A robust method for identifying behavioral changes in gappy animal movement datasets

Eli Gurarie
Quantitative Ecology and Resource Management
University of Washington, Seattle

April 6, 2008

Abstract

Individual animal movement data is collected at an increasing rate as remote-sensing technology develops. At its best, analysis of the data can suggest mechanisms by which organisms exploit a heterogeneous and variable environment. Unfortunately, movement data are multidimensional, non-independent and almost always suffer from the twin banes of measurement error and gappiness, making appropriate analyses far from straightforward. While error in the data can be accounted for by state-space models, an increasingly popular statistical approach, gappiness (i.e. irregular intervals between times of measurement) has not been well-addressed in the literature. Gappiness is particularly widespread in data on marine organisms, for which remote sensing depends on a combination of transmitter exposure and satellite presence. I suggest a method of dealing with gappiness by identifying a persistence component of movement that can be modelled as a continuous auto-correlated stochastic process with three parameters: a mean, a variance and a continuous auto-correlation coefficient. I then develop several methods for identifying changes between behavioral modes. A single breakpoint can be found with a maximum likelihood estimation over an entire gappy time-series, while multiple breakpoints can be found by sweeping a window and identifying which of the parameters change at any given breakpoint. After demonstating the robustness of the method with simulation, I apply the routine to GPS data collected on a northern fur seal (*Callorhinus ursinus*) in the Kuril Islands in Russia. The resulting model suggests a complex behavioral profile that can potentially be analyzed against environmental covariates.

Keywords: animal movement; gappy time series; structural shifts; behavioral signals; continuous autoregressive models; northernn fur seal; Callorhinus ursinus*; foraging trips*

Contents

1 Introduction **2**

2 Models of movement **4**

3 Defining correlation structure in gappy time-series **6**
3.1 Gappy time-series model 6
3.2 Example with data 7

4 Identifying structural shifts **7**
4.1 Identifying models 8
4.2 Simulation study 9
4.3 Multiple changepoints 10

5 Data **12**

6 Results **12**

7 Discussion **14**

8 Conclusions **16**

9 Acknowledgement **17**

A Tables **20**

B Figures **21**

C Code **30**
C.1 Estimating continuous correlation coefficient ρ. 30
C.2 Obtaining log-likelihood of single structural breakpoint τ. 30

1 Introduction

In recent years, there has been a rapidly growing body of work devoted to the detailed study of animal movements in the wild, mirroring the increase in the technological ability to accumulate data. Most fundamental ecological processes are a direct result of movement, including foraging success, breeding success, migrations and dispersals. Furthermore, individual animal movements are a measurable behavioral response to a combination of internal states, physiological constraints and environmental factors. Informative analysis of movement data can yield sophisticated insights

into the behavioral mechanisms that allow organisms to exploit temporally variable and spatially heterogeneous environments. This appears to be a common goal of many animal movement researchers.

Analysis of movement data, however, is far from straightforward, since the data is multi-dimensional and auto-correlated in space and time. A common approach to modeling a movement track is to apply some variety of a correlated random walk models (Skellam 1951, Turchin 1998, Okubo and Levin 2001), which typically hypothesizes some distribution of step-lengths and turning angles clustered around zero degrees. Over a long enough time-scale, multiple behaviors can be captured by a single dataset and the properties of mixed random walk models have been explored (Grünbaum 2000, Skalski and Gilliam 2003). Recently, models have been constructed that successfully relate changes in correlated random walk parameters to environmental and landscape features (Morales et al. 2004, Forester et al. 2007, Haydon et al. 2008, Aarts et al. 2008).

There are two common features of movement data which complicate the straightforward application of correlated random walk models. The first is error in the measurement process. A fruitful body of research has emerged recently that addresses measurement error with the use of state-space models (SSM's) (Jonsen et al. 2003; 2005, Royer et al. 2005, Patterson et al. 2008). State-space models are efficient frameworks for parsing movement data into a process model of movement and an observation model that accounts for error. They have bee, in particular, invaluable in the interpretation of ARGOS-satellite derived telemetry data.

The second significant very common issue with movement data is irregular timing of measurements, sometimes referred to as "gappiness". This issue is particularly important for tagged marine species such as pinnipeds, whales, penguins, sea turtles and large pelagic fish, where sending a signal requires that the tag be exposed to the air and that a connection be obtained with the receiving satellite. Clearly, a straightforward estimation of turning angle distributions and estimated velocities is compromised by the irregularity of the data. The solution to this problem is to model the movement data as an irregular subsampling of a continuous auto-correlated process (Johnson et al., *in press*).

In this paper, I present an efficient and robust method for identifying behavioral changes in a gappy movement dataset. The fundamental idea is to sweep an analysis-window over the movement data and identify the time and nature of any significant behavioral shifts, outputting estimates of descriptive parameters and aggregating them into a behavioral summary of the movement.

The complete analysis method involves several steps. I begin by suggesting a natural decomposition of velocity and turning-angle data into persistence and turning velocity components. These velocity components have the statistically attractive feature of being locally stationary and Gaussian, crucially allowing for the straightforward modeling of an auto-correlation structure. I consider these time-series to be samplings from a continuous auto-regressive process and present a straightforward likelihood-based method to determine a continuous auto-regression coefficient, a

parameter that has an important biological interpretation. The movement can then be described by three parameters for each of the two velocity components: a mean, a variance and the continuous auto-correlation. Next, I present a maximum likelihood method to estimate the location of a single change point, i.e. the likeliest time at which one or more of the three parameters that describe the time-series changes values. It is of interest to be sure that the change is a "real" one and not an artifact of randomness, as well as to identify which of the parameters are changing. Both of these challenges are met by applying an information-based criterion to test all possible permutations of model parameters changing. A simulation study demonstrates that the Bayesian Information Criterion (BIC) does an excellent job of identifying which of the parameters are really changing their value at a given most likely changepoint. I provide code for all of the relevant pieces of the analysis algorithm in the Appendix.

The complete analysis machinery was applied to GPS-derived data on movements of an adult Northern fur seal female (*Callorhinus ursinus*) in the Kuril Islands in Russia. After giving birth, female fur seals nurse their pups for several weeks before beginning to take extended foraging trips up to seven days in length. They can wander up to a hundred kilometers from the rookery of birth in pursuit of sufficient forage to fulfill both their own energetic requirements and that of the nursing pup. These foraging trips show distinct modes of movement, including rapid travel, periods of frequent diving and semi-somnolent drifting. The application of the analysis method allows for the identification of a complex behavioral profile.

2 Models of movement

We take a "Lagrangian" approach to modeling movement, that is, we place the organism at the center of our analysis. Thus, rather than analyze the absolute positions (X_i, Y_i) or compass orientation (Φ_i), we examine the two variables that the organism controls: speed V_i and turning angles (θ_i). Because we are specifically interested in analyzing gappy data, we have an additional time vector of measurements (T_i), and denote out velocities and angles as $V(T_i)$ and $\theta(T_i)$.

A very common method for modeling movement data is the apply a two-dimensional correlated random walk (CRW). A traditional CRW, which we might further specify as an *unoriented*, *homogenous* correlated random walk, is defined as one in which step-lengths R_t are independent and have some positive random distribution and the turning angles θ_t have some non-uniform distribution clustered around 0 degrees (Kareiva and Shigesada 1982, Marsh and Jones 1988). Common distributions for velocities include the Weibull and lognormal, while turning angles are often modelled with wrapped Cauchy distribution with zero mean and clustering parameter κ: When $\kappa = 0$ the angles are distributed uniformly between $-\pi$ and π, and when $\kappa = 1$, all the angles are concentrated at 0 (Fisher and Lee 1994). A CRW can generate smooth tracks that do a good job of resembling real animal movement data, and the parameters of the model are readily estimable. For example,

$\widehat{\kappa}$ is just the expected value of the cosine of the turning angles.

Whatever distribution for turning angles has been used to model movement, very rarely has the auto-correlation between subsequent tuning angles been discussed. Typically, temporal auto-correlation in movement data is "absorbed" into the clustering coefficient, though there are some fundamental differences between a movement that is highly clustered and one in which turning angles are auto-correlated. A related problem can arise with the asumption of independence for the velocities. Nonetheless, CRW's have proven to be useful and tractable models for unoriented, homogeneous movements.

It is, however, not entirely clear how to fit a CRW to gappy data, since neither $V(t_i)$ nor $\theta(t_i)$ are identically distributed when the interval between measurements changes. There are two feasible approaches for dealing with this problem. The first is to calculate or approximate the expected distributions of R and Θ as a function of the gap. The second, and the one that is addressed in this report, is to transform the data by decomposing every step into orthogonal components of persistence velocity $V_p(t)$ and turning velocity $V_t(t)$:

$$V_p(t) = V(t)\cos(\Theta(t)) \tag{1}$$

$$V_t(t) = V(t)\sin(\Theta(t)) \tag{2}$$

V_p captures the tendency of a movement to persist in a given direction, and the velocity of that movement while V_t captures the tendency of movement to head in a perpendicular direction at a given interval. Thus, the primary descriptive features of movement, namely speed, directional persistence, and variability are captured in these variables. A further fundamental advantage of these transformations is that the resulting variables are well-modeled by stationary, Gaussian, autoregressive time series models. Empirical explorations of movement data via histograms or qqnorm plots suggest that both V_p and V_t are well approximated by mixed normal distributions, with V_t notably always having a mean very close to zero. This appealing statistical property allows for the application of an arsenal of analysis techniques for characterizing auto-correlated data.

Though geometrically orthogonal, these two variables are not necessarily independent. Because, however, our eventual technique is ultimately descriptive and the interpretation of each of these variables is somewhat unique, we choose to analyze them separately. An investigation of their correlation structure is the subject of future work.

3 Defining correlation structure in gappy time-series

3.1 Gappy time-series model

We begin the analysis with the empirically supported assumption that the persistence velocity V_p (1) is a sample from a stationary, Gaussian continuous space, continuous time process $X(t)$ which has the following properties:

$$\begin{aligned} X(0) &= X_0 \\ \mathrm{E}\left[X(t)\right] &= \mu \\ \mathrm{Var}\left[X(t)\right] &= \sigma^2 \\ \mathrm{Corr}[X(t), X(t-\tau)] &= \rho^{\tau}, \end{aligned} \tag{3}$$

where $0 < \rho < 1$ is the first order auto-correlation at a time lag 1, i.e. at whatever units the time is measured. (Note, all subsequent discussion applies analogously to V_t, with the constraing that $\mu = 0$). Consider observations X_i that are made at times t_i, where $t_0 = 0$ and beginning at initial observation X_0. The series X_i can be described as

$$X_i = \mu + \rho^{\tau_i}(X_{i-1} - \mu) + \epsilon_i, \tag{4}$$

where $i \in \{1, ..., n\}$, $\tau_i = t_i - t_{i-1}$ is the time interval between subsequent observations, ρ^{τ_i} is the auto-correlation as a function of the time gap and ϵ_i is a stochastic error term. Given the constraints in 3, ϵ_i can be shown to have mean 0 and variance $\sigma^2(1 - \rho^{2\tau_i})$. The derivation of the variance is obtained as follows:

$$\begin{aligned} \mathrm{Var}\left[\epsilon_i\right] &= \mathrm{Var}\left[X_i - \rho^{\tau_i} X_{i-1}\right] \\ &= \sigma^2 + \rho^{2\tau_i}\sigma^2 - 2\rho^{\tau_i}\mathrm{Cov}[X_i, X_{i-1}] \\ &= \sigma^2(1 - \rho^{2\tau_i}) \end{aligned} \tag{5}$$

In order to estimate the continuous correlation function ρ, we need to express the likelihood function for the entire process. Since any X_i depends only on the previous observed value X_{i-1}, a total conditional likelihood can be written

$$L(\rho|\mathbf{X}, \mathbf{T}) = \prod_{i=1}^{n} f(X_i|X_{i-1}, \tau_i, \rho), \tag{6}$$

where $\mathbf{X}$ and τ represent the vector of observations and times of observation respectively and the

distribution function f is the probability density function of the conditional distribution $X_i|X_{i-1}$. For the process (4) with a Gaussian error structure, f is given by

$$f(X_i|X_{i-1}) = \frac{1}{\sigma\sqrt{2\pi(1-\rho^{2\tau_i})}} \exp\left(\frac{(X_i - \rho^{\tau_i}(X_{i-1}-\mu))^2}{2\sigma^2(1-\rho^{2\tau_i})}\right). \tag{7}$$

This likelihood (6) is smooth and easily maximized over the range of possible values for $0 < \rho < 1$. The estimate for ρ can be expressed as

$$\widehat{\rho} = \underset{\rho}{\operatorname{argmax}}\ L(\rho|\mathbf{X}, \mathbf{T}) \tag{8}$$

To verify this estimation routine, we simulated an AR(1) process with coefficient $\rho = 0.8$ and length $n = 1000$. We then randomly sampled 100 observations from this process and estimate $\widehat{\rho}$ based on the gappy time-series. Performing this operation 100 times yielded a mean estimate $\widehat{\rho} = 0.797$ with standard error 0.052. The MLE estimate works excellently, even for exremely gappy data.

3.2 Example with data

Magellanic penguins (*Spheniscus magellanicus*) in Argentina were tagged with sattelite transmitters in early 2005 and their movement during foraging trips tracked using the ARGOS tracking system (figure 2). Here we analyze a single foraging trip of a penguin (P1) performed over eight days from January 28 to February 4. There were 163 positions recorded and the interval between the position fixes ranges from 1 to 278 minutes.

We extracted a 'velocity tendency' time-series, i.e. $V_p = V(t_i)\cos(\theta(t_i))$ where $V(t_i)$ is the measured speed (distance travelled divided by time interval) and $\theta(t_i)$ is the turning angle at $t(i)$ from the previous direction. We obtained $\widehat{\rho}$ using the MLE estimation routine.

The resulting estimates for f movement for the penguin are $\widehat{\mu} = \bar{V}_P = 1.97$ km/h, $\widehat{\sigma} = 1.79$ km/day and $\widehat{\rho} = 0.448$ (figure 3). Thus, if the penguin's direction were sampled once an hour, the acf of the resulting time-series would be $\gamma(1) = 0.45$. A preliminary conclusion from this result is that the penguin exhibits multi-hour persistence in it's movement.

4 Identifying structural shifts

We now address the question of how to expand the methods outlined in section 3 in order to identify structural shifts. A *structural shift* is a change in the underlying continuous process, for example a behavioral change from a goal-oriented travelling mode of movement to a search and forage kind of movement. The challenge is to infer where a structural shift occurs within a gappy and error

ridden time series measurement of a process. We refer to the time at which the structural shift occurs as the *changepoint*.

Consider a continuous stochastic process $X(t)$ for $0 < t < T$ defined by parameter set $\mathbf{\Theta}(t)$ whose values change at an unknown time points τ such that:

$$\mathbf{\Theta(t)} = \begin{cases} \mathbf{\Theta_1} & \text{if} \quad 0 < t \leq \tau \\ \mathbf{\Theta_2} & \text{if} \quad \tau < t \leq T \end{cases} \tag{9}$$

We now take N samples from the continuous process $X(t)$ to obtain a time series X_i at times T_i. If n is defined as the number of measurements within the first regime, such that $T_n \equiv \max(T_i < \tau)$, then the likelihood of the parametrization $(\theta_1, \theta_2, \tau_1)$ given data X_i is given simply by the product of the two likelihoods:

$$L(\mathbf{\Theta}|\mathbf{X}, \mathbf{T}) = \prod_{i=1}^{n} f(X_i|X_{i-1}, \mathbf{\Theta}_1) \prod_{j=n+1}^{N} f(X_j|X_{j-1}, \mathbf{\Theta}_2) \tag{10}$$

This likelihood can be maximized by sweeping over all possible values of n (from 1 to N), and obtaining the MLE's for the remaining parameters after the split using the method described in section 3.

Thus, the parameter estimates can be expressed as:

$$\widehat{n} = \underset{n}{\text{argmax}}\ L(\mathbf{\Theta}|\mathbf{X}, \mathbf{T}) \tag{11}$$

$$\hat{\mu}_j = \bar{X}_j \tag{12}$$

$$\hat{\sigma}_j = S_j \tag{13}$$

$$\hat{\rho}_j = \underset{\rho}{\text{argmax}}\ L(\rho|\mathbf{X_j}, \mathbf{T_j}, \hat{\mu}_j, \hat{\sigma}_j) \tag{14}$$

where $j \in (1, 2)$ indexes the two regimes, such that if $j = 1$ the data and estimates are taken in the range $i = (1, 2, ..., \widehat{n})$ and if $j = 2$, they are taken from the second regime $(i = \widehat{n} + 1, \widehat{n} + 2, ...N)$. We term the estimate for the changepoint $(t_{\widehat{n}})$ the *most likely changepoint* MLCP.

The corresponding estimates for the orthogonal component of velocity V_t are simplified by the fact that the mean can safely be assumed to be 0. Indeed, analyses have verified that when this assumption is relaxed, the mean is almost never significantly different from zero

4.1 Identifying models

Each of the three parameters that characterize a movement can change, and a change in each of the values corresponds to a different behavioral interpretation. Thus, for $V\cos(\theta)$, an increase in μ corresponds to a combination of faster and more directed movement. An increase in σ indicates

more variable movement, e.g. more common stopping and moving or a greater increase in diving. A higher ρ indicates more directed and correlated movements, whether fast or slow. For $V\sin(\theta)$, higher values of σ indicate more turning or zig-zaggy movement, while higher values of ρ indicate longer turning radii. It is consequently of great interest to be able to identify which, if any, of the parameters actually change at the MLCP.

There are eight possible models to consider when analyzing a changepoint. These are defined as follows: M0 is the null hypothesis ($\mu_1 = \mu_2$, $\sigma_1 = \sigma_2$ and $\rho_1 = \rho_2$); M1, M2 and M3 have one inequality each while the other two parameters remain constant ($\mu_1 \neq \mu_2$, $\sigma_1 \neq \sigma_2$, $\rho_1 \neq \rho_2$ respectively); M4, M5 and M6 have one equality each while two other parameters change (μ and σ, μ and ρ, σ and ρ respectively); and M7 is the most "alternate" hypothesis, in which all parameter values change at the MLBP.

Because the conditional likelihood is well-defined, we can apply a consistent criterion to select our models. We compared two criteria, AIC (Akaike's Information Criterion) and BIC (Bayesian Information Criterion) defined as:

$$I_A(\mathbf{X},\mathbf{T}) = -2n\log\left(L(\hat{\theta}|\mathbf{X},\mathbf{T})\right) + 2d \quad (15)$$

$$I_B(\mathbf{X},\mathbf{T}) = -2n\log\left(L(\hat{\theta}|\mathbf{X},\mathbf{T})\right) + d\log(n) \quad (16)$$

where $L(\hat{X}|\theta)$ is the likelihood as defined in equation 10, d is the number of parameters in each of the eight models (ranging from $d = 3$ for M0 to $d = 6$ for M7). The model with the lowest value of the information criterion is chosen. The BIC is more conservative than the AIC at selecting models, since it additionally "penalizes" the criterion according to the size of the dataset.

4.2 Simulation study

Figure 4 illustrates an example of the single changepoint estimation routine. The simulated data in this example comes from an underlying process of length $N = 2000$ in which a discrete structural shift occurs at $t = 1000$ in which the mean drops from 5 to zero, the variance decreases from 10 to 5, and the autocorrelation coefficient drops from 0.9 to 0.2. This simulation represents a shift from a highly variable, positively biased, correlated process associated with directed movement to a zero mean, lower variance, less correlated process that mimics random foraging and feeding. We randomly sampled 100 points from the complete time series and obtained the MLCP according to the method described above. The resulting estimated parameters are listed in the figure caption.

The plots of the log-likelihoods given in figure 4 give an idea of how precise the τ estimate might be. The log-likelihood profile seems adequately peaked around the correct value. It should be noted that the time series in this example is particularly gappy and the total dataset is relatively short.

We explored the properties of this experiment for a variety of parameter value changes by simulating 100 gappy processes ($N = 400$, $n = 50$, $T_{br} = 200$) for each of eight different models (see table 1A) ranging from the null model of no change in parameter values to the most alternate model of change in all parameter values. Resulting parameter estimates appear unbiased and relatively precise (table 1B), all falling well within one standard deviation of the true value. Perhaps most importantly, the MLCP estimates are highly accurate, with means between 197 and 203. We also applied AIC and BIC to assess the eight models and report their results (table 1B and figure 5). The BIC performs far better than the AIC at identifying the true model, with, notably, a 78% rate of correctly identifying the null hypothesis and 72-94% probabilities of identifying parameters with a single parameter changing. For those models where two parameters shift, BIC tends to falsely favor the most complex model. In contrast, AIC performs miserably, never selecting the null model correctly and in other cases almost always choosing more complex models than necessary.

When analyzing actual data, a researcher may have greater interest in identifying the location and direction of significant, detectable structural shifts than in estimating the parameters themselves. For this reason, we prefer the more conservative criterion. A complete power analysis is very difficult to perform for the selection mechanism, since the ability of the model to correctly identify the model depends in complicated, non-independent ways on the magnitude of the difference between the paramters, the extent of the gappiness in the dataset and the length of the series. However, an effort was made in the simulation study to look at differences that are relatively comparable if not smaller than those in the actual data. The ability of the BIC to pick the appropriate level of complexity inspires confidence in this method of model selection.

4.3 Multiple changepoints

Heretofore, we have only discussed the identification of a single most likely breakpoint. Within the a single track, an organism likely exhibits multiple behaviors. Rigorously, a continuous process $X(t)$ can be defined such that for any interval between $0 < t < \tau_m$ defined by parameter set $\mathbf{\Theta}(t)$ whose values change at unknown time points $\mathbf{T} = \{\tau_1, ..., \tau_m\}$ such that:

$$\mathbf{\Theta(t)} = \begin{cases} \theta_1 & \text{if} \quad 0 < t \leq \tau_1 \\ \theta_2 & \text{if} \quad \tau_1 < t \leq \tau_2 \\ \vdots & \vdots \quad \vdots \\ \theta_m & \text{if} \quad \tau_{m-1} < t \leq \tau_m \end{cases} \tag{17}$$

Estimating the parametrization $\mathbf{\Theta}$ for the multiple change-point model is non-trivial problem that has generated some literature, though only for regular (non gappy) time-series and applied primarily to financial markets (see, for example Chib 1998). Furthermore, there are complications with model selection and inference in selecting the number of breakpoints in a longer dataset.

My approach to this problem is to pass a window over the complete time-series and apply the single changepoint analysis described above at each window. The algorithm can be summarized as follows:

1. Select a window of length $l << N$.
2. Find the most likely changepoint in a subsample of the data $X_1...X_l$.
3. Use some criterion to accept or reject the "significance" of the changepoint for each of the parameters μ, σ and ρ.
4. Based on the result of the test, log the location of the behavioral changepoint and the resulting estimated parameter values: $\hat{\mu}_1$, $\hat{\sigma}_1$, $\hat{\rho}_1$, $\hat{\mu}_2$, $\hat{\sigma}_2$, and $\hat{\rho}_2$.
5. Shift the window up one datapoint, and repeat steps 1-4.

Throughout the running of this algorithm, we output the estimates according to the model chosen by the BIC. Thus, if M0 is chosen, we estimate a single value for each of the parameters for the entire dataset. If M7 is chosen, we separately estimate parameters at each side of the breakpoint. All of the estimates are logged and averaged after the sweep is performed.

This method has the advantage of being able to capture arbitrarily many behaviors. Furthermore, by returning and averaging the model estimated parameters at every timestep multiple times, we allow for gradual shifts in behavioral parameters between the discrete structural shifts, and potentially dramatic jumps where the behavior does change suddenly and significantly The agreggated information including both the parameter estimates, the significant breakpoints and their nature and direction (i.e. increasing μ, decreasing ρ, etc.) can be considered a distilled behavioral model of movement.

It should be noted that the one tunable variable in this method is the size of the window. The greater the size of the window, the more consistent the estimates provided will be and the higher the power of the model selection (figure 5), but the probability increases that shifts at smaller scales will be lost. A smaller window size will reveal finer-scale structure in the data, but the risk of spurious responses increases. Furthermore, issues can arise when estimating means and variances with small sample sizes, especially when the correlation is high. Because of the behavioral complexity of the dataset in the application that follows, I chose a window of size 30 which is probably about the lower limit of an acceptable window size (see Results).

All analysis was performed using the freely available `R` programming package.[1]

[1] `http://www.R-project.org`

5 Data

I applied the movement analysis method to GPS data collected on several foraging trips taken by a nursing female northern fur seal (*Callorhinus ursinus*). Northern fur seals aggregate annually during the summer months at large rookeries. The females give birth usually in late June through July, and, after 8-10 consecutive days of nursing, begin to take foraging trips of up to a week in length to replenish energetic reserves (Gentry 1998). The data analyzed here was collected from a female tagged in summer of 2007 at Dolgaya Rock, one of the Lovushki Islands - a small group in the central Kuril Island chain in Russia which runs between the Pacific Ocean to the southeast and the Sea of Okhotsk to the northwest (see figure 6). The Lovushki Islands are the location of one of the largest northern fur seal rookeries in the northwestern Pacific (Loughlin et al. 1984, Burkanov *pers. comm.*).

The female discussed in this paper, labeled NFS07-03, was captured on June 21, when her pup was one to three days old, and instrumented with a Fastloc®GPS data-logging device (MK10-F, Wildlife Computers Inc.) The tag allows for quick fixes and relatively high accuracy, with at least 60% of the locations within 100 m (Bryant 2007). The MK10-F device was also equipped with dive depth, temperature and light sensors, and the animal was instrumented with a separate stomach temperature sensor. Here, we only consider the movement data; however, the existence of corroborating evidence for foraging behaviors, such as diving and prey capture, makes this an ideal system to explore what can be inferred from pure movement data.

The animal was monitored for 38 days in total, taking seven foraging trips in that time. The first trip occurred 9 days after tagging, on July 1, and lasted just over five days, while the last trip began on July 27 and lasted about two and a half days. On all trips, the fur seal headed in a northwesterly direction towards the Sea of Okhotsk (figure 6). A total of 763 position fixes were obtained over all seven trips, with individual trips ranging from 30 and 33 datapoints (trips 5 and 6) to 205 datapoints in length for trip 1. The time intervals between the fixes range widely from a less than a minute to 700 minutes, with the majority (over 80%) clustered around 15, 30 and 45 minutes. The data was filtered to exclude implausible swim speeds (> 11 km/h) and were georeferenced using standard methods. Velocities were estimated by dividing displacements by time intervals, and turning angles were calculated directly from the positions.

6 Results

I applied the analysis technique to five of the seven trips (those with more than 70 location fixes) with a window of size 30. At an average time interval between measurements of 37 minutes, the window covered on average a period of 18 hours, which seemed to allow for reasonable chance to pick up single behavioral shifts at an acceptable cost to power. The output of analyses with window

size 50, corresponding to about a 24 hour period, seemed too insensitive since it is reasonable to expect that behaviors can change significantly more than once in a day.

Model outputs for two of the trips (1 and 7) are presented graphically in figure 7. The evolution of the estimates for the V_p parameters in time is schematized in Figure 8.

Because the model output is somewhat complex, it is constructive to walk through a single track from beginning to end. The first trip is the longest (5.1 days long) and furthest (max. distance from rookery 96 km), as is typical for female fur seals taking their first feeding trip after a fasting period associated with birth. The fur seals's initial departure from the rookery is marked by high values of of V_p ($\widehat{\mu}$ around 5 km per hour) and a high estimated per-hour auto-correlation $\rho = 0.36$ (NB since the time data is in hours, the ρ is an estimate of what the first order auto-correlation coefficient at lag 1 would be if the movement data were collected once an hour exactly without gaps). The first significant changepoint, occuring at 01:53 in the middle of Night 1 was selected as the MLCP by every single one of the 25 windows that passed over it. The models chosen by the BIC were split about evenly between M4 (μ and σ change) and M7 (all three parameters change). The estimates for μ drop from near 4 km/h to around 2 km/h, estimates for σ drop from 1.5 to 0.7, estimates for ρ decrease more gradually from 0.3 to under 0.1. A similarly dramatic changepoint occurs at around 1900 before the fifth and last night of the trip, where fully 30 MLCP's were selected for at either 20:14 or 21:12. The BIC selected models were mixed between M2 (only σ changes), M4 and M7. The track indicates that this final changepoint is associated with a sudden turn south, and a fairly correlated, moderately fast (around 3 km/hour) journey home.

The interim period between these two travelling bouts is marked most significantly by lower autocorrelation, essentially staying below 0.14 during the entire period and reaching near zero values during the fourth night. The mean persistence velocity and deviances, however, vary considerably in this period, and, while significant changepoints are identified at evening/dusk and dawn on all three interim nights, the behavioral pattern at these times varies somewhat. Before Night 2 at 21:15 there is a significant increase in both μ and σ (11 of 15 MLCP's chose M4). At 00:20 there is a notable drop in ρ only (4 of 4 MLCP's chose M3), and at 4:18 there is a significant drop in μ only (11 of 11 MLCP's chose M1). The μ during the evening is fairly high, at 2.5-3.0 km/h. During the day the mean velocity plummets to 0-0.5 km/hour, while the deviance gradually increases until the evening before Night 3. Night 3 again shows an increase in μ and a decrease in σ, but to a lesser extent that Night 2, and again a drop in velocity during the day. Finally, Night 4 is marked by the lowest autocorrelations ($< .01$), low velocities (< 0.5 km/hour) but a moderate deviance (around $0.5 - 1.0$ km/hour). The final day shows the greatest tightening of the deviance with no appreciable increase in velocity, until the aforementioned burst home at 21:00.

The analysis of turning velocity V_t for the first track proves less informative than the analysis of V_p. Of the 175 MLCP's that were identified (one for each window), 121 returned M0, the null model of no significant changes in σ or ρ, (in contrast to only 23 of 175 M0's in V_p), 51 selected M2 (change

in σ), 3 selected M3 (change in ρ), and none selected M6 (change in both). The value for ρ were all quite low, with a mean value of 0.002 and about half of the estimates equalling zero. It is notable, however, that at the very beginning of the trajectory, the ρ estimates are quite a bit higher (up to 0.04) than in the remainder of the trip, when the animal was moving at its highest velocity. The higher correlation here, and generally negative values for V_t near the beginning indicates that the animal is conducting a persistent, wide left-turning arc. The deviance remained fairly consistent throughout this trajectory, around 1 km/h, tightening most significantly to about 0.4 km/h during the intensive, low correlation, low V_p deviance fourth night of foraging.

Trip 7 displayed some features that were unique compared to the other analyzed trips. Notably, the velocities the fur seal moved at during Trip 7 were much, higher with an aggregate V_p mean of 3.8 km/h (s.e. 1.1) compared to 2.0 (s.e. 1.25) for Trip 1. Somewhat atypically, it left in the evening, moving quickly (around 5 km/h) but with low autocorrelation in its persistence velocity. It made a hairpin-like change in direction at dawn after its first night, appearing to turn back to the rookery, turns again in the afternoon to head NW. During Night 2, the fur seal exhibits considerable slowing, low autocorrelations and loops on its own track several times, appears to head out further to sea second time, spends a night moving relatively fastly and with relatively high auto-correlations, and at dawn after Night 3 performs one last hairpin turn and heads back at a very high pace, raching the highest estimated persistence velocity of all trips at 6.4 km/h, likewise attaining the highest autocorrelations of any trips (0.73). It appears to be missing the rookery on its way home by several kilometers, but adjusts near the end. This return trip is likewise marked by some of the highest auto-correlations in the V_t timeseries (up to 0.11), once again indicating a large and consistent turning radius of movement

A comparison of all the aggregated parameter outputs of the model for all of the trips are presented in figure 9. Violinplots are used in order to highlight the multimodal nature of many of the parameters.

7 Discussion

The method presented here displays the ability to robustly identify complex behavioral patterns in irregurlarly measured movement data. The maximum likelihood method is sensitive enough to identify changepoints even in very gappy or noisy data, and the BIC model selection method is conservative enough to protect against excessive complexity. The data considered in the application here is highly accurate positionally, but marked with a "typical" level of gappiness for a marine species. In general, however, the method is applicable to many kinds of data, including ARGOS, archival tags or radio telemetry data. Furthermore, the algorithms are computationally quite tractable. While the output is somewhat sensitive to the analysis window size, this variable is easily tuned according to the temporal resolution of the dataset and the consistency of the model

output. Because the method is focused on identifying the location and direction of structural shifts, it is also robust to error, though this aspect is not explored in this paper.

An important innovation in the method presented here is the explicit and tractable analysis of the autocorrelation. This relies an analysis of persistence and turning transformations, which, while not as biologically intuitive as raw speed and turning angle data, have attractive statistical properties that allow for an explicit description of the autocorrelation structure. This autocorrelation is readily interpreted in terms of time-scales at which an organism's basic movement patterns (persistence and turning) change. For example, the most significant distinction between the traveling modes and non-traveling behavioral modes in the fur seal tracks was associated with a change in the auto-correlation. Often, this is an even more significant variable than mean velocities. The fur seal's return trips do not necessarily attain appreciably higher velocities than those during the "foraging" phase (see for example Trip 1, Night 2). Thus, the fur seal can have a high rate of persistence displacement in a given direction but have a much more erratic movement pattern, puctuated by feeding bouts, dives, turns and loops, whereas the steadiness of a directed movement is captured by a higher persistence velocity autocorrelation. A useful index to consider is the time-interval at which the first order correlation drops to some fixed number such as 0.5. Using the relationship $\tau_h = \log(0.5)/\log(\rho)$, we obtain an estimated "time to half-autocorrelation" ($\hat{\tau}_h$). For trip 1, this value is around ranges between 30 and 40 minutes for the travelling modes, and between 4 and 15 minutes during the remainder of the trip. Similar patterns, broadly, hold for the remaining trips.

It is worth making a distinction between the information contained in the autocorrelations of persistence and turning and the parameters of standard CRWs. Although this commonly applied mechanistic model of movement is termed the "correlated random walk", in fact almost every application of the model assumes independent turning angles and independent velocities and the smoothness of a CRW path is a result of the clustering of the turning angles around zero degrees. In fact, there is a fundamental distinction between independent turning angles and correlated turning angles, which are manifested in larger turning radii. This is well-captured by the auto-correlation coefficient on the turning component of velocity, as, for example, near the beginning or Trip 1 and the end of Trip 7, where the significant rise of the turning velocity auto-correlation is associated with a large scale arcs, as contrasted with the smaller, independent directions shifts that dominate the bulk of the movements. Similarly, any autocorrelation in the velocities is largely accounted for in the persistence velocity correlation.

Returning to Trip 1, the analysis of the movement data over the three nights of foraging indicate a higher behavioral complexity that might have been expected. The primary prey items for are cephalopods and smaller pelagic fish that engage in vertical diel migrations, particularly during the crepuscular hours. It is known that the majority of feeding occurs at night (Goebel et al. 1991.) and the dive data associated with NFS07-03 corrobarates this expectation, with most dives beginning at around 23:00 and continuing through the early morning (Andrews *pers. comm.*). During the day,

northern fur seals are known to sleep at sea (Gentry 1998). Thus, the movement during daylight hours, which tended to be the least autocorrelated in Trip 1 for the first three daylight periods is more likely to reflect a mixture of drifting with ambient currents and periods of directed swimming. The distinct changes identified before each of the evening bouts correspond to increased activity and changes in absolute orientation, implying an active pursuit of prey. Interestingly, the mean velocities were quite high, particularly during the first evening, implying perhaps that the prey patch itself was moving, or that the animal is actively seeking prey over a larger area. Again, the autocorrelations are relatively low throughout, indicating irregular movements and many changes. The behavior during the third night of foraging reflects the closest to the expectation of what active, successful foraging might look like, with the lower variance and the lowest autocorrelations. However, even here there is a distict, oriented, eastward movement.

The other important feature of the method presented here is the ability of the analysis not only to estimate values of parameters, but to identify significant shifts, such that the model output provides both gradual changes in the parameter values and more discrete differences that are presumably assiciated with behavioral choices. The analysis of these behavioral shifts provides a window into the complexity of an animal's movements.

The violinplots of several of the aggregated parameter values do indicate that types of movement are often clustered into several groups. Notably, the estimates for the mean persistence velocity for Trips 2, 3 and 4 indicate a clustering of low velocities (around 2 km/h) and a seperate clustering of high velocities (around 5 km/h). Generally, the higher velocities appear less often at night. This pattern is mirrored to a smaller extent in both the persistence and turning autocorrelation coefficients. However, the behavioral trace in figure 8 suggests a striking complexity in behavioral modes, especially as contrasted with the expectation of a few distinct behaviors. While a marine organism is free of many of the landscape-related obstacles that constraint terrestrial organisms, its movement, even at rest, can be confounded by ambient currents. However, the sudden and dramatic shifts in the parameter space trace strongly suggest that these signals are reflections of real behavioral choices.

Both a consideration of autocorrelations and the basic structure of behavioral shift models can be productively implemented on datasets which lack the unavoidable gappiness of marine organism tracking data.

8 Conclusions

A top predator like the fur seal is extremely well-adapted to exploiting the heterogeneous environment of the open ocean to fulfill its survival needs. Its movement is a complicated manifestation of an organism's internal state, access of information, physiological constraints and behavioral responses to environmental cues. The data collected on this movement is an irregular subsampling

of this structurally complex, continuously auto-correlated, multi-dimensional process. The analysis method presented here is a purely descriptive attempt to capture and distill the dynamics of the underlying process from noisy and gappy data. The output is a somewhat complex array of estimates, model selections and aggregated averages, but many of the nuanced patterns of movement behaviors emerge in a clear and tractable way.

The ultimate utility of the method, however, will be in its use for answering important questions. The greatest mystery is: How does an organism successfully exploit its environment? In order to approach this challenge, it is necessary to model whatever information about the behavior the analysis method produces against potentially informative covariates. In the study on fur seal movements in the Kuril Islands, for example, detailed data has been collected in parallel on diving events, ambient temperatures and foraging success via stomach temperature sensors. An analysis of these data against the movement model can yield higher level insights into the relationships between the estimated movement parameters and foraging strategies and success. On a larger scale, comparisons can be made between animals in different locations throughout the range, under varying environmental conditions, as well as between different groups (older and younger, male or female).

Both a consideration of autocorrelations and the basic structure of a behavioral shift models can be productively implemented on datasets which lack the unavoidable gappiness of marine organism tracking data. Indeed, the output of the analysis method as applied to terrestrial organisms can be modeled as a response landscape features which are more readily measurable on land than at sea.

9 Acknowledgement

Thanks to E. Skewgar and R. Andrews for the penguin and fur seal data and fruitful collaboration. Discussions with H. Nesse, T. Gneiting and V. Minin provided invaluable insights.

References

G. Aarts, M. MacKenzie, B. McConnell, M. Fedak, and J. Matthiopoulos. Estimating space-use and habitat preference from wildlife telemetry data. *Ecography*, 31(1):140–160, 2008.

E. Bryant. 2D Location Accuracy Statistics for Fastloc®Cores Running Firmware Versions 2.2 & 2.3, Technical Report TR01. Technical report, Wildtrack Telemetry Systems Ltd., September 2007.

S. Chib. Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86:221–241, 1998.

N.I. Fisher and A.J. Lee. Time series analysis of circular data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):327–339, 1994.

J. D. Forester, A. R. Ives, M. G. Turner, D. P. Anderson, D. Fortin, H. L. Beyer, D. W. Smith, , and M. S. Boyce. State-space models link elk movement patterns to landscape characteristics in yellowstone national park. *Ecological Monographs*, 77:285–299, 2007.

R. Gentry. *Behavior and Ecology of the Northern Fur Seal*. Princeton University Press, 1998.

J.L. Goebel, M.E.and Bengtson, R.L. DeLong, R.L. Gentry, and T.R. Loughlin. Diving patterns and foraging locations of female northern fur seals. *Fisheries Bulletin*, 89:171– 179, 1991.

D. Grünbaum. Advection-diffusion equations for internal state-mediated random walks. *SIAM (Society for Industrial and Applied Mathematics) Journal of Applied Mathematics*, 61:43–73, 2000.

D. T. Haydon, J. M. Morales, A. Yott, D. A. Jenkins, R. Rosatte, and J. M. Fryxell. Socially informed random walks: incorporating group dynamics into models of population spread and growth. *Proceedings of the Royal Society B-Biological Sciences*, 275(1638):1101–1109, 2008.

D. S. Johnson, J. M. London, M.-A. Lea, and J. W. Durban. Continuous-time correlated random walk models for animal movement data. in press.

I. D. Jonsen, R. A. Myers, and J. M. Flemming. Meta-analysis of Animal Movement Using State-Space Models. *Wildlife Research*, 21:149–161, 2003.

I. D. Jonsen, R. A. Myers, and J. M. Flemming. Robust state-space modeling of animal movement data. *Ecology*, 86:2874–2880, 2005.

P. M. Kareiva and N. Shigesada. Analyzing insect movement as a correlated random walk. *Oecologia*, 56:234–238, 1982.

T. R. Loughlin, D. J. Rugh, and C. H. Fiscus. Northern sea lion distribution and abundance: 1956-80. *The Journal of Wildlife Management,*, 48(3):729–740, 1984.

L.M. Marsh and R.E. Jones. The form and consequence of random walk movement models. *Journal of Theoretical Biology*, 133:113–131, 1988.

J. M Morales, D. T. Haydon, J. Frair, K. E. Holsinger, and J. M. Fryxell. Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology*, 85(9): 2436–2445, 2004.

A. Okubo and S. Levin. *Diffusion and Ecological Problems: Modern Perspectives*. Springer Verlag, New York, 2001.

T. A. Patterson, L. Thomas, C. Wilcox, O. Ovaskainen, and J. Matthiopoulos. State-space models of individual animal movement. *Trends in Ecology and Evolution*, 23(2):87–94, 2008.

F. Royer, J.-M. Fromentin, and P. Gaspar. A state-space model to derive bluefin tuna movement and habitat from archival tags. *Oikos*, 109:473–484, 2005.

G.T. Skalski and J.F. Gilliam. A diffusion-based theory of organism dispersal in heterogeneous populations. *The American Naturalist*, 161(3):441–458, 2003.

J. G. Skellam. Random dispersal in theoretical populations. *Biometrika*, 38:196–218, 1951.

P. Turchin. *Quantitative Analysis of Movement: measuring and modeling population redistribution in plants and animals*. Sinauer Associates, Sunderland, MA., 1998.

A Tables

Table 1: Results of a simulation study used to explore the efficiency of the estimation routine for a single structural break in a gappy time series. Underlying time series of length 400 were generated with a breakpoint at $T_{br} = 200$ for several models, ranging from a null model of no change in μ, σ and ρ, to a model where all three parameters change value. The parameters used in the simulation are tabulated in 1A, with those parameters which change at the breakpoint are boldfaced. From each timeseries, 50 datapoints were randomly selected to create "gappiness" and estimates were obtained using the estimation routine described in the text. Simulations were performed 100 times for each parameter set, and the resulting mean estimates and standard errors are reported in 1B. Essentailyl all estimates are within one standard error of the true value. The last two columns (N_A and N_B) are the number of times that AIC and BIC respectively select the correct model out of 100 attempts (see Figure 5).

A) True parameter values

	μ_1	μ_2	σ_1	σ_2	ρ_1	ρ_2
S0	0	0	1	1	0.5	0.5
S1	**-1**	**1**	1	1	0.5	0.5
S2	0	0	**0.5**	**2**	0.5	0.5
S3	0	0	1	1	**0.2**	**0.9**
S4	**-1**	**1**	**0.5**	**2**	0.5	0.5
S5	**-1**	**1**	1	1	**0.2**	**0.9**
S6	0	0	**0.5**	**2**	**0.2**	**0.9**
S7	**-1**	**1**	**0.5**	**2**	**0.2**	**0.9**

B) Estimates

	T_{br}	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\rho}_1$	$\hat{\rho}_2$	N_A	N_B
S0	206 (113)	0.0 (0.21)	0.0 (0.17)	1.0 (0.11)	1.0 (0.11)	0.48 (0.16)	0.47 (0.15)	0	78
S1	197 (13.2)	-1.0 (0.18)	1.0 (0.19)	1.2 (0.21)	1.2 (0.22)	0.59 (0.25)	0.63 (0.19)	1	84
S2	203 (10.3)	-0.03 (0.12)	-0.04 (0.27)	0.50 (0.04)	2.0 (0.19)	0.46 (0.16)	0.47 (0.16)	28	72
S3	201 (26.1)	-0.02 (0.22)	0.04 (0.43)	1.0 (0.08)	1.0 (0.12)	0.23 (0.21)	0.92 (0.056)	29	92
S4	199 (5.49)	-1.0 (0.08)	1.1 (0.33)	0.5 (0.04)	2.0 (0.17)	0.47 (0.2)	0.53 (0.17)	22	40
S5	199 (12.0)	-0.96 (0.17)	1.0 (0.51)	1.0 (0.12)	1.0 (0.14)	0.19 (0.18)	0.92 (0.033)	7	15
S6	200 (8.51)	0.0 (0.09)	0.02 (0.93)	0.49 (0.05)	2.0 (0.17)	0.20 (0.19)	0.92 (0.036)	22	40
S7	201 (15.6)	-0.99 (0.08)	0.94 (0.87)	0.50 (0.05)	2.0 (0.18)	0.19 (0.18)	0.93 (0.033)	97	97

B Figures

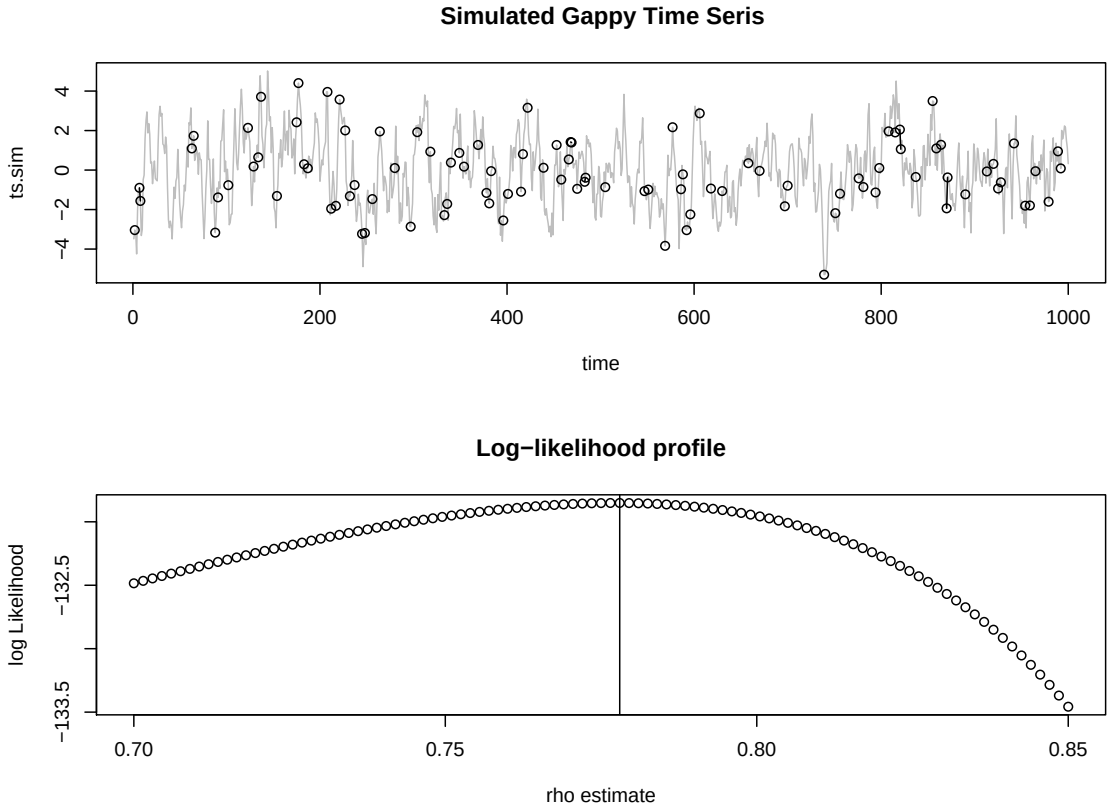

Figure 1: A simulated full AR(1) time series with auto-correlation coefficient $\rho = 0.8$ in plotted in grey on the top graph. The dots indicate a sampling of 100 points from this time-series. The lower plot shows is the log-likelihood profile over ρ for the sampled data. The estimated value for ρ in this simulation is 0.778

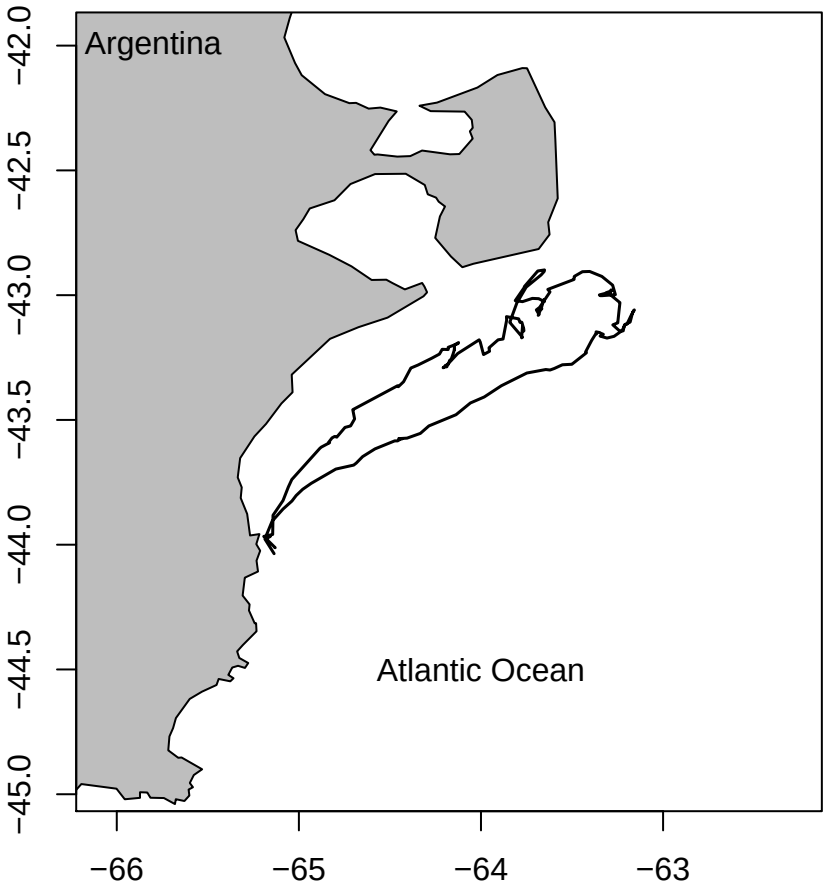

Figure 2: Penguin track

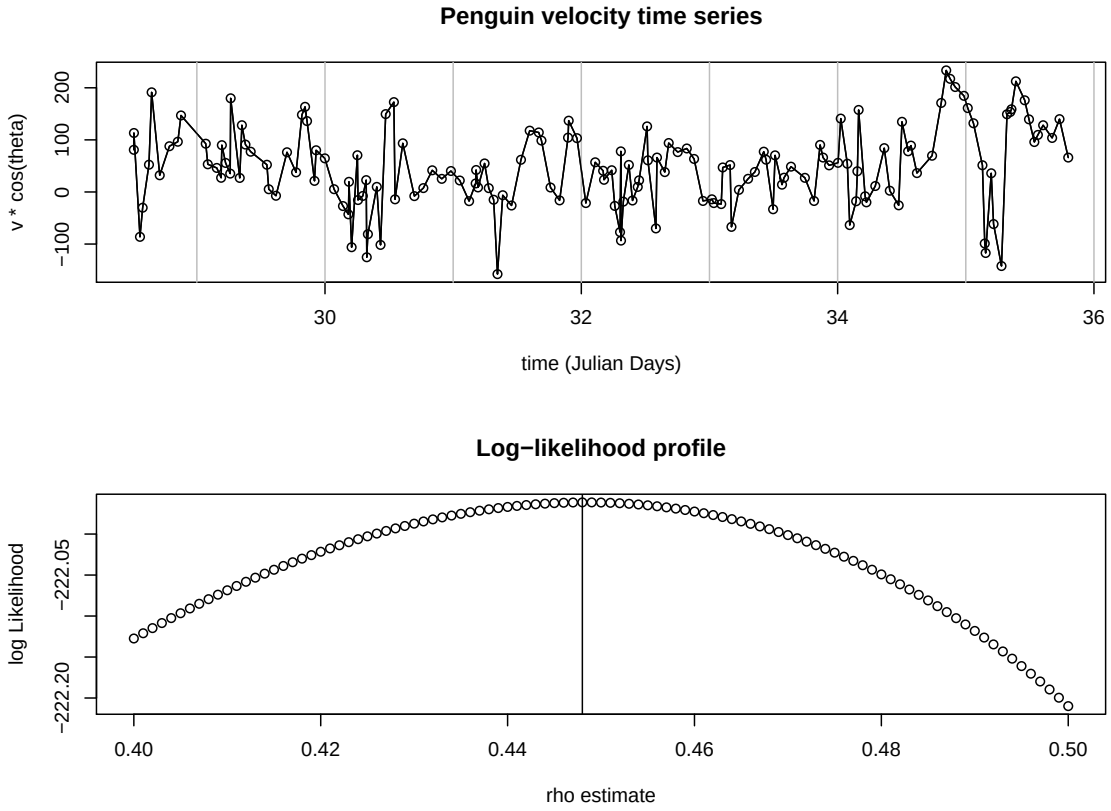

Figure 3: The estimation procedure for ρ is performed on $v\cos(\theta)$ time series (upper plot) from the track of penguin P1. The log-likelihood profile is plotted in the lower plot. The estimate $\widehat{\rho}$ is 0.448

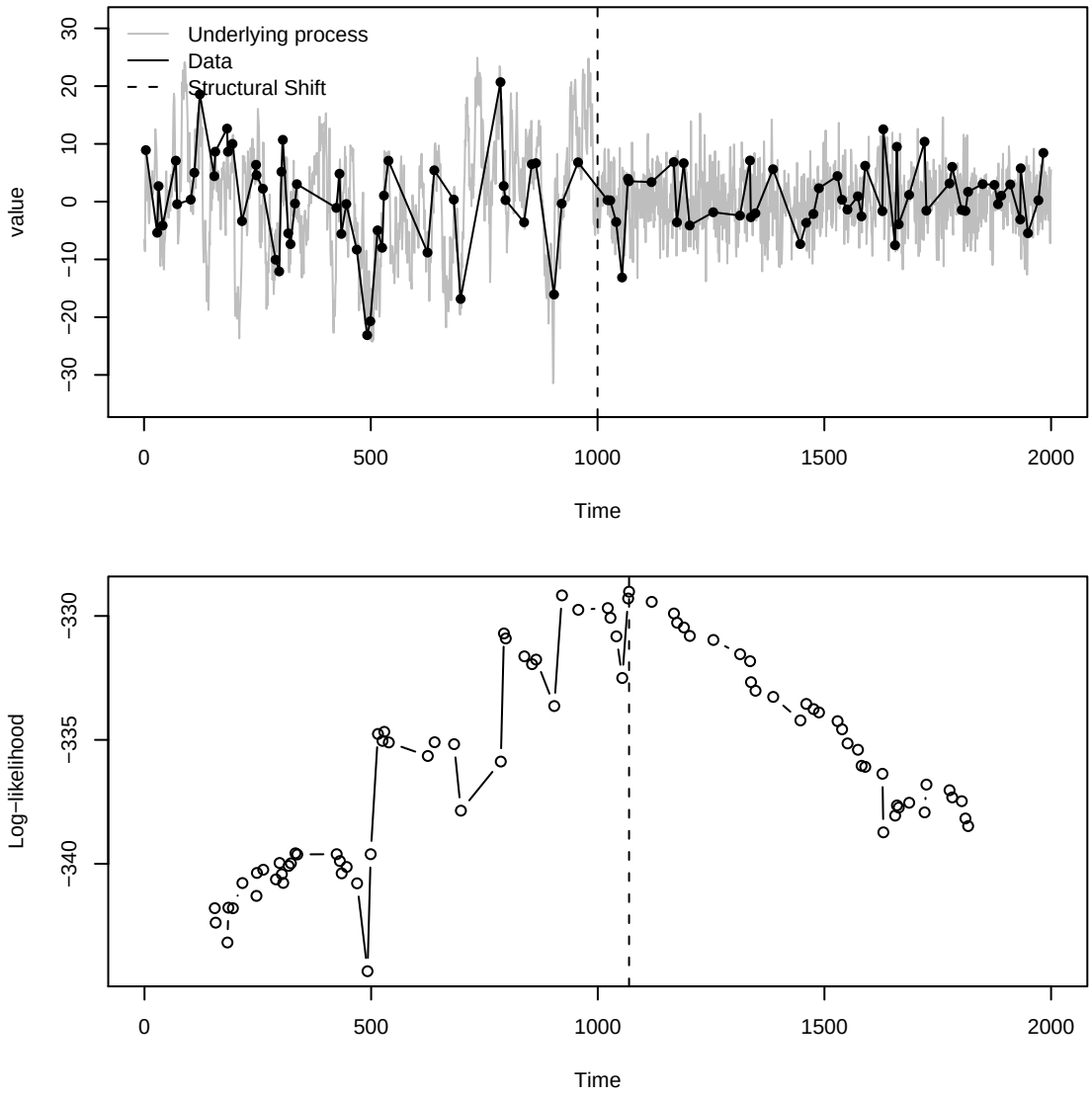

Figure 4: Simulation and estimation of gappy time-series with a single structural shift generated by sampling 100 values from a time-series of length 2000 with a breakpoint at $T_{br} = 1000$ represented by the vertical dotted line in the top graph. The true parameter values are $\mu_1 = 5$, $\mu_2 = 2$, $\sigma_1 = 10$ $\sigma_2 = 5$, $\rho_1 = 0.9$ and $\rho_2 = 0.2$. In this simulation, the resulting parameter estimates were: $\widehat{\mu}_1 = 1.6$, $\widehat{\mu}_2 = 0.24$, $\widehat{\sigma}_1 = 8.75$ $\widehat{\sigma}_2 = 4.73$, $\widehat{\rho}_1 = 0.91$, $\widehat{\rho}_1 = 0.35$, while the MLBP is at $t = 1060$.

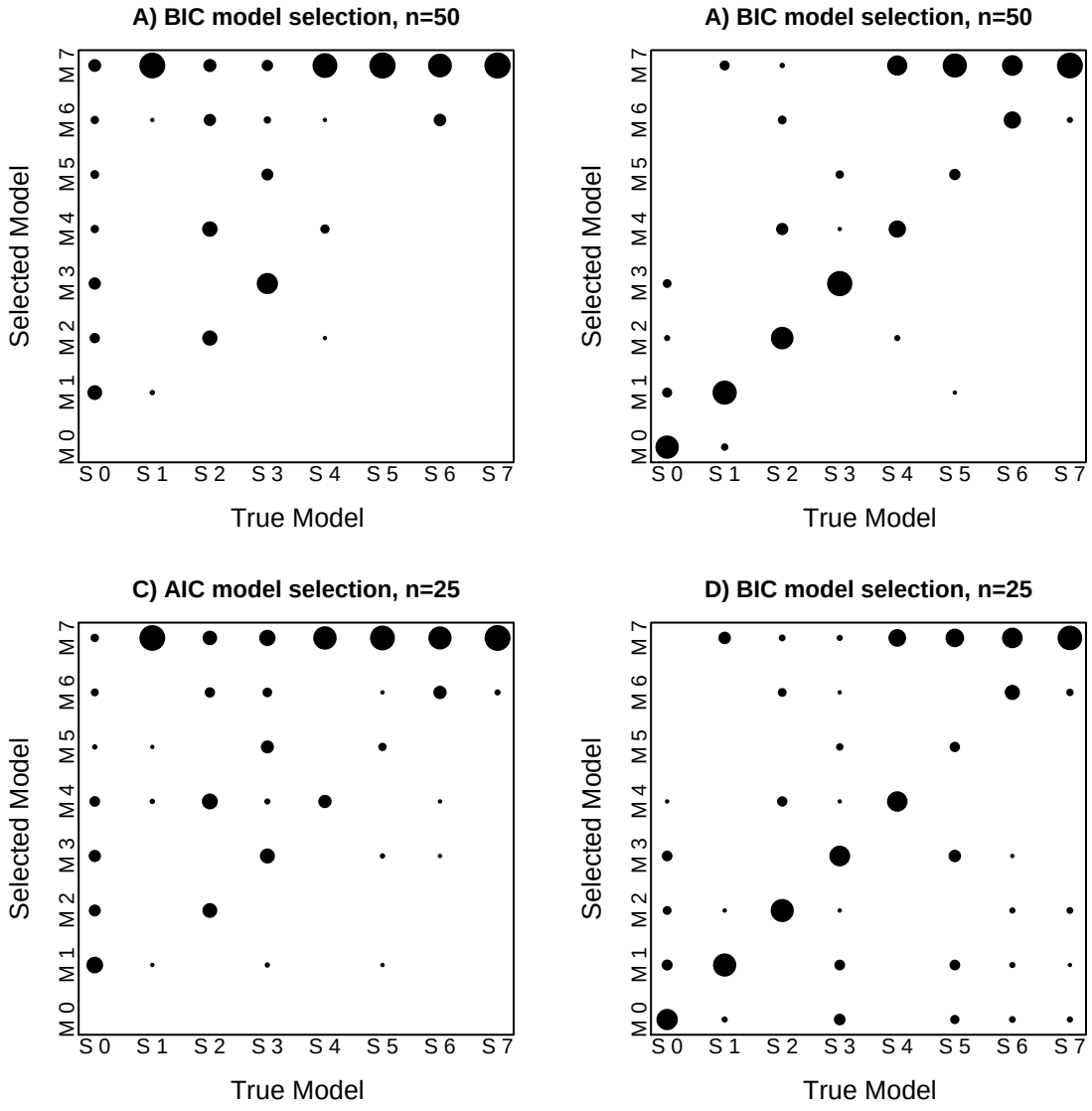

Figure 5: Results of model selection simulation using both AIC (A and C) and BIC (B and D) for breakpoints at $n = 25$ from a gappy time series of length $N = 50$ subsampled from a true process of length $T = 200$ (A and B) and $n = 50$, $N = 100$ and $T = 400$ (C and D). The true models are parameterized as in table 1, ranging from the null model (S0) of no difference before and after the gap to the extreme model (S7) with differences in all three parameters. 100 gappy time series were performed for each parameter set. The area of the circles is proportional to the number of time a model is selected for a given parameter set.

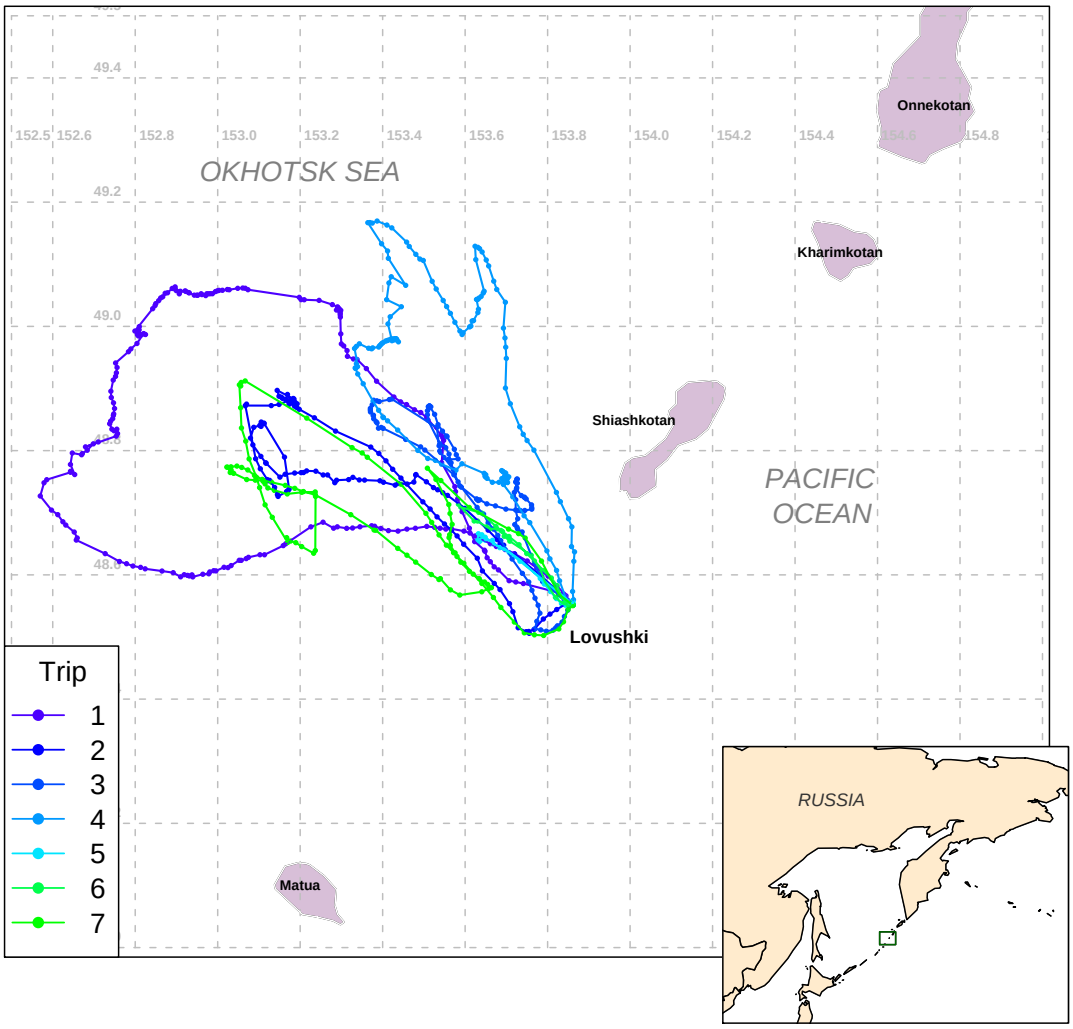

Figure 6: Map of all trips taken by NFS07-03 in Summer 2007. The square in the inset map indicates the study area.

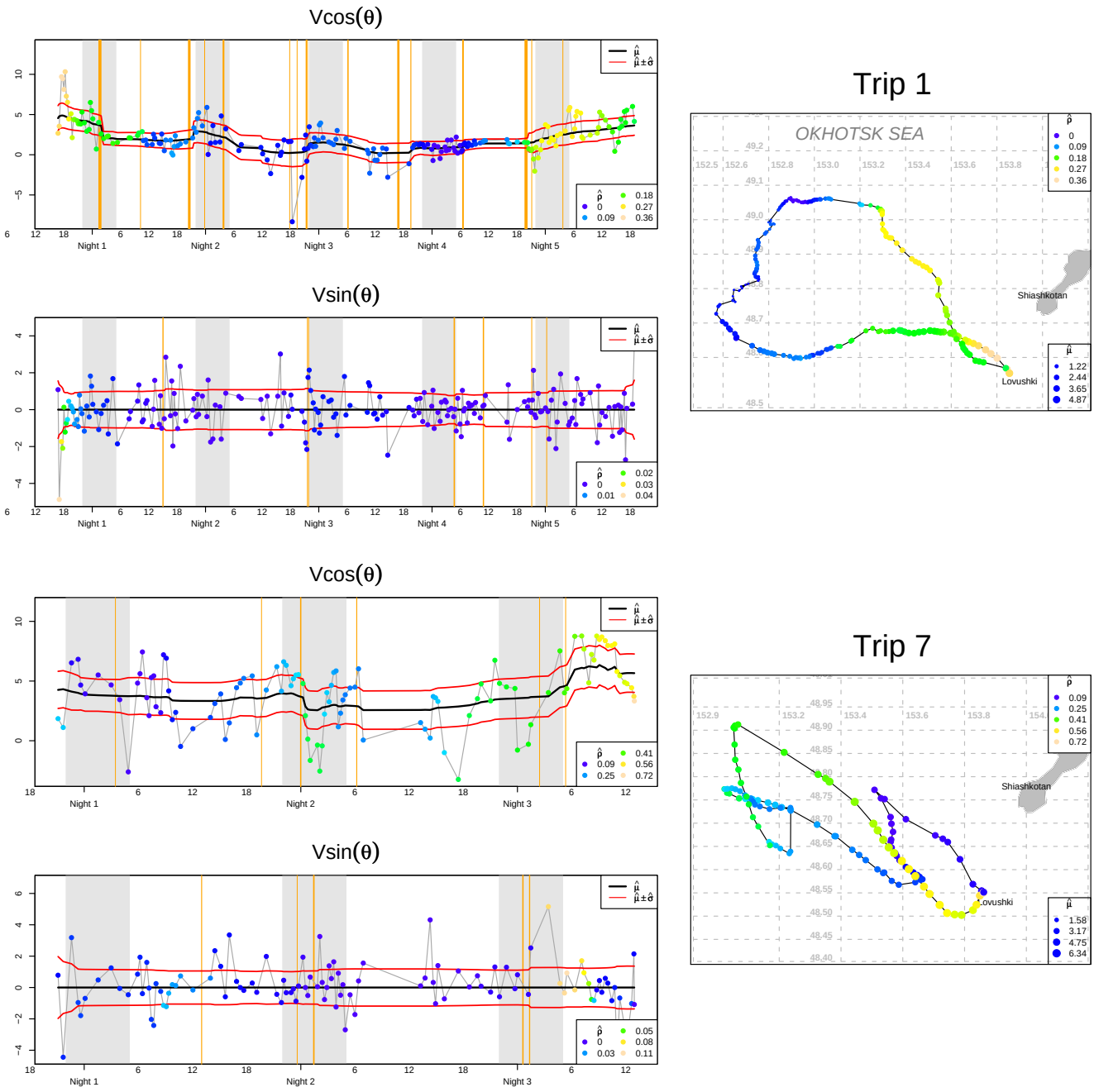

Figure 7: Plots of two NFS07-03 foraging trips. The plots on the left show the time-series of velocity and turning angles decomposed into persistence ($V_p = V\cos(\theta)$) and turning ($V_t = V\sin(\theta)$) components. The black line is the estimate for the mean $\widehat{\mu}$, the red line represents the estimate for the standard deviaion $\widehat{\sigma}$ around the mean, and the color reflects the auto-correlation $\widehat{\rho}$: bluer colors indicate less auto-correlated movement, while yellow colors indicate more auto-correlated movement. Vertical orange lines indicate changepoints, where the estimation routine suggests there was a significant behavioral shift; thicker lines correspond to a higher number of selected shifts. Wide grey bands indicate nighttime. To the right is a mapping of the track itself. Colors again indicate the auto-correlation estimate for the persistence component (V_p), while the size of the dots is proportional to the estimated local mean value V_p. For both of these analyses, a window of size 30 was swept over the data.

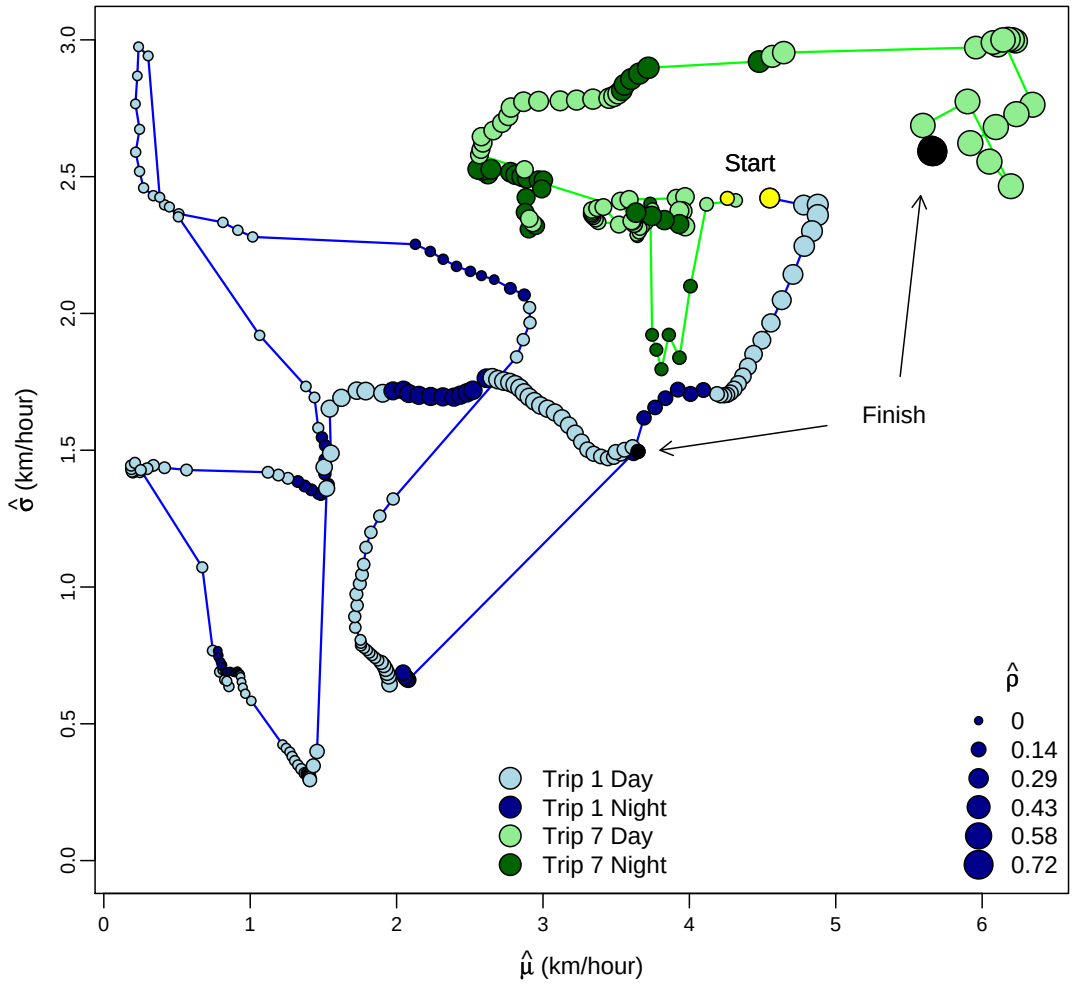

Figure 8: A mapping of all three parameters for V_p for trips number 1 (blue circles) and 7 (green circles). The size of the circles is proportional to $\hat{\rho}$. This connecting lines illustrate the progress of the walk through the parameter space, beginning at the yellow circles in the upper-right quadrant and finishing at the black dots. The long connections are associated with significant structural shifts.

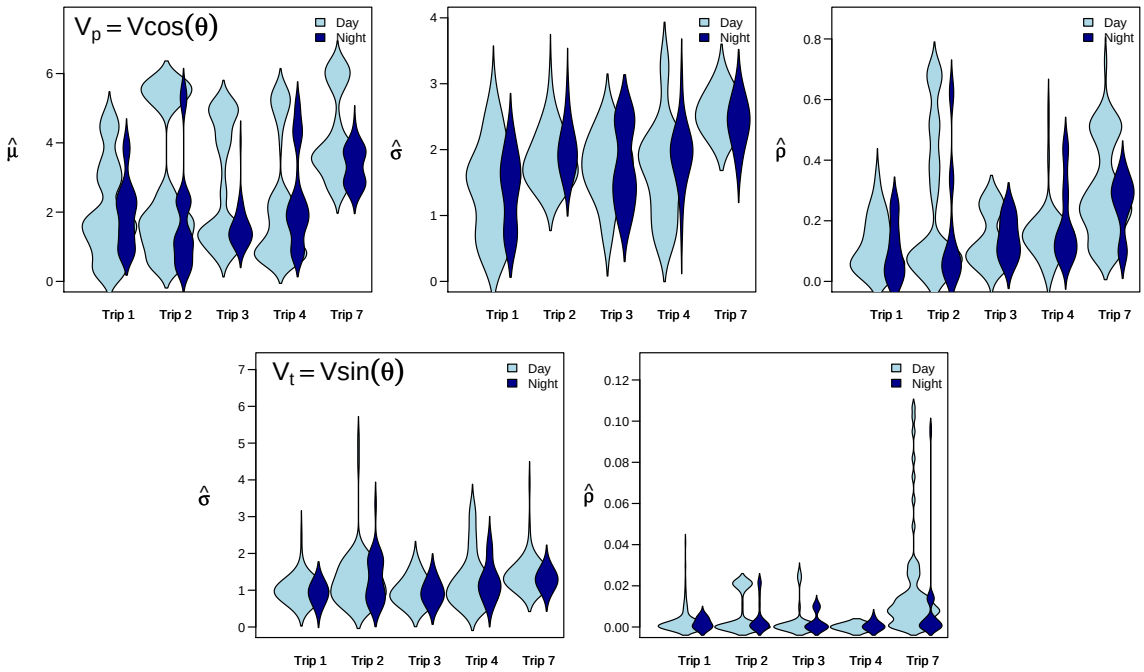

Figure 9: Violinplots of all parameter values for each of the five analyzed trips separated by daytime and nighttime estimates. The violin plot combined the information of a boxplot with a histogram by showing an estimated Gaussian kernel density of the data. It is particularly useful when the data is multimodal, as is the case for many of these estimates. The area of the "violins".

C Code

C.1 Estimating continuous correlation coefficient ρ.

```
GetRho <- function(x,t)
# This function returns the MLE estimate of rho for a gappy time-series
{
  # Negative log-Likelihood function
getL <- function(rho)
{
dt <- diff(t)
s <-sd(x)
mu <- mean(x)

n<-length(x)

x.plus <- x[-1]
x.minus <- x[-length(x)]

Likelihood <- dnorm(x.plus,mean=mu+(rho^dt)*(x.minus-mu),sd=s*sqrt(1-rho^(2*dt)))
logL <- sum(log(Likelihood))

return(-logL)
}

# Optimisation routine
o<-optim(0.5,getL,method="L-BFGS-B",lower=0,upper=0.999))
return(data.frame(rho=o$par,LL=-o$value))
}
```

C.2 Obtaining log-likelihood of single structural breakpoint τ.

```
  GetDoubleL <- function(x,t,tbreak){
  # This function obtains estimates for mu sigma and rho
  # before and after a given break

    x1<-x[1:tbreak]
    t1<-t[1:tbreak]
    x2<-x[(tbreak+1):length(x)]
    t2<-t[(tbreak+1):length(t)]

    o1<-GetRho(x1,t1)
    o2<-GetRho(x2,t2)
```

```
mu1 <- mean(x1)
sigma1 <- sd(x1)
rho1 <- o1$rho

mu2 <- mean(x2)
sigma2 <- sd(x2)
rho2 <- o2$rho

LL1 <- o1$LL
LL2 <- o2$LL

return(data.frame(mu=c(mu1,mu2),sigma=c(sigma1,sigma2),
 rho=c(rho1,rho2),LL=c(LL1,LL2)))}
```