1

The challenge of real-time automatic mapping for environmental monitoring network management

Edzer J. Pebesma[1], Gregoire Dubois[2], and Dan Cornford[3]

[1] Geosciences Faculty, Utrecht University, The Netherlands; e.pebesma@geo.uu.nl
[2] European Commission – DG Joint Research Centre, Institute for Environment and Sustainability, Ispra, Italy
[3] Neural Computing Research Group, Aston University, UK

Summary. The automatic interpolation of environmental monitoring network data such as air quality or radiation levels in real-time setting poses a number of practical and theoretical questions. Among the problems found are (i) dealing and communicating uncertainty of predictions, (ii) automatic (hyper)parameter estimation, (iii) monitoring network heterogeneity, (iv) dealing with outlying extremes, and (v) quality control. In this paper we discuss these issues, in light of the spatial interpolation comparison exercise held in 2004.

1.1 Introduction

Many environmental variables are monitored in a (semi-)continuous way; examples include air quality and background radiation levels. In order to utilize the network, maps of observed values are usually instantly available to network operators, but maps with interpolated values often need lengthy intervention by (spatial) statisticians before they become available. We believe that spatial interpolation can, and should, be automated to the extent that both in routine and emergency situations interpolated maps can become available in near real-time (i.e., within seconds up to tens of minutes) *without such intervention*. Of course there will always be a role for the spatial statistician in providing in depth analysis of a given data; our focus is on situations where decisions must be made quickly.

In a decision theoretic setting, a map with interpolated (predicted) values, is not sufficient information; knowledge of prediction errors and their probability distributions is necessary for optimal results. We explore some of the issues that the requirement for automatic, probabilistic, real-time, prediction raises.

This paper discusses issues in both algorithm development and their practical implementation in the form of a web service for operational monitoring

network management. It will review some of the submissions of the Spatial Interpolation Comparison (SIC) 2004 (Dubois and Galmarini, 2005; EUR, 2005). The issues we will address comprise

(i) quantifying and communicating exceedance probabilities for given threshold levels, in order to estimate risk of exposure.
(ii) the automated estimation of parameters describing the spatial variability in presence of extremes
(iii) dealing with heterogeneity of monitoring networks, e.g. across EU member state boundaries
(iv) detection of outliers in space and time
(v) quality control.

1.2 Communicating prediction error distributions

Interpolating in two dimensions can be relatively simple. In cases where the variogram is close to an exponential or spherical model, and the nugget variance is small, the inverse distance interpolation algorithm is hard to beat significantly with highly advanced geostatistical models, when the implementation is tuned to have a varying power in the distance weights, or a varying neighbourhood selection. One of the disadvantages of inverse distance methods is that they do not yield interpolation, or prediction errors when no variogram model is assumed. Interpolation errors can be large, and are of importance, if for example someone is faced with the decision whether an area, or how large an area should be evacuated based on the interpolation of measured radiation levels after a radioactive outbreak.

Ideally, an automatic prediction algorithm should provide a user with the full conditional cumulative distribution function (ccdf), which is for a random variable Z at arbitrarily chosen unobserved location s_0 the probability

$$F(Z(s_0), c) = \Pr(Z(s_0) < c \mid z(s_i), i = 1, ..., n) \tag{1.1}$$

with $Z(s_i), i \geq 1$ the observed data. Usually s_0 is chosen to be a large number of points (or square blocks) over a regular grid, and $F(Z(s_0), c)$, for a given level c, can be shown as a map. In risk studies, it may be more intuitive to map $1 - F(Z(s_0), c)$, which is the probability of exceeding c, but for the discussion here this is irrelevant. An alternative visualisation is that of the quantile function, obtained by inverting (1.1), which gives the Z values corresponding to a spatially constant given quantile value $q \in [0, 1]$:

$$F^{-1}(Z(s_0), q) = c \tag{1.2}$$

such as the median, or the 2.5 and 97.5 percentiles[4].

[4] Although not necessary for the discussion here, we want to note that in a considerable part of the geostatistical literature ccdf's are associated with, or discussed

For fixed, chosen values of c, the value of $F(Z(s), c)$, or alternatively one minus this value (the probability of exceeding c) may be shown as a static map. Usually but not necessarily, s is a collection of points on a regular grid covering the area studied. Accordingly, for fixed values of q a quantile map for $F^{-1}(Z(s), q)$ can be shown. Choosing these values ahead of time may be guided by regulatory guidelines, e.g. from maximum tolerated or established zero-risk concentrations, but threshold values found there often contain a certain or even considerable amount of arbitrariness in them, and a user may want to change them. An important issue to find out is how much a slight change in c results in a different assessment of the exceedence probability map.

Visual communication of the full functions $F(Z(s), c)$ and $F^{-1}(Z(s), q)$ is area of research. Pebesma et al. (accepted) describe a tool for the dynamic analysis of maps of (1.1) and (1.2) under different modelling or interpolation scenarios. In case of the ccdf (Figure 1.1) the value of c can be dynamically changed by dragging and dropping the vertical line in the ccdf widget, which is followed by immediate update of the corresponding maps of $F(Z(s), c)$. In case of the quantiles plot (Figure 1.2), the value of q (horizontal line in the ccdf widget) can be dynamically changed, to be followed by an update of the maps of $F^{-1}(Z(s), q)$.

1.3 An INTERPOLATE button, or web service?

Ideally, we would like to have a routine (let us say a button in a computer program or web client) which, given a set of measurements, provides near real-time maps of interpolated values, and / or their associated distribution or quantile views. This means that data have to be submitted, interpolated values computed, and ccdf's are returned (Figure 1.3).

While the implementation of the interpolation algorithm is clearly very important in determining the accuracy of the predictions, the usefulness of the system also depends on the ease of integration into the overall network management system. An exciting opportunity is presented by the adoption of a service oriented architecture (often called 'web services') with carefully defined interfaces offers an exciting prospect of developing an automatic interpolation service which any user capable of employing web services can utilise. This will necessitate the definition of standards for communicating uncertainty.

---

in the context of certain specific forms of kriging, notably indicator kriging and its descendents or generalisations. This is not necessary as ordinary, simple or universal kriging can provide ccdf's whenever a parametric distribution function (e.g. normal, lognormal, normal after Box-Cox transformation) is assumed. Such assumptions may be strong, but so are the assumptions about the identification of tail distributions and their spatial correlation in the indicator and related approaches.

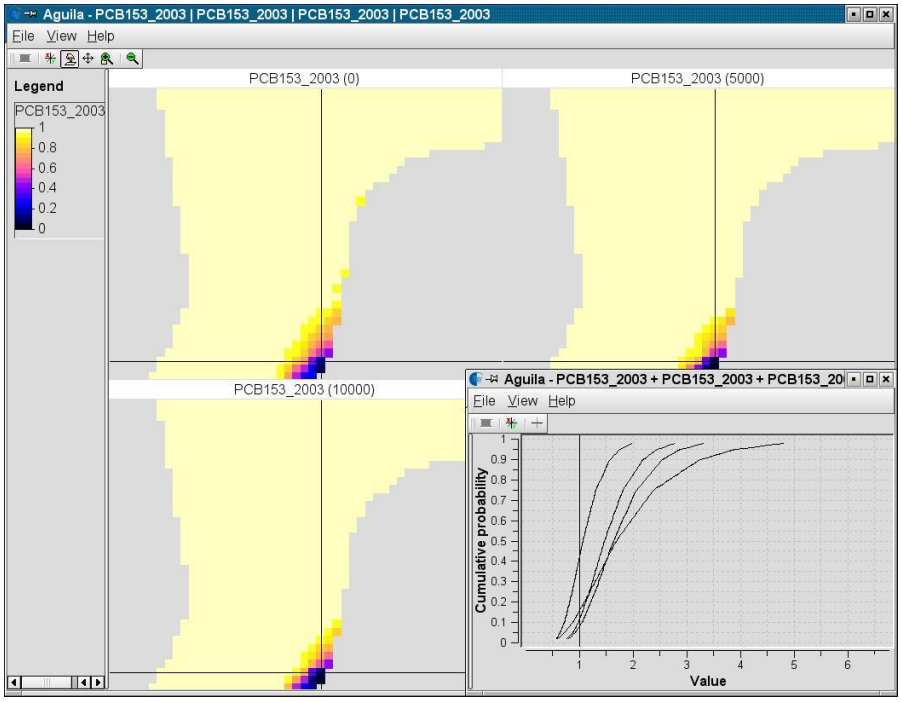

Fig. 1.1. Screen shot of a tool to visualize the distribution function $F(Z(s), c)$ for PCB-138 concentrations in North Sea floor sediment data, analyzed in Pebesma and Duin (2005). The maps show probabilities for the interpolated values of being below the PCB138 threshold value of 1 ppm (the legend caption misses this point). The scenarios refer to block size: (0) refers to point kriging, (5000) to kriging at 5 km × 5 km block, (10000) to kriging at 10 km × 10 km blocks.

1.4 Monitoring network heterogeneity

The idealized situation of Figure 1.3 discards much of the information that is usually available for monitoring networks. Besides the measurements themselves, the following, non-exhaustive list may be relevant for the interpolation routine:

- what do the measurements usually look like?
- are these measurements taken from a variable that can only take positive values (e.g. a concentration variable), and does it have an upper boundary?
- are the measurements obtained under identical conditions, or are there differences in measurement device, monitoring network design (e.g. between states or countries), standardization issues, or rules regarding the classification of a monitoring station? (e.g. is an air quality monitoring station classified as *rural* comparable to a likewise classified station in another country?)

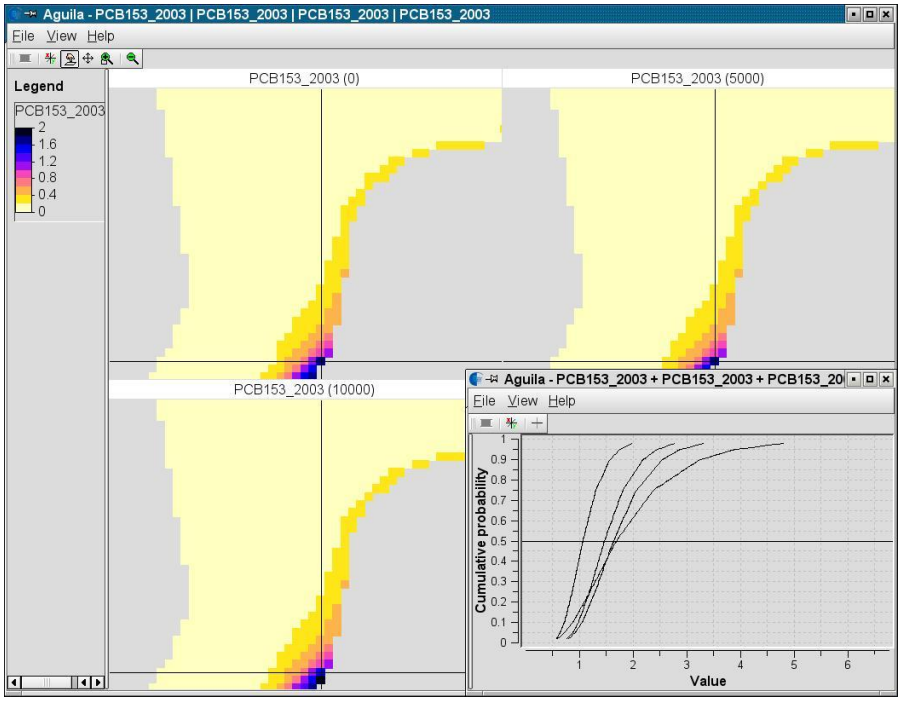

Fig. 1.2. Screen shot of a tool to visualize the quantile function $F^{-1}(Z(s), q)$ for PCB-138 concentrations in North Sea floor sediment data, analyzed in Pebesma and Duin (2005). The maps show quantiles for the interpolated values for the probability value 0.5 (i.e., the median).

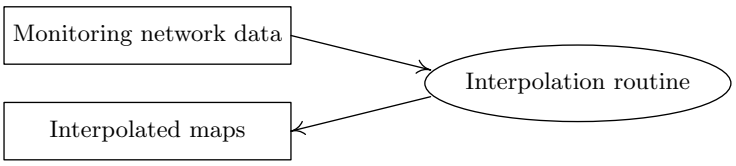

Fig. 1.3. Idealised data flow for an automatic mapping procedure. Implemented as a web service, the arrows may represent data flow over `http` connections

- are there variables available to which the monitored variable bears a relationship, that are useful for interpolation? (e.g. ozone may be related to altitude when looking a large region)
- is there any other prior knowledge available that needs to be taken into account in interpolation? (e.g. previous measurements, spatial correlation characteristics, prior beliefs)
- is there historic information that certain measurement stations behave anomalous, more often than others?

A naive interpolation procedure that does not take any of these issues into account may seem fairly easy to implement. When looking at interpolation as a stage in exploratory analysis of monitoring network data, such a naive interpolation procedure should be useful to detect some of the issues mentioned. As a dedicated system for decision support in emergency conditions, the requirements are different. Communicating any (or all) of the information to an interpolation procedure poses another interesting technical problem.

However, even if none of the above information is provided, an automatic interpolation should still be possible. The main issue is then (i) the modelling of the variogram (or covariance function), and (ii) the possible uncertainty about variogram model and/or model parameters. Given measurements and their locations, several issues require careful consideration. If we want to fit models to sample variograms, for example

1 how should we compute the sample variogram (maximum distance, lag interval width, directional or isotropic)?
2 which particular variogram model or group of models do we want to fit?
3 which criteria do we use for the actual fit?
4 which initial values do we provide for the fit, in case it involves non-linear parameters (such as range)?
5 how do we deal with the problem of an ill-fitting model or non-convergence in the fit?

Some of the above questions were discussed, but not typically "solved" by Pebesma (2005). When fitting by ML/REML, questions 2, 4 and 5 are relevant as well. In case of a Bayesian, so-called model-based approach (Diggle et al., 1998), two further questions are

6 which prior distributions should be chosen, automatically, for the variogram fitting procedure, and
7 how do we verify automatically that the Markov chain Monte Carlo algorithm has converged?

In the context of SIC2004, Palaseanu-Lovejoy (2005) has shown that this Bayesian procedure worked when the algorithm was applied to data that matched the prior assumptions, but failed in case of extreme, unexpected outlying data. Clearly further research is required to address these issues in the context of an automatic interpolation method.

1.5 Outliers in space and time

Outliers are of utmost importance, as they either need to be discarded as invalid measurements (monitoring network failure) or indicate extreme conditions, possibly notifying us of an emergency condition. An automatic interpolation routine should never automatically remove outliers in order to

remain useful for the second type of situation, but it is useful for network management, to provide a mechanism for deciding which case is true.

Interpolation in the presence of outliers (one or very few highly extreme values) is a major source of trouble for any interpolation procedure (e.g., EUR, 2005). One wonders whether single stationary random fields of whatever kind are useful as models for fields which really include outliers. We might also consider how one field (say, background concentrations) should be distinguished from the second (with outlying measurements) and in addition how the spatial correlation of the outlier field should be characterized on the basis of maybe one or two observations. Cornford (2005) suggested that in case of outliers that arise from real physical processes, we should work to probabilistic models that incorporate the physics of the phenomena modelled, using a data assimilation framework. This is a complex task and it remains to be seen whether one or two outlying observations are sufficient for successful assimilation of the outlying phenomena in absence other information on the magnitude and location of a source, but an integrated space-time analysis does seem indispensable for these cases.

1.6 Space-time approaches

One important issue for (near) real-time interpolation is whether past observations should be taken into account for the interpolation based on current observations. If measurements are taken with high frequency, this seems attractive because they may carry additional information when the process is temporally correlated. On the other hand, for certain processes sudden jumps in time (e.g. a radioactive outbreak) may not show up well in interpolated maps if these rely on the regular behaviour that nuclear radiation shows when there is no outbreak. For such emergency cases a space-time model should allow for sudden jumps in time. In any case, when interpolations are needed in near real-time, computation speed is an issue and this may currently be a challenge for space-time approaches, more than for spatial approaches alone.

1.7 Quality control and implementation issues

As in many other fields, software architectures in Geographic Information Systems are shifting from application oriented to service oriented paradigms. This means that algorithms are not implemented as a button in a stand-alone application, but rather operate as a web service facilitating their use from a client anywhere in the world. We envisage that interpolation is a service that can, and should be served this way. Among the motivations for this are (i) monitoring data are collected in real-time, but are not present in real-time on the client computer, but typically available after a service request, (ii) the network data may not be publicly available, but views on the data or

other derivative products may or may not be, or may be available for specific purposes, and (iii) the monitoring data may be served by a varying data base infrastructures and computer architectures.

How can we ensure that the code, or web service, does what it is supposed to do? At the base of software development, one should always build regression tests. Such tests provide input and verified output, such that in an automatic setting the code can be run to verify that it produces output identical to the verified output. As an example, package development in R (R development core team, 2006) stimulates package writers to supply their own tests, which are automatically run when porting packages to a new computing platform, or when R itself is upgraded. Developing regression tests for a wide variety of situations (not only success, but also failure situations) does harden the code, but is no guarantee for quality.

Another aspect is the use of legacy code. Software contains errors, or has undocumented features. Using code leads to errors being found and software that has been maintained for a long time may therefore be expected to contain fewer (unknown) bugs than freshly written code. Use of legacy code may also reflect the environment in which the code is written, e.g. low-level programming languages as C or Fortran, object-oriented languages such as C++ or java, or high-level environments such as R or Matlab. Code written in the latter environments may be easier to verify (by those who can read it), as it is 5-10 times as compact. The underlying numerical algebra is, at least for R, dealt with by legacy linear algebra libraries (lapack/linpack/blas). In addition, every aspect of R is open source, and as such fully verifiable by anyone.

The implementation as a web service facilitates the creation of a web testing client, which can subsequently be used to (automatically) test the performance of *any* other web service that implements the automatic mapping interface. This can give us more confidence in a new implementation since the automated regression tests will be extensive.

1.8 Lessons learnt from SIC2004

SIC2004 (Dubois and Galmarini, 2005; EUR 2005) was a spatial interpolation comparison, especially set up to test automated mapping routines, and to see how they performed in case of unexpected, rare extremes (a simulated local radioactive outbreak). Some lessons learned from this exercise are:

- SIC2004 used overall, average performance criteria. It did not take the probabilistic aspect of prediction (predictive distributions) into account, and did not evaluate as a performance criteria of the area above a certain cut-off value.
- In a comparison of automated mapping routines, one should never reach final conclusions based on comparison experience using a single data set only, and one should use criteria related to the emergency mapping context (Boogaart, 2004).

- All but one of the participants truly applied an automatic interpolation algorithm, meaning that manual intervention took place after discovery of the outliers and before submitting results (Myers, 2004).

Overall, the results of SIC2004 highlighted the need for further research before a truly automatic algorithm can be robustly deployed. The research issues include spatial statistics, algorithmic developments and software implementations, with their practical deployment requiring development of software architecture and standards for interoperability making this a truly interdisciplinary problem.

1.9 Discussion

Insurance companies know that knowledge about uncertainties pays off when taking decisions (setting insurance rates). However, they can spread risk because failure happens with some frequency. When taking decisions in environmental emergency conditions (such as treatment or evacuation of populations), the situation is totally different, because taking a wrong decision may worsen (or even cause) a disaster. This does not mean that we do not need the uncertainties, but rather that we (and the decision makers) have to learn how to use probabilistic information optimally.

When there is no direct emergency, the information about the distribution of prediction errors may also be of use for exposure assessment. As an example, it seems that black smoke has health effects for a considerable fraction of the population in parts of Europe and Northern America. Distribution functions obtained from spatial interpolation should be handled with care though; if a spatial interpolation algorithm suggests that in some region the probability of exceeding a critical level is 10%, this does not mean that 10% of the time the level is exceeded, nor that 10% of the population living there is affected. Despite that, interpolation, and analysing error distribution functions may help evaluating monitoring network management (e.g. monitoring network optimization), assess risk of exposure for populations and be instrumental to policy evaluation and development.

Automatic interpolation procedures seem to be far away now, but we expect them to become available, and envisage their use will be adopted by monitoring network management, risk assessments, and policy evaluation instruments.

Acknowledgments

This work was funded by the European Commission, under the Sixth Framework Programme, by the Contract N. 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.

References

Cornford, D. 2005. Are comparative studies a waste of time? SIC2004 examined. In: EUR, 2005. Automatic mapping algorithms for routine and emergency monitoring data. Editor: G. Dubois; Luxembourg: Office for Official Publications of the European Communities. EUR 21595 EN - Scientific and Technical Research Series; ISBN 92-894-9400-X

Diggle, P.J., J.A. Tawn, R.A. Moyeed, 1998. Model-based geostatistics. Applied Statistics 47(3), pp 299-350.

Dubois, G., S.Galmarini, 2005. Introduction to the spatial interpolation comparison (SIC) 2004 exercise and presentation of the datasets. Applied GIS 1 (2), p. 9-1 – 9-11.

EUR (2005). Automatic mapping algorithms for routine and emergency monitoring data. Editor: G. Dubois; Luxembourg: Office for Official Publications of the European Communities. EUR 21595 EN - Scientific and Technical Research Series; ISBN 92-894-9400-X

Myers, D.E. (2004) Spatial interpolation comparison exercise 2004: a real problem or an academic exercise. In: EUR, 2005. Automatic mapping algorithms for routine and emergency monitoring data. Editor: G. Dubois; Luxembourg: Office for Official Publications of the European Communities. EUR 21595 EN - Scientific and Technical Research Series; ISBN 92-894-9400-X

Palaseanu-Lovejoy, M. (2005). Bayesian Automating Fitting Functions for Spatial Predictions. Applied GIS 1 (2), pp 14.1–14.14.

Pebesma, E.J. (2005). Mapping radioactivity from monitoring data: automating the classical statistical approach. Applied GIS 1 (2), pp 11.1–11.17.

Pebesma, E.J., R.N.M. Duin (2005). Spatio-temporal mapping of sea floor sediment pollution in the North Sea. In: Ph. Renard, and R. Froidevaux, eds. Proceedings GeoENV 2004–Fifth European Conference on Geostatistics for Environmental Applications; Springer.

Pebesma, E.J., K. de Jong, D. Briggs, 2007. Interactive visualization of uncertain spatial and spatio-temporal data under different scenarios: an air quality example. Int.J. of GIS, in press.

R development core team, 2006. The R project. `http://www.r-project.org`

Van den Boogaart, K.G., 2005. The comparison of one-click mapping procedures for emergency. In: EUR, 2005. Automatic mapping algorithms for routine and emergency monitoring data. Editor: G. Dubois; Luxembourg: Office for Official Publications of the European Communities. EUR 21595 EN - Scientific and Technical Research Series; ISBN 92-894-9400-X