Automated pre-processing strategies for species occurrence data used in biodiversity modelling

Article

Accepted Version

Automated pre-processing strategies for species occurrence data used in biodiversity modelling

Marshall J. Heap[1] and Alastair Culham[1]

Center for Plant Diversity and Systematics, School of Biological Sciences, University of Reading, Whiteknights, PO Box 217, Reading, Berks, RG6 6AS, U.K.

Abstract. To construct Biodiversity richness maps from Environmental Niche Models (ENMs) of thousands of species is time consuming. A separate species occurrence data pre-processing phase enables the experimenter to control test AUC score variance due to species dataset size. Besides, removing duplicate occurrences and points with missing environmental data, we discuss the need for coordinate precision, wide dispersion, temporal and synonymity filters. After species data filtering, the final task of a pre-processing phase should be the automatic generation of species occurrence datasets which can then be directly 'plugged-in' to the ENM. A software application capable of carrying out all these tasks will be a valuable time-saver particularly for large scale biodiversity studies.

Key words: Biodiversity richness; environmental niche modelling; pre-processing species occurrence data; automated filtering and occurrence file generation

1 Introduction

The term *biodiversity* is used in literature to describe the variety of biological organisms present within a specific geographic extent. High biodiversity is generally considered synonymous with a healthy ecosystem [1]. Most biodiversity studies are concerned with understanding and mitigating biodiversity loss. The most common causes of biodiversity loss amongst plant species, include; land-use change, climate change, atmospheric gas composition change, soil damage and the spread of invasive species [2]. Species richness, the number of species present in a specified geographic area, is the most commonly used measure of biodiversity [3]. Where species richness is estimated using Environmental Niche Models (ENM's), the typical approach is to model distribution probabilities for each species and then sum these probabilities for each grid cell in the chosen geographic extent [4]. ENMs combine species occurrence data with layers of environmental data in raster format. Species occurrence data consists of geo-referenced occurrence points and the species' scientific name. Environmental data consists of raster layers that define both geographic extent and the spatial resolution of the modelled species probability distributions. A machine learning algorithm (e.g. neural network, genetic algorithm, maximum entropy etc.) first finds the environmental grid cell values corresponding to each species occurrence point and

then divides pattern space between environmentally suitable/unsuitable habitat. Computer-based ENM's are split between two funadamental approaches; namely, presence only (PO) and presence/absence (PA) models. They are distinguished by the inclusion of geo-referenced absence data in PA models. However, since absence data is often difficult to obtain, a pseudo presence/absence (PA) generative approach, is often used instead. MAXENT [5] is an example of an ENM that uses this approach - computing its own absence data by drawing random data samples from the models environmental layers excluding species occurrence locations.

Model accuracy is measured by omission and commission error. The extrinsic omission rate (also known as sensitivity) represents the fraction of test species samples located in an unsuitable environment i.e false positives. Commission error (also known as specificity) is the fraction of absences falling in a suitable environment i.e. false negatives [6]. Model performance is measured quantitatively by the area under the receiver operating characteristics curve (AUC of ROC curve, henceforth referred to as AUC) which is a threshold independent method. Creating the ROC requires division of species point data between training and test data sets. Essentially, AUC measures the probability that a presence location is ranked higher than a random background location [7]. Maximum AUC values are close to, but less than 1, with 0.5 representing a prediction no better than random. While the Test AUC score indicates good model performance, it does not indicate if the right model has been built. Here reliance is placed on expert knowledge in choosing the right combination of environmental layers combined with expert analysis of the resulting probability distribution to see if it is a fair representation of the species' fundamental niche.

An important source of species occurrence data for global and other spatially extensive ENM's is the Global Biodiversity Information Facility (GBIF) [8]. Some GBIF data statistics, as of 1st March, 2010, are:

- 198,721,699 Species occurrences for all kingdoms
- 51,572,239 Plantae occurrence records
- 39,184,950 Plantae occurrence records with a geo-reference
- 31,994,765 Plantae occurrence records with geo/temporal references
- 579,946 Plantae species with 1 or more occurrence records
- 186 Institutions contributing Plantae data to GBIF

In a recent European plant biodiversity study, although 1,350 species were modeled they represented only a fraction of the number of species to be found there [9]. To obtain a more accurate map of European plant diversity will therefore require modelling the highest possible number of species. GBIF offers various ways of downloading Plantae data including by country and by geographic extent. However, as there are limits on the maximum file size of downloads (up to 250,000 occurrence records, depending on file format selected), several such downloads maybe required which must then be consolidated into a single file. The consolidated csv file for a European plant diversity study will consist of several million records. This is too large to be opened by standard spreadsheet packages. In fact, due to the need to filter out several types of data error, it makes sense to

develop custom programs in Java ® [10], for example, to do this. Another motive for automating the extraction of species occurrence datasets from the consolidated file is that preparing these datasets manually takes at least one hour per species based on our experience. Lets say that we wish to do a biodiversity study modelling 6,000 species. To prepare species occurrence datasets manually would take one person around three years whereas automating the process will take around 200 hours to develop the custom programs and, at most, another few hours of computing time to execute them. Earlier, we mentioned the need to filter out data error. In this paper, we will describe these errors and stipulate requirements for custom species occurrence data extraction tools that will, to the extent possible, pre-process the raw data filtering out these errors before preparing species occurrence point data files that can be 'plugged-in' directly to an ENM.

2 Taxonomic disambiguation

The taxonomic structure for GBIF is the 2007 copy of the Catalogue of Life a database updated annually [11]. The catalogue has grown by quarter of a million species since 2007 but is not complete, nor can it ever be, due to the thousands of new names published each year. However the coverage is close to 66% at species level for all named life and much higher at genus level. The Global Strategy for Plant Conservation [12] has, as its first target: A widely accessible working list of known plant species, as a step towards a complete world flora to be established by 2010. This shows the importance placed on the completion of a catalogue of life. Construction of such a catalogue offers many challenges, not least the lack of funding to support taxonomic expertise to populate the catalogue. Even if the catalogue was complete there are complexities to the application of scientific names to living things. As well as the routine synonymy of names that form accepted species within a kingdom, there are separate codes of nomenclature for plants [13], animals [14] and microbes [15] and the consequence is that organisms in different kingdoms can have the same valid name providing they are covered by different codes. This can lead to problems of synonymy (the use of the same name more than once) in global databases of life because the component databases, if they are purely for plants, animals or microbes, may well not include information on Kingdom. Resolving these conflicts and ensuring names are applied to the right things is called disambiguation. Without this process the result is that erroneous records can be gathered in an automated search of a data portal. There are many examples of such synonymy causing problems. A study by Culham & Yesson (in press) [16] showed that automated retrieval of data for a family of tropical timber and fruit trees, Ebenaceae, yielded 21,000 data points of which 11,000 were in the Atlantic ocean! This was because the genus *Paralia* is used both in the Ebenaceae and for a genus of phytoplankton [16]. Other such examples are cited by Page [17, 18] and Chavan et al [19]. A second major issue is the inconsistent use of valid names as accepted names or as synonyms. The broad bean is widely referred to as *Vicia faba* L. but in some

parts of the world is known as *Faba vulgaris* Moench., an equally valid name that is not widely accepted. The use of taxonomically intelligent network services [20] may help automate the process of identification of inconsistent use of names so directing taxonomic expertise to the solving of these problems. Currently both a single scientific name can validly apply to as many as three different organisms (Plant, Microbe, Animal) and multiple scientific names can apply to a single species (through the application of different taxonomic systems). Such problems of ambiguity may reduce when the Catalogue of Life becomes more complete by providing a single reference point. Manual scrutiny of data by the CoL editorial board is needed to prevent the same taxon being placed under different names if it is found in different source databases.

GBIF has a facility for filtering species occurrence data by Kingdom and for returning species synonyms [8]. However, this facility is only presently available for single species occurrence data downloads.

3 Excluding extraneous data fields & non Plantae records

Species occurrence data files must be prepared in the precise format required by the chosen ENM. Typically, these are csv or txt files, that at a minimum, include the species name (including genus) followed by longitude and latitude in decimal degrees. Consequently, extraneous data fields must be removed from the consolidated GBIF csv file. We also recommend carrying out a simple check that each record belongs to the Kingdom Plantae and removal of any records that don't. After indexing records alphabetically, identically spelled species' names but with differences in lower/upper case lettering should be identified and harmonized to prevent them being treated by an ENM as separate species.

4 Temporal error

Ideally, species occurrence observations should fall within the time period covered by the environmental layers. In practice, doing this will eliminate a large number of observations, possibly even the majority of them. In some cases, georeferenced GBIF data has no temporal reference. However, where the observation date is known, there are records ranging from the early 19^{th} Century to date. Filtering out records falling outside the environmental temporal framework is a straightforward task but there is a trade-off. If by excluding these records, we no longer have the minimum number needed to model the species (see Section 6), it may be preferable to maximize the number of species modelled even if doing so means introducing temporal error. Also, we may be excluding occurrences where in fact the species was still present during the time period of our model.

5 Data duplication

Data duplication commonly occurs when the same occurrence records occur in different datasets, which could happen where two or more herbaria hold du-

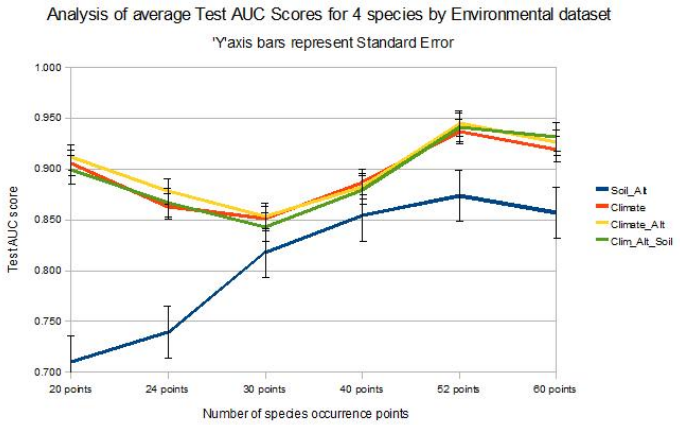

Fig. 1. Test AUC score variation by number of occurrences for 4 environmental datasets

plicates of the same collection, or because the same occurrence was observed at different times. Although, ENM's typically remove these duplications (e.g [5, 21]), this may reduce the number of records below the minimum threshold for an accurate species ENM. Consequently, there is a need to remove these duplications during species occurrence data pre-processing to establish whether modelling a species is practicable. In a survey of 544 mainland European plant species listed under the Berne Convention, 169 (31%) had no georeferenced data in GBIF and only 69 (13%) had 52 or more records [22] before filtering for duplication.

6 Minimizing Test AUC score variation

In practice, Test AUC scores will vary by species. This is because it is unlikely that more than one species will have exactly the same occurrence points. While we cannot control AUC Test score variation, we can minimize variances due to differences in dataset size. In a study of 4 plant species occurring in Italy, we analyzed Test AUC scores by different sized occurrence datasets (Figure 1). The results show that 52 point occurrence datasets performed best overall (regardless of environmental dataset type). Models with more than 60 occurrence points were not possible due to the lack of GBIF data for the 4 selected species at the chosen geographic extent. Our results are somewhat consistent with results obtained by other researchers who have observed that Test AUC scores achieved with 50 occurrence points tend to plateau at this level when larger numbers of occurrence points are considered [23–25]. Interestingly though, a recent American study of spotted knapweed, with species datasets averaging several hundred points saw test AUC score values ranging from 0.65 to 0.75 [26], indicating that the predictive ability of ENM's can also deteriorate when presented with large occurrence point datasets.

While no correlation was observed between test AUC score and species occurrence dataset size for smaller datasets, we believe that the score for 20 point datasets was excessively high due to the small number of test points used (4 or

5). Elith observes that Test AUC scores exceeding 0.75 are useful [27] while Baldwin adds that scores over 0.9 are very good [24]. Although, Hernandez [25] and Pearson [28] suggest that useful scores can be observed with sample sizes below 10, we recommend a minimum of 20. Our reason for this is that the smaller the number of test points, the less evidence there is that the probability distribution can be relied upon and where these test points are clustered (see Section 9), potential reliability diminishes further. For a biodiversity study, there is an obvious need to maximize the number of species being modelled. While it makes sense to minimize Test AUC score variance, due to different sized point datasets, we also need to maximize test AUC scores. Ultimately though, the choice of dataset size will depend on data availability.

A final observation here concerns the most appropriate mix of environmental layers. Figure 1 suggests that the best Test AUC results were obtained for the Climate_Altitude and Climate_Altitude_Soil environmental datasets. The 19 Climate layers are BIOCLIM [29] layers derived from average WorldClim [30] climate data for the period 1950-2000. These layers together with the altitude layer are highly correlated. In contrast, the 4 soil layers [31] included in the Climate_Altitude_Soil environmental dataset were uncorrelated. At 1km spatial resolution, it is likely that soil is a factor in a species probability distribution. Therefore, we recommend providing the ENM with the widest possible range of environmental data believed to be acting at the chosen spatial resolution provided this results in decent Test AUC scores. So, in the case of the experiments summarized in Figure 1, our choice for the best (52 occurrence point) model is that built with Climate Altitude Soil even though its 0.93 average AUC Test score was not significantly different from that obtained with Climate layers alone.

7 Spatial precision

The spatial precision of geo-referenced records is important because we need to ensure that this is not less precise than the spatial resolution of the environmental layers we use. Generally, the spatial precision of GBIF records varies in the range of 0 to 5 decimal places. For the sake of argument, lets say that we wish to conduct a biodiversity richness study at 30 Arc Seconds of spatial resolution (917 meters at the equator). Species occurrence point data stated to 2 decimal places equates to 1,110 metres at the equator. Therefore, only species occurrence records stated to 3 or more decimal places will be free of spatial precision error. At 2 decimal places, there is a small amount of spatial precision uncertainty but at 0 decimal places we are faced with not knowing which of 14,652 grid cells ($111{,}000^2/917^2$) in each of the environmental layers we should match the occurrence to? We use the qualifier 'possibly' because we may not be able to rely upon the positional accuracy of points whose coordinates are stated to a precision which is less than the spatial resolution of the environmental layers (some GBIF data is very old and may not have been accurately recorded). Another source of spatial precision uncertainty occurs when occurrence coordinates were originally derived from a raster dataset using grid cell centroid values [32]. For

example, some GBIF occurrence data is stated to 5 decimal places but came from 10km x 10km grid cell rasters! To resolve this problem, we must establish the spatial resolution of the original raster by consulting its metadata or, if necessary, the owner of the dataset. GBIF data also includes data where each coordinate is stated to a different number of decimal places. Consequently, for our study at approx. 1km resolution, we may wish to accept the minor spatial error of coordinates stated to 2 decimal places and we may also wish to include records where only one coordinate has been stated to 2 decimal places in the interests of preserving as many records as possible. Therefore, this filter should allow the user to specify the minimum number of decimal places of either both coordinates or one coordinate.

8 Points missing environmental data

Typically, ENM's (e.g [5, 21]) automatically exclude from the model, occurrence points for which any environmental data is missing. If we wish to control the size of datasets and minimize Test AUC score variation due to this factor, then it would pay to filter out these points first rather than have the ENM do this. The first category of points to eliminate are those falling outside the geographic extent of the environmental layers. This is a simple operation. The second category concerns points falling within the geographic extent of environmental layers. Here, two common reasons why occurrence points maybe missing environmental data are; coastal data and recording error [33]. In the first instance, errors occur when the observation was recorded to an insufficient number of decimal places and, as a result, appears in the sea, just off the coastline. Frequent causes of recording error include transposition of longitude/latitude values, and missing coordinates but they also include unexplainable errors. This filter will, therefore, need to find the appropriate grid cells in the environmental layers for each occurrence point and remove records where any *null* or *no data* values are returned.

9 Species dataset preparation

In Figure 1, generally the highest Test AUC scores were realized with 52 point datasets. It is interesting, therefore, to contrast, probability distributions for the bilberry (*Vaccinium myrtillus* L.) in the Italian region with 20 and 52 point datasets (Figure 2).

Note, in Figure 2 on the left, how the Slovenian probability distribution is almost non existent and the Alpine distribution of limited extent compared to the right image. Adding more occurrence points (white dots) in the right image, obviously increases the range of environmental data values available to the ENM. However, Test AUC scores were 0.967 for the 20 point dataset and and 0.95 for the 52 point dataset showing that these scores take on more importance when steps have been taken to maximize the range of environmental data values provided to the ENM. GBIF Plantae species occurrence data is frequently clumped.

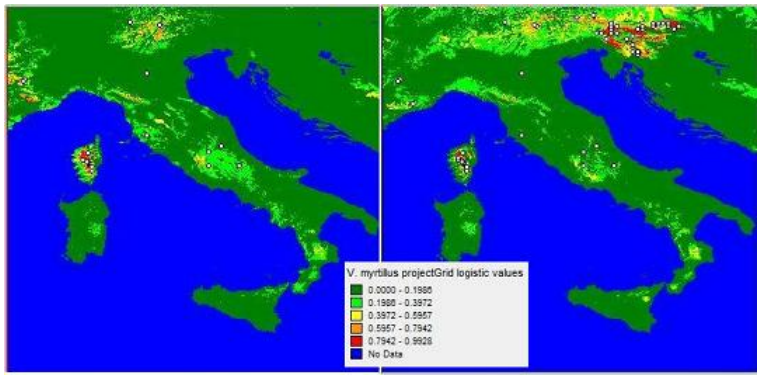

Fig. 2. *Vaccinium myrtillus* MAXENT logistic probability distribution Clim_Alt_Soil (20 points left, 52 points right)

This is almost certainly the result of the limited geographic extent of field trips to collect this data. We may be seeing this in Figure 2 and, of course, there are endemic plant species where we would expect to see clumped distributions. In his 2009 study of spatially autocorrelated sampling, Veloz concludes that the "AUC statistic is very sensitive to spatial autocorrelation between training and test points"[26]. ENM's attempt to overcome this issue by randomly allocating points between training and test data sets. In Section 6, we saw how generally a species occurrence data file size of 50 is probably the lowest file size choice if Test AUC scores are to be maximized. When more that 50 data points are available for a species from GBIF, it would make sense, therefore, to pick those points that provide the greatest geographic spread to avoid the spatial autocorrellation problem. Pythagoras's theorem [34] can help us here in calculating the distance d between two points (s_i,s_j):

$$d(s_i, s_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{1}$$

If we calculate the distance between each species occurrence point and every other point, then we can choose the 50 points showing the widest dispersion. Finally, after filtering out data uncertainty, species occurrence file generation is easily automated with Java ® which includes library classes for the production of species occurrence point datasets in csv or txt format as used by MAXENT and openModeller, for example.

10 Conclusions

Divorcing the pre-processing of species occurrence data (e.g. removing duplicate occurrences and points with missing environmental data) from the ENM,

gives the experimenter greater control over the size of species datasets and thus the ability to minimize test AUC score variance due to this factor. Including a widest dispersion filter for species with large numbers of occurrences similarly allows control over dataset size while providing the ENM with a wider range of environmental data values than that generated simply by random selection. The ability to filter out occurrences with coordinate precision lower than that required by the model's spatial resolution is an important pre-processing option. A separate species occurrence data pre-processing phase is also an opportunity to carryout temporal and synonymity filtering. After data filtering, the final task of the pre-processing phase is the automatic generation of species occurrence datasets which can then be directly 'plugged-in' to the ENM. A software application capable of carrying out all these tasks will prove to be a valuable time-saver particularly for biodiversity studies involving thousands of species. We are currently developing a software application implementing the species occurrence data pre-processing requirements described in this paper and we plan to present a tested version of this application at this special session of the KES 2010 conference.

References

[1] University of Idaho, college of Natural Resources, http://www.cnr.uidaho.edu/veg_measure/Modules/Lessons/Module%207/7_2_Biodiversity.htm
[2] Gurevitch J., Scheiner, S.M., Fox, G.A.: The Ecology of Plants, second edition. Sinauer Associates, Inc. Sunderland, MA, USA (2006)
[3] Gaston, K. J.: Species richness: Measure and measurement. In: Gaston K.J. (ed.). Biodiversity, a biology of numbers and difference. 396, pp. 77-113. Cambridge, Blackwell Science (1996)
[4] Escalante, T. et al.: Ecological niche models and patterns of richness and endemism of the southern Andean genus Eurymetopum (Coleoptera, Cleridae). Rev. Bras. entomol.[online] 53, 3, 379-385 (2009)
[5] The OpenNLP Maxent Homepage, http://maxent.sourceforge.net/
[6] Fielding, A.H., Bell, J.F.: A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38-49 (1997)
[7] Phillips, S.J., Anderson, R.P., Schapire, R.E.: Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231-259 (2006)
[8] GBIF Data Portal, 01/03/2010, http://data.gbif.org
[9] Thuiller, W. et al.: Climate change threats to plant diversity in Europe. In: Proc. Nat. Acad. Sci. 102: pp. 8245 8250, USA (2005)
[10] Oracle and Java |Technologies, http://www.oracle.com/us/technologies/java/index.html
[11] Bisby F.A., Roskov Y.R., Orrell T.M., Nicolson D, Paglinawan L.E., et al eds.: Species 2000 & ITIS Catalogue of Life: 2009 Annual Checklist Taxonomic Classification. CD-ROM; Species 2000: Reading, UK (2009)
[12] Johnson, B.: Understanding the Global Strategy for Plant Conservation. In: Annual meeting of the North American Association For Environmental Education (2005)
[13] McNeill, J., Barrie, F.R., Burdet, H.M., Demoulin, V., Hawksworth, D.J., et al.: International Code of Botanical Nomenclature (Vienna Code) adopted by the Seventh International Botanical Congress Vienna, Austria, July 2005. A.R.G. Gantner Velag, Ruggell, Liechtenstein. pp. 568 (2006)

[14] International Commission on Zoological Nomenclature, International Code of Zoological Nomenclature (4th edn), The Int. Trust for Zoological Nomenclature (1999)
[15] Sneath, P.H.A.: International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision. ASM Press (1992)
[16] Culham, A., Yesson, C.: Biodiversity informatics for climate change studies. pp XXX-YYY in Hodkinson et al. Climate Change and Systematics. Systematics Association, UK. (2010)
[17] Page, R. D. M.: A taxonomic search engine: Federating taxonomic databases using web services. BMC Bioinformatics 6, 48 (2005)
[18] Page, R. D. M.: Taxonomic Names, Metadata, And The Semantic Web. Biodiversity Informatics 3, 1-15 (2006)
[19] Chavan, V., Rane, N., Watve, A.: Resolving Taxonomic Discrepancies: Role Of Electronic Catalogues Of Known Organisms. Biodiversity Informatics 2, 70-78 (2005)
[20] Patterson, D.J., Remsen, D., Marino, W.A., Norton, C.: Taxonomic IndexingExtending the Role of Taxonomy Systematic Biology 55(3), 367-373 (2006)
[21] openModeller, http://openmodeller.sourceforge.net/
[22] Heywood V.H., Culham A.: The Impacts Of Climate Change On Plant Species In Europe. Report to The Convention On The Conservation Of European Wildlife And Natural Habitats, Standing Committee 29th Meeting Bern, Nov 09 (2010)
[23] Stockwell, D.R.B., Peterson, A.T.: Effects of sample size on accuracy of species distribution models. Ecol. Model. 148, 1-13 (2002)
[24] Baldwin, R.A.: Use of Maximum Entropy Modeling in Wildlife Research. Entropy 11, 854-866 (2009)
[25] Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L.: The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773-785 (2006)
[26] Veloz, S.D.: Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. J. Biogeogr. 36, 2290-2299 (2009)
[27] Elith, J.: Quantitative methods for modeling species habitat: comparative performance and an application to Australian plants. - In: Ferson, S. and Burgman, M. (eds). Quantitative methods for conservation biology. pp. 39-58. Springer (2002)
[28] Pearson, R.G., Raxworthy, C.J., Nakamura, M., Townsend-Peterson, A.: Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeopraphy 34, 102-117 (2007)
[29] Bioclim — WorldClim - Global Climate Data, http://www.worldclim.org/bioclim
[30] WorldClim - Global Climate Data, http://www.worldclim.org/
[31] European Soil Data Center (ESDAC), http://eusoils.jrc.ec.europa.eu/library/esdac/index.html
[32] Biodivertido.blogspot.com, http://biodivertido.blogspot.com/2009/02/grid-data-shared-as-point-data-errors.html
[33] Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., et al.: How Global Is the Global Biodiversity Information Facility?. PLoS ONE 2(11): e1124, doi:10.1371/journal.pone.0001124 (2007)
[34] OSullivan, D., Unwin, D.J.: Geographic Information Analysis. Wiley, USA (2003)