Biological Conservation 181 (2015) 102–110

Contents lists available at ScienceDirect

Biological Conservation

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

Uncertainty analysis of crowd-sourced and professionally collected field data used in species distribution models of Taiwanese moths

Yu-Pin Lin [a,*], Dongpo Deng [b,c], Wei-Chih Lin [a], Rob Lemmens [b], Neville D. Crossman [d], Klaus Henle [e], Dirk S. Schmeller [e,f,g]

[a] Department of Bioenvironmental Systems Engineering, National Taiwan University, Taiwan, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan
[b] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands
[c] Institute of Information Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan
[d] CSIRO Land and Water Flagship, PMB 2, Glen Osmond, South Australia 5064, Australia
[e] UFZ – Helmholtz Centre for Environmental Research, Department of Conservation Biology, Permoserstr. 15, 04318 Leipzig, Germany
[f] Université de Toulouse, UPS, INPT, EcoLab (Laboratoire Ecologie Fonctionnelle et Environnement), 118 route de Narbonne, 31062 Toulouse, France
[g] CNRS, EcoLab, 31062 Toulouse, France

ARTICLE INFO

Article history:
Received 17 June 2014
Received in revised form 7 November 2014
Accepted 8 November 2014

Keywords:
Social media
Citizen science
Volunteer survey
Prediction of species distribution
Uncertainty
Natural language
Large-scale monitoring program

ABSTRACT

The purposes of this study are to extract the names of species and places for a citizen-science monitoring program, to obtain crowd-sourced data of acceptable quality, and to assess the quality and the uncertainty of predictions based on crowd-sourced data and professional data. We used Natural Language Processing to extract names of species and places from text messages in a citizen science project. Bootstrap and Maximum Entropy methods were used to assess the uncertainty in the model predictions based on crowd-sourced data from the *EnjoyMoths* project in Taiwan. We compared uncertainty in the predictions obtained from the project and from the Global Biodiversity Information Facility (GBIF) field data for seven focal species of moth. The proximity to locations of easy access and the Ripley K method were used to test the level of spatial bias and randomness of the crowd-sourced data against GBIF data. Our results show that extracting information to identify the names of species and their locations from crowd-sourced data performed well. The results of the spatial bias and randomness tests revealed that the crowd-sourced data and GBIF data did not differ significantly in respect to both spatial bias and clustering. The prediction models developed using the crowd-sourced dataset were the most effective, followed by those that were developed using the combined dataset. Those that performed least well were based on the small sample size GBIF dataset. Our method demonstrates the potential for using data collected by citizen scientists and the extraction of information from vast social networks. Our analysis also shows the value of citizen science data to improve biodiversity information in combination with data collected by professionals.

1. Introduction

Data availability constrains the predictive accuracy of species distribution models (Phillips et al., 2009; Sardà-Palomera et al., 2012). Species distribution models (SDMs) are statistical models that exploit environmental and/or geographic information (data) to explain observed patterns of species distributions (Elith and Graham, 2009), and to answer questions in the fields of conservation biology, ecology, and evolution (Barbet-Massin et al., 2012). SDMs are influenced by various sources of uncertainty. Therefore the quantification of the uncertainty of SDMs and their predictions (Dormann et al., 2008; Wiens et al., 2009; Bagchi et al., 2013) as well as the improvement of their accuracy are major concerns. Uncertainty in SDM predictions has two main sources: (i) deficiencies of data such as missing covariates, small sample sizes, biased and absent data, and; (ii) errors in the specifications of the model (Barry and Elith, 2006; Dormann et al., 2008; Elith and Leathwick, 2009).

A general challenge to improve data quality and quantity for SDMs is the sharing of biodiversity data and the related need for a coordinated publishing and integration system (e.g. Hoffmann et al., 2014), especially for data that are collected from citizen science projects that engage non-professional volunteers to collect data and solve scientific problems (Silvertown, 2009; Dickinson

* Corresponding author. Tel.: +886 2 33663467.
E-mail addresses: yplin@ntu.edu.tw (Y.-P. Lin), dongpo.deng@gmail.com (D. Deng), b97602046@ntu.edu.tw (W.-C. Lin), lemmens@itc.nl (R. Lemmens), Neville.Crossman@csiro.au (N.D. Crossman), klaus.henle@ufz.de (K. Henle), dirk.schmeller@ufz.de (D.S. Schmeller).

http://dx.doi.org/10.1016/j.biocon.2014.11.012
0006-3207/

et al., 2010; Newman et al., 2010). Data repositories tend to be isolated from each other in the absence of standards and communication protocols. Additionally, the heterogeneity of terms and their meanings create obstacles in every aspect of data integration and use, including discovery, comparison, and assessment of quality (Wieczorek et al., 2012).

The difficulties of assembling sufficient distribution data for SDMs are paralleled by the heterogeneity of biodiversity monitoring protocols. To improve the quantity of data in biodiversity monitoring, distributional data are collected using a large variety of approaches, ranging from organized professional surveys to the *ad hoc* collection of observations from the general public (Higgins et al., 2012; Schmeller et al., 2009, 2012). A trade-off is made between data quality and cost efficiency of data collection (Schmeller et al., 2009; Sardà-Palomera et al., 2012). However, citizen science can solve some of the problem of the usually limited financial and human resources available for monitoring biodiversity (Schmeller et al., 2009; Devictor et al., 2010; Jetz et al., 2012). Volunteer-based Monitoring Schemes (VMS) have been applied in ecological monitoring in the last decade, such as in the eBird (http://ebird.org), BeeID (Stafford et al., 2010) and epicollect projects (Aanensen et al., 2009), which allow collaboration among volunteers in the collection of ecological data.

Recently, an alternative approach to field monitoring has involved the use of museum specimens, or opportunistic record collection by volunteers using web-based tools (Roberts et al., 2007; Munson et al., 2010; Conrad and Hilchey, 2011; Sullivan et al., 2014). Volunteers nowadays use online search tools to identify species, such as the "Cornell Lab of Ornithology's All About Birds" Web (Farnsworth et al., 2013). However, general search engines are not built as tools for species identification (the semantics differ), and can provide deficient or misleading results (Farnsworth et al., 2013). Web 2.0 technologies, such as social media (e.g. Twitter, Facebook, Flickr, and YouTube), are changing interactions among citizen scientists. Researchers can use social media to engage with many citizen scientists to increase the spatial and temporal scope of data collection. Although social media produces vast volumes of information, the transformation of such crowd-sourced information for scientific use is difficult because social media are not generally designed for scientific purposes (Bifet and Frank, 2010). Additionally, crowd-sourced data supplied by users through social media are often unstructured, comprising for example, short text messages and photographs. Such unstructured data are difficult to use in scientific analysis. Therefore, the effective exploitation of social media in citizen science depends on the development of a method for transforming unstructured data into structured data.

Using existing data to identify threshold dates and counts, e-bird has created accurate spatial and temporal filters for evaluating incoming check list submissions with machine learning analytical techniques (Sullivan et al., 2014). To exploit data from social media, Natural Language Processing (NLP) approaches can be used to extract large amounts of information from free text (Thessen et al., 2012) to identify meaningful entities such as locations and species. NLP can assist to obtain useful crowd-sourced data from presence-only data from museum or herbarium collections (Graham et al., 2004) or open access data from volunteer observation networks (Naimi et al., 2011).

One of the problems associated with crowd-sourced data is the uncertainty of the exact location of sampling sites (Naimi et al., 2011). This uncertainty is caused by various factors, including operator errors and inaccuracy in the measurement of the location because of failure to specify the geographical datum or other georeferencing errors (Graham et al., 2004; Naimi et al., 2011). Sardà-Palomera et al. (2012) found that predictions of SDMs based on standardized monitoring are more accurate than those based on crowd-sourced data with small sample sizes, especially when the modeling of common species is based on a combination of simple observations that are opportunistically obtained by volunteers and professional standardized monitoring surveys. Tulloch et al. (2013) indicated that professional sampling in survey gaps is needed to reduce bias in volunteer-collected datasets. However, cross-validations can be used to provide many additional opportunities for assessing species distribution records and flagging and filtering them based on quality, such as by using predicted presence probabilities for outlying points (Jetz et al., 2012). A cross-validation of such data can be achieved with Maximum Entropy models (Guisan et al., 2006). In Maximum Entropy modeling, the area under the receiver operating characteristic curve (AUC) characterizes the ability of models to distinguish presence records from background data (Cord and Rödder, 2011). Zipkin et al. (2012) demonstrated that AUC is helpful in quantifying the uncertainty in model predictions while explicitly accounting for detection biases. Liu et al. (2011) emphasized the need to determine the accuracy of AUC and suggested the use of bootstrapping and randomization methods for estimating confidence intervals. Such a bootstrapping approach is necessary when data are not mutually independent, for example when partitioning the dataset into training data and test data (Gibson et al., 2007).

Here we report on our use of social media in the monitoring of biodiversity in Taiwan. *EnjoyMoths*, with 1877 participants, is a citizen science project that involves identifying moths and their occurrences. The project uses Facebook and is hosted by the Endemic Species Research Institute, Council of Agriculture, Taiwan. Although users can easily contribute observations using Facebook, transforming their observations, usually text and photographs, into structured data for scientific purposes is difficult. For privacy and security reasons, most Exchangeable Image Format (EXIF) data are stripped when photographs are uploaded to Facebook. Without EXIF data, a photograph from Facebook is just a graphic record: the photograph itself does not indicate the time and location of where it was taken. The text that accompanies the photographs is the main source of ecological information about the species therein. In the *EnjoyMoths* project, to satisfy the requirements for scientific data, a Natural Language Processing (NLP) approach was developed to extract the names of species and places from the text. To assess the uncertainty in data collected through the *EnjoyMoths* project, we applied a bootstrap method and Maximum Entropy. Moreover, we tested the spatial bias of the national GBIF branch Taiwan Biodiversity Information Facility (TaiBIF) and the crowd-sourced samples using a proximity to easy access locations analysis. We tested if the uncertainty of data from volunteer based monitoring and from professional data sources are markedly different and if the combination of the two data sources improves our understanding of species distributions and the robustness of SDMs. Based on our results we developed recommendations for the use of our approach to complement existing data sets with citizen science data to yield a more robust understanding of the status and trend of biodiversity.

2. Material and methods

Taiwan is an island with an area of 36,000 km^2. It has a subtropical environment. Here, data of seven species, *Asota heliconia zebrina*, *Chrysaeglia magnifica*, *Asota egens indica*, *Biston perclarus*, *Cyana hamata*, *Eumelea ludovicata*, and *Lebeda nobilis* (see Appendix A.1 for more details; National Museum of Natural Science, http://www.nmns.edu.tw/index_eng.html; TaiBIF, http://www.taibif.org.tw/; Surprise Mountain Line, http://gaga.biodiv.tw) were selected from the *EnjoyMoths* project as focal species. Observations were posted by participants on the *EnjoyMoths* Facebook interest group with an observation provider, an observation location, an