7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.
Edited by M. Caetano and M. Painho.

Dynamic visualisation of spatial and spatio-temporal probability distribution functions

Edzer J. Pebesma[1], Derek Karssenberg[1] and Kor de Jong[1]

1 Dept of Physical Geography, Geosciences Faculty
PO Box 80.115, 3508 TC Utrecht, The Netherlands
Tel.: +31 30 2533051; Fax: +31 30 2531145
e.pebesma@geo.uu.nl

Abstract

In this paper we will present and demonstrate aguila, a tool for interactive dynamic visual analysis of gridded data that come as spatial or spatio-temporal probability distribution functions. Probability distribution functions are analysed in their cumulative form, and we can choose to visualize exceedance probabilities given a threshold value, or its inverse, the quantile values. Threshold value or quantile level can be modified dynamically. In addition, classified probabilities in terms of (1-alpha)x100% (e.g. 95%) confidence or prediction intervals can be visualized for a given threshold value. Different modelling scenarios can be compared by organizing maps in a regular lattice, where individual maps (scenarios) are shown in panels that share a common legend and behave identically to actions like zooming, panning, and identifying (querying) cells. Variability over time is incorporated by showing sets of maps as animated movies. We will demonstrate this tool using sea floor sediment quality predictions under different spatial aggregation scenarios (block sizes), covering the Dutch part of the North Sea. The tool is freely available in binary and source code form; source code is distributed under the Gnu GPL; grid maps are read from disc through the GDAL library, or from memory as e.g. in an R session.

Keywords: dynamic graphics, interactive graphics, exploratory data analysis

1 Introduction

Visualising spatial accuracy is not trivial. One of the issues is that accuracy can refer to various characteristics of the data, e.g. to the spatial coordinates or to measured attributes, and that measured attributes may be of categorical or continuous nature. Without the intention to undervalue accuracy issues we do not address, in this paper we limit ourselves to visualizing the accuracy measures of measured attributes rather than spatial coordinates, and to continuous (or ordered categorical) rather than categorical data. Also, at this stage we concentrate on the visualisation of fields represented on lattices in space or space-time, rather than data collected on (irregular) points or more irregular polygonal areas.

Data of this type typically accrue when some modelling is done, e.g. by interpolating from irregularly spaced observation locations to a regular grid, or e.g. when a dynamic model is run to predict solute transport through a permeable soil body. Discretizing space with a regular grid is a very common choice; hydrological, meteorological, soil physical, geophysical, geostatistical and environmental models are among the models who may do this.

Accuracy statements may come in different flavours, but we will limit our attention to those that have the form of probability distribution functions. Such functions may result e.g. from a regression prediction or kriging interpolation where the prediction error is assumed to be normally distributed (possibly after e.g. log-transforming the variable), or they may result from

error propagation studies (Heuvelink, 1998) where it is approximated based on a hopefully large Monte Carlo sample of model outputs.

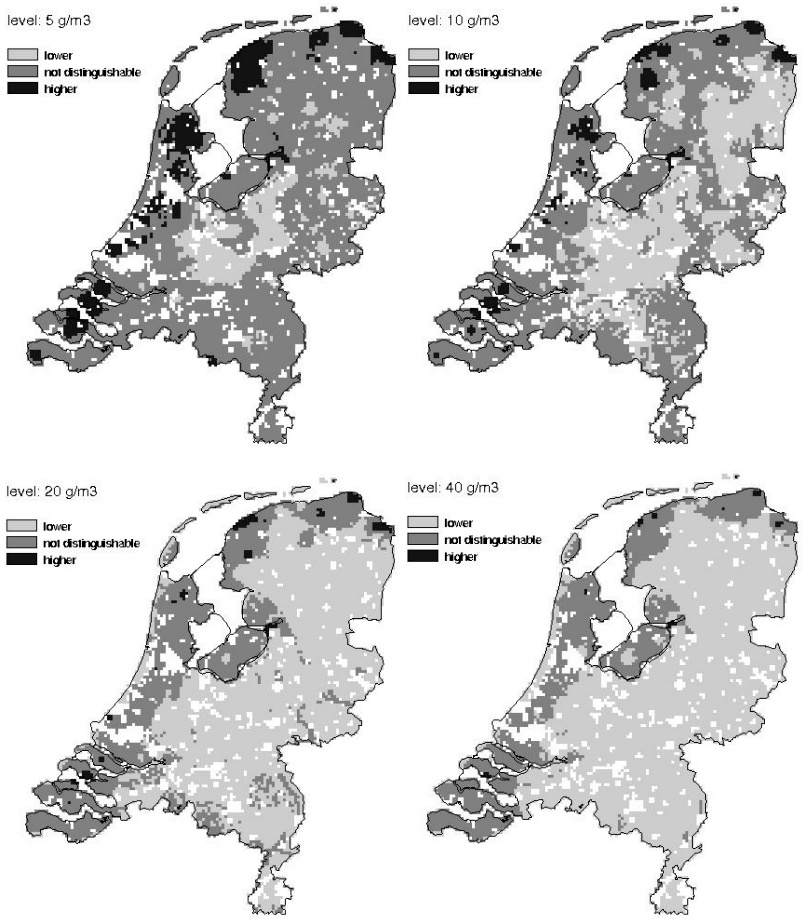

Figure 1 Maps for potassium (K) concentration in the groundwater between 5 and 15 meter depth in the Netherlands. Levels refer to classified approximate confidence intervals for 4 km x 4 km block median concentrations.

Pebesma and De Kwaadsteniet (1997) showed accuracy of spatial predictions for block median groundwater concentration variables by classifying approximate 95% prediction intervals with respect to four pre-specified (legislation related) threshold levels (Figure 1). If the complete interval was below the level, the cell was classified as *lower* than the level, if the interval was above the level the cell was classified as *higher*, if the interval straddles the level the cell was classified as *not distinguishable* from the level, given available information (monitoring network measurements, soil and land use maps to stratify). The current, dynamic visualisation, approach to analyse spatial probability distribution functions generalizes this by providing in addition to the static approach of Pebesma and De Kwaadsteniet the ability to (i) display

quantiles, such as the median or 2.5-percentile, dynamically adjusting the quantile level, (ii) display (one minus) exceedance probabilities, dynamically adjusting the threshold level, and (iii) display classified prediction intervals, dynamically adjusting the threshold level, or adjusting the confidence level (1 - α). The reason for providing a dynamic solution instead of a static one is that the function we want to communicate is high-dimensional; choosing static, fixed cross-sections through it limits the user the possibility to explore the complete information.

2 Cumulative probabilities and quantiles in space and time

For each point in space and time we can define a cumative distribution function, *P*, which gives for the random variable *Z* the probability of being below a given threshold *c*:

$$P_Z(c) = \Pr(Z < c)$$

Given that Z is a spatio-temporal field with spatial index s and temporal index t, then we can write

$$P_{Z(s,t)}(c) = \Pr(Z(s,t) < c)$$

We can think of cumulative probabilities as "one minus exceedance probabilities".

When we invert this function, we obtain the quantile function $q_{Z(s,t)}(P) = P^{-1}{}_{Z(s,t)}(c)$ of Z; e.g. the 0.5 quantile q(0.5) is the the value c for which $P_{Z(s,t)} = 0.5$ (i.e., the median). When space is two-dimensional, we now have two four-dimensional functions to analyze: either P as function of s, t and c, or q as a function of s, t and P. Note that these probability distribution functions are marginal functions, i.e. joint probabilities $\Pr(Z(s_1,t_1) < c_1, Z(s_2,t_2) < c_2)$ are in general not available from the marginal distributions only.

3 Scenarios

Adding somewhat to the complexity, in many cases when looking at spatial probability distribution functions we may want to compare several functions side by side, resulting from different scenarios. Scenarios may refer to future scenarios resulting from different choices for boundary conditions, but equally well to modelling scenarios resulting from different model choices. The latter may for example include functions resulting from different choices in interpolation model (predictors selected, variogram model used) or aggregation level (point support prediction, different block sizes). In a more general Monte Carlo study scenarios may refer to optimistic, pessimistic and intermediate levels of uncertainty in initial or boundary conditions or model parameters, or other varying choices for probability distributions chosen for them. Scenario is clearly a categorical variable, adding scenario index *k* to the distribution function: $P_{Z(s,t,k)}(c) = \Pr(Z(s,t,k) < c)$, and adding another dimension to the functions to analyze.

4 Example implementation: aguila

We developed a tool, called *aguila* (latin for eagle), for interactive display and analysis of spatio-temporal probability distribution functions, under different scenarios. The tool is written in C++, is cross platform (it uses the Qt widget set) and can read map information either

through the gdal library (http://www.gdal.org/; this library supports over 40 different grid map formats) or directly from memory (e.g. from within an R session). The source code of *aguila* is available under the GPL.

The tool will be demonstrated during the oral presentation of this paper. As a static paper, we can here only show a few screen shots; the interaction has to be imagined, and the interested reader can download the application from the internet (through http://pcraster.geo.uu.nl).

The information fed into aguila is a set of raster maps which have for each scenario and for each time step and for a sequence of quantiles (*P* values, e.g. 0.01, 0.05, 0.25, 0.5, 0.75, 0.95, 0.99) the corresponding *Z* value. Suppose we have 10 time steps, 4 scenarios and 7 quantiles, then we need 10 x 4 x 7 = 280 maps. Given that we work in an environment where it is easy to compute quantiles from samples or probability distributions (for example R), then generating such a set of maps is fairly trivial. The basic view aguila shows is the median (or 0.5 quantile) for each scenario (e.g., Figure 2). We can zoom and pan in any scenario window, or change legend properties, after which the modified view is adjusted for each of the scenarios.

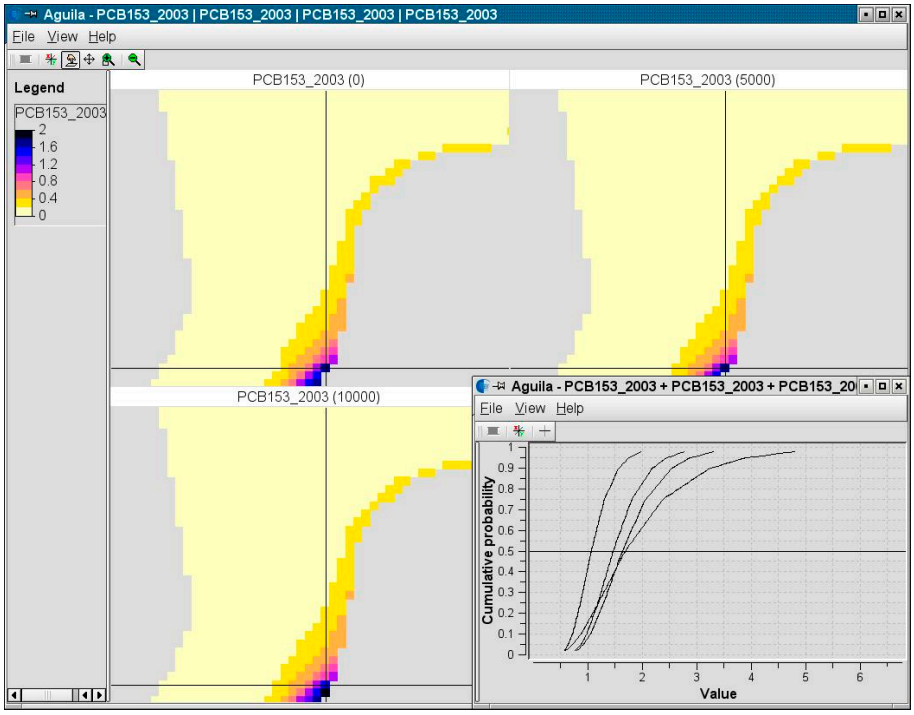

Figure 2 the quantile view.

The variable shown in Figure 2 is PCB-153, measured in µg/kg in the sediment fraction smaller than 63µm, interpolated for 2003 over the sea floor of the Dutch part of the North Sea; the missing values area on the right of the sub-maps reflects the Dutch North Sea shore, grid cell size is 5 km x 5 km; see Pebesma and Duin (2005) for more details. The four scenarios

(only 1-3 are shown) refer to four point kriging and 3 different block sizes (5000x5000 m, 10000x10000m are shown) for block kriging predictions. Probability distribution functions were approximated by assuming that on the log-scale, where linear kriging interpolation took place, the kriging prediction error is normally distributed. Another example, using air quality data over Europe, is shown in Pebesma et al., (accepted).

In a separate window, the probability distribution curves under the cursor (cross) are shown. The curves are formed by straight lines drawn between the pairs (P, Z) provided in the map. The horizontal line is at the median. Clicking and dragging this line has the effect that the maps for each scenario continuously update to the quantile at which the line is. This way, we can for a set of scenarios compare each of the quantiles given; quantiles not provided as maps are interpolated linearly between the nearest two quantiles; no quantiles beyond the extremes can be queried. At time of writing this paper (but hopefully not at time of presentation), the option to easily connect a specific curve to its corresponding scenario is missing.

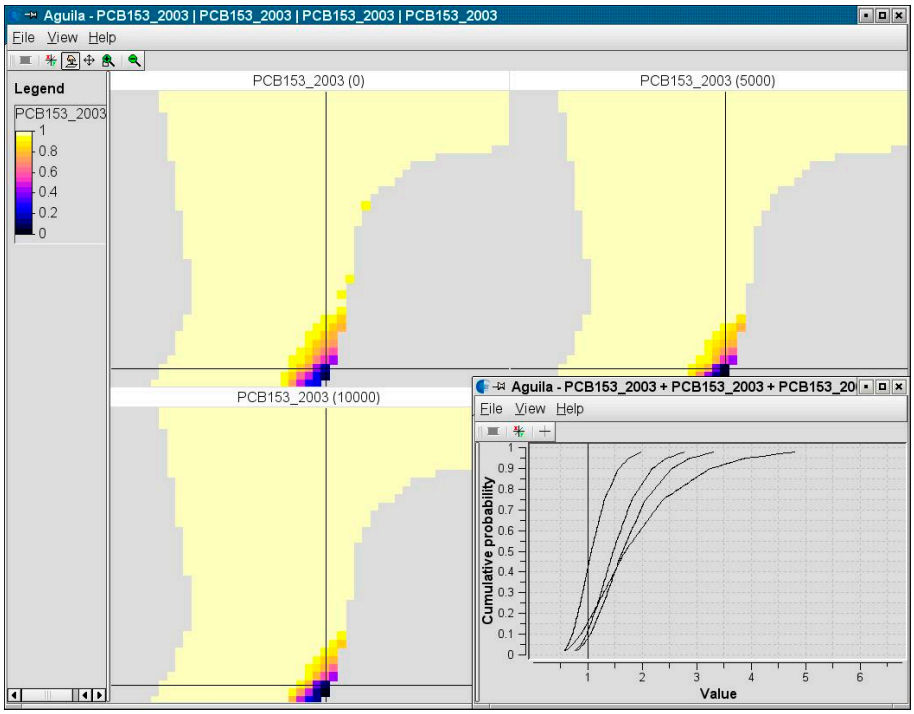

Figure 3 the probability view.

The complementary view of this window is obtained when the cross hair tool button is pushed in the probability distribution curve widget (the one with the +). In this case, shown in Figure 3, the vertical line indicates a given attribute value, say c, and the maps show the corresponding $P(Z(s,t,k) < c)$.

Interactively modifying *c* updates the maps for all scenarios to reflect the corresponding *P* value. The legend scale now also changes from *Z* values to *P* values, ranging from 0 to 1. The values shown can be interpreted as one minus the exceedance probability; some work still needs to be done to flip the legend to actually show exceedance probabilities as an alternative to probability distribution values.

A last view is derived from the probability view: the classified confidence interval view (which resembles Figure 1). One could argue that showing the wealth of probability values in Figure 2 is overkill, and that only (i) values larger than e.g. (1-α)/2, (ii) values smaller than α/2 and (iii) remaining values, suffices. This is equal to the classification in Figure 1, where the maps are shown for α = 0.05, for four threshold levels (*c* = 5, 10, 20, 40) and the classes were named (i) lower, (ii) higher, and (iii) not distinguishable. In aguila we can choose the value for α, but not (yet) dynamically modify it. We can still click and drag the threshold value *c*, after which the classified cells are continuously updated for this new value.

For each of these views, variability over time can be shown by animating a movie with values varying over time.

5 Comparison to other visualisation techniques

The proposed visualisation is a huge extension in terms of information content compared to the simple approach taken by Pebesma and De Kwaadsteniet (1997). It has the following additions: (i) we can now dynamically modify the cutoff level *c*, and choose any value, (ii) we can adjust the level of α, (iii) in addition we can show maps of the distribution values *P* for any value *c*, and its inverse, the quantile (*Z*) values corresponding to given cumultaive probability levels *P*. This, of course, comes at the cost that we cannot print the application in hardcopy form without loosing many options (or without using lots of paper).

Other approaches to visualize attribute accuracy are:

- Showing animated movies of simulations; our experience is that from such animations it is hard to infer quantitatively where the uncertainties are large and where they are small, or how the particular marginal distributions look like, and vary, spatially and temporally. Also, transitions between them may be very sudden, a problem discusses by Ehlschlager et al (1997).
- Using a two-dimensional legend (e.g. Hengl et al., 2004) where colour or hue expresses attribute estimate, and gray-ness or whiteness expresses prediction error; we think that this approach fails to give a quantitative assessment of the spatio(-temporal) probability distribution functions. These maps still separate estimate from estimation error. Kriging standard error maps are as old as the kriging technique, but have allways failed to provide information with operational value. The approach taken by Hengl et al. (2004) in addition needs arbitrarily thresholds to translate uncertainty into whiteness.

6 Discussion

A major strength of the application demonstrated here is that it allows us to infer exceedence probabilities for cutoff levels without being completely bound to a pre-determined set of cutoff levels that have inevitably been chosen with a certain degree of arbitrariness. The obvious question raised with a map that shows exceedance probabilities for the level 40 is: “so what, and what would this map look like if the level were 45?”, and can, using this tool, be answered

instantly, along with any other similar questions. In addition, we can now easily look at the consequences of shifting focus from e.g. the median to the 90-th percentile when developing or setting environmental regulation standards (such as in the OSPAR commission, http://www.ospar.org/).

We believe that probability distribution functions provide a more complete overview of the extent of our knowledge, or the accuracy, of attribute data than estimated values alone, estimation standard errors, or estimates and their standard errors. In the approach taken here, spatial and spatio-temporal random functions with *any* distribution function can be visualized, under different scenarios.

Of course we focus on the marginal probability distribution functions rather than the joint distributions, because visualizing the latter as static views is impossible. Joint distributions need to be taken into consideration when spatial and/or temporal interaction (e.g. spatial aggregation, such as averaging over an area) is considered; the variables obtained *after* this process can then be analyzed as described here.

Acknowledgements

The RIKZ, in particular Richard Duin, is acknowledged for using the North Sea floor sediment quality data in the illustration. Part of the development costs for *aguila* were obtained from the EU project Apmosphere (http://www.apmosphere.org/, EVK2-2002-00577), one of the preliminary GMES projects under FP5.

References

Ehlschlager, C.R., A.M. Shortridge, M.F. Goodchild, 1997. Visualizing spatial data uncertainty using animation. Computers & Geosciences 23(4): 387—395.

Hengl, T., G.B.M. Heuvelink, A. Stein, 2004. A generic framework for spatial prediction of soil variables based on regression kriging. Geoderma 122 (1-2): 75—93.

Heuvelink, G.B.M., 1998. Error propagation in environmental modelling with GIS. Taylor & Francis.

Pebesma, E.J., J.W. de Kwaadsteniet, 1997. Mapping groundwater quality in the Netherlands. Journal of Hydrology 200, 364—386.

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 (accepted). Interactive visualization of uncertain spatial and spatio-temporal data under different scenarios: an air quality example. Int.J. of GIS, accepted for publication.