Challenges and perspectives for species distribution modelling in the neotropics

Luciana H. Y. Kamino1,*,†, João Renato Stehmann1, Silvana Amaral2, Paulo De Marco Jr1, Thiago F. Range1,†, Marinez F. de Siqueira4, Renato De Giovanni5 and Joaquín Hortals6,*,†

1Depto Botânica, Instituto Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
2Divisão de Processamento de Imagens, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Brazil
3Depto Ecologia, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiás, Brazil
4Unidade de Botânica Sistemática, Instituto de Pesquisa Jardim Botânico do Rio de Janeiro, Rio de Janeiro, Brazil
5Centro de Referência em Informação Ambiental, Campinas, Brazil
6Depto Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales (CSIC), Madrid, Spain
7Authors for correspondence (lucianakamino@gmail.com; jhortal@mncn.csic.es).

The workshop ‘Species distribution models: applications, challenges and perspectives’ held at Belo Horizonte (Brazil), 29–30 August 2011, aimed to review the state-of-the-art in species distribution modelling (SDM) in the neotropical realm. It brought together researchers in ecology, evolution, biogeography and conservation, with different backgrounds and research interests. The application of SDM in the megadiverse neotropics—where data on species occurrences are scarce—presents several challenges, involving acknowledging the limitations imposed by data quality, including surveys as an integral part of SDM studies, and designing the analyses in accordance with the question investigated. Specific solutions were discussed, and a code of good practice in SDM studies and related field surveys was drafted.

Keywords: Wallacean shortfall; rare species; habitat suitability mapping; biodiversity databases; bias; experimental design

1. INTRODUCTION

The geographical distribution of most extant species is totally or partially unknown—the so-called Wallacean shortfall [1]. Species distribution models (SDMs) are commonly used to overcome this drawback. SDMs are tools that relate known species’ occurrences with a series of predictors; these relationships are projected into the geographical space to estimate current distributions or range dynamics [2]. Their wide applicability in theoretical and applied studies has made SDM one of the hottest and most rapidly growing fields of ecology [3]. However, their application is often questioned owing to problems in data quality and conceptual formulation [4–6]; these are common in the neotropics, where the use of SDM faces different challenges than in temperate areas [7,8].

The workshop ‘Species distribution models: applications, challenges and perspectives’ hosted by the Federal University of Minas Gerais (Belo Horizonte, Brazil, 29–30 August 2011) aimed to review the state-of-the-art in the application of SDM to research in evolution, biogeography and conservation in the neotropics. The discussions covered key aspects of the use of SDM in tropical and subtropical biomes, including gaps in knowledge and common mistakes, experimental design, and the particular challenges faced by SDM in these regions, which include working on a large number of rare species with few occurrence records and massive spatial biases [7]. In this report, we synthesize the conclusions of this debate, which include several solutions and recommendations about good practice in field surveys.

2. CONSTRUCTING HYPOTHESES AND MODELLING SPECIES DISTRIBUTIONS FROM FEW DATA IN MEGADIVERSE AREAS

The use of SDM should be based on adequate knowledge of their theoretical basis and practical applicability. Depending on the type of distributional data, predictors and modelling technique used, SDM results can correspond to either potential, actual or realized distributions [5,6]; each of these aspects has a different conceptual meaning and allows the answering of different kinds of questions. Unfortunately, studies applying SDM just for the sake of using them are increasingly common. Many SDM studies do not determine their exact purpose with clarity, thus failing to choose the most adequate technique and study design. To overcome this, the analytical design of SDM applications should be defined based on the purpose of the study and the available data. From the conceptual point of view, such design shall account for the following questions. What is the purpose of the analysis? Which aspect of the species distribution needs to be described? What will SDM results be used for? How good are data on species’ occurrences? Are additional records necessary? Is there reliable data on species’ absence? Is it necessary for the question at hand? if that is the case, which kind of absence data is needed? (see [3]), Which are the most adequate predictors? and which SDM technique is the most appropriate?

(a) Data quality

Data on species’ occurrences generally come from former surveys and are stored in biodiversity databases and/or natural history collections. These data typically offer an incomplete picture of species distributions, presenting spatial, temporal and taxonomic biases [1,9]. For most neotropical species, the data available are limited and plagued with imprecise geographical locations and inappropriate metadata, as exemplified by the database of the Brazilian National Centre for Plant Conservation (CNCFlora; http://cncflora.net/), which holds 1 917 217 records for 85 007 species’ names. Leaving apart eventual taxonomic misidentifications, 72 per cent of these species names are represented by less than 10 records and 37 per cent by just a single occurrence. Using such data to calibrate SDM would result in a high frequency of errors, in particular, for rare species.
Two main strategies can help to improve data quality. First, assess the reliability of each database record using automatic filters and rigorous revisions [9], discarding unreliable records but not actual occurrences that may appear to be outliers. Here, the quality and geographical precision of old records—which may be quite useful to assess temporal range shifts [8]—can be improved by reconstructing the route followed by the expeditions based on the collectors’ field notes, contrasting the collection dates in the original vouchers and assessing the field expertise of the recorders. Second, when data useful for the analysis are scarce, it is necessary to conduct additional field surveys.

(b) Predictor variables

Choosing appropriate predictors—e.g. environmental gradients, land cover, species interactions, historical or contingent factors—is of key importance because they are bound to particular hypotheses on the mechanisms regulating species distributions. Previous knowledge of the species’ ecology and requirements—and the history of the studied region—can inform such choice, improving the final quality of the work. Most studies discussed during the workshop used climatic, topographical and less frequently, soil variables. However, much of this work could have made use of remote sensing variables [10] (S. Amaral 2011, unpublished data). There is a diverse supply of satellite imagery products at different resolutions, including vegetation indices, leaf area indices, fraction of photosynthetically active radiation, soil water content, land cover, or high-resolution climate data. Much remote sensing information is available on the Internet, for example, publicly available 1 km resolution databases for the whole of Brazil and the Amazonian rainforest are hosted by the Brazilian National Institute for Spatial Research (INPE; http://www.dpi.inpe.br/Ambdata/). However, this information should be used with caution, taking into account spatial resolution, processing and georeferencing quality and interpretability. Remote sensing variables will be useful predictors only when the particular variable and its resolution are appropriate to describe a well-defined ecological phenomenon.

Collinearity among variables can lead to the misidentification of the most relevant predictors, producing inaccurate SDM predictions, in particular, at low sampling effort levels. The actual impact of this problem is unknown because the effect of such multi-collinearity is seldom assessed. To account for this problem, P. De Marco Jr, T. F. Rangel and M. F. de Siqueira suggested using the most explanatory axes extracted from a principal components analysis as predictors. An alternative would be using a random resampling procedure to identify the most consistent predictor(s) regardless of the calibration data used. Furthermore, it is also necessary to determine whether SDM results can be compromised by the spatial structure (i.e. spatial autocorrelation) in the recorded species’ occurrences. If such autocorrelation is an artefact of the distribution of the surveys, then it should be removed from the data (using a resampling strategy; N. M. Torres & P. De Marco Jr 2011, unpublished data). But if it corresponds to the actual spatial structure of the species’ range, then no data should be discarded; if a spatially explicit process is affecting the species’ distribution, then the spatial location of the samples should be included within the list of predictors.

Not all factors affecting species distributions can be adequately formulated as predictors in SDM. Variables affecting the spatial dynamics of the species such as connectivity or fragmentation cannot be used in most studies because SDM techniques produce static models—unable to make predictions on the species’ population dynamics. Modelling habitat suitability and spatial population dynamics altogether is a difficult task [11]; an alternative can be to use SDM to predict the species’ potential distribution to then filter its probability of occurrence according to the connectivity and/or characteristics of each habitat, and the temporal resolution of the analysis.

(c) Species distribution modelling technique

Some basic knowledge on how SDMs actually work may improve their use. SDMs are algorithms that identify a mathematical or a logical function linking species’ occurrences and a set of predictors, and the model described by such function. The choice of the SDM technique may be particularly difficult, because it is affected by data quality, and will in turn affect the type of distribution that is modelled [5]. T. F. Rangel (2011, unpublished data) presented a conceptual experiment comparing the performance SDM techniques in modelling the realized distribution of a virtual species using the known environmental determinants as predictors. Here, Bioclim, general additive models, and Genetic Algorithm for Rule-set Production (GARP) performed the worst and Euclidean environmental distance, multivariate adaptive regression splines, and neural networks were the best ones. However, SDM techniques cannot be chosen based solely on this kind of comparison, because they differ in the complexity—and thus number of parameters—of their solutions, their link to ecological mechanisms, their need for data on absences or their spatial and temporal transferability. Models are incomplete descriptions of reality, and their adequacy depends on whether the studied question needs more emphasis on realism, generality or precision; improving any of these aspects implies sacrificing the other two [12]. Rather than choosing the technique that best fits the data, we recommend selecting the most adequate for answering the specific question being investigated; most times, simpler techniques with a solid theoretical background may be preferable to complex techniques that overfit the data [5].

3. NEOTROPICAL PERSPECTIVES TO THE STUDY OF SPECIES DISTRIBUTIONS

During the workshop, it became clear that brand new solutions to describe neotropical diversity based on scarce data are being developed. M. F. de Siqueira (2011, unpublished data) presented a protocol to study rare species, which are typically known from a reduced number of localities. Here, SDMs are not used to generate hypotheses on the species’ distribution, but to identify potentially suitable areas that are unexplored (see also [13]). Distance-based SDM techniques calibrated with few data points (sometimes as few as one) are used to plan additional surveys in areas of high,
medium and low environmental similarity, in order to find additional populations and determine the environmental tolerance of the species.

The scarcity of distributional data and the often low quality of their metadata lead to the conclusion that following a code of good practice would maximize the effectiveness of field surveys. Such a code should include the following recommendations: design surveys to cover the environmental and geographical variability of the studied region and avoid accessibility bias; record in the field the geographical location, altitude and habitat (because of rapid habitat transformation); gather information on the accession route to each site, record as much ecological data as possible, describe the survey methods and sampling effort devoted to each collection and record confirmed absences—because they can inform on non-environmental effects on species’ distributions [3,5]. R. De Giovanni (2011, unpublished data) presented the SDM strategy planned for a new biodiversity information system for the Brazilian flora (INCT virtual herbarium of plants and fungi; http://inct.florabrasil.net/), which will provide tools to detect quality issues in herbarium records and help in designing new surveys. This will create feedback loops of increasing data quality and quantity to ultimately produce SDM that may be good enough for conservation [9].

We also observed a disparity in the formation of the recorders and users of biodiversity data. While students with a background in systematics hold a critical view about data quality and the associated drawbacks, many SDM users have a background in ecology or bioinformatics and little knowledge of the limitations of biodiversity data. We thus call for bridging in the gap between these two schools, using the expertise of taxonomists to improve the quality of SDM studies. Researchers should follow a critical approach to the design of SDM studies, recognizing their uncertainties but also the importance of formulating research questions in an appropriate way, avoiding the fascination for numbers and colour maps that may result in forgetting the actual biological significance of the results.

The workshop was organized by L.H.Y.K. and J.R.S. We thank all attendees for the instructive discussions held during the plenary and poster sessions and their feedback about the adequacy of particular methods and/or practical solutions. L.H.Y.K. holds a UFMG (Pró-Reitoria de Pós-Graduação; Instituto Ciências Biológicas; Pós-Graduação em Biologia Vegetal; Genética; Zoologia e Ecologia, Conservação e Manejo de Vida Silvestre), Conselho Regional de Biologia, Viamundi and CAPES grant and J.H. a CNPq Visiting Researcher grant.