

Visualizing uncertainty in spatial and spatiotemporal field data

Edzer Pebesma

uncert|web
uncertainty-enabled model web



ifgi
Institute for Geoinformatics
University of Münster

52north
exploring horizons

Perspectives on the visualization of uncertainty, ITC Sep 15, 2014

Visually-Supported Reasoning with Uncertainty Workshop, GIScience, Sep 23, 2014

Overview of the talk

1. What is field data?
2. What is uncertainty?
(and what is ambiguity?)
3. Where does uncertainty come from?
sampling – modelling – error propagation
4. Probability and dimensionality
5. Visualisation approaches
static – dynamic – web
6. Encoding ST probability distributions
UncertML – NetCDF-U – FieldGML – R

What is field data?

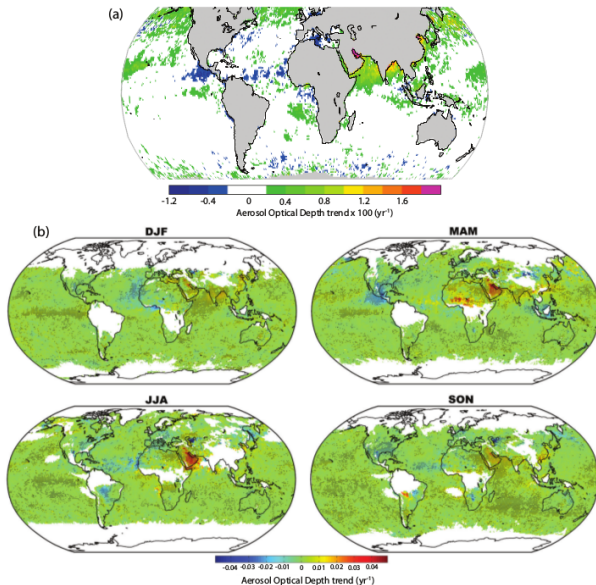


Figure 2.9 | (a) Annual average aerosol optical depth (AOD) trends at 0.55 μm for 2000–2009, based on de-seasonalized, conservatively cloud-screened MODIS aerosol data over oceans (Zhang and Reid, 2010). Negative AOD trends off Mexico are due to enhanced volcanic activity at the beginning of the record. Most non-zero trends are significant (i.e., a trend of zero lies outside the 95% confidence interval). (b) Seasonal average AOD trends at 0.55 μm for 1998–2010 using SeaWiFS data (Hsu et al., 2012). White areas indicate incomplete or missing data. Black dots indicate significant trends (i.e., a trend of zero lies outside the 95% confidence interval).

Fields are *functions* of space or space and time

$$z = f(s, t)$$

with

- ▶ s the location vector (e.g. long, lat in WGS84)
- ▶ t the time (e.g., seconds since 01-01-1970, 00:00 UTC)
- ▶ z , or $f(s, t)$ is the (observed or modelled) variable

Fields are *functions* of space or space and time

$$z = f(s, t)$$

with

- ▶ s the location vector (e.g. long, lat in WGS84)
- ▶ t the time (e.g., seconds since 01-01-1970, 00:00 UTC)
- ▶ z , or $f(s, t)$ is the (observed or modelled) variable

In this case: $f(s)$, time is not “visible”

- ▶ s locations on some grid (e.g. long, lat in WGS84)
- ▶ t time (caption): 2000-2009, aggregated out by computing trend
- ▶ $z(s, t)$ the annual average aerosol optical depth trend, computed over this time period

Functions, a reminder

We can write $z = f(s, t)$ or alternatively, $s \times t \rightarrow z$, or, by currying, $s \rightarrow t \rightarrow z$, with:

- ▶ s, t the *domain* of the function
- ▶ z the *value* of the function for a particular (s, t) pair
- ▶ for each pair (tuple) (s', t') we have *one and only one* value z
- ▶ the reverse is not true, but function inversion results in sets (sets of regions \times periods)
- ▶ if for $z(s)$, s is two-dimensional and regularly discretized, we have *raster data*

What is uncertainty?

Uncertainty arises when we are not *certain* about something. E.g., we are uncertain whether the statement

it will rain¹ tomorrow in this city

is true. We can express our uncertainty in terms of probabilities.

¹minimal 1.0 mm rainfall over at least 75% of the administrative boundaries of the city

What is uncertainty?

Uncertainty arises when we are not *certain* about something. E.g., we are uncertain whether the statement

it will rain¹ tomorrow in this city

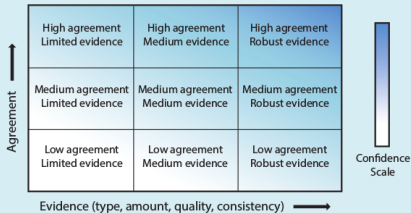
is true. We can express our uncertainty in terms of probabilities.
Probabilities

- ▶ range between 0 and 1, inclusive
- ▶ sum to 1, over all possible outcomes (e.g., “yes”, “no”)
- ▶ may be subjective, personal
- ▶ are often based on experiences (relative frequencies)
- ▶ are a way to make money for gamblers, traders, and insurance companies

¹minimal 1.0 mm rainfall over at least 75% of the administrative boundaries of the city

IPCC, AR5, Tech. Summary: *Treatment of Uncertainty*

The following summary terms are used to describe the available evidence: limited, medium, or robust; and for the degree of agreement: low, medium, or high. A level of confidence is expressed using five qualifiers very low, low, medium, high, and very high, and typeset in italics, e.g., *medium confidence*. Box TS.1, Figure 1 depicts summary statements for evidence and agreement and their relationship to confidence. There is flexibility in this relationship; for a given evidence and agreement statement, different confidence levels can be assigned, but increasing levels of evidence and degrees of agreement correlate with increasing confidence.



Box TS.1, Figure 1 | A depiction of evidence and agreement statements and their relationship to confidence. Confidence increases toward the top right corner as suggested by the increasing strength of shading. Generally, evidence is most robust when there are multiple, consistent independent lines of high quality. (Figure 1.11)

The following terms have been used to indicate the assessed likelihood, and typeset in italics:

Term*	Likelihood of the outcome
<i>Virtually certain</i>	99–100% probability
<i>Very likely</i>	90–100% probability
<i>Likely</i>	66–100% probability
<i>About as likely as not</i>	33–66% probability
<i>Unlikely</i>	0–33% probability
<i>Very unlikely</i>	0–10% probability
<i>Exceptionally unlikely</i>	0–1% probability

* Additional terms (*extremely likely*: 95–100% probability, *more likely than not*: >50–100% probability, and *extremely unlikely*: 0–5% probability) may also be used when appropriate.

... then what is ambiguity?)

Ambiguity refers to different opinions, not uncertainty. E.g.,

is the color of my eyes green?

Possible answers: yes, no, somewhat, mwah, kind-a-blueish-green. The issue under discussion is one that has no *crisp boundaries*, unlike *rain* as it was defined on the previous slide. To deal with the variety in answers, we can come up with class membership, the degree to which you believe a statement is true, or fuzzy numbers.

... then what is ambiguity?)

Ambiguity refers to different opinions, not uncertainty. E.g.,

is the color of my eyes green?

Possible answers: yes, no, somewhat, mwah, kind-a-blueish-green. The issue under discussion is one that has no *crisp boundaries*, unlike *rain* as it was defined on the previous slide. To deal with the variety in answers, we can come up with class membership, the degree to which you believe a statement is true, or fuzzy numbers. Why are fuzzy numbers **not useful for representing uncertainty**?

1. proper theory is lacking (they do not add up to 1)
2. they cannot deal with conditional probabilities, and hence dependence

Where does uncertainty come from?

sampling observation is incomplete, the function $z(s, t)$ is only observed at particular locations and times (gaps in time series, sparse soil samples, satellites always circling around); estimating aggregated values (e.g. global means) or interpolated values is due to *sampling error*.

modelling observing the wrong variable: we observe color, but want land use; we observe altitude, but want temperature; we observe GPS points, but want speed and direction. Models are (i) wrong, as they simplify, and (ii) usually have parameters that need to be estimated. (essentially, this is sampling error too)

error propagation we do forward modelling, say a rainfall-runoff model is formulated as $r = f(z_1, z_2, z_3, p_1, p_2)$ and z_1 , z_2 and p_2 are subject to uncertainty. Solution: propagate errors (e.g., by Monte Carlo simulation).

The stochastic dimension in a dynamic GIS

Edzer J. Pebesma, Derek Karssenbergh and Kor de Jong

Utrecht Centre for Environment and Landscape Dynamics, Faculty of
Geographical Sciences, Universiteit Utrecht, P.O. Box 80.115, 3508 TC
Utrecht; e.pebesma@geog.uu.nl

Abstract. Coping with random fields in a time-dynamic geographic information system (GIS) increases the computational burden and storage requirements with a large amount, and calls for a number of custom functions to enable easy analysis of the resulting random components, as well as specialised output reporting functions. This paper addresses the computational and implementation issues when a Monte Carlo approach is taken, and shows some results from a rainfall-runoff model running within a GIS.

Keywords. Geographical information systems, Monte Carlo, temporal GIS, stochastic modelling, geostatistics

1 Introduction

Geographical information systems (GIS, Burrough and McDonnel, 1998) liberate the end user from worrying about looping over all spatial entities by

Probability and dimensionality (COMPSTAT 2000)

The COMPSTAT 2000 paper argued that for fields, when z is not “known” but treated as a random variable Z , every scalar $z(s, t)$ needs to be replaced by a probability distribution function

$$F(z)(s, t) = \Pr(Z < z)(s, t)$$

which adds one dimension. But note:

Probability and dimensionality (COMPSTAT 2000)

The COMPSTAT 2000 paper argued that for fields, when z is not “known” but treated as a random variable Z , every scalar $z(s, t)$ needs to be replaced by a probability distribution function

$$F(z)(s, t) = \Pr(Z < z)(s, t)$$

which adds one dimension. But note:

- ▶ this only captures the *marginal* distribution, and not the joint probability

$$F(z_1, \dots, z_n)(s, t) = \Pr(Z_1 < z_1, \dots, Z_n < z_n)(s, t)$$

- ▶ for the latter, n grid nodes would need an n -dimensional distribution function...

The stochastic dimension in a dynamic GIS

Edzer J. Pebesma, Derek Karssen and Kor de Jong
Utrecht Centre for Environment and Landscape Dynamics, Faculty of
Geographical Sciences, Universiteit Utrecht, P.O. Box 80.115, 3508 TC
Utrecht • pebesma@geog.uu.nl

Abstract. Coping with random fields in a time-dynamic geographic information system (GIS) increases the computational burden and storage require-

Visualisation approaches

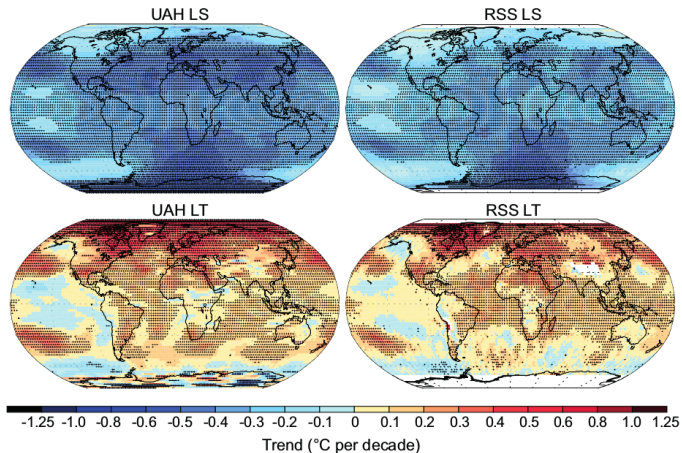
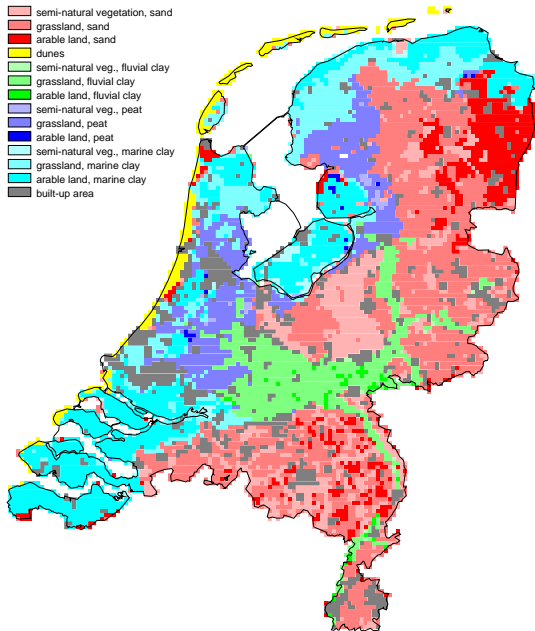


Figure 2.25 | Trends in MSU upper air temperature over 1979–2012 from UAH (left-hand panels) and RSS (right-hand panels) and for LS (top row) and LT (bottom row). Data are temporally complete within the sampled domains for each data set. White areas indicate incomplete or missing data. Black plus signs (+) indicate grid boxes where trends are significant (i.e., a trend of zero lies outside the 90% confidence interval).

20 years ago...

- semi-natural vegetation, sand
- grassland, sand
- arable land, sand
- dunes
- semi-natural veg., fluvial clay
- grassland, fluvial clay
- arable land, fluvial clay
- semi-natural veg., peat
- grassland, peat
- arable land, peat
- semi-natural veg., marine clay
- grassland, marine clay
- arable land, marine clay
- built-up area



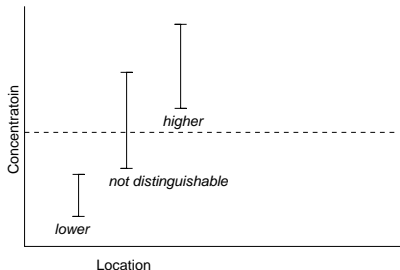
Groundwater wells (10-25 m; yearly)



Maps with 95% confidence intervals

Usually, decision (smaller/larger) is done on the basis of (e.g. 95%) confidence interval:

- ▶ c.i. above threshold: larger
- ▶ c.i. below threshold: smaller
- ▶ else: undecided ("*not distinguishable*")

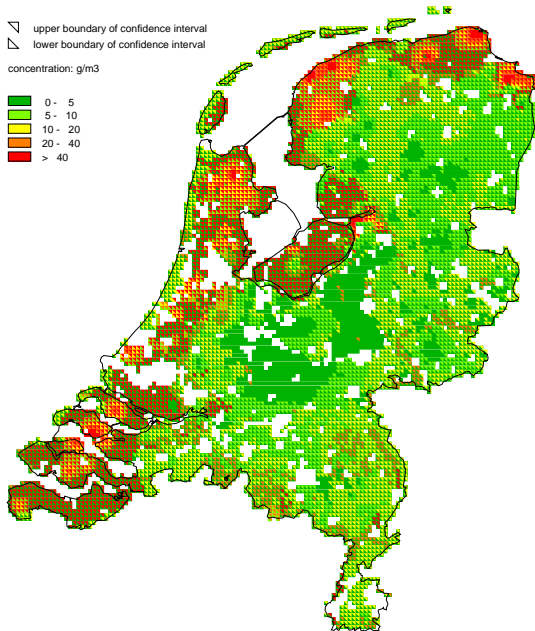


K, groundwater, 5-17 m depth

▽ upper boundary of confidence interval

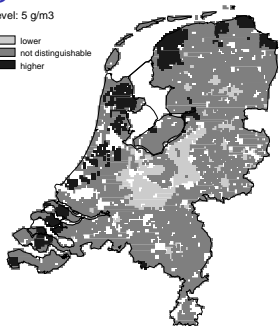
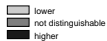
▽ lower boundary of confidence interval

concentration: g/m³

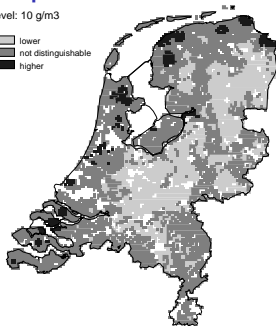
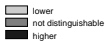


K, groundwater, 5-17 m depth

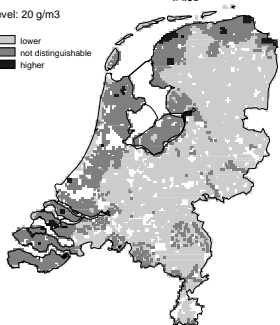
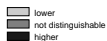
level: 5 g/m³



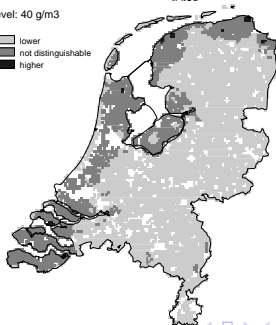
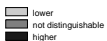
level: 10 g/m³



level: 20 g/m³



level: 40 g/m³



International Journal of Geographical Information Science
Vol. 21, No. 5, May 2007, 515–527



Interactive visualization of uncertain spatial and spatio-temporal data under different scenarios: an air quality example

EDZER J. PEBESMA[†], KOR DE JONG[†] and DAVID BRIGGS[‡]

[†]Faculty of Geosciences, Utrecht University, The Netherlands

[‡]Imperial College, London, UK

(Received 23 November 2005; in final form 13 October 2006)

This paper introduces a method for visually exploring spatio-temporal data or predictions that come as probability density functions, e.g. output of statistical models or Monte Carlo simulations, under different scenarios. For a given moment in time, we can explore the probability dimension by looking at maps with cumulative or exceedance probability while varying the attribute level that is

Probability distribution curves

the curve the less uncertainty the data have; certainty about the concentration value would result in a step from 0 to 1 at that particular value;

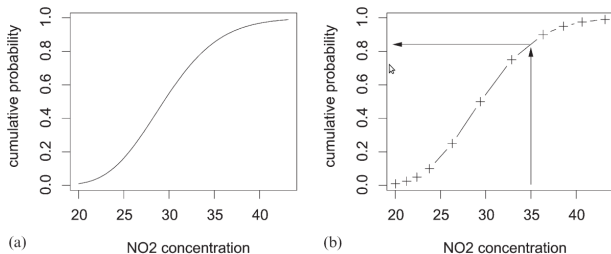
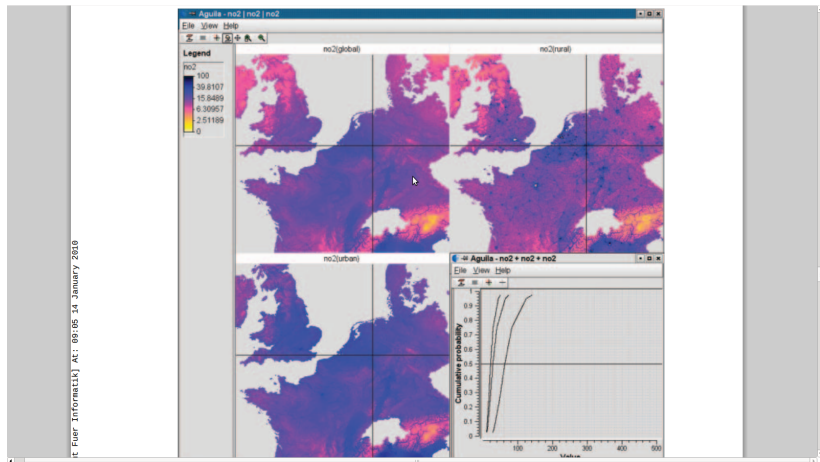


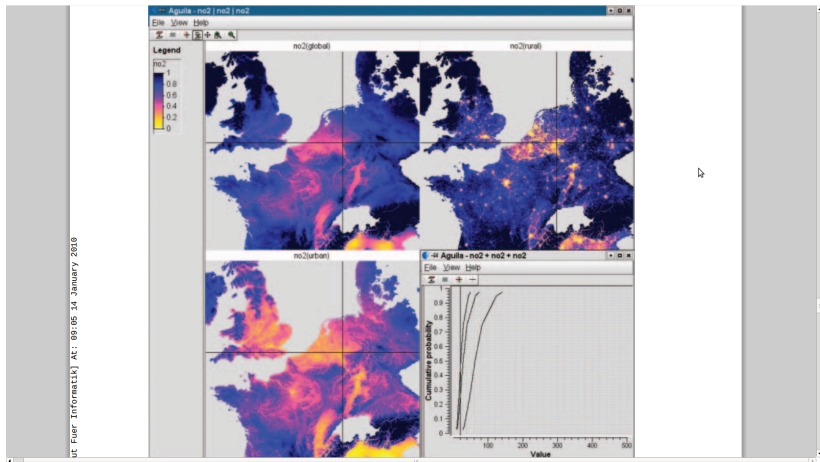
Figure 2. (a) cumulative probability plot for a given location s_0 and given scenario; values are obtained by assuming a normal distribution on the log-scale, and are drawn from P values ranging from 0.01 to 0.99; (b) deriving cumulative probability P from a discretized representation of the CDF, using linear interpolation.

Probability distribution curves

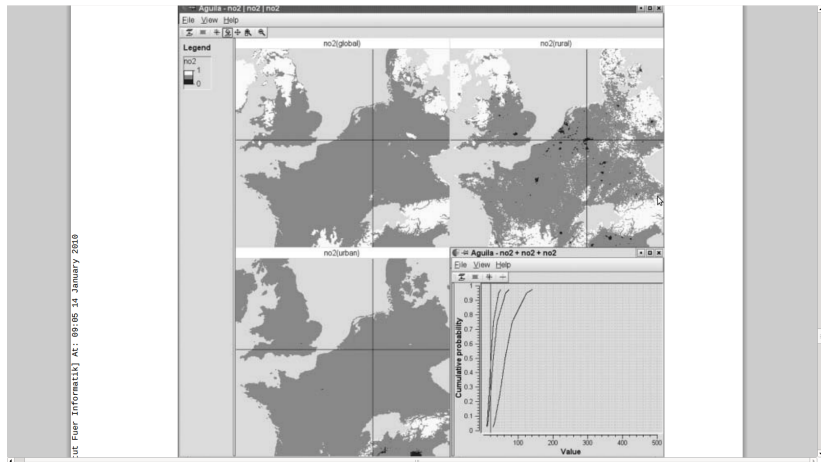


IT-Fluer_Informatik | AT: 09:05:14 January 2010

Probability distribution curves



Probability distribution curves



Usability of Spatio-Temporal Uncertainty Visualisation Methods

Hansi Senaratne^{1,2}, Lydia Gerharz¹, Edzer Pebesma^{1,2}, Angela Schwering¹

¹ Institute for Geoinformatics, University of Muenster, Germany
<http://www.ifgi.de>

² 52°North Initiative for Geospatial Open Source Software, Germany
<http://www.52north.org>

Abstract

The presented work helps users of spatio-temporal uncertainty visualisation methods to select suitable methods according to their data and requirements. For this purpose, an extensive web-based survey has been carried out to assess the usability of selected methods for users in different domains, such as GIS and spatial statistics. The results of the survey are used to incorporate a *usability* parameter in a categorisation design to characterise the uncertainty visualisation methods. This enables users to determine the uncertainty visualisation method(s) that are most suitable according to their domain of expertise. Finally, the categorisation design has been implemented and incorporated in a web-based tool as the *Uncertainty Visualisation Selector*. This web application can automatically recommend suitable uncertainty visualisation method(s) from user and data requirements.

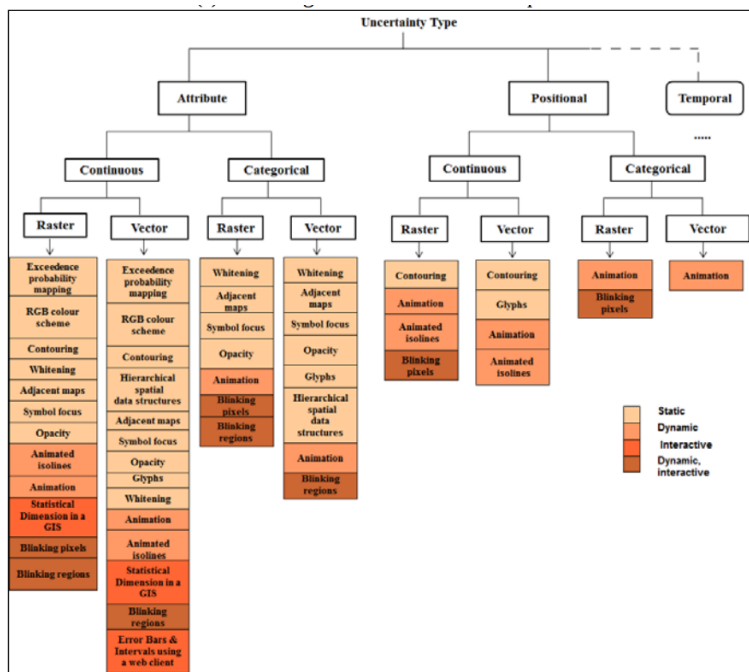


Figure 1. Categorisation of selected uncertainty visualisation methods (Senarathne & Gerharz 2011)

3.2.3 Symbols Visualisation

The Symbols method (Pang 2001) expresses the figurative similarities of objects based on shape or colour (Bertin 1983). Assigning colours to symbols was done with much caution as it needs to convey a realistic meaning to the users such that they can relate to it. Here, the uncertainties in the land use data set of Asia were depicted using circular symbols as seen in the foreground of Figure 4. Different land use classes over Asia are displayed in the background. The increasing uncertainty was shown by symbols of increasing size and varying colour. The colours green, yellow, orange and red were used in order, to represent increasing uncertainties, red communicating highest uncertainties.

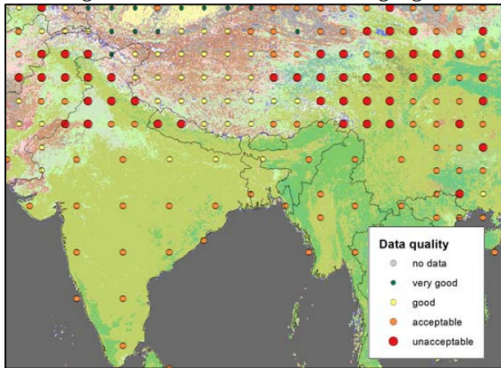
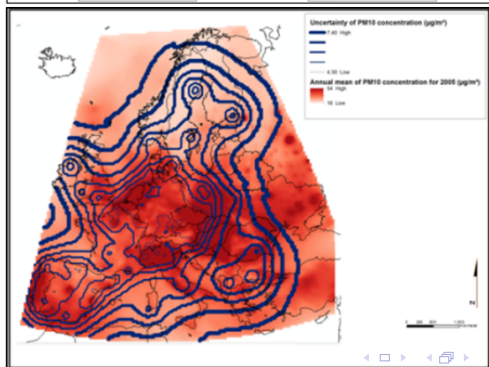
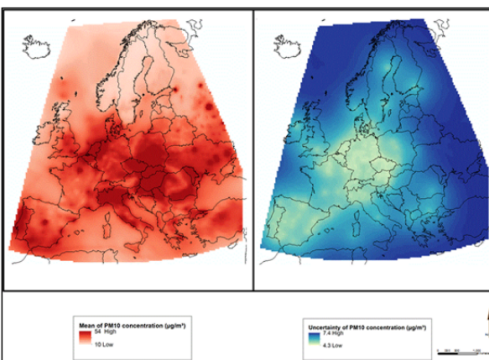


Figure 4. Uncertainty of land use classes represented through symbols of varying colour and size



UncertWeb 2010-2013: vis client

As *aguila* was written in C++, using Qt, it looked odd to use it as end point of a web-based model chain. So, we rewrote from scratch, in javascript:

1. starting from openlayers, adding multi-panel support
2. added reading resources from the web
3. supported the UncertWeb profiles (vector: O&M-U, raster: NetCDF-U)
4. -U implies UncertML: also parametric distributions, exceedance probabilities are understood
5. rendering of raster *data* uses a server component, vis-server (light-weight WMS)

<http://giv-uw.uni-muenster.de/vis/v2/>

greenland

GeoViQua/52North continued development of the *vis client*:


1. “branded” it to GREENLAND (google: greenland 52North)
2. development taken over by 52North (a spin-off of ifgi)
3. added GeoViQua requirements: support for ncWMS, WMS-Q etc
4. extended 2 linked windows to n linked windows
5. added use of glyphs (Jon Blower’s presentations)
6. added *persistent links*
7. added a web site with VERY NICE examples which you can click on to run

<https://wiki.52north.org/bin/view/Geostatistics/Greenland>

Encoding ST probability distributions

- ▶ UncertML.org: UncertML is a conceptual model and XML encoding designed for encapsulating probabilistic uncertainties
- ▶ NetCDF uncertainty conventions (NetCDF-U)
- ▶ FieldGML: implicit encoding (communicate data + prediction function)

Encoding ST probability distributions

- ▶ UncertML.org: UncertML is a conceptual model and XML encoding designed for encapsulating probabilistic uncertainties
- ▶ NetCDF uncertainty conventions (NetCDF-U)
- ▶ FieldGML: implicit encoding (communicate data + prediction function)
- ▶ 
 - ▶ (still) lacks explicit encodings for ST probability distributions
 - ▶ transparent: communicates what we do, in a reproducible way

Challenges

- ▶ open source, but how to continue development?
- ▶ optimize caching, deal with big data
- ▶ visualize uncertainty in categorical spatio-temporal fields
- ▶ integrate in some useful way with R