Diversity and Distributions A Journal of Conservation Biogeography

BIODIVERSITY RESEARCH

Presence-only versus presence–absence data in species composition determinant analyses

Rafi Kent* and Yohay Carmel

Faculty of Civil and Environmental Engineering, Technion, Israel Institute of Technology, Haifa, Israel

*Correspondence: Rafi Kent, Faculty of Civil and Environmental Engineering, The Technion, Haifa 32000, Israel.
E-mail: rkent@technion.ac.il.

ABSTRACT

Aim Studying relationships between species and their physical environment requires species distribution data, ideally based on presence–absence (P–A) data derived from surveys. Such data are limited in their spatial extent. Presence-only (P-O) data are considered inappropriate for such analyses. Our aim was to evaluate whether such data may be used when considering a multitude of species over a large spatial extent, in order to analyse the relationships between environmental factors and species composition.

Location The study was conducted in virtual space. However, geographic origin of the data used is the contiguous USA.

Methods We created distribution maps for 50 virtual species based on actual environmental conditions in the study. Sampling locations were based on true observations from the Global Biodiversity Information Facility. We produced P–A data by selecting ~1000 random locations and recorded the presence/absence of all species. We produced two P-O data sets. Full P-O set was produced by sampling the species in locations of true occurrences of species. Partial P-O was a subset of full P-O data set matching the size of the P–A data set. For each data set, we recorded the environmental variables at the same locations. We used CCA to evaluate the amount of variance in species composition explained by each variable. We evaluated the bias in the data set by calculating the deviation of average values of the environmental variables in sampled locations compared to the entire area.

Results P–A and P-O data sets were similar in terms of the amount of variance explained by the different environmental variables. We found sizable environmental and spatial bias in the P-O data set, compared to the entire study area.

Main conclusions Our results suggest that although P-O data from collections contain bias, the multitude of species, and thus the relatively large amount of information in the data, allow the use of P-O data for analysing environmental determinants of species composition.

Keywords
Biodiversity, canonical correspondence analysis, environmental determinants, GBIF, simulations, species composition, virtual species.

INTRODUCTION

Studying the relationships between species and their physical environment requires data on the distribution of species in space. Ideally, such analyses would be based on presence–absence data (P–A), collected through dedicated surveys. However, such data are scarce, and exist only for areas of small spatial extent, and are especially uncommon in the most diverse areas of the planet (Elith *et al.*, 2006; Ferrier & Guisan, 2006; Loiselle *et al.*, 2008; Sastre & Lobo, 2009). Presence-only (P-O) data have various shortcomings with regard to analyses of species–environment relationships, for example (1) they lack explicit information on unvisited locations and (2) they might contain errors and biases. Potential biases include spatial bias (concentration of

DOI:10.1111/j.1472-4642.2011.00755.x
http://wileyonlinelibrary.com/journal/ddi

observations in easily accessible locations and over sampling of species-rich areas, Ponder *et al.*, 2001; Kadmon *et al.*, 2004), taxonomic bias (over-representation of certain species, Hijmans *et al.*, 2000), and environmental bias (under-representation of areas at the edges of the environmental gradient, Loiselle *et al.*, 2008). In order to quantify the amount of bias in the data set, we carried out analyses to quantify the amount of environmental and geographical bias in a relatively large sample of observation locations in GBIF (GBIF data portal, http://data.gbif.org). The GBIF portal allows one to access data from multiple museum and university collections simultaneously, and to download the data in a uniform format. The data also go through a screening process to unify synonyms and omit obvious errors. In contrast to the shortcomings described earlier, data in portals like GBIF are readily available in large quantities and, because of an accelerating effort to digitize and publicize these data, are also highly accessible (Graham *et al.*, 2004). The validity of using P-O data in ecological analyses was studied several times. Results have been inconclusive, with some authors reporting sufficiency of P-O data (Elith *et al.*, 2006; Loiselle *et al.*, 2008), superiority of P–A data (Guisan & Zimmermann, 2000; Hirzel *et al.*, 2001; Graham & Hijmans, 2006), or differential success for different species (Elith *et al.*, 2006; Tsoar *et al.*, 2007). The focus of most studies was the distribution of a single species or modelling species-richness patterns. Ferrier & Guisan (2006) reviewed approaches to community-level modelling. They used both P–A and P-O data for their models, and concluded that P-O data are problematic for such analyses. They stated that owing to data limitations, analyses of species composition are limited to areas of small spatial extent. To the best of our knowledge, the value of P-O data for studying species composition determinants at large spatial scales was seldom evaluated before (but see Kadmon & Heller, 1998; Yom-Tov & Kadmon, 1998; Kadmon & Danin, 1999).

In this study, we attempt to determine whether, when considering a multitude of species over a large spatial extent, data type (P–A vs. P-O) has a significant effect on the results of analyses of species–environment relationships. A direct comparison of the effect of data type on the results of such analyses requires complete data sets of the two types, containing data on the same species and with the same spatial extent. Therefore, available data of actual observations is not optimal for such analyses, and a simulation study seems to be the most appropriate solution. We simulated P–A and P-O data sets of virtual species within the contiguous USA. We examined the effect of data type on the results of multivariate analyses of the environmental determinants of species composition. In order to keep the simulations as close to reality as possible, we used real environmental data to define species niches. We also used real locations of observations to create a P-O sampling scheme, incorporating real biases into our data sets.

METHODS

We created 50 distribution maps for virtual species within the land area of the contiguous USA. Species distributions were based on niches reflecting actual environmental conditions in the study area. The realized ecological niche of each virtual species was defined by selecting a random location within the study area to represent the niche centre in parametric space and recording the values of six environmental variables in this location: maximum temperature of the hottest month (MaxT), minimum temperature of the coldest month (MinT), annual precipitation (Prec), altitude (alt), normalized difference vegetation index (NDVI) and distance to nearest urban area (dtu). We used climatic and topographic variables from Worldclim (Hijmans *et al.*, 2005). We used an NDVI layer from MODIS (http://glcf.umiacs.umd.edu/data/ndvi) and produced a layer of distance to nearest urban area from a map of the urban areas of the USA [data was extracted from ESRI data files (ESRI, 1999)]. We tested the correlation level between each pair of variables at 200 random locations within the study area. The average correlation level was 0.36, and the maximum correlation was 0.84. All environmental layers were rescaled to a resolution of 0.0833° (~10 km). Niche breadth was set as a random fraction (between 0.05 and 0.5) of the true range of each variable in the study area, above and below the niche centre. Simulations were carried out in MatLab (MathWorks, Natick, MA, USA). Distribution maps were produced for each virtual species in ArcGIS (ESRI, 1999) by superimposing a grid with mesh size of 0.0833° (~10 km) over the entire study area. Grid cells were assigned a value of 1 where all environmental variables were within the specified realized niche, and zero otherwise (Fig. 1).

Presence–absence and presence-only data sets

To produce a P–A data set, we randomly selected 1072 locations from the geographic space of the entire study area (1072 is the median number of sampling locations found in 17

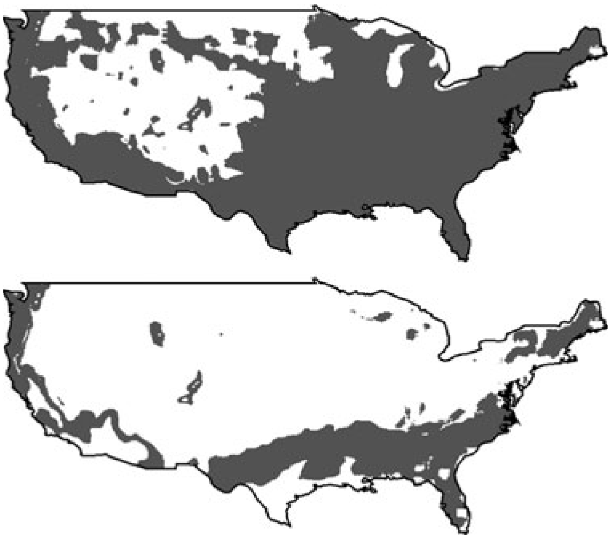

Figure 1 Distribution maps of two of the virtual species. One with a wide niche (top panel) and one with a relatively narrow niche (bottom panel).

studies that used presence–absence data). For each location, we recorded the presence/absence of all virtual species, producing a matrix of 50 columns and 1072 rows. The values of eight environmental variables [the same six variables used for defining niches plus temperature seasonality (standard deviation*100; TempS) and precipitation seasonality (coefficient of variance; PrecS)] were recorded in the same 1072 grid cells, resulting in an environmental data matrix of eight columns by 1072 rows. The two additional variables were expected to show a weaker relationship with species composition than the variables used to define the niches. As they are correlated to variables defining the species' niches, we expected to find some relation between them and the distribution of the species.

P-O data typically contain spatial bias towards easily accessible locations, as well as areas with high biodiversity (Hijmans *et al.*, 2000). In order to incorporate such bias into our data sets, we used the locations of real observations of a random selection of avian species in the contiguous USA using GBIF (GBIF, 2008). We compiled ~200,000 observation locations derived from real observations in the GBIF data set, hereafter the observation pool. The distribution range of virtual species *j* contained a subset of N_j records from the observation pool. In order to mimic taxonomic bias, as it exists in observations of real species, we selected at random 50 avian species and recorded the number of observations existing for them in GBIF. Each virtual species was randomly assigned a number of observations of one of the 50 avian species (n_j). For each species *j*, we randomly selected n_j observation locations, out of the N_j observations located within its occurrence range. We produced a matrix of 50 columns, denoting the 50 virtual species and, initially, 120,670 rows (the number of grid cells in the entire study area). Next, we deleted all empty cells from the matrix (cells with no species present). Mean size of the full P-O sets was 24,696 observations. As there was an order of magnitude difference between P–A and P-O data set sizes, we also produced partial sets of P-O data, consisting of a random choice of 1072 rows from the P-O matrix, resulting in P-O data sets of the same size as the P–A data sets. We produced sets of each data type (P–A, P-O and partial P-O) five times, independently.

Data type effect

We used canonical correspondence analysis (CCA) to examine whether data type affects the results of analyses of the relationships between species composition and environmental parameters (Ter Braak & Verdonschot, 1995) using CANOCO 4.5 (Ter Braak & Smilauer, 2002). CCA is an ordination technique that performs gradient analyses, constrained by species composition, iteratively (Ter Braak, 1986; Legendre & Legendre, 1998). Ordination is the simplification of a multidimensional space by reducing the number of axes in this space (Legendre & Legendre, 1998). The reduction is achieved by extracting the major gradients from the explanatory variables, which explain the largest amount of the variance in the independent variable distribution, and creating axes that represent these gradients. CCA assumes that the relationships between environmental variables and species composition are unimodal, rather than linear as do Principal Component Analysis and General Linear Models (Ter Braak, 1986; Legendre & Legendre, 1998). As the simulated species distributions do not have a unimodal response to the environmental variables, we expected the CCA will explain only a part of the variance in the data. We applied CCA analyses to the three data set types and compared the contribution of the various environmental variables as explanatory variables determining species composition. Each CCA analysis resulted in a λ value for each variable. $\lambda(x)$ is the proportional contribution of variable *x* to the eigenvalue of the first axis. Another element of the ordination is the relationships between the various variables, i.e. the level of correlation and the directionality of their effect on species composition (Ter Braak, 1986; Ter Braak & Verdonschot, 1995). We examined the ordination diagrams in order to qualitatively explore the relationships among the different variables and between them and the virtual species, within the ordination space.

Although CCA analyses are not normally repeated, and do not require replications, we repeated the analyses five times, to ensure the consistency of our results. Thus, we applied univariate analysis of variance in SPSS (SPSS, Chicago, IL, USA) using the different variables as covariates and the different data types as fixed factors, to determine whether the differences in the amount of variance explained by the environmental variables (λ values) obtained from P–A and P-O data, both full and partial.

Bias analysis

We performed analyses on the full P-O data set and its respective environmental data set, in order to quantify the amount of bias in the data. We calculated the environmental bias as the difference between values of each environmental variable in the entire study area (contiguous USA) derived from (1) all grid cells in the study area and (2) all the locations of the actual observations in the observation pool (~200,000). From the environmental bias, we also extracted spatial bias (difference in distance to nearest urban area, as described earlier).

RESULTS

Environmental determinants of species composition

Ordination diagrams produced for each data type revealed that environmental variables had similar effect on species composition, in all data sets regardless of data type (Fig. 2). For example, distance to nearest urban area and altitude was highly correlated in their effect on species composition. We examined the effect of data type on the amount of variance explained by each variable (λ value). We expected that the effect of distance to nearest urban area (dtu) would be less prominent when using P-O data, because the range of values of this variable was smaller in P-O data than in randomly

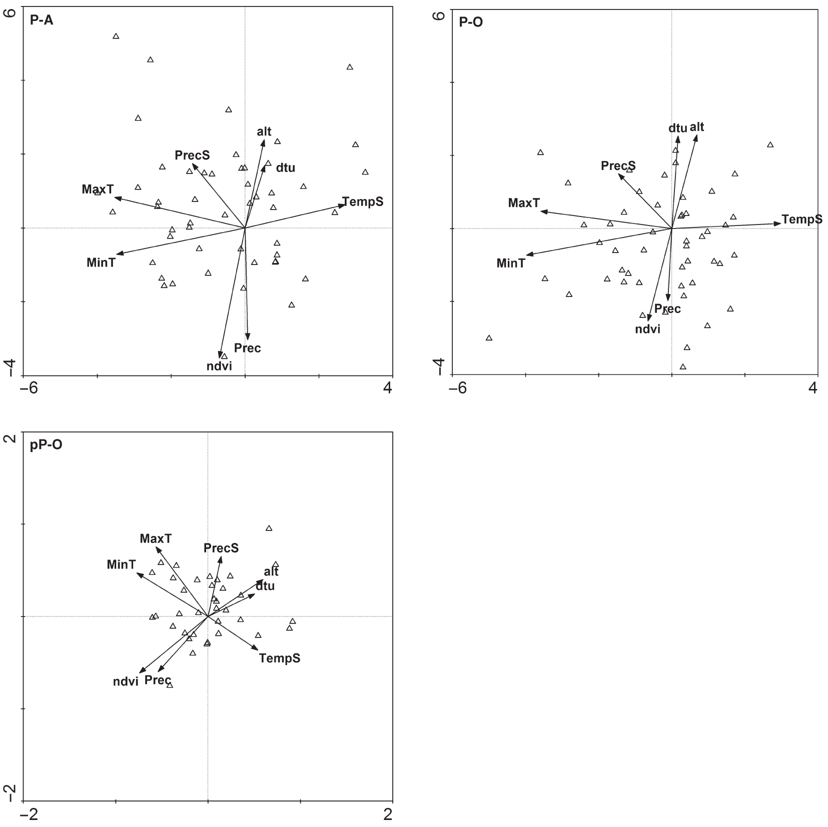

Figure 2 Three of the ordination diagrams. In the upper left corner is an ordination diagram of one canonical correspondence analysis (CCA) repetition applied to presence–absence (P–A) data for 50 virtual species in 1072 sites. In the upper right corner is an ordination of one CCA repetition of presence-only (P-O) data for 50 virtual species in ~25,000 sites. In the bottom left diagram is an ordination of partial P-O data (a subset of 1072 sites of ~25,000 P-O data). Relationships between the various variables (arrows) are similar in all diagrams, as well as the strength of their effects on species composition. Partial P-O is rotated around the origin of the axes, yet the relationships between the variables and the species, as well as among the different variables are similar to those in the P–A and P-O diagrams.

selected locations (P–A). In contrast, our results showed that there was no significant difference between the amount of variance explained by this factor in the two data types (Kruskal–Wallis ANOVA, $\chi^2 = 4.455$, $P = 0.108$). Univariate ANOVA, with environmental variables as covariates and data type as a fixed factor ($F = 2.553$, $P = 0.082$, Fig. 3) showed there was no significant effect of data type on the results of the CCA.

Bias analysis

Mean values of minimum temperature in the coldest month, distance to nearest urban area and altitude were lower in the locations of the observation pool than in the entire study area. In contrast, mean values of NDVI and annual precipitation were higher in these locations (Fig. 4). Ranges of all variables were similar between the sampled area and the entire study area.

DISCUSSION

Our results show quantitatively, for the first time, that P-O data can be used to characterize the relationships between environmental variables and species composition. We found, by using virtual species for which we have complete distributional information, that CCA is robust enough to identify the main environmental drivers of species composition despite the bias contained in such data.

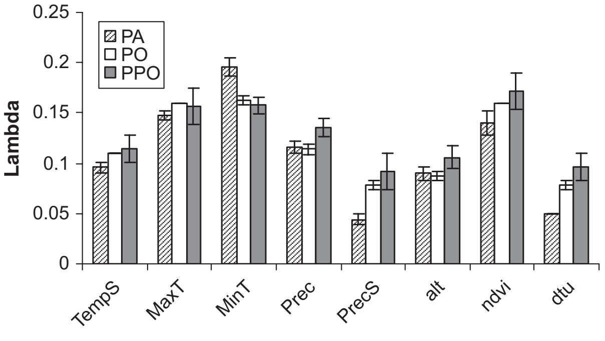

Figure 3 Canonical correspondence analysis λ values of the different environmental variables. Bars are average values over five repetitions. Error bars are standard deviations. Univariate ANOVA showed no significant effect of data type on λ values when the different variables were used as covariates and data type as a fixed factor (SPSS).

Results of the CCA analyses were highly consistent, showing similar effect of the various environmental variables on species composition, regardless of data type. This consistency implies that the method is not sensitive to data type, and that the bias in the GBIF data does not significantly affect the outcome of analyses, at least at large geographical extents.

As expected, the results only partially explained the variance in the data, because of the non-unimodal relationships of the simulated species with environmental variables. However, it has been suggested that such simple representations of species

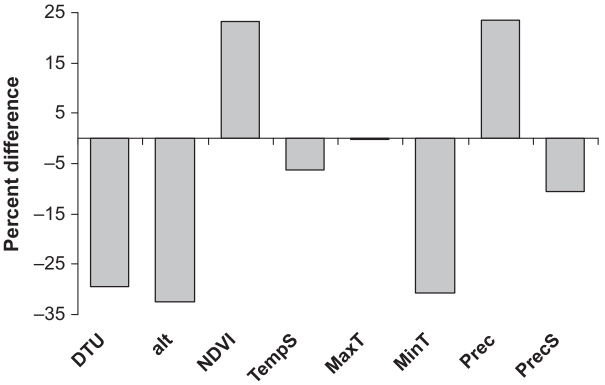

Figure 4 Difference, in percent, between the average values of the different environmental variables in the sampled area and the entire study area (contiguous USA). Positive values denote higher sample values and negative values denote higher study area values.

should be used to test the robustness of ordination techniques, such as CCA (Minchin, 1987). Our results suggest that CCA is robust enough to give consistent results despite the violation of the unimodal relationship assumption. We used partial P-O data in order to account for the difference in sample size between P–A and P-O (an order of magnitude). One might expect that the larger amount of data in the P-O set might compensate for its assumed relatively poor quality. Yet, partial P-O analyses results were very similar to those of P-O, suggesting that the amount of data had little effect on the results. On the other hand, there might be a convergence of P–A and P-O data at large sample sizes, as larger samples increase the probability of absences being true rather that false absences even in P-O data. All three data sets revealed similar relationships between environmental variables and species composition.

Given the disagreement among authors regarding the value of P-O data for species distribution modelling, and the conclusions of Ferrier & Guisan (2006) that P-O data are insufficient for community-level modelling, our results may seem surprising. One plausible explanation for our results is related to the amount of information within a data set. Analyses of species–environment relationship require dividing the studied area into grid cells. As typically most grid cells in a given study area are empty in P-O data sets (Ferrier & Guisan, 2006), the amount of information in the occupied cells is crucial for the success of the analysis. When analysing data from multiple species, each grid cell may contain data on more than one species. The cells may thus contain more information than in a single-species analysis. In addition, the number of occupied cells is dependent on the number of species, because of the larger amount of observations, as well as on the spatial distribution of the observations in the data set. Thus, using multiple species increases the amount of data available for the interpretation of the species–environment relationships. This may explain the doubts regarding P-O data for individual species modelling, as well as our success in using P-O data for community level analyses. We used 50 virtual species, which is a relatively small number of species, compared to actual species numbers found in such large areas, e.g. there are > 400 mammal species (Kays & Wilson, 2002) and > 900 bird species (http://www.birdlist.org/usa.htm) in the contiguous USA. Thus, analyses based on real species may be even more robust and reflect the true species–environment relationships. The effect of multiple species on the consistency of the results of the analyses is apparent in our results. There are differences in the locations of specific species in relation to the different environmental variables in the ordination space of the various data types, suggesting that single-species analyses may be more sensitive than multi-species analyses to data type.

Kadmon *et al.* (2004) incorporated roadside bias correction when modelling species distribution with bioclimatic models. They concluded that such corrections should be incorporated only posteriori to an examination of the amount of environmental variability between near-road locations and off-road locations. They also suggested that in an area of small climatic variance between the road network and the entire area, roadside survey data are appropriate without correction. In a simulation study, Reese *et al.* (2005) found that using data that contain roadside bias may produce model results that do not differ much from models based on systematic surveys.

Our analysis revealed that the observations in the GBIF database indeed included environmental and geographical biases. Observations were biased towards areas of high primary productivity, higher annual precipitation, higher minimum temperatures and lower altitudes. Seasonality had a small effect on observation frequency, probably due to the relatively low temporal resolution of the data. All the differences indicate that observers tended to look for species in productive areas, where conditions are relatively convenient, and avoid extreme environments. We represented geographical bias by the average distance from the nearest urban area, under the assumption that observations will be concentrated closer to urban areas than would be expected by chance. We found that the average distance of observations to an urban area was indeed ~30% smaller than the average distance in the study area, suggesting a strong bias towards sampling 'close to home'. In spite of these biases, the results of our analyses were robust and consistent. Our findings thus confirm that easily accessible, web-based data are indeed amenable for the study of large-scale species composition determinants.

ACKNOWLEDGEMENTS

This project was funded by the Israel Science Foundation (grant number 486-2010).

REFERENCES

Elith, J., Graham, C.H., Anderson, R.P. *et al.* (2006) Novel methods improve prediction of species' distributions from occurence data. *Ecography*, **29**, 129–151.

ESRI (1999) *ArcView GIS*. ESRI, Redlands, CA, USA.

Ferrier, S. & Guisan, A. (2006) Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, **43**, 393–404.

GBIF (2008) *GBIF training manual 1: digitisation of history collections data, version 1.0*. Global biodiversity information facility, Copenhagen.

Graham, C.H. & Hijmans, R.J. (2006) A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography*, **15**, 578–587.

Graham, C., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, **19**, 497–503.

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Hijmans, R.J., Garrett, K.A., Huaman, Z., Zhang, D.P., Schreuder, M. & Bonierbale, M. (2000) Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. *Conservation Biology*, **14**, 1755–1765.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolates climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hirzel, A.H., Helfer, V. & Mertal, F. (2001) Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, **145**, 111–121.

Kadmon, R. & Danin, A. (1999) Distribution of plant species in Israel in relation to spatial variation in rainfall. *Journal of Vegetation Science*, **10**, 421–432.

Kadmon, R. & Heller, J. (1998) Modelling faunal responses to climatic gradients with GIS: land snails as a case study. *Journal of Biogeography*, **25**, 527–539.

Kadmon, R., Farber, O. & Danin, A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, **14**, 401–413.

Kays, R. & Wilson, D. (2002) *Mammals of North America*. Princeton University Press, Princeton, NJ, USA.

Legendre, P. & Legendre, L. (1998) *Numerical ecology*, 2nd edn. Elsevier Science, Amsterdam.

Loiselle, B.A., Jorgensen, P.M., Consiglio, T., Jimenez, I., Blake, J.G., Lohmann, L.G. & Montiel, O.M. (2008) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography*, **35**, 105–116.

Minchin, P. (1987) Simulation of multidimensional community patterns: towards a comprehensive model. *Plant Ecology*, **71**, 145–156.

Ponder, W.F., Carter, G.A., Flemons, P. & Chapman, R.R. (2001) Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology*, **15**, 648–657.

Reese, G.C., Wilson, K.R., Hoeting, J. & Flather, C.H. (2005) Factors affecting species distribution predictions: a simulation modeling experimrnt. *Ecological Applications*, **15**, 554–564.

Sastre, P. & Lobo, J.M. (2009) Taxonomic survey biases and the unveiling of biodiversity patterns. *Biological Conservation*, **142**, 462–467.

Ter Braak, C.J.F. (1986) Canonical corespondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167–1179.

Ter Braak, C.J.F. & Smilauer, P. (2002) *CANOCO reference manual and CanoDraw for Windows user's guide: software for canonical community ordination (version 4.5)*. p. 500. Microcomputer power, Ithaca, New-York.

Ter Braak, C.J.F. & Verdonschot, P.F.M. (1995) Cannonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences*, **57**, 255–289.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D. & Kadmon, R. (2007) A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, **13**, 397–405.

Yom-Tov, Y. & Kadmon, R. (1998) Analysis of the distribution of insectivorous bats in Israel. *Diversity and Distributions*, **4**, 63–70.

BIOSKETCHES

Rafi Kent recently graduated his PhD studies in the Department of Civil and Environmental Engineering, in the Technion – Israel Institute of Technology. His research interests are patterns of biodiversity distribution from a theoretical, empirical and conservational point of view.

Yohay Carmel is an associate professor in the faculty of Civil and Environmental Engineering at the Technion – Israel Institute of Technology. His research interests cover spatial aspects of ecological and environmental phenomena, including biodiversity, vegetation dynamics, and air pollution.

Editor: Janet Franklin