FILLING IN THE GAPS OF THE NDFF: A MODEL-BASED APPROACH FOR CORRECTING OBSERVER-BIAS IN SPECIES DISTRIBUTION MODELLING

LUCAS GRAAS

A MASTER'S THESIS FOR THE APPLIED DATA SCIENCE PROGRAM

SUPERVISORS:
DR. AD FEELDERS
DR. MENNO STRAATSMA

IN COOPERATION WITH TAUW B.V.

Abstract

Biodiversity is essential for ecological balance and ecosystem functioning. The National Database of Flora and Fauna (NDFF) is a valuable resource for understanding biodiversity, but it suffers from observer bias, resulting in data gaps and incomplete knowledge about species distribution. To address this issue, this study employs Species Distribution Modelling (SDM) for the Plebajus Argus in the Utrecht and Veluwe regions by combining existing species occurrence data with relevant environmental variables. A probable cause for observer bias in the NDFF are differences in accessibility and visitation rates across the study area. Therefore, an additional set of 'observer bias variables' that describe these factors were added and their efficacy assessed. By fixing these variables to a constant when predicting, all locations are treated as if they have equal accessibility and consequently a distribution free from observer bias is predicted. Results show that a Logistic Regression, Random Forest and KNN predict the distribution of the Plebajus Argus with an accuracy of around 92%. Particularly the Logistic Regression and Random Forest show sensible corrections for observer bias.

Table of contents

Introduction

Biodiversity plays a crucial role in maintaining the ecological balance and functioning of ecosystems. The Netherlands, like many other regions, has experienced a decline in biodiversity over the past century. Biodiversity in the Netherlands lowered from well above 40% MSA (Mean Species Abundance) in 1900 to about 15% in 2000 (CBS, 2013). In comparison with Europe and the rest of the world this loss is notably more substantial. Efforts have been made in the last decade to slow down the loss of biodiversity by constructing new natural areas, and while these measures are promising, it is essential to have accurate information on the current state of biodiversity and the factors contributing to its preservation.

The National Database of Flora and Fauna (NDFF) is a valuable resource that collects observations of species in the Netherlands and is often used to provide insights to policy makers on the status of biodiversity. The NDFF is an aggregate of multiple sources, including citizen science initiatives, professional record centers and expert observations. However, due to differences in accessibility and research focus, some areas of the Netherlands have received more attention from these sources and therefore have a higher density of recordings. This phenomenon is known as “observer bias” and can impact research particularly with presence-only data (Chauvier et al., 2021). This uneven documentation of occurrences across different locations leads to data gaps and incomplete knowledge about the distribution of species.

To address the limitations imposed by observer bias, researchers and conservationists are employing species distribution modeling (SDM) where existing species occurrence from well documented regions are combined with relevant environmental variables such as climate, soil type and vegetation cover to estimate the occurrence of species to lesser documented regions. It has been shown that adding variables that try to account for observer bias, such as proximity to population clusters and roads, can improve the accuracy of a SDM (Warton et al., 2013). The underpinning idea in this approach is fixing these observer bias variables to a chosen value, rather than using the real observed values. With these fixed values, predictions can be made as if all locations had the same accessibility. An intuitive value is ‘0’, where all locations are treated as if their accessibility was ‘perfect’.

In collaboration with TAUW, a European consulting and engineering firm with a strong position in environmental consulting and sustainable development of the living environment, this thesis will focus on investigating the effectiveness of using a model-based approach to account for observer bias in combination with SDM to improve the insights that the NDFF offers on species distribution and therefore biodiversity. Specifically, the presence of the Plebejus Argus, commonly known as the ‘heideblauwtje’ in Dutch, will be modeled in the area of Utrecht and the Veluwe. The ‘heideblauwtje’ is a butterfly species which is not heavily reliant on a specific host plant making it well suited for modeling based on environmental characteristics.

In Chapter 2 the data and data preparation methods will be discussed, in Chapter 3 the methods of analysis will be discussed, Chapter 4 presents the results and analysis, in Chapter 5 and 6 the discussion and conclusion are found, respectively.

Data

Study Area Description

Though the NDFF offers recordings for the whole of the Netherlands, only a small portion of the Netherlands will be included in the model. The model serves as a proof of concept and the study area can be extended when computational costs and time are not constrained. All datasets used in this study are publicly available for the whole of the Netherlands. The study area covers an area of 2000 km^2 around the municipalities Utrecht, Utrechtse Heuvelrug and Ede, and includes a broad range of environmental conditions such as urban areas, wetlands, waterbodies, agriculture and grasslands, forests, and heather. The wide range of environmental conditions ensure that many different habitat types are included in the study area. See Figure 3 for a map of the study area.

Dutch National Database Flora and Fauna (NDFF)

The NDFF is the data warehouse that collects observations of plant and animal species in the Netherlands. It is used to provide the target variable in this model. The NDFF has more than 1.5 million observations, of which 10.446 are observations of the Heideblauwtje in the study area. The observations in the NDFF are collected via citizen science projects, professional record centers and expert observations. All observations reported to the NDFF are automatically validated. This means that the observation is checked on species, count, date and locations. When complete a substantive test is conducted. For example: an observation of a butterfly in December while this species only flies in August is marked as unreliable. Around 10% of the incoming recordings are marked as unreliable and are then checked manually by a validation team that decides on its acceptance (*ndff.nl*). The records often follow some protocol such that observations are easily comparable, and that the origin of the record is traceable. The frequency of observations in the study area along with their protocol are found in Figure 1, it shows that most of the records are individual observations from different organizations.

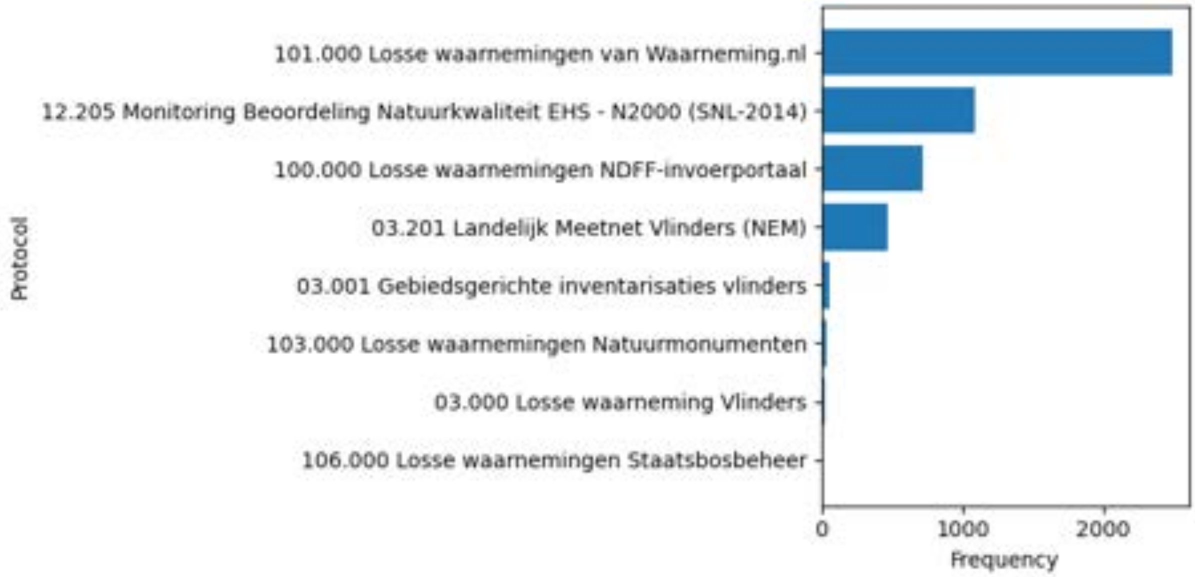

Figure 1 Frequency of observation per protocol: 101.00, 100.000 and 03.201 are protocols for opportunistic records from different organizations. Protocol 12.205 is a professional survey to monitor the quality of nature. (*ndff.nl*)

According to Tisja Daggers, an ecologist at TAUW, records older than 10 years are not likely to be relevant for predicting the current distribution of a species. Since there are no records of species in the NDFF of 2023, only records between 2012 and the end of 2022 are included, resulting in 5338 observations. However, it is worth noting that out of these 5338 observations, 403 were collected on an area larger than the resolution of the model and are therefore not included in the dataset for analysis. The remaining usable records are plotted in Figure 3.

The data for this period and area are claimed to represent precise counts. However, Figure 2 illustrates the distribution derived from these counts. The unusual spikes at 10 or multiples of ten are likely the effect of the 'round number bias', whereby individuals tend to round their estimates towards 10 or multiples of 10. This shows that the claimed precision is not always true. The distribution shows that nearly half of the recorded instances correspond to a count of 1. The frequency of observations where more than one Heideblauwtje is observed quickly diminish with a higher number of observed butterflies.

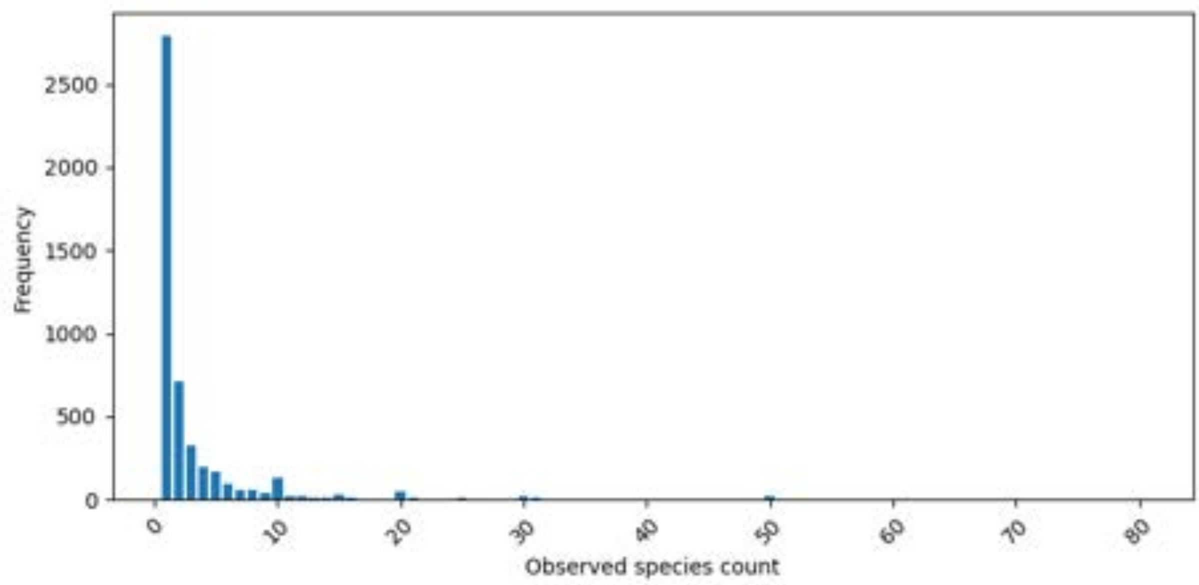

Figure 2 Distribution of observed count of the Heideblauwtje in the NDFF

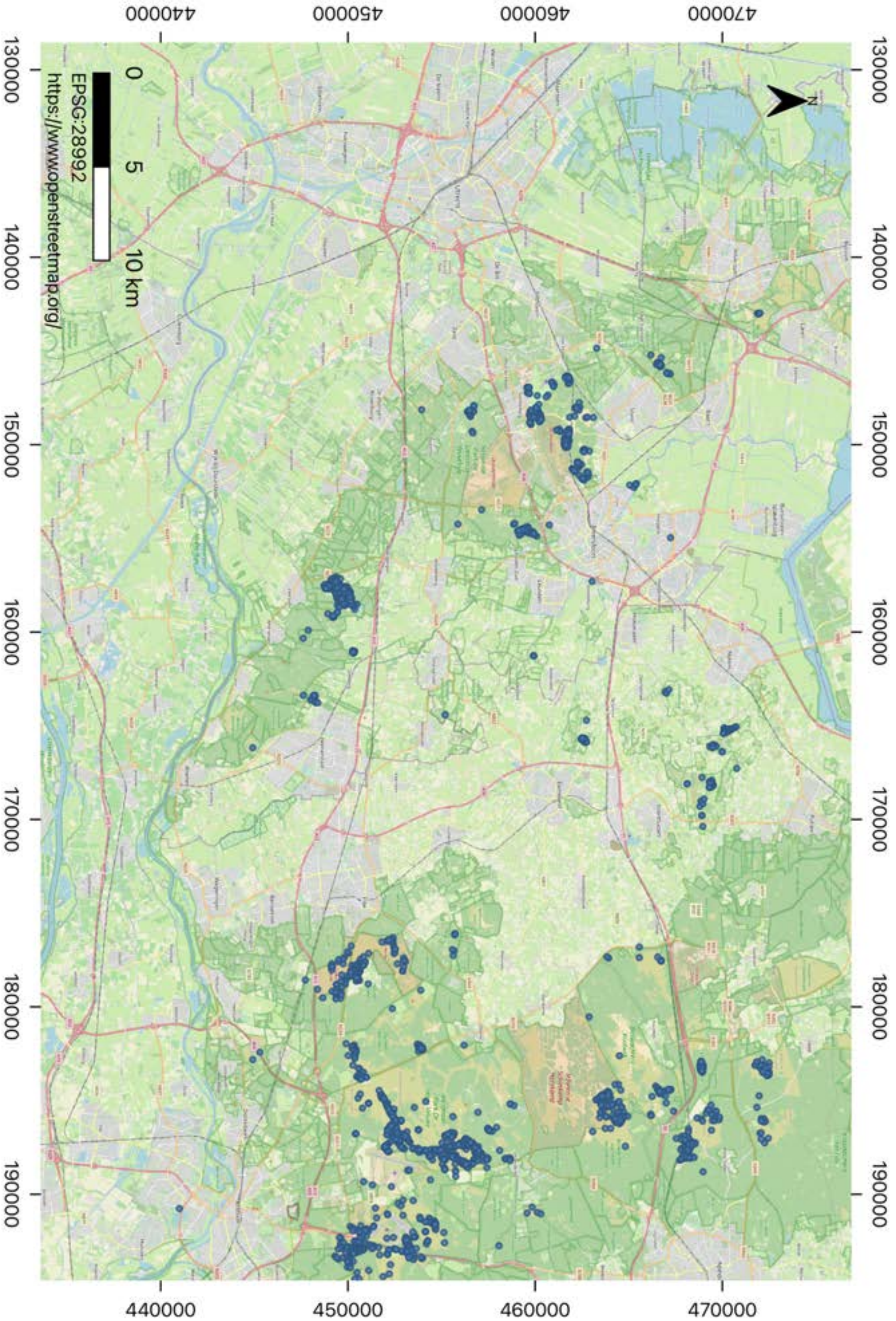

Figure 3 Locations of observations, marked by blue dots for the Heideblauwtje in the study area between 2012-2022.

Environmental data

The predictive variables in this model can be subdivided into two categories. First the environmental variables that try to describe a species habitat, and secondly variables that try to model the accessibility for individuals to account for observer bias in the NDFF.

The concept of a species' ecological niche provides the central theoretical basis for describing species-environment relationships in SDM (Miller, 2010). A species' niche, in context of an SDM, can best be described as an 'n-dimensional hypervolume' in environmental space in which a species can exist. This environmental space is constructed using the datasets in the following sections. The full list of data sources is added to reference section, while Appendix 1 presents the full list of features and cell count derived from the following datasets.

Agrarisch Areaal Nederland

The Agrarisch Areaal Nederland (AAN) dataset was obtained from PDOK (Public Services On the Map), and it provides spatial information on agricultural areas, permanent meadow, or the cultivation of perennial crops. The data were built using RGB remote sensing techniques.

Basisregistratie Gewaspercelen

The Basisregistratie Gewaspercelen dataset, sourced from PDOK contains geospatial information on crop fields in the Netherlands. The dataset is constructed using the previously mentioned Agrarisch Areaal Nederland, where the areas have been categorized in 29 types of crops. The user of the plots must declare yearly which crop they will cultivate. Each year a dataset is generated on the 15th of May. The features derived from this dataset can be found in Appendix 1 with the prefix 'BRP_gewas_'.

BRO Geomorfologische kaart

The Geomorfologische kaart (GMM) dataset, obtained from PDOK is part of the Basisregistratie Ondergrond (BRO). It provides two categorical maps for this analysis: BRO_genese and BRO_landform. For both variables, the derived features and their cell count are found in Appendix 1 with prefix 'BRO_genese_' and 'BRO_landform_'. The BRO_genese is an aggregate of the BRO_landform, based on similarities in the processes that created the landforms.

CBS Bestand Bodemgebruik

CBS (Centraal bureau voor de Statistiek) is a well-known provider of statistics and data in the Netherlands. The data was sourced from PDOK. The 'Bestand Bodemgebruik' is a comprehensive land-use dataset and includes detailed information in the Netherlands. The full list of features and cell count is found with prefix 'cbs_landuse_' in Appendix 1.

Fysisch Geografische Regios

The dataset Fysisch Geografische Regios was downloaded from PDOK and defines the physical geographical regions in the Netherlands. These regions are characterized by specific combinations of landforms, geology, climate and soil properties. The regions and their cell count are found in Appendix 1.

Nationale Parken

The Nationale Parken dataset was sourced from PDOK and contains a binary map demarcating the borders of the 20 national parks in the Netherlands. These parks are often characterized by significant amounts of natural landscapes and biodiversity. National parks and protected areas are often established to conserve and protect important habitats for various species. As a result, these areas can serve as proxies or indicators for the presence of certain species. The dataset and the cell count can be found in Appendix 1.

Natura2000

The Natura2000 was obtained from PDOK and is a dataset containing the spatial information of a network of protected areas across Europe. The network consists of 160 areas that aim to conserve and protect Europe’s most valuable and threatened species and habitats. The Natura2000 areas are protected under the ‘Nature Conservancy Act’ and are defined by the European Birds Directive and Habitats Directive. These areas can therefore serve as proxies for the distribution of species. The cell count can be found in Appendix 1.

SGM Ondergrondmodel

The SGM soil map was at acquired at PDOK and contains detailed information on the soil type in the Netherlands. It describes the soil up to a depth of 1.2 meters, and is created at a scale of 1:50,000. It includes data on soil type and deposition method, soil formation, composition as well as the calcium content. It contains 96 categories which can be found with prefix ‘SGM_ondergrond_’ in Appendix 1 along with their cell count.

Satellite Images

Using the planetary computer by Microsoft, satellite images of the study area were downloaded. The planetary computer provides images made by the Landsat satellites and are publicly available. A maximum cloud cover of 2% was used to minimize noise caused by moisture in clouds. The original resolution of the images is 30x30 meters. With the different bands, covering different parts of the light spectrum, a ‘normalized difference vegetation index’ (NDVI) image was constructed. The NDVI is a value between -1 and 1. Values close to -1 are likely to be water bodies, while values close to +1 are likely to be dense green leaves. Values of 0 are likely to be urban areas. The NDVI is therefore a good indicator for nature areas.

Stiltegebieden

The stiltegebieden dataset can be found at PDOK and provides geospatial information on areas where human activities cannot be louder that 40 dB. These areas can possibly be informative on species distributions as these quiet zones can be indicative for the absence of human activity, making it more suitable for certain species. The cell count can be found in Appendix 1.

Corine Land Cover

The Corine Land Cover dataset, version of 2018, sourced from the Copernicus website and is a widely used and comprehensive land cover classification system developed by the European Environment Agency (EEA). It provides detailed information about the different land cover types and their spatial distribution. After cleaning this dataset has 23 categories which can be found in Appendix 1 using prefix ‘clc2018’.

Wetlands

The Wetlands dataset, sourced from PDOK, contains geospatial information on wet areas in the Netherlands. These wet areas include different water types such as marshes, fens, peat, lake areas and stagnant or flowing water bodies including artificial water bodies. A total cell count is found in Appendix 1.

Observer bias variables

Observer bias refers to the potential distortion or influence of human observers on the data they collect, which can introduce biases in species occurrence records. An excellent example of observer bias found in the NDFF is given in Figure 4. This figure illustrates that observations of the Heideblauwtje show a high proximity to roads and that accessibility to an area is a probable cause. To address this bias and enhance the accuracy of SDM, multiple features that describe the accessibility and visitation rate to an area are incorporated. These data source will be discussed in the following sections.

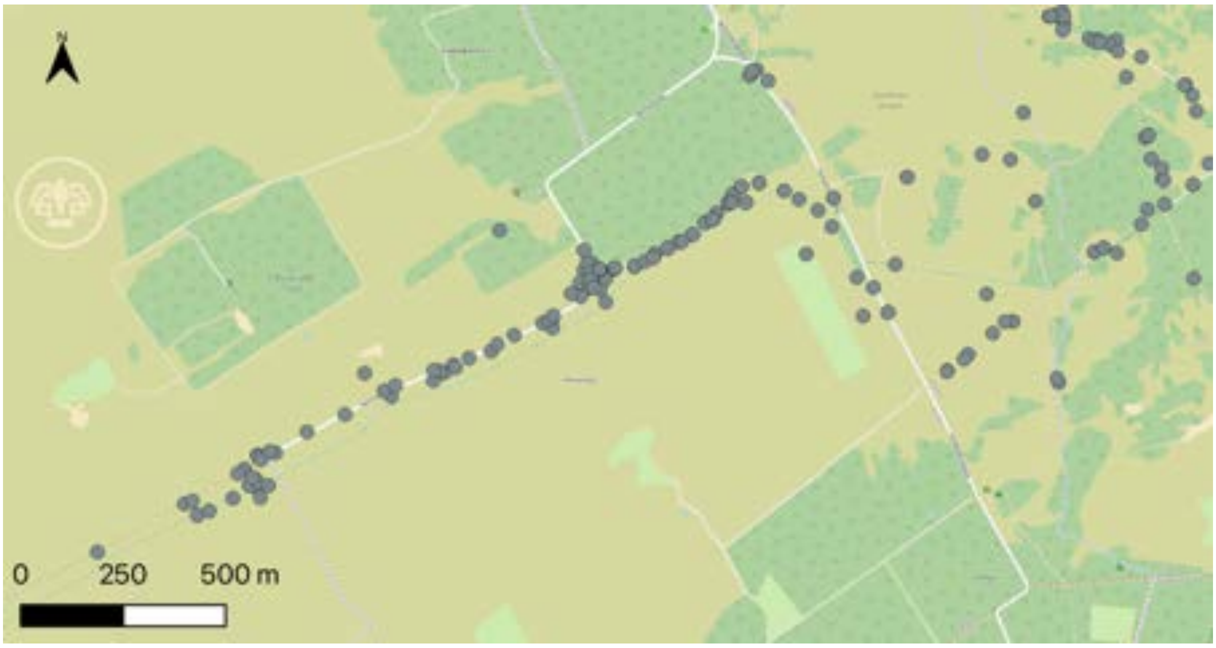

Figure 4 Observations of the Heideblauwtje showing a significant proximity to a nearby road.
(Base map from OpenStreetMap and OpenStreetMap Foundation (CC-BY-SA). © https://www.openstreemap.org and contributors.)

Roads

All highways, roads and hiking paths are queried from OpenStreetMap (OSM) resulting in a dataset of 20533 road sections in the study area. Combined they sum to a total length of around 3100 km. This data has been used to derive two variables for our analysis: a variable containing the distance to the closest road section and a variable containing a density of roads per raster cell.

Population

Another source of data that could account for observer bias in occurrence records of the NDFF is population. A population distribution dataset was acquired PDOK. This dataset was used to include 3 variables to our analysis: raw population distribution, a binary variable indicating population clusters above 1500 individuals per square kilometer, and a variable containing the distance to the closest population cluster.

Data Acquisition and Data Preparation

The preparation of data is a critical step in any data analysis or research project. In this section, we describe the operation undertaken to acquire and prepare the dataset for analysis. The original data includes both vector and raster formats and their contents and sources have been described in the 'Data' section. The desired format for the analysis-ready dataset is a CSV (Comma Separated Values) file, where each row corresponds to a specific cell in the raster datasets and each column represents a feature extracted from a dataset. Additionally, two columns with 'x' and 'y' values have been added, corresponding to the coordinates of the center of the cell. The abundance values of the NDFF records, have been included for each year separately, as well as a total.

When combining different datasets, it is essential that the raster cells of the different datasets are identical. To achieve this, a template raster is needed from which the resolution, coordinate reference system, cell origin and extent can be copied. In this study, we choose to project all data in the Amersfoort / RD new (EPSG:28992) coordinate reference system, the most used projection for Dutch datasets. The bounding box coordinates of the study area are found in Table 1. The resolution and cell origin were chosen to be copied from the reprojected CORINE Land Cover datasets (93x106 m), such that no resampling method was necessary for this dataset and thus no information was lost.

		Coordinate
North	Max	471634,8070
	Min	438840,1989
East	Max	193030,5756
	Min	130129,1596

Table 1 The table provides the coordinates of the study area's extent, representing the boundaries of the analyzed region. Coordinates are given in EPSG:28992.

Reprojecting

The first step to make all data analysis-ready is reprojecting the datasets to the same CRS. Most datasets were obtained from a Dutch source and were already in the correct CRS (EPSG:28992). Reprojecting and clipping to extent were done with QGIS 3.28.1. The clipping operation was done using the coordinates from Table 1.

Rasterizing

Rasterizing vector data is necessary to be able to store the data as a CSV. The 'Rasterize (vector to raster)' tool by GDAL was used for this operation. This tool assigns each cell the value found at its center. Since the file format used for rasters (.tif) does not support categories, all datasets containing categories were separated and each category was rasterized to a binary map using zero's and one's to indicate the absence or presence of that category. These binary maps were than stacked such that, for example, the dataset 'Basisregistratie Gewaspercelen' with 29 categories was rasterized to a TIFF file with 29 bands, each band containing a binary map for a particular category. When rasterizing the observations from the NDFF, the same operation was used, however this time the sum of all observations in the cell extent was taken.

Aligning

Since all rasters are clipped to the same extent and then rasterized to the same resolution, it is expected that they have the same cell origin. However, after inspection, it seems that cell centers are very slightly misaligned. Using the QGIS tool 'Align Rasters' the cells have been aligned perfectly through resampling using a 'Nearest Neighbour' algorithm. A 'Nearest Neighbor' algorithm was chosen primarily because of its preservation of the original values, where other methods often use some averaging or interpolation. The 'Nearest Neighbor' algorithm selects the value of the closest pixel in the source raster when determining the value of a new pixel in the resampled raster. Given that the source and resampled raster cells are only very slightly misaligned, this ensures that the new resampled raster contains the original values in their corresponding positions.

Combining data to CSV

To obtain a CSV from a raster file, a python function has been written. It loops over every TIF file in a folder and stores each band as a column in a pandas dataframe where each row is a specific cell. The coordinates of the cell center are added as separate columns ‘x’ and ‘y’. In case of a categorical variable, when the number of bands is larger than 1, the columns with binary values are combined to a single categorical variable. When all datasets are converted to csv another python code combines these datasets to one final CSV.

Methods

The main objective of this thesis is to investigate the effectiveness of integrating a model-based approach to address observer bias in conjunction with SDM. Consequently, the aim is to enhance the quality of insights provided by the NDFF, and to demonstrate a framework for predicting the spatial distribution of a species, while correcting for observer bias. To accomplish this, the presence and absence of the 'Heideblauwtje' will be predicted using a variety of machine learning algorithms. The spatial distribution of the 'Heideblauwtje' will be modelled using environmental variables and complementary to these variables, several 'observer bias variables' will be added. These variables aim to describe the spatial variation in observer bias, following a similar approach as proposed by (Warton et al., 2013). To get a true distribution of the 'Heideblauwtje', where observation bias is corrected for, the observer bias variables can be set to a fixed value. An intuitive example is setting the variable 'distance to paths' to 0, such that the predictions can be interpreted as the areas where the 'Heideblauwtje' would be observed when these areas would be perfectly accessible by paths. In this study we will use the mean of the observed accessibility variables as the fixed value. This way, locations where the real accessibility is very high, for example in urban areas, get assigned a lower value than their true accessibility, while inaccessible locations get assigned a higher value. The underlying idea is that a 'Heideblauwtje' would have been observed in highly accessible areas if it occurred there, and that assigning a lower value would decrease the probability of a positive prediction. Likewise, assigning a higher value to lesser accessible locations would increase the probability of an observation of a 'Heideblauwtje', corresponding to the idea that an observation is more likely if an area is more accessible.

Presence-only data

Though the NDFF collects absence data, the target variable for this study only contains records of the presence of the 'Heideblauwtje'. This leaves us with what is known as a One-Class Classification (OCC) problem. While there are various models that work with this type of data (Khan & Madden, 2014), a common method in the field of SDM is to artificially generate absence records.

To address this issue, we deploy an approach known as generating 'pseudo-absences', as described by (Barbet-Massin et al., 2012). Pseudo-absences are locations where there are no observations of the species and are handled as if they are true observed absences. In this project we created a balanced dataset by randomly selecting an equal number of pseudo-absences to match the number of presence records. To achieve this, locations where there are no presence records are randomly sampled in the study area. We will handle these randomly selected cells as if they were true absence records.

One-Hot-Encoding

Since many machine learning algorithms are not able to use categorical variables, all categorical variables have been one-hot-encoded. This creates a separate binary variable for each of the categories found in the original variable. To prevent perfect multicollinearity one of the encoded variables must be removed. The choice of which category to be removed does not impact the performance but does affect the interpretation of coefficients as the removed category acts as a baseline to which the other encoded variables are compared. Therefore, when a categorical variable contained a category in the likes of 'Other' or 'No category' it was chosen to be deleted. With the variables that did not have such a category, the most common category was deleted.

Standardization

Although standardizing continuous variables does not affect the statistical inference, centering and scaling allows for easy comparison of the model coefficients and therefore feature importance. Furthermore, it can improve model stability and convergence. In this study all continuous variables were standardized using z-scaling. This process involves transforming each value by subtracting the mean and dividing by the standard deviation.

Multicollinearity

The general approach for data collection in this thesis is collecting everything that could possibly be informative for prediction of species occurrence, and then using the machine learning algorithms to find the most important features. Following this approach will yield datasets that contain the same information, i.e., variables that are correlated with each other. When using a model only for prediction, multicollinearity is not so much of a problem. When the objective is also inference, e.g., the analysis of a models' coefficients and other trends, multicollinearity becomes a problem. When two variables are correlated it becomes difficult to determine the individual effect of each variable on the response variable as the predictors tend to change in unison. In an effort to reduce multicollinearity, some features have been deleted based on their Pearson correlation. For example: when predictors X1 and X2 are linearly correlated with a Pearson correlation larger than 0.65, one of these predictors is removed. The choice of which predictor to remove is based on its correlation with the target variable, where the predictor with smaller correlation is removed. The removed variables are found in Appendix 5.

Model selection and Parameter tuning

The goal of our study is to predict an unbiased distribution of species. This can be done by modelling abundance, an actual number of 'Heideblauwtjes' at a location, or by occurrence where only the presence of absence of a 'Heideblauwtje' is predicted. Despite the potential for abundance modeling to provide valuable insights to ecologists and policymakers, the exploratory data analysis using regression models yielded very discouraging results. Either the environmental variables are not informative enough to predict a count, or the NDFF data contains too much noise to predict a reliable abundance. Therefore, our target variable will be binary: per cell in the study area, we will predict the presence or absence of the 'Heideblauwtje'. The performance based on the test set will lead to the choice of final model. We will consider the following machine learning models: a Logistic Regression, Random Forest and K-Nearest Neighbors (KNN). The logistic regression is mainly added for its interpretability, while the other models are added for their possible predictive power. With the use of a grid search, the best parameters for each model are found by iteratively refining the grid guided by performance metrics. Table 2 shows the parameters used in grid search for each of the models and Appendix 2Appendix 3 and Appendix 4 show the different combination of the top 40 best combination of parameters.

Model	Parameters	Best value
Logistic Regression	C	0.5
	Penalty	L1
Random Forest	N_estimators	400
	Max_features	Log2(total number of features)
	Max_depth	300
	Min_samples_split	4
	Bootstrap	True
KNN	N_neighbors	7
	Weights	Distance
	Metric	Manhattan

Table 2 Parameters used in grid search.

Variable importance

For Logistic regression, variable or feature importance is easily extracted from the model. Since continuous variables have been standardized, the absolute value of the coefficient of each variable can be interpreted as its importance for the model. A random forest does not have coefficients and feature importance is based on the mean and standard deviation of the accumulated decrease in impurity per tree for each feature in a forest. KNN, a distance-based method, does not have coefficients or impurity and its feature importance is approximated using ‘permutation importance’. This approach approximates the importance of a variable by comparing the performance of an original model with a model where the values of the variable are randomly shuffled. For each of the features this was done 10 times to cancel out the chance that a feature has been shuffled to a state that happens to be contributing positively to the model.

Spatial cross validation

When working with spatial data it is known that using traditional Cross Validation (CV) will yield optimistically biased performance metrics. Normal CV randomly splits the data into ‘k’ folds using one of these folds for evaluation, while the other folds are used for training the model. However, when handling spatial data, near values are more related than distant values, which is known as spatial autocorrelation. Because of this phenomenon, features are not randomly distributed over space, and can therefore also not be sampled randomly as a randomly sampled observation in the training set is correlated with a randomly sampled observation in the test set, when these observations are close to each other in geographical space. A solution to this problem is to use spatial cross validation (SCV), where folds are not generated by random sampling, but rather created as all records in a distinct area. This way, when predicting in the test fold, we are sure that the model has not ‘seen’ observations from this area and will thus create a true estimate of the predictive performance in areas outside of the training area. However, since this thesis focusses on filling the data gaps in the NDFF, i.e., there will be no predicting outside the training area, the data leakage from normal CV is not such a problem. It could even be beneficial as the model learning from the nearest areas around a data gap would help fill in the gap more accurately. Since the method of this thesis is generalizable to the whole of the Netherlands and the NDFF data is found everywhere in the Netherlands, predictions will always be somewhere near observations and thus can regular CV be used as model evaluation.

Results and analysis

This section will describe the performance, variable importance and mapped predictions twice for each of the models: once for the dataset without the observer bias variables and once for the dataset with observer bias variables.

Model performance

The performance of a model is often described by its accuracy, precision and recall. These metrics are often some ratio of the correct or incorrect predictions of the positive or negative class. A table containing all sorts of classifications is called a 'confusion matrix' of which an example is given below.

		Predicted Condition	
		Positive	Negative
Actual Condition	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positve (FP)	True Negative (TN)

Table 3 Confusion Matrix

In context of this study, an analysis where pseudo-negatives were generated, we don't know the actual condition of the 'Negative' class. For example, it could be that a location randomly sampled as negative actually contains the presence of a Heideblauwtje. Therefore, all metrics involving the 'False Positive' or 'True Negative' have some uncertainty and need to be interpreted with some caution. The recall of the positive class, also called sensitivity or true positive rate, is defined as:

$$Sensitivity = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

and does not involve the uncertainty that follows from a 'False Positive' or 'True negative'. Furthermore, the use of random generated pseudo-absences introduces some uncertainty itself. It is very possible that one set of random pseudo-absences, by chance, is more in line with the presence data. This means that the pseudo-absences are not close to known observations, and thus not adding noise to the data. To quantify the effects of different sets of randomly generated pseudo-absences 50 training sets were created, each with the same presence records, but with a different set of randomly sampled pseudo-absences. A training set consists of 1988 records, of which 996 are pseudo-absences. The test set contains 498 records. A model, using the same hyperparameters, was trained on each of these training sets and tested on a test set that did not change. The performance metrics found in Table 4, Table 5 andTable 6 are the mean over these 50 results, accompanied with the standard deviation. The hyperparameters used for these model runs were found using grid search with 5-fold cross validation on one of the training sets. To compare the performance of a model with and without the inclusion of observer bias variables this was all done twice, once for a dataset without observer bias variables and once with the observer bias variables.

	Without observer bias variables		With observer bias variables	
	Mean Value	Standard Deviation	Mean Value	Standard deviation
Precision	0.908354	0.012157	0.925922	0.010875
Recall	0.910560	0.011285	0.928189	0.010155
F1-score	0.908816	0.012200	0.926470	0.010933
Accuracy	0.909197	0.012287	0.926787	0.010985
Sensitivity	0.925991	0.003560	0.944053	0.006182

Table 4 Performance metrics for Logistic Regression

	Without observer bias variables		With observer bias variables	
	Mean Value	Standard Deviation	Mean Value	Standard deviation
Precision	0.920525	0.010214	0.926321	0.009464
Recall	0.922588	0.009542	0.928845	0.008817
F1-score	0.921113	0.010173	0.926937	0.009549
Accuracy	0.921486	0.010217	0.927229	0.009596
Sensitivity	0.935066	0.006589	0.947137	2.242989e-16

Table 5 Performance metrics for Random Forest.

	Without observer bias variables		With observer bias variables	
	Mean Value	Standard Deviation	Mean Value	Standard deviation
Precision	0.904717	0.011988	0.914786	0.010791
Recall	0.907029	0.011483	0.917514	0.010557
F1-score	0.905125	0.012174	0.915033	0.011376
Accuracy	0.905502	0.012234	0.915301	0.011444
Sensitivity	0.924317	0.012234	0.942555	0.008057

Table 6 Performance metrics for KNN.

All models performed better when observer bias variables are included. The largest gain in performance was by Logistic Regression with 1.5 percent point. The Random Forest gained the least when observer bias variables were added, but also showed the smallest standard deviation of all the models over the different training sets. The performance of the three models differs only slightly, the random forest performs best with an accuracy of 0.93. As mentioned in the 'Methods' section, model evaluation was done using normal CV. It is however worth mentioning that when using spatial CV where the study area is separated into six geographically distinct folds, the accuracy drops to around 0.89.

Variable importance

Logistic Regression

For Logistic Regression the importance of features has been approximated using two methods: the interpretation of model coefficients as feature importance (Figure 5) and the effect of variable permutation on accuracy (Figure 6). Figure 5 and Figure 6 include all features used, i.e., features not included in the figure were shrunken to zero due to the lasso penalty.

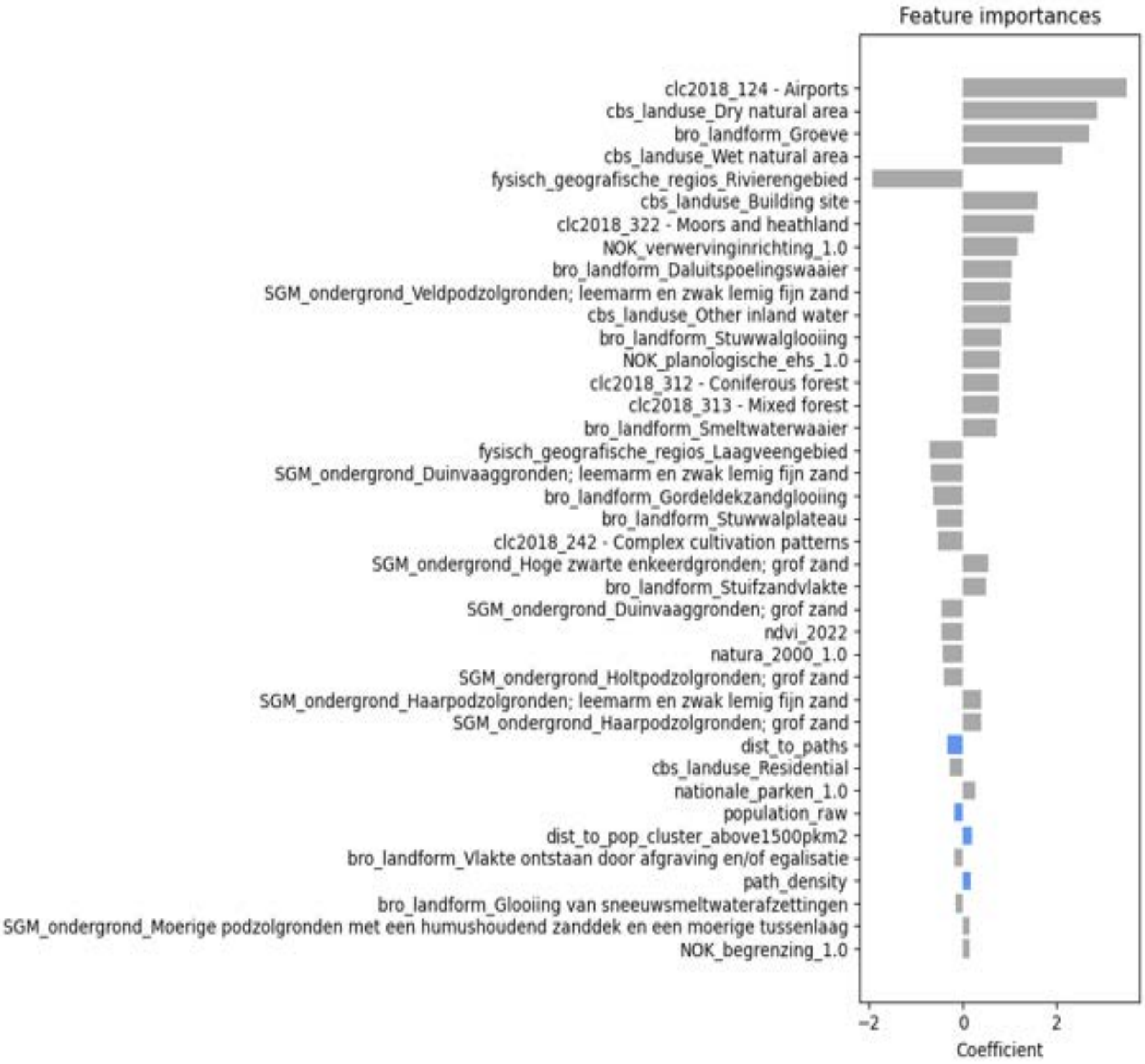

Figure 5 Coefficients of the Logistic Regression. Highlighted in blue are the variables that model observer bias.

The first thing to be noted is that the observer bias variables (blue) are relatively unimportant to the model. Almost all environmental variables (gray) have a larger coefficient. One advantageous aspect of interpreting coefficients as feature importance is that they not only provide information about the magnitude but also offer insights into the direction of the feature's effect on the outcome. Looking at coefficient directions it is remarkable that the features 'cbs_landuse_Dry natural area' and 'cbs_landuse_Wet natural area', seemingly two opposite types of environmental characteristics both contribute positively to the presence of a 'Heideblauwtje'. Furthermore, the feature 'fysich_geografische_regios_Rivierengebied' (areas that cover rivers), where environmental characteristics are expected to be somewhat similar to 'cbs_Landuse_Wet natural Area' have opposite signs. The observer bias variables 'dist_to_path' and 'path_density' have a negative and positive influence on the presence of the 'Heideblauwtje', respectively. This is as expected: an increase in distance to paths result in a decreased possibility that an individual made an observation of the

‘Heideblauwtje’. Similarly, an increase in ‘path_density’ makes it more likely that an individual passing by makes an observation. The effect of the other two observer bias variables is a bit more difficult to interpret. An increase in ‘Population_raw’, the number of individuals living in the cell extent, would decrease the possibility of an observation of the ‘Heideblauwtje’, while an increase in ‘dist_to_pop_cluster_above1500pkm2’ would increase the possibility of observing a ‘Heideblauwtje’. These variables were meant to complement the ‘dist_to_path’ and ‘path_density’ variables, expecting a similar impact on the likelihood of observation of the ‘Heideblauwtje’. That is, the idea that more accessible areas are more likely to have observations. This would mean that an increase in population and a decrease in the distance to population clusters would increase the likelihood of an observation. However, the unexpected “wrong” sign could be explained by the interplay between the distance to population clusters and the environmental characteristics of the area. While greater distances from urban areas tend to reduce the likelihood of an observation, since the area is less likely to be visited by an individual, areas further away from population clusters often possess more favorable environmental conditions. Consequently, although the ‘dist_to_pop_cluster_above1500pkm2’ variable suggests lower visitation rates, the environmental factors associated with greater distance could explain the unexpected sign. Likewise, the variable ‘population_raw’ was added following the idea that higher populated areas suggest higher visitation rates and would therefore increase the likelihood of an observation. However, the unsuitable environmental conditions that come with population areas cause a negative relation with respect to the occurrence of a ‘Heideblauwtje’.

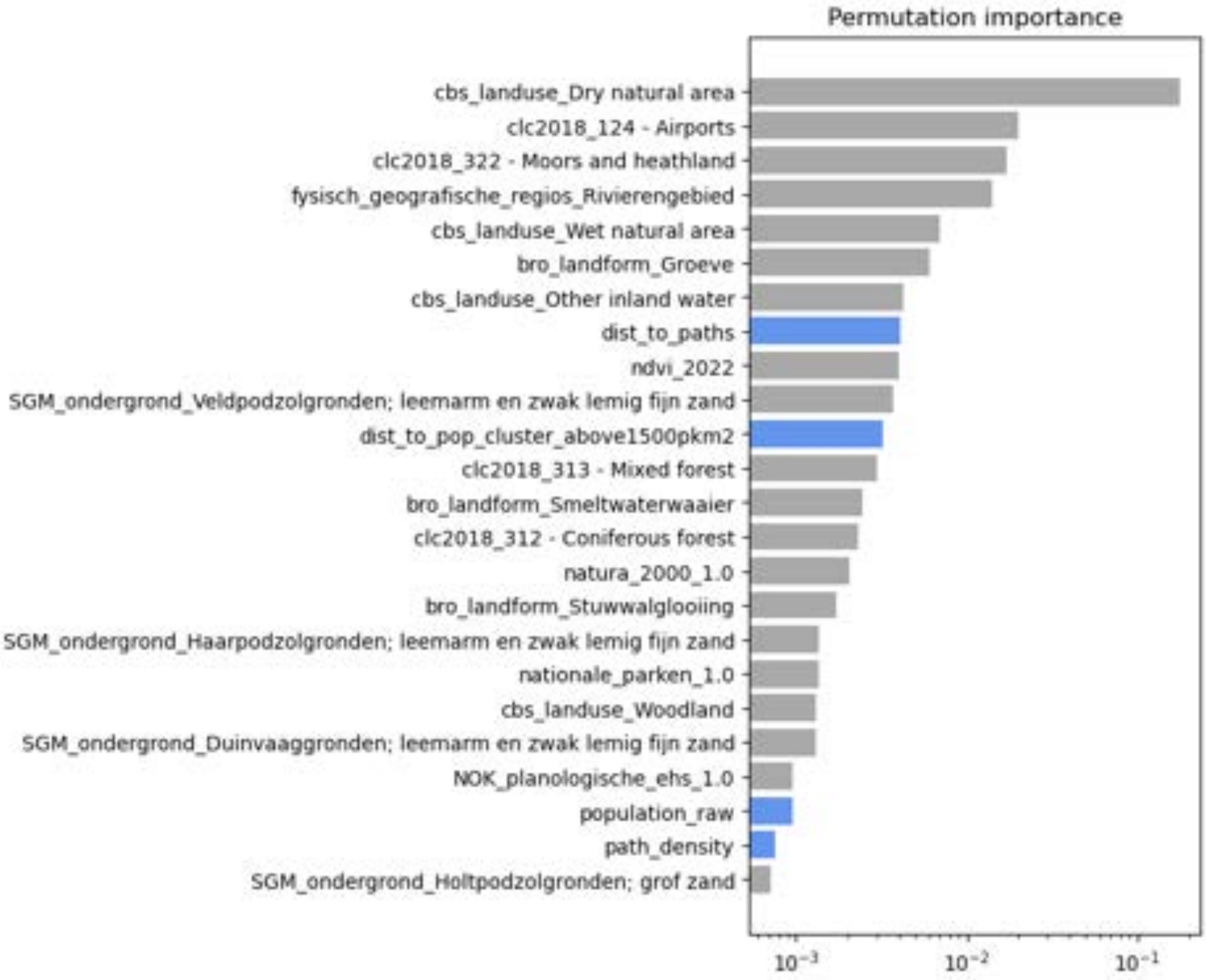

Figure 6 Feature importance for Logistic Regression, calculated using the permutation of variables. Highlighted in blue are the variables that model observer bias. Note that the x-axis is on a logarithmic scale.

The permutation importance is defined as the decrease in model accuracy when a the values of single feature are randomly shuffled. A feature with large decrease in performance is deemed to be more important for the model. Compared to Figure 5, there is almost no change in order for the 5 most important features. However, the approach using permutation marks the ‘cbs_landuse_Dry natural area’ far more important than the other features. Furthermore, the observer bias variables have moved up the list of important variables.

Random Forest

The feature importance for Random Forest is approximated twice using permutation importance (Figure 8) and using the decrease in Gini impurity (Figure 7) when splitting based on a feature. Both methods show that the Random Forest model gives more importance to the observer bias variables compared to the Logistic Regression. Notably, the continuous variables seem to be more important than the binary categorical features.

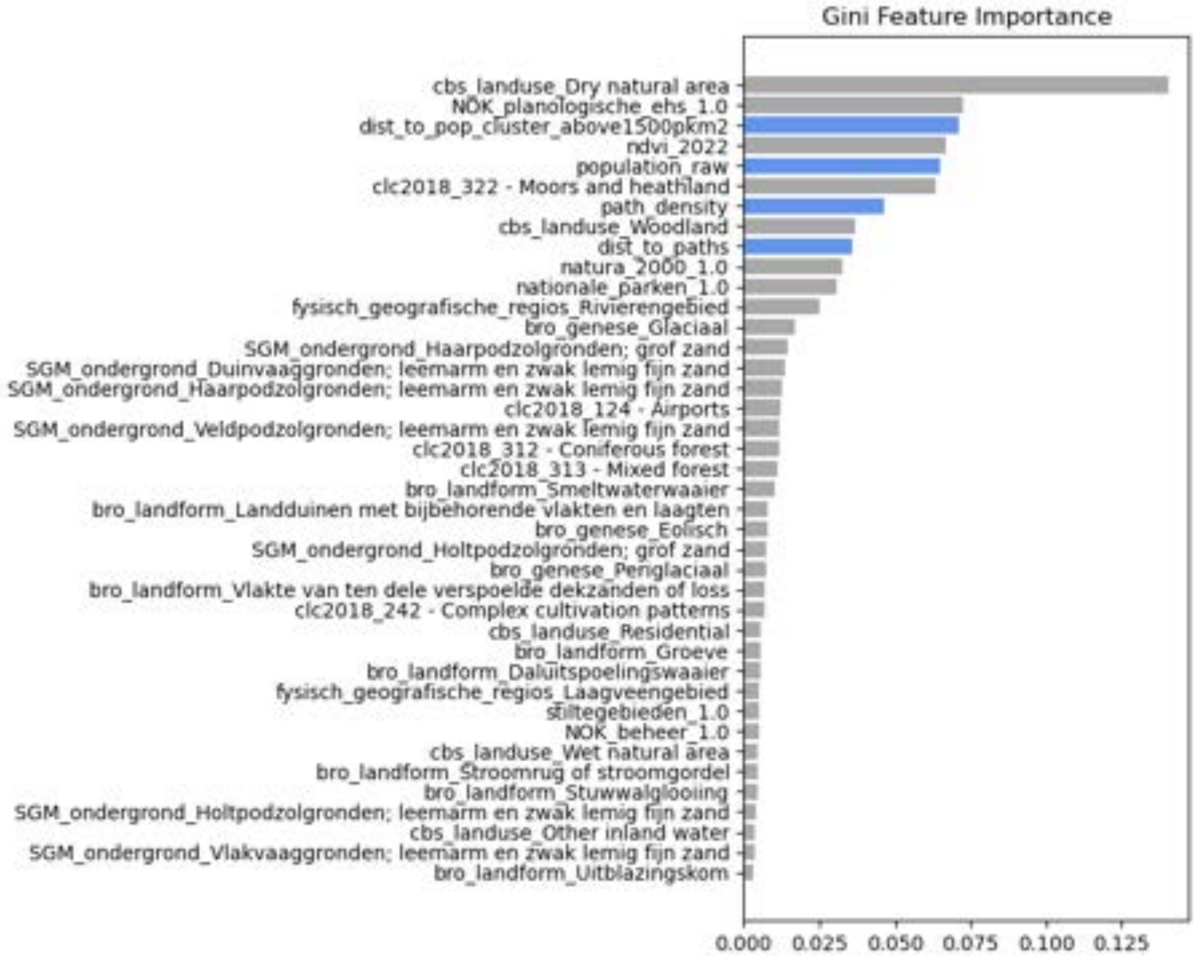

Figure 7 Feature importance for Random Forest with observer bias variables highlighted in blue. Approximated by decrease in Gini impurity of adding the variable.

The feature importance approximation using permutation importance even shows that the ‘path_density’ feature has the most influence on the accuracy of the model. Compared to the Logistic Regression the ‘cbs_landuse_Dry natural area’ is still important but has decreased from 0.15 in Figure 6 to 0.0125 in Figure 8.

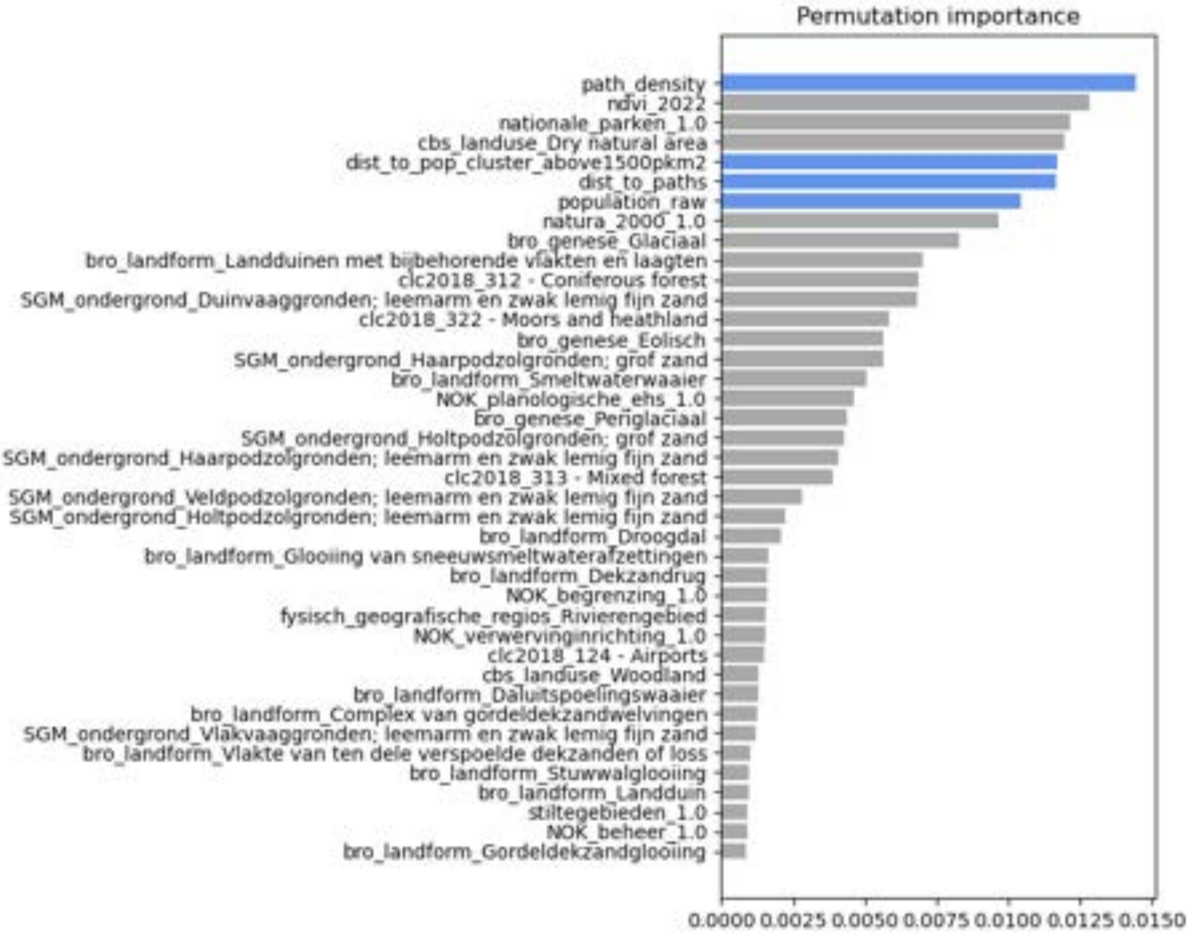

Figure 8 Feature importance for Random Forest with observer variables highlighted in blue. Approximated using permutation of variables.

K Nearest Neighbors

For K-Nearest Neighbors the only method for approximating feature importance is by the permutation of variables Figure 9. Similar as to the Random Forest, the observer bias variables are deemed very important to the model, although the decrease in accuracy is now twice as large for the ‘dist_to_pop_cluster_above1500pkm2’, ‘path_density’ and ‘dist_to_path’. The influence of ‘population_raw’ stays roughly the same in terms of decrease in performance but becomes less import relatively to the other variables.

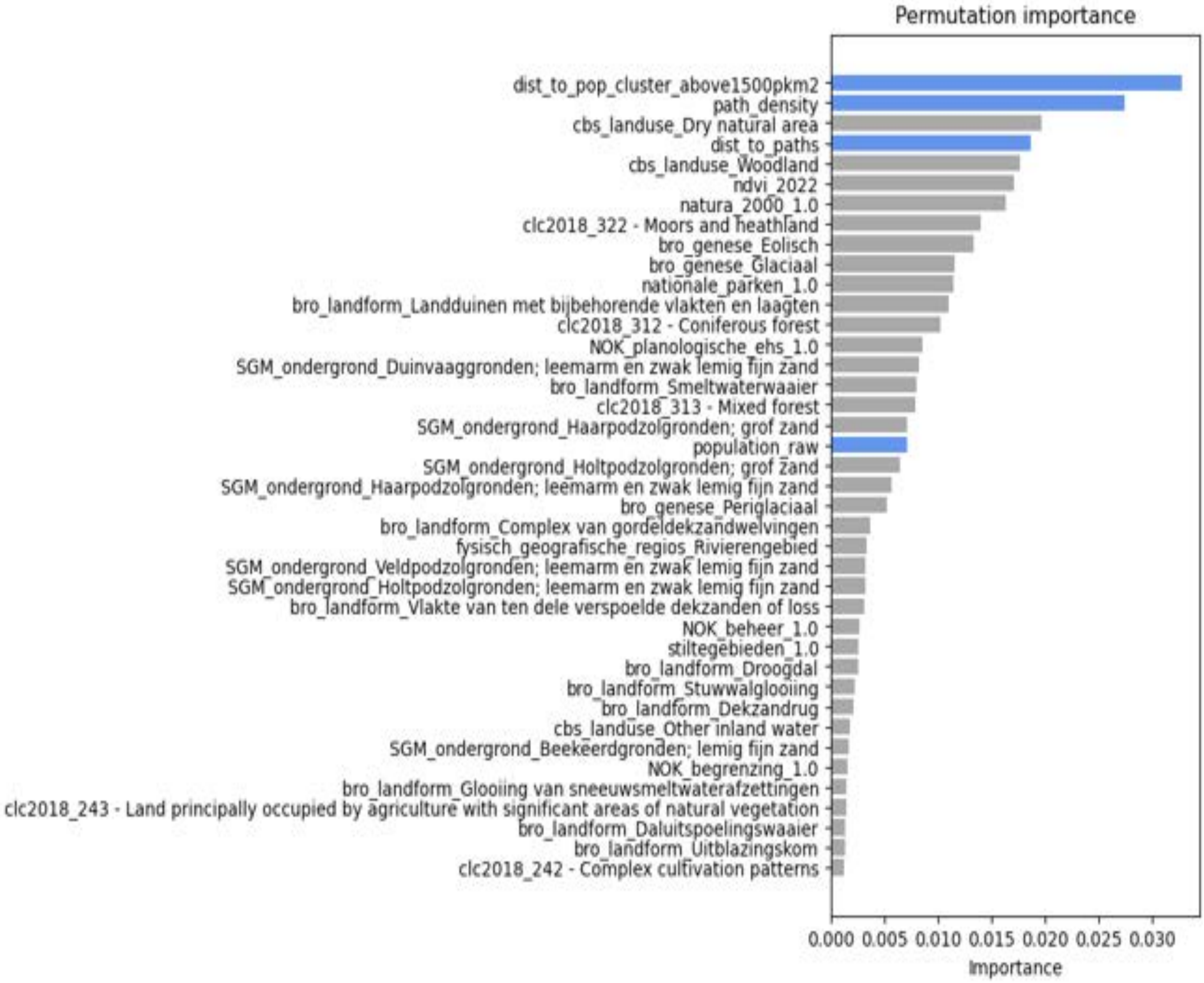

Figure 9 Permutation importance for the KNN model. Highlighted in blue are the observer bias variables.

Predictions

The following section show the difference in spatial distribution between an observer bias corrected and an observer bias uncorrected version of the model. An uncorrected model means that the observer bias variables were included, but with their real values, while a corrected version means that the mean of the observer bias variables were passed to the model for all locations.

Table 7 shows the actual amount of cells where the presence of the 'Heideblauwtje' is predicted by the uncorrected model and the corrected model, as well as the number of cells where an original NDFF observation was made.

	Uncorrected presence count	Corrected presence count
Logistic Regression	22588	19890
Random Forest	17614	19121
KNN	21844	23343
NDFF observations	1243	

Table 7 The number of cells where the presence of the 'Heideblauwtje' has been predicted with and without correction for observer bias. The number of cells where the 'Heideblauwtje' was present in the NDFF dataset has also been added.

The number of predicted presences by the Logistic decreases by roughly 2000, while the number of presences of Random Forest and the KNN increase with roughly 2000. Figure 10 shows the distribution of the standardized observer bias variables. Of all cells in the study area roughly 70% has a lower 'path_density' than the mean. Similarly, 62.8% of the cells have a lower distance to a population cluster than the mean. 63.4% of the cells have a lower distance to paths as the mean and 78.7% has a lower population than the mean. Thus, by passing the mean value, for most cells the 'path_density', distance to population clusters, the distance to paths and population will increase. If these variables influence the model as intended than an increase in 'path_density' and 'population_raw' would increase visitation rates and an increase in 'dist_to_paths' and 'dist_to_pop_cluster_above1500pkm' would decrease visitation rates. Consequently, we would expect an increase in predicted presence by observer bias correction. The unexpected sign for the coefficients 'population_raw' and 'dist_to_pop_cluster_above1500pkm2' in Figure 5 possibly explains why the bias corrected count of presences has decreased for the Logistic Regression.

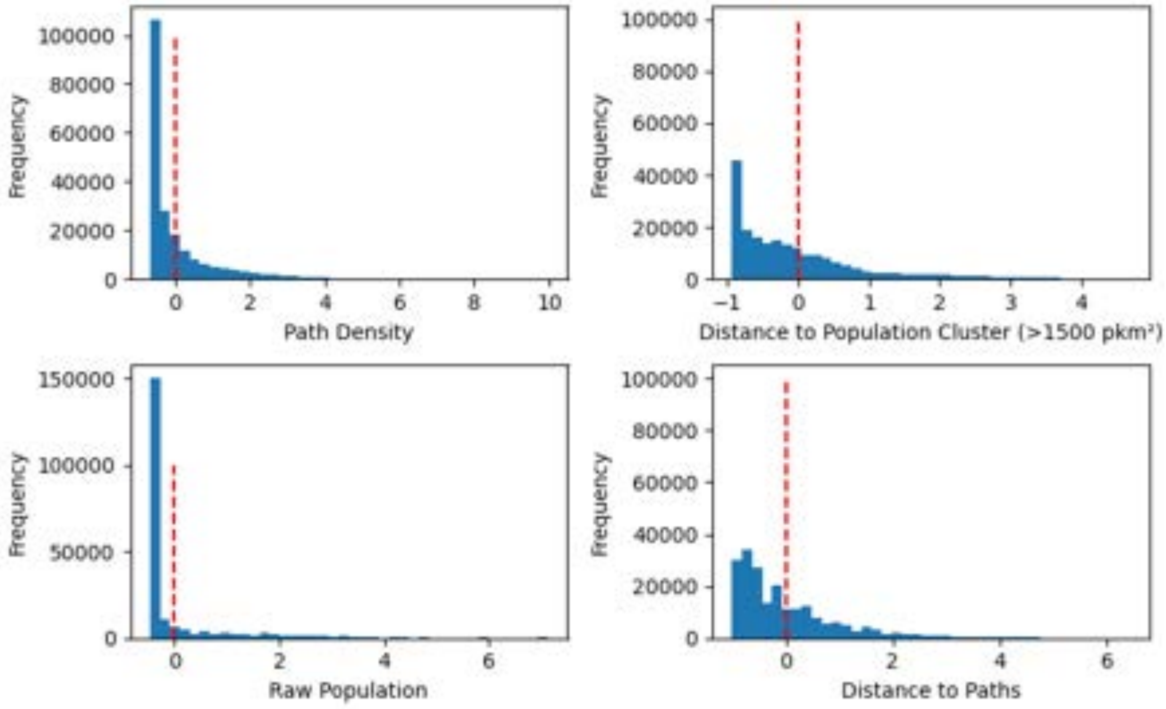

Figure 10 Distributions of observer bias variables. The vertical line at 0 indicates the mean of the standardized features.

Logistic Regression

Figure 11 shows the predictions made by the Logistic Regression in the study area. The predictions follow roughly the natural areas in the study area and no substantial presence has been predicted in agricultural or urban areas. Figure 12 shows a more detailed version of a part of the study area where the paths dataset used to construct the observer bias variables have been plotted in black. Figure 12 shows two examples of the effects by observer bias correction. The red area is the area where the correction of observer bias caused the model to change its prediction from presence to absence, while the blue areas show the locations where the model predicted absence before correction, and presence after correction of observer bias.

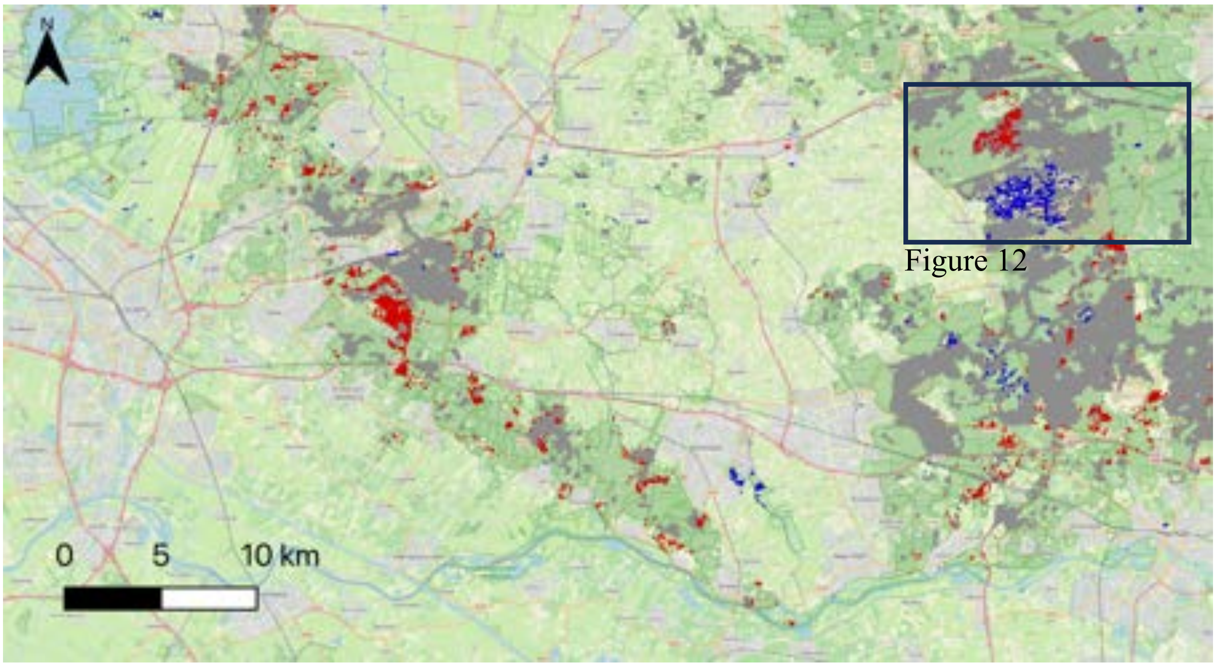

Figure 11 Mapped predictions using Logistic Regression with observer bias variables. Areas in blue are additions by the observer bias correction, areas in red are removals by observer bias correction and gray areas are presence locations unaffected by bias correction. The rectangle is referring to Fout! Verwijzingsbron niet gevonden.**.**

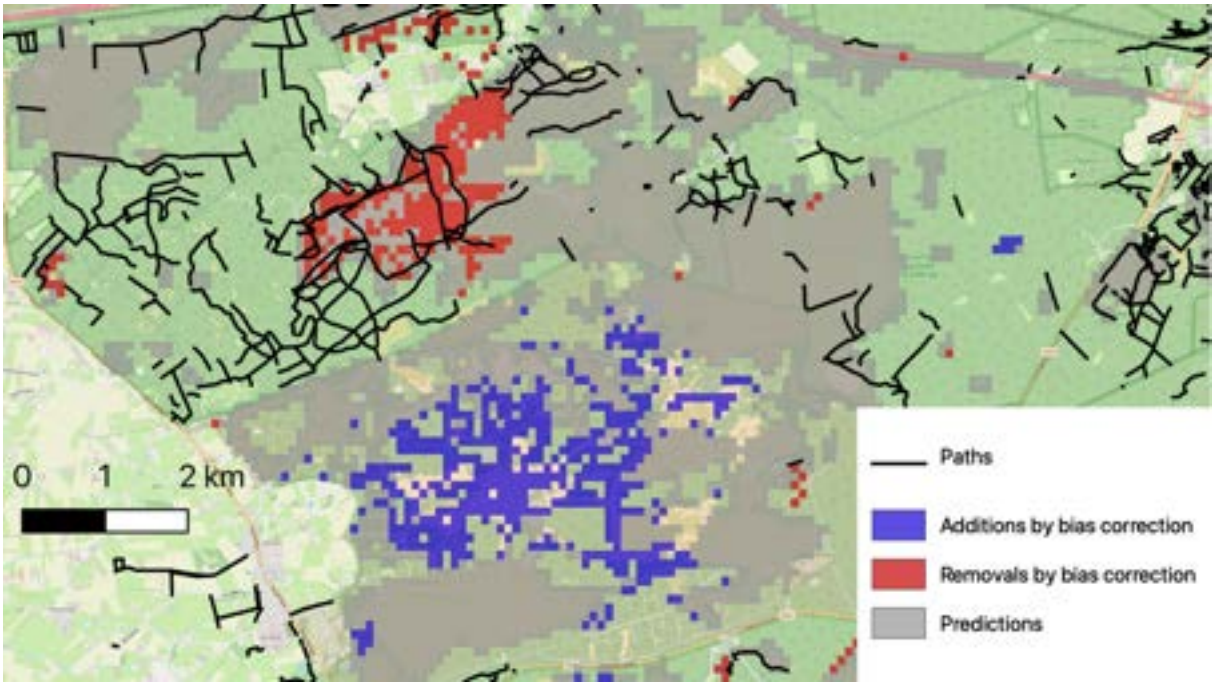

Figure 12 Predictions of the Logistic Regression on a small part of the study area. Black lines mark the paths dataset used in this analysis.

Notably, the blue areas do not intersect with paths, while the red areas do intersect with paths. By fixing the observer bias variables to their mean, the observer bias variables at the red areas have been assigned a lower value compared to their real value. This caused the model to alter its prediction to absence instead of presence. This in line with expectations as the visitation rate at these locations is likely to be high, but there is no observation of a 'Heideblauwtje' in the NDFF. If a 'Heideblauwtje' is present it would have likely been observed. On the other hand, the blue areas are areas where no paths were present in the data. These areas got assigned a higher accessibility when making predictions with correction and therefore the model altered its prediction from absence to presence. This is also in line with expectations. The environmental characteristics are suitable enough for the presence of a 'Heideblauwtje', and it is likely that a 'Heideblauwtje' would be observed when the area is more accessible.

Random Forest

Figure 13 shows the predictions in the study area using a Random Forest algorithm. Areas that have been added by observer bias correction are in blue, while areas that are removed are red. Overall, the predictions follow a very similar pattern compared with the Logistic Regression. That is, most predictions are in natural areas, and not in urban or agricultural areas. Figure 14 shows the same zoomed in area as in Figure 12. The pattern in this part agrees on the presence with the Logistic Regression for the bottom half, but the Random Forest predicts presence for the upper half of the figure while the Logistic regression predicts absence. The interpretation for the bottom half follows the same reasoning as discussed in the previous section. It is however hard to explain why exactly the Random Forest predicts something different than the Logistic Regression. It could be that at this location the population has a lot of influence on the predictions. External validation would provide a lot of insight on this situation.

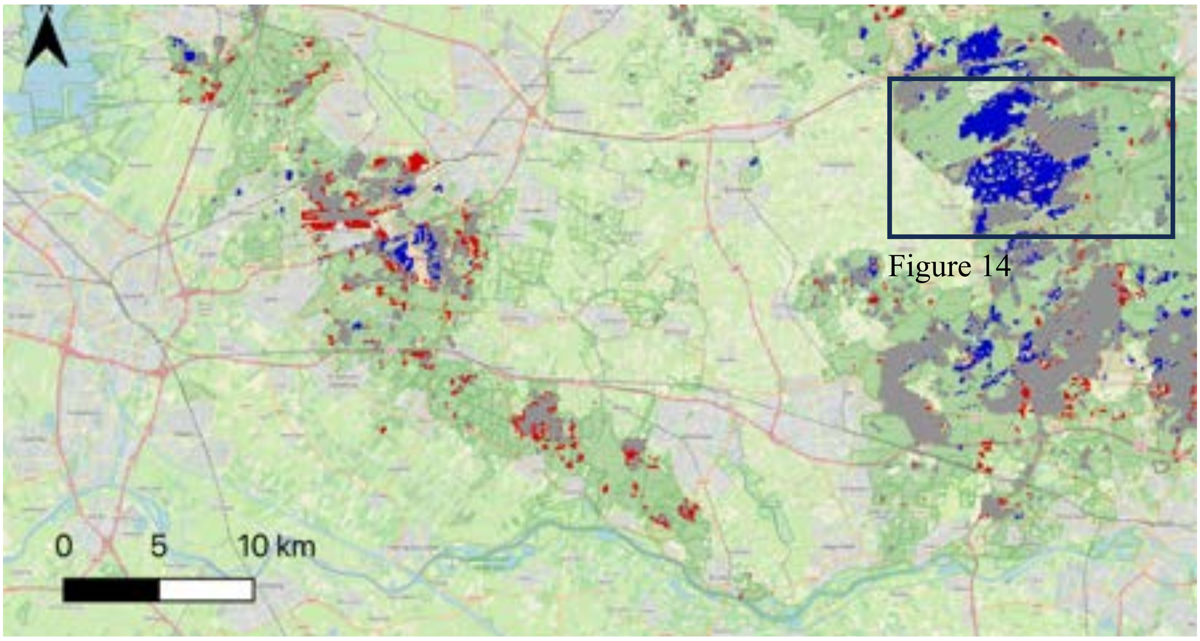

Figure 13 Mapped predictions using Random Forest with observer bias variables. The rectangle refers to Figure 14.

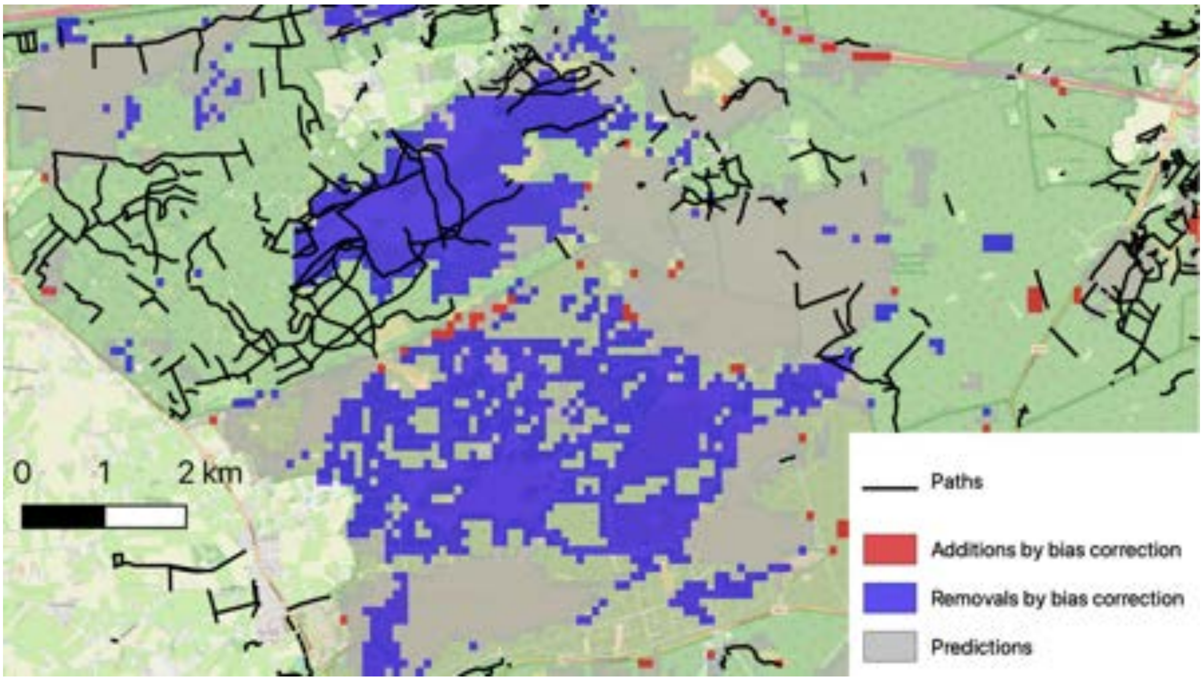

Figure 14 Predictions by the Random Forest on a small part of the study area.

KNN

Lastly, the predictions by the KNN algorithm have been depicted in Figure 15. In comparison to the previous two algorithms, the KNN model exhibits a greater variance in its corrections. In general, the distribution of the 'Heideblauwtje' follows the natural regions of the study area, similar to the other models' predictions. Once again, Figure 16 showcases a small part of the study area with the predictions made by the KNN. Whereas the upper part of this small area aligns more closely with the predictions made by the Random Forest, the lower part does not show any pattern similar to the previous algorithms. The added and removed areas do not seem to follow any specific pattern when examined in relation the paths.

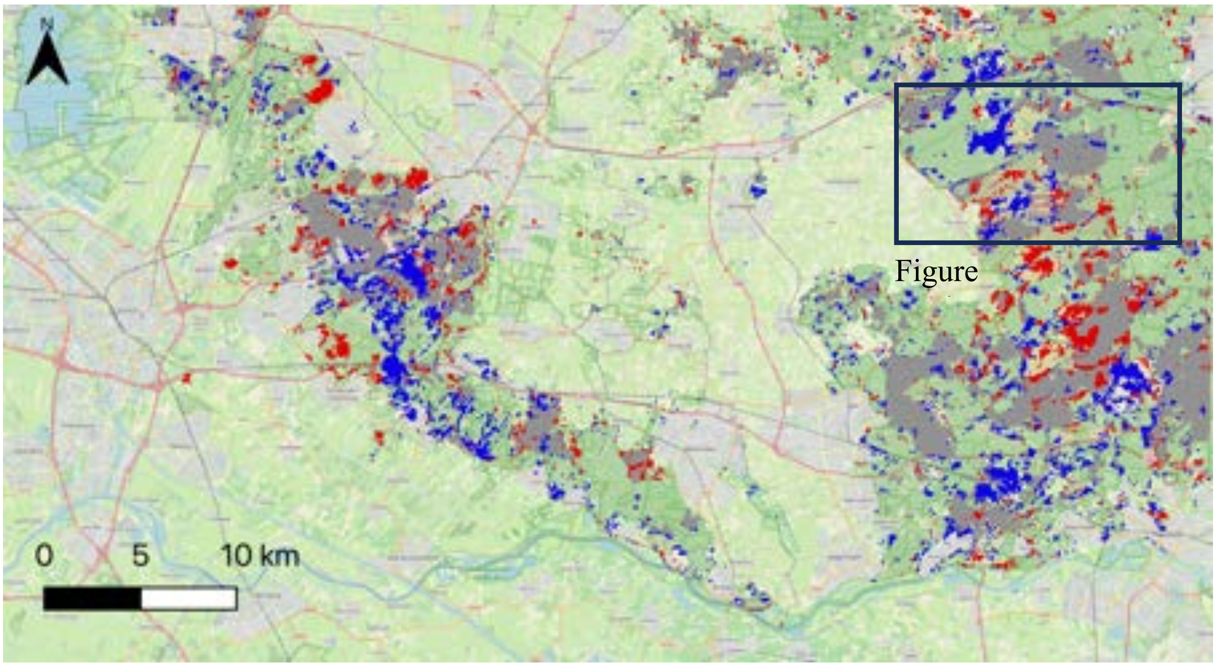

Figure 15 Mapped prediction by the KNN algorithm including observer bias variables.

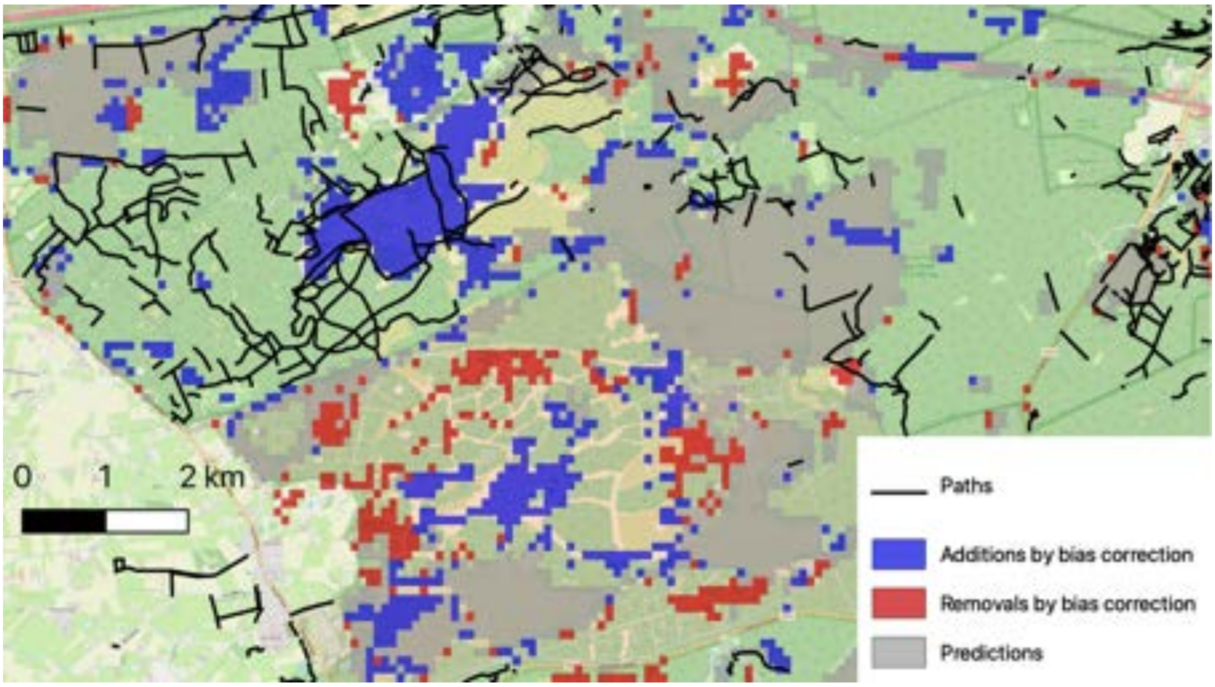

Figure 16 Predictions by the KNN algorithm on a small part of the study area.

Discussion

The results show that the performance of all models improve when adding observer bias variables, but the Logistic Regression and Random Forest perform slightly better than the KNN. However, deciding which model performs best is not only decided by its accuracy. Although all the algorithms show a similar trend in predictions over the whole study area, the Logistic Regression and Random Forest show corrections that are very sensible. The corrections for observer bias by the KNN on the other, show a lot more variation spatially and when viewed in relation to the accessibility of that region it does not show any obvious patterns. It must be noted that KNN suffers greatly from high dimensional data, such as used in this analysis, and would likely improve greatly from some kind of feature selection. Due to l1 penalty on Logistic Regression and the random feature selection per split on a Random Forest, feature selection is less necessary for these models and is not expected to yield great improvements. When comparing the bias corrections made by the Logistic Regression and the Random Forest model, large areas are corrected similarly, but some areas get difference predictions. This can be the result of differences in relative influence of environmental and observer bias variables for the models. Logistic Regression and Random Forest also differ in how they handle non-linearity. Logistic Regression is a linear model and is only able to model the effects of a feature in linear terms. On the other hand, a Random Forest is a non-parametric model and an ensemble learning method that combines multiple decision trees. Each decision tree is built on a random subset of the data and each split is made on a random subset of the features. By aggregating the predictions from multiple trees, a Random Forest is able to model complex interactions and non-linear relationships. It is possible that there is interaction between features and that a Logistic Regression is not able to capture these interactions while a Random Forest is. Another cause for the differences between the models can be multicollinearity. As effect of the data acquisition approach in this study, it is likely that correlated datasets are used in this study. It has been shown that the polarity of coefficients in regression models can be changed when features are highly correlated (Kalnins, 2018). Feature importance methods can also suffer from the effects of multicollinearity. For example, if two features are highly correlated and one of them is removed to assess the decrease in performance, the information can be retrieved by the other feature, making the features seemingly less important. However, this is not such a problem when making predictions, but rather when trying to infer from the model. Moreover, correlations between observer bias variables and environmental variables are shown to be problematic. For example, the distance to population clusters was added with the underlying idea that a shorter distance to population clusters would increase visitation rates. However, a confounding mechanism is that the environmental characteristics improve when the distance to cities increases. A possible solution would be to train a separate model using observations of many different species with different habitat types and only with observer bias variables. Then use this model to predict an 'observation likeliness' for the study, which in turn can be added as predictor in a SDM. The different habitat types in such an approach break the confounding mechanism between observer bias variables and environmental variables. This method assumes that observer bias is equal among different species.

Another thread to the validity of the model is the mismatch between the point in time of an NDFF observation and the environmental characteristics at that time. This study includes observations from the NDFF up to 11 years ago, while for example the NDVI values are retrieved from satellite images in 2022. An area that was used for agriculture ten years ago can in the meanwhile be covered by buildings. The corrections for observer bias seem to be greatly dependent on the 'paths' dataset. The dataset used in this study is gathered from OpenStreetMap, which is collaborative mapping product that is constructed by volunteers. The quality of this data is not guaranteed and can differ over time and space.

Unfortunately, prediction of species abundance showed little promise. However, it is possible to gain more insight in predictions by mapping the log odds of a Logistic Regression or a ratio between trees that predicted presence and trees that predict absence instead of the outcome of a majority vote.

When using a binary decision in a Logistic Regression, it is worth to investigate the effects of adjusting the threshold. Furthermore, the fixed value for observer bias variables used while predicting does not have to be the mean. It can be any value, and tweaking this value influences how strong the correction is. Another sensible value would be the mean value of cells where an original NDFF observation was made. Ideally, the tweaking of these parameters is done using external validation such as professional surveys. A good place to do such a survey would be at 'Kootwijkerzand'. This is a region where the Random Forest and Logistic Regression differ in their corrections for observer bias (Figure 12 and Figure 14) and absence or presence records at this location could provide valuable insights to model performance.

Conclusion

In this study, we set out to fill in the data gaps in the NDFF and investigate the efficacy of using a model-based approach to mitigate observer bias. The performance of Logistic Regression, Random Forest and KNN were examined using randomly sampled locations as pseudo-absences. These models were assessed on a dataset containing 231 features that describe the environmental conditions. In an effort to account for observer bias, an additional 4 features were added that aim to capture the varying accessibility and visitation rates across the study area. The addition of these features was motivated by the recognition that these factors are likely causes for the inconsistency in documentation completeness in the NDFF. Results show that the different models performed very similar with an accuracy of around 0.92 and improved slightly by adding the observer bias variables (+/- 1.5%). When predicting the occurrence of the 'Heideblauwtje' over the entire study area, the observer bias variables were set to their mean value, such that predictions can be interpreted as if each location has equal accessibility. Examining the spatial distribution of the bias corrected areas show very sensible results, particularly for the Logistic Regression and Random Forest. Highly accessible regions without NDFF observations were excluded from the predicted presences, while inaccessible regions were added. Consequently, a distribution free from observer bias is predicted. Though these results are very promising, external validation is necessary to shed light onto the differences and accuracy between models. An excellent way of validation would be surveying 'Kootwijkerzand', as environmental conditions seem to be favorable for the presence of a 'Heideblauwtje' and the models substantially differ in their corrections in this region. Other future research includes investigating the effects of the fixed value used for observer bias variables while predicting, and adjustments to the threshold used to classify the predictions of the Logistic Regression. Finally, this study serves as a proof of concept and since only nationally covered datasets have been used, this approach can be extended to the whole of the Netherlands for any species in the NDFF.

References

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327-338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

CBS, PBL, RIVM, & WUR. (2013). *Biodiversiteitsverlies in Nederland, Europa en de wereld, 1700-2010*. CBS, PBL, RIVM, WUR.

Home. (z.d.). Geraadpleegd 9 juni 2023, van https://www.ndff.nl/

Kalnins, A. (2018). Multicollinearity: How common factors cause Type 1 errors in multivariate regression. *Strategic Management Journal*, *39*(8), 2362-2385. https://doi.org/10.1002/smj.2783

Khan, S. S., & Madden, M. G. (2014). One-class classification: Taxonomy of study and review of techniques. *The Knowledge Engineering Review*, *29*(3), 345-374. https://doi.org/10.1017/S026988891300043X

Miller, J. (2010). Species Distribution Modeling. *Geography Compass*, *4*(6), 490-509. https://doi.org/10.1111/j.1749-8198.2010.00351.x

Protocollen. (z.d.). Geraadpleegd 1 juni 2023, van https://www.ndff.nl/overdendff/validatie/protocollen/

Warton, D. I., Renner, I. W., & Ramp, D. (2013). Model-Based Control of Observer Bias for the Analysis of Presence-Only Data in Ecology. *PLOS ONE*, *8*(11), e79168. https://doi.org/10.1371/journal.pone.0079168

Data sources

Dataset	Reference
Agrarisch Areaal Nederland	https://www.pdok.nl/introductie/-/article/agrarisch-areaal-nederland-aan
Basisregistratie Gewaspercelen	https://www.pdok.nl/introductie/-/article/basisregistratie-gewaspercelen-brp-
BRO Geomorfologische Kaart	https://www.pdok.nl/-/wms-service-voor-bro-geomorfologische-kaart
CBS Bestand Bodemgebruik	https://www.pdok.nl/introductie/-/article/cbs-bestand-bodemgebruik
Fysisch Geografische Regios	https://www.pdok.nl/introductie/-/article/fysisch-geografische-regio-s
Nationale Parken	https://www.pdok.nl/geo-services/-/article/nationale-parken
Natura2000	[https://www.pdok.nl/geo-services/-/article/natura-2000
SGM Ondergrondmodel	https://www.pdok.nl/introductie/-/article/bro-bodemkaart-sgm-
Satellite images	Landsat-7 image courtesy of the U.S. Geological Survey
Stiltegebieden	https://www.pdok.nl/geo-services/-/article/stiltegebieden
Corine Land Cover	https://land.copernicus.eu/pan-european/corine-land-cover/clc2018?tab=download
Paths	OpenStreetMap and OpenStreetMap Foundation (CC-BY-SA). © https://www.openstreemap.org and contributors
Population	https://www.pdok.nl/geo-services/-/article/cbs-bevolkingsspreiding-population-distribution-

Appendix

Feature	Count
dist_to_paths	-
ndvi_2022	-
path_density	-
population_raw	-
dist_to_pop_cluster_above1500pkm2	-
AAN_1.0	7535
BRP_gewas_Aardappelen, consumptie	66
BRP_gewas_Appelen. Aangeplant lopende seizoen.	10
BRP_gewas_Appelen. Aangeplant voorafgaande aan lopende seizoen.	37
BRP_gewas_Bieten, voeder-	48
BRP_gewas_Bos, blijvend, met herplantplicht	24
BRP_gewas_Bos- en haagplanten, open grond,	18
BRP_gewas_Gerst, zomer-	22
BRP_gewas_Gladiool, bloembollen en - knollen	18
BRP_gewas_Grasland, blijvend	3634
BRP_gewas_Grasland, natuurlijk. Hoofdfunctie landbouw.	283
BRP_gewas_Grasland, natuurlijk. Hoofdfunctie natuur.	57
BRP_gewas_Grasland, tijdelijk	611
BRP_gewas_Graszoden	10
BRP_gewas_Kersen, zoet	30
BRP_gewas_Laanbomen/parkbomen, onderstammen, open grond,	13
BRP_gewas_Laanbomen/parkbomen, opzetters, open grond,	66
BRP_gewas_Laanbomen/parkbomen, spillen, open grond,	34
BRP_gewas_Maïs, snij-	811
BRP_gewas_Natuurterreinen (incl. heide)	193
BRP_gewas_Overig	40
BRP_gewas_Peren. Aangeplant voorafgaande aan lopende seizoen.	125
BRP_gewas_Rand, grenzend aan bouwland, hoofdzakelijk bestaand uit een ander gewas dan gras. (EA: beheer)	83
BRP_gewas_Rand, grenzend aan bouwland, hoofdzakelijk bestaand uit een ander gewas dan gras. (EA: onbeheerd)	43
BRP_gewas_Rogge (geen snijrogge)	22
BRP_gewas_Sloot, grenzend aan beheerde akkerrand	11
BRP_gewas_Tarwe, winter-	32
BRP_gewas_Tarwe, zomer-	52
BRP_gewas_Vaste planten, open grond,	22
BRP_gewas_Vruchtbomen, overig, open grond,	26
cbs_landuse_Allotment garden	275
cbs_landuse_Building site	1758
cbs_landuse_Cemetery	512
cbs_landuse_Dry natural area	19057
cbs_landuse_Holiday recreation	870
cbs_landuse_Industrial area and offices	4716

cbs_landuse_Other inland water	16014
cbs_landuse_Park and public garden	4397
cbs_landuse_Public institutions	1051
cbs_landuse_Residential	13204
cbs_landuse_Rijn & Maas	2309
cbs_landuse_Socio-cultural facility	2032
cbs_landuse_Sports ground (incl. car parks)	2597
cbs_landuse_Water with recreational usage	1997
cbs_landuse_Wet natural area	1430
cbs_landuse_Woodland	57253
clc2018_121 - Industrial or commercial units	6413
clc2018_122 - Road and rail networks and associated land	1396
clc2018_123 - Port areas	99
clc2018_124 - Airports	639
clc2018_131 - Mineral extraction sites	68
clc2018_132 - Dump sites	124
clc2018_133 - Construction sites	509
clc2018_141 - Green urban areas	1409
clc2018_142 - Sport and leisure facilities	3375
clc2018_211 - Non-irrigated arable land	5238
clc2018_222 - Fruit trees and berry plantations	837
clc2018_242 - Complex cultivation patterns	20220
clc2018_243 - Land principally occupied by agriculture with significant areas of natural vegetation	7718
clc2018_311 - Broad-leaved forest	3438
clc2018_312 - Coniferous forest	29570
clc2018_313 - Mixed forest	18686
clc2018_321 - Natural grasslands	1332
clc2018_322 - Moors and heathland	9803
clc2018_324 - Transitional woodland-shrub	30
clc2018_331 - Beaches - dunes - sands	972
clc2018_411 - Inland marshes	2294
clc2018_511 - Water courses	2612
clc2018_512 - Water bodies	3353
fysisch_geografische_regios_Laagveengebied	9179
fysisch_geografische_regios_Rivierengebied	53612
fysisch_geografische_regios_Zeekleigebied	1473
nationale_parken_1.0	17908
natura_2000_1.0	48974
NOK_begrenzing_1.0	9248
NOK_beheer_1.0	2525
NOK_planologische_ehs_1.0	91481
NOK_verwervinginrichting_1.0	4590
SGM_ondergrond_Akkereerdgronden; grof zand	118
SGM_ondergrond_Beekeerdgronden; leemarm en zwak lemig fijn zand	1343

SGM_ondergrond_Beekeerdgronden; lemig fijn zand	11123
SGM_ondergrond_Duinvaaggronden; grof zand	554
SGM_ondergrond_Duinvaaggronden; leemarm en zwak lemig fijn zand	17338
SGM_ondergrond_Gooreerdgronden; grof zand	86
SGM_ondergrond_Gooreerdgronden; leemarm en zwak lemig fijn zand	4973
SGM_ondergrond_Gooreerdgronden; lemig fijn zand	1236
SGM_ondergrond_Haarpodzolgronden; grof zand	10829
SGM_ondergrond_Haarpodzolgronden; leemarm en zwak lemig fijn zand	7668
SGM_ondergrond_Hoge bruine enkeerdgronden; grof zand	416
SGM_ondergrond_Hoge bruine enkeerdgronden; leemarm en zwak lemig fijn zand	123
SGM_ondergrond_Hoge bruine enkeerdgronden; lemig fijn zand	859
SGM_ondergrond_Hoge zwarte enkeerdgronden; grof zand	1145
SGM_ondergrond_Hoge zwarte enkeerdgronden; leemarm en zwak lemig fijn zand	7894
SGM_ondergrond_Hoge zwarte enkeerdgronden; lemig fijn zand	340
SGM_ondergrond_Holtpodzolgronden; grof zand	17887
SGM_ondergrond_Holtpodzolgronden; leemarm en zwak lemig fijn zand	4880
SGM_ondergrond_Kalkarme drechtvaaggronden; zware klei, profielverloop 1	367
SGM_ondergrond_Kalkarme leek-/woudeerdgronden; zavel, profielverloop 2	53
SGM_ondergrond_Kalkarme nesvaaggronden; klei	55
SGM_ondergrond_Kalkarme poldervaaggronden; klei, profielverloop 3, of 3 en 4, of 4	55
SGM_ondergrond_Kalkhoudende nesvaaggronden; zavel en lichte klei	96
SGM_ondergrond_Kalkhoudende ooivaaggronden; lichte zavel	674
SGM_ondergrond_Kalkhoudende ooivaaggronden; zware zavel en lichte klei	5651
SGM_ondergrond_Kalkhoudende poldervaaggronden; klei, profielverloop 2	198
SGM_ondergrond_Kalkhoudende poldervaaggronden; lichte zavel, profielverloop 5	580
SGM_ondergrond_Kalkhoudende poldervaaggronden; zavel en lichte klei, profielverloop 3, of 3 en 4, of 4	966
SGM_ondergrond_Kalkhoudende poldervaaggronden; zavel, profielverloop 2	2797
SGM_ondergrond_Kalkhoudende poldervaaggronden; zware klei, profielverloop 5	109
SGM_ondergrond_Kalkhoudende poldervaaggronden; zware zavel en lichte klei, profielverloop 5	5715
SGM_ondergrond_Kalkhoudende vlakvaaggronden; grof zand	132
SGM_ondergrond_Kalkhoudende vlakvaaggronden; matig fijn zand	44
SGM_ondergrond_Kalkloze drechtvaaggronden; profielverloop 1	4337
SGM_ondergrond_Kalkloze nesvaaggronden; zavel en lichte klei	326
SGM_ondergrond_Kalkloze nesvaaggronden; zware klei	349
SGM_ondergrond_Kalkloze ooivaaggronden; lichte zavel	172
SGM_ondergrond_Kalkloze ooivaaggronden; zware zavel en lichte klei	3242
SGM_ondergrond_Kalkloze poldervaaggronden (bruine komgrond); zware klei, profielverloop 3, of 3 en 4, of 4	932
SGM_ondergrond_Kalkloze poldervaaggronden; zavel en lichte klei, profielverloop 2	864
SGM_ondergrond_Kalkloze poldervaaggronden; zavel en lichte klei, profielverloop 3, of 3 en 4	5489
SGM_ondergrond_Kalkloze poldervaaggronden; zware klei, profielverloop 2	91
SGM_ondergrond_Kalkloze poldervaaggronden; zware klei, profielverloop 3, of 3 en 4	5929
SGM_ondergrond_Kalkloze poldervaaggronden; zware klei, profielverloop 4	7670

SGM_ondergrond_Kalkloze poldervaaggronden; zware zavel en lichte klei, profielverloop 4	387
SGM_ondergrond_Kalkloze poldervaaggronden; zware zavel en lichte klei, profielverloop 5	1866
SGM_ondergrond_Kamppodzolgronden; leemarm en zwak lemig fijn zand	601
SGM_ondergrond_Kleiige beekdalgronden	48
SGM_ondergrond_Koopveengronden op (meestal niet-gerijpte) zavel of klei, beginnend ondieper dan 1.2 m	70
SGM_ondergrond_Koopveengronden op veenmosveen	41
SGM_ondergrond_Koopveengronden op zand, beginnend ondieper dan 1.2 m	852
SGM_ondergrond_Koopveengronden op zeggeveen, rietzeggeveen of (mesotroof) broekveen	90
SGM_ondergrond_Laarpodzolgronden; leemarm en zwak lemig fijn zand	7452
SGM_ondergrond_Laarpodzolgronden; lemig fijn zand	395
SGM_ondergrond_Lage enkeerdgronden; leemarm en zwak lemig fijn zand	815
SGM_ondergrond_Lage enkeerdgronden; lemig fijn zand	55
SGM_ondergrond_Leek-/woudeerdgronden; klei, profielverloop 3, of 3 en 4, of 4	136
SGM_ondergrond_Leek-/woudeerdgronden; zavel, profielverloop 3, of 3 en 4, of 4	140
SGM_ondergrond_Leek-/woudeerdgronden; zavel, profielverloop 5, of 5 en 2, of 2	246
SGM_ondergrond_Liedeerdgronden; klei, profielverloop 1	20
SGM_ondergrond_Loopodzolgronden; grof zand	900
SGM_ondergrond_Loopodzolgronden; leemarm en zwak lemig fijn zand	486
SGM_ondergrond_Loopodzolgronden; lemig fijn zand	21
SGM_ondergrond_Madeveengronden op zand met humuspodzol, beginnend ondieper dan 1.2 m	230
SGM_ondergrond_Meerveengronden op zand met humuspodzol, beginnend ondieper dan 1.2 m	419
SGM_ondergrond_Meerveengronden op zand zonder humuspodzol, beginnend ondieper dan 1.2 m	251
SGM_ondergrond_Meerveengronden op zeggeveen. rietzeggeveen of broekveen	36
SGM_ondergrond_Moerige eerdgronden met een moerige bovengrond op zand	1178
SGM_ondergrond_Moerige eerdgronden met een zanddek en een moerige tussenlaag op zand	568
SGM_ondergrond_Moerige eerdgronden met een zavel- of kleidek en een moerige tussenlaag op zand	608
SGM_ondergrond_Moerige podzolgronden met een humushoudend zanddek en een moerige tussenlaag	1370
SGM_ondergrond_Moerige podzolgronden met een moerige bovengrond	82
SGM_ondergrond_Moerige podzolgronden met een zavel- of een kleidek en een moerige tussenlaag	229
SGM_ondergrond_Overslaggronden	86
SGM_ondergrond_Petgaten	959
SGM_ondergrond_Stuifzandgronden	624
SGM_ondergrond_Tuineerdgronden; lichte zavel, profielverloop 5, of 5 en 2, of 2	54
SGM_ondergrond_Tuineerdgronden; zware zavel en klei, profielverloop 5, of 5 en 2, of 2	48
SGM_ondergrond_Veldpodzolgronden; grof zand	173
SGM_ondergrond_Veldpodzolgronden; leemarm en zwak lemig fijn zand	12560
SGM_ondergrond_Veldpodzolgronden; lemig fijn zand	235
SGM_ondergrond_Venige beekdalgronden	423

SGM_ondergrond_Vlakvaaggronden; grof zand	154
SGM_ondergrond_Vlakvaaggronden; leemarm en zwak lemig fijn zand	3614
SGM_ondergrond_Vlakvaaggronden; lemig fijn zand	2177
SGM_ondergrond_Vlierveengronden op bagger, verslagen veen, gyttja of andere veensoorten	113
SGM_ondergrond_Vlierveengronden op zand met humuspodzol, beginnend ondieper dan 1.2 m	143
SGM_ondergrond_Vlierveengronden op zand zonder humuspodzol, beginnend ondieper dan 1.2 m	36
SGM_ondergrond_Vlietveengronden	25
SGM_ondergrond_Waardveengronden op bosveen (of eutroof broekveen)	648
SGM_ondergrond_Waardveengronden op veenmosveen	125
SGM_ondergrond_Waardveengronden op zand, beginnend ondieper dan 1.2 m	497
SGM_ondergrond_Waardveengronden op zeggeveen, rietzeggeveen of (mesotroof) broekveen	847
SGM_ondergrond_Weideveengronden op zand, beginnend ondieper dan 1.2 m	280
SGM_ondergrond_Weideveengronden op zeggeveen, rietzeggeveen of (mesotroof) broekveen	155
SGM_ondergrond_Zandige beekdalgronden	502
stiltegebieden_1.0	10035
bro_genese_Antropogeen	7770
bro_genese_Denudatief	1073
bro_genese_Eolisch	84893
bro_genese_Glaciaal	34006
bro_genese_Marien	445
bro_genese_Periglaciaal	17950
bro_landform_Beekdalbodem	226
bro_landform_Complex van dekzandwelvingen	953
bro_landform_Complex van gordeldekzandwelvingen	10157
bro_landform_Daluitspoelingswaaier	2617
bro_landform_Dalvormige laagte	4641
bro_landform_Dekzandrug	18816
bro_landform_Dekzandwelving	645
bro_landform_Doorbraakwaaier	321
bro_landform_Droogdal	6090
bro_landform_Geulranddekzandrug	566
bro_landform_Glooiing van hellingafspoelingen	1073
bro_landform_Glooiing van sneeuwsmeltwaterafzettingen	3290
bro_landform_Gordeldekzandglooiing	2217
bro_landform_Gordeldekzandrug	717
bro_landform_Gordeldekzandvlakte	1600
bro_landform_Groeve	597
bro_landform_Kronkelwaardgeul	474
bro_landform_Kronkelwaardrug	620
bro_landform_Kunstmatig gecreeerd relief voor recreatiedoeleinden zoals golfbanen	437
bro_landform_Laagte ontstaan door afgraving	2028

bro_landform_Landduin	1058
bro_landform_Landduinen met bijbehorende vlakten en laagten	16757
bro_landform_Meanderruggen en -geulen	2202
bro_landform_Ondergraven stuwwalzijde	193
bro_landform_Ontgonnen veenvlakte met petgaten	3120
bro_landform_Overloop- of crevassegeul	187
bro_landform_Plateau-achtige storthoop	1245
bro_landform_Restgeul	1461
bro_landform_Rivier- of beekbedding	237
bro_landform_Rivierdalbodem	256
bro_landform_Rivierkom- en oeverwalachtige vlakte	4026
bro_landform_Rivierkomvlakte	11477
bro_landform_Smeltwaterheuvel	127
bro_landform_Smeltwaterwaaier	9145
bro_landform_Storthopen met grind-	449
bro_landform_Stroomrug of stroomgordel	14849
bro_landform_Stroomrugglooiing	2864
bro_landform_Stuifzandvlakte	2914
bro_landform_Stuwwalglooiing	2770
bro_landform_Stuwwalplateau	1910
bro_landform_Terp (wierd) of hoogwatervluchtplaats	176
bro_landform_Trechtervormig droogdal	1294
bro_landform_Uitblazingskom	1091
bro_landform_Veenrestvlakte	528
bro_landform_Vlakte ontstaan door afgraving en/of egalisatie	2832
bro_landform_Vlakte van rivierafzettingen	556
bro_landform_Vlakte van smeltwaterafzettingen	1700
bro_landform_Vlakte van ten dele verspoelde dekzanden of loss	27688
bro_landform_Welvingen in rivierafzettingen	500

Appendix 1 All 236 features and their cell count used in this study. The prefix, if present, corresponds to the source dataset.

C	max_iter	penalty	solver	Mean accuracy	Std. accuracy	rank
0,5	1000	l1	liblinear	0,90694404	0,00742352	1
0,5	5000	l1	liblinear	0,90694404	0,00742352	1
0,5	5000	l2	liblinear	0,90694404	0,006895347	3
0,5	1000	l2	liblinear	0,90694404	0,006895347	3
0,5	5000	l1	saga	0,906438996	0,006623447	5
0,5	1000	l1	saga	0,906438996	0,006623447	5
1	1000	l1	liblinear	0,906435199	0,005425983	7
1	5000	l1	liblinear	0,906435199	0,005425983	7
1	1000	l1	saga	0,90593142	0,005711901	9
1	5000	l1	saga	0,90593142	0,005711901	9
2	1000	l1	liblinear	0,905432705	0,010612158	11
2	5000	l1	liblinear	0,905432705	0,010612158	11
1	1000	l2	liblinear	0,905431439	0,007226765	13
1	5000	l2	liblinear	0,905431439	0,007226765	13
1	1000	l2	saga	0,904926395	0,006284735	15
1	1000	l2	sag	0,904926395	0,006284735	15
1	1000	l2	lbfgs	0,904926395	0,006284735	15
1	5000	l2	lbfgs	0,904926395	0,006284735	15
1	5000	l2	sag	0,904926395	0,006284735	15
1	5000	l2	saga	0,904926395	0,006284735	15
2	5000	l2	liblinear	0,904925129	0,007410227	21
2	1000	l2	liblinear	0,904925129	0,007410227	21
0,5	5000	l2	saga	0,904426414	0,006553993	23
0,5	5000	l2	newton-cg	0,904426414	0,006553993	23
0,5	5000	l2	sag	0,904426414	0,006553993	23
0,5	1000	l2	sag	0,904426414	0,006553993	23
0,5	1000	l2	saga	0,904426414	0,006553993	23
0,5	5000	l2	lbfgs	0,904426414	0,006553993	23
0,5	1000	l2	newton-cg	0,904426414	0,006553993	23
0,5	1000	l2	lbfgs	0,904426414	0,006553993	23
2	5000	l1	saga	0,904425148	0,010432902	31
2	1000	l1	saga	0,904425148	0,010432902	31
1	1000	l2	newton-cg	0,904423883	0,006779687	33
1	5000	l2	newton-cg	0,904423883	0,006779687	33
4	1000	l2	liblinear	0,904421351	0,009046498	35
4	5000	l2	liblinear	0,904421351	0,009046498	35
4	5000	l1	saga	0,903918839	0,012082706	37
3	5000	l1	saga	0,903918839	0,012491151	37
3	1000	l1	saga	0,903917573	0,012299455	39
2	5000	l2	saga	0,90341506	0,008562786	40

Appendix 2 Parameters and accuracy for the 40 best performing Logistic Regression models using 5-fold cross validation.

max_depth	max_features	min_samples_split	n_estimators	Mean accuracy	Std. accuracy	rank
300	log2	4	400	0,924542106	0,007997904	1
300	log2	4	600	0,924537043	0,010623665	2
None	log2	4	400	0,924038328	0,010272904	3
200	sqrt	4	600	0,924037062	0,009384304	4
None	log2	4	500	0,924035796	0,010045534	5
200	log2	4	400	0,924035796	0,009662436	5
200	log2	5	500	0,923534549	0,007459627	7
200	log2	4	600	0,923532018	0,009564396	8
None	sqrt	4	600	0,923032037	0,007622336	9
200	log2	5	400	0,923032037	0,008258375	9
None	log2	5	500	0,923032037	0,0076215	11
None	log2	5	600	0,923030771	0,007638363	12
300	sqrt	4	500	0,923030771	0,01019063	12
200	log2	4	500	0,923030771	0,009682365	12
300	sqrt	4	600	0,922532056	0,007035605	15
None	log2	4	600	0,92253079	0,010010426	16
300	sqrt	5	500	0,92253079	0,008362526	16
300	sqrt	5	600	0,922529524	0,00790877	18
300	log2	5	400	0,922528258	0,010034069	19
300	log2	4	500	0,922528258	0,010157866	19
300	log2	5	500	0,922028277	0,009434782	21
300	sqrt	4	400	0,922027012	0,007999533	22
200	log2	5	600	0,922027012	0,006805602	22
None	sqrt	4	400	0,922027012	0,010935488	24
None	sqrt	4	500	0,922027012	0,009037082	24
200	sqrt	4	400	0,922023214	0,009749828	26
200	sqrt	7	500	0,921525765	0,007038429	27
200	sqrt	4	500	0,921524499	0,008364749	28
300	log2	6	500	0,921524499	0,005883569	28
300	log2	6	400	0,921524499	0,007572522	28
None	sqrt	5	400	0,921523233	0,008961616	31
None	sqrt	5	600	0,921523233	0,00977045	31
None	sqrt	6	600	0,921021987	0,006718498	33
None	sqrt	7	400	0,921020721	0,007616013	34
200	sqrt	5	600	0,921020721	0,009397123	34
200	sqrt	5	500	0,921020721	0,008843369	34
None	log2	5	400	0,921019455	0,007792954	37
None	sqrt	6	400	0,920519474	0,008534127	38
300	sqrt	6	600	0,920518208	0,007610416	39
None	log2	6	500	0,920518208	0,007608742	39

Appendix 3 Parameters and accuracy of the 40 best performing Random Forests using 5-fold cross validation. The number of candidates considered for the best split were "log2" and "sqrt," which corresponded to taking the logarithm base 2 and the square root of the total number of features, respectively.

metric	n_neighbors	weights	mean accuracy	std. Accuracy	rank
manhattan	7	distance	0,919522043	0,009225753	1
manhattan	6	distance	0,918518284	0,009558798	2
minkowski	4	distance	0,918008177	0,007881441	3
euclidean	4	distance	0,918008177	0,007881441	3
manhattan	12	distance	0,917513259	0,011113841	5
manhattan	8	distance	0,917510727	0,012005914	6
manhattan	11	distance	0,917009481	0,011738827	7
manhattan	13	distance	0,9165095	0,011325461	8
manhattan	9	distance	0,916503171	0,010909878	9
manhattan	5	distance	0,916501905	0,009295191	10
manhattan	4	distance	0,916500639	0,00888903	11
manhattan	10	distance	0,916000658	0,008890856	12
euclidean	8	distance	0,915494348	0,01095484	13
minkowski	8	distance	0,915494348	0,01095484	13
euclidean	6	distance	0,914996899	0,013394301	15
minkowski	6	distance	0,914996899	0,013394301	15
manhattan	18	distance	0,913994405	0,012389645	17
manhattan	14	distance	0,913994405	0,011323127	17
minkowski	5	distance	0,913988076	0,012928104	19
euclidean	5	distance	0,913988076	0,012928104	19
manhattan	16	distance	0,913493158	0,01249998	21
euclidean	12	distance	0,913485564	0,010088406	22
minkowski	12	distance	0,913485564	0,010088406	22
manhattan	3	distance	0,913476703	0,010533664	24
manhattan	15	distance	0,91298938	0,01375345	25
manhattan	4	uniform	0,912984317	0,007317544	26
manhattan	19	distance	0,912486868	0,013644246	27
manhattan	17	distance	0,912484336	0,013467935	28
euclidean	7	distance	0,912478007	0,011805968	29
minkowski	7	distance	0,912478007	0,011805968	29
euclidean	4	uniform	0,91197676	0,00821917	31
minkowski	4	uniform	0,91197676	0,00821917	31
manhattan	2	distance	0,911975495	0,009233451	33
minkowski	6	uniform	0,911476779	0,015891965	34
euclidean	6	uniform	0,911476779	0,015891965	34
euclidean	10	distance	0,911472982	0,012010862	36
minkowski	10	distance	0,911472982	0,012010862	36
euclidean	3	distance	0,911466653	0,009477419	38
minkowski	3	distance	0,911466653	0,009477419	38
manhattan	3	uniform	0,911466653	0,009477419	40

Appendix 4 Parameters and accuracy of the 40 best performing KNN models using 5-fold cross validation.

Dropped variables
‘pop_clusters_above_1500pkm2_1.0’
‘clc2018_112 – Discontinuous urban fabric’
Wetlands_1.0
Bro_landform_Ontgonnen veenvlakte
Bro_genese_Fluvatiel
Bro_landform Stuwwal
Bro_landform Vlakte van getij-riviermondafzettingen
Bro_genese_Organogeen
Fysisch_geografische_regios_Hogere Zandgronden

Appendix 5 Variables that were dropped because of high correlation with other features.