SUPPLEMENTARY MATERIAL FOR

TRAVEL TIME ESTIMATION FOR AMBULANCES USING BAYESIAN DATA AUGMENTATION

By Bradford S. Westgate, Dawn B. Woodard, David S. Matteson and Shane G. Henderson

APPENDIX 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 $\Sigma_{\text{L-S}} = \begin{pmatrix} 111.6 & 0 \\ 0 & 111.6 \end{pmatrix}$ and $\Sigma_{\text{Std}} = \begin{pmatrix} 92.7 & 0 \\ 0 & 92.7 \end{pmatrix}$. 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 $\Sigma_{\text{Good}} = \begin{pmatrix} 84 & 0 \\ 0 & 84 \end{pmatrix}$, and $\Sigma_{\text{Bad}} = \begin{pmatrix} 312 & 0 \\ 0 & 312 \end{pmatrix}$.

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 (Section 2.2). 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 there exists an initial travel time estimate τ_j for each arc j. For example, in Section 7 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 2.3, $V_i^\ell \sim \mathcal{LN}\left(\log(v) - \zeta^2/2, \zeta^2\right)$. 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). 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 2.1 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^s to d_i^f, 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 by the principle that for a trip of average travel time, a driver is ten times less likely to choose a path that has 10% longer travel time. If $\bar{T}$ is the average travel time, then by Equation 2.1, this requires

$$0.1 = \frac{\exp\left(-C\left(1.1\bar{T}\right)\right)}{\exp\left(-C\bar{T}\right)} = \exp\left(-0.1C\bar{T}\right), \tag{A.1}$$

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

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 time for each dataset $\mathcal{D}$. We estimate a weighted mean time $T_W^{\mathcal{D}} = T^{\mathcal{D}} L_1^{\text{Tor}}/L_1^{\mathcal{D}}$ for dataset $\mathcal{D}$ for a trip of length L_1^{Tor}, and use the time $T_W^{\mathcal{D}}$ to set C by Equation A.1. This yields $C_{\text{L-S}} = 0.211$ and $C_{\text{Std}} = 0.110$.

APPENDIX B: REVERSIBILITY OF THE PATH UPDATE

The path $A_i = (A_{i,1}, \ldots, A_{i,N_i})$ takes values in the finite set $\mathcal{P}_i$. Conditional on A_i, the vector T_i takes values on the simplex

$$\mathcal{X}_{N_i} \triangleq \left\{ T_i \in \mathbb{R}^{N_i} : T_{i,j} > 0, \sum_{j=1}^{N_i} T_{i,j} = t_i^f - t_i^s \right\},$$

where $t_i^f - t_i^s$ is the known total travel time of trip i. For the reference measure on $\mathcal{X}_{N_i}$ we use $(N_i - 1)$-dimensional Lebesgue measure on the first $N_i - 1$ elements of the vector. Then

$$(A_i, T_i) \in \mathcal{C} \triangleq \bigcup_{A \in \mathcal{P}_i} \{A\} \times \mathcal{X}_{\text{len}(A)}$$

where $\text{len}(A)$ is the number of arcs in $A \in \mathcal{P}_i$. We claim that the move for (A_i, T_i) is reversible with respect to the conditional posterior density of (A_i, T_i) given the GPS data $G = \{G_{i'}\}_{i'=1}^{I}$, the parameters, and the paths and travel times $A_{[-i]}, T_{[-i]}$ for all other trips:

$$\begin{aligned} \nu(A_i, T_i) &\triangleq \pi\left(A_i, T_i \;\middle|\; G, A_{[-i]}, T_{[-i]}, \left\{\mu_j, \sigma_j^2\right\}_{j=1}^{J}, \zeta^2\right) \\ &\propto f_i\left(A_i, T_i, G_i \;\middle|\; \left\{\mu_j, \sigma_j^2\right\}_{j=1}^{J}, \zeta^2\right). \end{aligned} \tag{B.1}$$

Since the dimension of the unknown vector T_i depends on A_i, one can consider this to be a case of model uncertainty as in Green (1995), where the model index k corresponds to the value of $A_i \in \mathcal{P}_i$. Our context, which has an uncertain route for each trip, is slightly different from the context of Green (1995), which has a single uncertain model index k and corresponding parameter vector $\theta^{(k)}$. However, their argument can still be used to show reversibility of a move for (A_i, T_i) conditional on $A_{[-i]}, T_{[-i]}$ and the parameters $\{\mu_j, \sigma_j^2\}_{j=1}^J, \zeta^2$.

Conditional on $A_i^{(1)}$ and $A_i^{(2)}$, we show that our move from $T_i^{(1)} \in \mathcal{X}_{\text{len}\left(A_i^{(1)}\right)}$ to $T_i^{(2)} \in \mathcal{X}_{\text{len}\left(A_i^{(2)}\right)}$ satisfies the dimension-matching condition of Green (1995), Section 3.3. We need a bijection between an augmented vector $\left(T_i^{(1)}, u^{(1)}\right)$ and the corresponding augmented vector $\left(T_i^{(2)}, u^{(2)}\right)$, for some $u^{(1)}$ and $u^{(2)}$. Take $u^{(1)} \triangleq \left(T_i^{(2)}(p_1), \ldots, T_i^{(2)}(p_n)\right)$ and $u^{(2)} \triangleq \left(T_i^{(1)}(c_1), \ldots, T_i^{(1)}(c_m)\right)$ and recall that $u^{(1)}$ is drawn independently of $T_i^{(1)}$. Define the bijection $h\left(T_i^{(1)}, u^{(1)}\right) \triangleq \left(T_i^{(2)}, u^{(2)}\right)$ that simply rearranges the elements of the vector $\left(T_i^{(1)}, u^{(1)}\right)$. The absolute value of the Jacobian of such a transformation is one (since that of the identity transform is one, and since rearranging the elements corresponds to permuting the rows of the Jacobian, which only changes the sign of the determinant). Although for notational convenience we have included the redundant final elements of the vectors $u^{(1)}$, $u^{(2)}$, $T_i^{(1)}$, and $T_i^{(2)}$, the dimension-matching is on the non-redundant elements of the vectors; in the notation of Green (1995), $n_1 = N_i^{(1)} - 1$, $m_1 = n - 1$, $n_2 = N_i^{(2)} - 1$, and $m_2 = m - 1$.

For a dimension-matching move, the acceptance probability that ensures reversibility with respect to a density $\nu(A_i, T_i)$ is given by Equation 7 of Green (1995). It is equal to the absolute value of the Jacobian, times $\frac{\nu\left(A_i^{(2)}, T_i^{(2)}\right)}{\nu\left(A_i^{(1)}, T_i^{(1)}\right)}$, times the ratio of the proposal density of the reverse move relative to that of the proposed move. The probability of proposing a move to $A_i^{(2)}$, given that the current state is $\left(A_i^{(1)}, T_i^{(1)}\right)$, is $\frac{1}{N_i^{(1)} \min\{a^{(1)}, K\}}$ divided by the number of paths of length $\leq K$ from d' to d''. The probability of attempting the reverse move is $\frac{1}{N_i^{(2)} \min\{a^{(2)}, K\}}$ divided by the number of paths of length $\leq K$ from d' to d''. We propose $T_i^{(2)}$ by drawing the subvector $T_i^{(2)}(j) : j \in \{p_1, \ldots, p_n\}$

according to the density

$$\frac{1}{S_i^{n-1}} \operatorname{Dir}\left(\frac{T_i^{(2)}(p_1)}{S_i}, \ldots, \frac{T_i^{(2)}(p_n)}{S_i}; \alpha\theta(p_1), \ldots, \alpha\theta(p_n)\right)$$

on the simplex $\{T_i \in \mathbb{R}^n : T_{i,j} > 0, \sum_{j=1}^n T_{i,j} = S_i\}$, with respect to $(n-1)$-dimensional Lebesgue measure. The reverse move, from $T_i^{(2)} \in \mathcal{X}_{\text{len}\left(A_i^{(2)}\right)}$ to $T_i^{(1)} \in \mathcal{X}_{\text{len}\left(A_i^{(1)}\right)}$, proposes $T_i^{(1)}$ by drawing the subvector $T_i^{(1)}(j) : j \in \{c_1, \ldots, c_m\}$ according to the density

$$\frac{1}{S_i^{m-1}} \operatorname{Dir}\left(\frac{T_i^{(1)}(c_1)}{S_i}, \ldots, \frac{T_i^{(1)}(c_m)}{S_i}; \alpha\theta(c_1), \ldots, \alpha\theta(c_m)\right).$$

Plugging these quantities into Equation 7 of Green (1995) and using our Equation B.1 gives the acceptance probability in our Equation 3.1.

APPENDIX C: 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 4.1, it is critical whether the GPS readings are sampled by distance or by time. Sampling-by-distance could mean recording a GPS point every 100 m, and sampling-by-time could mean recording a GPS point every 30 s, for example. As discussed in Sections 1 and 4.1, 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, speeds are typically recorded by loop detectors at fixed locations on the road, which means that sampling is done by distance. In this context, 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 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

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

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 C.1. Again, the inequality is strict unless $v_i(\cdot)$ is constant.

REFERENCES

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

Rakha, H. and Zhang, W. (2005). 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.

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

Stein, W. E. and Dattero, R. (1985). Sampling bias and the inspection paradox. *Mathematics Magazine* **58** 96–99.

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

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

School of Operations Research
and Information Engineering
Cornell University
Rhodes Hall
Ithaca, NY 14853
USA
E-mail: bsw62@cornell.edu
woodard@cornell.edu
sgh9@cornell.edu

Department of Statistical Science
Cornell University
1196 Comstock Hall
Ithaca, NY 14853
USA
E-mail: matteson@cornell.edu