Diversity and Distributions, (Diversity Distrib.) (2015) 1–10

BIODIVERSITY RESEARCH

Scientists and software – surveying the species distribution modelling community

Sadia E. Ahmed[1], Greg McInerny[1,2], Kenton O'Hara[3], Richard Harper[3], Lara Salido[4], Stephen Emmott[1] and Lucas N. Joppa[5]*

[1]*Computational Science Laboratory, Microsoft Research, 21 Station Road, Cambridge CB1 2FB, UK,* [2]*Department of Computer Science, University of Oxford, Wolfson Building, Parks Road, Oxford OX1 3QD, UK,* [3]*Human Experience and Design Group, Microsoft Research, 21 Station Road, Cambridge CB1 2FB, UK,* [4]*National Museums Collections Centre, 242 West Granton Road, Granton, Edinburgh EH5 1JA, UK,* [5]*Microsoft Research, 14820 NE 36th Street, Redmond, WA 98052, USA*

*Correspondence: Lucas N. Joppa, Microsoft Research, 14820 NE 36th Street, Redmond, WA 98052, USA.
E-mail: lujoppa@microsoft.com

ABSTRACT

Aim Software use is ubiquitous in the species distribution modelling (SDM) domain; nearly every scientist working on SDM either uses or develops specialist SDM software; however, little is formally known about the prevalence or preference of one software over another. We seek to provide, for the first time, a 'snapshot' of SDM users, the methods they use and the questions they answer.

Location Global.

Methods We conducted a survey of over 300 SDM scientists to capture a snapshot of the community and used an extensive literature search of SDM papers in order to investigate the characteristics of the SDM community and its interactions with software developers in terms of co-authoring research publications.

Results Our results show that those members of the community who develop software and who are directly connected with developers are among the most highly connected and published authors in the field. We further show that the two most popular softwares for SDM lie at opposite ends of the 'use-complexity' continuum.

Main conclusion Given the importance of SDM research in a changing environment, with its increasing use in the policy domain, it is vital to be aware of what software and methodologies are being implemented. Here, we present a snapshot of the SDM community, the software and the methods being used.

Keywords
Scientific software, species distribution, survey.

INTRODUCTION

Understanding why species occur where they do is one of ecology's oldest questions, and predicting where species might go under changing climates is one of its newest and most important challenges. Much has changed since the work of early scientists such as Humboldt, whose qualitative descriptions of isothermal lines and altitudinal observations of plant communities in 1805 (Stepan, 2001) were an earlier form of what is now generally known as species distribution modelling (SDM). Various other terms include ecological or environmental niche modelling, bio-climate modelling, habitat suitability modelling and climate envelope modelling. Each term implies subtle yet important differences in approach (Araújo & Peterson, 2012; Peterson & Soberon, 2012), but we use the term SDM to cover all of these in a general sense. Regardless of nomenclature, perhaps the most important changes of all have been the integration of statistics, biology and geography (Elith & Leathwick, 2009). These developments have led to a rapid increase in the complexity and sophistication of SDM methods – and an increasing reliance on computing and emergence of diverse software to support this research.

SDM software has grown both in number and in complexity, from domain-specific platforms to more general platforms that can be adapted via self-coding or libraries. While SDM is highly diverse in terms of methods, it is relatively constrained in terms of implementation. Wrapping complex analyses into scientific software, be it a tool, library, or package, offers a variety of potential benefits such as extending

DOI: 10.1111/ddi.12305
http://wileyonlinelibrary.com/journal/ddi

A Journal of Conservation Biogeography
Diversity and Distributions

analyses, increasing scientific output and increasing accessibility. Accessibility expands the user base to less expert users who would otherwise be unable to implement the computer programming required to run the algorithms. This has led to SDM rapidly rising in prominence in the scientific literature and being used across a diverse range of applications (Guisan *et al.*, 2006; Kozak & Wiens, 2006; Moffett *et al.*, 2007; Esselman & Allan, 2011; Svenning *et al.*, 2011). SDM has been used in research branches across life and environmental sciences (Thuiller *et al.*, 2009), with an increase in peer-reviewed papers within the field of ecology [from 10 papers in 1992 to 2546 in 2011, accounting for 0.3% of all ecological papers (Guisan *et al.*, 2013)]. Today there is a strong focus on future predictions using SDM, where it is used to assess how the changing environment (climate, land use, etc.) will influence species distributions (Bellard *et al.*, 2013), with an aim to provide guidance for mitigation through policy (Reed & Rodda, 2009; Dawson *et al.*, 2011; Hof *et al.*, 2011; Pereira *et al.*, 2011; Guisan *et al.*, 2013).

Given the reliance on software in the SDM domain, it is important to understand the context in which SDM software is being used (Joppa *et al.*, 2013). There have been a variety of calls for unification and synthesis in the SDM literature (e.g. Elith & Leathwick, 2009; Aarts *et al.*, 2012) and discussions on topics ranging from the correct usage of methods (e.g. Merow *et al.*, 2013) to the terminology this domain uses (Elith & Leathwick, 2009; McInerny & Etienne, 2012; Warren, 2012; Soberón, 2014). Yet, these calls and debates are rarely based on information from the whole SDM community. SDM is one of the most widely reviewed topics in ecology (Araújo & Peterson, 2012); however, as noted by Austin (2007), any review can contain biases due to the sampling of references from a very large literature (also see Elith & Leathwick, 2009). Further, reviews may be based on particular perspectives or schools of thoughts. As a result, we rarely know what the broader SDM community is really doing and thinking, or what tools they are using. The latest, cutting-edge research trends may differ from the predominant trends in the whole field, and the latest research papers may not necessarily characterize the broader use of SDM. Many topics that are crucial to the development of SDM are not typically discussed in publications or are perceived to be outside of the scope of typical academic publications (e.g. terminology and ideas used to interpret models, or the characteristics of software). Further, not all SDM activities lead to academic publications, but are important to current usage (e.g. the grey literature and land management actions via NGOs or governmental agencies). Statements are often made that are not necessarily supported by evidence and are rather based on individual experiences, anecdotes or assumptions about the community (e.g. levels of expertise in various areas such as coding and statistical ability or the prevailing trends in software use). To address this issue, we carried out a large-scale survey (also see Joppa *et al.*, 2013) of the SDM community. In it, we report a number of findings of interest to SDM modellers, software developers, researchers and the community as a whole. Surveys are of course imperfect and will always be a biased sample of the community; however, we set out to capture a 'snapshot' of the SDM field by asking who is participating in SDM research, what types of questions they are trying to answer, which software and analysis tools they are using and how users feel about the software tools they use.

One major finding of this process, presented in Joppa *et al.*, 2013 on the treatment of black-box algorithms in software, was related to a very small subset of the overall survey (results used in Joppa *et al.*, 2013 are highlighted in SI Data). Here, we present the full survey, providing the first large-scale assessment of the SDM community with a focus on software use. Among our results, we find that the two most popular softwares for SDM lie at opposite ends of the 'use-complexity' continuum, MAXENT at the point-and-click (termed 'click' henceforth) end and R at the syntax driven end, and that these two platforms also standout from other software in terms of user satisfaction. We also find that those who develop SDM software and who are directly connected to developers are among the most highly published and connected authors in the field.

Methods

Eight face-to-face interviews were used to (1) provide an initial background to the perceptions and use of SDM software within the community and (2) inform the design of the web-based survey. The interviews with groups and individuals included 19 scientists working across the SDM domain, from researchers in training to established researchers and a software developer (for details, please see Data S1, Interviews). The survey design was further informed by a literature search into general scientific software development, specifically examining aspects of problem formulation, model design, implementation, calibration and uncertainty assessment. The resulting survey (presented in Data S3) consisted of 50 questions: a mixture of multiple choice, Likert rating and free-text answers. These questions covered four main areas of interest:

1. *Respondent background*, covering aspects relating to the respondent such as age, experience with SDM and organizational position.

2. *General software/statistics use*, contained questions pertaining to the general computational competency of the respondents.

3. *SDM research*, this section sought to understand the type of research that the respondents were conducting within the SDM domain. More specifically, this investigated issues relating to the types of questions explored (e.g. methodological vs. applied), the types of data used and the types of methods implemented. We also considered peripheral aspects of research such as what influences the questions asked and how methods are assessed (i.e. are the original technical papers consulted or do user guides suffice).

4. *SDM software use*, here questions related specifically to SDM software, covering aspects, such as what software is used, what it is used for, reasons for software choice, satisfaction with software and awareness of other SDM software.

The survey was distributed online via Qualtrics for 1 month (mid-June 2011–mid-July 2011). To minimize survey bias and maximize response coverage in terms of user type and location, we contacted over 130 people within the SDM domain based at universities, research institutes and ecological/zoological societies, from across the globe. We asked these individuals to both participate themselves and to enlist participation from others. In addition, links to the survey were posted on scientific blogs and software user group forums to reach an even wider audience.

From our literature search (see details below), we estimate that 34,779 people have published within the SDM literature. Based on this as a population size, a desired confidence interval of 95% and an error tolerance of 5%, we used two online survey power analysis tools (Qualtrics and Surveymonkey, SI Power analysis) to determine how many respondents would be required to be representative.

To assess the differences between respondents using point-and-click and syntax driven software, data were split into two groups based on users' preferred software. We use the same syntax/click classification as Joppa *et al.* (2013) – specifically, MAXENT, OPENMODELLER, MODECO, GARP, BIOMAPPER, CANOCO, Domain and Species are click and that R, MATLAB, WINBUGS, OPENBUGS and BIOMOD are syntax. Based on this grouping, we investigated the differences in SDM methods implemented for SDM and approaches used to evaluate SDM results divided by click and syntax software users.

The generic R function *princomp* from the 'stats' library (R Development Core Team, 2006), with default settings, was used to conduct principal components analysis (PCA) on survey responses relating to 13 key softwares that were identified during our preliminary interviews (MAXENT, R, BIOMOD, MATLAB, OPENMODELLER, MODECO, GARP, BIOMAPPER, CANOCO, WINBUGS, OPENBUGS, DOMAIN and SPECIES) in order to characterize and compare the software in terms of learning experience, system capabilities and overall satisfaction among our survey sample. PCA was conducted on average weighted scores for each software within each response category. For example, for the 'Learning PCA' each software user rated (1–5) how strongly they agreed with statements such as 'It took a short time to become effective and productive in using this software'. The counts for each score for each software were multiplied by the score and divided by the total number of responses across software for that question. This provided the weighted average for a given software for each question, and the PCA was performed on the resulting values.

To further investigate the characteristics of the SDM community and its interactions with software developers in terms of co-authoring research publications, we conducted an extensive literature search. All SDM papers published between 1990 and July 2011 found on ISI Web of Science were used to construct an SDM authorship network. We use the network as a proxy for collaborative behaviour of developers and users. Search phrases included: 'habitat suitability model', 'bio-climate model', 'climate suitability model', 'ecological niche model', 'environmental niche model' and 'species distribution model'. The authors were grouped into five categories. (1) *Developers*, the authors of the original publication detailing an SDM software/package. Note: for this group, we only focus on software developers rather than those who developed analytical methods as the vast majority of software appropriate and implement existing techniques (e.g. techniques form machine learning [e.g. Artificial Neural Networks (ANN), Maximum Entropy Methods (MAXENT)] and Statistics [e.g. General Additive Modelling (GAM), Random Forests)] rather than develop and implement new techniques. In addition, software developers will rarely have co-authored a paper with those who originally developed the original modelling technique. A list of developers considered for each software is presented in SI Data, developers list. (2) *Super-collaborators*, authors that link 3 or more developer groups, (3) *Ingroup collaborators*, authors that link two developer groups, (4) *Single-collaborators*, authors that have co-authored with a single developer group and (5) *Other*, authors who publish on SDM but are not directly linked with any developer groups (i.e. category 1, *developers*). We used an Excel plug-in, NodeXL (http://nodexl.codeplex.com/), to plot the connections between author categories 1–4 (authors not linked with a developer group were removed from the graph). Across author categories 1–5, the number of papers and number of unique collaborators were calculated for each author. We compared statistically if the level of connectedness with developers (i.e. the authors category) significantly affected the number of papers published or the number of unique co-authors through one-way analysis of variance (ANOVA) and t-tests.

RESULTS

Respondent background

A total of 364 people completed the survey. Of the respondents, most were associated with academic institutions (73%), based in Europe or North America, aged between 24 and 40, with 6–10 years' experience with scientific software. While the majority of survey respondents were from academic institutions, responses included those working at government institutions, non-governmental organizations (NGOs) and the private sector. The survey responses sampled the spectrum of roles SDM users may have. As expected, fewer 'team leaders' were represented in the responses due to their lower frequency (see SI section A). The majority of respondents had peer-reviewed published papers (85%), and of those, 82% had published in the SDM domain (SI B).

All survey responses except those pertaining to personal information are available in SI data, and complete results are available in SI A-SI W.

General software use

The vast majority of survey respondents reported competence with software; most respondents, 40%, considered themselves to be 'good and technical', and 34% considered themselves to be 'familiar but not technical' with 94% considering themselves to be at least familiar with software and its use (SI C). This is of course based on individualistic ratings of personal software familiarity and subjective interpretations of the scale presented in the survey. Without objective benchmarks to test responses, we assume a common scale and so the results, as with any survey, include this potential survey bias.

The use of general software (i.e. software not specifically designed for SDM research) was ubiquitous among respondents the most commonly used being ArcGIS, Microsoft Excel and R, with 317, 310 and 304 of 364 respondents using each (SI D). Interestingly, the most commonly used 'general' software is spatially oriented (ArcGIS), highlighting the necessity of geography to researchers within this field.

To assess the level of coding ability, we asked respondents how many and which languages they were comfortable working with. Of the 10 programming languages listed in the survey, R was the most commonly used with 73% of respondents comfortable coding in R. Few respondents reported being comfortable with any of the other nine languages (Visual basic, MATLAB, Fortan, C, C++, Java, C#, f#, Python) (0–16% of the respondents). 10% of respondents said that they used software not listed (i.e. 'other') citing 17 other coding platforms including SQL, Perl and Pascal (SI E).

SDM research

All the survey respondents worked on SDM at least some of the time, with 8% stating that they worked on SDM 'always' (SI F). In terms of SDM, less than a third of respondents considered themselves '*expert* '(99/364) and most respondents considered themselves familiar with the topic (189/364). This suggests that people are using software while developing expertise. There was a degree of variation in the preferred term for this type of modelling, and of the 10 options presented, the most popular was 'species distribution modelling' (284/363), followed by 'habitat suitability modelling' (144/363) and 'ecological niche modelling' (141/363) (SI G).

To assess what SDM is being used to investigate, we asked questions to characterize the major themes of research respondents conducted. SDM was most frequently used to *describe* the current distribution of species, while slightly less frequently used to *analyse* changes in observed distributions or *predict* future/past species distributions. Here we take, 'describe' to be simply mapping out the species range, while 'analysing changes' suggests finding a reason for observed changes in range and 'predicting' is to find the range under future or past, changing environmental variables. More striking than the differences in analysis purpose are the marked differences in the complexity of the study systems. Almost equal numbers of respondents investigated single and multiple species models across all three purposes (i.e. describe, analyse and predict), but less than half that number conducted studies that investigated the interactions between species in those models (Fig. 1). There were three types of questions that the respondents addressed with SDM: (1) applied questions which relate to informing a real-world problem, for example establishing invasive species risk (Reed & Rodda, 2009), (2) pure questions which relate to long-standing questions on the determinants of species distribution patterns and (3) methodological questions which are concerned with model properties or comparisons between results from several methodological approaches. We found researchers generally worked on more than one type of question, with almost one-third (27%) addressing all three types (Fig. 1). When asked 'what is the maximum number of species you have applied SDM (to) in a single study?' our respondents reported a range of between 1 and 120,000 species, with a median of 20 and a mean of 880 (SD = 8027). The distribution of the number of species to which SDM had been applied in a single study was heavily left skewed with single species models predominating

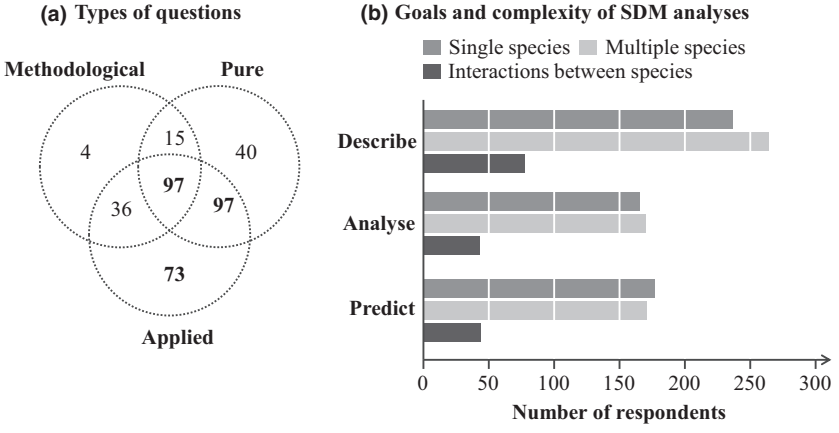

Figure 1 Types of questions SDM is applied too. (a) three key types of questions SDM is used to answer (pure, applied and methodological) and the degree of overlap between them according to respondents of the survey, and (b) three key methodological approaches used in SDM (descriptive, analytical and predictive approaches) divided by the complexity of the models used (single species, multiple species, with interactions).

(14%), and a quarter of studies (26%) focussed on 1–3 species (SI H).

Data are the foundation of all scientific enquiry, and the availability of data can determine the scope of a study and what methods are used. As such, we queried the data use of respondents with regard to their SDM research. Data used in analyses were commonly 'presence only' data (278/362) and 'presence/absence' data (213/362) (respondents could respond in the positive to both of these questions). Less than half the respondents said they used abundance data (158/362), and a small minority said they used other types of data including biomass, production, rate and proportion data. The sources of data for species occurrence and climate varied; climate data were more commonly 'downloaded' and 'bought' than species occurrence data (278/359 and 29/359 for climate data vs. 168/362 and 8/362), and species occurrence data were more commonly self-collected than climate data (117 vs. 49, SI I).

As highlighted in the introduction, there is a plethora of available SDM methods and we wanted to investigate (1) How SDM users assessed which methods to use and (2) Which of those chosen methods were most popular? With regards to the implementation and interpretation of the methods, the majority of respondents 281/358 said they referred back to the original methodological papers. 234/358 respondents referred to standard papers and 194/3584 also referred to reviews and guide papers. The most commonly used methods in SDM modelling were GLMs, GAMs and maximum entropy, with 65%, 44% and 52% of respondents saying that they use them. The next three most commonly used methods were approximately half as popular as the most popular three: random forests (classification trees), boosted regression trees and ensemble of models from different methods, with 30%, 27% and 26%, respectively. In general, we found little difference between the methods implemented by syntax and click users, except for maximum entropy, for which click users were almost three times as prevalent (27% click users compared to 10% of syntax users implemented this method, SI J). Finally, to assess model performance, the metric cited as best to judge model performance was 'AUC/ROC' (32%). We find little difference in model assessment metrics between click and syntax driven software, except in the case of AUC/ROC where 42% of click software users compared to 27% of syntax users use this metric (SI K). These differences in methods and metrics used between click and syntax users could be attributed to the prevalence of MAXENT (which primarily uses a maximum entropy method and AUC/ROC metrics) among click users.

SDM software

The most commonly used software specifically for SDM research were R (81%) and MAXENT (64%) (percentages assume respondents who did not provide an answer did not use the software). Generally, the more commonly any given software had been heard of, the more commonly it was used. For example, of the 340 respondents who had heard of R, 296 used it (SI L).

The reasons for SDM software choice (in descending order starting from most popular) were 'it is freely available', 'it is a versatile tool', 'it is the easiest way of implementing the analysis', 'it is the most recognized tool in my discipline' and 'it is the only tool I know how to use'. Further reasons for choice were 'it has been validated against other methods in peer-review', 'the method is entirely transparent' and 'it produces models that are ecologically relevant' (SI M).

With regards to how frequently respondents used each software over 50% of respondents (for each individual software) stated that they 'always' used a given software. Suggesting either a reliance on a given software or that the software is well suited to the task (or both). BIOMOD, MAXENT and R were the exception to this with far fewer respondents stating 'always' (46%, 15% and 4%, respectively) rather with the majority of these users stating that they 'regularly' or 'sometimes' use these softwares (SI N). Across respondents, there was generally software fidelity (few respondents claimed to have previously used software but do so no longer), with R having the highest fidelity (SI O). It should be noted, however, that while respondents claim to 'always' use a given software, many of them also provided responses for other software; thus, if a respondent provided responses for software other than their 'primary' software, we assumed that they also used this software (SI O table). Two notable exceptions to this trend were WINBUGS and OPENBUGS which had relatively high proportions of previous users who switched to other software. However, these two softwares had very few total users as such there is high degree of uncertainty around this.

From an SDM users' point of view, R and MAXENT stand out as best for overall software utility when judged on (1) usefulness, (2) learning, (3) system capabilities and (4) user satisfaction (Fig. 2 & SI P). In terms of usefulness, R and MAXENT stand out from the other software for the perceived level of control over the analysis that the software provides, with R viewed as providing the most control and MAXENT requiring the fewest actions needed to implement the analysis. With regard to learning, we find that R is judged to have easy to use documentation and MAXENT has a shorter time to effective use. For system capabilities, R and MAXENT are both rated highly as being easy to run batch analyses with and judged to be good at interfacing with other software, however, for both, R rates higher. It is interesting that the two 'best' softwares lie at either end of the use-complexity continuum: R being very general, flexible and syntax driven, while MAXENT is very specific with only one implemented method (albeit with adjustable settings) and can be 'click' based. We found that of our 364 respondents, the majority (217) were syntax software users. R and MAXENT were the two most popular software within their category of syntax (189/217) and click (109/147).

Some interesting points became apparent when respondents were asked to agree with statements (SI Q). For example, in general, respondents think SDM can provide valuable

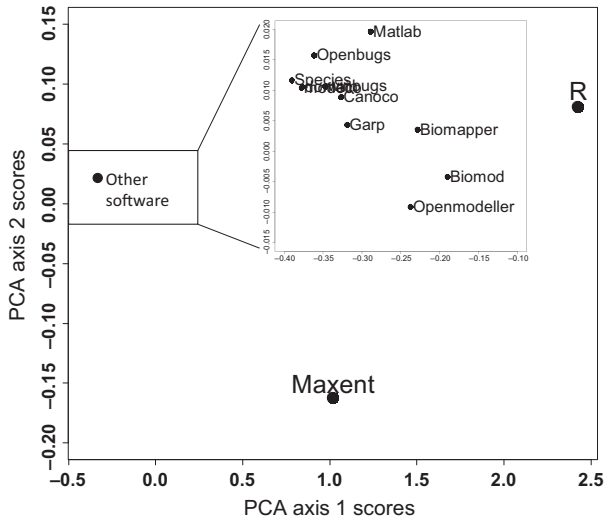

Figure 2 PCA of SDM software usefulness. Principal components analysis (PCA) based on the respondents' opinion of software usefulness; showing that R and MAXENT stand out from the other software (inset shows a close up of the phase space of the remaining software). Similar patterns are observed for opinions relating to 'Learning the software', 'system capabilities' and overall 'satisfaction' with the software. PCA was conducted on survey responses relating to the 13 key softwares (R, MAXENT, BIOMOD, MATLAB, OPENMODELLER, MODECO, GARP, BIOMAPPER, CANOCO, WINBUGS, OPENBUGS, domain and species). See S1 P for principal component contributions.

predictions (272/361 agreed), but at the same time, many feel that methods need to be improved (241/361), and few people agreed uncertainty was adequately taken into account (48/361) and very few agree with all the ways SDM models and software are used (8/361). These results suggest that while SDM software is widely used and useful, it requires further advancement/development.

Additional results from the survey may be found in SI R-SI W. Further, all responses are available in SI data.

SDM publication networks

A total of 15,536 papers were returned from the Web of Science search, written by various combinations of 34,779 unique author names. Of the papers returned, we found (1) 21 developers, (2) 13 super-collaborators, (3) 24 ingroup collaborators, (4) 436 single-collaborators and (5) 34,285 other authors (Fig. 3). Those people involved in either SDM software development or the sophisticated cross-analysis of different SDM software packages are among the most highly connected and published scientists within the domain. Developers, super-collaborators, intergroup collaborators and single collaborators (categories 1–4) have on average significantly more publications ($t = 10.6$, d.f. = 494, $P < 0.001$, 95% CI = 3.3–4.8) and co-authors (t = 14.2, d.f. = 494, $P < 0.001$, 95% CI = 12.5–16.8) within the SDM domain than scientists who have not published directly with develop-

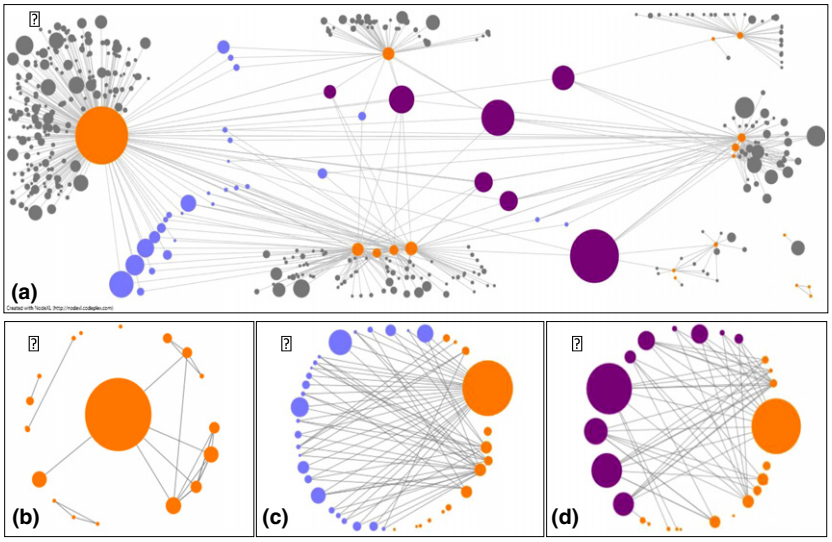

Figure 3 Citation networks highlighting the interactions between software developers and their co-authors in Species Distribution Modelling (SDM). Each circle represents an author. Authors are grouped by software and identified by colour for author category; orange: software developers (category 1); purple: super-collaborators (category 2); blue: intergroup collaborators (category 3); grey: single-collaborators (category 4). (a) Network of all authors who have developed software or co-authored a paper with a developer (authors who are not connected to a developer are not shown). (b) Network among software developers. Each software had a different number of 'developers' (as we classed them for this exercise; MAXENT (n developers = 3), GARP (n = 1), BIOMAPPER (n = 1), CANOCO (n = 1), DOMAIN (n = 3), SPECIES (n = 4), BIOMOD (n = 1), OPENMODELLER (n = 2) and MODECO (n = 2). (c) Network among intergroup collaborators and software developers. (d) Network among super-collaborators and software developers. The size of the node is relative to the total number of papers the author has published in SDM and does not represent solely the number of links shown. See Results for definitions of each author category. There appears to be a disinclination for software developers (category 1) to publish together (inset b). Connections between developers are generally mediated by intergroup and super-collaborators (inset c, d).

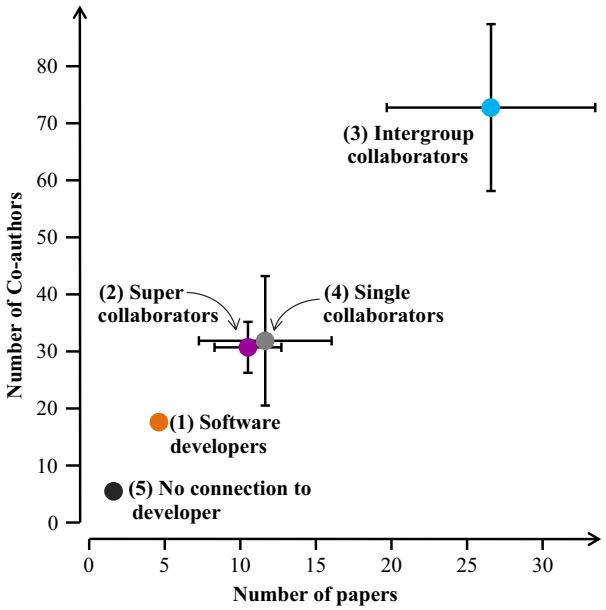

Figure 4 Does being connected to a software developer make you better connected and more published? A literature search returned a total of 34,779 authors within the SDM domain, and these were divided into five categories, where categories 1–4 correspond to categories 1–4 in Figure 3, and category 5 represents those authors that are not directly connected to an SDM software developer. The number of papers and co-authors per author were calculated. Authors who are either a developer or directly connected to a developer (categories 1–4) have on average more papers published and more co-authors than those who are not connected to a developer (category 5).

ers. Super-collaborators have on average significantly more papers (ANOVA, F = 1194, d.f. = 4, 34774, $P < 0.001$) and collaborators ($F = 659$, d.f. = 4, 34774, $P < 0.001$) than any of the other author categories (Fig. 4). There appear to be few instances of software developers publishing together (Fig. 3b). Connections between developers are generally mediated by intergroup and super-collaborators (Fig. 3c & d). Given this, software developers in the SDM community can exert a great deal of influence over the field, because the methods they implement in their software are the methods that (in general) are used across the field and this inevitably has repercussions for users.

DISCUSSION

SDM is a growing field that is used increasingly in the policy domain and that relies heavily on software. Here, we have presented the first survey of SDM software use. As with any survey, there is a degree of ambiguity both on the side of the respondents who have to interpret what the questions mean and in the interpretation of the results. Further, despite contacting 130 members of the SDM community to disseminate the survey, some degree of bias and non-representation is likely to remain as it is not feasible to survey the whole field. We know that 34,779 authors have published in the SDM domain, but our survey represents only 364 members of the community. Two survey power analysis tools suggest that for a population of 34,779, with a confidence interval of 95% and an error tolerance of 5%, 385–395 respondents would be required to be representative (SI), and our responses are just short of this. Even if we exceeded the required sample size, issues of bias would remain as we did not implement stratified sampling.

The expansion of SDM has undoubtedly been influenced by the availability of software and data. However, software is rarely discussed in terms other than which methods are available within software, with some recognition of the importance of functionality for visualization and summarizing of SDM analyses (Elith & Leathwick, 2009; Franklin 2010). In passing, a wide variety of descriptors are attached to software – such as 'popular' (Merow *et al.*, 2013), 'widely used' (Mesgaran *et al.*, 2014), 'easy to use' (Taylor & Kumar, 2013), 'user friendly' (Guo & Liu, 2010), 'useable' (de Souza Mun˜oz *et al.*, 2011) and 'de facto standard' (Dormann *et al.*, 2012) – yet there are few qualifiers for these descriptions. Our survey has shown that most users are currently willing to continue using their current tools given the options available. Respondents appear to show high fidelity to their tools (SI O) despite many claiming that methods need to be improved, with few agreeing that uncertainty is adequately taken into account or that SDM captures ecological reality (SI Q).

The two software that seem to be most successful within this field are R and MAXENT, with the most positive ratings in terms of software usefulness, learning to use the software, system capabilities and overall satisfaction. It is interesting that the two 'best' software lie at either end of the use-complexity continuum: R, syntax driven, with a vast array of methods are available, while MAXENT implements one method (albeit with adjustable settings) driven from a graphical user interface (GUI) based on point-and-click running (although the option to programmatically use MAXENT via other platforms, such as R and batch, does exist). Users' understanding of software defaults and algorithms have a variety of consequences for making scientific inferences. If methodological or implementation details are in anyway concealed or inaccessible, 'black-box' software is created. Alternatively, users may actively or inadvertently ignore the details and nuances of their models (voluntary or involuntary black-boxing). Concerns that click software does not facilitate an in-depth understanding of the methods, models and algorithms used, with a greater tendency towards black-boxing should be mirrored in syntax software that have preset defaults or that are not open source. For example, many R libraries have preset defaults which many users apply with little thought. Thus, having 'harder to use' syntax driven software does not necessarily mean that users have a more in-depth understanding of the methods, models and algorithms.

Broadly speaking, syntax software offers more potential for users to modify, tune and/or extend methods than click software. This potential control is rarely presented in click soft-

ware where workflows are more constrained. Moreover, a model that results from a series of user interactions in a GUI may not always be retrievable (although this is not the case in MAXENT). Whereas, a series of syntax operations are unavoidably encoded to give a retrievable model, leading to better reproducibility of work.

Our analysis using the classification of 'click' (GUI) and syntax-based software presents a number of important questions on whether users are building/modifying models with in-depth knowledge or using defaults. This is not only important to developers, but also to the training scientists receive, peer review and open science. To further investigate this issue, more explicit investigations pertaining to the degree of model, method and algorithm, understanding and implementation would be required to disentangle the interacting factors of software functionality and user expertise. It should be noted that the distinction between click and syntax software is largely artificial as software exists on a spectrum, as such rather than using click and syntax software as a proxy for whether users are building/modifying models with in-depth knowledge (syntax) or just using a GUI/parameter defaults (click) explicit questions pertaining to the degree of model, method and algorithm, understanding and implementation should be asked in future.

Users of scientific software in general put their trust in the software to be correct and in the developer to have implemented the methodology correctly. This may not always be the case as some software are coded and implemented differently from what is reported in peer-reviewed literature. Software may not be correct for a variety of reasons not associated with method implementation, for example programming errors are common with estimates suggesting 1–10 errors per thousand lines of code, numerical errors as a result of rounding are common and software may behave differently when run on different platforms (Ince *et al.*, 2012). Indeed, errors in software within other fields have led to 'nightmare' situations of flawed results which in turn have resulted in retractions of work (Miller, 2006). Sometimes as software is developed, the implementation changes from what is reported in the original publication, with each subsequent user adding to the code base, but the details of these new changes are rarely reported or published (Merali, 2010). Further, if code is not provided (in an understandable format), scientific rigour may be impacted, because even the best 'descriptors' of software function/methodology in a paper can suffer from ambiguity. While these are general problems applicable to all software, the issues of disparity between what is reported in publication and what the software is (or has become) is problematic for SDM as most respondents to our survey said that they rely on the original papers to understand the software.

From a methodology stand point, the type of software being used (click/syntax) is of little importance because the same correlative methods are being implemented by both click and syntax software. Very few respondents cited methods other than those listed as options (all correlative), those that did, cited correlative methods. This was unexpected because one of the benefits of syntax software is that it can implement mechanistic (rather than correlative) methods that none of the click software mentioned in the survey currently implements. This could be indicative of a dependency, of the SDM field as a whole, on simple correlative approaches. It should be noted, however, that while none of the respondents mentioned mechanistic modelling, this may be because mechanistic modellers may consider themselves apart from the SDM domain. Alternatively, respondents may have considered these types of models outside the scope of the survey. It has been acknowledged that existing SDM methods are mostly correlative (Booth *et al.*, 2014). This reliance on correlative methods is worrying given the growing concerns voiced in the literature and by our survey respondents on the failure of such methods to fully address the complexities of SDM. Attempts have been made to generate and use mechanistic SDM software, for example CLIMEX (Sutherst & Maywald, 2005; Lozier & Mills, 2011). A question that must be asked is: Are correlative methods more prevalent than mechanistic methods because more interesting and useful conclusions can be drawn from using these methods under conditions of limited information and system understanding? Or, is this preference driven by the fact that correlative methods can be relatively easily implemented in software as automated algorithms?

Software developers are of course influential on the SDM community because modelling methods would not otherwise be available. This reliance on developers inevitably has many repercussions for users and the science they produce. Our co-authoring study shows that developers have a great deal of influence over the SDM field and authors who are not directly connected to at least one developer have significantly fewer publications and collaborators than those who are connected to a software developer. This highlights two points (1) the central role of SDM software to the field and (2) the level of power that developers could potentially have to influence the fields' development via publication/co-authoring networks. Of course, then the question must be raised as to whether developers actually have a responsibility to the field's development beyond implementing methods into an accessible software format. Our results further suggest that super-collaborators and intergroup collaborators play a key role in bridging the user–developer divide. We did not investigate any form of modularity in the co-authoring network (tendency to preferentially co-author within a subset of co-authors) beyond these simple classifications. However, given the assumption that co-authorship with a developer concerns the software, without super- and intergroup collaborators, there is the suggestion that the SDM community would be fragmented around particular software with a reduced level of examination, validation and comparison of the different software and methods.

A particular interest of the survey was to find out what software is being used and for what kinds of tasks. Not all SDM methods and algorithms are available in all software, and so,

software selection impacts what kinds of analyses are being used and how they are being used. However, this topic is rarely discussed. Yet, at the same time, statements are made regarding particular methodological issues, user expertise and skills, software interfaces and scientific quality (e.g. Elith & Leathwick, 2009; Thuiller *et al.*, 2009; Richardson, 2012; Merow *et al.*, 2013) which explicitly or implicitly refer to the computational or statistical literacy of users (Elith & Leathwick, 2009; Thuiller *et al.*, 2009; Dormann *et al.*, 2012), or the interactions of users and methods through the software interface (Pimm, 2008; Joppa *et al.*, 2013; Merow *et al.*, 2013). Our survey supports fruitful discussions on all these topics.

This survey provides a variety of benefits for different parts of our community. For the whole community: a stimulus for discussion on how what kinds of systems are being studied, and how SDM research directions interact with software characteristics. For developers: feedback on software from users and information on those users and the user community which justify their development work and approaches taken, from operating systems, to number of species, to user characteristics. For users: greater awareness of what people are actually doing, and prevailing practices, what software is available and being used, where and for what kinds of data. For research purposes: to stimulate new research questions and reflections on SDM.

ACKNOWLEDGEMENTS

We would like to thank all those who gave their time to take part in the survey and interviews. We would also like to thank Microsoft Research for funding this project. The authors thank Jane Elith, Townsend Peterson and two anonymous referees whose comments and suggestions greatly improved this manuscript.

REFERENCES

Aarts, G., Fieberg, J. & Matthiopoulos, J. (2012) Comparative interpretation of count, presence–absence and point methods for species distribution models. *Methods in Ecology and Evolution*, **3**, 177–187.

Araújo, M. & Peterson, A.T. (2012) Uses and misuses of bioclimatic envelope modelling. *Ecology*, **93**, 1527–1539.

Austin, M. (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1–19.

Bellard, C., Thuiller, W., Leroy, B., Genovesi, P., Bakkenes, M. & Courchamp, F. (2013) Will Climate change promote future invasions? *Global Change Biology*, **19**, 3740–3748.

Booth, T.H., Nix, H.A., Busby, J.R. & Hutchinson, M.F. (2014) BIOCLIM: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions*, **20**, 1–9.

Dawson, T.P., Jackson, S.T., House, J.I., Prentice, I.C. & Mace, G.M. (2011) Beyond predictions: biodiversity conservation in a changing climate. *Science*, **332**, 53–59.

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B. & Singer, A. (2012) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, **39**, 2119–2131.

Elith, J. & Leathwick, J.R. (2009) Species distribution models: ecological explanation and prediction across time and space. *Annual Review of Ecology, Evolution and Systematics*, **40**, 677–697.

Esselman, P.C. & Allan, J.D. (2011) Application of species distribution models and conservation planning software to the design of a reserve network for the riverine fishes of north-eastern Mesoamerica. *Freshwater Biology*, **56**, 71–88.

Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N.G., Lehemann, A. & Zimmermann, N.E. (2006) Using niche-based models to improve sampling of rare species. *Conservation Biology*, **20**, 501–511.

Guisan, A., Reid, T., Baumgartner, J.B. *et al.* (2013) Predicting species distributions for conservation decisions. *Ecology Letters*, **16**, 1424–1435.

Guo, Q. & Liu, Y. (2010) ModEco: an integrated software package for ecological niche modeling. *Ecography*, **33**, 637–642.

Hof, C., Araujo, M.B., Jetz, W. & Rahbek, C. (2011) Additive threats from pathogens, climate and land-use change for global amphibian diversity. *Nature*, **480**, 516–521.

Ince, D.C., Hatton, L. & Graham-Cumming, J. (2012) The case for open computer programs. *Nature*, **428**, 485–488.

Joppa, L.N., McInerny, G., Harper, R., Salido, L., Takeda, K., O'Hara, K., Gavaghan, D. & Emmott, S. (2013) Troubling trends in scientific software use. *Science*, **340**, 814–815.

Kozak, K.H. & Wiens, J.J. (2006) Does niche conservatism promote speciation? A case study in North American Salamanders *Evolution*, **60**, 2604–2621.

Lozier, J.D. & Mills, N.J. (2011) Predicting the potential invasive range of light brown apple moth (*Epiphyas postvittana*) using biologically informed and correlative species distribution models. *Biological Invasions*, **13**, 2409–2421.

McInerny, G.J. & Etienne, R.S. (2012) Ditch the niche - is the niche a useful concept in ecology or species distribution modelling? *Journal of Biogeography*, **39**, 2096–2102.

Merali, Z. (2010) Why scientific programming does not compute. *Nature*, **467**, 775–777.

Merow, C., Smith, M.J. & Silander, J.A. (2013) A practical guide to MaxEnt for modelling species' distributions: what it does, and why inputs and settings matter. *Ecography*, **36**, 1058–1069.

Mesgaran, M.B., Cousens, R.D. & Webber, B.L. (2014) Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Diversity and Distributions*, **20**, 1147–1159.

Miller, G. (2006) A scientists nightmare: software problem leads to five retractions. *Science*, **314**, 1856–1857.

Moffett, A., Shackleford, N. & Sarker, S. (2007) Malaria in Africa: vector species' niche models and relative risk maps. *PLoSOne*, **9**, 1–18.

Pereira, M., Sergurado, P. & Neves, N. (2011) Using spatial network structure in landscape management: a case study with pond turtles. *Landscape and Urban Planning*, **100**, 67–76.

Peterson, A.T. & Soberon, J. (2012) Species distribution modelling and ecological niche modelling: getting the concepts right. *Brazilian Journal of Nature Conservation*, **10**, 102–107.

Pimm, S.L. (2008) Biodiversity: climate change or habitat loss – which will kill more species? *Current Biology*, **18**, R117–R119.

R Development Core Team (2006) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Available at: http://www.R-project.org/, https://stat.ethz.ch/R-manual/R-devel/library/stats/html/princomp.html

Reed, R.N. & Rodda, G.H. (2009) Giant constrictors: Biological and management profiles and an establishment risk assessment for nine large species of Pythons, Anacondas, and the Boa Constrictor, U.S. Geological Survey Open-File Report 2009–1202. Available at: http://pubs.usgs.gov/of/2009/1202/pdf/OF09-1202.pdf (accessed 22 October 2013).

Richardson, D.M. (2012) Conservation biogeography: what's hot and what's not? *Diversity and Distributions*, **18**, 319–322.

Soberón, J. (2014) Commentary on Ditch, Stitch and Pitch: the niche is here to stay. *Journal of Biogeography*, **41**, 414–417.

de Souza Mun˜oz, 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.

Stepan, N.L. (2001) *Picturing tropical nature*, pp. 38–39. Reaktion Books, London.

Sutherst, R.W. & Maywald, G.F. (2005) A climate-model of the red imported fire ant, *Solenopsis invicta* Buren (Hymenoptera: Formicidae): implications for invasion of new regions, particularly, *Oceania*. *Environmental Entomology*, **34**, 317–335.

Svenning, J.C., Flojgard, C., Marske, K.A., Nogues-Bravo, D. & Normand, S. (2011) Applications of species distribution modelling to paleobiology. *Quaternary Science Reviews*, **30**, 2930–2947.

Taylor, S. & Kumar, L. (2013) Potential distribution of an invasive species under climate change scenarios using CLIMEX and soil drainage: a case study of *Lantana camara* L. in Queensland, Australia. *Journal of Environmental Management*, **114**, 414–422.

Thuiller, W., Lafourcade, B., Engler, R. & Araujo, M. (2009) BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.

Warren, D. (2012) In defence of 'niche modeling'. *Trends in Ecology & Evolution*, **27**, 497–500.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Data S1 Details on the pre-survey interviewees, details of power analysis used, and full survey results SI A- SI W.

Data S2 Survey response data.

Data S3 A copy of the survey questions and answer choices that were presented online to respondents.

BIOSKETCH

The Computational Sciences (L.N.J., S.E., S.E.A., – G.M., L.S. at time of project) and Human Experience and Design Group (K.O'H, R.H.) groups at Microsoft Research are interested in fundamental questions at the intersection of science and technology. Members of the groups pursue new methods and systems for analysing ecological data and designs for improving how users interact with computational systems.

Author contributions: L.N.J., S.E. and G.M conceived the ideas; K.O'H., R.H., L.S., L.N.J. and G.M. collected the data; S.E.A., L.N.J, L.S. and G.M. analysed the data; and S.E.A. and L.N.J. led the writing.

Editor: Jane Elith