INTAMAP: the design and implementation of an interoperable automated interpolation web service

Edzer Pebesma[a,*], Dan Cornford[b], Gregoire Dubois[c], Gerard B.M. Heuvelink[d], Dionisis Hristopoulos[e], Jürgen Pilz[f], Ulrich Stöhlker[g], Gary Morin[h], Jon O. Skøien[i]

[a] *Institute for geoinformatics, University of Münster*
[b] *NCRG, Engineering and Applied Science, Aston University*
[c] *Joint Research Centre of the European Commission*
[d] *Wageningen University*
[e] *Technical University of Crete*
[f] *University of Klagenfurt*
[g] *Bundesamt fur Strahlenschutz*
[h] *Keynetix Systems*
[i] *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 integrated, open source solution. The system couples an open-source Web Processing Service (developed by 52°North), accepting data in the form of standardised XML documents (conforming to the OGC Observations and Measurements standard) with a computing back-end realized in the R statistical environment. The probability distribution of interpolation errors is encoded with UncertML, a markup language designed 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 anisotropy, extreme values, and data with known error distributions. Besides a fully automatic mode, the system can be used with different levels of user control over the interpolation process.

Key words: 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

*Weseler Strasse 253, 48151 Münster, Germany; edzer.pebesma@uni-muenster.de

is a problem for which no simple, 'one-fits-all' solution exists. In an experiment where several experts were confronted with interpolating the same data set (EUR, 2005), 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, who work in fields such as mining, petroleum industry, environmental monitoring, or risk assessment, 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 that are too simplistic for the application at hand.

Motivated on one hand by the increasing availability of sensor data offered in near real time, and on the other by the need to take quick decisions in cases such as disaster management, where there is no time to consult interpolation experts, the INTAMAP[1] project has built an automated interpolation web service that provides interpolation without requiring any specialised skills. This was realized employing open standards together with using and providing an open source software solution[2]. As interpolation cannot be done without introducing interpolation 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 define probabilities of exceeding a number of given thresholds. Such error information may 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 – think of evacuating areas not in danger, or not evacuating areas that should have been evacuated), or deciding where monitoring efforts needs to be increased or can be decreased.

The paper is organised as follows. First, the interpolation challenges faced when developing an automated mapping system will be discussed. Next, the statistical methods and decisions underlying the system will be described. Then, the technical realization and system architecture are presented. Issues of performance and embedding it in a service oriented environment are addressed. Finally, the discussion will provide a perspective on how this service might be used and extended, along with ideas for future developments of environmental management systems based on service oriented architectures (SOA).

[1] http://www.intamap.org/

[2] http://www.sourceforge.net/projects/intamap/

2. The interpolation challenge

The interpolation problem we want to solve is the following: given a set of measurements of a continuous process, compute the best prediction at one or more unmeasured locations, along with characteristics of the interpolation error distribution such as the variance or quantiles.

Spatial interpolation can be seen as consisting of three stages. In the first stage, a model for the spatial variability has to be selected. In the second, its parameters are estimated. In the third stage, given this model and these estimates, a prediction of the measured process is used to interpolate, and a prediction error is characterized.

2.1. A general model

In geostatistics, typically models of the form

$$Z(s) = m(s) + e(s) \tag{1}$$

are deployed (Cressie, 1993), with $Z(s)$ the measured process at spatial location s, $m(s)$ the spatially varying (or constant) trend component, which could be modelled as a linear in parameters regression model of the form $m(s) = X(s)^T\beta$ with $X(s)$ often derived from layers in a GIS (Pebesma, 2006), ^{T} denoting matrix transpose, β unknown regression coefficients, and $e(s)$ usually a second order or intrinsically stationary residual process. Non-linear geostatistical models may extend this to

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

with $f(\cdot)$ some non-linear function, such as the Box-Cox transform,

$$f(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(x) & \text{if } \lambda = 0 \end{cases} \tag{2}$$

and then aim to predict $Y(s)$, carefully using the back transformation to finally predict $Z(s)$.

The first stage in a geostatistical analysis entails the choice of a trend function, a covariance function for the residual process, and enables the estimation of parameters of both components in stage two. The third stage involves, given this model and the observations, the spatial interpolation (prediction) using the estimated model for new locations s_0, for the linear case:

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

with new locations s_0 usually taken over a grid covering the region of interest. Initially, the interpolation challenge addressed here ignores availability of trend variables $X(s)$, i.e. $X(s) = 1$ and $m(s) = \beta_0$ is an unknown constant. More general cases will be mentioned in the discussion.

Spatial prediction under the linear case of model (1) would lead to universal or ordinary kriging. Three special cases will be discussed next, namely

the detection of geometric anisotropy, the presence of extreme values, and the possibility to deal with known measurement errors and sensor models.

This project did not attempt to realize spatio-temporal prediction in an automated fashion. The motivation for this is discussed in section 5.3.

2.2. Geometric anisotropy

Many environmental variables are subject to geometric anisotropy, meaning the degree of spatial continuity is in some direction stronger than in others. This phenomenon is typical for atmospheric pollutants diffusing like a plume in a particular direction.

Hristopulos (2002) and Chorti and Hristopulos (2008) have developed methodology for the detection of geometric anisotropy directly from point data. The method, referred to as CTI, is based on the covariance tensor identity (Swerling, 1962), which links sample-based gradients with the Hessian of the covariance function. Assuming differentiability, CTI can estimate geometric anisotropy without specifying a covariance model. This method is used in the INTAMAP interpolation service.

2.3. 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[3], collects radiological monitoring data from around 4000 sensors spread over most European countries and is available to decision-makers in near real-time. The main purpose of this network is motivated by emergency cases, where the exchange of these data among contributing countries 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 previously (EUR, 2005), are characterised by many low observations and very few observations with extreme, outlying measured values. Such data sets violate the assumptions behind ordinary kriging and universal kriging, model (1). The INTAMAP automated interpolation service deals with data containing extreme outliers, and primarily deploys an interpolation method based on spatial copulas to interpolate these data (Kazianka and Pilz, in press, 2009). Spatial copulas are flexible models that combine separate specification of correlation structure and spatial process marginal distributions, thus allowing very general non-Gaussian kriging to be employed.

Kazianka and Pilz (in press) present a copula-based spatial modelling and interpolation approach that works with both continuous and discrete marginal distributions and which makes it possible to include covariates e.g. a spatial

[3] http://eurdep.jrc.ec.europa.eu/

trend or elevation. They show that their model generalizes trans-Gaussian kriging (Cressie, 1993) and provides an alternative to generalized linear geostatistical models Diggle and Ribeiro (2007). A Bayesian extension of the spatial copula model is given in Kazianka and Pilz (2009c).

Experimenting with the spatial copulas for interpolation has taught us that their run time requirements may be large: for large data sets and/or a large number of prediction locations, computation may require hours or even days. As a fall-back method for the case where copulas require too much time, trans-Gaussian kriging has been implemented with the Box-Cox family (2) of power and log transformations.

2.4. Observations from sensors 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 or elaborate and complicated calibration. An example of this are the atmospheric chemistry measurements from satellites such as OMI (Boersma et al., 2004). Interpolation of data with considerable, known measurement error and / or known sensor models should take these sensor models and errors into account.

In the INTAMAP interpolation service if error characteristics of the observations are specified a sequential interpolation method (Csató and Opper, 2002) based on projected sequential Gaussian processes (Ingram et al., 2008a) is available to optimally interpolate the spatial field. The main benefits of the *sequential* approach are that they permit the treatment of non-Gaussian observation errors and non-linear sensor models without requiring high dimensional integrals to be computed (Cornford et al., 2005). The *projected* nature of the algorithm, which can be related to the fixed rank approach of Cressie and Johannesson (2008), makes it possible to control the computational complexity of the posterior approximation (Ingram et al., 2008b) which makes the treatment of large data sets possible, despite the use of likelihood based parameter inference within the algorithm.

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 that of emergency evacuation: we cannot evacuate single points, but decide whether neighbourhoods, regions, villages, towns, or flood plain sections will be evacuated. In addition, methods were developed and employed to correct for systematic errors between measurements from different networks (Skøien et al., 2009).

3. Statistical implementation: the interpolation decision tree

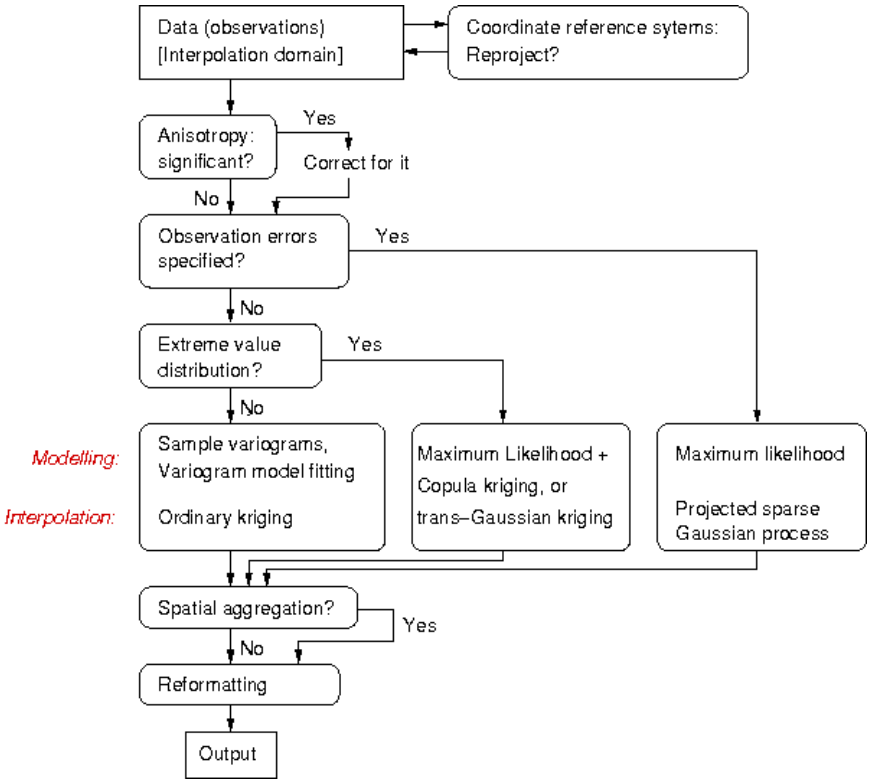

Figure 1: Decision tree for the interpolation method choices in the interpolation process that takes place in R. References in text.

The decision tree for choosing an interpolation method automatically is shown in Figure 1. In the context of the INTAMAP project, dedicated interpolation methods have been implemented for (i) detecting and correcting for anisotropy, (ii) dealing with skewed and extreme value distributions, (iii) dealing with known measurement errors. Details of the major choices will be given now.

3.1. Coordinate reference systems

Interpolation on the sphere, i.e., on longitute/latitude coordinates, is not supported by the software developed. It is, however, allowed that the measurement data come with longitute/latitude coordinates, and the prediction locations (e.g. the prediction grid) are in a certain projection. Before interpolation, in that case, the measurement data coordinates are projected to this target projection. Alternatively, for more user control a target projection system can be specified for both.

3.2. Anisotropy significant?

Anisotropy is taken into account in each of the interpolation procedures of Figure 1, if the anisotropy ratio, i.e. the ratio between the correlation ranges in the major and minor direction, is significantly different from one. A procedure to efficiently obtain an approximate confidence interval for the CTI estimate of the anisotropy ratio is given in Petrakis and Hristopulos (submitted). The confidence interval estimate is used to test for significant anisotropy in the sample, which would require the use of an anisotropic variogram model. If the latter is required, CTI estimates of the anisotropy parameters (anisotropy ratio and principal axes orientation) are used in the variogram model. In maximum likelihood fitting of variogram functions, the CTI estimates can provide initial values of anisotropy.

3.3. Extreme value distribution?

The decision whether to consider a variable as having an extreme value distribution is also taken automatically (Figure 1). For a given variable $z = (z_1, ..., z_n)$, the data were considered strongly non-Gaussian when for the derived variable z', defined by

$$z' = \begin{cases} z - \min(z) + \sigma_z & \text{if } \min(z) \leq 0, \\ z & \text{if } \min(z) > 0, \end{cases}$$

with σ_z the standard deviation of z, if one or more of the following conditions holds:

1. more than 10% of the values of z' lie beyond the extremes of the whiskers of a boxplot (i.e., are further than 1.5 IQR away from the nearest quartile, with IQR the inter-quartile range of z')
2. $Q_{.5} - Q_{.25} < IQR/3$, with $Q_{.5}$ and $Q_{.25}$ the median and first quartile of z', respectively.
3. $Q_{.75}(z') - Q_{.5} < IQR/3$, with $Q_{.75}$ the third quartile of z'
4. the identity transform $\lambda = 1$ in the Box-Cox transform (2) falls outside the 90% confidence interval for the estimated Box-Cox parameter for z'.

The first condition checks whether more than 10% of the data are "outlying", i.e. outside the whiskers of the boxplot. The second and third inequalities check for asymmetry/skewness of the data. The last checks whether a Box-Cox transformation makes the data distribution significantly closer to normality.

It should be noted that criteria 1-3 are in principle insensitive to sample size. Criterion 4 however involves significance testing. This means that for very large samples, very small, asymmetric deviations from normality will lead to a decision that non-linear transformation (i.e., using copula or trans-Gaussian kriging) is advisable. Of course, for λ values close to one, this non-linearity is modest and trans-Gaussian kriging is practically similar to ordinary kriging.

3.4. Automated variogram modelling

Automated, omnidirectional variogram model fitting takes place according to the following steps:

1. Let D be 0.35 times the length of the diagonal of the box that spans the data, i.e. $D = 0.35\sqrt{dx^2 + dy^2}$ with dx the range of x coordinates and dy the range of y coordinates of the data points;
2. for p distance intervals with boundaries 0, 2, 4, 6, 9, 12, 15, 25, 35, 50, 65, 80, and 100% of D, compute the classical, omnidirectional sample variogram values $\hat{\gamma}(h_i)$ with h_i the average distance of all point pairs available in distance interval i;
3. while the number of point pairs for lag i, N_i, in the first lag interval is smaller than 5, merge the first two distance intervals, recompute the sample variogram, and lower p with one;
4. choose the following variogram model parameters as initial values for the fit:
   - sill: mean of the maximum and median value of the $\hat{\gamma}(h_i)$
   - nugget: minimum value of the $\hat{\gamma}(h_i)$
   - range: $D/3.5$, which is 10% of the length of the diagonal that spans the data
   - candidate values for the smoothness parameter κ of the Matern variogram model: 0.05, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0, 5.0, and 10.0;
5. for each of the variogram model types *spherical*, *exponential*, *Gaussian*, and for the *Matern* model for all κ values, fit the variogram model with weighted least squares, using a Gauss-Newton algorithm, by minimizing the weighted error sum of squares

$$\mathrm{SS}_{err} = \sum_{i=1}^{p} \frac{N_i}{h_i^2} (\hat{\gamma}(h_i) - \gamma(h_i))^2$$

   with $\gamma(h_i)$ the variogram model value for distance h_i;
6. select the model (model type, nugget, sill, range, κ in case of the *Matern* model) with the smallest SS_{err}.

These heuristic steps have been presented earlier in Hiemstra et al. (2009) and some of them in Pebesma (2004, 2005) and Pebesma and Wesseling (1998).

For the methods that used maximum likelihood fitting of variogram parameters, the variogram model and fitted values obtained by the procedure above were taken as initial values. No further iteration over various variogram model types took place by maximum likelihood.

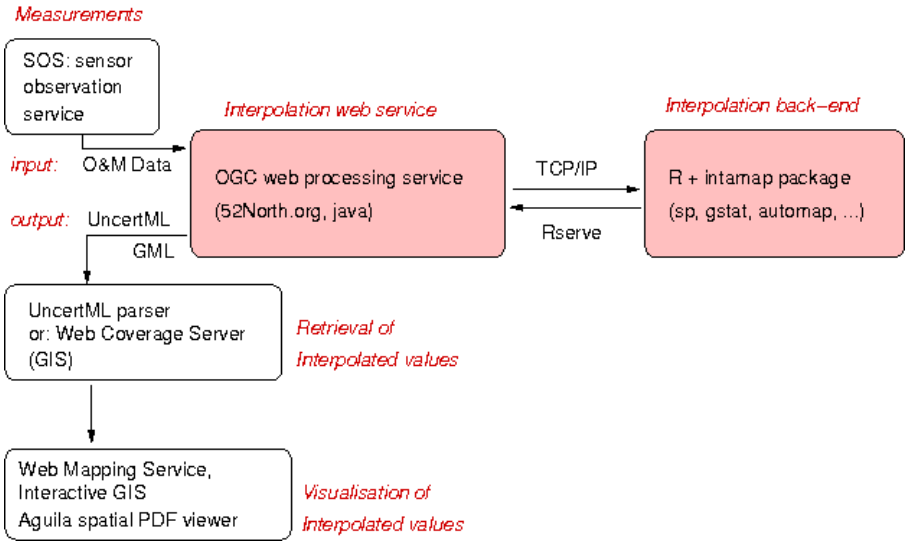

Figure 2: Technical set up of the automatic interpolation service. UncertML stands for uncertainty markup language (see text); O&M stands for observations and measurements, an XML standard for encoding monitoring network data.

3.5. Spatial aggregation

Spatial aggregation can in some cases be done by simple aggregation of a series of point predictions, but other methods are necessary for estimation of the associated error distributions and for non-linear aggregates. The INTAMAP interpolation service uses the most efficient aggregation method for the problem at hand: block kriging where this is allowed and available, and simulation with numerical aggregation otherwise.

4. Technical realisation

4.1. OGC Web Services

Web service standards as agreed upon by standard bodies such as ISO TC211[4], OGC[5], and adopted by the European INSPIRE directive (EUR, 2007) are the basis for useful generic services to exchange geographic data. INTAMAP has delivered an interpolation Web Processing Service (WPS, Schut (2007)) that is based on the open source 52°North implementation[6]) and is schematically shown in Figure 2. It accepts sensor data from a Sensor Observation

[4]http://www.isotc211.org
[5]http://www.opengeospatial.org/
[6]http://www.52north.org/

Service (SOS, Na and Priest (2007)), i.e. encoded as an Observations & Measurements document (O&M, Cox (2007)), and returns the interpolation result e.g. a GML document of a coverage encoded as a `gml:RectifiedGrid`. To encode the interpolation error UncertML, a markup language for specifying uncertain information that is represented probabilistically, has been developed within the project, which OGC has currently released as an OGC Discussion Paper[7] (Williams et al., 2009).

4.2. The interpolation back-end in R

The procedures for the statistical analysis of the data are implemented in extension packages for R, the major open source environment for analysing statistical data. As Figure 2 shows, this is not apparent to the user of the INTAMAP Web Processing Service, since R runs only at the back-end. Interfacing R from the Web Processing Service by using the http protocol (i.e., as a web service, using the Rserve package, Urbanek (2009)) has the advantage that the R process, doing the numerical work, may be running on a dedicated computing cluster behind a firewall. Coordinate transformations are also done in R (Bivand et al., 2008). Multiple interpolation requests at the same time will be executed in parallel.

4.3. Clients

To better illustrate the flexibility of the architecture from a users' point of view, the following clients have been developed to interface the INTAMAP interpolation service:

1. A thick, yet still web based, client based on the mapguide open source[8] client has been developed for use in geotechnical applications, e.g. for the interpolation of soil geochemical variables and borehole information. In the client, one can request all properties of the interpolation process: mean, variance, quantiles or probabilities of exceeding a threshold (Figure 3).
2. An application for the visualization and interpolation of gamma dose rate data on the European scale. The application is based on the deegree-framework[9] and uses a Web Map Service (WMS 1.3) interface to show monitoring network locations along with grid maps returned from the INTAMAP WPS. It transforms the interpolation results into maps according to national guidelines for radiation mapping. In addition a client based on mapbender[10] is under development. Internally BfS, the Germany radiological monitoring authority, also uses an application that directly calls the INTAMAP R package for interpolation.

[7] http://xml.coverpages.org/OGC-UncertML.html
[8] http://mapguide.osgeo.org/
[9] http://www.deegree.org
[10] www.mapbender.org

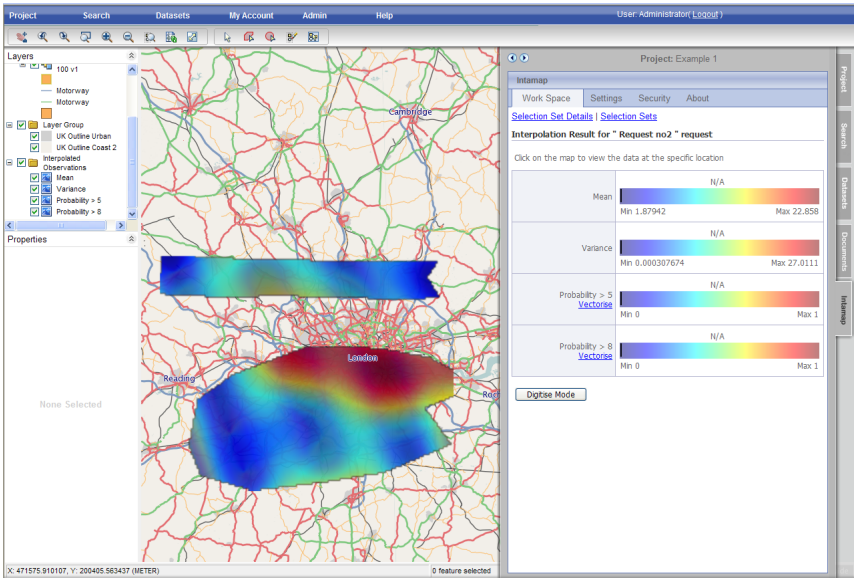

Figure 3: The INTAMAP client based on mapguide OS; on the right, the user can select and blend various properties of the probability distribution of the interpolated value.

3. A dedicated client for the interpolation of European near real-time air quality data from a SOS (Henneböhl et al., 2010), that uses the aguila (Pebesma et al., 2007) interactive viewer of spatio-temporal data encoded as probability distributions (Figure 5).
4. The tryIntamap web client, a simplified web form[11] that allows to paste comma-separated data with `x,y,value` on each line and interpolate this; results are shown as images (Figure 4), and on a Google Earth plugin.
5. A Java based mobile phone client that can query observations from a simplified SOS interface, or allow users to add observations in the field, which can then connect to the INTAMAP service through a simplified interface developed to minimise bandwidth usage and shows interpolated values, and uncertainties on a mobile phone or other mobile device.

5. Discussion

5.1. Spatial interpolation

The INTAMAP project has contributed to change the status of geostatistical interpolation from a playing field where one needed to have expertise, knowledge of jargon, be able to use complicated routines and a lot of experience to

[11] available at http://www.intamap.org/

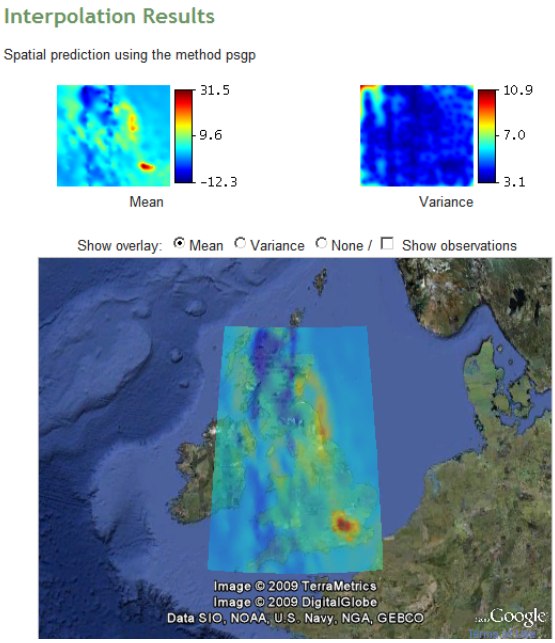

Figure 4: An example of graphical output of the tryIntamap simple web page client.

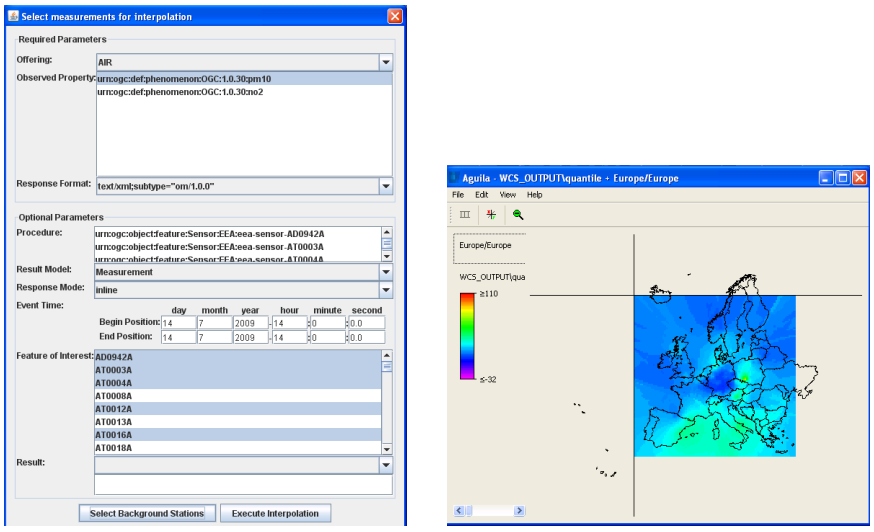

Figure 5: Example graphical output of the dedicated SOS-WPS client for interpolating air quality data; selection screen (left), interpolation result in aguila (right).

compute and model variograms, to a status where one has a single function or web service where data are sent, and interpolation along with interpolation error characteristics are returned. As a result automatic mapping functions based on advanced algorithms are available for real-time applications *and* these functions can be used further for benchmarking exercises by the research community. This outcome has been realised by finding consensus within a limited group of researchers. A number of issues can potentially lead to a better interpolation system, in particular:

- the decisions, especially the inclusion of condition 4, when to change to using non-Gaussian (copula or trans-Gaussian) kriging,
- the exact steps taken for methods-of-moments variogram computation and model fitting, notably the lack of adjustment of D and the choice of lag intervals.

Other issues result from the fact that software development was carried out by project partners with different backgrounds. Two items can be mentioned:

- 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. The projected sequential (psgp) method is designed to work efficiently for large data sets, and is able to work with global neighbourhoods.
- Developing software with a group of developers across project partners from different countries inevitably leads to some extent to isolated decisions being taken by contributers, and inhomogeneity in the final software system. Among these inhomogeneities, we should mention (i) two of the three interpolation methods use maximum likelihood for estimating model parameters, the other uses the method of moments estimator; (ii) during maximum likelihood estimation, the copula method further adjusts anisotropy parameters, whereas the psgp method does not, (iii) only the ordinary kriging method can deal with external predictors of the form of (1), in which case it changes to universal/external drift kriging.

As this is the outcome of a research project from several partners that have a lasting interest in the methodology and technology, it is hoped that involvement of a larger group of users, software developers and geostatisticians will help continue improving on the current state of this open source project.

The INTAMAP interpolation procedure can be seen as a statistical model, with several sub-models. As for any statistical model, it is possible to come up with cases where the model will not work. Kriging will not work with duplicate observations; variogram modelling will fail in the presence of strong outliers. Removing errors or duplicates is not dealt with automatically, as we feel it is the responsibility of the user (or client) to address them. There are special cases

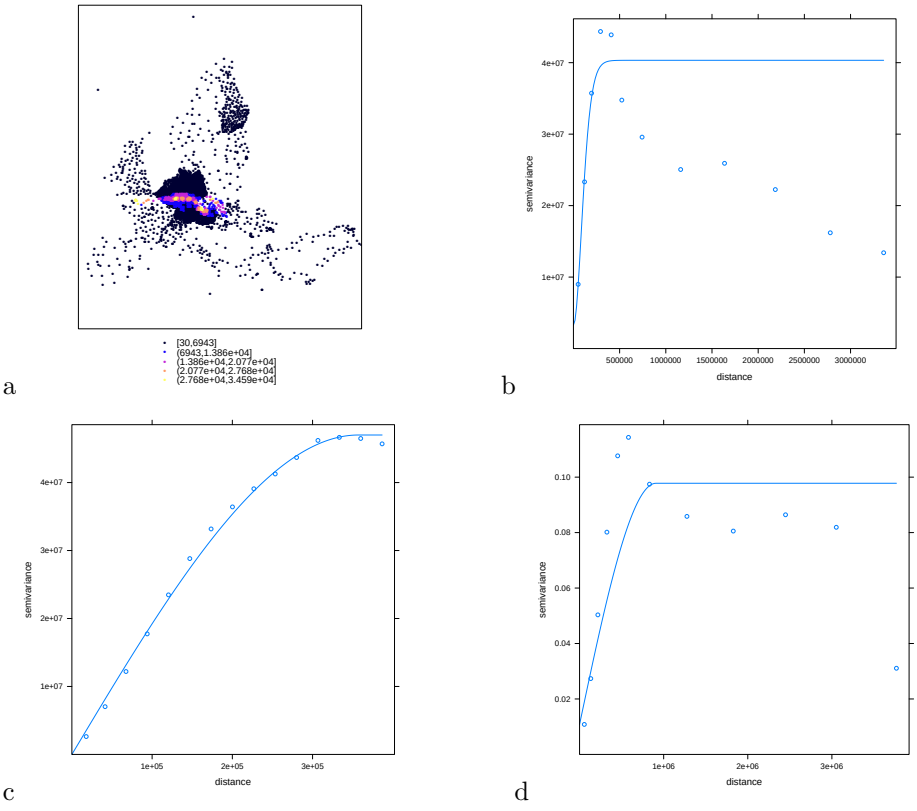

Figure 6: a. Simulated data of measured gamma dose rate values after a modelled release; b. Automatically modelled and fit variogram on raw data; c. Idem, but after manual adjustment of the maximum distance of the variogram, d. Automatically modelled and fit variogram on Box-Cox transformed data with $\lambda = -0.3$

of interpolation problems for which the interpolation service developed will not work well. Examples are: (i) there are fewer than 30 observations (an error will be raised), (ii) the process $Z(s)$ contains an extreme, but the data carries very little information about it, and (iii) the process is mixed discrete-continuous, e.g. daily rainfall in convective systems. For special cases like (iii), dedicated procedures (e.g., Schuurmans et al. (2007)) could be developed and added to the current system. An example where the automatic variogram fitting for untransformed data did not work well, is shown in Figure 6. For this case, the automatic procedure would have chosen the copula or trans-Gaussian methods, which appears to give a better fit in Figure 6.)

As opposed to a generic interpolation algorithm, the interpolation of real variables with known characteristics would in addition to measurement data typically use further information: for air quality one would like to use remotely sensed data, land use and/or traffic information, for environmental radioactiv-

ity it might make sense to use geology and altitude. Although such information might be 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), unless one includes the (reference to the) additional information in the service request and provides generic models only, e.g. the universal kriging/external drift model (Hengl et al., 2004). The generic interpolation service developed here can be used as a first major component to build such a specific interpolation service. An example is given in Henneböhl et al. (2010).

It should be noted that several of the models implemented at the R level (copula, ordinary kriging) do provide for the inclusion of trend variables.

5.2. Statistical methods beyond interpolation

In addition to interpolation methods, methods for monitoring network harmonisation (bias estimation and removal) 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 and experience gained during INTAMAP, it would be relatively simple to customize the INTAMAP web service and perform these manipulations.

5.3. Extensions: space-time

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. For example, Hiemstra et al. (2009) and Henneböhl et al. (2010) used a spatial nugget over time that was fixed over time to suppress some of the variation that otherwise occurred while interpolating subsequent time steps of gamma dose rate and PM_{10}, respectively.

One motivation for not addressing time was that in space-time modelling through a spatio-temporal stationary covariance function 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 problem. 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 approximate reality and introduce further 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. The development of UncertML (Williams et al., 2009) within the INTAMAP project has been a major first step in this direction.

5.4. Interfaces and technology

In the current implementation, in an interpolation service request the user is able to specify that a specific method, and optionally a specific parameter setting is used. The default is automatic interpolation (automated method selection, variogram model selection and parameter fitting).

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, which defaults to 30 seconds. The timing is estimated by fitting a loess function for each interpolation method to a large set of combinations of number of observations and number of prediction locations. When installing the R INTAMAP package, the user can choose to recalibrate the time estimation functions.

Asynchronous use of the service has not been implemented. Although the WPS 1.0 standard does provide the infrastructure for this, it would have been a considerable challenge to retrieve the progress status of the R process, as asynchronous communication with R would have to be set up as well, and all R functions would require extension to report their progress status. The current service allows approximate time evaluation prior to execution, and the advantages of having asynchronous interaction with the WPS were not considered large enough to give it high priority.

The observations read by the INTAMAP interpolation service need to be contained in an O&M document, 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, which are cases that make little sense to send to an interpolation service. Availability of SOS profiles for particular application areas (such as meteorology or seismology) should make the adoption more easy. The INTAMAP package will attempt to take geographic projections into account, if possible. This includes the possibility of supplying observations and prediction locations in different projections.

Several instances of the INTAMAP interpolation service are publicly accessible for testing and research purposes. Users may also download the software and install and run the service in their own public or private environment.

Besides interpolated values, the interpolation R process can return meta-information, such as: which method was used, what the values of the fitted parameters are, and maybe even some relevant diagnostic plots, e.g. of the sample variogram and fitted model. This information is currently not passed through the web service interface.

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. Availability, scaling and load balancing has not been addressed, but may become an issue. Solutions for this are found in the area of grid and cloud computing (Woolf and Shaon, 2009; Baranski, 2009).

6. Conclusions

The generic automatic interpolation service developed in the INTAMAP project can be used, installed, deployed, extended and/or modified free of charge. It copes with a number of "difficult" cases that include data with strong anisotropy, data with skewed or extreme value distributions, and data with known measurement errors.

All the interpolation software has been developed as R packages that can be downloaded from CRAN[12] and can in addition to the web service be used interactively or be used from other environments, e.g. from python or shell scripting, data bases, through SOAP (W3C, 2007), or on a mobile device, meaning that potential users of the interpolation technology are not forced to use the WPS protocol. Further uptake and usage as well as further development of suitable clients will extend the success of the interpolation Web Processing Service.

The current version should be seen as a starting point that is open for further improvement and joint development with users and specialists.

7. Where to find the software

All software developed, as well as several test data sets are available from sourceforge[13], project intamap, and are distributed under the GPL version 2[14] or higher. Working instances of the interpolation service are mentioned on the intamap web site[15]. The R packages produced by the INTAMAP project that are now on CRAN are:

intamap provides the basic interpolation routines and classes, and all code for anisotropy estimation and spatial copulas

psgp provides projected sparse Gaussian process code; it requires the C++ library IT++ to be installed on the host system, and is therefore only available as a (linux) source package, at this moment not as binary package for Windows or MacOS.

intamapInteractive provides code for bias estimation, bias removal, and monitoring network optimization.

Acknowledgements

This work has been 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

[12] http://cran.r-project.org/
[13] http://www.sourceforge.net/projects/intamap
[14] http://www.gnu.de/documents/gpl-2.0.en.html
[15] http://www.intamap.org/

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 http://www.intamap.org/ and, http://www.uncertml.org/ , respectively.

References

Baranski, B., 2009. Ogc OWS-6 WPS grid processing profile engineering report. OGC document 09-041r3.
URL http://portal.opengeospatial.org/files/?artifact_id=34977

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., 2008. Applied Spatial Data Analysis with R. Use R. Springer, New York.

Boersma, K. F., Eskes, H. J., Brinksma, E. J., 2004. Error analysis for tropospheric NO2 retrieval from space. Journal of Geophysical Research: Atmospheres 109, D04311.

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

Cornford, D., Csato, L., Opper, M., 2005. Sequential, Bayesian Geostatistics: A principled method for large data sets. Geographical Analysis 37, 183–199.

Cox, S., 2007. Observations and Measurements – part 1 - observation schema. OGC document 07-022r1.
URL http://portal.opengeospatial.org/files/?artifact_id=22466

Cressie, N., 1993. Statistics for Spatial Data, Revised edition. John Wiley and Sons, Inc.

Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70, 209–226.

Csató, L., Opper, M., 2002. Sparse online Gaussian processes. Neural Computation 14, 641–669.

Diggle, P., Ribeiro, P., 2007. Model-based Geostatistics. Springer, New York.

EUR, 2005. 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, EUR 21595, EN.

EUR, 2007. Directive 2007/2/ec of the european parliament and of the council of 14 march 2007 establishing an infrastructure for spatial information in the european community (inspire). Official Journal of the European Union L 108/1 EN.
URL http://eur-lex.europa.eu/JOHtml.do?uri=OJ:L:2007:108:SOM:EN:HTML

Hengl, T., Heuvelink, G., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120 (1-2), 75–93.

Henneböhl, K., Gerharz, L. E., Pebesma, E. J., 2010. OGC web services for near real-time air quality modeling at a European scale. Computers and Geosciences this issue.

Hiemstra, P., Pebesma, E., Twenhöfel, C., Heuvelink, G., 2009. Real-time automatic interpolation of ambient gamma dose rates from the dutch radioactivity monitoring network. Computers and Geosciences 35 (8), 1711–1721, dOI: http://dx.doi.org/10.1016/j.cageo.2008.10.011.

Hristopulos, D. T., 2002. New anisotropic covariance models and estimation of anisotropic parameters based on the covariance tensor identity. Stochastic Environmental Research and Risk Assessment 16 (1), 43–62.

Ingram, B., Cornford, D., Csato, L., 2008a. A projected process kriging algorithm for sensor networks with heterogeneous error characteristics. In: Ortiz, J., Emery, X. (Eds.), Geostats 2008 - 8th International Geostatistics Congress. p. in press.

Ingram, B. R., Cornford, D., Evans, D. J., 2008b. Fast algorithms for automatic mapping with space-limited covariance functions. Stochastic Environmental Research and Risk Assessment 22, 661–670.

Kazianka, H., Pilz, J., 2009. Spatial interpolation using copula-based geostatistical models. In: Atkinson, P. (Ed.), geoENV VII - Geostatistics for Environmental Applications. p. in press.

Kazianka, H., Pilz, J., 2009c. Bayesian spatial modeling and interpolation using copulas. Computers and Geosciences, this volume.

Kazianka, H., Pilz, J., in press. Copula-based geostatistical modeling of continuous and discrete data including covariates. Stochastic Environmental Research and Risk Assessment.

Na, A., Priest, M., 2007. Sensor Observation Service. OGC document 06-009r6. URL `http://portal.opengeospatial.org/files/?artifact_id=26667`

Pebesma, E., 2005. Mapping radioactivity from monitoring data, automating the classical geostatistical approach. Applied GIS 1 (2), 1–10, dOI: http://dx.doi.org/10.2104/ag050011.

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

Pebesma, E., Wesseling, C., 1998. Gstat, a program for geostatistical modelling, prediction and simulation. Computers and Geosciences 24 (1), 17–31.

Pebesma, E. J., 2004. Multivariable geostatistics in S: the gstat package. Computers and Geosciences 30 (7), 683–691.

Pebesma, E. J., de Jong, K., Briggs, D., 2007. Interactive visualization of uncertain spatial and spatio-temporal data under different scenarios: an air quality example. International Journal of Geographical Information Science 21 (5), 515–527.
URL http://dx.doi.org/10.1080/13658810601064009

Petrakis, M., Hristopulos, D., submitted. On the joint probability density function of geometric anisotropy statistics for two dimensional differentiable random fields and a non-parametric test of statistical isotropy. IEEE Transactions on signal processing.

Schut, P., 2007. Opengis Web Processing Service. OGC document 05-007r7.
URL http://portal.opengeospatial.org/files/?artifact_id=24151

Schuurmans, J. M., Bierkens, M. F. P., Pebesma, E. J., Uijlenhoet, R., December 2007. Automatic prediction of high-resolution daily rainfall fields for multiple extents: The potential of operational radar. Journal of Hydrometeorology 8 (6), 1204–1224.
URL http://dx.doi.org/10.1175/2007JHM792.1

Skøien, J., Baume, O., Pebesma, E. J., Heuvelink, G., 2009. Identifying and removing heterogeneities between monitoring networks. Environmetrics in press.

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

W3C, 2007. Soap version 1.2 part 1: Messaging framework (second edition). http://www.w3.org/TR/soap12-part1/.

Williams, M., Cornford, D., Bastin, L., Pebesma, E., 2009. Uncertainty Markup Language (UncertML). OGC Discussion Paper, Document Number: 08-122r1.
URL http://portal.opengeospatial.org/files/?artifact_id=33234

Woolf, A., Shaon, A., 2009. An approach to encapsulation of grid processing within an ogc web processing service. Workshop on Grid Technologies for Geospatial Applications, AGILE 2009, Hannover.
URL http://portal.opengeospatial.org/files/?artifact_id=35975