Towards combining occurrence and abundance distribution models of Great Bustard for conservation: A global research template from Bohai Bay?

Chunrong Mi [1] , **Falk Huettmann** [2] , **Rui Sun** [3] , **Yumin Guo** [Corresp. 1]

[1] College of Nature Conservation, Beijing Forestry University, Beijing, China
[2] Department of Biology and Wildlife, Institute of Arctic Biology, University of Alaska - Fairbanks, Fairbanks, Alaska, United States
[3] Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, University of Chinese Academy of Sciences, Beijing, China

Corresponding Author: Yumin Guo
Email address: guoyumin@bjfu.edu.cn

Species distribution models (SDMs) have become important and essential tools in conservation and management. However, SDMs built with count data, commonly referred to as species abundance models (SAMs), are still less used so far. SDMs are increasingly used now in conservation decisions, whereas SAMs are still not widely employed. Species occurrence and abundance do not frequently display similar patterns, often they are not even well correlated. This leads to an insufficient or misleading conservation. How to combine information from SDMs and SAMs all together for unified conservation remains a challenge. In this study, we put forward for the first time a priority protection index (PI). The PI combines the prediction results of occurrence and abundance models. We used the best-available presence and count records for an endangered farmland species, Great Bustard (*Otis tarda dybowskii*) in Bohai Bay, China, as a case study. We then applied the advanced Random Forest algorithm (Salford Systems Ltd. implementation), a powerful machine learning method, with eleven predictor variables to forecast the spatial occurrence as well as the abundance distribution. The results show that the occurrence model had a decent performance (ROC: 0.77) and the abundance model had a RMSE 26.54. It is of note that environmental variables influenced bustard occurrence and abundance differently. We found that occurrence and abundance models display different spatial distribution patterns. Still, combining occurrence and abundance indices to produce a priority protection index (PI) used for conservation could guide the protection of the areas with high occurrence and high abundance (e.g. in Strategic Conservation Planning). Due to the widespread use of SDMs and the rel. easy subsequent employment of SAMs these findings have a wide relevance and applicability, worldwide. We promote and strongly encourage to further test, apply and update the priority protection index (PI) elsewhere in order to explore the generality of these findings and methods readily

available now for researchers.

Towards combining occurrence and abundance distribution models of Great Bustard for conservation: A global research template from Bohai Bay?

Mi Chunrong[1,2], Huettmann Falk[3], Sun Rui[2], Guo Yumin[1,*]

[1]College of Nature Conservation, Beijing Forestry University, P.O. Box 159, Beijing 100083, China;

[2]Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, University of Chinese Academy of Sciences, Beijing 100101, China

[3]EWHALE Lab, Department of Biology and Wildlife, Institute of Arctic Biology, University of Alaska Fairbanks (UAF), 419 Irving I, P.O. Box 757000, AK 99775, USA

Corresponding author: Yumin Guo guoyumin@bjfu.edu.cn

College of Nature Conservation, Beijing Forestry University, P.O. Box 159, Beijing 100083, China;

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3240v1 | CC BY 4.0 Open Access | rec: 11 Sep 2017, publ: 11 Sep 2017

ABSTRACT

Species distribution models (SDMs) have become important and essential tools in conservation and management. However, SDMs built with count data, commonly referred to as species abundance models (SAMs), are still less used so far. SDMs are increasingly used now in conservation decisions, whereas SAMs are still not widely employed. Species occurrence and abundance do not frequently display similar patterns, often they are not even well correlated. This leads to an insufficient or misleading conservation. How to combine information from SDMs and SAMs all together for unified conservation remains a challenge. In this study, we put forward for the first time a priority protection index (PI). The PI combines the prediction results of occurrence and abundance models. We used the best-available presence and count records for an endangered farmland species, Great Bustard (*Otis tarda dybowskii*) in Bohai Bay, China, as a case study. We then applied the advanced Random Forest algorithm (Salford Systems Ltd. implementation), a powerful machine learning method, with eleven predictor variables to forecast the spatial occurrence as well as the abundance distribution. The results show that the occurrence model had a decent performance (ROC: 0.77) and the abundance model had a RMSE 26.54. It is of note that environmental variables influenced bustard occurrence and abundance differently. We found that occurrence and abundance models display different spatial distribution patterns. Still, combining occurrence and abundance indices to produce a priority protection index (PI) used for conservation could guide the protection of the areas with high occurrence and high abundance (e.g. in Strategic Conservation Planning). Due to the widespread use of SDMs and the rel. easy subsequent

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3240v1 | CC BY 4.0 Open Access | rec: 11 Sep 2017, publ: 11 Sep 2017

employment of SAMs these findings have a wide relevance and applicability, worldwide. We promote and strongly encourage to further test, apply and update the priority protection index (PI) elsewhere in order to explore the generality of these findings and methods readily available now for researchers.

Keywords: conservation decision, occurrence model, abundance model, Great Bustard (*Otis tarda dybowskii*), machine learning method, Random Forest

INTRODUCTION

The knowledge of species occurrence and abundance distribution makes for a fundamental information for conservation biology (VanDerWal et al., 2009; Drew et al., 2011; Primack, 2012; Johnston et al., 2015). Understanding how environmental factors are related to species occurrence and abundance distribution explicit in time and space represent a priority in current biodiversity conservation (Drew et al., 2011; Martín et al., 2012).

Species distribution models (SDMs) are empirical ecological models that relate species observations to environmental predictors (Guisan & Zimmermann, 2000); usually that is done with machine learning algorithms (Drew et al., 2011, see Mi et al., 2017 for an application). They have become important and essential tools in ecology, biogeography, climate change research, conservation and management based on their spatial occurrence prediction (Peterson et al., 2002; Guisan & Thuiller 2005; Elith et al., 2006; Araújo & New 2007; Mi et al., 2016). SDMs built with count data are called species abundance models (SAMs) (Elith & Leathwick 2009; Barker et al., 2014; see Yen et al 2004 for an application). SAMs are still less commonly used yet, despite their greater information for conservation and management. But increasing attention has been paid to these problems in recent years (e.g. Yen et al., 2004; Martín et al., 2012; Howard et al., 2015; Ashcroft et al., 2017; Fox et al., 2017).

In the past, spatial conservation decisions and plans are usually just based on SDMs (e.g. Suárez-Seoane et al., 2008; Gray et al., 2009; Adams et al., 2016; Mi et al., 2016). However, despite statements by Newton (2008), many scholars found species occurrence and abundance distribution not to display similar patterns (Yen et al., 2004; Karlson et al., 2011; Yin & He 2014; Johnston et al., 2015). Therefore, conservation decisions only based on SDMs predictions are insufficient and may even be misleading, so do SAMs. In the future, one time-critical challenge and associated progress will be centered how to combine the useful information that SDMs and SAMs each offer for conservation.

In this study we chose the endangered Great bustard (*Otis tarda dybowskii*) wintering in Cangzhou at the North China Plain near Bohai bay as a case study. This area is one of the most important wintering grounds for this species (about 300 individuals, c.13.6~20.0 % of China's total wintering population (Goroshko 2010; Meng 2010). Using the Great Bustard as a case study would contribute to our conservation knowledge about habitat use of a threatened farmland species and for a better policy. By studying not only the spatial occurrence and the abundance patterns, but also combining these two model types together as a role model for predictive modeling and its inference would potentially have wider conservation implications. Our overall objective of this research was to (1) assess and develop models to predict accurately the patterns of bustard occurrence and abundance; (2) infer on environmental variables that influence occurrence and abundance of this species; (3) combine occurrence and abundance models as a new contribution to conservation decisions; and (4) investigate the overall relationship among predicted occurrence, predicted abundance and observed abundance. Well-tested and suited methods from this research could be useful for the conservation of Great Bustard, but also other rare species and biodiversity in general where SDMs and SAMs can be employed.

MATERIALS AND METHODS

Study area

This study was conducted at the wintering grounds of endangered Great Bustards in Cangzhou, southeast of the Heibei Province in the wider Bohai Bay (Fig. 1). It is located at 38°12′57″ - 38°36′51″ latitude and at 116°50′48″ - 117°24′03″ longitude in the warm temperate, semi-humid monsoon climate zone, which features the slightly marine climatic characteristic of the Bohai Sea region. The topographical and climate condition varies little in the study area. The total study area is 2,191.4 km^2, consisting of farmland (1,675.1 km^2; 76.4%), residential area (330.5 km^2, 15.1%), open water (23.5 km^2; 1.1%) and other unspecified land uses (e.g. home lots, sheds).

Put Fig. 1 Here

Most of the farms in this region produce cereal, which is grown in a 2-year rotation system. In the first year, winter cereal is cultivated from early September to the end of April the following year. Then, corn is cultivated between the end of April to early September of the same year. The study area was chosen (Fig. 1) because of its large numbers (about 300 individuals, c.13.6 ~20 % of China's total wintering population (Goroshko 2010; Meng 2010). This area is the world's largest wintering ground of the endangered *O. t. dybowskii*. This area is representative of the typical farmland situation in the North China Plain. In addition, accurate Great Bustard census data, geographic information system (GIS) data coverages and satellite imagery were readily available.

Bird census data

Spatial occurrence and abundance data for Great Bustards were used to develop models. A Great Bustard census was winter survey conducted during November 2013 to March 2014. In the study area, we travelled with a small four-wheel-drive tractor along fixed routes, using speeds of 10-30 km/hour. Our team consisted of two experienced observers (one surveyor and one local resident) counting bustards and with a good knowledge of the area to be surveyed. When a flock was found, we drove slowly and stopped on the location at a 100 - 500 m distance from bustard flocks, recording its size, location, habitat type and basic behavior. This resulted in a good detection of birds and flocks in the study area because birds can be seen already from long distances (~3km) but also when flying away. The actual animal coordinates were obtained by Google Earth when combing it with our recorded location. Each census was done from dawn to dusk. During the study, we identified 94 bustard sites in the study area. To our knowledge, this census data were the best available ones in China for bustards.

GIS environmental layers

Based on environmental conditions in our study area, we selected eleven habitat and landscape (environmental) variables to construct models predicting occurrence and abundance (Table 1). In order to obtain these variables, we acquired the basemap from Google Earth (using Daogle, an open source software made by a Chinese individual http://www.daogle.com/; as used and explained in Mi et al., 2014) and derived otherwise unavailable high resolution landscape inventory information about open water pools, rivers, residential areas, national roads, provincial roads, expressway, farmland road, ditch and farmland areas from the base map. Next we constructed a distance layer for these variables (except for the farmland area) using the Euclidean Distance tool in ArcGIS 10.1 with a 30 m×30 m spacing. This high pixel resolution was chosen in order to be consistent with remote sensing variable resolution we used.

Satellite images

A range of the best cloud-free HJ-1A/B (HuanJing (HJ)) satellite images (http://218.247.138.121/DSSPlatform/index.html#) at a 30 m×30 m resolution was obtained for each month for November 2013 to March 2014 in order to calculate the normalized difference vegetation indices (NDVI) signature for each pixel. The HJ-1A/B CCD data were run for radiometric calibration, atmospheric correction and geometric correction in order to obtain surface reflectance data and subsequent NDVI data. Radiometric calibration was finished using 2014 HJ-1A/B CCD absolute radiometric calibration coefficients provided by the China Centre for Resources Satellite Data and Application. For this study, we used maximum and mean NDVI to represent the vegetation condition (Osborne et al., 2001).

Put Table 1 Here

Model development

We employed an advanced machine learning technique, Random Forest, to model the occurrence as well as abundance distribution of Great Bustards. Breiman (2001)'s Random Forest implementation in SPM7 by Salford Systems Ltd is robust to over-fitting and is widely recognized to produce very good predictive models. Hence, it is increasingly applied to species distribution modelling (Cutler et al., 2007; Drew et al., 2011; Mi et al., 2016 for an application using bustards in China). Though Random Forest performed the best to predict abundance itself (see Appendix 1), testing the feasibility for other data was essential for good certainty. So for an assessment on the robustness of the model we pooled data from 2013 and 2014, and then used 80% abundance data as training data and the remaining 20% as testing data. When we constructed initial abundance models with all eleven environmental predictors, model performance is not so good (R^2 was small). Likely that has to do with the regression settings in Random Forest algorithm. For a better outcome we assessed a "stepwise" setting in SPM for whole abundance data (100%) to re-run models, and found better results. In that way, we identified a multivariate set of four environmental predictors (distance to expressway, distance to national road, distance to pool, MNNDVI), which have the best performance (biggest R^2). Using these four predictors, we re-constructed the abundance model based on the training data (80%) and validated it with testing data (20%). We found that the regression model performance was acceptable but fair (R^2 = 0.551) between observation and simulation abundance. Thus, we constructed the final abundance model based on the above four selected variables and with the entire observation data. In order to obtain an abundance index more close to observations we adjusted the simulation abundance according to the linear regression between observation and simulation abundance (Fig. 2a).

Put Fig. 2 Here

Further, Random Forest was also applied to rank the relative importance of environmental variables. In SPMv7, partial dependence plots are not directly implemented in Random Forest yet, but can easily be obtained from R or are mimicked in TreeNet model as a Random Forest run. Thus, we used TreeNet with bagging settings to create partial dependence plots for each variable of the occurrence and abundance models.

About 10,000 pseudo-absence points were taken by random sampling across study areas using the freely available Geospatial Modeling Environment (GME) software (http://www.spatialecology.com/gme/) for distribution models. In SPMv7 we set balanced class weights, grew each model to 1,000 classification trees for occurrence model and 1,000 regression trees for abundance model, and used all other software default settings. We extracted the habitat information from the environmental layers for presence and pseudo-absence points for Great Bustards in GME, and then created a model file in SPM7 called a 'grove' containing the algorithm quantifying patterns of occurrence for scoring all pixels in the study area. We also extracted the habitat information from the same environmental layers for abundance points, and then generated a 'grove' file for abundance to score abundance estimates for each pixel in the study area.

For spatial occurrence and abundance distribution visualization, we applied the SPM7 grove files to a regular lattice of points (pixels; also attributed to the environmental variables) spaced at 30 m intervals across the study area. Model outputs generated relative indices of occurrence (RIO; an index of pixels from 0 to 1 representing a relative index belonging to the 'occurrence' class) and a relative abundance index (simulation abundance) for each point in the regular lattice based on its underlying environmental variables. We also adjusted the predicted abundance based on a linear regression as constructed in the previous model development steps (Fig. 2a). For a better continuous spatial visualization, the RIO and predicted abundance values were smoothed between neighboring points across the extent of the study area using the Inverse Distance Weighting (IDW) tool in ArcGIS 10.1. This yielded spatially continuous predictive distribution and abundance raster maps of Great Bustard.

Model validation

The Random Forest performance was first assessed internally using a set of 'out-of-bag' (OOB) training points (OOB; a specific concept used with Random Forest models to describe a subset of points not used initially for model fitting; Breiman 1996, Breiman 2001). Using this out-of-bag dataset, the receiver-operating characteristic (ROC) and RMSE were used to calculate predictive performance of occurrence and abundance models, respectively (Zweig and Campbell 1993; Fielding and Bell 1997; Huettmann and Gottschalk 2010).

Priority protection analysis

In order to have a more suitable and scientific protection plan for the endangered Great Bustard, in this study we put forward for the first an index called the priority protection index (PI), which combines the predicted results of SDM and SAM. This index is calculated by the following equation for each site:

$$\boldsymbol{PI} = \frac{\boldsymbol{RIO} \times \boldsymbol{RA}}{\mathbf{max}\,(\boldsymbol{RIO} \times \boldsymbol{RA})} \qquad (1)$$

where ***PI*** = Priority protection index (an index of pixels from 0 to 1 representing the priority of conservation), ***RIO*** = relative index of occurrence, and ***RA*** = relative abundance (simulation abundance). In our study, we computed the PI for the whole study area based on RIO and the adjusted RA value of each grid cell by spatial occurrence and abundance maps. Then we used the IDW tool in ArcGIS 10.1 to generate spatially continuous priority protection index (PI) raster maps. In this equation we did not consider the weighting of biotic and socioeconomic variables.

So the justification and use of the PI should be explained a little more: When combining SDM with SAM one will not find a straight forward relationship between occurrence and abundance (see Yen et al. 2004 for an example). What the PI will do, but what has not been achieved before much, is to essentially model that relationship and provide a combined view of occurrence index and abundance index explicit in space and time. Achieving this can thereby help to prioritize pixels better with let's say high occurrence index and low abundances on pixels etc.

RESULTS

Model performance

Our distribution model obtained a decent performance (ROC: 0.77) according to Fielding and Bell (1997), and the abundance model had RMSE 26.54 (RMSE is unit-less). Such model predictions allow us to infer from such models and how they are built.

Variable importance

Table 2 presents the variable importance ranking of occurrence and abundance models obtained from the Random Forest method. We found that the area of farmland, distance to residential area (buildings), to ditch and to expressway were the top four most important variables influencing bustard occurrence. Those come as a multivariate package. The NDVI which represents vegetation condition was less important than the other nine predictors. As for the abundance model, the most important factors were distance to national road and to expressway, followed by water factors (distance to pool) and food-related factors (MNNDVI)

Put Table 2 Here

Partial dependence plots

Partial dependence plots could interpret the functional relationships and effects of each variable by representing a variable's marginal effects on the response (Elith et al., 2008; Johnstone et al., 2010). It helps to find the signal in the data; Fig. 3a indicated that the occurrence preference of bustards for farmland area was between 0.6 and 7.5 km^2. Distance to residential area ranging from 250 to 2,500 m (Fig. 3b), distance to ditch ranging from 100 to 4,500 m (Fig. 3c), and distance to expressway from 6,000 to 19,000 m (Fig. 3d) were bustard preferences. While for abundances, more individuals would occur beyond 2,300 m, but less than 9,500 m away from national roads (Fig. 3e), and be found in the range between 7,000 and 11,000 m away from expressway (Fig. 3f). Moreover, this species kept themselves away from pools (larger than 1,500 m, Fig. 3g) and with more vegetation (mean NDVI during the investigation larger than 0.13, Fig. 3h). The information for other variables, more marginal, can be found in Appendix 2.

Put Fig. 3 Here

Occurrence, abundance distribution patterns and priority protection

Fig. 4 shows the maps of RIO (relative index of occurrence), adjusted RA (relative abundance) and PI (priority protection index). From the RIO map (Fig. 4a), we found that the distribution area of high RIO of bustards is high. The regions of high occurrence possibility of bustards were concentrated in the south-central study area; and the whole habitats represented a fragmented distribution. The abundance distribution had a different pattern, showing high populations occurring in the central and northwestern study area (Fig. 4b). Based on the occurrence and abundance distribution results, we used equation (1) and obtained the result of Fig. 4c. It displays that a high PI is located in the center, north and northeast of the study area and it shows a sporadic fragmented distribution which would be the priority protection site if a conservation decision is to be made.

Put Fig. 4 Here

DISCUSSION

The occurrence and abundance models of Great Bustard developed here were designed to identify relevant locations for where to prioritize conservation, and to assess the effects of each variable that influenced this species occurrence and abundance (Fig. 3). Area of farmland, distance to residential area, distance to ditch and to expressway were among the top four most important predictors for bustard occurrence in a multivariate perspective; while for the abundance model they consisted of another multivariate package of distance to national road, distance to expressway, distance to pool and mean NDVI (Table 2). We found that high RIO habitats had a fragmented distribution throughout the entire study area (Fig. 4a). The abundance model showed that high population usually occurred in the central and northwestern part of our study area (Fig. 4b). The center, north and northeast of the study area with a high priority protection index (PI) and with a severely fragmented distribution should be the priority site for protection (Fig. 4c). This not only confirms our own records but with the help of the PI can now be quantified and modeled further for an effective conservation!

In our study area, human disturbance was very strong, such as density of roads and residential areas (Fig. 1). During our study we also found other threats to this endangered species: farmers grazed their sheep; famers sprinkled poison baits in the wheat field to avoid sheep entering; some bird photographers pursued bustards by walking or following on motor vehicles to take photos, which they wanted to show off to others; hunters with dogs chasing hare and ring-necked pheasant during day and night; some local people hunted bustards; increasing power lines setting in agriculture land, bustards sometimes collided with wires and were injured or even died when starting to fly in foggy days or when in a hurry (Janss & Ferrer 2000); and the interference of firecracker sounds during Chinese Spring Festival as well as oil rigs and wind farms. Though carrying a high disturbance and for a stress synthesis (e.g. "death by thousand cuts"), still, a large number of wintering bustards (about 300, c. 13.6 ~20.0 % of China's total wintering population; Goroshko 2010; Meng 2010) wintered in this area. In times of climate change, it can be assumed the population widens (Mi et al., 2016). Thus, this is an area of essential importance for bustards in China either way. A feasible conservation plan should therefore be made, based on our model prediction result, combined with local public customs and financial support and a wider buy-in. In our opinion, improved education on animal protection to local people as we usually did over the years would be useful. The same applies to increasing budgets, enforcement and frequency of patrol by the local management and conservation NGOs in the region with high PI value and the local community, with corresponding government financial support. Patrol route designation in the field should avoid getting too close to bustards though, so as not to disturb and stress the regular wintering activities of bustards. For the benefit of this species and its habitats we suggest to not change crop farmland into nursery farmland; and we encourage farmers to harvest their crops with a machine, which is a more beneficial harvesting method for bustards based on our previous research results (Mi et al., 2014). We also highly recommend, if possible, to bury power lines into the ground and to collect hunting guns from local public.

In this study, occurrence and abundance did not display identical spatial distribution patterns which was reported in some previous studies (Conlisk et al., 2007; Karlson et al., 2011; Yin & He 2014; Johnston et al., 2015). There is actually no reason to assume a presence site just shows one animal individual, or a linear relationship between RIO and abundance. Technically-speaking, 'presence' can mean 1-infite animals and details depend on the actual pixel set-up and how it fits into the obtained model. So while the relationship is not automatically clear this could be due to several reasons and depending on specific habitat details: Firstly, environmental variables that contributed to occurrence and abundance were different, as Table 2 indicated. Secondly, predictor preference in bustard occurrence and abundance models were different. For instance, bustards occur at a distance to expressway from 6,000 to 19, 000m (Fig. 3d), while most populations occur between 7,000 and 11,000m from expressway for abundance (Fig. 3f) (see more details in Fig. 3 and Appendix 2). Thirdly, they differed in their spatial distribution for occurrence and abundance (Fig. 4a, b). Based on the analysis of overlaying observation sites with RIO and observation abundance (Fig. 5a, b), estimated relative index of occurrence (RIO) do not consistently relate with the relative index of abundance (Fig. 5a). All locations of observed abundance had high RIO (Fig 5a), and the relationships between occurrence and abundance estimates were nonlinear (Fig. 5b). These differences may represent a mixture of effects reflecting differences between the underlying biological processes that give rise to specific abundance and occurrence at a pixel, as well as limitations imposed by the data and methodology to estimate these patterns (Johnston et al., 2015; see Buckland et al., 2016 for Distance Sampling and detectability problems). In addition, how to understand the inconsistency between these two indices of plant prediction is a problem waiting to be resolved further. For instance, between crop occurrence index (equal to habitat suitability index) and crop abundance (e.g. production).

Put Fig. 5 Here

When treating all presences as equal in species distribution models (SDMs; occurrence model, habitat niche model) -regardless of the abundance of individuals that the habitat supports - this could provide us with information on the suitability of habitat loss (Howard et al., 2014). Applying models based on abundance data even at a relatively coarse scale can help to predict spatial patterns of occurrence modelled with even greater refinement (Howard et al., 2014). Conservation decision-making should use as much knowledge and information as possible to optimize the benefits of conservation action (Sutherland et al., 2004; Segan et al., 2011). The use of species distribution models (SDMs) of occurrence has been an important tool in optimizing the selection of protected areas (Franklin 2013; Guisan et al., 2013, Mi et al., 2016; Han et al., 2017) based on the ecological niche space (Drew et al., 2011), but relative abundance is often perceived a more relevant metric because it can quantify animals on a pixel, and thus, populations (Johnston et al., 2015). Modeling abundance requires methods that can handle large numbers of zero counts as well as the rare, but important, high counts (Welsh et al., 1996) without a solid research design, according to frequentist statistics. However, Yen et al., (2004), Magness et al., (2008) and Fox et al., (2017) showed already how machine learning can change this perspective and provide very powerful solutions.

High counts and their locations are particularly important because the pixels with the highest densities of animals are potentially of greatest interest for conservation planning (Johnston et al., 2015). In our study, we found that the regressions in Random Forest performed imperfectly for low and high counts (Fig. 2b) although it showed a highly linear relationship between observed and simulation abundance (R^2=0.844; Fig 2a). Therefore, we argue that the regression method in Random Forest algorithm should optimize low and high count predictions. We recommend to classify abundances in bins (e.g. high, medium, low with associated abundance estimates) because Random Forest is exceptionally strong for classification problems. This remains an open field of research, for now. However, we find our progress remains substantial.

Abundance data could also provide valuable baselines against which to assess future changes (Cumming 2007) (e.g. climate change, land use change). Such changes in abundance will be much more rapidly apparent, and hence more rapidly detected than changes in presence-absence patterns across ranges (Gregory et al., 2005). However, only a few spatial distribution modelers derived models with the collection of abundance data (e.g. Yen et al. 2004, Fox et al. 2017). This may be because collection of abundance data is more cost or resource demanding than collecting presence - absence data especially for highly mobile animals. Such data are sophisticated in structure and research design, and still they are rarely shared (see in GBIF.org). We therefore recommend that abundance data could be collected (easily to be turned into presence-absence data, too), even at only relatively coarse numerical scales because the benefits are considerable (as stated by Howard et al., 2014). One thing that should be mentioned is that plenty of abundance data and (non-linear regression) models did not perform well and abundance were extremely difficult to predict (Oppel et al., 2012). Finding the underlying causes that influence abundance model accuracy and constructing more accurate models would be extreme important and useful in future applications towards individual-based policy applications.

For a spatial priority protection of mobile species, one should note that high numbers of individuals are not always present in the same habitats and pixels, instead low numbers may occur in one place many times. And this may have implications for spatial priority protection for mobile species. Previous studies have used analytical approaches to deal with some of these challenges (e.g. Nichols et al., 2009; Kery & Andrew Royle 2010; Oppel et al., 2012; Jiguet et al., 2013). However, no general modeling framework has been proposed for dealing with all of these analytical challenges simultaneously. This is exactly where our PI offers progress. We also thought the situation of mobile species selecting habitats could be divided into five scenarios: higher numbers and multi frequency, higher numbers and lower frequency, low numbers and multi frequency, low numbers and low frequency, none. When a conservation plan is made for a species, one should consider not only occurrence index and frequency, but also abundance. Here we proposed the priority protection index (PI; equation (1) and Fig. 4) based on the distribution of occurrence and abundance pattern as more helpful for a fast priority protection plan than indices and it's only based on distribution of occurrence or abundance.

To date, quantitative estimates of population size during global and local changes have actually proven to be difficult to forecast. This is a major hindrance for effective management, as population size and trend are considered among the best correlates of extinction risk (O'Grady et al., 2004). Such measures are commonly used in determining the conservation status of a species (e.g. IUCN (2001)). We argue that habitat loss remains the one and only powerful metric that can be obtained quickly on a landscape-scale in the absence of proper trends and abundances (e.g. Drew et al. 2011). The relationship between predicted environmental suitability and abundance - as presented here - may provide us with a possible method for predicting population size and its changes associated with distributional changes, particularly appropriate for non-mobile species (e.g. plants, fungi). However, this method is not particularly suitable for mobile species, especially for highly mobile species such as many birds, bats, and flying insects. They may move over a large landscape within just a single day, and abundance and the environment can vary seasonally and spatially. When computing population size or population density using abundance, the primary task will be how to determine the unit area of investigation and for conservation management.

This study is the first that has combined model-predicted occurrence (representing species distribution model) and abundance indices (representing species abundance model) to produce a priority protection index (PI), which may contribute to spatial conservation and management decisions worldwide. We strongly encourage other researchers to test, apply and update the priority protection index (PI) to explore the generality of these findings further.

Acknowledgements

We thank Liu Min for his hard work in the field. Thanks also to a shared field survey among the authors. Further thanks to Salford Systems Ltd. for providing the free trial version of SPM

software.

REFERENCES

Adams MP, Saunders MI, Maxwell PS, Tuazon D, Roelfsema CM, Callaghan DP, Leon J, Grinham AR, and O'Brien KR. 2016. Prioritizing localized management actions for seagrass conservation and restoration using a species distribution model. Aquatic Conservation Marine and Freshwater Ecosystems 26:639-659.

Araújo MB, New M. 2007. Ensemble forecasting of species distributions. Trends in Ecology and Evolution 22:42-47.

Ashcroft MB, King DH, Raymond B, Turnbull JD, Wasley J, and Robinson SA. 2017. Moving beyond presence and absence when examining changes in species distributions. Global Change Biology 23:2929-2940.

Barker NKS, Cumming SG, Darveau M. 2014. Models to predict the distribution and abundance of breeding ducks in Canada. Avian Conservation and Ecology 9:7.

Breiman L. 1996. Bagging predictors. Machine learning 24:123-140.

Breiman L. 2001. Random forests. Machine learning 45:5-32.

Buckland ST, Rexstad EA, Marques TA, and Oedekoven CS. 2016. Distance Sampling: Methods and Applications. Methods in Statistical Ecology 63:152-153.

Conlisk E, Conlisk J, Harte J. 2007. The impossibility of estimating a negative binomial clustering parameter from presence - absence data: a comment on He and Gaston. The American Naturalist 170:651-654.

Cumming GS. 2007. Global biodiversity scenarios and landscape ecology. Landscape Ecology 22:671-685.

Cutler DR, Edwards JrTC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler, JJ. 2007. Random forests for classification in ecology. Ecology 88:2783-2792.

Drew CA, Wiersma Y, Huettmann F. 2011. Predictive species and habitat modeling in landscape ecology: concepts and applications. Springer, London.

Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton JMM, Peterson AT, Phillips SJ, Richardson K, Scachetti-Pereira R, Schapire RE, Soberón J, Williams S, Wisz MS, and Zimmermann NE. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129-151.

Elith J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77: 802-813.

Elith J, Leathwick JR. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40:677-697.

Fielding AH, Bell JF. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental conservation 24:38-49.

Fox CH, Huettmann F, Harvey GKA, Morgan KH, Robinson J, Williams R, Paquet PC. 2017. Predictions from machine learning ensembles: marine bird distribution and density on Canada's Pacific coast. Marine Ecology Progress Series 566:199-216.

Franklin J. 2013. Species distribution models in conservation biogeography: developments and challenges. Diversity and Distributions 19:1217–1223.

Goroshko OA. 2010. Present status of population of Great Bustard (*Otis tarda dybowskii*) in Dauria and other breeding grounds in Russia and Mongolia: distribution, number and dynamics of population, threats, conservation. First International Symposium on Conservation of Great Bustard, Beijing (in Chinese).

Gray TN, Borey R, Hout SK, Chamnan H, Collar NJ, and Dolman PM. 2009. Generality of models that predict the distribution of species: conservation activity and reduction of model transferability for a threatened bustard. Conservation biology 23:433–439.

Gregory RD, Strien AV, Vorisek P, Meyling AWG, Noble DG, Foppen RP, Gibbons DW. 2005. Developing indicators for European birds. Philosophical Transactions of the Royal Society of London B: Biological Sciences 360:269-288.

Guisan A, Thuiller W. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8:993-1009.

Guisan A, Tingley R, Baumgartner JB, Naujokaitis-Lewis I, Sutcliffe PR, Tulloch AIT, Regan TJ, Brotons L, Mcdonald-Madden E, Mantyka-Pringle C. 2013. Predicting species distributions for conservation decisions. Ecology Letters 16:1424-1435.

Guisan A, Zimmermann NE. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135:147-186.

Han X, Guo Y, Mi C, Huettmann F, and Wen L. 2017. Machine Learning Model Analysis of Breeding Habitats for the Black-necked Crane in Central Asian Uplands under Anthropogenic Pressures. Scientific Reports 7:6114.

Howard C, Stephens PA, Pearce-Higgins JW, Gregory RD, Willis SG. 2014. Improving species distribution models: the value of data on abundance. Methods in Ecology and Evolution 5:506-513.

Howard C, Stephens PA, Pearce - Higgins JW, Gregory RD, Willis SG. 2015. The drivers of avian abundance: patterns in the relative importance of climate and land use. Global Ecology and Biogeography 24:1249-1260

Huettmann F, Gottschalk T. 2010. Simplicity, Model Fit, Complexity and Uncertainty in Spatial

Prediction Models Applied Over Time: We Are Quite Sure, Aren't We? Springer New York.

IUCN. 2001. UCN red list categories and criteria: version 31 Prepared by the IUCN Species Survival Commission.

Janss GFE, Ferrer M. 2000. Common Crane and Great Bustard Collision with Power Lines: Collision Rate and Risk Exposure. Wildlife Society Bulletin 28:675-680.

Jiguet F, Thomas C, Cook A, Newson S, Ockendon N, Rehsch M, Roos S, Thaxter C, Brown A, Crick H. 2013. Observed and predicted effects of climate change on species abundance in protected areas. Nature Climate Change 3:1055-1061.

Johnston A, Fink D, Reynolds MD, Hochachka WM, Sullivan BL, Bruns NE, Hallstein E, Merrifield MS, Matsumoto S, Kelling S. 2015. Abundance models improve spatial and temporal prioritization of conservation resources. Ecological Applications 25:1749-1756.

Johnstone JF, Hollingsworth TN, Chapin FS, Mack MC. 2010. Changes in fire regime break the legacy lock on successional trajectories in Alaskan boreal forest. Global Change Biology 16:1281-1295.

Karlson RH, Connolly,SR, Hughes TP. 2011. Spatial variance in abundance and occupancy of corals across broad geographic scales. Ecology 92:1282-1291.

Kery M, Andrew RJ. 2010. Hierarchical modelling and estimation of abundance and population trends in metapopulation designs. Journal of Animal Ecology 79:453-461.

Magness DR, Huettmann F, Morton JM. 2008. Using Random Forests to provide predicted species distribution maps as a metric for ecological inventory and monitoring programs. pp 209-229 In Smolinski TG, Milanova MG, Hassanien AE Applications of Computational Intelligence in Biology: Current Trends and Open Problems Studies in Computational Intelligence, Vol 122, Springer-Verlag Berlin, Heidelberg.

Martín B, Alonso JC, Martín CA, Palacín C, Magaña M, Alonso J. 2012. Influence of spatial heterogeneity and temporal variability in habitat selection: A case study on a great bustard metapopulation. Ecological Modelling 228:39-48.

Meng D. 2010. Study on the Rescue to Great Bustard in Cangzhou, Hebei. First International Symposium on Conservation of Great Bustard, Beijing (in Chinese).

Mi C, Huettmann F, Guo Y. 2014. Obtaining the best possible predictions of habitat selection for wintering Great Bustards in Cangzhou, Hebei Province with rapid machine learning analysis. Chinese Science Bulletin 59:4323-4331.

Mi C, Huettmann F, Guo Y. 2016. Climate envelope predictions indicate an enlarged suitable wintering distribution for Great Bustards (*Otis tarda dybowskii*) in China for the 21st century. PeerJ 4:e1630.

Mi C, Huettman F, Guo Y, Han X, Wen L. 2017. Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. PeerJ 5:e2849.

Newton I. (2008) Migration Ecology of Birds. London, UK: Academic Press.

Nichols JD, Thomas L, Conn PB. 2009. Inferences about landbird abundance from count data: recent advances and future directions pp 201-235 In Thomson DL, Cooch EG, Conroy MJ Modeling demographic processes in marked populations Springer, US.

O'Grady JJ, Reed DH, Brook BW, Frankham R. 2004. What are the best correlates of predicted extinction risk? Biological Conservation 118:513-520.

Oppel S, Meirinho A, Ramírez I, Gardner B, O'Connell AF, Miller PI, Louzao M. 2012. Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. Biological Conservation 156:94-104.

Osborne PE, Alonso J, Bryant R. 2001. Modelling landscape - scale habitat use using GIS and remote sensing: a case study with great bustards. Journal of Applied Ecology 38:458-471.

Peterson AT, Ortega-Huerta MA, Bartley J, Sánchez-Cordero V, Soberón J, Buddemeier RH, Stockwell DR 2002. Future projections for Mexican faunas under global climate change scenarios. Nature 416:626-629.

Primack RB. 2012. A primer of conservation biology. 5th Edition Sunderland, Massachusetts. Sinauer Associates.

Segan DB, Bottrill MC, Baxter PWJ, Possingham HP. 2011. Using conservation evidence to guide management. Conservation Biology 25:200-202.

Suárez-Seoane S, Morena ELGDL, Prieto MBM, Osborne PE, and Juana ED. 2008. Maximum entropy niche-based modelling of seasonal changes in little bustard (Tetrax tetrax) distribution. Ecological Modelling 219:17-29.

Sutherland WJ, Pullin AS, Dolman PM, Knight TM. 2004. The need for evidence-based conservation. Trends in Ecology and Evolution 19:305-308.

VanDerWal J, Shoo LP, Johnson CN, Williams SE. 2009. Abundance and the environmental niche: environmental suitability estimated from niche models predicts the upper limit of local abundance. The American Naturalist 174:282-291.

Welsh AH, Cunningham RB, Donnelly CF, Lindenmayer DB, 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecological Modelling 88:297-308.

Yen PPW, Huettmann F, Cooke F. 2004. A large-scale model for the at-sea distribution and abundance of Marbled Murrelets (*Brachyramphus marmoratus*) during the breeding season in coastal British Columbia, Canada. Ecological Modelling 171:395-413.

Yin D, He F. 2014. A simple method for estimating species abundance from occurrence maps.

Methods in Ecology and Evolution 5:336-343.

Zweig MH, Campbell G, 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry 39:561-577.

Table 1(on next page)

Table

1 **Tables**

2 Table 1 Comparison of features around 94 sites occupied by great bustards and 10 000 random points. Values are means ± standard deviations.

Layer	Variable	Description	Bustard sites	Random points
1	Distance to pool	Distance to pool in meter	1179.0 ± 734.5	1378.0 ± 910.3
2	Distance to river	Distance to river in meter	2302.0 ± 1751.2	2630.0 ± 2483.0
3	Distance to residential	Distance to residential in meter	935.0 ± 586.8	980.2 ± 723.8
4	Distance to national road	Distance to national road in meter	5280.0 ± 4234.2	5855.0 ± 4036.9
5	Distance to provincial road	Distance to provincial road in meter	8730.0 ± 5928.7	9217.0 ± 6112.4
6	Distance to expressway	Distance to expressway in meter	10010 ± 5750.0	9585.0 ± 6666.7
7	Distance to farmland road	Distance to farmland road in meter	477.4 ± 385.3	524.9 ± 455.8
8	Distance to ditch	Distance to ditch in meter	1522.0 ± 1722.7	2120.0 ± 2078.1
9	Area of farmland	Area of farmland in kilometers	3.3 ± 3.2	5.3 ± 6.2
10	MNNDVI	The average value of the normalized difference vegetation index from November, 2013 to March, 2014	0.14 ± 0.04	0.13 ± 0.05
11	MAXNDVI	The maximum value of the normalized difference vegetation index from November, 2013 to March, 2014	0.23 ± 0.06	0.21 ± 0.07

3

Table 2 Variables importance ranking of occurrence and abundance models

Ranking	Occurrence model	Abundance model
1	Area of farmland	Distance to national road
2	Distance to residential	Distance to expressway
3	Distance to ditch	Distance to pool
4	Distance to expressway	MNNDVI
5	Distance to pool	--
6	Distance to river	--
7	Distance to provincial road	--
8	Distance to national road	--
9	Distance to farmland road	--
10	MAXNDVI	--
11	MNNDVI	--

Figure 1

Figure 1

Study area and bird abundance and occurrence data for Great Bustard in Cangzhou, China. Photograph of Great Bustard by Jianguo Fu.

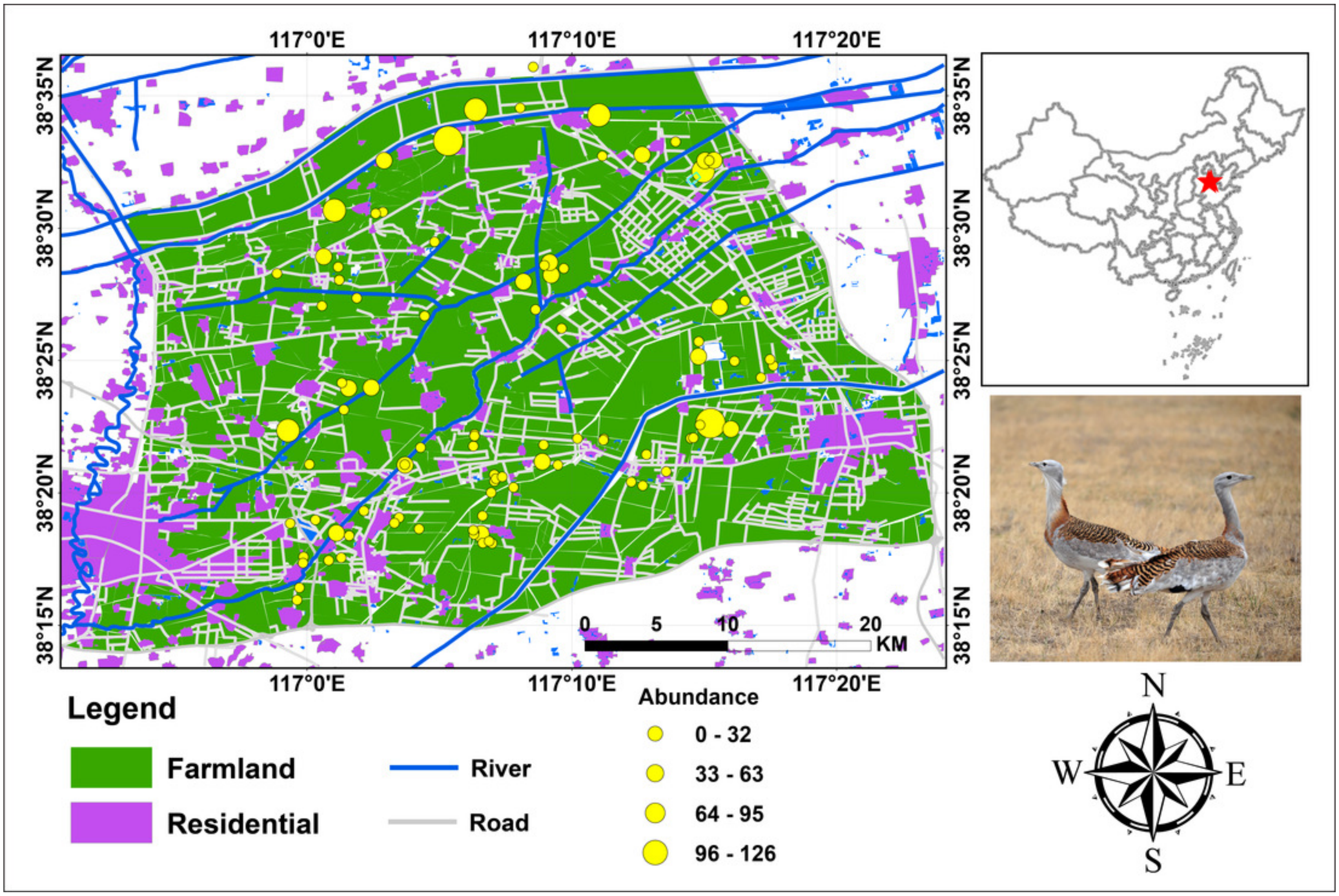


Figure 2(on next page)

Figure 2

Figure 2 The relationship between observation and prediction abundance using Random Forest for Great Bustards. (a) Scatter plot of observation abundance with prediction and adjustment prediction abundance, and (b) lines and points plot of observation, prediction and adjustment prediction abundance.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3240v1 | CC BY 4.0 Open Access | rec: 11 Sep 2017, publ: 11 Sep 2017

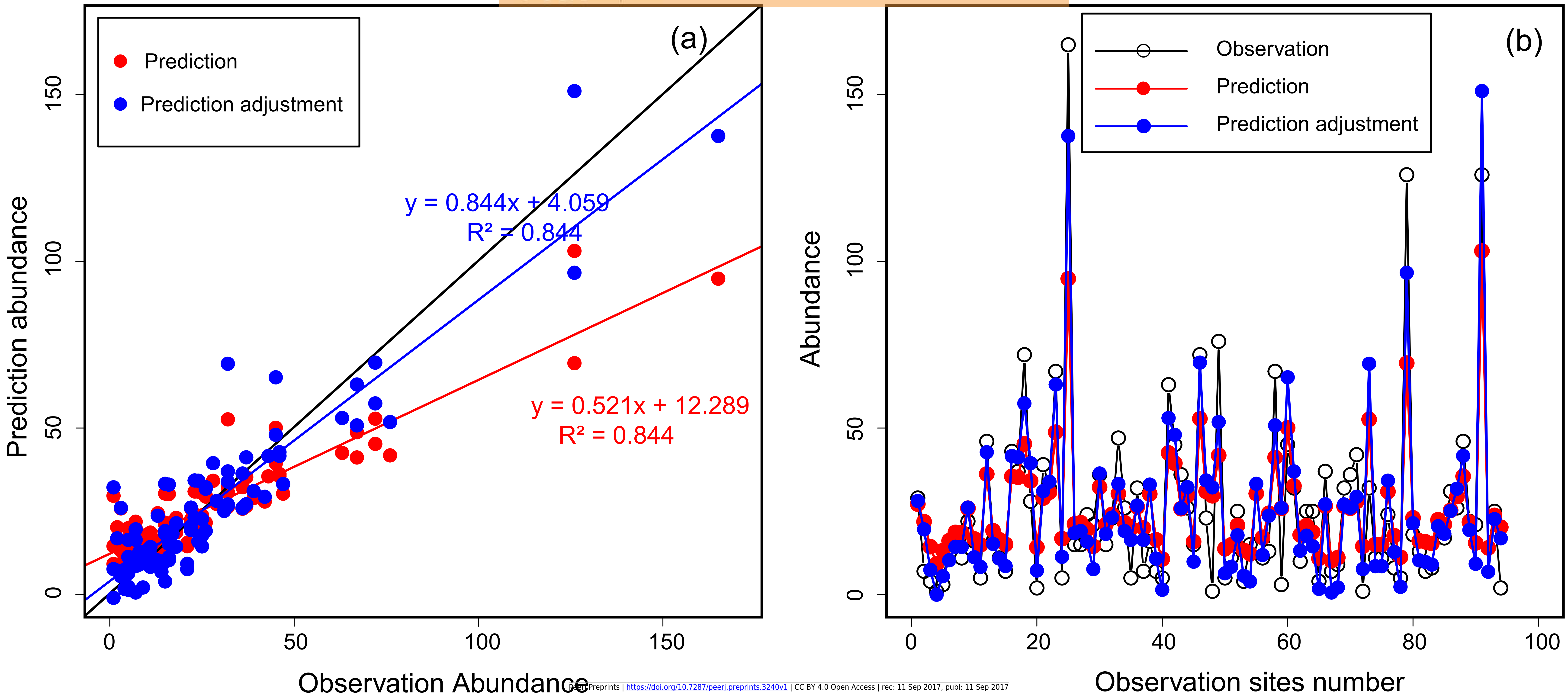
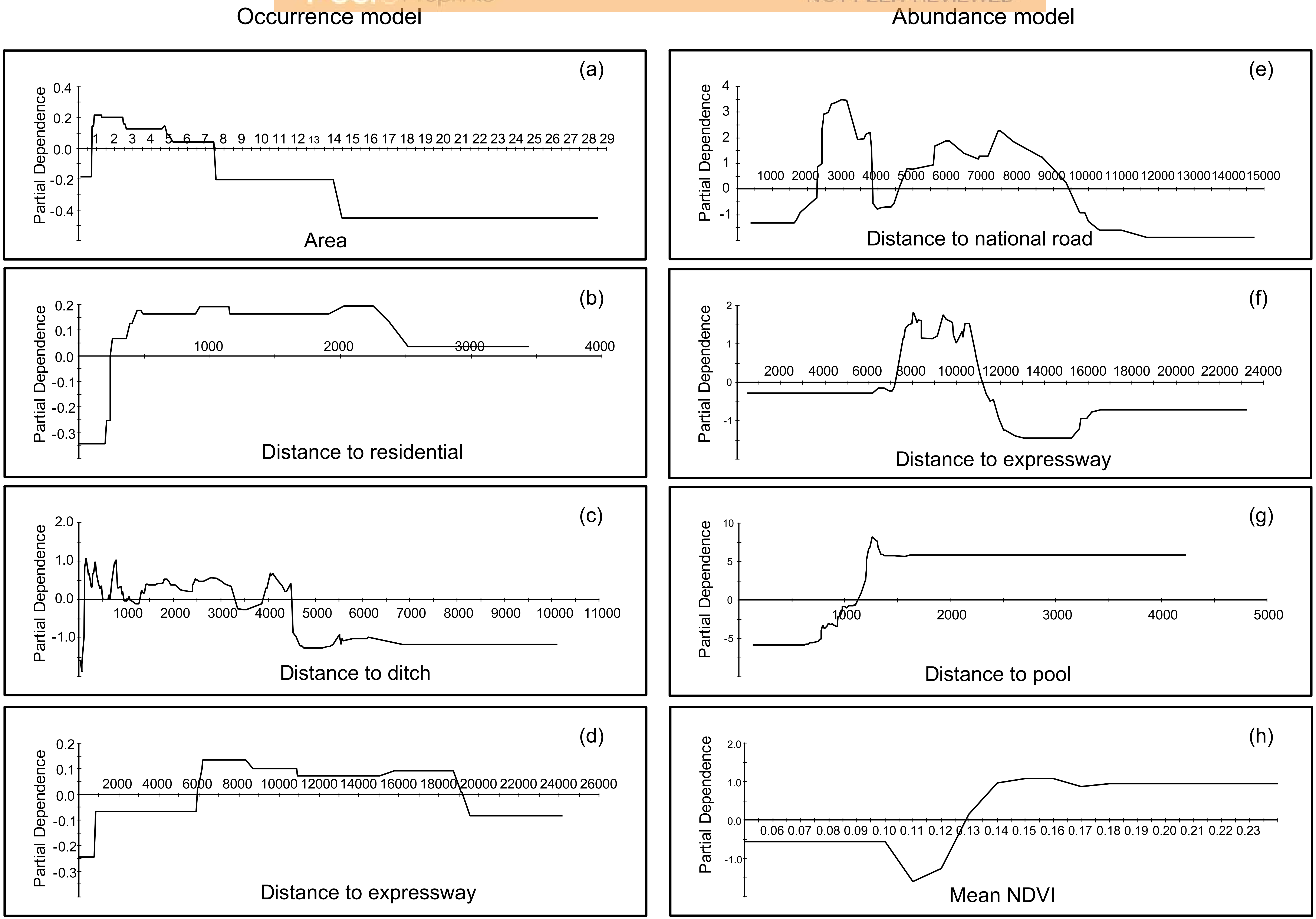

(a)
Prediction
Prediction adjustment
y = 0.844x + 4.059
R² = 0.844
y = 0.521x + 12.289
R² = 0.844
Prediction abundance
Observation Abundance
0
50
100
150
(b)
Observation
Prediction
Prediction adjustment
Abundance
Observation sites number
0
20
40
60
80
100

Figure 3 (on next page)

Figure 3

Partial dependence plots for the top four most influential variables in the occurrence and abundance distribution models for Great Bustards, respectively**:** (a) area of farmland in occurrence distribution model; (b) distance to residential in occurrence distribution model; (c) distance to ditch in occurrence distribution model; (d) distance to expressway in occurrence distribution model; (e) distance to national road in abundance distribution model; (f) distance to expressway in abundance distribution model; (g) distance to pool in abundance distribution model; and (h) mean NDVI in abundance distribution model.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3240v1 | CC BY 4.0 Open Access | rec: 11 Sep 2017, publ: 11 Sep 2017



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3240v1 | CC BY 4.0 Open Access | rec: 11 Sep 2017, publ: 11 Sep 2017

Figure 4

Figure 4

Spatial distribution map of relative index of occurrence (RIO), relative abundance (RA) and priority protection index (PI). (a) Map of relative index of occurrence (RIO); (b) map of adjusted relative abundance (RA); and (c) map of priority protection index (PI).

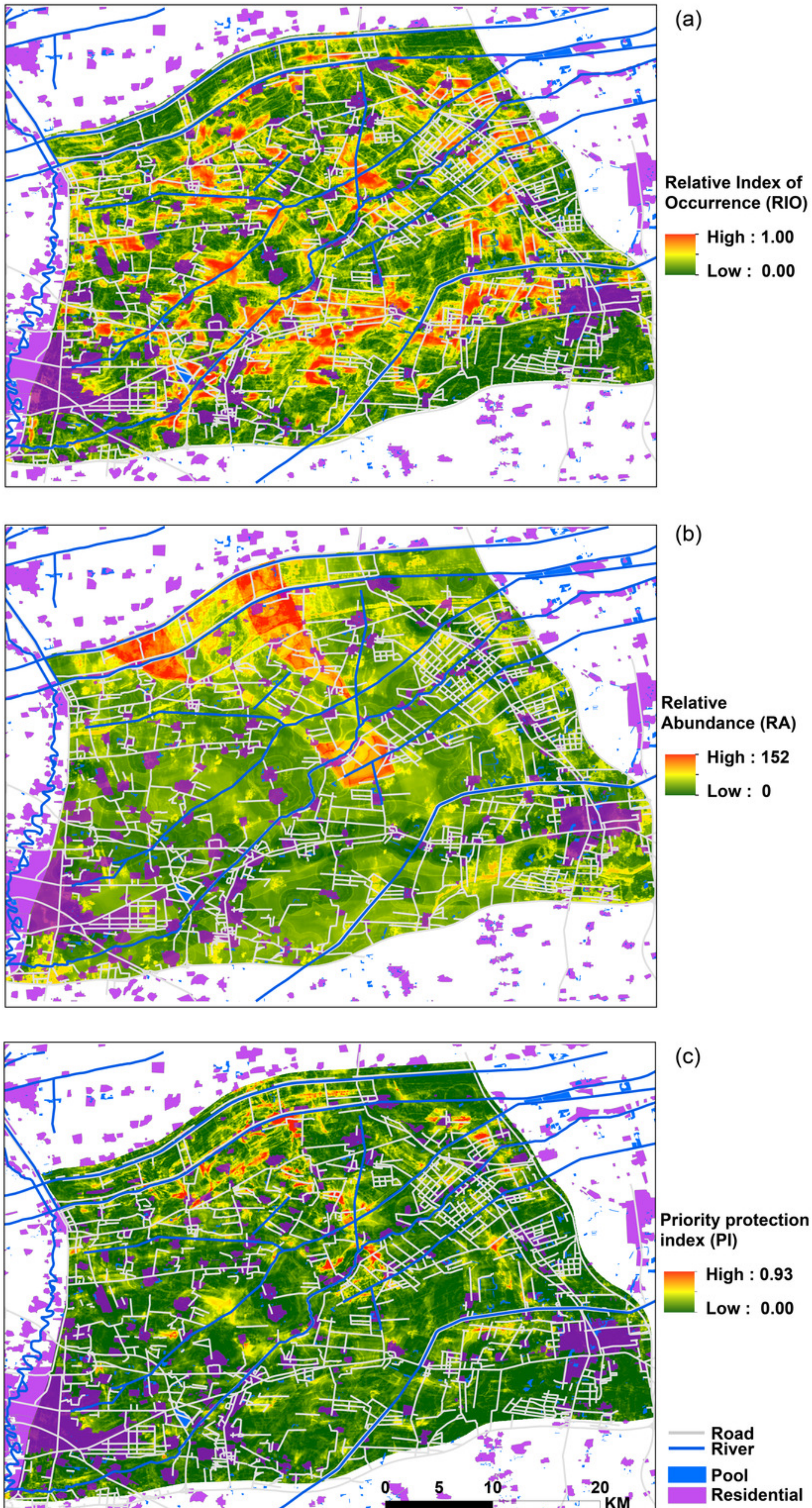
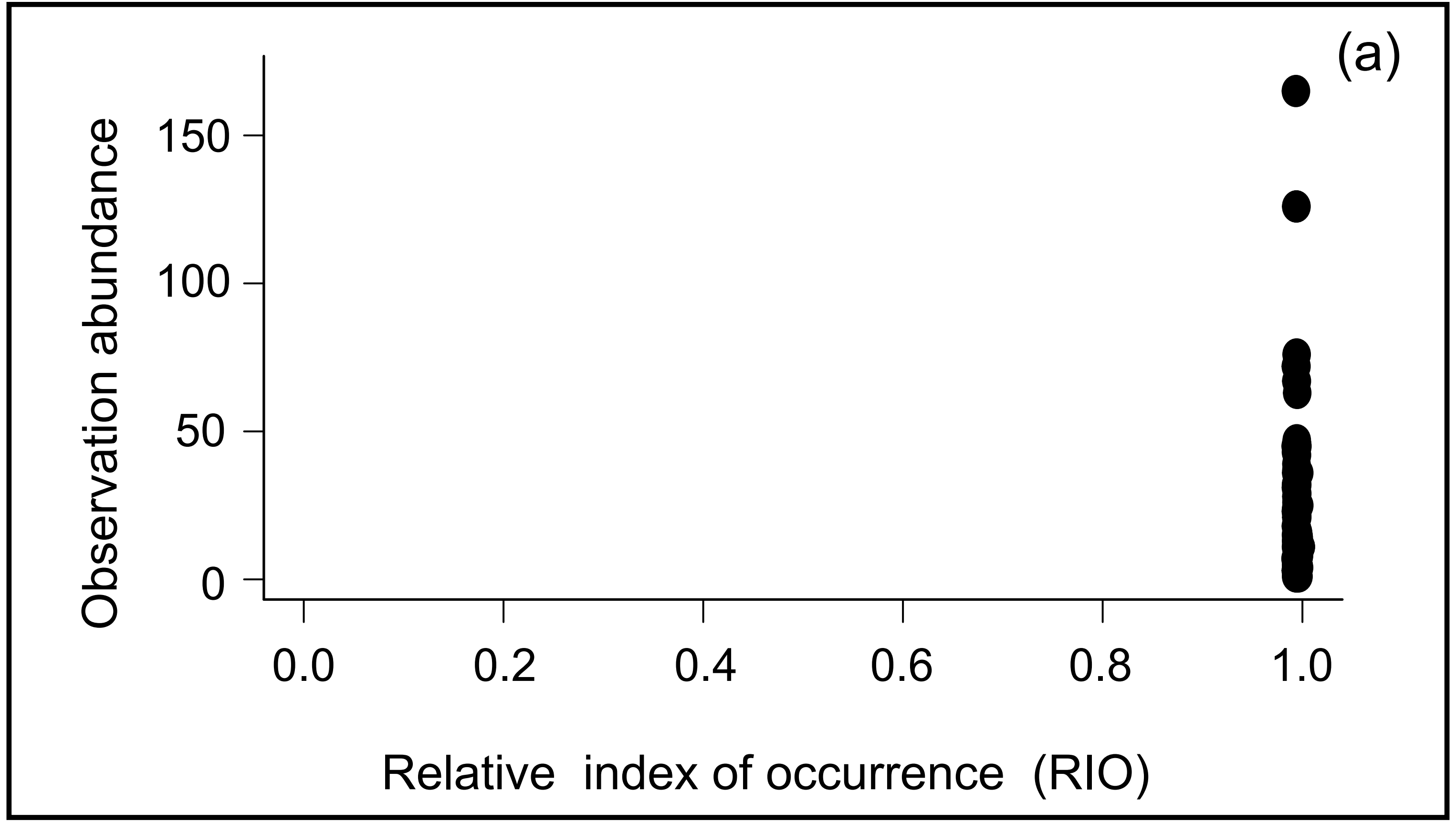
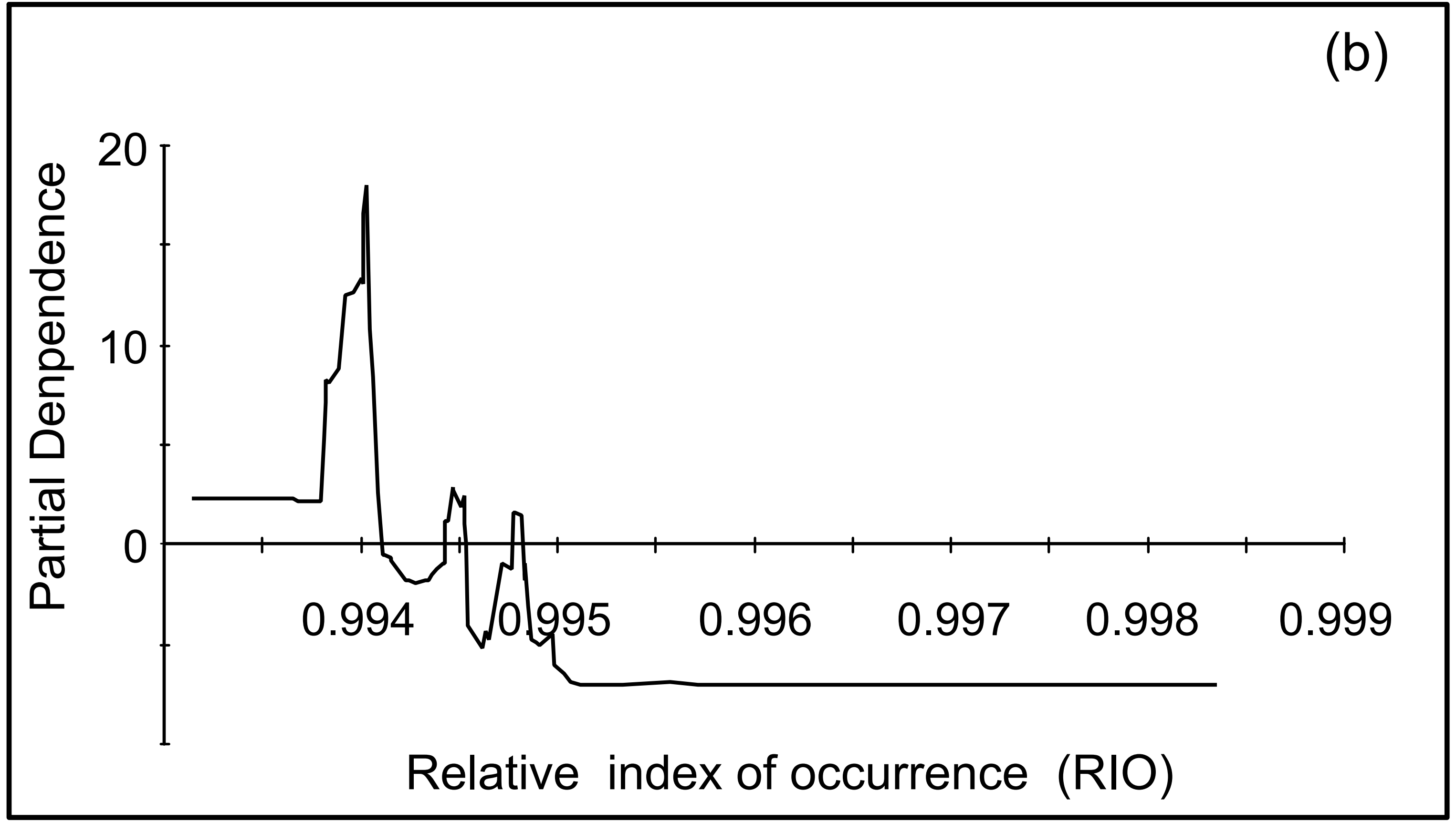

(a)
Relative Index of Occurrence (RIO)
High : 1.00
Low : 0.00
(b)
Relative Abundance (RA)
High : 152
Low : 0
(c)
Priority protection index (PI)
High : 0.93
Low : 0.00
Road
River
Pool
Residential
0 5 10 20
KM

Figure 5(on next page)

Figure 5

Plots of the relationship between relative index of occurrence (RIO) and observation abundance. (a) Scatter plot between relative index of occurrence (RIO) and observation abundance; and (b) partial dependence plot between relative index of occurrence (RIO) and observation abundance (obtained from TreeNet, non-parametric method).


(a)
Observation abundance
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0
Relative index of occurrence (RIO)


(b)
Partial Denpendence
20
10
0
0.994
0.995
0.996
0.997
0.998
0.999
Relative index of occurrence (RIO)