Process Imputation of Animal Telemetry Data

Josh Cullen

03 December 2019

Imputation Approaches

The snail kite telemetry data, as well as many telemetry datasets, are sampled irregularly in time. While this would not be a problem in some analyses (point process models, continuous-time movement models), the irregular sampling frequency precludes the consistent measurement of movement parameters (step length, turning angle, first passage time, directional persistence, etc) for discrete-time movement models. One possible solution would be to remove all observations that are not sampled at the time step of interest, although this can result in large gaps within the time series. Another simple solution is the use of linear interpolation (i.e. connecting the dots), which assumes a constant speed across consecutive observations and estimates these missing data on a linear path between true observations. This method (1) provides a biased view of the latent trajectory because it is not based on a mechanistic process of the movement dynamics, (2) the observed data are assumed to represent completely accurate positions of the latent trajectory, and (3) uncertainty in the trajectory is not accounted for at the observations or missing data. A more robust method that can account for these limitations is multiple imputation.

Multiple imputation is an iterative form of stochastic imputation, which generates a distribution from which inference can be made on the missing values. With respect to animal telemetry data, this includes the application of a process model (e.g. continuous-time correlated random walk using an Ornstein-Uhlenbeck process) as a representation of the latent trajectory to impute missing locations multiple times. This process accounts for the assumed underlying movement dynamics as well as measurement error of the location estimates. Multiple imputation depends on the ability to evaluate the complete-data posterior (i.e. model parameters given the observed and missing data), and the ability to sample missing data sets from the imputation distribution (i.e. the missing data given the observed data). This can be written as:

$$\begin{aligned}[\boldsymbol{\theta}|\mathbf{s}] &= \int [\boldsymbol{\theta}, \mathbf{s}_m|\mathbf{s}]\, d\mathbf{s}_m, \\ &= \int [\boldsymbol{\theta}|\mathbf{s}, \mathbf{s}_m]\, [\mathbf{s}_m|\mathbf{s}]\, d\mathbf{s}_m,\end{aligned}$$

where $\boldsymbol{\theta}$ is a vector of the model parameters, $\mathbf{s}$ is a vector of the data, and $\mathbf{s}_m$ is a vector representing the missing data. In this case, $[\boldsymbol{\theta}|\mathbf{s}]$ represents the desired posterior distribution, $[\mathbf{s}_m|\mathbf{s}]$ represents the imputation distribution, and $[\boldsymbol{\theta}|\mathbf{s}, \mathbf{s}_m]$ represents the complete-data posterior distribution. If K samples are drawn from the imputation distribution, the posterior expectation $\mathbb{E}(\boldsymbol{\theta}|\mathbf{s})$ can be approximated by averaging the imputed data $\mathbf{s}_m$ across all draws. By similarly averaging the variances across all K draws, these can provide a close approximation of the true posterior variance $\text{Var}(\boldsymbol{\theta}|\mathbf{s})$.

In terms of animal movement analyses, the telemetry data are defined as $\mathbf{s} \equiv (\mathbf{s}(t_1)', \ldots, \mathbf{s}(t_n)')'$, a $2n \times 1$ vector, for observation times $t_1, \ldots, t_n$. The true latent position process is represented as $\boldsymbol{\mu} \equiv (\boldsymbol{\mu}(t_1)', \ldots, \boldsymbol{\mu}(t_m)')'$, a $2m \times 1$ vector, where each $\boldsymbol{\mu}(t_j)$ represents a true, but unknown position at time t_j. The true position process is often modeled as a distribution conditioned on a set of model parameters $\boldsymbol{\theta}$, such that $\boldsymbol{\mu} \sim [\boldsymbol{\mu}|\boldsymbol{\theta}]$. In a hierarchical framework, the observed telemetry data are conditioned on the latent trajectory as well as the observation parameters $\boldsymbol{\psi}$, such that $\mathbf{s} \sim [\mathbf{s}|\boldsymbol{\mu}, \boldsymbol{\psi}]$. In some circumstances however, fitting a hierchical model of animal telemetry data can exceed computational resources, especially when applying an MCMC algorithm. A method proposed by Scharf et al. (2017) can be used to avoid this computational burden using an approximation method termed "process imputation".

Process Imputation

Process imputation is proposed as a method that is "similar in spirit" to multiple imputation, but applies an approximate model-fitting procedure that was originally developed by Hooten et al. (2010) and Hanks et al. (2011). The motivation for this approach arises from a decomposition of the posterior distribution of the process parameters. Using standard properties of conditional probability, the posterior distribution is written as:

$$[\boldsymbol{\theta}|\mathbf{s}] = \int [\boldsymbol{\theta}|\boldsymbol{\mu}, \mathbf{s}]\,[\boldsymbol{\mu}|\mathbf{s}]\,d\boldsymbol{\mu}.$$

This form is similar to that of the first integral which showed an approach for multiple imputation, except that $\mathbf{s}_m$ is replaced by $\boldsymbol{\mu}$. While evaluating the posterior distribution $[\boldsymbol{\theta}|\boldsymbol{\mu}, \mathbf{s}]$ up to a proportionality is relatively straightforward, it is more challenging to sample from the process imputation distribution $[\boldsymbol{\mu}|\mathbf{s}]$. This is because sampling from the process imputation distribution $[\boldsymbol{\mu}|\mathbf{s}] \propto \int [\mathbf{s}|\boldsymbol{\mu}, \boldsymbol{\psi}]\,[\boldsymbol{\mu}]\,[\boldsymbol{\psi}]\,d\boldsymbol{\psi}$ requires that we have the marginal distribution $[\boldsymbol{\mu}]$, which also requires that we evaluate $[\boldsymbol{\mu}] = \int [\boldsymbol{\mu}|\boldsymbol{\theta}]\,[\boldsymbol{\theta}]\,d\boldsymbol{\theta}$. For animal telemetry models, this integral is intractable as a result of either noninvertibility or computational burden.

To avoid the issues of sampling from the process imputation distribution $[\boldsymbol{\mu}|\mathbf{s}]$, we can instead sample from another conditional parameter $\boldsymbol{\mu}^*$. We work under the assumption that $\boldsymbol{\mu}^*$ is sufficiently similar to $\boldsymbol{\mu}$ that draws from $[\boldsymbol{\mu}^*|\mathbf{s}]$ can be used to approximate $[\boldsymbol{\theta}|\mathbf{s}] = \int [\boldsymbol{\theta}|\boldsymbol{\mu}, \mathbf{s}]\,[\boldsymbol{\mu}|\mathbf{s}]\,d\boldsymbol{\mu}$. Scharf et al. (2017) refer to $[\boldsymbol{\mu}^*|\mathbf{s}]$ as the approximate imputation distribution (AID), which is used to discern the true posterior distribution $[\boldsymbol{\theta}^*|\mathbf{s}]$. To perform process imputation, you must first specify a model for $[\boldsymbol{\mu}^*|\mathbf{s}, \boldsymbol{\phi}]$, which is parameterized by $\boldsymbol{\phi}$. Estimates of $\boldsymbol{\phi}$ are intially determined by fitting $[\boldsymbol{\mu}^*|\mathbf{s}, \boldsymbol{\phi}]$ to the data. While multiple models can be used to fit the AID, the one used in this demonstration is the Ornstein–Uhlenbeck velocity process (Johnson et al., 2008). Additional details regarding process imputation can be found in the publication by Scharf et al. (2017) and a vignette in R for fitting the model can be found in the Supplementary Material.

Example of Process Imputation of Snail Kite Data

In this example, I will be using the *crawl* package in R to fit the process imputation model based on a Ornstein-Uhlenbeck velocity process within a continuous-time correlated random walk model (CTCRW). Within this framework, I will include an observation error of 30 m in both the x and y directions for each location estimate at a regular time interval of 1 h. Upon fitting the CTCRW model, I will draw 20 samples from the AID to visualize the variability of the trajectory for each of the IDs. From this AID, I will calculate the mean position for the missing locations on the time interval of 1 h, which can be used in all further analyses for deriving movement parameters. These will be used to segment time of the time series for the movement parameters and the clustering of these segments for behavior classification.

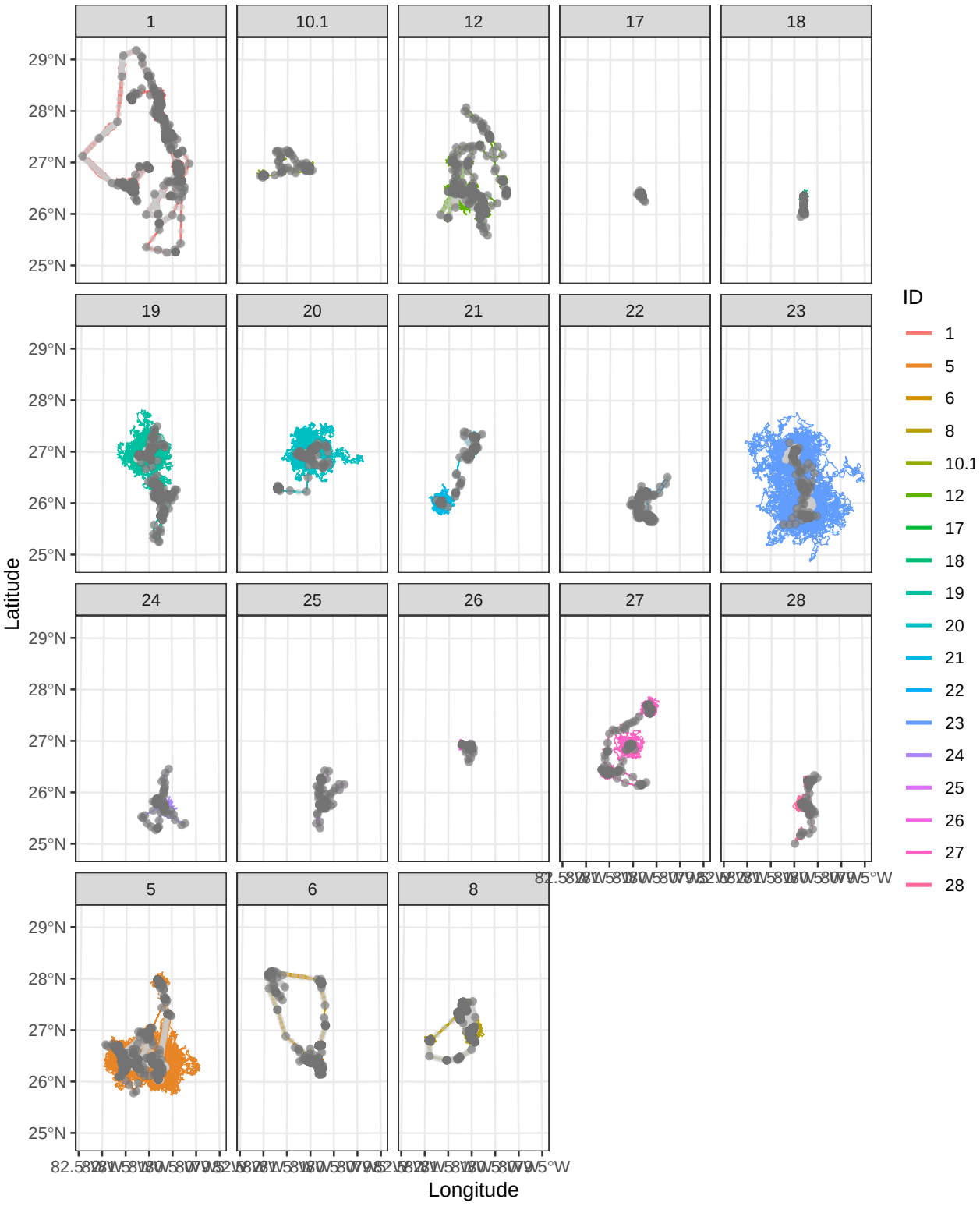

Figure 1: Simulated tracks are shown in different colors by ID for all 20 draws from the AID. Dark grey points represent known observations at irregular time intervals, while light grey points represent at a regular time interval of 1 h.

Comparison of Observed vs Imputed Data

Given occasional large time gaps in the data for all IDs, it is necessary to evaluate the proportion of imputed vs observed data. This is ensures that statistical inference is primarily made on the observed data instead of on predicted data that may not be representative of the true latent trajectory. To quickly evaluate this, two plots are shown below that display the proportion of observed or imputed data based upon the maximum time interval used to perform process imputation (1, 2, 12, 24 hrs). The total number of observations by ID (N) is shown for reference. While there is a relatively wide range in the proportion of observed/imputed data relative to N, it appears that the time interval with the greatest boost in observed data and lowest proportion of imputed data occurs for the **2 hr time interval**. However, this would still result in the loss of data that are not observed at 1 or 2 hr time intervals, which may or may not represent important steps along the trajectory.

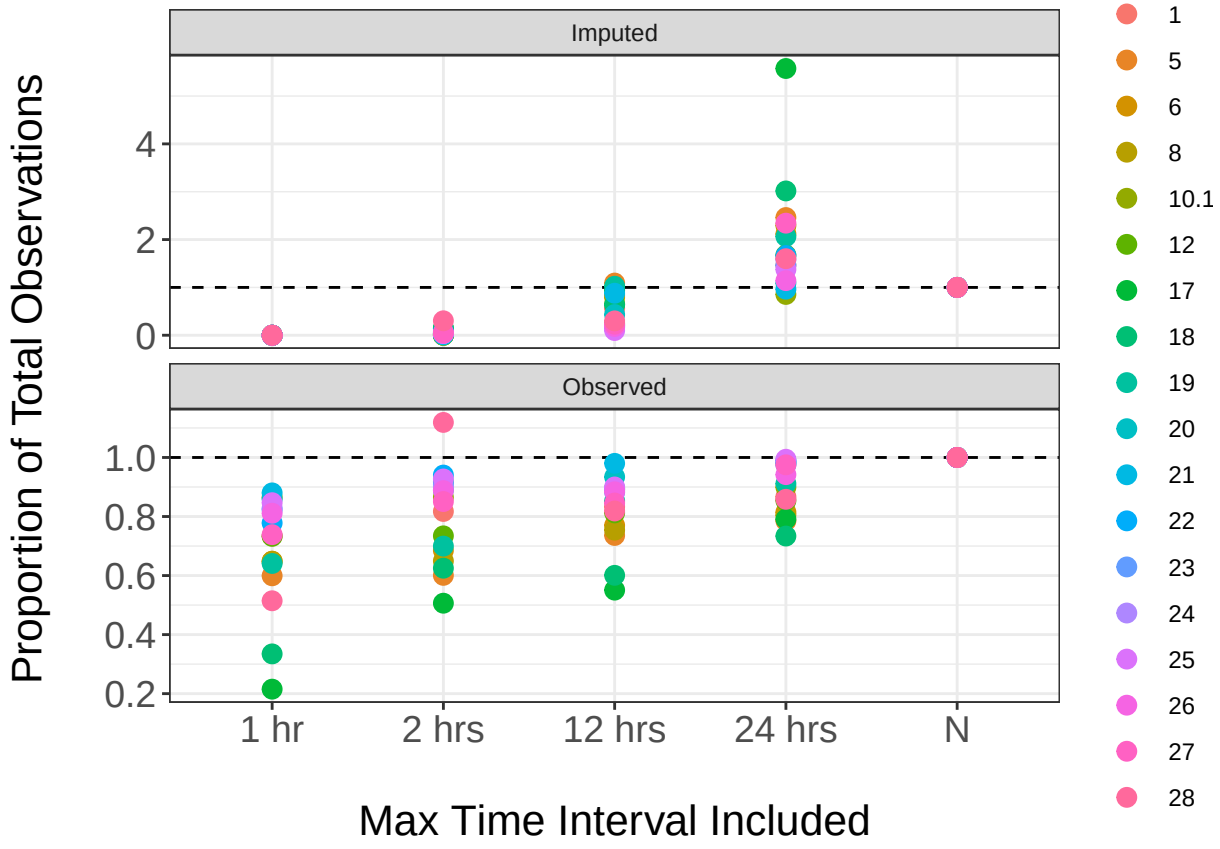

Figure 2: Observed and imputed data are scaled to the total number of observations by ID (N). N varies widely by ID, ranging from 158 to 9877 total observations. Dashed horizontal line denotes where counts are equal to N.

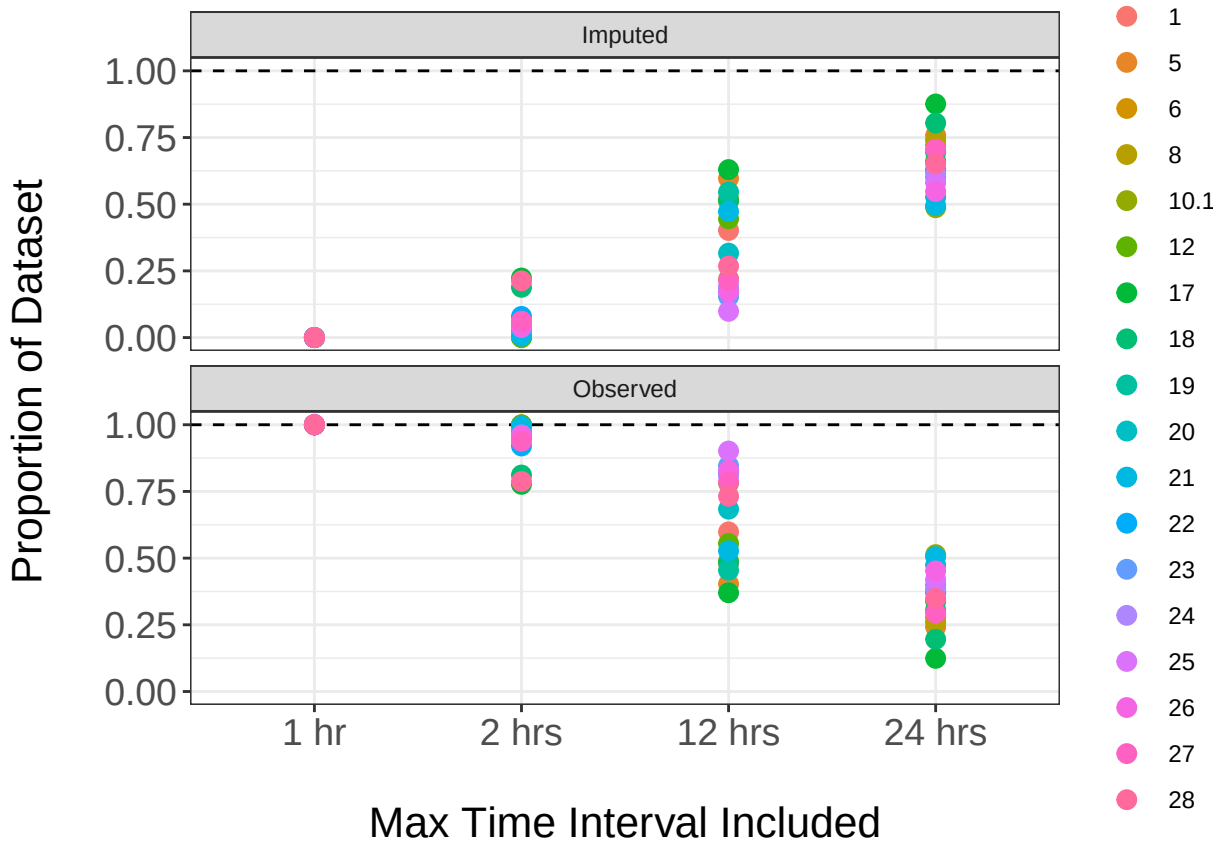

Figure 3: Observed and imputed data are scaled to the total count of both observations and imputations for each max time interval. This does not take into account the original sample size of the observed dataset. Dashed horizontal line denotes where counts are equal to total count of observations and imputations.

Evaluating Imputed Tracks with High Variability

As seen in Figure 1, simulated tracks vary considerably in how meandering they are when missing data are imputed. This is likely a result of relocations that are not very far apart, but are separated by a large time interval (e.g. $dt >> 1$ h). The following figures investigate whether the clouds of simulated tracks observed for some IDs, such as **IDs 19 and 23**, are a result of very large time intervals. For comparison, I will also be displaying a plot for **ID 1**, which did not show this high level of variability among simulations. These plots display the mean trajectory resulting from 20 imputed paths. Red points denote the beginning and end of time intervals ≥ 15 h.

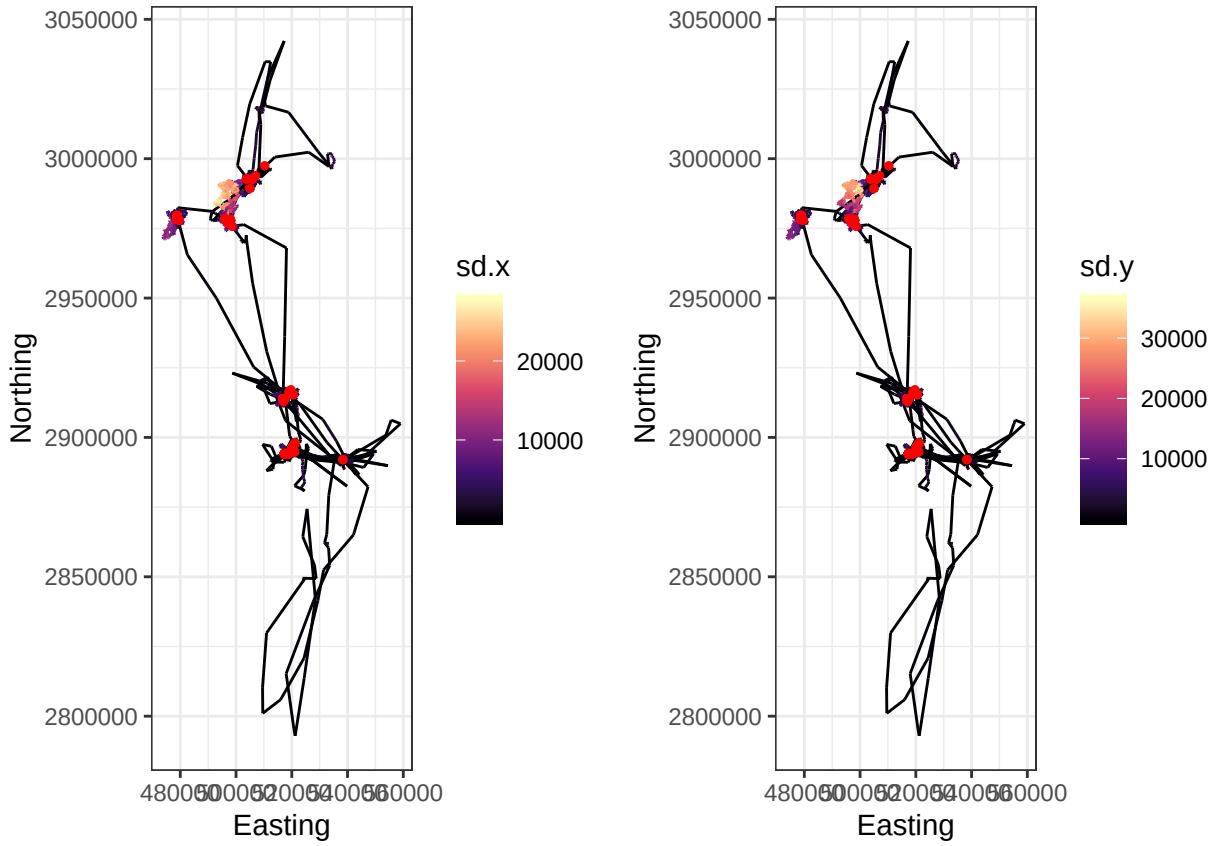

Figure 4: Mean trajectory of 20 imputed paths for ID 19. Color ramp denotes standard deviation across all 20 trajectories in the x or y direction. Red points indicate beginning and end of large time intervals.

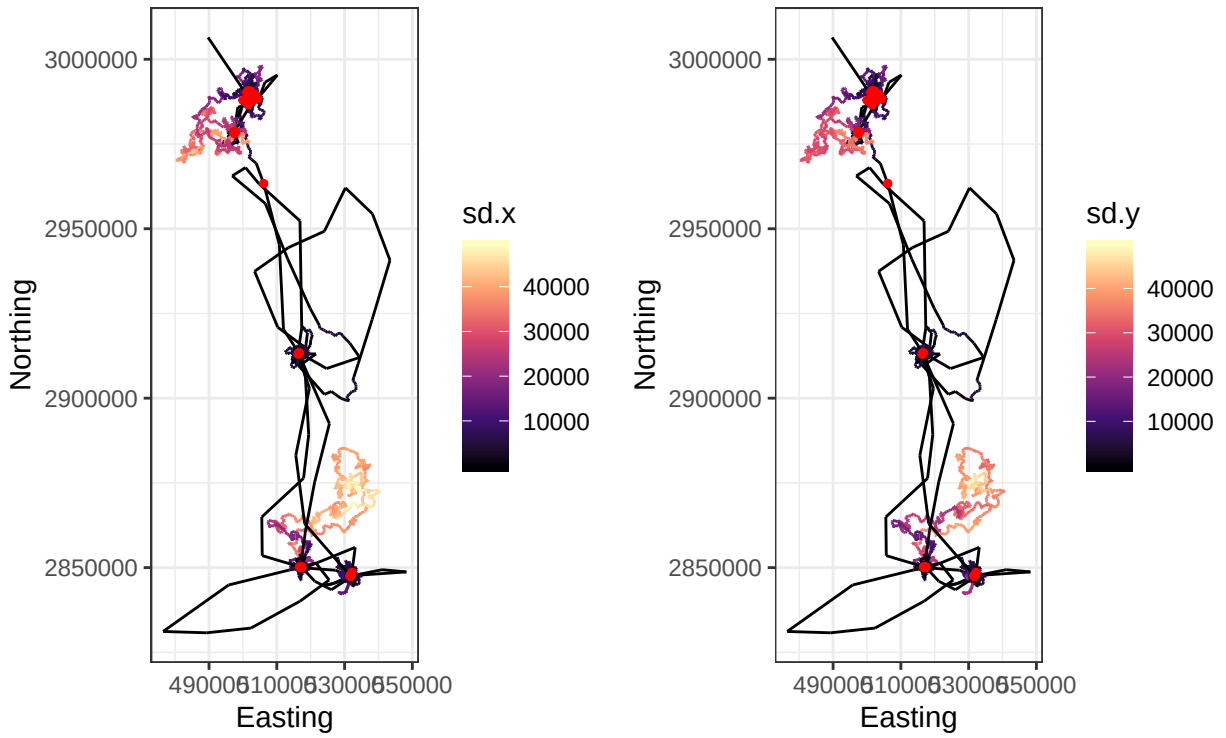

Figure 5: Mean trajectory of 20 imputed paths for ID 23. Color ramp denotes standard deviation across all 20 trajectories in the x or y direction. Red points indicate beginning and end of large time intervals.

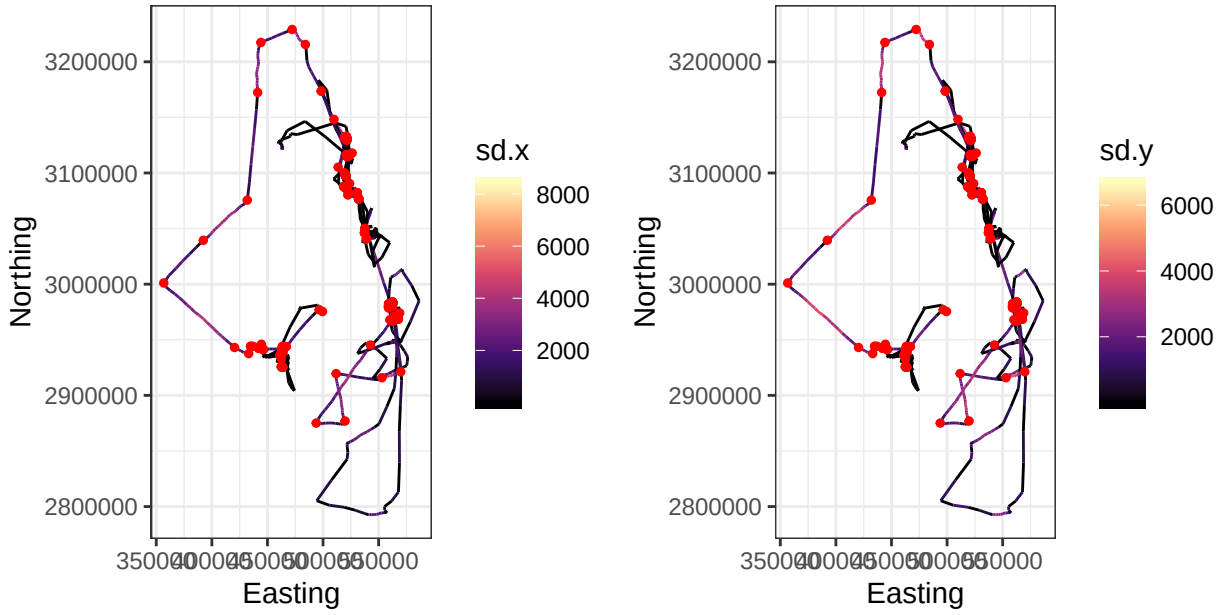

Figure 6: Mean trajectory of 20 imputed paths for ID 1. Color ramp denotes standard deviation across all 20 trajectories in the x or y direction. Red points indicate beginning and end of large time intervals.

Given these comparisons, it appears that these very large time gaps appear to be the primary cause of high variability in the imputed tracks. Although there are some large time intervals for **ID 1**, these imputed tracks appear relatively straight. This is likely a result of the large distances covered for this ID, which would necessitate the track be relatively straight to reach the next observed location.

Conclusions

These exploratory analyses demonstrate the utility of process imputation for recovering observations at other time intervals than the primary one (1 h), but also identify some shortcomings for the datasets where time intervals may be highly irregular. For the snail kite telemetry data, the use of a maximum time interval of 12 h would result in imputed data comprising nearly 50% of the total dataset and approximately 75% if set to 24 h. Therefore, the use of a 2 h maximum time interval for imputation would provide the best tradeoff for a larger sample size while minimizing the proportion of imputed data in further analyses. However, there is still the problem of the remaining observations that will be excluded from further analyses. These excluded data represent 30% of the whole dataset on average (range: 14 - 67%). Cutting the trajectories above a threshold of some time interval would not necessarily help if imputation is already conducted on a relatively short max time interval. For the snail kite data, this is because many of these irregularly sampled data are sampled at intervals shorter than 1 h (~ 4200 observations; 8% of entire dataset). While this does not necessarily need to be addressed, it would need to be recognized as a limitation of our method if using telemetry data that are highly irregular.