00000100000
00001010000
00010001000
00100000100
01000000010
10000000001
Software notes

Ecography 38: 541–545, 2015
doi: 10.1111/ecog.01132

Subject Editor: Thiago Rangel. Editor-in-Chief: Miguel Araújo. Accepted 18 November 2014

spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models

Matthew E. Aiello-Lammens, Robert A. Boria, Aleksandar Radosavljevic, Bruno Vilela and Robert P. Anderson

M. E. Aiello-Lammens (matt.lammens@gmail.com), Dept of Ecology and Evolutionary Biology, Univ. of Connecticut, Storrs, CT 06269, USA, and Dept of Ecology and Evolution, Stony Brook Univ., Stony Brook, NY 11794, USA. – R. A. Boria, A. Radosavljevic and R. P. Anderson, Dept of Biology, City College of the City Univ. of New York, New York, NY 10031, USA. AR present address: Plant Biology and Conservation, Northwestern Univ., Evanston, IL 60208, USA, and Dept of Plant Science, Chicago Botanic Garden, Glencoe, IL 60022, USA, and Dept of Botany, National Museum of Natural History, Smithsonian Inst., Washington, DC 20560, USA. RPA also at: Graduate Center of the City Univ. of New York, New York, NY 10016, USA, and Division of Vertebrate Zoology (Mammalogy), American Museum of Natural History, New York, NY 10024, USA. – B. Vilela, Depto de Ecologia, Inst. de Ciências Biológicas, Univ. Federal de Goiás, Goiânia, Goiás, Brazil, and Depto de Ciencias de la Vida, Univ. de Alcalá, ES-28805 Alcalá de Henares, Madrid, Spain.

Spatial thinning of species occurrence records can help address problems associated with spatial sampling biases. Ideally, thinning removes the fewest records necessary to substantially reduce the effects of sampling bias, while simultaneously retaining the greatest amount of useful information. Spatial thinning can be done manually; however, this is prohibitively time consuming for large datasets. Using a randomization approach, the 'thin' function in the spThin R package returns a dataset with the maximum number of records for a given thinning distance, when run for sufficient iterations. We here provide a worked example for the Caribbean spiny pocket mouse, where the results obtained match those of manual thinning.

Correlative techniques for modeling species niches and their associated geographic distributions (often termed ecological niche modeling, ENM; or species distribution modeling, SDM) are an important component of many biogeographic, evolutionary, and conservation-related studies (Elith and Leathwick 2009, Peterson et al. 2011, Araújo and Peterson 2012, Warren 2012). However, addressing the effects of sampling bias remains an important outstanding issue. For many datasets of occurrence records (especially from museums and herbaria), geographic sampling bias is pervasive (Hijmans et al. 2000, Reddy and Dávalos 2003, Graham et al. 2004, Kadmon et al. 2004, Hijmans 2012). Such biases can lead to environmental bias as well, resulting in an over-representation of environmental conditions associated with regions of higher sampling (Williams et al. 2002, Kadmon et al. 2004, Anderson and Gonzalez 2011). ENMs constructed with such data may fit the environmental signal of the bias, in addition to that of the niche, hindering model interpretation and application (Araújo and Guisan 2006, Wintle and Bardos 2006). Furthermore, environmental biases lead to inflated estimates of model performance (Veloz 2009, Hijmans 2012).

Several approaches can ameliorate the effects of sampling bias. Ideally, sampling effort across geography is quantified either directly or via indices derived from the results of sampling (i.e. via records of a target group; Anderson 2003), and integrated into model calibration to correct for associated biases in environmental space (Phillips et al. 2009). However, such information is frequently unavailable, leaving researchers with a quandary: how to reduce the effects of biased sampling without reducing the signal of the species' niche (Anderson 2012). Viable solutions in such cases include thinning (also known as 'filtering') occurrence records either in environmental space or geographic space. Thinning in environmental space directly addresses the problem that proximally affects model calibration (de Oliveira et al. 2014, Varela et al. 2014). In contrast, thinning in geographic space, or spatial thinning, acts in the dimensions in which the original bias occurred – e.g. the collection of occurrence records (Reddy and Dávalos 2003, Kadmon et al. 2004, Anderson 2012).

Here we consider spatial thinning (i.e. in geographical space), which has been applied frequently and can result in species occurrence data that yield better performing ENMs (Pearson et al. 2007, Veloz 2009, Kramer-Schadt et al. 2013, Syfert et al. 2013, Verbruggen et al. 2013, Boria et al. 2014, Fourcade et al. 2014). Current spatial thinning methods generally fall into one of two categories, either employing stratified random sampling or thinning based on nearest neighbor distance. One method in the first category entails

overlaying a grid on the study region and randomly sampling a set number of occurrence records (e.g. one) from each grid cell (Hijmans and Elith 2011), where grid cells should have equal area. Other methods involve stratifying based on the density of occurrence records, randomly selecting records for inclusion based on the density of sampling in geographic space (Verbruggen et al. 2013).

The second category involves removing occurrence records so that no two are closer than a linear distance *x* (Pearson et al. 2007), resulting in a minimum nearest neighbor distance (NND) greater than or equal to *x*. To retain the greatest amount of useful information (i.e. niche signal), records should be thinned such that the largest possible number of records is retained (Anderson and Raza 2010, Radosavljevic and Anderson 2014). This method presents several challenges. Like all thinning approaches (both geographic and environmental), the optimal degree of thinning remains subject to empirical determination. Specifically, the optimal NND (*x*) likely varies across species and study regions. Some methods for estimating this distance have been developed (Veloz 2009), but further research is needed.

Additionally, this category presents serious computational challenges. Determining the optimal (i.e. maximum) number of occurrence records that meet the NND constraint can be viewed as the classic set-packing problem in computational complexity theory, which is considered nondeterministic polynomial-time (NP) hard (Johnson 1982). While solutions to such problems can be checked quickly, it remains unclear whether a solution can be found quickly (Garey and Johnson 1979). Furthermore, it is possible but seldom feasible to manually thin records. Such thinning requires human inspection of a network of distances for each cluster of records violating the NND constraint. This is time consuming and prohibitive for species with more than a small number of records (Shcheglovitova and Anderson 2013). To address these shortcomings, we developed an automated randomization approach implemented as an R package that should facilitate: 1) spatial thinning for a user-specified NND; and 2) empirical experiments that vary that distance in order to determine the best balance between bias removal and signal weakening (e.g. the distance that maximizes performance in spatially independent evaluations).

Description of the spatial thinning algorithm underlying spThin

We developed a spatial thinning method that takes a set of occurrence records and identifies multiple new subsets that meet the minimum NND constraint. From these new datasets, one (or more) retaining the largest number of records can be selected and used to construct an ENM. At the core of this method is an algorithm implemented in the R programming environment (R Core Team) that randomly removes records violating the minimum NND constraint.

Algorithm steps (for a single repetition of the function 'thin'): 1) a thinning distance (i.e. minimum NND) *x* is specified by the user. 2) Pair-wise distances between all records are calculated. 3) For each record, the number of occurrence records within distance *x* is identified. 4) The record(s) with the greatest number of neighboring occurrences within the NND is determined. 5) One of the records identified in step 4 is removed at random. 6) Steps 3 to 5 are repeated until no record in the dataset has a nearest neighbor closer than *x*.

Pair-wise distances between records are calculated using the function 'rdist.earth' in the fields package (Furrer et al. 2012), which calculates distances (in km) between geographic locations, correcting for the decreasing length of units of latitude toward the poles. R code for both our algorithm, 'thin.algorithm', and the wrapper function, 'thin', were compiled as a package named spThin. The 'thin' function provides various options described below to facilitate spatial thinning for ecological modeling. This package is provided as source code in the Supplementary material Appendix 1 and is available on Comprehensive R Archive Network (CRAN).

A single repetition of the algorithm returns one spatially thinned dataset, however for all but the smallest datasets, multiple sets of records will meet the minimum NND constraint. The user specifies the number of independent algorithm repetitions (*n*), resulting in multiple thinned datasets, which can vary in the number of records retained. The default setting of 'thin' is to save up to five datasets that yield the maximum number of records retained (compared across all repetitions). These datasets are saved as comma separated values (csv) files containing the columns: species name, latitude, and longitude. Other important arguments include options to save information for all of the datasets constructed (within an R session, not as files written to disk), to change the maximum number of csv files saved, to change the name of the log file created upon execution of 'thin', and to turn off log file creation.

Example application of spThin

As an empirical case to test the algorithm's ability to produce thinned datasets comparable to a hand-thinned dataset, we applied the spThin 'thin' function to a set of occurrence records for the Caribbean spiny pocket mouse, *Heteromys anomalus.* This dataset contains 201 verified, georeferenced occurrence records that were spatially thinned manually in a previous study, using a thinning distance of 10 km (Radosavljevic and Anderson 2014). We used this same distance in applying 'thin' to the dataset. The occurrences lie along the coastal mainland (hereafter, mainland) of northern South America (174) and on three nearby Caribbean islands: Trinidad (21), Tobago (4), and Margarita (2).

Because the algorithm includes a random element, the maximum number of records retained from any repetition of a given run may not match the optimal number of records. To investigate how many repetitions, *n*, are necessary to achieve the optimal number, we applied 'thin' to each of the four regions independently. Then, to compare with an even larger dataset (and explore the relative efficacy of splitting a complex problem into simpler independent constituent problems), we ran the algorithm on the four regions combined. Spatial thinning by hand (Radosavljevic and Anderson 2014) yielded 110 occurrence records for the mainland, and 12, 1, and 1 for Trinidad, Tobago, and Margarita, respectively (total = 124). We examined the number of repetitions required to achieve at least one thinned dataset with the optimal number of records. For the mainland, we ran 'thin' with

n equal to 10 and 100. We ran it with 10 repetitions each for Trinidad, Tobago, and Margarita. For the combined dataset, we ran 'thin' with *n* equal to 10 and 100.

The 'thin' function returned datasets with occurrences that clearly mitigated the effects of clustered sampling (Fig. 1). This was particularly noticeable in areas that we expect to present biased sampling, such as in reserves and near roads and research centers, illustrating issues characteristic of the kinds of sampling that lead to biases in biodiversity datasets (Fig. 1B). In this illustrated region of north-central Venezuela, the 11 easternmost records follow the path of a road (from El Limón to Ocumare de la Costa) that traverses the Parque Nacional Henri Pittier; this flagship reserve lies near major research centers (Museo de la Estación Biológica de Rancho Grande; and Museo del Inst. de Zoología Agrícola, Univ. Central de Venezuela).

The overall performance of 'thin' depended on the total number of records in the dataset (Table 1). Datasets with larger numbers of records should require a greater number of repetitions to consistently achieve the optimal number of records. Similarly, the number of thinned datasets containing the maximum number of retained records also depended on the total number of unthinned records and the number of repetitions. Generally, more repetitions resulted in a greater number of such datasets. Applied to the mainland occurrence dataset, 'thin' produced thinned datasets with the optimal number of occurrence records (110) when *n* was set to as few as 10 repetitions (Table 1A). Similarly, it produced thinned datasets with the optimal number of occurrences on the islands of Trinidad (12), Tobago (1), and Margarita (1) with *n* set to 10 repetitions, requiring 0.06, 0.03, and 0.01 s of computation time, respectively. Running 'thin' on the combined dataset (i.e. mainland and islands) with *n* equal to 10 and 100 also produced thinned datasets with the optimal maximum occurrence records, 124 (Table 1B).

On a standard desktop computer (specifications in Table 1 legend), run time depended on the size of the dataset and the number of repetitions, but was trivial in all cases. The optimal number of occurrence records for the mainland dataset was achieved with both 10 and 100 repetitions, requiring less than 10 s of computation time. Running 'thin' on a dataset of the four regions combined, the optimal number of records also was achieved using 10 and 100 repetitions, requiring approximately 2 and 13 s of computation time, respectively (Table 1B). Combining the computation time required to thin the mainland dataset using 100 repetitions with that needed for the three individual island datasets using 10 repetitions each yielded a total of 9.04 s, an improvement over the 12.73 s required to thin the combined dataset. These results demonstrate the utility of separating occurrence records into regional clusters, treated as independent datasets (i.e. where no records from one cluster lie within the NND of any records in any other cluster). However, given the relatively modest performance increase, this may only be important when working with very large datasets and carrying out many repetitions (e.g. tens of thousands) of the

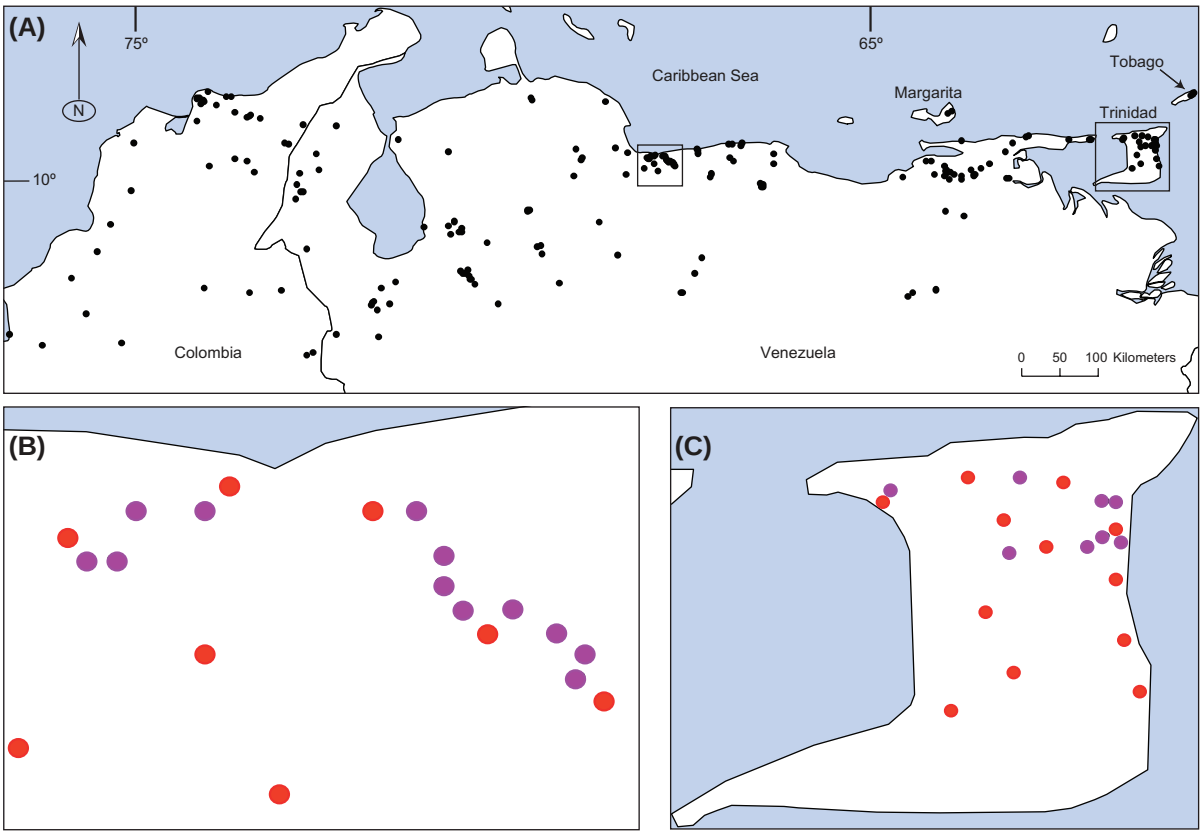

Figure 1. Occurrence records of unthinned and spThin-derived datasets for *Heteromys anomalus* in coastal northern South America and nearby islands. (A) (full figure) full spatial extent of unthinned records (black circles). (B) (left figure) and (C) (right figure) sub-regions (indicated with black outlines in (A)) showing records removed by 'thin' (purple circles) and those retained by the function (red circles).

Table 1. Summary of spThin results executed with different numbers of repetitions (*n*) for *Heteromys anomalus* in (A) the coastal mainland of northern South America and (B) the combined dataset holding records from both the coastal mainland and Caribbean islands. Shown are the maximum number of records retained, number of datasets with the maximum number of records retained, and the user run-time. User run-time based on execution of 'thin' on a Dell Optiplex 980 with Intel (c) i5 CPU (3.2 GHz) and 8 GB RAM.

(A) Coastal mainland		
spThin – repetitions (*n*)	**10**	**100**
Maximum number of occurrence records retained	110	110
Number of thinned datasets with maximum number of occurrence records	9	59
spThin run time	0.86 s	8.94 s
(B) Combined coastal mainland and islands		
spThin – repetitions (*n*)	**10**	**100**
Maximum number of occurrence records retained	124	124
Number of thinned datasets with maximum number of occurrence records	3	45
spThin run time	1.36 s	12.73 s

algorithm. Future work should focus on automated division of occurrence records into independent clusters prior to spatial thinning, which would increase the function's efficiency.

This spatial thinning method returns datasets containing the optimal number of occurrence records if run for sufficient repetitions. In our example, both 10 and 100 repetitions resulted in datasets with the optimal number (110) of occurrence records for the mainland dataset. These datasets differed in the particular records that were retained, which we expect given the random elements of the algorithm. However, a comparison of ENMs constructed using these differing datasets demonstrated that they yield similar model results (Supplementary material Appendix 1).

Practicalities and future directions

Here, we had the advantage of knowing the optimal number of occurrence records; however, this is likely not the case in most applications. To help the user determine a sufficient number of repetitions for 'thin', we recommend a visual inspection of plots of the number of maximum records found versus number of repetitions on both arithmetic and logarithmic scales, which are returned by the 'plotThin' function (Supplementary material Appendix 1, Fig. A2). This latter plot should increase linearly if a sufficient number of repetitions has not been reached (similar to the species–area relationship, species-accumulation curves, or other phenomena that follow power laws), but then have an extended plateau for higher number of repetitions after a sufficient number have been carried out.

Determining an appropriate NND is another challenge when spatially (or environmentally) thinning an occurrence dataset. In our example, we chose a value that we estimated to be reasonable based on our knowledge of the species' biology, understanding of the environmental heterogeneity of the region, and previous research on this system, including general knowledge of patterns of sampling in the region (Anderson and Raza 2010, Radosavljevic and Anderson 2014). However, such information will not always be available, and operational procedures hold important benefits. Veloz (2009) provided one method for determining an appropriate spatial thinning distance – i.e. examining semivariograms to determine the distance at which occurrence records are spatially independent. An alternative approach proposed recently is to determine the number of occurrences representing spatially independent information in a given dataset (de Oliveira et al. 2014). This is done by first fitting a simultaneous autoregressive model to the occurrence data. The autoregressive coefficient can then be used to calculate the effective number of degrees of freedom in the dataset, which is treated as the number of occurrences representing spatially independent information. In that paper, de Oliveira and colleagues (2014) calculated the distances between occurrences (separately in both geographic and environmental space), and then selected the most-distant points recursively until their dataset reached this value. We propose that this value can be used in conjunction with spThin, using the 'thin' function with multiple NND values to find the distance associated with that number of occurrences.

Even with the considerations outlined above, spatial thinning of occurrence records provides an easy-to-implement and relatively straightforward method to alleviate the effects of sampling bias (Kramer-Schadt et al. 2013, Boria et al. 2014, Radosavljevic and Anderson 2014). However, recent work has also demonstrated the utility of thinning occurrence records in environmental space, showing that under some circumstances it results in more accurate models than those produced with spatially thinned data (de Oliveira et al. 2014, Varela et al. 2014). Determining the generality of these findings and establishing best practices for minimizing the effects of sampling bias require further research. The spThin package will facilitate part of this investigation. Additionally, the algorithm underlying the 'thin' function could be applied toward environmental filtering in the future.

In sum, the spThin package provides an easy-to-implement spatial thinning method, which can be used to process occurrence records for use in constructing and evaluating ENMs, as well as in other spatial analyses. It should facilitate spatial thinning and enable research into the optimal level of thinning for various species in varying environments.

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

Aiello-Lammens, M. A., Boria, R. A., Radosavljevic, A., Vilela, B. and Anderson, R. P. 2015. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. – Ecography 38: 000–000 (ver. 0).

Acknowledgements – This research was supported by the U. S. National Science Foundation (NSF DEB-0717357 and DEB-1119915; to RPA) and the Luis Stokes Alliance for Minority Participation (Bridge to Doctorate Fellowship; to RAB). Darla M. Thomas assisted with hand thinning. Katherine St John provided critical guidance in framing the mathematical challenges encountered here.

References

Anderson, R. P. 2003. Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. – J. Biogeogr. 30: 591–605.

Anderson, R. P. 2012. Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. – Ann. N. Y. Acad. Sci. 1260: 66–80.
Anderson, R. P. and Raza, A. 2010. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. – J. Biogeogr. 37: 1378–1393.
Anderson, R. P. and Gonzalez, I. J. 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. – Ecol. Model. 222: 2796–2811.
Araújo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. – J. Biogeogr. 33: 1677–1688.
Araújo, M. B. and Peterson, A. T. 2012. Uses and misuses of bioclimatic envelope modeling. – Ecology 93: 1527–1539.
Boria, R. A. et al. 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. – Ecol. Model. 275: 73–77.
de Oliveira, G. et al. 2014. Evaluating, partitioning, and mapping the spatial autocorrelation component in ecological niche modeling: a new approach based on environmentally equidistant records. – Ecography 37: 637–647.
Elith, J. and Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. – Annu. Rev. Ecol. Evol. Syst. 40: 677–697.
Fourcade, Y. et al. 2014. Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. – PLoS One 9: e97122.
Furrer, R. et al. 2012. fields: tools for spatial data. – R package ver. 6.7, <http://CRAN.R-project.org/package=fields>.
Garey, M. R. and Johnson, D. S. 1979. Computers and intractability. – Freeman.
Graham, C. H. et al. 2004. New developments in museum-based informatics and applications in biodiversity analysis. – Trends Ecol. Evol. 19: 497–503.
Hijmans, R. J. 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. – Ecology 93: 679–688.
Hijmans, R. J. and Elith, J. 2011. Species distribution modeling with R. – <http://cran.r-project.org/web/packages/dismo/vignettes/>.
Hijmans, R. J. et al. 2000. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. – Conserv. Biol. 14: 1755–1765.
Johnson, D. S. 1982. The NP-completeness column: an ongoing guide. – J. Algorithms 3: 182–195.
Kadmon, R. et al. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. – Ecol. Appl. 14: 401–413.
Kramer-Schadt, S. et al. 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. – Divers. Distrib. 19: 1366–1379.
Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – J. Biogeogr. 34: 102–117.
Peterson, A. T. et al. 2011. Ecological niches and geographic distributions. – Princeton Univ. Press.
Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecol. Appl. 19: 181–197.
Radosavljevic, A. and Anderson, R. P. 2014. Making better Maxent models of species distributions: complexity, overfitting and evaluation. – J. Biogeogr. 41: 629–643.
Reddy, S. and Dávalos, L. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. – J. Biogeogr. 30: 1719–1727.
Shcheglovitova, M. and Anderson, R. P. 2013. Estimating optimal complexity for ecological niche models: a jackknife approach for species with small sample sizes. – Ecol. Model. 269: 9–17.
Syfert, M. M. et al. 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. – PLoS One 8: e55158.
Varela, S. et al. 2014. Environmental filters reduce the effects of sampling bias and improve predicitons of ecolgical niche models. – Ecography 37: 1–8.
Veloz, S. D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. – J. Biogeogr. 36: 2290–2299.
Verbruggen, H. et al. 2013. Improving transferability of introduced species' distribution models: new tools to forecast the spread of a highly invasive seaweed. – PLoS One 8: e68337.
Warren, D. L. 2012. In defense of "niche modeling". – Trends Ecol. Evol. 27: 497–500.
Williams, P. H. et al. 2002. Data requirements and data sources for biodiversity priority area selection. – J. Biosci. 27: 327–338.
Wintle, B. A. and Bardos, D. C. 2006. Modeling species–habitat relationships with spatially autocorrelated observation data. – Ecol. Appl. 16: 1945–1958.

Supplementary material (Appendix ECOG-01132 at <www.ecography.org/readers/appendix>). Appendix 1.