International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

Automatic modelling and continuous map generation from georeferenced species census data in an interoperable GIS environment

Lluís Pesquer [a], Ester Prat [a], Ricardo Díaz-Delgado [b], Joan Masó [a], Javier Bustamante [b] and Xavier Pons [c]

[a] *CREAF, Cerdanyola del Vallès 08193, Spain. l.pesquer@creaf.uab.cat*
[b] *Remote Sensing and GIS Lab (LAST-EBD). Estación Biológica de Doñana, CSIC. rdiaz@ebd.csic.es*
[c] *Departament of Geography, Universitat Autònoma de Barcelona. xavier.pons@uab.cat*

Abstract: The Natural Processes Monitoring Team from the Doñana Biological Station (EBD), systematically acquires data on more than 100 indicators of ecological processes and the status of many fauna and flora species in Doñana National Park, one of the most important protected wetlands in Europe, covering 54000 Ha. This information is available on a website as tabular data and trend charts. A detailed analysis is necessary in order to interpret this information and provide decision-making criteria for the management of the natural area.
The purpose of this paper is to improve public access to the information collected in the monitoring program and at the same time increase its quality. The proposed methodology integrates spatial interpolation methods, multivariate linear and logistic regression models (including the use of remote sensing images as predictors) and hybrid tools of these methodologies into a Geographic Information System (GIS) based model to generate predictive maps of ecological parameters. The information on the distribution, abundance, population structure and densities of different terrestrial and aquatic species, biophysical parameters and also their corresponding validation methodologies were used in the automatic generation of continuous maps of the distribution and abundance of the species in the study region.

***Keywords**:* GIS; automatic modelling, interoperable web map.

1 INTRODUCTION

Public administrations, institutions and research centres are currently devoting considerable effort to projects that collect large datasets of environmental geoinformation. For example, the Natural Processes Monitoring Team of the Doñana Biological Station (EBD) systematically acquires information on more than 100 indicators of ecological processes and on the status of many fauna and flora species (Díaz-Delgado [2010]). This information is available on a website in the form of tabular data and trend charts. Therefore, specific analyses are needed to interpret this information (Bonham-Carter [1994]) so that it can be used to provide decision-making criteria for research and management purposes (Scotts and Drielsma [2003]). A continuous map representation of the spatial distribution of these datasets would increase the number of potential users and facilitate analytical applications, and thus these public resources would be used more effectively.

The present work describes a methodology for automatically generating maps from species census data, which are then published on a web map server. Chain process automation, which includes acquiring and filtering data, map generation and web map publication, requires a well-analysed design and suitable tools that include self-decision procedures. Metadata in this context have an important function, and only with accurate knowledge of the quality indicators of the different steps of the chain process allows making automatic decisions.

Some previous works (Kiehle *et al.* [2007], Walter *et al.* [2011]) have emphasized the key role played by metadata in web map services, and here, in the present work, metadata plays an additional consequential role in the steps before web publication: the metadata are used to improve the models as well as map generation. The models and methods selected here have already been applied in similar environments (Valley *et al.* [2005], Hancock and Hutchinson [2006]). The aim of this work, however, is not to determine an optimal prediction method but rather to design and implement an entire automatic solution integrated into a GIS environment following standard and interoperable protocols.

2 METHODOLOGY

The proposed methodology is a chain of automatic processes that goes from downloading data on different servers to adding the corresponding map to a web portal for its publication. Figure 1 summarizes a flowchart of the entire chain process.

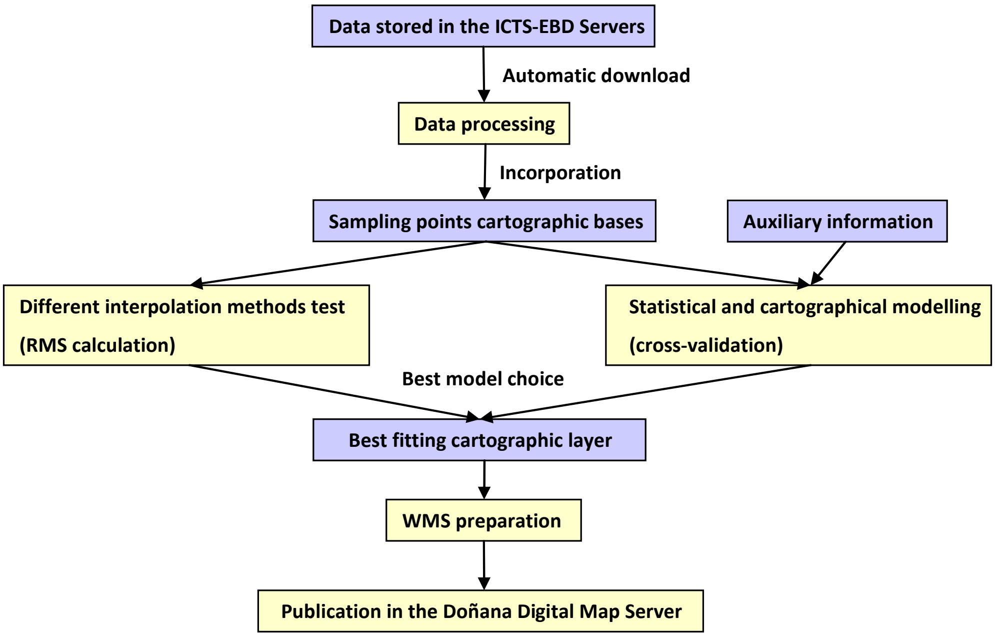

Data stored in the ICTS-EBD Servers
Automatic download
Data processing
Incorporation
Sampling points cartographic bases
Auxiliary information
Different interpolation methods test (RMS calculation)
Statistical and cartographical modelling (cross-validation)
Best model choice
Best fitting cartographic layer
WMS preparation
Publication in the Doñana Digital Map Server

Figure 1. Flowchart of the proposed methodology.

All of these processes are integrated in the MiraMon GIS software (Pons [2000]), which allows chaining and controlling the flow execution by metaprogramming commands in a BATCH Windows environment (Microsoft [2010]). The main characteristic of this software is that the metadata are managed accurately, which allows, at every partial result, to program automatic decisions based on quality information from the previous step in the process chain.

It is important to note that the species and models were chosen for map generation solely in order to demonstrate the feasibility of automating the entire procedure, and the objective was not to obtain the most suitable inference method for each species. This work focuses on automatic decisions that depend on a specific dataset and are based on the comparisons of the results of a large number of tests.

The study region is the Doñana National Park, one of Europe's most important wetland reserves (Díaz-Delgado [2010]) and a major site for migrating birds. It covers an area of over 54000 Ha and is located in the south of the Iberian Peninsula, southwestern Europe (Figure 2).

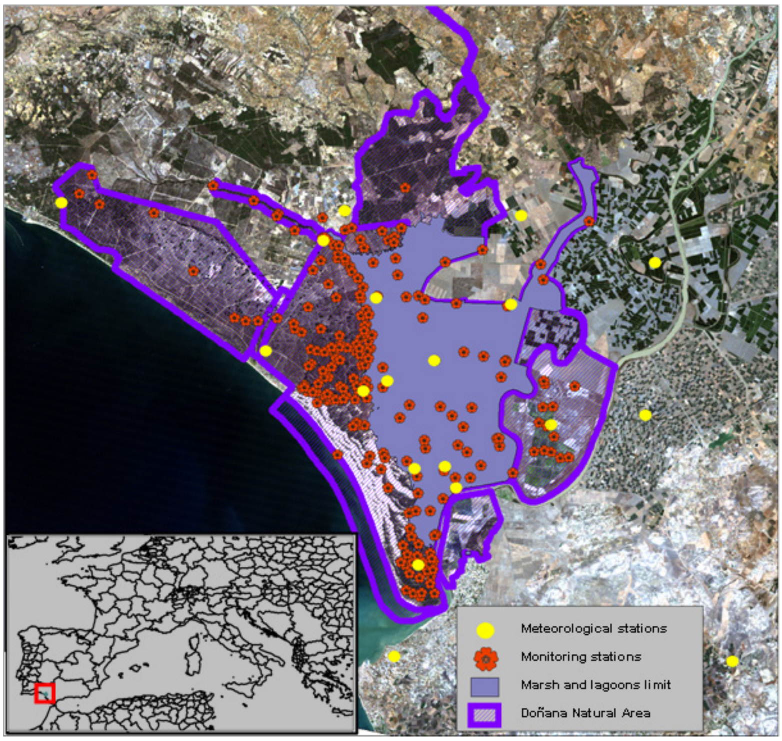


Figure 2. Location of Doñana National Park and the sampling stations.

2.1 Materials and data processing

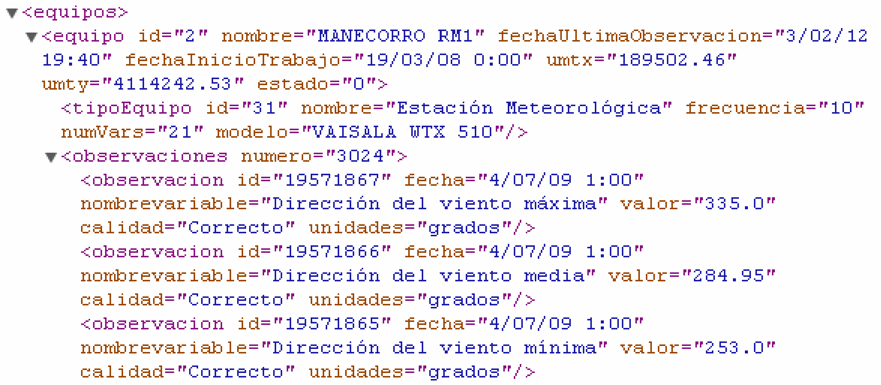
Four different types of data are downloaded:

- Meteorological variables, such as mean air temperature, maximum air temperature, minimum air temperature and precipitation;
- Abundance of aquatic fauna, for example, the number of Lousiana crayfish (*Procambarus clarkii*) individuals;
- Presence/absence of aquatic vegetation; for example, the salt-marsh bulrush (*Boloboschoenus maritimus)*;
- Hydrological variables, such as water level, water temperature, minimum water temperature and maximum water temperature.

Meteorological variables are obtained from two different sites:

- Automatic weather stations from the Singular Scientific and Technological Infrastructure (ICTS) of the Doñana Biological Reserve: http://icts.ebd.csic.es/GeneradorDatosXMLGeneralServlet.
- Agrometeorological stations of the Research and Training Institute for Agriculture and Fisheries (IFAPA): http://www.juntadeandalucia.es/agriculturaypesca/ifapa/ria/servlet/FrontController .

Hydrological variables are also downloaded from the ICTS website at http://icts.ebd.csic.es/GeneradorDatosXMLGeneralServlet. The ICTS server provides data in XML format (Gutiérrez et al. [2003]), as shown in Figure 3, while the IFAPA server provides data in plain text format. These datasets require several post-processing procedures to prepare them for modelling, such as detecting and erasing wrong data, fusing coherently from different origins, grouping and average calculating.



```
▼<equipos>
  ▼<equipo id="2" nombre="MANECORRO RM1" fechaUltimaObservacion="3/02/12
   19:40" fechaInicioTrabajo="19/03/08 0:00" umtx="189502.46"
   umty="4114242.53" estado="0">
     <tipoEquipo id="31" nombre="Estación Meteorológica" frecuencia="10"
     numVars="21" modelo="VAISALA WTX 510"/>
   ▼<observaciones numero="3024">
       <observacion id="19571867" fecha="4/07/09 1:00"
       nombrevariable="Dirección del viento máxima" valor="335.0"
       calidad="Correcto" unidades="grados"/>
       <observacion id="19571866" fecha="4/07/09 1:00"
       nombrevariable="Dirección del viento media" valor="284.95"
       calidad="Correcto" unidades="grados"/>
       <observacion id="19571865" fecha="4/07/09 1:00"
       nombrevariable="Dirección del viento mínima" valor="253.0"
       calidad="Correcto" unidades="grados"/>
```

Figure 3. Example of meteorological variables in an XML file from the ICTS server http://icts.ebd.csic.es/GeneradorDatosXMLGeneralServlet?idEstacion=3&fechaInicio=040720090000&fechaFin=040720092359

For the data on aquatic fauna abundance and the presence/absence of aquatic vegetation, the download site is http://icts.ebd.csic.es/GeneradorDatosSeguimientoXMLServlet and similar processes of filtering and aggregating data are involved.

2.2 Map and model generation

Previous studies have tested different methodologies for generating a continuous representation of a quantitative variable from the values observed in specific locations (Lloyd [2006]). The prediction and modelling methods implemented were selected from the most usual methods (Burrough and McDonnell [1998]) and those that are most suitable for automation, for example kriging interpolation was discarded due to the difficulty involved in obtaining an automatic variogram (Pesquer *et al.* [2011]). Continuous map generation methods include different kinds of strategies: univariate and multivariate, and based on spatial patterns or statistical regressions, among others.
The specific methodologies tested were:

- Two spatial interpolation methods, inverse distance weighting (IDW) (Bartier and Keller [1996]) and splines (Mitasova and Mitas [1993]) used for meteorological and hydrological variables and for aquatic fauna abundance.
- Logistic regression (Kleinbaun [1994]) to determine the probability of the presence of aquatic vegetation.
- Multivariate regression + residual spatial interpolation (Ninyerola *et al.* [2000]), for meteorological variables and aquatic fauna abundance.

The final map is chosen by comparing the accuracy of the results generated by using different parameter sets for the same method or comparing the results of different methods. This estimate of the accuracy is obtained with an independent test validation subset or by using a leave-one-out cross-validation (Isaaks and Srivastava [1989]) for small samples. In addition, these validations and comparisons are integrated automatically into the entire flowchart.
The maps generated have a spatial resolution of 150 m, which is based on the distribution of the sampling locations, (Hengl [2006]) and they are georeferenced in the UTM-29N reference system.

2.3 Web publication

Publishing continuous maps for the different variables studied in a web environment is an essential step for the widest possible dissemination of results. Furthermore, using OGC standard protocols (Open Geospatial Consortium [2008]), such as WMS map servers, makes it possible to integrate data generated in the emerging Spatial Data Infrastructures and provides interoperability with other available information.

The automated integration of maps into the existing Doñana Biological Station server (http://mercurio.ebd.csic.es/seguimiento) is the final step in the project. The CreaMMS tool (Maso and Pons [2005]) from the MiraMon software makes it possible to prepare layers that will be served later, and which will therefore be accessible to any OGC client.

3 RESULTS

3.1 Maps and models

A collection of continuous maps for each annual hydrological cycle (from September to August) and within the period 2007-2011 have been obtained. The maps include meteorological variables, hydrological variables, abundance of aquatic fauna and the probability of the presence of aquatic vegetation.

Three representative examples are shown below:

a) Spatial interpolation of precipitation (2010-2011).

Table 1 compares different quality results (by RMS) at different exponent parameter values for the IDW method. A range of tension parameter values have been used in the splines method.

IDW		Splines	
Exp.	**RMS**	**Ten.**	**RMS**
1	202.71	25	332.25
1.25	203.99	50	201.08
1.50	206.04	75	190.53
1.75	208.59	100	189.12
2.00	211.39	***125***	***189.09***
2,25	214.23	150	189.19
2.50	217.02	175	189.28
2.75	219.71	200	189.37

Table 1: RMS for each exploration parameter: exponent for IDW and tension for splines.

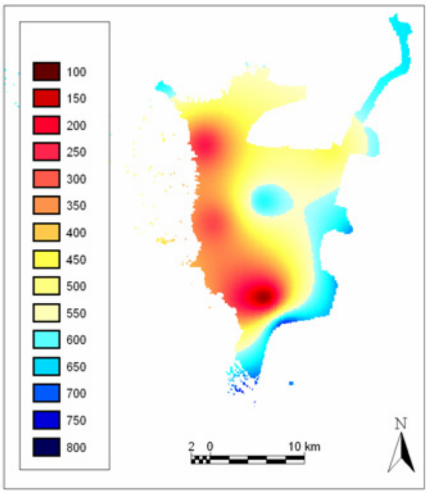


Figure 4: Spline interpolation for precipitation in the period 2010-2011.

In this example, most spline interpolations are better than the best IDW. A spline tension of 125 was selected for map generation (see Figure 4) as it has a small RMS.

b) Multivariate linear regression and residual spatial interpolation for Lousiana crayfish abundance (2007-2008).

Three variables were initially entered into the regression model in this example: hydroperiod (number of rainy days during a complete cycle) was automatically excluded, and the probability of the presence of aquatic vegetation and the average maximum temperature were selected, as shown in Table 2. This regression generated residual values for sampling locations were spatially interpolated until a minimum RMS was obtained. Figure 5 shows the final result, the regression model + spatial interpolation of the regression residuals.

c) The probability of the presence of salt-marsh bulrush (2007-2008) using logistic regression.

Table 3 provides a case study that models the probability of the presence of an aquatic species. Three remote sensing products were introduced as possible auxiliary variables: the maximum and average NDVI for the entire period (Rouse *et al*. [1973]) and the marsh water turbidity (Bustamante *et al.* [2009]).

Independent Variables	**Signi-ficant**	**Coefi-cient**
Hydroperiod	No	
Aquatic vegetation probability	Yes	227.288
Average maximum temperature	Yes	49.219
Intercept		1235.588

Table 2: Independent variables introduced into the linear regression model and corresponding results.

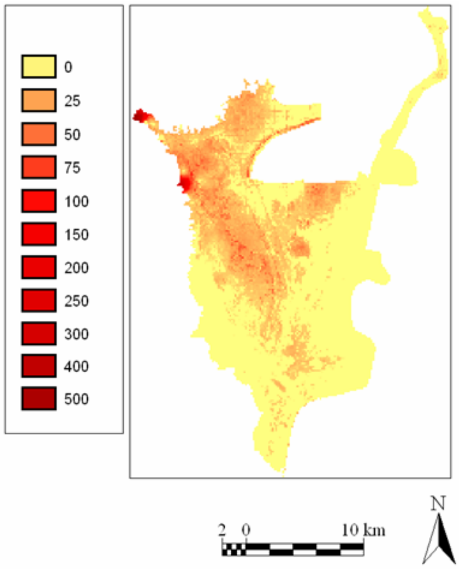


Figure 5: Map of the regression model + residual interpolation for *Procambarus clarkii* abundance

Two meteorological variables were used, hydroperiod and mean maximum temperature for the complete period, and finally the average distance to the flooded area was also introduced (Díaz-Delgado *et al*. [2006]). In this example hydroperiod and maximum NDVI were not significant and were not used in the final regression model. Figure 6 shows the resulting map generated from the Table 3 regression.

Independent Variables	**Signifi-cant**	**Coefi-cient**
Hydroperiod	No	-
Maximum NDVI	No	-
Turbidity	Yes	-0.001
Average distance to the flooded area	Yes	0.001
Average NDVI	Yes	-4.842
Average of maximum temperature	Yes	1.327
Intercept		32.715

Table 3: Independent variables introduced into the logistic regression model and corresponding results

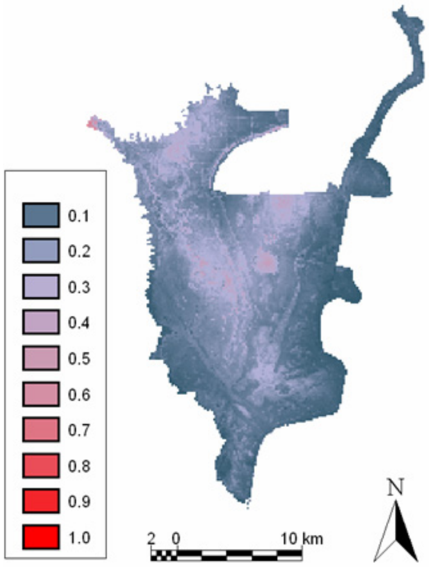


Figure 6: Logistic regression map for the probability of the presence of the salt-marsh bulrush (2007-2008).

3.2 Web portal

The maps and models generated have been integrated automatically into an existing Web Service: Servidor de Cartografía Digital de Seguimiento del Parque Nacional de Doñana (http://mercurio.ebd.csic.es/seguimiento). The existing interoperability protocols of the OGC (Open Geospatial Consortium [2008]) have been preserved, and a new tool specifically developed as a Web Processing Service (WPS) (Schut [2007]) has been added: analytic statistical overlay between a user's layer of polygon features and the map results generated in the present work. Figure 7 shows this portal.

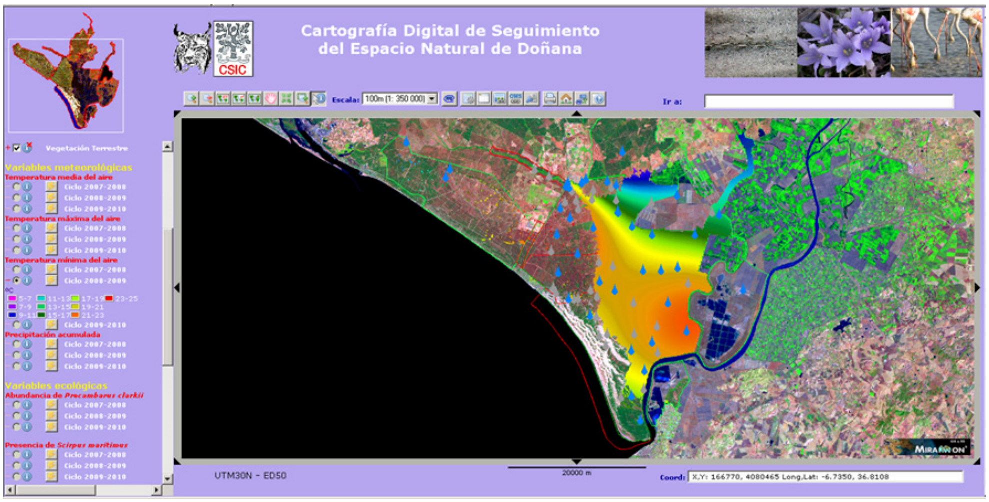


Figure 7. Web portal of the *Servidor de Cartografía Digital de Seguimiento del Parque Nacional de Doñana*.

4 CONCLUSIONS

The main contribution of this work is the automation and integration in a GIS environment of the entire proposed methodology, from downloading selected census data to the final publication of maps in a Web Map Server on the Internet using Open Geospatial Consortium standards. A service for invoking remote Web Processing Services for generating maps on demand is also provided.

The selected pilot cases, which represent different types of species census data and involve different estimation methods, demonstrate real implementations of the present work and a test-bed evaluation for extending this proposal to other scenarios.

ACKNOWLEDGMENTS

This work was supported by the Catalan Government (Spain) under grant SGR2009-1511 and the ICREA-Acadèmia award, by the Instituto de Cartografía de Andalucía (Spain) under grant "Ayudas a la investigación en materia de información geográfica" and uses data provided by the project CGL2009-09801 financed by FEDER funds.

REFERENCES

Bartier, P. M., Keller, C. P., Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). *Computers & Geosciences*, 22(7): 795-799, 1996.

Bonham-Carter G.F, *Geographic information systems for geoscientists modelling with GIS*, Pergamon, 398 pp., 1994.

Burrough, P.A.,McDonnell,R.A., *Principles of Geographical Information Systems*. Oxford University Press, 333 pp, 1998.

Bustamante, J., Pacios, F., Díaz-Delgado R., Aragonés, D., Predictive models of turbidity and water depth in the Doñana marshes using Landsat TM and ETM+ images. *Journal of Environmental Management*. 90:2219-2225, 2009.

Díaz-Delgado, R., Bustamante, J., Aragonés, D. and Pacios, F., Determining water body characteristics of Doñana shallow marshes through remote sensing. In *Proceedings of the 2006 IEEE International Geoscience & Remote Sensing Symposium* (IGARSS2006), Denver, Colorado, EE.UU., 3662-3664. 2006.

Díaz-Delgado, R., An integrated monitoring programme for Doñana Natural Space: The set-up and implementation. In C. Hurford, M. Schneider e I. Cowx, (Ed.), *Conservation Monitoring in Freshwater Habitats: A Practical Guide and Case Studies.* Springer, Dordrecht, 375-386 pp., 2010.

Gutiérrez Martínez, J. M., Palacios, F. and Gutiérrez de Mesa, J.A., *El estándard XML y sus tecnologías asociadas*. Ed. Danysoft, pp 506, 2003.

Hancock, P.A., Hutchinson, M.F., Spatial interpolation of large climate data sets using bivariate thin plate smoothing splines. *Environmental Modelling and Software* 21, 1684-1694, 2006.

Hengl T. Finding the right pixel size. *Computers & Geosciences* 32:1283–1298, 2006.

Horning, N., Fosnight, E., eds. Secretariat of the Convention on Biological Diversity, Montreal,Technical Series no. 32, 201 pages. Pp. 83-102. ISBN: 92-9225-072-8

Isaaks, E.H., Srivastava, R.M. *Applied Geostatistics*. Oxford University Press, New York, 1989.

Kiehle, C., Greve, K., Heier, C. Requirements for next generation spatial data infrastructures-standardized web based geoprocessing and web service orchestration *Transactions in GIS* 11(6): 819-834, 2007

Kleinbaun, D.G, *Logistic regression*. New York, Springer-Verlag, 1994.

Lloyd, C. D., *Local Models for Spatial Analysis*. CRC Press, 244 pp, Belfast, 2006.

Masó J., Pons. X., Adding functionalities to WMS-WCS Clients: Download And Animation, *International Cartographic Conference*, A Coruña, 9-16, 2005.

Mitasova, H., Mitas, L., Interpolation by Regularized Spline with Tension. *Mathematical Geology*, 25 (6) 641-655 pp, 1993.

Microsoft Corporation Using BATCH files http://www.microsoft.com/resources/documentation/windows/xp/all/proddocs/en-us/batch.mspx?mfr=true [Date of access: 2-2-2012], 2010.

Ninyerola M, Pons X, Roure JM, A methodological approach of climatological modelling of air temperature and precipitation through GIS techniques. *International Journal of Climatology*, 20:1823-1841, 2000.

Open Geospatial Consortium: *OGC Reference Model*, Open Geospatial Consortium Inc. Reference number: OGC 08-062r4 Version: 2.0, 2008.

Pesquer L., Cortés A., Pons X., Parallel ordinary kriging interpolation incorporating automatic variogram fitting, *Computers & Geosciences* 37, 464-473, 2011.

Pons, X., *MiraMon. Geographical Information System and Remote Sensing Software*. Centre for Ecological Research and Forestry Applications, CREAF, ISBN: 84-931323-4-9 In Internet: http://www.creaf.uab.es/MiraMon, 2000.

Rouse J.W., Haas, R.H., Schell, J.A., Deering, D.W., Monitoring Vegetation Systems in the Great Plains with ERTS. *Third ERTS Symposium*, NASA SP-351, I:309-317, 1973.

Scotts, D., Drielsma, M. Developing landscape frameworks for regional conservation planning; an approach integrating fauna spatial distributions and ecological principles. *Pacific Conservation Biology* 8(4): 235-254, 2003

Valley, R.D., Drake, M.T., Anderson, C.S. Evaluation of alternative interpolation techniques for the mapping of remotely-sensed submersed vegetation abundance. *Aquatic Botany* 81(1):13-25, 2005

Walker Johnson, G., Gaylord, A.G., Franco, J.C., Cody, R.P., Brady, J.J., Manley, W., Dover, M., Garcia-Lavigne, D., Score, R., Tweedie, C.E., Development of the Arctic Research Mapping Application (ARMAP): Interoperability challenges and solutions *Computers & Geosciences* 37(11): 1735-1742, 2011.