European conference of the Czech Presidency of the Council of the EU
TOWARDS eENVIRONMENT (Challenges of SEIS and SISE: Integrating Environmental Knowledge in Europe)
http:/www.e-envi2009.org/proceedings/
J. Hřebíček, J. Hradec, E. Pelikán, O. Mírovský, W. Pilmmann, I.Holoubek, R. Legat (eds.)
Masaryk University, 2009

Automated mapping of environmental variables from a SEIS or SISE perspective

Edzer Pebesma[a], Gregoire Dubois[b] and Dan Cornford[c]

[a] Institute for Geoinformatics, University of Münster, Germany (edzer.pebesma@uni-muenster.de); [b] Joint Research Centre of the European Commission, Ispra, Italy (gregoire.dubois@jrc.it); [c] Computer Science and NCRG, Aston University, Birmingham B4 7ET, UK (d.cornford@aston.ac.uk).

Abstract

The INTAMAP FP6 project has developed an interoperable framework for real-time automatic mapping of critical environmental variables by extending spatial statistical methods and employing open, web-based, data exchange protocols and visualisation tools. This paper will give an overview of the underlying problem, of the project, and discuss which problems it has solved and which open problems seem to be most relevant to deal with next. The interpolation problem that INTAMAP solves is the generic problem of spatial interpolation of environmental variables without user interaction, based on measurements of e.g. PM_{10}, rainfall or gamma dose rate, at arbitrary locations or over a regular grid covering the area of interest. It deals with problems of varying spatial resolution of measurements, the interpolation of averages over larger areas, and with providing information on the interpolation error to the end-user. In addition, monitoring network optimisation is addressed in a non-automatic context.

KEYWORDS: *Environmental data; Environmental information; In-situ sensors, Spatial interpolation, Geostatistics, OGC, SOA.*

1. INTRODUCTION

Spatial interpolation of in situ sensed variables such as meteorological variables, air quality variables, groundwater quality, or environmental radioactivity is a problem for which no single solution exists. In an experiment where several experts were confronted with interpolating the same data set (EUR 21595, 2005), the approaches differed wildly, and best results were obtained by machine learning techniques as well as geostatistical methods. One of the reasons behind this variety is that one needs to choose a model of spatial variability before one

can interpolate, and experts disagree on which models are most useful – a case that is not uncommon whenever modelling is involved.

A lack of generally accepted solutions has led to a situation where interpolation experts with highly domain-specific expertise work in fields such as mining, oil exploration, environmental monitoring, or risk assessment and use highly specialised tools. A side effect is that in several domains where interpolation might be useful it is either not applied because of a lack of expertise, or applied using algorithms so simplistic that it undermines the quality of the results.

Motivated on one hand by the increasing availability of sensor data in near real time, and on the other by the need to take decisions in disaster management frameworks without having time to consult interpolation experts, the INTAMAP FP6 project aims to build an automated interpolation service that should provide useful interpolation without requiring any specialised skills. This should be realized using open standards, and under an open source software license. As interpolation cannot be done without introducing errors, the experts in the project consortium considered the word "useful" to mean that the interpolation comes with meaningful information about the interpolation error to characterise the uncertainty in the result. This information might be in the form of an interpolation standard error or prediction variance, the specification of a full conditional probability distribution, or e.g. define probabilities of exceeding a number of given thresholds. Such error information might be ignored by some, but might help others to optimise decision making in the presence of uncertainty, e.g. weighting the risks and costs of type I and type II errors (false negatives or false positives such as evacuating areas not in danger, or not evacuatiing areas that required evacuation).

The paper is organised as follows. First, the statistical considerations underlying automated mapping will be discussed, and the challenges faced outlined. The technical realization and system architecture will be described. Issues of performance and embedding it in a service oriented / service chained environment will be discussed. Finally, we provide a perspective on how this service might be extended along with ideas for future developments of environmental management systems based on service oriented architectures (SOA).

2. STATISTICAL CONSIDERATIONS

Spatial interpolation basically consists of two steps. First, a model for the spatial variability has to be selected, and its parameters have to be estimated. In geostatistics, models of the form

$$Z(s) = m(s) + e(s)$$

are usually deployed (Cressie, 1993), with $Z(s)$ the measured process at spatial location s, $m(s)$ the spatially varying (or constant) trend component usually modelled as a linear in parameters regression model of the form $m(s) = X(s)\beta$ with $X(s)$ often layers in the GIS (Pebesma, 2006) and β unknown regression coefficients, and $e(s)$ usually a second order stationary residual process. This

first step then boils down to the choice of a trend function, a covariance function for the residual process, and the estimation of all parameters involved in both components.

The second step involves, given this model and the observations, the spatial interpolation (prediction, evaluation) of this model for new observation locations s_0

$$\hat{Z}(s_0) = \hat{m}(s) + \hat{e}(s)$$

where s_0 is usually taken over a grid covering the region of interest.

2.1 The emergency case: spatial extremes

The original motivation for INTAMAP came from the monitoring of environmental radioactivity at a European scale. EURDEP, the European radiological data exchange platform (see http://eurdep.jrc.ec.europa.eu/), makes unvalidated radiological monitoring data coming from around 4000 sensors spread over most European countries available in near real-time to decision-makers. The main purpose of this network is motivated by emergency cases, where the exchange of these data between EU member states greatly facilitates the monitoring in near real-time of the spread of a radioactive release over Europe. The first stage of an emergency, with a very localised but significant release, is however one of the most difficult problems to interpolate. Several approaches to this have been compared, and developed, within this project. Early stages of a release, such as tested in the interpolation comparison exercise mentioned before (EUR 21595, 2005), are characterised by many low observations and very few observations with extremely outlying measured values. Interpolating such variables is extremely difficult from a statistical perspective.

The INTAMAP automated interpolation service deals with data containing extreme outliers, and deploys dedicated methods, based on spatial copulas, to form a model for spatial variability and interpolate these data (Pilz et al., 2008; Kazianka and Pilz, 2009).

2.2 Uncertainty

Interpolation requires modelling, and modelling involves approximation. Scientist will rarely claim that an interpolated value equals the true value. Statistical models can help to quantify the interpolation error. To interpret the interpolation results in the right way, this information should be transmitted, along with the maps produced. This can be done in several ways, e.g. providing standard errors, probabilities of exceeding thresholds, or by sampling from the statistical model. When the interpolated map is meant to serve as an input to a next processing stage, e.g. to compute exposure of a population over a certain region, this interpolation error specification is indispensable.

The INTAMAP automated interpolation service addresses the communication of errors associated with the interpolation, either in a simple form (standard error, variance) or in a more complete form, ranging from specification of the full parametric distribution (distribution form, parameters), the approximation of this distribution by a number of statistics (e.g. quantiles, or distribution

function values), or by a sample from the multivariate distribution (a Monte Carlo sample). The section on technical implementation will detail how this was done.

2.3 Anisotropy detection

Many environmental variables are subject to anisotropy, meaning that in some direction the degree of spatial continuity, or spatial correlation, is stronger than in others. This phenomenon is e.g. present when point sources diffuse, and one transport direction (e.g. due to wind) dominates, e.g. East-West.

The INTAMAP automatic interpolation service automatically detects anisotropy, tests whether it is significant (Chorti and Hristopoulos, 2008), and if it is, corrects for this anisotropy before further steps are taken (modelling of spatial correlation; spatial interpolation).

2.4 Observations with known errors

All observations on continuous variables are measured with some degree of measurement error. Often, this error is unknown, or believed to be very small according to the specifications of the producer of the sensor used. In other cases however, the error magnitudes are known and considerable in size, e.g. because they result from indirect sensing and elaborate and complicated calibration. An example of this are the atmospheric chemistry measurements from satellites such as OMI.

Interpolation of data with considerable, known measurement error should take these errors into account. In the INTAMAP interpolation service, error characteristics of the observations can be specified and a sequential interpolation method based on Gaussian processes (Ingram et al., 2008) is used to optimally interpolate the spatial field in this case.

2.5 Spatial aggregation: estimating areal averages

Besides the usual interpolation to points (on a grid) in space, one may decide to estimate average (or differently spatially aggregated) values, e.g. for complete grid cells, or for larger areas. This may be convenient when decision making does not take place for points, but rather for areas of some size, typically defined by administrative boundaries. An example of this is evacuation: we don't evacuate points, but rather neighbourhoods, regions, villages, towns, or flood plain sections.

The need to consider spatial aggregation in the interpolation process is that although interpolated values can easily be aggregated by averaging them after interpolation took place, the associated errors or error distributions cannot be obtained this way, but need to be quantified during the interpolation process.

2.6 Monitoring network harmonisation and optimisation

Integrating measurements across EU Member States, or even within Member States, e.g. by applying an interpolation procedure often reveals harmonisation issues: different sensor types or different treatment by sensor operation bodies

result in constant or random biases. Several possible bias types have been identified, and procedures have been implemented to estimate their magnitude from monitoring network data for the case where they are not reported (Skøien et al., 2009).

Monitoring network optimisation involves the placement, removal or moving of monitoring network stations. Part of this problem is obviously political, as it involves the monetary costs and benefits of a network station, and deals e.g. with questions concerning which variables a society wants to monitor at all. The scientific contribution to this problem involves the monetary assessment of benefits or losses that addition, removal or moving of stations will result in. Implementing a generic, domain independent solution to this is difficult as monitoring networks usually serve different goals (Müller, 2007).

The INTAMAP automated interpolation service does not automatically correct for statistical biases, because a complete understanding and agreement on the magnitude of such biases should be found before it can be part of an automated analysis system. The software delivered does provide the tools to estimate biases from network data. From a similar argument, network optimisation code has been developed but is not interfaced through a web service, because it will mostly be evaluated off-line, most likely in a non-automated setting where many more constraints play a role than an automated service can ever consider.

3. TECHNICAL REALISATION

3.1 OGC Web Services

Web service standards as agreed upon by e.g. ISO TC211, OGC and INSPIRE are the basis for useful generic services that can be part of SISE or SEIS. INTAMAP delivers an interpolation web processing service schematically shown in Figure 1. It accepts sensor data from a sensor observation service (as an observations & measurements document), and returns the interpolation result e.g. as a GML document or as a web coverage service (Williams et al, 2007). To encode the interpolation error information, UncertML, a markup language for specifying information that is represented probabilistically as a representation of a random variable, has been developed within the project, and proposed as a standard to the OGC (Williams et al., 2008).

3.2 The R back end and interpolation decision tree

The procedure for the statistical analysis of the data are implemented in R, the major open source environment for analysing statistical data. As figure 1 shows, this is not noticeable for the user of the INTAMAP web processing service, as R is run in the back end. Interfacing R from the web processing service by using the TCP/IP protocol (i.e., as a web service, using the Rserve package; Urbanek, 2009) has the advantage that the R process, doing the hard numerical work, may be running on a highly dedicated computing cluster that

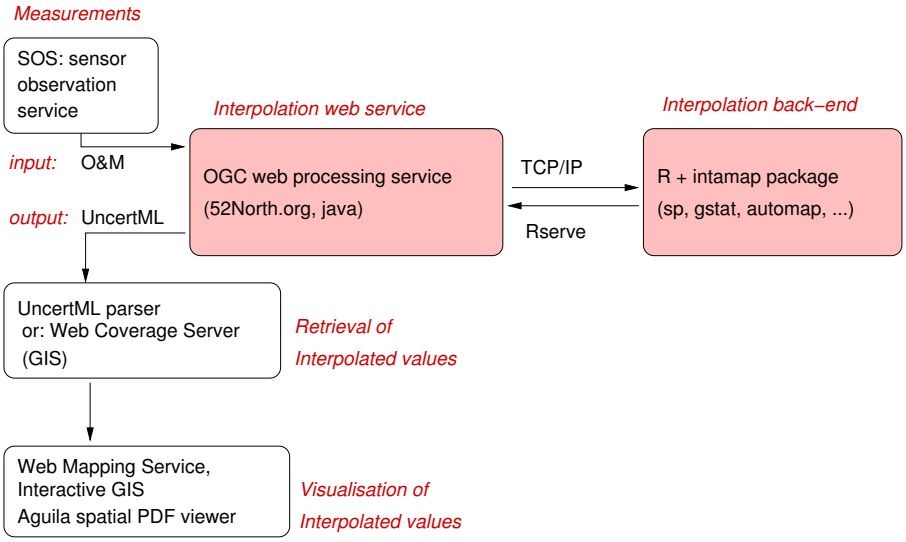

Figure 1: Technical set up of the automatic interpolation service. UncertML stands for uncertainty markup language (see text); O&M stands for observations and measurements, an XML standard for encoding monitoring network data.

is not directly connected to the internet. A second advantage of having all statistical routines in the R environment is that it can be re-used independently from the WPS interface, e.g. interactively on a PC, from a python or SOAP interface, or on a mobile device.

The decision tree for choosing an interpolation method automatically is shown in figure 2. In the context of the INTAMAP project, dedicated interpolation methods have been implemented for (i) detecting and correcting for anisotropy, (ii) dealing with extreme value distributions, (iii) dealing with known measurement errors.

Methods for network harmonisation were also developed, but are not part of the automated interpolation framework, as this should be done before interpolation takes place. The same is true for outlier removal and monitoring network optimisation. With the software developed for INTAMAP, it would be relatively simple to customize the INTAMAP web service and perform these manipulations.

4. OPERATIONAL PERFORMANCE

At the stage of writing this paper, the INTAMAP interpolation service is fully functional, and open for testing. Under this testing framework, the following issues need to be resolved before setting up a robust public service that allows everyone to use it:

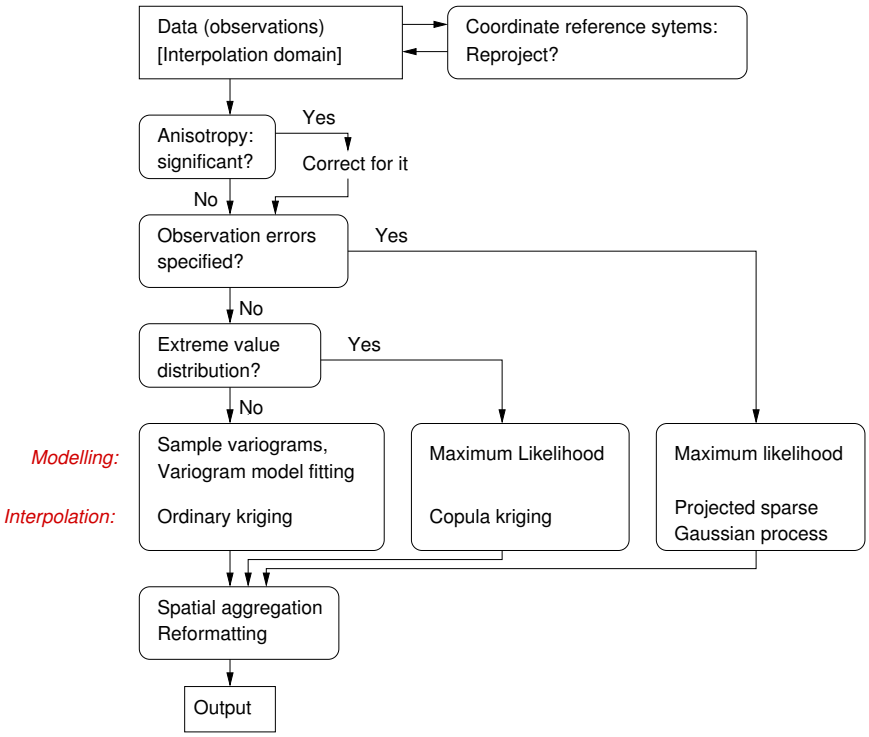

Figure 2: Decision tree for the interpolation method choices in the interpolation process that takes place in R. References in text.

- Both maximum likelihood and (global) ordinary kriging need to solve systems of linear equations of size $n \times n$, with n the number of observations. When n becomes large, say over 1000, then this process takes very long. For ordinary kriging this can be solved by reducing the system by default to only address the nearest m observations, with e.g. m in the range of 50.

- When running a web service, it is hard to be certain that the service or server will not at some stage get overloaded when many server requests arrive at the same time.

- Some of the interpolation methods implemented need a considerable amount of time to process, of the order of hours or more; an interpolation request should then specify the (maximum) amount of time available, and the service should be able to discard some of the methods based on that requirement. In case of long-term processes, asynchronous protocols are needed, and these are implemented in the reference WPS used.

- After using the automatic method selection of the INTAMAP service, we envisage that experienced or expert users will want to have more control over the interpolation method chosen. Thus we allow the WPS to accept parameters that are passed to the R process, to control this.

- The observations read by the INTAMAP interpolation service need to be an O&M document (observations and measurements), but not every O&M document will be accepted. This is because O&M accommodates practically every possible observation scenario, including time series data and imagery data – cases that make little sense to send to an interpolation service.

- When used in a controlled environment, e.g. to a restricted domain such as air quality or environmental radioactivity, the R web service can also be constrained to always use the same method, in order to get results that are easier comparable across different interpolation requests.

- Besides interpolated values, the interpolation should return some information about which method was used, what the values of the fitted parameters are, and maybe even some relevant diagnostic plots, such as the variogram and fitted model.

A few observations made here are common to SEIS or SISE. The availability of the web services, the computational time that needs to be accounted for when requests are made in parallel to different environmental web services, the propagation or errors, the tracking and documentation of manipulations in chained service environments are all challenges SEIS and SISE will have to address.

5. DISCUSSION AND OUTLOOK

The automated interpolation web service, the main deliverable of INTAMAP, is an important asset to SEIS or SISE – it takes monitoring data, interpolates to arbitrary points, grids, or averages over polygons, and yields information on the interpolation approximation errors made. It deals with anisotropy, with errors in observations, and with outliers/extreme value distributions. In addition, in a number of application areas (air quality, environmental radioactivity, meteorology) the use of the service will be shown in use cases and demonstrations. The implementation uses open OGC standards and is completely open source. Technology for network optimisation and harmonisation has been developed as well.

The generic interpolation service does just that: automatic interpolation. Clearly, interpolation of real variables with known characteristics would typically not only use measured data, but additional information: for air quality one would like to use remotely sensed data, land use and/or traffic information, for environmental radioactivity it makes sense to use geology and altitude. Although such information is readily available, the appropriate interpolation service would become domain specific (only relevant for a specific variable) and location specific (only useful for a specific region). The generic interpolation service developed here could, however, very well be used as a first major component to build such a specific interpolation service.

In the same thread, phenomena for which near real-time interpolation is relevant are usually dynamic in time, and the interpolation service set up currently ignores time. The step from spatial interpolation to spatio-temporal interpolation is not a trivial one, and again the current development can be used as a first building block for it. One reason not to address time was that in space-time modelling some kind of gradual development of the spatial field over time is usually assumed. In case of unexpected extremes (a nuclear accident), such assumptions may lead to underestimation of the real problems. Further, the behaviour of many variables is subject to transport and diffusion, and involving a transport model would again make the approach domain specific.

For all extension directions: including static GIS information, including dynamic mechanistic models, and including the temporal component, the real challenge lies in developing a method (one or more service) that acknowledges that data are subject to errors, models are subject to errors, and as a consequence spatio-temporal interpolations and model predictions are subject to error as well. These errors should be informative to, and used by, the next level of information uptake, be it modelling or decision making.

Acknowledgements

This work is funded by the European Commission, under the Sixth Framework Programme, by the Contract No. 033811 with the DG INFSO, action Line IST-2005-2.5.12 ICT for Environmental Risk Management. The views expressed herein are those of the authors and are not necessarily those of the European Commission. More information on INTAMAP and UncertML can be found on the internet: http://www.intamap.org/ , http://www.uncertml.org/

References

Chorti, A. and D.T. Hristopulos, 2008. Non-parametric Identification of Anisotropic (Elliptic) Correlations in Spatially Distributed Data Sets, IEEE Transactions on Signal Processing, vol. 56(10), pp. 4738-4751, 2008.

N. A. Cressie, "Statistics for spatial data", Wiley, New York, 900 p., 1993.

Cressie, N.A.C., 1993. Statistics for Spatial Data; Revised Edition. John Wiley & Sons, Inc., New York

EUR 21595 EN, Automatic mapping algorithms for routine and emergency monitoring data. Report on the Spatial Interpolation Comparison (SIC2004) exercise. Dubois G. (Ed), European Commission, Office for Official Publications, Luxembourg, 150 p., 2005.

B. Ingram, D. Cornford, and L. Csato, 2008. A projected process kriging algorithm for sensor networks with heterogeneous error characteristics. In: Geostats 2008 - 8th International Geostatistics Congress, J. Ortiz & X. Emery (Eds), GECAMIN Ltd., Santiago de Chile, Chile, December 2008.

Kazianka, H. and Pilz, J. (in press), Spatial Interpolation Using Copula-Based Geostatistical Models, In: geoENV VII - Geostatistics for Environmental Applications (P. Atkinson et al. eds.), Springer, Berlin

Müller, W.G., 2007. Collecting Spatial Data: Optimum Design of Experiments for Random Fields. Springer, New York.

Pebesma, Edzer J., 2006. The Role of External Variables and GIS Databases in Geostatistical Analysis. Transactions in GIS Vol. 10 No. 4, 615-632.

Pilz, J., Kazianka, H. and Spöck, G. (2008), Interoperability - Spatial Interpolation and Automated Mapping, In: Proc. 4th Int. Conference on Information and Communication Technologies in Bio and Earth Sciences HAICTA 2008 (T. Tsiligiridis ed.), Agricultural University of Athens, Athens, 110-118

Skøien, J.O., G. Blöschl, E.J. Pebesma (2008) Geostatistics for automatic estimation of environmental variables - some simple solutions. Georisk, published online.

Skøien, J.O., O. Baume, E. J. Pebesma, and G. B. M. Heuvelink, 2009. Identifying and removing heterogeneities between monitoring networks, Environmetrics, accepted for publication.

Simon Urbanek, 2009. Rserve: Binary R server, R package version 0.4-7. http://www.rosuda.org/Rserve/

Williams, M., Cornford, D., Bastin, L. and B. Ingram. (2008). UncertML: an XML schema for exchanging uncertainty. In: "Proceedings of the GISRUK 2008", pp. 275-279, April 2008, Manchester, UK.

Williams, M., Cornford, D., Ingram, B., Bastin, L., Beaumont, T., Pebesma, E. and G. Dubois (2007). Supporting interoperable interpolation: the INTAMAP approach. Presented at: International Symposium on Environmental Software Systems (ISESS 2007), Prague, Czech Republic, 22-25 May 2007.