Ecological Informatics 34 (2016) 139–145

Contents lists available at ScienceDirect

Ecological Informatics

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

Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models

Tomer Gueta *, Yohay Carmel

Department of Civil and Environmental Engineering, The Technion – Israel Institute of Technology, Haifa 32000, Israel

ARTICLE INFO

Article history:
Received 14 April 2016
Received in revised form 29 May 2016
Accepted 1 June 2016
Available online 07 June 2016

Keywords:
Biodiversity informatics
Data-cleaning
SDM performance
MaxEnt
Australian mammals
Big-data

ABSTRACT

The recent availability of species occurrence data from numerous sources, standardized and connected within a single portal, has the potential to answer fundamental ecological questions. These aggregated big biodiversity databases are prone to numerous data errors and biases. The data-user is responsible for identifying these errors and assessing if the data are suitable for a given purpose. Complex technical skills are increasingly required for handling and cleaning biodiversity data, while biodiversity scientists possessing these skills are rare. Here, we estimate the effect of user-level data cleaning on species distribution model (SDM) performance. We implement several simple and easy-to-execute data cleaning procedures, and evaluate the change in SDM performance. Additionally, we examine if a certain group of species is more sensitive to the use of erroneous or unsuitable data. The cleaning procedures used in this research improved SDM performance significantly, across all scales and for all performance measures. The largest improvement in distribution models following data cleaning was for small mammals (1 g–100 g). Data cleaning at the user level is crucial when using aggregated occurrence data, and facilitating its implementation is a key factor in order to advance data-intensive biodiversity studies. Adopting a more comprehensive approach for incorporating data cleaning as part of data analysis, will not only improve the quality of biodiversity data, but will also impose a more appropriate usage of such data.

1. Introduction

The recent availability of species occurrence data from numerous sources, standardized and connected within a single portal, has the potential to answer fundamental ecological questions (Peterson et al., 2015). This integration and analysis of massive amounts of data is timely, as researchers increasingly address questions at broader scales (Hackett et al., 2008; Peterson et al., 2015). Until recently, biodiversity data were scattered in different formats in natural history collections, survey reports, and in the literature (Guralnick and Hill, 2009; Michener and Jones, 2012). In the last fifteen years, efforts were made to establish essential standardization in the biodiversity database structure. To-date, there are several centralized portals that aggregate large volumes of biodiversity records from around the world and publish them in common formats (Wieczorek et al., 2012). Among these networks of biodiversity databases, the Global Biodiversity Information Facility (GBIF) is the largest and best known (Otegui, 2012; Yesson et al., 2007). At present, the GBIF network provides access to 653 million biodiversity records from 15,781 different data sources, including museum collections, scientific studies, citizen science, surveys, and atlas data. Since the year 2008, over 1286 peer-reviewed articles have reported using GBIF-mediated data in analyses (GBIF, 2015). The subject areas covered by these studies include climate change, human health, food security, community ecology, biogeography, evolutionary ecology and conservation biology (GBIF, 2013).

Large distributional databases as GBIF are prone to data errors, due to incomplete or erroneous information at the publisher level (e.g. the observer), errors during the publishing processes (e.g. formatting of date information), as well as errors during the central harvesting and indexing procedures (Otegui, 2012; Wieczorek et al., 2012). These problems have raised concerns that GBIF data cannot be reliably used for biodiversity research (Mesibov, 2013; Yesson et al., 2007). Data cleaning is a process used to determine inaccurate, incomplete, or unreasonable data, and improve the quality through correction of detected errors and omissions. The cleaning process may include format checks, completeness checks, reasonableness checks, limit checks, etc. (Chapman, 2005a). These processes usually result in flagging, documenting, and subsequent correcting or eliminating suspect records (Chapman, 2005a; Mathew et al., 2014). Other cleaning approaches may include the review of the data to identify geographic, temporal or environmental outliers (Bennett, 2012), and visualization of the data to unveil patterns and detect data anomalies (Chapman, 2005b; García-Roselló et al., 2013; Geng et al., 2011; Otegui and Ariño, 2012). Complex technical skills are increasingly required for handling and cleaning

* Corresponding author.
E-mail addresses: tomer.gu@gmail.com (T. Gueta), iyohay@tx.technion.ac.il (Y. Carmel).

http://dx.doi.org/10.1016/j.ecoinf.2016.06.001
1574-9541/© 2016 Elsevier B.V. All rights reserved.

biodiversity data, while biodiversity scientists possessing these skills are rare (Peterson et al., 2015).

In addition to error cleaning procedures, another set of cleaning routines could be conceived, which would select and remove data that are not erroneous, but are unsuitable for a particular application or purpose (Belbin et al., 2013; Boakes et al., 2010; Otegui et al., 2013a,b; Yesson et al., 2007). This case-specific cleaning approach could enable scientists to further improve the quality of biodiversity data with espect to the specific research. For example, data with low spatial resolution may be faulty when constructing high-resolution species distribution model (Hefley et al., 2014; Maldonado et al., 2015; Velásquez-Tibatá et al., 2015). Several studies that assess the quality of biodiversity data exist (Ballesteros-Mejia et al., 2013; García-Roselló et al., 2014; Mesibov, 2013; Otegui et al., 2013b; Vandepitte et al., 2015). Yet, studies that actually quantify the effect of data cleaning are scarce (e.g. Feeley and Silman, 2010; Maldonado et al., 2015). Although procedures for data quality assessment are clearly vital, comprehensive and practical tools facilitating it are still missing (Otegui and Guralnick, 2016). Species Distribution Modeling (SDM) is a commonly used analytical method that estimates the relationship between species records at sites, and environmental and spatial characteristics of those sites, in order to estimate the response function and contribution of environmental variables to the observed species distribution (de Souza Muñoz et al., 2011; Elith et al., 2011; Franklin, 2009). The performance of a distribution model could be a proxy for the strength of environmental factors in affecting species distribution, assuming that we select the appropriate environmental variables and use an appropriate spatial scale (Fei and Yu, 2015; Franklin, 2009; Peterson et al., 2011; Soininen and Luoto, 2014). A Maximum Entropy SDM approach (MaxEnt) developed by Phillips et al. (2006) is the most widely used SDM algorithm (Fourcade et al., 2014); due to its high performance (Elith et al., 2010), capability to deal with presence-only data (Elith et al., 2011), and low sensitivity to small sample sizes (Elith et al., 2010). The value of data-cleaning can be estimated indirectly via modeling species-environment relationship; it is expected that when erroneous or unsuitable data are removed, species affinity to environmental factors will increase, hence, the distribution model will perform better (Fei and Yu, 2015; Hefley et al., 2014; Velásquez-Tibatá et al., 2015).

The goals of this study are to estimate the effect of user-level data cleaning on SDM performance, and to exemplify the value of more intensive and case-specific data cleaning, which are rarely conducted by GBIF data users. We implement several relatively simple and easy-to-execute data cleaning procedures, and test SDM performance improvement, using GBIF occurrence data of Australian mammals, and in various different spatial scales. In addition, we examine if a certain group of species is more sensitive to erroneous or not suitable data using various species grouping.

2. Methods

2.1. Study area and taxon

The focal group in this study is Australian mammals, due to the high-resolution environmental data and relatively large number of mammalian occurrence records in this continent.

2.2. Data retrieval

Occurrence data for all Australian mammals (1,041,941 records) were downloaded in April 2014 from the Australian GBIF node (Atlas of Living Australia, see Appendix A for a list of data sources). The query used to download records was all records with class “Mammalia”. In parallel, 24 raster layers of environmental variables in Australia (elevation, land use, NDVI, and 21 climatic variables) were compiled at a spatial resolution of 1 km^2 (Table 1).

Table 1
Environmental variables used in the MaxEnt model. Elevation was derived from Diva-GIS (Hijmans et al., 2012). NDVI, solar exposure and evaporation were derived from the Australia Bureau of Meteorology (http://www.bom.gov.au). Land use was derived from the Australian Department of Agriculture and Water Resources (http://www.agriculture.gov.au). All BIO variables were taken from Worldclim (Hijmans et al., 2005).

Variable name	Description
Elevation	SRTM30 dataset. CGIAR-SRTM data aggregated to 30 s
NDVI	Six-monthly NDVI Average for Australia from Dec. 2013–May 2014
Land Use	Land Use of Australia, Version 4, 2005/2006 (September 2010 release)
Solar exposure	Annual global solar exposure over Australia for the period 1990 to 2011.
Evaporation	Average amount of water which evaporates from an open pan annually
BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp–min temp))
BIO3	Isothermality (BIO2/BIO7) (* 100)
BIO4	Temperature Seasonality (standard deviation *100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

2.3. Data cleaning

Prior to data analysis, three essential cleaning procedures were carried out (hereafter, ‘essential data cleaning’), in order to remove erroneous data: (a) Species taxonomic level cleaning: removal of records with insufficient taxon rank identification (not identified at the species level). (b) Removal of records with unrecognized species names, based on the Atlas of Living Australia species backbone. (c) Removal of records with missing or non-Australian coordinates. This data cleaning represents the typical level of cleaning conducted by researchers. In order to evaluate the specific value of user-level data cleaning, we conducted an additional, more advanced and research-specific data cleaning phase (hereafter, ‘user-level data cleaning’), which was designed to the specific question of using GBIF data for building SDMs. The user-level cleaning was aimed at removing records that are not necessarily erroneous, but are unsuitable for a specific application, which is, in our case, high resolution species distribution models. Additionally, it included fixing erroneous coordinates. Thus, these procedures included the essential data cleaning mentioned above, and the following data checks: coordinate data checks and filtering, to remove records with insufficient spatial accuracy: (a) specific data checks to salvage records with badly formatted coordinates (e.g. Degree Minute Second format, a string instead of a number), switched longitude and latitude, and numerical sign confusion. (b) Removal of coordinates located exactly at the center of Australia (may suggest incorrect georeferencing). (c) Removal of records of domesticated or extinct species due to its discrepancy with our research question. (d) Removal of records taken before the year 1990 due to high potential of insufficient spatial accuracy. (e) Removal of records with unknown year. (f) Removal of records with longitude and latitude precision with less than three decimal digits. The effectiveness of the data in building SDMs before- and after the user level data cleaning was compared.

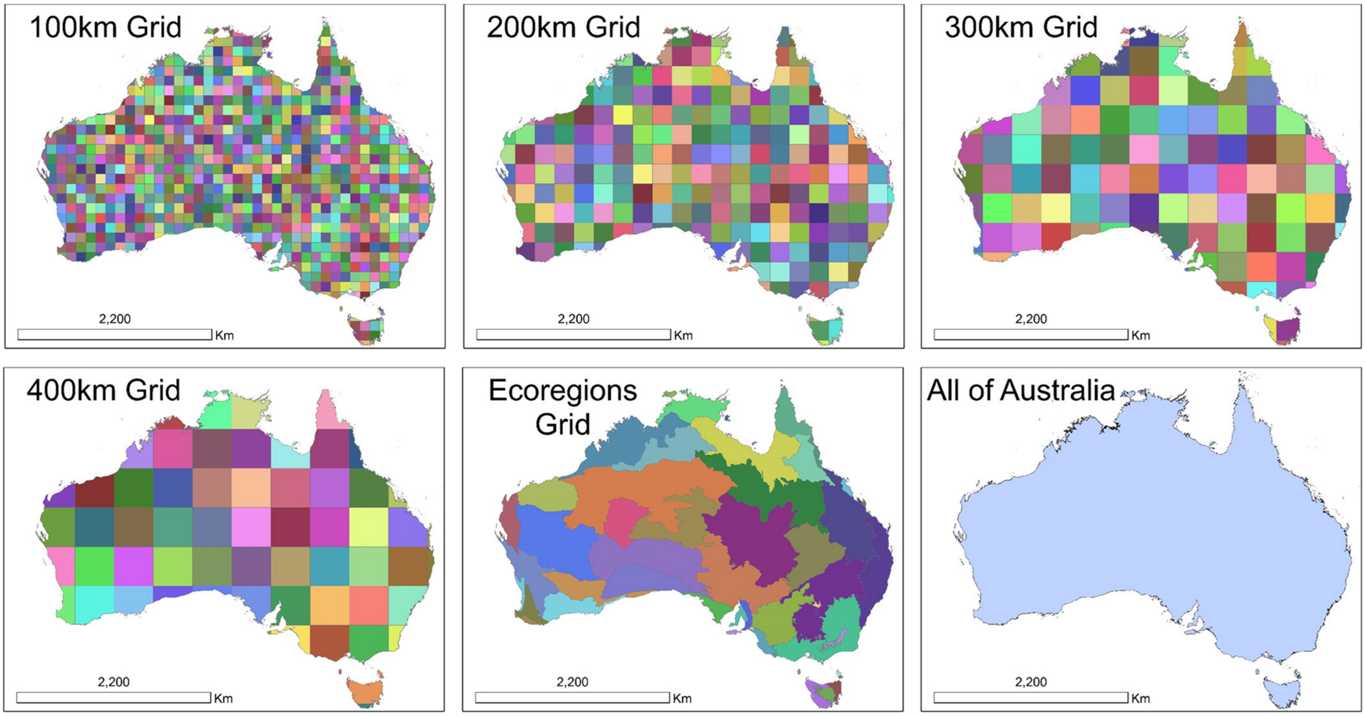

Fig. 1. Grids used in this study: (a) 100 km, (b) 200 km, (c) 300 km, (d) 400 km, (e) WWF Ecoregions (Olson et al., 2001) and (f) all of Australia.

Reconciliation to different web services, for coordinate validation, taxonomic identification, and the retrieval of biological traits for all Australian mammals was carried out using Open Refine (Verborgh and De Wilde, 2013). Australia was dissected to five different grids (100 km, 200 km, 300 km, 400 and WWF eco-regions) using QGIS (QGIS Development Team, 2013), in order to test the effect of cleaning at six different spatial scales (five grids and all of Australia, see Fig. 1).

2.4. Species Distribution Model (SDM)

A MaxEnt model (Phillips et al., 2006) was performed for each species in each grid cell. Only cells/regions above 7000 km^2 were included in the analysis. The following procedures were conducted separately for each cell/region: For each mammal species present in the cell/region, 70–100 occurrence data points were randomly sampled. A species with less than 70 occurrence data points (with unique coordinates) in a specific cell/region was omitted from further analysis in that particular cell. In addition, 1000 background points for each species were randomly sampled. For each presence/background record, the values of the 24 environmental variables were extracted and recorded. A 10-fold cross-validation procedure was used to estimate errors around predictive performance on held-out data (Elith et al., 2011). The gain and AUC values were recorded for each model. Gain is a measure for goodness of fit of models. It represents the likelihood of presence records compared to background records (Phillips, 2008). A gain of 1.1 means that an average presence location has a relative probability of e$^{1.1}$, which is three times

Table 2
Details of essential- and the user-level data cleaning. Essential data cleaning are cleaning procedures often applied by researchers. User-level data cleaning contains more advanced data checks seldom applied by researchers.

Type of cleaning	Issue (data check)	# of Records	Action
Essential data cleaning	Initial number of records downloaded from ALA[1]	1,041,941	–
	Species name was not recognized by ALA[1]	3502	Removal
	Insufficient taxon rank identification	89,775	Removal
	Records with missing or non-Australian coordinate	76,735	Removal
	Total	A dataset of 939,198 records	Removal of 102,743 records (9.86%)
Essential + User level data cleaning	Initial number of records downloaded from ALA[1]	1,041,941	–
	Wrong coordinate systems	8	Repair
	coordinates as string	270	Repair
	Switched Longitude & Latitude	0	Repair
	Numerical sign confusion	2	Repair
	Records with missing or non-Australian coordinate	76,455	Removal
	Coordinates exactly in center of Australia	30	Removal
	Longitude & Latitude precision less than 3 digits	292,541	Removal
	Records collected before the year 1990	350,403	Removal
	Records with unknown year	54,481	Removal
	Species name was not recognized by ALA	3502	Removal
	Insufficient taxon rank identification	89,775	Removal
	Domesticated species	16,190	Removal
	Extinct species	1269	Removal
	Total	A dataset of 515,479 records	Removal of 526,462 records (50.52%)

[1] Atlas of Living Australia.

Table 3
The average increase (in percentage) in performance measures after user level cleaning procedures, across six spatial scales. Asterisks indicate levels of statistical significance of a one-tailed paired t-test (***P < 0.001). All five performance measurements exhibit significant improvement across various scales.

	Regularized training gain	Unregularized training gain	Test gain	Training AUC	Test AUC
Grid	Mean (±s.d)	Mean (±s.d)	Mean (±s.d)	Mean (±s.d)	Mean (±s.d)
100 km	9.95*** (33.86)	7.67*** (22.67)	10.72*** (40.04)	0.91*** (3.00)	1.74*** (7.11)
200 km	11.21*** (25.65)	8.79*** (20.18)	16.48*** (41.11)	1.17*** (2.75)	2.26*** (5.58)
300 km	13.05*** (29.56)	10.62*** (23.17)	17.53*** (41.44)	1.27*** (2.94)	2.28*** (5.51)
400 km	15.89*** (29.43)	12.91*** (23.45)	23.71*** (41.04)	1.59*** (3.04)	3.17*** (5.62)
Ecoregions	24.43*** (45.34)	18.58*** (31.07)	27.85*** (44.62)	2.22*** (3.53)	4.41*** (7.47)
All Australia	8.28*** (13.66)	6.42*** (10.73)	11.65*** (18.50)	0.47*** (1.16)	1.14*** (2.24)

higher than an average background point. AUC is the area under the curve of the receiver operating characteristic plot (Swets, 1988). It measures the overall discriminatory ability of the model by quantifying the probability that the model correctly ranks a random presence locality higher than a random background pixel (Phillips et al., 2006). AUC ranges between 0.5 (model that performs no better than random) and 1 (model with perfect discrimination).

2.5. SDM performance: essential cleaning vs. user-level cleaning

MaxEnt model generates three gain measures and two AUC measures: *Regularized training gain* accounts for the number of predictors in the model to address overfitting; *Unregularized training gain* has no compensation for the number of predictors in the model; and *Test gain* is calculated from presence records held out to test the model. *Training AUC* calculates AUC using the training data; and *Test AUC* calculates AUC using the test data. In order to estimate the improvement of the SDM performance after the user level data cleaning, all five measures were compared for each species in each cell, after essential data cleaning, and after the user level cleaning, using one-tailed paired Z-test.

2.6. Null model (random thinning model)

A null model was constructed in order to test the effects of non-random thinning of the data resulting from data cleaning (hereafter, random thinning model). User level data cleaning often involves loss of large amount of data (such as in the present study). Such data thinning can cause ‘shrinking’ of the observed niche of a species, and lead to a stronger affinity between the inferred species distribution and environmental factors, which, in turn, may artificially increase SDM performance (Peterson et al., 2011). The user level data cleaning reduced the database by 53% to 515,479 records; for this model, the database was reduced to the same size by filtering out randomly selected records. It is expected that if thinning of the data has no effect on SDM performance measures, the results of the random thinning model would yield similar results to the original model prior user level cleaning.

2.7. Group analysis

We examined if data cleaning affects certain groups of species more than others. The grouping was based on body size and trophic level, resulting in six groups: small herbivores (1 g–100 g), medium size herbivores (100 g–5000 g), large herbivores (5000 g+), small carnivores (1 g–100 g), medium size carnivores (100 g–5000 g) and all bats. We conducted a Friedman test (Friedman, 1937) between the groups for each performance measure. If a test was found significant, a pairwise post-hoc Mann–Whitney test was performed with Benjamini & Hochberg corrections methods (Benjamini and Hochberg, 1995).

3. Results

The essential data cleaning procedures filtered out 9.9% of the downloaded data, leaving 939,198 records and 291 species. The user level cleaning procedures filtered out 50.5% of the downloaded data, leaving 515,479 records representing 242 species. Table 2 provides details on all cleaning procedures.

One-tailed paired Z-test was used to compare the results after essential data cleaning vs. after user level cleaning. All paired Z-tests showed a significant increase in performance ($\alpha \ll 0.001$) after the user level data cleaning, in all spatial scales and for all performance measures (Table 3).

When examining the effect size of the change, gain measures yielded an average increase of 7.67%–27.85% across the different grids, and the AUC measures showed an increase of 0.91%–4.4% after data cleaning (Table 3). AUC measures were apparently less sensitive to data cleaning, presumably due to the relatively high AUC values even before data cleaning (average AUC was 0.88). All five measures were highly correlated (Pearson-$r > 0.89$, p-value $\ll 0.001$ in all cases). At the scale of the entire continent, 109 (74.1%) of the mammal SDMs showed an

Table 4
Unregularized training gain analysis for different spatial grids. Gain values were compared after essential- and user-level data cleaning, in each species/grid cell combinations. # species-cell refers to the number of unique species-cell combination (sample size).

	Mean gain values		gain increase		gain decrease	
Grid	Essential cleaning	User-level cleaning	# species-cell (%)	Mean % gain increase	# species-cell (%)	Mean % gain decrease
100 km	1.11	1.23	407 (64.2%)	+17.31%	227 (35.8%)	−9.63%
200 km	0.96	1.06	373 (67%)	+17.40%	184 (33%)	−8.63%
300 km	1.26	1.35	345 (67.6%)	+19.88%	165 (32.4%)	−8.73%
400 km	1.28	1.41	325 (72.4%)	+21.17%	124 (27.6)	−8.72%
WWF Ecoregions	1.32	1.48	271 (76.3%)	+26.70%	84 (23.7%)	−8.40%
All Australia	2.10	2.22	109 (74.1%)	+10.29%	38 (25.9%)	−4.67%
Average (%)			70.3%	+18.8%	29.73%	−8.13%

Table 5
The average values of all performance measures across all spatial scales, for three model types: after essential data cleaning; after random thinning of the data; and after user level data cleaning. The random thinning models present values closer to the essential data cleaning models and not to the user level data cleaning.

		100 km	200 km	300 km	400 km	Ecoregions	All Australia
Regularized training gain	Essential cleaning	0.895	0.974	0.982	1.020	1.023	1.917
	Random thinning	0.823	0.940	0.963	1.030	1.027	1.926
	User level cleaning	0.960	1.059	1.079	1.145	1.207	2.050
Unregularized training gain	Essential cleaning	1.158	1.243	1.248	1.278	1.283	2.100
	Random thinning	1.078	1.203	1.23	1.296	1.297	2.109
	User level cleaning	1.229	1.332	1.354	1.414	1.475	2.220
Test gain	Essential cleaning	0.764	0.836	0.846	0.879	0.879	1.865
	Random thinning	0.679	0.793	0.810	0.880	0.879	1.868
	User level cleaning	0.847	0.947	0.966	1.045	1.115	2.037
Training AUC	Essential cleaning	0.873	0.886	0.888	0.889	0.893	0.951
	Random thinning	0.864	0.881	0.885	0.892	0.891	0.950
	User level cleaning	0.882	0.896	0.898	0.903	0.910	0.960
Test AUC	Essential cleaning	0.801	0.817	0.821	0.825	0.825	0.932
	Random thinning	0.758	0.812	0.815	0.826	0.825	0.930
	User level cleaning	0.813	0.835	0.839	0.850	0.859	0.942

improvement in Unregularized training gain, while 38 (25.9%) mammal SDMs revealed a decrease in performance, following data-cleaning. For the ecoregion-specific SDMs, 271 (76.3%) of the SDMs increased performance (mean gain increase of 26.70%), and 84 (23.7%) decreased their performance. The results for the other spatial scales were similar (Table 4).

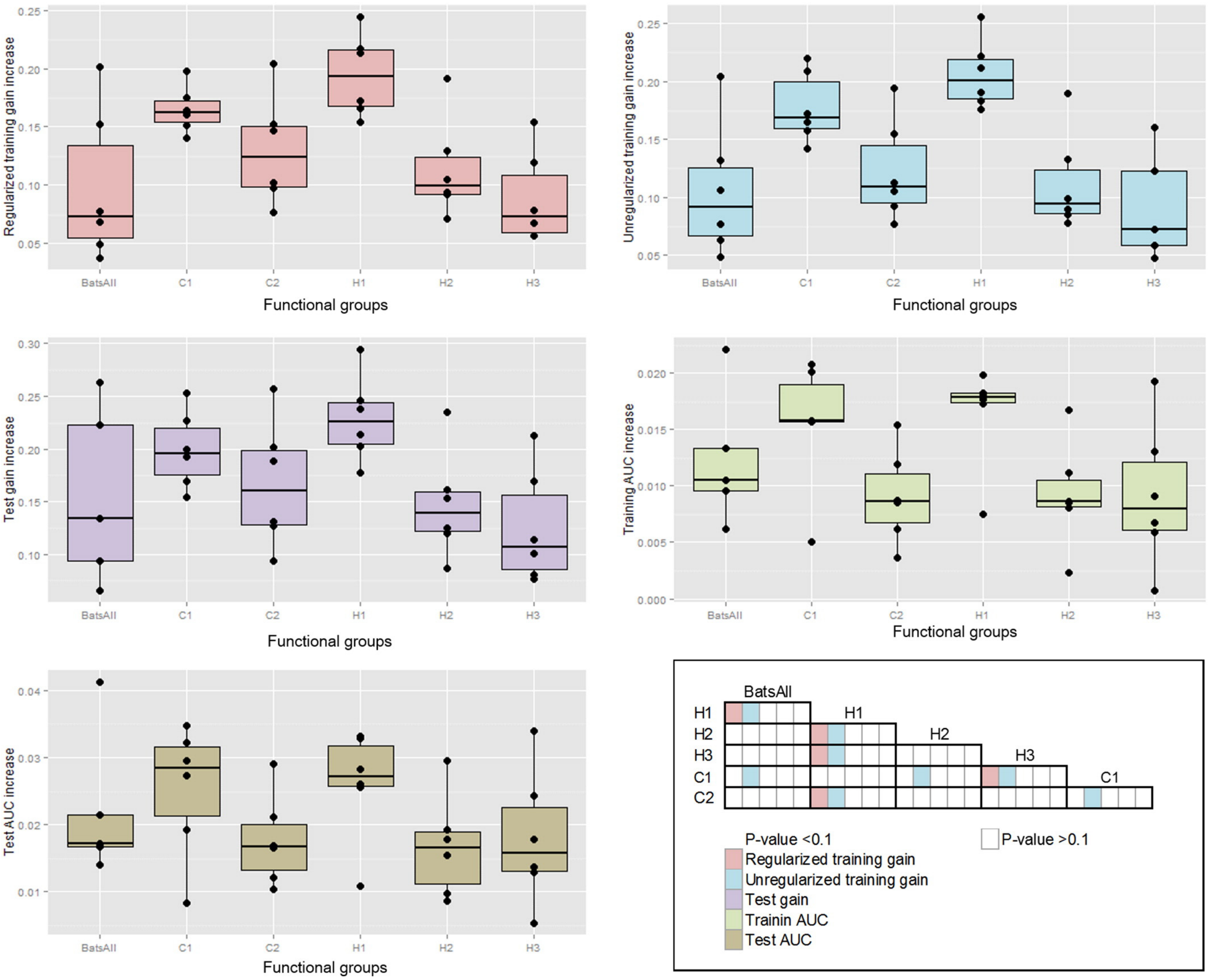

Fig. 2. Relative performance of SDMs of different functional groups. BatsAll — all bats, H1 — small herbivores (1 g–100 g), H2 — medium size herbivores (100 g–5000 g), H3 — large herbivores (5000 g+), C1 — small carnivores (1 g–100 g), C2 — medium size carnivores (100 g–5000 g). In black frame a pairwise comparison of all groups using a Mann–Whitney test with Benjamini and Hochberg (1995) correction method, *P*-value <0.1.

Performance measures of the post-thinning SDMs (the random thinning model) were similar to the respective pre-thinning SDMs, and lower than those of the SDMs built following user level data cleaning (Table 5). The increase in SDM performance was higher in small herbivores and small carnivores than in other groups (Fig. 2).

4. Discussion

We estimated the effect of tailored and easy-to-execute data cleaning on SDM performance at different spatial scales, using occurrence data of Australian mammals (over 1,000,000 records). The study-specific cleaning procedures improved SDM performance significantly across all the studied scales. SDMs and cleaning procedures were simple, basic, and designed for all Australian mammals as one group; fine-tuning them to a specific species or functional group would probably result in higher value of data cleaning and bigger improvements in performance.

Quantifying the effectiveness of data cleaning using SDMs is novel. Datasets like GBIF are frequently used to develop species distribution models, which are widely used in a range of fields and applications (Elith and Leathwick, 2009; Naimi and Ara, 2016; Peterson et al., 2011). Several methodological approaches have been developed to improve SDMs built with biased data (Bierman et al., 2010; Botts et al., 2012; Colwell and Coddington, 1994; Kadmon et al., 2004; Kent and Carmel, 2011; Rocchini et al., 2011; Syfert et al., 2013), yet quantitative tools to evaluate data quality are still scarce (Fei and Yu, 2015; Otegui and Guralnick, 2016). Our study may facilitate the development of different data quality indices (e.g. Apparent Quality Index develop by GBIF Spain; Representativeness and Completeness Index develop by Fei and Yu, 2015).

AUC is one of the most commonly used statistics to characterize model performance (Yackulic et al., 2013). However, its use has been highly criticized, especially in a presence-only modeling framework (Jiménez-Valverde et al., 2013; Lobo et al., 2008; Yackulic et al., 2013), since it ignores the predicted probability values and the goodness-of-fit of the model (Yackulic et al., 2013). Here, the use of all of MaxEnt performance measures was valid, since all comparisons were made between models of the same single species, in the same grid cell, and using the same model characteristics (predictors, background data, etc.). Nevertheless, our results show that the low informative value, and thus the low sensitivity of AUC limit its use in presence-only modeling.

Choosing the appropriate spatial configuration when evaluating a species distribution or an ecological niche is imperative (Elith and Graham, 2009). Therefore, we evaluated the effect of data cleaning across six different spatial scales. The results suggest that cleaning procedures were effective regardless of spatial grid configuration. This finding showcase the value of user-level data cleaning for big data, regardless of spatial scale.

We found that small mammals (1 g–100 g) were most affected by data cleaning (Fig. 2), possibly because retaining only coordinates with high spatial accuracy has a stronger effect on organisms with lower movement capabilities (Farjalla et al., 2012; Pöyry et al., 2008). High spatial accuracy and fine scale is crucial for studying distribution of low-mobility organisms.

In a typical research, data are very expensive, and filtering/removing big proportion of the data is inconceivable. In contrast, in the big-data world, data are plentiful and relatively inexpensive, and it is sometimes worthwhile to dispose large volumes of data for the sake of data quality. Here, for example, we disposed half a million records, which consisted 50% of the database, in order to increase data quality. Thus, tools for easy yet advanced query of the data are as important as tools for detecting and correcting errors (Vandepitte et al., 2015). The results of our study stress the need for data validation and cleaning tools that incorporate customizable techniques, for example by developing an R package. This will enable biodiversity researchers a much better understanding and control on data retrieved from large distributional databases as GBIF.

Improving the quality of biodiversity research, in some measure, is based on improving users-level data cleaning tools and skills. Adopting a more comprehensive approach for incorporating data cleaning as part of data analysis will not only improve the quality of biodiversity data, but will impose a more appropriate usage of such data. This can greatly serve the scientific community and consequently our ability to address more accurately urgent conservation issues.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ecoinf.2016.06.001.

Acknowledgments

TG was supported by a scholarship from The Technion – Israel Institute of Technology.

References

Ballesteros-Mejia, L., Kitching, I.J., Jetz, W., Nagel, P., Beck, J., 2013. Mapping the biodiversity of tropical insects: species richness and inventory completeness of African sphingid moths. Glob. Ecol. Biogeogr. 22, 586–595.

Belbin, L., Daly, J., Hirsch, T., Hobern, D., Salle, J.L., 2013. A specialist's audit of aggregated occurrence records: an "aggregator"s' perspective. Zookeys 76, 67–76.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300.

Bennett, S., 2012. Notes on methods for detecting spatial outliers in species occurrence data. Atlas of Living Australia.

Bierman, S.M., Butler, A., Marion, G., Kühn, I., 2010. Bayesian image restoration models for combining expert knowledge on recording activity with species distribution data. Ecography 33, 451–460.

Boakes, E.H., McGowan, P.J.K., Fuller, R.a., Chang-qing, D., Clark, N.E., O'Connor, K., Mace, G.M., 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biol. 8, e1000385.

Botts, E.a., Erasmus, B.F.N., Alexander, G.J., 2012. Methods to detect species range size change from biological atlas data: a comparison using the south African frog atlas project. Biol. Conserv. 146, 72–80.

Chapman, A.D., 2005a. Principles of Data Quality, Version 1.0. Global Biodiversity Information Facility, Copenhagen.

Chapman, A.D., 2005b. Principles and Methods of Data Cleaning — Primary Species and Species-Occurrence Data, Version 1.0. Global Biodiversity Information Facility, Copenhagen.

Colwell, R.K., Coddington, J.A., 1994. Estimating terrestrial biodiversity through extrapolation. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 345, 101–118.

de Souza Muñoz, M.E., De Giovanni, R., de Siqueira, M.F., Sutton, T., Brewer, P., Pereira, R.S., Canhos, D.A.L., Canhos, V.P., 2011. openModeller: a generic approach to species' potential distribution modelling. Geoinformatica 15, 111–135.

Elith, J., Graham, C.H., 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography 32, 66–77.

Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods Ecol. Evol. 1, 330–342.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 17, 43–57.

Farjalla, V.F., Srivastava, D.S., Marino, N.A.C., Azevedo, F.D., Dib, V., Lopes, P.M., Rosado, A.S., Bozelli, R.L., Esteves, F.A., 2012. Ecological determinism increases with organism size. Ecology 93, 1752–1759.

Feeley, K.J., Silman, M.R., 2010. Modelling the responses of Andean and Amazonian plant species to climate change: the effects of georeferencing errors and the importance of data filtering. J. Biogeogr. 37, 733–740.

Fei, S., Yu, F., 2015. Quality of presence data determines species distribution model performance: a novel index to evaluate data quality. Landsc. Ecol. 31, 31–42.

Fourcade, Y., Engler, J.O., Rödder, D., Secondi, J., 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.

Franklin, J., 2009. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge University Press.

Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. 32, 675–701.

García-Roselló, E., Guisande, C., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., Manjarrás-Hernández, A., Vaamonde, A., Granado-Lorencio, C., 2013. ModestR: a software tool for managing and analyzing species distribution map databases. Ecography 36, 1202–1207.

García-Roselló, E., Guisande, C., Heine, J., Pelayo-Villamil, P., Manjarrés-Hernández, A., González Vilas, L., González-Dacosta, J., Vaamonde, A., Granado-Lorencio, C., 2014. Using modestr to download, import and clean species distribution records. Methods Ecol. Evol. 5, 708–713.

GBIF, 2013. GBIF science review. 33 pages. Copenhagen. Global biodiversity information facility. (Available online at) http://www.gbif.org/resources/2261.

GBIF, 2015. GBIF annual report 2014, Copenhagen: global biodiversity information facility. (34 pp. Available online at) http://www.gbif.org/resource/annual_report_2014.

Geng, Z., Laramee, R., Loizides, F., Buchanan, G., 2011. Visual analysis of document triage data. IMAGAPP/IVAPP 151–163.

Guralnick, R., Hill, A., 2009. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. Bioinformatics 25, 421–428.

Hackett, E.J., Parker, J.N., Conz, D., Rhoten, D., Parker, A., 2008. Ecology Transformed: The National Center for Ecological Analysis and Synthesis and Changing Patterns of Ecological Research. Scientific Collaboration on the Internet. The MIT Press, pp. 277–297.

Hefley, T.J., Baasch, D.M., Tyre, A.J., Blankenship, E.E., 2014. Correction of location errors for presence-only species distribution models. Methods Ecol. Evol. 5, 207–214.

Hijmans, R., Guarino, L., Mathur, P., 2012. DIVA-GIS Version 7.5 Manual 71.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978.

Jiménez-Valverde, A., Acevedo, P., Barbosa, a.M., Lobo, J.M., Real, R., 2013. Discrimination capacity in species distribution models depends on the representativeness of the environmental domain. Glob. Ecol. Biogeogr. 22, 508–516.

Kadmon, R., Farber, O., Danin, A., 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecol. Appl. 14, 401–413.

Kent, R., Carmel, Y., 2011. Presence-only versus presence–absence data in species composition determinant analyses. Divers. Distrib. 17, 474–479.

Lobo, J.M., Jiménez-valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. 17, 145–151.

Maldonado, C., Molina, C.I., Zizka, A., Persson, C., Taylor, C.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? Glob. Ecol. Biogeogr. 24, 973–984.

Mathew, C., Güntsch, A., Obst, M., Vicario, S., Haines, R., Williams, A., de Jong, Y., Goble, C., 2014. A semi-automated workflow for biodiversity data retrieval, cleaning, and quality control. Biodivers. Data J. 2, e4221.

Mesibov, R., 2013. A specialist's audit of aggregated occurrence records. Zookeys 293, 1–18.

Michener, W.K., Jones, M.B., 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends Ecol. Evol. 27, 85–93.

Naimi, B., Ara, M.B., 2016. sdm: a reproducible and extensible R platform for species distribution modelling. Ecography (Cop.). 39, 368–375.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., D'amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P., Kassem, K.R., 2001. Terrestrial ecoregions of the world: a new map of life on earth. Bioscience.

Otegui, J., 2012. Quality and Fitness-for-Use Assessments on the Primary Data Indexed at the Global Biodiversity Information Facility (GBIF). University of Navarra.

Otegui, J., Ariño, A.H., 2012. BIDDSAT: visualizing the content of biodiversity data publishers in the global biodiversity information facility network. Bioinformatics 28, 2207–2208.

Otegui, J., Ariño, A.H., Chavan, V., Gaiji, S., GBIF, 2013a. On the dates of GBIF mobilised primary biodiversity records. Biodivers. Inform. 173–184.

Otegui, J., Ariño, A.H., Encinas, M.A., Pando, F., 2013b. Assessing the primary data hosted by the Spanish node of the global biodiversity information facility (GBIF). PLoS One 8, e55144.

Otegui, J., Guralnick, R.P., 2016. The Geospatial Data Quality REST API for Primary Biodiversity Data. Bioinformatics btw057.

Peterson, A.T., Soberón, J., Krishtalka, L., 2015. A global perspective on decadal challenges and priorities in biodiversity informatics. BMC Ecol. 15, 15.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M., Araújo, M.B., 2011. Ecological Niches and Geographic Distributions. Princeton University Press.

Phillips, S., 2008. A Brief Tutorial on Maxent. AT&T Research, AT&T Research.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259.

Pöyry, J., Luoto, M., Heikkinen, R.K., Saarinen, K., 2008. Species traits are associated with the quality of bioclimatic models. Glob. Ecol. Biogeogr. 17, 403–414.

Development Team, Q.G.I.S., 2013. QGIS Geographic Information System. Open Source Geospatial Foundation Project (http://qgis.osgeo.org).

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jimenez-Valverde, A., Ricotta, C., Bacaro, G., Chiarucci, A., 2011. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. Prog. Phys. Geogr. 35, 211–226.

Soininen, J., Luoto, M., 2014. Predictability in species distributions: a global analysis across organisms and ecosystems. Glob. Ecol. Biogeogr. 23, 1264–1274.

Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. Science 80-. (240), 1285–1293.

Syfert, M.M., Smith, M.J., Coomes, D.a., 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. PLoS One 8, e55158.

Vandepitte, L., Bosch, S., Tyberghein, L., Waumans, F., Vanhoorne, B., Hernandez, F., De Clerck, O., Mees, J., 2015. Fishing for data and sorting the catch: assessing the data quality, completeness and fitness for use of data in marine biogeographic databases. Database 2015, 1–14.

Velásquez-Tibatá, J., Graham, C.H., Munch, S.B., 2015. Using measurement error models to account for georeferencing error in species distribution models. Ecography (Cop.) 38, 001–012.

Verborgh, R., De Wilde, M., 2013. Using OpenRefine. PACKT Publising.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D., 2012. Darwin Core: an evolving community-developed biodiversity data standard. PLoS One 7, e29715.

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Campbell Grant, E.H., Veran, S., 2013. Presence-only modelling using MAXENT: when can we trust the inferences? Methods Ecol. Evol. 4, 236–243.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.a., Culham, A., 2007. How global is the global biodiversity information facility? PLoS One 2, e1124.