An Object Model for Biodiversity Analysis

BORIS MILAŠINOVIĆ¹, TONI NIKOLIĆ², KREŠIMIR FERTALJ¹
¹Faculty of Electrical Engineering and Computing, ²Faculty of Science
University of Zagreb
¹Unska 3, 10000 Zagreb, ²Marulićev trg 9a, 10000 Zagreb
CROATIA
{boris.milasinovic, kresimir.fertalj}@fer.hr, toni.nikolic@botanic.hr

Abstract: - Biodiversity as the variety within the living species is commonly defined in relation to specific spatial unit. The number of geocoded localities rises significantly and considerable research has been devoted to mapping and analyzing distribution of flora. The paper describes an object model that addresses spatial distribution of biodiversity at any taxonomic rank. The model can be used as a part of a web service or as a layer in an application. The model enables pairing data from taxa localities and data from GIS layers thus allowing new knowledge about biodiversity. Types of analysis and results, input and output interface have been described.

Key-Words: - Biodiversity, Spatial analysis, Alpha diversity, Ecological niche, Object model

1 Introduction

"Biological diversity" or "biodiversity" is variety within the living world, genes, individuals, species, ecosystems [8]. This variety is commonly defined in relation to specific spatial unit, from very small area up to the whole World. One of the basic information on biodiversity and probably the most commonly used refers to the number of taxa that occur in the defined area [10]. The biodiversity is essential for ecological stability on planet Earth, maintenance of the biosphere in a state supportive of human life, and between others, as necessary source of material basis for humans (i.e. food, shelter, medicine, chemicals, pharmaceuticals, etc.). Information on species richness is often used for defining the so-called hot-spot areas [17][18][24], the areas particularly important as a background towards the implementation of conservation programs [2][23] and other biogeographic analysis (i.e. [14]). The same information, however, has a great significance in the scientific interpretation and understanding of natural laws that affect the distribution of biodiversity (i.e. [1][3][5]).

At the same time it is believed that each year 40,000 species disappear, among which there are mainly those that have not even been described yet. Together with them the possibilities of their application in medicine, agriculture or forestry, so far unexplored, disappear too. The historical United Nations Conference on Environment and Development (UNCED) held in Rio de Janeiro in 1992 was undoubtedly the turning-point in the approach to the protection of nature and environment. On that occasion 157 countries signed the Convention on Biological Diversity - the most recent global step towards a comprehensive protection and sustainable use of natural resources. Its basic objectives are preservation and improvement of the existing biological diversity, as well as the economical use of natural resources on the principles of sustainability [16]. Owing to its unprecedented response in the world the Conference gave a powerful encouragement to nature protection and contributed to a proper appreciation of this problem area. A vast number of legislative acts around the world are adopted at different levels, and an incalculable number of practical activities around the world have been launched (see World Conservation Monitoring Centre or IUCN web sites, www.iucn.org) with the aim of preserving biodiversity, sustainable use and equitable and uniform distribution of benefits arising from the use of biodiversity as sources.

The object model described in this paper addresses spatial distribution of biodiversity at any taxonomic rank (i.e. species, genus, family, etc.). The model uses geocoded data on species findings, prepares input data, analyzes species distribution, calculates diversity, helps to determine species ecological profile and ecological niche and enables spatial analysis using GIS tools. As a study area the Croatian territory has been chosen which is further elaborated in the second section. The model has defined input and output types and interfaces and can be part of a web service or a layer of an
application. Model structure is described in the third section followed by description of types of analysis and results in the fourth and the fifth section. The paper concludes with guidelines for future development and research in this area.

2 Study area
As a study area the Croatian territory was selected. Located within 4 of the 11 recognized European bio geographical regions: Alpine, Mediterranean, Continental and Pannonian [22] it is characterized by outstanding biodiversity [13][16][19][20]. As a test data, data on distribution of vascular plants is used from Flora Croatica Database (FCD) [6]. Currently more than a half million of localities data is geocoded although geocode precision varies within eleven categories ranging from low accuracy (i.e. county level) to GPS precision. It is expected that this number will rise significantly as only minor parts of herbarium collections in Croatia’s museums have been digitalized due to lack of appropriate hardware and lack of funding. Also field experts would be equipped with newly developed Android application (in stage of testing) for support of field observations that will allow them to enter data from field observations during observations and directly copy data to FCD. Gradually increasing the amount of accumulated geocoded findings has provided the basis for the spatial evaluation of biodiversity [3][13][20] but until the model have been developed the largest part of analysis had to be been done manually.

Species ecological indices (anatomy, life form, pollination type, humidity dynamics, etc.) have been used for analysis. Several climate layers were used providing information about temperature, rain, etc. also as several layers with regional data like county borders, geographical regions, etc. producing new knowledge about species leading to a few research papers in review process or already published (i.e. [21]).

3 Model Types and Interfaces
Fig. 1 shows main class diagram of the model. It contains both classes used for data manipulation and input and output. Input and output are modeled using interfaces thus making model available in various usage scenarios as a web service or a layer in an application. As it only defines structure of the input data, the model is independent of concrete data loaders. An analysis can be performed on one or more thematic GIS layers either already stored on a server, or those uploaded by users, where currently only ESRI shape format [4] with polygons is supported. Each polygon from an ESRI shape file (shp) is paired with corresponding record in a database file (dbf) having several columns (in further text attributes) that provide additional information about the polygon. Only one attribute simultaneously can be chosen for an analysis.

As a prerequisite for an analysis, localities of species findings have to be joined with polygons from the layer(s) thus forming JoinedShapeRecords. This task of spatial join is equivalent to determining whether a particular point is inside a polygon. As summarized in [9][12] two basic concepts for solving this problem are known in literature: the even–odd rule (ray-crossing method) and the winding number (angle summation algorithms). For its simplicity an algorithm using angle summation algorithm from [12] have been chosen during model implementation. As significant number of findings could have same coordinates (especially those points of lower precision) it would be inefficient to determine multiple times if the same point belong to a polygon. Due to this all finding points are merged in a list of Localities. A Locality inherits Point and contains a reference to all finding points at the same spot. A finding point contains coordinates (inherited from Point class), taxon identifier, geocode precision, finding source, finding identifier and a year of the finding. Term taxon is used instead of species allowing model to work not only with species but with any taxonomy type.

After spatial join had been done, several values for a polygon or set of polygons from a layer are calculated. Results are returned as one or more data matrices and/or by extending dbf file of the GIS layer used in the analysis. A user can make two different types of analyses: ecological niche analysis and biodiversity analysis explained in detail in the fourth and the fifth section. Depending on the request type and the type of analysis, two categories of functions are used to produce results: single record functions and grouped record functions.

Single record functions return data on a single polygon from a shape file. Calculated values are included in the result and also can be used to extend dbf file with new attributes, i.e. how many field observation have been performed in the polygon or how many different species exists in the polygon, etc. The number of attributes varies on chosen options and number of possible values, i.e. on number of ecological indices and number of possible indices values. The extended dbf file is returned to the user (with unmodified shp file and spatial index file) thus making a new layer that can be used in GIS tools.
Grouped records functions are functions that return summary data for a group of polygons having same attribute value. I.e., a habitat type can be formed of multiple polygons in the shp file and multiple records in the corresponding dbf file with the same attribute value in column that defines habitat type. If a user is interested to find how many findings were noted in each habitat type he/she wants summarized data from all polygons belonging to the habitat. These data are returned only in matrices form.

![Diagram of object model](image)

**Fig. 1.** Main classes of the object model

### 4 Ecological niche analysis

Purpose of ecological niche analysis is to distribute chosen species per attributes from one or more GIS layers. For the simplicity of the first version of model only layers stored on server can be used. For each layer the chosen attribute would be used for grouping data and producing summary values. Except names of chosen GIS layers and attribute names, input parameters for the calculation are findings filters consisting of species names or part of the names, endemic and invasive status, year of finding, locality source and geocode precision. For each input layer localities for chosen species are joined with polygons from the layer. Afterwards model implementation produces summary data per attribute value. Result of this type of analysis is collection of matrices where all matrices have same number of columns and column names represents
species names. Number of matrices in collection is equal to the number of input layers. Row names in a matrix represent possible attribute values of a chosen attribute in layer. A value in the matrix produced for layer \( l \) in row \( a \) and column \( s \) counts how many polygons from layer \( l \) having attribute value \( a \) contain species \( s \).

As values for chosen attribute can be continuous this type of analysis is usually paired with further data processing where attribute values are grouped in user defined ranges and then shown on a graph like in Fig. 2 where average temperature is chosen as an attribute for analysis of three species. Rather than observing absolute numbers, distribution pattern have to be examined, because not all species have same outspread or not all of them are noted so frequently.

![Fig. 2. Example of data from ecological niche analysis grouped in user-defined ranges](image)

In addition to that, it has to be noted that interpretation of these results depends on type of layer used for analysis. For instance, if each attribute is consisted of only one polygon then described function is characteristic function (thus can have only values 0 or 1 on y axes) showing whether a species is present in an area with the particular attribute value. Another useful result can be produced if a layer is consisted of environmental data divided in grid cells of equal size. In this case matrix value represents how many cells with environmental value \( a \) contain species \( s \). If there are several weather and land type layers this provides foundation for forming ecological niche of a species. Misinterpretation occurs if polygons are of different size. I.e., an attribute \( X \) can be consisted of one large polygon and an attribute \( Y \) can be consisted of many small polygons with a total area even smaller than those covered by \( X \). The maximum result for species \( s \) and the attribute \( X \) can only be 1 and for species \( s \) and the attribute \( Y \) can be equal to number of polygons of \( Y \) and such values cannot be compared and used for any analysis.

5 Biodiversity analysis

Biodiversity analysis works with one GIS layer. Input parameters for biodiversity analysis are similar to those in ecological niche analysis with remark that, beside publicly available layers, a user can upload own layer. Data returned to a user contains modified ESRI shape files and three matrices: BioDiversityData, CheckList and TaxaData matrix. Modified ESRI dbf file (with a polygon data in each row) and a BioDiversityData matrix where row names are all possible attribute values are extended/filled with integer values returned by set of functions belonging to the following categories: alpha diversity, research intensity and categories data, where all three include both single record and grouped records functions. CheckList matrix represents a species checklist where row names are formed of attribute values and column names are formed of species names. Each value in the matrix represents number of findings of a species in an area belonging to an attribute. The third matrix gives additional data about species and it is not related to a particular polygon but rather to a whole layer and species that are noted in the layer.

5.1 α - diversity

\( \alpha \)-diversity measures the number of different species in an area. Biologists have used several slightly different definitions for \( \alpha \)-diversity. In this paper, the term means species diversity in a single spatial unit [13], i.e., number of species occurring within an area of a given size [11] or the species richness of a single sampling unit [7]. This means that \( \alpha \)-diversity has to be calculated both for each polygon from a shape file (and it would be calculated as a number of different species in that polygon) and for each attribute (number of species noted in all polygons of same attribute). For the latter calculation distinct union of all findings has to be done as a species can be present in more than one polygon of the same attribute but have to be counted only once per attribute value.

5.2 Research intensity

Research intensity (number of records per spatial unit) is calculated separately for each locality source: field observations, literature references, herbarium collections and users’ photos. It is calculated both for a single polygon and as a summary for all polygons having same attribute value. Research intensity for a source in a particular area is formed of number of findings and count of unique finding identifiers from the source in that area. I.e. in an area it could be \( m \) findings from \( n \)
field observations and research intensity for observations in that area is pair \((m, n)\). Thus research intensity for a locality source \(ls\) and a polygon \(p\) is pair \((x, y)\) where \(x\) is number of findings from source \(ls\) inside polygon \(p\) and \(y\) is number of unique finding identifiers among those findings.

Using research intensity relation between number of terrain expeditions and number of findings can be established. Such information can help field experts to determine whether is worth going in some areas where someone has already been and to determine how many time field experts have to go to terrain to be sure (within a statistical error) that already all species in some area have been noted.

To calculate value \(x\) for research intensity of a source \(ls\) and an attribute \(a\) sum of finding counts of all polygons having attribute value \(a\) is taken. However, the number \(y\) cannot be calculated by summing \(y\) values of each polygon from the group as one field observation could be done in more than one polygon. Therefore, distinct union of all finding identifiers has to be done, and \(y\) is cardinality of the union.

5.3 Categories data

Functions related to categories data provides analysis similar to analyzing ecological niche of a species, but this time the accent is not on species. Purpose of this analysis is to join each polygon with each possible value from a set of categories (i.e. categories can be formed of possible ecological indices values). Assigning a number \(N\) for a polygon \(p\) and a value \(V\) where \(V\) is one of the possible distinct values in a category \(C\) means that in the polygon \(p\) there are \(N\) species that have value \(V\) for category \(C\).

Similar calculation can be done for a set of polygons having same attribute value with notable difference that there are two valid approaches. The first one is to count different species and the second is to sum values already assigned to a particular polygon from a set. Both approaches have biological explanation. Data about number of different species that are located in some area maybe will not give enough information as information gained by summing individual values of each polygon from the group. I.e. suppose that a layer with climate data is formed of \(1\text{km}^2\) cells and that area where temperature is -5 is present in \(m\) cells and \(m\) species is distributed in those \(m\) cells in such way that neither two species are in the same cell. Then, the number of species is \(m\) and total sum is also \(m\). But, if every species is present in every of \(m\) cells then number of species is \(m\) but the total sum is \(m^2\) which brings additional information to interpretation of results.

6 Conclusion

In recent years, considerable research has been devoted to mapping the flora distribution, spatial analysis and biodiversity calculation. Complex analyses were performed manually and thus work intensive. Developed model enables pairing data from species localities and data from GIS layers, makes some analysis automated and eases complex analysis by providing many new aggregated data. Thus it enables new knowledge about species in Croatia. There are currently a few new research papers in revision process addressing spatial distribution of endemic, threatened and invasive species in Croatia and its relationship to conservation efforts, determining ecological niche of the species, locating hot spots and under investigated parts of country and finding patterns for species distribution. These parts of botanic research in Croatia have been insufficiently examined and this model enabled further research in those fields. However, it is important to note that interpretation of these results depends on type of layer used for analysis and species outspread.

Improvements of the proposed model could be achieved by developing various input methods using web interfaces and exposing model as a service to wider specter of users. Due to processor intensive calculation, a cloud solution would be appropriate if the number of users would rise.

References:


