XVII[th] World Congress of the International Commission of Agricultural and Biosystems Engineering (CIGR)

Hosted by the Canadian Society for Bioengineering (CSBE/SCGAB)
Québec City, Canada June 13-17, 2010

CHALLENGES IN ECOLOGICAL NICHE MODELLING

FABIANA SOARES SANTANA[1], ANTONIO MAURO SARAIVA[2]

[1] F.S. SANTANA, Federal University at ABC – Center of Mathematics, Computer and Cognition – SP – Brazil, fabiana.santana@gmail.com.
[2] A.M. SARAIVA, University of São Paulo SP – School of Engineering, Department of Computer and Digital Systems Engineering – SP – Brazil, saraiva@usp.br.

CSBE101644 – Presented at the 8th World Congress on Computers in Agriculture (WCCA) Symposium

ABSTRACT Ecological niche modelling combines species occurrence points with environmental raster layers, in georeferrenced coordinates, so as to obtain models that describe the probabilistic distribution of a biological species in a pre-defined area. This problem involves many challenges in different areas of knowledge, such as systems interoperability and architecture, distributed database integration, algorithm research, modelling strategies, statistics and parallel computing solutions. This paper presents the main challenges related to the ecological niche modelling problem and discusses the relevance of solving these problems in a broader context. In addition, this work also presents suggestions in order to apply the obtained solutions to solve other related problems.

Keywords: Sustainable Biosystems, Ecological Niche Modelling, Species Distribution, System Modelling.

INTRODUCTION

Nowadays, conservation and sustainable use of natural resources are highly important research themes for many different purposes. The geographic distribution of a species, considering the spatial, the ecological and the evolutionary perspectives, is a very relevant aspect of this broader problem. The study of species distribution can be very important for their survival, specially considering the environmental problems that are, potentially, being introduced by global warming (Guralnick & Neufeld, 2005). Techniques to obtain such distributions are the main purpose of ecological niche modelling.

A species fundamental niche represents the conditions that allow a species to survive (Soberon & Peterson, 2005). A species realized niche is a subset of the fundamental niche considering external factors, such as human influence, biotic interactions and geographic barriers, which may impede a population from growing, even within an adequate ecological niche area. Despite their relevance, many of those factors are difficult to evaluate without extra information. So, an ecological niche-based model is closer to the fundamental niche of a species than to the realized one.

The purpose of ecological niche modelling is to obtain areas environmentally similar to those where the species actually occurs. The hypothesis is that, if a species can be found under certain conditions, then it should be able to survive and reproduce in any place with the same conditions.

The result of the ecological niche modelling process is a niche model, which represents the probability of finding a species under the time-space conditions described by the input data. This model can be projected onto a map of the same study region, or onto different regions or periods in time, considering scenarios in past, present and future.

The process to generate an ecological niche modelling was described in Santana et al. (2008a). The input data of the process are mainly species presence and absence data, represented as single points in the geographical space, and layers of data about the environmental conditions that are relevant to that species survival. The layers must include, at least, the area where the presence and absence points were found. The input data must be georeferrenced so as to allow probabilistic analysis. After receiving the input data, specific computational modelling algorithms must be applied in order to obtain a niche model. After being generated, the model is projected onto a geographic region, presenting, in a map, a function corresponding to the species distribution, according to the probabilistic method applied to generate the model. So, it is possible to tell the probability of finding the studied species in each georeferrenced coordinate of the area where the model was projected.

This technique is very important and was already applied to, for instance,: 1) Propose scenarios for sustainable use of the environment (Chapman et al., 2005); 2) Evaluate the potential of invasive species (Peterson et al., 2003); 3) Evaluate climatic changes impacts on biodiversity (Huntley et al., 1995; Magana et al., 1997; Peterson et al., 2002; Sala et al., 2000; Siqueira & Peterson, 2003; Thomas et al., 2004); 4) Delineate potential routes of infections and diseases (Petersen & Roehrig, 2001; Peterson et al. 2007a); and 5) Indicate potential priority areas for conservation (Chen & Peterson, 2002; Ortega-Huerta & Peterson, 2004).

CHALLENGES IN ECOLOGICAL NICHE MODELLING AT THE DATA, MODEL AND COMPUTACIONAL TOOL LEVELS

Though, at the first sight, ecological niche modelling may be seen as a simple problem to be solved by computational tools, it is remarkable the amount of challenges that are involved in the problem. Figure 1 presents a summary of these challenges. Each of them will be detailed in the following text.

The study of the challenges may start by discussing algorithm research for ecological niche modelling. The following algorithms are already available: 1) Bioclim (Nix, 1986); Climate Space Model – Broken-Stick [http://openmodeller.sourceforge.net]; 2) Envelope Score (Nix, 1986); 3) Environmental Distance [http://openmodeller.sourceforge.net]; 4) GARP – Genetic Algorithm for Rule-set Production (Stockwell & Peters, 1999; Stockwell & Noble, 1992), in two different versions: DesktopGarp e openModeller; 5) GARP BestSubsets – Genetic Algorithm for Rule-set Production with Best Subsets

Procedure (Anderson et al., 2003); 6) MaxEnt (Philips et al., 2006); and 7) Support Vector Machine - SVM (Giovanni & Lorena, 2007). Most of them are non-deterministic solutions to the same modelling problem which are based on a variety of computational methods. However, it is not always possible to say that an algorithm is more precise than another, since there are no reliable automatic techniques for evaluating models. So, from the theoretical and from the practical point of view, the problem of model precision is still open.

Directly related to this problem, there is a highly important and even more difficult challenge, which is to define a reliable method to evaluate the accuracy of a model. This problem, obviously, involves statistical and computer methods, because it is necessary to establish parameters and algorithms to allow an accurate evaluation.

Besides evaluating the model itself, it is also necessary to evaluate each sample, because there are ecological conditions that must also be considered. For instance, what does an absence point really mean? Does it mean that the species does not *exist* in some specific area or does it simply mean that the species was *not yet found* in the area? Other important challenge is related to the evaluation of the quality of the sample. Even when the data set available is composed of valid and verified presence points only, it is still necessary to decide if the amount of points in the set is enough and if their distribution is representative of the evaluated area.

Since modelling is heavily dependent on data, an important issue is that of discovering, accessing and obtaining the data. Biological specimens' occurrence data typically comes from biological collections, from small personal collections to large collections in museums. A global effort is under way to digitize biological collections data. As this is done with very different software tools, with different database schemas and many other levels of heterogeneity, there are initiatives to develop standards and protocols to address specific needs of the community that deals with biodiversity data. Institutions and organisms such as BIS (Biodiversity Informatics Standards, formerly TDWG – Taxonomic Database Working Group), and GBIF (Global Biodiversity Information Facility) and others have played an essential role on the definition and adoption of the Darwin Core Schema (DwC) for biological data, and of the TAPIR protocol, to mention the most important. Based on DwC and TAPIR many databases are becoming available online and more easily accessible on integrating portals such as those from GBIF (www.gbif.org) and IABIN (pollinators.iabin.net). However, currently a small percentage of the estimated total of specimens ever collected in the world is digitized and available on line, according to GBIF. A challenge is to increase the pace of digitization and to cover collections and datasets from all over the world. Besides the obvious problems with funding for that activity there is also the problem of willingness to share data, something that encounters barriers that can be personal and national (due to the fear of biopiracy, for instance).

Yet another point is to deal with data quality issues. A series of problems can be listed on this respect: from digitization errors of the complicated Latin scientific names, to misclassification of the specimens, to changes in classification; from the lack of geographical coordinates and locality data, to the uncertainty of the data and digitization errors, etc. The process of reviewing the data for quality is very time-consuming, requires computer tools to speed it up and experts to identify less evident errors.

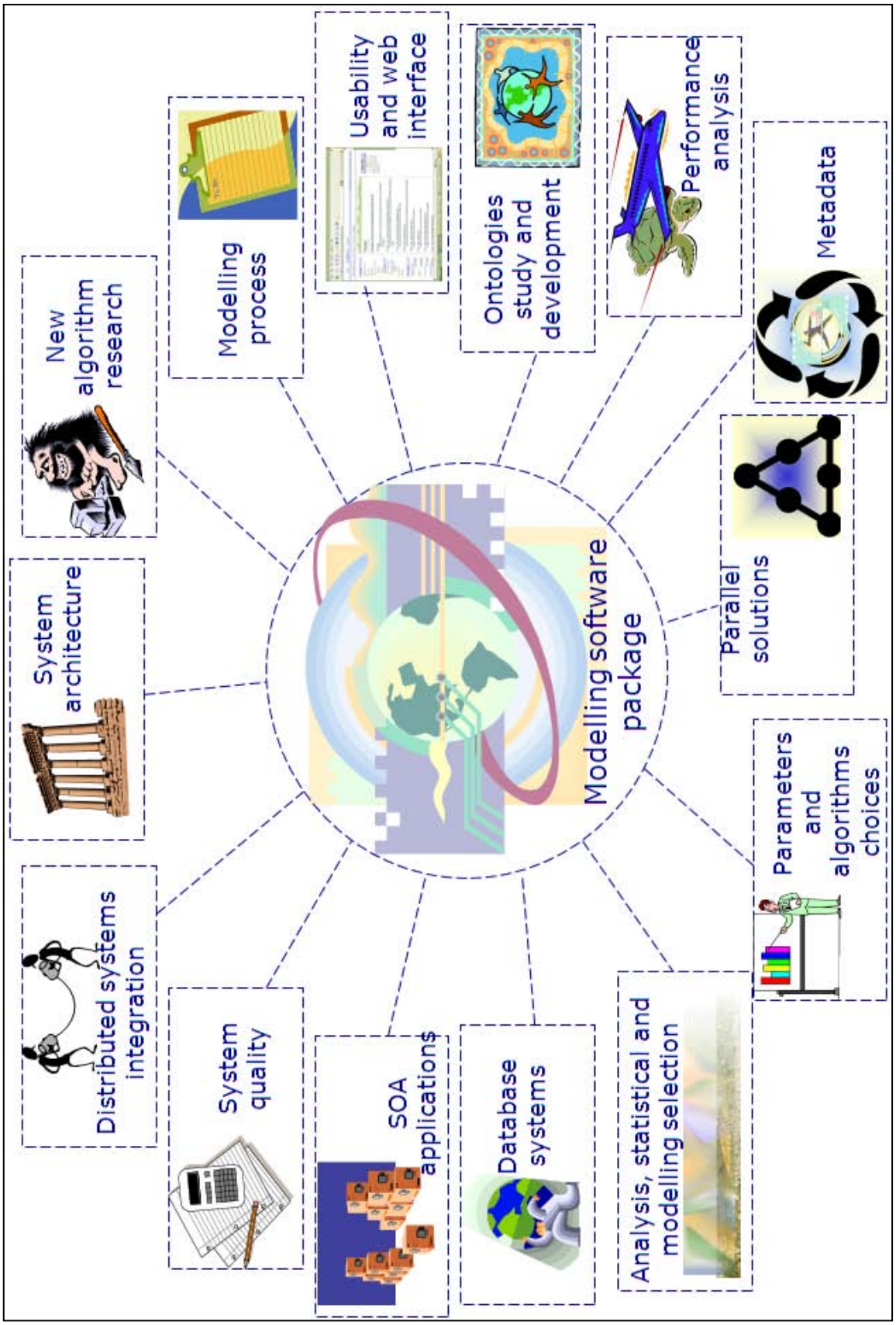

Figure 1. Computational and ecological challenges in ecological niche modelling.

Another issue is the evolution of the standards to facilitate the digitization of more data types to enrich the data currently available for modeling and analysis. The development of ontology-based standards is a promise to allow easier discovery and seamless integration of data sources and is already being pursued at BIS. A special case are ontologies based on web services technologies, such as the alternative presented in Michener et al. (2007) and Daltio and Medeiros (2008).

Once a model is generated, it should be possible to store it, so as to allow it to be share and compared to other models for further analysis and research improvements. A models data base is, thus, necessary.

This problem becomes, then, the definition of the requirements of the models data base and the communication with this database. Besides, it is also necessary to define policies for storage. For instance, some generated models may be part of unpublished works and the authors may not be interested in their divulgation at the moment they were generated, only in its storage. A standard to describe models is therefore necessary: a workgroup to work on this topic was established during the latest TDWG meeting, in 2009.

Performance improvements are other big challenge, since depending aspects such as the number of species, of presence points, on the spatial resolution and amount of layers, a model can take from hours to days to be generated. A few works into this direction can be mentioned: the development of parallel algorithms for computer clusters, such as the P-GARP of openModeller project (Santana et al., 2008b) and other initiatives in the LifeMapper project [http://www.lifemapper.org/]. Nevertheless, there is still a long path to obtain sustainable and scalable results. Due to the performance improvements requirements, parallel algorithm research is therefore another important challenge, directly related to the previous. In addition, tools to evaluate and carry out the analysis of performance are directly related to this challenge and they must be adopted, in order to measure the evolution of parallel solutions.

Considering the environmental modelling process (Santana et al., 2008a), there are many degrees of liberty in the choice of parameters for algorithms. These choices can results in completely different models, which is undesirable for automatic processes, such as the LifeMapper project approach. Studies applying statistics, adaptive techniques and sensitivity analysis could contribute to turn this choice into something less experimental than it actually is. The same problem appears in the algorithm choice, but as most of them are non-deterministic, the problem here is even more complicated.

Still related to the same context, other pre- and post-analysis tools are also very desirable. The software Maxent, for instance, uses the jackknife method (Efron, 1979) to help select the environmental layers. openModeller, another important niche modeling tool, also offers jackknife and a data cleaning option to remove invalid entries from input data bases. Since data are not always very reliable, as previously mentioned, data cleaning is an important improvement.

Other series of algorithms could be developed to help these tasks. For instance, an automatic process of Internet data base scan and data importing could help the researcher in the pre-processing work, which is a very time-consuming task (Chapman et al., 2005). Filters to evaluate the quality of such data would also be very helpful. The same would

apply to GIS integration and distributed database integration, using services furnished by many different laboratories. A partial integration with GIS systems was proposed in Santana et al. (2007c).

Other relevant research area is directly related to the usability of the systems for ecological niche modelling. Sometimes, the user has many problems to use these systems because of their complex interfaces and because modeling is in itself a complex process; so studies to enhance the usability would be very helpful. There are already monolithic solutions with relatively good interfaces, but more effort must be made for web solutions, so as to allow a seamless integration among different systems and data sources, which will lead to another level in ecological niche modelling technique. A web interface which has the same facilities of monolithic systems and which is able to include tools with better performance, cluster access and database integration, for instance, would increase the success of such systems and elevate them to a new scale of quality. An on-line help would also be very helpful for beginners.

At last, but certainly one of the most important, is the issue of building architectural solutions to easily incorporate and adequate all the discussed challenges. This must incorporate the concept of software services, mainly in the form of web services, since the integration of so many different solutions and platforms would be very difficult using monolithic solutions only or proprietary approaches. A reference architecture was proposed in Santana et al. (2007a) and an infrastructure to integrate all the different challenges of the ecological niche modelling problem was developed in Santana (2009). This solution incorporates reusable modules and is based on SOA (service oriented architecture) principles, open patterns, platform independence, and it is also compatible with many different interfaces, so as to offer the ability to integrate new algorithms and solutions, offering security and access control resources.

However, even the best solution will still have to conquer users and overcome further barriers, including convincing developers of the relevance of services and metadata adoption in order to provide plain integration and more chances to achieve success, such as it happens in other important projects. This kind of data base could also be distributed world wide, under access and dissemination rules established by a consortium (e.g. Protein Data Bank [http://www.rcsb.org/pdb]).

DISCUSSION

The challenges presented have very different levels of complexity and thus can be met according to different strategies. Some require an integrated approach by the international community while others can be met by more isolated efforts from a single research group or individual. The former case is that of the standards and protocols for data digitization and sharing, and for models sharing, as well as the development of data integration portals and other tools. The latter case is that of algorithms research and of data digitization of a specific dataset.

An interesting example of an integrated effort is that of Elith et al. (2007), in which a big modelling experiment was developed so as to compare models generated by different

algorithms, for a set of species under related conditions. It required 27 collaborators to discuss 16 modelling methods over 226 species from 6 regions of the world. At least the following software packages were applied: DIVA-GIS [http://www.diva-gis.org/]; Desktop GARP [http://www.nhm.ku.edu/desktopgarp/]; R [http://www.r-project.org/]; MaxEnt [http://www.cs.princeton.edu/~schapire/maxent/]; and openModeller [http://openmodeller.sourceforge.net/].

The experiment shows that there is a need for methods and parameters to help choose the tools, algorithms and parameters to be used for a specific problem. It also highlights the need for integration of the software tools being developed world wide. openModeller seems to be the more open effort of all, and will benefit from having more collaborators and an effective open collaborative software development process.

The definition of the ecological niche modelling process (Santana et al., 2008a), was an attempt to formalize the process, aiming at serving as a guide to software developers and to modellers, especially newcomers. Many issues related to the modeling process were highlighted there.

Other important point to stress is that some challenges presented here are not exclusive for ecological niche modelling. Some are broad computational problems pertinent to many areas. In Santana et al. (2007b), the relationship among ecological niche modelling and precision agriculture systems were demonstrated, as both demand modeling spatial data distribution, based on many layers, involving biological processes. The same could be applied to other areas of knowledge or to other projects, such as the Lifemapper [http://www.lifemapper.org/], ViNCES, Virtual Networking Center of Ecosystem Services [http://www.ib.usp.br/vinces/], and IABIN - Pollinators Thematic Network [http://pollinators.iabin.net/],which involve data integration and sharing, for instance.

Until integrate solutions become available, researchers of ecological niche modelling will continue highly dependent of the software packages and of each implementation. So the international community must understand the relevance of the usage of standards and the discussion of global and effective solutions.

CONCLUSION

There are many challenges that have to be faced to further develop ecological niche modelling tools, and to allow a more effective, scientifically-based and broad use of niche models. As it was showed, though there are efforts to meet some of them, many others still remain open problems.

The proposal of this discussion was to motivate and impel more collaborators to contribute to the solution of some of the presented problems. Despite the relevance of the ecological niche modelling technique itself, some results or solutions obtained for this research area may probably be applied in the solution of other related research areas and problems.

Modeling is an essential technique to develop decision support systems, and these are very much in need nowadays to help protect the environment, and promote sustainable agricultural production systems.

Acknowledgements. Authors are grateful to FINEP – Financiadora de Estudos e Projetos – from Ministério da Ciência e Tecnologia – MCT/Brazil, and to FAPESP – Fundação de Amparo à Pesquisa do Estado de São Paulo/Brazil, for the support to the openModeller (04/11012-0) and BioAbelha (04/15801-0) projects.

REFERENCES

Anderson, R.P., Lew, D., Peterson, A.T. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological Modelling 162, 211-232.

Chapman, A. D., Muñoz, M. E. S., Koch, I. 2005. Environmental information: placing biodiversity phenomena in an ecological and environmental context. Biodiversity Informatics 2, 24-41.

Chen, G. and Peterson, A. T. 2002. Prioritization of areas in China for biodiversity conservation based on the distribution of endangered bird species. Bird Conservation International 12, 197-209.

Daltio, J. and Medeiros, C. B. 2008. An Ontology Web Service for Interoperability across Biodiversity Applications", Information Systems.

Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, Vol. 7, No. 1., pp. 1-26.

Elith, J., Graham, C. H., Anderson, R. P., Dudı´k, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. McC., Peterson, A. T., Phillips, S. J., Richardson, K. S., Scachetti-Pereira, R., Schapire, R. E., Sobero´n, J., Williams, S., Wisz, M. S., Zimmermann, N. E. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129-151.

Giovanni, R. & Lorena, A.C. 2007. Support Vector Machine. [http://openmodeller.sourceforge.net/].

Guralnick, R., Neufeld, D. 2005. Challenges building online GIS services to support global biodivesity mapping and analysis: lessons from the mountain and plains database and informatics project. Biodiversity Informatics 2, 56-69.

Huntley, B., Berry, P. M., Cramer, W., and McDonald, A. P. 1995. Modelling present and potential future ranges of some European higher plants using climate response surfaces. Journal of Biogeography 22, 967-1001.

Magana, V., Conde, C., Sanchez, O., and Gay, C. 1997. Assessment of current and future regional climate scenarios for Mexico. Climate research 9, 107-114.

Michener, W. K., Beach, J. H., Jones, M. B., Ludäscher, B., Pennington, D. D., Pereira, R. S., Rajasekar, A., Schildhauer, M. 2007. A knowledge environment for the biodiversity and ecological sciences. Journal of Intelligent Information Systems. Volume 29, Number 1 / August, 2007. 111-126.

Nix, H.A. 1986. A biogeographic analysis of Australian elapid snakes. In: Atlas of Elapid Snakes of Australia. Ed. R. Longmore, pp. 4-15. Australian Flora and Fauna Series Number 7. Australian Government Publishing Service: Canberra.

Ortega-Huerta, M. A. and Peterson, A. T. 2004. Modelling spatial patterns of biodiversity for conservation prioritization in North-eastern Mexico. Diversity and Distributions 10, 39-54.

Petersen, L. R. and Roehrig, J. T. 2001. West Nile virus: A reemerging global pathogen. Emerging Infectious Diseases 7, 611-614.

Peterson, A. T., Ortega-Huerta, Miguel A., Bartley, Jeremy, Sanchez-Cordero, Victor, Soberón, Jorge, Buddemeier, R. H., and Stockwell, David R. B. 2002a. Future projections for Mexican faunas under global climate change scenarios. Nature 416,

626-629.
Peterson, A. T., Papes, M., and Kluza, D. A. 2003. Predicting the potential invasive distributions of four alien plant species in North America. Weed Science 51, 863-868.
Peterson, A. T., Benz, B. W., Papeş, M. 2007. Highly pathogenic H5N1 avian influenza: Entry pathways into North America via bird migration. PLoSONE 2(2): e261. doi:10.1371/journal.pone.0000261.
Phillips, S. J., Anderson, R. P., and Schapire, R. E. 2006. Maximum entropy modelling of species geographic distributions. Ecological Modelling 190, 231-259.
Piñeiro, R., Aguilar, J. F., Munt, D. D. & Feliner, G. N. 2007. Ecology matters: Atlantic-Mediterranean disjunction in the sand-dune shrub Armeria pungens (Plumbaginaceae). Molecular Ecology. 16, 2155-2171
Sala, O. E., Chapin-III, F. S., Armesto, J. J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L. F., Jackson, R. B., Kinzig, A., Leemans, R., Lodge, D. M., Mooney, H. A., Oesterheld, M., Poff, N. L., Sykes, M. T., Walker, B. H., Walker, M., and Wall, D. H. 2000. Global biodiversity scenarios for the year 2100. Science 287, 1770-1774.
Santana, F. S. Uma Infraestrutura Orientada a Serviços para a Modelagem de Nicho Ecológico. PhD Thesis in Engineering (Engineering of Computer and Digital Systems). Escola Politécnica da Universidade de São Paulo. April, 2009. 141p. (In portuguese.)
Santana, F. S., Murakami, E., Saraiva, A. M., Bravo, C. & Correa, P. L. P. 2007a. Uma arquitetura de referência para sistemas de informação para modelagem de nicho ecológico. Anais do 6º Congresso Brasileiro de Agroinformática – SBIAgro 2007, Campinas: Embrapa Informática Agropecuária, 2007. Editors: S.Tiernes, L.H.A. Rodrigues. p. 101-105.
Santana, F. S., Murakami, E., Saraiva, A. M. & Correa, P. L. P. 2007b. A comparative study between precision agriculture and biodiversity modelling information systems. 6th Biennal Conference of the European Federation of IT in Agriculture, Glasgow: C.Parker, S.Skerratt, C.Park, J.Shields, 2007. v. 1. p. 1-6.
Santana, F. S., Pinaya, J. L. D., Saraiva, A. M., Correa, P. L. P., Becerra, J. L. R. & Bravo, C. 2007c. Aplicação de SOA para identificação de serviços em sistemas de modelagem de nicho ecológico e GIS. I2TS'2007 Proceedings of the 6th International Information and Telecommunication Technologies Symposium, Brasília: IEEE R9, 2007. Editors: Fundação Bardall de Educação e Cultura; Boukerche, A, Loureiro, A.A.F., Melo, A.C.M.A. and Gondim, P.R.L.
Santana, F. S., Siqueira, M. F., Saraiva, A. M. & Correa, P. L. P. 2008a. A reference business process for ecological niche modelling. Ecological Informatics, Vol. 3, Issue 1, pp 75-86.
Santana, F. S., Bravo, C., Saraiva, A. M. 2008b. Parallel Genetic Algorithms for Rule-set Production. Environmental Modelling and Software (submitted).
Siqueira, M. F. and Peterson, A. T. 2003. Consequences of Global Climate Change for Geographic Distributions of Cerrado Tree Species. Biota Neotropica 3.
Soberon, J. and Peterson, A. T. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiversity Informatics 2, pp 1-10.
Phillips, S. J., Dudík, M. & Schapire, R. E. 2004. A maximum entropy approach to species distribution modeling. In Proceedings of the Twenty-First International Conference on Machine Learning, pages 655-662.
Phillips, S. J., Anderson, R. P., Schapire, R. E. 2006. Maximum entropy modeling of

species geographic distributions. Ecological Modelling, 190:231-259.
Stockwell, David R. B. and Noble, I. R. 1992. Induction of sets of rules from animal distribution data: A robust and informative method of analysis. Mathematics and Computers in Simulation 33, 385-390.
Stockwell, D. R. B. and Peters, D. 1999. The GARP modelling system: Problems and solutions to automated spatial prediction. International Journal of Geographic Information Systems 13, 143-158.
Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F. N., Siqueira, M. F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., Jaarsveld, A., Midgley, G. F., Miles, L., Ortega-Huerta, M. A., Peterson, A. T., Phillips, O. L., and Williams, S. E. 2004. Extinction risk from climate change. Nature 427, 145-148.