Biostatistics (2018), **0**, 0, *pp.* 1–15
doi:10.1093/biostatistics/2018-08-31˙SupplementaryMaterials

Inferring Mobility Measures from GPS Traces with Missing Data: Supplementary Materials

IAN BARNETT*

Department of Biostatistics, University of Pennsylvania, 423 Guardian Drive, Philadelphia, PA 19104

JUKKA-PEKKA ONNELA

Department of Biostatistics, Harvard University, 677 Huntington Avenue, Boston, MA 02115

ibarnett@pennmedicine.upenn.edu

*To whom correspondence should be addressed.

1. MOBILITY MEASURES

Mobility measures are calculated on a daily basis (midnight-to-midnight). We consider the following mobility measures, some selected from the same measures used by **?**, in our analysis:

1.1 *Distance travelled*

Denoted **DistTravelled**. The DistTravelled variable for a given day is the sum of the flight lengths (meters) over all flights that occurred on that day.

1.2 *Radius of gyration*

Denoted **RoG**. Let D be the set of all times in the 24-hour period defining a particular day. Let

$$L(D) = \sum_{j=1}^{n} I_{\{t_j \in D\}} I_{\{e_j = flight \text{ or } e_j = pause\}}(t_{j+1} - t_j)$$

be the total time in a flight or a pause during day D. Let

$$C(D) = \frac{1}{L(D)} \sum_{j=1}^{n} I_{\{t_j \in D\}}(t_{j+1} - t_j) \left(I_{\{e_j = flight\}}(G(t_j) + G(t_{j+1}))/2 + I_{\{e_j = pause\}} G(t_j) \right)$$

be the average (or center) spatial location on day D. The RoG variable (meters) for that day is

$$\sqrt{\frac{\sum_{j=1}^{n} I_{\{t_j \in D\}}(t_{j+1} - t_j) \left(I_{\{e_j = flight\}} ||(G(t_j) + G(t_{j+1}))/2 - C(D)||_2^2 + I_{\{e_j = pause\}} ||G(t_j) - C(D)||_2^2 \right)}{L(D)}}$$

1.3 *Maximum diameter*

Denoted **MaxDiam**. The MaxDiam variable for a day is the maximum pairwise distance (meters) between any two pause locations that both occur on that day.

1.4 *Significant locations visited*

Denoted **SigLocsVisited**. This represents the number of significant locations a person visits in a day. In order to determine the set of all significant locations for a person, K-means is performed on the set of all pause locations with a minimum duration of 10 minutes. Longer pauses are given additional weight: suppose e_j is a pause, then the point $G(t_j)$ is represented by $\lfloor (t_{j+1} - t_j)/600 \rfloor$ replicated points in the analysis. The clusters number, K, is maximized under the condition that no two cluster centers are within 400 meters of one another. We use the Hartigan-Wong algorithm (**?**) for K-means which initializes starting locations randomly so that the results are stochastic. After K-means is completed, any pauses within 200 meters of a cluster center are assumed to belong to that cluster. The SigLocsVisited variable is the number of significant locations that a person is within 200 meters of at any point during the day. The choice of distance parameters, 200 meters and 400 meters, can be adjusted to the specific scientific question under investigation.

1.5 *Time spent at home*

Denoted **Hometime**. The significant location with the largest total amount of time during the night hours between 9PM and 6AM over the course of the study period is assumed to be the location of the person's home. The Hometime variable is the amount of time (minutes) during the day spent within a 200 meter radius of home.

1.6 *Maximum distance from home*

Denoted **MaxHomeDist**. After determining the location of home, the distance between home and every other pause location over the course of the day is calculated. The MaxHomeDist variable (meters) is the maximum of these distances.

1.7 *Average flight length*

Denoted **AvgFlightLen**. The AvgFlightLen variable for a given day is the average length (meters) of all flights that occur on that day.

1.8 *Standard deviation of flight length*

Denoted **StdFlightLen**. The StdFlightLen variable for a given day is the standard deviation of flight length (meters) over all flights that occur on that day.

1.9 *Average flight duration*

Denoted **AvgFlightDur**. The AvgFlightDur variable for a given day is the average duration (seconds) of all flights that occur on that day.

1.10 *Standard deviation of flight duration*

Denoted **StdFlightDur**. The StdFlightDur variable for a given day is the standard deviation of flight duration (seconds) over all flights that occur on that day.

1.11 *Fraction of time stationary*

Denoted **FracPause**. For any given day, D, the FracPause variable is the fraction of time, or probability of being paused, relative to the amount of time spent moving, or in flight, and defined by:

$$\frac{\sum_{j=1}^{n} I_{\{t_j \in D\}} I_{\{e_j = pause\}} (t_{j+1} - t_j)}{\sum_{j=1}^{n} I_{\{t_j \in D\}} I_{\{e_j = flight \text{ or } e_j = pause\}} (t_{j+1} - t_j)}$$

1.12 *SigLocEntropy*

Denoted **SigLocEntropy**. Suppose there are K significant locations. Let ts_k be the time spent on a given day within a 200 meter radius of the kth significant location, and let $p_k = ts_k / \sum_{k=1}^{K} ts_k$. Then the SigLocEntropy variable for that day is given by $-\sum_{k=1}^{K} p_k \log p_k$. Large values indicate a person spreading their time out across many different locations fairly evenly for that day, whereas small values indicate a concentration at few significant locations.

1.13 *MinsMissing*

Denoted **MinsMissing**. The MinsMissing variable for a given day is the number of minutes of missing data for that day. Note that there are some missing intervals which are inferred to be pauses if both before and after the missing interval a pause occurs at the same location (within a 50 meter radius). These inferred pauses are not included in the MinsMissing variable.

1.14 *Circadian routine*

Denoted **CircdnRtn**. For two different days D_i and D_j, let $d(D_i, D_j)$ be the fraction of time that the person is in the same place (within a 200 meter radius) at the same time of day. This is approximated by taking the location of a person at 12:30AM on day D_i, and seeing if the person is ever within 200 meters of that location between the hours of 12:00AM and 1:00AM on day D_j. This is repeated for each of the 24 one-hour periods of the day and $d(D_i, D_j)$ is reported as the fraction of those one-hour periods that the person's location was synced (up to the 200 meter buffer). Supposing there are m days of GPS data collected by the person, for a given day, D, the CircdnRtn variable takes a value of:

$$\frac{1}{m-1} \sum_{D_j \neq D} d(D, D_j)$$

Low values of CircdnRtn indicate a break from routine on day D, whereas high values indicate that the person followed their daily routine on day D.

1.15 *Weekend/Weekday circadian routine*

Denoted **WkEndDayRtn**. This is computed the same way as CircdnRtn, except weekdays and weekend days are stratified and treated separately. Therefore, to calculate the WkEndDayRtn index on a weekday, all of the weekend days are omitted from the analysis completely. Similarly, to calculate the WkEndDayRtn index on a weekend day, all of the weekdays are omitted from the analysis completely.

2. EXPECTED GAP BETWEEN A MOBILITY TRACE AND ITS SURROGATES: DERIVATIONS

First, we let $(\bar{\Delta}^x, \bar{\Delta}^y) = \frac{1}{n}\sum_{i=1}^{n}(\Delta_i^x, \Delta_i^y)$ and $\bar{D}_x = \frac{1}{n}\sum_{i=1}^{n}(\Delta_i^x - \Delta_{(i)}^x)$ and $\bar{D}_y = \frac{1}{n}\sum_{i=1}^{n}(\Delta_i^y - \Delta_{(i)}^y)$.

2.1 *Continuity Assumptions*

Because $F^{(f)}$, $F^{(p)}$, and Ψ are unknown, we must rely on empirical estimates of these functions from the observed data. With the goal of resampling from observed events, or hot-deck imputation, we must make several continuity-like assumptions on $F^{(f)}$, $F^{(p)}$, and Ψ in order to enable local resampling from observed events to impute missing events. We assume Ψ is continuous, and that $\forall \boldsymbol{z}$ and for every $\epsilon > 0$, there exists $\delta > 0$ such that

$$\left\| F^{(f)}(\cdot|\boldsymbol{Z} = \boldsymbol{z} + \boldsymbol{\delta}) - F^{(f)}(\cdot|\boldsymbol{Z} = \boldsymbol{z}) \right\|_\infty \leqslant \epsilon \tag{2.1}$$

$$\left\| F^{(p)}(\cdot|\boldsymbol{Z} = \boldsymbol{z} + \boldsymbol{\delta}) - F^{(p)}(\cdot|\boldsymbol{Z} = \boldsymbol{z}) \right\|_\infty \leqslant \epsilon \tag{2.2}$$

for all $0 \leqslant \delta_x < \delta$, $0 \leqslant \delta_y < \delta$, and $0 \leqslant \delta_t < \delta$, where $\boldsymbol{\delta} = (\delta_x, \delta_y, \delta_t)$. This condition ensures that the distribution of flights and pauses are similar locally with respect to location and time.

Let $E_f = \{i \in \{1, ..., n\} : B_i = 1\}$ be the indices of flights with $n_f = |E_f|$. Suppose we wish to resample from the observed set of flights to impute a trajectory at some new time and location $\boldsymbol{z}_{\text{new}} = (x_{\text{new}}, y_{\text{new}}, t_{\text{new}})$. The empirical distribution, giving $w_k(\cdot)$ weight to the kth flight, is

$$\hat{F}^{(f)}\left\{\boldsymbol{\Delta} = (\Delta^{(x)}, \Delta^{(y)}, \Delta^{(t)}) | \boldsymbol{Z} = \boldsymbol{z}_{\text{new}}\right\} = \frac{\sum_{k \in E_F} w_k(\boldsymbol{z}_{\text{new}}) I_{\{\Delta_k^{(x)} < \Delta^{(x)}, \Delta_k^{(y)} < \Delta^{(y)}, \Delta_k^{(t)} < \Delta^{(t)}\}}}{\sum_{k \in E_F} w_k(\boldsymbol{z}_{\text{new}})}.$$

In addition to the continuity assumption of equation (2.1), $\epsilon = \epsilon(\delta)$ where $\epsilon(\cdot)$ is non-decreasing and $\epsilon(0) = 0$, we consider the first two moments of the asymptotic distribution of the empirical distribution function:

$$\begin{aligned}
&\left| E\left[\sqrt{n_f}\left(F^{(f)}(\boldsymbol{\Delta}|\boldsymbol{Z} = \boldsymbol{z}_{new}) - \hat{F}^{(f)}(\boldsymbol{\Delta}|\boldsymbol{Z} = \boldsymbol{z}_{new})\right)\right]\right| \\
&= \left| \sqrt{n_f} \frac{\sum_{k \in E_F} w_k(\boldsymbol{z}_{new}) \left(F^{(f)}(\boldsymbol{\Delta}|\boldsymbol{Z} = \boldsymbol{z}_{new}) - F^{(f)}(\boldsymbol{\Delta}|\boldsymbol{Z} = \boldsymbol{z}_k)\right)}{\sum_{k \in E_F} w_k(\boldsymbol{z}_{new})} \right| \\
&\leqslant \sqrt{n_f} \frac{\sum_{k \in E_F} w_k(\boldsymbol{z}_{new}) \epsilon\left(\|\boldsymbol{z}_{new} - \boldsymbol{z}_k\|_\infty\right)}{\sum_{k \in E_F} w_k(\boldsymbol{z}_{new})}
\end{aligned} \tag{2.3}$$

$$\begin{aligned}
&\text{Var}\left[\sqrt{n_f}\left(F^{(f)}(\boldsymbol{\Delta}|\boldsymbol{Z} = \boldsymbol{z}_{new}) - \hat{F}^{(f)}(\boldsymbol{\Delta}|\boldsymbol{Z} = \boldsymbol{z}_{new})\right)\right] \\
&= n_f \frac{\sum_{k \in E_F} w_k^2(\boldsymbol{z}_{new}) F^{(f)}(\boldsymbol{\Delta}|\boldsymbol{Z} = \boldsymbol{z}_k) \left(1 - F^{(f)}(\boldsymbol{\Delta}|\boldsymbol{Z} = \boldsymbol{z}_k)\right)}{\left(\sum_{k \in E_F} w_k(\boldsymbol{z}_{new})\right)^2}.
\end{aligned} \tag{2.4}$$

If the distribution function of flights is independent of time and location, then $F^{(f)}(\cdot|\boldsymbol{Z}) = F^{(f)}(\cdot)$, and so resampling can be performed as in the case of an independent and identically distributed sample by letting $w_i(\cdot) = 1/n_f$ for each event. In this case $\epsilon(\delta) = 0 \ \ \forall\ \delta > 0$, so the expectation in Equation (2.3) reduces to 0 and the variance in Equation (2.4) simplifies to the binomial variance, $F^{(f)}(\boldsymbol{\Delta})(1 - F^{(f)}(\boldsymbol{\Delta}))$.

However, it is unlikely that the distribution of flights or pauses is identically distributed across all times and locations. In this case, the empirical distribution function will be biased, with a bound for the magnitude of this bias specified in Equation (2.3). The bound for this bias is minimized by giving higher weight to events that are closer in time and location to the new event

$\boldsymbol{z}_{\text{new}}$. This can be achieved by specifying $w_k(\boldsymbol{z}_{new})$ so that it is inversely related to $\|\boldsymbol{z}_{\text{new}} - \boldsymbol{z}_k\|_\infty$. In this extreme, letting $w_k(\boldsymbol{z}_{\text{new}}) = I_{\{\boldsymbol{z}_k = \boldsymbol{z}_{\text{new}}\}}$ would eliminate the bias completely, but is impractical is it as it wouild assign a weight of 0 every other observed event. To minimize the variance in Equation (2.4), the weights are spread out equally across all events $w_k(\boldsymbol{z}_{\text{new}}) = 1/n_f$, but this will lead to an inflated bias. Instead, a balance must be achieved when selecting weights so that both the bias and variance of the empirical distribution function are kept low. This can be done by selecting weights from a unimodal function centered on $\boldsymbol{z}_{\text{new}}$. To this effort, we choose a t-distribution function with ν degrees of freedom in order to allow for both spread and kurtosis to be controlled as tuning parameters. The same principles in selecting weights can be applied to the empirical approximations of $F^{(p)}(\cdot)$ and $\Psi(\cdot)$. The empirical approximation to $\Psi(\cdot)$ is

$$\hat{\Psi}(\boldsymbol{z}_{\text{new}}) = \frac{\sum_{j=2}^{n} B_{j-1} B_j w_j(\boldsymbol{z}_{\text{new}})}{\sum_{j=2}^{n} B_{j-1} w_j(\boldsymbol{z}_{\text{new}})}.$$

In order to improve resampling further, in addition to the continuity assumptions of Equations (2.1) and (2.2) we also consider several potentially realistic assumptions on human mobility:

i.*Temporally local* (TL) weights: Events close in time tend to have similar mobility patterns.

$$\begin{aligned} F^{(f)}(\cdot|\boldsymbol{Z}) &= F^{(f)}(\cdot|T) \\ F^{(p)}(\cdot|\boldsymbol{Z}) &= F^{(p)}(\cdot|T) \\ \Psi(\boldsymbol{Z}) &= \Psi(x, y, T) \quad \forall x, y \in \mathbb{R}. \end{aligned}$$

Resampling weights corresponding to this assumption are:

$$w_j(\boldsymbol{z}_{new}) = \psi_\nu \left(c \cdot (t_{new} - t_j) \right),$$

where $\psi_\nu(\cdot)$ is the t-distribution density function with ν degrees of freedom and c is a scaling constant.

ii. *Geographically local* (GL) weights: Events close in space tend to have similar mobility pat-

terns.

$$\begin{aligned} F^{(f)}(\cdot|\boldsymbol{Z}) = &\quad F^{(f)}(\cdot|X,Y) \\ F^{(p)}(\cdot|\boldsymbol{Z}) = &\quad F^{(p)}(\cdot|X,Y) \\ \Psi(\boldsymbol{Z}) = &\quad \Psi(X,Y,t) \qquad \forall t \in \mathbb{R}. \end{aligned}$$

Resampling weights corresponding to this assumption are:

$$w_j(\boldsymbol{z}_{\text{new}}) = \psi_\nu \left(c \cdot \sqrt{(x_{\text{new}} - x_j)^2 + (y_{\text{new}} - y_j)^2} \right).$$

iii. *Geographically local with circadian routine* (GLC) weights: Events close in space and close in the time of day have similar mobility patterns. Considering time to be measured in hours:

$$\begin{aligned} F^{(f)}(\cdot|X,Y,T=t) = &\quad F^{(f)}(\cdot|X,Y,T=t+24k) \quad \forall k \in \mathbb{Z}, \forall t \in \mathbb{R} \\ F^{(p)}(\cdot|X,Y,T=t) = &\quad F^{(p)}(\cdot|X,Y,T=t+24k) \quad \forall k \in \mathbb{Z}, \forall t \in \mathbb{R} \\ \Psi(X,Y,T) = &\quad \Psi(X,Y,T+k\cdot 24 \text{ hours}) \quad \forall k \in \mathbb{Z}. \end{aligned}$$

Letting s represent 24 hours (in the units of time of t) and letting c_1 and c_2 be the scaling constants, the resampling weights corresponding to this assumption are:

$$\begin{aligned} w_j(\boldsymbol{z}_{new}) = &\psi_\nu \left\{ c_1 \cdot \sqrt{(x_{new} - x_j)^2 + (y_{new} - y_j)^2} \right\} \\ &\cdot \psi_\nu \left[c_2 \cdot \min \left\{ |t_{new} - t_j| \pmod s, s - |t_{new} - t_j| \pmod s \right\} \right]. \end{aligned}$$

There are reasonable arguments for any of the TL, GL, or GLC assumptions. Human mobility patterns may be a function of location. For example, people may be more stationary when at home than when outside the home. In this case the GL assumption would be able to ensure low values of $\Psi(\boldsymbol{z}_i)$ when (x_i, y_i) are the coordinates of home a person's home. The GLC assumption adds to this a circadian component which would be better at recovering information about, say, a regular commute. The TL assumption would do well to model bursty human movement where flights tend to occur in bunches over time. This assumption captures what is likely the most general and robust pattern of human behavior.

2.2 *Expected gap between $\boldsymbol{L}(t)$ and $\hat{\boldsymbol{L}}(t)$*

$$
\begin{aligned}
E\left[\left\|\boldsymbol{L}(t)-\hat{\boldsymbol{L}}(t)\right\|^2\right] &= E\left[\left\|\begin{bmatrix} \sum_{i=1}^t(\Delta_i^x-\Delta_{(i)}^x-\bar{D}_x) \\ \sum_{i=1}^t(\Delta_i^y-\Delta_{(i)}^y-\bar{D}_y)\end{bmatrix}\right\|^2\right] \\
&= \textstyle\sum_{l\in\{x,y\}} E\left[\left(\sum_{i=1}^t(\Delta_i^l-\Delta_{(i)}^l-\bar{D}_l)\right)^2\right] \\
&= \textstyle\sum_{l\in\{x,y\}} E\left[\sum_{i=1}^t(\Delta_i^l-\Delta_{(i)}^l-\bar{D}_l)^2+2\sum_{i<j}^t(\Delta_i^l-\Delta_{(i)}^l-\bar{D}_l)(\Delta_j^l-\Delta_{(j)}^l-\bar{D}_l)\right] \\
&= \textstyle\sum_{l\in\{x,y\}}\Big\{\sum_{i=1}^t\left(E\left[(\Delta_i^l-\Delta_{(i)}^l)^2\right]-2E\left[\bar{D}_l(\Delta_i^l-\Delta_{(i)}^l)\right]+E\left[\bar{D}_l^2\right]\right) \\
&\quad \textstyle+2\sum_{i<j}^t E\left[(\Delta_i^l-\Delta_{(i)}^l-\bar{D}_l)(\Delta_j^l-\Delta_{(j)}^l-\bar{D}_l)\right]\Big\} \\
&= \textstyle\sum_{l\in\{x,y\}}\left\{\sum_{i=1}^t\left(2\sigma_l^2-\frac{4}{n}\sigma_l^2+\frac{2}{n}\sigma_l^2\right)+2\sum_{i<j}^t E\left[(\Delta_i^l-\Delta_{(i)}^l-\bar{D}_l)(\Delta_j^l-\Delta_{(j)}^l-\bar{D}_l)\right]\right\} \\
&= \textstyle\sum_{l\in\{x,y\}}\Big\{\sum_{i=1}^t 2(1-\frac{1}{n})\sigma_l^2+2\sum_{i<j}^t E\left[\Delta_i^l-\Delta_{(i)}^l\right]E\left[\Delta_j^l-\Delta_{(j)}^l\right] \\
&\quad \textstyle-E\left[\bar{D}_l(\Delta_i^l-\Delta_{(i)}^l)\right]-E\left[\bar{D}_l(\Delta_j^l-\Delta_{(j)}^l)\right]+E\left[\bar{D}_l^2\right]\Big\} \\
&= \textstyle\sum_{l\in\{x,y\}}\left\{\sum_{i=1}^t 2(1-\frac{1}{n})\sigma_l^2+2\sum_{i<j}^t -\frac{2}{n}\sigma_l^2-\frac{2}{n}\sigma_l^2+\frac{2}{n}\sigma_l^2\right\} \\
&= \textstyle\sum_{l\in\{x,y\}}\left\{2t(1-\frac{1}{n})\sigma_l^2-t(t-1)\frac{2}{n}\sigma_l^2\right\} \\
&= \textstyle\sum_{l\in\{x,y\}} 2t\sigma_l^2(1-\frac{t}{n}) \\
&= 2t(1-\tfrac{t}{n})(\sigma_x^2+\sigma_y^2)
\end{aligned}
$$

2.3 *Continuity assumptions and the expected gap between $\boldsymbol{L}(t)$ and $\tilde{\boldsymbol{L}}(t)$*

$$
\begin{aligned}
E\left[\left\|\boldsymbol{L}(t)-\tilde{\boldsymbol{L}}(t)\right\|^2\right] &= E\left[\left\|\begin{bmatrix}\sum_{i=1}^t(\Delta_i^x-\bar{\Delta}_x)\\ \sum_{i=1}^t(\Delta_i^y-\bar{\Delta}_y)\end{bmatrix}\right\|^2\right]\\
&= \textstyle\sum_{l\in\{x,y\}} E\left[\left(\sum_{i=1}^t(\Delta_i^l-\bar{\Delta}_l)\right)^2\right]\\
&= \textstyle\sum_{l\in\{x,y\}}\left\{\sum_{i=1}^t E\left[(\Delta_i^l-\bar{\Delta}_l)^2\right]+2\sum_{i<j}^t E\left[(\Delta_i^l-\bar{\Delta}_l)(\Delta_j^l-\bar{\Delta}_l)\right]\right\}\\
&= \textstyle\sum_{l\in\{x,y\}}\Big\{\sum_{i=1}^t\left[E[(\Delta_i^l)^2]-2E[\Delta_i^l\bar{\Delta}_l]+E[\bar{\Delta}_l^2]\right]\\
&\quad \textstyle+2\sum_{i<j}^t\left[E[\Delta_i^l]E[\Delta_j^l]-E[\Delta_i^l\bar{\Delta}_l]-E[\Delta_j^l\bar{\Delta}_l]+E[\bar{\Delta}_l^2]\right]\Big\}\\
&= \textstyle\sum_{l\in\{x,y\}}\Big\{\sum_{i=1}^t[\sigma_l^2+\mu_l(i)^2-\frac{2}{n}\left(\sigma_l^2+\sum_{j=1}^n\mu_l(i)\mu_l(j)\right)\\
&\quad \textstyle+\frac{1}{n^2}\left(\sum_{k=1}^n(\sigma_l^2+\mu_l(k)^2)+2\sum_{k<j}^n\mu_l(k)\mu_l(j)\right)]\\
&\quad \textstyle+2\sum_{i<j}^t[\mu_l(i)\mu_l(j)-\frac{1}{n}\left(2\sigma_l^2+\sum_{k=1}^n\mu_l(k)(\mu_l(i)+\mu_l(j))\right)\\
&\quad \textstyle+\frac{1}{n^2}\left(\sum_{k=1}^n(\sigma_l^2+\mu_l(k)^2)+2\sum_{k<s}^n\mu_l(k)\mu_l(s)\right)]\Big\}\\
&= \textstyle\sum_{l\in\{x,y\}}\Big\{t\left(1-\frac{t}{n}\right)\sigma_l^2+\sum_i^t\mu_l(i)^2+\frac{t^2}{n^2}\sum_{i=1}^n\sum_{j=1}^n\mu_l(i)\mu_l(j)\\
&\quad \textstyle-\frac{2}{n}\sum_{i=1}^t\sum_{j=1}^n\mu_l(i)\mu_l(j)\\
&\quad \textstyle+2\sum_{i<j}^t\left[\mu_l(i)\mu_l(j)-\frac{1}{n}\sum_{k=1}^n\mu_l(k)(\mu_l(i)+\mu_l(j))\right]\Big\}\\
&= \textstyle t\left(1-\frac{t}{n}\right)(\sigma_x^2+\sigma_y^2)+\sum_{l\in\{x,y\}}\Big[\sum_i^t\mu_l(i)^2+\frac{t^2}{n^2}\sum_{i=1}^n\sum_{j=1}^n\mu_l(i)\mu_l(j)\\
&\quad \textstyle-\frac{2}{n}\sum_{i=1}^t\sum_{j=1}^n\mu_l(i)\mu_l(j)\\
&\quad \textstyle+2\sum_{i<j}^t\left[\mu_l(i)\mu_l(j)-\frac{1}{n}\sum_{k=1}^n\mu_l(k)(\mu_l(i)+\mu_l(j))\right]\Big]\\
&= t\left(1-\tfrac{t}{n}\right)(\sigma_x^2+\sigma_y^2)+M(t)
\end{aligned}
$$

where

$$
\begin{aligned}
M(t) = \textstyle\sum_{l\in\{x,y\}}\Big\{&\textstyle\sum_i^t\mu_l(i)^2+\frac{t^2}{T^2}\sum_{i=1}^n\sum_{j=1}^n\mu_l(i)\mu_l(j)-\frac{2}{n}\sum_{i=1}^t\sum_{j=1}^n\mu_l(i)\mu_l(j)\\
&\textstyle+2\sum_{i<j}^t\left[\mu_l(i)\mu_l(j)-\frac{1}{n}\sum_{k=1}^n\mu_l(k)(\mu_l(i)+\mu_l(j))\right]\Big\}
\end{aligned}
$$

is a function of the $\mu(i)$.

3. R PACKAGE: *GPSmobility*

The GPSmobility R package implements the approaches presented in this paper. After installing the package, one can follow CodeDemonstration.R by using either the "ExampleFull"

or the "ExampleMissing" as the *filename* argument. The *ExampleFull.Rdata* file contains an example week of GPS trace, recorded near-continuously. The *ExampleMissing.Rdata* file contains the exact same example week of GPS trace, except recorded over 2 minute intervals with 10 minute off-periods of missingness in between.

Table S 1. **Mobility measures compared across different missing data imputation approaches for the trajectories.** GPS was collected continuously to establish the ground truth. For the missing data imputations, a Cauchy kernel was used with scale factor denoted by the number following the period. Larger scale factors give increased weight on nearby observations during resampling. For the TL, GL, and GLC approaches, the margin of error represents the standard deviation over 100 repeated simulations.

Measures	**TL.1**	**TL.10**	**TL.20**	**GL.1**	**GL.10**	**GL.20**	**GLC.1**	**GLC.10**	**GLC.20**	**LI**	**Truth**
Hometime	831.5 ±2.3	832.3 ±2.4	833.4 ±2.2	830.3 ±2.2	830.5 ±2.8	829.8 ±1.9	829.1 ±2.1	832.1 ±2.2	831.3 ±2.5	826.7	882.8
DistTravelled	22184 ±969.7	22446 ±843.5	22569 ±811.6	18801 ±466.3	18801 ±337.5	18779 ±369.4	21791 ±969.9	22380 ±712.1	22444 ±645.6	17236	19344
RoG	2787.3 ±2.3	2791.3 ±2.6	2791.2 ±1.9	2783.0 ±1.6	2783.0 ±1.9	2783.3 ±2.5	2785.6 ±1.3	2787.0 ±1.5	2787.5 ±1.8	2779.4	2781.3
MaxDiam	6717 ±169	6745 ±129	6727 ±98	6494 ±44	6483 ±8	6496 ±34	6516 ±55	6517 ±55	6562 ±94	6479	6467
MaxHomeDist	6372 ±165	6410 ±123	6379 ±93	6160 ±49	6147 ±16	6153 ±39	6144 ±30	6152 ±5	6163 ±24	6149	6129
SigLocsVisited	2.96 ±0.73	3.20 ±0.58	3.20 ±0.71	3.16 ±0.69	3.00 ±0.76	2.96 ±0.79	3.28 ±0.61	3.12 ±0.60	3.20 ±0.65	2	3
AvgFlightLen	172.7 ±10.7	160.2 ±7.6	158.6 ±7.4	200.2 ±23.2	193.2 ±19.2	191.7 ±18.1	129.9 ±13.6	122.8 ±6.1	127.1 ±7.6	478.8	251.2
StdFlightLen	152.9 ±30.8	125.8 ±10.1	123.2 ±5.5	213.4 ±51.5	205.8 ±36.3	202.7 ±43.5	151.0 ±30.0	134.2 ±8.4	137.1 ±9.0	639.6	223.3
AvgFlightDur	79.0 ±9.3	69.4 ±5.8	68.8 ±5.6	119.0 ±17.9	115.2 ±13.4	113.5 ±13.7	65.4 ±10.5	57.2 ±4.1	60.0 ±5.1	340.6	77.0
StdFlightDur	131.7 ±17.0	115.3 ±9.0	113.5 ±10.2	170.3 ±22.0	168.7 ±14.8	166.7 ±14.4	103.7 ±18.2	85.0 ±10.9	91.7 ±13.1	289.8	55.2
FracPause	0.88 ±0.01	0.89 ±0.01	0.89 ±0.01	0.87 ±0.01	0.87 ±0.01	0.87 ±0.01	0.87 ±0.01	0.88 ±0.01	0.88 ±0.01	0.86	0.93
SigLocEntropy	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63	0.63
MinsMissing	1243	1243	1243	1243	1243	1243	1243	1243	1243	1243	92
CircdnRtn	0.64 ±0.02	0.63 ±0.01	0.63 ±0.02	0.67 ±0.01	0.67 ±0.01	0.67 ±0.01	0.65 ±0.02	0.66 ±0.01	0.66 ±0.02	0.69	0.66
WkEndDayRtn	0.76 ±0.02	0.76 ±0.01	0.76 ±0.01	0.78 ±0.01	0.77 ±0.01	0.78 ±0.01	0.76 ±0.02	0.76 ±0.01	0.77 ±0.01	0.81	0.79

[*Received XXXXX; revised XXXXX; accepted for publication XXXXX*]

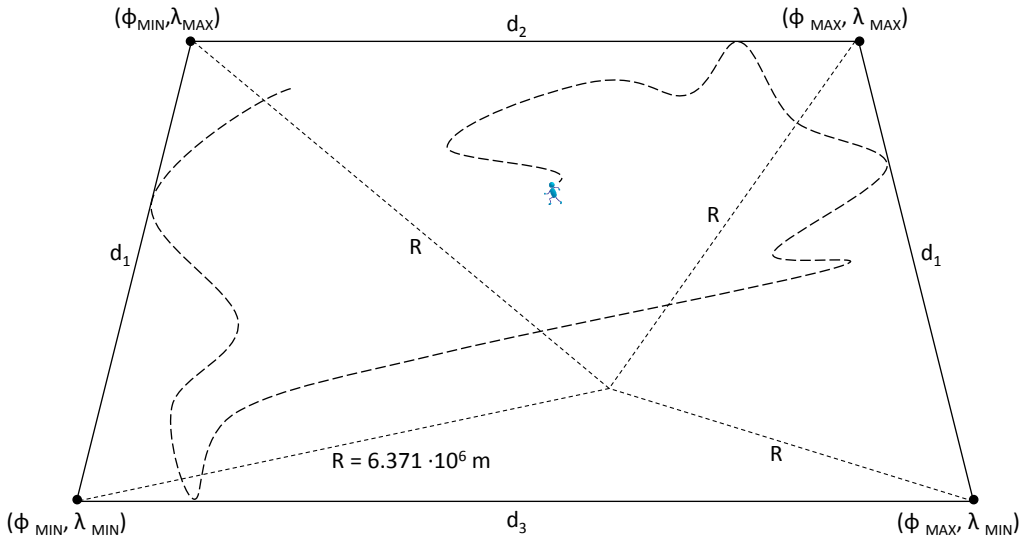

Figure S 1. **Schematic of longitude-latitude projection to X-Y plane.** The isosceles trapezoid contains the projection of the mobility trace for a particular individual. In the northern hemisphere $d_2 < d_3$ while in the southern hemisphere $d_2 > d_3$. The long dashed curve represents a person's example mobility trace.

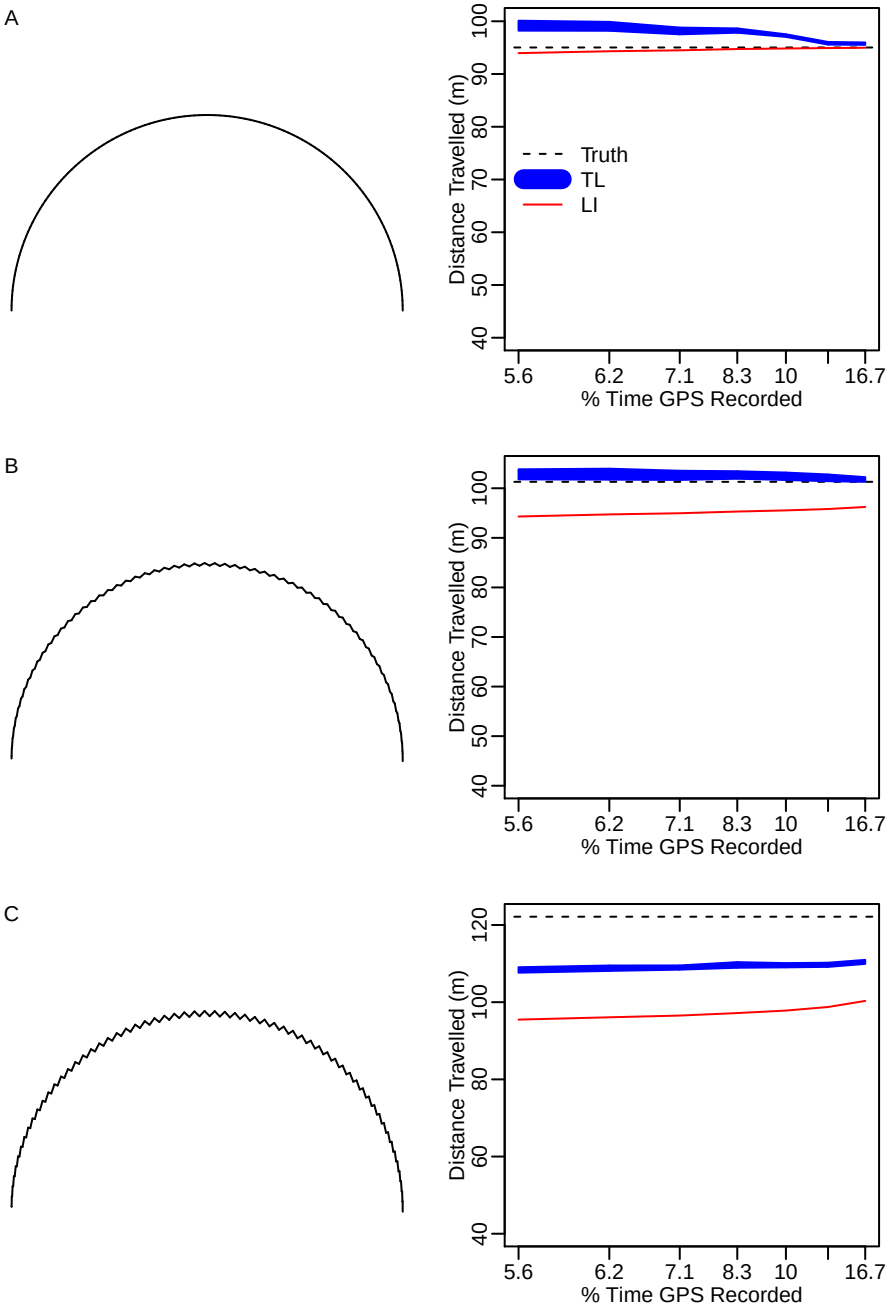

Figure S 2. **Bias in the estimation of mobility metrics as the amount of missingness changes.** On the left, three different trajectories are displayed. The right of each panel shows the estimates of distance travelled for various levels of missingness in the trajectory on the left. Missingness is generated by taking evenly spaced intervals of different sizes (for different levels of missingness) out of the semicircular trajectories. For TL, each level of missingness is repeated for 100 simulated trajectories to obtain the 95% confidence band. This is repeated for three different types of movement: a smooth trajectory (A), small jitters (B), and larger jitters (C). This demonstrate that both the direction and magnitude of a surrogate's bias in approximating the true trajectory can vary significantly depending on the true trajectory.