Process Imputation of Animal Telemetry Data

Josh Cullen

25 November 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.

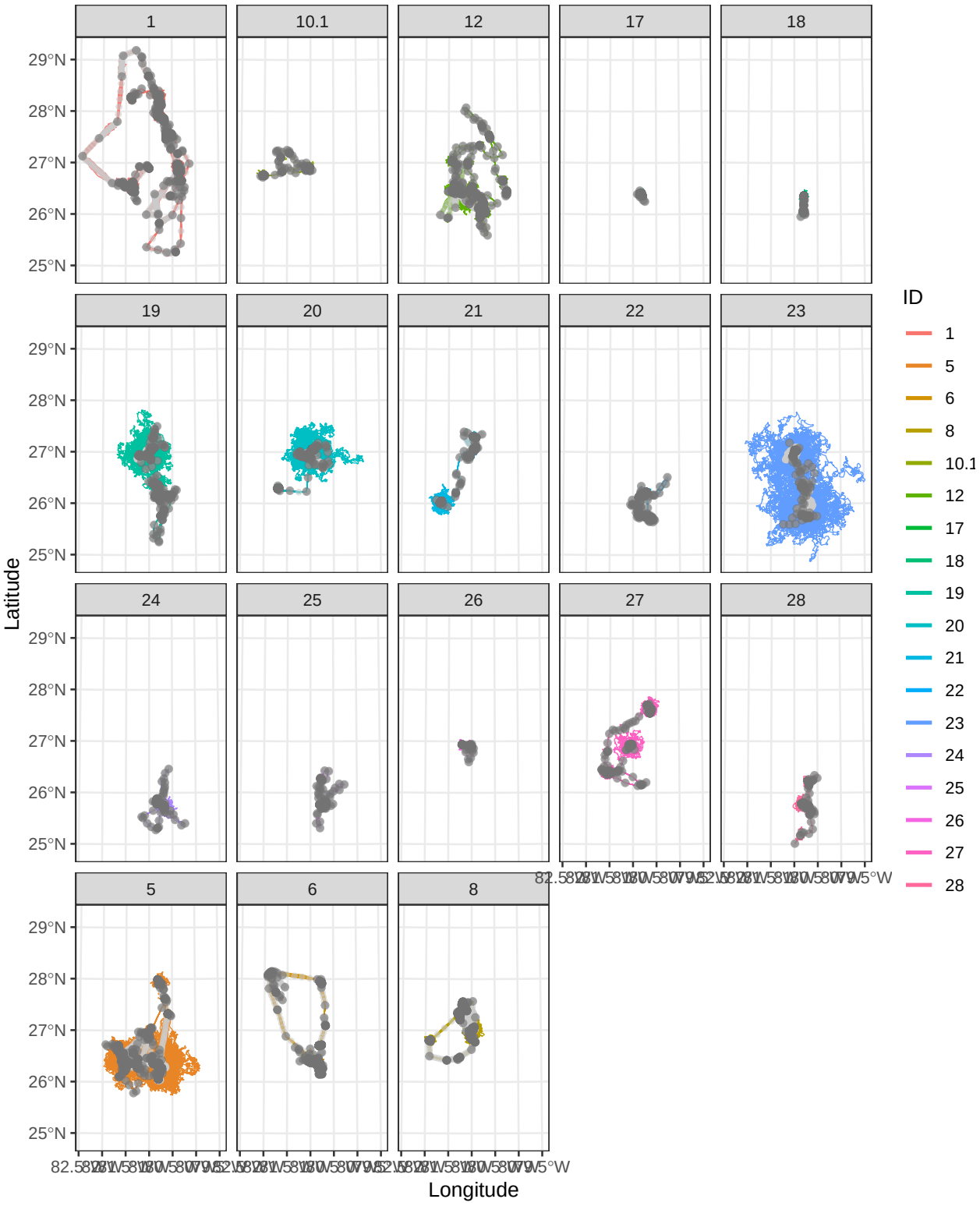

Figure 1: Simulated tracks are shown in different colors by ID for all 20 draws from the AID. Dark grey points represent known observations at irregular time intervals, while light grey points represent at a regular time interval of 1 h.