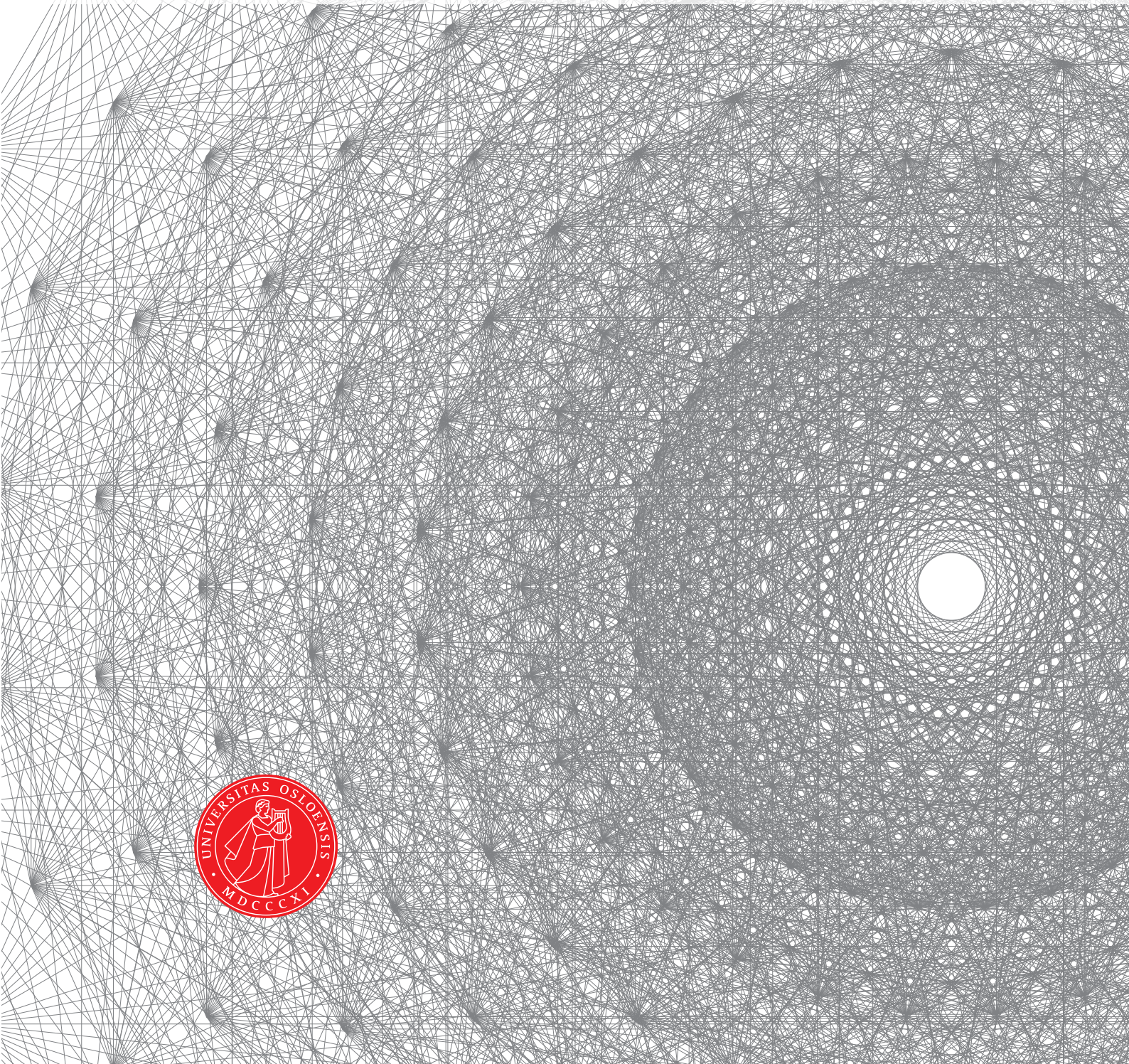
UiO : Department of Mathematics
University of Oslo

Estimating animal movements from light-logger data using Bayesian state-space modeling

Maunya Doroudi Moghadam

Master's Thesis, Spring 2016



Cover design by Martin Helsø

The front page depicts a section of the root system of the exceptional Lie group E_8, projected into the plane. Lie groups were invented by the Norwegian mathematician Sophus Lie (1842–1899) to express symmetries in differential equations and today they play a central role in various parts of mathematics.

Abstract

Estimating the movements of animals and obtaining information about their behaviors are fundamental subjects in many ecological studies because many ecological processes are related to movement. Movements and locations during migration are often not directly observable and must be inferred from indirect data and statistical analysis methods. The use of state-space modeling approaches leads to inference of hidden (unknown) locations and movements. In my Master thesis I will use bio-physical features corresponding to light levels. As the light levels and day lengths, based on sunrise and sunset, change in different sites and different times, they are suitable indicators to find locations. Light-level geolocation is one of the currently most used methods in these types of studies. In this thesis I developed an state-space model to estimate unknown locations of a bird by the use of data from light loggers attached to migrating birds.

Acknowledgement

I would like to thank the Department of Mathematics and Department of Biosciences at the University of Oslo, for all of the opportunities I was given as an international student to carry out this master thesis.

I would like to express my gratitude to my supervisors, Geir Storvik, Torbjørn Ergon and Morten Helberg for their motivations, patience and all academical helps.

I would like to thank my mother, Minoo Asadi Khomami, my father, Shaahin Doroudi Moghadam and my especial husband, Mohammad Mehdi Shahri, for their dedications and wise counsels.

Maunya Doroudi Moghadam

Oslo, Spring 2016

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Specification

Estimating the movements of animals and obtaining information about their locations and behaviour are fundamental subjects in many ecological studies because many ecological processes are related to movement. There are two prevalent methods to record data about locations which can be generally grouped as *remote* and *archival*. In the *remote* methods, techniques such as radio and satellite telemetry are used to locate the tags which are attached to animals. The *Archival* methods use special tags to record properties of the animal's environment over time such as the *light intensity* and the *water temperature*. As the light intensity levels and the day lengths, based on the sunrise and sunset, change at different times and rate in different locations, they are suitable indicators to find locations. Analyzing the *light intensity levels* which are recorded by a data-logging device to estimate the geographical locations is called *light-level geolocation*.

Many environmental factors exist, such as weather condition, shading from vegetation and animals' behaviors and routines such as nesting, influence the natural light intensity and causes many noises on recorded light intensities. Although high level of precision and accuracy for *light-level geolocation* method are not always guaranteed, archival tags are used for variety of animal species because these tags have considerably lower weights and are more affordable.

As the movements and the behavioral states of animals are not always directly observ-

able, they must be inferred from indirect data by use of statistical analysis methods. In addition, as mentioned, the indirect data such as light intensity data is not always pure and clean data to use, therefore some sophisticated methods are needed to cope with unexpected noises.

1.2 Approaches

The most frequently used method to determine light-level geolocation is named the *Threshold Method*. As mentioned before, the times of sunrise and sunset vary in different locations and different days of year, therefore a table of sunrise and sunset times will help us to distinguish the locations on the Earth. The times of sunrise and sunset can be specified from the times that the light intensity passes a certain threshold. Although it seems that the general principle and structure are simple, it is not very easy and straightforward to do the analysis in high accuracy.

There is an R package called *Geolight* which has been developed by *Simon Lisovski* and *Steffen Hahn* to analyze light intensity data based on the Threshold Method [5]. This package includes basic functions that use light intensity measurements over time (typically several times per hour) to calculate sunrise and sunset times of each day which are then used to calculate locations. Based on experience, sometimes the relevant function gives more than one pair of sunrise and sunset in one day which are unrealistic and affect the result. The package also includes a distance filter function that uses a maximum distance in a certain time unit to partly filter out unrealistic estimates of coordinates.

TripEstimation is the name of another package in R that is used for *light-level geolocation*. This package is developed by *M. Summer* and *S. Wotherspoon* and provides estimation algorithms and a supporting code which result in estimation of 2 fixes, i.e. one at dawn and another at dusk. However, its application is somewhat confusing since it lacks sufficient help and examples [3].

TripEstimation package includes functions that calculate elevation of astronomical objects such as sun or moon and solar position parameters. There is also a function for calibration that uses a set of light intensity data from a known location and given solar elevation to return the expected light intensity level. By this package we can create a solar model object by mainly using a vector for identifying twilight segment and vectors of *light intensity* and *time*, etc, and then there is a function that uses the Bayesian analysis and Markov Chain Monte Carlo method (MCMC) to provide a direct implementation of the *Metropolis algorithm* to calculate marginal posterior of locations and full-track estimates.

One of the latest study in *light-level geolocation* analysis is provided in a paper by Eldar Rakhimberdiev et al. [6]. In this paper they develop a hidden Markov chain model for analysis of geolocator data and estimates tracks for animals with complex migratory behaviour by combining (1) a shading-insensitive, template-fit physical model, (2) an uncorrelated random walk movement model that includes migratory and sedentary behavioural states, and (3) spatially explicit behavioural masks. They implement their model in an R package named *FlightR*.

The approach that I mainly utilized and developed in this project is based on *Bayesian State-Space Modeling* which includes a *Process Model* and an *Observation Model*. In this method of the modeling I considered the *location* as a time dependent variable. The time-step in this project is *day* and each location at day d can be estimated by the location at day $d-1$ via a dynamic model named *Process Model*. Then I constructed a model called the *Observation model* to connect the *location* variable which is unknown to the *light intensity* variable which is observed during the time. To make inference and estimate the parameters of the models and unknown locations I applied a Bayesian framework and used Markov Chain Monte Carlo methods via a statistical software named *JAGS* and some packages in *R* named *R2jags*, and *dclone* and *coda*. The main reason that I used the program JAGS, is transparency and flexibility of implementing models in this kind of programs. Details and structures about the modeling and methods of inference are represented in chapter 3 and chapter 4.

In chapter 2, I explain the available data for this project. In chapter 5, I provide es-

timates of the model parameters and also unknown locations, by mapping the estimated tracks and other relevant plots, for a specific bird in a period of time. In addition I compare the results from our state-space model with the results obtained from the GeoLight package for the same data. Then we give a conclusion and discussion in the last chapter.

Chapter 2

Data

In this chapter I describe the data that were available for this project. The data have been collected by Morten Helberg, who was supported by the Norwegian Research Counsel grant "Animal Movements". He used *Biotrack Geolocator model MK15* which is a device weighted 2.5 gr and has a light sensor. The device can be deployed on birds to record every 10 minutes of environment's *light intensity* with exact date and time for up to five years (depends on battery life and memory capacity). The records can be transferred to the computer software named *transEdit* to get the information and see them visually. The figure 2.1 is a sample of displaying recorded light intensities via *transEdit*.

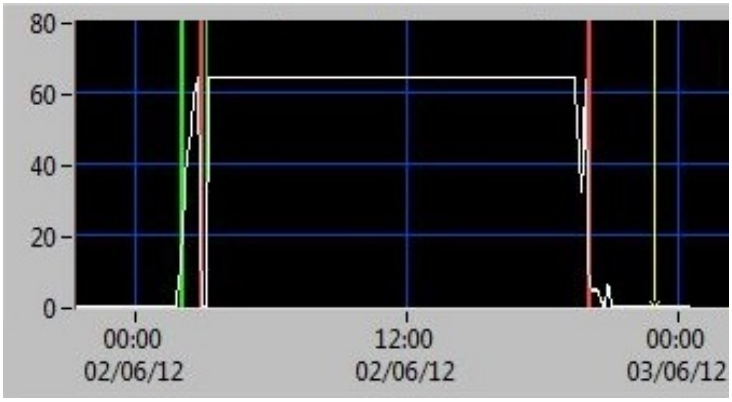


Figure 2.1: *The curve of recorded light intensities in 24 hours*

The values of recorded *light intensities* vary from 0 (completely darkness) to 64 (maximum lightness) but due to several environmental effects or birds' behaviours mentioned before, the recorded *light intensities* are not always vary in a natural way, therefore it is possible to have unexpected lack of light during a midday or artificial light at night. We

can see the curves of recorded *light intensities* with occurred noises in the Figure 2.2.

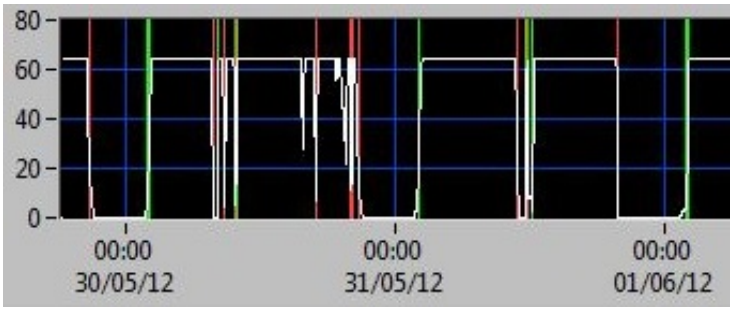


Figure 2.2: *The curves of recorded light intensities for the days include noises*

when we have data which has curves like Figure 2.1, estimating sunrise and sunset are relatively easy by using a threshold method, but analysis of data with curves such as Figure 2.2 is much more difficult, and more sophisticated statistical modeling is needed.

I chose a data set for one Lesser black-backed gull (Larus fuscus, (Figure 2.3)) that contained records from 24/05/2011 to 04/06/2014 and I had normally 144 records for each day. Then I selected the data for 181 days of that time period, from 30/05/2012 to 27/11/2012 which included an annual migration from breeding area to the wintering area. However this bird is still followed.

During the study period sometimes the bird which is identified by its field readable ring code, were sighted and the corresponding dates and locations (by *latitude* and *longitude*) were recorded. I also used this known locations data for some parts of our analysis. In addition I had information such as times and locations of deploying the loggers on the birds, breeding and nesting times and locations, etc that were useful to become more familiar with the bird, but I didn't use them all directly in the analysis. The information about recorded locations and some other technical issues, birds' behaviours and futures are reported in the website http://www.ringmerking.no regularly.

Figure 2.3: *The picture of the study bird, Lesser black-backed gull (Larus fuscus)*

Chapter 3

State-Space Modeling

3.1 Model description

In this section I briefly introduce ***State-Space Models (SSM)*** and then describe the models that we apply in this project.

State-Space Model (SSM) is a class of time-series models that predicts the future state of a system from its previous states formed by coupled stochastic models, a ***Process Model*** and an ***Observation Model***.

The ***Process Model*** is a model of the dynamics of the movement through time and space [1].

The ***Observation Model*** is a model that specifies how the observed data relate to the states in the process model.

In other words, the ***Sate-Space Models*** are hierarchical models that decompose an observed time series of counts or other observed responses into a process variation and an observation error component. They are suitable for description of Markovian, that is, auto-regressive, processes that are latent or hidden, because they are observed imperfectly [4].

3.1.1 Process Model: Model of a Bird's Movement

Here I specify my process model by using geographical coordinates of birds' ***locations*** in each day and two main elements of birds' movement which are ***movement direction*** and ***movement distance***. Commonly, the geographical coordinates of locations are identified by ***latitude*** and ***longitude***.

The ***latitude*** refers to the imaginary circles drawn parallel to the Equator that specifies the North-South position of a location on the Earth. The Equator represents 0 degrees latitude, while the North and South Poles represent 90 degrees North (+90) and 90 degrees South (-90) respectively.

The ***longitude*** refers to the imaginary circles drawn vertically to the Equator that specifies the East-West position of a location. The circles are also called meridians. The prime meridian is assigned the value of 0 degree, and runs through Greenwich, England. The valid range of longitude in degrees, is -180 and +180 for the western and eastern hemisphere respectively.

Now I consider $\boldsymbol{lat_d}$ and $\boldsymbol{lon_d}$ which refer to the *latitude* and the *longitude* respectively on *day* $\boldsymbol{d}$. I assume the bird only move during the day-time which is biologically plausible for the chosen bird. Hence, d refers to success intervals from one midday to the next midday (noon to noon). In fact I consider one geographical location per day and define the process model as below

$$lat_d = lat_{d-1} + cos(\theta_d)dis_d \tag{3.1}$$

$$lon_d = lon_{d-1} + \frac{sin(\theta_d)dis_d}{cos(\frac{\pi}{180}lat_d)} \tag{3.2}$$

$\boldsymbol{\theta_d}$ implies the ***movement direction*** for each *day* and $\boldsymbol{dis_d}$ implies the ***movement distance*** for each *day*. I divide "$sin(\theta_d)dis_d$" by "$cos(\frac{\pi}{180}lat_d)$" in (3.2), because the Earth is spherical and the distances between vertical lines of *longitude* change from the

North to the South.

Then the ***movement distance*** variable can be decomposed as below

$$dis_d = m_d Dis_d \tag{3.3}$$

$\boldsymbol{m_d} \in \{0, 1\}$ and refers to the bird's movement as a binary variable (0 for moving and 1 for not moving) and it is modeled as a ***Markov Chain*** with a ***Transition-Probability Matrix***

$$\Psi = \begin{pmatrix} 1 - \psi_{01} & \psi_{01} \\ 1 - \psi_{11} & \psi_{11} \end{pmatrix}$$

We assume that $Prob(m_1 = 0) = 1$.

Here the ***Transition-Probability Matrix*** $\Psi = (\psi_{i,j} : i, j \in \{0, 1\})$ is a matrix which means each $\psi_{i,j} \geq 0$ and $\sum_{j \in \{0,1\}} \psi_{i,j} = 1$ and the matrix represents probabilities of transit from i to j. Hence ψ_{01} is the probability that "the bird did not move in one day, will move in the next day" and ψ_{11} is the probability that "the bird moved in one day, will also move in the next day".

The ***Markov chain*** is a chain of events during a discrete period of time that indicates tendency of an event to be followed by the one in the next step. Here it is claimed that $\boldsymbol{m_d}$ depends on $\boldsymbol{m_{d-1}}$.

$\boldsymbol{Dis_d}$ refers to the ***movement distance*** and is considered to be normally distributed with a range between 0 and 25 degree latitudes as below

$$Dis_d \sim truncNorm(\mu_0, \sigma_0, 0, 25) \tag{3.4}$$

μ_0 and σ_0 are specified in the next section (Table 3.1). $\boldsymbol{\theta_d}$ which was mentioned as the ***movement direction***, is considered to be uniformly distributed as below

$$\theta_d \sim Unif(0, 2\pi) \tag{3.5}$$

3.1.2 The Observation Model : A Model for matching recorded light intensities to the locations

In this section first I specify the general form of my observation model. The distribution of recorded ***light intensity*** ($y_{d,t}$), given the ***location*** (lat_d,lon_d), is Normal truncated to the range of 0 to 64,

$$y_{d,t} \sim truncNorm(\mu_{d,t}, \sigma_{d,t}, 0, 64) \tag{3.6}$$

For $\boldsymbol{t}$ from 1 to the number of *light intensity* records in each *day* and $\boldsymbol{d}$ from 1 to the number of *days* is the study. I consider only one location per *day* to estimate, not as many as the number of the recorded *light intensity* per day.

Now I show how $\boldsymbol{\mu_{d,t}}$ and $\boldsymbol{\sigma_{d,t}}$ are obtained; First I divide 24 hours of a day in three ***phases*** (ϕ) of ***Night***, ***Twilight*** and ***Day*** based on three intervals of the ***Sun Elevation Angles*** ($e_{d,t}$) as below (I assign 1, 2 and 3 for *Night time*, *Twilight time* and *Day time* respectively)

$$\phi_{d,t} = \begin{cases} 1 & e_{d,t} < -6 \\ 2 & -6 < e_{d,t} < -2 \\ 3 & e_{d,t} > -2 \end{cases}$$

The ***Twilight*** period refers to the time during the sunrise and sunset that the sun is not directly visible, and the *light intensity* is poor but variable.

The ***Sun Elevation Angle*** is the angular height of the sun that measured from the horizontal in degree. The elevation is 0 degree at sunrise and 90 degrees when the sun is directly overhead. This angle is calculated as a function of *latitude*, *longitude*, *date* and *time*. In addition the *light intensity* can directly depend on the *sun elevation angle* via regression model(s).

As mentioned before the range of *light intensity* in this study varies from 0 to 64. It is more likely to have a value of 0 during the *Night* and value of 64 during the *Day* but the probabilities of having artificial periods of light or darkness during the *Night* and *Day* respectively are not zero. There can also be a wide range of *light intensities* during the *Twilight*, and I can define a regression of *light intensity* on the *sun elevation angle* for these periods of the day (morning and evening). I therefore use three possible ***classes*** of light intensities in each of the three ***phases***.

I define $\boldsymbol{P_{\phi,\kappa}}$ as the relevant probability of having specific $\mu_{d,t}$ and $\sigma_{d,t}$ for the mentioned normal distribution in (3.6), at *phase* ϕ and *class* κ ($\phi = 1, 2, 3$ and $\kappa = 1, 2, 3$ and $\sum_{\kappa=1}^{3} P_{.,\kappa} = 1$). Therefore I assign $\mu_{\phi,\kappa}$ and $\sigma_{\phi,\kappa}$ to the $\mu_{d,t}$ and $\sigma_{d,t}$ respectively and they can be obtained from the following table.

	$class1$	$class2$	$class3$
$Night$	$P_{1,1}$ $\mu_{1,1} = 0$ $\sigma_{1,1} = 10^{-6}$	$P_{1,2}$ $\mu_{1,2} = \alpha_1 + \delta_d$ $\sigma_{1,2}$	$P_{1,3}$ $\mu_{1,3} = 64$ $\sigma_{1,3} = 10^{-6}$
$Twilight$	$P_{2,1}$ $\mu_{2,1} = 0$ $\sigma_{2,1} = 10^{-6}$	$P_{2,2}$ $\mu_{2,2} = \alpha_2 + \beta_2 e_{d,t} + \delta_d$ $\sigma_{2,2}$	$P_{2,3}$ $\mu_{2,3} = 64$ $\sigma_{2,3} = 10^{-6}$
Day	$P_{3,1}$ $\mu_{3,1} = 0$ $\sigma_{3,1} = 10^{-6}$	$P_{3,2}$ $\mu_{3,2} = \alpha_3 + \delta_d$ $\sigma_{3,2}$	$P_{3,3}$ $\mu_{3,3} = 64$ $\sigma_{3,3} = 10^{-6}$

$\boldsymbol{\delta_d}$ is a random effect of each day, assumed distributed as $\delta_d \sim Norm(0, \sigma_\delta)$. I include a random day-effect to partly remove possible dependence among the observations espe-

cially in the twilight period. The corresponding *means* and *standard deviations* for *class2* are considered as stochastic variables and specified in Table 3.1.

3.2 Prior Distributions

In this project I utilized the Bayesian analysis approach for the intended statistical inference which are explained in the next chapter. As a basic requirement of the Baysian analysis I need to set a relevant prior distribution for each parameter of my process model and observation model that I want to estimate. In table 3.1, I present The prior distributions of the parameters.

Prior Distributions	Descriptions
$\mu_0 \sim Unif(0.5, 20)$	Mean of the movement distance (given movement) among days
$\sigma_0 \sim Unif(0.01, 2)$	Standard deviation of the movement distance (given movement)*
$\psi_{01} \sim Unif(0, 1)$	$\psi 0_{01} = Pr$(moving at d | not moving at d-1)
$\psi_{11} \sim Unif(0, 1)$	$\psi 0_{11} = Pr$(moving at d | moving at d-1)
$P_{\phi,\cdot} \sim Dir(1, 1, 1)$	Vector of mixture class probability in sun elevation phase
$\alpha_\phi \sim N(0, 10)$	The intercept of the regression model in $class2$ for each sun elevation phase
$\beta_2 \sim TruncNorm(0, 10, 0, \infty)$	The slope of the regression model in $class2$ for the Twilight sun elevation phase
$\tau_{\phi,2} \sim Gamma(0.1, 0.1)$	$\sigma_{\phi,2} = \sqrt{\frac{1}{\tau_{\phi,2}}}$
$\tau_\delta \sim Gamma(0.1, 0.1)$	$\sigma_\delta = \sqrt{\frac{1}{\tau_\delta}}$

Table 3.1: *Prior distributions for the parameters of the process model and the observation model.*

* The prior distributions $\tau_0 \sim Unif(1, 20)$ ($\sigma_0 = \sqrt{\frac{1}{\tau_0}}$) and $\sigma_0 \sim Unif(1, 20)$ are also tried which are explained in chapter 5.

In the next chapter I will explain more about the steps and structures.

Chapter 4

Statistical Inference

Frequentist and *Bayesian* methods are the most common methods to utilize for data analysis and statistical inferences. Therefore to fit the *state-space models* to data and estimate parameters and hidden states we can use either the *Frequentist* or *Bayesian* approaches. To fit nonlinear models, models with non-Gaussian errors, and models with a combination of discrete and continues states, often simulation-based Bayesian techniques are used.

4.1 The Baysian Approach and MCMC

In a Bayesian model, the parameters are considered as random and have a probability distributions that indicate accessible information about those parameters before any data collected and these probability distributions are called the *priors*. After data are collected and modeled we can use Bayes theorem to update the *priors* knowledge and detect the *posterior* distributions of the parameters, then the inference is based on the *posterior* distributions of the parameters and latent variable(s) and hidden states.

If I set a likelihood as $f(Z|\Theta)$ with data Z and parameter Θ and specify the prior distribution $\pi(\Theta)$, then the *posterior* distribution of a parameter given the data can be obtained by Bayes Theorem as below

$$p(\Theta|Z) = \frac{f(Z|\Theta)\pi(\Theta)}{\int f(Z|\Theta)\pi(\Theta)d\Theta}. \tag{4.1}$$

The integral in the denominator of equation (4.1) is the marginal distribution of Z with dimension equal to the number of parameters. This type of integral is not always easy to solve numerically, so we need other methods to cope with. One approach is using Monte Carlo integration based on Markov Chains that is shortly described in chapter 3.

Markov Chain Monte Carlo (MCMC) methods contain algorithms to draw samples from the *posterior* distribution to approximate the intended *posterior*. Then all the properties of the *posterior* are approximated by the corresponding properties of the samples. The Monte Carlo integration, which is the well known method to approximate the complicated integral by sampling, can be done with independent samples. However in MCMC methods which are based on the Markov chain rule, the samples are dependent. The two most common MCMC methods are *Metropolis-Hastings* and *Gibbs sampling* [2].

Based on theory, an infinitive number of samples in a MCMC method can exactly represent the *posterior*. However in practice only finite samples can be drawn, therefore desired accuracy can be reached by increasing the number of samples. It takes time and need a large number of samples until the *posterior* distribution of a parameter *converge* to its stationary distribution. The *convergence* happens when a chain moves stationarily around a target value. There are also two other key words named *mixing* and *burn-in*. *Mixing* refers to fluctuation of a chain around a target value and we have a *good mixing* when there is low autocorrelation and the chain samples more rapidly from the entire *posterior* distribution (Figure 4.1). To check the *convergence* and *mixing* we can set one chain for a long time (with more samples) or several chains with different starting points. *Burn-in* is the period before the chain reaches the *convergence* (Figure 4.2).

The open source software such as WinBUGS/OpenBUGS and JAGS are used for Bayesian analysis using MCMC methods. They are not very complicated to use and a vast variety of models can be programmed by providing code to specify the model structure and then the software process automatically to sample from the *posterior*. In addition there are R packages such as *R2OpenBUGS* and *R2jags* for easy implementation in R and make a connection between R and the mentioned software.

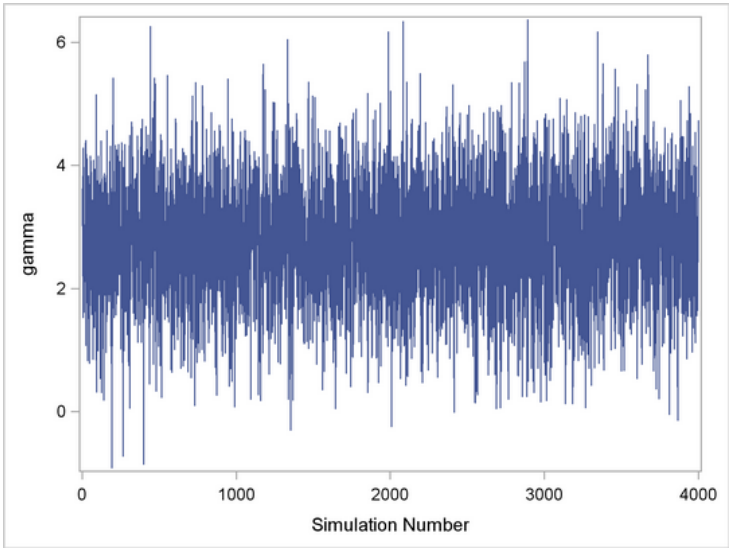


Figure 4.1: *An example of a perfect MCMC trace plot; The chain is mixing well and the posterior can be sampled efficiently.*

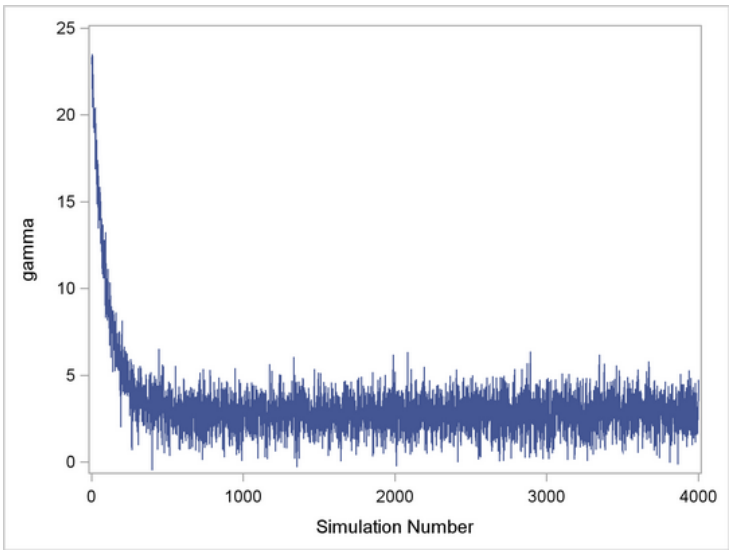


Figure 4.2: *An example of a trace plot that indicates the first few hundred iterations should be discarded, in other word the burn-in sample size should be increased.*

In this project I mainly use JAGS via the *R2jags* and *dclone* packages in R. JAGS codes can be run in Unix based environments. *R2jags* is used to call JAGS from R. The R package *dclone* also includes functions to call JAGS from R and parallel chains are run on parallel workers (for example on different cores of a computer's CPU), thus computations can be done faster for long MCMC runs.

4.1.1 Computational Challenges

Working with large data sets and complex models using MCMC, require very long computational time and powerful computer system. Therefore, in addition to manage the data before starting analysis I need to set reasonable number of samples by specifying the number of *iterations* and *burn-in* and check the convergence of the *posterior* distribution of parameters to their true distribution. For this project I could get an access to the Abel computer cluster and send my jobs to implement in that computer system. The maximum time that I were allowed to spend for each job was 168 hours (one week).

To start the analysis in this project, first I decided to set up the appropriate prior distributions which were explained in chapter 3 for the parameters of my observation model. Then I used the known location's coordinates of times that the bird were sighted during three years and corresponding *light intensities*, both as data, to estimate the *posterior* distributions of the observation model's parameters.

Now I set $y_{d,t}$ as the observations of the *light intensities* at record t of day d and x_d as the *locations* on day d, and Θ_o as the set of the observation model's parameters with the prior distribution $\pi(\Theta_o)$. If I define S_1, X and S_2 as below

$$X = \{(x_d, x_{d-unknown}\} \tag{4.3}$$

$x_{d-unknown}$ refers to the unknown locations.

$$S_1 = \{(y_{d,1}, ..., y_{d,T}), x_{d-unknown}\} \tag{4.4}$$

then we have

$$p(\Theta_o|S_1) = \frac{f(S_1|\Theta_o)\pi(\Theta_o)}{\int f(S_1|\Theta_o)\pi(\Theta_o)d\Theta_o}. \tag{4.5}$$

In the next step I utilized the obtained *posterior* means as the constant values in the observation model and also consider prior distributions for the parameters of the process model (chapter 3). Then I used the *light intensity* data with corresponding *dates* and *times* for 181 days of the study and used the whole *state-space model* to obtain the *posterior* distributions of *the process model's parameters* and *unknown locations* for each day.

Now if I consider θ_p as the set of parameters in the process model we have

$$p(X, \Theta_p | S_2, \Theta_o) = \frac{f(S_2|X, \Theta_p, \Theta_o) g(X|\Theta_p) \pi(\Theta_p)}{\int\int f(S_2|X, \Theta_p, \Theta_o) g(X|\Theta_p) \pi(\Theta_p) dX d\Theta_p}. \tag{4.6}$$

f and g correspond to the observation model and the process model respectively. As I explained, the mentioned *posterior* distributions (4.5) and (4.6) are obtained via MCMC methods.

To save time I managed my data set in different ways. First I thinned the data to every forth record of the *light intensity* and also excluded the records between 10:00 and 15:00 o'clock which normally should be at the maximum level of light intensity (64) and may not be very informative. Then to have more informative data and more accurate results, I restricted the data to include only 2 hours before sunrise to 1 hour after sunrise plus 1 hour before sunset to 2 hours after sunset, based on the preliminarily *posterior* means of the *locations*. The exact times of the sunrises and sunsets of the days of the study were unknown, so I approximated them via a function in R named *RAtmosphere::suncalc* which needs corresponding *days of year* and values of *latitudes* and *longitudes* of those days, so I used set of estimated *latitudes* and *longitudes* that obtained from the previous run (The run with using every forth record and excluding 5 hours records as data).

In addition to the explained data managing, I also did two other types of analyses. As mentioned in chapter 2, I have some recorded *locations* for the time that the bird is sighted during the whole times of the study. Hence I also tried to use the *known locations* in the period of 181 days as data (which belonged to the first days and last days) to check if it results in accurate estimates. The obtained plots and maps with the details of each step are represented in the next chapter.

The set of all JAGS and R code that I constructed and utilized are attached in the appendix.

Chapter 5

Results

In this chapter I present some parts of our results from the Bayesian analysis via MCMC methods that have been done in this study to make inference about our state-space model and to estimates unknown states. Simulation-based Bayesian inference needs to use simulated sampling to obtain the posterior distribution or any relevant required quantities. Generally there are two issues that I should consider in MCMC analysis. First, I need to check whether the Markov chain has reached its stationary, or in other word the posterior distributions of the parameters converge to the desired distributions. Then I should determine the number of iterations to run after the Markov chain has reached stationary.

The trace plot (plot of successive values of the MCMC chains) is a useful tool to assess convergence and the density plot is used to visualize the density of the samples that are used for posterior estimates. In the following sections I show posterior estimates of parameters of both the observation model and the process model via their trace plots and density plots. To reach my main goal of this project I estimated posterior estimates of the states which are the coordinates of locations of the bird migration for 181 days of the study (from end of May til end of November). I then present the estimated tracks of the bird migration on plots and geographic maps. In addition I visualize the estimated coordinates of locations obtained via the *GeoLight* package to make a comparison.

5.1 Posterior Estimates for the Parameters of the Observation Model

As explained in the previous chapter, first I used only data from the days where the bird had been sighted during the three years (June 2011 - June 2014), and fitted only the observation model to these data. The total number of MCMC iterations that I set for this job to run, was 7000, the number of iterations that we discarded as burn-in was 4000, and I used 5 chains, so I have 3000 samples per chain. The results are represented in table 5.1 and figures 5.1, 5.2, 5.3, 5.4 and 5.5.

Parameter	Mean	St. Dev.	Q. 2.5%	Median	Q. 97.5%
α_1	-14.107	5.625	-26.038	-13.622	-4.530
α_2	9.518	10.486	-10.618	9.585	30.152
α_3	-17.686	6.817	-31.965	-17.277	-5.426
β_2	28.745	6.065	17.887	28.312	41.779
$P_{1,1}$	0.979	0.001	0.976	0.979	0.982
$P_{1,2}$	0.021	0.001	0.018	0.021	0.024
$P_{1,3}$	0.0002	0.0002	0.000	0.0002	0.0006
$P_{2,1}$	0.498	0.014	0.471	0.497	0.524
$P_{2,2}$	0.399	0.013	0.373	0.399	0.425
$P_{2,3}$	0.104	0.008	0.088	0.104	0.120
$P_{3,1}$	0.062	0.002	0.058	0.062	0.065
$P_{3,2}$	0.058	0.002	0.055	0.058	0.062
$P_{3,3}$	0.880	0.003	0.875	0.880	0.885
$\sigma_{1,2}$	7.027	1.163	4.864	6.997	9.464
$\sigma_{2,2}$	48.730	5.271	38.645	48.640	59.203
$\sigma_{3,2}$	39.239	2.974	33.752	39.109	45.408
σ_δ	14.153	2.966	9.063	13.956	20.458

Table 5.1: *Summary statistics for the parameters of the observation model from corresponding posterior distributions based on data with known locations.*

Table 5.1 contains posterior means of the parameters of the observation model, which I used them as constant values in our further analysis. It also contains other summary statistics that give us general insight about the posterior distributions of the parameters. The mean estimates of α_1, α_2 and α_3 represent the intercepts of regression model between the *light intensity* and the *elevation* for three time phases and β_2 it is the regression coefficient for *twilight* phase. As mentioned before, I considered β_1 and β_3 as zero (no

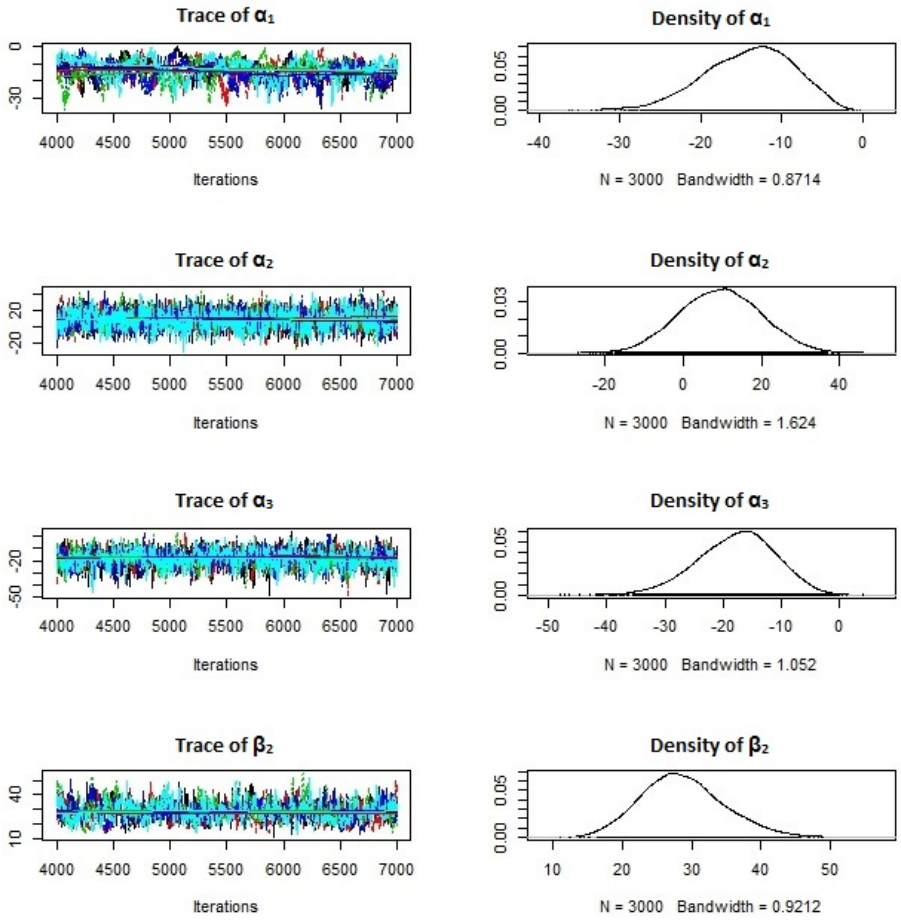


Figure 5.1: *Trace plots and Density plots for the parameters* $\alpha_1, \alpha_2, \alpha_3, \beta_2$

effect of *elevation* on *light intensity* at *night* and *day* phases). The estimated means of $P_{\phi,\kappa}$ for (i=1,2,3 and j=1,2,3) confirm that it is unlikely to have maximum light at night or minimum light at midday and vise verse, and also give the corresponding probabilities for the mentioned regression models at twilight times. I also see more variation for the values of the *light intensity* in *twilight* and *day* phases rather than *night* phase as expected.

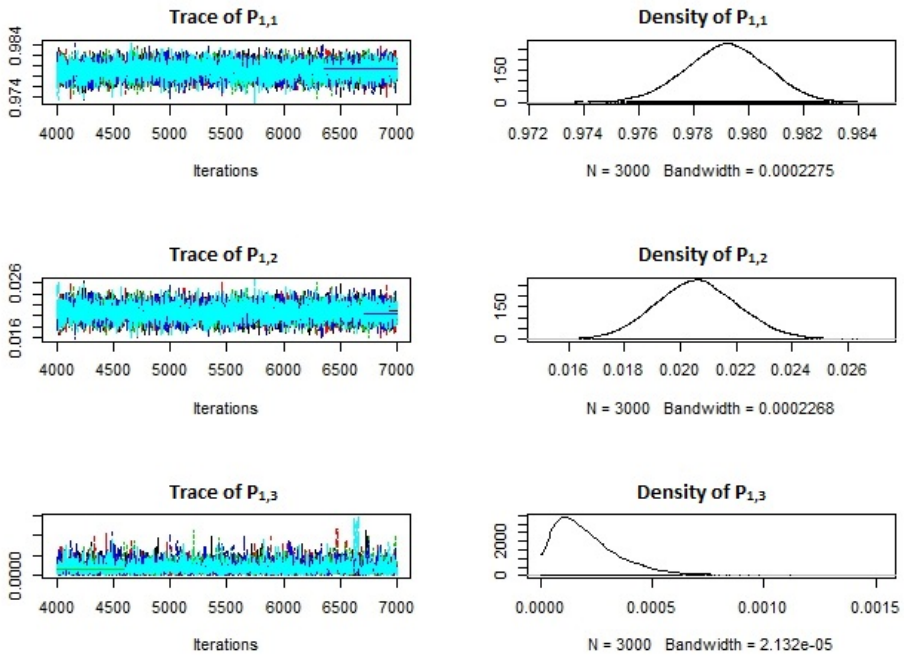


Figure 5.2: *Trace plots and Density plots for the parameters* $P_{1,1}, P_{1,2}, P_{1,3}$

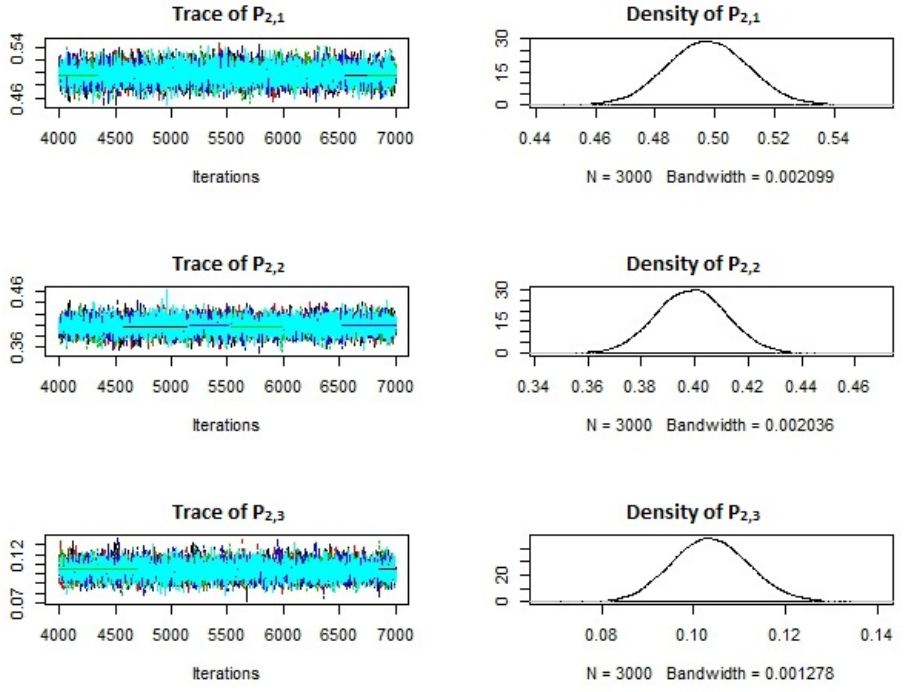


Figure 5.3: *Trace plots and Density plots for the parameters* $P_{2,1}, P_{2,2}, P_{2,3}$

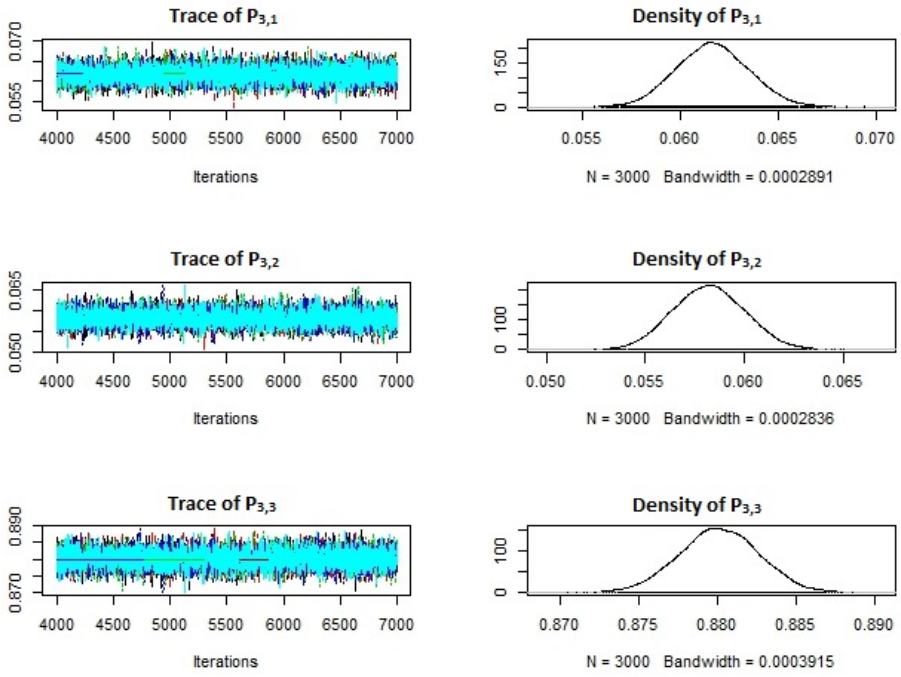


Figure 5.4: *Trace plots and Density plots for the parameters* $P_{3,1}, P_{3,2}, P_{3,3}$

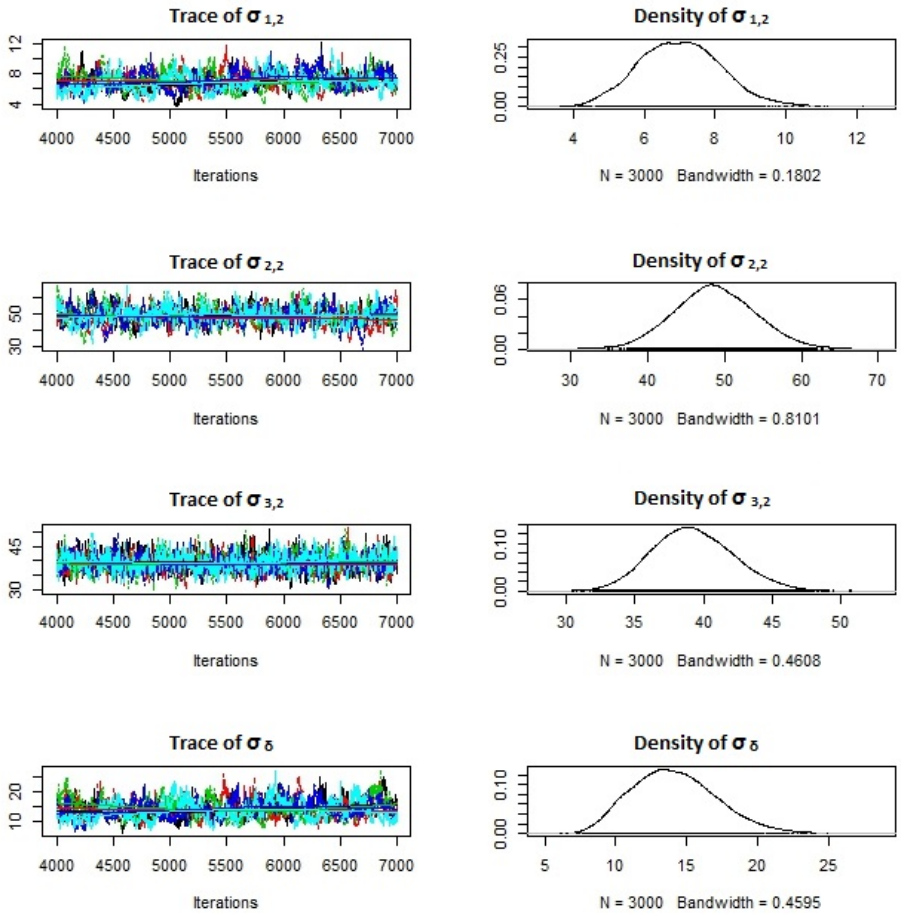


Figure 5.5: *Trace plots and Density plots for the parameters* $\sigma_{1,2}, \sigma_{2,2}, \sigma_{3,2}, \sigma_{\delta}$

Figures 5.1, 5.2, 5.3, 5.4 and 5.5 show trace plots and density plots for each parameter of our observation model from a MCMC performance with 5 parallel chains and 7000 iterations with 4000 burn-in samples. The parallel chains in each trace plot for each parameter show they converge to the almost same target estimate and the mixing is good; The density plots visualise almost symmetric distributions for the parameters, so I used the means of the distributions in further analysis.

5.2 Posterior Estimates for The Parameters of The Process Model and States

After I obtained the posterior distributions of the parameters of the observation model, I replaced each of those parameters with corresponding posterior means and then fit my whole state space model to the *light intensity* data for 181 days of the study (from 30/05/2012 to 27/11/2012). However because of the time limitation as explained in the previous chapter, first thinned the *light intensity* data to every forth record and excluded the records between 10:00 and 15:00 o'clock and checked the results such as posterior estimates of the process model parameters and the hidden states which are in fact unknown coordinates of *locations*. Then I used the *light intensity* records belonging to the times around the Twilight periods (sunrise and sunset times) during the 181 days of the study. Another challenge that I had, was related to set an appropriate prior distribution for the variation of the mentioned *movement distance* in the process model (chapter 3), so I set different prior distributions that are mentioned in the following (shown by standard deviation σ_0 or *precision* τ_0). In addition, I used known *locations* belonging to the first days and last days of the period of 181 days as data and checked the results.

The results are provided in four sections with four settings as below

Setting 1:

1-1) Selecting every forth record of the *light intensity* data.

1-2) Excluding the records between 10:00 and 15:00 o'clock.

1-3) Setting $\tau_0 \sim Unif(1, 20)$ as a prior distribution for precision of the *movement distance* variable (hence, as $\sigma_0 = \sqrt{\frac{1}{\tau_0}}$, the range of the prior distribution for σ_0 becomes $\sqrt{\frac{1}{20}} < \sigma_0 < 1$ which is too small).

Setting 2:

2-1) Selecting every forth record of the *light intensity* data.

2-2) Excluding the records between 10:00 and 15:00 o'clock.

2-3) Setting $\sigma_0 \sim Unif(1, 20)$ as a prior distribution for standard deviation of the *movement distance* variable.

2-3) Using known *locations* as data.

Setting 3:

3-1) Selecting the *light intensity* records around the Twilight periods.

3-2) Setting the $\sigma_0 \sim Unif(0.01, 2)$ as a prior distribution for the standard deviation of the *movement distance* variable.

3-3) Using known *locations* as data.

Setting 4:

3-1) Selecting the *light intensity* records around the Twilight periods.

3-2) Setting the $\sigma_0 \sim Unif(0.01, 2)$ as a prior distribution for the standard deviation of the *movement distance* variable.

5.2.1 Setting 1

Here I provide the results belonging to the setting 1.

Figure 5.6 and Table 5.2 show the posterior estimates of the process model parameters via corresponding trace plots, density plots and summary statistics that are obtained from MCMC with 3 parallel chains and 7000 iterations with 4000 burn-in samples.

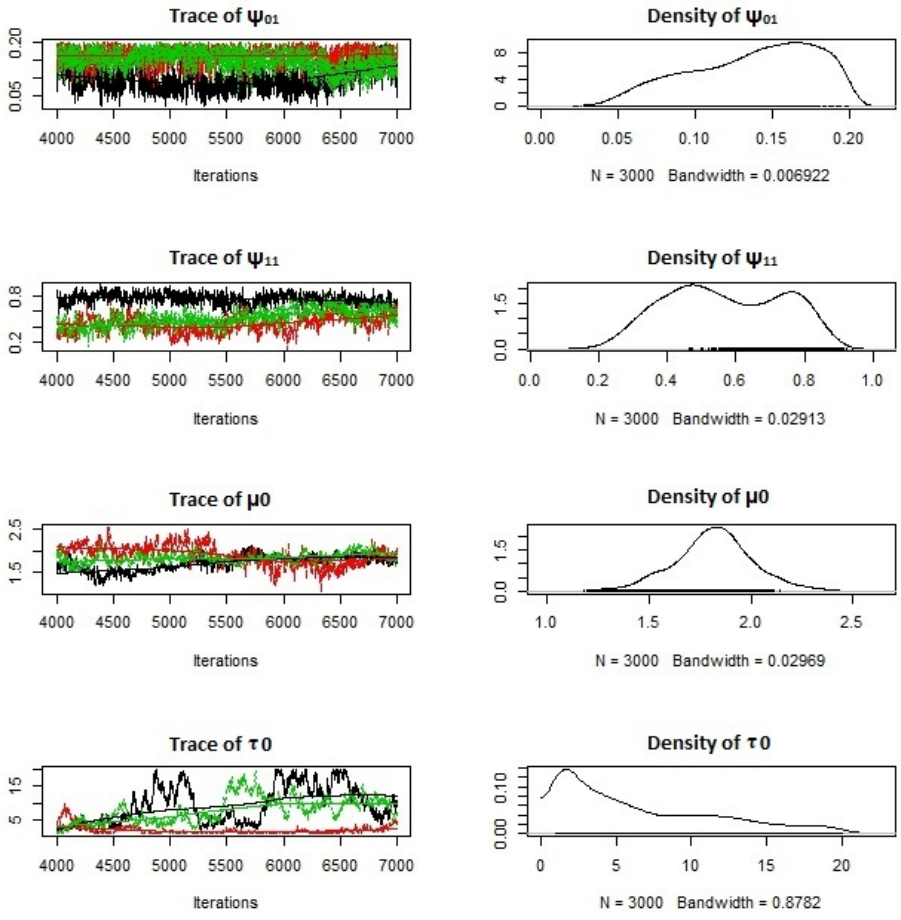


Figure 5.6: *Trace plots and Density plots for the parameters of the process model for Setting 1*

Figure 5.6 shows that there are no perfect mixing in chains of each trace plot, especially the one belongs to τ_0 which confirm that our prior distribution for the *precision* is far from the true distribution and needs more sample to reach the convergence. However the posterior distribution of the parameter μ_0 (mean of the *distance* variable) is acceptable

Parameter	Mean	St. Dev.	Q. 2.5%	Median	Q. 97.5%
ψ_{01}	0.137	0.040	0.055	0.144	0.196
ψ_{11}	0.568	0.170	0.271	0.556	0.854
μ_0	1.818	0.196	1.422	1.822	2.219
τ_0	6.430	5.119	1.037	4.705	18.555

Table 5.2: *Summary statistics for the parameters of the process model from Setting 1*

because of the better mixing in corresponding chains and the quite symmetric density plot. Contents of Table 5.2 also confirm the symmetry or the lack of symmetry in the density of each parameter by comparing the corresponding means and medians.

Then the estimated coordinates of the bird's locations (*latitudes* and *longitudes*) from setting 1 are displayed on the geographical map (Figures 5.7), and also via separated plots versus *day* (5.8).

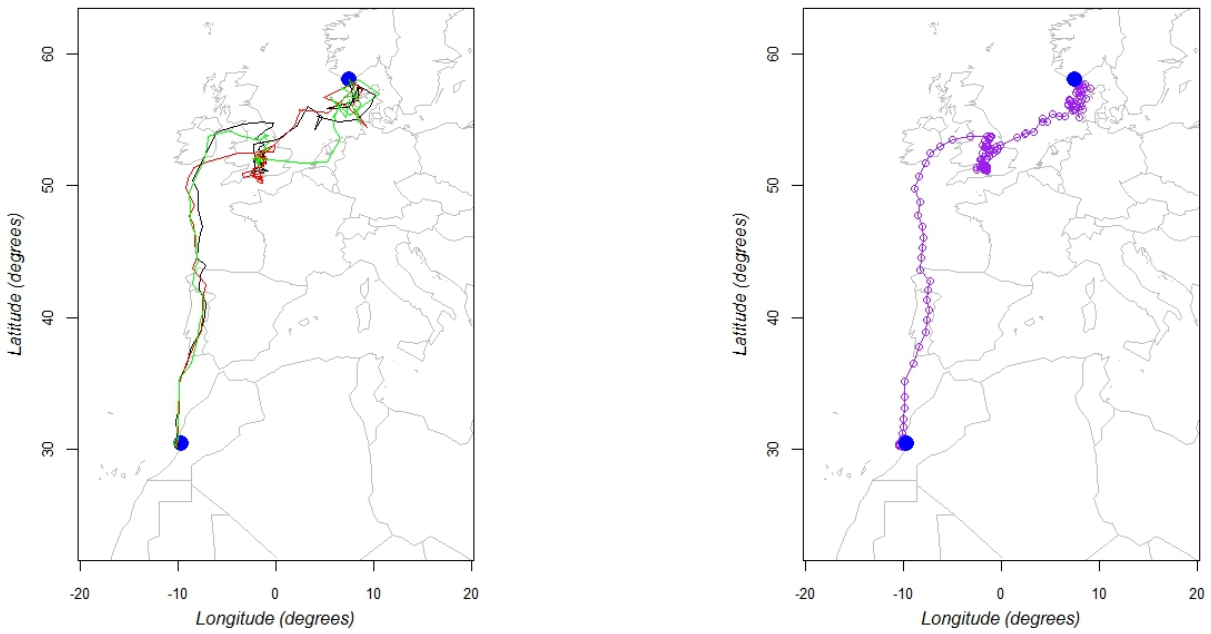


Figure 5.7: *The estimated track from Setting 1 (selecting every forth record, excluding 5 hours of data, allowing small variation for the movement distances). The left panel shows the means of the estimated locations from samples of the three chains separately. The right panel shows the means of all samples of the estimated locations. The blue spots just indicate the known locations but they are not used as data.*

Figure 5.7 shows the estimated track(s) of migration of the bird during the 181 days of the study that is from Norway to Morocco via North sea, UK, Atlantic ocean, Spain

and Portugal. The left panel displays the estimated tracks from each chain separately and I see the good chains mixing from the time that the bird migrated to the south directly. The right panel shows the total estimated track from all three chains. As I allowed small variation for the *movement distance* between days, I see that there are no considerable variations in most of the distances between the estimated locations in the days that the bird moved. However it is not biologically plausible. Generally it is expected that the bird fly much faster and pas longer distances above the water than above the lands.

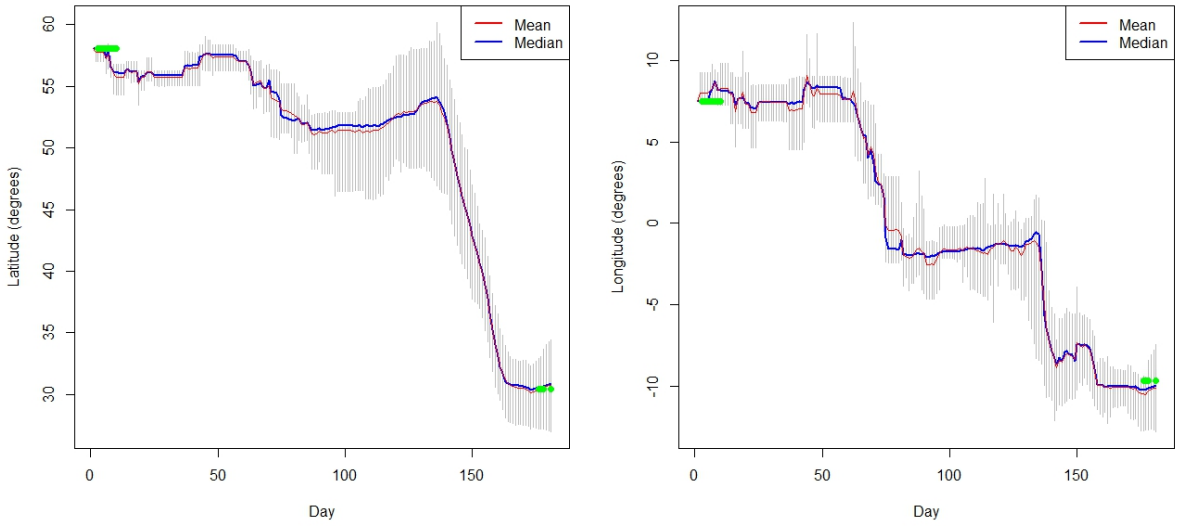


Figure 5.8: *Curves of the posterior means and medians of the samples of the estimated latitudes and longitudes versus day with 95% credibility intervals (gray vertical lines) from setting 1 (selecting every forth record, excluding 5 hours of data, allowing small variation for the movement distances). The green spots just indicate the known locations but were not used as data.*

Figure 5.8 illustrates the posterior means (red line) and medians (blue line) of the samples of the *latitudes* and *longitudes* in 181 days of the study with 95% credibility intervals from setting 1. The *Latitude* plot shows that the bird migrated sharply latitudinal to the south from approximately 130th day of the study when the winter started and the *Longitude* plot indicates two considerable longitudinal jumps in the bird migration, one at summer (around July) and one at winter (around October).

It seems that the posterior means and medians of the samples of the estimated coordinates almost match in most days and it confirms that the density of the posterior estimates of

the *latitudes* and *longitudes* are approximately symmetric. Green spots refers to known locations for eight first and four last days of the period. The location of the green spots are close to the corresponding estimated locations and they especially confirm that I have obtained good estimates at the end days of the study.

5.2.2 Setting 2

Here I provide the results belonging to the setting 2. I fit the model to every forth record of the *light intensity* data and used known locations as data. The records between 10:00 and 15:00 o'clock are also excluded. In addition I set $\sigma_0 \sim Unif(1, 20)$ as a prior distribution for the standard deviation of the *movement distance*.

Figure 5.9 and Table 5.3 show posterior estimates of the process model parameters via corresponding trace plots, density plots and summary statistics that are obtained from MCMC with 3 parallel chains and 7000 iterations with 4000 burn-in samples.

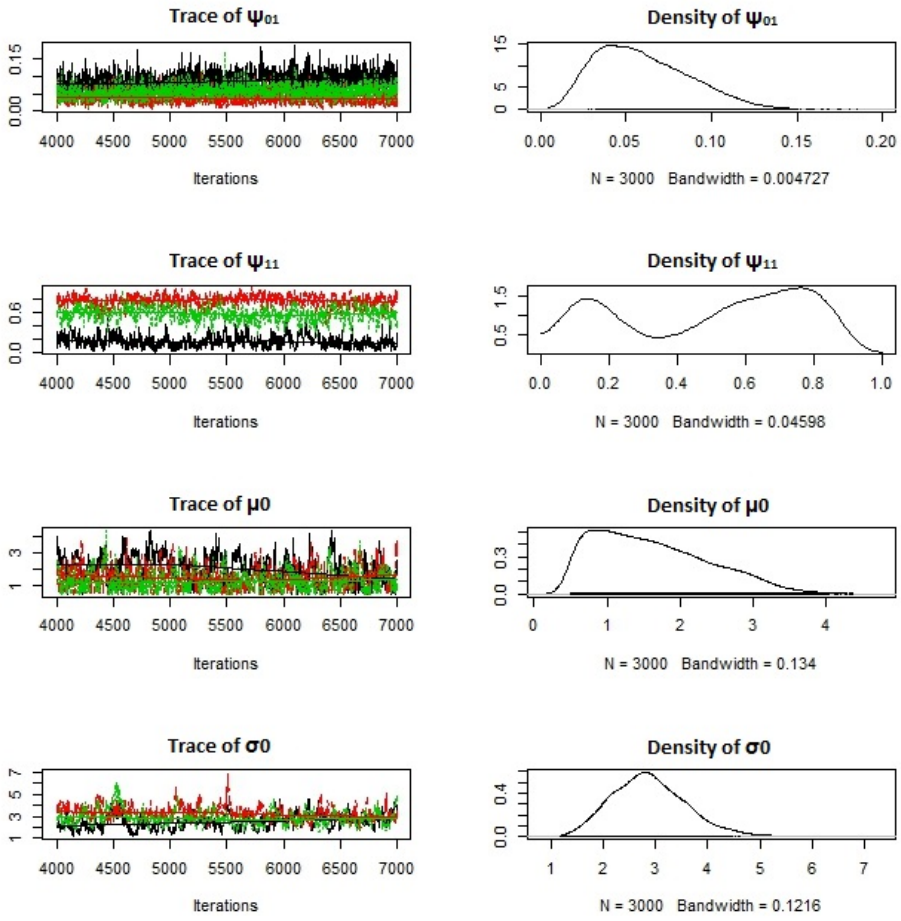


Figure 5.9: *Trace plots and Density plots for the parameters of the process model from Setting 2*

Figure 5.9 shows that, by the specified number of iterations, there are no perfect mixing in the chains of each trace plot, especially the one belonging to ψ_{11}. However the posterior

Parameter	Mean	St. Dev.	Q. 2.5%	Median	Q. 97.5%
ψ_{01}	0.06	0.03	0.02	0.06	0.12
ψ_{11}	0.50	0.27	0.06	0.57	0.88
μ_0	1.62	0.78	0.54	1.50	3.29
σ_0	2.89	0.71	1.66	2.84	4.45

Table 5.3: *Summary Statistics for the parameters of the process model from Setting 2*

distribution of parameter σ_0 (the standard deviance of the *distance* variable) is almost acceptable because of the better mixing in corresponding chains and the quite symmetric density plot. Contents of the Table 5.3 also confirm the symmetry or lack of symmetry in the density of each parameter by comparing the corresponding means and medians.

Then the estimated coordinates of the bird's locations (*latitudes* and *longitudes*) from setting 1 are displayed on the geographical map (Figures 5.10), and also via separated plots versus *day* (Figures 5.11).

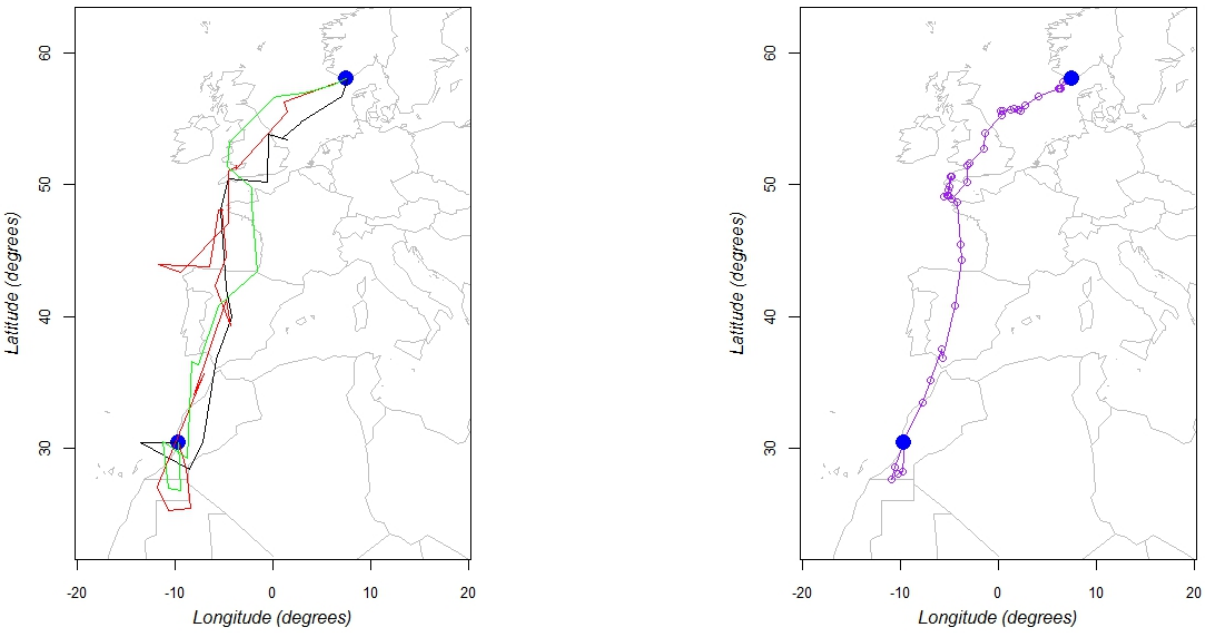


Figure 5.10: *The estimated track(s) from Setting 2 (selecting every forth record, excluding 5 hours of data, allowing large variation for the movement distances and using known locations as data). The left panel shows the means of the estimated locations from samples of the three chains separately. The right panel shows the means of all samples of the estimated locations. The blue spots indicate the known locations.*

Figure 5.10 displays the estimated track(s) of migration of the bird during the 181 days of the study that is from Norway to Morocco via North sea, UK, Atlantic Ocean, Spain. The left panel displays the estimated tracks from each chain separately and I see that the chains mixing is not satisfiable. The right panel shows the total estimated track from all three chains. As I allowed a large variation for the *movement distances* between days, I see that the distances between estimated locations are varied considerably. The track also shows that the bird migrated to the south before turning north again, because the blue spot in the bottom of the map belongs to the last location of the bird during this period of time.

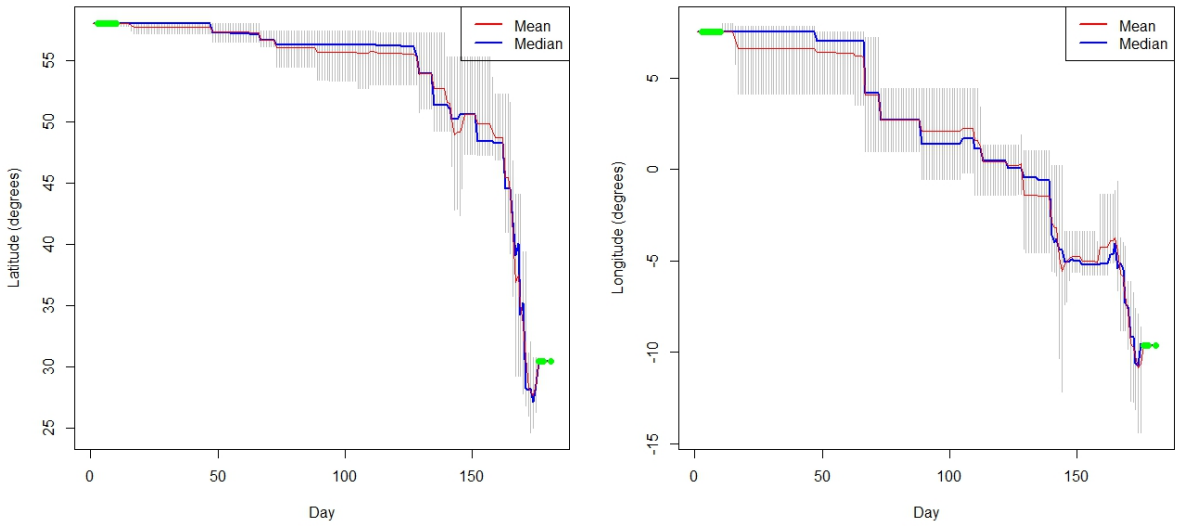


Figure 5.11: *Curves of the posterior means and medians of the samples of the estimated latitudes and longitudes versus day with 95% credibility intervals (gray vertical lines) from setting 2 (selecting every forth record, excluding 5 hours of data, allowing large variation for the movement distances, using known locations as data). The green spots just indicate the known locations but were not used as data.*

Figure 5.11 shows the posterior means and medians of the samples of the estimated *latitudes* and *longitudes* in 181 days of the study with 95% credibility intervals for the posteriors. The *latitude* plot confirms the turn to the north during the last days in this period and the *longitude* plot indicates a turn towards east. The general track of the bird's migration is similar to the previous model fit (setting 1), but the patterns are slightly

different, especially in the longitudinal movement.

It seems the posterior means and medians almost match in many days and but not in all days. I see the large uncertainties in almost all estimates belonging to the longitude (grey lines in Figure 5.11) except for first and last few days which I used corresponding locations as data.

5.2.3 Setting 3

In this step I analyse the *light intensity* records around Twilight periods which can be more informative (data between 2 hours before and 1 hour after sunrise times and data between 1 hour before and 2 hours after sunset times). These records are more informative because during twilight periods the values of *light intensity* change more that other periods. As explained in the previous chapter sunrise and sunset times for specified locations can be approximately calculated by a function in R named *RAtmosphere::suncalc*, Which needs *days* of year and corresponding *latitudes* and *longitudes* to approximately calculate pairs of sunrise and sunset times. For this purpose I used estimated *latitudes* and *longitudes* from the results of setting 2 (The estimated locations from both setting 1 and setting 2 are not very different for the purpose approximating sunrise and sunset times). As I did not see the satisfiable chains mixing for the means of the sample of the estimated locations from setting 2, I decided again to decrease the range of the prior distribution of the standard deviation for the daily *movement distance*, but not as small as before. I tried $\sigma_0 \sim Unif(0.01, 2)$. In addition I used the known locations as data.

Table 5.4 and Figure 5.12 show posterior estimates of the process model parameters (form setting 3) via corresponding summary statistics, trace plots and density plots that are obtained from MCMC with 3 parallel chains and 7000 iterations with 4000 burn-in samples.

Parameter	Mean	St. Dev.	Q. 2.5%	Median	Q. 97.5%
ψ_{01}	0.08	0.02	0.04	0.08	0.14
ψ_{11}	0.67	0.10	0.46	0.68	0.84
μ_0	1.47	0.54	0.56	1.50	2.45
σ_0	1.50	0.29	0.93	1.51	1.97

Table 5.4: *Summary Statistics for the parameters of the process model from Setting 3*

Figure 5.12 shows that again, by the specified number of iterations, I don't have a perfect mixing in the chains of each trace plot, but this time the posterior estimates of $\psi_0 1$ and $\psi_1 1$ can be more reliable than posterior estimates of the μ_0 and σ_0 because of better mixing in chains and quiet symmetric density plots. The summary statistics of the table 5.4 also confirm the symmetry or lack of symmetry in the density of each parameter by

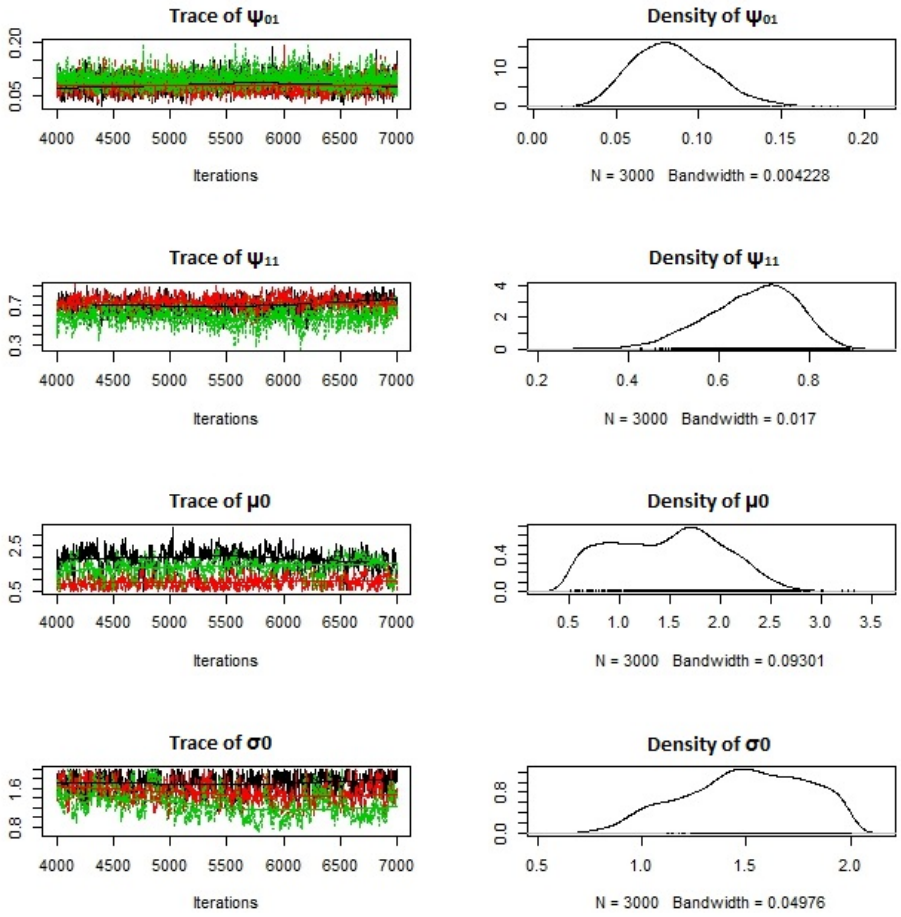


Figure 5.12: *Trace plots and Density plots for the parameters of the process model from Setting 3*

comparing the corresponding means and medians.

Then the estimated coordinates of the bird's locations (*latitudes* and *longitudes*) from setting 3 are displayed on the geographical map (Figures 5.13), and also via separated plots versus *day* (5.14).

Figure 5.13 displays the estimated track(s) of migration of the bird during the 181 days of the study that are partly similar to the previous estimated tracks (setting 1 and 2). The left panel displays the estimated tracks from each chain separately and I see that the chains mixing is not partly satisfiable which is probably caused by limited number of iterations. The right panel shows the total estimated track from all three chains. This track is generally more reliable than the tracks from previous model fits (setting 1 and 2) because I used more informative records of *light intensity* and set more reasonable prior

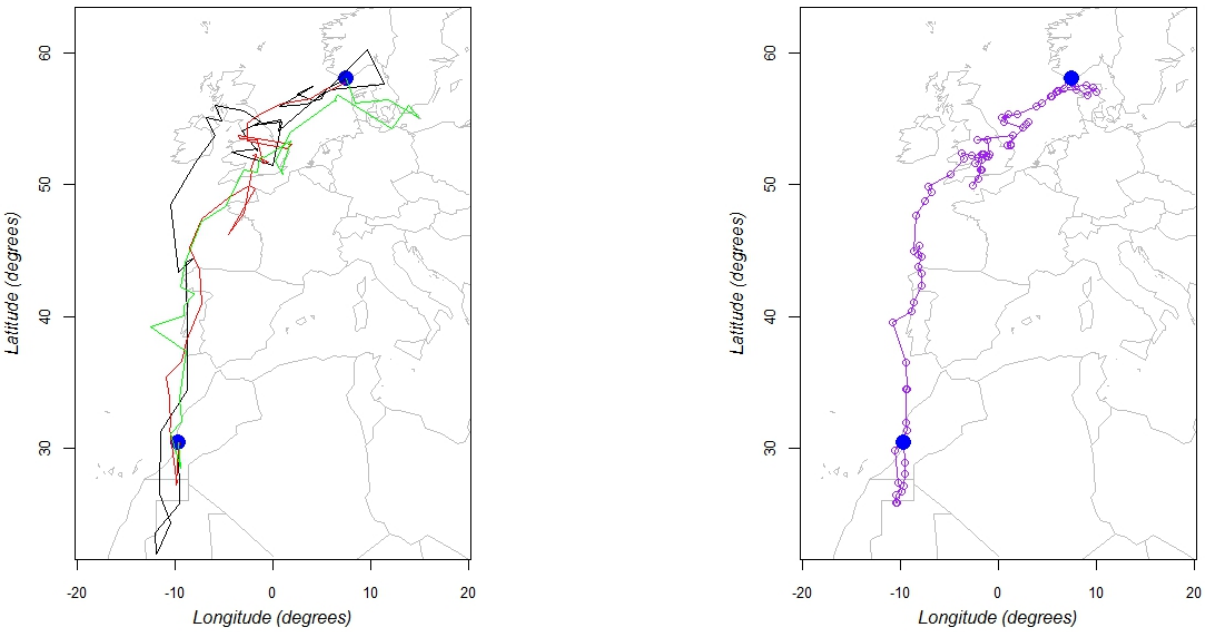


Figure 5.13: *The estimated track(s) from Setting 3 (selecting the records around twilight times, using the known location as data, allowing medium variation for the movement distances (0.01,2)). The left panel shows the means of the estimated locations from samples of the three chains separately. The right panel shows the means of all samples of the estimated locations. The blue spots indicate the known locations.*

distribution for the standard deviation of the *movement distance* variable.

Figure 5.14 illustrates the posterior means (red line) and medians (blue line) of the samples of the *latitudes* and *longitudes* in 181 days of the study with 95% credibility intervals from setting 3. I see less uncertainty in most daily estimated. However, usually for the periods that the bird does not move much, better estimates can be obtained because the information during these periods are accumulated.

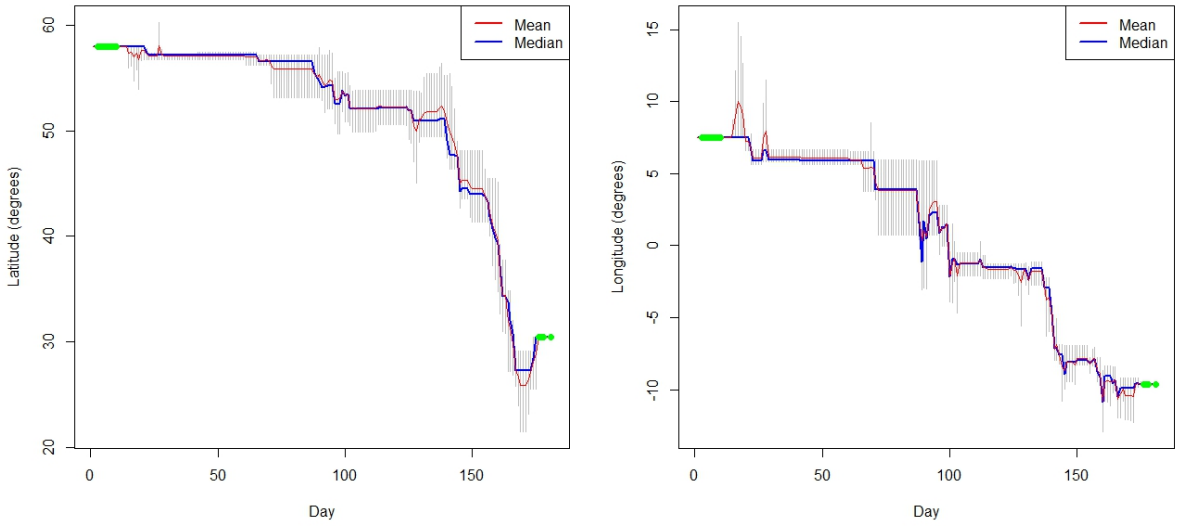


Figure 5.14: *Curves of the posterior means and medians of the latitudes and longitudes versus day with 95% credibility intervals (gray vertical lines) from setting 3 (selecting the records around Twilight periods, using the known location as data, allowing medium variation for the movement distances (0.01,2)). The green spots indicate the known locations.*

5.2.4 Setting 4

In the last step I again selected the *light intensity* records around Twilight periods and used the same prior distribution for the mentioned standard deviation ($\sigma_0 \sim Unif(0.01, 2)$), but I did not use the known locations. Here I only represent maps and the plots for *latitudes* and *longitudes* vursus *days*. Figure 5.15 and Figure 5.16.

By comparing the results from setting 3 and setting 4, I mainly see that the estimated locations belonging to the middle parts of the tracks, around UK, are partly different from each other (the right panels of Figure 5.13 and Figure 5.15). In addition there are no satisfiable chains mixing in the mentioned part of tracks in both chains map (the left panels of Figure 5.13 and Figure 5.15). Therefore one reason of some differences between estimated tracks from setting 3 and setting 4 can be convergence problem which is not far from the expectation due to the limited number of iterations. On the other hand by checking the *latitude* plots versus days in Figure 5.14 and Figure 5.16, I see that there are considerable uncertainties (grey vertical lines) in the period between days 110th and 140th which are around the equinox time (21 September). Usually in this period which the lengths of days and nights are close to each other, it is difficult to use *light intensity* data to estimate *locations*.

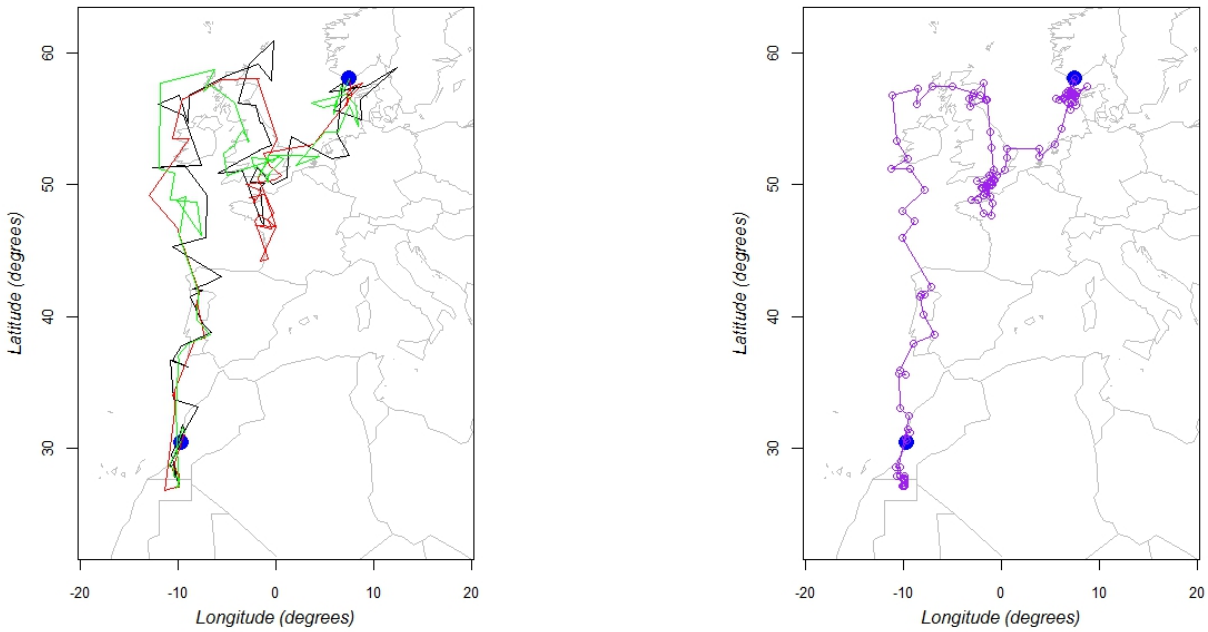


Figure 5.15: *The estimated track(s) from Setting 4 (using the records around twilight times and allowing medium variation for the movement distances (0.01,2)). The left panel shows means of estimated locations from samples of the three chains separately. The left panel shows the means of the estimated locations from the samples of the three chains separately. The right panel shows the means of all samples of the estimated locations. The blue spots indicate the known locations but were not used as data.*

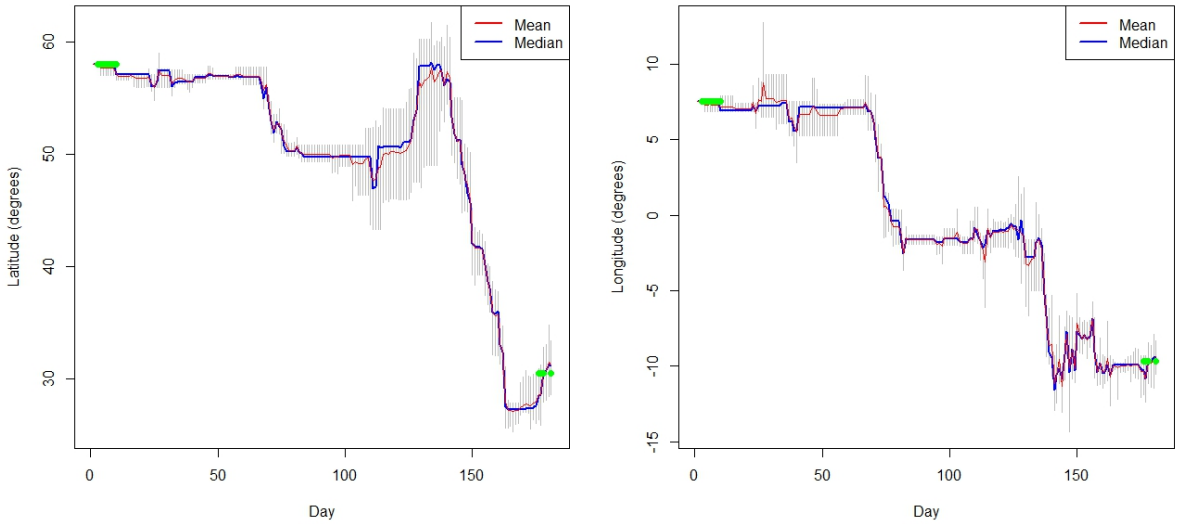


Figure 5.16: *Curves of the posterior means and medians of the latitudes and longitudes versus day with 95% credibility intervals (gray vertical lines) from setting 4 (selecting the records around Twilight periods and allowing medium variation for the movement distances (0.01,2)). The green spots indicate the known locations but were not used as data.*

5.2.5 Comparing my result with the result from GeoLight Package

I also used the GeoLight package functions to analyse the *light intensity* data for the same bird, at the same time period, to calculate the corresponding locations. I used all data belonging to 181 days of the study (without any thinning) because it did not take long time to run. The calculated locations have been plotted in Figure 5.17.

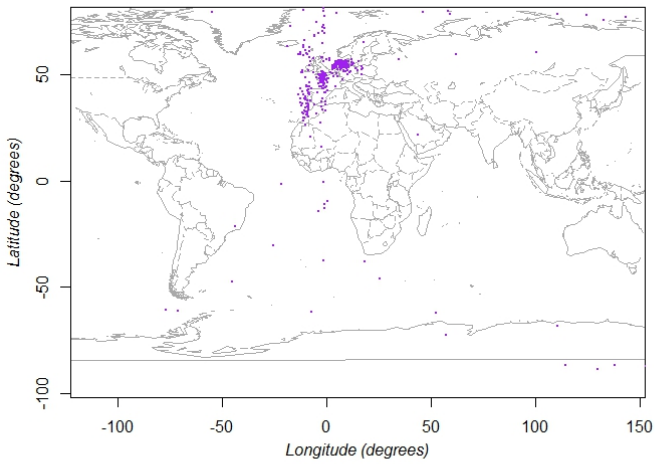


Figure 5.17: *Calculated locations by GeoLight package*

Figure 5.17 show many unrealistic calculated locations which are not reliable.

Then I used a distance filter from this package to partly filter out unrealistic calculated coordinates by determining the maximum distance in a certain time unit for the bird of this study (Figure 5.18). However it seems that there are still unexpected calculated locations.

I also provided the estimated locations that I obtained from my Bayesian state-space model in the world map; I used the results from setting 3 (Figure 5.19).

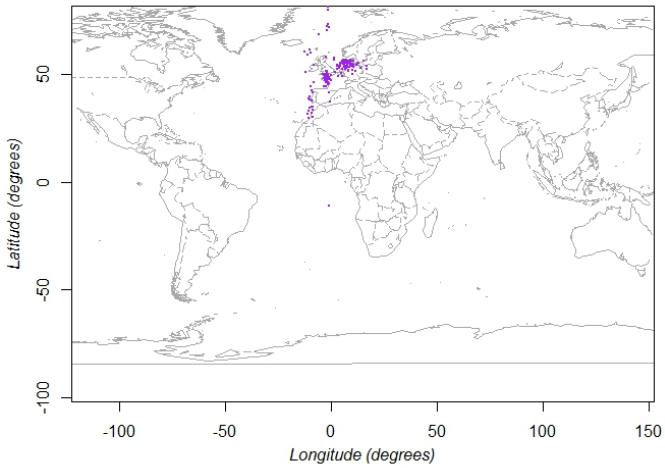


Figure 5.18: *Calculated locations by GeoLight package with using a distance filter*

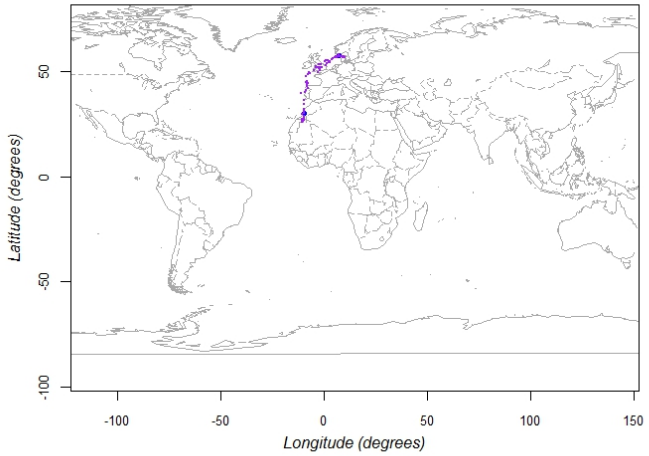


Figure 5.19: *The estimated locations by the Bayesian state-space model*

Chapter 6

Conclusion and Discussion

In this project I mainly aimed to estimate locations of a migrating bird and obtain tracks of its movements during a specific period of time based on recorded light intensities. The data available for this project has been collected by using of a device called Geolocator, which is deployed on birds and records light intensities in every 10 minutes. Generally, due to the environmental effects and the birds behaviours, these kinds of data are typically noisy. In addition known locations for the times that the bird was captured or just sighted were available.

To achieve the main purpose of this project I constructed a *state-space model* which includes a process model and an observation model. My process model refers to the dynamics of the bird movement through the study period. I used a 24-hour period from midday to midday as the time step and specified by d, because I assumed the bird moves during the day time and not during the night time. My observation model specifies how the light intensity data relates to the locations which are in fact hidden states.

I used a *Bayesian* approach and MCMC methods, to fit my *state-space model* to the data and estimate the unknown parameters and the hidden locations. To implement the model transparently and flexibly in a computer software I used JAGS which is a software to do MCMC and mostly uses Gibbs sampling. JAGS is not very complicated to use and a vast variety of models can be programmed by providing code to specify the model structure. In addition for easier and faster implementation of JAGS code, I used R packages namely *R2jags* and *dclone* that call JAGS from R.

Working with the large data sets and complex models with many latent variables using MCMC requires very long computational time. Although when I used powerful a computer system (The Abel computer cluster), I had time limitation. Therefore it was necessary to utilize some strategies to save time. The maximum time for each run was 7 days, and each setting of my model and data with 3000 iterations (samples) after 4000 burn-in took around 4-5 days.

Based on the Bayesian approach, I needed to set appropriate *prior* distributions for parameters of my process model and observation model, and then obtain the *posterior* distribution of each parameter. To reach more robust results and save computational time, first I fit only the observation model on a set of light intensity data with the known locations (from whole the data set) to estimate the parameters of the observation model via MCMC methods. Then, by using the *posterior* means from this fit as the fixed parameters, I fit the whole *state-space model* on the light intensity for a specific period of time and obtained *posterior* estimates of the process model parameters and hidden locations.

Another strategy that I utilized, was thinning the light intensity data to every forth record and also exclude the *light intensity* records between 10:00 and 15:00 o'clock from each day. In this setting my *prior* was allowed only small variation for the *distances* that the bird can pass in each day, given that it changes location. I saw partly acceptable convergences and good mixing for the MCMC iterations of the locations especially as it was closing to the end days. The estimated track of the bird was from southern of Norway to Morocco via North sea, UK, Atlantic ocean, Spain and Portugal. However, the closeness of estimated locations for day to day was not biologically plausible because generally it is expected that migrating birds flight faster and pass longer distances over water than lands.

Then I analyzed the same data but allowed the *movement distances* to have more variation among days (up to the maximum possible standard deviation of the *movement distance* for this migrating bird). I used also the known locations belonging to the first days and last days of the period. The corresponding results showed a good convergence and chains mixing for the estimate of the between day standard deviation in *movement distance*.

The *distances* between estimated locations seemed more reasonable, but the convergence and the chains mixing of other parameters was not perfect. It was not satisfiable for the most of the estimated locations. However the general track was similar to the one from the previous setting but I saw that the bird migrated further south before turning north again at the end of the track.

In the next step, I tried a new strategy to thin the data; I selected the light intensity data which had been recorded around the twilight periods (sunrise and sunset times) in each day, based on the estimated locations in the previous fit, in order to use more informative data. As I did not see a satisfiable convergence and chains mixing for the estimated locations of the previous fit, I decided again to decrease the variation of the daily *movement distances* by specifying a smaller interval for the prior distribution of the corresponding standard deviation although not as small as the one in the first setting. Then I fit the model on the new version of thinned data in two different settings (runs), one using the known locations (in the first and last days of the period) and one without using the mentioned known locations. When I got the results and compared them, I saw that the general estimated tracks are similar but there were different movement patterns in the middle days which was mainly caused by convergence problems. I did not have satisfiable convergence and chains mixing at those days. However the desired convergence and chains mixing were obtained close to the end days of the period, which was considerable.

Generally from the different model fits, we obtained partly similar tracks for the migration of the bird, started from southern Norway to Morocco via UK, the North Sea, Spain, and the Athletic Ocean. Based on the ecological knowledge about migrating of Lesser Black-backed Gulls, the estimated tracks are not far from the expectations, and they reached the final known location. However, mainly due to insufficient number of iterations in the MCMC simulations, I could not gain perfect posterior distributions for all the parameters and locations. Due to the long time needed to run the model via JAGS, using more number of iterations to obtain better precision was beyond the scope of this master thesis. Nevertheless, the model and prior distributions can also be improved.

For future works, first it is recommended to repeat all the mentioned steps in this project with more numbers of MCMC iterations if the required computational facilities are provided. It would also be a good idea to try to find more efficient ways to fit the model (e.g. more efficient MCMC sampling or other methods, such as the use of particle filters).

Then it would be useful to implement the model on different light intensity data that were obtained from different birds to see if the behaviours of migration birds which affect on the recorded light intensities (such as breeding time and nesting), can all result in realistic estimates of locations. It is also useful to analyse the data for this individual and other individuals over several years.

Another suggestion is using the information from the obtained posterior estimates to construct more sufficient prior distributions for fitting model to new data sets.

Appendix

R and JAGS Code

Here the main R code and JAGS code which were used in this project, are provided.

```
#### The main R code ####

# Necessary libraries

library(parallel)
library(dclone)
library(rgdal)
library(tripEstimation)
library(RAtmosphere)

# Reading the data set and changing the format of dates and times

my24249$=$read.table("Gull$-24249_000.lig",sep=$",",
col.names$=$ c("Ok", "Date", "DateNum", "Light"))

my24249\$Dpos <- as.POSIXct(strptime(as.character(my24249\$Date)
,"\%d/\%m/\%y \%H:\%M:\%S"), "GMT")

# Making the data from midday to midday
```

```
my24249r <- my24249[(my24249$Dpos >
as.POSIXct('2012-05-30 12:00:00', tz = "GMT")) &
(my24249$Dpos < as.POSIXct('2012-11-27 12:00:00', tz = "GMT")), ]

# Keeping every 4th record

my24249r <- my24249r[c(T,F,F,F),]

# Excluding data between 10 and 15 o'clock

my24249r$time.dd <- as.numeric(format(my24249r$Dpos, "%H"))
+ as.numeric(format(my24249r$Dpos, "%M"))/60 +
as.numeric(format(my24249r$Dpos, "%S"))/3600
my24249r <- my24249r[-which(my24249r$time.dd > 10
& my24249r$time.dd < 15),]

# Making data and other required components for analysis in JAGS

table(table(floor(julian(my24249r$Dpos))))
N.days = length(unique(floor(julian(my24249r$Dpos)))) - 1
N.per.day = nrow(my24249r)/N.days
cat('Number of days: ', N.days,
'\nObservations each day: ', N.per.day, '\n')

sun.pos = solar(my24249r$Dpos)

my24249r.matrix = matrix(my24249r$Light, nrow=N.days,
ncol=N.per.day, byrow=TRUE)

#using known locations

day1 = as.POSIXct('2012-05-30')
```

```
sights.days = floor(c(
  as.POSIXct('2012-06-01'),
  as.POSIXct('2012-06-02'),
  as.POSIXct('2012-06-03'),
  as.POSIXct('2012-06-04'),
  as.POSIXct('2012-06-05'),
  as.POSIXct('2012-06-06'),
  as.POSIXct('2012-06-07'),
  as.POSIXct('2012-06-08'),
  as.POSIXct('2012-11-21'),
  as.POSIXct('2012-11-22'),
  as.POSIXct('2012-11-23'),
  as.POSIXct('2012-11-26')) - day1 + 1)

sights.lon = c(
  7 + 30/60 + 33/3600,
  7 + 30/60 + 33/3600,
  7 + 30/60 + 33/3600,
  7 + 30/60 + 33/3600,
  7 + 30/60 + 33/3600,
  7 + 30/60 + 33/3600,
  7 + 30/60 + 33/3600,
  7 + 30/60 + 33/3600,
  -(9 + 38/60 + 54/3600),
  -(9 + 38/60 + 54/3600),
  -(9 + 38/60 + 54/3600),
  -(9 + 38/60 + 54/3600))

sights.lat = c(
  58 + 1/60 + 4/3600,
  58 + 1/60 + 4/3600,
  58 + 1/60 + 4/3600,
```

```
  58 + 1/60 + 4/3600,
  58 + 1/60 + 4/3600,
  58 + 1/60 + 4/3600,
  58 + 1/60 + 4/3600,
  58 + 1/60 + 4/3600,
  30 + 26/60 + 15/3600,
  30 + 26/60 + 15/3600,
  30 + 26/60 + 15/3600,
  30 + 26/60 + 15/3600)

sights = cbind(sights.days, sights.lon, sights.lat)

Data = list(
  Y = my24249r.matrix,
  SUN.POS = as.matrix(as.data.frame(sun.pos)),
  sights = sights,
  N.days = nrow(my24249r.matrix),
  N.per.day = ncol(my24249r.matrix),
  N.sights = nrow(sights)
)

str(Data)

init.FUN = function(){
        list(
                psi01 = runif(1, 0.11, 0.17),
                psi11 = runif(1, 0.43, 0.72),
                mu0 = runif(1, 1.7, 1.9),
                sigma = runif(1, 2, 10)
        )
}
Inits <- list(init.FUN(), init.FUN(), init.FUN())
```

```
str(Inits)

monitor <- c("lon", "lat", "psi01", "psi11", "mu0",
"sigma", "theta", "pi01", "pi11", "move", "Dis")


# Running JAGS through dclone:

(t1 <- Sys.time())
parfit <- jags.parfit(cl, data=Data, params=monitor,
model="m_sights.txt", inits=Inits, n.chains=numWorkers,
n.adapt=3000, n.update=1000, thin=1, n.iter=3000)
(t2 <- Sys.time())
t2-t1
summary(parfit)
save.image(file = "J9MA.sights.thin4.results.RData")

# Selecting records around Twilight period

load("J9MA.sights.thin4.results.RData")
PS = as.mcmc(as.matrix(parfit)) # Posterior Samples
means = apply(PS, 2, mean)
lat.means =  means[substring(names(means), 1, 3) == "lat"]
lon.means =  means[substring(names(means), 1, 3) == "lon"]


my24249 <- read.table("24249_000.lig",sep=",",col.names=
c("Ok", "Date", "DateNum", "Light"))
my24249$Dpos <- as.POSIXct(strptime(as.character(my24249$Date)
,"%d/%m/%y %H:%M:%S"), "GMT")
```

```
my24249r <- my24249[(my24249$Dpos >
as.POSIXct('2012-05-30 12:00:00', tz = "GMT"))
& (my24249$Dpos < as.POSIXct('2012-11-27 12:00:00', tz = "GMT")), ]

Jdays.tmp = unique(as.POSIXlt(my24249r$Dpos, tz="UTC")[["yday"]])

# Calculating sunset and sunrise times

set.times = as.POSIXct('2012-01-01 00:00:00', tz = "GMT")
+ Jdays*24*60*60 + suncalc(Jdays, Lat=lat.means,
Lon =lon.means, UTC=TRUE)
rise.times = as.POSIXct('2012-01-01 00:00:00', tz = "GMT")
+ (Jdays+1)*24*60*60 + suncalc(Jdays+1, Lat=lat.means,
Lon =lon.means, UTC=TRUE)$sunrise*60*60

solar.set = solar(set.times)
solar.rise = solar(rise.times)

# Extracting data (by twilight periods) to use

use.data = NULL
for(i in 1:N.days){
        set.diff = difftime(my24249r$Dpos,
        set.times[i], units = "hours")
        rise.diff = difftime(my24249r$Dpos,
        rise.times[i], units = "hours")
        set.data = my24249r[set.diff>(-1) & set.diff<2,]
        rise.data = my24249r[rise.diff>(-2) & rise.diff<1,]
        use.data = rbind(use.data, set.data, rise.data)
}

#### The main JAGS Code ####
```

```
model{

# Estimated parameters of the observation model
(I provided the approximations at the result tables)

P[1,1] <- 0.9792
P[1,2] <- 0.02061
P[1,3] <- 0.00019

P[2,1] <- 0.4976
P[2,2] <- 0.3987
P[2,3] <- 0.1037

P[3,1] <- 0.06167
P[3,2] <- 0.05823
P[3,3] <- 0.8801

# Slope (B2), intercept (A2) in class 2
# ...night:
A2[1] <- -14.11
B2[1] <- 0
# ...twilight:
A2[2] <- 9.518
B2[2] <- 28.75
# ...daytime:
A2[3] <- -17.69
B2[3] <- 0

# Precision for each time-of-day and class
Tau[1,1] <- 1000000              # night, class 1
Tau[1,2] <- 0.0202516351         # night, class 2
```

```
Tau[1,3] <- 1000000              # night, class 3
Tau[2,1] <- 1000000              # twilight, class 1
Tau[2,2] <- 0.0004211213         # twilight, class 2
Tau[2,3] <- 1000000              # twilight, class 3
Tau[3,1] <- 1000000              # daytime, class 1
Tau[3,2] <- 0.0006494444         # daytime, class 2
Tau[3,3] <- 1000000              # daytime, class 3
# For monitoring:
Sigma[1,2] <- sqrt(1/Tau[1,2])
Sigma[2,2] <- sqrt(1/Tau[2,2])
Sigma[3,2] <- sqrt(1/Tau[3,2])

# Random day-to-day variation in class 2
tau.delta <- 0.004994444
sigma.delta <- sqrt(1/tau.delta)

# Constants
Mu1 <- 0                         # Mean in class 1
Mu3 <- 64                        # Mean in class 3
piD180 <- 0.01745329 # = pi/180

# Priors for the parameters of the Observation model and
  used variables

psi01 ~ dunif(0, 0.2)
psi11 ~ dunif(0,1)
mu0 ~ dunif(0.5,20)
sigma ~ dunif(0.01, 2)
tau.Dis <- 1/(sigma*sigma)
move[1] <- 0

## Process Model
```

```
# 1st lon and lat:
lon[1] <- 7.509167 # 7+(30/60)+(33/3600)
lat[1] <- 58.01778 # 58+(1/60)+(4/3600)
for(d in 2:N.days){
    pi01[d] ~ dbern(psi01)
    pi11[d] ~ dbern(psi11)
    move[d] <- move[d-1]*pi11[d] + (1-move[d-1])*pi01[d]
    Dis[d]  ~ dnorm(mu0, tau.Dis)T(0,25)
    dis[d] <- move[d]*Dis[d]  # Movement distance
    theta[d] ~ dunif(0,6.28)

        lat[d] <- lat[d-1] + cos(theta[d])*dis[d]
        lon[d] <- lon[d-1] + sin(theta[d])*dis[d]/cos(piD180*lat[d])
}


# Observation Model:

for(d in 1:N.days){
 for(t in 1:N.per.day){

        #Computing elevation (from elevation{tripEstimation})

        hourAngle[d,t] <- SUN.POS[(d-1)*N.per.day + t, 1] +
        lon[d] - 180
        cosZenith[d,t] <- (sin(piD180 * lat[d]) *
        SUN.POS[(d-1)*N.per.day + t,2]
        + cos(piD180 * lat[d]) * SUN.POS[(d-1)*N.per.day + t,3] *
        cos(piD180 * hourAngle[d,t]))
        gt1[d,t] <- step(cosZenith[d,t]-1)
        ltm1[d,t] <- step(-1-cosZenith[d,t])
```

```
        cosZ[d,t] <- (1-gt1[d,t])*(1-ltm1[d,t])*cosZenith[d,t]
        + gt1[d,t] - ltm1[d,t]
        elevation[d,t] <- 90 - arccos(cosZ[d,t])/piD180

        # Computing time of day (1 = night, 2 = twilight, 3 = daytime)

        c1[d,t] <- step(elevation[d,t] + 6)
        c2[d,t] <- step(elevation[d,t] + 2)
        tod[d,t] <- 1 + c1[d,t] + c2[d,t]
        # Here tod[] was used for time phases

        # Mixture class
        class[d,t] ~ dcat(P[tod[d,t],])

        # Intercept and slope in class 2 depending on time-of-day

        a2[d,t] <- A2[tod[d,t]]
        b2[d,t] <- B2[tod[d,t]]

        Mu[d,t,1] <- Mu1
        Mu[d,t,2] <- a2[d,t] + b2[d,t]*elevation[d,t] + delta[d]
        Mu[d,t,3] <- Mu3

        mu[d,t] <- Mu[d,t,class[d,t]]
        tau[d,t] <- Tau[tod[d,t], class[d,t]]

        # Observation likelihood

        Y[d,t] ~ dnorm(mu[d,t], tau[d,t])T(0,64)
        }
        delta[d] ~ dnorm(0, tau.delta)
}
```

```
# Additional Observation likelihood for known locations

for(i in 1:N.sights){
        sights[i,2] ~ dnorm(lon[sights[i,1]], 1000)
        sights[i,3] ~ dnorm(lat[sights[i,1]], 1000)
}

}

#Priors for the observation model parameters

# Dirichlet priors for mixture proportions
for(k in 1:3){ # time of day
        for(i in 1:3){ # mixture class
                Pg[k,i] ~ dgamma(1,1)
                P[k,i] <- Pg[k,i]/sum(Pg[k,])
        }
}

# Slope (B2), intercept (A2) in class 2
# ...night:
A2[1] ~ dnorm(0, 0.01)
B2[1] <- 0
# ...twilight:
A2[2] ~ dnorm(0, 0.01)
B2[2] ~ dnorm(0, 0.01)T(0,)
# ...daytime:
A2[3] ~ dnorm(0, 0.01)
B2[3] <- 0
```

Bibliography

[1] Toby A Patterson et al. "State–space models of individual animal movement". In: *Trends in ecology & evolution* 23.2 (2008), pp. 87–94.

[2] Peter D Hoff. *A first course in Bayesian statistical methods*. Springer Science & Business Media, 2009.

[3] Jean-Baptiste Thiebot and David Pinaud. "Quantitative method to estimate species habitat use from light-based geolocation data". In: *Endangered Species Research* 10.1 (2010), pp. 341–353.

[4] Marc Kéry and Michael Schaub. *Bayesian population analysis using WinBUGS: a hierarchical perspective*. Academic Press, 2012.

[5] Simeon Lisovski and Steffen Hahn. "GeoLight–processing and analysing light-based geolocator data in R". In: *Methods in Ecology and Evolution* 3.6 (2012), pp. 1055–1059.

[6] Eldar Rakhimberdiev et al. "A hidden Markov model for reconstructing animal paths from solar geolocation loggers using templates for light intensity". In: *Movement ecology* 3.1 (2015), p. 1.