Comparing unequal probability sampling with random stratified sampling with respect to finding rare species and making distribution models

Øyvind Lynne

Master of Science thesis
Department of Bioscience
Natural history museum

UNIVERSITY OF OSLO

December 2019

Comparing unequal probability sampling with random stratified sampling with respect to finding rare species and making distribution models

A case study of rare species in Oslo

Øyvind Lynne

Oyvlyn@gmail.com

+47 92857580

Subervisors:

Olav Skarpaas

olav.skarpaas@nhm.uio.no

Rune Halvorsen

rune.halvorsen@nhm.uio.no

Geo-Ecology research group

Natural History Museum

Master in Biology: Ecology and Evolution

Department of Bioscience

University of Oslo

December 2019

Abstract

Unequal probability sampling (UPS) is a sampling method for observing a phenomenon, where the drawing observation units are not equal, but vary e.g according to predicted probability of presence. UPS can in theory work well in situations where the targeted phenomenon is rare. This study aims to test UPS as a sampling method for observing a group of species that are all connected to the same rare, environment conditions. Specifically, to compare the UPS data with a data set sampled in a different way (stratified random sampled), with respect to the prevalence, and for gaining response variable data when building distribution models. In addition, I also tested how the size of the response variable data, the inclusion of indirect variables and the specification of the model selection criteria can affect which variables are chosen when building distribution models.

The study were conducted in Oslo and the surrounding area. There were in total 200 ten by ten meters plots that were surveyed as a part of the study design. Observation of 29 species were registered, all of which are associated with the presence of limestone in the substrate. The the probability of drawing a particular 10x10 m cell as a survey plot were weighted on the predicted probability of the 29 species studied. Predicted probability were calculated using a poisson regression distribution model based on citizen science data. The UPS data were then compared with a stratified random sampled data (SRS). The SRS data are observation of the same targeted species, within the same geographical. The prevalence between these data were compared and the difference were tested for significance. The UPS and the SRS data were also used to build distribution models using the same statistical tool. Twelve models were made in total. The models differed in three ways: The sampling method, if indirect variables were included or not and the variable selection criteria, i.e alpha value in the variable selection part that produced the model. The subsequent models were then compared, with emphasis on the difference in selected variables.

The prevalence of the UPS data were significantly higher than the SRS data. In addition, the distribution models made with the UPS data were all able to identify variables that were directly associated with limestone, while the models made from the SRS data did not identify such variables. All the models with no a priori modification on the model selection process chose elevation in the first round of the forward selection process. The models with a more conservative forward selection process had fewer significant explanatory variables. This demonstrates the effect of the different choices one have to make during a model building process. Additionally, the results shows the importance of adapting the sampling based on the prevalence of the targeted phenomenon.

This study have shown that UPS can work as a sampling method if the goal is to observe rare phenomenon more frequently. It is reasonable to assume that the SRS data are affected by bias because of autocorrelation. In addition, the data had comparably few observation units where limestone were present. More research are therefore needed here. A more conclusive result could be obtained if one could compare the UPS data with an equal random sampled data set. Several of the UPS models identified the presence of exposed limestone as an important predictor, which may make UPS a good choice for modelling rare species.

Acknowledgements

I would firstly like to thank my two supervisors Olav Skarpaas and Rune Halvorsen. They steered me in the right direction whenever I needed it, and at the same time allowed me to find my own path in the research and writing process. I would also like express my very profound gratitude to my family, friends and fellow students. They have provided me with encouragement and support throughout my studies, and last but not least through the process of writing this master's thesis.

Contents

1. Introduction

When searching for any species, especially if the species is rare, it is then of particular interest to try to predict where the species will be present. This was solved for a long time by so-called *site-based sampling*, a method that involves going to locations (often called *sentinels sites*) where the targeted species has already been observed (Yoccoz, Nichols, & Boulinier, 2001). Many also used so-called *judgment sampling*, which is sampling on location where it is assumed to be presences of the target species based on expert knowledge (Olsen et al., 1999). However, both these sampling methods are based heavily on subjective assumption about the underlying causes of the species' distribution, which in most cases can lead to biases, meaning the potential observation with these methods cannot be considered independent (Smith, 1983). The reasons for choosing such "subjective" sampling methods are not directly tied to any assertion about the underlying causes of the distribution of the target species. Because of this, *site-based* and *judgement* sampling methods can be inadequate when one infers or tests the hypothesis about the causes of its distribution. In general, it is preferred to make inferences from the observation about the whole population (in the statistical sense) for a given area (R. Halvorsen, 2012), which becomes problematic in the presence of dependency in the data (Lohr, 2019).

It is reasonable to assume that in many cases, one of the more important variables that influence the distribution of a species is one (or several) variables, which is often treated as a gradient in which the species have an optimum on a specific point along this variable (Austin & Gaywood, 1994). In many cases, this assertion forms the basis for distribution modelling. This is the science on predicting the distribution of in theory any observable natural phenomenon by knowing the distribution of the explanatory variables that conditions the phenomena in question (R. Halvorsen, 2012), often by designing statistical models made with a model selection tool and a response variable data. Distribution modelling is now a popular branch in biology (particularly in ecology) and with better software and more readily available environmental data, it will presumably become more relevant over time, especially with respect to finding rare species and/or interesting location in a conservation effort context.

I assert that distribution modelling will in most cases serve better than the aforementioned "subjective" sampling methods if the goal is to predict where a natural phenomenon will be. This approach is less vulnerable to the formation of false conclusion about the species-environment relationships because of bias. However, the interpretation of the prediction calculated from such distribution models will depend on several factors of the response variable data.

The size of N, the model selection criteria and the environment variables

The effect of sample size N (total number of observation units) can be important. By not adapting the model selection criteria to the data, a large N can cause model overfitting. By overfitting, I mean the case in which the model is overly complex by including parameters that reflects qualities of the data, rather than generalization of the species-environment relationships. One can account for this by being more conservative in the model selection process (Aho, Derryberry, & Peterson, 2014).

Environmental variables (EV) included in the model selection process and how they interact with each other can be important for the interpretation of the distribution models. Both in terms of correlation between the different environment variables (Yoccoz et al., 2001), and the response variable single-effect response curve for each of selected EV in a given distribution model (Irvine, Rodhouse, Wright, & Olsen, 2018). If two or several environment variables are correlated, and one of them are deemed significant in the model selection process, a confounding effect can occur (Yoccoz et al., 2001). Such variables are often called *indirect* variables (Austin, 1980). Elevation is often an example of an indirect variable (Austin, 1980). Indirect variables or gradients can often be strong predictors for explaining the variation in species distribution (Whittaker & Peet, 1985). Nevertheless, if one wishes to produce a distribution model that are more widely applicable, then the exclusion of such proxy variables is a more desired setup for the model selection process.

Sampling method

One of the first things the modeler have to consider is what the response variable should consist of, and how it should be sampled. This is mainly decided by the purpose of the modelling, as both the study design and the methods used in the analysis must be adapted based on the eventual goal with the study (Irvine et al., 2018).

Maybe the most straight forward sampling method is simple random sampling, which places a certain number of observation units randomly on the study area as a whole (Meng, 2013). This method will sample any possible observation unit with the same probability, and as such will in most cases ensure independency between the observation units. Simple random sampling can be suitable if one does not have any prior knowledge and/or hypothesis about the causes of the distribution of the species (Gravetter & Forzano, 2009).

There are also more deliberate sampling approaches, which are still for the most part unbiased. For instance, stratified random sampling, or **SRS**, which divide the data or population into sub-groups where each group share a specific trait. For each group, a predefined number of observation units are sampled randomly (White, 2020). In ecology, such methods are often used to test hypothesis about the importance of a specific environment variable (J. B. Halvorsen, 2019). The gradsect sampling method is similar to stratified random sampling, as it is a sampling methods that seeks to capture the variation of the species along the whole range of the gradient (Guisan & Zimmermann, 2000) (Austin & Heyligers, 1989).

There are also examples of study design using several sampling methods (where the methods used was among the abovementioned), where the goal is to capture most of the relevant EV variation, while still retaining independency between the observation units (Wollan, 2011).

Whether one seeks to find specific areas that are interesting in a conservation effort setting (Yoccoz et al., 2001) or one attempts to test environment-species relationship (Wollan, 2011), the problem of getting too few observations of the species in question can often occur especially if the species in question are rare (Skarpaas O, 2019). By too few, I mean in the sense that accurate inferences about the population as a whole become difficult.

Arguably the most common type of response variable consists of a single species (called single species distribution model (Henderson, Ohmann, Gregory, Roberts, & Zald, 2014)). However, if one wishes to predict the distribution of a specific nature type or community, one could achieve this by combining several species which one suspects are connected to the same environment conditions (Wollan, 2011). The benefit of this approach is that several rare species that all have low prevalence, can "share" their distribution with each other, which ensures that the total number of observations is high enough for a meaningful relationship between the response variable and the environment variable can be recognized (Ovaskainen & Soininen, 2011).

In addition to combining several species to one response variable, the modeler can also use unequal probability sampling or **UPS** (Yoccoz et al., 2001) to account for rarity. This method does not directly violate the assumption of independence, while obtaining more observation of a rare phenomenon by oversampling, compared to the other randomized sampling methods (Olsen et al., 1999). A prerequisite for using an UPS method is knowledge about the conditions that influence the distribution of the target phenomena. In essence, an UPS method weights the distribution of the observation units based on this a priori knowledge. A higher concentration of plots are placed on areas with higher predicted probability of observed presence (Olsen et al., 1999; Skarpaas O, 2019).

Comparing unequal random sampling with other sampling methods

The proposition up to this point has mainly been that UPS is a suitable method for acquiring a response variable data set for the specific purpose of building a distribution model. However, one could also use UPS method to infer the distribution of a rare species or a rare community in a conservation efforts context (Edvardsen, Bakkestuen, & Halvorsen, 2011). And as (R. Halvorsen, 2012) points out, there is a difference between observation data in itself versus observation data that are used in analysis. I will argue that this is especially true for the UPS, given the nature of the data.

One does not have to make many assumption if the data is used as it is, without any specific analytic treatment (R. Halvorsen, 2012). Examples of this could be inferring the population size from the sampled data, or comparing an unequal probability sampled data set with a stratified random sampled data set in terms of the prevalence. However, if one uses the data for distribution modelling, the assumption of no bias is important, as the more direct consequence of sampling bias is uneven distribution of the observation units with respect to

the range of the environment variables. The resulting distribution model will in most cases not be able to infer the actual relationship between the response variable and the environment variable if sufficient amount of environment data are missing along the environment variable. This can be explored further by using frequency of observed presence plots (FOP-plots) (Støa, Halvorsen, Mazzoni, & Gusarov, 2018).

Oslo is an area where a lot of research on the ecology and natural variation already has been conducted (J. B. Halvorsen, 2019; Wollan, 2011). There have also been observed several rare species here that are adapted to high concentration of calcium in the soil (Wollan, 2011). Despite all the research that already exist on Oslo and on the more rare species that have been observed within this study area, how unequal probability sampled data on these species compares to data sampled in another way are yet to be tested. Additionally, how the size of the data set, the model selection criteria, and the variation included (if they are indirect variables or not) in the model selection process are questions that can be interesting to be highlighted within a distribution modelling context.

An interesting question then is how a DM made with a UPS response variable data set would compare with a distribution model made with response variable sampled with different methods, assuming the study design, study area and the statistical analysis are all the same. In addition, is UPS truly a better sampling method compared to other sampling methods if the goal is simply to infer the total population?

I assert that if one wish to test this comparison, then the access to two data sets sampled with two different sampling methods within the same study area where the same rare target phenomenon were registered. This study therefore aim to test UPS as a method to observe and model a collection of rare species that all are connected to the same environmental conditions, in this case, species that are adapted to a substrate high in limestone. The inner Oslo selected as a suitable study area. To achieve this, I will compare different sampling methods, i.e compare an unequal random sampled data set with a stratified random sampled data set that were sampled within the same geographical area and within the same field season (inner Oslo and the surrounding forests, summer of 2018). I will also adapt the model selection process to emphasize how the size of the response variable data, the model selection criteria and the inclusion of indirect environment variables can affect the outcome of a model selection.

Aims

1. How does an unequal probability sampled data and stratified random sampled data differ in terms of prevalence, i.e the proportion of presences in the data? Or, how do they differ in terms of finding rare phenomenon?

2. When using the same distribution modelling tool on both the data sets, how do the resulting distribution models differ in terms of selected variables, variation explained and predictability?

More specifically, how is the outcome of a model selection affected by;

3. The population size N, or total number of observation units?

4. Different variable selection criteria? In this case, different alpha values?

5. A priori decision about which variables to include in the model selection, specifically the inclusion of proxy variables that are known to be not directly linked to the investigated species?

2. Materials and method

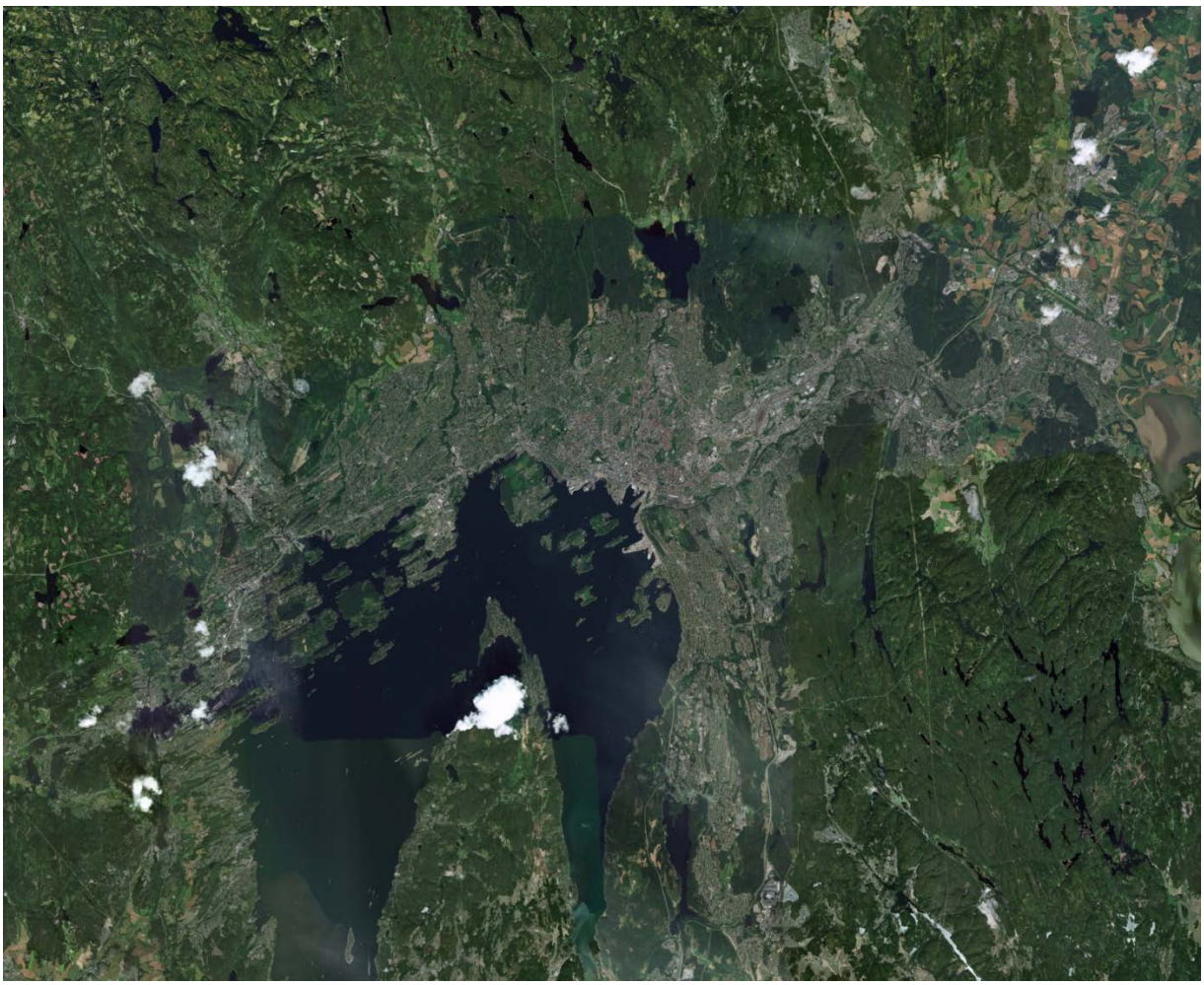
2.1 Study area

Oslo is the capital of Norway, and lies in the southeastern part of the country. The bedrock in large areas contain a lot of limestone (Pedersen, Nyhuus, Blindheim, & Krog, 2004). More specifically, the geology of the islands in the inner fjord and the northern part of Oslo consist mainly of late Palaezoic sediments, while the southern parts of Oslo are made by Precambrian basement (W, 2009). The annual mean temperature for the study area is 6 ℃, which is based on measurements from 1961 to 1990 (Aune, 1993). Oslo lies within the boreonemoral vegetation zone (Moen, 1998).

The surrounding area is quite divers in regards to both animal and plant species (Pedersen et al., 2004). The coastline and the many islands in the inner part of the fjord in particular are considered bio diversity hot spots. This is assumed to be caused mainly by the presence of the limestone in the substrate (Pedersen et al., 2004). However, it has been recorded more pressure on these high diversity area in recent years, mainly in the form of human activity (Wollan, 2011). Some forms of activity is not as devastating as others, but the sum has still caused a reduction of the nature type that these rare species are adapted to (Wollan, 2011). As such, many of the species that is highly connected to these hot spots has been classified as endangered by the red list (Pedersen et al., 2004).

Oslo was chosen as the study area for this project because a lot of research has already been done in this area with regards to species distribution and the effect on human activity on the natural environment (Wollan, 2011). Furthermore, Oslo is also the study area for the URBAN EEA-project (Barton et al., 2017) where the supporting data material that is used in this study (J. B. Halvorsen, 2019), and the DM model that the UPS data set derives from is a part of (Skarpaas, et al. in prep).

Figure 2.1: The extent of the study area, which mainly consists of Oslo city and the surrounding forests. The map is obtained from the QGIS 3.2 QuickMapServices (QGIS development Q. D. Team, 2009) with the ESRI Satellite layer (ESRI, 2017).



2.2 Supporting material

The poisson regression model

The DM that the UPS data derives from were made with poisson regression (Skarpaas, et al. in prep). The RV consists of observation of the targeted species and observation of other species. Total frequency of species observation are used as the offset.

The stratified random sampled data

I used a stratified random sampled data to compare with the data I have gathered. The stratified random data consist of a primary stratum based on an urbanization gradient, and a secondary stratum based on area of coverage (J. B. Halvorsen, 2019).

The environment variables

I used the following variables in the model selection. The same ones were used when making the poisson model, except for traffic and surface temperature (Skarpaas, et al. in prep).

Table 2.1: The environmental variables that were a part of the model selection process in the analysis.

Name in the analysis	*Variable*	*Definition and sources*
aspect	Aspect	Terrain aspect calculated from DEM
building_AW	Building density	Area-weighted density of buildings, calculated from building map (kartverk, 2017a)
curvature	Curvature	Terrain curvature calculated from DEM
elevation	DEM	Digital elevation model (DEM) in meters (kartverk, 2018)
road_distance	Distance to road	Distance to nearest road (m), calculated from road map (kartverk, 2017b)
slope	Slope	Terrain slope calculated from DEM
sun_exposure	Sun exposure	Exposure to sunlight, based on topography and latitude
temp_surface	Surface temperature	Surface temperature 2nd Jul 2015, based on air temperature measurements and urban heat island effect (Blumentrath, 2016)
TPI	TPI	Terrain position index (difference in elevation between focal pixel and mean of surrounding pixels), calculated from DEM
traffic	Traffic	Road traffic: cars per day (Statens vegvesen, 2017)
geo_substrate	Bedrock & soil	Combination of geology (Limestone) and substrate (Soil cover): (1) exposed nutrient-poor bedrock, (2) exposed limestone, (3) soil cover
landcover_sat	Land cover	Land cover classification based on Sentinel 2 satelite images: (1) agriculture edge, (2) grass, (3) built-up, (4) tree canopy, (5) water edge (Nowell, 2017)
limestone	Limestone	Indicator variable for calcareous bedrock (1) or not (0), based on the presence of the word "kalk" in the discriptor field of the national geological map (NGU, 2016)
substrate	Soil cover	Simplified soil cover map in two classes: (1) no or little soil, (2) deep soil cover, made by merging classes in the national soil cover map (NGU, 2017)

2.3 Study design

I conducted the field work during the summer of 2018, and partly during the summer of 2019. It should be mentioned that the summer season of 2018 were unusually hot.

Investigated species

All the investigated species have some attributes in common; the most important one in this case is that they all are adapted to soil/substrate with relative high calcium content. Most of them are quite small, and all of them with three exceptions are herbs. A more detailed table showing additional information for each species can be found in *Appendix 3*.

Table 2.2: The Norwegian and latin name of the targeted species.

Norwegian name	**Latin name**
Blodstorkenebb	Geranium sanguineum
Knollmjødurt	Filipendula vulgaris
Bergskrinneblom	Arabis hirsute
Flatrapp	Poa compressa
Bakketimian	Thymus pulegiodes
Nakkebær	Fragaria viridis
Aksveronika	Veronica spicata
Krattalant	Inula salicina
Hundetunge	Cynoglossum officinale
Flekkgrisøre	Hypochaeris maculata
Berggull	Erysimum strictum
Fagerknoppurt	Centaurea scabiosa
Nikkesmelle	Silene nutans
Dvergmispel	Cotoneaster intergerrimus.
Liguster	Ligustrum vulgare
Smaltimotei	Phleum phleoides
Stjernetistel	Carlina vulgaris
Vårrublom	Draba verna
Ormehode	Echium vulgare
Fjellrapp	*Poa alpine*
Nyresildre	Saxifraga granulate
Vårstarr	Carex caryophyllea
Dragehode	Dracocephalum ruyschiana
Oslosildre	Saxifraga osloensis
Kanelrose	Rosa majalis
Trefingersildre	Saxifraga tridactylites
Flerårsknavel	Scleranthus perenni
Legesteinfrø	Lithospermum officinale
Vill-lin	Linum catharticum

Sampling design

I were to survey in total 200 plots. They were drawn randomly weighted on the probabilities calculated from a distribution model (Skarpaas, et al. in prep), thus obtaining an unequal random sampled data. The size of all the observation units used in the analysis were 10x10 meters, meaning that I recorded the species in fine to medium local scale (R. Halvorsen, 2012). The surveyed plots were each 30x30 meters divided into nine 10x10 meter zones. The observation units used in the analysis were in the middle, surrounded by eight 10x10 meter zones. In these zones, I only registered the presence for each target species. The purpose for this setup were to (in situations were no observation were registered in the observation unit) improve the ability to evaluate the cause of why no observation were made in the observation unit itself. For instance, if no observation were registered in the observation unit, but observation of one or several of the targeted species were registered in one of the surrounding zones, then the absence in the observation unit may be because of some random coincidence.

To speed up the field work, I first evaluated the plots based on the surrounding ecology. I were not as thorough if the surrounding area implied a low to no probability of presence of one of the target species. I first did a fast sweep of the whole plot (30x30 meter). I also registered some things that I deemed relevant for the interpretation of the results:

- Species outside of the species list. These species were never deliberately searched for, and I only registered species that I knew and that could tell something about the nature of the plot.
- Decide the potential reasons for the absence of the species. This included signs of human influence.

When I evaluated the surrounding area as ideal (primarily if limestone were present or not) for the target species, I did these following things:

- Ascertain the corner points of the plot, and for each point place a pin to delineate the survey area.
- First do a quick but deliberate search of the whole plot, record all the species that were most obvious.
- Then do a more thorough search, in an attempt to record the more inconspicuous plants. This search was done in a more systematic way, in which I trailed back and forth between the opposite sides of the plot.
- Register only the presences of the targeted species in the buffer zone.
- The species that was so abundant that to count them by individual was not expedient in regards to the time needed was instead noted as coverage in square metres.

In my field protocol I noted the coverage as intervals using this system:

- Less than 1 square meter; meaning that there were so many individual plants that to count them would take too long, but the coverage still consisted of less than 1 square meter (approximately).
- Between 1 and 2 square metres.
- Between 2 and 3 square metres, and so on.

Processing of the field data

I recorded the binary presence of at least one of the targeted species (or absence of all the species) for each observation unit. The resulting data were used as response variable in the distribution modelling. I did the same with the stratified random sampled data set (J. B. Halvorsen, 2019).

See *Appendix 1* for details on the processing of some of the NA plots and absence plots.

2.4 The analysis

I did all the statistical analysis in R, version 3.6.0 (R Core Team, 2019b).

2.4.1 Correlation between the variables

I tested correlation between the numerical environmental variables for the UPS and SRS (and UPS+SRS) using the spearman correlation test (Hollander & Wolfe, 1973. Pages 185--194). I also tested for dependency between the numerical variables and the categorical variables using the kruskal wallis test (Hollander & Wolfe, 1973. Pages 115--120). To test for dependency between the categorical variables, I used Pearson's qhi-square test (Agresti, 2007).

2.4.2 Testing the prevalence difference

I used the Pearson's qhi-sqaure statistical test (Agresti, 2007) to check if the proportion of presence points (known as *prevalence*) between the UPS dataset and the SRS were significantly different. Specifically, I used the prop.test function (R core Team, 2019a, prop.test, 20.12.19).

2.4.3 Logistic regression with MIAmaxent

I chose to use the R-package MIAmaxent in the analysis of the presence-absence datasets, i.e the UPS and SRS data (Vollering, Halvorsen, & Mazzoni, 2019). This tool in particular is adapted for distribution modelling, as it gives the modeler the ability to both test the qualities of the RV-data and test the predictive ability of the designed models. If the data for many environmental data are provided, it can also handle the selection of the best model quite efficiently by using the stepwise forward selection (R. Halvorsen, 2013). MIAmaxent can handle presence/only-data (PO), which is the type of RV that is used in maxent (Mazzoni, Halvorsen, & Bakkestuen, 2015). It can also handle presence/absence-data (PA) using logistic regression, which is why I chose this method for building the distribution models for this study. Logistic regression was carried out by specifying "algorithm = LR" whenever necessary.

I made Twelve models with MIAmaxent, four for each data set (UPS, SRS and UPS+SRS, see part *Combining the data sets*), six with elevation included and six with elevation excluded (see part *Exclusion of the elevation variable*), six with alpha value = 0.05 and 6 with alpha value = 0.001 (see part *Alpha value*).

Four R-commands from MIAmaxent were used in the analysis of the presence/absence-data:

- readData(), which takes provided presence/absence (or presence-only) data with the coordinates and the EV data in the form of raster stacks, and makes a table shoving the OUs and all their corresponding EV value. This table are used in the subsequent commands.
- deriveVars(), that makes transformation of each EV. There are in total seven different transformations that are subjected on the EVs, six of them only relevant for numerical variables (linear (L), monotonous (M), deviation (D), forward hinge (HF), reverse hinge (HR), threshold (T)) and one of them only relevant for categorical variables (binary (B)). All variable transformation are applied by default. However, an endless number of different transformation are possible for the spline types (forward hinge, reverse hinge and threshold). That is why the function produces 20 of each, and chooses the one that explains the most variation (Vollering, 2019). The numbers of DVs for each EV that are included in the output will depend on the preselection of threshold and hinge transformation of the numerical variables and the numbers of types in the categorical variables (Vollering et al., 2019).
- selectDVforEV, subjects a stepwise forward selection for each group of DV from a single EV, and selects DVs that explains a significant amount of variation. This is achieved by using a likelihood ratio test that accounts for sample size (R. Halvorsen, 2013; Vollering et al., 2019). EVs that have no significant DVs are rejected from the selection.
- selectEV, subjects a stepwise forward selection for each significant DV, and in this case selects the model with only EV (which now consist of one or several significant DVs) that explains a significant amount of variation, using the same likelihood-ratio test from selectDVforEV. The interaction argument were specified as “true” for all the models shown in this study, to test the hypothesis that limestone in combination with other variables can comparably explain more variation.

In addition, I used four commands to visualize the results:

- FOPplot(), which makes a frequency of observed presence for a specified EV. This is usually done a priori, to look for trends in the RV data within an EV. The command gives a plot where the x-axis shows the range of the specified EV, and the y-axis shows the expected probability of presence. The numerical EVs includes black dots on the plot that represents the binned presence frequency, and a red line that is a local regression following the binned frequencies. The FOP-plot for the categorical variables is a bar-plot, where the grey bars represents the background data density and the transparent bars represents the occurrence frequencies. Since the FOP-plots shown in the results are made from P/A-data, they are instead called *frequency of empirical presence*. The number of intervals into which the continuous EV is divided is determined by the argument *intervals*. Irrelevant for categorical variables.
- plotResp(), which plots the output from a selected EV in a model across the range of said EV. The response plot for one specific EV is called a single-effect response plot.

- testAUC(), plots the ROC-curve and the corresponding AUC-value. All the ROC-curves for the logistic regression models (model 1-12) were obtained with this command.
- projectModel(), which determines the prediction of the model for any spot where the EV is known. With this command one can make a predicted distribution map. Such a map is shown for each model made in the analysis in the results.

Table 2.3: All the models shown in this study.

Model name	Type of RV data (all are PA)	Elevation included	Alpha value for selectEVforEV() and selectEV()
Model 1	UPS	yes	0.05
Model 2	UPS	no	0.05
Model 3	SRS	yes	0.05
Model 4	SRS	no	0.05
Model 5	UPS+SRS	yes	0.05
Model 6	UPS+SRS	no	0.05
Model 7	UPS	yes	0.001
Model 8	UPS	no	0.001
Model 9	SRS	yes	0.001
Model 10	SRS	no	0.001
Model 11	UPS+SRS	yes	0.001
Model 12	UPS+SRS	no	0.001

Alpha value

The goal here was to test the effect of different threshold for how much variation a variable must explain for it to be retained. Alpha is in the case for the selectDVforEV()-command the p-value threshold in which the derived variables are treated as significant (id erst captures a significant amount of the variation in the RV), using the likelihood-ratio test (R. Halvorsen, 2013; Nordhausen, 2009). In the case for the selectEV(), the alpha is the p-value threshold in which the explanatory variable (which is now described by the DV of the EV) are treated as significant (id erst explains a significant amount of variation in the RV) using the same likelihood-ratio test.

I placed the alpha value at 0.05 for model 1-6, in accordance with (Fisher, 1925). For model 7-12, I placed the alpha value at 0.001, in accordance with (Vollering, 2019).

Exclusion of the elevation variable

The elevation variable correlates a lot with actual important environmental variables, while one can assume the variable itself is ecologically meaningless concerning these particular species and this particular study area, i.e it is an indirect variable.

As such, I decided to include both models with elevation in the model selection process and models without elevation. This gives the opportunity to demonstrate the differences between models made with MIAmaxent that were subjected to some a priori adjustments (meaning not

to include the elevation variable in the stepwise forward selection) and models also made with MIAmaxent where no a priori modifications were done.

Combining the data sets

I made distribution models with the combination of the UPS and the SRS data, to test the effect of the total number of observation unit on the model selection process in MIAmaxent. The combined data is denoted as UPS+SRS.

ROC-plots and AUC-values

As part of the model evaluation and comparison, I decided to plot the receiver operating characteristic curve (ROC) and show the corresponding AUC-value (area under the curve) for all the models made in the analysis and the poisson regression model. ROC-curve with corresponding AUC-values can be a suitable tool when checking the model for how widely applicably it is, as it test the models ability to extinguish false positives from true positives and false negatives from true negatives (Fawcett, 2006). In the case of SPD, a ROC-curve will tell if the SPD are able to adequately predict the presence or absence for a given test data set.

A ROC-curve shows (in this case) the distribution models ability to distinguish between a presence OU and an absence OU, using (ideally) independent test data. The sensitivity are plotted against 1-sensitivity, where the line shows the sensitivity vs 1-sensitivity (which is the inverse of the specificity) for each possible cutoff. The AUC-value shows the area under the line, which gives a value of how able the model is to distinguish true positives from false positives and true negatives from false negatives. The value can go from 0 to 1, where 1 is a perfect model (interpret all OUs correctly), 0 is a model that treats all presence observations as absence observations and vice versa for absence observations. An AUC-value of 0.5 means that the model cannot extinguish presence from absence points at all. This will give the opportunity to compare the DM models concerning their ability to accurately predict the presences of the targeted species (Fawcett, 2006).

I used the GBIF data that were used as response variable data for the poisson regression model (Skarpaas, in prep.) as the test data for all the ROC-curves.

3. Results

I surveyed 187 of the 200 plots that were drawn for this study. By plots, I mean the 30 by 30 meters units. Of all the observation units that were surveyd, 28 contained at least one of the species, meaning that I observed none of the targeted species in 157 observation units. By observation units, I mean the 10 by 10 meter units placed in the middle of the plots. *Geranium sanguineum*, *Fragalia viridis* and *Filipindula vulgaris* where the three most common species (figure 2a). Thirteen of the 29 investigated species had no registered observation. The two most abundant plots were 76 and 97, where I observed ten of the 29 investigated species. Both plots were placed at Gressholmen (figure 2b).

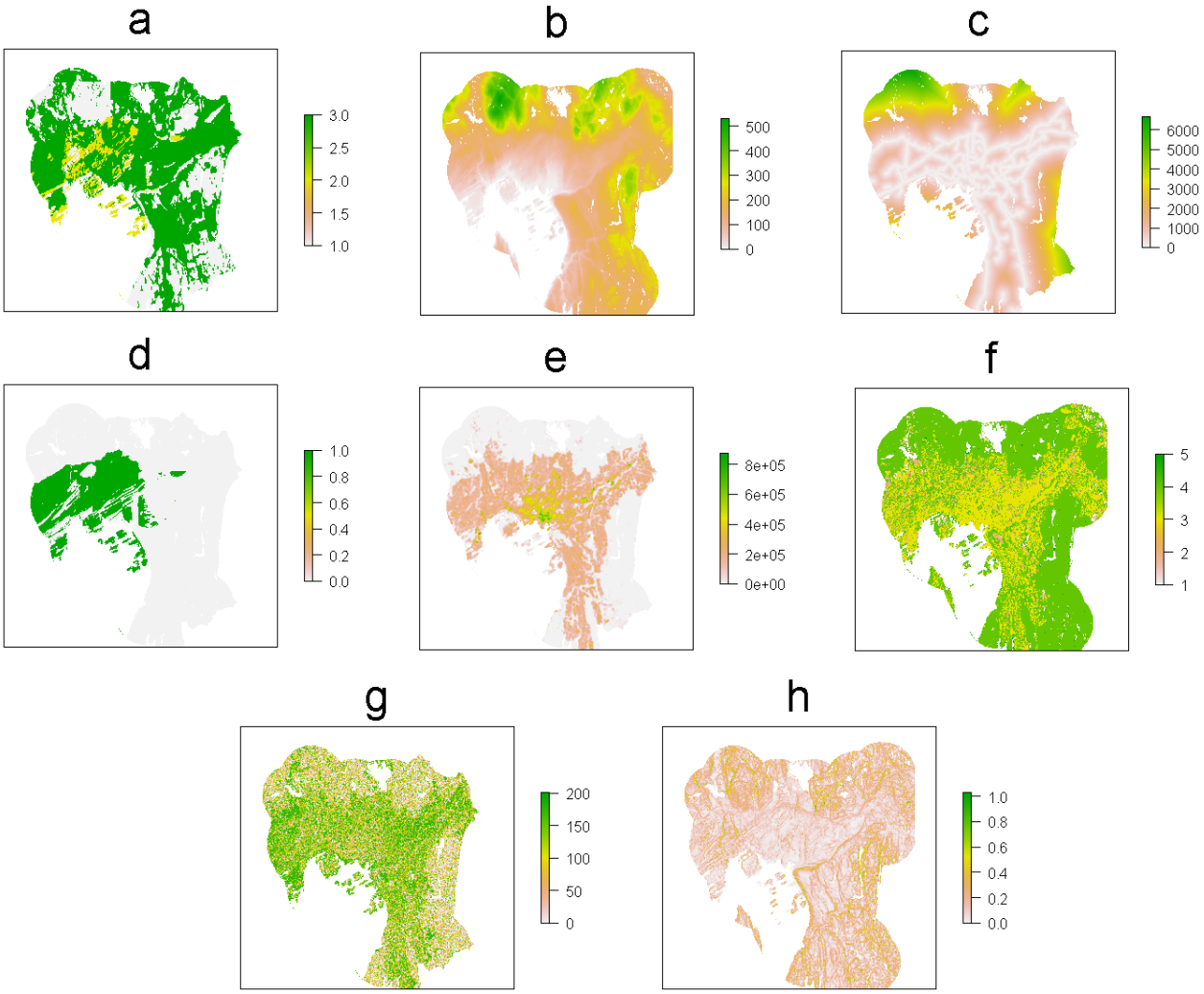


Figure 3.1: a) geo.substrate, b) elevation, c) traffic, d) limestone, e) building.AW, f) landcover.sat, g) sun.exposure, h) slope.

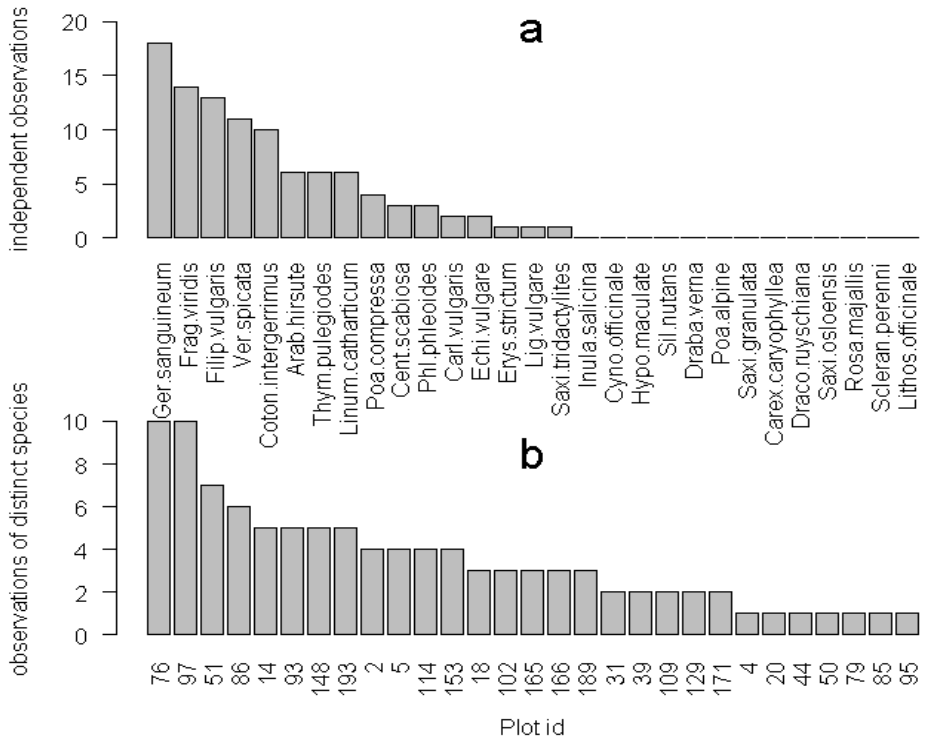


Figure 3.2: a) Barplot showing the numbers of independent observation for each investigated species. b) Barplot of the occurence observation units that show the numbers of distinct species observed for all the presence plots, in which at least one species was found.

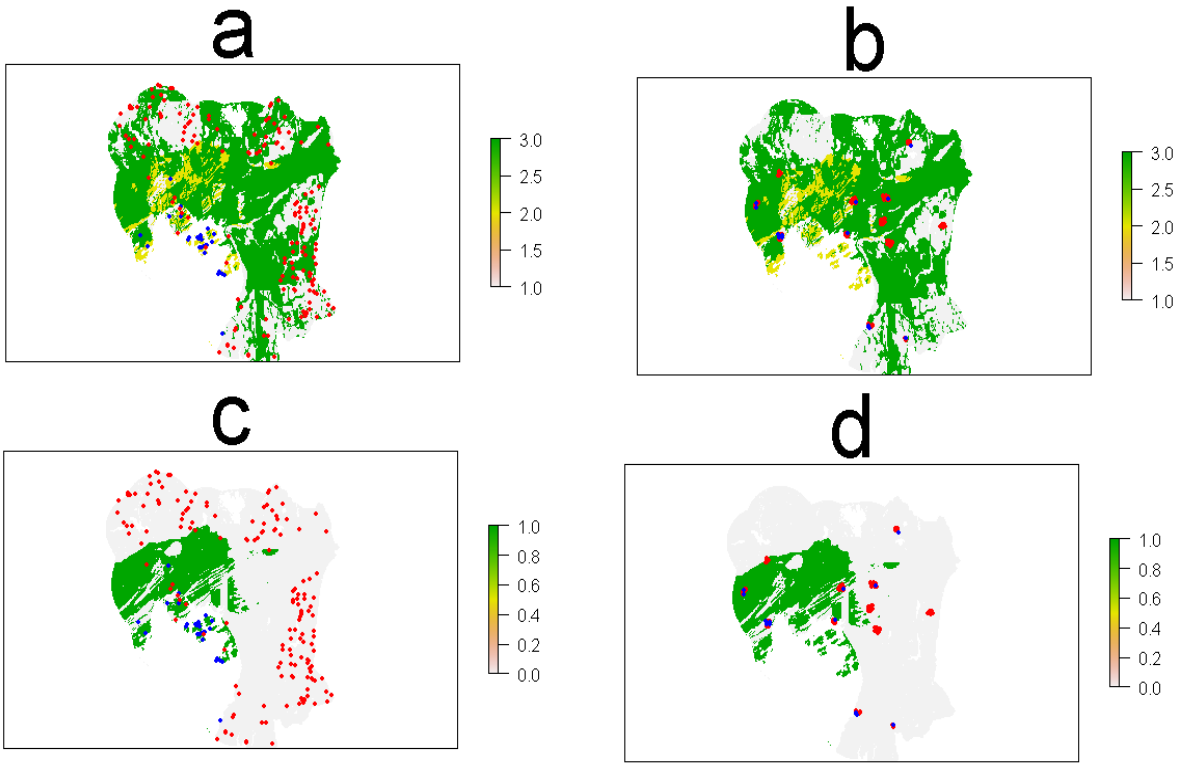


Figure 3.3: *a)* The distribution of geo.substrate with the UPS PA-points. *b)* The distribution of geo.substrate with the SRS PA-points. *c)* The distribution of limestone with the UPS PA-points. *d)* The distribution of limestone with the SRS PA-points. Blue dots represents presences, while red dots represents absences.

3.1 Characteristics of the data sets

Comparing the unequal probability sampled data with the random stratified sampled data in terms of prevalence and frequency of empirical presence

The proportion of presence points for the UPS data set (0.1497) were higher than the proportion of presence points for the SRS data set (0.0462), and they differed significantly ([Chi-sq.test: $\chi^2 = 18.101$, p-value < 0.0001]).

There are high similarity for each variable when comparing the FOP-plots between the UPS data and the UPS+SRS data, while the FOP-plots for the SRS data deviates. Here I show the FOP-plots for the significant variables for UPS and SRS. The FOP-plots for UPS+SRS are shown in *Appendix 4*.

The interval argument in the FOP plots is not the same for all the variables. The interval argument is relevant only for numerical variables

Table 3.1: The interval value for all the FOP-plots for the UPS data and the SRS data

	UPS	SRS
Elevation	50	30
Traffic	20	25
Building.AW	50	28
Sun.exposure	30	25
Slope	28	21

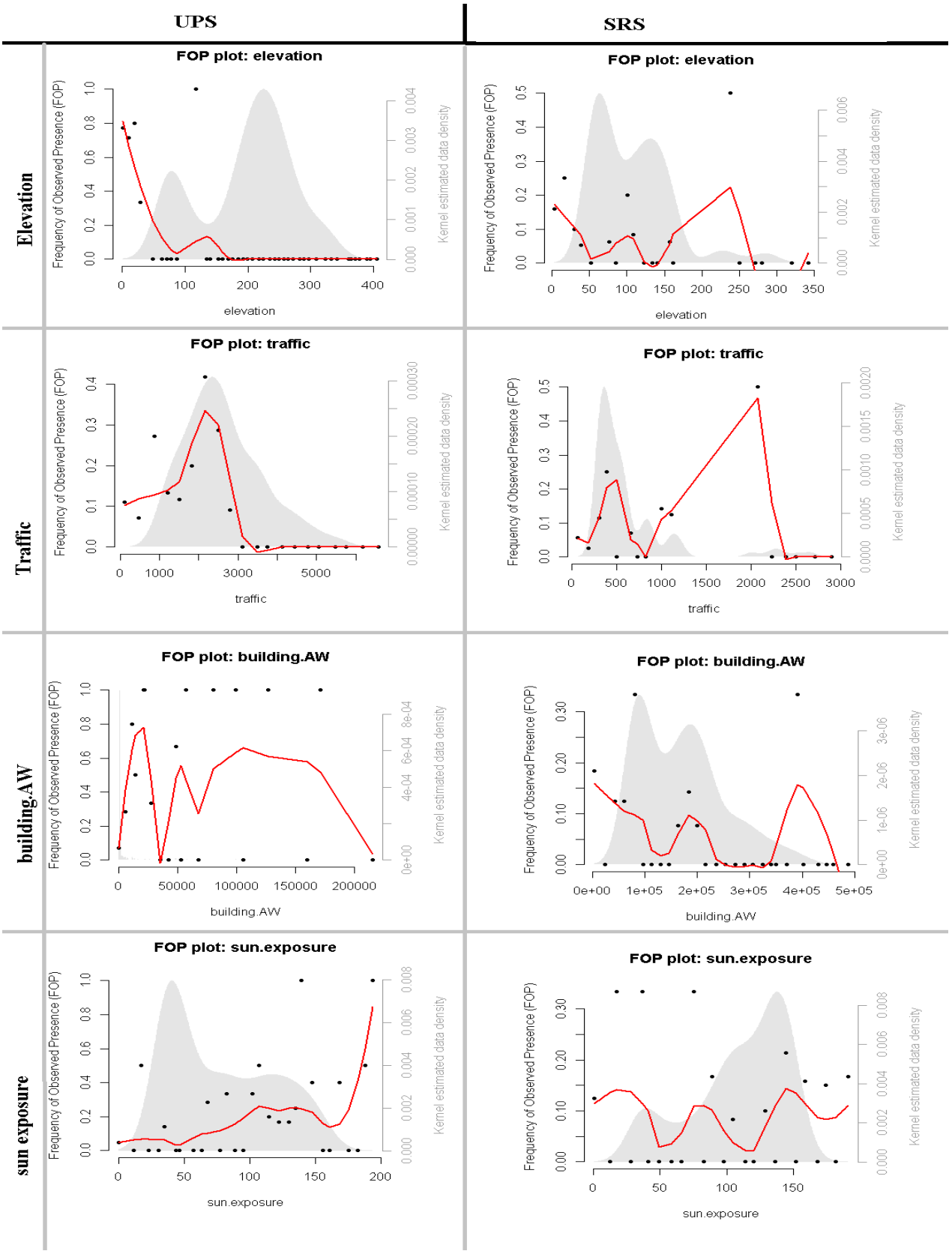


Figure 3.4: The FOP-plots for elevation, traffic, building.AW and sun exposure for the UPS and the SRS data.

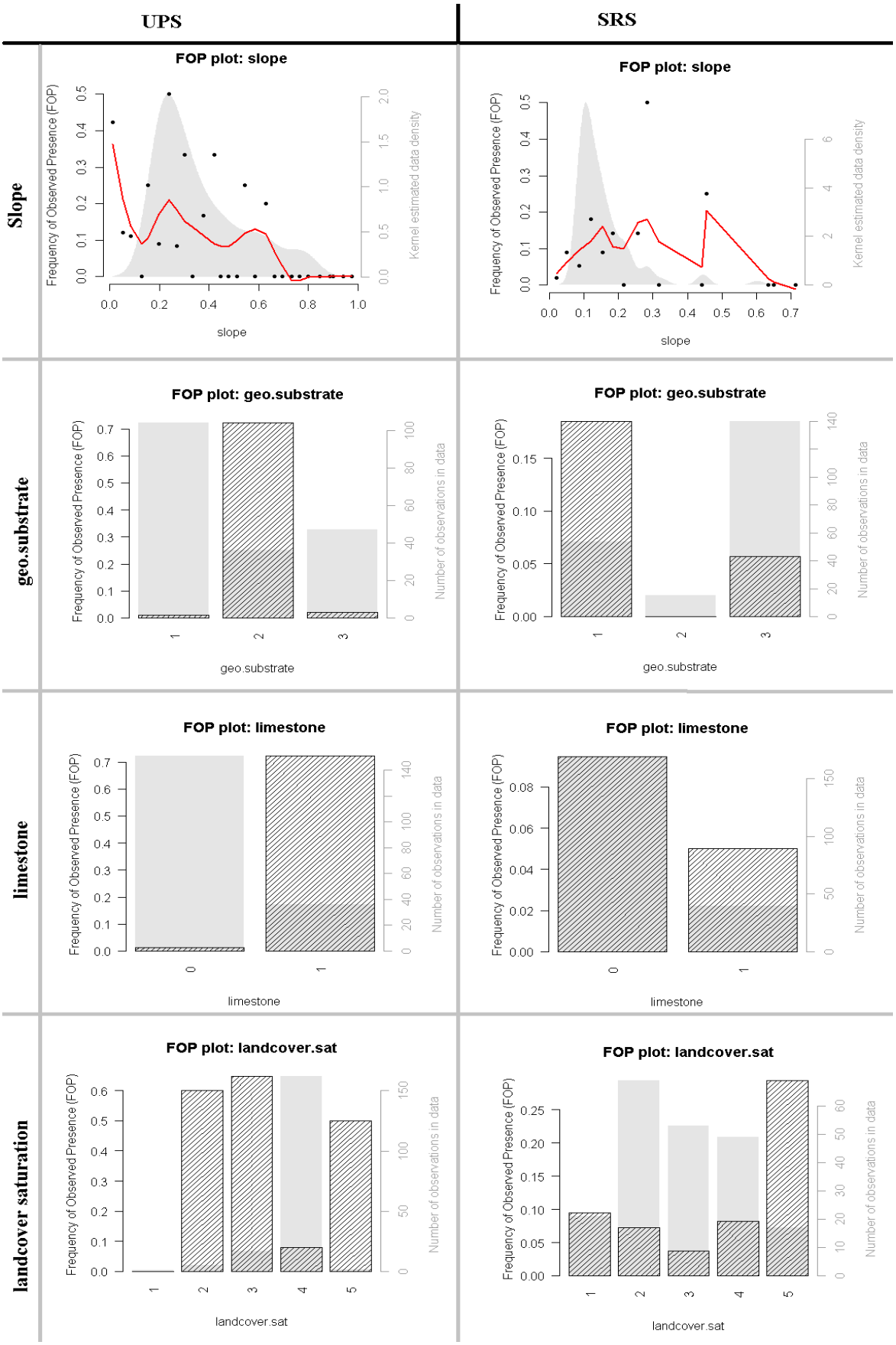


Figure 3.5: The FOP-plots for slope, geo.substrate, limestone and landcover saturation for the UPS and the SRS data.

The elevation FOP smoothed line for the UPS data a negative exponential form (figure 2a and 2c). The SRS data shows a quite different form (figure 2b).

The smoothed line for the traffic variable in the UPS data have a quite sharp unimodal shape that are a bit skewed to the left. The traffic variable for the SRS data (figure 2e) does also have an approximate unimodal shape. However, it is more skewed to the right and also have a smaller “peak” on the left side. The UPS data set FOP for the building.AW (figure 2g) has an unimodal shape. The EV range for the UPS data are shorter compared to the range of the SRS data.

The FOP for sun.exposure in both the UPS increases with higher sun.exposure unit values. The FOP for SRS-data (*figure 2k*) shows a quite different form where the shape is not as obvious.

The FOP slope for the UPS data starts with around 35 % of predicted empirical presence and drops unevenly with increasing slope values, while the FOP for the SRS data (figure 2n) have an intimation of an unimodal shape.

The UPS data has a high frequency of empirical presence in the goe.substrate category 2 (exposed limestone) and the limestone category 1 (limestone present). The SRS data has the highest frequency in category 1 (exposed nutrient poor bedrock) in the geo.substrate variable and category 2 in the limestone variable (limestone absent).

Category 5 in the landcover.sat (water edge) variable have the highest frequency of empirical presence compared to the other categories in the SRS data, while frequency in category 2, 3 and 4 are high in the UPS plot.

The shape of the FOP plots for both the UPS data and the SRS data were inconsistent when changing the interval value.

Correlation between the environmental variables

Table 3.2: Correlation table of the numerical variables for the UPS data. The lower triangle shows the spearman's tau coefficients and the upper triangle shows the p-values. The same is true for all the numerical vs numerical correlation tables.

	Aspect	B.AW	Curv	Elevation	Road.dis	Slope	Sun.exp	Temp	TPI	traffic
Aspect	1	0.1051	0.0557	0.7368	**0.0002**	0.2496	0.1327	0.1561	0.0586	0.1905
b.AW	0.1188	1	0.9594	**<0.0001**	**<0.0001**	**0.01791**	**<0.0001**	**<0.0001**	0.9731	**<0.0001**
Curv	0.1402	0.0037	1	0.4485	0.1182	**0.0075**	0.6266	0.5715	**<0.0001**	0.7061
Elev	0.0247	**-0.5237**	0.0558	1	0.2610	**0.0018**	**0.0001**	**<0.0001**	0.4638	**<0.0001**
Road	**- 0.2681**	**-0.3130**	-0.1146	0.0826	1	0.5695	0.4216	**0.0252**	0.1193	**<0.0001**
Slope	0.0846	**-0.1730**	**0.1950**	**0.2262**	-0.0419	1	0.0789	**0.0025**	0.0603	0.8257
Sun.e	-0.12	**0.3848**	0.0358	**-0.2798**	-0.0591	-0.1288	1	**0.0002**	0.2187	0.5925
Temp	-0.1041	**0.4412**	-0.0416	**-0.5523**	**-0.1636**	**-0.2191**	**0.2706**	1	0.5461	**<0.0001**
TPI	0.1385	-0.0025	**0.3320**	0.0539	-0.1143	0.1376	-0.0904	0.0444	1	0.7579
Traffic	-0.0962	**-0.2974**	-0.0277	**0.5089**	**0.3809**	0.0162	-0.0394	**-0.4558**	0.0227	1

Table 3.3: Correlation table of the numerical variables for the SRS data.

	Aspect	B.AW	Curv	Elevation	Road.dis	Slope	Sun.exp	Temp	TPI	traffic
Aspect	1	0.7849	0.8308	0.0962	0.0800	**0.0090**	**<0.0001**	0.1573	0.6767	0.6514
b.AW	-0.0190	1	**0.0306**	**0.0157**	**0.0002**	**<0.0001**	0.1893	**<0.0001**	0.1805	**0.0002**
Curv	-0.0149	**0.1496**	1	0.2773	0.6272	0.9061	0.4275	0.1036	**<0.0001**	0.2221
Elev	-0.1154	**-0.1670**	0.0755	1	**<0.0001**	0.9654	0.5534	0.4923	**0.0062**	**0.0006**
Road	-0.1214	**-0.2544**	-0.0338	**0.2846**	1	0.5968	0.2233	**<0.0001**	0.1276	**<0.0001**
Slope	**0.1803**	**-0.3804**	-0.0082	-0.0030	0.0368	1	**0.0271**	**<0.0001**	0.9982	0.0776
Sun.e	**-0.2690**	0.0912	0.0552	0.0412	-0.0846	**-0.1528**	1	**<0.0001**	0.5636	0.2236
Temp	-0.0982	**0.6972**	0.1129	-0.0477	**0.2984**	**0.4208**	**0.2664**	1	0.0902	**<0.0001**
TPI	-0.0290	0.0923	**0.3526**	**0.1887**	0.1057	0.0001	0.0402	0.1175	1	0.5921
Traffic	-0.0314	**-0.2581**	-0.0848	**0.2370**	**0.3604**	0.1223	0.0845	**-0.2667**	0.0373	1

Table 3.4: Correlation table of the numerical variables for the UPS+SRS data.

	Aspect	B.AW	Curv	Elevation	Road.dis	Slope	Sun.exp	Temp	TPI	traffic
Aspect	1	0.1872	0.1959	0.2675	**0.0009**	**0.0362**	**0.0031**	0.7759	0.2646	**0.03326**
b.AW	0.0664	1	0.1065	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	0.0680	**<0.0001**
Curv	0.0651	-0.0812	1	**0.0075**	0.1421	**0.0058**	0.4774	0.0901	**<0.0001**	0.1608
Elev	-0.0559	**-0.6050**	**0.1342**	1	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	0.0001	**<0.0001**
Road	**-0.1655**	**-0.6936**	0.0739	**0.5087**	1	**<0.0001**	**<0.0001**	**<0.0001**	**0.0310**	**<0.0001**
Slope	**0.1053**	**-0.4660**	**0.13825**	**0.3373**	**0.2780**	1	**<0.0001**	**<0.0001**	**0.01937**	**<0.0001**
Sun.e	**-0.1485**	**0.4480**	-0.0358	**-0.3876**	**-0.3555**	**-0.2889**	1	**0.0001**	0.0686	**<0.0001**
Temp	-0.0143	**0.8386**	-0.0853	**-0.5540**	**-0.6950**	**-0.4785**	**0.4668**	1	0.2039	**<0.0001**
TPI	0.0562	-0.0918	**0.3460**	**0.1899**	**0.1085**	**0.1175**	-0.0916	-0.0640	1	**0.0047**
traffic	**-0.107**	**-0.6819**	0.0706	**0.6067**	**0.7472**	**0.3210**	**-0.2973**	**-0.7379**	**0.1419**	1

Table 3.5: Reported test H statistics and p-values from kruskal wallis rank sum test between the categorical variables and the numerical variables in the UPS data.

	Geo.substrate		Landcover.sat		Limestone		substrate	
	H	p-value	*H*	p-value	*H*	p-value	*H*	p-value
Aspect	0.6286	0.7303	3.955	0.4121	0.3454	0.5567	0.1185	0.7307
Building.AW	**67.96**	**<0.0001**	**46.31**	**<0.0001**	**67.96**	**<0.0001**	**5.845**	**0.0156**
Curvature	**6.648**	**0.0360**	4.440	0.3498	0.479	0.4888	**4.7805**	**0.0288**
Elevation	**70.77**	**<0.0001**	**55.69**	**<0.0001**	**70.312**	**<0.0001**	2.967	0.085
Road.distance	4.8854	0.08693	8.1369	0.08669	0.0065	0.9358	**4.585**	**0.0322**
Slope	**21.27**	**<0.0001**	**19.57**	**0.0006**	**13.67**	**0.0002**	2.553	0.1101
Sun.exposure	**20.11**	**<0.0001**	**44.59**	**<0.0001**	**19.769**	**<0.0001**	3.312	0.0688
Temp.surface	**16.81**	**0.0002**	**23.0**	**0.0001**	**16.80**	**<0.0001**	1.135	0.2868
TPI	3.039	0.2189	4.0195	0.4034	0.6963	0.404	1.517	0.218
Traffic	**12.21**	**0.0022**	**12.91**	**0.0117**	**11.86**	**0.0006**	0.1665	0.6833

Table 3.6: Reported test H statistics and p-values from kruskal wallis rank sum test between the categorical variables and the numerical variables in the SRS data.

	Geo.substrate		Landcover.sat		Limestone		substrate	
	H	p-value	*H*	p-value	*H*	p-value	*H*	p-value
Aspect	1.179	0.5545	19.01	0.0008	1.24	0.2655	0.9944	0.3187
Building.AW	**46.69**	**<0.0001**	**23.68**	**<0.0001**	2.453	0.1173	**21.67**	**<0.0001**
Curvature	0.1234	0.9402	**12.71**	**0.0128**	0.0402	0.841	0.0034	0.9535
Elevation	**11.82**	**0.0027**	**9.853**	**0.0430**	0.2754	0.5997	**9.814**	**0.0017**
Road.distance	5.299	0.0707	**23.26**	**0.0001**	1.775	0.1827	0.1329	0.7154
Slope	**20.16**	**<0.0001**	4.532	0.3388	0.0950	0.7579	**12.23**	**0.0005**
Sun.exposure	**9.624**	**0.0081**	**40.78**	**<0.0001**	1.567	0.2106	**9.432**	**0.0021**
Temp.surface	**22.72**	**<0.0001**	**38.75**	**<0.0001**	**4.013**	**0.0451**	**20.42**	**<0.0001**
TPI	1.223	0.5426	**25.16**	**<0.0001**	0.0062	0.9374	1.164	0.2807
Traffic	3.588	0.1663	**13.13**	**0.0107**	0.4395	0.5074	1.453	0.2281

Table 3.7: Reported test H statistics and p-values from kruskal wallis rank sum test between the categorical variables and the numerical variables for the UPS+SRS data.

	Geo.substrate		Landcover.sat		Limestone		substrate	
	H	p-value	*H*	p-value	*H*	p-value	*H*	p-value
Aspect	1.196	0.5498	**12.56**	**0.0136**	1.826	0.1766	0.0331	0.8555
Building.AW	**97.87**	**<0.0001**	**169.9**	**<0.0001**	**20.39**	**<0.0001**	**62.60**	**<0.0001**
Curvature	**7.622**	**0.0211**	**22.57**	**0.0001**	0.2417	0.623	**6.834**	**0.0089**
Elevation	**81.82**	**<0.0001**	**134.8**	**<0.0001**	**51.58**	**<0.0001**	1.701	0.1922
Road.distance	**52.02**	**<0.0001**	**121.4**	**<0.0001**	1.865	0.172	**49.19**	**<0.0001**
Slope	**62.13**	**<0.0001**	**61.40**	**<0.0001**	**6.724**	**0.0095**	**36.33**	**<0.0001**
Sun.exposure	**33.39**	**<0.0001**	**130.7**	**<0.0001**	**7.461**	**0.0063**	**17.57**	**<0.0001**
Temp.surface	**70.65**	**<0.0001**	**157.6**	**<0.0001**	2.312	0.1283	**62.11**	**<0.0001**
TPI	**8.797**	**0.0123**	**29.62**	**<0.0001**	0.1823	0.6694	**7.765**	**0.0053**
Traffic	**36.57**	**<0.0001**	**109.7**	**<0.0001**	1.731	0.1883	**31.48**	**<0.0001**

Table 3.8: Reported test-statistic from χ^2-test between categorical variables in the lower triangle, with the p-values in the upper triangle, for the UPS data.

	Geo.substrate	Landover.sat	Limestone	substrate
Geo.substrate	1	**<0.0001**	**<0.0001**	**<0.0001**
Landcover.sat	**56.13**	1	**<0.0001**	0.8855
Limestone	**187**	**53.65**	1	**0.0002**
substrate	**187**	1.1545	**13.36**	1

Table 3.9: Reported test-statistic from χ^2-test between categorical variables in the lower triangle, with the p-values in the upper triangle, for the SRS data.

	Geo.substrate	Landover.sat	Limestone	substrate
Geo.substrate	1	**0.0002**	**<0.0001**	**<0.0001**
Landcover.sat	**30.58**	1	**0.0005**	**0.0011**
Limestone	**76.30**	**19.82**	1	0.6284
substrate	**209**	**18.30**	0.2342	1

Table 3.10: Reported test-statistic from χ^2-test between categorical variables in the lower triangle, with the p-values in the upper triangle, for the UPS+SRS data.

	Geo.substrate	Landover.sat	Limestone	substrate
Geo.substrate	1	**<0.0001**	**<0.0001**	**<0.0001**
Landcover.sat	**79.68**	1	**<0.0001**	**<0.0001**
Limestone	**256.3**	**23.83**	1	**0.0079**
substrate	**396**	**58.33**	**7.051**	1

The UPS+SRS data sets have more significant correlation compared to the other data sets. We see a strong correlation between geo.substrate and elevation in both the UPS data and the UPS+SRS data.

3.4 The MIAmaxent logistic regression models

Table 3.11: The formula and variation explained for all the models made in the analysis.

	Mod 1	Mod 2	Mod 3	Mod 4	Mod 5	Mod 6
Model formula	RV ~ elevation_D05 + geo.substrate_BX2	RV ~ geo.substrate_BX2 + traffic_HF8 + traffic_L	RV ~ elevation_HR3 + geo.substrate_BX1	RV ~ building.AW_T4 + landcover.sat_BX5	RV ~ elevation_HR4 + traffic_D05 + traffic_HF9 + limestone_BX0 + traffic_D05:limestone_BX0+traffic_HF9:limestone_BX0	RV ~ geo.substrate_BX2 + traffic_D05 + traffic_HF9 + building.AW_D05 + landcover.sat_BX5 + sun.exposure_D05 + sun.exposure_M + slope_HF13 + geo.substrate_BX2:trafficD05 + geo.substrate_BX2:traffic_HF9
Null deviance	157.9	157.9	122.7	122.7	284.5	284.5
Residual deviance	43.98	53.58	107.9	110.4	157.5	167.4
Variation explained	0.721	0.66	0.121	0.100	0.446	0.411

See *table 2.3* .for further details on what the difference is between the models.

In the first round of the subset selection for model 1, the model with elevation was selected. In the second round, the model with geo.substrate included explained significantly more than the other possible models. Additional variables in the third round did not produce any models that explained significantly more variation. As such, model 1 ended up with elevation and geo.substrate as the only significant explanatory variables.

For model 2, the model with geo.substrate got selected in the first round, while in the second round the model with geo.substrate and traffic got selected.

Subset selection for model 3: Elevation in the first round, geo.substrate in second round. The selection steps for model 4: building.AW in the first round followed by landcover.sat in the second round.

The subset selection for model 5: Elevation in the first round, traffic in the second round, limestone in the third round and an interaction term between traffic and geo.substrate in the fourth round. Including additional variables did not produce models that explained significantly more variation.

The subset selection that produced model 6 consisted of seven steps: Geo.substrate in the first round, traffic in second, building.AW in third, landcover.sat in fourth, sun.exposure in fifth, slope in sixth and an interaction term between traffic and geo.substrate in seven.

Table 3.12: The model formula and variation explained for the models made in MIAmaxent where the alpha value were specified at 0.001. Note that the model formula for model 9 and 10 could not be obtained. Variation explained are determined by the null deviance and the residual value ((Null-residual)/Null deviance).

	Mod 7	Mod 8	Mod 9	Mod 10	Mod 11	Mod 12
Model formula	RV ~ elevation_D05	RV ~ geo.substrate_BX2	NA	NA	RV ~ elevation_HR4 + trafficD05	geo.substrate_BX2 + traffic_D05 + geo.substrate_BX2:traffic_D05
Null deviance	157.9	157.9	NA	NA	284.5	284.5
Residual deviance	52.12	63.81	NA	NA	174	190.8
Variation explained	0.67	0.596	NA	NA	0.39	0.33

There were no significant DV from the selectDVforEV()-function for model 9 and 10, which is the reason why model formula and variation explained could not be acquired for these models.

Single-effect response plots

For the models with elevation as an explanatory variable do we see that the highest probability along elevation gradient is around 5 meters. From 5 meters, the probability drops substantially. The same pattern is true also for mod 5. The peak of the increase in predicted probability of presence for model 1 were around 80 %, for model 3 the increase peaked at around 25 % while for model 5 the increase peaked at around 45 %.

For the four models with geo.substrate as one of the significant explanatory variables, category two (exposed limestone) were the category that gave significant increase in predicted presence. Geo.substrate were also a significant predictor in model 5. However, category 1 (exposed nutrient-poor bedrock) were the category that gave an significant increase in predicted probability.

Traffic, which were chosen as a significant explanatory variable in model 2,3 and 4 does also share a similar shape in the response curve. For all the models that included traffic as a predictor, the relative increase in predicted presence peaked with around 40 % increase.

For model 3, which has limestone instead of geo.substrate as the significant variable that accounts for calcium, we see that the presence of limestone increased the predicted probability of presence with around 35 %.

Model 4 was the only model that chose both slope and sun.exposure as significant explanatory variables. Sun.exposure peaks at 70 % increased relative predicted presence. The slope variable gives 12 % increased predicted presence up to a certain point.

Only model 4 and 6 has building.AW and landcover..sat as one of their predictors. The response gradient for building.AW peaks at around 35 % relative increased predicted probability of success. The shape of the response for model 6 are similar. For the landcover.sat variable, category 5 (water edge) gives a 20 % increased relative predicted probability of presence compared to the other categories, while the same category for model 6 gives around 25 % relative increased probability of presence.

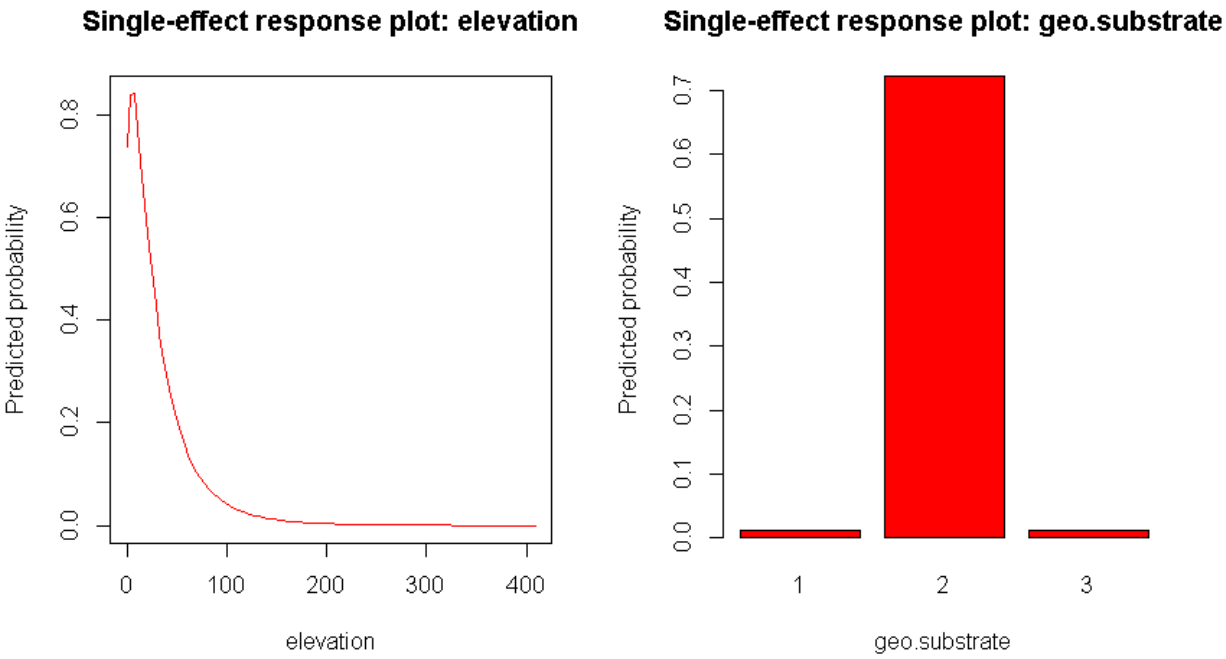


Figure 3.6: The single effect response plots for model 1 (elevation and traffic).

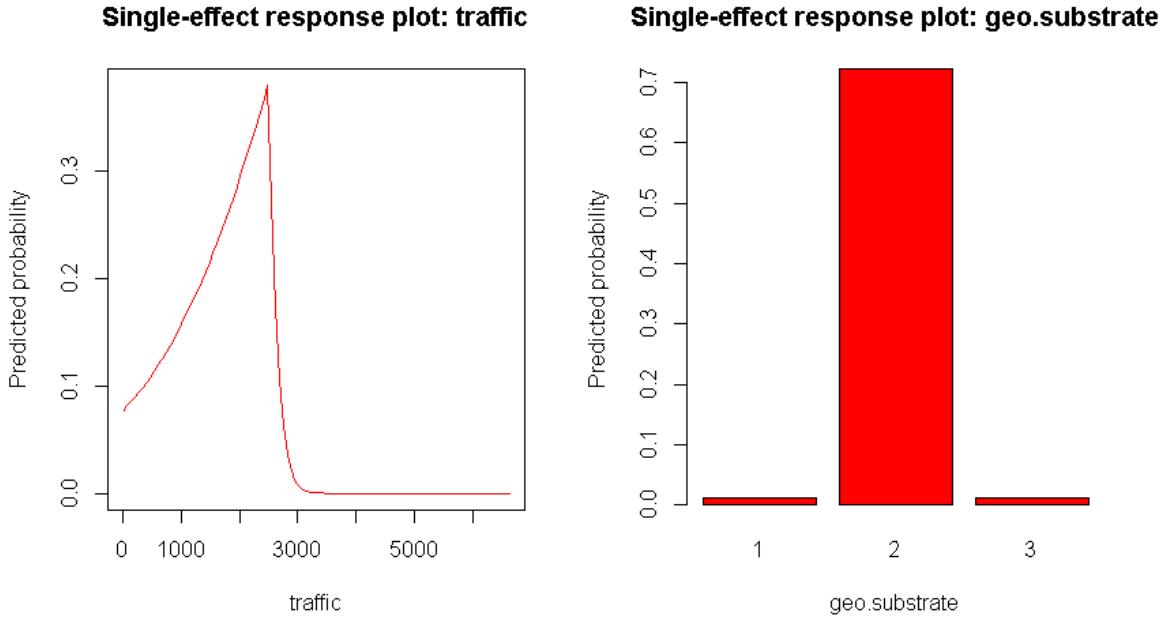


Figure 3.7: The single-effect response plots for model 2 (traffic and geo.substrate).

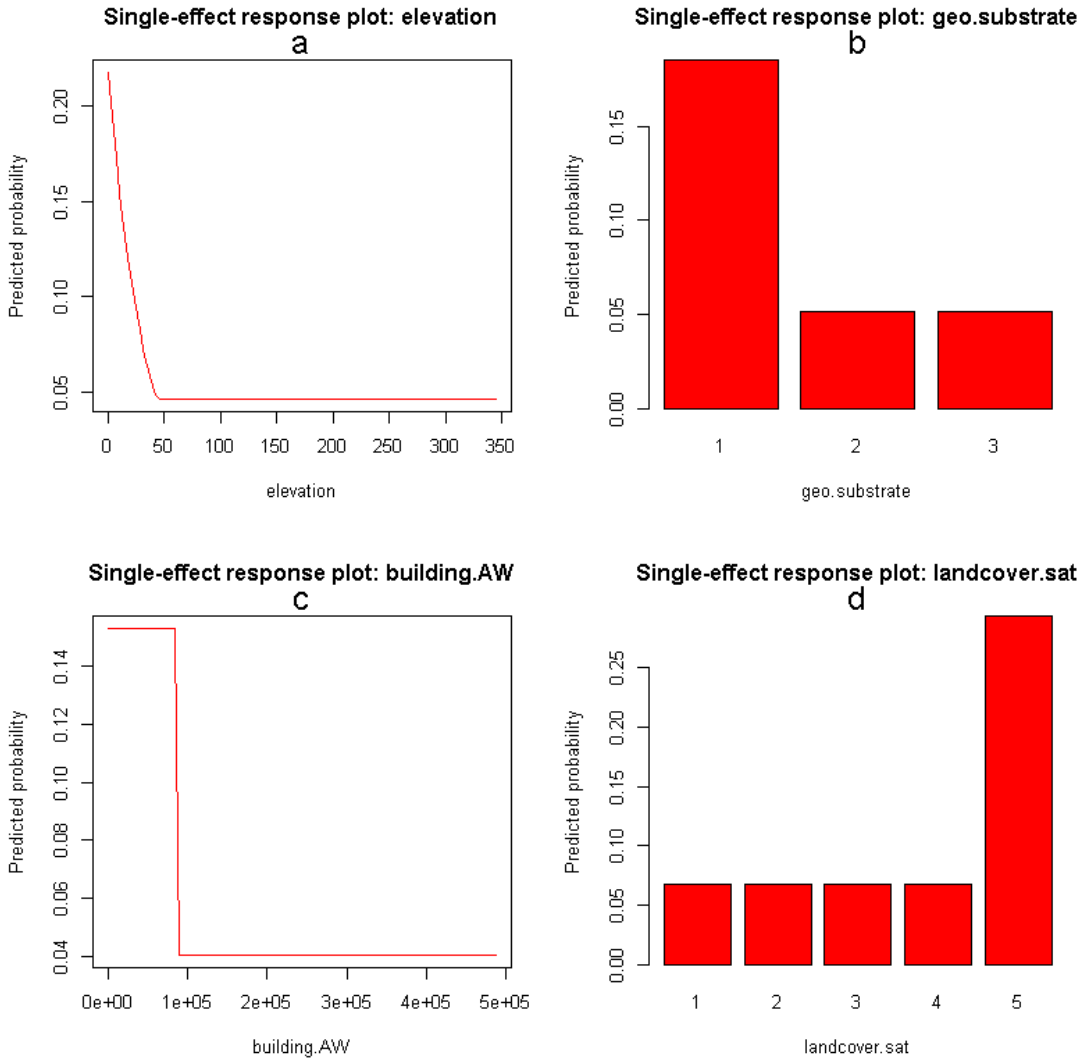


Figure 3.8: The single-effect response plots for model 3 and 4 (model with the SRS data as response variable). *a)* The response for elevation in model 3. *b)* The response for geo.substrate in model 3. *c)* Response for building.AW selected in model 4. *d)* The response for landcover.sat selected in model 4.

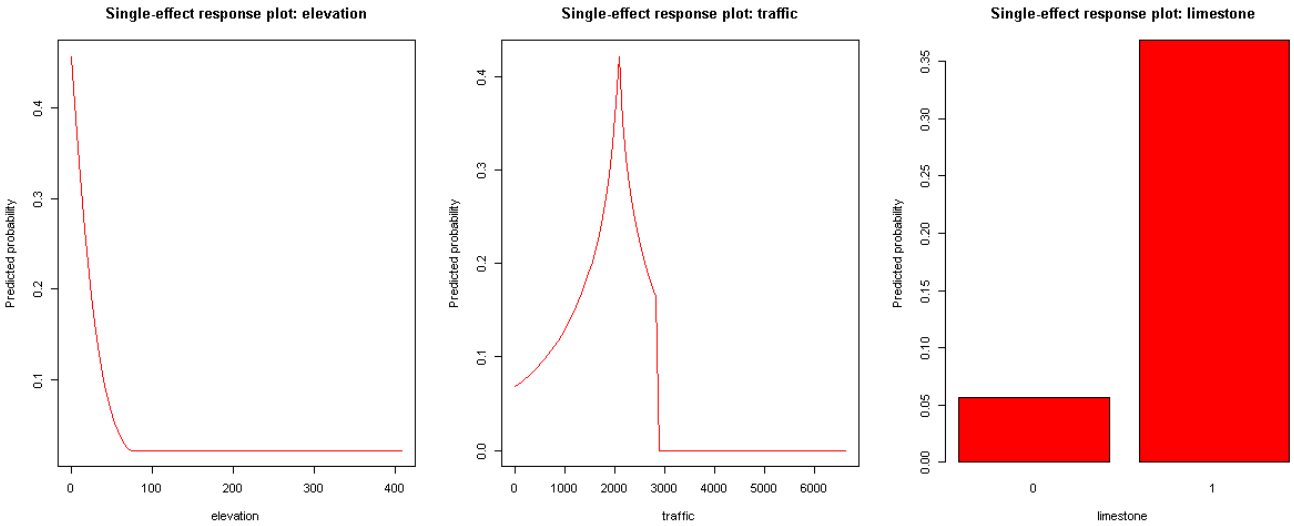


Figure 3.9: The single-effect response plots for model 5 (elevation, traffic and limestone). Model 5 and 6 had the UPS+SRS as the response variable.

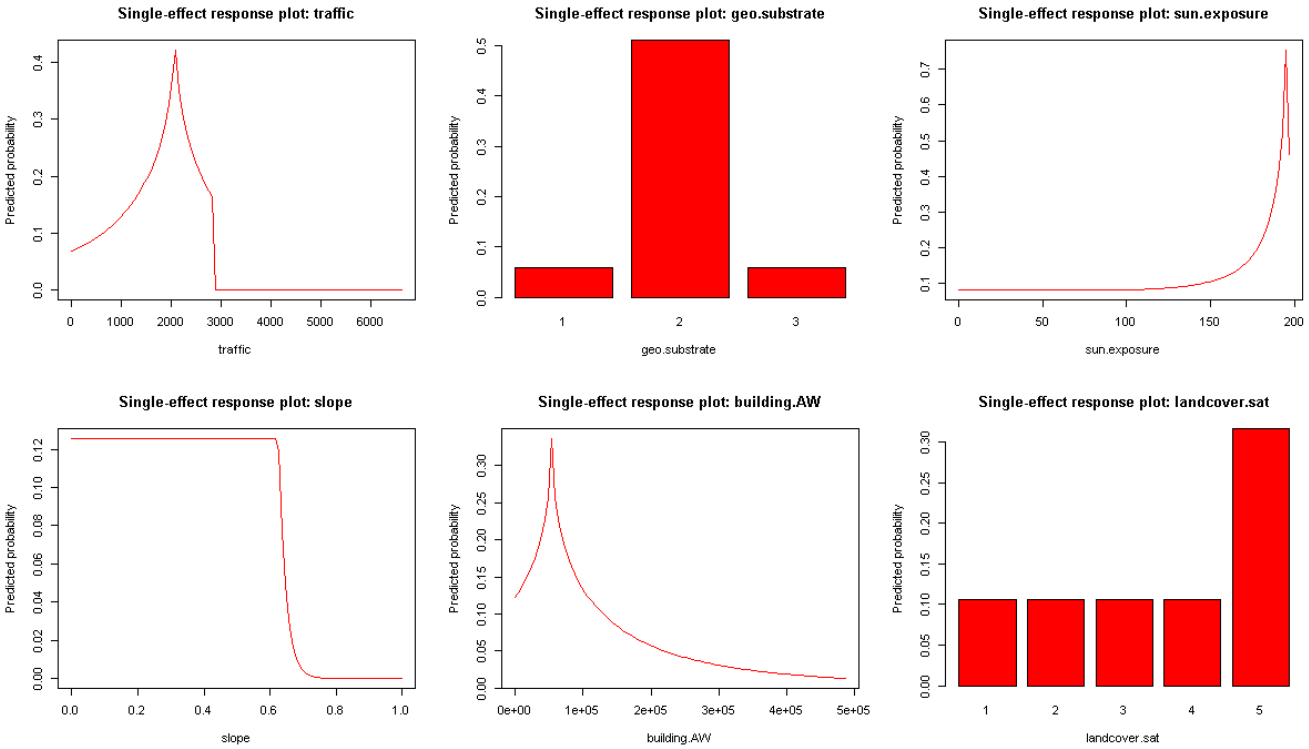


Figure 3.10: The single-effect response plots for model 6 (traffic, geo.substrate, sun.exposure, slope, building.AW, landcover.sat).

Both elevation response curves from model 7 and model 11 (3.11a) and 3.11c) respectively) have an approximately negative deviation form. However, 3.11b) does start off with a considerably higher predicted probability of presence (around 80 % increased predicted probability of presence at its highest) compared to 3.11c) (around 40 % increase in predicted probability of presence at its highest). The two response plots for geo.substrate are also quite similar; Presence of either category one or three does not give any increase in predicted probability while category two gives around 70 % increase in predicted probability, according to model 8. In model 12 both category two gives around 45 % relative increase in predicted probability. The response curve for traffic in both model 11 and 12 have an identical shape; A quite sharp almost uni-modal shape, with a peak around 2000 cars per day at which the model assumes a 40 % increase in predicted probability of presence.

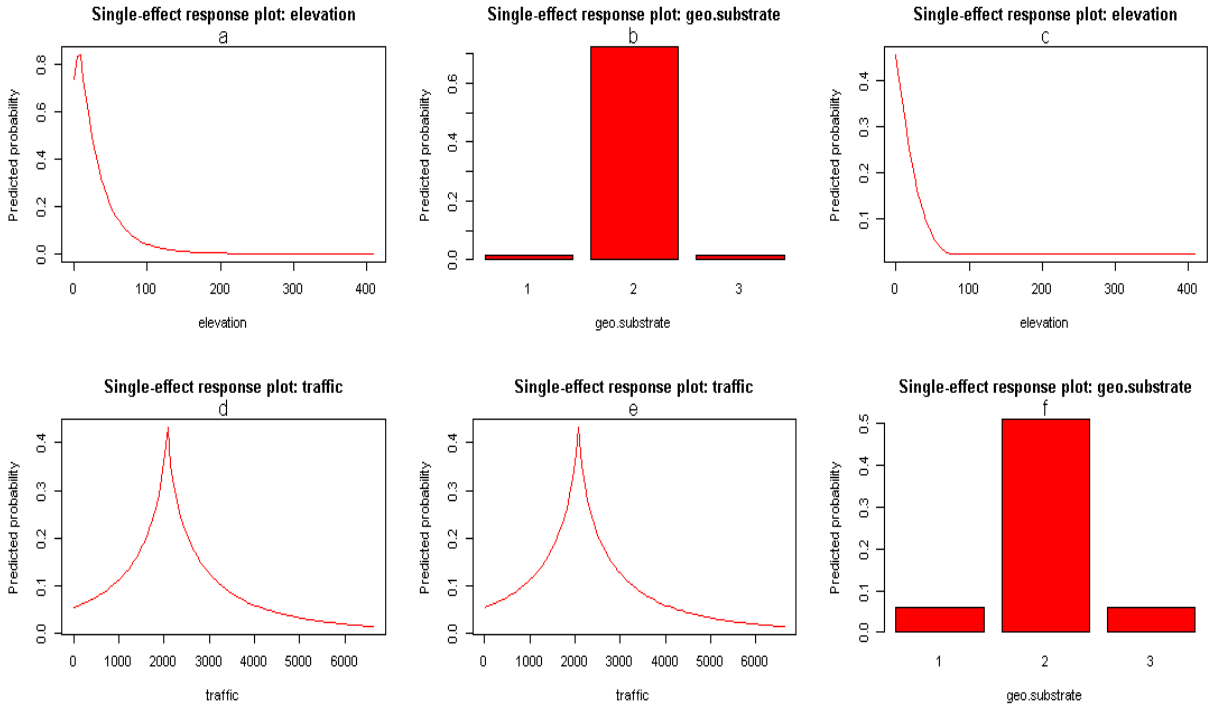


Figure 3.11: a) The response plot for the selected variable in model 7, which in this case is only elevation. b) The response curve for the selected variable in model 8, which is only geo.substrate. c) The response curve for elevation in model 11. d) The response curve for traffic in model 11. e) The response curve for traffic in model 12. f) The response plot for geo.substrate in model 12.

ROC plots and AUC values

Model 1,2 and 3 have quite similar AUC-value (around 0.75), while model 3 and 4 are the least predictive ones with an AUC around 0.7.

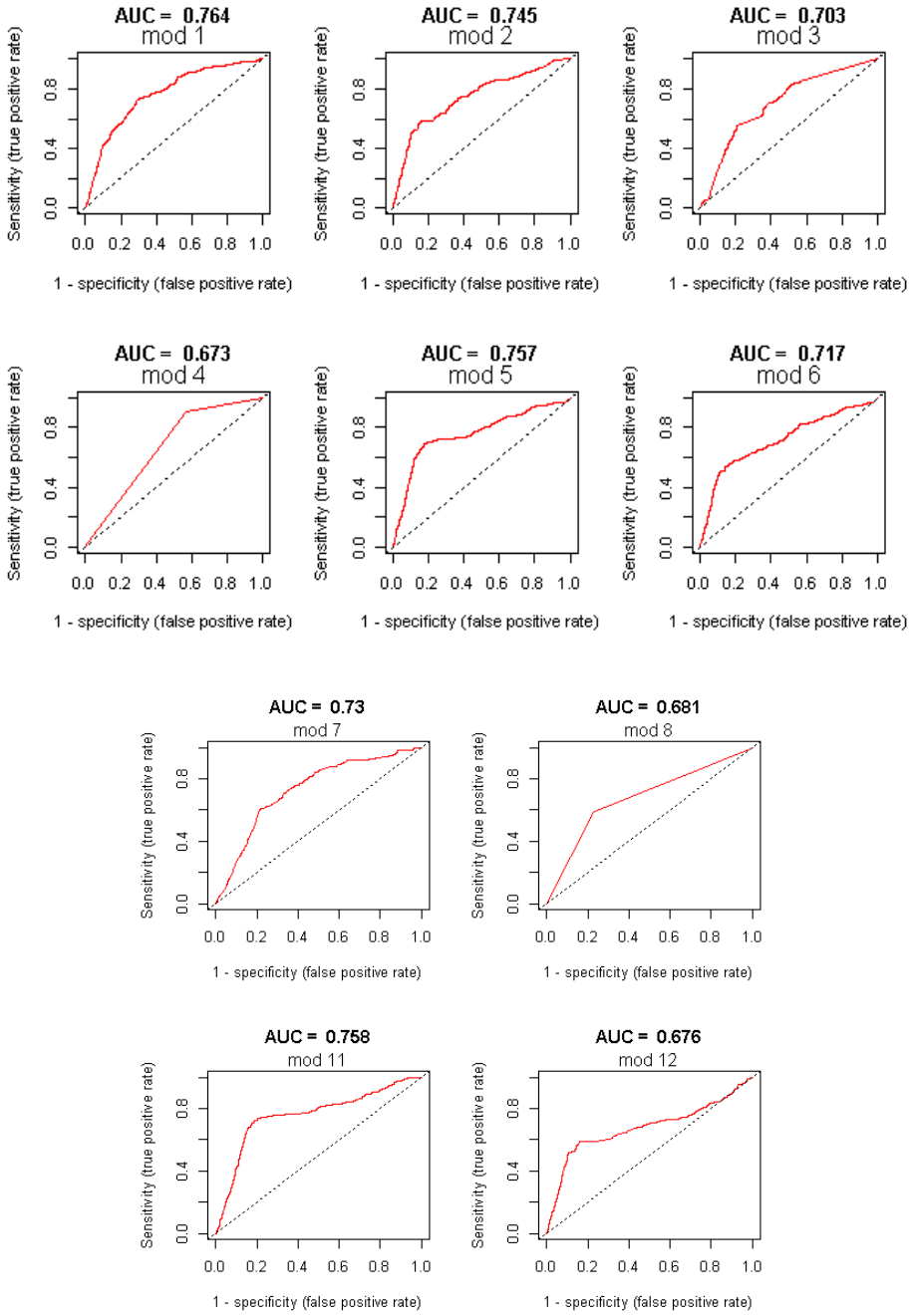


Figure 3.12: The ROC-curve and corrisponding AUC-value for model 1-8 and mod 11-12.

Predicted distribution maps

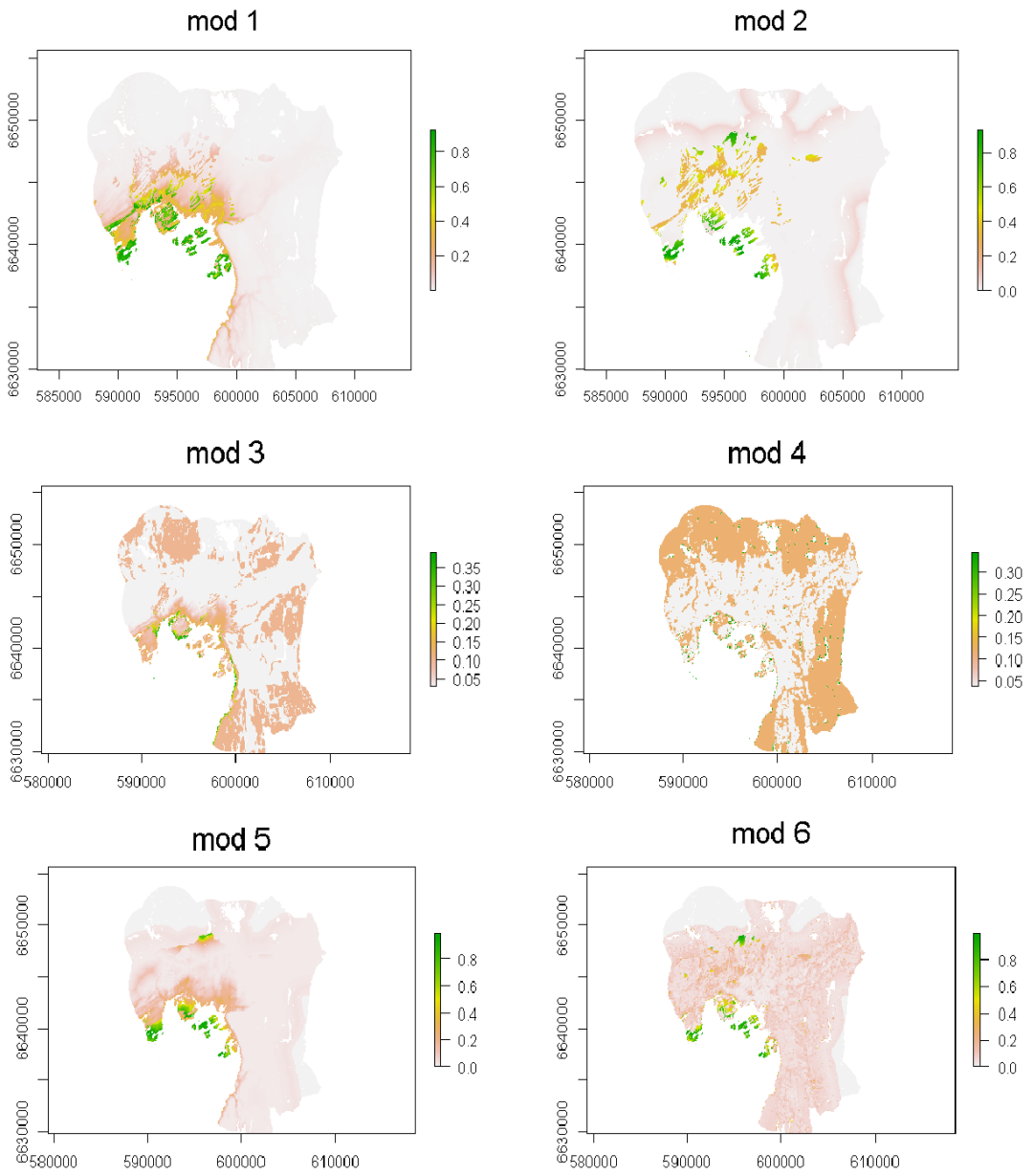


Figure 3.13: The predicted distribution for model 1, 2, 3, 4, 5 and 6.

We see the presence of the elevation variable for model 1 and 5, as there is a clear gradient of probability of presence from the ocean to the forests. The islands in the inner fjord, Fornebu, and Bygdøy are all location where the probability of observing one of the targeted species according to model 1 and 5. We can also observe that model 1-4 considers the forests around the main city as low to 0 % probability of presence zones.

We see that the predicted distribution for the models with alpha value 0.001 “mirror” the predicted distribution of their equivalent models with alpha value 0.05. The main difference is that these have a more restricted distribution of high predicted probability.

The map for both model 2 and 3 shows large areas that are low to moderate probability of presence. They do not treat the islands, Bygdøy and Fornebu as high probability areas in the same manner as the other models. For model 1, 2, 5 and 6 the scale of predicted probability are from zero percentage to around 90 % (0.8-0.9), whereas the scale for model 3 and 4 goes from zero to only around 30 % (0.3).

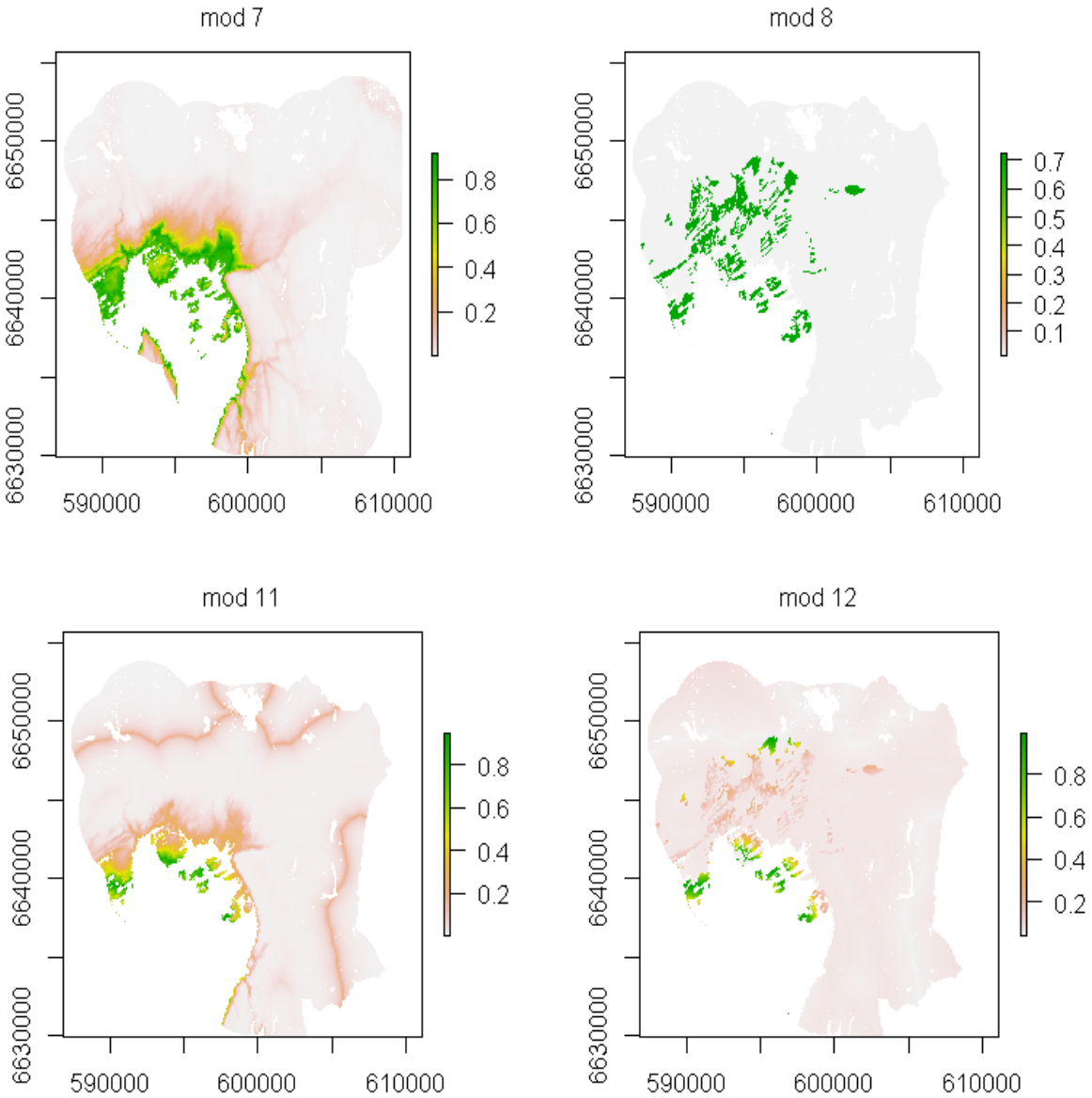


Figure 3.14: The predicted distribution for model 7, 8, 11 and 12.

Both model 7 and 8 ended up with only one significant variable (elevation and geo.substrate respectively). Model 11 ended up with elevation and traffic as significant EVs, elevation first followed by traffic. Model 12 ended up with geo.substrate, traffic and an interaction term between those two, geo.substrate in the first round, traffic in the second and the interaction term in the third round. Model 9 and 10 had no significant derived variables.

4. Discussion

Unequal probability sampling as a method to find rare species

The non-parametric proportion test showed a significant difference in the proportion of presence point between the UPS data set and the SRS data set. I will therefore argue that using UPS to find rare species is a more appropriate sampling method than stratified random sampled method. I do think that this comparison is not an ideal situation when inferring which sampling method suits best for finding rare species. As we can see from *figure 3.2b*, the SRS data does not have many observation units on the part of study area where geo.substrate category 2 (exposed limestone) were present. Presumably, this means that the SRS data does not include many observation units with limestone present and ragged substrate, which is already established as the most important EV for the targeted species.

This is further exemplified by the FOP-plot for geo.substrate. The background data for category two is less than the background data density for category one and three. We can also observe that the background density for category two in the FOP-plot for the UPS data are comparably higher. In other words, the UPS data have more observation units where exposed limestone is present. I think two factors are at play here; Firstly that the UPS data drew more observation units on areas with exposed limestone because of the nature of the sampling method, secondly that the SRS data, indirectly because how the sampling scheme is structured, drew fewer observation units with exposed limestone than it otherwise would.

I assert that the main problem in this situation is that the SRS data were gathered specifically for a different purpose (J. B. Halvorsen, 2019). I therefore argue more research is needed before one can be confident in the conclusion that UPS as a method for finding rare species is the preferred sampling method. Specifically, I think the most preferred scenario is the comparison of an unequal probability sampled data with an equal or simple random sampled data collected in the same scale with the same number of observation units (Meng, 2013). However, the results of this study at least partly demonstrates that unequal probability sampling method in case of this study area and these investigated species will be able to over-sample observation units that have environment conditions (exposed limestone) that favors the species in question.

The data set size and variable selection criteria

The effect of the total number of observation units was demonstrated with the UPS+SRS models (model 5, 6, 11 and 12). Model 5 and 6 (*table 2.3*) were the two models with the most significant variables. However, we can see from model 11 and 12 that a more conservative model selection criterion, i.e lower alpha value, did produce models with only two significant variables. I think this shows how important it is to adapt the alpha value in these situations, as that will directly affect which variables are chosen. I think in the case for the models made with the UPS+SRS data sets, the selection process for model 4 and 5 (UPS+SRS with alpha = 0.05) could have been more conservative.

Model 6 had in total six significant explanatory variables, which for me seems excessive. I do think it is reasonable to assume that some of the significant EVs in the final version of model 6 got selected because they accidently had a non-random pattern with regards to the

UPS+SRS data set. If we compare model 6 with model 12, we can see that fewer EVs were deemed significant in the model selection process. No derived variables for the SRS data were identified as significant when alpha were specified as 0.001. This further supports the notion that there were weak to no environment-species pattern in the SRS data.

Indirect variables

As we see from this study; EVs can be selected and deemed significant and yet still be ecologically meaningless. Elevation is an example of this. It was the most significant variable in model 1, and second most significant in model (see *table 3.11*). Yet, for these particular targeted species, I will argue that elevation is not an ecologically important variable, as the range (i.e the size of the study area) of the elevation difference is so small that the usual effects from this variable are not present (atmospheric pressure and temperature).

I found it challenging to infer which variable actually describes the distribution of the targeted species elevation. We can see from the models without elevation included (model 2, 4, 8 and 12) that elevation were often "replaced" with traffic, i.e traffic was the variable that explained the most variation among the variables that significantly correlated with elevation. I think the traffic is an indirect variable, as it is a manmade phenomenon. This variable is defined as cars per day (The environment variables), which for me makes the interpretation of this variable challenging. I suspect that traffic is somehow tied with the substrate, all though I find it hard to prove it directly.

Using unequal probability sampling in distribution modelling

It was not clear which of the twelve models were the most predictive one. As we can see from the AUC-value and the corresponding ROC-curve, they all had near the same prediction ability. This can mean that the test data used was not optimal. It is therefore not straight forward to evaluate which model is the best one in predicting the presence or absence of the targeted species. It is not obvious to compare the variation explained, as it is a measure that tells more about the data itself rather than how general the model is. I think the test data were insufficient here, as it is not a true PA-data set and is based on the same study area. The ROC-curves and the AUC-values may have been different if the test data were better (Vollering et al., 2019). A test data from another study area is desired if one wishes to test how general the model is (Araújo et al., 2019).

In this case I will argue that it is better to look at the selected EV within each model, as they can tell (as long as one has an idea of which EVs are the most important) how widely applicable the model is with regard to the species ecology. One can therefore disregard model 1 and 3, as they both included elevation as one of their significant EVs (since elevation is in fact does not predict the species' distribution directly). With model 1 and 3 excluded, we can then compare model 2 and 4. Of the two models, model 2 has fewer explanatory variables than model 4.

From the predictive distribution map, we see that both model 2 and model 4 has correctly identified the islands, Fornebu and Bygdøy as high probability areas. However, we see that

model 4 has a lot more moderate to low probability areas compared to model 2. This is not that surprising, as model 4 has more significant EVs compared to model 2. Based on the UPS data set and prior study, one can argue that model 4 is too generous with regard to what it deems as moderate to high probability areas. It may be right to assume that among the models presented in this study, model 2 is the best one for predicting the investigated species' distribution. If we compare model 2 and 12 with (Wollan, 2011), we can see that exposed limestone is important. This means that the procedure presented here, will indeed produce models that at least in part reflects the actual species-environment relationship.

In addition, there is a potential problem with the distance of the observation units themselves. If the study area is relatively constrictive with respect to location with moderate to high probability of success, then a situation where there are many OUs grouped together can arise. This can lead to bias caused by autocorrelation, i.e observed patterns in the data that are due to dispersion mechanics and/or historic events as opposed to environmental factors (Irvine et al., 2018). A few of the observation units from the UPS data may be affected by autocorrelation. For instance observation unit 97 and 76; in both these observation units, a total of 10 distinct species were observed, which is the highest number among all observation units visited. Both of these observation units were on the same island (Gressholmen).

The degree of autocorrelation is potentially higher in the SRS data. We can see from figure *figure 3.3* that the SRS observation units have a more clustered distribution. This is not that surprising, given the nature of the sampling method. However, I think that this clustering is probably causing autocorrelation. This means that some (if not all for model 3 and 4) of the selected variables in model 3, 4, 5, 6, 11 and 12 (the models where the SRS data were a part of the response variable data) represents qualities of the data rather than species-environment relationship.

What this means for the species, the study area and distribution modelling

Overall I will argue that the UPS data captured a better representation of the environment-species relationship. The proportion test (*henvise til den delen av resultater som viser prop.testen*), and the comparison of model 1,2 with 3,4, do support this claim. However, as we have seen, there are problems with the SRS when using it for this particular comparison. If the UPS data were compared with a data sampled equally across the whole study area, one could probably infer more conclusively the relative performance of the unequal probability sampling method for species adapted to exposed limestone substrate in Oslo and surrounding area.

Yet, I do think unequal probability sampling is a good approach if one wishes to find the species studied here. UPS could also be combined with a gradsect style sampling approach (Guisan & Zimmermann, 2000), to capture variation along what one might suspect is the most important environment gradients (for instance how much calcium is available in the soil, and how rugged it is). This all depends on what information is already available for the area one wishes to study, but for Oslo at least, this approach seems better suited than the sampling scheme for the stratified random sampled data.

Conclusion

The UPS data and the SRS data differed significantly in terms of the prevalence, where the UPS data had a higher proportion of presences of the investigated species than the SRS data. This implies that UPS as a sampling method is better suited to finding rare species. However, because the category exposed limestone were comparably poorly represented in the SRS data the results are not as conclusive as one could hope for. In addition, the potential presence of autocorrelation in both data sets makes the interpretation of the results somewhat challenging. The conclusion could maybe be more robust if the UPS data was compared to a simple random sampled data set.

There were clear differences between the models made with the UPS data and the models made with the SRS data. All the models associated with this data set had variables connected to limestone. Even when the model selection criteria were more conservative, three of the four models made with the UPS data (model 8, 11, 12) selected exposed limestone as a significant predictor for the presences of the investigated species. The models made from the SRS data did not select any variables associated with limestone, which indicates that the UPS data captured a better representation of the environment-response relationship for the investigated species.

The effect of larger population size were strongest when not adapting the model selection criteria, as model 6 had the most significant variables. I suspect some of those variables represent qualities of the data rather than actual responses from the targeted species.

The models with elevation captured more variation in the response variable than the other models. Nevertheless, correlation and statistic tests showed that the variable was associated with many of the other variables included in the model selection. This mean that if one wishes to predict the presences of the targeted species in this study, using a distribution model where elevation is included might not be a bad idea. However, the same models might not be applicable for other study areas

There are still some questions that remain unanswered. For instance, what does the traffic variable actually represent in the models where it was present? Is it an indirect variable, like elevation, or does numbers of cars per day actually have an impact on the distribution of the investigated species? I suspect that traffic is an indirect variable, but more research may be needed to understand the relationship between the traffic variable and the investigated species.

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*: Wiley.

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology, 95*(3), 631-636. doi:10.1890/13-1452.1

Araújo, M. B., Anderson, R. P., Márcia Barbosa, A., Beale, C. M., Dormann, C. F., Early, R., . . . Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances, 5*(1), eaat4858. doi:10.1126/sciadv.aat4858

Aune, B. (1993). *Temperaturnormaler normalperiode 1961-1990*. Retrieved from

Austin, M. P. (1980). Searching for a model for use in vegetation analysis. *Vegetatio, 42*(1), 11-21. doi:10.1007/bf00048865

Austin, M. P., & Gaywood, M. J. (1994). Current Problems of Environmental Gradients and Species Response Curves in Relation to Continuum Theory. *Journal of Vegetation Science, 5*(4), 473-482. Retrieved from <Go to ISI>://WOS:A1994PN43700005

Austin, M. P., & Heyligers, P. C. (1989). Vegetation Survey Design for Conservation - Gradsect Sampling of Forests in Northeastern New-South-Wales. *Biological Conservation, 50*(1-4), 13-32. doi:Doi 10.1016/0006-3207(89)90003-7

Barton, D., Grimsrud, K., Greaker, M., Heyman, A., Chen, X., Garnåsjordet, P., & Aslaksen, I. (2017). *Monetary valuation methods in urban ecosystem accounting - examples of their relevance for municipal policy and planning in the Oslo metropolitan area*.

Blumentrath, S. (2016). Land Surface Temperature in Oslo at 2015-07-02. In L. S. T. i. O. a. 2015-07-02 (Ed.), (pp. LandSurfaceTemperature in Oslo at 2015-2007-2002 (start of the periode with the warmest day in 2015), estimated from Landsat2018 imagery using the i.landsat2018.swlst (https://grass.osgeo.org/grass2070/manuals/addons/i.landsat2018.swlst.html) module in GRASS GIS 2017 WITHOUT correction for land cover type. http://urban.nina.no/layers/geonode%2013Alc81980182015183lgn81980182015100_lst_const_lc).

Edvardsen, A., Bakkestuen, V., & Halvorsen, R. (2011). A fine-grained spatial prediction model for the red-listed vascular plant Scorzonera humilis. *Nordic Journal of Botany, 29*(4), 495-504. doi:10.1111/j.1756-1051.2010.00984.x

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. doi:https://doi.org/10.1016/j.patrec.2005.10.010

Fisher, R. A. (1925). *Statistical Methods for Research Workers*: Oliver and Boyd.

Gravetter, F. J., & Forzano, L. A. B. (2009). *Research Methods for the Behavioral Sciences*: Wadsworth Cengage Learning.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling, 135*(2-3), 147-186. doi:Doi 10.1016/S0304-3800(00)00354-9

Halvorsen, J. B. (2019). *Characterisation and typification of urban ecosystem types – A test of the NiN system.* (Master). University of Oslo, University of Oslo.

Halvorsen, R. (2012). A gradient analytic perspective on distribution modelling. *Sommerfeltia, 2012 v.35*, pp. 1-165. doi:10.2478/v10208-011-0015-3

Halvorsen, R. (2013). A strict maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. *Sommerfeltia, 36*, 1-132. doi:10.2478/v10208-011-0016-2

Henderson, E. B., Ohmann, J. L., Gregory, M. J., Roberts, H. M., & Zald, H. (2014). Species distribution modelling for plant communities: stacked single species or multivariate modelling approaches? *Applied Vegetation Science, 17*(3), 516-527. doi:10.1111/avsc.12085

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*: Wiley.

Irvine, K. M., Rodhouse, T. J., Wright, W. J., & Olsen, A. R. (2018). Occupancy modeling species-environment relationships with non-ignorable survey designs. *Ecological Applications, 28*(6), 1616-1625. doi:10.1002/eap.1754
kartverk, S. (2017a). FKB-Bygning. In (pp. available online: https://kartkatalog.geonorge.no/metadata/8b4304ea-4304fb4300-4479c-a4324d-fa4225e4302c4306e4397). Geonorge kartkatalog.
kartverk, S. (2017b). FKB-Veg. In (pp. available online: https://kartkatalog.geonorge.no/metadata/4920b4452-4975cc-4945f4922-4964c-3378204c3373517). Geonorge kartkatalog.
kartverk, S. (2018). DTM 10 Terrengmodell. In (pp. available online: https://kartkatalog.geonorge.no/metadata/19cf1687-1684ed1686-1645ec-1689f1685b-fae1613a1661e1671b). Geonorge kartkatalog.
Lohr, S. L. (2019). *Sampling: Design and Analysis: Design and Analysis*: Chapman and Hall/CRC.
Mazzoni, S., Halvorsen, R., & Bakkestuen, V. (2015). MIAT: Modular R-wrappers for flexible implementation of MaxEnt distribution modelling. *Ecological Informatics, 30*, 215-221. doi:https://doi.org/10.1016/j.ecoinf.2015.07.001
Meng, X. (2013). *Scalable simple random sampling and stratified sampling*. Paper presented at the Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, Atlanta, GA, USA.
Moen, A. (1998). *Vegetasjon*. [Hønefoss]: Norges geografiske oppmåling.
Mossberg, B., Båtvik, S. T., Stenberg, L., & Moen, S. (2010). *Gyldendals nordiske feltflora*. [Oslo]: Gyldendal.
NGU. (2016). Berggrunn N250. Norges geologiske undersøkelser. In (pp. available online: https://kartkatalog.geonorge.no/metadata/7c39be66-77b36-34b74-b58d-53b36bee90067). Geonorge kartkatalog.
NGU. (2017). Løsmasser. Norges geologiske undersøkelser. In (pp. available online: https://kartkatalog.geonorge.no/metadata/3de4ddf6-d6b8-4398-8222-f4395c47791a47757). Geonorge kartkatalog.
Nordhausen, K. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition by Trevor Hastie, Robert Tibshirani, Jerome Friedman. *International Statistical Review, 77*(3), 482-482. doi:10.1111/j.1751-5823.2009.00095_18.x
Nowell, M. (2017). Sentinel 2 land cover Oslo Akershus 2017. In (pp. available online: http://urban.nina.no/layers/geonode%3As2_lc_oaf_08_2017_2010m_25833.). URBAN EEA geonode.
Olsen, A. R., Sedransk, J., Edwards, D., Gotway, C. A., Liggett, W., Rathbun, S., . . . Young, L. J. (1999). Statistical issues for monitoring ecological and natural resources in the United States. *Environmental Monitoring and Assessment, 54*(1), 1-45. Retrieved from <Go to ISI>://WOS:000078748100001
Ovaskainen, O., & Soininen, J. (2011). Making more out of sparse data: hierarchical modeling of species communities. *Ecology, 92*(2), 289-295. doi:10.1890/10-1251.1
Pedersen, Å. Ø., Nyhuus, S., Blindheim, T., & Krog, O. M. W. (2004). Implementation of a GIS-based management tool for conservation of biodiversity within the municipality of Oslo, Norway. *Landscape and Urban Planning, 68*(4), 429-438. doi:10.1016/S0169-2046(03)00148-8
Skarpaas, i. p. Poisson regression model
Skarpaas O, H. E., Framstad E & H R. (2019). When to switch from simple random to probability-based sampling in mapping and monitoring of rare habitats and species? doi:10.31219/osf.io/mgkwn

Smith, T. M. F. (1983). On the Validity of Inferences from Non-Random Samples. *Journal of the Royal Statistical Society Series a-Statistics in Society, 146*, 394-403. Retrieved from <Go to ISI>://WOS:A1983SA76900004
SNL. (2019). https://snl.no/. Retrieved from https://snl.no/
Støa, B., Halvorsen, R., Mazzoni, S., & Gusarov, V. (2018). Sampling bias in presence-only data used for species distribution modelling: theory and methods for detecting sample bias and its effects on models. *Sommerfeltia, 38*, 1-53. doi:10.2478/som-2018-0001
Team, Q. D. (2009). QGIS Geographical Information System. Retrieved from http://qgis.osgeo.org
team, R. c. (2019a). prop.test. Retrieved from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prop.test
Team, R. C. (2019b). R: A Language and environment for statistical computing. Retrieved from http://www.R-project.org
vegvesen, S. (2017). Trafikkmengde. In (pp. available online: https://kartkatalog.geonorge.no/metadata/af2c4a0a-1978-1974e1962-b1908d-ed1971f1936bd5023). Geonorge kartkatalog.
victoria2. (2017). ESRI Satellite (ArcGIS/World_Imagery). Retrieved from http://server.arcgisonline.com/ArcGIS/rest/services/World_Imagery/MapServer/tile/{z}/{y}/{x}
Vollering, J. (2019, 30.05.19). A modeling example. Retrieved from https://cran.r-project.org/web/packages/MIAmaxent/vignettes/a-modeling-example.html#references
Vollering, J., Halvorsen, R., & Mazzoni, S. (2019). The MIAmaxent R package: Variable transformation and model selection for species distribution models. *Ecology and Evolution, 9*(21), 12051-12068. doi:10.1002/ece3.5654
W, J. (2009). Geological Guide to Oslo and District. Edited by O. Holtedahl and J. A. Dons. 2nd. Edition. 118 pp., 42 figures, separate geological map, 1: 50,000. (Scandinavian University Books). Universitetsforlaget, Oslo, 1966. Price N.Kr. 25.00. *Geological Magazine*, *104*, 206. doi:10.1017/S0016756800040899
White, S. E. (2020). Probability & Related Topics for Making Inferences About Data. In *Basic & Clinical Biostatistics, 5e*. New York, NY: McGraw-Hill Education.
Whittaker, R. H., & Peet, R. K. (1985). *Plant community ecology: papers in honor of Robert H. Whittaker*: W. Junk.
Wollan, A. K. (2011). *Åpen grunnlendt kalkmark i Oslofjordområdet - et hotspot-habitat. Sluttrapport under ARKO-prosjektets periode II*. Retrieved from
Yoccoz, N. G., Nichols, J. D., & Boulinier, T. (2001). Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution, 16*(8), 446-453. doi:Doi 10.1016/S0169-5347(01)02205-4

Appendices

Appendix 1

Re-registration of The NA- and absence observation units

Some of the plots were not surveyed during the first field trip. There were several reasons for this. For instance, some were placed on water, and to minimize the problem of these (at first) NA-plots, a new field trip was planned in the summer of 2019. The decision was made to move them to nearest landmass, under the assumption that the plots on the water were placed there because of the probability-based distribution.

Furthermore, some plots were placed on steep cliffs, and as such were unavailable through normal means. These plots were therefore treated as absence points. There were also some plots that were not directly unavailable, but were still not accessible because they were placed within private grounds. In addition to the above-mentioned plots that were unavailable, there were several plots were placed on islands or islets that were only accessible by boat. Lastly, Fragaria viridis, one of the targeted species and can be confused with a close relative, *Fragaria vesca*, as they are quite similar with regards to the vegetative parts. Since all the targeted species in the analysis are treated as one RV, then the case of *viridis* and *vesca* being almost identical could be a source of error.

Four reasons in total that some of the plots were not visited:

- Some plots were placed on water
- Some plots were placed along steep cliffs
- The area in which the plot were placed was within private property
- Some plots were placed on unavailable islands

Plots on water

To minimize the problem of NA-plots, a new field trip was planned in the summer of 2019. The decision was made to move them to nearest landmass, under the assumption that the plots on the water were placed there because of the probability-based distribution.

The reason for the misplacements of these plots is still currently unknown. It might have something to do with the transfer of coordinates between different devices. It should be noted that the placements of the plots both within the qgis-software and the handheld gps were the same.

The observation units that were originally on water, but due to being moved were re-registered as presence: 2, 31, 39, 102, 166, 189 and 193.

The observation units that was re-registered to absence that was originally on water: 7, 45, 90, 111, 117, 125, 147 and 194.

Plots on or along cliffs

The steepness was so extreme that they were unavailable through normal means. It was then decided to write these as NA These plots were primarily in the forest area around the main city area. They were then treated as presence points, as there were every indication that none of the targeted species would be present if visited. The plots that were re-registered from NA to absence using this argument: 16, 89 and 122.

Plots on private property

Some of the plots were unavailable due to being placed within private property.The plot 171 was at first registered as NA, because it was mostly within a private lawn. However, a small part was available because of a nearby road. Both *bergskrinneblom* and *viridis/vesca* were present on that part of the plot. This plot was therefore treated as a presence plot, although the abundance of each species was not decided for this plot.

Plot 86 was also unavailable for the same reason as 171, but one of its buffer zone were placed on an open field in between several houses. A lot of the targeted species were present in this particular buffer zone, and as such the decision to move the plot to that spot were made.

The plots that are still registered as NA: 6, (50), 71, 81, 112, 120, 145, 177

Plots on islands or islets

Several plots were placed on islands or islets that were only accessible by boat. Only two plots that were on an island was surveyed. The rest of the plots in this category were on islands or islets that were so remote, that a bigger boat with all the appropriate equipment was necessary to make the trip. These were registered as NA. The plots in question are: 7, 36, 160, 173 and 188.

Fragaria viridis and Fragaria vesca

Fragaria viridis, one of the targeted species and can be confused with a close relative, *Fragaria vesca*, as they are quite similar with regards to the vegetative parts. The main difference is in the fruit/berry. *Fragaria vesca* is quite common in Norway and is associated with rich soil.

Since all the targeted species in the analysis are treated as one RV, then the case of *viridis* and *vesca* being almost identical could be a source of error. There were a total of 5 presence points that inly had Fragaria viridis as observation. For all the registration of *Fragaria viridis* only the vegetative parts of the plant were present, or more importantly the fruit were absent. Because of this, it is not 100 % certain that these observation units should be regarded as present points.

Several of these 5 present points had observation of some of the other targeted species in the buffer zone. But the main reason for labeling these plots as presence points were their surroundings, as all either had presence points close to them or were placed on areas that was considered to be high probability regions.

Specifically, the 5 plots are:

- Plot 4. This is one of the plots placed on Bygdøy, which can be considered a high probability region. There were several individuals of *Fragaria* viridis/Fragaria *vesca* present.
- Plot 20. This is also one of the Bygdøy plots. There were no species other than *Fragaria viridis/vesca* that were registered in the plot itself. But both *Geranium sanguineum* and *Conoeaster intergerrimus* were present in the buffer zone.
- Plot 79. Both *Geranium sanguineum* and *Filipendula vulgaris* were present in the buffer zone.
- Plot 85. There were registered no observations in the buffer zone. Mainly registered as a present point because of the location, as it was placed on a high probability region (*Lindøya*).
- Plot 95. Both *Filipendula vulgaris* and *Veronica spicata* were registered in the buffer zone.

Fragaria viridis vs Fragaria vesca

Assuming that observation of vegetative parts of either *Fragaria viridis* or *Fragaria vesca* are all observation of *Fragaria viridis* in certain areas where the environmental conditions "favors" *Fragaria viridis* could lead to false conclusion about the investigated plot. Especially when one consider the possibility of an ecological condition that causes the observed absence of the investigated species (beside *Fragaria viridis*/*Fragaria vesca*).

By stating that these particular observation units were presence points, their values concerning the environmental variables could potentially skew the forward selection process on environmental variables. This consequence gives emphasis on variables that are in fact irrelevant to the targeted species.

Appendix 2

The distribution maps for the rest of the tested environmental variables are shown here.

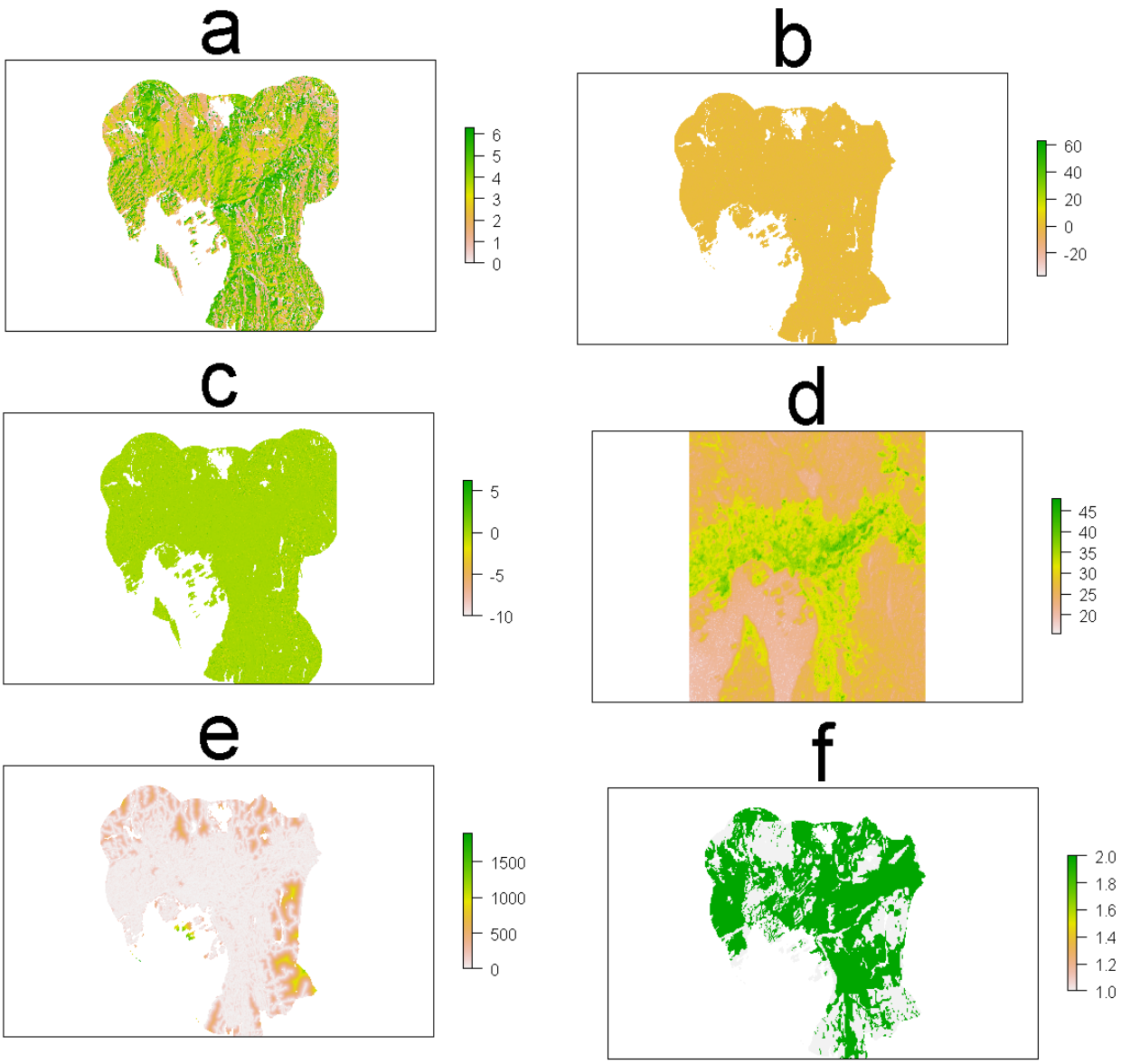


Figure A.1: *a)*Distribution of aspect variable. *b)*Distribution of curvature variable. *c)*Distribution of TPI variable. *d)* Distribution of surface temperature variable. *e)*Distribution of road distance variable. *f)* Distribution of substrate variable.

Appendix 3

A more detailed overview of the investigated species are shown here.

Table A.1: Table showing the name of the targeted species and other relevant information about the individual species (Mossberg, Båtvik, Stenberg, & Moen, 2010; SNL, 2019).

Latin name	Length of life cycle and type	Size	Reproductive	Vegetative	Comments
Gernaium sanguineum	Perennial	15 – 60 cm tall.	The flowers are radial and bright purple, making the plant quite noticeable when flowering.	The leaves are lobed and narrow.	Arguably the least rare species among the investigated species, and can be found both in open terrain and near forest.
Filipendula vulgaris	This species is herbaceous and perennial.	They can grow up to 50 cm.	The flowers are numerous, white with a purple hue underneath.	The leaves are lobed and are arranged in a rosette near the ground.	
Arabis hirsute	An herbaceous plant that is both biennial and perennial.	They can be between 20 and 60 cm tall.		The plant has whole, narrow leaves both near the ground in the form of a rosette, and along the stalk.	They are typically found on bare rocks, hence the Norwegian name (berg=rock).
Poa compressa	Perennial grass.	Can be 10 to 40 cm tall.		The straw is blue green with narrow leaves that are flattened.	They are often found on dry, ragged places.
Thymus pulegiodes	A small subshrub.	Can be 5 to 25 cm tall.	The leaves are small, and quite aromatic.	Flowers are pink.	
Fragaria viridis	A perennial herb		Flowers are round and white. Fruit similar to *vesca*, only "wraps around" more.	Leaves similar to *vesca*, but the tip are more blunted.	Can be confused with *Fragaria vesca(Markjordbær)*
Veronica spicata	Perennial herb	Can be 5 to 40 cm tall.	They have many small purple flowers placed quite densely on a long spike		This species is considered endangered in Norway, but mainly because of the nature type (T2-C8) is rare and declining in prevalence.
Inula salicina	Perennial herb	Can be 20 to 70 cm tall.	Two to four cm wide, yellow flowers on the top.	The leaves are placed quite densely along the stem.	
Cynoglossum officinale	Biennial herb	They can be 30 to 80 cm tall.		Have leaves formed like tongues.	The leaves have hair glands that produces a foul-smelling odour.
Hypochaeris maculata		They can be 20 to 60 cm tall.	The flowers are yellow and forming a capitula.	The leaves are arranged in a rosette, with big black spots.	
Erysimum strictum	Biennial herb	Can be 30-90 cm tall.	Bright yellow flowers. Silique fruits.	Lancelet shaped leaves.	Mainly found on dry places.
Centaurea scabiosa	Perennial herb	Can be 30-100 cm tall.	Sphere shaped capitula.	Lobed leaves.	Found on dry places with high calcium-content.
Silene nutans	A perennial herb	They can be 20 to 40 cm tall.	The flowers are white, with the cap leaning towards the side.	The leaves are arranged in a rosette near the bottom of the plant.	
Cotoneaster intergerrimus.	Shrub	can grow up to 2 meters tall	The flowers are small and bright red.	The leaves are oblong shaped, the inferior side covered in white hairs.	The second largest one of the targeted species, and are relatively common in forests.
Ligustrum vulgare	Deciduos shrub	Up to 3 meters	Flowers are white and arranged in panicles that are 3-6 cm long.	Leaves are opposite in pairs along the stem, whit a shiny green coloration and an oval shape	The largest one of the targeted species. The stem is brown-grey and stiff.
Phleum phleoides	Perennial herb	Up to 60 cm tall	The axis is quite dense and "splits apart" when bent.	The straw has a violet hue.	Considered endangered in Norway
Carlina vulgaris	Herb	Can grow between 10 to 60 cm.	The flowers (or the crown) are yellow woth a red tint, while the outer support		

			structures are pointing outwards.		
Draba verna	Annual herb	Can be 1 to 10 cm tall.	White flowers		Flowers early in the spring, and whiter shortly afterwards.
Echium vulgare	Biennial herb	Can be 20 to 80 cm tall	Bright blue flowers, and the stamens have a red-pink coloration.		
Poa alpine	Perennial herb	Can be 15 to 40 cm tall	The axis are broad and have a red-violet hue.	Dark green leaves.	They can reproduce sexually (with flowers), or with gemmae.
Saxifraga granulate	Perennial herb	Up to 40 cm.	Many white flowers, arranged in tassels.	The leaves are near the bottom of the stem, in the form of a rosette.	
Carex caryophyllea	Herb	Can be 10 to 30 cm tall	Male axis at the top with two female axis right below.	Stiff leaves.	
Dracocephalum ruyschiana	Perennial herb	Can be 5 to 15 cm tall	The flowers are bisymmetrical and around 3 cm long. The color can differ between dark blue, blue-violet and deep blue.	Leaves are lancet shaped and are placed in paired opposite of each other along the stem.	
Saxifraga osloensis	Annual herb		The flowers are white and are 6 to 10 mm wide.	The stem is erect and the leaves consist of three leaflets, where the middle one is the largest.	The stem have glandular hair
Rosa majalis	Shrub	Grow up to 1 meter tall.	The flowers are relatively small with pink coloration.	The bark is red-brown.	
Saxifraga tridactylites	Annual herb	Can be 3 to 15 cm tall.		Fleshy leaves, with three distinct leaflets. Stem is red.	The whole plant (beside the flowers) are covered in adherent glands.
Scleranthus perenni	Perennial herb	Can be 5 to 15 cm tall.	They have no pedals, but the sepals are white and function as pedals.	The whole plant has a blue green hue, while the leaves have ha pointy end.	They flowers early in the summer.
Lithospermum officinale	Herb	Can be up to 80 cm.	The flowers consist of large bract around green or white pedals.	The leaves have an oval shape, while the stem is erect and branching at the top.	
Linum catharticum	Annual herb	Can be 5 to 20 cm tall.	The flowers are white, with glands on the sepals.	The leaves are opposite with one nerve.	

Appendix 4

The FOP-plots for the non-selected variables are shown here, in addition to the FOP-plots for selected variables in the UPS+SRS models.

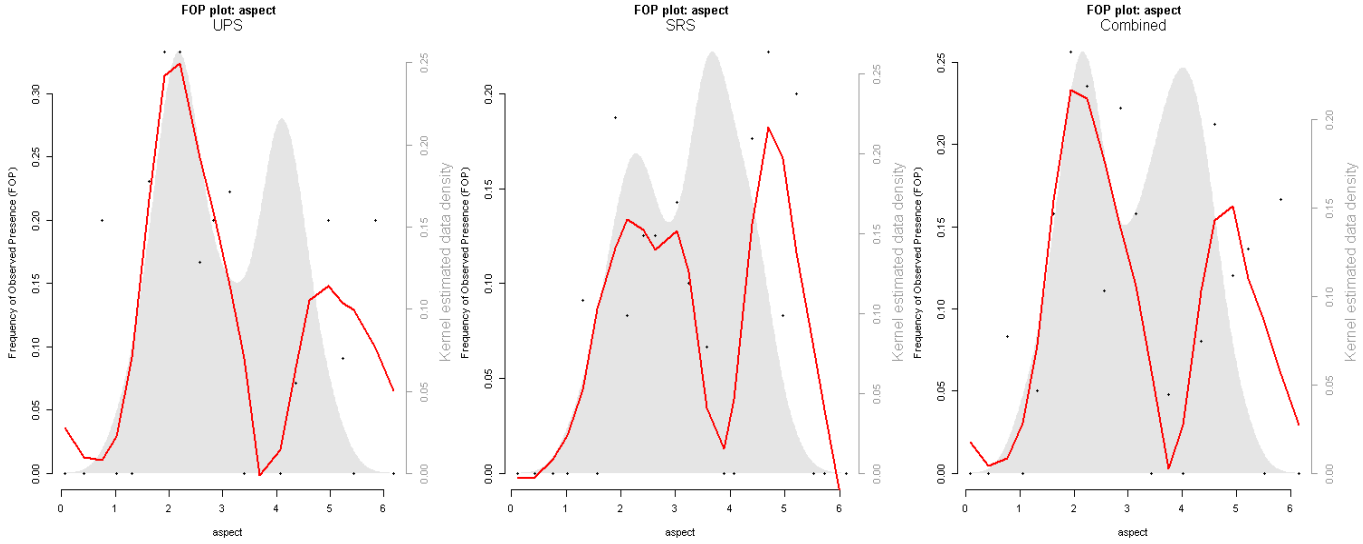


Figure A.2: The frequency of empirical presence for the apsect variable (for UPS, SRS and UPS+SRS, i.e combined). The black dots represents the proportion of points for their respective interval that contains presences. The red line represents a local regression line that gives a representation of the patterns in the frequency of presence for the investigated EV. The grey field represents an approximation of the data density along the EV range. The interval = 20 for all three FOP of aspect.

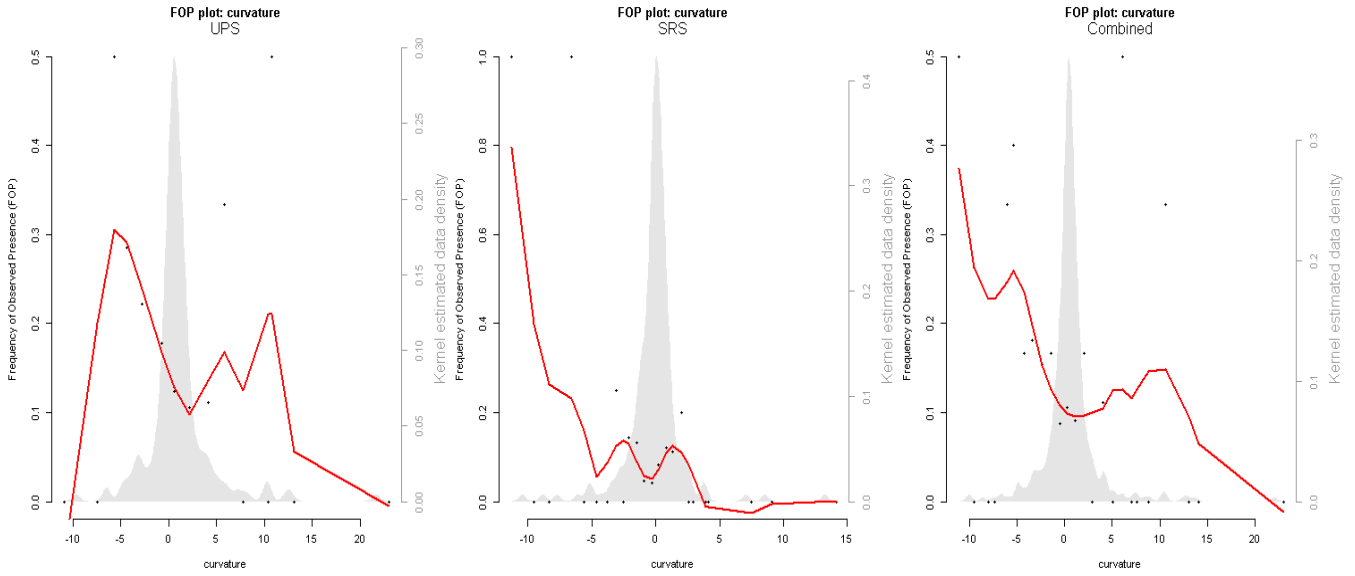


Figure A.3: The frequency of the empirical presence for the curvature variable. The interval equals 20 for all the FOP plots (UPS, SRS and UPS+SRS).

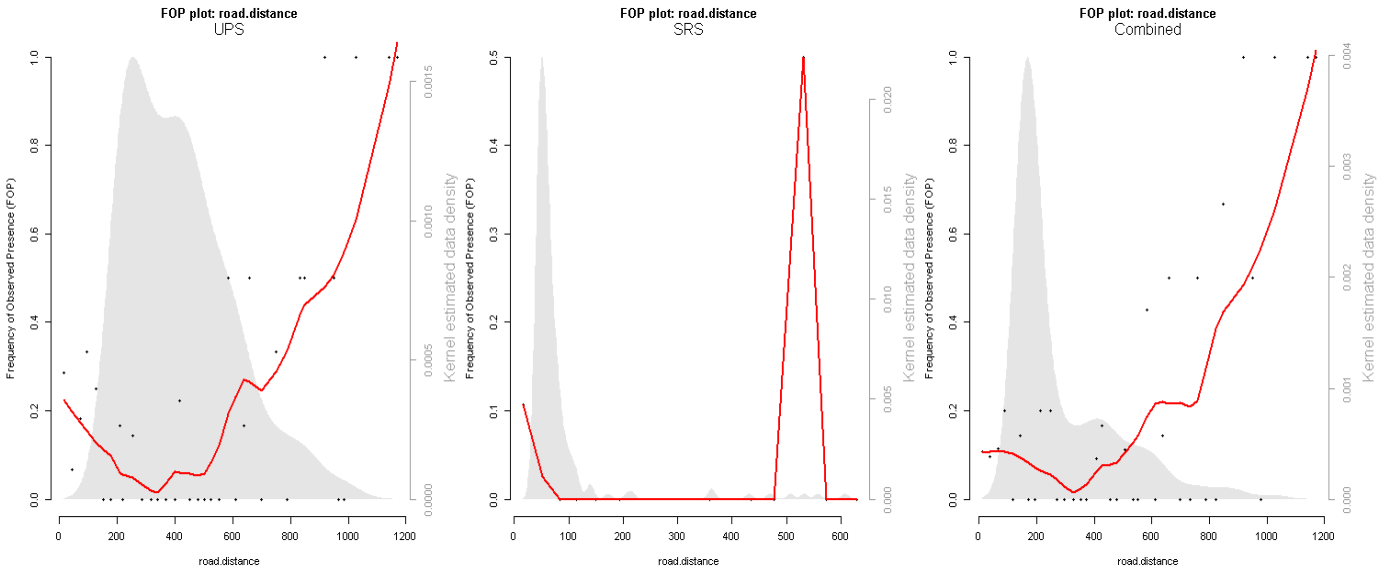


Figure A.4: The frequency of empirical presence for the road distance variable. Interval = 20 for all the data sets (UPS, SRS and UPS+SRS).

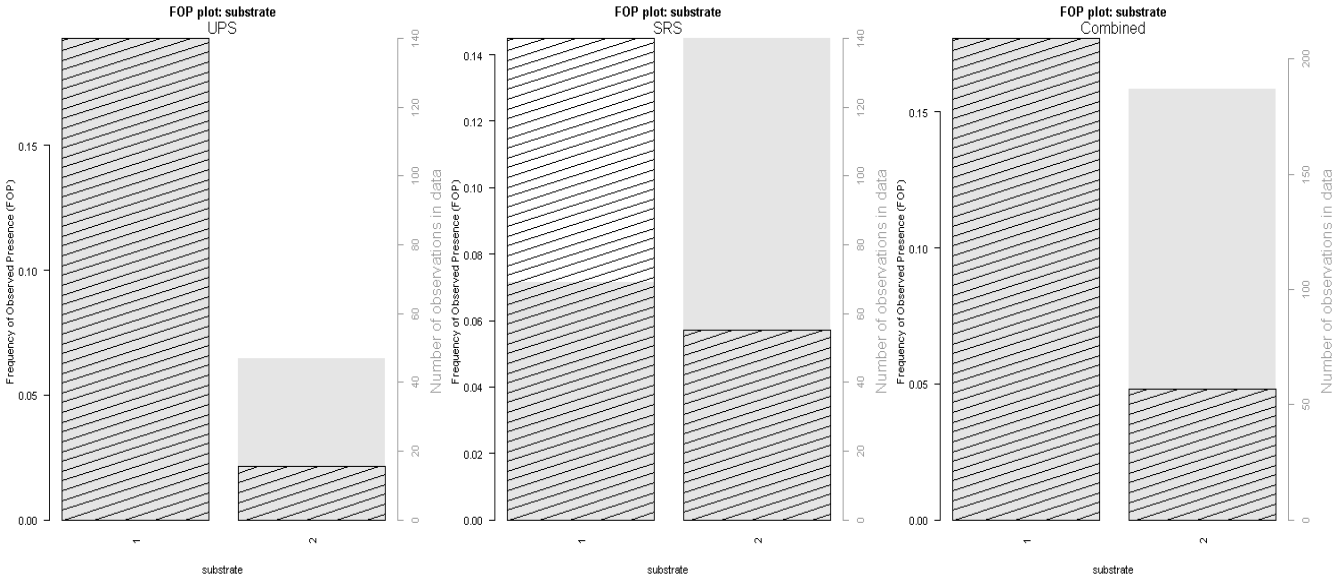


Figure A.5: The frequency of the empirical presence for the substrate variable. The striped bar represent the proportion of presences for its respective category. Similarly to the FOP for numerical variables, the gray represents the data density for its respective category.

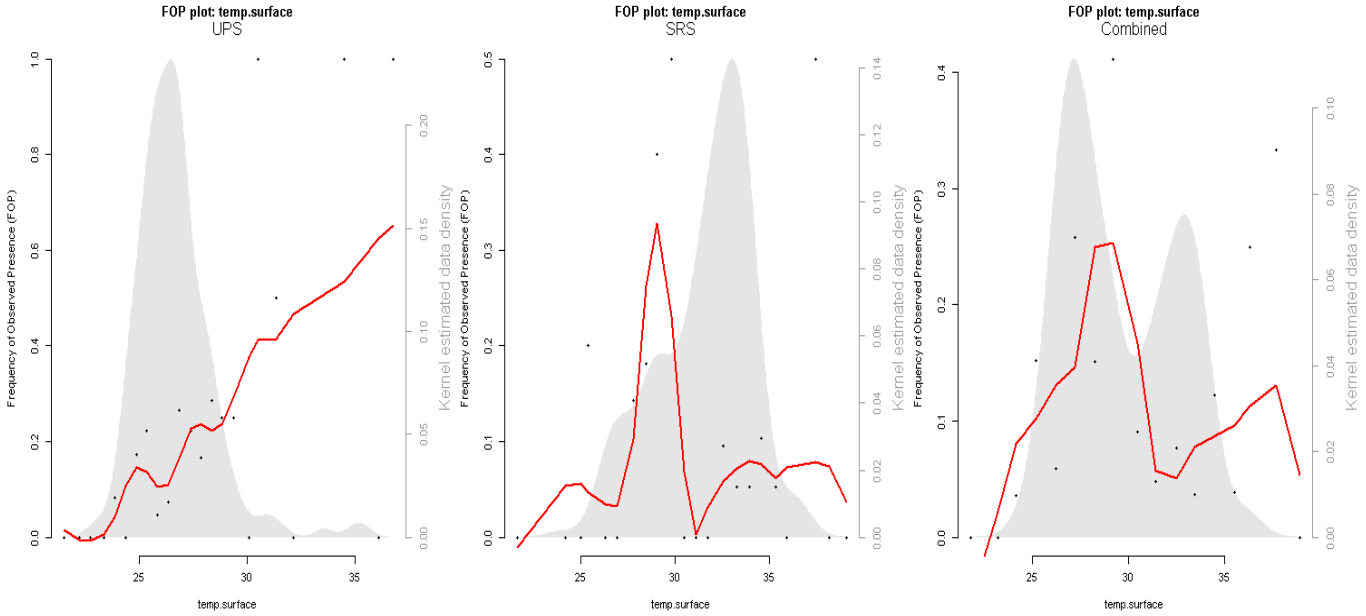


Figure A.6: The frequency of the empirical presence for the temperature surface variable. For the UPS FOP-plot the interval = 26. For the SRS FOP-plot the interval = 17. For the combined FOP-plot the interval = 17.

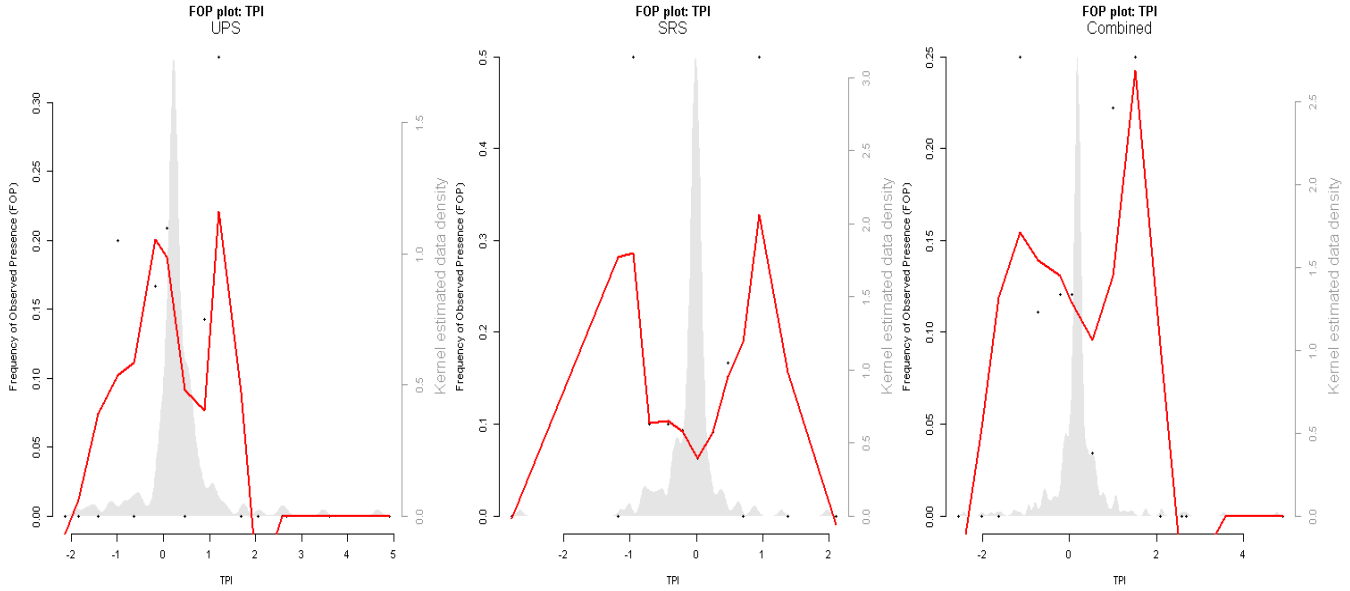


Figure A.7: The frequency of the empirical presence for the TPI variable. For the UPS FOP-plot the interval = 19. For the SRS FOP-plot the interval = 20. For the combined FOP-plot the interval = 17.

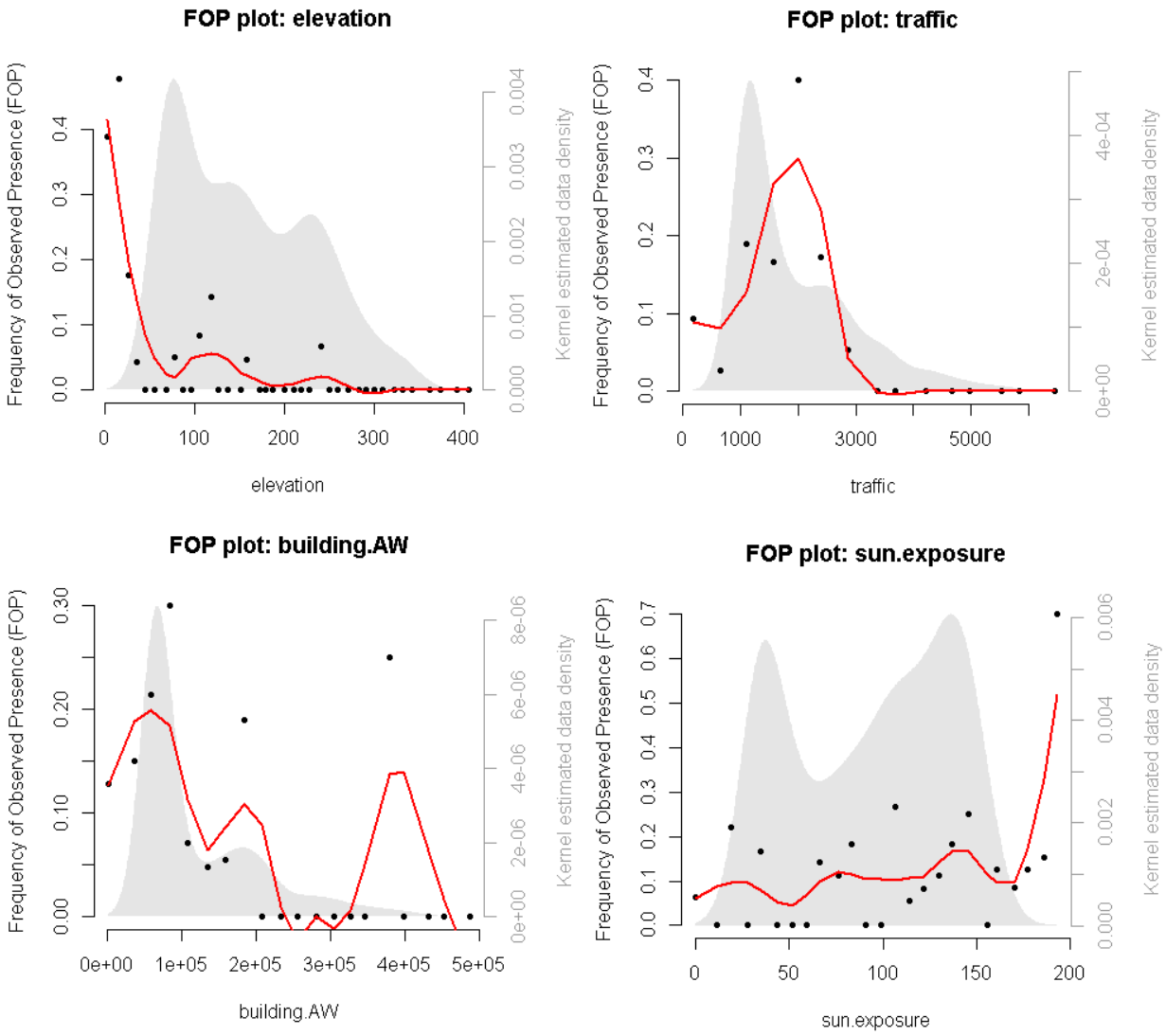


Figure A.8: The FOP-plots for elevation (int = 40), traffic (int = 15), building.AW (int = 20) and sun.exposure (int = 25) for the UPS+SRS data.

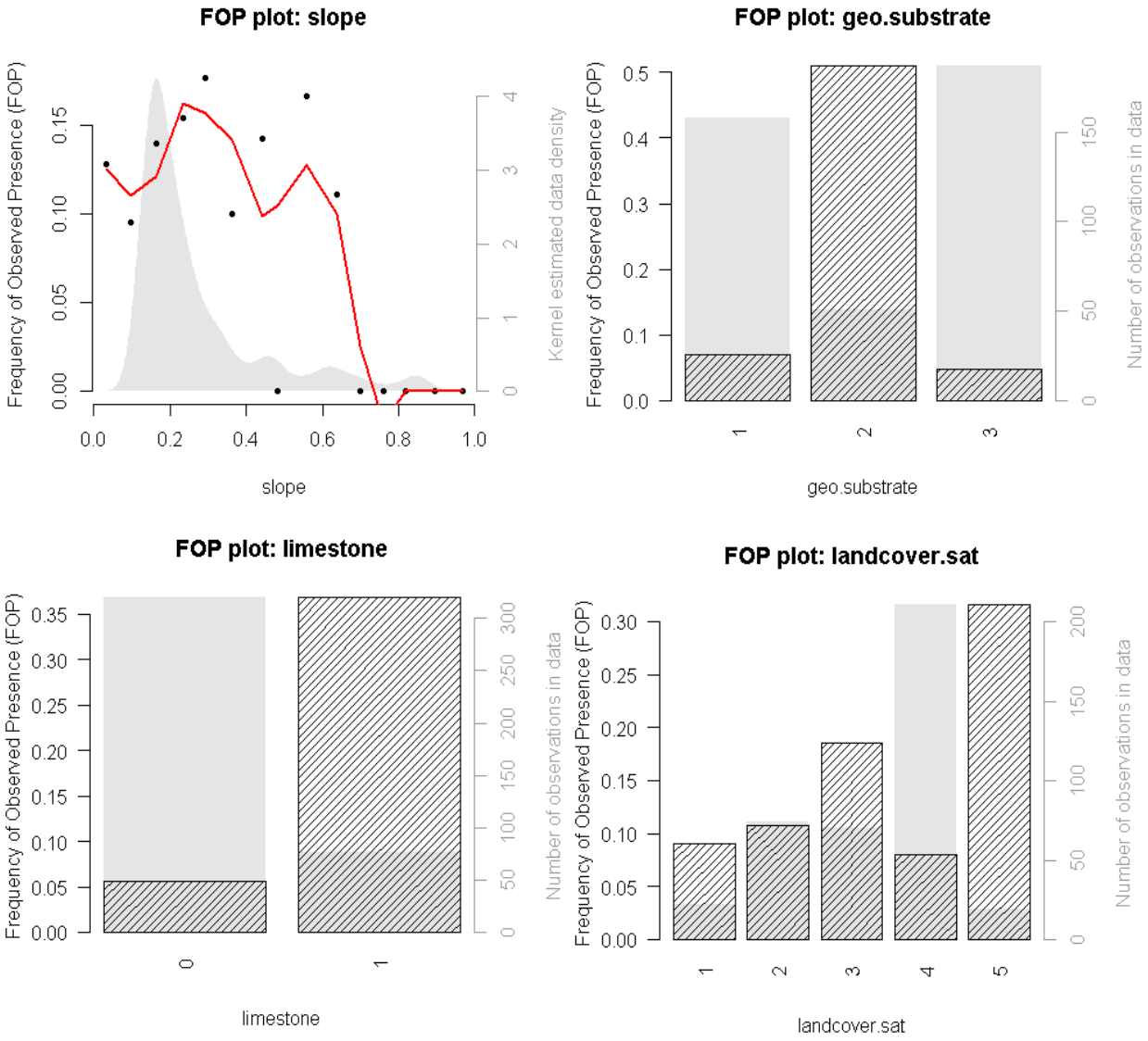


Figure A.9: The FOP-plots for slope (int = 15), geo.substrate , limestone and landcover.sat for the UPS+SRS data.

Appendix 5

A more detailed description of the models themselves are shown here. Specifically, the output from the *selectEV$selectedmodel* are shown for model 1-12 (see table 2.3 for the model discriptions).

Model 1:

```
Call:  stats::glm(formula = formula, family = "binomial", data = data)

Coefficients:
      (Intercept)      elevation_D05  geo.substrate_BX2
          -0.2901            -8.0307             2.5769

Degrees of Freedom: 186 Total (i.e. Null);  184 Residual
Null Deviance:      157.9
Residual Deviance: 43.98        AIC: 49.98
```

Model 2:

```
Call:  stats::glm(formula = formula, family = "binomial", data = data)

Coefficients:
      (Intercept)  geo.substrate_BX2        traffic_HF8          traffic_L
           -6.228              5.316            -31.502              9.386

Degrees of Freedom: 186 Total (i.e. Null);  183 Residual
Null Deviance:      157.9
Residual Deviance: 53.58        AIC: 61.58
```

Model 3:

```
Call:  stats::glm(formula = formula, family = "binomial", data = data)

Coefficients:
      (Intercept)      elevation_HR3  geo.substrate_BX1
           -3.516              1.682              1.321

Degrees of Freedom: 208 Total (i.e. Null);  206 Residual
Null Deviance:      122.7
Residual Deviance: 107.9        AIC: 113.9
```

Model 4:

```
Call:  stats::glm(formula = formula, family = "binomial", data = data)

Coefficients:
      (Intercept)     building.AW_T4  landcover.sat_BX5
           -2.008             -1.244              1.363

Degrees of Freedom: 208 Total (i.e. Null);  206 Residual
Null Deviance:      122.7
Residual Deviance: 110.4        AIC: 116.4
```

Model 5:

```
Call:  stats::glm(formula = formula, family = "binomial", data = data)

Coefficients:
              (Intercept)              elevation_HR4                 traffic_D05
                    1.841                      2.837                       -8.922
              traffic_HF9              limestone_BX0  traffic_D05:limestone_BX0
                -30474.605                     -5.065                       7.692
traffic_HF9:limestone_BX0
                       NA

Degrees of Freedom: 395 Total (i.e. Null);  390 Residual
Null Deviance:      284.5
Residual Deviance: 157.5        AIC: 169.5
```

Model 6:

```
Call:  stats::glm(formula = formula, family = "binomial", data = data)

Coefficients:
                    (Intercept)              geo.substrate_BX2
                          5.599                          7.013
                     traffic_D05                    traffic_HF9
                          1.574                     -33389.812
                 building.AW_D05              landcover.sat_BX5
                         -2.489                          1.831
               sun.exposure_D05                 sun.exposure_M
                         -8.396                         -6.185
                     slope_HF13  geo.substrate_BX2:traffic_D05
                        -21.504                         -9.635
geo.substrate_BX2:traffic_HF9
                             NA

Degrees of Freedom: 395 Total (i.e. Null);  386 Residual
Null Deviance:      284.5
Residual Deviance: 167.4        AIC: 187.4
```

Model 7:

```
Call:  stats::glm(formula = formula, family = "binomial", data = data)

Coefficients:
  (Intercept)  elevation_D05
        2.053        -11.208

Degrees of Freedom: 186 Total (i.e. Null);  185 Residual
Null Deviance:      157.9
Residual Deviance: 52.17        AIC: 56.17
```

Model 8:

```
Call:  stats::glm(formula = formula, family = "binomial", data = data)

Coefficients:
      (Intercept)  geo.substrate_BX2
           -4.311              5.266

Degrees of Freedom: 186 Total (i.e. Null);  185 Residual
Null Deviance:      157.9
Residual Deviance: 63.81        AIC: 67.81
```

Model 11:

```
Call:  stats::glm(formula = formula, family = "binomial", data = data)

Coefficients:
  (Intercept)  elevation_HR4    traffic_D05
       -1.290          4.040         -5.711

Degrees of Freedom: 395 Total (i.e. Null);  393 Residual
Null Deviance:      284.5
Residual Deviance: 174  AIC: 180
```

Model 12:

```
Call:  stats::glm(formula = formula, family = "binomial", data = data)

Coefficients:
                   (Intercept)              geo.substrate_BX2
                       -3.1994                         7.5234
                   traffic_D05  geo.substrate_BX2:traffic_D05
                        0.7697                       -10.6567

Degrees of Freedom: 395 Total (i.e. Null);  392 Residual
Null Deviance:      284.5
Residual Deviance: 190.8        AIC: 198.8
```

Appendix 6

Abundance data and coordinates for the UPS data set

Plot id, coordinates and abundances (for each targeted species) for all the presence points:

plot_id	P/A	POINT_X	POINT_Y	Ger sang	Filip vul	Arab hirs	Poa com	Thym pule
2	1	596574	6639955	1	0	0	0	0.2
4	1	594175	6642375	0	0	0	0	0
5	1	597755	6637595	0	1	0	0	0
14	1	597015	6640215	2	1	0.1	1	0
18	1	596415	6639285	1	1	0	0	0
20	1	594215	6643255	0	0	0	0	0
31	1	597988	6632248	1	0	0.1	0	0
39	1	596547	6639935	2	0	0	0	0
44	1	590385	6640815	0	0	0	0	0
51	1	596035	6640555	0	2	0.1	0	2
76	1	596525	6639915	5	1	0	0	0.4
79	1	593105	6642385	0	0	0	0	0
85	1	595835	6640535	0	0	0	0	0
86	1	594874	6640427	1	1	0	0	1
93	1	598015	6637455	3	0.3	0	1	0
95	1	595715	6640555	0	0	0	0	0
97	1	595945	6639715	3	2	0.1	1	1
102	1	591098	6639868	5	0.2	0	0	0
109	1	596115	6640395	0.1	0	0	0	0
114	1	594925	6640605	1	1	0	0	0
129	1	596045	6640815	0	2	0	0	0
148	1	597295	6640885	2	0	0	2	0
153	1	596705	6641335	1	0.2	0	0	0
165	1	596255	6639785	1	0	0	0	0
166	1	598213	6637313	2	0	0	0	0
171	1	593246	6645661	0	0	0.3	0	0
189	1	597656	6637432	0	1	0	0	1

plot_id	Frag vir	Ver spic	Inula sali	Cyno of	Hypo mac	Erys stric	Cent scab	Silen nut
2	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0
5	2	0.2	0	0	0	0	0	0
14	0	0.4	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0
20	1	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0
51	0	1	0	0	0	0	0.1	0
76	0	0.4	0	0	0	0.1	0	0
79	1	0	0	0	0	0	0	0
85	1	0	0	0	0	0	0	0
86	1	1	0	0	0	0	0	0
93	1	0	0	0	0	0	0	0
95	1	0	0	0	0	0	0	0
97	1	1	0	0	0	0	0	0
102	0	0	0	0	0	0	0.1	0
109	0	0.1	0	0	0	0	0	0
114	2	1	0	0	0	0	0	0
129	0	1	0	0	0	0	0	0
148	0	0.4	0	0	0	0	2	0
153	1	0	0	0	0	0	0	0
165	1	0	0	0	0	0	0	0
166	0	0	0	0	0	0	0	0
171	1	0	0	0	0	0	0	0
189	0	0	0	0	0	0	0	0

plot_id	Coton inte	Ligus vul	Phl phle	Car vul	Draba ver	Echi vul	Poa alp	Saxi gran
2	0.1	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0.1	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0
18	0.1	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0
39	0	0	0	0	0	1	0	0
44	0	0	0	0	0	0.2	0	0
51	0.1	0	0.1	0	0	0	0	0
76	2	1	1	0.2	0	0	0	0
79	0	0	0	0	0	0	0	0
85	0	0	0	0	0	0	0	0
86	0.1	0	0	0	0	0	0	0
93	1	0	0	0	0	0	0	0
95	0	0	0	0	0	0	0	0
97	2	0	1	0	0	0	0	0
102	0	0	0	0	0	0	0	0
109	0	0	0	0	0	0	0	0
114	0	0	0	0	0	0	0	0
129	0	0	0	0	0	0	0	0
148	0	0	0	0	0	0	0	0
153	0.1	0	0	0	0	0	0	0
165	0.2	0	0	0	0	0	0	0
166	0	0	0	0.1	0	0	0	0
171	0	0	0	0	0	0	0	0
189	0	0	0	0	0	0	0	0

plot_id	Car caryo	Draco ruy	Saxi oslo	Rosa maja	Saxi trida	Scler per	Lithos of	fLinum catl
2	0	0	0	0	0	0	0	0.1
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	0
76	0	0	0	0	0	0	0	1
79	0	0	0	0	0	0	0	0
85	0	0	0	0	0	0	0	0
86	0	0	0	0	0	0	0	0
93	0	0	0	0	0	0	0	0
95	0	0	0	0	0	0	0	0
97	0	0	0	0	0	0	0	1
102	0	0	0	0	0	0	0	0
109	0	0	0	0	0	0	0	0
114	0	0	0	0	0	0	0	0
129	0	0	0	0	0	0	0	0
148	0	0	0	0	0	0	0	1
153	0	0	0	0	0	0	0	0
165	0	0	0	0	0	0	0	0
166	0	0	0	0	0	0	0	0.3
171	0	0	0	0	0	0	0	0
189	0	0	0	0	0	0	0	0.1

Plot id and coordinates for all the absence points:

plot_id	POINT_X	POINT_Y	plot_id	POINT_X	POINT_Y	plot_id	POINT_X	POINT_Y
0	603565	6648335	57	596805	6650915	111	596514	6639745
1	607805	6648645	58	606805	6633665	113	596745	6651115
3	606345	6642125	59	606055	6634335	115	588505	6650325
8	603245	6649995	60	594065	6651345	116	605795	6643175
9	604895	6643696	61	588655	6651365	117	594853	6642334
10	604535	6637125	62	591465	6653445	118	599465	6635265
11	605325	6633647	63	605755	6642835	119	597775	6630575
12	602865	6630265	64	594405	6649025	121	604775	6635955
13	590105	6651924	65	605555	6637705	122	605475	6641705
15	600485	6647905	66	589485	6652045	123	605715	6644375
16	600465	6647975	67	593215	6653605	124	604905	6637035
17	605465	6641435	68	604895	6640135	125	605355	6648875
19	604195	6649715	69	607465	6635885	126	605435	6641235
21	592215	6653725	70	602525	6652255	127	595285	6651265
22	605645	6633975	72	599095	6630975	128	595255	6648985
23	595735	6651895	73	606755	6635705	130	589085	6649465
24	603075	6648835	74	606615	6642905	131	603775	6638425
25	593897	6640952	75	603375	6636835	132	604425	6638405
26	605215	6643215	77	591615	6652745	133	604145	6633475
27	603805	6638805	78	598605	6640735	134	605285	6640975
28	591305	6651915	80	605265	6642305	135	605635	6639735
29	594855	6650485	82	606495	6635865	136	594865	6652145
30	594495	6649715	83	589385	6649165	137	605665	6639045
32	589625	6651845	84	605715	6633795	138	589845	6649095
33	592345	6652995	87	601455	6648075	139	592105	6653835
34	598445	6638405	88	597545	6649975	140	604515	6637895
35	605775	6634035	89	605485	6641595	141	606655	6639995
37	591195	6648635	90	593431	6643636	142	595085	6650675
38	593395	6653535	91	606425	6636985	143	603175	6652485
40	603685	6636085	92	602085	6650785	144	606305	6644505
41	604445	6640105	94	597985	6648055	146	604625	6636045
42	605115	6642795	96	606075	6637495	147	606625	6639556
43	593645	6644075	98	589475	6651845	149	592225	6651365
45	606246	6636774	99	602305	6650955	150	599245	6632745
46	606345	6636455	100	592325	6651015	151	602995	6648765
47	602505	6647035	101	602095	6633475	152	591215	6645805
48	606105	6634315	103	592685	6652525	154	598505	6631325
49	606055	6637275	104	600445	6630345	155	595385	6649425
52	604765	6642555	105	606045	6634945	156	601215	6647885
53	594115	6643115	106	599765	6634585	157	590715	6647645
54	600425	6630405	107	605805	6633935	158	598455	6631195
55	592375	6651155	108	607795	6634785	159	608335	6634435

plot_id	POINT_X	POINT_Y	plot_id	POINT_X	POINT_Y	plot_id	POINT_X	POINT_Y
162	606135	6637455	176	607905	6633865	187	603755	6651025
163	589555	6648455	178	606085	6635945	190	606595	6641405
164	606555	6635775	179	606145	6641785	191	605365	6639335
167	601905	6648525	180	606815	6650215	192	602445	6652005
168	599135	6632515	181	602285	6632045	194	601010	6649530
169	595685	6648765	182	603615	6637255	195	601935	6648635
170	606675	6638255	183	604785	6643465	196	603285	6649735
172	606955	6644985	184	604585	6639345	197	590885	6650845
174	606425	6638905	185	604785	6642745	198	601855	6632045
175	589525	6651895	186	602075	6649325	199	594275	6642705

Plot id and coordinates for all the NA points:

plot_id	POINT_X	POINT_Y	plot_id	POINT_X	POINT_Y
6	591026	6639360	120	593435	6642475
7	599103	6638788	145	593735	6642855
36	590825	6638425	160	598625	6638365
50	602955	6632675	173	598335	6638955
71	599115	6641765	177	597935	6650835
81	596565	6646525	188	592745	6642625
112	602065	6647335			

Appendix 7

A general R-script for making distribution models in MIAmaxent

```
install.packages("MIAmaxent")

library(MIAmaxent)

#First I did were to make a table showing the PA-points and their corresponding EV-value, using the
#readData()-command.

readData_table <- readData(

occurrence="C:/ occurrence_data_csv",

contEV="C:/folder_with_numeric_var",

catEV="C:/folder_with_categoric_var",

maxbkg=20000, PA=TRUE)

#The occurrence argument must be specified as a destination to a csv file with the occurence data. The data need
#to consist of three columns: First column should consist of PA (denoted 1 or 0), second and third should consist
#of the x and y coordinates respectively.

#Argument contEV and catEV should both lead to a target folder. The contEV folder should contain the
#numerical variables; while the catEV folder should contain the categorical variables. All variables should have
#the asci file type.

#The maxbkg arguments specifies the maximum number of grid cells randomly selected as unknown
#background points for the response variable. Irrelevant for PA data.

#PA argument affects how the occurrence data is interpreted. For this case (PA=TRUE), the 0 in the first row of
#the occurrence data are treated as absence points, 1 are treated as presence while NAs are excluded.


#The FOP-plots were obtained with the plotFOP-function:

FOP_variable <- plotFOP(readData_table, "variable_name", span=0.4, interval=20)

#readData_table (data frame that usually are obtained from the readData()function, but not necessarily) must
#contain the variable one wish to plot.

#"variable_name" must correspond to the name or column index of the variable in the "readData_table".

#The span-argument specifies the the neighborhood of the smoother.

#The interval-argument specifies the number of intervals one uses to calculate the FOP-plot.

#Note that the FOP-plot for categorical variables are a bit different. Neither span nor interval are relevant for
#categorical variables.

#Frequency of observed presence are in the case for presence/absence-data called frequency of empirical
#presence.
```

#The transformation were obtained with the deriveVars()-function:

transformed_data <- deriveVars(readData_table, algorithm = “LR”)

#See part 2.4.4 MIAmaxent.

#The algorithm specifies how the variables are transformed (the transformations

#are data-dependant). The algorithm = “LR” for PA-data (default is “maxent”).

transformed_data consists of two parts: data frames of the DVs for each EV (named “*dvdata”*), the #transformation functions used to produce each DV (named “transformations”)

#The selection of the most explanatory derived variables (DVs) were achieved with the *selectDVforEV*-function.

selectedDVs <- selectDVforEV (*transformed_data*$dvdata, alpha = 0.05,

algorithm = «LR», quiet = TRUE)

#The alpha argument sets the threshold for how much variation a DV must explain to be retained. The algorithm #argument must equal “LR” when using PA-data. The quiet argument, if true, suppresses the progress bar #(*henivse til hjelpesiden for selectDVforEV*). The *selectedDVs* object consist of two part: The dvs that were #selected for each EV (named “dvdata”), and the trails of nested models that were build and compared for each #EV during the selection process (named “selection”).

#The selection of the most explanatory environmental variables (EVs) were obtained with the *selectEV*-function:

selectedEVs <- selectEV(*selectedDVs*$dvdata, alpha = 0.05, interaction = TRUE, algorithm = “LR”)

#This function uses the selected DVs from the selectDVforEV-object. Alpha argument specifies the threshold for #how much the EVs (now consisting of DVs) need to explain for it to be retained. Interaction, if true, performs #testing of interaction terms between selected EVs. Only first –order interaction are tested. Algorithm specifies #the fitting algorithm. Algorithm = “LR” for the case where one uses PA-data. The *selectedEVs* object consist of #three parts: the EVs that were selected (“dvdata”), the trail of nested models that were build and compared #during the selection process (“selection”) and the selected full model under the specified alpha value #(“selectedmodel”).

#The response plots of the selected EV from each model were obtained with the plotResp()-function:

plotResp (*selectedEVs*$selectedmodel, *transformed_data*$transformations, “variable_name”)

#The *selectedEVs*$selectedmodel represents the final model. The *transformed_data*$transformations gives #access to the relevant transformed EV values. The “variable_name” specifies which selected variable one wish #to plot.

#The prediction distribution maps were obtained with the projectModel()-function:

projectModel(model = *selectedEV*$selectedmodel, transformations = *transformed_data*$transformations, data = *predictors*)

#The model argument specifies which model one wish to predict the distribution. For this case, only selected #models form the *selectEV* object were explored. The transformations argument specifies the transformation #functions. For the models shown here, I used only the transformations from the deriveVars object. The data #argument specifies the variable information needed to make the prediction. If the provided data is in the form of #a RasterStack or RasterBrick (for this study, the provided data were in the form of a raster stack), the #projectModel function automatically plots a predicted distribution map.

#To obtain the ROC curves and the AUC values for each model, I used the testAUC()-function:

testAUC(model = *selectedEV*$selectedmodel, transformations = *transformed_data*$transformations, data=*test_data*, plot=TRUE)

#The model argument specifies the model one wish to plot the ROC-curve for. In this case, I used the final #selected model from the *selectedEV* object. The transformation argument provides the relevant transformation #functions to create the derived variables. In this case, I used the transformations obtained with the deriveVars()-#function. The data argument specifies the test data, which should consist of PA in the first column and #corresponding EV values in the following columns (in the same form as a readData table in other words).