INTAMAP: an interoperable automated interpolation web service

Edzer Pebesma[1], Dan Cornford[2], Gregoire Dubois[3], Gerard Heuvelink[4], Dionisis Hristopoulos[5], Juergen Pilz[6], Ulrich Stoehlker[7], Jon Skoien[8].

[1]Institute for geoinformatics, Univ. of Münster, edzer.pebesma@uni-muenster.de
[2]Computer Science, Aston University, [3]Joint Research Centre of the European Commission, [4]Wageningen University, [5]Technical University of Crete, [6]University of Klagenfurt, [7]Bundesamt fur Strahlenschutz, [8]Dept of Physical Geography, Utrecht University.

Abstract. INTAMAP is a web processing service for the automatic interpolation of measured point data. Requirements were (i) using open standards for spatial data such as developed in the context of the open geospatial consortium (OGC), (ii) using a suitable environment for statistical modelling and computation, and (iii) producing an open source solution. The system couples the 52-North web processing service, accepting data in the form of an observations and measurements (O&M) document with a computing back-end realized in the R statistical environment. The probability distribution of interpolation errors is encoded with UncertML, a new markup language to encode uncertain data. Automatic interpolation needs to be useful for a wide range of applications and the algorithms have been designed to cope with anisotropies and extreme values. In the light of the INTAMAP experience, we discuss the lessons learnt.

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 simple solution exists. In an experiment where several experts were confronted with interpolating the same data set [3], the approaches differed strongly, and best results were obtained by machine learning techniques as well as geostatistical methods. One of the reasons behind this variety was that one needs to choose a model of spatial variability before one can interpolate, and experts disagree on which models are most useful.

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, 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 too simplistic for the application at hand.

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 has built an automated interpolation web service that provides useful interpolation without requiring any specialised skills. This was realized using open standards, and using and providing an open source software solution. As interpolation cannot be done without introducing errors, the interpolation service returns meaningful information about the interpolation error characterising 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 evacuating areas that required evacuation), or deciding where monitoring needs to be increased or decreased.

The paper is organised as follows. First, the statistical considerations and challenges underlying automated mapping will be discussed. The technical realization and system architecture will be described. Issues of performance and embedding it in a service oriented / service chained environment are addressed. 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 estimated. In geostatistics, models of the form $Z(s) = m(s) + e(s)$ are usually deployed [2], 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 $m(s) = X(s)^T\beta$ with $X(s)$ often layers in a GIS [6], β unknown regression coefficients, and $e(s)$ usually a second order stationary residual process. This first step entails the choice of a trend function, a covariance function for the residual process, and the estimation of parameters of both components. The second step involves, given this model and the observations, the spatial interpolation (prediction) of this model for new locations s_0: $\hat{Z}(s_0) = \hat{m}(s_0) + \hat{e}(s_0)$ where s_0 is usually taken over a grid covering the region of interest.

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 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 among contributing countries 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 [3], 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 an interpolation method based on spatial copulas to interpolate these data [5]. Spatial copulas are flexible models which combine separate specification of correlation structure and spatial process marginal distributions, thus allowing very general non-Gaussian kriging to be employed.

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

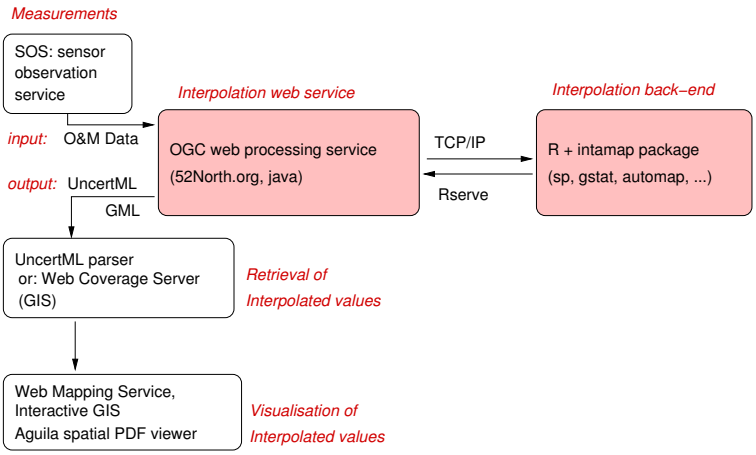

Figure 1: Technical set up of the automatic interpolation service. UncertML stands for uncertainty markup language (see text); O&M stands for observations and measurements, an XML standard for encoding monitoring network data.

[1], and if it is, corrects for anisotropy by transforming coordinates to an isotropic space, before further steps are taken.

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 if error characteristics of the observations are specified a sequential interpolation method based on projected sequential Gaussian processes [4] is used to optimally interpolate the spatial field.

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. Spatial aggregation can in some cases be done by simple aggregation of a series of point predictions, but particular methods are necessary for estimates of the associated error distributions and for non-linear aggregates. The INTAMAP interpolation service will use the best aggregation method for the problem at hand.

3 TECHNICAL REALISATION

OGC Web Services Web service standards as agreed upon by e.g. ISO TC211, OGC and INSPIRE are the basis for useful generic services to exchange geographic data. INTAMAP delivers an interpolation web processing service (built on the open source 52North implementation) 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 coverage. To encode the interpolation error UncertML, a markup language for specifying information that is represented probabilistically, has been developed within the project, which OGC has currently released as an OGC Discussion Paper [8].

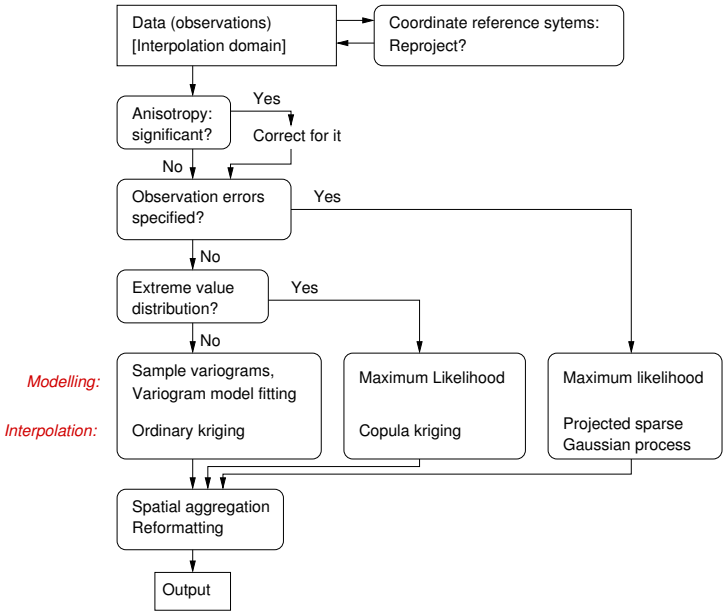

Figure 2: Decision tree for the interpolation method choices in the interpolation process that takes place in R. References in text.

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[7]) has the advantage that the R process, doing the numerical work, may be running on a dedicated computing cluster not directly connected to the internet. Multiple interpolation requests at the same time will be executed in parallel. A second advantage of having all statistical routines in the R environment is that it can be re-used independently of the WPS interface, e.g. interactively on a PC, from a 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. During the testing period, the the following issues need to be further

investigated before setting up a robust public service that allows everyone to use it: (i) 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 is currently solved by reducing the system by default to only address the nearest 50 observations. We will examine the value of this number and the possibility of choosing n in a more flexible way. We note the projected sequential method is less affected by the size of the observation set. If necessary the user is able to specify that a specific method and parameter setting is used, although the default is automatic interpolation. (ii) Some of the interpolation methods implemented need a considerable amount of time to process, of the order of hours or more; the interpolation service has been set up to only select methods that are estimated to finish within a time limit defined in the request, defaulting to 30 seconds. The estimation of processing time can still be improved. Asynchronous protocols are implemented in the reference WPS used. (iii) 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. (iv) Besides interpolated values, the interpolation service is able to return more information, such as: which method was used, what the values of the fitted parameters are, and maybe even some relevant diagnostic plots, e.g. as the variogram and fitted model. We have to consider which one would be most useful. (v) The observations read by the INTAMAP interpolation service need to be contained in 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.

5 DISCUSSION AND OUTLOOK

The automated interpolation web service, the main deliverable of INTAMAP, 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 for off-line use.

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 can be used as a first major component to build such a specific interpolation service.

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 de-

velopment can be used as a first building block for it. One motivation for not addressing 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 services) 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

[1] A. Chorti and D.T. Hristopulos. Non-parametric identification of anisotropic (elliptic) correlations in spatially distributed data sets. *IEEE Transactions on Signal Processing*, 56(10):4738–4751, 2008.

[2] N. Cressie. *Statistics for Spatial Data, Revised edition*. Wiley, 1993.

[3] 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, 2005.

[4] B. Ingram, D. Cornford, and L. Csato. A projected process kriging algorithm for sensor networks with heterogeneous error characteristics. In J. Ortiz and X. Emery, editors, *Geostats 2008 - 8th International Geostatistics Congress*, 2008.

[5] H. Kazianka and J. Pilz. Spatial interpolation using copula-based geostatistical models. In P. Atkinson, editor, geoENV VII - Geostatistics for Environmental Applications, 2009.

[6] E.J. Pebesma. The role of external variables and GIS databases in geostatistical analysis. *Transactions in GIS*, 10(4):615–632, 2006.

[7] S. Urbanek. Rserve: Binary R server, R package version 0.4-7. 2009.

[8] M. Williams, D. Cornford, L. Bastin, and E. Pebesma. Uncertainty Markup Language (UncertML). *OGC Discussion Paper, Document Number: 08-122r1*, 2009.