Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Clara Panchaud

Accounting for GPS Error in Animal Movement Studies

Master's thesis in Mathematical Sciences
Supervisor: Stefanie Muff
Co-supervisor: Johannes Signer
July 2021

NTNU
Kunnskap for en bedre verden

Clara Panchaud

Accounting for GPS Error in Animal Movement Studies

Master's thesis in Mathematical Sciences
Supervisor: Stefanie Muff
Co-supervisor: Johannes Signer
July 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU

Kunnskap for en bedre verden

Preface

This thesis is written for the course MA3911 *Master Thesis in Mathematical Sciences* as part of the Master's Degree in Mathematical Sciences at the Norwegian University of Science and Technology. It was carried out during autumn 2020 and spring 2021, supervised by Associate Professor Stefanie Muff and co-supervised by Dr. Johannes Signer.

Clara Panchaud
Trondheim, July 2021

Acknowledgments

I want to thank Stefanie Muff, as I am deeply grateful for her kindness, understanding, support and feedback throughout this thesis. I also want to thank her for suggesting the topic of this thesis, which shaped my future. Further thanks to my co-supervisor Johannes Signer for his precious help, and to David Wolfson and John Fieberg for their valuable collaboration. In addition, thanks to Benedikt Gehr for the lynx dataset.

Further thanks to Anders, Andréa, Céline, Koda, Nathalie, Nina, Nora, Ophélie and Orianna for their love and support through the years. Finally, the biggest of thanks to my parents for being my biggest motivation and bringing me to where I am today, more specifically to my mom for always being there for me and taking care of me, and to my dad who encouraged my scientific mind since I was a child, and who I know would have loved to read this thesis and ask me a million questions on it.

Funding for the crane dataset

Funding for this project was provided by the U.S. Fish and Wildlife Service and U.S. Geological Survey through Research Work Order No. 101 at the U.S. Geological Survey, Minnesota Cooperative Fish and Wildlife Research Unit; by the Minnesota Environmental and Natural Resources Trust Fund as recommended by the Legislative-Citizen Commission on Minnesota Resources (LCCMR); by the U.S. Fish and Wildlife Services Webless Migratory Game Bird Program; and by the Minnesota Department of Natural Resources.

Abstract

Animal movement studies aim to understand animal behavior by analyzing data consisting of locations visited by the animals, often collected by GPS collars. As the positioning technologies have been improving quickly over the last decades, new opportunities have arisen in animal movement studies, accompanied by challenges. In this thesis, we focus on the issue of GPS error in the animal positions collected by GPS collars. We propose to use an existing error correction method called Simulation Extrapolation, or SIMEX, in order to understand GPS error in animal movement studies, and try to account for them.

We start with a review of the common frameworks used in animal movement studies, which are Resource-Selection Functions (RSFs) and Step-Selection Functions (SSFs), with a focus on the latter. We describe these methods and their challenges, as well as their associated models. An SSF is commonly formulated as a conditional logistic model, but we also recall that it can be reformulated as a Poisson model. Furthermore, we review the need for random effects in both formulations. To display the differences between the existing methods to fit an SSF, we apply them to a dataset on lynx. Finally, we introduce the use of SIMEX in animal movement studies with simulations as well as a case study on a dataset of sandhill cranes.

The analysis on the lynx dataset revealed that the Poisson reformulation with random effects is generally the preferred model. Furthermore, the simulations showed that SIMEX is an interesting method to understand GPS error and reduce bias in the estimated parameters. The crane analysis further emphasized the conclusion from the simulations, even though some limitations of SIMEX appeared. Overall, the results suggested that GPS error is an issue that should not be ignored in animal movement studies, and that SIMEX is an easy and intuitive approach to consider as a potential solution to the issue.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Animal movement and habitat selection studies represent an important part of statistical ecology. This field aims at understanding animals' movement behaviors and resource selection, which can depend on many factors such as the weather, the presence of predators or the availability of food (Fortin et al., 2005; Thurfjell et al., 2014; Gehr et al., 2017). The understanding of those behaviors is key to answering many questions in wildlife conservation. For example it can be useful to learn more about a certain species: Do they like to stay in herds or are they solitary? Do they change territory often or are they sedentary? What resources makes them select a certain habitat? Are they easily disturbed by the presence of humans? Answering those question will bring us a better understanding of the species, but it will also help with the planning of wildlife management actions, with consequences such as the protection of endangered species and the reduction of human-wildlife conflicts (Rosenzweig, 1991; Gaillard et al., 2010; Chapron et al., 2014; Raynor et al., 2017).

The recent progress in positioning technologies is making it possible to collect data on the locations of animals in a new efficient way (Cagnacci et al., 2010; Tomkiewicz et al., 2010). With the use of Global Positioning System (GPS) collars, it is now easy to collect positions at a fine temporal scale. This makes it possible to obtain enormous amounts of data, meaning that we might be able to get a better understanding of animal movement behavior (Hebblewhite and Haydon, 2010). The simplified access to radio telemetry data has been beneficial for research, but it has also created new challenges.

This thesis focuses on a challenge that inevitably arises with data collected by GPS collars: the GPS inaccuracies. Indeed, the observed locations are not fully accurate, which can lead to erroneous estimated parameters (Jerde and Visscher, 2005; Ganskopp and Johnson, 2007; Lewis et al., 2007; Frair et al., 2004). This may be problematic, as the whole basis of animal move-

ment studies is to get a better understanding of animals, in order to take certain actions for their benefits. However, biased parameters can lead to wrong conclusions about the animal's behavior.

Some other challenges also considered in this thesis concern the models used to analyze the data collected by radio telemetry. The frameworks most commonly used are called Resource-Selection Functions (RSFs) and Step-Selection Functions (SSFs), whose main idea is the same for both, to compare the resources at the locations that an animal visited with the ones of other randomly sampled locations (Manly et al., 2002; Fortin et al., 2005). We focus on SSFs, as they have the advantage to account for temporal correlations better, by considering the animals' steps in chronological order.

We will discuss some of the challenges brought by the SSFs, which appear from the design of the experiment to the final results. For example, when designing an experiment, one starting question is how often should locations be collected. Then, when defining the SSFs, we need to decide how to sample the new locations, and how many of them are needed per observed location. Those are some of the many decisions that need to be taken by the designer of the experiment. Even though literature can assist with this decision-making, the choices are expected to depend on the specific study and can have a negative influence on the results if chosen poorly (Fortin et al., 2005; Thurfjell et al., 2014). Once the data is ready, other challenges arise. This data contains many observations on each individual, which creates a correlation that needs to be taken into account to avoid pseudoreplication (Hurlbert, 1984; Gillies et al., 2006; Fieberg et al., 2010; Muff et al., 2020). Furthermore, while the individuals of the same population tend to act similarly, there is still individual variability that needs to be considered (Fieberg et al., 2009). In order to account for this, it has been suggested to include random effects in SSFs (Duchesne et al., 2010). However, SSFs are usually fitted by conditional logistic regression (Fortin et al., 2005), which are not easy to fit when random effects are included. Nevertheless, Muff et al. (2020) proposed a reformulation to a log-linear Poisson model, which can be fitted both in a likelihood-based and in a Bayesian framework. We discuss the different methods to fit and compute the results of SSFs, in order to understand their differences and which one to choose. This topic has been studied for some years already, and we wanted to give an overview of what has been done so far.

As the main aim of this thesis is to address the issue of GPS error in animal movement studies, we propose to take advantage of an existing error correction method, that is called Simulation Extrapolation (SIMEX) (Cook and Stefanski, 1994). The method consists of two steps, a simulation step and an extrapolation step. We assume that data of animals' locations has been collected, and that a model was already selected and fitted to the data. From this point, the simulation

step consists of manually adding error to the original locations, and re-fitting the model to this blurred data. This procedure is executed with different magnitudes of error, in order to observe what effects the GPS error truly have on the estimated parameters. Once we have parameters corresponding to different error levels, we can move on to the extrapolation step. For each variable of interest, this step consists of extrapolating the parameters obtained from blurred data to obtain the coefficient that would correspond to data without error. This approach is quite easy to implement and intuitive to understand, making it a good potential option to approach the GPS error problem. This SIMEX method has already been used in different fields, including ecology with a study on pedigrees (Ponzi et al., 2019), but never on animal movement to our knowledge. We want to present a method that is easy to exploit and gives effective results.

In order to introduce SIMEX in the context of animal movement data, we used simulations as well as actual datasets, one on lynx and the other on cranes (Gehr et al., 2017; Wolfson et al., 2017, 2020). The simulations were used to test the SIMEX method in a controlled environment and understand which factors affect the analysis. The lynx data allowed us to compare different models used to fit SSFs, but SIMEX could unfortunately not be used as we had hoped, because we did in the end not gain access to the raw data. In order to illustrate how the SIMEX procedure impacts the estimated parameters, we applied it to the crane data.

The thesis is built up as follows. First, the necessary background theory is introduced in Chapter 2. This includes the existing models to study animal movement, but also the challenges linked to GPS collected data, and the SIMEX algorithm. In Chapter 3 the methods used for the different analyses are described, before presenting their results in Chapter 4. The results are then discussed in Chapter 5. Finally, we find the conclusion in Chapter 6.

Chapter 2

Background

We will start by introducing the popular frameworks used in animal movement studies, as well as the statistical models that will be used in later chapters.

We are considering animal locations data from wildlife radio telemetry. The animal positions are collected at regular time intervals by Global Positioning System (GPS) collars. We note that in this thesis we use the term GPS to mean any global positioning system, not only the American one, branded GPS. One observation in the dataset will usually contain the animal ID number, the GPS location, the time and some environmental variables at this location such as the temperature, altitude, presence of food, etc. Ideally, we would like to to have data of where an animal has been and has not been. However, it is not realistic to survey an animal's territory at all times to get this kind of information. Nevertheless, as mentioned previously, with the GPS collars we can easily access the locations where animals have indeed been. This type of data is called *presence-only* (Fithian and Hastie, 2013). The dataset is then extended by sampling random location where the animal might have been, called *available locations*, that we will consider as pseudo-absences (Warton et al., 2010). A response variable is then added to the data, which takes value 1 for observed positions and value 0 for available ones. Considering those "pseudo-absences" is a necessity in order to obtain valid results. Indeed, if we only look at the visited locations, we will see what kind of resources the animals select but this omits the fact that an animal's usage of a resource can depend on the other available resources (Manly et al., 2002).

2.1 Resource-selection functions

Resource-selection functions (RSFs) are statistical models that compare the environmental covariates between the visited locations and the available ones (Manly et al., 2002). In this case the available locations are sampled from the animal's home range. The definition of this home range is not obvious and would ideally be discussed with a specialist on the considered specie. It is important to once again mention that we do not know if a random location drawn from the home range was ever visited by the animal or not, which sets us in a use-availability design. Comparing those locations is usually done with *logistic regression* which yields similar results as the inhomogeneous Poisson process model proposed for this kind of data (Warton et al., 2010; Fithian and Hastie, 2013). What we call an RSF is the exponential of the linear predictor (Manly et al., 2002). If at each location we consider n covariates $\mathbf{x} = (x_1, ..., x_n)$ the RSF takes the following form:

$$RSF = exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)$$

where the $\boldsymbol{\beta}^\top = (\beta_1, \ldots, \beta_n)$ are the coefficients to be estimated. For $m = 1, ..., M$ animals and $j = 1, ..., J_m$ locations for animal m, the probability that a location y_{mj} with covariate vector $\mathbf{x}_{mj}$ is observed is $P(y_{mj} = 1|\mathbf{x_m}.) = \pi_{mj}$, then the logistic regression model is given by

$$y_{mj} \sim Bern(\pi_{mj}),$$

where

$$\pi_{mj} = \frac{exp(\boldsymbol{\beta}^T \mathbf{x}_{mj})}{1 + exp(\boldsymbol{\beta}^T \mathbf{x}_{mj})},$$

which corresponds to a generalized linear model. RSFs are not the main tool that will be used in this thesis, but introducing them was important in order to understand step selection functions that are discussed below, since they are closely related.

2.2 Step Selection Functions

As mentioned previously, the home range of animal can be hard to define, which can result in available locations in RSFs that are not possibly reachable by the animal because of time and distance issues. This difficulty is part of what led to the design of step selection functions (SSFs), a similar method introduced by Fortin et al. (2005). In SSFs the movement of animals is considered. The data consists of steps instead of locations, where a step consists of the line between two consecutive locations (Fortin et al., 2005). In terms of data, this means that one observation

in the dataset will contain the time, both the starting and ending location of a step, as well as the length between those two locations and the turning angle. The environmental variables of interest can be measured at different places along the step, depending on what is wanted. In SSFs, the available steps are created by sampling a new location from the turning angle and step length distributions. From one observed step a chosen amount of random steps with the same starting location is sampled, representing places where the animal could have gone instead (Fortin et al., 2005). The objective of the SSF method is to compare the environmental attributes of the observed steps with the environmental attributes of the sampled available steps. The advantage of the SSF design is that the available steps can now be considered as true absences since the animal cannot have visited more than one location at a given time. Furthermore, sampling steps can be easier than sampling locations from an animal's home range, which as we saw earlier is one of the drawbacks of RSFs (Duchesne et al., 2010).

Let's clarify the shape of the full dataset. For each animal the data is collected at regular time intervals, usually a few hours, over many days, weeks or months, depending on the study. At each time step we have a data stratum made of J observations: one realized step and $J-1$ available ones. Some studies can focus on herds of animals, in which case we will have m used steps and $J-m$ available ones in each stratum (Craiu et al., 2011). The response variable y associated with each observation is a binary variable that takes value 1 if it is a used step, and value 0 if it is an available one. All the strata generated by one animal are called a cluster. Such data containing many observations on each animal can be assumed to have some within-cluster correlation and between-cluster heterogeneity (Fieberg et al., 2010). We will later see how to account for that correlation with cluster-level random effects.

When we talk about SSF we usually mean the exponential of the linear predictor, which is the same as for a RSF, namely:

$$SSF = exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n) \ ,$$

where $\mathbf{x}^\top = (x_1, \ldots, x_n)$ are the covariates and $\boldsymbol{\beta}$ are the coefficients which in this case are estimated by conditional logistic regression, which will be discussed in Section 2.2.1. The idea is that if a step has a higher SSF value, it has higher odds of being chosen by the animal compared to the other steps that are available (Fortin et al., 2005).

Many aspects of SSFs are still being discussed and will need more time before reaching a consensus. This includes questions such as

- At which frequency should we collect locations?
- Does the frequency have to be constant or not?
- What are the temporal and spatial scales?
- Where along the step should the covariates be measured?
- How many random steps do we need to produce for one observed step?
- How do we sample available steps?

The choices made in this thesis are based on how those questions were answered in different papers. The choice of distribution that we sample the step length and turning angle from is an important factor that has to be decided and is a complicated topic in SSFs (Thurfjell et al., 2014). A method commonly used is to draw the step lengths and turning angles independently from distributions built from data on other animals of the population (Fortin et al., 2005). However, both variables could be correlated, so it can be useful to first estimate the correlation between them, to know if the respective correlation needs to be taken into account (Thurfjell et al., 2014).

2.2.1 Conditional logistic regression

The sampling design of SSFs leads to a formulation that requires a conditional logistic regression model (Compton et al., 2002; Fortin et al., 2005; Boyce, 2006). We therefore introduce this model here, and start by establishing the notation. For simplicity we will assume that the same amount of strata has been observed for each animal, that each stratum contains the same amount of observations, and that among them we only have one observed location. So, we have N animals, each of them observed over S strata. For each stratum i, $i = 1, ..., S$ we have J locations, leading to a vector of binary responses $Y_{ni} = (Y_{ni1}, ..., Y_{niJ})$ where $\sum_j Y_{nij} = 1$. There is also a $J \times p$ matrix of covariates $\mathbf{X}_{ni} = (\mathbf{x}_{ni1}, ..., \mathbf{x}_{niJ})$.
The probability that the animal n chooses location j in stratum i given $\mathbf{X}_{ni}$ is

$$P(Y_{nij} = 1|\mathbf{X}_{ni}) = \pi_{nij} = \frac{exp(\boldsymbol{\beta}^T \mathbf{x}_{nij})}{\sum_{k=1}^{J} exp(\boldsymbol{\beta}^T \mathbf{x}_{nik})},$$

where we want to estimate $\boldsymbol{\beta}$. This model can also be seen as a special case of a multinomial distribution (McCullagh and Nelder, 1989). Since for each stratum i we have the vector $\mathbf{Y}_{ni}$ that

is a single multinomial observation such that $\sum_{j=1}^{J} y_{nij} = 1$, we then have

$$P(\mathbf{Y}_{ni} = y_{ni} | \pi_{ni1}, ..., \pi_{niJ}, \sum_{j=1}^{J} y_{nij} = 1) \propto \pi_{ni1}^{y_{ni1}} ... \pi_{niJ}^{y_{niJ}} .$$

Now we can calculate the log likelihood for each stratum in our case,

$$\begin{aligned} l_{Y_{ni}}(\boldsymbol{\beta}) &= \sum_j y_{nij} log\left(\frac{exp(\boldsymbol{x_{nij}^T \beta})}{\sum_l exp(\boldsymbol{x_{nil}^T \beta})}\right) \\ &= \sum_j \left(y_{nij} \boldsymbol{x_{nij}^T \beta} - y_{nij} log\left(\sum_l exp(\boldsymbol{x_{nil}^T \beta})\right)\right) \\ &= \sum_j y_{nij} \boldsymbol{x_{nij}^T \beta} - log\left(\sum_l exp(\boldsymbol{x_{nil}^T \beta})\right) . \end{aligned}$$

This is the likelihood of stratum i of animal n. In order to get the full likelihood, we need to sum up this expression for each stratum of each animal.

2.2.2 Equivalence to Poisson model

The multinomial model of the conditional logistic regression presented above is likelihood equivalent to a log-linear Poisson model (McCullagh and Nelder, 1989). Let's have a look at why. We will focus on stratum i and location j of an animal. The animal's subscript is omitted here to keep notation simple. The log-linear poisson model can be written as follows

$$log(\mu_{ij}) = \phi_i + \mathbf{x_{ij}^T \boldsymbol{\beta}} ,$$

where ϕ_i and $\boldsymbol{\beta}$ are coefficients to be estimated, $\mathbf{x_{ij}}$ are the covariates and μ_{ij} is the parameter of a Poisson distribution such as $Y_{ij} \sim Poisson(\mu_{ij})$. We calculate the log likelihood over the strata, omitting the terms not depending on $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$

$$\begin{aligned} l_Y(\boldsymbol{\phi}, \boldsymbol{\beta}) &= \sum_{ij} (y_{ij}(\phi_i + \boldsymbol{x_{ij}^T \beta}) - exp(\phi_i + \boldsymbol{x_{ij}^T \beta})) \\ &= \sum_i \phi_i + \sum_{ij} y_{ij} \boldsymbol{x_{ij}^T \beta} - \sum_{ij} exp(\phi_i + \boldsymbol{x_{ij}^T \beta}) . \end{aligned}$$

Let's introduce a transformation of the parameters such that $\tau_i = \sum_j \mu_{ij}$. Following McCullagh and Nelder (1989), Chapter 6.4, Pages 210, we rewrite the likelihood by also adding and sub-

stracting a new term,

$$\begin{aligned} l_Y(\boldsymbol{\phi}, \boldsymbol{\beta}) &= \sum_i \phi_i + \sum_{ij} y_{ij} \boldsymbol{x_{ij}^T \beta} - \sum_{ij} e^{\phi_i + \boldsymbol{x_{ij}^T \beta}} + \sum_i log\left(\sum_j e^{\boldsymbol{x_{ij}^T \beta}}\right) - \sum_i log\left(\sum_j e^{\boldsymbol{x_{ij}^T \beta}}\right) \\ &= \sum_i log\left(\sum_j \mu_{ij}\right) + \sum_{ij} y_{ij} \boldsymbol{x_{ij}^T \beta} - \sum_{ij} \mu_{ij} - \sum_i log\left(\sum_j e^{\boldsymbol{x_{ij}^T \beta}}\right) \\ &= \sum_i (log(\tau_i) - \tau_i) + \sum_i \left(\sum_j y_{ij} \boldsymbol{x_{ij}^T \beta} - log\left(\sum_j e^{\boldsymbol{x_{ij}^T \beta}}\right)\right) \\ &= l_m(\boldsymbol{\tau}; m) + l_{Y|m}(\boldsymbol{\beta}; y) \ . \end{aligned}$$

This separates the likelihood in two parts. Only the second term depends on $\boldsymbol{\beta}$ so it can be used to estimate the coefficients, where that term corresponds to the likelihood of the previously defined multinomial model. This proves that those two models are equivalent to estimate $\boldsymbol{\beta}$. The advantage of this reformulation is that it will allow the model to be fitted as a generalized linear model, which will be computationally much more efficient than the conditional logistic regression formulation.

Note that in this new formulation there are now stratum-specific intercepts ϕ_i that need to be estimated, which can be a large computational disadvantage. However, it is actually more efficient to consider them as random effect from a normal distribution $N(0, \sigma_\phi^2)$ (Muff et al., 2020). The variance of this distribution can be set at a large value to avoid shrinkage.

2.2.3 Random Effects in Step Selection Functions

It is crucial to consider random effects in SSFs. Even though individuals of the same species can be expected to have somehow similar behaviors, there is still some heterogeneity between them that cannot be ignored (Fieberg et al., 2009). In a fixed-effect model we will obtain coefficients that apply to the whole population. Nevertheless, individuals might respond differently from one another to a change in a covariate (Gillies et al., 2006). Furthermore, there is some correlation between the different strata of one animal which is also taken into account by a model with individual-level random effects (Fieberg et al., 2010). Not considering this correlation can lead to underestimated standard errors and confidence intervals, because of a phenomenon called pseudoreplication (Hurlbert, 1984). Adding individual-level random coefficients will allow variations between individuals, while still using the information on the whole population from the data (Gillies et al., 2006; Muff et al., 2020).

Another reason to include random effects is linked to the independence from irrelevant alternatives (IIA) assumption, which in the context of habitat selection states that an animal's preference for an habitat over another does not depend on the other available habitats. However

this assumption is often violated by the habitat selection behavior of animals, so adding animal-specific random-effects to the model will relax the IIA assumption at the population level (Duchesne et al., 2010). Furthermore the availability of a resource might also have an influence on its usage, an issue once again resolved by the addition of random-effects (Gillies et al., 2006).

Thurfjell et al. (2014) noted that, despite the recommendation to use mixed effect models by Duchesne et al. (2010), almost no publications in between have used them. Then Muff et al. (2020) once again realized that the mixed effect models used in the literature were usually only using random intercepts and not random slopes. This might be due to the fact that those models have been lacking good software to fit them. A method called the two-step estimation is a solution to fit a random effects SSF, but it only gives an approximation of the results (Craiu et al., 2011). This is where the previously discussed Poisson reformulation comes into play, as it provides a new approach that allows random effects models to be directly and easily fitted (Muff et al., 2020).

Let's have a look at the random effects model. Let the random effects b_n have density $f(\boldsymbol{b_n;\theta})$ with $\boldsymbol{\theta}$ a vector of unknown parameters. The probability that animal n chooses location j in the i-th stratum is now

$$P(Y_{nij}=1|\mathbf{X})_{ni} = \int \frac{exp(\boldsymbol{x_{nij}^T\beta + b_n^t z}_{nij})}{\sum_l^k exp(\boldsymbol{x_{nil}^T\beta + b_n^t z}_{nil})} f(\mathbf{b_n};\boldsymbol{\theta}) d\mathbf{b_n} \,,$$

where $\boldsymbol{z}_{nil}$ is usually a vector containing a subset of covariates (Duchesne et al., 2010). The Poisson reformulation takes the form

$$log(\mu_{nij}) = \phi_i + \boldsymbol{\beta}^T \boldsymbol{x_{nij}} + \boldsymbol{b_n^T z_{nij}} \,.$$

Since fixed-effects models are more efficient and easier to interpret, we should not be using mixed-effects models when they are not needed. Since a mixed-effects model with zero variance-covariance in $f(\boldsymbol{b_n;\theta})$ is a fixed-effects model, an idea can be to perform a likelihood-ratio test to evaluate the need for random effects.

Two-step Estimation

Let's explain how to fit a random effect SSF using the two-step estimation introduced by Craiu et al. (2011). As its name indicates, the two-step estimation consists of two different steps. The first one consists of estimating cluster-specific parameters, by fitting a classical regression model to each individual. If there are a large number of observations for each individual, the regression coefficient estimates are approximately normal (Craiu et al., 2011). Therefore if we consider one covariate, its population-level parameter corresponds to the mean of a multivari-

ate normal distribution from which the regression coefficients previously obtained come from. The second step consists of applying the expectation-maximization algorithm, a well known method that can be used here to approximate the population-level coefficients.

The two-step approach is a good method when the number of steps per animal is large, but it often fails when individuals do not have enough variability in their movements, meaning when they do not encounter all categories of a categorical predictor (Craiu et al., 2011; Muff et al., 2020). Moreover, the two-step estimation is an approximation and does not provide exact results. It is still a popular approach, as fitting random effects SSFs is a complex task.

2.3 Integrated Step Selection Analysis (iSSA)

Integrated Step Selection Analysis (iSSA) is an extension of SSF which takes into account the dependence between the movement parameters and resource selection parameters (Avgar et al., 2016). The idea of iSSA is to simultaneously estimate both sets of parameters.

As mentioned in Section 2.2, it is common in SSFs to draw step lengths and turning angles from distributions built on data from other individuals of the same population. Then, we fit the model and estimate the resource selection parameters. This implies an independence between movement and resource selection that can be questioned. Indeed, shouldn't habitat availability affect the animal's movement patterns and its movement capacity affect its use of resources? Ignoring this dependence leads to biased estimates, according to Avgar et al. (2016).

iSSA solves the problem of dependency by including parameters representing movement in the model. Let's assume that the step length follows a Gamma distribution and the turning angle are Von Mises distributed (Avgar et al., 2016). The Von Mises distribution is often used to sample angles since it is similar to a normal distribution but on the circle. The probability density function of a Gamma distribution with shape parameter k and scale parameter θ is

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} \exp\left(\frac{-x}{\theta}\right),$$

which we can also write in the following form

$$\exp(\ln(f(x))) = \exp\left(-\ln(\Gamma(k)) - k\ln(\theta) + (k-1)\ln(x) - \frac{x}{\theta}\right),$$

so as the exponential of a linear combination of x and $\ln(x)$, where in our case x is the step length. This shows that by including the step length and log step length in the model, we will obtain their coefficients, which will correspond to a transformation of the shape and scale parameters of the Gamma distribution. Adding the cosine of the turning angle to the model will similarly lead to the Von Mises concentration parameter (Avgar et al., 2016).

In the iSSA procedure, we start by sampling available points as in SSF, with step length and turning angle distributions obtained from observed data. As in an SSF, we assign a variable with value 1 to the observed data and value 0 to the sampled data. Then, we fit a conditional logistic regression to a model containing the three movement covariates in addition to the environmental covariates. What we called SSF is now

$$SSF = \exp(\boldsymbol{\beta_1 x} + \beta_2 sl + \beta_3 \log(sl) + \beta_4 \cos(ta)).$$

Once the new parameters for the step length and turning angle distributions are obtained,they

can be combined with the tentative parameters used to sample the available steps. iSSA is very useful as it allows to test for more hypotheses on the relationship between the animal's movement and its environment.

2.4 Challenges of GPS collars

The recent use of GPS collars in animal movement studies has extended the capacity to collect data at a finer temporal scale and spatial resolution. These are great advantages but they also come with drawbacks. The first one is the costs of the collars, which can limit their usage (Morris and Conner, 2017). Those costs refer to the collars themselves but also to their batteries. A short battery life is therefore undesirable, but in the same time it is restricted by technology and by weight (Dewhirst et al., 2016). Indeed, bigger batteries would last longer, but they are too heavy to be worn by an animal. This has however improved over the past few years and will hopefully continue to do so. In any case, we would like to change the GPS collar as little as possible, since it costs money, effort and it intrudes on the animals' lives (Dewhirst et al., 2016). Collecting locations less frequently is a solution to make the battery last longer, but reducing the amount of available data can be detrimental to the statistical analyses.

Another drawback of GPS data is the errors that it contains in the form of missing data and inaccurate locations (Frair et al., 2004). We focus on the second type of error in this thesis. There are many factors contributing to the inaccuracy of a collected location, such as the vegetation, canopy cover, terrain, satellite geometry, atmospheric conditions and animal movement (Lewis et al., 2007; Frair et al., 2004; Montgomery et al., 2011; Muminov et al., 2019). Even though the error does vary according to those characteristics of the locations, it is possible to approximate the error variance of a given GPS collar, by collecting data on the collar set on the floor at a given location. Once we have an order of magnitude for the error we can ask ourselves some important questions: Does the GPS error affect the parameter estimates in animal movement studies to an extend that we have to worry? And if so, how can we account for this error? These are the questions that we will investigate.

2.4.1 GPS Error in Animal Movement Studies

GPS error influences the results in the context of animal movement studies, as has been discussed by, for example, Jerde and Visscher (2005); Ganskopp and Johnson (2007); Lewis et al. (2007); Montgomery et al. (2011); Muminov et al. (2019). Let's look at a few practical cases where GPS error was an issue. In a study on cattle behavior, Ganskopp and Johnson (2007) found out that the GPS error was not significant for moving animals, but that it was detrimental for the more static ones. Indeed, when classifying what an animal is doing, any error in the location of

a static animal will most likely misclassify it as active, but not the other way round. A paper by Muminov et al. (2019) considers GPS error in the context of virtual fence collars. The idea is to make goats wear GPS collars that can transmit a stimuli of some kind to the animal when it goes outside of its allowed grazing area. In this context even small errors can be problematic. Measurement errors also have an influence on the estimation of turning angles and step lengths (Jerde and Visscher, 2005), and the influence depends on the distance traveled by the animal between two collected locations. If locations are collected every 4 hours, it is possible that the animal walked a few kilometers in that time, and therefore, an error of 10m will not have much effect on the results in that case. Jerde and Visscher (2005) urge to take into account the known error variance of the collar in the sampling design.

It is common to think that with the recent improvement in GPS systems, the errors can be ignored. However, according to Montgomery et al. (2011), there is a complex relationship between the GPS error and the way the habitat characteristics are mapped, namely the raster resolution and patch size. So, any GPS error, even when small, coupled with non matching patch sizes and resolutions can lead to significant errors in the results of statistical inferences if ignored.

There are a few proposed solutions to account for GPS error. One of them is to use accelerometer and magnetometer data from a sensor on the collar to predict if a collected position is plausible or not(Dewhirst et al., 2016; Muminov et al., 2019). Inaccurate locations could then be rectified. The positional dilution of precision (PDOP) is a measure of the satellite geometry linked to the location error that can also be used to detect invalid locations(Lewis et al., 2007). However, screening data is likely to lead to additional biases so we have a trade-off situation between eliminating inaccurate locations and retaining the maximum amount of information.

Overall, we saw that the possible and needed accuracy depends on many factors like the way habitat characteristics are mapped, the goal of the study and the specie considered. The methods addressing GPS error that we have seen consist mostly of removing inaccurate data.

2.5 Simulation Extrapolation (SIMEX)

We will consider an existing method called Simulation Extrapolation (SIMEX) (Cook and Stefanski, 1994) to account for GPS error in the context of resource selection studies. SIMEX is a heuristic method used to account for measurement error when the error variance is known, or at least well estimated (Cook and Stefanski, 1994). Error correction usually requires the error model and its parameters to be known, but SIMEX presents a useful alternative, because it only requires one information on the error: its variance (Ponzi et al., 2019). Furthermore, it is an intuitive and easy to implement method, so it is overall an attractive option.

When we are using a regression analysis, error in the covariates is expected to introduce bias in the resulting parameters (Carroll et al., 2006). This concern lead to the first application of SIMEX, at the time where it was introduced by Cook and Stefanski (1994). It is common to think the bias caused by the error is always an attenuation, dragging the parameter towards zero. This makes sense in the context of animal movement as errors will make it seem like an individual is acting more randomly, drawing the coefficients representing resource selection to zero. Even though attenuation happens quite often, the effect of the measurement error on the parameters is determined by more factors, such as the other variables, their correlations, the model itself and the measurement error distribution (Carroll et al., 2006). Let's consider the effect of additive error on a simple linear regression $Y = \beta_0 + \beta_x X + \epsilon$, with $\epsilon \sim \mathrm{N}(0, \sigma_\epsilon^2)$, where instead of observing the covariate X we observe $W = X + U$, where U is independent of X and Y, with mean zero and variance σ_u^2. The model considered is then $Y = \beta_0 + \beta_w W + \epsilon$ and the estimated parameter from the data containing error is $\beta_w = \lambda \beta_x$, where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} < 1$$

(Carroll et al., 2006). For this specific model and error, the estimated parameter is indeed attenuated to zero.

SIMEX is based on the idea that adding more error to the data will introduce more bias (Cook and Stefanski, 1994). Let's look at how SIMEX works for one parameter of the model. SIMEX consists of first a simulation step where we add more and more error to the data and compute the estimate corresponding to each added error level. This means that we first select the variances of the errors we will add, and then for a variance α, we generate an error with this variance, add it to the data and estimate the parameter $\hat{\beta}_b(\alpha)$. For each error level we repeat this B times and find the mean

$$\hat{\beta}(\alpha) = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_b(\alpha) \, .$$

In the extrapolation step, the goal is to find a pattern between the error variances α and the parameters $\hat{\beta}_b(\alpha)$ in order to extrapolate back to zero error to obtain $\hat{\beta}_{SIMEX} = \hat{\beta}(\alpha = 0)$. A quadratic extrapolant is often used for this step, as it is considered quite stable (Carroll et al., 2006). The extrapolated parameter is an approximately consistent estimator, which means that it converges in probability to an approximation of the true parameter(Cook and Stefanski, 1994). The previously discussed attenuation pattern that appears quite often is presented in Figure 1, where we see a decrease in the parameter magnitude as more error is added. Even without the extrapolation step, the SIMEX plot offers very useful information on the relationship between the error variance and the parameter value.

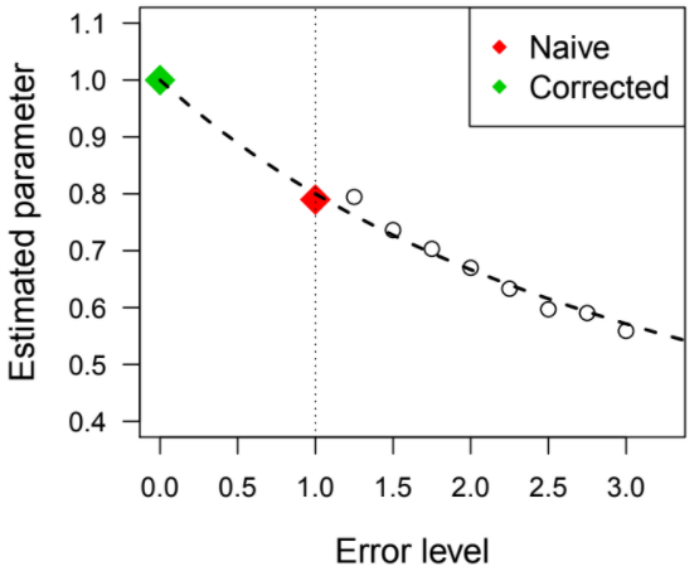


Figure 1: The general SIMEX idea, where adding more error draws the parameter to zero. The naive parameter has value 0.8, and the corrected parameter found from quadratic extrapolation has value 1.0.

The SIMEX method is quite simple, but one drawback is the difficulty to obtain standard errors for the error-corrected estimates. A solution is to use bootstrapping, but the computing costs can be too large (Carroll et al., 2006). Another way to do this is to find the two error components that make up the SIMEX parameter variance (Apanasovich et al., 2009; Ponzi et al., 2019). The first one is the variance of the estimated parameter itself. Let's say that we are at iteration b of error level α, and with the chosen estimation method we get the estimated parameter $\hat{\beta}_b(\alpha)$. The model also yields the standard error of this parameter, and so by squaring it we get the variance $\text{SD}^2(\hat{\beta}_b(\alpha))$. A way to get this variance for the SIMEX parameter is actually to apply the SIMEX procedure on it. We simply store this variance at each step and find the mean of the variance for each error level, $\text{Var}(\hat{\beta}(\alpha)) = \frac{1}{B}\sum_{b=1}^{B}\text{Var}(\hat{\beta}_b(\alpha))$. We can then use the same extrapolation method as with the parameter in order to find $\text{Var}(\hat{\beta}(\alpha = 0))$.

The second component is due to the difference in variance between each simulation for a fixed α, namely $\text{Var}(\hat{\beta}_b(\alpha)) - \text{Var}(\hat{\beta}(\alpha))$. An approximation of this value can be found as follows

$$s(\alpha) = \frac{1}{B-1}\sum_{b=1}^{B}(\text{Var}(\hat{\beta}_b(\alpha)) - \text{Var}(\hat{\beta}(\alpha)))^2$$

and again we can extrapolate in order to find $s(\alpha = 0)$. The total variance is then the difference

$\text{Var}(\hat{\beta}_{SIMEX}) = \text{Var}(\hat{\beta}(\alpha = 0)) - s(\alpha = 0)$.

SIMEX is overall a straightforward method and intuitive to understand, which allows us to understand how measurement error affect the resulting estimated parameters. The main advantage of SIMEX is that the error model for the covariates $\boldsymbol{x}$ does not need to be explicitly stated, instead, we act directly on the GPS measurement, which is where the error occurs. Furthermore, it is easy to add more error to the GPS data if we have access to the landscape variables, and the starting error variance can be estimated from repeated measurements on a collar set at a known location. Therefore, data collected by GPS collars creates a good setting to use SIMEX.

Chapter 3

Methods

Our exploration of the SIMEX method in animal movement studies is separated in three parts. We start with some simulations to see if the method worked in a controlled simulation environment and then move on to case studies on lynxes and sandhill cranes. The analyses were done with the programming language `R` (R Core Team, 2021).

3.1 Simulations

The idea of the simulations was to generate an animal moving through a landscape with a known preference for some characteristics of the habitat. The first step was to generate an environmental variable x_e from a landscape, and for this we used a Gaussian random field (GRF). The main features of a GRF are its resolution, its range of spatial autocorrelation, and the magnitude of variation of this autocorrelation. We set those values to 10, 20 and 0.001, respectively, in order to create a continuous landscape. We also created a GRF with values 10, 10, 0.001 and made a binary landscape from it. The landscape variable is the only environmental variable in the model. We considered simulations using the continuous landscape from Figure 2, and some simulations with the categorical landscape from Figure 3. In the continuous setting, we can imagine that this variable represents the temperature, while in the categorical setting, it could for example constitute the habitat type, such as forest or non forest.

Once we had a landscape, we needed to define how the animal moves in it. We followed the iSSA procedure, and defined the model as an SSF containing the three movement parameter (step length, logarithm of step length and cosine of turning angle) as well as the environmental variable of interest x_e, thus

$$SSF = \exp(\beta_1 x_e + \beta_2 sl + \beta_3 \log(sl) + \beta_4 \cos(ta)) \, . \tag{3.1}$$

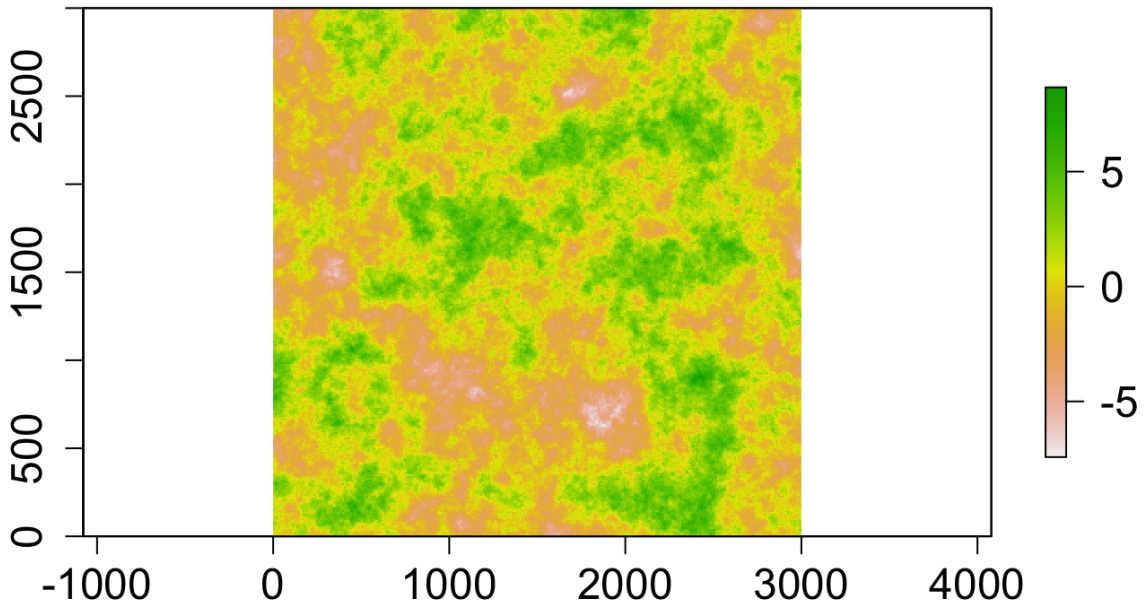


Figure 2: Simulated continuous Landscape used as the variable x_e in some simulations.

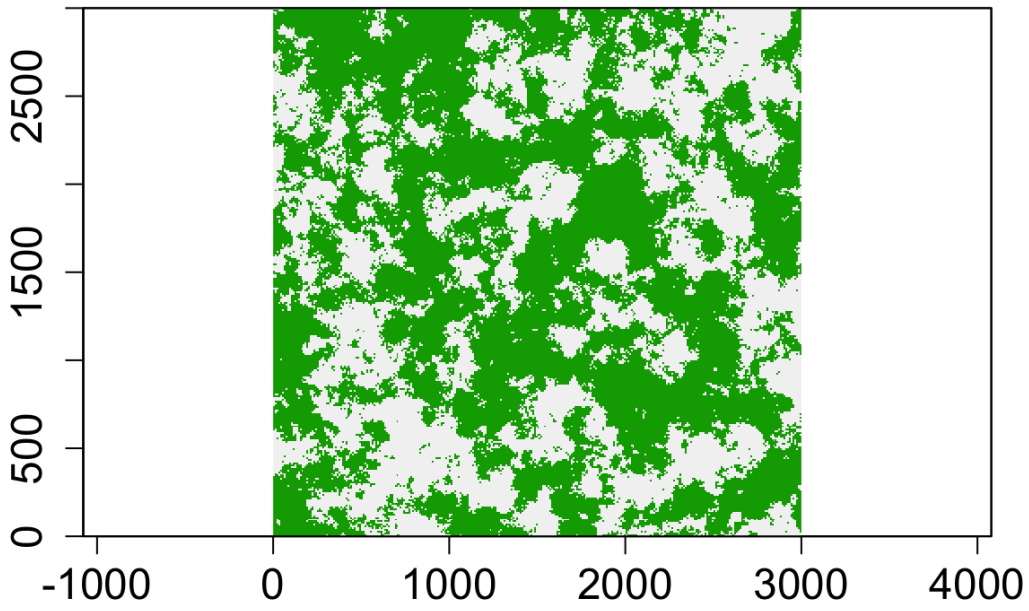


Figure 3: Simulated categorical Landscape used as the variable x_e in some simulations.

The next step was to come up with true $\boldsymbol{\beta}$ coefficients for each term. For the habitat variable, a higher coefficient means that the animal has a preference for higher values. For the continuous landscape we considered two cases, one with $\beta_1 = 0.5$ and the other with $\beta_1 = -0.5$. We called them continuous case 1 and continuous case 2 respectively. There were also two cases for the categorical landscape, with coefficients $\beta_1 = 1$ and $\beta_1 = -1$, categorical case 1 and categorical case 2. For the animal's movement through the landscape, we assigned to the step length a Gamma distribution with a shape parameter of 10 and a scale parameter of 15, which yields the probability density function

$$f(x) = \frac{1}{\Gamma(10)15^{10}} x^9 \exp\left(\frac{-x}{15}\right),$$

shown in Figure 4, indicating a typical step length of roughly 100-150m. For the turning angle, we assumed that the animal does not have a preferred direction and set the concentration parameter of the Von Mises distribution to 0. Once we had all those parameters we simulated a track with the function `simulate_track()`, from the `amt` package (Signer et al., 2019).

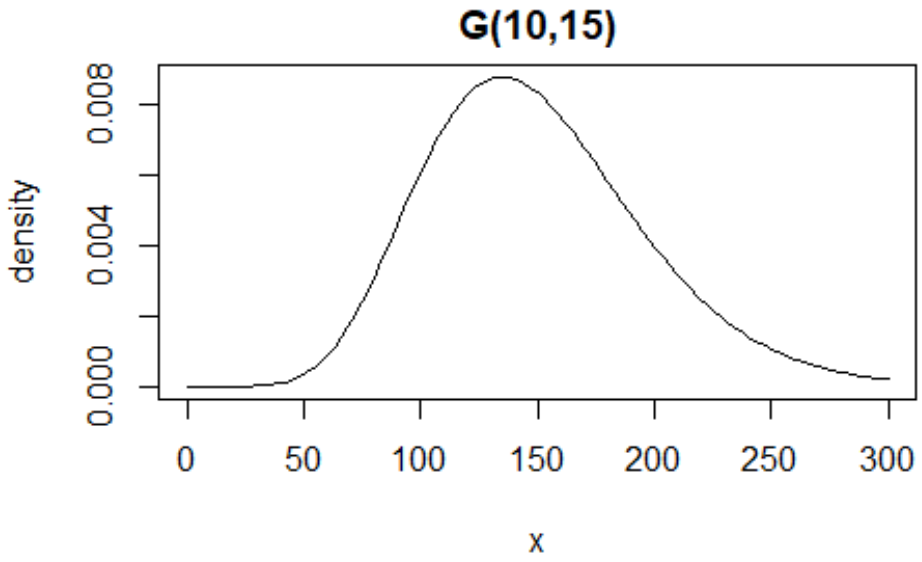


Figure 4: Gamma Distribution for the step length, with shape parameter $k = 10$ and scale $\theta = 15$.

In Figure 5 we show the simulations of 500 steps starting in the center of the landscape. The Figures (a) and (b) represent the two cases of the animal moving in the continuous landscape. The two categorical cases are displayed in Figures (c) and (d). It is quite clear what the preferred habitat is in each Figure.

With the defined landscape and coefficients, we could then start the simulation part of SIMEX, which required to add some error. We recall that an advantage of SIMEX is that the explicit

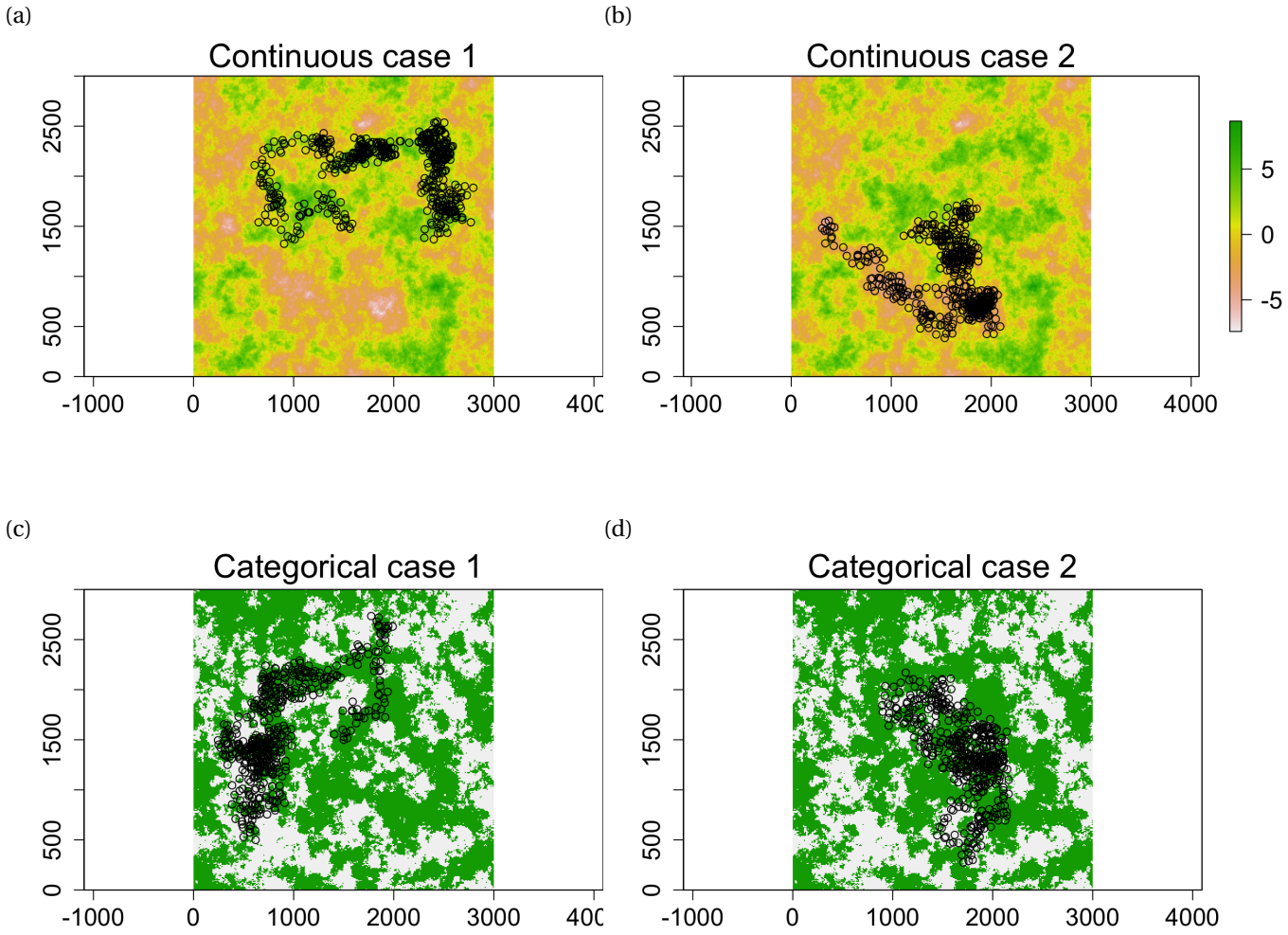


Figure 5: Plot of four different simulations, each representing 500 locations visited by the simulated animal. (a) and (b) are set in a continuous landscape, while (c) and (d) in a categorical one. Going from (a) to (d), β_1 takes the values $0.5, -0.5, 1, -1$.

error model at the level of the covariates is not necessarily required. Instead, we only needed to know the error-generating mechanism on the lowest level of the data-generating mechanism, which in our situation was the animal locations. The error-generating mechanism we used was as follows: We started by picking the variances of the errors to be added as $\boldsymbol{\sigma^2} = (5, 10, 15, 20, 25, 30, 35, 40, 45, 50)$. Let's only consider the continuous case 1 from now on, as the other cases were processed in a similar way. As a start, we simulated 50 tracks of 501 locations (500 steps), using Equation 3.1 with $\beta_1 = 0.5$. In Figure 5 (a), we find an example of locations visited by a simulated animal. Each of the simulated tracks was then blurred once with each $\sigma_i^2 \in \boldsymbol{\sigma^2}$. Blurring a track with σ_i^2 was done by sampling two independent variables

$\epsilon_j \sim N(0,\sigma_i^2)$, $j \in \{1,2\}$ for each location of the track. Since a location is defined by its $x-$axis and $y-$axis components p_x and p_y , the blurring was done by replacing those components by $p_x+\epsilon_1$ and $p_y+\epsilon_2$. This way of blurring tracks was suggested by Jerde and Visscher (2005). At the end of this blurring process, we obtained $10 \cdot 50 = 500$ blurred tracks, so a total of 550 tracks if we also count the 50 original ones.

We then fitted an SSF model to each of those 550 tracks. This was done by first sampling 10 available steps for each used one, then extracting the covariate values from the landscape, and finally fitting a conditional logistic regression using the function `fit_clogit` from the `amt` package (Signer et al., 2019). This function is used to fit a conditional logistic regression in the absence of random effects. As we considered only one individual, we indeed did not need to include random effects.

We then applied the extrapolation part of the SIMEX method to the environmental variable x_e. In order to simulate a real situation, we ignored all parameters that were estimated from the tracks with no error. We then wanted to simulate two different situations, one where the starting error variance would be 5, and one where it would be 15. We first calculated the mean parameter estimates per error level from error 5 and higher and tried a linear, quadratic and cubic extrapolation on it. The best extrapolation according to the AIC criteria was selected, following Ponzi et al. (2019). We then did the exact same thing, but considering the parameter estimates only from error variance 15 and higher. For each case, we then used the corresponding extrapolation to obtain the SIMEX parameter, corresponding to 0 error. We wanted to know if SIMEX could improve the parameter estimates obtained from the blurred data.

3.2 Lynx Data

The first dataset that we considered is the one used by Gehr et al. (2017) to investigate how the habitat selection of the Eurasian lynx is influenced by human disturbance and the availability of roe deer. We decided to reproduce the analysis of this paper and compare it to the results obtained by the Poisson reformulation. Then, we wanted to use the SIMEX approach on it.

Let's start by describing the data and the model as in Gehr et al. (2017). The lynx dataset contains 19128 locations from 13 lynx, collected in the northwestern Swiss Alps between 2011 and 2014. The GPS standard deviation had been estimated at 8.8m. The model used for the analysis is an SSF that contains two environmental covariates: the altitude and the habitat type. There are two types of habitat, open and cover, with the latter being the reference category. The two other main covariates are human disturbance and deer availability. The human disturbance index was built from the building density and the distance to the closest road. The deer availability variable is

a prediction of the probability of deer occurence, obtained from a previously fitted RSF on a roe deer dataset. In order to include temporal dynamics in the model, harmonics of time of day and day of year were created. The significant interactions between those harmonics and the four main predictors are added to the model. With the exception of habitat type, quadratic terms are also added for the main covariates. Furthermore, an interaction term between step length and human disturbance is added, as well as one between habitat type and human disturbance. The data we had access to had already been processed; it contained 13185 observed steps, each of them matched with 10 sampled available steps.

We first wanted to confirm that both models were equivalent. In order to do so, we fitted a conditional logisitic regression with `clogit` and a generalized linear model with `glmmTMB`, both with no random effects, on the lynx data. When random effects are omitted, both models are expected to give the same results. However, we are looking at multiple individuals, so a model with random effects is actually needed. The previous models were just used to confirm the equivalence between them. Therefore, we continued with the two-step estimation `glmmTMB`, which was the approach used in Gehr et al. (2017) to fit the random effects SSF. However, the two-step estimation is an approximation, so the Poisson reformulation giving precise estimates might be preferred. In order to investigate the differences between both approaches, we also fitted the Poisson reformulation with random effects using `glmmTMB`. We chose to work in a likelihood set up, as a Bayesian approach would have been less computationally efficient. The focus of this analysis was on comparing the results of the four models in order to understand their differences, and not on analyzing the parameter values, as that was already done by Gehr et al. (2017).

We then wanted to apply SIMEX on this data. However, in order to do this we needed to blur the original locations and retrieve the covariates of the blurred tracks. Unfortunately, we did not obtain the necessary landcover file containing the covariates, and could therefore not proceed forward. We had to switch to another dataset on sandhill cranes, that allowed us to try the SIMEX method.

3.3 Crane Data

The following analysis was done on a dataset containing locations of sandhill cranes. This data comes from a study of breeding sandhill crane populations in Minnesota (Wolfson et al., 2017, 2020). 34 individuals were recorded using GPS/GSM transmitters (Cellular Tracking Technologies models CTT-1060a-LB and CTT-1060-LM-BT3). We selected one individual for the analysis, with id *7J (Melby colt 1)*, as it visited a wide range of locations. The variable that we included in the model is a categorical variable that comes from a land cover layer of Minnesota. Therefore,

we filtered the data to consider only the locations in Minnesota. This left us with 7104 locations from the 17th of April 2016 until the 1st of December 2016, collected approximately every 15 minutes from sunrise to sunset.

In order to apply the SIMEX algorithm, we had to know the variance of the measurement error. Thankfully, we also had data collected at different time intervals from 12 collars put down at a known location. The error was quite different in each of the 12 collars, with standard errors from 5m to 90m. We chose to define the starting error with a standard error of 35m, which lead to a variance of $1200m^2$ when rounded.

For the simulation part of SIMEX, we had to blur the locations by adding more errors to them. Since the starting error variance was $1200m^2$, we decided to increase the variance in steps of $500m^2$ until $7000m^2$. This resulted in 14 different levels of added error variances σ_i^2. For each of those σ_i^2, we blurred the original locations 50 different times. Blurring a location was done by sampling $\epsilon_j \sim N(0, \sigma_i^2)$, $j \in \{1,2\}$ and replacing the x and y component of the location by $x+\epsilon_1$ and $y+\epsilon_2$. After the blurring was done, we had $50 \cdot 14 = 700$ blurred tracks.

We could then fit all the blurred tracks as well as the original one, in a similar way as in the simulations. For each track, we turned the locations into steps and sampled 10 available steps for each observed steps. We then extracted the covariate x_{land} from the landcover and reduced it to 7 categories of interest: wetlands, barren, developed, forest, herbaceous, planted_cultivated and water, with wetlands as the reference category. The SSF also included the step length, logarithm of step length and cosine of turning angle. It is given by the following equation:

$$SSF = \exp(\beta_1 x_{land} + \beta_2 sl + \beta_3 \log(sl) + \beta_4 \cos(ta)) \, . \tag{3.2}$$

The estimated β coefficients were obtained with the function `fit_clogit` from the `amt` package (Signer et al., 2019). The Poisson reformulation was not needed here, because we were only looking at one individual, which meant no need for random effects.

The extrapolation part was done by first finding the mean estimated parameter for each error level. Counting the parameter from the original data, this added up to 15 values. We then extrapolated the coefficient corresponding to 0 error, once again using the best out of linear, quadratic and cubic extrapolation, according to the AIC criteria.

Chapter 4

Results

4.1 Simulations

The results of the simulations indicate good results when the data without error is used (Table 1). Indeed, the mean of the parameters estimated by conditional logistic regression using `clogit` from the original tracks, $\beta_{estimated}$, is in each case close to the true value β_{true}. This confirms that the model is accurate. Adding error to the simulated animal's location caused an attenuation of the parameters towards zero for all cases, as we can see in Figure 6. This is the pattern that was expected to be found. The attenuation looks like it converges to a value higher than zero.

The parameter that results from blurring the locations with an error of variance 5 is called $\beta_{\text{err}}^{(5)}$, and similarly $\beta_{\text{err}}^{(15)}$ for a variance of 15. In the continuous case 1, the true parameter was 0.5, and the estimated parameter from the true data was 0.508 ± 0.0976. Adding errors with variance 5 made this parameter drop to 0.455 ± 0.0958, while a variance of 15 gave the parameter 0.434 ± 0.0964 (Table 1). These decreases of about 10% and 15% show that the more error we add, the more the estimate diminishes, as we had expected. In the categorical case 1, where the true value was 1, adding error had an even stronger effect on the estimated parameters. Indeed, errors with a variance of 5 lead $\beta_{\text{err}}^{(5)}$ taking the value 0.830 ± 0.331, a decrease of about 20%, and $\beta_{\text{err}}^{(15)}$ the value 0.734 ± 0.339, a decrease of about 25%. The continuous and categorical case 2 are very similar to the cases so far described, except that their numbers are negative. Furthermore, we see that the standard error do not seem affected by the added error. They are, however, much larger in the categorical cases. This might be due to the type of landscape, or to the magnitude of the true parameter.

Two extrapolations were performed for the two categorical cases as well as for the two contin-

	β_{true}	$\beta_{estimated}$	$SD_{estimated}$	$\beta_{err}^{(5)}$	$SD_{err}^{(5)}$	$\beta_{SIMEX}^{(5)}$	$SD_{SIMEX}^{(5)}$	$\beta_{err}^{(15)}$	$SD_{err}^{(15)}$	$\beta_{SIMEX}^{(15)}$	$SD_{SIMEX}^{(15)}$
continuous case 1	0.5	0.508	0.0498	0.455	0.0489	0.471	0.0496	0.434	0.0492	0.497	0.0493
continuous case 2	-0.5	-0.503	0.0499	-0.452	0.0491	-0.469	0.0498	-0.432	0.0494	-0.507	0.0496
categorical case 1	1	1.04	0.179	0.830	0.169	0.875	0.178	0.734	0.173	0.854	0.171
categorical case 2	-1	-1.007	0.133	-0.842	0.130	-0.902	0.132	-0.755	0.131	-0.930	0.131

Table 1: Estimates and standard errors of the environmental variable in the simulations. β_{true} is the true parameter used to simulate the animal's movement. $\beta_{estimated}$ is the mean of the estimated parameters from the 50 simulated tracks without error, while $\beta_{\text{err}}^{(5)}$ and $\beta_{\text{err}}^{(15)}$ are the means of the estimated parameters from the tracks blurred with an error variance of respectively 5 and 15. Finally, $\beta_{\text{SIMEX}}^{(5)}$ and $\beta_{\text{SIMEX}}^{(15)}$ are the corresponding parameter obtained from the SIMEX procedure starting from the error levels 5 and 15.

uous ones. Out of the two extrapolations. one included the parameters estimated from tracks containing error of variance 5 and above, and one with variance 15 and above. Choosing these two variances aimed to simulate two situations, where the estimated parameter from the observed data would have been β_{error5}, and $\beta_{error15}$, respectively. We tested a linear, quadratic and cubic extrapolation, and selected the best one according to the AIC. This resulted in the cubic extrapolation being selected most of the time, with the exception of the two extrapolations of the categorical case 1, and the extrapolation starting at variance 5 of the continuous case 2, which were quadratic extrapolations. When performing the extrapolation, we obtained the SIMEX parameters corresponding to a starting error of 5 and 15, $\beta_{\text{SIMEX}}^{(5)}$ and $\beta_{\text{SIMEX}}^{(15)}$. In all cases, both $\beta_{\text{SIMEX}}^{(5)}$ and $\beta_{\text{SIMEX}}^{(15)}$ are closer to $\beta_{estimated}$ than their corresponding β_{err} (Table 1). The standard errrors of the SIMEX estimates are of the same magnitude as the β_{err}s, but they are a little larger in most cases. Since the SIMEX parameters reduced some of the bias, they were expected to increase the variance. However, we expected a larger increase of the variances than what we found.

An interesting point to mention is that in the two continuous cases, $\beta_{\text{SIMEX}}^{(15)}$ is much closer to the true parameter than $\beta_{\text{SIMEX}}^{(5)}$ is. In the continuous case 1 for example, the SIMEX parameter $\beta_{\text{SIMEX}}^{(5)}$ is 0.471 ± 0.0972, while $\beta_{\text{SIMEX}}^{(15)}$ is 0.497 ± 0.0966 which is closer to 0.5. This is intriguing, as we thought that the smallest errors increments would provide the most information, and therefore lead to more accurate SIMEX parameters. However, in this case the extrapolation is better when starting further away from the true track. In any case, both SIMEX parameters still help accounting for some of the simulated GPS error. The continuous case 2 gives almost the same results, suggesting that the sign of the parameter does not play an important role. In the categorical case 1, the SIMEX parameter starting from the variance 5 errors is closer to the real parameter than the one from variance 15. Nonetheless, it is the opposite in the categorical case 2.

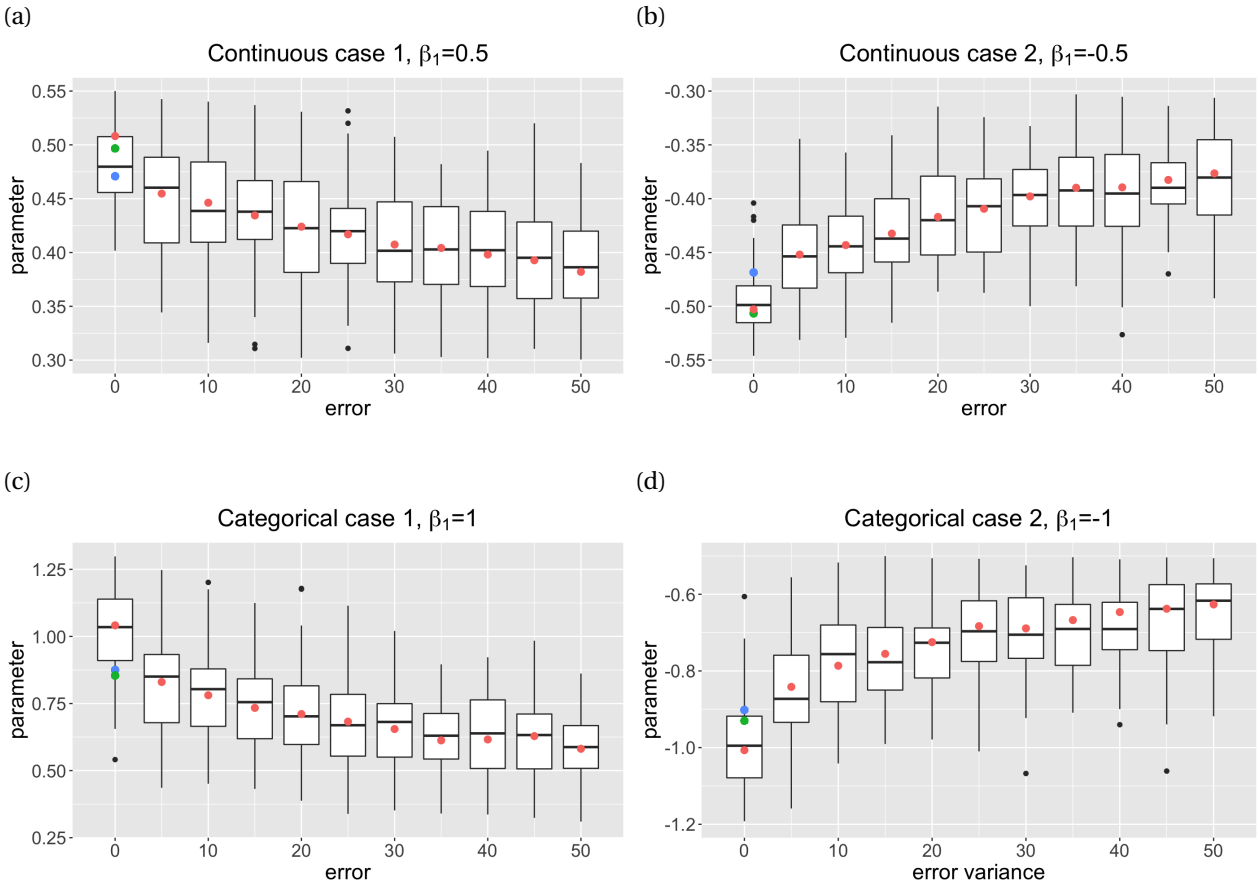


Figure 6: Effects of adding error to the locations of an animal moving through a landscape. The red points are the mean estimated parameters per error level. In blue and green we find the SIMEX estimates starting from error level 5 and 15 respectively. The results are displayed in Table 1. (a) and (b) are set in a continuous landscape, while (c) and (d) in a categorical one. Going from (a) to (d), β_1 takes the values $0.5, -0.5, 1, -1$. The error variance is increased by 5 at a time until reaching 50.

4.2 Lynx Data

The analysis of the lynx data using the Poisson model yielded similar results to those from the approximate two-step procedure (Table 2). The intercept was included in both the fixed effects and the mixed effects Poisson model, and the variable

`Deer availability:dcos` was removed from the models without random effects, as it introduced an error. The conditional logistic regression and the Poisson reformulation yield very similar results in the case of fixed effects, as expected from the Poisson reformulation. If we look at all βs in absolute values, the resulting parameters when using random effects are larger than without, suggesting an underestimation of the parameters when random effects are omitted.

	Fixed Effects				Mixed Effects			
Variable name	β_{clr}	SD_{clr}	$\beta_{Poisson}$	$SD_{Poisson}$	β_{clr}	SD_{clr}	$\beta_{Poisson}$	$SD_{Poisson}$
Intercept			-24.484	1.831			-24.768	1.827
Habitat type	-0.359	0.0235	-0.359	0.0235	-0.406	0.0627	-0.415	0.0549
Altitude	0.229	0.0255	0.229	0.0255	0.325	0.0802	0.244	0.0658
Human dist. Index	-0.172	0.0158	-0.172	0.0158	-0.147	0.0338	-0.179	0.0403
Deer availability	0.632	0.0342	0.632	0.0342	0.732	0.0659	0.676	0.0423
Altitude sq	-0.234	0.0131	-0.234	0.0131	-0.306	0.0402	-0.286	0.0336
Human dist. Index sq	-0.0518	0.00396	-0.0518	0.00396	-0.0542	0.00711	-0.0548	0.00644
Deer availability sq	-0.0837	0.920	-0.0837	0.00679	-0.0903	0.0138	-0.0819	0.00960
Step length	0.0624	0.0103	0.0624	0.0103	-0.0897	0.0873	-0.0991	0.0834
Habitat type:ycos2	-0.147	0.0304	-0.147	0.0304	-0.153	0.0339	-0.149	0.0315
Habitat type:dsin	-0.206	0.0353	-0.206	0.0353	-0.162	0.0402	-0.179	0.0367
Habitat type:dsin2	0.166	0.0357	0.166	0.0357	0.197	0.0561	0.169	0.0422
Habitat type:dcos2	0.000271	0.0322	0.000269	0.0322	-0.0984	0.0482	-0.0921	0.0404
Altitude:ycos	-0.232	0.0322	-0.232	0.0322	-0.209	0.0635	-0.253	0.0633
Altitude:ysin2	-0.231	0.0288	-0.231	0.0288	-0.289	0.0588	-0.259	0.0424
Altitude:dcos	-0.0924	0.0347	-0.0924	0.0348	-0.0623	0.0660	-0.0629	0.0690
Altitude:dsin2	-0.0799	0.0254	-0.0799	0.0254	-0.121	0.0280	-0.117	0.0271
Human dist:Step length	0.0383	0.00624	0.0383	0.00624	0.0623	0.00886	0.0628	0.00877
Human dist:ysin	0.0851	0.0113	0.0851	0.0113	0.0949	0.0244	0.0843	0.0214
Human dist:ycos	0.0823	0.0155	0.0823	0.0155	0.104	0.0248	0.0980	0.0209
Human dist:ycos2	-0.0217	0.0132	-0.0217	0.0132	-0.0297	0.0246	-0.0319	0.0183
Human dist:dsin	-0.0974	0.0129	-0.0974	0.0129	-0.123	0.0346	-0.112	0.0360
Human dist:dcos	0.117	0.0165	0.117	0.0165	0.122	0.0395	0.122	0.0368
Human dist:Habitat type	-0.0747	0.0157	-0.0747	0.0157	-0.0936	0.0326	-0.0850	0.0271
Deer availability:ysin2	0.0800	0.0216	0.0800	0.0216	0.0675	0.0355	0.0756	0.0384
Deer availability:ycos2	0.0368	0.0207	0.0369	0.0207	0.0474	0.0281	0.0428	0.0247
Deer availability:dsin	-0.108	0.0237	-0.108	0.0237	-0.0797	0.0263	-0.0839	0.0245
Deer availability:dcos					0.0953	0.0480	0.123	0.0400

Table 2: Results of the different analyses on the lynx data. For both the fixed effects and the mixed effects approaches we have the parameter β_{clr} estimated with a conditional logistic regression, the parameter $\beta_{Poisson}$ obtained from the Poisson reformulation, and their corresponding standard deviations SD_{clr} and $SD_{Poisson}$.

Furthermore, the parameters estimated with the Poisson model with random effects tend to be smaller than the ones from the two step estimation. The Poisson model is an exact reformulation of the conditional logistic regression model, so it is expected to give precise estimates, whereas the two-step estimation is an approximation. This explains the difference between both sets of parameters, and suggests that the Poisson model estimates should be preferred. The standard error are larger when using random effects.

4.3 Crane Data

For the crane data, the analysis displayed interesting results (Table 3). The SIMEX estimates are all larger, in absolute value, than their corresponding naive parameter. This would again suggest that the parameters were originally underestimated. The SIMEX standard errors are quite similar to the naive ones, which is surprising since we expected them to be larger because of the bias-variance tradeoff that is supposed to takes place. Furthermore, the attenuation pattern that we had seen in the simulations did appear for some categories, but not all of them, as can be seen in Figure 7.

	β_{naive}	SD_{naive}	β_{SIMEX}	SD_{SIMEX}
Barren	0.692	0.319		
Forest	-0.510	0.0775	-0.523	0.0784
Planted Cultivated	-0.133	0.0372	-0.150	0.0372
Herbaceous	0.263	0.233	0.727	0.225
Developed	-2.042	0.293	-2.661	0.297
Water	1.006	0.0863	1.055	0.0836

Table 3: Estimates and standard errors of the landcover variable in the crane data analysis. For each category we have β_{naive}, the estimated parameter from the original data, β_{SIMEX}, the parameter obtained from the SIMEX procedure, as well as their standard deviations SD_{naive} and SD_{SIMEX}. SIMEX was not applied on barren, because of the pattern it displayed.

At first sight, the categories planted cultivated, herbaceous, developed and water in Figure 7 (c), (d), (e) and (f), display similar patterns to the ones in the simulations. However, let's look at them more closely. We start with the category planted cultivated, in Figure 7 (c). The parameters are drawn towards zero as more error is added, which coincides with the theory. It seems that as more error is added, the parameter estimates will converge around the value -1. However, we notice that the naive parameter seems to be equal to the parameter estimated from data blurred

with an error of variance $500m^2$. This is interesting, as it might suggest that adding error has an effect on the parameter estimates only from a certain error level.

The herbaceous category presents an interesting situation in Figure 7d. Indeed, we seem to notice the same pattern as in the simulations, but actually adding more error leads the estimated parameters to cross the zero threshold before stabilizing itself around the value -0.3. The parameter estimated from the blurred data with an error variance of $7000m^2$ is around the same magnitude as the starting parameter, but with its sign reversed. So far we had seen that GPS error causes an underestimation of the parameters, but now we see a new possible effect of GPS error that goes as far as reversing the sign of the estimated parameter. We note that the herbaceous category is also the one that has the most difference between the SIMEX and the naive parameter(Table 3), with the SIMEX parameter being more than twice as large as the naive parameter.

The categories developed and water show the behavior that we had expected, similar to the simulations. The patterns are clear and adding more error drags the parameters towards zero. The scale of the effects of SIMEX is however different for those two categories. For the water category, the naive parameter is 1.006 ± 0.169 and the SIMEX parameter 1.055 ± 0.164. Even though this suggests that the naive parameter was underestimated, the SIMEX parameter is only about 5% larger. In the developed category, the naive parameter takes the value -2.042 ± 0.574 and the SIMEX parameter -2.661 ± 0.582. SIMEX increases the parameter by about 30% for the developed category, but we also note that the confidence intervals are much larger than for the water category.

We decided to not include a SIMEX estimate for the barren category, because there did not seem to be a clear pattern appearing from adding error to the data, as we can see in Figure7 (a). The estimated barren parameter takes the value $\beta_{naive} = 0.692 \pm 0.625$, but once we add some error the parameters seem to jump up and gather between 1 and 1.1. The unclear pattern and the large confidence interval made it so that it did not seem informative to perform the extrapolation. Furthermore, barren is a category that was not selected by the animal very often, which could explain what we observed.

The effect of adding error to the data influenced the forest category in a similar way than the barren category. However, the naive estimate is -0.510 ± 0.152 and all the parameters estimated from blurred tracks seem to lie between -0.48 and -0.45. We decided to extrapolate the error-prone estimates for the forest category because the confidence interval was reasonable and a pattern might still exist despite being unclear. We obtained a SIMEX parameter of -0.523 ± 0.154, so about 2.5% larger than the naive parameter. We are questioning the use of SIMEX for

this category, and we cannot claim that the SIMEX parameter is better than the naive one. It is possible that some correlations between the levels of the landcover variable are leading to unusual patterns such as the ones displayed by the barren and forest category.

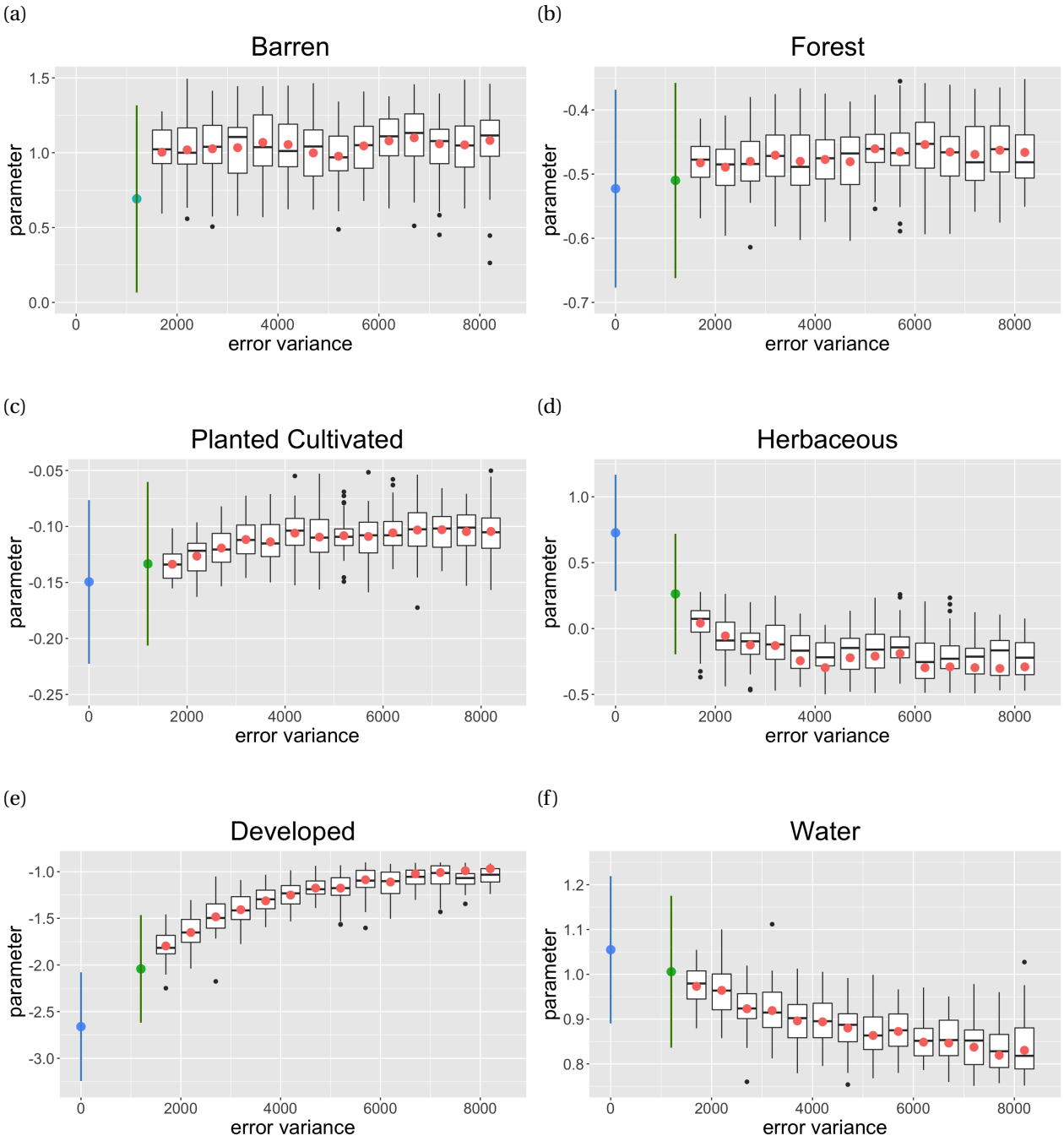


Figure 7: Results of the SIMEX procedure on the crane data. In red we find the mean estimated parameters per error level, in green the naive parameters and their confidence intervals, and in blue the SIMEX parameters and their confidence intervals. The starting error variance is $1200m^2$ and we increased that by increments of $500m^2$ until reaching $8200m^2$. Table 3 displays the results.

Chapter 5

Discussion

In the following, we will start by discussing the lynx results, as they allowed us to gain some insight on how to model animal habitat selection. Then, we consider the SIMEX results in the simulations and on the crane data. We also discuss some topics such as the challenges of designing experiments, GPS technologies and other sources of error.

Lynx analysis

We start by acknowledging the fact that the analysis on the lynx data might not seem coherent with the rest of the thesis, as it does not include an application of SIMEX. However, it is included because the analysis of the different models had already been performed when we faced the fact that it would not be able to apply SIMEX on this dataset. Furthermore, this analysis does provide an overview of different models used in animal movement studies, which is relevant information.

We wanted to reanalyze the lynx data from Gehr et al. (2017), with a similar analysis as Muff et al. (2020). In Section 2.2.2, we have seen that a conditional logistic regression is equivalent to a log linear Poisson model, when no random effects are added. This is illustrated by the results found in Table 2, showing that the resulting parameters and standard errors of these two models are almost identical. This confirms that the Poisson reformulation works well, and that considering the stratum specific intercepts as random effects with a large fun is a good way to avoid estimating all the stratum-specific intercepts as fixed effects. The `clogit` function computed the results faster than `glmmTMB`, so the former should be preferred when fitting fixed effects models.

However, as discussed in Section 2.2.3, random effects are often needed, especially with this lynx dataset where many individuals are considered. Therefore, we also fitted the same two models

with random effects. In this case, we can see some differences between the two sets of parameters. The differences might be explained by the fact that the conditional logistic regression with random effects had to be fitted with a two-step estimation, which is an approximation. On the other hand, the Poisson model with random effects could directly be fitted, which is expected to give unbiased estimates, given the modeling assumptions are met. Overall, the two-step estimation is computationally more efficient, but can give questionable results, and it might fail to run in some circumstances. Therefore, we would suggest to mainly keep the two-step estimation as a solution for when the Poisson reformulation takes too long to run, but otherwise use the Poisson model. This was already suggested by Muff et al. (2020).

We also compared the parameter estimates from the models with fixed effects to those from the random effects models. We compared both sets of results from the Poisson approach. The parameters differ a little, with the parameter estimates from the random effects model tending to be larger than those from the fixed effects model, so there might have been an underestimation of the parameters by the fixed effects model. The standard errors of the fixed effects model are way smaller than with the random effects. This can be interpreted as a sign of pseudoreplication, which was expected to result in too optimistic standard errors and biased parameters (Gillies et al., 2006; Duchesne et al., 2010; Fieberg et al., 2010; Muff et al., 2020).

To summarize the analysis of these four models, we suggest to use the `clogit` approach when fitting fixed effects models, and `glmmTMB` when fitting random effects models. Since the mixed effects models take an additional amount of time to compute, the need for random effects should first be investigated. We had mentioned in Section 2.2.3 that this could be done with a likelihood-ratio test, for example. Random effects are also usually not needed when only one individual is considered, therefore, if not many individuals are observed it can be better to fit a separate fixed effects model for each individual, as already suggested by Thurfjell et al. (2014).

Simulation analysis

With a better understanding of the models used to fit SSFs, we then moved on to our main focus on GPS error. The simulations allowed us to introduce SIMEX as an innovative method to account for GPS error in animal movement studies. The simulations considered four different cases, with differences being the type of landscape (continuous or categorical) and the value of the true parameter. We did not exactly follow the simulation part of the SIMEX method. Instead, we simulated 50 tracks that we considered as the original ones, and then blurred each of them once for each selected error variance. This resulted in 50 blurred track per error level, where each of them corresponded to exactly one original track. This was a way for us to observe what

happens when error gets added to location data.

Blurring the tracks corresponded to the simulation part of the SIMEX method. From the blurred tracks we obtained biased estimated parameters, which showed that the errors do significantly affect the results, confirming our expectations. Furthermore, for all the simulation cases, we saw a clear pattern in Figure 6, suggesting an underestimation of the parameters as more error is added. In Section 2.5, it was discussed that adding error to an animal's track makes it seem like it is acting randomly, which can be translated to habitat selection parameters being drawn to zero. Therefore, the pattern of the results is also as expected. The differences in how errors affects the results between the different simulations might be due to the type of landscape, or to the fact that the magnitude of the parameters is larger in the categorical cases. Furthermore, with the categorical landscape we have a binary variable, so a misclassified observation might have a big impact on the results, while on the continuous landscape the impact may be smoother. Moreover, the larger parameters of the categorical case represent a stronger preference for a certain type of habitat, which is then more likely to be disrupted by errors in the locations. From this simulation step, we gained useful knowledge on how the extrapolation function should look like, which motivated us to explore the extrapolation step.

The resulting SIMEX parameters reduce the bias of the estimates obtained from data containing error (Table 1). However, the parameters are not fully unbiased, which can be explained by the fact that SIMEX is ,under the right assumptions, an approximately consistent estimator 2.5. Moreover, there could be some error associated with the extrapolation. Finally, we recall that we did not exactly follow the SIMEX procedure in these simulations, so it is not surprising to not have fully unbiased SIMEX parameters. Overall, even though we cannot claim to have unbiasedness in the simulations, SIMEX does help to partially correct the underestimated parameters. The bias-variance tradeoff is surprising in these results, as we expected much larger variances for the SIMEX parameters. Our only explanation for this unexpected result is that the method to calculate the SIMEX standard error contained some approximation error.

In these simulations, we chose the values of 5 and 15 to represent starting error variances that we could have had in a real situation. We have seen that the smaller starting error does not always lead to the best SIMEX estimate, which can seem unexpected. Indeed, we thought that the tracks blurred with the smallest error would provide the most information, as they were closer to the real track. It is hard at this point to give an explanation to these differences between the SIMEX estimates obtained from starting variance 5 and 15 other than chance. However, this phenomena could suggest that extrapolating is not always the best option when the GPS error is small.

Overall in the simulations, the effects of GPS error appear to act as expected, drawing the parameters towards zero in a clear continuous pattern. We have considered two different landscapes, but it is still not quite clear how the landscape settings such as smoothness, resolution, or the type of landscape affect the results. It seems like the sign of the true parameter was not of much importance here, but its magnitude was. Some other factors had to be decided in the simulations, and would need to be thought of when designing an experiment. This includes the choice of errors that we add. In these simulations, we selected errors with variances going from 5 to 50 with increments of 5. This choice lead to clear patterns as more error is added, but we cannot claim that these numbers would work for any experiment. This is a choice that needs to be made with respect to the dataset, the landscape we are studying, and maybe some still unknown factors. We also decided that each simulated animal would walk 500 steps. Can the length of the track have an influence on the results as well? We suggest more research to be done to analyze how these different factors may influence the effects of GPS error in animal habitat selection studies. Most of them are probably correlated, and should therefore be analyzed together.

Crane analysis

The crane data allowed us to investigate SIMEX in a real setting. First of all, we obtained interesting information about how the GPS errors affect the parameter estimates from the simulation part of SIMEX. Then, we estimated the SIMEX parameters, which were all larger than the naive ones, indicating that the parameters were originally underestimated, as we had expected(Table 3). By looking at the plots of Figure 7, we observed different patterns for the different categories. In Figures 7c, 7e and 7f, we observed one type of situation that can happen, where the pattern is clear and similar to what we found in the simulations (see Figure 6). When this type of pattern appears, we suggest to go on with the extrapolation, as it seems to account for some of the GPS error. This does not necessarily mean that we will use the SIMEX parameter instead of the naive parameter in further analyzes, but it is worth it to obtain and discuss the SIMEX parameters in this type of situation. As the SIMEX parameters for the planted cultivated and the water category are not very different from their naive parameters (Table 3), we can conclude that those categories were not terribly affected by the GPS error in the original dataset, but SIMEX still helps correcting the estimates. However, the developed category's SIMEX parameter is -2.661 ± 0.582, while the naive parameter is -2.042 ± 0.574. This large difference indicates that this category is quite affected by GPS error, and that SIMEX provides a way to account for some of this error.

The herbaceous category presents an interesting situation in Figure 7d. At first sight it looks similar to the categories planted cultivated, developed water. However, despite having a pattern that is going down as more error is added, the sign of parameters also get reversed. This brings

up a new effect of GPS error, as so far, we have discussed the importance of not underestimating the parameters in an analysis, but getting a parameter of the wrong sign can have even worse consequences. Indeed, when trying to understand how an animal behaves, the management actions that will be taken depend on its habitat preferences, therefore it is important to know if the species has a preference for a herbaceous landscape or if it prefers to stay away from it. It is not obvious to understand why the sign of the parameter got reversed here, it might be due to how the landscape is laid out, or to some complex correlations among the different categories of the variable. The fact that GPS error affects the sign of a parameter is a problem that the simulation part of SIMEX can help us discover.

Another situation that can happen is observed In Figures 7a and 7b, where adding error with variance $500m^2$ or $7000m^2$ seems to have the same effect on the parameters. These Figures are not very informative in terms of analyzing the effects of GPS error. Therefore, we cannot justify the use of an extrapolation for this type of situation. We did try with the forest category, but the result is not convincing enough to suggest the use of that SIMEX parameter for further analyzes. We hypothesize that what we see in Figures 7a and 7b is the result of some complex correlations between the levels of the landcover variables.

Altogether, the results from the crane analysis show that GPS error do have an influence in a real situation. We get useful information from just the simulation step, which is why we suggest to carry out this step of the SIMEX method in all situations where it is possible. Depending on the pattern that this first step, we can then decide if we want to proceed with the extrapolation. We observed three situations that can arise in this crane dataset, but correlation in the dataset, types of landscapes and other factors might lead to different patterns.

We estimated the starting error variance to be $1200m^2$, but would like to remain critical of this choice. Indeed, to find that variance, we used validation data from 12 GPS collars that were laid out on the ground, collecting data at different time intervals, from 30 seconds to 30 minutes, over 2 days. We then calculated the average standard error of the GPS error in both the x and y direction for each of those collars. The results were varied, from 5m to 90m, so we then chose a value in between, 35m, and rounded its variance to $1200m^2$. However, the way we calculated the standard errors could contain some error, and the final choice of variance could have been different. Therefore, more attention should be given to that part of the process in the future.

General discussion

We have now seen an application of SIMEX on both simulated data and a real dataset. This is just a starting point for the use of SIMEX in animal movement studies, but it seems promising. Different parameters have to be chosen when using SIMEX. First of all, the starting error vari-

ance needs to be selected. As previously stated, this can be done by collecting data from a collar set on the floor, but new challenges rise from this experiment. Indeed, the error obtained from a GPS set on the floor will not necessarily be the same as the GPS error on the data. Some external factors will affect the error, such as the speed of the animal, the type of terrain, the vegetation, the altitude, etc. (Tomkiewicz et al., 2010; Cagnacci et al., 2010). Therefore, determining the starting error variance is not such an easy task. As already mentioned in the simulations, another important choice is the magnitude of the error variance increments that will be added to the original data. The error increments should be adapted to the scale of the starting variance, which is why Cook and Stefanski have proposed to use a factor λ, for example $\lambda \in \{1.1, 1.2, ..., 2.5\}$, which gives the different error levels when multiplied with the starting error variance Cook and Stefanski (1994). Another decision is the number of times we will blur the track for each error level. We selected the number 50, because it had been used a few times by, Cook and Stefanski (1994), when introducing SIMEX. Finally, we need to select an extrapolation function. We proceeded with the extrapolation in a similar way as Ponzi et al. (2019), but maybe it would be worth to investigate other ways to do it. Overall, those design choices should be adapted to each situation. We hope that more research will be done in order to give recommendations on how to make those choices.

There are also design choices to be made when the data is collected. In our case we obtained data to analyze, but a researcher who wants to study a certain specie will have to start by collecting data. With the advancement of GPS technologies, data can be collected at higher frequency more easily. So people tend to collect positions more often, because it gives more information about the animal's movement. However, if we look at animal's positions at a very fine time scale, GPS error can be expected to have larger effects on the results (Jerde and Visscher, 2005). Indeed, let's say we collect positions every 30 seconds, on a slow animal that moves on average 5m in 30 seconds. With a standard GPS error of 30m, we are likely to obtain steps of length more than 50m, which will show an implausible behavior for the individual. What is important is to decide on the time step between collection of locations depending on characteristics like the speed of the animal, and the GPS error of the collars that will be used. As seen in Section 2.4.1, it is also possible to screen some of the observed positions using already known information on the specie of interest.

GPS technologies are constantly improving, and will continue to do so until very high accuracy is reached. At the moment, three systems are in use: GPS, GLONASS and Galileo, who can even be combined for better precision (Kiliszek and Kroszczyński, 2020). This will hopefully lead to data containing neglectable error, which would make error correction methods, including SIMEX, obsolete. However, there is still quite some time until the GPS systems can give us

highly accurate data, and many existing datasets do contain significant GPS error. GPS error therefore remains a nuisance to animal movement studies, and a method like SIMEX seems like an interesting solution to explore.

Our main focus is on GPS error, but let's not forget that there are other types of error that can appear in this type of study. First of all, there will be some error that appears during the collection of covariates such as the temperature, the precipitation, etc. Moreover, when we obtain covariates through a landcover file, it is specified at a given scale and resolution, which are important factors. Indeed, if the GPS error is much smaller than the resolution, it might not have much influence on the results. Furthermore, representing the landscape with pixels also introduces some error. In order to understand this, we consider an example of a categorical variable that represents two types of landscapes: forests and clearings. One pixel of the landcover data might be named forest, when it actually consists of 80% forest and 20% clearing. Therefore, if the animal we are studying is resting in the clearing, we will obtain an error in the data. It could be interesting to look into using SIMEX to solve this issue. Another source of error is in the computations, as we used different models in our analyses, with some being approximations. Those are likely to lead to some approximation errors. Finally, we have considered GPS error, but have not talked about missing data, which also happens when collecting data from GPS collars. All sources of error are avenues for future research.

To conclude our anaylsis of SIMEX on both the simulated and the crane data, we give our suggestions for future work. We have already mentioned various factors that could be studied in order to come up with a concrete framework giving instructions on how to use SIMEX in animal movement studies. Even without these instructions, we encourage researchers to experiment with SIMEX in animal movement studies when the data allows it, meaning when we possess a dataset with animal locations, landscape variables and know the starting GPS error. Even without extrapolating, observing the effects of GPS error on a dataset can provide useful information. Then, it can be decided if extrapolating seems actually useful or not. In this thesis we obtained promising results that will hopefully encourage the investigation of the use of SIMEX in animal movement studies.

Chapter 6

Conclusion

In this thesis we have studied measurement error in GPS-based animal telemetry studies. We adapted an error correcting method called SIMEX to the case of GPS error. By following this approach, introduced by Cook and Stefanski (1994), we managed to obtain a better understanding of the effect of GPS error on the resulting parameters of an SSF. We saw how adding more error to the data can make a pattern appear, that can be used to extrapolate the parameter corresponding to zero error. This was illustrated with simulations, and the application to a real data example. We have also covered some tools commonly used to study animal movement and habitat selection. This included the SSF and the need for random effects to be included, as well as the reformulation to an equivalent Poisson model. We obtained a dataset on Lynx on which we applied those tools. However, the original plan was also to test the SIMEX procedure on this Lynx data, but since we did not get access to the geoinformation from the respective landcover files, we could not generate new blurred trajectories and thus not apply SIMEX.

SIMEX proved to be a simple method with great potential to correct for GPS error in animal movement studies. A great advantage is that we can use SIMEX without needing to formulate an explicit error model. Its only prerequisite is for the error-generating mechanism and the starting error to be known. Therefore, we suggest to use it whenever we have the possibility to. This means when we know the starting GPS error variance, and when the dataset contains the locations as well as the landcover, so that we can extract the covariates from it after blurring the positions. The simulation part of SIMEX will always give some information on how GPS error is potentially disturbing the results. However, we would not suggest to use the extrapolation on any dataset without checking the results from the simulation step, and the specific situation. Our simulations and case studies provided a useful starting point to go on for further analyses, but a lot of questions still need to be answered. We would need more examples of SIMEX with similar data in order to confirm its efficiency. Furthermore, since SIMEX requires the user

to make a lot of choices, future work could aim at establishing a framework for those choices. Once a framework is put into place, it could be quite simple to use SIMEX in animal movement studies.

The current progress in positioning technologies is creating new and exciting opportunities in animal movement studies. GPS error is certainly a drawback, but the GPS accuracy keeps improving and will lead to data containing less and less location error. If very high accuracy is reached, the SIMEX approach presented in this thesis might not be needed. However, in the meantime, we believe that it is important to find methods such as SIMEX to account for telemetry error in animal movement studies.

Bibliography

Apanasovich, T. V., R. J. Carroll, and A. Maity (2009). SIMEX and standard error estimation in semiparametric measurement error models. *Electronic Journal of Statistics 3*, 318.

Avgar, T., J. R. Potts, M. A. Lewis, and M. S. Boyce (2016). Integrated step selection analysis: bridging the gap between resource selection and animal movement. *Methods in Ecology and Evolution 7*(5), 619–630.

Boyce, M. S. (2006). Scale for resource selection functions. *Diversity and Distributions 12*(3), 269–276.

Cagnacci, F., L. Boitani, R. A. Powell, and M. S. Boyce (2010). Animal ecology meets GPS-based radiotelemetry: a perfect storm of opportunities and challenges.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Chapron, G., P. Kaczensky, J. D. Linnell, M. Von Arx, D. Huber, H. Andrén, J. V. López-Bao, M. Adamec, F. Álvares, O. Anders, et al. (2014). Recovery of large carnivores in europe's modern human-dominated landscapes. *science 346*(6216), 1517–1519.

Compton, B. W., J. M. Rhymer, and M. McCollough (2002). Habitat selection by wood turtles (clemmys insculpta): an application of paired logistic regression. *Ecology 83*(3), 833–843.

Cook, J. R. and L. A. Stefanski (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association 89*(428), 1314–1328.

Craiu, R. V., T. Duchesne, D. Fortin, and S. Baillargeon (2011). Conditional logistic regression with longitudinal follow-up and individual-level random coefficients: a stable and efficient two-step estimation method. *Journal of Computational and Graphical Statistics 20*(3), 767–784.

Dewhirst, O. P., H. K. Evans, K. Roskilly, R. J. Harvey, T. Y. Hubel, and A. M. Wilson (2016). Improving the accuracy of estimates of animal path and travel distance using GPS drift-corrected dead reckoning. *Ecology and evolution 6*(17), 6210–6222.

Duchesne, T., D. Fortin, and N. Courbin (2010). Mixed conditional logistic regression for habitat selection studies. *Journal of Animal Ecology 79*(3), 548–555.

Fieberg, J., J. Matthiopoulos, M. Hebblewhite, M. S. Boyce, and J. L. Frair (2010). Correlation and studies of habitat selection: problem, red herring or opportunity? *Philosophical Transactions of the Royal Society B: Biological Sciences 365*(1550), 2233–2244.

Fieberg, J., R. H. Rieger, M. C. Zicus, and J. S. Schildcrout (2009). Regression modelling of correlated data in ecology: subject-specific and population averaged response patterns. *Journal of Applied Ecology 46*(5), 1018–1025.

Fithian, W. and T. Hastie (2013). Finite-sample equivalence in statistical models for presence-only data. *The annals of applied statistics 7*(4), 1917.

Fortin, D., H. L. Beyer, M. S. Boyce, D. W. Smith, T. Duchesne, and J. S. Mao (2005). Wolves influence elk movements: behavior shapes a trophic cascade in yellowstone national park. *Ecology 86*(5), 1320–1330.

Frair, J. L., S. E. Nielsen, E. H. Merrill, S. R. Lele, M. S. Boyce, R. H. Munro, G. B. Stenhouse, and H. L. Beyer (2004). Removing GPS collar bias in habitat selection studies. *Journal of Applied Ecology 41*(2), 201–212.

Gaillard, J.-M., M. Hebblewhite, A. Loison, M. Fuller, R. Powell, M. Basille, and B. Van Moorter (2010). Habitat–performance relationships: finding the right metric at a given spatial scale. *Philosophical Transactions of the Royal Society B: Biological Sciences 365*(1550), 2255–2265.

Ganskopp, D. C. and D. D. Johnson (2007). GPS error in studies addressing animal movements and activities. *Rangeland ecology & management 60*(4), 350–358.

Gehr, B., E. J. Hofer, S. Muff, A. Ryser, E. Vimercati, K. Vogt, and L. F. Keller (2017). A landscape of coexistence for a large predator in a human dominated landscape. *Oikos 126*(10), 1389–1399.

Gillies, C. S., M. Hebblewhite, S. E. Nielsen, M. A. Krawchuk, C. L. Aldridge, J. L. Frair, D. J. Saher, C. E. Stevens, and C. L. Jerde (2006). Application of random effects to the study of resource selection by animals. *Journal of Animal Ecology 75*(4), 887–898.

Hebblewhite, M. and D. T. Haydon (2010). Distinguishing technology from biology: a critical review of the use of gps telemetry data in ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences 365*(1550), 2303–2312.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological monographs 54*(2), 187–211.

Jerde, C. L. and D. R. Visscher (2005). GPS measurement error influences on movement model parameterization. *Ecological Applications 15*(3), 806–810.

Kiliszek, D. and K. Kroszczyński (2020). Performance of the precise point positioning method along with the development of GPS, GLONASS and Galileo systems. *Measurement 164*, 108009.

Lewis, J. S., J. L. Rachlow, E. O. Garton, and L. A. Vierling (2007). Effects of habitat on GPS collar performance: using data screening to reduce location error. *Journal of applied ecology 44*(3), 663–671.

Manly, B., L. McDonald, D. L. Thomas, T. L. McDonald, and W. P. Erickson (2002). *Resource selection by animals: statistical design and analysis for field studies*. Springer Science & Business Media.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. London: Chapman Hall / CRC.

Montgomery, R. A., G. J. Roloff, and J. M. V. Hoef (2011). Implications of ignoring telemetry error on inference in wildlife resource use models. *The Journal of Wildlife Management 75*(3), 702–708.

Morris, G. and L. M. Conner (2017). Assessment of accuracy, fix success rate, and use of estimated horizontal position error (EHPE) to filter inaccurate data collected by a common commercially available GPS logger. *PLoS One 12*(11), e0189020.

Muff, S., J. Signer, and J. Fieberg (2020). Accounting for individual-specific variation in habitat-selection studies: Efficient estimation of mixed-effects models using Bayesian or frequentist computation. *Journal of Animal Ecology 89*(1), 80–92.

Muminov, A., O. Sattarov, C. W. Lee, H. K. Kang, M.-C. Ko, R. Oh, J. Ahn, H. J. Oh, and H. S. Jeon (2019). Reducing GPS error for smart collars based on animal's behavior. *Applied Sciences 9*(16), 3408.

Ponzi, E., L. F. Keller, and S. Muff (2019). The simulation extrapolation technique meets ecology and evolution: A general and intuitive method to account for measurement error. *Methods in Ecology and Evolution 10*(10), 1734–1748.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raynor, E. J., H. L. Beyer, J. M. Briggs, and A. Joern (2017). Complex variation in habitat selection strategies among individuals driven by extrinsic factors. *Ecology and Evolution 7*(6), 1802–1822.

Rosenzweig, M. L. (1991). Habitat selection and population interactions: the search for mechanism. *The American Naturalist 137*, S5–S28.

Signer, J., J. Fieberg, and T. Avgar (2019). Animal movement tools (amt): R package for managing tracking data and conducting habitat selection analyses. *Ecology and Evolution 9*, 880–890.

Thurfjell, H., S. Ciuti, and M. S. Boyce (2014). Applications of step-selection functions in ecology and conservation. *Movement ecology 2*(1), 4.

Tomkiewicz, S. M., M. R. Fuller, J. G. Kie, and K. K. Bates (2010). Global positioning system and associated technologies in animal behaviour and ecological research. *Philosophical Transactions of the Royal Society B: Biological Sciences 365*(1550), 2163–2176.

Warton, D. I., L. C. Shepherd, et al. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics 4*(3), 1383–1402.

Wolfson, D., J. Fieberg, J. S. Lawrence, T. R. Cooper, and D. E. Andersen (2017). Range overlap between mid-continent and Eastern sandhill cranes revealed by GPS-tracking. *Wildlife Society Bulletin 41*(3), 489–498.

Wolfson, D. W., J. R. Fieberg, and D. E. Andersen (2020). Juvenile Sandhill Cranes exhibit wider ranging and more exploratory movements than adults during the breeding season. *Ibis 162*(2), 556–562.

Appendix A

Code

The essential of the code used in the analyses is displayed in the next pages. The more detailed and latest version (which may be modified after completion of this thesis) can be found at https://github.com/clarapasu/Master_Thesis.

Simulations

Johannes Signer and Clara Panchaud

05/07/2021

```
library(sf)
library(raster)
library(amt)
library(tidyverse)
library(NLMR)
library(lubridate)
library(glmmTMB)
library(ggplot2)
library(gridExtra)
library(ggpubr)
source("/Users/clara/Documents/Master Thesis/Simulations/functions.R")
```

First, we generate a landscape on which the animal moves. We will use a Gaussian field. We also need to set the coefficients that will be used to simulate the trajectory. The ones for step length and the logarithm of step length are transformations of the parameters of a Gamma distribution G(10,15). We pick a concentration parameter of 0 for the Von Mises distribution of the turning angle. A transformation of this parameter will be the coefficient of the cosine of the turning angle in the model. We define different sets of coefficients, where the variable coefficient takes the values 0.5,-0.5,1,-1. We will only show the code used to perform the analysis on the continuous landscape with the variable coefficient 0.5, as the other cases are done similarly.

```
set.seed(124)
formula <- ~ var_end + log_sl_ + sl_ + cos_ta_
coefs <- c("var_end" = 0.5, scale_to_sl(15), shape_to_log_sl(10), kappa_to_cos_ta(0))
coefs2 <- c("var_end" = -0.5, scale_to_sl(15), shape_to_log_sl(10), kappa_to_cos_ta(0))
coefs3 <- c("var_end" = 1, scale_to_sl(15), shape_to_log_sl(10), kappa_to_cos_ta(0))
coefs4 <- c("var_end" = -1, scale_to_sl(15), shape_to_log_sl(10), kappa_to_cos_ta(0))
```

Here is how to generate the categorical landscape that we will use.

```
lscp2 <- NLMR::nlm_gaussianfield(300, 300,nug=0.001,
         resolution = 10,autocorr_range=20,user_seed = 3,rescale = FALSE)
lscp2 <- stack(lscp2)
names(lscp2) <- "var"

plot(lscp2,cex.main=2, cex.axis=1.5,legend.width=0.8,
     legend.shrink=0.7,axis.args=list( cex.axis=1.5))
```

And how to generate a track on this landscape.

```
set.seed(7)

trk <- simulate_track(formula, coefficients = coefs,
       start = c(1500,1500), spatial.covars = lscp2,max.dist = 100, n = 500)
plot(lscp2,xlim=c(0,3000),main=expression(paste("Continuous case 1")),
     cex.main=2, cex.axis=1.5,legend=FALSE)
points(trk$x_, trk$y_)
```

Here is just a little test to show that the conditional logistic regression model and the Poisson model yield the same results.

```
set.seed(123)

m<-trk %>% steps %>% random_steps() %>%
    extract_covariates(lscp2) %>%
    mutate(log_sl_ = log(sl_), cos_ta_ = cos(ta_))

m2<-fit_clogit(case_ ~ var + sl_ + log_sl_ + cos(ta_) + strata(step_id_),data=m)

TMBStruc = glmmTMB(case_ ~ var + sl_ + log_sl_ + cos(ta_) + (1|step_id_),
                   family=poisson,
                   data=m,
                   doFit=FALSE)

TMBStruc$parameters$theta[1] = log(1e3)
TMBStruc$mapArg = list(theta=factor(c(NA)))
m1 = glmmTMB:::fitTMB(TMBStruc)

c(summary(m2)$coef[1,1],summary(m1)$coef$cond [2,1] ,
  summary(m2)$coef[1,3],summary(m1)$coef$cond [2,2] )
```

```
## [1] 0.45664795 0.45664854 0.04032624 0.04032623
```

We then simulate and blur 50 trajectories, to observe the simulation step of SIMEX, even though we do it a little differently than in the original algorithm here.

```
set.seed(123)
run=1

if (run==0){
  param <- data.frame(matrix(ncol = 3, nrow = 0))
  names <- c("variable","sd","error")
  colnames(param) <- names
  start_variance<-2.5
  variance<-c(2.5,5,7.5,12.5,17.5,22.5,27.5,32.5,37.5)
  for (i in 1:50) {

    trk <- simulate_track(formula, coefficients = coefs,
           start = c(1500, 1500), spatial.covars = lscp2,max.dist = 50, n = 500)
    param[nrow(param) + 1,]=c(fit(trk,lscp2),0)

    trk_blur<-blur(trk,start_variance)
```

```
    param[nrow(param) + 1,]=c(fit(trk_blur,lscp2),start_variance)

    for (j in 1:length(variance)) {
      trk_new<-blur(trk_blur,variance[j])
      param[nrow(param) + 1,]=c(fit(trk_new,lscp2),variance[j]+start_variance)
    }
  }
  write.csv(param,"/Users/clara/Documents/Master Thesis/Simulations/simex1.csv",
            row.names = FALSE)
}
if (run==1){
  param=read.csv("/Users/clara/Documents/Master Thesis/Simulations/simex1.csv")
}
```

Now, we can group the results by error level and find the mean parameter. Then, we extrapolate back to 0 error, once starting from error variance 5 and once 15. For each extrapolation we use the AIC criteria to pick the type of extrapolation.

```
set.seed(123)
df<-param %>% group_by(error) %>% summarise(variable=mean(variable))
df2<-subset(df,error!=0)
df20<-subset(df,error!=0 & error!=5 & error!=10)

standard_error<-param %>% group_by(error) %>% summarise(variable=mean(sd))

fit = lm(variable ~ error, data = df2)
fit2 = lm(variable ~ error+I(error^2), data = df2)
fit3 = lm(variable ~ error+I(error^2)+I(error^3), data = df2)

c(AIC(fit),AIC(fit2),AIC(fit3))

fits = lm(variable ~ error, data = df20)
fit2s = lm(variable ~ error+I(error^2), data = df20)
fit3s = lm(variable ~ error+I(error^2)+I(error^3), data = df20)

c(AIC(fits),AIC(fit2s),AIC(fit3s))

new_df<-data.frame(error=0)
p<-predict(fit3, newdata = new_df, interval = "confidence", type = "response")
p2<-predict(fit3s, newdata = new_df, interval = "confidence", type = "response")

plot1<-ggplot(data=param, aes(group=error,x=error,
       y=variable)) + geom_boxplot()+geom_point(aes(y = p[1],
       x = 0,colour="SIMEX parameter"),size=2.5)+geom_point(aes(y = p2[1],
       x = 0,colour="SIMEX 2" ),size=2.5)+
       theme(legend.position="none",legend.text = element_text(size=11))+
       ggtitle(expression(paste("Continuous case 1, ",
       beta[1],"=0.5")))+theme(axis.title=element_text(size=18),
       axis.text=element_text(size=15),plot.title = element_text(size=20,
       hjust = 0.5))+ylim(0.30,0.55)+geom_point(data = df,
       mapping = aes(x = error, y = variable,
       color="Mean estimated parameters per error"),size=2.5)+ylab("parameter")
```

Lynx analysis

Benedikt Gehr, Stefanie Muff and Clara Panchaud

05/07/2021

We start by loading the lynx dataset.

```
library(survival)
library(glmmTMB)
library(TwoStepCLogit)
load("/Users/Clara/Documents/Master Thesis/Data/Lynx/data/lynx.RData")
dat <- lynx_table
```

First, we fit the models with fixed effects only. We want to check that the conditional logistic regression model gives the same results as the poisson reformulation. We start by fitting the condition logistic regression model.

```
run_clog=0
if (run_clog==1){
r.clogit <- clogit(formula=use~cover_swisstopo +altitude_swisstopo+
                     cover_swisstopo:ytcos2 +
                     cover_swisstopo:tsin +
                     cover_swisstopo:tsin2 +
                     cover_swisstopo:tcos2 +
                     altitude_swisstopo +
                     hum_indx +
                     prey_avail +
                     I(altitude_swisstopo^2) +
                     I(hum_indx^2) +
                     I(prey_avail^2) +
                     altitude_swisstopo:ytcos +
                   altitude_swisstopo:ytsin2 +
                   altitude_swisstopo:tcos +
                   altitude_swisstopo:tsin2 +
                   dist2 +
                   hum_indx_loc2:dist2 +
                   hum_indx:ytsin +
                   hum_indx:ytcos +
                   hum_indx:ytcos2 +
                   hum_indx:tsin +
                   hum_indx:tcos +
                   hum_indx:cover_swisstopo +
                   prey_avail:ytsin2 +
                   prey_avail:ytcos2 +
                   prey_avail:tsin +
                   strata(loc_id), data=dat)
```

```
clog<-data.frame(summary(r.clogit)$coef)

write.csv(clog,"/Users/Clara/Documents/Master Thesis/Data/Lynx/clogit.csv",
          row.names = TRUE)
}
if (run_clog==0){
  clog=read.csv("/Users/Clara/Documents/Master Thesis/Data/Lynx/clogit.csv",
                row.names = 1)
}
```

Now, we fit the Poisson model and compare the results.

```
run_glmm=0
if (run_glmm==1){

TMBStruc = glmmTMB(use~cover_swisstopo +altitude_swisstopo+
                   cover_swisstopo:ytcos2 +
                    cover_swisstopo:tsin +
                    cover_swisstopo:tsin2 +
                    cover_swisstopo:tcos2 +
                    altitude_swisstopo +
                    hum_indx +
                    prey_avail +
                    I(altitude_swisstopo^2) +
                    I(hum_indx^2) +
                    I(prey_avail^2) +
                    altitude_swisstopo:ytcos +
                    altitude_swisstopo:ytsin2 +
                    altitude_swisstopo:tcos +
                    altitude_swisstopo:tsin2 +
                    dist2 +
                    hum_indx_loc2:dist2 +
                    hum_indx:ytsin +
                    hum_indx:ytcos +
                    hum_indx:ytcos2 +
                    hum_indx:tsin +
                    hum_indx:tcos +
                    hum_indx:cover_swisstopo +
                    prey_avail:ytsin2 +
                    prey_avail:ytcos2 +
                    prey_avail:tsin +
                    #prey_avail:tcos+
                    (1|loc_id),
                    family=poisson,
                    data=dat,
                    doFit=FALSE)



TMBStruc$parameters$theta[1] = log(1e3)

TMBStruc$mapArg = list(theta=factor(c(NA)))
m1 = glmmTMB:::fitTMB(TMBStruc)
```

```
summary(m1)
glmm=summary(m1)$coef[1]
write.csv(glmm,"/Users/Clara/Documents/Master Thesis/Data/Lynx/glmm.csv",
          row.names = TRUE)
}
if (run_glmm==0){
  glmm=read.csv("/Users/Clara/Documents/Master Thesis/Data/Lynx/glmm.csv",
                row.names = 1)
}

round(clog[1],digit=2)==round(glmm[2:27,1],digit=2)
```

We can then move on to the random effects models. We start with the two-step estimation in order to fit a conditional logistic regression with random effects, which gives the same results as the paper that was published on this data.

```
run_twostep=0

if (run_twostep==1){
lynx_model <-Ts.estim(formula = use~cover_swisstopo +
                        cover_swisstopo:ytcos2 +
                        cover_swisstopo:tsin +
                        cover_swisstopo:tsin2 +
                        cover_swisstopo:tcos2 +
                        altitude_swisstopo +
                        hum_indx +
                        prey_avail +
                        I(altitude_swisstopo^2) +
                        I(hum_indx^2) +
                        I(prey_avail^2) +
                        altitude_swisstopo:ytcos +
                        altitude_swisstopo:ytsin2 +
                        altitude_swisstopo:tcos +
                        altitude_swisstopo:tsin2 +
                        dist2 +
                        hum_indx_loc2:dist2 +
                        hum_indx:ytsin +
                        hum_indx:ytcos +
                        hum_indx:ytcos2 +
                        hum_indx:tsin +
                        hum_indx:tcos +
                        hum_indx:cover_swisstopo +
                        prey_avail:ytsin2 +
                        prey_avail:ytcos2 +
                        prey_avail:tsin +
                        prey_avail:tcos +
                        strata(loc_id)+cluster(id_anim), data = dat)
#lynx_model$r.effect
twostep<-cbind(beta=lynx_model$beta, se =lynx_model$se)

write.csv(twostep,"/Users/Clara/Documents/Master Thesis/Data/Lynx/twostep.csv",
          row.names = TRUE)
}
```

```
if (run_twostep==0){
  twostep=read.csv("/Users/Clara/Documents/Master Thesis/Data/Lynx/twostep.csv",
                   row.names = 1)
}
```

And finally, the Poisson model with random effects.

```
run_random=0

if (run_random==1){
TMBStruc = glmmTMB(use~cover_swisstopo +altitude_swisstopo+
                     cover_swisstopo:ytcos2 +
                     cover_swisstopo:tsin +
                     cover_swisstopo:tsin2 +
                     cover_swisstopo:tcos2 +
                     hum_indx +
                     prey_avail +
                     I(altitude_swisstopo^2) +
                     I(hum_indx^2) +
                     I(prey_avail^2) +
                     altitude_swisstopo:ytcos +
                   altitude_swisstopo:ytsin2 +
                   altitude_swisstopo:tcos +
                   altitude_swisstopo:tsin2 +
                   dist2 +
                   hum_indx_loc2:dist2 +
                   hum_indx:ytsin +
                   hum_indx:ytcos +
                   hum_indx:ytcos2 +
                   hum_indx:tsin +
                   hum_indx:tcos +
                   hum_indx:cover_swisstopo +
                   prey_avail:ytsin2 +
                   prey_avail:ytcos2 +
                   prey_avail:tsin +
                   prey_avail:tcos+
                     (1|loc_id)+
                    (0+cover_swisstopo|id_anim) +
                     (0+altitude_swisstopo|id_anim)+
                     (0+hum_indx|id_anim)+
                     (0+prey_avail|id_anim)+
                     (0+cover_swisstopo:ytcos2|id_anim)+
                     (0+cover_swisstopo:tsin|id_anim)+
                     (0+cover_swisstopo:tsin2|id_anim)+
                     (0+cover_swisstopo:tcos2|id_anim)+
                     (0+I(altitude_swisstopo^2)|id_anim)+
                   (0+I(hum_indx^2)|id_anim)+
                   (0+I(prey_avail^2)|id_anim)+
                     (0+altitude_swisstopo:ytcos|id_anim)+
                   (0+altitude_swisstopo:ytsin2|id_anim)+
                   (0+altitude_swisstopo:tcos|id_anim)+
                   (0+altitude_swisstopo:tsin2|id_anim)+
                     (0+hum_indx_loc2:dist2|id_anim) +
```

```
                      (0+hum_indx:ytsin|id_anim) +
                      (0+hum_indx:ytcos|id_anim) +
                      (0+hum_indx:ytcos2|id_anim) +
                      (0+hum_indx:tsin|id_anim) +
                      (0+hum_indx:tcos|id_anim) +
                      (0+hum_indx:cover_swisstopo|id_anim) +
                      (0+prey_avail:ytsin2|id_anim) +
                      (0+prey_avail:ytcos2|id_anim) +
                      (0+prey_avail:tsin|id_anim) +
                      (0+prey_avail:tcos|id_anim) +
                    (0+dist2|id_anim),
                    family=poisson,
                    data=dat,
                    doFit=FALSE)



TMBStruc$parameters$theta[1] = log(1e3)
TMBStruc$mapArg = list(theta=factor(c(NA,1:27)))
m = glmmTMB:::fitTMB(TMBStruc)

random=summary(m)$coefficients[1]
random
write.csv(random,"/Users/clara/Documents/Master Thesis/Data/Lynx/glmmrandomwithintercept.csv",
          row.names = TRUE)
}
if (run_random==0){
  random=read.csv("/Users/Clara/Documents/Master Thesis/Data/Lynx/glmmrandomwithintercept.csv",
                  row.names = 1)
}
```

Crane analysis

David Wolfson and Clara Panchaud

5/7/2021

```
source("/Users/clara/Documents/Master Thesis/Simulations/functions.R")
library(raster)
library(rgdal)
library(here)
library(tidyverse)
library(sf)
library(amt)
```

We start by importing the raster that represents the environmental variable. It is a general land cover layer with the extent of Minnesota.

```
ras<-raster(here("/Users/clara/Documents/Master Thesis/Code/Crane/nlcdmnutm15.tif"))
```

We then bring in GPS points, select one individual, and filter to keep only the locations in Minnesota.

```
df1 <- read_csv(here("df16.csv"))
df <- subset(df1,id=="7J (Melby colt #1)" )
df<-df %>% filter(location.long>(-96),
                  location.long<(-93),
                  location.lat>45,
                  location.lat<48)
```

We make the data into a track.

```
track<-make_track(df,location.long,location.lat,loctime, crs = CRS("+init=epsg:4326"))
track <- transform_coords(track, CRS(proj4string(ras)))
```

We can now prepare the track to be fitted, by among other things sampling available steps and extracting the covariate. The land cover gives a lot of categories, so we regroup into 7 categories, with 'wetlands' as the reference category. We then fit the conditional logistic regression.

```
set.seed(123)
trk<-prepare_track(track)
fit<-trk%>% fit_clogit(case_ ~ category + sl_ + log_sl_ + cos(ta_) + strata(step_id_))
```

We can then apply SIMEX on the data. We vector *variance* contains the variances of the error that will be used to blur the original track.

```
set.seed(123)
start_variance<-1200
run=1
if (run==0){
param <- data.frame(matrix(ncol = 13, nrow = 0))
names <- c("categorybarren","categorydeveloped","categoryforest",
           "categoryherbaceous","categoryplanted_cultivated",
           "categorywater","SDbarren","SDdeveloped","SDforest",
           "SDherbaceous","SDplanted_cultivated","SDwater","error")
colnames(param) <- names
variance<-c(500,1000,1500,2000,2500,3000,3500,4000,4500,5000,5500,6000,6500,7000)
param[nrow(param) + 1,]=c(summary(fit)$coef[1:6,1],
                          summary(fit)$coef[1:6,3]^2,start_variance)

for (j in 1:length(variance)) {
  for (i in 1:50){
    trk_blur<-blur(track,variance[j])
    trk_blur<-prepare_track(trk_blur)
    fit_blur<-trk_blur%>% fit_clogit(case_ ~ category + sl_ +
                             log_sl_ + cos(ta_) + strata(step_id_))
    param[nrow(param) + 1,]=c(summary(fit_blur)$coef[1:6,1],
                             summary(fit_blur)$coef[1:6,3]^2,start_variance+variance[j])
  }
}
}

if (run==1){
param<-read.csv("/Users/clara/Documents/Master Thesis/Code/Crane/crane_results4.csv")
param[1,]=c(summary(fit)$coef[1:6,1],summary(fit)$coef[1:6,3]^2,start_variance)
}
```

We realised that there was a mistake in the above code, as we saved the variances and not the standard deviations under the name SD. Instead of running the code again, we change the names of the columns to fix that mistake.

```
names <- c("categorybarren","categorydeveloped","categoryforest",
           "categoryherbaceous","categoryplanted_cultivated",
           "categorywater","Variancebarren","Variancedeveloped","Varianceforest",
           "Varianceherbaceous","Varianceplanted_cultivated","Variancewater","error")
colnames(param) <- names
```

Now that we have the results from the simulation part of SIMEX, we plot the results for each category and proceed with the extrapolation to obtain the SIMEX parameter. We extrapolate with a linear, quadratic and cubic function, and pick the best one according to the AIC criteria. We also find the standard errors of the SIMEX parameter. Here is the code for one category, as they are all similar.

```
df<-param %>% group_by(error) %>%
  summarise(categoryplanted_cultivated=mean(categoryplanted_cultivated))

fit = lm(categoryplanted_cultivated ~ error, data = df)
fit2 = lm(categoryplanted_cultivated ~ error+I(error^2), data = df)
fit3 = lm(categoryplanted_cultivated ~ error+I(error^2)+I(error^3), data = df)
```

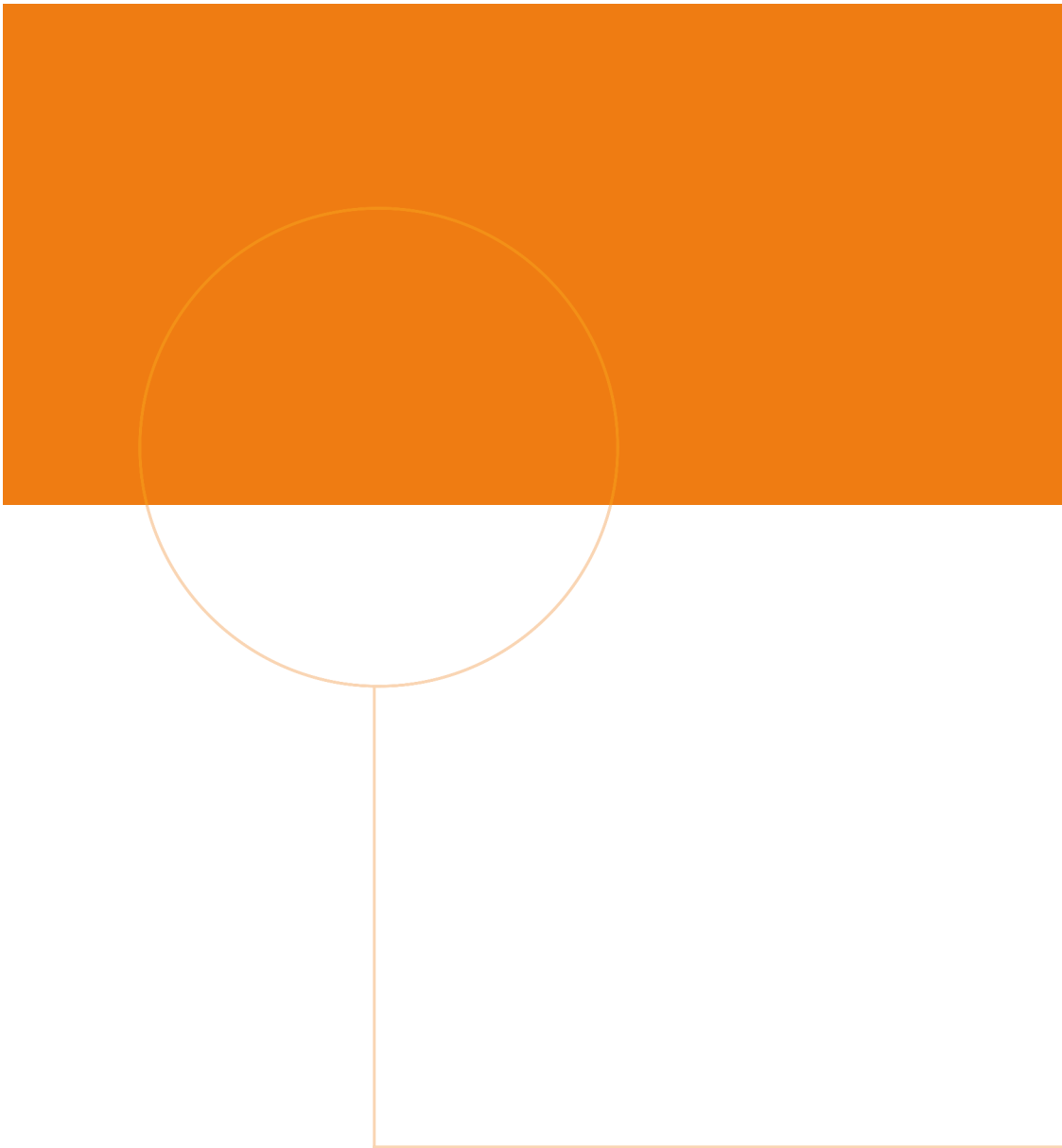
```
c(AIC(fit),AIC(fit2),AIC(fit3))

new_df<-data.frame(error=0)
p<-predict(fit2, newdata = new_df, interval = "confidence", type = "response")
naive<-as.double(subset(df,error==1200)[2])

variable<-pull(param,categoryplanted_cultivated)
variance<-pull(param,Varianceplanted_cultivated)
error<-pull(param,error)
sd<-sqrt(variance)
dfvar<-data.frame(variable,sd,error)
sdplanted<-sqrt(find_variance(subset(dfvar,error!=1200)))

confsimex<-ConfidenceInt(p[1],sdplanted)
conf<-ConfidenceInt(param[1,5],sqrt(param[1,11]))

plot<-ggplot(data=subset(param,error!=1200), aes(group=error,x=error,
        y=categoryplanted_cultivated)) + geom_boxplot()+xlim(0,8500)+
        ylim(-0.25,-0.05)+geom_point(aes(y = p[1], x = 0,
        color="SIMEX parameter from quadratic extrapolation",
        label="hello"),size=3.5)+xlab("error variance")+
        ylab("parameter")+
        ggtitle(label="Planted Cultivated")+
        theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))+
        labs(color="")+theme(legend.position="none")+
        geom_point(data = subset(df,error!=1200), mapping = aes(x = error,
        y = categoryplanted_cultivated,color="mean parameter per error level"),size=3.5)+
        theme(axis.title=element_text(size=20),axis.text=element_text(size=15),
        plot.title = element_text(size=26,hjust = 0.5))+
  geom_point(aes(y = naive, x = 1200, color="naive parameter",label="blop"),size=3.5)+
  geom_segment(aes(x=1200,y=conf[2],xend=1200,yend=conf[1]),color="chartreuse4",size=0.7)+
  geom_segment(aes(x=0,y=confsimex[1],xend=0,yend=confsimex[2]),color="steelblue3",size=0.7)
plot
```



NTNU

Kunnskap for en bedre verden