Jun 17th, 3:40 PM - 5:20 PM

Modeling and Implementation of a Geospatial Database for Environmental Niches and Potential Geographic Distributions

Gerardo José Zárate
Universidade Federal de Vicosa, gerardo.zarate.m@gmail.com

Jugurta Lisboa-Filho
Universidade Federal de Vicosa, jugurta@ufv.br

Carlos Frankl Sperber
Universidade Federal de Vicosa, sperber@ufz.br

Flavia Maria da Silva Carmo
Universidade Federal de Vicosa, fmcarmoufv@gmail.com

Modeling and Implementation of a Geospatial Database for Environmental Niches and Potential Geographic Distributions

Gerardo José Zárate[a], Jugurta Lisboa-Filho[a], Carlos Frankl Sperber[b], Flavia Maria da Silva Carmo[b]
[a]*Departamento de Informática, Universidade Federal de Viçosa, Viçosa, MG, Brazil (gerardo.zarate.m@gmail.com, jugurta@ufv.br)*
[b]*Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa, MG, Brazil (sperber@ufv.br, fmcarmoufv@gmail.com)*

Abstract: The ecological niche of a species is defined by an array of biotic and abiotic requirements that allow organisms to live and reproduce in a geographic region. Abiotic data from the ecological niche in combination with occurrence data can be used to predict the potential geographic distribution of a species in different regions. Potential geographic distributions are useful in predicting the extent of invasive species, predict distributions as preparation for climate change and find appropriate regions for endangered species, among others. Geographic entities and environmental variables can be represented with a high-level abstraction in diagrams using formalism dedicated to model geospatial databases. The schemas produced by these formalisms can later be transformed into implementation code using software-design approaches such as the Model-Driven Architecture (MDA). This work describes the stages of MDA to generate geospatial databases for ecological niches and potential geographic distributions data. The paper also presents a study case, estimating the potential distribution of the tree *Myracrodruon urundeuva*, to show how professionals and researches can use the proposed schema to implement a repository for ecological niche and potential distributions data.

***Keywords**:* Environmental niche; Potential Geographic Distributions; Model-Driven Architecture; Conceptual Modelling; Myracrodruon urundeuva.

1. INTRODUCTION

The ecological niche of a species includes all biotic interactions and abiotic variables of the environment in which organisms can survive and reproduce (Polechová and Storch, 2008). Both biotic interactions and abiotic variables can be used to construct an ecological niche in the manner of an n-dimensional hypervolume, as first proposed by Hutchinson (Hutchinson, 1957).

Over the past few years, scientists and researchers have been modeling ecological niches in order to predict distributional areas (also known as Potential Geographic Distributions) of certain species (Soberón, 2007). Potential Geographic Distributions are useful to predict scenarios of climate change projection, outbreak of diseases, and invasion of species, among others (Blackburn, 2010; Peterson and Vieglais, 2001).

Mathematical tools, such as BioMaper and the Genetic Algorithm for Rule-Set Production (GARP), use occurrence data of a species and a set of environmental layers as inputs to model the ecological niche and project possible distributions of the species (Ortega-Huerta and Peterson, 2008). Both inputs and outputs can be imported into a Geographic Information System (GIS), such as QuantumGIS, to display the data or to perform further analysis. Unfortunately, tools, like the mentioned before, work with separate files and does not provide a unified repository.

The final aim of our work is to implement a unified repository for ecological niches (represented by an n-dimensional hypervolume) and potential geographic distribution of species in a Database

Management System (DBMS) with geospatial capabilities, i.e., with means to handle geographic information. To accomplish this, first we need a conceptual data schema capable of representing environmental and geographic data, that capability is the major advantage over other attempts to model niches and distribution data from a conceptual standpoint. To construct the data schema we used UML GeoProfile, a formalism dedicated to model geospatial databases from a conceptual standpoint, and its support to the Model Driven Architecture (MDA) approach (Lisboa-Filho et al., 2013). An implementation of the proposed data schema would be capable of storing the inputs and outputs of mathematical algorithms, such as GARP.

The rest of this paper is structured as follow. Section 2 reviews related works. Section 3 describes the stages of the MDA approach to implement a geospatial database. Section 4 presents a study case, in which we use available data for the *Myracrodruon urundeuva* species and environmental layers to estimate an ecological niche model and to produce a potential geographic distribution. Section 5 provides some final considerations.

2. RELATED WORK

Previous studies have attempted to provide means to model niche and geographic distribution information from a conceptual standpoint. This section summarizes prior efforts found in the literature regarding databases as well as other works related to projections of potential geographic distributions.

Blackburn (2010) emphasizes the importance of databases in GIS applications stressing their storage capabilities. Moreover, Blackburn (2010) provides a six-step guide for using ecological niche data to predict potential geographic distributions using the GARP algorithm. The Canopy Database Project (McIntosh et al., 2007) is a tool developed to help ecologists to design databases. The focus of this work is to simplify the design process for ecologists with no experience in database theory. Models created in this tool can later be exported to a DBMS. OntoCrucible was developed as a framework to provide guidelines for representing ecological niches in a conceptual model (Semwayo and Berman, 2004). The authors propose an ontological engineering approach to model ecological data; the focus of their study is modeling the relationships between humans and their environment. Keet (2006) presents an Object-Role Modeling (ORM) diagram of the ecological niche. The proposed ORM diagram includes entities such as species, conditions, hyper-volumes, fundamental niche and realized niche.

Regarding previous work related to projection of potential geographic distributions, there are several studies that can be taken into account, here we mention only a few of them which use predictive algorithms to generate possible geographic distributions. Ward (2007) used occurrences and environmental data to predict the geographic extent of invasive ants in New Zealand. Roura-Pascual, et al, (2004) took advantage of the GARP algorithm to predict the potential geographic distribution of Argentine ants in the face of global climate change. Similarly, Blackburn (2010) used the GARP algorithm to project the potential geographic distribution of the Anthrax Angent (Bacillus Anthracis). Finally, Ortega-Huerta and Peterson (2008) compared the results of six algorithms regarding the prediction of the potential geographic distribution of ten Mexican birds.

3. DATA SCHEMA FOR ENVIRONMENTAL NICHES AND POTENTIAL GEOGRAPHIC DISTRIBUTIONS USING MDA

Zárate et al., (2014) proposed a conceptual data schema for environmental niches and potential geographic distributions using the formalism UML GeoProfile (Lisboa-Filho et al., 2013). Being an extension of the Unified Modeling Language (UML), UML GeoProfile uses classes, associations (relationships) and other UML features. UML GeoProfile was preferred over other formalisms for its capacity to model both object and field phenomena, as well as for the implementation of international standards. Other characteristic of UML GeoProfile is its support of the MDA approach. MDA is a design technique that emphasizes the use of models in the software development process (Mellor et al., 2002). In MDA, software is first modeled in a Computation Independent Model (CIM); CIM models are later transformed to a Platform Independent Model (PIM). The third stage of the process is the Platform Specific Model (PSM), which is later converted to implementation code. Details of UML GeoProfile and other formalisms for geospatial databases can be found in Lisboa-Filho et al. (2013) and Pinet (2012).

Figure 1 shows the CIM stage of the conceptual data schema modeled using UML GeoProfile. Some classes are marked with a pictogram (stereotypes in UML GeoProfile), which indicates a class being geospatial (polygon, point, field objects) or affected by time (temporal objects).

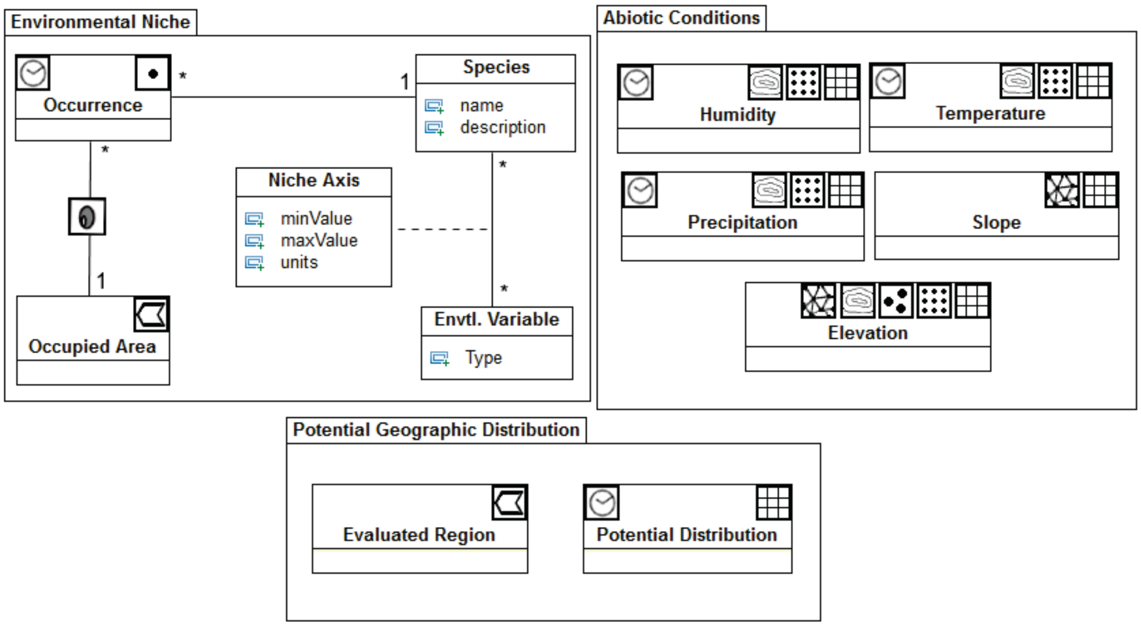

Figure 1. MDA CIM level for Environmental Niches and Potential Distributions

Occurrences of a species are modeled as geospatial points that are inside a specific area. It is possible to spot the same organism in two or more areas in a different time; this is solved by assigning the *Temporal Object* stereotype to the *Occurrence* class, this indicates the instant in which the organism was spotted. Dimensions of the hypervolume are represented with the class *Niche Axis*, which is dependent of the relationship between the *Species* and the *Environmental Variable* classes. Multiple instances of the *Niche Axis* class form the n-dimensional hypervolume of a species.

Environmental data in the form of GIS coverage types are shown in the package *Abiotic Conditions*. These conditions can be stored in a variety of GIS types, such as TIN, Grid of Points and Grid of Cells, among others. Those types are specified with the UML stereotypes (pictograms). Notice that there is not relationship between geographic areas and coverage layers, the reason behind this is that they belong to different conceptual views (Geographic Objects and Fields) and there is no topological relationship between them. Finally, the *Evaluated Region* and *Potential Distribution* classes (i.e. the potential distribution of the evaluated region of interest) are modeled as Polygon and Grid of Cells respectively. Further explanation of the CIM stage is presented in Zárate et al., (2014).

The next level of abstraction is the PIM level, which is still independent of any implementation technologies. The transformation process from CIM to PIM consists in the inclusion of identifiers that differentiate the instances of a class and the use of standard ISO types. UML GeoProfile stereotypes change to ISO data types, e.g., the Polygon stereotype is transformed to a GM_Surface attribute (Lisboa-Filho et al., 2013). Figure 2 exhibits the PIM level of MDA.

Finally, for a PSM model it is necessary to use a technology capable of storing geospatial phenomena. Some wide known technologies are Oracle Spatial and PostgreSQL (with the PostGIS geospatial extension), both able to store and manage geospatial data. Figure 3 exhibits the PSM level of the proposed schema using PostgreSQL and PostGIS. The PSM level presents implementation details like the use of Primary and Foreign Keys. Notice the use PostgreSQL specific basic data type such as NUMERIC or CHARACTER VARYING. ISO standard geospatial types are transformed to PostGIS implementations, e.g., the GM_Surface ISO type is transformed to the generic GEOMETRY type. The PSM stage can later be implemented in PostgreSQL using basic SQL statements.

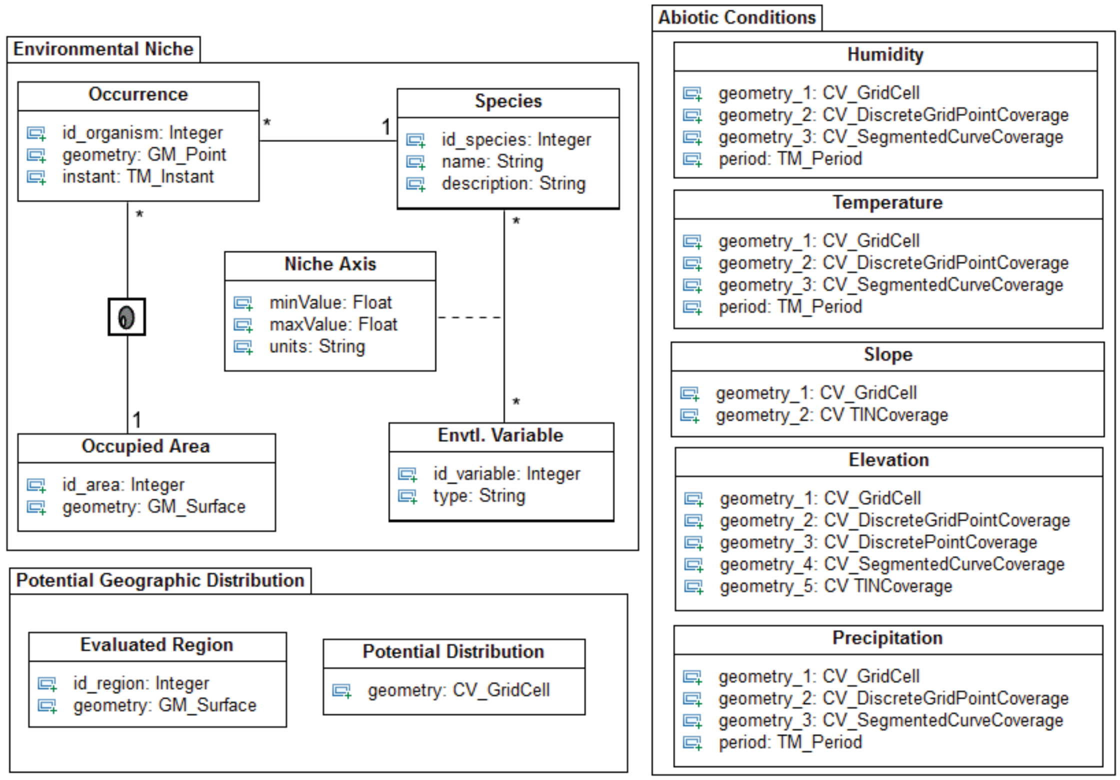

Figure 2. MDA PIM level.

4. STUDY CASE AND IMPLEMENTATION

As mentioned before, one of the applications of the potential geographic distributions is the projection of the geographic extent of invasive species. *Myracrodruon urundeuva* is a native Brazilian tree that grows in the northeast, southeast and central-west regions of Brazil. *M. urundeuva* has both economic and medical importance: it is a high-quality hardwood that can be used in constructions and its tannins are useful in pharmacology (Carmello-Guerreiro and Paoli, 1999). Furthermore, *M. urundeuva* has shown potential in therapeutic treatments of neurological disease patients (Souza et al., 2007).

Up to the year 2008, *M. urundeuva* was on the Official List of Endangered Brazilian Flora, which prevents the exploitation of the species. However, in some regions of Brazil there is a notable monodominance (a tree species occupies more than 50% of the forest area), which causes soil erosion preventing the growth of other species. Furthermore, the monodominance leads to environmental, economic and social damage (Murta et al., 2012). Therefore, we consider of utmost importance to know which regions have the appropriate environment for *M. urundeuva* for two reasons. Because it is listed as an endangered species in Brazil; and, contrastingly, because it can potentially be an invasive species if it presents monodominance in the region. We attempt to propagate the usage of potential geographic distribution techniques, combined with the data schema presented in this paper, for professional and researchers interested in the species.

DesktopGarp and OpenModeller are two well-known free tools for ecological niche modeling and geographic distribution predictions. While DesktopGarp is an implementation of the GARP algorithm, OpenModeller offers a wide selection of algorithms, including GARP. For the test presented in this paper, we used OpenModeller as an arbitrary choice. Notice that comparing both tools is out of the scope of this paper.

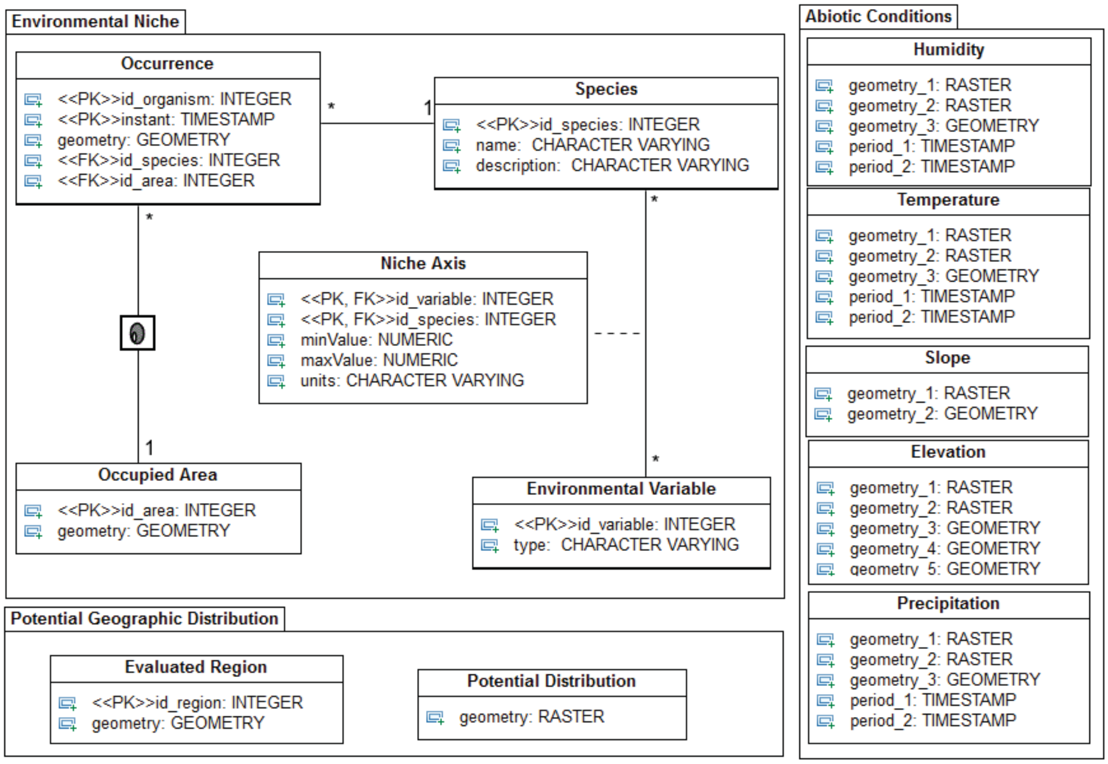

Figure 3. MDA PSM level using PostgreSQL and PostGIS.

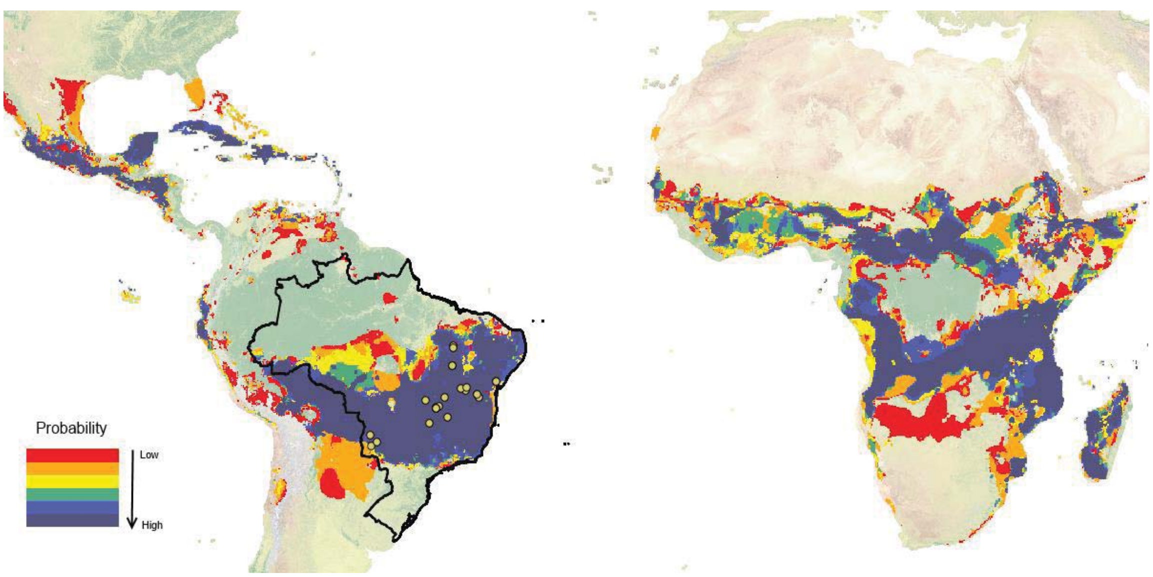

Figure 4. A possible potential geographic distribution of *M. urundeuva.*

As mentioned before, algorithms work with occurrence data and environmental layers to construct an ecological niche model and to predict geographic distributions. We obtained occurrence data for *M. urundeuva* from GBIF.org (www.gbif.org), where every occurrence is registered with source details, which includes the institution that registered the occurrence. Environmental layers were retrieved from

WorldClim (http://www.worldclim. org/bioclim). WorldClim offers global environmental layers that are commonly used to construct ecological niche models and potential geographic distributions. Figure 4 exhibits the potential geographic distribution obtained with an instance of the GARP algorithm included in OpenModeller. The prediction was made using occurrences within Brazil and global environmental conditions. For illustration purposes, the background shows a relief map instead of a world map.

Based on some occurrences registered in Brazil, GARP produced an ecological niche model and a potential geographic distribution for the rest of the World. Both inputs and outputs were stored in a data schema, created in PostgreSQL, using basic SQL statements for non-geospatial data and the tools *raster2pgsql* and *shp2pgsql* for geospatial data. Figure 5 shows an example of the usage of the tool raster2pgsql via command in a computer console. A detailed reference of PostGIS is available at http://postgis.net/docs/.

```
C:\>raster2pgsql -s 4236 distribution1.asc  -F -a public.potential_distribution
| psql -U postgres -d nichedb -h localhost -p 5432_
```

Figure 5. Usage of the tool raster2pgsql.

Notice that algorithms can be configured with parameters that might improve the quality of the prediction and that the purpose of this example is only to show the usage of the algorithm. More accurate and detailed tests can be done with the appropriate configurations and datasets.

5. CONCLUDING REMARKS

This paper presented the MDA stages of a database design for Ecological Niches and Potential Geographic Distributions of species. The use of MDA provides models with different levels of detail that can be transformed to a variety of lower-end implementations. In this work, we use UML GeProfile for the CIM level and the PSM level was implemented with the data types offered by PostgreSQL and PostGIS, a variation of the PSM level can be implemented with Oracle Spatial. A database implementation provides the benefit of having data for multiple species stored in a single place instead of different files. Furthermore, our approach can exploit all the advantages of a DBMS. We also presented a study case for the ecological niche and potential geographic distribution of *M. urundeuva*, an important tree species in Brazil that is both an invasive and endangered species. Occurrence data of the species and environmental layers were used as inputs for a predictive algorithm and both inputs and outputs were stored in the proposed data schema. The storage process is carried out with SQL statements and commands in the console of the computer. QuantunGIS or similar software can be used to retrieve the stored data from the database in the form of maps. The models presented in this paper could serve as a starting point for research related to ecological niches and geographic distributions in the Computer Science field. On the other hand, professionals in the field of Ecology may found in this work a suitable introduction to database and GIS technologies. Future work includes the development of software that facilities the storage process for professional and researches that are not familiar with database technologies.

6. ACKNOWLEDGMENTS

This project is partially financed by CAPES-Coordenação de Apoio ao Ensino Superior, OAS-Organization of American States, CNPq-Conselho Nacional de Desenvolvimento Científico e Tecnológico and FAPEMIG-Fundação de Amparo à Pesquisa do estado de Minas Gerais.

7. REFERENCES

Blackburn, J.K., 2010. Integrating geographic information systems and ecological niche modeling into disease ecology: a case study of Bacillus anthracis in the United States and Mexico. In Emerging and Endemic Pathogens, Springer Netherlands, 59-88.

Carmello-Guerreiro, S.M., Paoli, A., 1999. Aspectos morfológicos e anatômicos da semente de aroeira (Myracrodruon urundeuva Fr. Allem.-Anacardiaceae), com notas sobre paquicalaza. Revista Brasileira de Sementes, 21(1), 222-228.

Hutchinson, G., Concluding remarks. *Cold Spring Harbour Symposium on Quantitative Biology 22*, pp. 415–427, 1957.

Keet, C. M, 2006. Representations of the ecological niche. In WSPI 2006: Contributions to the Third International Workshop on Philosophy and Informatics, 75-88.

Lisboa-Filho, J., Nalon, F.R., Peixoto, D.A., Sampaio, G.B., Borges, K.A, 2013. Domain & Model Driven Geographic Database Design. In Domain Engineering: Product Lines, Languages, and Conceptual Models. New York: Springer Berlin Heidelberg, 375-399.

McIntosh, A., Cushing, J.B., Nadkarni, N.M., Zeman, L., 2007. Database design for ecologists: Composing core entities with observations. Ecological informatics, 2(3), 224-236.

Mellor, S.J., Scott, K., Uhl, A., Weise, D, 2002. Model-driven architecture. *In Advances in Object-Oriented Information Systems*, *Springer Berlin Heidelberg*, 290-297.

Murta, R.D.O., Mafra, R.L.M., Oliveira, F.P.D., & Coelho, F.M.G., 2012. Regional development in midle Rio Doce/MG: analysis of the legal possibility of sustainable management of aroeira. Revista Direito GV, 8(2), 455-483.

Ortega-Huerta, M.A., Peterson, A.T., 2008. Modeling ecological niches and predicting geographic distributions: a test of six presence-only methods. Revista Mexicana de Biodiversidad, 79(1), 205-216.

Peterson, A.T., Vieglais, D.A., 2001. Predicting Species Invasions Using Ecological Niche Modeling: New Approaches from Bioinformatics Attack a Pressing Problem. BioScience, 51(5), 363-371.

Pinet, F., 2012. Entity-relationship and object-oriented formalisms for modeling spatial environmental data. Environmental Modelling & Software, 33, 80-91.

Polechová, J., and Storch, D., 2008. Ecological niche. Encyclopedia of ecology, 2, 1088-1097.

Roura-Pascual, N., Suarez, A.V., Gómez, C., Pons, P., Touyama, Y., Wild, A.L., Peterson, A.T., 2004. Geographical potential of Argentine ants (Linepithema humile Mayr) in the face of global climate change. Proceedings of the Royal Society of London. Series B: Biological Sciences, 271(1557), 2527-2535.

Semwayo, D.T., 2004. Berman, S., Representing ecological niches in a conceptual model. In Conceptual Modeling for Advanced Application Domains, Springer Berlin Heidelberg, 31-42.

Soberón, J., 2007. Grinnellian and Eltonian niches and geographic distributions of species. Ecology letters, 10(12), 1115-1123.

Souza, S.M.C., Aquino, L.C.M., Milah, AC Jr., Bandeira, M A.M., Nobre, M.E.P., Viana, G.S.B., 2007. Antiinflammatory and antiulcer properties of tannins from Myracrodruon urundeuva Allemão (Anacardiaceae) in rodents. Phytotherapy Research, 21(3), 220-225.

Ward, D., 2007. Modelling the potential geographic distribution of invasive ant species in New Zealand. Biological Invasions 9.6, 723-735.

Zárate, G., Lisboa-Filho, J., Sperber, C.F, 2014. Conceptual modeling for environmental niches and potential geographic distributions using UML GeoProfile. In: International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing) 6, Barcelona, 31-37.