International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

Species distribution modelling and open source GIS: why are they still so loosely connected?

Anne Ghisla [a], **Duccio Rocchini**[a], **Markus Neteler**[a], **Michael Förster**[b], **Birgit Kleinschmit**[b]

[a] *Fondazione Edmund Mach, Research and Innovation Centre, Department of Biodiversity and Molecular Ecology, GIS and Remote Sensing Unit, Via E. Mach 1, 38010, S. Michele all'Adige (TN), Italy (anne.ghisla@fmach.it, duccio.rocchini@fmach.it, markus.neteler@fmach.it)*

[b] *Technische Universität Berlin - Fakultät VI, Institut für Landschaftsarchitektur und Umweltplanung, Fachgebiet Geoinformation in der Umweltplanung, Sekr. EB 5 - Raum 235a, Straße des 17. Juni 145 , D-10623 Berlin (michael.foerster@tu-berlin.de, birgit.kleinschmit@tu-berlin.de)*

Abstract: Species Distribution Models (SDMs) are correlative models that use environmental and/or geographical information to explain patterns of species occurrences. Those models are being used in various fields including climate change, invasive species research, evolutionary biology and epidemiology. Thanks to the availability of increasing computational resources, new methods continue to be developed. However, software packages that include the SDM algorithms usually focus on one or few methods, and have different degrees of integration with other geographical and statistical software. Specifically, SDM implementations are often standalone programs developed by university laboratories, either as extensions to statistical software. In few cases they are written as extensions for the most common proprietary Geographic Information System (GIS) products, despite the strong geographical component present in the data. On the other hand, open source GIS software has loose connection with SDM implementations, and usually requires more effort to build a complete and well connected software stack. This paper investigate on the possible causes of the present separation of open source GIS and Species Distribution Modelling and on the benefits of a closer integration, and lists a selection of candidates for future joint development. A further step would be the adoption of open source principles in the implementation process of SDM. This will enable a peer-review mechanism on the computational code, that will strongly reduce the risk of attaining biased results due to inaccurate implementations.

Keywords: Species Distribution Modelling; GIS; open source; integration

1 INTRODUCTION

An increasing range of applied and theoretical questions regarding species' occurrences are taking benefit from the application of Species Distribution Models (SDMs). SDMs are currently being used in a variety of fields including evolutionary biology, where they are used to study topics such as speciation or hybrid zones (Kozak et al. [2008]) and epidemiology, where they are used to predict the spread of disease (e.g., Peterson et al. [2002]). As a result of these diverse uses of SDMs that have been spurred on by advances in

GIS (Foody [2008]) and data analysis (Breiman [2001b]), new and more complex modelling methods continue to be implemented and compared with existing ones (Elith et al. [2006]; Phillips et al. [2006]; Jolma et al. [2012]).

However, the expertise needed to write robust and usable software for species potential distribution encompasses mathemathics, statistics, computer science, ecology and geography. Ideally, these applications should be able to deal with a range of tasks such as transforming between different geospatial reference systems, handling geospatial data in different scales and extents, reading and writing geospatial data in different file formats, and facilitating data visualization. The final software package should also ideally offer pre-analysis and post-processing tools, providing support for a range of protocols and data standards for sharing and retrieving occurrence and environmental data (Muñoz et al. [2009]). To date, unfortunately, most software development has been carried out in isolation and by small teams of researchers, producing separate software packages that are targeted to a single algorithm. This is the case for Domain (Carpenter et al. [1993]) and Maxent (Phillips et al. [2006]). There are drawbacks in having a different software package for each algorithm, if interoperability is disregarded. In particular, users need to learn multiple software applications to use different algorithms. In addition, most implementations have little geospatial functionality (e.g. they are unaware of projections and common spatial data formats) and are released as binaries only, so that the implementation is not available for review.

On the contrary, open source software like R (R Development Core Team [2011]), Quantum GIS (`http://www.qgis.org`) and GRASS GIS (Neteler and Mitasova [2008]) are developed by a wide community of researchers and users from different education backgrounds, who extend, check and adapt the program to their specific needs. The open source development paradigm requires the source code to be maintained in a public repository, where each change is digitally tracked and subject to public peer review of code style, functionality, and quality. In a scientific context, the reproducibility of results and quality assessment of methods is greatly facilitated since full access to the underlying algorithms is guaranteed (Neteler et al. [2012]).

In the remainder of the paper, some of the most common software packages for SDM are examined, with focus on the connection with GIS and/or open source development. Possible causes of the separation between them are advanced, together with suggestions towards the adoption of common open source development practices in ecological research.

2 SDM SOFTWARE: AN INTEROPERABILITY-CENTERED REVIEW

The increasing number of software packages implementing SDMs complicates the selection of the most appropriate one for a specific use case; unfortunately, the criteria and advice that would enable informed choice of method are currently scattered throughout the literature, and are incomplete (Elith and Graham [2009]). In this section, common SDM implementations are examined, with particular attention to the present and potential interoperability with geostatistical software.

2.1 Maximum Entropy

The Maximum Entropy (**Maxent**) model is a general-purpose method for making predictions or inferences from incomplete information, that minimizes the relative entropy between two probability densities (one estimated from the presence data and one, from the landscape) defined in covariate space (Phillips et al. [2006]). Maxent is

also a Java implementation of this model, distributed by Princeton University (`http://www.cs.princeton.edu/~schapire/maxent/`) since 2004, free of charge for research and educational purposes. It is a general-purpose machine learning method with a simple and precise mathematical formulation, designed to accept presence-only data as input. It has been the first implementation of the Maxent model, and gained popularity among ecologists thanks to its robustness and usability . It has limited geostatistical functionality, so pre-processing of the data within a GIS is required, as well as post-processing of the outputs. The users' community is working on the connection with GIS by providing GRASS GIS addon modules `r.maxent.lambdas` and `r.out.maxent_swd` (see the full list at `http://grass.osgeo.org/wiki/Addons`), but unless Maxent source code becomes publicly available, it is unlikely that more developers will be able to participate to its development.

2.2 Random Forest

A RandomForest (Breiman [2001a]) is a collection of Classification And Regression Trees (CART) that are used to predict via a consensus or voting mechanism, where each tree is grown at least partially at random. A large number of large trees are grown and results can be remarkably accurate. A noticeable property is that random forests do not overfit as more trees are added.
There are several implementations of random forests, one being the proprietary software **RandomForests** (`http://www.salford-systems.com/products/randomforests/overview.html`). The original source code is licensed under the terms of General Public License (GPL, see `http://www.gnu.org`) , which allowed the reimplementation as open source variants in R (among others, **randomForest** package from Liaw and Wiener [2002]), Matlab, Python, Ruby, C++ and C# programming languages. This variety makes easier to integrate random forest computation in an existing script or software stack without forcing the use of a specific language besides R.

2.3 Genetic algorithms

LifeMapper (`http://lifemapper.org/`) is a web service that aims to achieve the construction and maintenance of an extensive predicted species habitat map archive, and the exposure of spatial data and analysis services based on this archive. Lifemapper implements the openModeller species niche modeling platform, and uses open source software and standards extensively. The user is able to generate distribution models and retrieve generated data from a web interface (Stockwell et al. [2006]).

openModeller (Muñoz et al. [2009]) is an open source modelling framework that currently implements two different implementations of GARP (for individual runs and GARP Best Subsets [4]), as well as Artificial Neural Networks (ANN), AquaMaps, Bioclim , Ecological Niche Factor Analysis (ENFA), Climate Space Model (CSM), a generic distance-based algorithm (Environmental Distance), Envelope Scores, Random Forest and Support Vector Machines (SVM). This wide choice within the same software package allows the comparison of different models on the same dataset without having to switch among software packages, that have specific data preparation procedures and requirements that can interfere with the comparison of the models.
The openModeller framework clearly separates the core functionality (model implementations) from the user interface and data input/output. This choice allows to add new techniques exposing a consistent interface. It makes use of open standards and libraries, is designed for integration with different applications such as openModellerDesktop and QuantumGIS. The input and output data are in full-featured GIS formats, ready to be merged with other spatial data.

2.4 Generalised Linear Models

The R package **geoRglm** (Christensen and Ribeiro Jr. [2002]) includes functions for inference in generalised linear spatial models. The posterior and predictive inference is based on Markov chain Monte Carlo methods. geoRglm is based on geoR, a R package containing geostatistical functions (Ribeiro Jr and Diggle [2001]). The development of the latter started in 2001, and is one of the native implementations of GIS functionality within R. R contains several spatial classes (see CRAN Spatial View `http://cran.r-project.org/web/views/Spatial.html`), that implement GIS-like functionality: the set of sp classes (Pebesma and Bivand [2005]) is becoming the standard implementation of GIS capabilities in R and is well connected with GIS data formats via rgdal (Keitt et al. [2010]), but alternate implementations like PBSmapping and geoR are equally valid and accepted.
Therefore, it is appropriate to state that R has developed spatial functionalities independently from GIS software since years, so that the implementations have reached a good degree of stability and can easily be linked to GIS data formats. These characteristics lead to forecast that R internal geodata formats will not be superseded by the integration with external GIS packages.

CONEFOR Sensinode v2.2 (CS22, Saura and Rubio [2010]) is a simple program for the quantification of the importance of habitat patches for maintaining landscape connectivity through graph structures and habitat availability indices, such as the integral index of connectivity (IIC) (Pascual-Hortal and Saura [2006]) and especially the probability of connectivity (PC). It is conceived as a tool for decision-making support in landscape planning and habitat conservation (Saura and Pascual-Hortal [2007]).
CS22 is written in Borland C++ and has recently been relicensed under GPL. The decision to release the code under an open license is farsighted, as it opens the development to other researchers and programmers. However, only the small Windows open source developers community can start contributing on the project, as the code compiles and runs only on Windows. The authors provide extensions for ArcGIS that help in preparing input data for Conefor, that are in a peculiar text format.

Circuitscape (Shah and McRae [2008]) is an open source program that uses circuit theory to predict connectivity in heterogeneous landscapes for individual movement, gene flow, and conservation planning. Landscapes are represented as conductive surfaces, with resistance proportional to the ease of dispersion or gene flow. Effective resistances, current densities, and voltages calculated across the landscapes can then be related to ecological processes (McRae et al. [2008]).
Circuitscape is being actively developed (the last version has been released in January 2012) and has a significant record of publications. The code is available under the LGPL license and is written in Python, with the use of NumPy, SciPy and wxPython libraries. The connection with GIS is performed by the interchange data formats, that are ASCII grids, and by the means of an ArcGIS plugin that prepares all raster layers for direct import into Circuitscape. The binding with other GIS will be established as soon as one or more users will write and publish more plugins. It is likely that this will happen, as Circuitscape is available for all main operating systems and is already making use of GIS formats. In addition, a tighter coupling can be obtained by embedding Circuitscape inside a GIS, in the same way that GRASS GIS and SAGA GIS are available into QGIS as toolboxes (Sherman [2008]).

A summary of the software packages examined in the present paper, focused on relevant information about code license and interoperability with GIS, is presented in Table 1.

Table 1: Overview of the presented software packages' properties

Software	Models	OS	License	GIS connection
Maxent	Maxent	All (Java)	Freeware	GRASS Addon
randomForest	Random Forest	All	GPL	via rgdal
openModeller	11 models	All	GPL	via open data formats
geoRglm	GLM	All	GPL	via rgdal
Conefor	GLM	Windows	GPL (from 2010)	ArcGIS plugin - no standard format
Circuitscape	GLM	All	GPL	ArcGIS plugin + open data formats

3 OPEN SOURCE GIS

3.1 History

In the 1990s a series of Open Source GIS software projects for both desktop and server systems was established in various GIS sectors, including software libraries for map reprojection and data format conversion, desktop GIS, Web mapping/Web GIS, spatial SQL databases, geostatistics, and metadata catalogues (Neteler et al. [2012]). In the same years, proprietary GIS products were in a more advanced development stage, noticeably on user interface: the first release of ArcView took place in 1991, after the launch of ARC/INFO in 1981 (`http://www.esri.com/about-esri/about/history.html`). In the same timeframe, GRASS was the earliest Open Source GIS to reach production status and the first to support both raster and vector data models. Development began in 1982 by the United States Army Corps of Engineers (Construction Engineering Research Laboratory, CERL) with software distributed through academic and public administration channels. The user interfaces of open source GIS were developed upon several GUI toolkits (Tcl/Tk, Qt, wxWidgets) that became popular in the mid 1990s. Therefore, until then, only closed source GIS was providing a working environment that was powerful and appealing at once.

3.2 Present and future

Up to date, a wide selection of desktop GIS packages are available, both as closed and open source. However, only open source is designed for the true interoperability due to its founding paradigm of reuse. It can also be customised at need, either by the user himself, either by a designed developer. On the opposite side, closed source software is not modifieable by the user, and the development directions are usually traced by the producer. This fundamental difference can be determinant, in the situations where there is equivalence of functionalities and usability.

4 OPEN SCIENCE

The open source philosophy is permeating into science (Barnes [2010]; Ince et al. [2012]). Reproducibility of a scientific paper's central finding, the central paradigm of modern science, is impossible when it uses closed source software. The defects in open source code are regularly found and corrected, and the same happens for closed source code, with the important difference that, in the latter, the errors and fixes remain unaccessible and of unknown severity. Declaration of algorithms implemented by a software

is only partially useful when determining the source of differences among software packages. Even when the algorithms descriptions are available, reproducibility and reliability of the software is not guaranteed (see "Failure of code descriptions" within Ince et al. [2012]). Hence, with some exceptions, anything less than release of actual source code is an indefensible approach for any scientific results that depend on computation, because not releasing such code raises needless, and needlessly confusing, roadblocks to reproducibility (Ince et al. [2012]). There are many obstacles to the publication of the whole code base of a scientific paper. Many science fields consider software of secondary importance with respect to the addressed issue, so that there is no reward for those who spend time and resources in polishing and documenting the code. This makes the code far from publishing standards, and hard to reuse, in a vicious circle that further weakens the reliability of results (Merali [2010]). Scientific research can improve the quality of the software used for publications by adopting well-tested open source development practices, such as documentation of the code, cooperative development of the code on version control systems and appropriate licensing. This is an overhead for scientists who have little computing background, but is necessary and should not be relied upon scientists alone. Governments, agencies and funding bodies have all called for transparency and have to concretely support it. "But the most important change must come in the attitude of scientists. If you are still hesitant about releasing your code, then ask yourself this question: does it perform the algorithm you describe in your paper? If it does, your audience will accept it, and maybe feel happier with its own efforts to write programs. If not, well, you should fix that anyway" (Barnes [2010]).

5 How to connect all this?

The convergence among SDM, open source GIS and open science is highly encouraged. The benefits for SDM will be a richer and stronger set of spatial tools inherited from open source GIS, and a better quality boosted by peer review of the code within open science philosophy. Open source GIS will have one more ecological analysis framework closely integrated to the existing functionalities. Good examples of this integration are openModeller, geoRglm, the Maxent connection for GRASS GIS, and the prototyped integration of QGIS with Conefor and Circuitscape (QGIS Ecological Toolbox and QGIS Frameworks).

There is an active progress in this area, for example on GRASS GIS Toolboxes, QGIS frameworks that will enable easy coupling with other software, and single developer activities like the GRASS-Maxent addon. These efforts could look poorly planned on a global perspective, but still they represent actual advances, and are open to further improvement.

What are the costs of this integration? Relicensing the code under an open license implies the acceptance of potential critics, and a lot of work to reshape the code in order to allow integration with other software - i.e. creating good interfaces for other statistical and GIS packages. This means that some parts will be rewritten entirely, based on the joint experience of both programmers and ecologists. The first step is to open the code for review, like Conefor Sensinode has done recently. The subsequent collaboration can be entirely delegated to interested researchers, programmers or institutions, or stimulated by specific financements. The migration to open source software always requires investment, mainly for users' training, data and functionality transfer. Over time, the migration costs will be lower than the annual costs of software licenses and potentially required hardware upgrades. An even more important advantage of open source software and data formats is the independence from single vendor (so called vendor lock-in), that should be considered at least toghether with economical aspects of migration.

It is important that scientists, researchers and students understand the importance of high quality programming in science, and are given training and support accordingly. This is especially true in ecological research, that has benefited from the exponential, though sometimes too enthusiastic, use of computer-driven analysis in the last decades. Much has been clearly said about the topic, but not enough on the means to reach the objective, and what are the issues perceived by researchers (Tse [2010]). This work identifies several software packages which are potential candidates for a complete SDM workflow, and the main guidelines of open source development, with the hope that it will encourage a better coupling of SDM software and GIS.

REFERENCES

Barnes, N. Publish your computer code: it is good enough. *Nature*, 467:753, 2010.

Breiman, L. Random forest. *Machine Learning*, 45:5–32, 2001a.

Breiman, L. Statistical modeling: the two cultures. *Statistical Science*, 16:199–215, 2001b.

Carpenter, G., A. N. Gillison, and J. Winter. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, 2:667–680, 1993. 10.1007/BF00051966.

Christensen, O. F. and P. J. Ribeiro Jr. geoRglm: A package for generalised linear spatial models. *R-NEWS*, 2(2):26–28, 2002.

Elith, J. and C. H. Graham. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, 32(1): 66–77, 2009.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC. M. Overton, A. Townsend Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberòn, S. Williams, M. S. Wisz, and N. E. Zimmermann. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151, 2006.

Foody, G. M. GIS: biodiversity applications. *Progress in Physical Geography*, 2(32): 223–235, 2008.

Ince, D. C., L. Hatton, and J. Graham-Cumming. The case for open computer programs. *Nature*, 482:485–488, 2012.

Jolma, A., D. Ames, N. Horning, H. Mitasova, M. Neteler, A. Racicot, and T. Sutton. *Springer Handbook of Geographic Information, Part C*, chapter Open-Source Tools for Environmental Modeling, pages 597–619. Springer, 2012.

Keitt, T. H., R. Bivand, E. Pebesma, and B. Rowlingson. *rgdal: Bindings for the Geospatial Data Abstraction Library*, 2010. R package version 0.6-33.

Kozak, K. H., C. H. Graham, and J. J. Wiens. Integrating GIS-based data into evolutionary biology. *Trends in Ecology and Evolution*, 23:141–148, 2008.

Liaw, A. and M. Wiener. Classification and Regression by randomForest. *R News*, 2(3): 18–22, 2002.

McRae, B., B. G. Dickson, T. H. Keitt, and V. B. Shah. Using circuit theory to model connectivity in ecology and conservation. *Ecology*, 10:2712–2724, 2008.

Merali, Z. Computational science: ...Error. *Nature*, 467:775–777, 2010.

Muñoz, M., R. Giovanni, M. Siqueira, T. Sutton, P. Brewer, R. Pereira, D. Canhos, and V. Canhos. openModeller: a generic approach to species' potential distribution modelling. *GeoInformatica*, pages 18–22, 2009.

Neteler, M., M. H. Bowman, M. Landa, and M. Metz. GRASS GIS: A multi-purpose open source GIS. *Environmental Modelling & Software*, 31(0):124 – 130, 2012.

Neteler, M. and H. Mitasova. *Open Source GIS: A GRASS GIS Approach*. Springer, New York, 3rd edition, 2008.

Pascual-Hortal, L. and S. Saura. Comparison and development of new graph-based landscape connectivity indices: towards the priorization of habitat patches and corridors for conservation. *Landscape Ecology*, 21(7):959–967, 2006.

Pebesma, E. J. and R. S. Bivand. Classes and methods for spatial data in R. *R News*, 5 (2):9–13, November 2005.

Peterson, A. T., V. Sánchez-Cordero, C. B. Beard, and J. M. Ramsey. Ecological niche modeling and potential reservoirs for Chagas disease, Mexico. *Emerging Infectious Diseases*, 8:662–667, 2002.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4):231–259, 2006.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

Ribeiro Jr, P. J. and P. J. Diggle. geoR: a package for geostatistical analysis. *R-NEWS*, 1(2):14–18, June 2001. ISSN 1609-3631.

Saura, S. and L. Pascual-Hortal. *Conefor Sensinode 2.2 User's Manual: Software for quantifying the importance of habitat patches for maintaining landscape connectivity through graphs and habitat availability indices*, 2007.

Saura, S. and L. Rubio. A common currency for the different ways in which patches and links can contribute to habitat availability and connectivity in the landscape. *Ecography*, 33:523–237, 2010.

Shah, V. and B. McRae. Circuitscape: a tool for landscape ecology. In *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, pages 62–66. G. Varoquaux, T. Vaught, J. Millman (Eds., 2008.

Sherman, G. *Desktop GIS: Mapping the Planet with Open Source Tools*. Pragmatic Bookshelf, Raleigh, 2008.

Stockwell, D. R., J. H. Beach, A. Stewart, G. Vorontsov, D. Vieglais, and R. S. Pereira. The use of the GARP genetic algorithm and Internet grid computing in the Lifemapper world atlas of species biodiversity. *Ecological Modelling*, 195(1–2):139 – 145, 2006.

Tse, H. Computer code: more credit needed. *Nature*, 468:37, 2010.