00000100000
00001010000
00010001000
00100000100
01000000010
10000000001
Software notes

Ecography 39: 001–008, 2016
doi: 10.1111/ecog.02593

Subject Editor: Thiago Rangel. Editor-in-Chief: Miguel Araújo. Accepted 5 October 2016

Spind: a package for computing spatially corrected accuracy measures

Gudrun Carl and Ingolf Kühn

G. Carl (gudrun.carl@ufz.de) and I. Kühn, Helmholtz Centre for Environmental Research – UFZ, Dept of Community Ecology, Halle, Germany. IK also at: Martin Luther Univ. Halle-Wittenberg, Geobotany and Botanical Garden, Halle, Germany, and German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany.

Using an appropriate accuracy measure is essential for assessing prediction accuracy in species distribution modelling. Therefore, model evaluation as an analytical uncertainty is a challenging problem. Although a variety of accuracy measures for the assessment of prediction errors in presence/absence models is available, there is a lack of spatial accuracy measures, i.e. measures that are sensitive to the spatial arrangement of the predictions. We present 'spind', a new software package (based on the R software program) that provides spatial performance measures for grid-based models. These accuracy measures are generalized, spatially corrected versions of the classical ones, thus enabling comparisons between them. Our method for evaluation consists of the following steps: 1) incorporate additional autocorrelation until spatial autocorrelation in predictions and actuals is balanced, 2) cross-classify predictions and adjusted actuals in a 4×4 contingency table, 3) use a refined weighting pattern for errors, and 4) calculate weighted Kappa, sensitivity, specificity and subsequently ROC, AUC, TSS to get spatially corrected indices. To illustrate the impact of our spatial method we present an example of simulated data as well as an example of presence/absence data of the plant species *Dianthus carthusianorum* across Germany. Our analysis includes a statistic for the comparison of spatial and classical (non-spatial) indices. We find that our spatial indices tend to result in higher values than classical ones. These differences are statistically significant at medium and high autocorrelation levels. We conclude that these spatial accuracy measures may contribute to evaluate prediction errors in presence/absence models, especially in case of medium or high degree of similarity of adjacent data, i.e. aggregated (clumped) or continuous species distributions.

Background

Accuracy measures such as Cohen's kappa coefficient (or Kappa for short) are coefficients useful to assess prediction errors in presence/absence models (such as species distribution models). In a spatial context, however, the traditional non-spatial measures are not appropriate and can thus be misleading in species distribution modelling (Fielding 2002). The reason is that a false prediction has simply the quality of being false regardless of its distance to an appropriate actual value and thus true prediction. One can argue, though, that a false prediction of presence in close proximity to a true (observed) presence is better than a false presence far away from an observed presence (Fielding and Bell 1997, Fielding 2002).

This is particularly the case when sampling at nearby locations leads to sample values that are not statistically independent from each other. If so, then it is to be expected that predictions have the same nature. This phenomenon of statistical dependence caused by spatial dependence should be considered as relevant. This applies particularly to sampling on raster maps, where original data maps are sectioned into grids (Hagen-Zanker 2009). Due to a relatively arbitrary specification of cell size and grid orientation, discretization will generally cause a loss of information. Occurrences at grid cell boundaries, for instance, must be allocated to a specific grid cell (ignoring proximity to the neighbour cell) (Shekhar et al. 2002). This is one reason for analysing spatial neighbourhoods and incorporating spatial dependence into accuracy assessment.

There are also ecological reasons for integrating spatial context. Given a species range, searching for new off-range occurrences, one would look probably more frequently and expect more likely to find a new occurrence close to its range margin than further away from its range margin. One of the reasons is that many expanding species have more frequently range advances close to its range margin than those rare long-distance dispersals resulting in occurrences far away (Nathan and Muller-Landau 2000, Nathan et al. 2002). Another is dispersal limitation, resulting in new occurrences close to known occurrences even under suitable environmental conditions further away (Svenning et al. 2006). Hence models exist to account for sampling bias, giving location further away from currently know occurrences a lower likelihood of being occupied (Bierman et al. 2010, Manceur and Kühn 2014).

Classical accuracy measures do not take into consideration the spatial context of any mispredictions. They neglect the degree of similarity of adjacent data. In reality, however, maps of both actual and predicted values have some degree of spatial autocorrelation (Hagen-Zanker 2009). In the presence of spatial autocorrelation of model residuals, the use of methods accounting for this is recommended (Carl and Kühn 2007, 2010, Dormann et al. 2007). These approaches account for problems in parameter estimation and realized degrees of freedom resulting in non-autocorrelated residuals. Hence they make sure that no fundamental assumptions of hypotheses testing and statistical approaches are violated. They do not yield, though, uncorrelated predictions. Hence the results of such models, when using traditional, non-spatial measure of accuracy, can potentially also suffer from the problems outlined above. Therefore, the use of spatial metrics of accuracy is even necessary when using methods to account for autocorrelation in model calibration.

For illustration purposes, the maps in Fig. 1 show details of the results of two grid-based models. Although the prediction in Fig. 1b is located in closer proximity to actuals than the prediction in Fig. 1a, classical performance measures assign both predictions to the class of false positive errors. In other words, classical measures suffer from the problem that accuracy is not a function of spatial arrangement. Instead, all falsely predicted positive errors rank equally as well as all falsely predicted negative errors, independent of the distance to actual (observed) values.

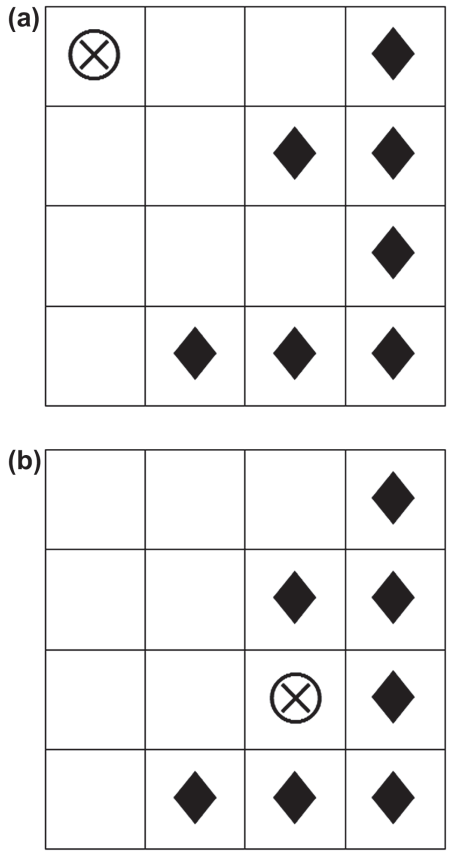

Figure 1. Example showing a prediction of presence as a result of two different models in relation to the same actual values, i.e. observed presences. Cells with/without diamond ◆ indicate presence/absence of actual values and cells with/without circlemultiply ⊗ refer to presence/absence of predicted values. (a) Locations in the first model, (b) locations in the second model. In spatial sense, the prediction in (b) might be more accurate than the prediction in (a).

Here, we present and describe the new software package 'spind', which introduces several spatial accuracy measures that are a) sensitive to the spatial arrangement of predictions and b) comparable to classical measures.

a) As alternative measures for the evaluation of grid-based models, they will take into account that a false prediction may not be completely wrong if it is in a certain spatial proximity to the correct result. The degree of dependency can be measured and analysed by correlograms, i.e. computations of spatial autocorrelation of both predicted and actual values. Moreover, a new classification and weighting scheme for predictions is needed.

b) We are not interested in developing totally new spatial measures. Such spatial measures already exist, as for instance, the average distance to nearest prediction (ADNP) and the Spatial Accuracy Measure (SAM) (Shekhar et al. 2002). They have the disadvantage that their results cannot be compared to those of non-spatial measures. Instead, the aim of our study is to generalize classical measures. To enable efficient comparisons, we modify and improve well-known measures (i.e. Kappa, as well as sensitivity, specificity, true skill statistic and other ones) to spatially corrected versions.

Methods

Spatially corrected method

The performance of a presence/absence model is often summarized in a confusion matrix (Table 1). This is a 2×2 contingency table that cross-classifies observed occurrences (i.e. actual presence/actual absence) and predicted ones according to two classes (i.e. predicted presence/predicted absence). Several classical measures are based on a calculation and evaluation of this confusion matrix. The threshold dividing into classes of predicted presences and absences has frequently the value threshold = 0.5, but any other threshold value within the interval from 0 to 1 could be chosen, e.g. based on prevalence or maximizing traditional accuracy measures such as Kappa or true skill statistic. When setting the threshold to 0.5, then the probability of presences is the same as the probability of absences.

Table 1. Confusion matrix as a 2×2 contingency table. Threshold is the threshold used to transform predicted probability of occurrence of species distribution models into 1's and 0's, for instance, for presence/absence maps.

	Actual (presence) 1	Actual (absence) 0	Total
Predicted (presence) 1 – threshold	True positive n_{11}	False positive n_{12}	$n_{1.}$
Predicted (absence) threshold – 0	False negative n_{21}	True negative n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

Fielding and Bell (1997) used two simple approaches of weighting in a spatial framework. These are methods that weight false positive errors n_{12} by a function of their distance/proximity to actual positive locations and thus provide adjusted false positive errors. In this way, the roughly weighted proximity relationships reflect autocorrelation for locations in a two-dimensional gridded dataset. As a result the ratio of adjusted errors to actual errors is recommended for assessment. The magnitude of weights (and thus the strength of autocorrelation), however, was chosen relatively arbitrarily. To circumvent this problem, one can propose new map similarity measures without any weights. One of such measures is the average distance to nearest prediction (ADNP) (Shekhar et al. 2002). This value (i.e. arithmetic mean of distances), however, is not related to a confusion matrix and its corresponding evaluation measures. Conversely, one can try to incorporate a spatial weights matrix reflecting the real proximity relationships into the confusion matrix. Shekhar et al. (2002) developed the spatial accuracy measure (SAM) based on such a generalized confusion matrix. Because a direct combination of different distance measures within one confusion matrix is problematic, the spatial weights matrix is incorporated into all elements of the confusion matrix. But as a consequence of this, all totals change in comparison to the classical confusion matrix and thus renormalization limiting their comparability to the classical confusion matrix is necessary. Hagen-Zanker (2009) introduced an improved Kappa statistic with particular focus on neighbour cells. This extension of the weighted Kappa takes the effect of spatial autocorrelation into consideration, however, without directly quantifying spatial autocorrelation. Instead, the approach tries to estimate its effects by counting adjacent neighbour cells and distinguishing between different degrees of belonging.

To overcome all these problems, we 1) implement proximity as the same amount of spatial autocorrelation in both actual and predicted values and 2) summarize the results in a weighted 4×4 contingency table.

1) For spatial data, the amount of spatial autocorrelation can be calculated by means of the Moran's I (Lichstein et al. 2002). This formula measures the strength of two-dimensional autocorrelation based on the assumption that it is isotropic (i.e. independent of direction). Autocorrelation is computed as a function of 'lag distance', therefore, one has to introduce lag distance intervals for the spatial structure under consideration. For a square grid underlying all maps used here, the first distance class can be defined to comprise lags between 0 and 1 and thus be assigned to nearest neighbours, i.e. to the (generally) four adjacent grid cells located at distance unit 1 (in relation to coordinates of cell centres) in the cardinal directions. Autocorrelation at lag distance 1 is generally higher than that at greater distances because close observations are more likely to be similar to one another than those far away from each other. Therefore, the autocorrelation value ac(1) is most important. It is noteworthy that the spatial autocorrelation ac(1) of predicted values (i.e. predictions before dividing into groups by a threshold) is generally higher than that of actual values. The reason is that predictions are continuous values varying between the extremes 0 and 1, whereas actual values simply consist of 0's and 1's. This autocorrelation deficit of actuals can be considered as a measure to what extent actual values can be adjusted to reflect a spatial context. Therefore, we generate 'adjusted actuals' having the same amount of autocorrelation as predictions. These adjusted actual values are softened compared to the original ones and, accordingly, appear widened in spatial mapping. Therefore, a prediction at a single location can be registered to be in the proximity (i.e. widened area) of an actual value. It is to remark, that, computationally, it is difficult to increase the autocorrelation of actuals in one step to a certain level. Here, we use a step-by-step procedure incorporating autocorrelation until it is balanced with the autocorrelation of predictions.

2) For evaluation, one has to summarize the results for predicted and adjusted actual values in a generalized confusion matrix (Table 2). In order to ensure that the additional information captured in adjusted actual values is not completely lost again, it is necessary to make the contingency table 'finer'. If we cross-classify the distributions of the variables in a 4×4 contingency table then we are able to distinguish different kinds of misclassification. Therefore, the predicted values have to be classified into 4 classes separated at the following 3 levels: 1) upper split: $us = (1 + \text{threshold})/2$, 2) threshold: $th = \text{threshold}$, and 3) lower split: $ls = \text{threshold}/2$. Since the total of elements remains constant, a comparison to the results of a 2×2 contingency table is possible. Three cells n_{ij} in the upper right corner (for: $j - i \geq 2$: n_{13}, n_{14}, n_{24}, displayed in dark-grey) contain false positive errors, whereas three cells in the bottom left corner (for: $i - j \geq 2$: n_{31}, n_{41}, n_{42}, displayed in dark-grey) contain false negative errors. This refined weighting pattern can simply be written in matrix notation, i.e. by means of the weighting matrix W

$$W = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

Having specified the values of this refined cross-classification, we can calculate measures such as weighted Kappa, sensitivity,

Table 2. Generalized confusion matrix as a 4×4 contingency table. As in Table 1, dark grey cells are considered as false while light grey ones as true. Please note that n_{32} and n_{23} would be classified as false in the classical approach but as true here due to the close match. *us*: upper split, *ls*: lower split, *th*: threshold used.

	Adjusted actual 1–0.75	Adjusted actual 0.75–0.5	Adjusted actual 0.5–0.25	Adjusted actual 0.25–0	Total
Predicted 1 – *us*	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
Predicted *us* – *th*	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
Predicted *th* – *ls*	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$
Predicted *ls* – 0	n_{41}	n_{42}	n_{43}	n_{44}	$n_{4.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	n

and specificity for evaluation of prediction accuracy. The weighted Kappa κ is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where $p_o = \sum_i \sum_j w_{ij} p_{ij}$ and $p_e = \sum_i \sum_j w_{ij} p_{i.} p_{.j}$ with $p_{ij} = n_{ij} / n$ (Fleiss and Cohen 1973, Fleiss 1981, Sachs and Hedderich 2006). Accordingly, the formulas for the weighted *sensitivity* and weighted *specificity* can be given by

$$sensitivity = \left(\sum_i \sum_k w_{ik} n_{ik}\right) \Big/ \sum_i \sum_k n_{ik} \text{ for } k = 1,2$$

and

$$specificity = \left(\sum_i \sum_l w_{il} n_{il}\right) \Big/ \sum_i \sum_l n_{il} \text{ for } l = 3,4$$

By computing sensitivity and specificity as functions of threshold, other measures such as receiver operating characteristic (ROC), the area under the ROC curve (AUC), and maximum true skill statistic (TSS) can be calculated as usual (Hanley and McNeil 1982, Franklin 2009).

In summary, our new method for evaluation of prediction accuracy consists of the following steps: 1) incorporate additional autocorrelation into binary observation data until spatial autocorrelation in predictions and actuals is balanced, 2) cross-classify predictions and adjusted actuals in a 4×4 contingency table, 3) use a refined weighting pattern for errors, and 4) calculate weighted Kappa, sensitivity, specificity and subsequently ROC, AUC, TSS to get spatially corrected indices.

Package overview

All statistical analyses were performed using the R x64 software ver. 3.1.2 (R Core Team). We provide all tools for calculating spatially corrected indices in our newly created package 'spind'. It is open-source software (published under the GPL public license, ver. 2), and is available as both a package spind_1.0.zip (windows version) and a source package spind.1.0-1.tar.gz. Both R packages, together with documentation, are available on GitHub (<https://github.com/carl55/spind>).

The R package depends on the package 'lattice', which produces Trellis graphics for **R**, as well as 'splancs' with function areapl, which calculates an area of a polygon (Rowlingson and Diggle 1993, Bivand and Gebhardt 2000). Spind contains four functions. Function th.dep calculates threshold-dependent metrics (kappa and confusion matrix), i.e. it depends on a cutoff value used for splitting predictions, whereas function th.indep calculates threshold-independent metrics (ROC, AUC, and (max) TSS). Both functions are based on a 2D analysis taking the grid structure of datasets into account (for a regular gridded dataset, grid cells are assumed to be square). Therefore, another two functions are used internally. Function adjusted.actuals provides adjusted actual values reflecting spatial autocorrelation balanced to predictions. Function acfft calculates spatial autocorrelation. Moreover, an example data set (Fig. 2) is given to demonstrate how one can use the functions.

Illustration and validation

Application to simulated data

Just in order to visualize the effect of step (1) in our analysis, we firstly present an example of simulated data based on a small grid. The model predictions (Fig. 2a) as well as the actual values (Fig. 2b) are displayed within their spatial context, i.e. the 10×10 grid. When we calculate spatial autocorrelation of predicted and actual values and increase the autocorrelation in actuals (Fig. 3), we produce adjusted actuals (Fig. 2c). Figure 2c shows that grid cells in the immediate proximity of the original agglomeration presented in Fig. 2b have now increased values, whereas a few actual presences are slightly reduced. In our example (Fig. 2b, in the bottom right hand corner), a hook-shaped group of adjoined actuals is to be found just as displayed in Fig. 1b. The prediction for the cell surrounded by this hook has the value 0.52. If we use, for instance, a threshold of 0.5, such a value is classified as false positive error in classical theory. For spatially corrected measures, however, we compute an adjusted actual value of 0.35 at this position. In the 4×4 contingency table (Table 2), therefore, this prediction is assigned to element n_{23} and thus is no longer considered an error.

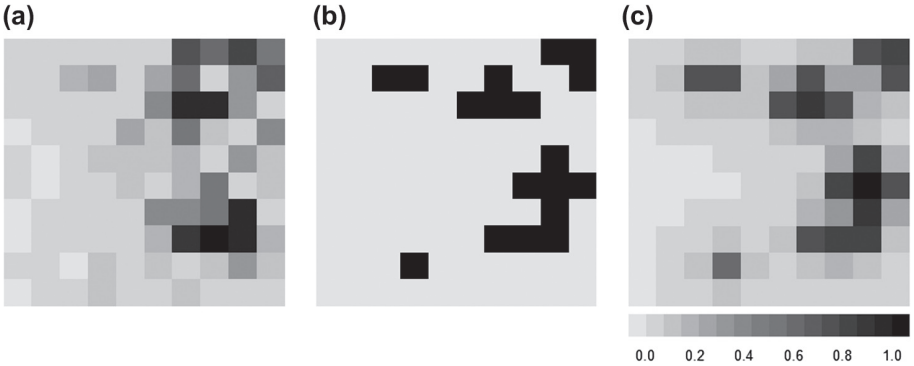

Figure 2. Example map of simulated data. (a) Predicted values, (b) actual values, and (c) adjusted actual values within their spatial context of a 10×10 grid.

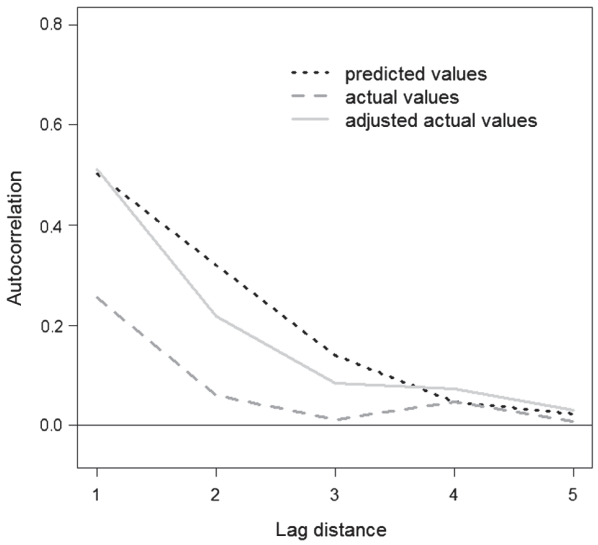

Figure 3. Example correlograms of simulated data. Spatial autocorrelation of predicted, actual, and adjusted actual values.

Application to real macroecological data

Secondly, we compute these spatially corrected indices for a real macroecological dataset. Therefore, we selected data for presence/absence of the plant species *Dianthus carthusianorum* across Germany. This is an example already used in a previous paper (Carl and Kühn 2008). The distribution of actual values of *D. carthusianorum* is given in Fig. 4b. To produce predicted values (Fig. 4a), we related environmental variables (which need not be the most appropriate ones) to these actual values and performed a logistic regression. Information on species distribution is available from FLORKART (<www.floraweb.de>) which contains species location in a grid of 2995 grid cells. The cells of this lattice are 10′ longitude × 6′ latitude, i.e. about 11 × 11 km², and therefore almost square cells. Moreover, we extracted climate variables (temperature, precipitation) provided by the 'Deutscher Wetterdienst, Dept Klima und Umwelt", elevation data from the ARCDeutschland500 dataset provided by ESRI, land use data from Corine Land Cover (1990) raster data, and geology digitized from data provided by the Bundesanstalt für Geowissenschaften und Rohstoffe (1993). As explained above, the spatial method modifies actuals until autocorrelation of actuals and predictions is balanced. In our example, the value of spatial autocorrelation of the actuals is 0.63 at lag distance 1, whereas this value for predictions is 0.87. Due to this difference, the method has to produce adjusted actual values of nearly the same magnitude of autocorrelation, i.e. $ac(1) \approx 0.87$. These adjusted actuals softened compared to the original ones are given in Fig. 4c.

Statistics

Lastly, several steps are undertaken to verify our spatial indices, in detail. At first we generate data of the kind given in Fig. 2a. For this purpose, values for two predictors and errors are randomly generated and provided with a certain degree of spatial autocorrelation. They are linearly combined using specified parameters (intercept and two slopes). This linear combination is scaled and transformed into outcomes ranging from 0 to 1. For normally distributed variables, the mean value of outcomes is 0.5 on average. Hence the prevalence of the simulated species is set to 0.5. If we subsequently split the values in 0's and 1's using a threshold value 0.5, then, of course, these ex-post created 'actuals' match perfectly with the 'predictions' generated prior to this. The values for classical Kappa, AUC, and TSS are 1 in this case of perfect match. This should also be valid for spatial indices. Additionally, in order to check the effect of a certain mismatch, we modify the map of actuals by shifting all columns to the left adjacent position (except for the leftmost column, which is shifted to the rightmost position). One can expect that such a displacement or pattern of 'shifted match' will result in lower fitting accuracy and thus lower values for Kappa, AUC, and TSS. The values depend on the index used for evaluation and, in addition, the degree of spatial autocorrelation. This is because the degree of adjacent similarity is relevant. If values in neighbourhoods are similar, then shifting may be less problematic than if values are randomly distributed and independent. To compare classical measures with our spatially corrected ones (i.e. spatial indices), we generate 30 × 30 maps as described above for both perfect and shifted match at 10 different levels of autocorrelation. Using 100

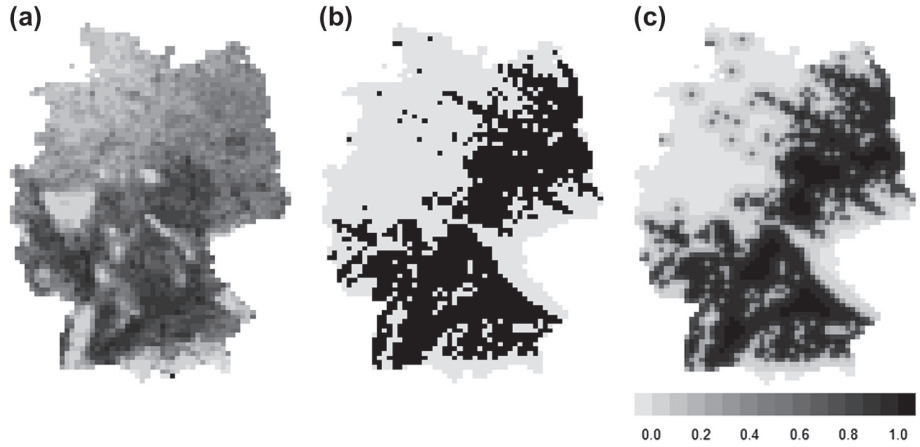

Figure 4. Example of real macroecological data, i.e. the distribution of *Dianthus carthusianorum* in Germany. (a) Predicted values, (b) actual values, and (c) adjusted actual values.

randomly generated datasets in each case, we run all settings 100 times to produce 100 solutions for the indices in each case.

Results

Application to real macroecological data

To demonstrate the impact of our method in a real-world example, the results for classical and spatially corrected measures for presence/absence data of the plant species *Dianthus carthusianorum* are given in Table 3. One can clearly see that the numbers of both false positive errors and false negative errors are less for spatial indices compared to those of classical ones. As a consequence, the values for Kappa (threshold = 0.5), AUC, and TSS increase when incorporating spatial corrections.

Statistics

Using 100 randomly generated datasets to compare classical and spatial indices, we find that the values for both classical and spatial indices reach their maximum of approximately 1 if we use data of perfect match. In case of shifted match, all indices are functions of autocorrelation. Starting at autocorrelation level 0, all indices increase as a function of autocorrelation. As expected and methodologically intended, spatial indices are, on average, equal or higher than classical ones. Higher values occur at medium and high autocorrelation levels due to the increasing degree of adjacent similarity being taken into consideration by spatial indices. We can see that especially when having strong autocorrelation, spatial indices tend to result in higher values and classical indices would indicate a poorer fit. Mean values and error bars for Kappa, AUC, and TSS are given in Fig. 5a. For testing the null hypothesis that the value for classical Kappa is equal to the mean value of spatial Kappa values, we use a 95% confidence interval. It is obtained by $\left(\hat{\kappa}-1.96\cdot\sigma\left(\hat{\kappa}\right),\hat{\kappa}+1.96\cdot\sigma\left(\hat{\kappa}\right)\right)$, where $\hat{\kappa}$ is the mean value of classical Kappa and $\sigma\left(\hat{\kappa}\right)$ is its standard deviation (Kanga et al. 2013). We found that for an autocorrelation value of 0.7, the null hypothesis is rejected and the difference for Kappa values is thus statistically significant. Accordingly, hypothesis tests can be used to evaluate differences between classical and spatial AUC values and between classical and spatial TSS values. In both cases, we found statistically significant differences at autocorrelation values of 0.6, 0.7, and 0.8.

Table 3. Predictions for plant species *Dianthus carthusianorum* across Germany. Results for classical measures and spatially corrected measures (i.e. spatial index).

	Classical index	Spatial index
False positive errors	397	331
False negative errors	462	406
Kappa	0.42	0.46
AUC	0.80	0.85
TSS	0.48	0.57

One might still ask whether spatial autocorrelation of predicted values is appropriate to estimate the autocorrelation deficit of actuals and thus to define their neighbourhoods. More specifically, the question arises whether the predictions are appropriate as a basis for adjusting the observations. We can respond with a counter question: how can one get better estimates for actual values than model predictions? Note that we do not use the predictors themselves such as environmental variables. Instead, outcomes predicted by statistical models such as species distribution models are used here. As a consequence, predictors that are not significant will usually have no impact, or only a minor impact, on predictions. To gain deeper insight into the adjustment of actuals and to discuss the risks of our method, we present a further example of simulated data. For this purpose, values for two non-autocorrelated predictors and for a non-autocorrelated error are randomly generated. They are linearly combined using specified parameters (intercept and two slopes). This linear combination is scaled and transformed into outcomes ranging from 0 to 1. Subsequently, we split the values in 0's and 1's as above. To produce (non-autocorrelated) predicted values, we relate the two predictors to these actual values and perform a logistic regression. Having non-autocorrelated data, we find the same values for both classical and spatial indices. If we instead regress these actuals on predictors affected by a certain degree of autocorrelation, then the fitting accuracy decreases. Note that, in this case, autocorrelation acts as a disturbing factor. The classical indices are the correct ones, and the spatial indices, which falsely impose autocorrelation, result in higher values. Therefore, this example investigates to what extent our method adjusts the observations wrongly towards an incorrect pattern. Mean values and error bars for Kappa, AUC, and TSS are given for 100 randomly generated datasets (Fig. 5b). As can be seen from Figure 5b, the differences in fitting accuracy are less than the standard deviations of classical indices and thus not significant.

Discussion

Our results show that especially under medium to high levels of spatial autocorrelation of predicted data spatial measures of accuracy yielded different results compared to classical measures. We therefore advocate the use of the proposed metrics.

There were several assumptions we made: 1) we corrected the actuals by adding autocorrelation to the same degree as that of the predicted values in order to change binary data to continuous data and to be able to define a neighbourhood. For technical reasons, it is impossible to do it the other way round. Still, this frequently results in lower values than 1 (being absolutely present). 2) While we used quartiles to classify adjusted actuals, we used varying thresholds and, dependent on them, upper and lower splits to classify predicted values. The flexibility for predictions is needed, because it is not always useful to use a threshold of 0.5 (Liu et al. 2005, Hanberry and He 2013). Using the same flexibility for actuals, though, turned out not to be useful when developing the method. 3) It is in the logic of the spatial index method to regard cells with close by values (n_{23} and n_{32} in the generalized confusion matrix) as true, rather

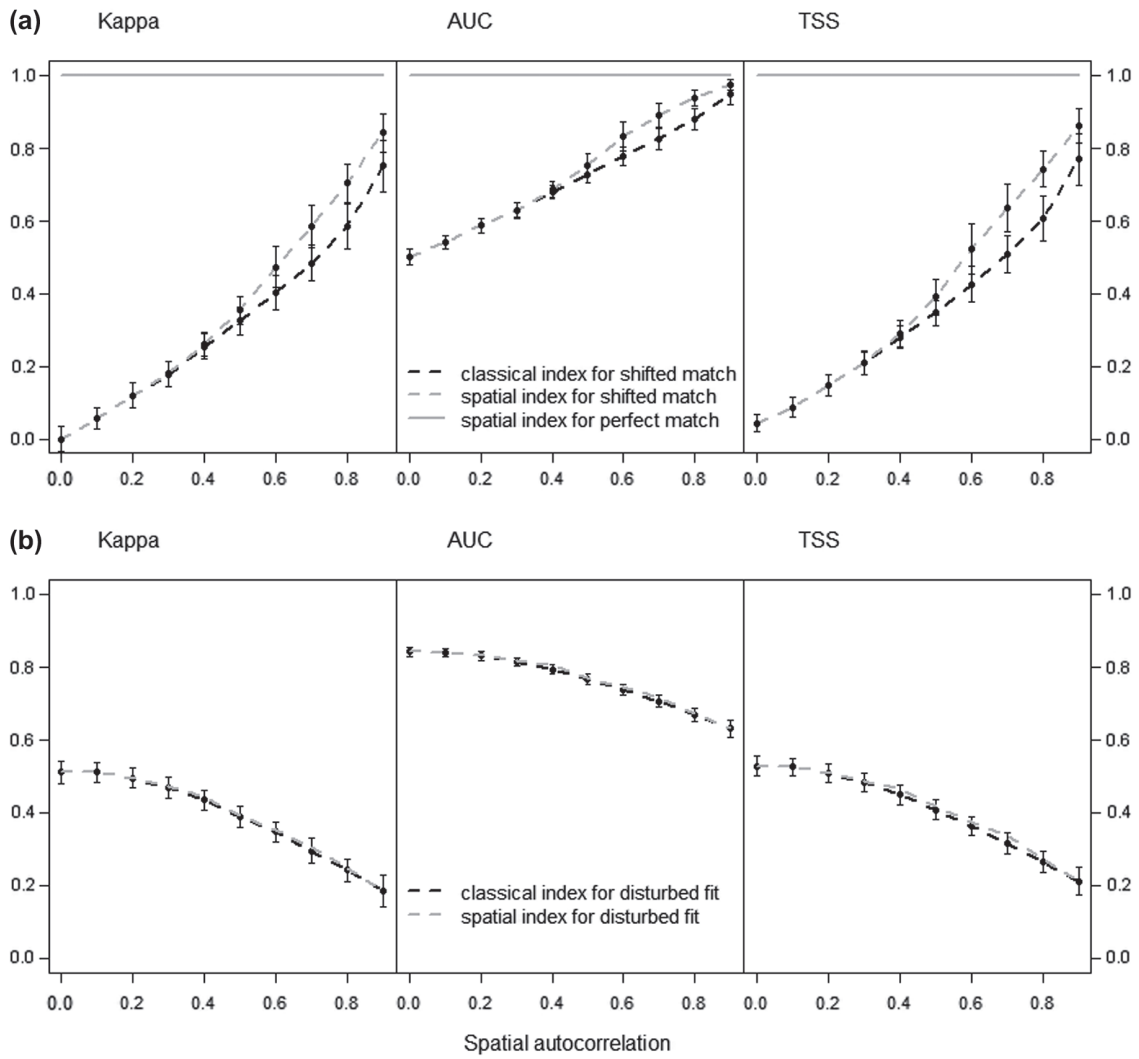

Figure 5. Statistic for comparing classical measures to spatially corrected ones (i.e. spatial indices). The indices Kappa, AUC, and TSS are given as a function of autocorrelation (a) for both perfect and shifted match and (b) for disturbed fit. Spatial autocorrelation is measured as Moran's *I*. The strength of autocorrelation is indicated by ac(1), i.e. the value of autocorrelation related to nearest neighbours. The error bars indicate the interval delimited by mean value ± standard deviation.

than false. But this decision is arbitrary. Not doing so would result in the classical measures of accuracy. 4) The generalized confusion matrix could in principle also have more elements than 4×4 cells. It turned out, however, that this gets computationally difficult, especially with varying thresholds. Further, defining 'true' and 'false' would get very arbitrary. Still, due to defining the spatial neighbourhood, the 4×4 cell confusion table might result in a higher susceptibility to very high or very low prevalences. This means that at small prevalences the number 'present' adjusted actuals (n_{32}) might increase and at very high prevalences the number 'absent' adjusted actuals (n_{23}) might increase at disproportionate rate compared to prevalence (as an effect of an unfavourable edge/area ratio). In such cases, though, with just very few observed presences or observed absences, robust models are inherently difficult to fit. Hence it is warned against the parameterisation of data deficient models, anyhow (Coudun and Gégout 2006, Franklin 2009, p. 63).

As briefly outlined in the methods section, there are measures available that consider spatial proximity. Fielding and Bell (1997) weight the false positives errors by a distance function to actual positive locations. The advantage of our approach is that the degree of spatial weighting is estimated as autocorrelation deficit rather than set arbitrarily. Further, in our approach the marginal sums of the confusion matrix remain the same compared to non-spatial metrics and we also consider the distance of false negatives to actual negatives. The methods of Shekhar et al. (2002) introduced completely different metrics which cannot be compared to the classical metrics, by design. It is our utmost aim to retain comparability between spatial and non-spatial metric but minimize arbitrary decisions.

The results of our simulations (Fig. 5) suggest that the proposed spatial measures of model accuracy only increase accuracy and do not decrease accuracy. This, however, is not inevitable. Because known presences get down-weighted and known absences get up-weighted by adjusting the actuals, in principle fit could (slightly) decrease, but will probably very rarely happen. So on average model accuracy increases when using spatial metrics. One could then argue that this is not

helpful and does not warrant using the new approach. Using our metrics will help to formalize spatial uncertainty and may even account (partially) for unobserved, though present, actuals, i.e. occurrences that indeed are there but were not yet observed. To some degree observer bias (Manceur and Kühn 2014) can thus be minimized when assessing accuracy. Further, our approach increases the comparability of results between autocorrelated and non-autocorrelated data. And lastly we argue that the use of spatial measures of accuracy is better, since we think it is the correct measure, compared to the use of non-spatial measures, in case of autocorrelated data.

Having non-autocorrelated data, our simulations suggest that there is no difference between the spatial and the classical measures of accuracy. So one could use them but it is not necessary. In non-autocorrelated situations, therefore, spatial arrangements of predictions and actuals become irrelevant. In the presence of autocorrelated data, however, one is advised to already use spatial metrics of accuracy.

One issue that still remains unsolved is to properly measure accuracy when having presence-only data. Since both, classical and spatial metrics, need presence as well as true absence data, they are inappropriate when using presence-only data. The results heavily depend on the choice of the algorithm used to select pseudo-absences (Barbet-Massin et al. 2012). This is a fruitful and rewarding topic for future research.

We conclude that these spatial accuracy measures are useful, especially in case of medium or high degree of similarity of adjacent data. They are primarily intended as goodness-of-fit measures for the evaluation of species distribution models based on high resolution maps.

To cite Spind or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for 'version 0':

Carl, G. and Kühn, I. 2016. Spind: a package for computing spatially corrected accuracy measures. – Ecography 39: 000–000 (ver. 0).

Acknowledgements – We acknowledge support by the EU Collaborative project 'EU BON: Building the European Biodiversity Observation Network' (grant 308454) funded under the European Commission Framework Programme (FP) 7. We are grateful to Thiago Rangel and two referees for constructive critiques of the manuscript.

References

Barbet-Massin, M. et al. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? – Methods Ecol. Evol. 3: 327–338.

Bierman, S. M. et al. 2010. Bayesian image restoration models for combining expert knowledge on recording activity with species distribution data. – Ecography 33: 451–460.

Bivand, R. and Gebhardt, A. 2000 Implementing functions for spatial statistical analysis using the R language. – J. Geogr. Syst. 2: 307–317.

Bundesanstalt für Geowissenschaften und Rohstoffe 1993. Geologische Karte der Bundesrepublik Deutschland 1:1 000 000. – Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover.

Carl, G. and Kühn, I. 2007. Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. – Ecol. Model. 207: 159–170.

Carl, G. and Kühn, I. 2008. Analyzing spatial ecological data using linear regression and wavelet analysis. – Stoch. Environ. Res. Risk Assess. 22: 315–324.

Carl, G. and Kühn, I. 2010. A wavelet-based extension of generalized linear models to remove the effect of spatial autocorrelation. – Geogr. Anal. 42: 323–337.

Corine Land Cover 1990. raster data. – <www.eea.europa.eu/data-and-maps/data/corine-land-cover-1990-raster>.

Coudun, C. and Gégout, J.-C. 2006. The derivation of species response curves with Gaussian logistic regression is sensitive to sampling intensity and curve characteristics. – Ecol. Model. 199: 164–175.

Dormann, C. F. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. – Ecography 30: 609–628.

Fielding, A. H. 2002. What are the appropriate characteristics of an accuracy measure? – In: Scott, J. M. et al. (eds), Predicting species occurrences. Issues of accuracy and scale. Island Press, pp. 271–280.

Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.

Fleiss, J. L. 1981. Statistical methods for rates and proportions. – Wiley.

Fleiss, J. L. and Cohen, J. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. – Educ. Psychol. Measur. 33: 613–619.

Franklin, J. 2009. Mapping species distributions. Spatial inference and prediction. – Cambridge Univ. Press.

Hagen-Zanker, A. 2009. An improved Fuzzy Kappa statistic that accounts for spatial autocorrelation. – Int. J. Geogr. Inform. Sci. 23: 61–73.

Hanberry, B. B. and He, H. S. 2013. Prevalence, statistical thresholds, and accuracy assessment for species distribution models. – Web Ecol. 13: 13–19.

Hanley, J. A. and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristics curve. – Radiology 143: 29–36.

Kanga, C. et al. 2013. Kappa statistic for the clustered dichotomous responses from physicians and patients. – Stat. Med. 32: 3700–3719.

Lichstein, J. W. et al. 2002. Spatial autocorrelation and autoregressive models in ecology. – Ecol. Monogr. 72: 445–463.

Liu, C. et al. 2005. Selecting thresholds of occurrence in the prediction of species distributions. – Ecography 28: 385–393.

Manceur, A. M. and Kühn, I. 2014. Inferring model-based probability of occurrence from preferentially sampled data with uncertain absences using expert knowledge. – Methods Ecol. Evol. 5: 739–750.

Nathan, R. and Muller-Landau, H. C. 2000. Spatial patterns of seed dispersal, their determinants and consequences for recruitment. – Trends Ecol. Evol. 15: 278–285.

Nathan, R. et al. 2002. Mechanisms of long-distance dispersal of seeds by wind. – Nature 418: 409–413.

Rowlingson, B. and Diggle, P. 1993. Splancs: spatial point pattern analysis code in S-Plus. – Comput. Geosci. 19: 627–655.

Sachs, L. and Hedderich, J. 2006. Angewandte Statistik, Methodensammlung mit R. – Springer.

Shekhar, S. et al. 2002. Spatial contextual classification and prediction models for mining geospatial data. – IEEE Trans. Multimedia 4: 174–188.

Svenning, J.-C. et al. 2006. Range filling in European trees. – J. Biogeogr. 33: 2018–2021.