Network of Biodiversity Information Database System for Area-based Research, West Thong Pha Phum Project

Krisanadej Jaroensutasinee* and Mullica Jaroensutasinee
Walailak University, Nakhon Si Thammarat
*krisanadej@gmail.com

Abstract: The Network of Biodiversity Information Database System (NBIDS) for area-based research, West Thong Pha Phum project has been developed for collecting Thai biodiversity data, and providing advanced tools for querying, analyzing, modeling, and visualizing patterns of species distributions for researchers and scientists. Google Earth KML and ArcGIS were used as tools for map visualization. *webMathematica* was used for simple data visualization and also for advanced data analysis and visualization, e.g., spatial interpolation, and statistical analysis.

Key words: Thong Pha Phum, ArcGIS, Google Earth, biodiversity, database system, *webMathematica*

Introduction

A Biodiversity Database is a database for collecting biodiversity data. Biodiversity data refers to scientific information, primarily about biological species and specimens. At the species level, such data would include the scientific names of a species and all of its synonyms, the common name(s) of the species, and other information about the species, such as a description of the species, its physiological properties, genetics, geographic distribution, phylogenetic relationships, role in the dynamics of ecosystem processes including cases of invasions, applications, etc. Specimen-level data including samples for molecular analysis, would include the scientific name of the species to which the specimen belongs, information on where, when and by whom the specimen was collected, where the specimen is currently located, who identified it, what is the specimen number, and other associated information derived from the specimen (e.g., living culture, frozen tissues, photographs, parasites, and hosts) and any other related field notes written by the collector of the specimen.

Because of humanity's dependence on natural systems, information about biodiversity and ecology is vital to a wide range of scientific, educational, commercial, and governmental uses. Biodiversity and ecosystems are themselves interdependent. Ecosystems and the diversity of species they support underpin our lives and our economies in very real, though often underappreciated, ways. The living things with which we share the planet provide us with clean air, clean water, food, clothing, shelter, medicines, and aesthetic enjoyment. Yet, increasing human populations and their activities are disturbing species and their habitats, disrupting natural ecological processes, and even changing climate patterns on a global scale. There are greater stresses on the natural world than humanity has ever generated in the past. Since biodiversity is arguably the most precious resource on Earth, it is becoming more and more important that we actively conserve biodiversity and protect natural ecosystems in order to preserve the quality of human life. As human populations and their demands on the natural world grow, our accumulated knowledge about biodiversity and the environment will become ever more important in the effort to develop a sustainable world.

Recognition of this has led to the National Biological Information Infrastructure in the United States, to the Environmental Resources Information Network in Australia, and to a number of regional biodiversity information networks (NABIN, IABIN, EIOnet, and others). Indeed, the recommendation by an international working group established by the Global Science Forum (formerly Megascience Forum) of the Organization for Economic Cooperation and Development (OECD) that the nations of the world establish and maintain a Global Biodiversity Information Facility (GBIF), which is poised to become a reality in early 2001, is a direct outgrowth of both concern about the environment and the

economy, and the acknowledgment that the complexity of biodiversity and ecological datasets reflects the complexity of natural systems. It has become apparent that practitioners in the computer science and information technology fields must become as invigorated by and invested in the biodiversity and ecological information domain as are the biologists, who collect, generate, query, and interpret the data (Chefaoui et al., 2005; Lane et al., 2000).

The Network of Biodiversity Information Database System (NBIDS) has been developed by a Walailak University team and funded by the Biodiversity Research Training Program (BRT). The goal of this project is to provide advanced tools for querying, analyzing, modeling, and visualizing patterns of distributions of species found in Thailand for researchers and scientists.

Methodology

NBIDS is a web based system designed with four main features: database, data analysis tool, data visualization tools, and GIS tools. NBIDS database is developed using SQL technology. We have developed web-based tools for data entry and data access. NBIDS data record two types of datasets: biodiversity data and environmental data. Biodiversity data are species presence data and species status. The attributes of biodiversity data can be further classified into two groups: universal and project-specific attributes. Universal attributes are attributes that are common to all of the records, e.g. X/Y coordinates, year, and collector name. Project-specific attributes are attributes that are unique to one or a few projects, e.g., flowering stage. Environmental data include atmosphere, hydrology, soil, and land cover data using GLOBE protocols.

Data analysis tools for NBIDS are statistical analysis tools and computational modules for each research project. Examples of computation modules' outputs are biodiversity index, mosquito house index, coral data and fish morphometric data.

The data visualization tool was developed using *webMathematica* technology (Wolfram, 2003). This tool is an interactive tool for visualizing graphs using the high performance computing power of *Mathematica* software.

Google Earth is well used for GIS visualization. Other information is added to the default Google Earth map. This information

Figure 1. Google Earth used for GIS visualization

includes landmarks, administration maps, transportation, LandSat images, and geo-computing data. Geo-computing data have been developed using the ArcGIS program. These data include DEM, aspect, and flow direction (Fig. 1).

NBIDS has five user types: system manager, project manager, researcher, senior scientist, and system administrator. A project manager is a principle investigator of each project. A project manager collects data in the field and inputs data on the website. A project manager views and makes changes to his own data. A system manager and senior scientist have access to and view data from all projects. However, only a system manager and system administrator can make some changes and modification to all the data and system.

Results

A prototype NBIDS is now online at URL http://www.nbids.org since November 2005. Data from west Thong Pha Phum projects have been uploaded on to the NBIDS website.

Data

Now NBIDS west Thong Pha Phum project contains 51 sub-projects.

Web Tools

We can search for information from NBIDS such as study site, species name, common name, family, physical parameters and date (Fig. 2).

We can do data visualization in NBIDS using *webMathematica* in bar charts, line and pie charts in terms of the number of species, the number of common names and the number of families at study sites (Fig. 3A-C).

NBIDS shows the latitude and longitude of study sites on Google Earth, the number of species and the species present at the study sites (Fig. 4). NBIDS offers online data cleaning. The researchers can make some corrections on the web.

NBIDS and Assessing Habitat-Suitability Models

Prediction of species distribution is an important element of conservation biology. Management for endangered species (Palma et al., 1999; Sanchez-Zapata and Calvo, 1999), ecosystem restoration (Mladenoff et al., 1997), species re-introductions (Breitenmoser et al., 1999), population viability analyses (Akçakaya and Atwood, 1997; Akçakaya et al., 1995) and human–wildlife conflicts (Lay et al., 2001) often rely on habitat-suitability modeling. Multivariate models are commonly used to define habitat suitability and, combined with geographical information systems (GIS), allow one to create potential distribution maps (Guisan and Zimmermann, 2000). Numerous multivariate analyses have been developed for building habitat suitability or abundance models in the past decade (Lek et al., 1996; Manel et al., 1999; Özesmi and Özesmi, 1999; Hirzel et al., 2002).

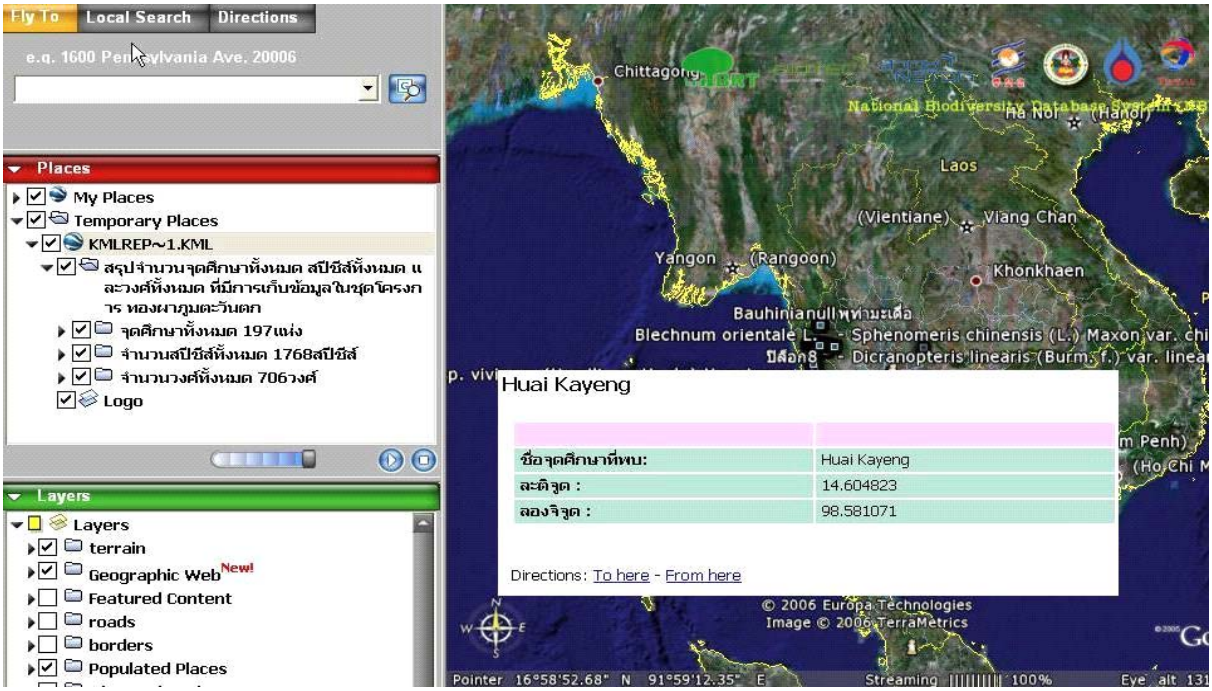
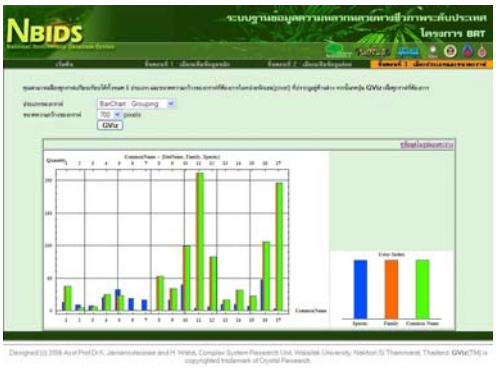
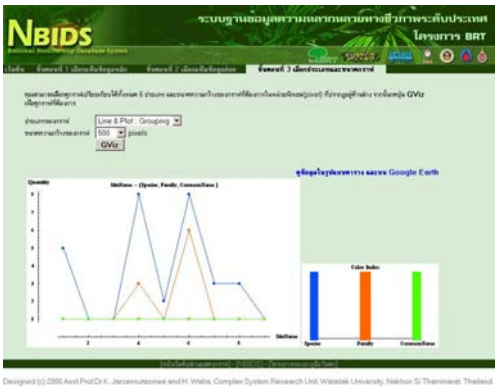


Figure 2. Example of NBIDS search page

(A)



(B)



(C)

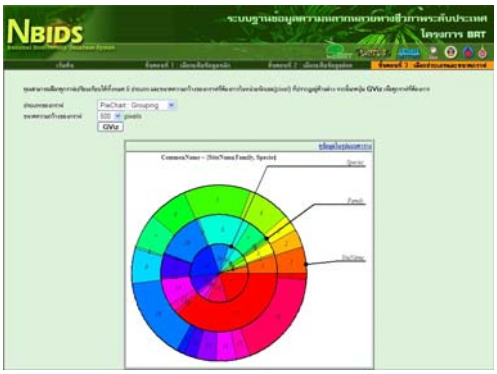


Figure 3. NBIDS visualization tool, interactive graph and descriptive statistics using *webMathematica* (A) Bar chart, (B) Line, and (C) Pie Chart of species, common name and family name

Ecological niche factor analysis (ENFA) (Hirzel et al., 2002) is a heuristic modeling approach recently developed to predict potential species distributions from presence-only data. Modeling with ENFA is usually done by using the software Biomapper (Hirzel et al., 2002; Hirzel et al., 2004; Chefaoui et al., 2005). In this study, we rewrote the ENFA program with *Mathematica* (Wolfram, 2003) which is a mathematical and statistical package with visualization tools. We tested our ENFA program with virtual species data and real eco-geographical and climatic data.

Virtual ecological niche: the 'true' habitat suitability map

On this spatial canvas, the virtual species was generated by creating a simulated ecological niche in an *n*-dimensional space (Hutchinson, 1957). It was modeled by a niche coefficient H ($H \in [0,1]$), which can be viewed as the probability that each cell belongs to the niche; note that H is a de facto habitat suitability index. This value was built as summarized in Equation (1).

$$H = \frac{1}{\sum w_i} \sum w_i H_i + \varepsilon \qquad (1)$$

where

H is the habitat suitability of the focal cell

H_i is the value of the i^{th} partial niche coefficient

w_i is the weight assigned to the i^{th} partial niche coefficient, and

ε is a random value.

Global habitat suitability is composed of a weighted average of partial niche coefficients (H_i) and a stochastic coefficient (ε). The partial niche coefficients are the habitat suitability engendered by each predictor value. They are computed from four predictors that are picked out of the nine available predictors by four niche functions (i.e. elevation with Gaussian function, aspect with Gaussian function, the amount of rainfall with truncated linear, and minimum air temperature with decreasing linear function).

Three types of functions are used to model three types of environmental optima: 1) a Gaussian function to model a median optimum, 2) a linear function to model an extreme optimum, and 3) a truncated linear function to model a buffer zone effect. Each of these H_i values is then weighted by a w_i factor and the global niche coefficient is calculated as their weighted average. Finally, a random term ε, generated from a uniform distribution in the range [−0.05, 0.05], is added. The niche-function parameters and the weights are arbitrarily tuned in order to generate about 50% of cells with $H \geq 0.5$.

This produces the 'true' habitat suitability map (Fig. 5), representing the 'real' intrinsic preferences of our virtual species. By 'true' map, we mean that it represents the kind of information usually unreachable by ecologists, the information they are trying to reveal through field sampling and statistical analysis. The 'true' map will be constantly used

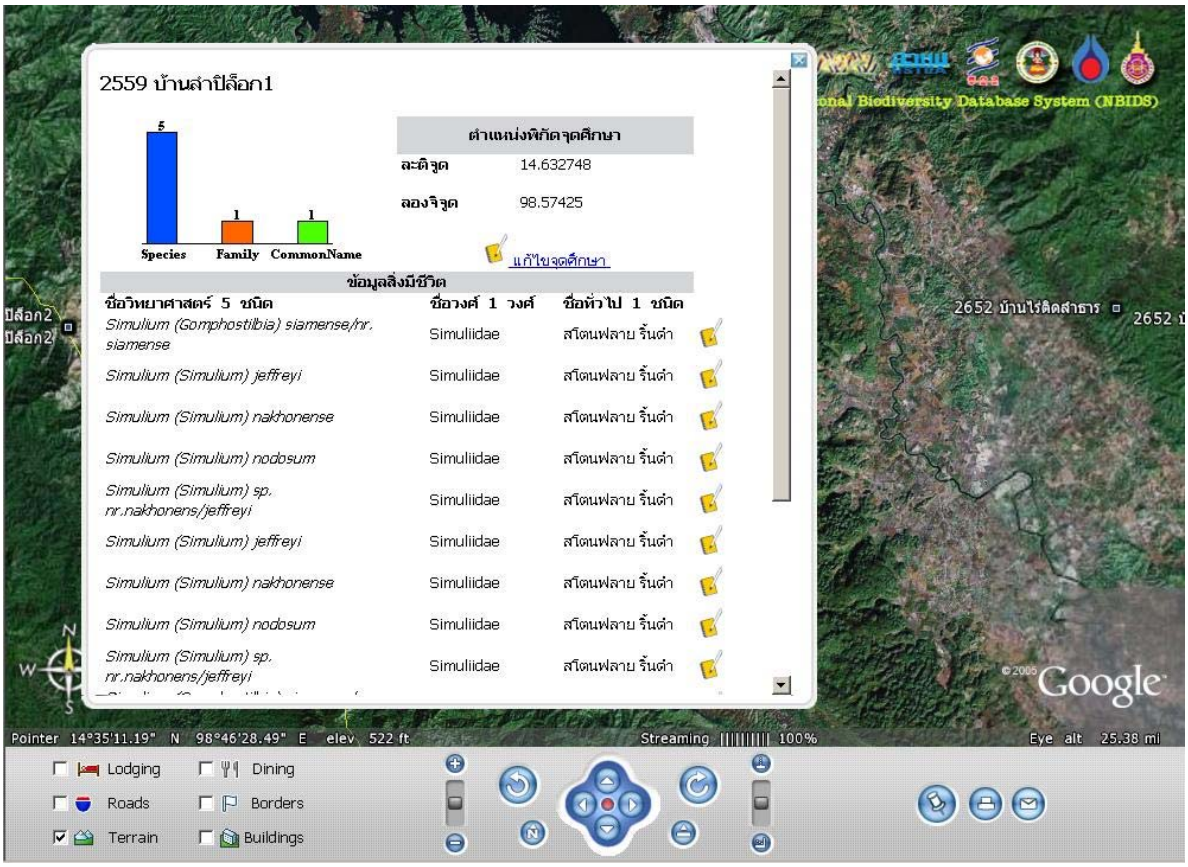


Figure 4. NBIDS visualization tool demonstrating coordinates and species present at a study site

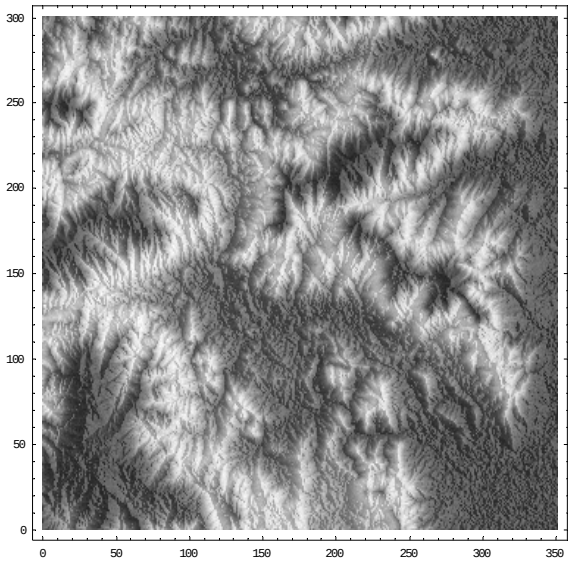


Figure 5. The 'true' habitat suitability map generated to model the ecological niche of the virtual species. High suitability areas are indicated by white pixels.

as a basis to generate data and as a reference to assess the accuracy of habitat suitability analyses.

Distribution map

Distribution maps are computed on the basis of the 'true' map; the distribution maps give the 'true' presence/absence of the virtual species, information usually unavailable to field ecologists. Three distribution scenarios are addressed in order to determine the advantages and drawbacks of each habitat suitability analysis. They can be viewed as three historical phases of colonization—the fundamental niche does not change but the realized one does:

1) a 'spreading phase' showing a density gradient from the north-west corner of the map to the south-east corner

2) an 'equilibrium phase' where the species are abundant enough to occupy all the available suitable areas

3) an 'overabundance phase' where the species are so numerous that it has to spread in to less suitable areas (Fig. 6).

The 'equilibrium' distribution map is computed as follows. To each cell of the 'true' habitat suitability map is added a random value taken in the range [−0.2, 0.2] (uniform distribution); this is made in order to introduce some stochasticity into the model. If the resulting habitat suitability coefficient is larger

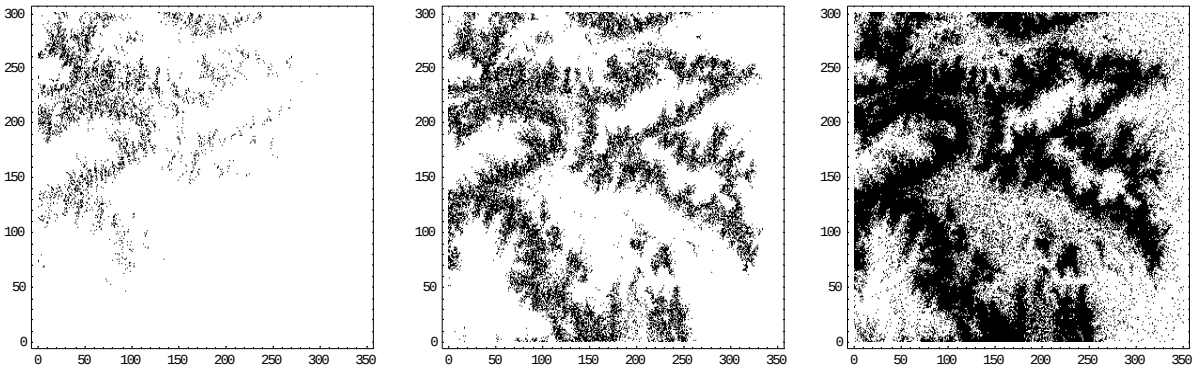


Figure 6. distribution maps of the virtual species for three colonization scenarios. Black points are the cells where the species is present and the white ones are those where the species is absent. Map (A) represents the 'spreading' scenario: the species entered the area from the northwest and is currently propagating in all directions, settling down in the most suitable areas. Map (B) represents the 'equilibrium' scenario in which the species occupies uniformly all the suitable areas. Map (C) represents the 'overabundance' scenario in which very high densities force the species to occupy less adequate areas.

than 0.7, the cell is marked as occupied.

The 'overabundance' distribution map is computed in a similar way but with a 0.5 habitat suitability threshold to simulate the overflowing density.

The 'spreading' distribution needs an additional operation: each cell of the 'true' habitat suitability map is beforehand multiplied by a value decreasing as $1/d^2$, d being the distance to a point arbitrarily placed north-westward to south-eastward corner of the map. This gradient function is tuned to produce values ranging from 0 to 1, 0.5 lying approximately in the middle of the map. This new gradient map is then submitted to the same operations as the 'equilibrium' scenario (habitat suitability threshold = 0.7). This generating method allow us to obtain distribution maps with a presence density correlated with area suitability.

Discussion and Conclusion

GIS tools of NBIDS can help scientists and researchers to plan their research because the tool developed is compatible with Google Earth which is easy to use. This Google Earth can demonstrate maps, and LandSat images. With these pictures, NBIDS can help researchers to understand an area, select their study sites effectively and plan their experiments appropriately. When scientists are doing their research, they can use this GIS tool for observing and constructing some relationship between geographical data, environmental data, and species presence data. Furthermore, scientists could model niche characterization and potential distribution of species using some mathematical and computational methods. Tools for mathematical modeling are planned to be added to NBIDS in the near future. NBIDS is an effective tool for studying the relationships among species. All NBIDS data are stored with the same universal attributes that make these data comparable. For example, coordinates of species occurrence are collected in the same units that make the study possible and luminous. NBIDS data are stored in a security system. Only permitted users can access their own data. However, when scientists need to compare the relations among species, permission for accessing data can be granted by the Principle Investigators of the projects.

Acknowledgements

This work was supported by the TRF/BIOTEC Special Program for Biodiversity Research and Training grant, PTT Pubic Company Limited, and Complex System Research Unit, the Institute of Research and Development, Walailak University.

References

Akçakaya, H.R. and J.L. Atwood. 1997. A habitat-based metapopulation model of the California Gnatcatcher. *Conservat. Biol.* 11: 422-434.

Akçakaya, H.R., M.A. McCarthy and J.L. Pearce. 1995. Linking landscape data with population viability analysis: management options for the helmeted honeyeater *Lichenostomus melanops cassidix*. *Biol. Conservat.* 73: 169-176.

Breitenmoser, U., F. Zimmermann, P. Olsson, A. Ryser, C. Angst, A. Jobin and C. Breitenmoser-Würsten. 1999. Beurteilung des Kantons St. Gallen als Habitat für den Luchs, KORA, Bern.

Chefaoui, R.M., J. Hortal and J.M. Lobo. 2005. Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. *Biol. Conservat.* 122: 327-338.

Guisan, A. and N.E. Zimmermann. 2000. Predictive habitat distribution models in ecology. *Ecol. Model.* 135: 147-186.

Hirzel, A.H., J. Hausser and N. Perrin. 2004. Biomapper 3.1, Lab. of Conservation Biology, Department of Ecology and Evolution, University of Lausanne. Available: http://www.unil.ch/biomapper

Hirzel, A.H., J. Hausser, D. Chessel and N. Perrin. 2002. Ecological-niche factor analysis: how to compute habitat suitability maps without absence data? *Ecology* 83(7): 2027-2036.

Hutchinson, G.E. 1957. Concluding remarks, Cold Spring Harbor Symposium. *Quantitative Biol*. 22: 415-427.

Lane, M.A., J.L. Edwards and E.S. Nielsen. 2000. The Challenge of Rapid Development, Large Databases and Complex Data. *Proc. 26th Inter. Conf. Very Large Databases*, Cairo, Egypt.

Lay, G. Le, P. Clergeau and L. Hubert-Moy. 2001. Computerized map of risk to manage wildlife species in urban areas. *Environ. Manage.* 27: 451-461.

Lek, S., M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga and S. Aulagnier. 1996. Application of neural networks to modelling nonlinear relationships in ecology. *Ecol. Model.* 90: 39-52.

Manel, S., J.M. Dias, S.T. Buckton and S.J. Ormerod. 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *J. Appl. Ecol.* 36: 734-747.

Mladenoff, D.J., R.C. Haight, T.A. Sickley and A.P. Wydeven. 1997. Causes and implications of species restoration in altered ecosystems. A spatial landscape projection of wolf population recovery. *Bioscience* 47: 21-23.

Özesmi S. L. and U. Özesmi. 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecol. Model.* 116: 15-31.

Palma, L., P. Beja and M. Rodgrigues. 1999. The use of sighting data to analyse Iberian lynx habitat and distribution. *J. Appl. Ecol.* 36: 812-824.

Sanchez-Zapata, J.A. and J.F. Calvo. 1999. Raptor distribution in relation to landscape composition in semi-arid Mediterranean habitats. *J. Appl. Ecol.* 36: 254-262.

Wolfram, S. 2003. The Mathematica Book, 5th ed., Wolfram Media.