \frac{1}{n_j}\sum_{k=1}^{n_j} \log\left(V_j^k\right), \qquad \hat{s}_j^2 = \frac{1}{n_j}\sum_{k=1}^{n_j}\left(\log\left(V_j^k\right) - \hat{m}_j\right)^2$$

to estimate m_j and s_j^2. Our point travel time estimator is

$$\hat{T}_j^{\text{MLE}} = E\left(T_j \middle| \hat{m}_j, \hat{s}_j^2\right) = \exp\left(\log(L_j) - \hat{m}_j + \frac{\hat{s}_j^2}{2}\right).$$

This second local method also provides a natural distribution estimate for the travel times via the estimated lognormal distribution for T_j^k. Correcting for zero-speed readings is again done by thresholding, to avoid log(0), but the results are less dependent on the threshold, because the speeds are not inverted.

The MLE method is biased towards overestimating the travel times, because of extra variation in the observed speeds, caused by GPS speed errors [Witte and Wilson 2004] and departures from the assumption of constant speed on each road segment. Consider modeling this extra variation with a lognormal error. Let the observed speed $O_j^k = V_j^k \mathcal{E}$, where V_j^k is now the true speed at the GPS time, and $\mathcal{E} \sim \text{Log-}N\left(-\psi^2/2, \psi^2\right)$ (so that $E(\mathcal{E}) = 1$). Then $O_j^k \sim \text{Log-}N\left(m_j - \psi^2/2, s_j^2 + \psi^2\right)$. The MLE method actually estimates $E\left(L_j/O_j^k\right)$ instead of $E\left(L_j/V_j^k\right) = E\left(T_j^k\right)$, causing overestimation, because $E\left(L_j/O_j^k\right) = E\left(T_j^k\right)\exp\left(\psi^2\right)$. There is no way to adjust for this bias if the parameters are estimated independently for each arc, as in our local methods. The speed variability s_j^2 and the error ψ^2 are not separately identified; only their sum is identified.

Some small residential arcs have no assigned GPS points in the Toronto EMS dataset (see Figure 1). In this case, we use a breadth-first search [Nilsson 1998] to find the closest arc in the same "road class" that has assigned GPS points. The road classes are described in Section 5; by restricting our search to arcs of the same class we ensure that the speeds are comparable.

5 Simulation Experiments

Next we test the Bayesian and local methods described in Sections 2-4 on simulated data. We simulate ambulance trips on the road network of Leaside, Toronto, shown in Figure 1. The region's area is almost 4 square kilometers. In this region, a value $K = 6$ (see Section 3.5) guarantees that the Markov chain is irreducible and thus valid. This region has four road classes. We define the highest-speed class to be "primary" arcs, the two intermediate classes to be "secondary" arcs, and the lowest-speed class to be "tertiary" arcs (see Figure 1). We compare the accuracy of our three methods for predicting durations of test trips with known paths. We also evaluate the map-matching estimates from the Bayesian method for example simulated paths.

5.1 Generating Simulated Data

We simulate data as follows. We generate ambulance trips with true paths, travel times, and GPS readings. For each trip i, we uniformly choose start and end nodes. We then construct the true path A_i arc-by-arc. At each node, beginning at the start node, we uniformly choose an adjacent arc out of those that lower the expected time to the end node. We repeat this process until the end node is reached. This method differs from the Bayesian prior (see Section 2.2), and can lead to a wide variety of paths traveled between two nodes.

The arc travel times are lognormal: $T_{i,k} \sim \text{Log-}N(\mu_{A_{i,k}}, \sigma^2_{A_{i,k}})$. To set the true travel time parameters $\left\{\mu_j, \sigma_j^2\right\}$ for arc j, we uniformly generate a speed between 20-40 mph. We set μ_j and σ_j^2 so that the arc length divided by the mean travel time equals this speed. To give the arcs a range of travel time variances, we draw $\sigma_j \sim \text{Unif}\left(0.5\log\left(\sqrt{3}\right), 0.5\log(3)\right)$. These two constraints define the required value of μ_j. The range for σ_j is narrower than the prior range (see Appendix A), but still generates a wide variety of arc travel time variances. Comparisons between the estimation methods are invariant to moderate changes in the σ_j range.

We simulate datasets with two types of GPS data: "good" and "bad." The "good" GPS datasets are designed to mimic the conditions of the Toronto EMS dataset. Each GPS point is sampled at a travel distance of 250m after the previous point (straight-line distance is 200m in the Toronto EMS data, but we simulate data via the longer along-path distance). The GPS locations are drawn from a bivariate normal distribution with $\Sigma = \left(\begin{smallmatrix} 100 & 0 \\ 0 & 100 \end{smallmatrix}\right)$. The GPS speeds are drawn from a lognormal distribution with $\zeta^2 = 0.004$, which gives a mean absolute error of 5% of speed, approximately the average result seen by Witte and Wilson [2004].

The "bad" GPS datasets are designed to be sparse and have GPS error consistent with the high error results seen by Chen et al. [2005] and Witte and Wilson. GPS points are sampled every 1000m, which is still more frequent than the rate observed in the Tokyo taxi data [Lou et al. 2009]. The constant $\Sigma = \left(\begin{smallmatrix} 465 & 0 \\ 0 & 465 \end{smallmatrix}\right)$, which gives mean distance of 27 m between the true and observed locations, the average error seen in Hong Kong by Chen et al.. The parameter $\zeta^2 = 0.01575$, corresponding to mean absolute error of 10% of speed, which is approximately the result from low-quality GPS settings tested by Witte and Wilson.

5.2 Travel Time Prediction

We simulate ten good GPS datasets and ten bad GPS datasets, as defined in Section 5.1, each with a training set of 2000 trips and a test set of 2000 trips. Taking the true path for each test trip as known, we calculate point and 95% predictive interval estimates for the trip durations using the three methods. To obtain a "gold standard" for performance, we implement an "Oracle" method. In this method, the true travel time parameters $\left\{\mu_j, \sigma_j^2\right\}$ for each arc j

are known. The true expected travel time for each test trip is used as a point estimate. This implies that the Oracle method has the lowest possible root mean squared error (RMSE) for realized travel time estimation.

We compare the predictive accuracy of the point estimates from the four methods via the RMSE (in seconds), the median absolute error ("MdAE," in seconds), the mean bias as a percentage of mean trip duration ("Bias %;" see Equation 7) and the correlation between the predicted and true durations ("Cor."). We compare the interval estimates using the the percentage of 95% predictive intervals that contain the true trip duration ("Cov. %") and the arithmetic mean width of the 95% predictive intervals ("Width"). The "MdAE" metric is used for consistency with the Toronto EMS experiments (Section 6), where there are severe outliers in the real datasets. The "Bias %" metric is the following, if $\{T_i\}_{i=1}^{M}$ are true durations and $\left\{\hat{T}_i\right\}_{i=1}^{M}$ are predicted durations from a given method for M test trips:

$$\text{Bias \%} = \frac{\sum_{i=1}^{M}\left(\hat{T}_i - T_i\right)}{\sum_{i=1}^{M} T_i} \times 100\%. \tag{7}$$

Table 1 gives means for these metrics over the ten replications of good and bad datasets.

Good GPS data (Mean over ten datasets)						
Estimation method	RMSE (s)	MdAE (s)	Corr.	Bias %	Cov. %	Width (s)
Oracle	14.6	8.5	0.954	0.13	-	-
Bayesian	14.9	8.6	0.953	0.17	96.0	59.0
Local MLE	15.5	9.1	0.949	0.65	94.8	58.5
Local Harm.	15.5	9.1	0.949	0.59	94.4	57.9
Bad GPS data (Mean over ten datasets)						
Estimation method	RMSE (s)	MdAE (s)	Corr.	Bias %	Cov. %	Width (s)
Oracle	14.8	8.5	0.954	0.07	-	-
Bayesian	15.5	9.1	0.950	1.34	96.2	63.1
Local MLE	16.6	9.8	0.943	1.59	92.3	60.5
Local Harm.	16.6	9.7	0.943	1.45	91.1	57.9

Table 1: Out-of-sample trip travel time estimation performance on simulated data.

In both dataset types, the point estimates from the Bayesian method greatly outperform the estimates from the local methods. The Bayesian estimates closely approach the Oracle estimates, especially in the good GPS datasets. In the good datasets, the Bayesian method has 75% and 79% lower error than the local methods in RMSE and MdAE, respectively, after eliminating the unavoidable randomness of the oracle method. In the bad datasets, the Bayesian method outperforms the local methods by 62% and 53% in RMSE and MdAE, relative to the oracle method. The Bayesian method is almost unbiased on the good datasets. All three methods have over 1% bias on the bad data. The interval estimates are very similar across methods for the good GPS data. For the bad GPS data, the Bayesian intervals are slightly wider and have higher coverage percentage. The Bayesian estimates worsened by 4% in RMSE from good data to bad, while the local methods worsened by 7%. Because the bad GPS data are so sparse, the overall quantity of data is low. The Bayesian method appears to be more robust to this low-data situation. The Oracle method performs equally well on the two dataset types (up to simulation error).

5.3 Map-Matched Path Results

Next we assess map-matching estimates from the Bayesian method. Figure 2 shows two example ambulance paths. The black points show the GPS locations and the white nodes follow the true path taken. The starting node is marked in green and the ending node in red. Each arc is colored by the marginal posterior probability it is traversed in the path. Arcs with probability less than 1% are uncolored. The left-hand path is from a good GPS dataset (as defined in Section 5.1). The Bayesian method easily identifies the correct path. Every correct arc has close to 100% probability, and only two incorrect detours have probability above 1%. This is typical performance for simulated trips with good GPS data. The right-hand path is from a bad GPS dataset. The sparsity in GPS readings makes the path very uncertain. Near the beginning of the path, there are five routes with similar expected travel times, and the GPS readings do not distinguish between them, so each has close to 20% posterior probability. Near the end of the path, there are two routes with roughly 50% probability. The Bayesian method is very effective at identifying alternative routes when the true path is unclear.

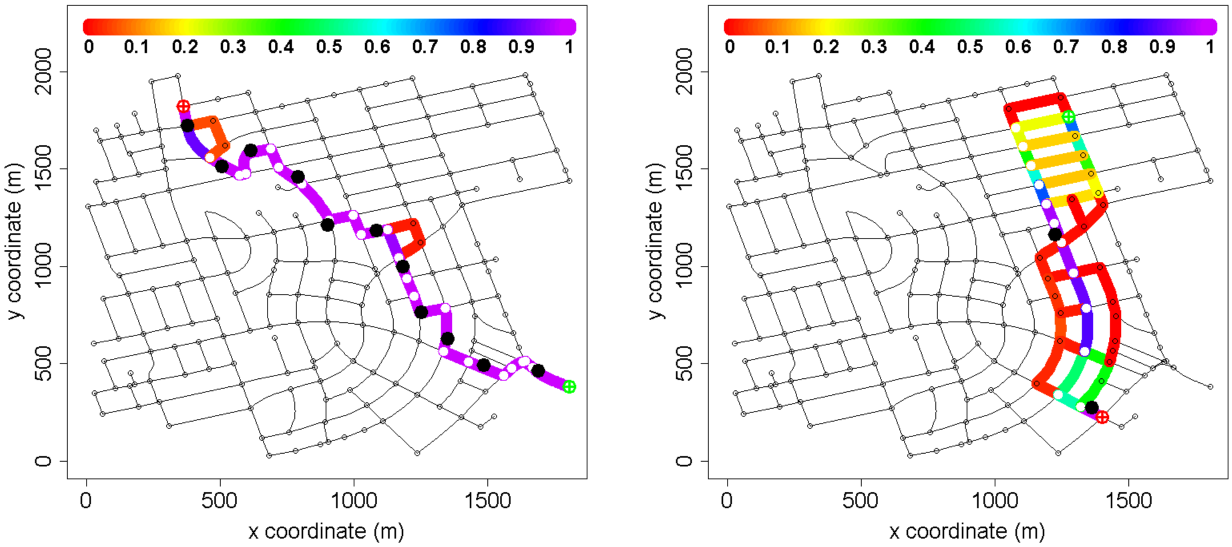

Figure 2: Map-matching estimates for two simulated trips, colored by the probability each arc is traversed.

6 Analysis of Toronto EMS Data

In this section, we compare the Bayesian and local methods on the dataset from Toronto EMS.

6.1 Data

The Toronto data consist of GPS data and trip information for ambulance trips with one of two priority levels: "lights-and-sirens" (L-S) or "standard travel" (Std). We address these separately, again focusing on the Leaside subregion of Toronto. The right plot in Figure 1 shows the GPS locations for the Std dataset. This dataset contains 3989 ambulance trips and almost 35,000 GPS points. The primary roads tend to have a large amount of data, the secondary roads a moderate amount, and the tertiary roads a small amount. The L-S dataset is smaller (1930 trips), with a similar spatial distribution of points.

We use only the portion of trips where the ambulance was driving to scene in response to an emergency call, and discard trips for which this cannot be identified. We also discard some trips (roughly 1%) that would impair estimation: for example, trips where the ambulance turned around or where the ambulance stopped for a long period (not at a stoplight or in traffic). Finally, most of the trips in the dataset do not begin or end in the subregion, they simply pass through, so we define start and end nodes and times appropriate for this subregion as follows. We use the closest node to the first GPS location as the approximated start node, and the time of the first GPS reading as the start time. Similarly, we use the last GPS reading for the end node. This produces some inaccuracy of estimated travel times on the boundary of the region. This could be fixed by applying our method to overlapping regions and discarding estimates on the boundary.

Table 2 compares rush hour (6-10 AM and 3-7 PM weekdays), non-rush hour, and overall mean GPS speeds, for the L-S and Std datasets. L-S speeds are higher than Std speeds. The mean speed decreases by roughly 8% during rush hour in L-S data, while the mean speed decreases by roughly 13% in Std data.

Mean GPS speed	All data	Rush hour	Non-rush hour
L-S	33.5 mph	31.3 mph	34.2 mph
Std	24.0 mph	21.7 mph	25.1 mph

Table 2: Mean observed GPS speeds for L-S and Std datasets.

6.2 Arc Travel Time Estimates

Here we report the travel time estimates from the Bayesian method. Toronto EMS has existing estimates of the travel times, which we use to set the prior $\{m_j\}_{j=1}^J$ hyperparameters (Appendix A). These estimates are different for L-S and Std trips, but are the same for the two travel directions of parallel arcs. We have also tested the Bayesian method with the data-based hyperparameters described in Appendix A and have observed similar performance. Figure 3 shows prior and posterior speed estimates (length divided by mean travel time) from the Bayesian method on the L-S dataset. Each arc is colored based on its speed estimate, so most roads have two colors, corresponding to the parallel arcs in each direction.

The posterior speed estimates from the Bayesian method are reasonable; primary arcs tend to have high speed estimates, and estimated speeds for consecutive arcs on the same road are typically similar. Arcs heading into major intersections (intersections between two primary or secondary roads, as shown in Figure 1) are often slower than the reverse arcs. This

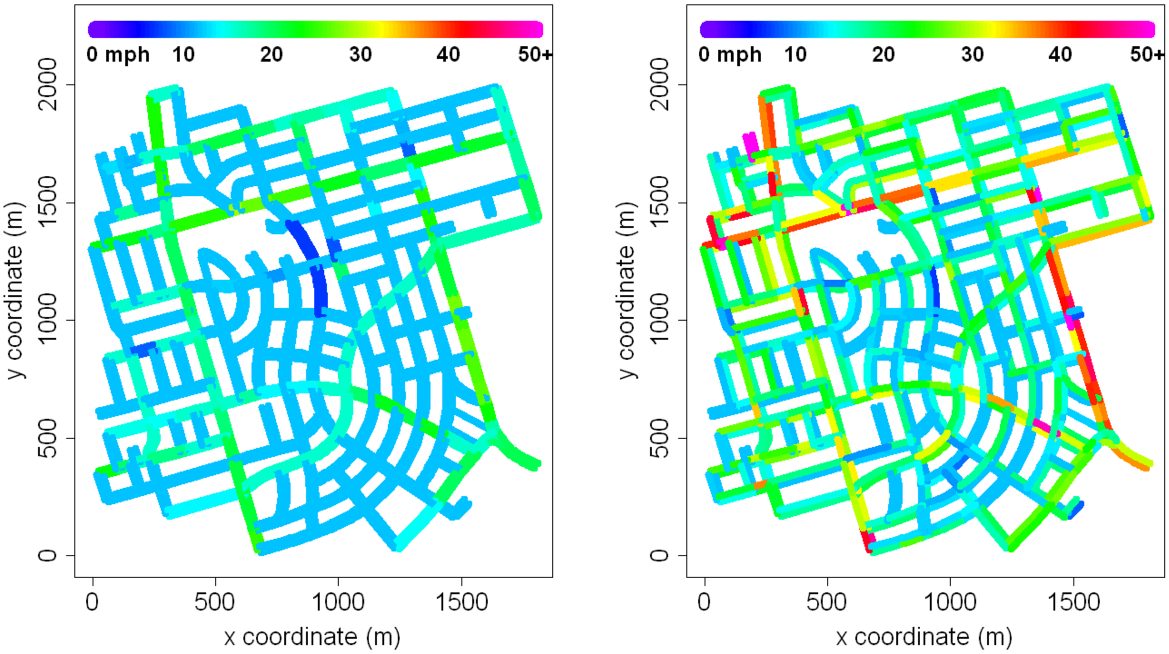

Figure 3: Prior (left) and posterior (right) speeds from the Bayesian method, for Toronto L-S data.

effect cannot be captured by the local methods, because both travel directions have the same estimates. Arcs with little data usually have posterior estimated speeds close to the prior estimates. For most arcs the posterior estimate of the speed is higher than the prior estimate, suggesting that the existing road speed estimates used to specify the prior are underestimates.

There are a few arcs that have poor estimates from the Bayesian method. For example, parallel pink arcs in the top-left corner have poor estimates due to edge effects. Also, some short interior arcs have unrealistically high estimates, likely because there are few GPS points on these arcs. This undesirable behavior could be reduced or eliminated by using a random effect prior distribution for these roads [Gelman et al. 2004], which has the effect of pooling the available data.

6.3 Travel Time Prediction

Next we evaluate the accuracy of travel time predictions from the Bayesian and local methods. We divide the set of trips from each dataset randomly into two halves to create training and test sets, fit the models to the training data, and predict the travel times for the trips in the test data. We repeat this five times. We also test whether the predictive accuracy is improved by fitting the model separately to rush-hour and non-rush hour training data.

We compare the known duration of each trip in the test data with the point and 95% interval predictions from each method. Unlike the simulated test data in Section 5, the true paths are not known. We use point estimates for the path taken; for each estimation method, we assume the path taken is the the expected fastest path, using the mean travel time estimates for each arc. This measures the ability of the methods to estimate both the fastest path and the travel time distributions accurately.

As in Section 5.2, we compare the point estimates from the three methods on the test data using RMSE, MdAE, Cor. and Bias %, and compare the interval estimates using Width and Cov. %. Mean results from the five replications (ten total test sets) are presented in Table 3. Because the true travel time distributions are unknown, we cannot use the Oracle method as in Section 5.2. However, we still wish to estimate "gold standard" performance, so we implement an "Estimated Oracle" method, in which we assume that the parametric model and MLE estimates from the Local MLE method are the truth. We simulate realized travel times on the fastest path for each test trip (in expectation, as estimated by the Local MLE method), and compare these to the expected value point estimates used by that method. To avoid simulation error, we use Monte Carlo mean estimates from 1000 replications of simulated travel times. Standard errors for these mean estimates are roughly 0.03 for RMSE, 0.01 for MdAE and 10^{-4} for Cor. and Bias % (for both L-S and Std data).

For the L-S data, the Bayesian method outperforms the local methods, suggesting that it is effectively utilizing whole-trip information, unlike the local methods which treat the GPS readings as isolated observations. The Bayesian method is substantially better in MdAE (12%, 16%) and slightly better in RMSE than the two local methods. The improvements increase to 27% and 35% in MdAE after subtracting the baseline error from the Estimated Oracle method. Partitioning into time bins did not substantially change performance of any of the methods.

The Bayesian method also substantially outperforms the local methods in interval estimates. The Bayesian intervals have much higher coverage percentage than the intervals from

L-S data. Averages over 5 replications (10 test sets).						
Estimation method	RMSE (s)	MdAE (s)	Cor.	Bias %	Cov. %	Width (s)
Est. oracle (1 time bin)	15.9	8.6	0.950	0.0	-	-
Bayesian (1 time bin)	38.1	13.6	0.777	-2.8	85.6	76.3
Local MLE (1 time bin)	38.6	15.4	0.771	0.2	69.9	58.2
Local Harm. (1 time bin)	38.8	16.2	0.770	2.2	72.7	80.6
Bayesian (2 time bins)	37.9	13.9	0.779	-1.4	84.9	76.6
Local MLE (2 time bins)	38.7	15.2	0.769	-1.4	68.2	54.6
Local Harm. (2 time bins)	38.8	15.9	0.768	0.3	69.0	73.3
Std data. Averages over 5 replications (10 test sets).						
Estimation method	RMSE (s)	MdAE (s)	Cor.	Bias %	Cov. %	Width
Est. oracle (1 time bin)	30.1	16.5	0.927	0.0	-	-
Bayesian (1 time bin)	130.0	42.6	0.610	-17.2	72.2	157.0
Local MLE (1 time bin)	136.4	42.5	0.600	-23.2	55.2	110.7
Local Harm. (1 time bin)	134.8	42.7	0.601	-21.5	64.1	138.5
Bayesian (2 time bins)	124.4	38.9	0.647	-15.3	74.1	158.3
Local MLE (2 time bins)	135.2	41.5	0.631	-24.8	54.5	104.9
Local Harm. (2 time bins)	133.7	41.1	0.631	-23.3	62.3	129.0

Table 3: Out-of-sample trip travel time estimation performance on Toronto EMS data.

the local methods. The Bayesian method takes the uncertainty in travel time parameters into account when generating predictive intervals, unlike the Local MLE method. The intervals from the MLE method are narrow and have poor coverage percentage. Therefore, the Local MLE method does not adequately account for the variability of travel times. This suggests that the Estimated Oracle method may underestimate the baseline error, in which case the Bayesian method would outperform the local methods by an even larger amount, relative to the baseline error.

The Bayesian method has larger bias on the L-S data than the local methods. This is a result of outliers in the data. There are several trips with abnormally large travel times. We believe that these are Std trips that were erroneously recorded to be L-S. This is why we report MdAE as well as RMSE. The local methods do not show the same bias, because this bias is counteracted by inherent biases in the positive direction (see Section 4 and Appendix B).

For the Std data, the Bayesian method outperforms the local methods by roughly 5% in RMSE, relative to the Estimated Oracle error, when using one time bin. The three methods are effectively equal in MdAE. The Bayesian method has higher correlation and lower bias.

Again, the interval estimates show substantial improvement. The Std dataset is more complex, with wider travel time distributions than the L-S dataset. There are many trips with large travel times, caused by slow traffic. This accounts for the large negative bias in all methods. Most of the slowest times are from rush hour. When applied to rush hour and non-rush hour separately, the Bayesian method improves substantially, outperforming the local methods by roughly 10% in both RMSE and MdAE, again relative to the Estimated Oracle performance.

6.4 Response Within Time Threshold

In this section, we consider the probability an ambulance completes its trip within a certain time threshold. These probabilities are critical for EMS providers (see Section 1). In Figure 4, we assume that an ambulance begins at the node marked with a black "X" and follows the fastest paths (in expectation) to each other node. The other nodes are colored by the probability the ambulance arrives within 2 minutes (left) or 3 minutes (right), using posterior travel time distribution estimates from the Bayesian method for each arc.

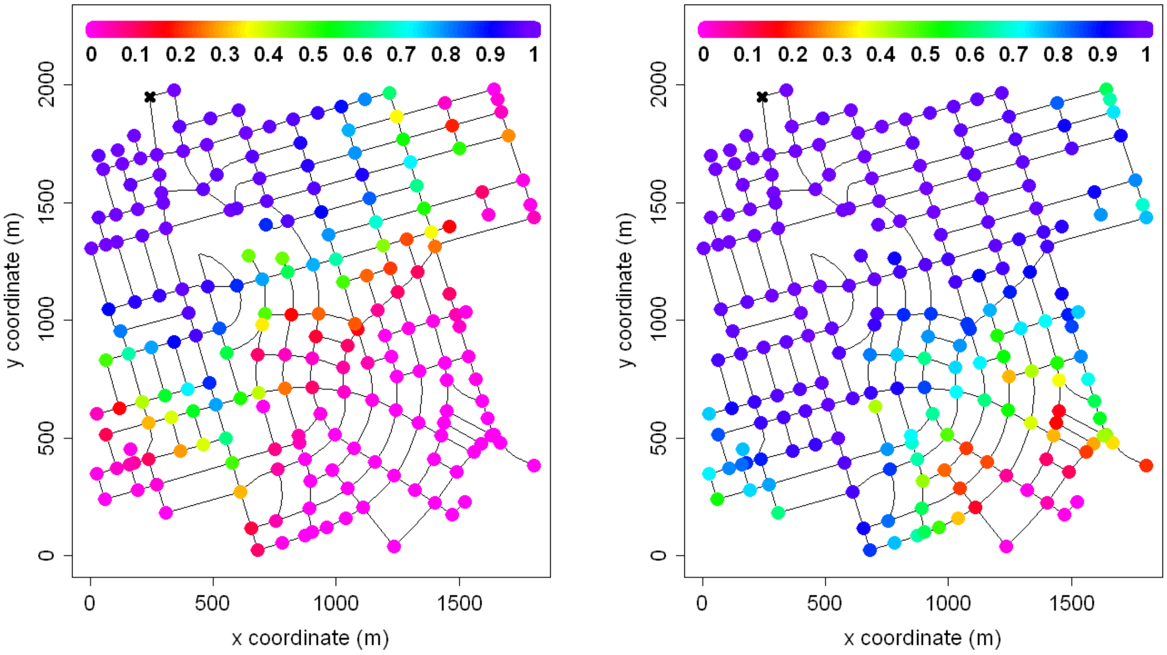

Figure 4: Bayesian estimates of probability of reaching each node in 120 seconds (left), 180 seconds (right), from the location marked "X."

The probabilities appear reasonable: high for nodes close to the start node, decreasing for nodes further from the start node, and typically higher for nodes on primary and secondary roads than for nodes on the slower tertiary roads (roughly the same distance from the start node). A method such as the one presented in Budge et al. [2010] that only uses travel distance cannot account for the speed difference between primary and tertiary roads.

Results from the local MLE method (not shown) are similar, but the decrease in probability is less gradual; there are more nodes with very high or very low probabilities. This is because the predictive intervals for the MLE method are generally narrower, with worse coverage than the predictive intervals from the Bayesian method (see Table 3). Table 4 shows the number of nodes with probabilities less than 5% or greater than 95% in Figure 4 for the two methods. In all cases, the Bayesian method has fewer nodes at these extremes than the MLE method.

	120 seconds		180 seconds	
	$\leq 5\%$	$\geq 95\%$	$\leq 5\%$	$\geq 95\%$
Bayesian	70	46	4	102
Local MLE	89	52	10	103

Table 4: Number of nodes with very high or low probabilities of being reached in a time threshold (out of 203 total nodes).

6.5 Map-Matched Path Accuracy

Finally, we assess map-matching estimates from the Bayesian method, for the Toronto L-S data. Figure 5 shows two example ambulance paths from the L-S dataset. The first GPS point is colored green, the last red, and the others black. As in Section 5.3, each arc is colored by its marginal posterior probability, if it is greater than 1%. In the left-hand path, there are two occasions where the path is not precisely defined by the GPS readings. On both occasions, most of the posterior probability ($\approx 90\%$) is given to a route following the main road, which is estimated to be faster. The final two GPS readings appear to have location error. However, the fastest path is still given roughly 100% posterior probability, instead of a detour that would be slightly closer to the second-to-last GPS reading. In the right-hand path, for an unknown reason, there is a large gap between GPS points. Almost all the posterior probability is given to the fastest route (see Section 5.3), following a primary and then secondary road. This illustrates the robustness of the Bayesian method to sparse GPS data.

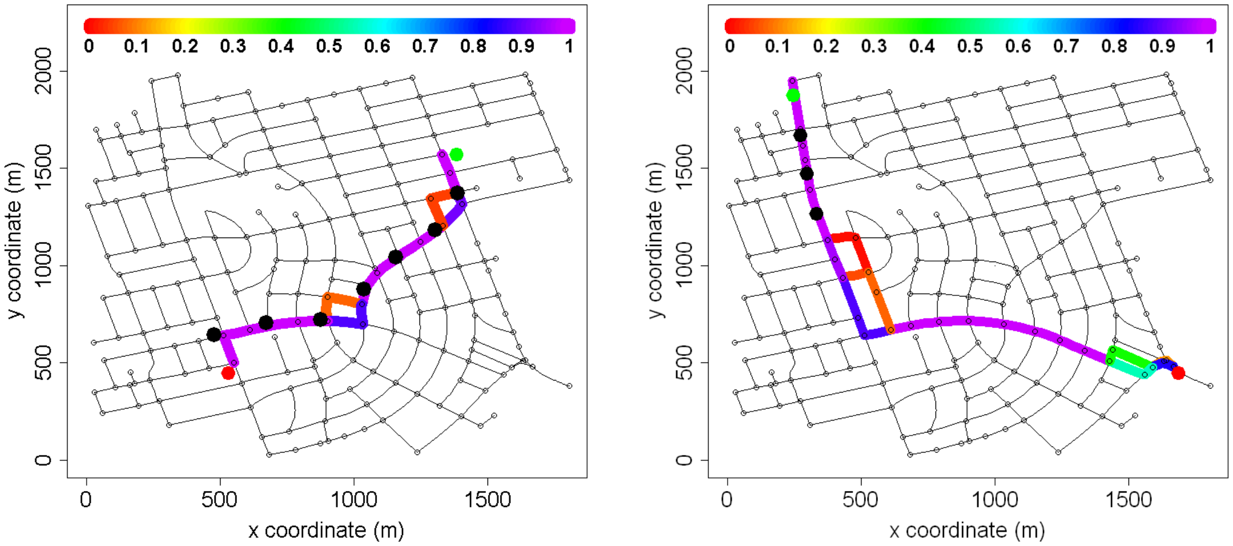

Figure 5: Map-matching estimates for two Toronto L-S trips, colored by the probability each arc is traversed.

7 Conclusions

We proposed a Bayesian method to estimate ambulance travel time distributions on each road segment in a city, using sparse and error-prone GPS data. We simultaneously estimated the ambulance paths and the parameters of the travel time distributions on each segment. We also introduced two “local” methods based on mapping each GPS reading to the nearest road segment. The first method used the harmonic mean of the GPS speeds; the second performed maximum-likelihood estimation for a lognormal distribution of travel speeds on each segment.

We applied the three methods to simulated data and data from Toronto EMS, in a subregion of Toronto. In simulations, the Bayesian method substantially outperformed the local methods in estimating out-of-sample trip durations, for both point and interval estimates. The estimates from the Bayesian method remained excellent even when the GPS data had high error. On the Toronto EMS data, the Bayesian method provided reasonable travel time estimates for the road segments in the network, and again outperformed the competing methods in out-of-sample prediction.

We also investigated the effect of applying these methods to data binned by time of day. There was little effect on predictive accuracy for “lights-and-sirens” data, but accuracy on “standard travel” data improved. One could also consider smooth estimation of the travel times as a function of time of day, using a functional data analysis approach.

A number of other extensions are possible, for further improving predictive accuracy. Random effect modeling of travel time parameters within a road class could provide more smoothing [Gelman et al. 2004]. Dependence between arc travel times within each trip could capture local traffic congestion effects or an ambulance driver’s speed preference. A heavier-tailed distribution could be used in place of the lognormal, to account for very large travel times more naturally. This could be particularly effective for “standard travel” data. Alternatively, the lognormal travel time assumption of the Bayesian method could be weakened to a mixture of lognormals. This could capture the different travel time distributions at intersections, depending on the light cycle and direction the ambulance turns.

In this paper, we have limited our discussion to the Leaside subregion of Toronto. Analysis of a very large dataset with ambulance trips for an entire city may require substantially more computing resources. In future work, we will introduce approximate computational methods for the Bayesian method suitable for this setting.

A Constants and Hyperparameters

There are several constants and hyperparameters to be specified in the Bayesian model. To set the GPS position error covariance matrix Σ, we calculate the minimum distance from each GPS location in the data to the nearest arc. Assuming that the error is radially symmetric, that the vehicle was on the nearest arc when it generated the GPS point, and approximating that arc locally by a straight line, this minimum distance should equal the absolute value of one component of the 2-dimensional error, i.e. the absolute value of a random variable

$\mathcal{E}_1 \sim N(0, \sigma^2)$, where $\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$. Since $E(|\mathcal{E}_1|) = \sigma\sqrt{2/\pi}$, we take $\hat{\sigma} = \hat{E}(|\mathcal{E}_1|)\sqrt{\pi/2}$, where $\hat{E}(|\mathcal{E}_1|)$ is the mean minimum distance of each GPS point to the nearest arc in the data. In the Toronto EMS datasets, we have $\hat{E}(|\mathcal{E}_1|) = 8.4$ m for the L-S data and 7.7 m for the Std data, yielding $\Sigma_{\text{L-S}} = \left(\begin{smallmatrix} 111.6 & 0 \\ 0 & 111.6 \end{smallmatrix}\right)$ and $\Sigma_{\text{Std}} = \left(\begin{smallmatrix} 92.7 & 0 \\ 0 & 92.7 \end{smallmatrix}\right)$. In the simulated data, a typical dataset has $\hat{E}(|\mathcal{E}_1|) = 7.3$ m for "good" GPS data and 14.1 m for "bad" GPS data, yielding $\Sigma_{\text{Good}} = \left(\begin{smallmatrix} 84 & 0 \\ 0 & 84 \end{smallmatrix}\right)$, and $\Sigma_{\text{Bad}} = \left(\begin{smallmatrix} 312 & 0 \\ 0 & 312 \end{smallmatrix}\right)$.

The hyperparameters b_1, b_2, s^2, and m_j control the prior distributions on the travel time parameters μ_j and σ_j^2 (Equation 6). We set b_1 and b_2 by estimating the possible range in travel time variation for a single arc. Some arcs have very consistent travel times (for example an arc with little traffic and no major intersections at either end). We estimate that such an arc could have travel time above or below the median time by a factor of 1.1. Taking this range to be a two standard deviation σ_j interval (so that $1.1\exp(\mu_j) = \exp(\mu_j + 2\sigma_j)$) yields $\sigma_j \approx 0.0477$. Other arcs have very variable travel times (for example an arc with substantial traffic). We estimate that such an arc could have travel time above or below the median time by a factor of 3.5, corresponding to $\sigma_j \approx 0.6264$. Thus, we set $b_1 = 0.0477$ and $b_2 = 0.6264$.

We assume we have an initial travel time estimate τ_j for each arc j (for example, in Section 6 we use previous estimates from Toronto EMS). We expect this estimate to be typically correct within a factor of two. Thus, we specify m_j and s^2 so that the prior distribution for $E[T_{i,j}]$ is centered at τ_j and has a two standard deviation interval from $\tau_j/2$ to $2\tau_j$. This gives

$$\begin{aligned} \tau_j &= E\left(\exp\left(\mu_j + \sigma_j^2/2\right)\right) \\ &= \exp\left(m_j + s^2/2\right) E\left(\exp\left(\sigma_j^2/2\right)\right), \\ \frac{\tau_j}{2} &= \exp\left(m_j + s^2/2 - 2s\right) E\left(\exp\left(\sigma_j^2/2\right)\right), \\ 2\tau_j &= \exp\left(m_j + s^2/2 + 2s\right) E\left(\exp\left(\sigma_j^2/2\right)\right), \end{aligned}$$

where the final equation is redundant. Therefore,

$$m_j = \log\left(\frac{\tau_j}{E\left(\exp\left(\sigma_j^2/2\right)\right)}\right) - \frac{s^2}{2}, \qquad s = \frac{\log(2)}{2}.$$

When τ_j is not available, as in our simulation study, one can use the following data-based choice for τ_j: find the harmonic mean GPS speed reading in the entire dataset and convert this speed to a travel time for each road.

Results are very insensitive to the hyperparameters b_3 and b_4, as long as the interval $[b_3, b_4]$ does not exclude regions of high likelihood. This is because the entire dataset is used to estimate ζ^2 (unlike for the parameters σ_j^2). We fix $b_3 = 0$ and $b_4 = 0.5$. For observed GPS speed V_i^ℓ, suppose the true speed at that moment is v. By Equation 4, $V_i^\ell \sim \text{Log-}N\left(\log(v) - \zeta^2/2, \zeta^2\right)$. If $\zeta = 0.5$, we estimate by simulation that

$$\frac{E\left(\left|V_i^\ell - v\right|\right)}{v} \approx 0.4,$$

which is much higher than any mean absolute error observed by Witte and Wilson [2004]. Thus, it is not realistic that the error could be greater than this.

The hyperparameter C governs the multinomial logit choice model prior distribution on paths. While the results of the Bayesian method are generally insensitive to moderate changes in the other hyperparameters, changes in the value of C do have a noticeable effect, so we obtain a careful data-based estimate. Equation 2 implies that the ratio of the probabilities of two possible paths depends on their difference in expected travel time. For example, let $C = 0.1$ and consider paths $\tilde{a}_i$ and $\dot{a}_i$ from d_i^1 to d_i^2, where the expected travel time of $\tilde{a}_i$ is 10 seconds less than the expected travel time of $\dot{a}_i$. Then path $\tilde{a}_i$ is $e \approx 2.72$ times more likely.

We specify C using the principle that for a trip of average duration, a driver is ten times less likely to choose a path that has 10% longer travel time. If $\bar{T}$ is the average trip duration, then by Equation 2, this requires

$$0.1 = \frac{\exp\left(-C\left(1.1\bar{T}\right)\right)}{\exp\left(-C\bar{T}\right)} = \exp\left(-0.1C\bar{T}\right), \tag{8}$$

giving $C = -\log(0.1)/\left(0.1\bar{T}\right)$. For our simulated data, this yields $C_{\text{Sim}} = 0.24$.

On the real data, we make a small adjustment to pool information across the L-S and Std datasets. Observing that the route choices are very similar in visual inspection of these datasets, we ensure that the prior distribution on the route taken between two fixed locations is the same for the L-S and Std datasets. To do this, we combine all the L-S and Std data to calculate an overall mean L_1 trip length L_1^{Tor} (change in x coordinate plus change in y coordinate) for the Toronto EMS data, which is $L_1^{\text{Tor}} = 1378.8$m. Let $L_1^{\mathcal{D}}$ and $T^{\mathcal{D}}$ be the mean L_1 length and mean trip duration for each dataset $\mathcal{D}$. We estimate a weighted mean duration $T_W^{\mathcal{D}} = T^{\mathcal{D}} L_1^{\text{Tor}} / L_1^{\mathcal{D}}$ for dataset $\mathcal{D}$ for a trip of length L_1^{Tor}, and use the duration $T_W^{\mathcal{D}}$ to set C by Equation 8. This yields $C_{\text{L-S}} = 0.211$ and $C_{\text{Std}} = 0.110$.

B Harmonic Mean Speed and GPS Sampling

When estimating road segment travel times via speed data from GPS readings, as in the local methods of Section 4, it is critical whether the GPS readings are sampled by distance or by time. Sampling-by-distance could mean recording a GPS point every 100m, and sampling-by-time could mean recording a GPS point every 30s, for example. As discussed in Sections 1 and 4, most EMS providers use a combination of distance and time sampling. If both constraints are satisfied frequently (unlike in the Toronto EMS dataset, where most points are sampled by distance), this could create a problem for estimating travel times via these speeds.

In the transportation research literature, where sampling is done by distance (because speeds are recorded at loop detectors at fixed locations on the road), it is well known that the harmonic mean of the observed speeds (the "space mean speed") is appropriate for estimating travel times [Rakha and Zhang 2005, Soriguera and Robuste 2011, Wardrop 1952]. Under a simple probabilistic model of sampling-by-distance, without assuming constant speed, we confirm that the harmonic mean speed gives an unbiased and consistent estimator of the mean travel time. However, we also show that if the sampling is done by time, the harmonic mean is biased towards overestimating the mean travel time.

Consider a set of n ambulance trips on a single road segment. For convenience, let the length of the road segment be 1. Let the travel time on the segment for ambulance i be T_i, and assume that the T_i are iid with finite expectation. Let $x_i(t)$ be the position function of ambulance i, conditional on T_i, so $x_i(0) = 0$ and $x_i(T_i) = 1$. Assume that $x_i(t)$ is continuously differentiable, with derivative $v_i(t)$, the velocity function, and that $v_i(t) > 0$ for all t. Each trip samples one GPS point. Let V_i^o be the observed GPS speed for the ith ambulance.

First, consider sampling-by-distance. For trip i, draw a random location $\xi_i \sim \text{Unif}(0, 1)$ at which to sample the GPS point. This is different from the example of sampling-by-distance above. However, if the sampling locations are not random, we cannot say anything about the observed speeds in general (the ambulances might briefly speed up significantly where the reading is observed, for example). Assuming that the ambulance trip started before this road segment, it is reasonable to model sampling-by-distance with a uniform random location.

Conditional on T_i, $x_i(\cdot)$ is a cumulative distribution function, with support $[0, T_i]$, density $v_i(\cdot)$, and inverse $x_i^{-1}(\cdot)$. Thus, $\tau_i = x_i^{-1}(\xi_i)$, the random time of the GPS reading, has distribution function $x_i(\cdot)$ and density $v_i(\cdot)$, by the probability integral transform. The observed speed $V_i^o = v_i(\tau_i)$, so the GPS reading is more likely to be sampled when the ambulance has high speed than when it has low speed. This is called the inspection paradox (see e.g. Stein and Dattero [1985]). Mathematically,

$$E(\,V_i^o|\,T_i) = E(\,v_i(\tau_i)|\,T_i) = \int_0^{T_i} v_i(t)v_i(t)dt \geq \frac{\left(\int_0^{T_i} v_i(t)dt\right)^2}{\int_0^{T_i} 1^2 dt} = \frac{1}{T_i},$$

by the Cauchy-Schwarz inequality, with strict inequality unless $v_i(\cdot)$ is constant. However, if we draw a uniform time $\phi_i \sim U(0, T_i)$, then

$$E(v_i(\phi_i)\,|T_i\,) = \int_0^{T_i} v_i(t)\frac{1}{T_i}dt = \frac{1}{T_i}. \tag{9}$$

In particular this implies that the speeds summarized in Table 2 are biased high. The inspection paradox has a greater impact in the Toronto Std data than in the L-S data, because ambulance speed varies more in standard travel.

Consider estimating the mean travel time $E(T_i)$ via the estimator $\hat{T}^H = 1/\bar{V}_H^o$, where $\bar{V}_H^o$

is the harmonic mean observed speed. We have

$$
\begin{aligned}
E\left(\hat{T}^H\right) &= E\left(E\left(\hat{T}^H \middle| \{T_i\}_{i=1}^n\right)\right) = E\left(\frac{1}{n}\sum_{i=1}^n E\left(\frac{1}{v_i(\tau_i)}\middle| T_i\right)\right) \\
&= E\left(\frac{1}{n}\sum_{i=1}^n \int_{t=0}^{T_i} \frac{1}{v_i(t)} v_i(t)dt\right) = E\left(\frac{1}{n}\sum_{i=1}^n T_i\right) = E(T_i),
\end{aligned}
$$

and so it is unbiased. Moreover, it is consistent as $n \to \infty$, by the Law of Large Numbers.

Next, suppose the sampling is instead done by time. To model this, let $\tau_i \sim \text{Unif}(0, T_i)$ be a random time to sample the GPS point for ambulance i. In this case, we have

$$
\begin{aligned}
E\left(\hat{T}^H\right) &= E\left(\frac{1}{n}\sum_{i=1}^n E\left(\frac{1}{v_i(\tau_i)}\middle| T_i\right)\right) \\
&\geq E\left(\frac{1}{n}\sum_{i=1}^n \frac{1}{E\left(v_i(\tau_i)\middle| T_i\right)}\right) \\
&= E\left(\frac{1}{n}\sum_{i=1}^n \frac{1}{\frac{1}{T_i}}\right) = E(T_i),
\end{aligned}
$$

by Jensen's Inequality and Equation 9. Again, the inequality is strict unless $v_i(\cdot)$ is constant.

C Calculations for Updating the Paths

Here we calculate the ratio of proposals q and Jacobian $|\text{Ja}|$ from Section 3.2. First, for the ratio of proposals. In Part 1 of the proposal in Section 3.2, the node d' is chosen with probability $1/N_i$, where N_i is the number of arcs in A_i. In Part 2, the node d'' is chosen with probability $1/\min(a, K)$. In Part 3, the number of routes of length up to K between d' and d'' is the same for the reverse proposal, so this probability cancels. Finally, the ratio of travel time densities can be calculated easily. Letting Dir denote the Dirichlet density,

$$
\frac{q\left(A_i, T_i \middle| A_i^*, T_i^*, \left\{\mu_j, \sigma_j^2\right\}_{j=1}^J\right)}{q\left(A_i^*, T_i^* \middle| A_i, T_i, \left\{\mu_j, \sigma_j^2\right\}_{j=1}^J\right)} = \frac{N_i \min(a, K)}{N_i^* \min(a^*, K)} \frac{\text{Dir}\left(\frac{T_i(c_1)}{S_i}, \ldots, \frac{T_i(c_m)}{S_i}; \alpha\theta(c_1), \ldots, \alpha\theta(c_m)\right)}{\text{Dir}\left(r_1, \ldots, r_n; \alpha\theta(p_1), \ldots, \alpha\theta(p_n)\right)} S_i^{n-m}.
$$

To calculate the Jacobian $|\text{Ja}|$, define random variables $U_\ell = r_\ell S_i$, for $\ell \in \{1, \ldots, n-1\}$

(emphasizing that the space of travel time proposals has dimension $n-1$). To take the same role for the reverse proposal, define $W_\ell = T_i(c_\ell)$, for $\ell \in \{1,\ldots,m-1\}$. Thus, we have a transformation between two spaces of dimension $m+n-1$, with Jacobian

$$|\mathrm{Ja}| = \begin{vmatrix} \frac{\partial T_i(c_1)}{\partial T_i^*(p_1)} & \cdots & \frac{\partial T_i(c_m)}{\partial T_i^*(p_1)} & \frac{\partial U_1}{\partial T_i^*(p_1)} & \cdots & \frac{\partial U_{n-1}}{\partial T_i^*(p_1)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial T_i(c_1)}{\partial T_i^*(p_n)} & \cdots & \frac{\partial T_i(c_m)}{\partial T_i^*(p_n)} & \frac{\partial U_1}{\partial T_i^*(p_n)} & \cdots & \frac{\partial U_{n-1}}{\partial T_i^*(p_n)} \\ \frac{\partial T_i(c_1)}{\partial W_1} & \cdots & \frac{\partial T_i(c_m)}{\partial W_1} & \frac{\partial U_1}{\partial W_1} & \cdots & \frac{\partial U_{n-1}}{\partial W_1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial T_i(c_1)}{\partial W_{m-1}} & \cdots & \frac{\partial T_i(c_m)}{\partial W_{m-1}} & \frac{\partial U_1}{\partial W_{m-1}} & \cdots & \frac{\partial U_{n-1}}{\partial W_{m-1}} \end{vmatrix} = 1,$$

by cofactor expansion.

Acknowledgements

We would like to thank The Optima Corporation and Dave Lyons of Toronto EMS. This research was partially supported by the National Science Foundation under Grant CMMI-0926814.

Affiliations

Bradford S. Westgate
Cornell University, School of Operations Research and Information Engineering
294 Rhodes Hall, Ithaca, NY, 14853
Email: bsw62@cornell.edu
Website: `https://confluence.cornell.edu/display/~bsw62/`

Dawn B. Woodard
Cornell University, School of Operations Research and Information Engineering
228 Rhodes Hall, Ithaca, NY, 14853

David S. Matteson
Cornell University, Department of Statistical Science
293 Ives Faculty Building, Ithaca, NY, 14853

Shane G. Henderson
Cornell University, School of Operations Research and Information Engineering
230 Rhodes Hall, Ithaca, NY, 14853

References

K. Aladdini. EMS response time models: A case study and analysis for the region of Waterloo. Master's thesis, University of Waterloo, 2010.

R. Alanis, A. Ingolfsson, and B. Kolfal. A Markov Chain model for an EMS system with repositioning. 2010. Working paper.

L. Brotcorne, G. Laporte, and F. Semet. Ambulance location and relocation models. *European Journal of Operational Research*, 147:451–463, 2003.

S. Budge, A. Ingolfsson, and D. Zerom. Empirical analysis of ambulance travel times: The case of Calgary emergency medical services. *Management Science*, 56:716–723, 2010.

W. Chen, Z. Li, M. Yu, and Y. Chen. Effects of sensor errors on the performance of map matching. *The Journal of Navigation*, 58:273–282, 2005.

S.F. Dean. Why the closest ambulance cannot be dispatched in an urban emergency medical services system. *Prehospital and Disaster Medicine*, 23:161–165, 2008.

E. Erkut, A. Ingolfsson, and G. Erdoğan. Ambulance location for maximum survival. *Naval Research Logistics (NRL)*, 55:42–58, 2008. ISSN 1520-6750.

J.J. Fitch. *Prehospital Care Administration: Issues, Readings, Cases.* St. Louis: Mosby-Year Book, 1995.

A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472, 1992.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis.* London: Chapman & Hall, 2004.

J.B. Goldberg. Operations research models for the deployment of emergency services vehicles. *EMS Management Journal*, 1:20–39, 2004.

P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

S.G. Henderson. Operations research tools for addressing current challenges in emergency medical services. In *Wiley Encyclopedia of Operations Research and Management Science.* New York: Wiley, 2010.

A. Ingolfsson, S. Budge, and E. Erkut. Optimal ambulance location with random delays and travel times. *Health Care Management Science*, 11:262–274, 2008. ISSN 1386-9620.

J. Krumm, J. Letchner, and E. Horvitz. Map matching with travel time constraints. In *Society of Automotive Engineers (SAE) 2007 World Congress*, 2007.

Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang. Map-matching for low-sampling-rate GPS trajectories. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 352–361. ACM, 2009.

F. Marchal, J. Hackney, and K.W. Axhausen. Efficient map matching of large Global Positioning System data sets: Tests on speed-monitoring experiment in Zurich. *Transportation Research Record: Journal of the Transportation Research Board*, 1935:93–100, 2005.

A.J. Mason. Emergency vehicle trip analysis using GPS AVL data: A dynamic program for map matching. In *Proceedings of the 40th Annual Conference of the Operational Research Society of New Zealand. Wellington, New Zealand*, pages 295–304, 2005.

D. McFadden. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, pages 105–142. New York: Academic Press, 1973.

N.J. Nilsson. *Artificial Intelligence: A New Synthesis*. San Francisco: Morgan Kaufmann, 1998.

H. Rakha and W. Zhang. Estimating traffic stream space mean speed and reliability from dual-and single-loop detectors. *Transportation Research Record: Journal of the Transportation Research Board*, 1925:38–47, 2005. ISSN 0361-1981.

S. Richardson and P.J. Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59: 731–792, 1997.

C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 2004.

G.O. Roberts and J.S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16:351–367, 2001.

F. Soriguera and F. Robuste. Estimation of traffic stream space mean speed from time aggregations of double loop detector data. *Transportation Research Part C: Emerging Technologies*, 19:115–129, 2011. ISSN 0968-090X.

W.E. Stein and R. Dattero. Sampling bias and the inspection paradox. *Mathematics Magazine*, 58: 96–99, 1985. ISSN 0025-570X.

S. Syed. Development of map aided GPS algorithms for vehicle navigation in urban canyons. Master's thesis, University of Calgary, 2005.

M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987.

L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22:1701–1728, 1994. ISSN 0090-5364.

J.G. Wardrop. Some theoretical aspects of road traffic research. *Proceedings of the Institute of Civil Engineers*, 2:325–378, 1952.

T.H. Witte and A.M. Wilson. Accuracy of non-differential GPS for the determination of speed over ground. *Journal of Biomechanics*, 37:1891–1898, 2004. ISSN 0021-9290.