Ecological Informatics 47 (2018) 67–72

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

On the usefulness of prediction intervals for local species distribution model forecasts

Christian Kampichler*, Henk Sierdsema

Sovon Dutch Centre for Field Ornithology, Natuurplaza (Mercator 3), Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

ARTICLE INFO

Keywords:
Species distribution model
Prediction interval
Quantile regression forest
Random forest

ABSTRACT

The maps produced by species distribution models (SDMs) are increasingly used by decision-makers for supporting local and regional land-use as well as landscape planning issues. While ecologists generally are interested in large-scale patterns and the overall quality of SDMs, decision-makers and conservationists focus on the reliability of localized predictions relevant for specific projects. Here, we use the machine learning methods Random Forest and Quantile Regression Forest to predict local abundance of the black-tailed godwit *Limosa limosa* with prediction intervals, a measure of the probability that a future observation will lie between certain limits. Although the confidence intervals for local predictions are very narrow, the corresponding prediction intervals are very wide. Therefore, the actual numbers of the black-tailed godwit expected at a given point in the field may vary from virtually absent to high density. We conclude that practitioners should lower their expectations of maps based on the currently available SDMs and to be careful when utilizing them for supporting local management decisions.

1. Introduction

In the last decades, research aimed at biodiversity and conservation has leaned increasingly on the contributions of the proliferating field of citizen science (Dickinson et al., 2010; Greenwood, 2007). The application of modern community technology has enabled volunteers to collect and provide large amounts of observational data, by participating in standardized monitoring schemes or by simply recording casual observations through internet-based data portals such as ebird.org or observado.org, and by cell-phone based facilities which allow georeferenced data input directly in the field. Worldwide, citizens feed scientific databases with their observations and allow their use for scientific investigation. However, even in a small country like the Netherlands (42,000 km^2) with a large number of birders (approx. 7000) contributing to national monitoring and atlas schemes, it is not possible to accomplish 100% coverage and many spatial data gaps have to be interpolated by species distribution models (SDM). In parallel to the ever-increasing availability of data, SDMs thus have experienced a large increase in popularity among ecologists in the last 15 years (Elith and Leathwick, 2009). SDMs combine observations (presence only, presence-absence or abundance) with environmental predictors such as climate characteristics, landscape and land-use features or vegetation, and are used to make predictions for localities that have not (sufficiently) been sampled.

With the growing availability of maps based on the predictions of SDMs, they have become increasingly used, not only among ecologists but also among local, regional and national decision-makers and conservationists. Maps showing the distributions and abundances of birds or other biota are used for various aims, such as the evaluation of nature reserve allocation, the designation of important bird areas at the national or regional scale, for urban and regional planning or the allocation of new infrastructure (roads, industrial plants) (e.g. Cabeza et al., 2004; Sierdsema et al., 2013). In the Netherlands public authorities have indicated a preference to consult maps based on SDMs and to target funding for agri-environment schemes in areas with the highest predicted abundance (Anon., 2017). With these new applications of SDMs, also new requirements are emerging. Ecologists are generally interested in large-scale patterns and the overall quality of SDMs. Decision-makers and conservationists, however, tend to focus on the reliability of local predictions relevant for a specific project. If the map delivered by an SDM is to be of practical use to them, they need some indication of how close the predictions for a particular location could come to future observations. One way to approach this demand by the practitioners is to calculate local prediction intervals, a measure of the probability that a future observation will lie between certain limits. Here, we describe how we use a Quantile Regression Forest (QRF)

Abbreviations: OOB, "out-of-the-bag"; QRF, quantile regression forest; RF, random forest; SDM, species distribution model

* Corresponding author.

E-mail addresses: christian.kampichler@sovon.nl (C. Kampichler), henk.sierdsema@sovon.nl (H. Sierdsema).

http://dx.doi.org/10.1016/j.ecoinf.2017.07.003
Received 21 February 2017; Received in revised form 19 June 2017; Accepted 21 July 2017
Available online 24 July 2017

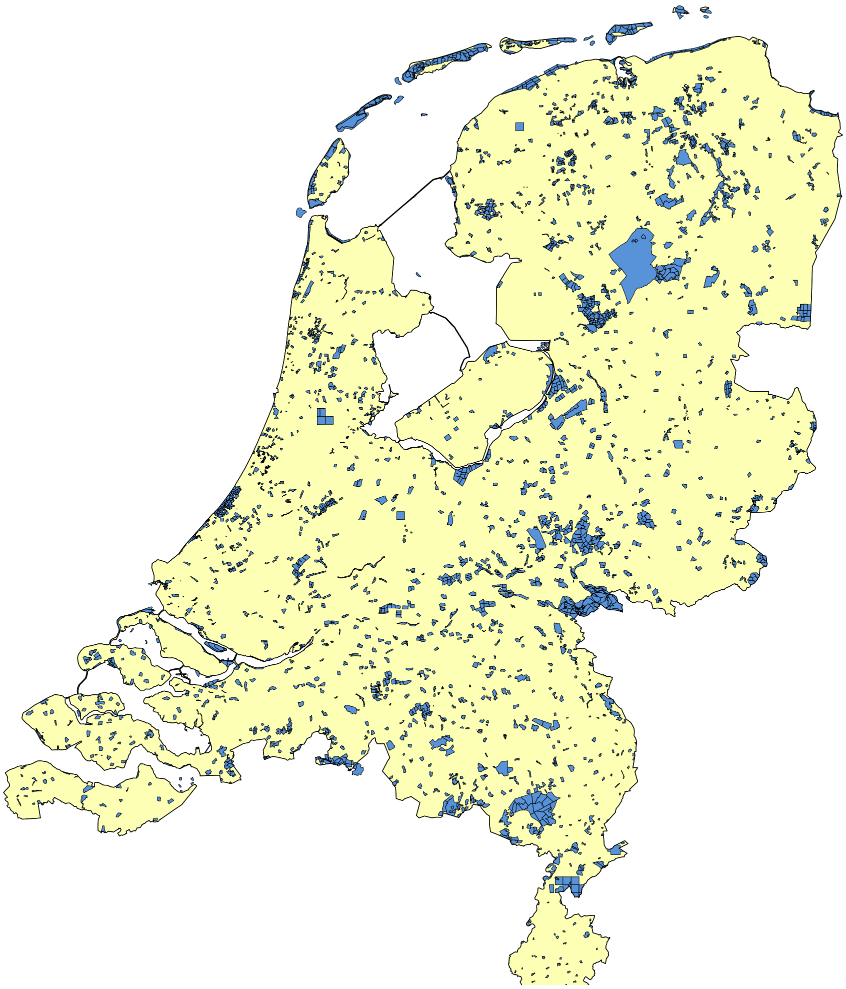

Fig. 1. Plots of the Dutch breeding bird monitoring programme that were surveyed at least once between 2010 and 2014.

(Meinshausen, 2006)—a novel machine learning method—for providing the prediction of local abundance with prediction intervals. Will prediction intervals, however, measure up to the expectations of practitioners and yield useful information for taking decisions on a local or regional scale? To evaluate this question we modelled the abundance of the flagship species of Dutch meadow bird conservation, the black-tailed godwit *Limosa limosa*, which has been elected as the national bird of the Netherlands (https://en.wikipedia.org/wiki/List_of_national_birds), and determined its local prediction intervals.

2. Theory

Generalised linear models were initially the most prominent tool (Schröder and Reineking, 2004) used in species distribution modelling for spatial interpolation. However, recently methods from the field of machine-learning, such as artificial neural networks or boosted regression trees, have become increasingly popular. This is due to the fact that they can better cope with high-dimensional, non-linear and collinear datasets. Especially one machine-learning technique, Random Forests (RF) (Boulesteix et al., 2012; Breiman, 2001a; Liaw and Wiener, 2002), has become particularly successful. It combines the above-mentioned characteristics with a highly efficient approach for ranking variables according to their ability to predict the response, with a low proneness towards over-fitting. It also provides the opportunity to incorporate large numbers of sometimes strongly correlated explanatory variables. Particularly this last feature of RF has often led to a heated debate between researchers in machine-learning and statisticians using stochastic data models. The latter generally advise to restrict the number of predictors used in the modelling process as much as possible. The machine-learning approach, in contrast, has turned the “curse of dimensionality” into a “blessing of dimensionality” (Breiman, 2001b).

Briefly, RF are based on the idea of training a large number *n* of single classification or regression trees, a machine learning method introduced three decades ago (Breiman et al., 1984). For each of the *n* trees only a bootstrapped sample of the cases is used—in the context of SDMs, a ‘case’ corresponds to the unit of spatial interpolation, typically a cell in a gridded representation of the area under study—and in each node in the trees only a random subsample of the explanatory environmental variables is used. Thus each of the trees grown in the forest will yield another set of predictions dependent on the cases and variables chosen. Finally, in case of classification all trees assign each case into a category according to a majority vote, or in case of regression calculate their mean value of predictions for each case. The cases not used for generating a tree—the so-called “out-of-the-bag” (OOB) case-s—are used for the evaluation of the respective single trees and for the determination of overall fitting quality and variable importance. RF have been applied in many SDM and similar modelling contexts (e.g., Benito-Garzon et al., 2006; Brandt et al., 2017; Cutler et al., 2007; Evans et al., 2011; Kampichler et al., 2010; Mascaro et al., 2014) and notably in several recent bird atlas projects such as the atlas of breeding and wintering birds of Britain and Ireland (Balmer et al., 2013) and the atlas of common breeding birds in Poland (Kuczyński and Chylarecki, 2012).

The mean of the predictions of all the trees in a random forest illustrates just one aspect of the distribution of the response variable, all other features of possible interest, however, are neglected. In statistics,

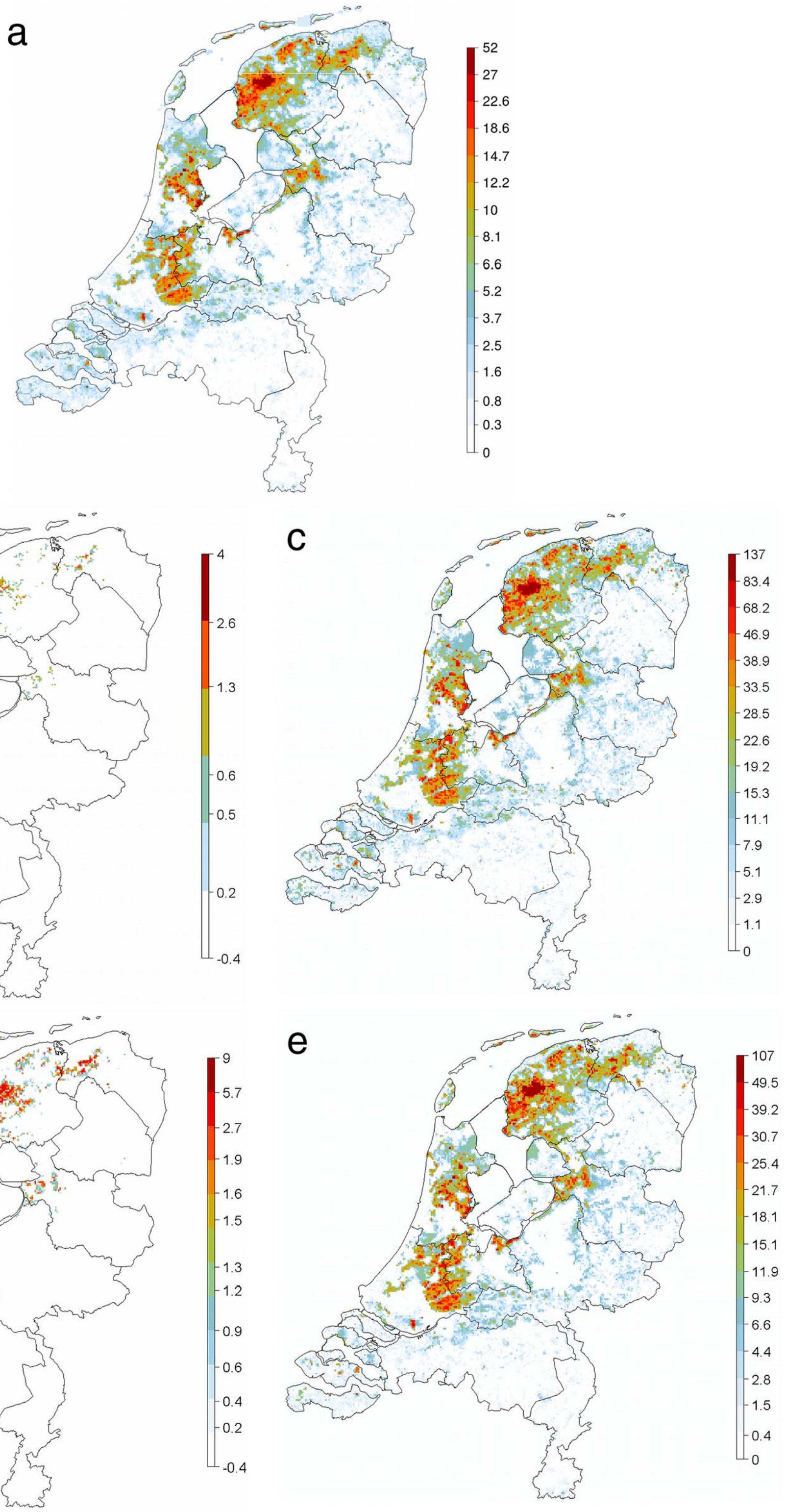

Fig. 2. Predicted abundance of breeding pairs of black-tailed godwits *Limosa limosa* per km^2 in the Netherlands based on a random forest model (a) and the lower (b, d) and upper (c, e) limits of the 90% (b, c) and 75% (d, e) prediction interval as determined by a quantile regression forest model. The colour legends for each panel were constructed using Jenks natural breaks classification method to guarantee the best arrangement of abundance values into different classes for optimal visualisation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the idea of quantile regression seeks to extend the regression concept to the estimation of conditional quantile functions (Koenker, 2005; Koenker and Bassett, 1978). This approach was adopted by Meinshausen (2006) and applied to the random forest technique, leading to the development of Quantile Regression Forests (QRF). First, QRF are grown using the standard random forest algorithm; second, the complete conditional distribution of the response variable is used for quantile determination instead of only using the conditional mean. The appealing feature of QRF is that they can be used for determining prediction intervals and thus provide a means to evaluate the reliability of local predictions of abundance, matching the requirements of decision-makers and conservationists outlined in the introduction. A prediction interval is the statistical interval that will contain a future observation—in the case of an SDM a missing observation from a location that has not been sampled—with a certain level of confidence (Hahn and Meeker, 1991). In QRF, it is determined by the interval between the quantiles $Q_{(1-\alpha)/2}$ and $Q_{(1+\alpha)/2}$ of the distribution of single tree predictions for a given grid cell and a given specified probability α. For example, the 90% prediction interval for a future observation in a grid cell would lie between the 5%-quantile and the 95%-quantile.

3. Material and methods

We applied RF for modelling the spatial abundance patterns of the black-tailed godwit *Limosa limosa* in the Netherlands and QRF for determining the prediction intervals for the local predictions. Abundance data were obtained from the Dutch breeding bird monitoring programme (van Dijk and Boele, 2011; Vergeer et al., 2016), a territory-mapping scheme on fixed counting plots throughout the country. The godwit data comprised territory data from 2010 to 2014 and were based on 2816 plots that were surveyed at least once in this time period (Fig. 1). For plots that were surveyed more than once in this time period, abundance was averaged. For generating the SDM we applied the software package TRIMmaps (Kampichler et al., 2016) developed at the Sovon Dutch Centre for Field Ornithology based on the statistical language R (R Development Core Team, 2016) and applying the R packages randomForest (Liaw and Wiener, 2002) and quantregForest (Meinshausen, 2012). We used 82 environmental features (percentage cover of different land-use types, openness of the landscape, quantity of natural structures such as hedges or reed-belts, variables characterising the soil-water balance, among others) as explanatory variables at a scale of 1000 × 1000 m. In the RF modelling run, a nested cross-validation (function rfcv of the randomForest package) was used to identify the subset of explanatory variables with a signal and to remove spurious variables. This reduces the votes that account for noise and can lead to an overall reduction in error (Evans et al., 2011). The procedure left 18 explanatory variables in the analysis, which were also used in the QRF modelling run. Furthermore, we used the function tuneRF to find the optimal number of explanatory variables (in terms of minimized OOB error) used in each split, both in the RF and the QRF run. Both RF and QRF consisted of 500 trees each. Model quality of the RF was characterised as the percentage of explained variation and calculated as

$$1 - \frac{MSE_{OOB}}{\hat{\sigma}_y^2}$$

where MSE_{OOB} is the mean of the squared residuals and itself was calculated as

$$MSE_{OOB} = n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i^{OOB})^2$$

where $\hat{y}_i^{OOB}$ is the average of the out-of-bag predictions for the i^{th} observation (Liaw and Wiener, 2002). With the QRF model we calculated the 90% and 75% prediction intervals of black-tailed godwit abundance for each cell in the grid.

4. Results and discussion

The abundance of the black-tailed godwit could satisfactorily be modelled with an explained variation of 47.2% and a Spearman correlation coefficient between observed and predicted abundance of 0.719. The largest part of the black-tailed godwit population in the Netherlands is concentrated in the northern and western provinces (Fig. 2a). The confidence intervals of the local means are very narrow and maps depicting them are virtually indistinguishable from the mean predictions (not shown). According to expert knowledge, the overall abundance pattern predicted by the model seems very realistic. However, the map can cause entirely misleading expectations for making local management decisions when abundances are observed at the scale of single grid cells for making local management decisions. Available habitat space is almost never saturated and environments that seem to fulfil all ecological requirements for a species may be underpopulated (or even unoccupied) due to unobserved processes such as dispersal limitations, demographic processes, biotic interactions or simply chance. Due to the same reason, environments that do not appear to be optimal can show higher densities than expected. A more realistic picture, thus, would be to display expected maximum and minimum numbers of black-tailed godwits per grid cell. The limits of the 90% prediction interval (Fig. 2b and c) make clear that the actual numbers of the black-tailed godwit to be expected at a given point in the field may vary from virtually absent to high density. The lower and upper limits converge at the 75% prediction interval (Fig. 2d and e) but the interval is still very wide. Fig. 2 is designed using Jenks natural breaks classification method to guarantee the best arrangement of abundance values into different classes for optimal visualisation (Jenks, 1967) of each panel. The width of the prediction intervals becomes even clearer when all panels are shown using the classification for the panel with widest range (that of Fig. 2c) (Fig. 3).

Clearly, such wide prediction intervals are of little practical use for decision-makers if they are interested in quantitative rather than mere presence-absence information. So the question arises if and how the prediction intervals can be narrowed. One could loosen the demands on probability, for example, and get narrower prediction intervals by choosing only very moderate levels of confidence such as 70% or lower, depending on which planning decision has to be taken or which level of protection the species under study deserves. This approach swiftly reaches a limit, however, since a prediction limit of 50% is uninformative (the expected local abundance lies either within or without the interval, both with the same probability). Additionally, the effects of various decisions in the modelling process must be investigated. We expect, for example, that the scale chosen for the representation of environmental factors and the grid chosen for making predictions will affect the prediction intervals. The smaller the size of the prediction grid, the larger will be the gap between the observation extremes, which will lead to more uncertainty in the predictions. The larger the grid (and thus coarsening what we called “local” prediction) the more environmental variation and heterogeneity of species distribution will level out and lead to narrower prediction limits. How scale affects the outcome of an SDM is a current topic of spatial modelling research (e.g. Sardá-Palomera et al., 2012; Suárez-Seoane et al., 2014).

Ideally, the best way to be able to make better and more precise predictions is to have more and better observational data available. The current atlas project of the Sovon Dutch Centre for Field Ornithology (Altenburg et al., 2017; Schekkerman et al., 2012; van Turnhout et al., 2012) will produce a wealth of information going far beyond the data available based on a breeding bird monitoring scheme alone (as used in this study). We will thus further explore the possibilities and limitations of RF/QRF for deriving local abundance estimates including prediction intervals.

5. Conclusions

Many authors state that prediction is the key to ecological understanding (e.g. Drew et al., 2011; Houlahan et al., 2017) and models should aim for high predictive accuracy and precision. Nonetheless, it has to be remembered that correlative models have their predictive limits if unobserved processes, such as dispersal limitations or biotic interactions, are not included. An explained variation of 47.2% simply means that more than half of the abundance variation is not yet explained. For many species in the Netherlands we have to be content with models explaining considerably less variation. Thus large uncertainty regarding the local predictions is really not surprising. Integrating biotic interactions, for example, into correlative SDMs is a

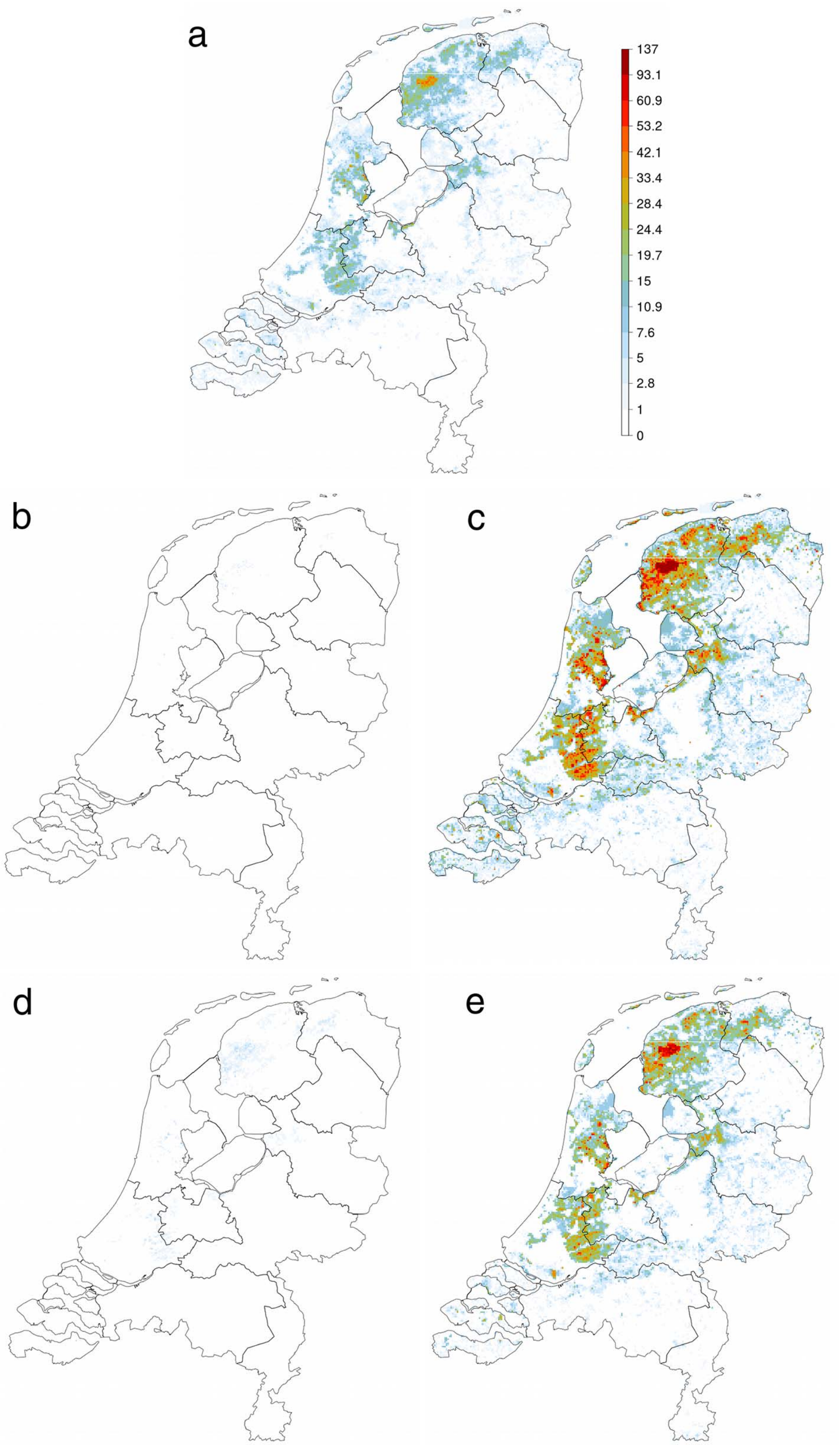

Fig. 3. Same as Fig. 2 but all panels use the classification of Fig. 2c for direct comparability of the panels.

current topic of research and there are no clear guidelines for how it should be achieved (Anderson, 2017). Therefore, we have to advise practitioners to lower their expectations on maps based on the currently available SDMs and to take care when using them for supporting local management decisions.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We thank Maaike Avé for checking the English of the manuscript.

References

Altenburg, J., van Diek, H., Foppen, R., Kampichler, C., Sierdsema, H., Troost, G., van Winden, E., van Turnhout, C., 2017. Fieldwork completed for the fourth Dutch bird atlas, a bonanza of counts and estimates to be utilised. Vogelwelt 137, 23–28.

Anderson, R.P., 2017. When and how should biotic interactions be considered in models of species niches and distributions? J. Biogeogr. 44, 8–17. http://dx.doi.org/10.1111/jbi.12825.

Anon., 2017. Geen subsidie voor akkerranden zonder broedvogels. Provinciale Zeeuwse Courant (Newspaper) January 17th, 2017. http://www.pzc.nl/zeeuws-nieuws/geen-subsidie-voor-akkerranden-zonder-broedvogels~a3416f002/ (accessed on February 20th, 2017).

Balmer, D., Gillings, S., Caffrey, B., Swann, B., Downie, I., Fuller, R., 2013. Bird Atlas 2007–2011 — The Breeding and Wintering Birds of Britain and Ireland. BTO Books, Thetford.

Benito-Garzon, M., Blazek, R., Neteler, M., Sanchez de Dios, R., Sainz, Ollero H., Furlanello, C., 2006. Predicting habitat suitability with machine learning models: the potential area of *Pinus sylvestris* L. in the Iberian Peninsula. Ecol. Model. 197, 383–393. http://dx.doi.org/10.1016/j.ecolmodel.2006.03.015.

Boulesteix, A.-L., Janitza, S., Kruppa, J., König, I.R., 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. WIREs Data Min. Knowl. Discovery 2, 493–507. http://dx.doi.org/10.1002/widm.1072.

Brandt, L.A., Benscoter, A.M., Harvey, R., Speroterra, C., Bucklin, D., Romañach, S.S., Watling, J.I., Mazzotti, F.J., 2017. Comparison of climate envelope models developed using expert-selected variables versus statistical selection. Ecol. Model. 345, 10–20. http://dx.doi.org/10.1016/j.ecolmodel.2016.11.016.

Breiman, L., 2001a. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., 2001b. Statistical modelling: the two cultures. Stat. Sci. 16, 199–231.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. CART: Classification and Regression Trees. Wadsworth, Belmont.

Cabeza, M., Araujó, M.B., Wilson, R.J., Thomas, C.D., Cowley, M.J.R., Moilanen, A., 2004. Combining probabilities of occurrence with spatial reserve design. J. Appl. Ecol. 41, 252–262. http://dx.doi.org/10.1111/j.0021-8901.2004.00905.x.

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

Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010. Citizen science as an ecological research tool: challenges and benefits. Annu. Rev. Ecol. Evol. Syst. 41, 149–172. http://dx.doi.org/10.1146/annurev-ecolsys-102209-144636.

Drew, C.A., Wiersma, Y.F., Huettmann, F., 2011. Conclusion: an attempt to describe the state of habitat and species modeling today. In: Drew, C.A., Wiersma, Y., Huettmann, F. (Eds.), Predictive Species and Habitat Modeling in Landscape Ecology, pp. 291–298. http://dx.doi.org/10.1007/978-1-4419-7390-0_15.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697. http://dx.doi.org/10.1146/annurev.ecolsys.110308.120159.

Evans, J.S., Murphy, M.A., Holden, Z.A., Cushman, S.A., 2011. Modeling species distribution and change using random forest. In: Drew, C.A., Wiersma, Y., Huettmann, F. (Eds.), Predictive Species and Habitat Modeling in Landscape Ecology, pp. 139–159. http://dx.doi.org/10.1007/978-1-4419-7390-0_8.

Greenwood, J.J.D., 2007. Citizens, science and bird conservation. J. Ornithol. 148 (Suppl. 1), S77–S124. http://dx.doi.org/10.1007/s10336-007-0239-9.

Hahn, G.J., Meeker, W.Q., 1991. Statistical Intervals. John Wiley & Sons, New York.

Houlahan, J.E., McKinney, S.T., Anderson, T.M., McGill, B.J., 2017. The priority of prediction in ecological understanding. Oikos 126, 1–7. http://dx.doi.org/10.1111/oik.03726.

Jenks, G.F., 1967. The data model concept in statistical mapping. In: International Yearbook of Cartography. 7. pp. 186–190.

Kampichler, C., Wieland, R., Calmé, S., Weissenberger, H., Arriaga-Weiss, S., 2010. Classification in conservation biology: a comparison of five machine-learning methods. Eco. Inform. 5, 441–450. http://dx.doi.org/10.1016/j.ecoinf.2010.06.003.

Kampichler, C., Hallmann, C., Sierdsema, H., 2016. TRIMmaps: An R Package for the Analysis of Species Abundance and Distribution Data – Extended Manual. Sovon Vogelonderzoek Nederland, Nijmegen, Netherlands.

Koenker, R., 2005. Quantile Regression. Cambridge University Press, Cambridge.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.

Kuczyński, L., Chylarecki, P., 2012. Atlas pospolitych ptaków lęgowych Polski — Rozmieszczenie, wybiórczość siedliskowa, trendy. GIOŚ, Warszawa.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2 (3), 18–22.

Mascaro, J., Asner, G.P., Knapp, E.E., Kennedy-Bowdoin, T., Martin, R.E., Anderson, C., Higgins, M., Chadwick, K.D., 2014. A tale of two “forests”: random forest machine learning aids tropical forest carbon mapping. PLoS One 9 (1), e85993. http://dx.doi.org/10.1371/journal.pone.0085993.

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

Meinshausen, N., 2012. quantregForest: Quantile Regression Forests. R Package Version 0.2-3. http://CRAN.R-project.org/package=quantregForest.

R Development Core Team, 2016. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project/org.

Sardá-Palomera, F., Brotons, L., Villero, D., Sierdsema, H., Newson, S.E., Jiguet, F., 2012. Mapping from heterogeneous biodiversity monitoring data sources. Biodivers. Conserv. 21, 2927–2948.

Schekkerman, H., van Turnhout, C., van Kleunen, A., van Diek, H., Altenburg, J., 2012. Naar een nieuwe vogelatlas: achtergronden van de veldwerkopzet. Limosa 85, 133–141.

Schröder, B., Reineking, B., 2004. Modellierung der Art-Habitat-Beziehung – ein Überblick über die Verfahren der Habitatmodellierung. In: Dormann, C.F., Blaschke, T., Lausch, A., Schröder, B., Söndgerath, D. (Eds.), Habitatmodelle – Methodik, Anwendung, Nutzen, UFZ-Berichte, 9/2004. UFZ, Leipzig, Germany.

Sierdsema, H., Schotman, A.G.M., Oosterveld, E.B., Melman, T.C.P., 2013. Weidevogelkerngebieden Noord-Holland; vergelijking van vier scenario's. In: Sovon-rapport 2013/23. Sovon Vogelonderzoek Nederland, Nijmegen.

Suárez-Seoane, S., Virgós, E., Terroba, O., Pardavila, X., Barea-Azcón, J.M., 2014. Scaling of species distribution models across spatial resolutions and extents along a biogeographic gradient. The case of the Iberian mole *Talpa occidentalis*. Ecography 37, 279–292. http://dx.doi.org/10.1111/j.1600-0587.2013.00077.x.

van Dijk, A.J., Boele, A., 2011. Handleiding Sovon Broedvogelonderzoek. Sovon Vogelonderzoek Nederland, Nijmegen.

van Turnhout, C., Altenburg, J., Foppen, R., 2012. A new Dutch atlas project on the go. Bird Census News 25, 22–29.

Vergeer, J.W., van Dijk, A.J., Boele, A., van Bruggen, J., Hustings, F., 2016. Handleiding Sovon Broedvogelonderzoek: Broedvogel Monitoring Project en Kolonievogels. Sovon Vogelonderzoek Nederland, Nijmegen.