Acta Soc. Zool. Bohem. **82**: 89–96, 2018
ISSN 1211-376X

Open access data and species distribution modelling: A case study of *Lithobius erythrocephalus* (Chilopoda: Lithobiidae)

Žan Kuralt & Ivan Kos

Department of Biology, Biotechnical Faculty, University of Ljubljana, Jamnikarjeva 101, 1000 Ljubljana, Slovenia

Received 24 August 2018; *accepted 26 October 2018*
Published 27 December 2018

Abstract. Open access data has made it easy to model the suitability of habitats and potential distributions of different species of plants and animals. However, such data need to be handled with care as it can include erroneous records (e.g. bad georeferencing, misidentification) and a considerable spatial bias. Building a model with open access data thus requires extensive data cleaning. Here, we present a modelling approach for open access data and provide an example of a habitat suitability model for a widespread European centipede, *Lithobius erythrocephalus* Koch, 1847, based exclusively on open access data.

Key words. Maxent, GBIF, open access data, WorldClim, Envirem.

INTRODUCTION

The amount of data on the distribution of species available in publicly accessible databases is steadily increasing. One of the largest databases is the Global Biodiversity Information Facility (GBIF 2018) with nearly a billion occurrence records. It offers an invaluable insight into the patterns of species distribution. But one should bear in mind that (1) this data could include erroneous records (misidentification, faulty georeferencing) and that (2) the spatial bias in the records can be substantial, as countries and institutions invest different amounts of resources into digitizing and publishing their collections online (Beck et al. 2013, Maldonado et al. 2015, Anderson et al. 2016).

Another type of data that has recently become available, consists of numerous gridded layers of interpolated environmental conditions based on temperature, precipitation and solar radiation. Some of the most frequently utilized environmental datasets are the Worldclim (Fick & Hijmans 2017) and two Worldclim derived sets, Bioclim (Fick & Hijmans 2017) and Envirem (Title & Bemmels 2016). These datasets often include layers with current environmental conditions as well as those of past (e.g. LGM) and projected future conditions.

Species distribution modelling (SDM) is a fast-growing field at the intersection of species occurrence and environmental data (Guisan & Thuiller 2005, Elith & Leathwick 2009). It takes both types of information and attempts to determine the environmental preferences of species, their so-called environmental niches or envelopes, and then locates areas with matching conditions. One of the most commonly used approaches in SDM is Maxent (Steven et al. 2017), although other algorithms are also used (e.g. DOMAIN, ENFA, GARP etc.). Species distribution modelling enables researchers to infer a species' distribution from known occurrences, thus discovering areas where a species could reside (Elith & Leathwick 2009, Elith et al. 2011). It has also been used to define possible locations of refuges for species during the Pleistocene (Waltari et al. 2007, Svenning et al. 2008, Vega et al. 2010) and even for predicting the effect of a change in climate on the

Presented at the 14th Central European Workshop on Soil Zoology, České Budějovice, Czech Republic, 16–18 April 2018

distribution of species (Jeschke & Strayer 2008, Loarie et al. 2008, Elith et al. 2010, Milanovich et al. 2010; Khanumet al. Kumar 2013, Krehenwinkel et al. 2016, Theodoridis et al. 2018). SDM has already been used to define the possible distributions of 20 species of centipedes in Norway (Georgopoulou et al. 2016).

Lithobius erythrocephalus Koch, 1847 is a centipede with a wide European distribution (Zapparoli 2003, Bonato et al. 2016) and as a consequence of its polymorphism a number of subspecies of uncertain taxonomic identity are described (Zapparoli 2003). Despite being widely distributed, there are only a few records for localities in Slovenia, which are all at high altitudes, in frost hollows and cave entrances (Ravnjak and Kos 2015). Its psychrophilic preferences indicate it is a Dinaric glacial relict. The Dinaric Arc served as a pleistocene refuge for a number of different animals (Hewitt 2000, Miracle et al. 2010) and plants (Brus 2010). While Alpine and Dinaric populations seem to be adapted to cold conditions (Stöckli 2009, Leśniewska et al. 2015) collected it in thermophilous thickets in eastern Poland.

We retrieved *L. erythrocephalus* occurrence data from GBIF and performed a systematic data cleaning and spatial thinning of the dataset. We then selected environmental layers based on this species biology and further tested them for multicollinearity using the variance inflation factor (VIF). The aforementioned Maxent algorithm was employed in building a species distribution model. We used the known occurrences of the species in Slovenia that were not used in model training, to validate the model's prediction. This prediction will enable a more efficient collection of specimens of *L. erythrocephalus*.

METHODS

We used R in all steps of the modelling procedure. The script used is also available at rpubs.com/zkuralt/litho_erythro_sdm.

Retrieving and cleaning species occurrence data

We retrieved 1163 occurrence records from GBIF using the "rgbif" package (Chamberlain 2017) and then followed the data cleaning procedure proposed by Hijmans & Elith (2017). The first step consisted of removing duplicate records and those with erroneous coordinates (when plotted on a map, these records were near the equator). We also removed records anchored to country centroids, which is probably a consequence of replacing the missing coordinates with coordinates of a country's centroid. In order to deal with the spatial bias, we applied spatial thinning to the dataset. We used "spThin" package (Aiello-Lammens et al. 2015) with thinning parameter set to 70 km. The final dataset consisted of 68 occurrence records.

We retrieved occurrences for model validation from the ChiloBio database of Slovenian centipedes (Ravnjak & Kos 2015). These occurrences were not used in the building of the model.

Obtaining and preparing the environmental data

We downloaded 37 different environmental layers (19 Worldclim bioclimatic environmental layers using "raster" R package (Hijmans 2017) and 18 ENVIREM layers directly from the ENVIREM website) with a resolution of 2.5 arc minutes (~5 km). Six environmental layers were then selected, based on the biology and physiology of *L. erythrocephalus*

Table 1. Description and variance inflation factor (VIF) of the different environmental variables

variable	variable description	VIF
bio10	mean temperature in warmest quarter	2.512623
bio17	precipitation in driest quarter	1.704614
bio3	isothermality (mean diurnal range / temperature annual range)	4.718597
PETDriestQuarter	mean monthly PET in driest quarter	3.016144
continentality	average temp. in warmest month – average temp. in coldest month	3.393825
tri	terrain roughness index	1.507418

Table 2. Settings used for building the models using the ENMeval package. Letters denote feature classes allowed (L = linear, Q = quadratic, P = product, T = threshold, H = hinge)

feature classes allowed	regularization multiplier values	data partitioning method
L	from 1 to 4 in 0.5 steps	block
LQ	from 1 to 4 in 0.5 steps	block
LT	from 1 to 4 in 0.5 steps	block
LQH	from 1 to 4 in 0.5 steps	block
LPQ	from 1 to 4 in 0.5 steps	block
LQHP	from 1 to 4 in 0.5 steps	block
LTQH	from 1 to 4 in 0.5 steps	block
LTPHQ	from 1 to 4 in 0.5 steps	block

(see Table 1). Selected rasters were aligned and cropped to the extent of the occurrence records. We used "HH" package (Heiberger 2017) to test layers for multicollinearity using the variance inflation factor (VIF) and discarded those with a VIF>10 (Table 1).

Building, evaluation and validation of Maxent model

Using the ENMeval package (Robert Muscarella et al. 2014), we built 56 different Maxent models with a range of settings (see Table 2).

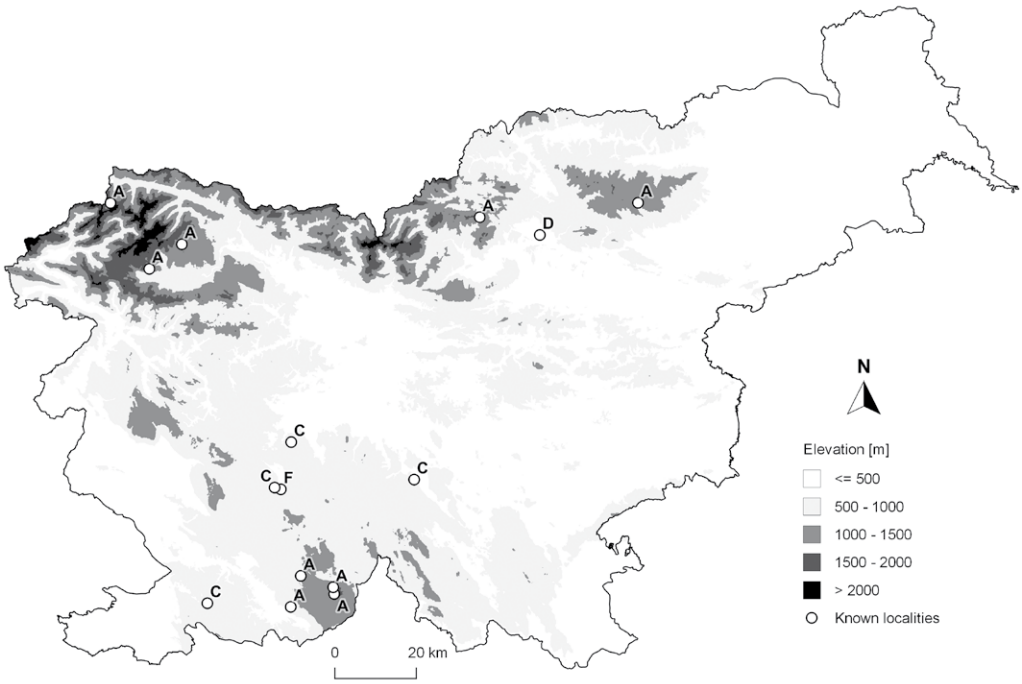

Fig. 1. Map showing the sites where *Lithobius erythrocephalus* was collected in Slovenia with altitudinal zones displayed in the background. Letters at each locality indicate the following: A – alpine, C – cave, F – frost hollow and D – construction waste dump site.

ENMeval package calculates a variety of evaluation statistics. Our optimal model was selected based on a combination of dAICc and mean AUC values. According to Muscarella et al. (2014), models with dAICc<2 generally have substantial support. We thus omitted models with a dAICc>2.

RESULTS

Occurrence data from Slovenia

Data on *L. erythrocephalus* in Slovenia is quite scarce with only 15 known localities. Specimens were collected at high altitudes (Julian Alps, Pohorje, Mt. Snežnik, Smrekovec, Pokljuka), in the frost hollow Unška Koliševka and in caves (Planinska jama, Podpeška jama, Hrencova jama).

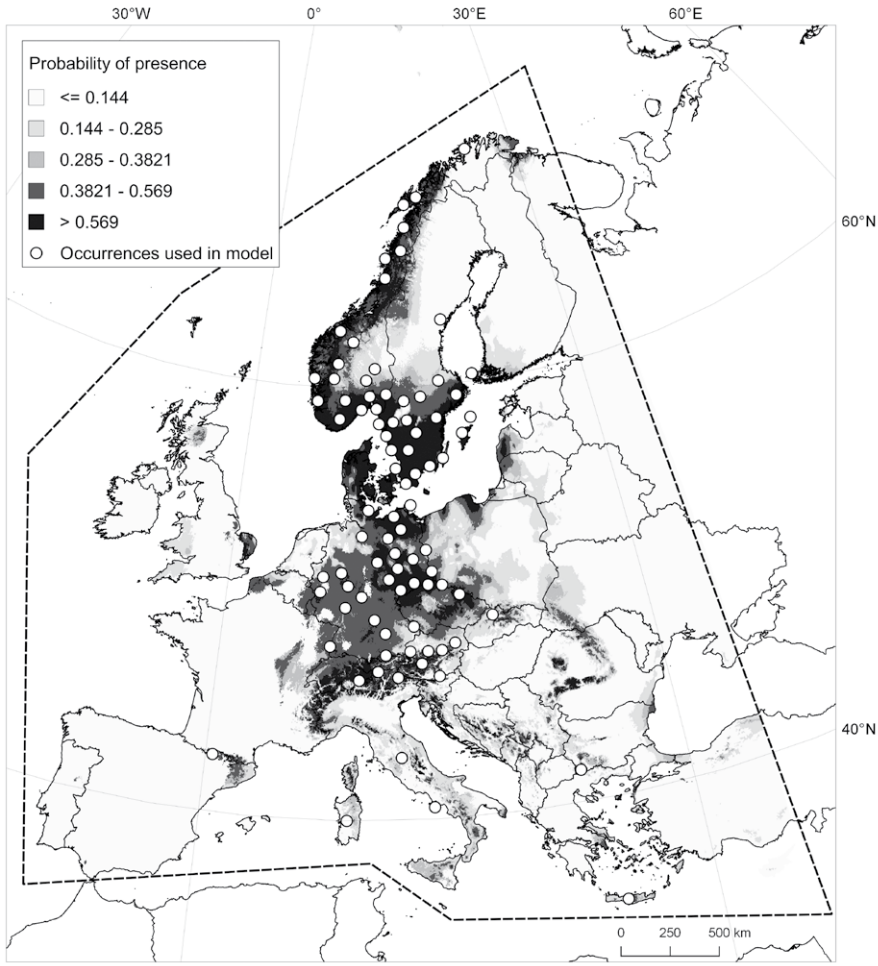

Fig. 2. The optimal model's prediction of the potential distribution of *Lithobius erythrocephalus*. Dashed line depicts the area studied.

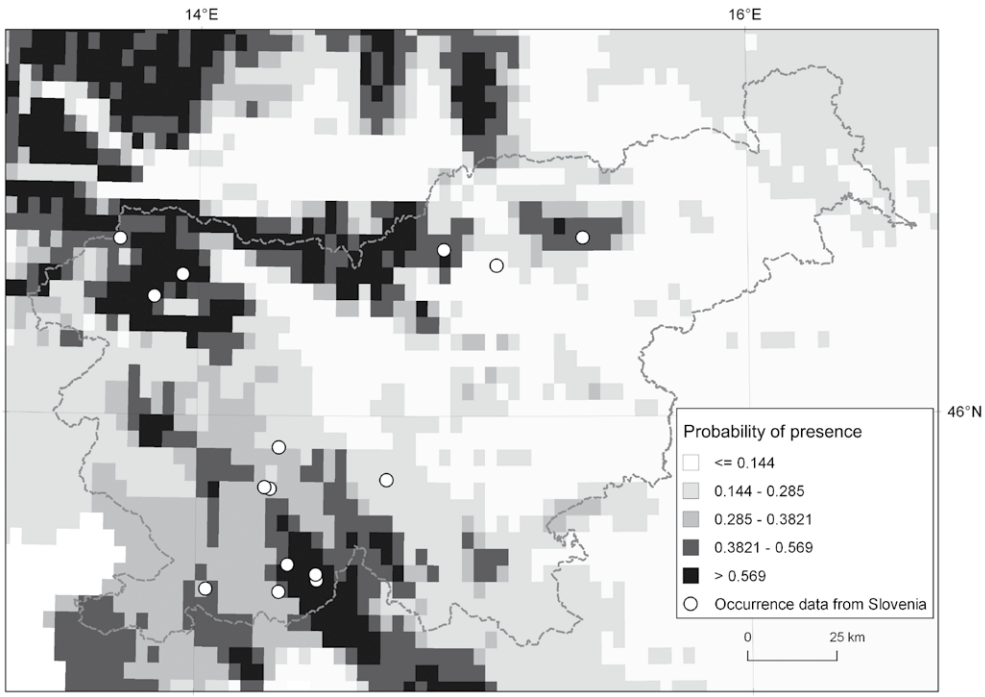

Fig. 3. A closer look at the model's prediction for Slovenia. Currently known collection sites are shown as white dots.

Lithobius erythrocephalus was also collected at a rather unusual locality (Fig. 1, D), a construction waste dumpsite near Velenje lake (370 m a. s. l.).

The optimal model

Model selection procedure described in the previous section yielded a model based on a combination of Linear, Threshold, Product, Hinge and Quadratic (LTPHQ) feature classes (FC) and a regularization multiplier (RM) of 2, as the one with the best support (see rpubs.com/zkuralt/litho_erythro_sdm for more)

DISCUSSION

Modelling approach

As demonstrated in this paper extensive systematic data cleaning is essential when working with data from public databases. The initial dataset should contain an abundance of records as many occurrences get discarded during the cleaning procedure. For example, we started with 1,163 records, whereas the final dataset consisted of 68 records. This can be a limiting factor for distribution models of species with just a few publicly available records.

In addition, as already stated by Anderson et al. (2016), there are some potential drawbacks that one cannot address when using open access data. While it is possible to account for spatial bias and faulty georeferencing, species misidentification is a major concern. Furthermore, specimens

from GBIF are (mostly) identified to species level, although several subspecies could exist. For example, there are nine accepted subspecies of *Lithobius erythrocephalus*, some of them with narrow and isolated distributions (Zapparoli 2003, Bonato et al. 2016). Model predictions based on such datasets can therefore result in wider potential distributions.

Selection of environmental layers is another crucial step when modelling species distributions. As Fourcade et al. (2018) demonstrate, the model may seem well supported, whilst its biological value is questionable, due to the poor choice of environmental variables. Selecting biologically and physiologically relevant variables is therefore of great importance. Nevertheless, there is often little knowledge of the environmental factors most profoundly affecting organisms, which makes variable selection a bit of a hit or miss. In addition, environmental layers should be selected according to the size of the area studied and the species' biology (Mackey & Lindenmayer 2001). Spatial resolution of environmental layers also affects the model's performance (Guisan et al. 2007). In some cases, rasters are unable to describe environmental variability due to coarse resolution. For example, a 5×5 km raster grid could fail to account for the microenvironment required by a species.

Not only environmental factors, but also biotic interactions, affect species distribution. This is especially the case with eurytopic organisms, such as *L. erythrocephalus*, that are able to tolerate a wide range of habitats and environmental conditions. Hence, the distributions of eurytopic species are probably highly affected by biotic interactions. Open access data on biotic interactions that one could incorporate into a model, have only recently been collected and published online by the Global Biotic Interactions project (Poelen et al. 2014). Building more comprehensive species distribution models, including both abiotic and biotic components, using solely open access data will unquestionably become feasible in the future.

Lithobius erythrocephalus collection sites throughout Europe and especially in the Balkans indicate that this species is a glacial relict (Ravnjak & Kos 2015). Although further investigations are needed to test this hypothesis, its habitat requirements indicate adaptation to cold conditions. With this assumption in mind we selected the environmental variables limiting for *L. erythrocephalus*. Variables bio10, bio17 and PETDriestQuarter thus deal with the temperature, precipitation and PET (potential evapotranspiration) in the warmest and driest periods, whereas bio3 and continentality variables deal with the stability of temperatures (see Table 1). We additionally included a variable tri (terrain roughness index), as there are probably more different microhabitats present in areas with a high terrain roughness index.

About the prediction

The model's prediction fits the training data and indicates that there are also areas of high probability of presence in the Western Alps, Dinaric Alps and Carpathian Mountains, for which there are no occurrence records (see Fig. 2). Data from these areas is either not publicly available or has not been collected. A closer look at the model's prediction for Slovenia (see Fig. 3) reveals a good fit to known collection sites. It also uncovers some areas where *Lithobius erythrocephalus* could potentially reside. There is a high probability of this species being present in mountainous regions in Slovenia (Jelovica, Trnovo Forest Plateau, Nanos Plateau, Kočevski Rog, Kamnik-Savinja Alps and Gorjanci) where it has so far not been collected. This prediction will assist in the planning of fieldwork and enable a more efficient collection of specimens in the field.

CONCLUSIONS

As presented, it is possible to create a well supported ecological niche model (ENM) using solely open access data. Provided enough data points are available and rigorous data cleaning is

employed, species distribution modelling can provide informative insights and augment current knowledge.

Acknowledgements

We thank Matej Križnar, Prishnee Bissessur and Roman Luštrik for proof reading the manuscript, the organizers of Central European Workshop on Soil Zoology for the opportunity to present our work and Stylianos Simaiakis for his constructive comments on the manuscript.

REFERENCES

Aiello-Lammens M. E., Boria R. A., Radosavljevic A., Vilela B. & Anderson R. P. 2015: spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* **38**: 541–545.

Anderson R. P., Araújo M., Guisan A., Lobo J. M., Martínez-Meyer E., Peterson A. T. & Soberón J. 2016: *Final Report of the Task Group on GBIF Data Fitness for Use in Distribution Modelling*. Geneva: Global Biodiversity Information Facility. http://www. gbif.org/resource/82612.

Beck J., Ballesteros-Mejia L., Nagel P., & Kitching I. J. 2013. Online solutions and the "Wallacean Shortfall": What does GBIF contribute to our knowledge of species ranges? *Diversity and Distributions* **19**: 1043–1050.

Bonato L., Chagas A. Jr, Edgecombe G. D. , Lewis J., Minelli A. , Pereira L. A. , Shelley R. M. , Stoev P. & Zapparoli M. 2016: *ChiloBase 2.0-A World Catalogue of Centipedes* (*Chilopoda*). Available Online at: Http://chilobase. Biologia. Unipd. It [Accessed 09/04/2018].

Brus R. 2010: Growing evidence for the existence of glacial refugia of European beech (Fagus sylvatica L.) in the south-eastern Alps and north-western Dinaric Alps. *Periodicum Biologorum* **112**: 239–246.

Chamberlain S. 2017: *Rgbif*: *Interface to the Global Biodiversity Information Facility API*. https://CRAN.R-project.org/package=rgbif.

Elith J., Kearney M. & Phillips S. 2010: The art of modelling range-shifting species. *Methods in Ecology and Evolution* **1**(4): 330–342.

Elith J. & Leathwick J. R. 2009: Species Distribution Models: Ecological explanation and prediction across space and time. *Annual Review of Ecology*, *Evolution*, *and Systematics* **40**: 677–697.

Elith J., Phillips S. J., Hastie T., Dudik M., Yung En Chee & Yates C. J. 2011: A Statistical Explanation of {MaxEnt} for Ecologists. *Diversity and Distributions* **17**: 43–57.

Fick S. E. & Hijmans R. J. 2017: WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* **37**: 4302–4315.

Fourcade Y., Besnard A. G. & Secondi J. 2018: Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography: A Journal of Macroecology* **27**: 245–256.

GBIF 2018: *GBIF Home Page. 2018*. Available from: https://www.gbif.org.

Georgopoulou E., Djursvoll P. & Simaiakis S. 2016: Predicting species richness and distribution ranges of centipedes at the northern edge of Europe. *Acta Oecologica* **74**: 1–10.

Guisan A. Graham C. H., Elith J., Huettmann F. & the NCEAS Species Distribution Modelling Group 2007: Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions* **13**: 332–340.

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

Heiberger R. M. 2017: *HH*: *Statistical analysis and data display*: *Heiberger and Holland*. https://CRAN.R-project.org/package=HH.

Hewitt G. M. 2000: The genetic legacy of the Ice Ages. *Nature* **405**: 907–913.

Hijmans R. J. 2017: *Raster*: *Geographic Data Analysis and Modeling*. https://CRAN.R-project.org/package=raster.

Hijmans R. J. & Elith J. 2017: *Species Distribution Modeling with R*. 78 pp.

Jeschke J. M. & Strayer D. L. 2008: Usefulness of bioclimatic models for studying climate change and invasive species. *Annals of the New York Academy of Sciences* **1134**: 1–24.

Khanum R., Mumtaz A. S. & Kumar S. 2013: Predicting impacts of climate change on medicinal asclepiads of Pakistan using maxent modeling. *Acta Oecologica* **49**: 23–31.

Krehenwinkel H., Rödder D., Năpăruş-Aljančič M., & Kuntner M. 2016: Rapid genetic and ecological differentiation during the northern range expansion of the venomous yellow sac spider Cheiracanthium punctorium in Europe. *Evolutionary Applications* **9**(10): 1229–1240.

Leśniewska M., Jastrzębski P., Stańska M. & Hajdamowicz I. 2015: Centipede (Chilopoda) richness and diversity in the Bug river valley (eastern Poland). *ZooKeys* **510**: 125–139.

Loarie S. R., Carter B. E., Hayhoe K., McMahon S., Moe R., Knight Ch. A. & Ackerly D. D. 2008: Climate change and the future of California's endemic flora. *Public Library of Science One* **3**(6): 1–10.
Mackey B. G. & Lindenmayer D. B. 2001: Towards a hierarchical framework for modelling the spatial distribution of animals. *Journal of Biogeography* **28**: 1147–1166.
Maldonado C., Molina C. I., Zizka A., Persson C., Taylor Ch. M., Albán J., Chilquillo E., Rønsted N. & Antonelli A. 2015: Estimating species diversity and distribution in the era of big data: To what extent can we trust public databases? *Global Ecology and Biogeography*: *A Journal of Macroecology* **24**: 973–984.
Milanovich J. R., Peterman W. E., Nibbelink P. P. & Maerz J. C. 2010. Projected loss of a salamander diversity hotspot as a consequence of projected global climate change. *Public Library of Science One* **5**(8): 1–10.
Miracle P. T., Mauch Lenardić J. & Brajković D. 2010: Last glacial climates, "refugia", and faunal change in southeastern Europe: Mammalian assemblages from Veternica, Velika Pećina, and Vindija caves (Croatia). *Quaternary International*: *The Journal of the International Union for Quaternary Research* **212**: 137–148.
Muscarella R., Galante P. J., Soley-Guardia M., Boria R. A., Kass J. M., Uriarte M. & Anderson R. P. 2014: ENMeval: An R Package for conducting spatially independent evaluations and estimating optimal model complexity for maxent ecological niche models. *Methods in Ecology and Evolution* **5**: 1198–1205.
Phillips S. J., Dudík M. & Schapire R. E. 2017: *Maxent Software for Modeling Species Niches and Distributions*. *Version 3.4.1*.
Phillips S. J., Anderson R. P. & Schapire R. E. 2006: Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190**: 231–259.
Poelen J. H., Simons J. D. & Mungall Ch. J. 2014: Global biotic interactions: An open infrastructure to share and analyze species-Interaction datasets. *Ecological Informatics* **24**: 148–159.
Ravnjak B. & Kos I. 2015: The current knowledge on centipedes (Chilopoda) in Slovenia: Faunistic and ecological records from a national database. *ZooKeys* **510**: 223–231.
Stöckli E. 2009: Literature-based survey on the Swiss fauna of Chilopoda. *Soil Organisms* **8**: 647–669.
Svenning J. C., Normand S. & Kageyama M. 2008: Glacial refugia of temperate trees in Europe: Insights from species distribution modelling. *Journal of Ecology* **96**: 1117–1127.
Theodoridis S., Patsiou T. S., Randin C & Conti E. 2018: Forecasting range shifts of a cold-adapted species under climate change: Are genomic and ecological diversity within species crucial for future resilience? *Ecography* **41**: 1357–1369.
Title P. O. & Bemmels J. B. 2016: ENVIREM: An expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Methods in Ecology and Evolution* **1**: 1–48.
Vega R., Fløjgaard C. & Searle J. B. 2010: Northern glacial refugia for the pygmy shrew Sorex minutus in Europe revealed by phylogeographic analyses and species distribution modelling. *Ecography* **33**: 260–271.
Waltari E., Hijmans R. J., Peterson A. T., Nyár A. S., Perkins S. L. & Guralnick R. P. 2007: Locating Pleistocene refugia: Comparing phylogeographic and ecological niche model predictions. *Public Library of Science One* **2**(6): 1–11.
Zapparoli M. 2003: The present knowledge on the European fauna of Lithobiomorpha (Chilopoda). *Bulletin of the British Myriapod and Isopod Group* **19**: 1882–1903.