ACCOUNTING FOR SAMPLING BIAS IN SPECIES DISTRIBUTION MODELLING USING RAREFACTION

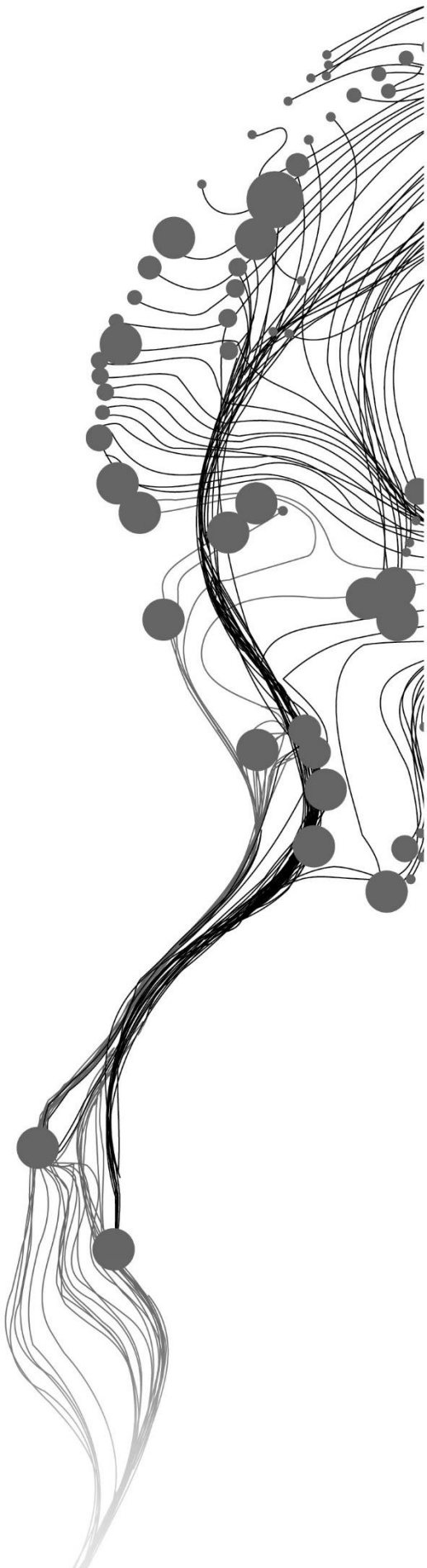
A case study of wild boar (*Sus scrofa*) in the Province Overijssel, The Netherlands

GICHANGI DOUGLAS MUTUOTA

July 2021

SUPERVISORS:

Dr.ir. T.A. Groen
Dr. P. Nyktas



ACCOUNTING FOR SAMPLING BIAS IN SPECIES DISTRIBUTION MODELLING USING RAREFACTION

A case study of wild boar (*Sus scrofa*) in the Province Overijssel, The Netherlands

GICHANGI DOUGLAS MUTUOTA

Enschede, The Netherlands,

July 2021

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-Information Science and Earth Observation

SUPERVISORS:

Dr.ir. T.A. Groen
Dr. P. Nyktas

THESIS ASSESSMENT BOARD:

Dr. T. Wang (Chair)
Dr. Nikolaos Fyllas (External examiner)

ABSTRACT

Prediction of the spatial distribution of species is vital for conservation planning. For accurate predictions, an appropriate sampling design should be used. Most ecological data that originate from unscientific sources are often biased towards the areas that are most accessible such as roads and nature parks. Besides, most methods of species distribution modelling (SDM) assume random and uniformly distributed samples. Thereby, spatially biased samples may lead to over or underprediction, therefore, making distribution models unreliable. Mostly presence/absence data are preferred to presence only because they contain more information about species' habitat. However, to collect presence/absence occurrence data, laborious field surveys and enormous resources are required, making it rare. However, plenty of presence-only (PO) observations exist in herbaria, museums, and online databanks, most of which are electronically accessible. In many instances, remedial treatment is required to make PO data reliable, for example, to correct sampling bias effects in data.

Ordinarily, clustering of data may lead to model's overprediction in areas that are intensively sampled. This effect can be mitigated by attempting to de-cluster the data, for example, rarefaction, or introducing randomly distributed background samples. The average nearest neighbour method was used to test two different wild boar observation datasets for spatial bias. Spatial rarefication was used to de-cluster presence-only data. Then a dataset with a similar number of observations (n) was selected from the original dataset. Five different methods of species distribution modelling (Boosted Regression Trees, Random Forests, Maximum Entropy, Support Vector Machine and Generalized Linear Models) were fitted with the two datasets and environmental predictors. The environmental predictors included 50m resolution Euclidean distance maps from the roads, nature reserves, heath &moor, farmlands, forest, and artificial surface. Randomization of models was undertaken by replicating the models twenty times for each method using bootstrapping. To check for consistency in the model predictions across methods was assessed by computing standard deviation in spatial prediction and comparing the zonal statistics for the various environmental variables.

It was found that the FBE dataset was more clustered than volunteer observations which are consistent with the way different observers are distributed. For all methods, the models from rarefied datasets were significantly different from the clustered ones. The machine learning method (RF, SVM, MaxEnt) tends to be more tolerant of survey bias because of clustering compared to empirical models (BRT, GLM). This was demonstrated through t-test statistic whereby the machine learning models were less significant compared with the rest. However, all the models performed well for both datasets, with a mean AUC above 0.8. Regarding variable importance to model permutations, the distance to nature reserves contributed most while distance to water was the least. However, the variability was more evenly distributed for the rarefied dataset compared to the clustered one. More so, all models depicted high prediction uncertainty in water areas while cultivated areas had the least. Therefore, bias correction demonstrated significant improvement to species distribution models' performance.

Keywords: Wild boar, species distribution modelling, biased, Random forest, boosted regression tree, support vector machine, generalized linear model, Maximum entropy, rarefaction, presence/absence, presence only, area under the curve, average nearest neighbour, random, clustered

ACKNOWLEDGEMENT

Firstly, I thank the Almighty God for always keeping me healthy during my study.

I'm grateful to the Kenya Electricity Generating Company Plc 9 (KenGen) and The University of Twente for sponsoring my study. At the time of this study (KenGen) permitted an off-duty and frequently checked to me to ensure I was comfortable.

My supervisor, Dr Ir. T.A. Groen & Dr P. Nyktas, offered unwavering support through my MSc Thesis phase from the inception of several research topics to make sure I have fantastic research. You helped me build ideas, facilitated data acquisition, and reviewed my work. I can't forget the interactive sessions we had during the internship through which I acquired the data for this research. I appreciate you to the chair of my MSc board Dr. T. Wang who always provided feedback and encouragement during my research proposal and review sessions.

More gratitude goes to Dr M.M. van den Berg (Maya), Mr Sjors Van der Graaf and the entire team I interacted with during the wild boar research project. To the team from NDFF and FBE who provided the wild boar observations data, I say thank you.

To Drs. R.G. Nijmeijer ensured that I was adequately informed on my academic expectations, checked on progress and welfare. More gratitude to the teaching team right from the core module to the specialization. The knowledge you imparted to me is immeasurable; I thank you all.

To my family, thank you for offering unwavering support, love and hope, the reason I always worked harder.

TABLE OF CONTENTS

1 INTRODUCTION ---------- 1

1.1 Background information ---------- 1
1.1.1 Sampling bias is species distribution modelling ---------- 1
1.1.2 Species distribution modelling ---------- 2
1.1.3 Human-wild boar conflicts ---------- 2

1.2 Problem statement ---------- 3

1.3 Research objectives, Questions & Hypotheses ---------- 4

2 MATERIALS & METHODOLOGY ---------- 7

2.1 Study area ---------- 7

2.2 Materials ---------- 8
2.2.1 Data ---------- 8

2.3 Methodology ---------- 9
2.3.1 Data pre-processing ---------- 10
2.3.2 Sampling bias analysis ---------- 12
2.3.3 Species distribution modelling ---------- 14
2.3.4 Assessing effects of sampling bias on species distribution models ---------- 17
2.3.5 Variable importance ---------- 17

3 RESULTS ---------- 18

3.1 To investigate variation in the distribution pattern of wild boar observations with varied survey bias -- 18

3.2 The effect of survey bias on the accuracy of species distribution modelling methods ---------- 19

3.3 Which environmental variables are relevant to predict the distribution of wild boars in Overijssel Province? ---------- 23

4 DISCUSSION ---------- 28

5 CONCLUSIONS ---------- 30

5.1 To investigate variation in the distribution pattern of wild boar observations with varied survey bias -- 30

5.2 The effect of survey bias on the accuracy of species distribution modelling ---------- 30

5.3 Which environmental variables are relevant to predict the distribution of wild boars in Overijssel Province? ---------- 30

6 RECOMMENDATIONS ---------- 31

LIST OF FIGURES

Figure 1: Study area 8
Figure 2: Overall flowchart 9
Figure 3: Flow chart demonstrating preparation of environmental variables 10
Figure 4: Flowchart demonstrating sample bias analysis 12
Figure 5: Comparative analysis of distribution patterns of datasets of different origins (Clustered-Blue, Random-Orange) 19
Figure 6: Comparison of model accuracy based on mean AUC 20
Figure 7: Wild boar distribution maps showing performance of various methods on clustered and rarefied datasets based on AUC metric 21
Figure 8: Performance of various models fitted with randomized and rarefied datasets based on AUC metric 22
Figure 9: Summary t statistic (two-sample assuming unequal variance) to test if there is a significant difference between SDMs' accuracy based on AUC fitted with the same method using clustered and rarefied datasets 22
Figure 10: Box plot showing variability in relative variable importance in the prediction of suitable habitats for wild boars based on AUC metric of clustered and rarefied datasets 23
Figure 11: gg-plot illustrating how different landcover predict the occurrence of wild boars based on: (a) randomly selected, (b) rarefied dataset for various environmental gradients based on Euclidean distance from LU/LC: Eudist_artificial (artificial), Eudist_cultivated (cultivated), Eudist_Heath_moor (Heath Moor), Eudist_reserve (reserve), Eudist_roads(roads), Eudist_water(water) 25
Figure 12: Wild boar's distribution maps (a) clustered & (b) rarefied and the respective zonal statistics plots (c) &(d) showing areas of model uncertainties depicted by the standard deviation 27

LIST OF TABLES

Table 1: Predictor variables significance and their respective hypotheses responses 5
Table 2: Data 9
Table 3: Landcover legend at the extent of Europe and Overijssel 11
Table 4: Summary statistics of data thinning and normality test (study area=3420840181 m²) 18

1 INTRODUCTION

1.1 Background information

1.1.1 Sampling bias is species distribution modelling

Prediction of the spatial distribution of species is vital for conservation planning (Elith et al., 2006). Moreover, a suitable sampling regime is essential for making accurate environmental predictions (Rocha et al., 2020). Phillips *et al.* (2009) defined sampling biased data as species occurrence localities that select ecological variables in a manner that is not proportional to the study area. Ecological samples are not primarily independent and identically distributed over study areas (Rocha et al., 2020). The places that are most accessible and most frequented often get intensely surveyed since observation of species is based on the chance that it is present and the locality is visited by an observer (Fernández & Nakamura, 2015).

Most species distribution modelling methods are designed to assume that the sampling effort was random in the study area (Komori et al., 2020). The accuracy of environmental models is dependent on the quality and quantity of data, for example, assessing the spatial distribution of species observations (Rocha et al., 2020). While partial sampling may fail to cover the habitat variability for species with a wide range, biased sampling efforts may overpredict the model towards the areas of high survey intensity (Elith et al., 2006). Therefore, the risk of transmitting the bias from species observations to the spatial distribution predictions need to be avoided (Phillips et al., 2009).

The performance of species distribution prediction can be improved through the effective remedy of survey bias in occurrence data and the critical selection of environmental predictors (Phillips et al., 2009). Elith et al. (2006) suggested that improving the quality of the training sample may enhance the model's performance. Previous studies proposed methods to reduce sampling bias such as 'mask layer' (Fernández & Nakamura 2015) ', quasi-linear Poisson point process', (Komori *et al.,* 2020) and 'background samples bias file' (Phillips *et al.,* 2009). The rationale of the mask layer is to create a subset barrier by increasing the cost for the geographical areas that the target species does not occur. On the contrary, the bias layer is a kernel density map delineating the areas where the species is likely to occur, which is used to constrain the selection of background points to the areas with high habitat suitability. For the Quasi-linear-Poisson point process method, it is assumed that sampling bias is high in the locations where the species abundance is high. Thus, the effect of sampling bias and environmental gradients are empirically modelled for separability. Brown (2014) developed a spatial rarefy toolbox that removes multiple spatially autocorrelated observations within a defined grid cell. The method was selected for use in this study because it can be applied to different species distribution modelling methods.

1.1.2 Species distribution modelling

Understanding the population distribution patterns of species is crucial for designing an ecological management program (Saunders & Kay, 1991). Species distribution modelling (SDM) constitute essential biodiversity variables (EBVs. These are measurements necessary for managing, assessing, and reporting variation in biodiversity (Pereira *et al.,* 2013). The focus of EBV is to take repeated measures of the same species then analyse them to make simplified indicators of change (e.g., species distribution models). Species distribution models predict the spatial spread of suitable habitats by establishing the empirical relationship between the occurrence or density of biodiversity and the environmental gradients (Elith et al., 2006).

Unlike the environmental variables, which have recently become more available due to advancements in remote data acquisition, reliable biodiversity observations are often expensive to acquire (Nakashima et al., 2018). However, a wealth of data, mainly presence only (PO), is archived in museums, herbaria, and online databases (Elith et al., 2006). Also, access to these data has been facilitated by online electronic transfers capabilities. Some methods of SDM (e.g. Generalized linear models and Boosted regression trees) use presence/absence (PA) data that's often collected from systematically designed field surveys, which is costly (Moeller et al., 2018). However, many methods have been developed that utilize PO data, which is the most available species occurrence secondary data (Philips et al., 2006, Cutler *et al.* 2007, Elith, Leathwick & Hastie 2008, Bruzzone & Persello 2009, Field 2011). Besides, the presence-only data do not follow systematic data collection protocols, and their acquisition intentions are unknown, making them prone to sampling biases (Komori et al., 2020). Moreover, it is challenging to infer absence from the PO data (Elith et al., 2006). Besides, Elith et al. (2006) indicated that different SDM methods are necessary to improve prediction accuracy in areas with limited data.

1.1.3 Human-wild boar conflicts

Human-wildlife conflicts have existed for many decades since the civilization of humankind (Messmer, 2000). The modification of natural environments by humans altered the ecological balance leading to competition for resources. Furthermore, the ever-increasing human populations have further exerted pressure on natural ecosystems due to the demand for food and raw materials. As a result, there is competing demand for conservation and farmland land use. Therefore, the conflict between humans and wildlife is more pronounced in the interface between farmlands and the marginal animal reserves where animals roam searching for food and water (Franz, Markus, Peter, 2020).

Consequently, there are losses, such as agriculture damage, diseases transmission, car-animal collisions, and animals attacks (Messmer, 2000). Unfortunately, many countries lack inventory of the magnitude and frequency

of damage caused by wildlife (Messmer, 2000). Understanding target species ecology is necessary for informing how to select and improve occurrence and environmental data and interpretation of model predictions (Elith et al., 2006). Wild boar are some of the most abundant mammalian species in terms of distribution within Europe (Franz, Markus, Peter, 2020). In the habitats that are close to agriculture fields, the species have various impacts on their natural ecosystem and agriculture (Hegel & Marini, 2013). They often move between the forest where they shelter and pasture land forage, often resulting in human-wildlife conflict (Saunders & Kay, 1991). Moreover, they are vectors of African Swine Fever (ASF) vectors, affecting them and domestic pigs (Blome et al., 2013). Similarly, Overijssel Province in the Netherlands hosts diverse ecosystems rich in biodiversity, including an unknown number of wild boars (*Sus scrofa).* The boars roam between a nature reserve and neighbouring farms and forests. As a result, some crop damage claims reported to the Overijssel Fauna Management Unit (FBE) (FBE Overijssel, 2019). Consequently, there are losses, such as agriculture damage, diseases transmission, car-animal collisions, and animals attacks (Messmer, 2000).

Besides being ecological generalists, wild boars are often free-ranging beyond their natural habitats (Fernando et al., 2019). Saunders and Kay (1991) reported the respective home range of male and female wild boar as 10.7 km^2 and 4.9 km^2. Besides, Saunders & Kay (1991) found that the distribution of wild boars varies between various habitats and seasons in response to the availability of resources (food & water). Moreover, sexually active males tend to have a more expansive home range than the breeding females (Saunders & Kay, 1991). Besides, response to hunting pressure and summer temperature also affect the daily activity pattern of wild boars since they lack sweat glands for thermoregulation (Dexter, 1999). Hence, wild boars are primarily nocturnal and spend hot days wallowing in mud/water or under shade (Fernando et al., 2019). The home range often increases in winter since the animals needed to travel widely, searching for scarce pasture and to raise body temperature, thus conserving energy required for thermoregulation (Dexter, 1999). High reproduction rates characterize wild boars if favourable conditions such as sufficient forage availability are present (Bieber & Ruf, 2005). According to Lowe et al. (2000) cited in (Fernando et al., 2019), the species lacks natural predators in the Netherlands. Consequently, in cases where wild boar population densities are high and predators are absent, hunting is the leading method to control their densities (Franz et al., 2020). Therefore, hunting is a pivotal factor in wild boars' activity and distribution patterns outside protected areas.

1.2 Problem statement

Species distribution modelling (SDM) is an essential tool for inferring and predicting the habitat suitability for flora and fauna. The relationship between species observations and predictor environmental gradients is modelled to predict the spatial distribution of species. Due to advancements in remote sensing, the

environmental predictors have become readily available at high resolution. However, it is difficult and expensive to obtain field data especially using traditional survey methods, which used visual detections of wildlife through aerial or ground counts. Presence only (PO) biological data is easy to obtain from databases, atlases, and museums worldwide, leveraging the much-needed data. One limitation of the presence-only data is that it is prone to sampling errors since the sampling effort is not controlled, unlike the traditional sampling surveys. The severity of sampling bias may vary from one dataset to another depending on how the data was collected.

In most cases, the PO data is corrected in the areas that are easily accessible such as settlements, roads, rivers, and parks, thereby defying randomness. It's worth comparing the occurrence datasets of different origins to unravel the sources of sampling bias and test if it's significant. Various methods, including rarefaction, bias file, and extrapolation, are used for sampling bias correction.

Moreover, most methods of species distribution modelling assume a random distribution of samples in the study area. Besides, Elith et al. (2006) indicated that using multiple SDM can enhance prediction for regions with partial or limited data. This study aims to account for the sampling bias in species distribution modelling through rarefaction.

1.3 Research objectives, Questions & Hypotheses

1. **Objective 1:** To investigate the spatial distribution of wild boar observations and related bias from different data sources

1.1. **Question 1.1:** Are wild boar occurrence datasets from different origins, hunters (FBE) or citizen observers (NDFF), randomly distributed?

1.1.1. Hypotheses

It is suspected that the hunters' data is more biased to the areas with high chances of encountering wild boars. Similarly, the citizen volunteered data may also be biased since more observations are logged from the areas that are most visited.

1.2. **Question 1.2:** Does the distribution patterns of wild boar occurrence datasets vary based on origin?

1.2.1. Hypotheses

The distribution patterns of wild boar occurrence datasets from different origins are different

2. **Objective 2:** To investigate the effect of survey bias on the accuracy of species distribution modelling

2.1. **Question 2.1**: Which methods of species distribution modelling are most responsive to sampling bias correction?

2.1.1. Hypotheses

The empirical modelling models such as BRT and GLM will be more responsive to bias correction than the machine learning methods that fit complex algorithms and are therefore less affected by bias.

3. **Objective 3:** To determine which environmental variables are relevant to predict the distribution of wild boars in Overijssel Province

3.1. **Question 3.1:** Which environmental variables are more important in predicting the distribution of wild boars?

3.1.1. Hypotheses

Table 1: Predictor variables significance and their respective hypotheses responses

	Variable	Significance	Hypothesized response
1	Distance to Nature reserve	A suitable area where various essential habitat requirement for wild boars is available. Hunting is not allowed in the reserves. Citizens visit nature reserves for recreation and are likely to capture more data than other areas	The shorter the distance from the nature reserve, the higher the probability of wild boar occurrence
2	Distance to water	Wild boars are water-dependent for metabolism and thermoregulation	The closer it is to the water bodies, the higher the probability of occurrence
3	Distance to forest	Suitable for shelter and food, especially fruits from oak	The closer it is to the forest, the higher the probability of wild boar occurrence
4	Distance to artificial surfaces	May act as a source of disturbance and barriers to migration	The further away from artificial surfaces, you would expect higher chances to find wild boars
5	Distance to cultivated areas	Cultivated crops are food for boars. Shooting by hunters frequent in cultivated areas	The closer it is to the cultivated areas, the higher the probability of wild boar occurrence
6	Distance to heath and moorland	Shelter and feeding habitat	Areas closer to heath and moorland have a higher probability of being inhabited by wild boars

7	Distance to roads	Most volunteer data are collected from accessible areas and road kills which may lead to false presences	The probability of wild boar occurrence may tend to be higher in the areas closer to the road when models are fitted with biased citizen observer's data. However, if the data is randomized, the contrary will happen

3.2. **Question 3.2.** Which environmental variables are most consistent in predicting the distribution of wild boars in Overijssel?

3.2.1. Hypotheses

The distance to the nature reserve is more important in predicting wild boar using clustered dataset. More so, rarefaction will reduce the variability in the importance of various predictors.

2 MATERIALS & METHODOLOGY

2.1 Study area

The study is undertaken in the province of Overijssel in the Netherlands (Figure 1). According to Malinowski *et al.* (2020), the landcover of Overijssel can be classified into 13 different categories (Table 3). Moreover, Overijssel hosts many natural areas covering approximately 62500 acres (Albers & Hoekstra, 2019). Out of 145 nature areas in the Netherlands, 11 are in Overijssel, which harbours diverse flora and fauna. The forests that constitute most nature reserves are categorized as follows: Coniferous forests, mixed forests, beach-oak forests, stream-conducting forests, river-conducting forests, and low moor forests (Albers & Hoekstra, 2019). Forest are essential habitats for wild boars. The study area has a variety of unique landscapes namely: peat bog and raised moor (e.g., Engbertsdijksvenen), unique water ways (e.g., Rivers Regge, Dinkel), Moraines (e.g., Lemeler & Oldenzal) and the blue grasslands. Among all flora and fauna in the Netherlands, 50% of some 12 unique species are found in Overijssel (Albers & Hoekstra, 2019). In terms of surface geology: the south-eastern region is dominated by sand and rivers such as Regge, to the northwest are sediments of clay, Vecht and Overijsselse, the northern part has remnants of veens (bog) such as Aamsveen, Engbertsdjiksvenen and Witteveen, while the north-western region is lakes which comprise depressions of peat mining (*Overijssel*, 2021).

According to the Koppen classification, the area climate is the oceanic climate like the rest of the Netherlands, although the winters are more severe than the rest of the country since it is far from the sea (The Royal Netherlands Meteorological Institute (KNMI), 2020). The warmest month is July, with an average daily mean temperature of 17.6℃, while the coldest month is January with an average maximum temperature of 2.3℃ (The Royal Netherlands Meteorological Institute (KNMI), 2020). The wettest month is July, with an average precipitation of 74.5 mm, while the driest month is February, with an average rainfall of 51.6 mm (The Royal Netherlands Meteorological Institute (KNMI), 2020).

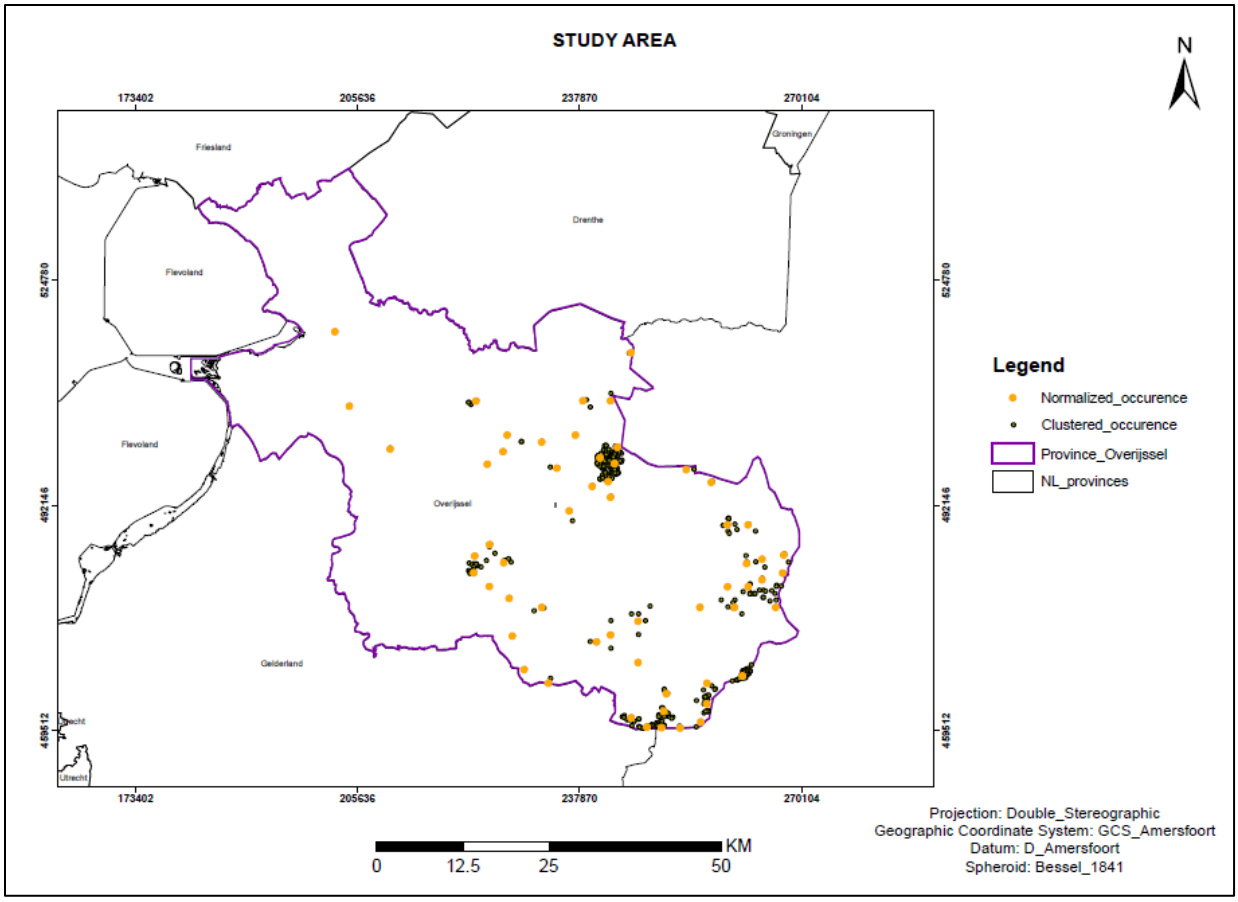


Figure 1: Study area

2.2 Materials

2.2.1 Data

The data comprised a pre-processed 10m resolution landcover map and a land-use shapefile for Overijssel province. The landcover map was extracted from Sentinel 2 Global Land Cover (S2GLC), which was developed by classifying Sentinel 2 imagery (Malinowski *et al.*, 2020). In addition, the land use map was downloaded from the Europe Geofabrik repository (geofrabrik, 2021). Additionally, the landcover and land use datasets were further processed to generate predictor variables. The wild boar's occurrence data were obtained from two repositories: Nationale Databank Flora en Fauna (NDFF) and Fauna Management Unit Overijssel (FBE). The data are tabulated below (

Table 2).

Table 2: Data

	Dataset	Extent	Type	Source
1.	10m resolution landcover map	Europe	raster	Opens street map
2	Land use map	Overijssel	Shapefile (Polygon & Lines)	(Malinowski et al., 2020)
3	432 wild boar occurrences (presence only)	Overijssel	Shapefile (points)	Nationale Databank Flora en Fauna (NDFF)
4	281 wild boar occurrences (presence only)	Overijssel	Shapefile (points)	Fauna Management Unit Overijssel (FBE)

2.3 Methodology

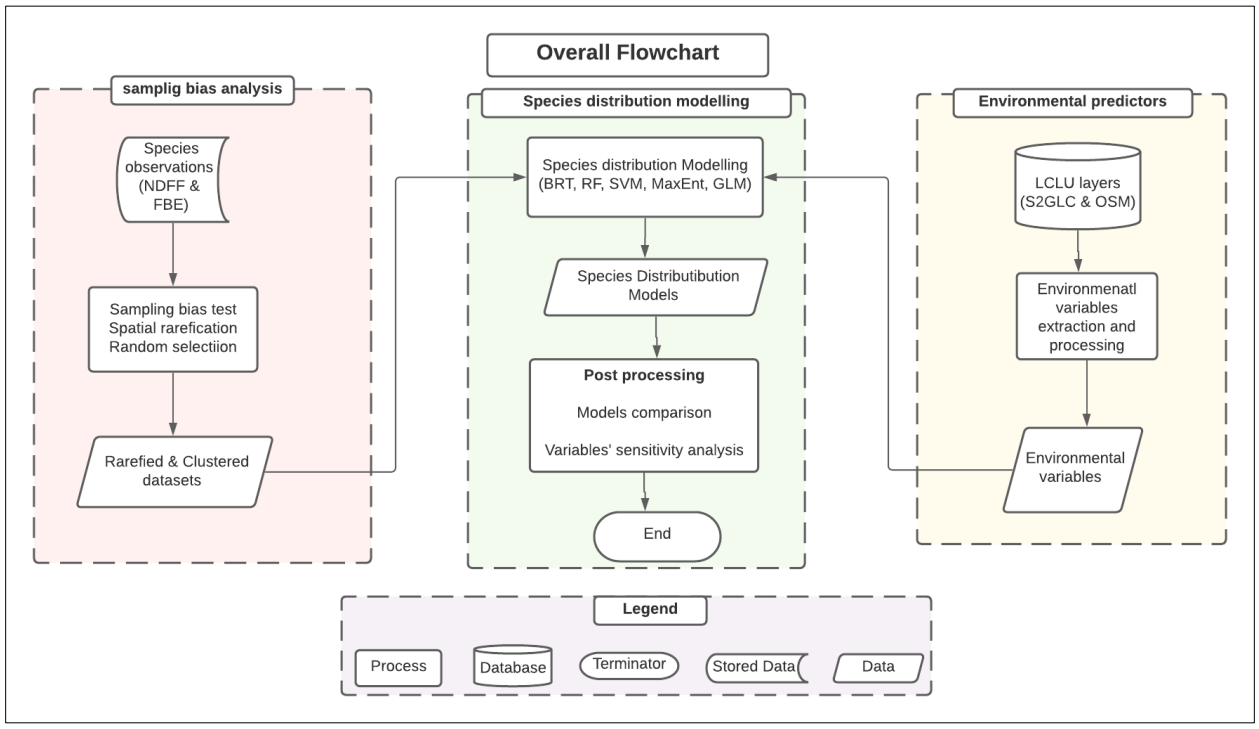


Figure 2: Overall flowchart

The flow chart above (Figure 2) represent the overall methodology used to analyse the data. Sampling bias analysis was undertaken using: an average nearest neighbour method for normality testing, a random selection of uncorrected observations, and spatial rarefaction for sample de-clustering. On the other hand, the environmental variables were extracted from land cover and land use maps. The rarefied and randomly selected observations and the environmental variables were used as inputs to species distribution modelling.

2.3.1 Data pre-processing

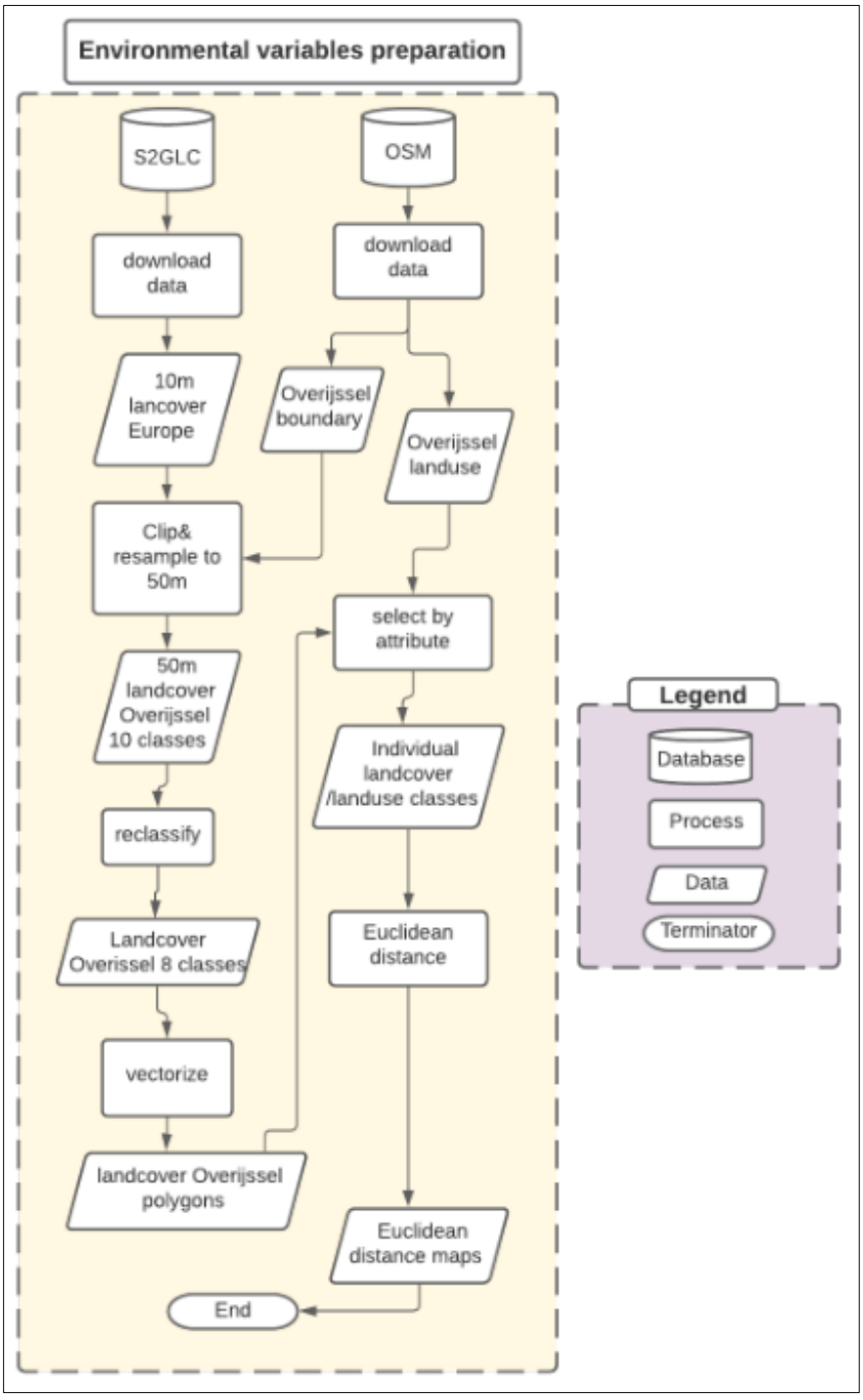


Figure 3: Flow chart demonstrating preparation of environmental variables

The preparation of predictor variables entailed projecting the individual layers to Dutch national triangulation system, clipping to the extent of Overijssel boundary and resampling the land cover map from 10m to 50m spatial resolution (Figure 4). Although a 10m resolution map would provide information necessary for creating environmental gradients, the computer processing capacity was highly diminished, therefore the necessity to resample to 50m. The original landcover map had 13 classes excluding the clouds and 'no data', although only ten types were represented in the study area as shown in the legend (Table 3). Further, the categories were reclassified into eight classes (Table 3). Then, the landcover map was vectorized for extracting the individual landcover/land use (LULC) classes. Ancillary predictor variables (Table 1) were created by undertaking Euclidean distance analysis at 50 meters intervals for the respective LULC classes.

The FBE data was provided in a comma-separated values (CSV) format, while the NDFF data was in shapefile format. Therefore, all the data were converted to shapefile and projected to the Dutch national triangulation system.

Table 3: Landcover legend at the extent of Europe and Overijssel

Class Code	**Europe**	**Overijssel**	**Overijssel _reclassified**
0	Clouds	Clouds	No data
62	Artificial surfaces and constructions	Artificial surfaces and constructions	Artificial
73	Cultivated areas	Cultivated areas	Cultivated
75	Vineyards	-	-
82	Broadleaf tree cover	Broadleaf tree cover	Forest
83	Coniferous tree cover	Coniferous tree cover	
102	Herbaceous vegetation	Herbaceous vegetation	Herbaceous
103	Moors and Heathland	Moors and Heathland	Moors and Heathland
104	Sclerophyllous vegetation	-	-
105	Marshes	Marshes	Water
162	Water bodies	Water bodies	
106	Peatbogs	Peatbogs	Peatbogs
121	Natural material surfaces	Natural material surfaces	Natural surfaces
123	Permanent snow-covered surfaces	-	-
255	No data	No data	No data

2.3.2 Sampling bias analysis

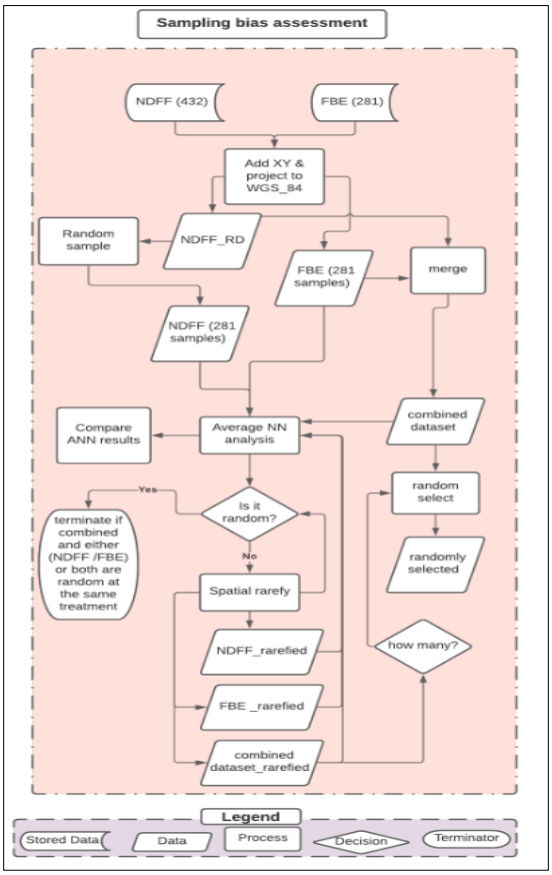


Figure 4: Flowchart demonstrating sample bias analysis

The wild boar occurrence data were analysed to investigate if it is affected by survey bias or not (Figure 3). Several treatments were performed in order attempt to satisfy the Poisson point process. Firstly, the FBE (hunters records) and NDFF (Citizen observations) were standardized to the same sample size to allow for unbiased comparison. Since the NDFF dataset had more samples (432) than FBE (281), an equal selection was randomly selected from the data. Then, the two standardized datasets (281 observations each for FBE & NDFF) were tested for randomness using the average nearest neighbour method. The average nearest neighbour method was used for testing whether the features in a defined area are because of a random process. The distribution pattern of point features was determined using the nearest neighbour ratio method.

2.3.2.1 Average nearest neighbour method

This method analyses the spatial distribution of two-dimensional point (coordinates) data such as species occurrences. According to Fortin and Dale (2009), the methods aim to test if the data conform to complete

spatial randomness (CSR) distribution. The average nearest neighbour method was selected for this study. This method calculates the distance between one feature and the other and tests if it differs significantly from the expected depending on the study plot scale. The nearest neighbour ratio (NNR) is used to distinguish three possible patterns on the scale of "1" (random), ">1" (dispersed) and "1>" clustered under the null hypothesis. The nearest neighbour ratio is computed using the equation below (Equation 1) developed by Pielou (1959) cited in (Fortin & Dale, 2009).

$$Q = \pi\lambda\left(\sum_{i=1}^{n} W_{i1}^{2}\right)/n$$

Equation 1

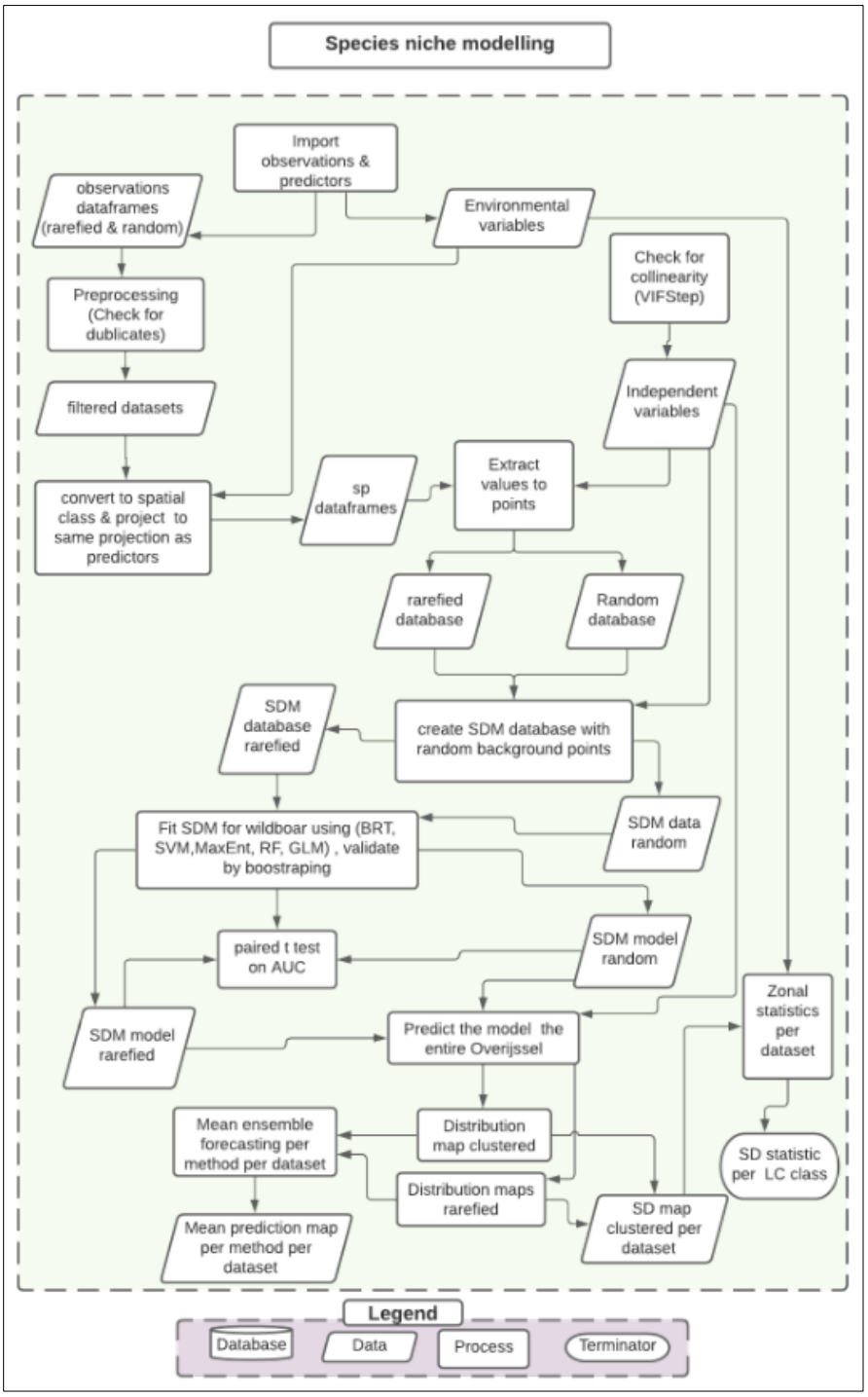
Where, Q=Nearest neighbour ratio, *Wi1*=distance between events (*i*) and λ=density of events and *n*=total number of occasions.

The nearest neighbour ratio depicts random distribution at the value 1, whereas values below and above 1 mean that the features are clustered and dispersed, respectively (Fortin & Dale, 2009). The nearest neighbour ratio is computed by dividing the observed mean distance between values separated by the expected mean distance between neighbouring features (Fortin & Dale, 2009).

2.3.2.2 Samples de-clustering

If the samples are clustered after the average nearest neighbour analysis, de-clustering was done using the spatial rarefy method (Brown, 2014). This method removes multiple occurrence points within a specific user-defined Euclidean distance (Brown, 2014). Again, the average nearest neighbour analysis is repeated on the spatially rarefied samples to verify the distribution pattern. Since spatial rarefy reduces the number of occurrence points, it was prudent to use the entire dataset (FBE & NDFF) in modelling. Therefore, a similar treatment was performed on the combined dataset. In essence, the combined dataset was initially analysed for randomness. If spatial clustering was found, the data was rarefied appropriately. Finally, a random sample equal to the resultant rarefied dataset was randomly selected for comparison with the latter after fitting species distribution models. The average nearest neighbour ratios were compared for all the data treatments.

2.3.3 Species distribution modelling



The inputs for the species distribution modelling were as follows: 58 rarefied and clustered observations each, and environmental variables including 50m resolution Euclidean distances maps from roads, forests, nature reserves, heath & moorland, water, artificial surface, and cultivated areas. The species distribution modelling analysis was performed in R Studio software using the SDM package.

The datasets (Occurrence & predictors) were loaded into the R studio. All the environmental variables were stacked a raster brick. Pre-processing included checking the occurrence datasets for duplicates and collinearity test for predictor variables. Variance Inflation Factor (VIF) at a threshold of 10 was performed using the stepwise elimination method. All the predictors were retained since the collinearity was below the threshold. The occurrence points were converted to a spatial data frame (sp) and projected to the same spatial reference system as the predictor variables. A data frame with corresponding values of predictor variables at the respective occurrence points was created. A unique database used by the SDM package was formed comprising of the training data (presence only), predictor variables (raster brick) and a set of randomly selected background points equivalent to training data. A couple of models were fitted using five different methods: Boosted Regression Trees (BRT), Random Forests (RF), Maximum Entropy (MaxEnt), Support Vector Machines (SVM) and Generalized Linear Models (GLM). For each method above, twenty (20) replications were done using the bootstrapping procedure. Model evaluation was also performed using bootstrapping; essentially, the method samples with replacement whereby the observations not drawn for training are used for validation. Finally, the models were extrapolated to the predictor maps to generate spatial predictions. Furthermore, the species distribution maps were fine-tuned by calculating weighted averages and standard deviations.

2.3.3.1 Description of SDM methods

1. Generalized Linear Model

Ordinarily, generalized linear models (GLM) comprise a class of statistical analysis that fits exceptional cases of linear models (McCullagh & Nelder, 2019). However, these models are related in that a predictor variable is combined using a linear function. However, the method is insufficient in fitting smooth curves akin to statistical methods such as Generalized Linear Models. More so, the models use maximum likelihood to predict the response variable for binormal variables such as the presence/absence data for species occurrence, logit function (logistic regression) fits a logistic model's parameters. By calculating the logarithm of odds ratio (probability of occurrence or not), the predictions are converted to linear probabilities (0-1).

2. Maximum Entropy

Maximum Entropy (MaxEnt) is a method that infers the probability of species distribution by finding the most spread out (near-uniform) constraints (Philips et al., 2006). The method fits a model that presents the maximum information gain from a set of measured features. For each sample observation, the predicted value should be equal to its average taken from the normal distribution. In species distribution modelling, the target features for modelling are sampled observations, and constraints are environmental gradients of the locality. Moreover, the boundary for fitting the probability distribution is defined by the pixels of the study area. MaxEnt employs a generative prediction approach. For unknown distribution probability in X (pixels of the study area), assigns a non-negative distribution to each observation (x) that sums to 1 (Equation2). Some salient features that make this method preferable are that it requires presence-only data. It has deterministic algorithms that meet optimal distribution and suitable for a limited amount of training data.

Probability distribution= $-\sum_{i=1}^{k} Piln(Pi)$

Equation 2

Where, i=1-observations where species are present, Pi-probability of occurrence in pixel 'i', k-total observations and, 'ln' is the natural logarithm to base 2

3. Boosted Regression Trees

Boosted regression trees (BRT) techniques combine the individual modelling methods decision trees and regression trees, enhancing the outputs through clustering (Elith et al., 2008). For decision trees, a rule-based partitioning of the variables identifies the most dominant tree fitted as constant. Then, using regression trees, a mean response algorithm is fit to the samples on that partition, assuming a normal distribution of deviations. The structure of the trees is hierarchical, whereby the response to the lower variables depends on the one above them. This model is more suited for modelling interactions between variables.

4. Support Vector Machines

Support vector machines (SVM) is a binary machine-learning classification technique. It separates multi-dimensional feature space data into two subclasses by fitting a geometric hyperplane to optimally split a binary dataset (Bruzzone & Persello, 2009). Assuming a hyperplane H: $y = (w.x) + b = 0$, the binary data (e.g., presences/absences) is separated based on the distance (vector w), of predictors at sample (x), from the hyperplane (H), and fitted scalar origin (b). Fitting of the hyperplane and the origin entails finding an optimal

position for separating the training data. This method is effective for training models with minimal samples and for fitting simple linear functions which are rather non-linearly inseparable using statistical methods (Bruzzone & Persello, 2009).

5. Random Forest

Like the BRT, random forest (RF) fits several decision trees and later ensemble the predictions (Cutler et al., 2007). In contrast, though, RF does not model regression function akin to BRT. Bootstrapping (sample with replacement) is used to sample several observations and then fit the classification algorithms. For every bootstrap sample, the out-of-bag observations (the observations that do not participate in model training) are used for validation. The best class is selected based on the majority vote, whereby the ties are chosen randomly (Cutler et al., 2007).

2.3.4 Assessing effects of sampling bias on species distribution models

The effect of sampling bias was assessed by comparing model performance for the different methods based on the area under the receiver operating curve (AUC) statistics. The AUC metric ranges between 0 and 1 where a score of 1 depicts the best discrimination, 0.5 discriminates randomly, and <0.5 less than random. The variation in the performance of different methods was compared using boxplots. The paired t statistic was also used to test if there is significance in the performance of other methods based on AUC.

2.3.5 Variable importance

Variable importance was compared using the percentage contribution of each variable to models' permutations. Another way of determining the contribution of variables was through the response curve. Indeed, the response curves can also indicate the sensitivity of variables by showing the standard deviation range. Normal deviation maps were also used to show the stability of various variables in predicting wild boar distribution.

3 RESULTS

3.1 To investigate variation in the distribution pattern of wild boar observations with varied survey bias

The different wild boars' observation datasets showed variation in distribution due to how they were surveyed. Although both datasets are significantly clustered (p-0.001), the 281-occurrence data acquired from citizen observers (NDFF) was less clustered (NNR-0.32) compared to the similar data obtained from hunters' records (FBE) (NNR-0.13). However, when rarefied, 44 NDFF randomly distributed observations are left at (NNR-0.99), while the 21 FBE records remain clustered (NNR-0.58). The findings agree as hypothesized that hunters' data is more clustered since the hunting sites are targeted mainly in areas where a higher probability of shooting an animal are preferred. On the contrary, the NDFF data is from volunteer observers distributed widely, and their observation is likely to be stochastic. Like, the NDFF data, the combined dataset showed clustering for the randomly selected sample (NNR-0.36). In comparison, the rarefied one was randomly distributed (NNR-0.96). The summary statistics of data thinning and normality tests are shown in (Table 4) and (Figure 5).

i) Summary statistics of data thinning and test sampling bias analysis

Table 4: Summary statistics of data thinning and normality test (study area=3420840181 m²)

	NDFF uncorrected	NDFF rarefied (2km)	FBE uncorrected	FBE rarefied (2km)	NDFF&FBE uncorrected	NDFF&FBE rarefied (2km)	NDFF&FBE randomly selected
N	281.00	44.00	281.00	21.00	713.00	58.00	58.00
Observed mean d (m)	550.09	4359.33	226.92	3678.50	243.06	3675.89	1383.31
Expected mean d (m)	1744.55	4408.70	1744.55	6381.56	1095.20	3839.92	3839.92
NN-Ratio	0.32	0.99	0.13	0.58	0.22	0.96	0.36
Z-score	-21.96	-0.14	-27.90	-3.71	-39.75	-0.62	-9.32
P-value	0.00	0.89	0.00	0.00	0.00	0.53	0.00
Pattern	clustered	Random	Clustered	Clustered	Clustered	Random	Clustered

ii) Comparison of distribution patterns of various datasets based on the average nearest neighbour ratio

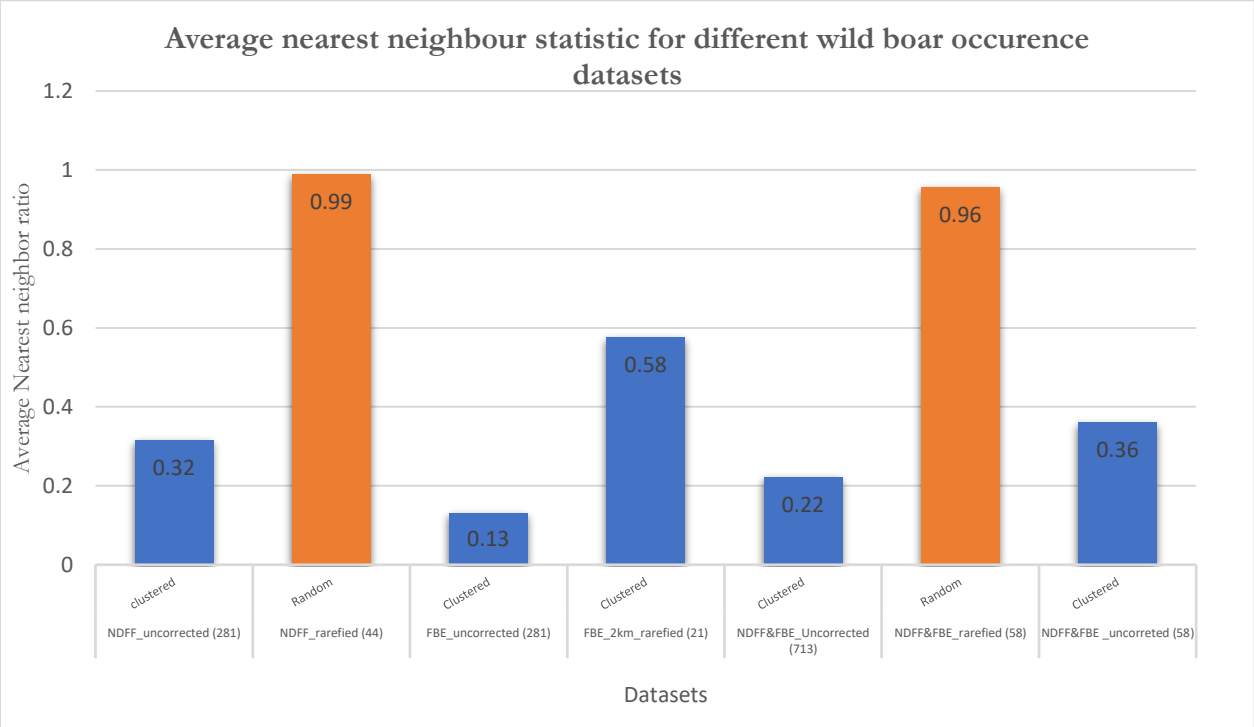


Figure 5: Comparative analysis of distribution patterns of datasets of different origins (Clustered-Blue, Random-Orange)

3.2 The effect of survey bias on the accuracy of species distribution modelling methods

Based on the area under the receiver operating curve (AUC), random forests (0.94) outperformed the other modelling methods, followed by boosted regression trees (0.92), support vector machines (0.91), maximum

). The performance of all models registered high accuracies when fitted with the clustered datasets (BRT-0.92, RF-0.94, MaxEnt-0.89, SVM-0.91, GLM-0.89) compared to rarefied datasets (BRT-0.84, RF-0.9, MaxEnt-0.84, SVM-0.84, GLM-0.8). High accuracies were expected since the clustered datasets tend to overfit the model towards the bias. More so, as demonstrated in (Figure 7), apart from RF (0.91) and GLM (0.8), all the other models had a similar AUC (0.84), portraying a high level of unity among models.

In terms of sensitivity to sampling bias, all methods show a significant difference in model performance for rarefied and randomly selected data (Figure 8 & Figure 9). These two methods are suitable for fitting complex models with limited training samples. Similarly, as shown in the box plot (Figure 8), RF and MaxEnt models

have fewer outliers for both datasets than the other methods. GLM and BRT show the most variability in predicted models (Figure 8). Both GLM and BRT share similarities in that they fit empirical models based on regression which are data intensive. In addition, the two methods utilize presence and absence data; thus, the background data used in this study may affect the model performance.

i) Spatial prediction of wild boars' niche distribution in Overijssel

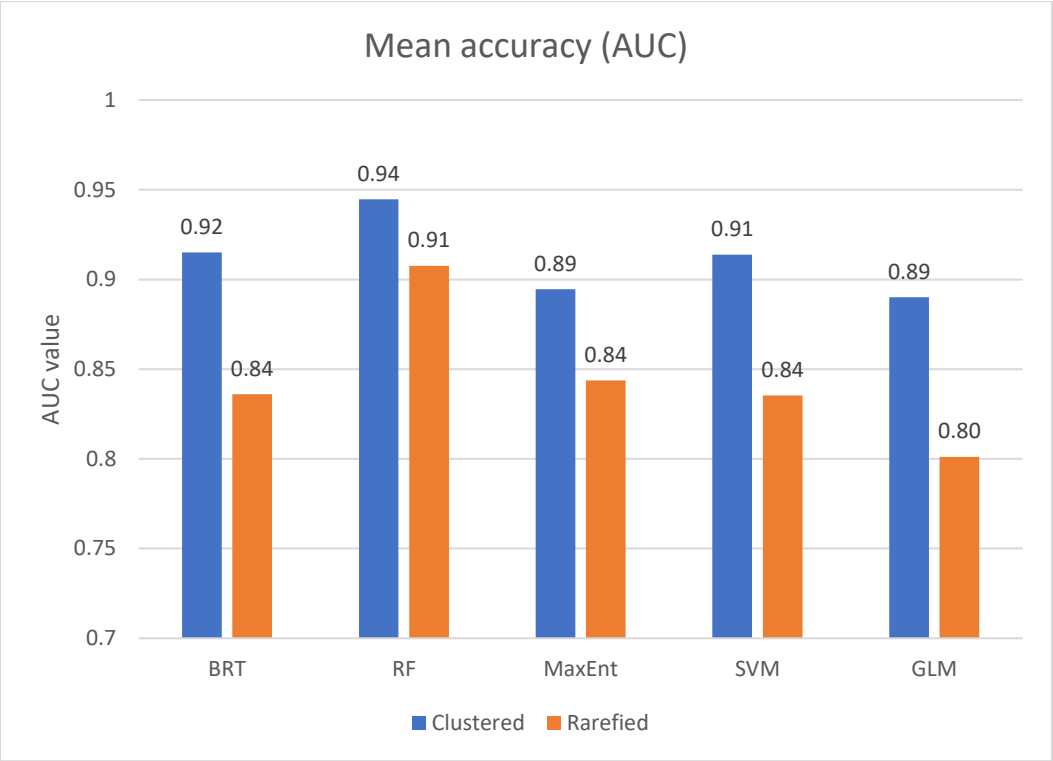


Figure 6: Comparison of model accuracy based on mean AUC

ii) Variability in performance of various SDM methods based on AUC metric

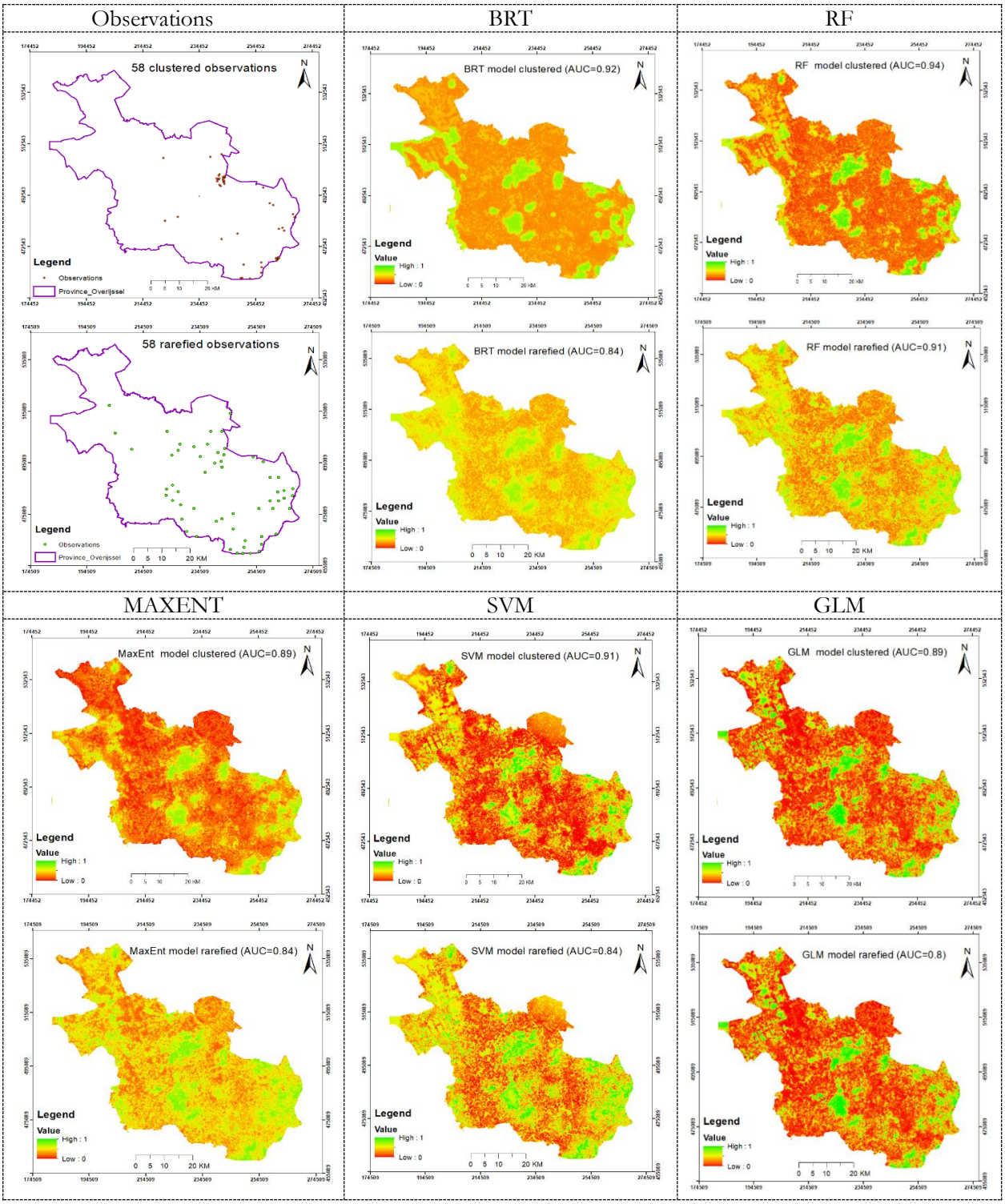


Figure 7: Wild boar distribution maps showing performance of various methods on clustered and rarefied datasets based on AUC metric

iii) Variability in performance of various SDM methods based on AUC metric

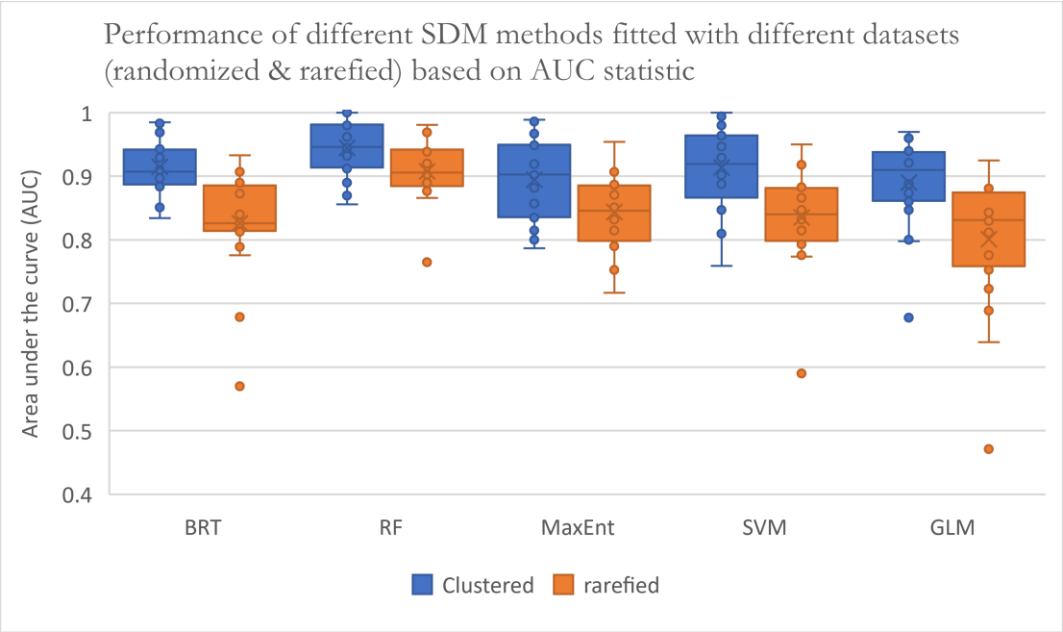


Figure 8: Performance of various models fitted with randomized and rarefied datasets based on AUC metric

iv) Comparison of model performance using clustered and randomized datasets

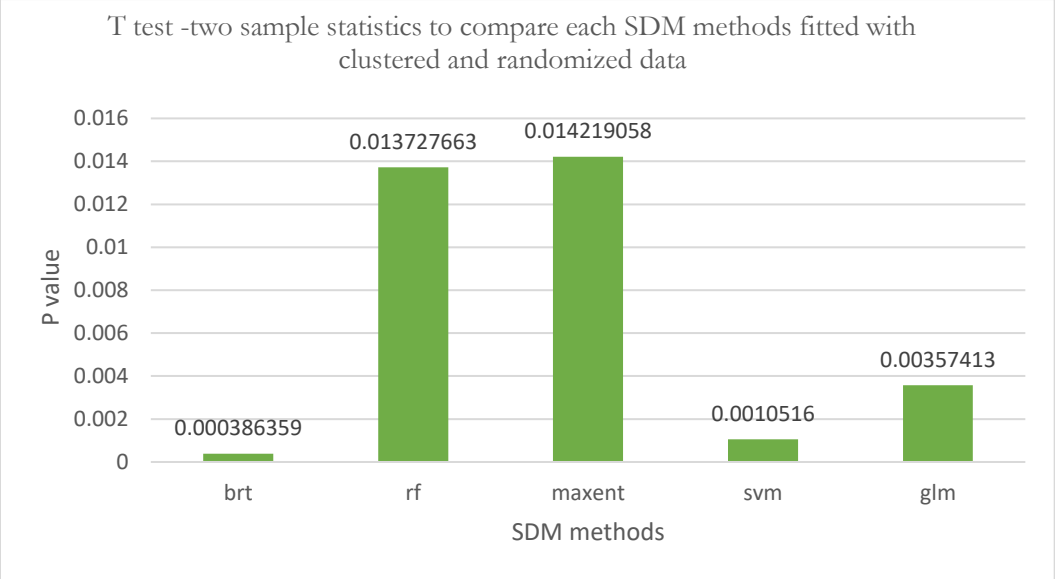


Figure 9: Summary t statistic (two-sample assuming unequal variance) to test if there is a significant difference between SDMs' accuracy based on AUC fitted with the same method using clustered and rarefied datasets

3.3 Which environmental variables are relevant to predict the distribution of wild boars in Overijssel Province?

Based on the AUC metric, distance from nature reserves contributes most to models' permutations in the clustered and the rarefied datasets (Figure 10). Similarly, the distance to nature reserves demonstrates high variability among different models for both rarefied and clustered datasets. On the other hand, the distance to the water contributes the least to model permutations and has the least variability in among models. Moreover, distance to roads ranks second in model importance for the clustered dataset, which depicts biased observations in the areas that are easily accessible as hypothesised. In addition, there is high variability amongst predictors in variable importance based on the clustered dataset (Figure 10). However, the variation in variable importance is relatively low in the rarefied datasets because the observations are clumped in some land covers (e.g., nature reserves and roads), which are the most visited.

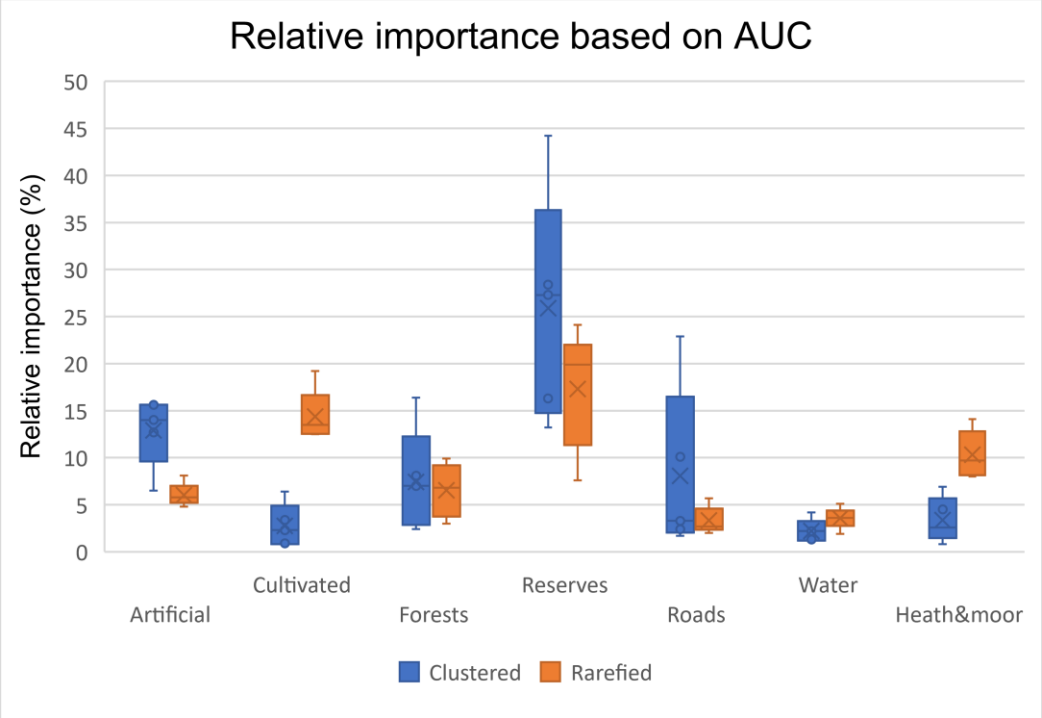
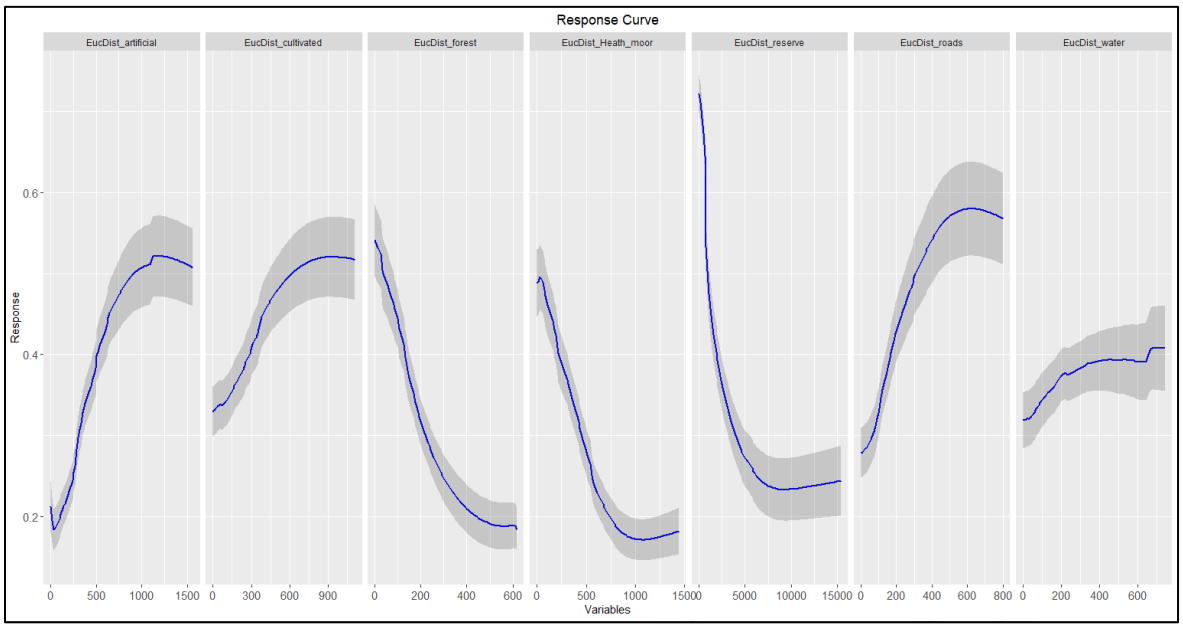


Figure 10: Box plot showing variability in relative variable importance in the prediction of suitable habitats for wild boars based on AUC metric of clustered and rarefied datasets

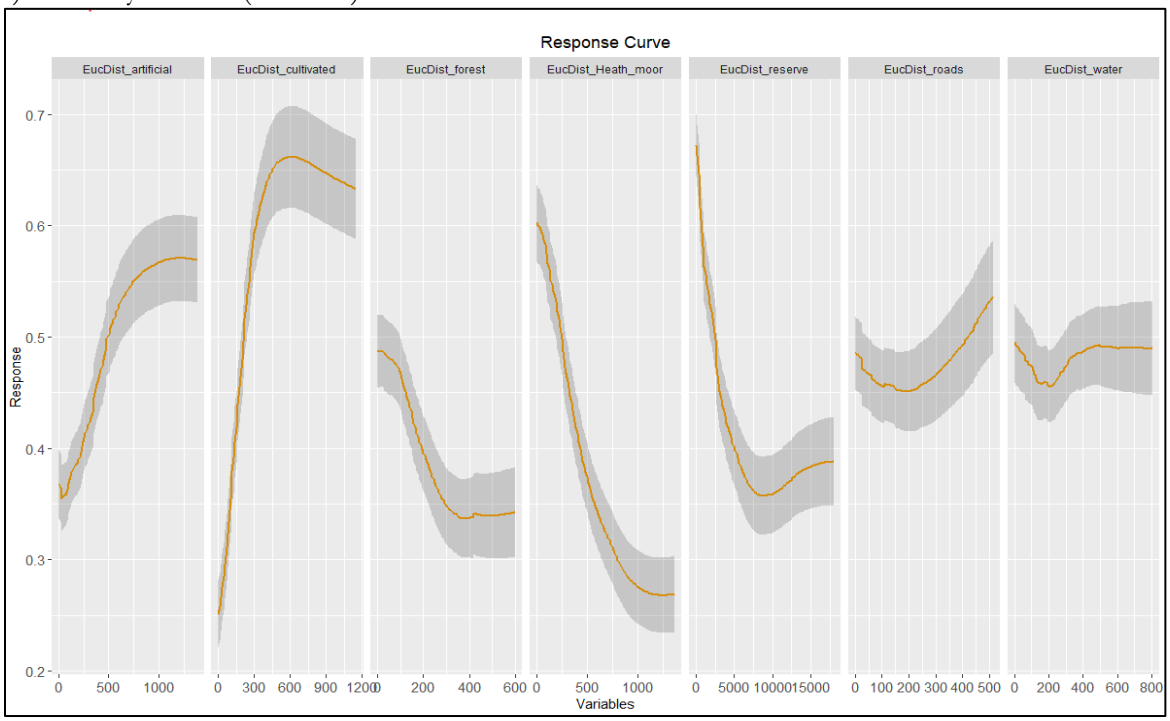
In terms of variable stability demonstrated in (Figure 11), the distance to forests, nature reserves and heath & moorland depict a direct relationship with the occurrence of wild boars. These are the known habitats of wild boars, therefore the direct correlation. On the other hand, distance to roads, artificial areas, cultivated land and water are inversely related to the occurrence of wild boar as depicted by the randomly selected and partly by the rarefied data. However, the distances to water and roads predict the opposite, although with high uncertainty. The direct relationship of wild boar to distance to roads is possible due to easy access to observers and that some observations may be recorded from accessible areas.

Similarly, wild boars are water-dependent to aid in thermal regulation since they lack sweat glands. The distance to water, forest and roads showed high uncertainties in predicting wild boars (Figure 11). The respective zonal statistics per land use were retrieved from the standard deviation maps (Figure 12). However, roads as landcover were not used in zonal statistics analysis because they constituted the artificial areas, and their scale was below the resolution (50m) used. For both datasets, the areas close to water bodies had consistently high standard deviation, unlike cultivated areas with the lowest. However, the variations in predictions between the datasets in the other classes (nature reserve, forest, heath & moor, and artificial areas) did not show a consistent trend. As shown in (Figure 12), the area northwest of the study area mainly covered by wetlands (water) had the smallest number of observations for both datasets. However, the clustered dataset variables' response curves are more consistent than the rarefied ones (Figure 11). This disparity can be explained by the possibility of spatial autocorrelation amongst neighbouring observations in the clustered dataset, which is essential for establishing an empirical relationship between occurrence locations and environmental variables. The environmental gradient, a distance related dependence, is crucial for developing a model consistent with the first law of geography "All things are related to each other but near things are more related than the distant ones".

i. Response curves



a) randomly selected (clustered)



b) Rarefied

Figure 11: gg-plot illustrating how different landcover predict the occurrence of wild boars based on: (a) randomly selected, (b) rarefied dataset for various environmental gradients based on Euclidean distance from LU/LC: Eudist_artificial (artificial),

Eudist_cultivated (cultivated), Eudist_Heath_moor (Heath Moor), Eudist_reserve (reserve), Eudist_roads(roads), Eudist_water(water)

ii. Standard deviation maps based on combined model predictions

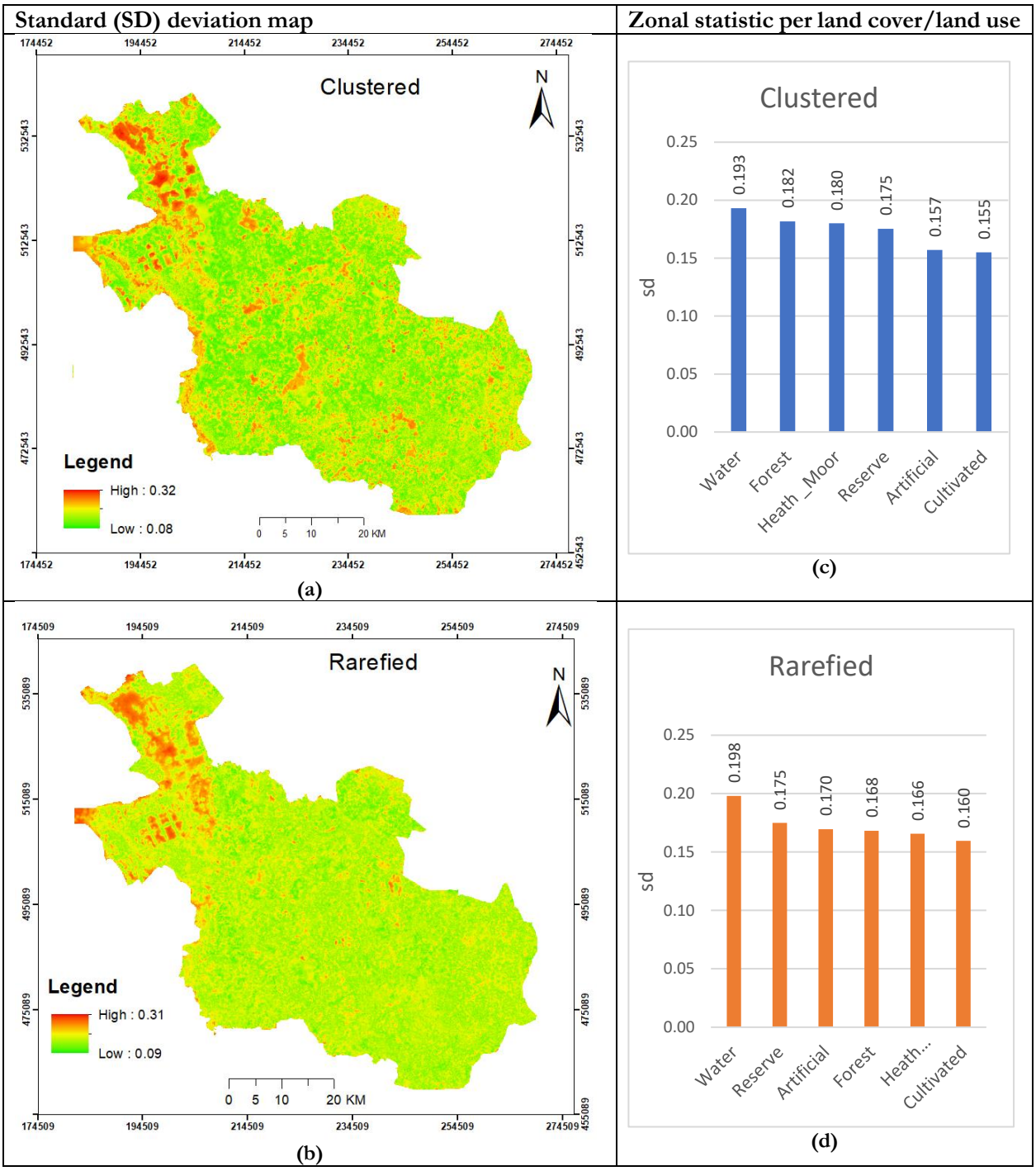


Figure 12: Wild boar's distribution maps (a) clustered & (b) rarefied and the respective zonal statistics plots (c) &(d) showing areas of model uncertainties depicted by the standard deviation

4 DISCUSSION

The distribution pattern of flora and fauna observations is dependent mainly on the means the data was acquired. Moreover, high variability is imminent where flora and fauna occurrence data is sourced from stochastic observations during recreation, nature walks, hunting and game drives. On the contrary, for research projects with a detailed field study, the distribution pattern of the observations conforms to a systematically structured data collection design. For this case study, the distribution patterns of two datasets of wild boar observations acquired from volunteer observers (NDFF) and hunters (FBE) were compared (Figure 5). For a densely inhabited country like the Netherlands, it would be expected that volunteer observers visit specific areas such as parks and reserves therefore the observations are also clustered to these areas. Similarly, the hunters visit specific areas where they anticipate high chances of finding a wild boar in the permitted areas. Therefore, as expected, both datasets (281 samples) were significantly clustered although the NDFF data was sufficiently de-clustered after the first rarefication at 2km, unlike FBE, which remained clustered.

Similarly, when the datasets (FBE & NDFF) were combined, the first rarefication at 2km was sufficient to randomize the data significantly. While the NDFF data was expected to be randomly distributed akin to the observers, high sampling density in the more visited areas such as parks and roadsides was the reason for the initial clustering. A similar pattern was observed in a similar study of Passerine birds undertaken in sub-Saharan Africa, where clustering was observed in the areas close to the roads, cities and rivers (Reddy & Dávalos, 2003).

Based on the AUC metric of accuracy, the comparison of individual species distribution modelling methods (BRT, RF, MaxEnt, SVM & RF) in predicting wild boar habitats within Overijssel for clustered and rarefied data depicted a significant difference. Models based on the clustered dataset, demonstrated high discrimination for wild boars' habitat suitability because of model overfitting due to the higher sampling effort in the area most visited and easily accessible areas. In agreement with the findings of Reddy & Dávalos (2003), high sampling intensity was found to increase the density probability of observing species in sub-Saharan Africa. More so, RF and MaxEnt demonstrated higher tolerance to sampling bias, consistent with the findings of Elith et al. (2006); the two modelling methods can fit complex models through regularization (MaxEnt) and ensemble boosting (RF). In terms of model performance based on AUC, the models utilizing machine learning algorithms (BRT, RF, SVM & MAXENT) outperform simpler methods such as regression models (GLM) (Elith et al., 2006). Also, the GLM method is designed to use presence-absence data; therefore, presence-background data may yield erroneous predictions. However, all the methods performed well with datasets by predicting the

distribution of wild boars with AUC above 0.8. Conversely, Elith *et al.* (2006) remarked that species with a wide habitat range akin to wild boars tend to have low accuracy. The high accuracies realised are due to the limited areas for dispersal of wild boars outside protected areas hence the species are confined to nature reserves and the adjacent farmlands. More so, bootstrapping method of replication performs well for even with limited data, but the independent testing may be affected unlike sub-sampling.

As demonstrated in section (3.3) the areas nearest to the nature reserves, forests, heath, and moorland consistently depicted a positive relationship with wild boar density. These variables form part of critical habitats where wild boar forage, shelter and are protected from hunting (Dexter 1999; Fernando et al. 2019, Franz, Markus & Peter 2020). Conversely, in clustered and partially rarefied datasets, the model demonstrated a negative relationship in wild boar habitats to cultivated areas and water & wetlands (water), otherwise known as essential habitats (Dexter 1999; Fernando et al. 2019, Franz, Markus & Peter 2020). The cultivated areas serve as foraging habitats, while wetlands provide shelter and thermal regulation since wild boars lack sweat glands. According to Fernando et al. (2019), wild boars tend to avoid areas with high human disturbance (e.g. farmlands) or change activity patterns to nocturnal, thus the negative relationship. The samples were very low or missing for the water areas, possibly because revellers do not prefer them for recreation. However, the distance to roads variable partially points to a positive relationship with wild boar occurrence. Like Reddy & Dávalos (2003) findings, areas that are more accessible and are already designated for conservation tend to be intensely sampled. Especially for recreations, game watchers and targeted biodiversity studies, areas known to harbour target species are often visited (Reddy & Dávalos, 2003).

Predominantly, distance to the nature reserves contributed to model permutations for either dataset. However, the inconsistency in the importance for all was lower and more even for the rarefied dataset. Higher sampling intensities in most visited areas lead to clustering, thus overpredicting species distribution (Reddy & Dávalos, 2003). Therefore, bias correction through rarefaction solved reduced contribution of intensely surveyed and redistributed to the others.

In terms of model consistency for various land cover or land use, water exhibited the highest standard deviation; the cultivated had the lowest while the others varied in-between. As stated earlier, areas covered by water and wetlands had low training samples, explaining the variability consistent with Stockwell and Peterson (2002) study that demonstrated an increase in accuracy of an SDM with the increase in sample size. On the other hand, observations in the cultivated areas may be more random since farming is a prevalent land use all over the study area.

5 CONCLUSIONS

Below the main conclusions are summarised for each of the objectives set in the present study.

5.1 To investigate variation in the distribution pattern of wild boar observations with varied survey bias

This study demonstrated that location data of flora and fauna observations acquired through volunteer observers and hunters are affected by clustering. Moreover, hunters' shooting and observation data are more clustered than volunteer sourced observations. Consequently, rarefaction coupled with an appropriate spatial statistical testing method such as the average nearest neighbour is satisfactory to correct the sampling bias introduced by uneven survey effort in ecological studies.

5.2 The effect of survey bias on the accuracy of species distribution modelling

It was demonstrated that most niche distribution modelling methods are affected by sampling bias, as in this study, five methods (BRT, RF, MaxEnt, SVM & RF) were investigated. Although all models fitted with clustered and rarefied datasets for the same method were significantly different, the machine learning-based techniques such as RF and MaxEnt were less significant than regression-based modelling, for example GLM & BRT. For all methods, rarefaction of species observations reduced the accuracy of the models based on the AUC metric

5.3 Which environmental variables are relevant to predict the distribution of wild boars in Overijssel Province?

In ecological niche modelling, areas that are more accessible are often intensely surveyed, thereby contributing more to model permutations. However, after bias correction through rarefaction, the contributions of the other environmental variables are more evenly shared. The environmental variables that characterize the ecology of wild boars (e.g., shelter, thermal regulation, stress) demonstrate consistent response, either positive or negative, in modelling their distribution. The biased survey causes variation in models' predictions whereby the areas that have more samples predict species distribution more consistently than the otherwise. Likewise, the randomly distributed predictors in nature (e.g., farmlands) are equally consistent in the model's permutations.

6 RECOMMENDATIONS

When using secondary sourced occurrence data for niche distribution modelling, utmost care is required to avoid transferring the sampling errors in the observations to the final predictions. Sampling effort standardization will adduce more confidence in model interpretation and application to conservation decisions. It is prudent to generate multiple models either by replication or/ and using different methods for comparison and more reliable application of the outputs for decision making. Combining several datasets will enhance predictions to avoid over predictions in areas affected by samples gaps. Since more high spatial resolution environmental predictors are available due to remote sensing and open data science, it is advisable to use several variables and critically assess their contribution to model performance to uncover sources of model variabilities. Consequently, this will enable the selection of the optimal variables for species distribution modelling.

Future studies may focus on:

- comparing the efficacy of survey bias-corrected models with presence-absence data collected from structured surveys.
- Replicate a similar study to test sampling bias sensitivity for the other existing and novel methods of niche distribution modelling (e.g., Convolution Neural Networks).
- Testing the performance of the method using other methods of replication such as subsampling.

REFERENCES

Albers, & Hoekstra. (2019). *The state of the biodiversity in Overijssel* (Issue April).

Bieber, C., & Ruf, T. (2005). Population dynamics in wild boar Sus scrofa : ecology , elasticity of growth rate and implications. *Journal of Applied Ecology*, *42*(Silvertown 1980), 1203–1213. https://doi.org/10.1111/j.1365-2664.2005.01094.x

Blome, S., Gabriel, C., & Beer, M. (2013). Pathogenesis of African swine fever in domestic pigs and European wild boar. *Virus Research*, *173*(1), 122–130. https://doi.org/10.1016/j.virusres.2012.10.026

Brown, J. L. (2014). SDMtoolbox: A python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses. *Methods in Ecology and Evolution*, *5*(7), 694–700. https://doi.org/10.1111/2041-210X.12200

Bruzzone, L., & Persello, C. (2009). Approaches based on support vector machine to classification of remote sensing data. *Handbook of Pattern Recognition and Computer Vision, Fourth Edition*, 329–352. https://doi.org/10.1142/9789814273398_014

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Dexter, N. (1999). The influence of pasture distribution, temperature and sex on home-range size of feral pigs in a semi-arid environment. *Wildlife Research*, *26*(6), 755–762. https://doi.org/10.1071/WR98075

Elith, H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., … E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

FBE Overijssel. (2019). *BIJ12 Faunazaken – schadecijfers 2019 – Faunabeheereenheid Overijssel*. https://overijssel.faunabeheereenheid.com/bij12-faunazaken-schadecijfers-2019/

Fernández, D., & Nakamura, M. (2015). Estimation of spatial sampling effort based on presence-only data

and accessibility. *Ecological Modelling*, *299*, 147–155. https://doi.org/10.1016/j.ecolmodel.2014.12.017

Fernando, Guilherme, Zilca, & Pellegrin, V. (2019). Activity pattern and habitat selection by invasive wild boars (Sus scrofa) in Brazilian agroecosystems. *Mastozoologia Neotropical*, *26*(1), 129–141. https://doi.org/10.31687/saremMN.19.26.1.0.08

Fortin, M.-J., & Dale, M. R. T. (2009). Spatial analysis of population data. In *Spatial Analysis* (pp. 32–110). https://doi.org/10.1017/cbo9780511542039.003

Franz , Markus, Peter, C. & J. (2020). Adaptation of wild boar (Sus scrofa) activity in a human-dominated landscape. *BMC Ecology*, *20*(1), 1–14. https://doi.org/10.1186/s12898-019-0271-7

geofrabrik. (2021). *Overijssel*. OpenStreetMap. https://download.geofabrik.de/europe/netherlands/overijssel.html

Hegel & Marini. (2013). Impacto do javali Europeu, Sus scrofa, em um fragmento da Mata Atlântica Brasileira. *Neotropical Biology and Conservation*, *8*(1), 17–24. https://doi.org/10.4013/nbc.2013.81.03

Komori, O., Eguchi, S., Saigusa, Y., Kusumoto, B., & Kubota, Y. (2020). Sampling bias correction in species distribution models by quasi-linear Poisson point process. *Ecological Informatics*, *55*(October 2019), 101015. https://doi.org/10.1016/j.ecoinf.2019.101015

Malinowski, R., Lewiński, S., Rybicki, M., Gromny, E., Jenerowicz, M., Krupiński, M., Nowakowski, A., Wojtkowski, C., Krupiński, M., Krätzschmar, E., & Schauer, P. (2020). Automated Production of a Land Cover/Use Map of Europe Based on Sentinel-2 Imagery. *Remote Sensing*, *12*(21), 3523. https://doi.org/10.3390/rs12213523

McCullagh, P., & Nelder, J. A. (2019). *Generalized Linear Models*. Routledge. https://doi.org/10.1201/9780203753736

Messmer, T. A. (2000). The emergence of human ± wildlife con ¯ ict management : turning challenges into opportunities. *International Biodeterioration & Biodegradation*, *45*, 97–102.

Moeller, A. K., Lukacs, P. M., & Horne, J. S. (2018). Three novel methods to estimate abundance of unmarked animals using remote cameras. *Ecosphere*, *9*(8). https://doi.org/10.1002/ecs2.2331

Nakashima, Y., Fukasawa, K., & Samejima, H. (2018). Estimating animal density without individual recognition using information derivable exclusively from camera traps. *Journal of Applied Ecology*, *55*(2),

735–744. https://doi.org/10.1111/1365-2664.13059

Overijssel. (2021). https://www.visitholland.nl/overijssel

Pereira, Ferrier, Walters, Geller, Jongman, Scholes, Bruford, Brummitt, Butchart, Cardoso, Coops, Dulloo, Faith, Freyhof, Gregory, Heip, Höft, Hurtt, Jetz, … Wegmann. (2013). Essential Biodiversity Variables. *Science*, *339*(6117), 277–278.

Philips, Anderson, & Schapire. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197. https://doi.org/10.1890/07-2153.1

Reddy, S., & Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, *30*(11), 1719–1727. https://doi.org/10.1046/j.1365-2699.2003.00946.x

Rocha, A. D., Groen, T. A., Skidmore, A. K., & Willemen, L. (2020). *Dependent Ecological Data With Remote Sensing*. *59*(1), 1–12.

Saunders, G., & Kay, B. (1991). Movements of feral pigs (Sus scrofa) at sunny corner, New South Wales. *Wildlife Research*, *18*(1), 49–61. https://doi.org/10.1071/WR9910049

Stockwell, D. R. B., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, *148*(1), 1–13. https://doi.org/10.1016/S0304-3800(01)00388-X

The Royal Netherlands Meteorological Institute (KNMI). (2020). *Twenthe, long-term averages, time period 1981-2010 (in Dutch)*. The Royal Netherlands Meteorological Institute (KNMI). http://www.klimaatatlas.nl/tabel/stationsdata/klimtab_8110_290.pdf

APPENDICES

Appendix 1: Conditions of sharing NDFF data

Terms of use for the provision of NDFF natural data for educational purposes

NDFF
Toernooiveld 1
(gebouw Mercator III)
6525 ED Nijmegen

NATIONALE DATABANK
FLORA EN FAUNA

I, the undersigned, on behalf of University of Twente, hereinafter referred to as 'User', declare signing this form to agree to the following terms, including the data provided by the NDFF:

1. NDFF grants User the right to use the files in question.
2. User may use the natural data exclusively for the project purpose as mentioned in the tender Details from the National Database Flora and Fauna according to 33050_Wild_Boar_Overijssel and any subsequent publication.
3. User may publish or otherwise disclose the natural data in an aggregated manner (1x1 km), unless otherwise agreed in writing with NDFF.
4. Use of Nature Data for other purposes, including all business or other business activities of the User, is expressly prohibited unless written permission has been granted by NDFF.
5. It is not a User's right to transfer to third parties the rights and obligations arising from this disclosure, NDFF.
6. User will not provide or give access to third parties, with or without permission from NDFF, the aforementioned data files or parts thereof, whether or not added to or enriched with other information.
7. NDFF reserves the right to change the original data files. NDFF is not obliged to provide these modified files to User.
8. NDFF can never be held liable for damage resulting from or related to the use and / or interpretation of the data provided (files).
9. When publishing results obtained on the basis of the data files provided, a clear source entry to the NDFF, including date of requests from the NDFF, must be mentioned. User is responsible for sending a copy (pdf) of the relevant publication to NDFF.
10. Upon termination of the above-described purpose, the provided data files, whether or not added or enriched with other information, must be destroyed by User.
11. In case of breach of one or more of the above conditions, the User shall pay an immediately payable fine of € 1,000, - or so much more if the business benefit is that the violation has been achieved. The forfeiture of the fine referred to here does not affect the right of the NDFF to claim full compensation in this regard.

User data: **Signature:**

Place and date: 3/12/2020 Enschede
Name signatory: Douglas Gichangi
Employment: Student
Organization: University of Twente
Address: Hengelosestraat 99, 7514
Plaats: Box 14
Phone: +31685172333
Email: d.m.gichangi@student.utwente.nl