Open Access

Spatially Predictive Habitat Modeling of a White Stork (Ciconia Ciconia) Population in Former East Prussia in 1939

Claudia Wickert*[,1], Dieter Wallschläger[2] and Falk Huettmann[3]

[1]*Ecoethology Lab, University of Potsdam, Maulbeerallee 2a, 14469 Potsdam, Germany;* [2]*Ecoethology Lab, University of Potsdam, Maulbeerallee 2a, 14469 Potsdam, Germany;* [3]*EWHALE Lab, Biology and Wildlife Department, Institute of Arctic Biology, University of Alaska-Fairbanks, Fairbanks Alaska 99775-7000, USA*

Abstract: Historic information is often crucial for assessing changes and drivers for wildlife and habitat changes although it is often plagued with statistically poor quality. Here we developed three habitat models on two different scales for 1939 for the white stork (Ciconia ciconia) in the region of former East Prussia. We used a geographical information system and a statistical modeling algorithm that comes from the disciplines of machine-learning and data mining (TreeNet). The occurrence of white stork nesting grounds is mainly defined by the variables 'distance to forest', 'distance to/density of settlement', 'distance to pasture' and 'distance to coastline'. The models present for the first time a quantitative predictive distribution estimate for East Prussia. They are a sound foundation but could be further improved by more data regarding the structure of the habitat and more exact spatially explicit information on the location of white stork nesting sites.

Keywords: Stochastic gradient boosting, TreeNet, Predictive GIS habitat model, White stork (Ciconia ciconia), East Prussia.

INTRODUCTION

White storks that nest in the former German province of East Prussia (equals app. the current Russian oblast Kaliningrad and the Polish voivodship Warmia-Masuria) are part of the Masurian-Baltic core population. This region shows a very high density in nesting (more than 10 nesting pairs per km^2 in the overall region and even 30 nesting pairs per 100 km^2 in the Polish section [1]). This finding contrasts the situation of the white stork in large areas of Western Europe where the number of successfully nesting white storks has been severely diminished due to human interference into nature by irrigation, monocultures, intensive agriculture and other practices that are widely known to be unsustainable (e.g. [2-5]). Thus, the question arises which environmental factors and conditions in East Prussia (in the past and in the present) have made this region such a successful nesting ground for the white stork? Secondly, are there differences for this productive population within the region or when exploring different scales?

Man has long shown a keen interest in the white stork [6, 7]. The first large-scale inventory in parts of Poland was exercised as early as 1876 [8]. In the region of former East Prussia a first inventory was done in 1905 [9, 10]. A first international census of the white stork population of many European countries took place in 1934, and then was repeated almost continuously at decadal intervals. One should mention in this context that animal surveys are statistically not trivial and that the methods for obtaining reliable estimates with high confidence have greatly improved by now [11, 12]. However, they have not been applied at that time, resulting into somewhat incomplete, biased, imprecise estimates with unknown but likely very high confidence estimates.

The impressive and comprehensive data material on the development of white stork populations that is available has further been extended by a variety of surveys, e.g. by studies on the ecology of food supply (e.g. [13, 14]) or on the migration behavior of the white stork (e.g. [15-17]).

An increasing digital availability of data sets (see [18]) on the biotic and abiotic characteristics of the landscapes, and in connection with GIS and progressive statistical methods provide a unique modeling opportunity of the environmental requirements for this species. But only in the last few years powerful methods in statistics were developed which show promising results in their applications, especially in regards to predicting species-habitat-relationships [19, 20] and when using poor quality, or sometimes even faulty data [21]. These methods originate, in part, from the disciplines of ecological niche modeling, machine-learning and data mining. That makes them important tools in the interpretation of historic data, for which the method of assuring the data details often can not be completely reconstructed, and which often do not adhere to high quality data collection methods. That is why faults, and a valid inference from such data, are so difficult to assess [20, 22], harming our knowledge on historical bird information.

In this study, for the first time, we develop a historic habitat model for the white stork using GIS and advanced modeling techniques in order to overcome inherent problems

*Address correspondence to this author at the Ecoethology Lab, University of Potsdam, Maulbeerallee 2a, 14469 Potsdam, Germany;
Tel: (907) 474-7882; Fax: (907) 474-6716;
E-mail: cla_wi@hotmail.com

in historical data for generalization. It is our goal to present general methods how such information can be derived, and a first application and model as a sound foundation for assessing changes in wildlife and habitats over time. Our data, code and methods are freely available for further improvement.

METHODS

Study Area in the 1930s

The region of former East Prussia is situated between the 53rd and the 56th northern latitude, and the 18th and 23rd eastern longitude. In the north it borders on the Baltic Sea and the River Neman ('Memel'), in the West on the Rivers Nogat and Vistula ('Weichsel') and in the East on Lithuania. In the South, the former province stretched to the southern tip of the Masurian Lakes Platteau ('Masurische Seenplatte'). From 1922 to 1939 East Prussia covered an area of 36,992 km^2 having approx. 2.5 million inhabitants in the year 1939 [23]. The capital of the province was Kaliningrad (former Königsberg). In former East Prussia the main sources of income and business were in agriculture and forestry sectors. Cattle farming became another important source of income, since it was widely independent of unexpected climatic impairment which traditionally led to a collapse in the crop yield. Even with the beginning of industrialization 1860s onwards little changed due to the lack of natural resources (coal deposits) and the dominance of the agricultural structure. In 1936, 47.2 % of the total area of East Prussia was used as crop land, 20 % as pastures and 19.3 % as forest acreage [24].

Modeling Approach to Generalize from Incomplete, Historic Data Sets

For overcoming the inherent problems in historic bird and habitat data, a powerful and progressive machine learning algorithm called TreeNet was used, the software is offered by Salford Systems Ltd. (http://www.salford-systems.com). TreeNet is a tree-based computational method within the realms of data mining, and presents one of the many modeling algorithms in the statistical modeling toolbox (see [20] for overview). One of its strength is that it is non-parametric and that it can be used for regression as well as classification problems, with continuous and/or categorical predictors. The underlying mechanism used for building the model is called stochastic gradient boosting which was developed and is described by [25]. The behavior of these algorithms is well known [26], adding confidence to these methods. They prove very powerful and are likely playing a major role in future modeling applications worldwide [20, 27], and for the natural resource management and many other applications.

Due to its complex software algorithm, TreeNet is often referred to as a 'black-box', and therefore it was widely rejected by biologist so far. However, there are several advantages of TreeNet compared with other techniques that are often used for building habitat models such as Generalized Linear Models (GLM) or Discriminant Function Analysis (DFA): if applied carefully, (i) it automatically selects the important predictor variables, thus no prior variable selection or data reduction is required, (ii) the results are invariant with regards to modifications of the data such as transformation or rescaling, (iii) the approach can handle missing values automatically and in the best possible, predictive way, and (iv) it is immune to outliers in predictors or the target variable, i.e. if samples are coded incorrectly and the model prediction starts to diverge substantially from observed data, that data will not be used in further updates, and (v) it can be learned and applied very quickly by users allowing us to focus on the real biological questions underneath. TreeNet constructs models conveniently and without time-consuming pre-processing of the data. Furthermore, it is remarkably resistant to overfitting. Hastie [26] calls the approach of multiple additive regression trees, which TreeNet is based on, an effective off-the-shelf procedure for data mining. Further, and because TreeNet can be used in a huge variety of applications beyond data mining, modeling and multiple regressions Salford Systems Ltd. refers to TreeNet as "the closest tool we have ever encountered to a fully automated statistician" [28].

Spatial Scale

The question of spatial scale plays an important role for many wildlife research projects [29, 30] and in particular for the white stork as a wide range species. For a valid assessment, modeling should be executed on different scales, since important influencing factors might be ignored or rather the influence of factors can vary with the scale [31, 32]. For example, the influence of climate data or vegetation on a broad scale can be overridden by competition or other biological processes on local scale [29, 32]. Also influenced by the data availability in our study the modeling was done on two different scales: (i) on a point scale using spatially exact nest locations and (ii) on an administrative district (polygon) scale modeling the density of white storks per district. The definition of the scales was influenced by the data availability but also biology. However, it might not necessarily represent a perfect choice for the white stork. However, exact scales for this research question are not known, yet. And using a small and a large scale is traditionally used and allows for a first assessment of scale [30]. This question has not been addressed before for White Storks.

Habitat Model at Point Scale

White Stork Data at the Point Scale

Two data sets describing the occurrence of white storks in East Prussia were available for the modeling at the point scale:

Data set 1 consisted of 418 individual banding locations derived from the white stork banding data set hosted and maintained by the German ornithological station 'Vogelwarte Radolfzell' (Fig. **1**) which cover the period 1908 to 1944. White storks are usually banded as juveniles in their nest assuring that there were nest sites close to the banding locations [33, 34]. The given coordinates for the nest locations showed an accuracy of +/- 1 spatial minute (corresponds to approx. 1.2 km from east to west and 1.8 km from north to south). More than ¼ of the nest sites (148 of 418) were located in the district of Insterburg due to intensive banding of white storks in this region in the years 1931 to 1942 (and possibly beyond that [33]). When using all 418 presence points to build a model, the variable 'distance to coastline' emerged to be the only relevant important predictor variable for the distribution of nest sites of white storks. In order to better assess for an overestimation of habitat features of the district of Insterburg due to human banding activities, 19 points were chosen at random from the district of Insterburg. This number represented the maximum number of points where banding was exercised in one of the other districts (which was Königsberg-Land). So the number of nest sites used for the modeling as presence locations reads 289, which allows for a more representative presence picture. For absence locations 578 points were distributed at random over the entire study area using the extension 'Hawth's Tools' [35] in ArcMap. These points did not represent a confirmed absence; however, instead they represent pseudo-absences and are the points of the habitat available to nesting white storks. Using pseudo-absences is a common method applied in such studies (see e.g. [22, 32]).

Data set 2 is derived from an inventory of the white stork breeding population done in 1931 (Fig. **2**). Unfortunately, the original data which was hosted by the ornithological station 'Vogelwarte Rossitten' was lost during the Second World War so that the survey at hand can only be traced from the map of the inventory, published in [10]. The number of white stork breeding sites, summarized by community, is shown in the community center. So the map does not show exact nest locations, meaning no real presence/absence reading, but merely an outline on the level of communities [10]. Nevertheless, this presents currently the best available map for this subject. After georeferencing the map was converted into a raster and overlaid with a point shapefile in which points were regularly arranged at a distance of 500 m for the entire study area (for details see [36]). All points on a raster cell indicating a nest location were coded as presence points (value '1'), all other points as absence points (value '0'). Since the representation of the white stork count does not correspond with the exact locations of the nests, but rather places them in the centers of the communities, only points at a distance of 2 km from the next presence point were used as absence points in the modeling. A selection of 5,000 locations from all the presence as well as absence points (a total of 10,000) was randomized, allowing for representative presence and pseudo-absence locations of white stork nest sites.

Environmental Data at Point Scale

For the historical model at point scale, the characterization of the landscape structure was achieved by using reprints of 15 official topographical maps with a scale of 1:100,000 that originated for the main part from the year 1939 (except for two maps from 1941 and one map from 1942). After georeferencing the maps, different habitat features relevant for white storks (forest, lake, watercourse and settlement) were digitized manually using the Editor feature in ArcMap (see Fig. **3**). Land use classes like wetland, pasture/meadow or cropland, which are potentially important for the distribution of the white stork as well, could not be digitized, since they were not clearly defined and distinguishable. For the layers 'forest' and 'lake' all forests and lakes marked on the maps showing an expanse of more than 1 km in any direction were digitized as individual objects. It was assumed that smaller objects have no relevant influence on the large-scale selection of breeding sites for the white stork

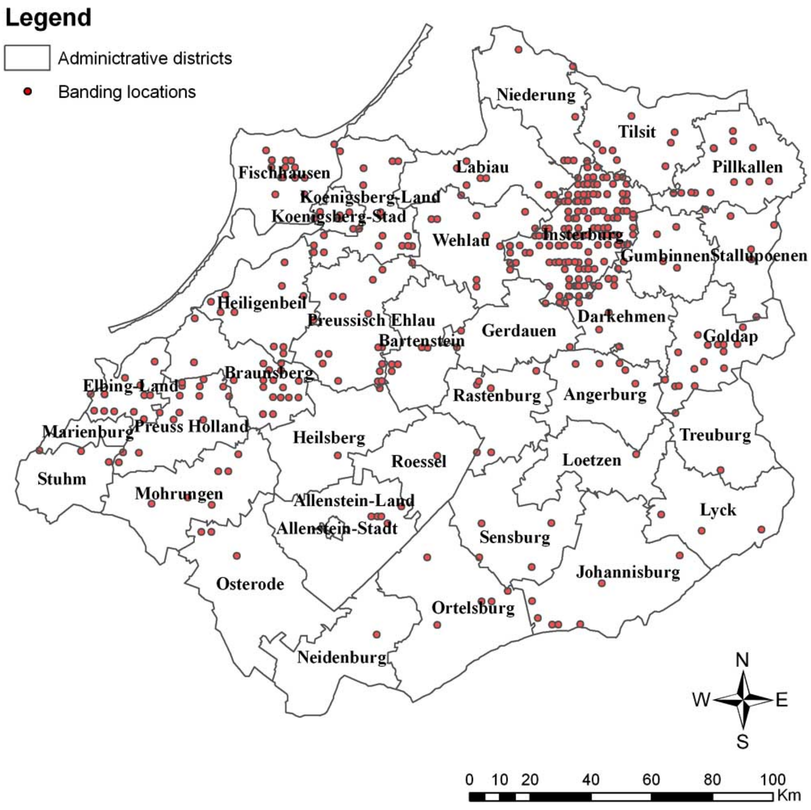

Fig. (1). White stork data set 1: Banding locations (418) from the banding data base of the German ornithological station 'Vogelwarte Radolfzell'.

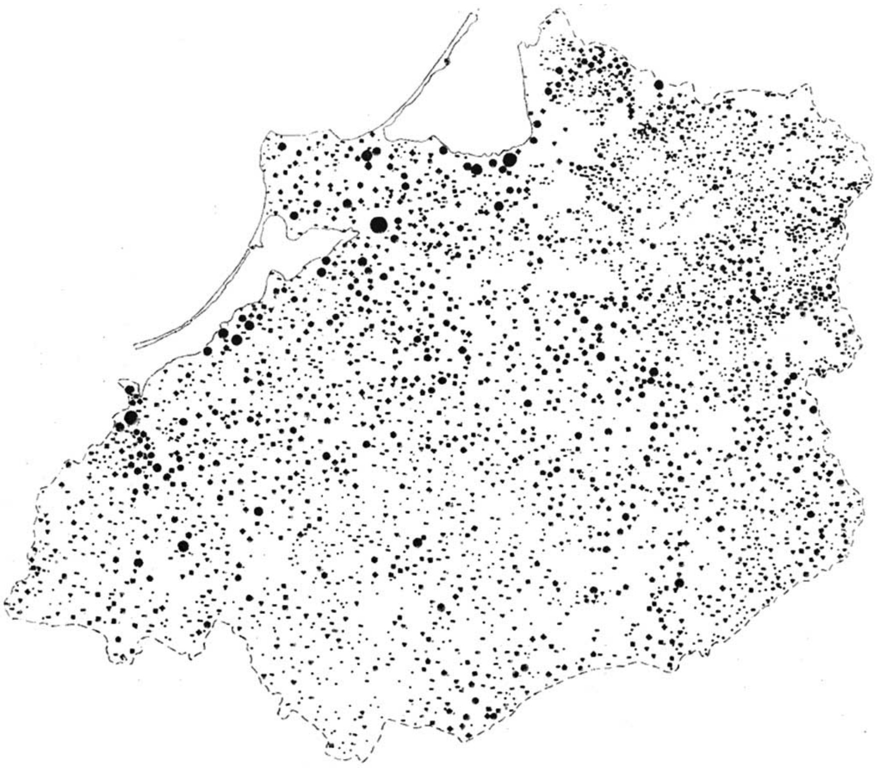

Fig. (2). White stork data set 2: Breeding grounds in East Prussia as surveyed in 1931, published in Schüz (1933).

because foraging trips can reach a radius up to 5 km from the nest [5]. For the layer 'watercourse' any waterbody represented on the maps by a double line (indicating a wider area) was digitized. In order to digitize the settlements according to their size, the settlements were divided into six different categories according to the different font sizes representing the names and the population size of the settlements. To minimize edge and fringe effects in the forthcoming modeling component, all objects that fitted one of the four land use classes were digitized up to about 4 km beyond the boundaries of the survey area (for more details see [36]). As the topographical maps also showed the division of the province of East Prussia in its districts, a layer could be digitized containing 37 administrative districts as polygon features. The districts of Rosenberg and Marienwerder, which are located in the south west of former East Prussia, had to be excluded from the investigation due to a lack of the relevant topographical maps.

Furthermore, a digital elevation model (DEM) derived from the Global Land One-Kilometre Base Elevation (GLOBE) data set of 1999 (available under www.ngdc.noaa.gov/mgg/ topo/globe.html) with a horizontal grid spacing of 30 arc seconds and a digital coastline data set extracted from the freely available online World Vector Shoreline data set of 1990 (WVS, available at the U.S. National Geographic Data Center (NGDC) website http://rimmer.ngdc.noaa.gov/coast/) was used as additional layers for the modeling. Despite the difference of about 50 and 60 years respectively when compared to the year 1939, these data sets were used because they represent the best available data for this study as it can be assumed that any changes which may have occurred will have taken place on a scale which will have no influence on white storks and the creation of the model.

For the modeling a raster with a cell size of 100 m * 100 m was created for each of the layers 'forest', 'lake', 'watercourse' and 'coastline'. Each raster cell contained a value which showed in meters the distance to the next object of the layer. For the layer 'settlement' a classification according to different size categories was taken and the following distance raster created: 'size range 1 to 3', 'size range 1 to 4', 'size range 1 to 5' and 'size range 1 to 6'. For each of the raster the value of the individual cells indicated the proximity to the next settlement of one of the included size categories. This was done because we assumed that localities of different size will have a varying impact on the distribution of white stork nesting sites, and deserves to be tested for its predictive performance. Further, several density raster with a cell size of 100 m * 100 m were generated for the layer 'settlement' using the Density-Tool 'Kernel' of the Spatial Analyst with a radius setting of 5 km. According to different size categories the following raster were generated: 'size range 1 to 4', 'size range 1 to 5' and 'size range 1 to 6'. Each raster cell contained a value indicating how close the settlements were located to one another. So it was possible to test which pool of size ranges and whether density of or distance to settlements showed the greatest effect explanatory on the distribution of breeding white stork. During the modeling process, each time only one of the created distance or density raster was included in the model and the one with the strongest effect on the response variable was selected to create the final model.

For the modeling in TreeNet a TXT file each was created containing all the presence and absence locations for the two

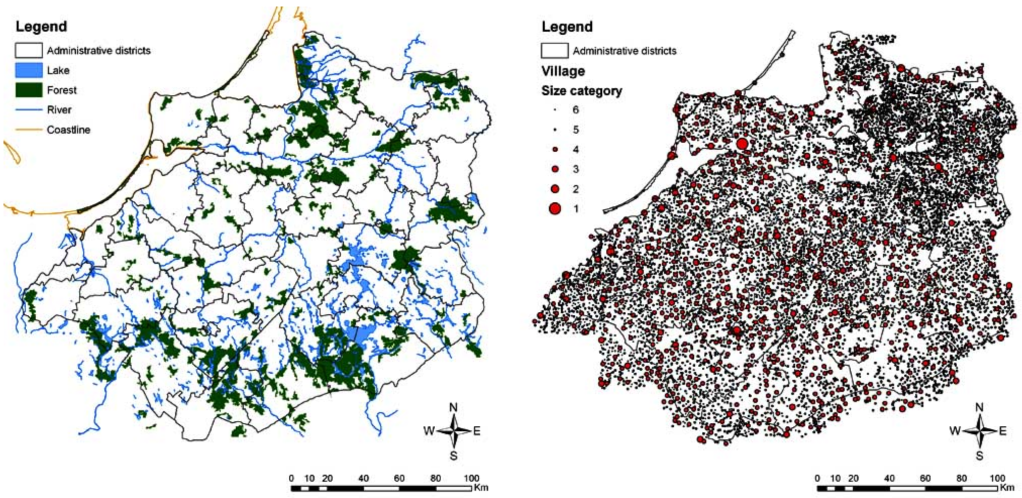

Fig. (3). Study area with the GIS layers (except for the DEM) applied for the modeling: left: Lake-, Forest, River- and Coastline-Layer; right: Settlement-Layer.

white stork data sets. For each predictor variable a column was added specifying a value for every location.

Modeling in TreeNet at Point Scale

Since the target variable was binomial (value ‘1’ for nest locations and value ‘0’ for available/absence points), the algorithm for 'Binary Logistic Models’ was selected in TreeNet.

As output, TreeNet provided a TXT file in which a value between 0 and 1 was listed for every point. To make predictions for the complete study area, a point shapefile was established in which points were regularly placed throughout the study area at a distance of 1 km and the value of the appropriate set of variables was allotted to each of these points. By applying a previously established model in TreeNet (termed grove file), a prediction of an index of relative importance could be made showing for each point a suitability as a white stork breeding site.

Model 1 was created using the predictor variables ‘distance to forest’, ‘distance to lakes’, ‘distance to watercourse’, ‘distance to coastline’, ‘elevation’ and ‘density of settlements, size range 1 to 5’. For the evaluation, 2000 randomly chosen points (1,000 presence and absence points each) of the data set of the white stork count from 1931 (Data set 2) were used as test data set. For the test data set a prediction was carried out using TreeNet and the area under the ROC-Curve (Area under the curve - AUC) was calculated. For that a Delphi program provided by B. Schröder was used (version January 2004), downloadable from the internet under (http://-brandenburg.geoecology.uni-potsdam.de/users/schroeder/download.html). With that program it was also possible to calculate bootstrapped confidence intervals for the AUC values with the percentile method referring to [37].

To create Model 2 the predictor variables ‘distance to forest’, ‘distance to lakes’, ‘distance to watercourse’, ‘distance to coastline’ and ‘elevation’ and ‘distance to settlement, size range 1 to 5’ were used. In order to evaluate Model 2 all 418 nest locations from Data set 1 were used. Since this data set only dealt with presence points, no values for AUC, sensitivity or specificity, could be established as for the evaluation with presence-absence data. That is why instead a Spearman-rank Correlation was used as described in [38]. A prediction was created on a regular point grid (distance between points = 1 km) which covered the complete study area. The predicted values were allotted to ten bins of identical size as described in [38]. In ArcMap the value of the original values of every point was converted to the number of the appropriate bins (1 to 10). Then, the area could be calculated that the individual bins covered (number of points per bin = area in square kilometer per bin). In the following step, the number of the 418 nest locations per bin was calculated and the area-adjusted frequency established (number of points of the testing data sets per bin divided by the area of the respective bin). A Spearman-rank Correlation was calculated with the bin number as the categorical variable and the area-adjusted frequency as the continuous variable. The calculation was carried out with the program R (version 2.3.1). To establish the deviation within the model the data set was divided randomly into five even-sized subsets for which the area-adjusted frequency each was calculated [38]. Another measure to evaluate the performance of a model when only opportunistic data (presence-only) is available is the calculation of the minimal predicted area (MPA). The MPA is the area obtained by considering all raster cells of the study area showing an occurrence. Based on the rule of parsimony the smaller the minimal predicted area the better is the performance of the model (see also [39]).The value above which 90 % of the locations of the test data set were observed is used as a threshold to transform the predicted values for the whole study area into occurrence and non-occurrence (as in [22]).

Habitat Model at Administrative District (Polygon) Scale

White Stork Data at the Administrative District Scale

In the year 1934 an international population assessment was carried out. Figures of the census for the individual administrative districts of the province of East Prussia were published in [40]. The counted number of white stork breeding pairs per 100 km^2 was given for every administrative district (Fig. **4**).

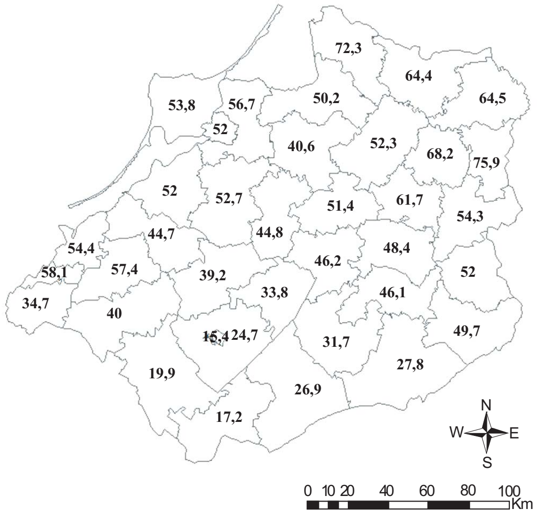

Fig. (4). White stork data set 3: Density of breeding pairs per 100 km^2 for the 37 administrative districts.

Environmental Data at Administrative District Scale

The area per administrative district of the land use types ‘arable land’, ‘pasture land’ and ‘forest’ in the year 1936 could be taken from [24]. In addition the digitized layers ‘settlement’ and ‘lake’ derived from the historical topographic maps as well as the information on the coastline and the DEM were used. The lake layer was converted to a raster covering the whole study area with a cell size of 100 m * 100 m. Subsequently, the percentage of area which was lake (‘water’) could be calculated for every administrative district. For the layer ‘settlement’, the numbers of settlements in the districts were counted. Here, a sub-division was taken, creating three different variables: ‘number of settlements, size range 1 to 4’, ‘number of settlements, size range 1 to 5’, and ‘number of settlements, size range 1 to 6’. The average reading of all the raster cells of the DEM in the individual administrative district was calculated as the variable ‘elevation’. For defining the variable ‘distance to coastline’ the average of all raster cells in the created raster showing the distance to coastline was calculated.

Modeling in TreeNet at Administrative District Scale

The variable ‘number of storks per km^{2}’ served as target variable. Since it concerns a continuous variable, in TreeNet the algorithm for ‘Logistic Regression Models’ was chosen. Because no independent data set was available for evaluating the models, the testing was done in TreeNet using a 10-fold cross-validation. This is recommended for small data sets when one cannot afford to reserve some data for testing [41]. The data set is partitioned into 10 bins. Then, a model is calculated for nine bins while the 10th bin serves as test data set. This is repeated 10 times until every bin was once used as test data. After all 10 folds are completed, the results from each fold are averaged to get a fair test estimate of the all-data model performance [28].

For Model 3 the variables ‘arable land’, ‘pasture land’, ‘forest’, ‘water’, ‘distance to coastline’, ‘elevation’ and ‘number of settlements per km^2, size range 1 to 5’ served as predictor variables. To evaluate the model the mean absolute error (MAE) between observed and predicted values was calculated. MAE is often used as a similar measure than to determine the goodness-of-fit of models [42, 43]:

$$MAE = N^{-1}\sum_{i=1}^{N}| O_i - P_i | \qquad \text{(Eq. 1)}$$

In addition the Coefficient of Efficiency (E) was calculated for better interpreting the results:

$$E = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2} \qquad \text{(Eq. 2)}$$

The Coefficient of Efficiency can have values between minus infinite and 1 (perfect model).

RESULTS

Habitat Models at Point Scale

For each predictor variable, TreeNet offers a relative importance score (Table **1**). „The relative importance score provides a relative measure of each variable’s contribution to the model’s predictive power. The raw importance scores are rescaled so that the most important variable always gets a score of 100. The raw variable importance score is computed as the cumulative sum of improvements of all splits associated with the given variable across all trees up to a specific model size.” [28]. So TreeNet allows a ranking of the used predictor variables according to their importance in the model. However, one should keep in mind that such a ranking is different from using p-values or AICs, because it is derived from a different method and distinct underlying statistical assumption than log-likelihood and parsimony for instance. Table **1** shows the variable importance for the predictor variables used to create the two models at the point scale. The three most important predictors are ‘distance to coastline’, density of settlement’/‘distance to settlement’ and ‘distance to forest’ in both models, but the ranking is exactly reversed.

One option for the interpretation of the model is offered by the partial dependence plots, provided by TreeNet. They show the effect of the respective predictor on the response including the interdependency with the other predictors applied.

In Figs. (**5**) and (**6**) the partial dependence plots for the three most important predictor variables for Model 1 and Model 2 are shown. Comparing the influence of each predictor in the two different models it can be seen that they show similar effects.

The predictor ‘distance to coastline’ (Figs. **5a** and **6c**) shows a positive effect on locations which are located up to a distance of about 30 km (Model 2) to 60 km (Model 1) from the coastline. With increasing distance to the coastline the partial dependence reads negative. Especially in Model 1, a strong negative effect arises. Regarding Model 2 a re-increasing positive influence of the predictor can be seen on locations showing a distance of 130 km to 140 km from the

Table 1. Importance of the Predictor Variables of Model 1 and Model 2

	Relative Importance Score	
Environmental variable	Model 1	Model 2
Distance to coastline	100.00	78.69
Density of/distance to settlement(s)	92.26	97.44
Distance to forest	76.82	100.00
Elevation	57.65	50.19
Distance to lake	47.64	58.25
Distance to watercourse	44.48	56.65

coast with a second maximum at a distance of about 150 km. Regarding the influence of the predictor derived from the layer ‘settlement’ on the response (Fig. **5b** and **6b**), it can be established that the effect on locations located in an area with a high density of settlements (Model 1) or in proximity to a settlement (Model 2) is positive. The partial dependence gets negative if the density of the settlements is lower than 0.1 (Model 1) or if the distance to the next settlement is higher than 1.5 km. Note that for Model 1 the density of settlements was used as predictor variable in contrast to the distance to the next settlement in Model 2, and therefore the plots seem to show a reverse development although the same effect is predicted.

The correlation of the predictor ‘distance to forest’ and the predicted response (Fig. **5c** and. **6a**) result in negative readings at a short distance of about 1 km (Model 2) to 2 km (Model 1) to the next forest. With increasing distance the partial dependence shows positive values.

Evaluation of the Models at Point Scale

For Model 1 the calculated AUC value for the test data set (per 1,000 randomly chosen presence and absence locations of Data set 2) represent 0.790 with a confidence interval of 0.771 - 0.809. An AUC of 0.790 means that when arbitrarily selecting a breeding pair from a presence and an absence location the predicted reading for the presence location is higher than the predicted value for the absence location by 79%. According to [44] a model with an AUC of between 0.7 and 0.8 is to be considered as ‘acceptable’.

For Model 2 no calculation of an AUC value was possible because as independent test data set only the presence-only locations of Data set 1 were available. Therefore, a Spearman-Rank Correlation had to be carried out which showed a value of 0.976, indicating a high performance of the model to predict the observed nest locations. For the determination of the variance within the test data set it was divided into 5 subsets and the medium area adjusted fre-

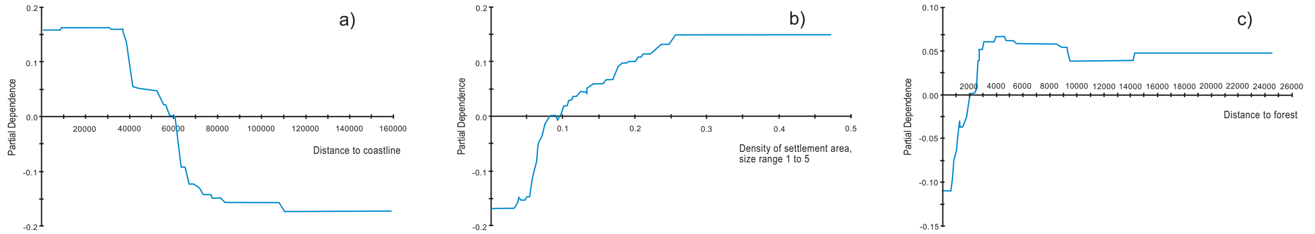

Fig. (5). Partial dependence plots for predictor variables employed in Model 1; **a**) ‘distance to coastline’, **b**) ‘density of settlements, size range 1 to 5’ and **c**) ‘distance to forest’.

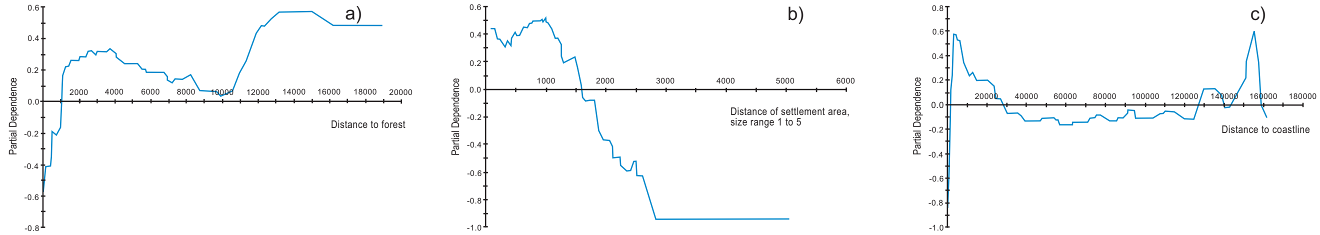

Fig. (6). Partial dependence plots for predictor variables employed in Model 2; **a**) ‘distance to forest’, **b**) ‘distance to settlement, size range 1 to 5’ and **c**) ‘distance to coastline’.

quency as well as the standard deviation were calculated (Fig. **7**).

When calculating the MPA for Model 2 (as in [22]) 75 % of the total survey area lay within the occurrence area (minimal predicted area). This shows that the white stork is predicted to be a widespread species within the study area.

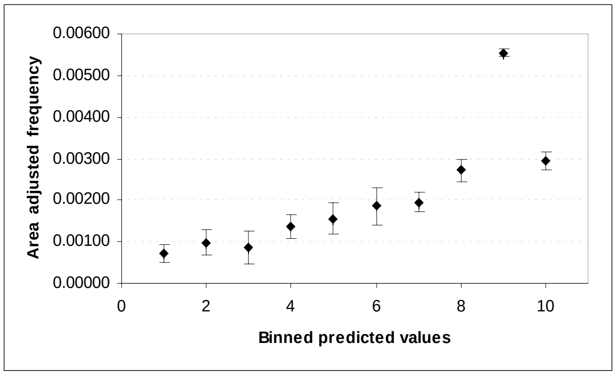

Fig. (7). Average area adjusted frequency with standard deviation per bin for the predicted values of the test data set (divided into five sub-samples).

Fig. (**8**) shows the index of relative importance calculated for the entire study area based on Model 1 and Model 2 respectively. Because the environmental data used to characterize the locations originated mainly from the year 1939 the predictions are related to this year. The predicted relative occurrences are located between 0.29 and 0.74 for Model 1 and between 0.001 and 0.972 for Model 2. Fig. (**8**) explains once again the evidence of the partial dependence plots and the relative importance score of the predictor variables: In Model 1 the variable 'distance to coastline' is the most important predictor and is most influential in the distribution of nest locations. Thus, in a strip of about 60 km along the coastline high occurrence indices of nest locations of the white stork are predicted. In contrast for the southern section of the survey area, situated at a greater distance from the coastline, low occurrence indices were predicted. Applying Model 2 high values were predicted in a strip of about 20 km along the coastline as well as at a distance of about 150 km from the coastline at the southern edge of the study area. A further region with high indices of relative importance is predicted in the northwest of the survey area by Model 2. Throughout the complete study area, locations on lakes and in forests or rather in their close proximity, showed as being virtually unsuitable for breeding in both models. Extensive forests and water bodies are situated especially in the south of the study area and correspond with the areas of a low predicted index of the presence of nest sites of white storks (compare Fig. **3**). In these regions the concentration of the settlements is minimal, too.

Habitat Model at Administrative District Scale

Table **2** shows the variable importance of the applied predictor variables calculated by TreeNet. Both predictor variables 'number of settlements per km^2, size range 1 to 5' and 'percentage of pasture' had considerable bearing on the modeling. Notably less important were the two variables 'percentage of arable land' and 'distance to coastline'.

Table 2. Importance of the Predictor Variables of Model 3

Predictor Variable	Relative Importance Score
Number of settlements per km^2	100.00
Percentage of pasture	90.67
Percentage of forest	54.40
Percentage of water bodies	53.97
Elevation	40.16
Percentage of arable land	27.29
Distance to Coastline	23.77

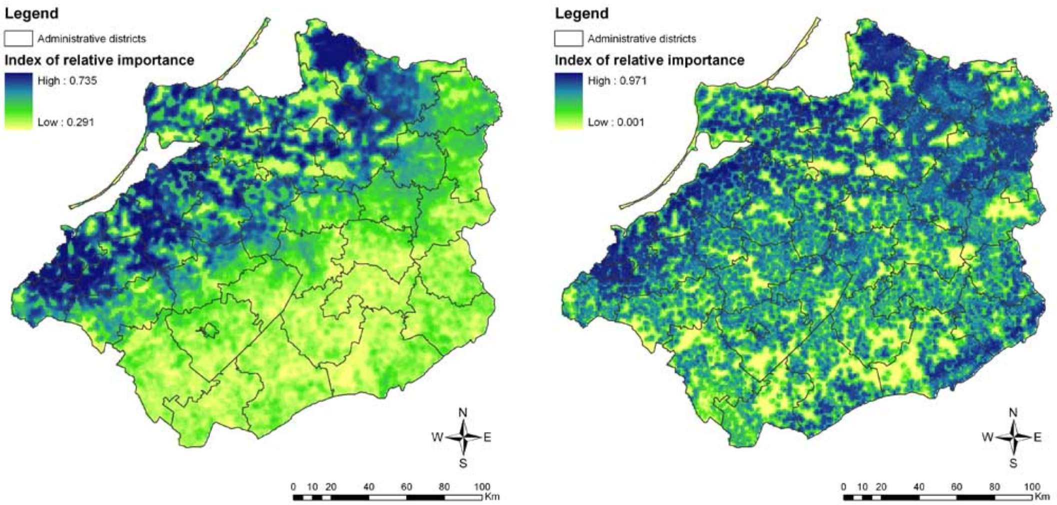

Fig. (8). Predicted index of relative importance for the year 1939 applying Model 1 (left) and Model 2 (right).

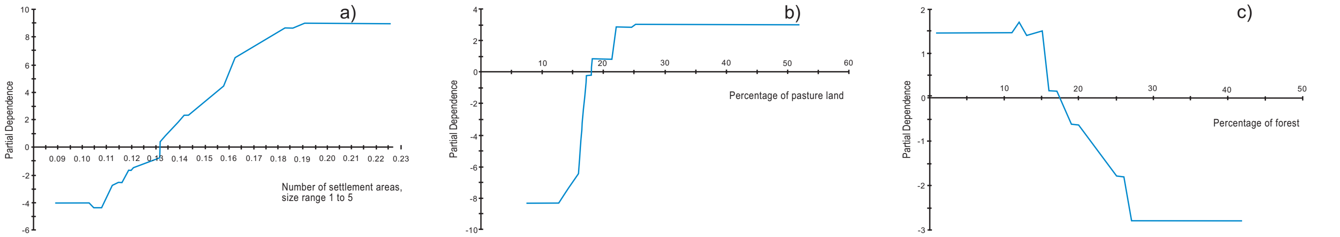

Fig. (9). Partial dependence plots for the three most important predictor variables employed in Model 3; **a**) ‘number of settlements, size range 1 to 5’, **b**) ‘percentage of pasture land’ and **c**) ‘percentage of forest’.

For the graphic visualization of the effect which the individual predictor has on the predicted response the partial dependence plots are shown in Fig. (**9**). The largest influence on the predicted response was contributed by the predictor ‘number of settlements per km^2 (Fig. **9a**). Up to a number of 0.11 settlements per square kilometer a negative partial dependence of -4 was found. With a growing number of settlements the partial dependence increased continuously reaching zero point at about 0.13 settlements per square kilometer and had a high positive reading of 9 from about 0.19 onwards. The predictor ’percentage of pasture’ (Fig. **9b**) showed a high negative partial dependence of -0.8 for districts that showed a reading lower than 15 %. From there the partial dependence rose considerably, reaching zero point at about 17 % and showing a positive partial dependence of 3 as from a percentage of 22 % pasture in the district. The predictor variable ’percentage of forest’ (Fig. **9c**) presents us with a high partial dependence for districts with a percentage of up to 15 % forest. The partial dependence declined with increasing percentage, reaching zero point in districts with roughly 17 % forest and showed a negative reading of nearly -3 as from 25 % forest in the district.

Model 3 had a mean absolute error (MAE) of 5.22. The Coefficient of Efficiency had a value of 0.80.

DISCUSSION

The White Storks’ Selection of Nesting Sites in Former East Prussia

Our knowledge, so far, about habitat selection of white storks in former East Prussia has been very limited. Here we offer for the first time results and methods on how to overcome these problems towards (spatial) generalization. We are proposing quantitative habitat associations and predictions to be tested and improved by the scientific community for better management of this precious resource.

In Model 1 and Model 2 the three environmental variables ‘forest’, ‘coastline’ and ‘settlement’ are most influential in the white stork’s choice of nesting place in East Prussia (Table **1**). In Model 3, which models the density of white stork breeding pairs in the different administrative districts, the variables ‘settlement’ and ‘forest’ also show a great effect on the modeling with a relative importance score of 100.00 or rather 54.40 (Table **2**). The variable ’coastline’ plays only a subordinate role. The variable ’pasture’, which depicts how much area per administrative district is exploited agriculturally as pasture or meadow, is the second most important predictor in Model 3 having a relative importance score of 90.67. For the Model 1 and Model 2 no data on agricultural usage was available and so this aspect has to go unconsidered for now but deserves more study. Comparing the models on the two scales (Model 1 and 2 vs. Model 3) no relevant difference could be found about the effect of the used predictor variables.

Surveys on the feeding ecology of the white stork in agriculturally effected areas have shown that to a great extent the white stork selects pasture and meadow with a low height of vegetation for foraging [5, 13 ,45]. Agricultural crop land appear to play an inferior role and shows only a greater availability of food supply at the time of working the land and after the harvest, i.e. at the end of the nestling phase [14, 46].

The ‘forest’ predictor is also of major importance in this model. Thus, the vicinity to forests has a negative effect on the probability of the white stork occurrence. According to the created models dense woodland must be regarded as inadequate habitat for the white stork, since the optically orientated stalking predator has apparently difficulties in finding food supply there [10, 47]. Aerodynamic and predation-related reasons could further contribute to this pattern.

Beside the availability of food supply the choice of nesting ground is also determined by the vicinity to human settlements, since nests are often built on emerging buildings [48]. Thus, in the year 1934 in former East Prussia (except for the districts of Rosenberg and Marienwerder) 92 % of the storks nested on roof tops [40]. This explains the great influence the variable ‘settlement’ has in the three models presented here. Nesting white storks are associated with a ‘light’ human footprint.

The influence of the distance to coastline in the white stork’s choice of nesting ground (in Model 1 and Model 2) has not been described yet. In particular in Model 1 this variable influences the distribution of the white stork considerably and leads to the prediction of their high occurrence near the coastline. However, considering the result of Model 2, there is a ‘high abundance band’ with a high predicted occurrence ‘probability’ for a distance of 160 km from the coastline. This is described for the first time. The great influence of the coastline could be based on a combination of several factors. For instance, in East Prussia a specific landscape composition parallel to the coastline can be found. About 150 km from the coastline, the Baltic land ridge with an elevation of maximum 313 m above sea level parallels the coast. South of the Baltic land ridge extensive sandy territory extends with broad regions of lake-land and forest probably unsuitable for breeding, whereas in the north of East Prussia loamy soils can be found with more pastures and meadows

[10, 49] providing a good food supply for the white stork. Therefore, the distance to the coastline could be exploited as a proximal factor for the interaction of the different variables that influence the white stork’s choice of nesting ground. However, this may only be applicable to East Prussia with its characteristic landscape described here and should not be applied to other regions without care. The specific structure of the data could be a further reason for the extremely high influence of the ’coastline’ predictor on the occurrence of the white stork in Model 1. Half of all the items of Data set 1 are within a distance of 41.5 km to the coast (the maximum distance to the coast is 164 km). This could be due to a more intensive banding activity in northern East Prussia and might lead to an over-estimation of the quality of the nesting grounds near the coast. However, the modeling approach chosen in this study should consider the representative ecological niche and help to overcome such problems of survey effort in space (see [21]).

Restrictions and Constraints of the Models

Both data sets applied for the modeling on a point scale (Data sets 1 and 2) contained no real absence locations to show at which locations the white stork is actually not nesting (=confirmed absences). Data set 1 only showed presence locations which registered as retraced locations by banded white storks and expanded into the banding data set of the ornithological station ‘Vogelwarte Radolfzell’. For the modeling, 578 locations were randomly scattered over the study area (pseudo-absence locations). The pseudo-absence locations may also arbitrarily contain locations suited as breeding sites for the white stork [50].

Although Data set 2 is based on a nearly complete mapping of all the white stork breeding grounds in East Prussia in 1931, a representation was chosen, however, in which the number of nesting grounds were both combined and presented in one community centre. This is because the exact locations (in coordinates) can only be deduced with some deviation. Locations were chosen at random that were at a distance of at least 2 km from the registered nesting site. Confirmed absence locations may be at sites where a pair of white storks has been breeding. This can lead to a decrease in the performance of the model, if absence locations turn out in fact to be suitable nesting grounds [41, 50].

Furthermore, and as typical in such models [32], it must be assumed that not all biologically relevant variables are included in the modeling (e.g. see [31]). This may happen because important variables are not recognized as such or also because they cannot be surveyed (e.g. due to complexity, time and effort). This is a common feature in current GIS models (e.g. see [51]) but supposed to improve with further data availability. Here, modeling methods were initiated, and a digital culture was set up within the white stork community emphasizing on the importance of such freely available GIS and white stork-related data. Especially in historic modeling only a limited choice of variables is commonly possible to apply, since only data can be used that is already available. A survey of data which have proven to be relevant for the species to be modeled is not always possible with hindsight. Instead, historical data sets often originate from museums or archives (see [52] for overview and applications). Often, it is not anymore known who did the surveys and how the data was surveyed [20, 22]. This includes a multitude of sources of error concerning the quality of data applied [20, 53], and therefore limits a direct inference. The predictors are thus primarily showing correlations [32], and more hypothesis and on the ground work is required to assess their biological validity and mechanisms further.

CONCLUSION

For the first time, a quantitative prediction model on white storks was presented, and assessed for its predictive performance. Based on freely available data, this study shows that opportunistic, historical data sets can be used successfully with GIS and with a robust machine-learning model method (TreeNet) to derive species habitat relation models and new biological knowledge.

The models generated here demonstrate that the vicinity of human settlements as well as the availability of sufficient feeding habitats (grass land with low vegetation, no close forest regions) have a strong influence on the nest site presence of the white stork in this region. This finding matters for future planning and management of landscapes if white storks are to be maintained.

Without further testing, the models generated are only valid for the region of former East Prussia so far. However, in order to transfer the relations found here between the distribution of the white stork and the state of the habitat onto other regions, the models would need to be further assessed using available data sets surveyed in these other regions. Model assessments are crucial for model validity, trustworthiness and improvements. The applicability and value of this approach is rather large because a global and quantitative white stork nesting model could be achieved, which would likely improve much of its sustainable and future management.

The models can be improved using additional variables. Thus, the quality of the habitat can be more closely described from the ecological perspective of the white stork with data on prey availability, the level of precipitation in June (when the young white storks are especially susceptible to wet conditions) or the temperature for example.

To create species distribution models with the highest possible accuracy it is important that all relevant stakeholders (e.g. public authorities, archives, nature conservation agencies, NGOs and scientists) co-operate and mutually exchange data with the public: only then can it be achieved that already existing data is exploited more effectively than currently done. The ideal state would be a freely accessible online data bank in which relevant data could be queried in digital form. This would further contribute to an improved and sustainable management of white storks, their habitats world-wide and natural resources as a whole.

ACKNOWLEDGMENTS

We are grateful to Wolfgang Fiedler, head of Vogelwarte Radolfzell, who provided us with data from the banding database, to Kai-Michael Thomsen of the Michael-Otto-Institut at NABU - Naturschutzzentrum Bergenhusen, (Germany), to Christoph Hinkelmann, managing director of the Ostpreußisches Landesmuseum (East Prussian County Museum) in Lüneburg (Germany) and to Gerd Dahms. We thank the team of Salford Systems Ltd. for the provision of the software TreeNet and their flexible support, as well as Dr. Boris Schröder (University of Potsdam, Germany) for the gratuitous provision of his personally created program 'ROC_AUC' online. FH acknowledges S. Linke, A. Jacob, EWHALE lab and IAB/UAF for kind support, ideas and communications.

REFERENCES

[1] Schimkat J. Untersuchungen der Populationsdynamik von Regionalbeständen Ostziehender Weißstörche (Ciconia ciconia) mittels eines Simulationsmodells. Dissertation, Universität Potsdam: Germany 2006.

[2] Dallinga JH, Schoenmakers S. Regional decrease in the number of white storks (Ciconia ciconia) in relation to food resources. Col Waterbirds 1987; 10: 167-77.

[3] Daly HE. Beyond growth: the economics of sustainable development. Boston, Massachusetts, USA: Beacon Press 1997.

[4] Czech B, Krausman PR, Devers PK. Economic associations among causes of species endangerment in the United States. Bioscience 2000; 50: 593-601.

[5] Johst K, Brandl R, Pfeifer R. Foraging in a patchy and dynamic landscape: human land use and the white stork. Ecol Appl 2001; 11: 60-9.

[6] Blotzheim Gv. Handbuch der Vögel Mitteleuropas. Aula-Verlag: Wiesbaden 1987.

[7] Creutz G. Der Weißstorch. Westarp Wissenschaften: Ziemsen 1988.

[8] Profus P. Bestandsveränderungen des Weißstorchs Ciconia ciconia in Polen. Charadrius 2005; 41: 12-20.

[9] Braun M. Die Nistweise des Storches. Schr Phys-Ökonom Ges Königsberg 1908; 49: 280-90.

[10] Schüz E. Der Bestand Des Weißen Storches (Ciconia c. ciconia*)* in Ostpreußen 1931. Verh Orn Ges Bay 1933; XX: 191-225.

[11] Seber GAF. A review of estimating animal abundance. Biometrics 1986; 42: 267-92.

[12] Buckland ST, Anderson DR, Burnham KP. Introduction to distance sampling: estimating abundance of biological populations. New York, USA: Oxford University Press 2001.

[13] Böhning-Gaese K. Zur Nahrungsökologie des Weißstorches (Ciconia ciconia) in Oberschwaben: Beobachtungen an zwei Paaren. J Ornithol 1992; 133: 61-71.

[14] Bairlein F, Henneberg HR. Der Weißstorch (Ciconia ciconia) im Oldenburger Land. Isensee: Oldenburg 2000.

[15] Van Den Bossche W, Berthold P, Kaatz M, Nowak E, Querner U. Eastern european white stork populations: migration studies and elaboration of conservation measures. BfN Scripten 66: Bonn 2002.

[16] Berthold P, Kaatz M, Querner U. Long-term satellite tracking of white stork (Ciconia ciconia) migration: constancy versus variability. J Ornithol 2004; 145: 356-9.

[17] Chernetsov N, Berthold P, Querner U. Migratory orientation of first-year white storks (Ciconia ciconia): inherited information and social interaction. J Exp Biol 2004; 207: 937-43.

[18] Huettmann F. Research and management viewpoint: databases and science-based management in the context of wildlife and habitat: toward a certified ISO standard for objective decision-making for the global community by using the internet. J Wildlife Manage 2005; 69: 466-72.

[19] Guisan A, Thuiller W. Predicting species distribution: offering more than simple habitat models. Ecol Lett 2005; 8: 993-1009.

[20] Elith J, Graham H, Anderson P, *et al*. Novel methods improve prediction of species' distributions from occurrence data. Ecography 2006; 29: 129-51.

[21] Kadmon R, Farber O, Danin A. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecol Appl 2004; 14: 401-13.

[22] Engler R, Guisan A, Rechsteiner L. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J Appl Ecol 2004; 41: 263-74.

[23] Barran FR. Städte-Atlas Ostpreußen. Rautenberg-Verlag: Leer 1988.

[24] Grenzlandverlag G. Statistisches Handbuch für die Provinz Ostpreußen 1938. Boettcher: Schloßberg & Leipzig 1938.

[25] Friedman JH. Stochastic gradient boosting. In: technical discussion of treenet 1999. Available from: http://www.salford-systems.com/treenet.html

[26] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer 2003.

[27] Popp JN, Neubauer D, Paciulli L, Huettmann F. Using treenet for identifying management thresholds of mantled howling monkeys' habitat preferences on ometepe island. Nicaragua, on a tree and home range scale. J Med Biol Sci 2007; 1: 1-14.

[28] Salford Systems Ltd. Treenet 2.0 user manual - software treenet. San Diego, CA 2003 (cited 2008 Feb 12). Available from: http://www.salfordsystems.com/

[29] Wiens JA. Spatial scaling in ecology. Funct Ecol 1989; 3: 385-97.

[30] Huettmann F, Diamond AW. Large-scale effects on the spatial distribution of seabirds in the northwest Atlantic. Landscape Ecol 2006; 21: 1089-108.

[31] Gottschalk, T, Huettmann F, Ehlers M. Thirty years of analysing and modelling avian habitat relationships using satellite imagery data: a review. Int J Remote Sens 2005; 26: 2631-56.

[32] Manly BFJ, McDonald LL, Thomas DL, McDonald TL, Erickson WP. Resource selection by animals: statistical analysis and design for field studies. Boston: Kluwer 2002.

[33] Hornberger F. Einige Ergebnisse Zehnjähriger Planarbeit Im "storchforschungskreis Insterburg" der Vogelwarte Rossitten. J Ornithol 1943; 91: 341-55.

[34] Sproll A. Zugverhalten und Mortalität des Weißstorches (Ciconia ciconia) nach ringfunden der Vogelwarte Rositten (Ostpreussen). Diploma thesis, fachhochschule für technik. Wirtschaft und Sozialwesen Zittau/Görlitz 2000.

[35] Beyer HL. Hawth's analysis tools for ArcGIS 2004. Available from: http://www.spatialecology.com/htools

[36] Wickert C. Breeding white storks (Ciconia ciconia) in former east prussia: comparing predicted relative occurrences across scales and time using a stochastic gradient boosting method (treenet), GIS and public data. Diploma thesis, Universität Potsdam: Germany 2007.

[37] Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. Biometrics 1997; 53: 603-18.

[38] Boyce MS, Vernier PR, Nielsen SE, Schmiegelow FKA. Evaluating resource selection functions. Ecol Mod 2002; 157: 281-300.

[39] Guisan A, Broennimann O, Engler R, *et al.* Using niche-based models to improve the sampling of rare species. Conserv Biol 2006; 20: 501-11.

[40] Tischler F. Die Vögel Ostpreußens. Ost-Europa-Verlag: Königsberg & Berlin 1941.

[41] Fielding AH, Bell JF. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 1997; 24: 38-41.

[42] Legates DR, McCabe JW Jr. Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. Water Resour Res 1999; 35: 233-41.

[43] Hall MJ. How well does your model fit the data? J Hydroinform 2001; 3: 49-55.

[44] Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons 2000.

[45] Struwe B, Thomsen KM. Untersuchungen zur Nahrungsökologie des Weißstorches (Ciconia ciconia, L. 1758) in Bergenhusen 1989. Corax 1991; 14: 210-38.

[46] Bäßler R, Schimkat J, Ulbricht J. Artenschutzprogramm Weißstorch in Sachsen. Sächsisches Landesamt für Umwelt und Geologie. Dresden 2000.

[47] Latus C, Kujawa K, Glemnitz M. The influence of landscape structure on white stork's Ciconia ciconia nest distribution. Acta Ornithol 2000; 35: 97-102.

[48] Schulz H. Der Weißstorch - Lebensweise und Schutz. Naturbuch-Verlag: Augsburg 1993.

[49] Hinkelmann C. Der Weißstorch (Ciconia ciconia) im Ehemaligen Ostpreußen. Blätter Naumann-Museum 1995; 15: 24-52.

[50] Pearce J, Boyce MS. Modelling distribution and abundance with presence-only data. J Appl Ecol 2006; 43: 405-12.

[51] Guisan A, Zimmermann NE. Predictive habitat distribution models in ecology. Ecol Mod 2000; 135: 147-86.

[52] Graham CH, Ferrier S, Huettmann F, Moritz C, Peterson AT. New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol Evol 2004; 19: 497-503.

[53] Huettmann F. Constraints, suggested solutions and an outlook towards a new digital culture for the oceans and beyond: experiences from 5 predictive GIS models that contribute to global management, conservation and study of marine wildlife and habitat. In: Berghe VE, Appeltans W, Costello MJ, *et al.*, Eds. Proceedings 'ocean biodiversity informatics' an international conference on marine biodiversity data management, Hamburg, Germany, 29 November-1 December 2004 IOC workshop report BSH/ VLIZ special publication 37, 2007; pp. 49-61.

Received: August 03, 2009 Revised: October 10, 2009 Accepted: October 15, 2009