Thesis Proposal
Spatial Occupancy Models: Reducing Run-Times

Yovna Junglee
Taru Singhal

May 2019

1 Background

Occupancy models are used to estimate the probability that a site is occupied by a species under investigation by collecting presence-absence data from multiple visits at multiple sites over time (Mackenzie et al., 2002). These models have many applications to help deal with challenges in the field of ecology, such as "large-scale monitoring programs" which require predictions of occurrence of a species (Mackenzie et al., 2002). The two underlying processes which define these models are the occupancy and detection process. The *occupancy* refers to whether or not a species is present at a particular site, and *detection* refers to whether or not the species was observed when it occupied the site. The detection probability of a species under investigation at a specific site is conditional on the occupancy, the model is therefore hierarchical. One advantage of the occupancy model is that it accounts for the ambiguity of non-detection of a species of interest at a specific site, indicating that non-detection can occur due to non-occurrence or if a species is present but undetected (Dorazio and Rodriguez, 2012).

The occupancy model as defined in Mackenzie et al. (2002), does not take into consideration auto-correlation present between sites that are neighbouring each other. This is overcome by spatial occupancy models which account for this spatial auto-correlation. This implies that the probability of occupancy is likely to increase if the neighbouring sites are occupied.

2 Problem Statement

A number of packages have been developed to model the spatial occupancies of species, although the implementation of the code using the packages are time consuming. This can limit the usefulness of this model when analysing large datasets with numerous species (Clark and Altwegg, 2019). For this reason, we have to integrate different techniques into the spatial occupancy modelling process to reduce its run-times.

3 Aim and Objectives

The aim of this research is to explore different algorithms and ideas that will improve the run time efficiency of the spatial occupancy models. The objectives are:

1. To implement parallelisation in the existing *Rcpp* codes. This involves reviewing existing code and determining points where independent tasks are taking place, and can therefore be run simultaneously.
2. To implement models that improve the run-times of the initial spatial occupancy model and evaluate the performance of these models.

4 Methodology

4.1 Main Papers

The primary paper which we will be reviewing and building our base model off is by Clark and Altwegg (2019). Further papers will be consulted, such as Mackenzie et al. (2002) and Johnson et al. (2013). Scott et al. (2013) will also be reviewed for details on consensus Monte Carlo Markov Chain (MCMC) methods for Bayesian techniques which aim to split the data into smaller groups, run the analysis simultaneously and then get a combined inference.

4.2 Packages

The current packages that exist to fit the spatial occupancy models are: *Rcppocc* and *stocc* which are built in `R` and make use of *Rcpp and Rcpparmidillo. OccuSpytial* also exists, and is written in `Python`. As part of our research, we will use the current code and see where it can be optimised. We will also write functions which will incorporate implementations of the altered models. These methods use Markov Chain Monte Carlo (MCMC) algorithms such as Gibbs sampling and consensus MCMC to implement models and estimate parameters (Scott et al., 2013). Methods to improve the efficiency of these algorithms will be explored with the help of existing literature.

4.3 Dataset

The 2^{nd} Southern African Bird Atlas Project (SABAP2) database will be used to demonstrate the reduced-run time methods. The Southern Africa region, consisting of South Africa, Lesotho and Swaziland, are divided into 2002 Quarter-Degree Grid Cells (QDGC) (Loftie-Eaten, 2015). The QDGC can further be divided into nine pentads (Loftie-Eaten, 2015). For the purposes of our investigation, we will utilize the QDGC initially, and then run the analysis on the pentads.

The dataset for several specific bird species will be chosen over a specific time period. This dataset will contain the location of the site, whether the species was detected or not, the time of visit and other covariates found at the site which can impact detection and occupancy. Covariates which can affect occupancy include geographical factors and the related ecosystem. Covariates which affect detection include length of observation at site.

Pre-processing and structuring of the data will be conducted to ensure it follows a format to build a model. Birds which are resident will be considered. Resident birds are those which are present all year long and do not make seasonal migrations (Willis et al., 2008). This will allow us to use a full year as the time period to be considered, as seasonal occurrence and migration of the bird species will no longer be an issue. A longer time period is not chosen to ensure that probability of occurrence will not change.

Two types of birds will be chosen, this allows us to test the different models on more than one dataset. The two birds chosen are the Black-headed Heron (*Ardea melanocephala*) and the Black Stork (*Ciconia nigra*).

4.4 Proposed Models

After the data retrieval process, different models will be built. Only the data for the Black-headed Heron will be focused on in the beginning, as this species is more abundant. The unaltered spatial occupancy model, referred to as the original model, will be run using the model as established by Clark and Altwegg (2019). The run time will be recorded, as well as the ease with which model was built and packages were to use. After this, altered models will be implemented and their run-times and results will be examined and compared.

We suggest reducing the complexity of the model by assuming that the probability of detection at $site_i$ across n visits is constant. This will result in a binomial distribution, under the assumption that probability of detection for a site remains constant over the visits. Covariates at different visits can be restructured in a way to contain all information over visits. Summary statistics of the covariates can be obtained over the visits at a specific site and be used to fit the models. Suitable posterior distributions of the parameters will be derived and used in conjunction with the Gibbs sampling algorithm (Clark and Altwegg, 2019). Although some information will be lost pertaining to specific site visits, it may result in reduced run-times.

As a second method, we will consider splitting the entire study region into k different data sets. The existing algorithm will be run on the split data independently and then the results will be combined. One of the challenges of this model is to find an optimal value for k. This idea is an adaptation of consensus Monte Carlo where the goal is to subset the data into groups, run a full Monte Carlo simulation from a posterior distribution for each subset on separate cores and then combine the simulations from each core to obtain a consensus posterior inference (Scott et al., 2013).

4.5 Model Comparison

To compare the altered models, one possibility is to construct a discrepancy measure by calculating the absolute difference in occupancy probabilities over all sites between the original and altered model. Visually, maps of occupancies for different models and their run-times can be compared. If a model produces a map sufficiently close to the original map, it will be deemed successful. A more statistical method will be used to further examine the models. This involves using Bayesian model selection methods as proposed by Hooten and Hobbs (2015). After fitting multiple models, they will be compared using these Bayesian evaluation methods to determine which model is preferred. There will be a trade-off between accuracy of model and its run-time.

5 Plan of Work

This is a rough guideline of how we plan on proceeding with the research.

TABLE 1 Timeline

5 June	Complete literature review. This includes understanding derivations of the original model as built by Clark and Altwegg (2019), our primary model.
20 June	Complete review of existing code. Under the review, attempts at parellelising sections of code will also be conducted.
30 June	Complete mathematical and statistical derivations for the altered models.
12 July	Complete model building for existing packages and analyse their run-times. Complete one altered model and its code.
15 July	Progress report due. This will include the original model and package, a parellelised model and an altered model.
10 August	Second altered model should be completed and implemented.
30 August	Comparisons of the different models using both bird species.
30 September	First draft completed.
7-11 October	Presentations.
21 October	Final hand-in.

6 Outcomes

We expect to report results based on the run-times of the different models, their performance, accuracy and ease of implementation. The revised models may lose some accuracy, but should result in improved run-times.

References

Clark, A. E. and R. Altwegg
2019. Efficient bayesian analysis of occupancy models with logit link functions. *Ecology and Evolution*, 9(2):756–768.

Dorazio, R. M. and D. T. Rodriguez
2012. A gibbs sampler for bayesian analysis of site-occupancy data. *Methods in Ecology and Evolution*, 3:1093–1098.

Hooten, M. B. and T. N. Hobbs
2015. A guide to bayesian model selection for ecologists. *Ecological Monographs*, 85(1):3–28.

Johnson, D. S., P. B. Conn, M. B. Hooten, J. C. Ray, and B. A. Pond
2013. Spatial occupancy models for large data sets. *Ecology and Evolution*, 94(4):801–808.

Loftie-Eaten, M.
2015. Geographic range dynamics of south africa's bird species. Master's thesis, University of Cape Town.

Mackenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, A. J. Royle, and C. A. Langtimm
2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255.

Scott, S. L., A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch
2013. Bayes and big data: The consensus monte carlo algorithm.

Willis, C. K., O. E. Curtis, and M. D. Anderson
2008. *Bird Checklist for South Africa's National Botanical Gardens.* South African National Biodiversity Institute.