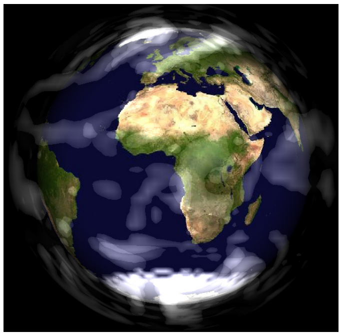
Westfälische
Wilhelms-Universität
Münster

inaugural speach – Antrittsvorlesung

Open geostatistics for global change



Edzer Pebesma

June 25, 2010

ifgi
Institute for Geoinformatics
University of Münster

1 Introduction

Open geostatistics for global change – about these five words I will try to talk for about 45 minutes. In reverse order: I will start off with *global change*, then discuss what the word *geostatistics* means and what we need it for, and then try to explain why I think this whole thing should be *open*. After that, I will explain what we, as a research group at the institute for geoinformatics, have been doing, are doing and would like to do over the next few years. At the end of this talk I hope to share a drink with you else in Room 72, which is next door.

2 Global change

The earth is changing – I hope everyone agrees with me on this one – and most of us would agree that it probably has been changing all the time from the early days on, and that luckily change seems to be modest compared to changes shortly after the big bang.

One question that keeps us busy Today is to which extent this change is caused by human action.

You may have noticed the image of the globe on the cover page of the invitation for this talk [the cover of this paper], and wondered what it represents. The image of the earth is coloured in a way that we might *imagine* we could see rather than what we would ever realistically see; figure 1 shows: how the earth looks when seen from space – indeed the area *around* the earth is black.

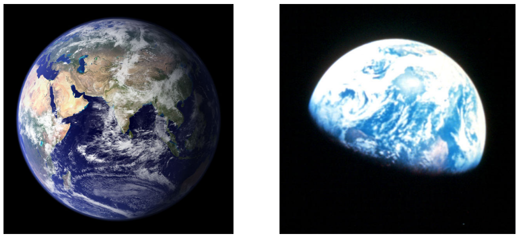


Figure 1: Left: “The blue marble”, another fake earth image but at least a true color composite, taken from `http://visibleearth.nasa.gov/`; right: earth rise seen from Apollo 8

On the cover image, the clouds draped over the earth are an artist impression of a climate simulation. As the atmosphere modelled here extends considerably the earth surface, it is blended into the black background that we expect to see as surrounding the earth when seen from outer space.

Given Todays realism, this blending with black would be easily associated with an artists impression, or model simulation, of the recent Deepwater Horizon oil spill after having spilled for a year or so. To close this still uncontrolled oil spill, the Russians have reported that in Soviet times such leaks were plugged with underground nuclear blasts: *... the underground explosion moves the rock, presses on it, and, in essence, squeezes the wells channel*[1]. An example of fighting one environmental disaster with the next.

Vladimir Lagowski, the author of the Russian newspaper article mentioned, was also kind enought to provide an estimate of failure: *"Total probability of failure in the Gulf of Mexico - 20 percent. Americans could take a chance."* As this was based on five experiments, out of which one failed, a simple analysis of his evidence, shown here in coloured R output:

```
> prop.test(x = 1, n = 5)

1-sample proportions test with continuity correction

data:  1 out of 5, null probability 0.5
X-squared = 0.8, df = 1, p-value = 0.3711
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.01052995 0.70120895
sample estimates:
  p
0.2

Warning message:
In prop.test(1, 5) : Chi-squared approximation may be incorrect
```

This indicates that based on these data instead of believing the chance of failure *is* 20%, we can only *have confidence* that the probability of failure lies between 1% and 70%. This, together with the lack of experience with deep water top kills (acknowledged by the Russian reporter) leaves some room for second thought – given the evidence of the sample data only, the effect might as well be negative.

Most probably, the resemblence of the cover image to oil spill was not thought of as a possibility at the time of creating the graphic. The picture of the black earth used on the invitation is a screen saver that shows climate simulations you can run on your own computer to help climate research. The site motivates us doing so by saying:

> *Climateprediction.net is a distributed computing project to produce predictions of the Earth's climate up to 2100 and to test the accuracy of climate models. To do this, we need people around the world to give us time on their computers – time when they have their computers switched on, but are not using them to their full capacity*[2].

[1] http://www.kp.ru/daily/24482/640124/, Komsomoloskaya Pravda, the best-selling Russian daily; translation from http://trueslant.com/juliaioffe/2010/05/04/nuke-that-slick/

[2] http://climateprediction.net/ claims to be the world's largest climate forecasting experiment for the 21st century

In contrast to many computers from the 20th century, modern computers use more energy when their computing capacity is increased. Although some argue that this energy is mostly converted into heat, so it might safe you costs in heating your house during the winter months, a complicated question remains whether or not this initiative it is a good thing to support. If we do so, we help climate research to better understand climate change, and by doing that, we might help society to take better decisions about whether and which measures are needed to mitigate climate change. If we do not, we safe energy, and potentially reduce global warming and directly mitigate climate change.

There is a difficult moral trade-off here. Likewise with choices what to eat, transportation, waste, health, and energy consumption, freedom that comes with wealth strongly increases the number of choices. Many of these are hard to make using moral arguments. The increase of information availability seems to make life more complex, rather than simple, and this is not only the case for scientists. Geostatistics tries to deal with the limits of inormation availability, or lack of evidence, when decisions need to be made regarding global change.

3 Geostatistics

Geostatistics is a branch of statistics that is concerned with the spatial and spatio-temporal layout of a particular phenomenon. Originating in the areas of meteorology and mining, its application areas now include agriculture, ecology, environment, health, disaster management, remote sensing, radioecology, and many others. A typical problem is the question *where* a particular event occurs, such as the problem to estimate precipitation at arbitrary locations, or identify locations where (and when) some environmental variable exceeds a critical threshold, based on limited measurement data (figure 2).

Since the early linear least-squares solutions to this problem was provided in parallel by pioneers working in the domains of meteorology, mining and forestry [3, 5, 6], over the last three decades the field of geostatistics has gone a long way into accomodating problems where data are multivariate [8], where GIS or remote sensing provides additional layers that cover the full area [7], where data are spatio-temporal [2], follow extreme value distributions [4], and where covariance functions and their parameters are considered uncertain [1]. Data assimilation frameworks [9] that integrate spatio-temporal data with physical model representations today nicely match geostatistical theories for residual processes, the difference between model predictions and observed values.

One could argue that, after 5 decades of success, geostatistics does not have enough selling points to keep it up as a tradition on its own, and I would be the first one to drop the somewhat esoteric jargon it occasionally it favours over the mainstream statistical alternatives. Yet, in the broader context of environmental modelling it deals well, and has not been surpassed by a different theory on two related aspects: *aggregation level* (or *support*) and *uncertainty*. The association that people who are not aware of the history have with the term *geostatistics* is that of statistics for geographical or geoscientific data, for which mainstream

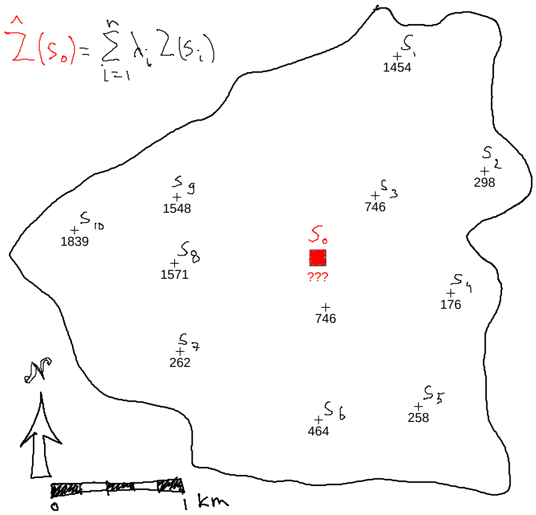


Figure 2: Archaic problem in geostatistics: estimating value $Z(s_0)$ from obervations $Z(s_i)$.

statistics reserves the term *spatial statistics*. In an era where research becomes more complex and more integrated accross domains, some convergence of the two terms might be a good thing.

The *aggregation level* of a certain quantity is strongly related to its variability. Yearly average temperatures fluctuate less, over a ten-year period, then daily average values. Country-averaged incomes for European countries fluctuate less than individual incomes with Europe. As such, when an aggregated quantity needs to be estimated from limited sample data, estimation errors typically decrease when the aggregation area or period is larger. At the cost of (spatial and/or temporal) detail, we can gain statistical accuracy.

Geostatistics then provides the methods and tools to quantify errors, accuracy and uncertainty, explicitly taking *aggregation level* into account. It provides a key ingredient for assessing *risk*, which is for a particular failure event (a disaster, a flood, a nuclear accident etc.) defined as the expected costs (probility of failure times costs of failure). Risk is typically the key criterium when evaluating whether, how much, when and where we need to collect data to support a particular action or decision. Challenges in local, national or environmental monitoring programmes are often to quantify the risk of not detecting a failure event. But in order to apply the precautionary principle[3], risk and hence uncertainties need to be quantified in the first place.

[3] `http://en.wikipedia.org/wiki/Precautionary_principle`, a statutory requirement in the European Union

4 Open

Risk evaluations that have been done by insurance companies are typically not open, and hopefully for good reasons – their competition could lead to sharper rates, benefiting citizens and society at large. But no insurance company insures the world against global change.

In order to take the typically unpopular measures against global change (from increasing fuel taxes to support reforestation, to increasing income taxes to support global change research), a broad public support is needed. In order to build such support, it is necessary that the research that motivated a particular measure is *open* in all possible aspects, including open access to

- data – data should be directly accessible through open standards from the authority that manages the data; appropriate versioning and semantic annotation are a minimum requirement
- software – only open source software available under licenses that allow redistributing modified source code provide sufficient freedom to share and discuss implementation details, and to share improved versions
- well-documented, reproducable procedures – modelling involves the combination of programs with data, as opposed to long sequences of mouse clicks, documented procedures are readable and repeatable
- web-based architectures – sharing ideas, data and algorithms can only work in distributed systems; web-based, interoperable architectures form a minimum requirement to realize this
- explicitly quantified significance and accuracy levels – only when one is explicit and open about the degree to which a research finding can be justified or known, does one have a clue about its value and quality
- well-managed user and developer communities and communication platforms – instead of blogs for dissemination and self-promotion, shared, collaborative research communities need modern means for open communication, that need to be taken care of.

When all this is realized, anyone should be able to carry out slight modifications to the data used, and to improve or replace individual model components to find out how robust the research findings are to slight modifications of data and procedures.

In order to reach this goal, as a research group, and as the open geostatistics[4] initiative, we are involved in the following activities that support all these:

- R, an open source environment for statistical and graphical data analysis, which now has become the *lingua franca* of statistics,

[4] http://www.opengeostatistics.org/

- Applied Spatial Data Analysis with R, a succesful book that explains how to use R for spatial data analysis; the book is supplied by a web site with all the example data and scripts to reproduce every analysis mentioned,
- r-sig-geo, a special interest group mailing list for discussing the development and use of R functions and packages for handling and analysis of geographical data
- ai-geostats, the oldest internet community on geostatistics,
- the Open Geospatial Consortium, a non-profit, international, voluntary consensus standards organization that is leading the development of standards for geospatial and location based services, where ifgi leads a number of standardisation initiatives,
- 52°North, an international research and development company whose mission is to promote the conception, development and application of free open source geo-software for research, education, training and practical use, in which ifgi is the primary founding institute.

Concrete actions of the open geostatitics initiative in particular domains (health, air quality, land use) include actively participating in

- the European Topic Centre on Land Use and Spatial Information (ETC-LUSI)
- the call for the European Topic Centre on Air Pollution and Climate Change Mitigation (ETC-ACM)
- several FP6 and FP7 projects involving integrated assessment and interoperability (HEIMTSA, INTAMAP, UncertWeb, GeoViQua)
- national and international standardization activities, including OWS-7 and AIP-3

5 What we do

This section lists the research topics of (mostly) PhD students currently under my supervision.

Integrated health assessment

When deciding on health policies at a European level, it is necessary to have an idea how large effects will be of measures taken. These need to be expressed in quantities that are comparable accross environmental themes (such as: noise, chemicals, air quality). For exposure assessment, uncertainties can be controlled to some extent by averaging over space, time, or space and time. But do we know enough about e.g. air quality to give reasonable assesments at the individual level? (PhD student: Lydia Gerharz; FP6 Heimtsa)

Optimizing in situ sensor networks

Chernobyl has learned us that radioactive accidents don't happen frequently, but have large and long-lasting consequences when they happen. In situ sensing networks are active all over Europe to detect emergencies and to monitor fall out after accidents. Are these networks sufficient for our current needs? Which cases do they not detect fast enough? Can relocation adapt them to our current needs or do we have to extend them? (PhD student: Kristina Helle; FP7 Detect)

Planning mobile sensor trajectories

The ash cloud and sub-sea level oil spills are two cases of dynamic cloud-like phenomena for which the spatial extent and dynamics are hard to asses. When we have a mobile device, such as an plane, autonomous underwater vehicle, or a drone, which instructions can we give it to delineate the cloud as fast as possible? How can the sensor information be used to optimize the route? (PhD student: Juliane Exeler; IRTG)

Generalizing geosensor networks

Geosensor networks provide point measurement information to internet clients. A first step is to visualize this information, which works for cases up to a few hundred sensors. But which visualisations do we choose when we are looking at an area for which we have hundreds of thousands of sensors available? A generalization strategy is needed, but which generalization algorithms are available for this case? And how can the measured values be used to improve generalization? (PhD student: Christoph Stasch; FP7 UncertWeb)

The uncertainty enabled model web

Currently, the internet is in a transition from providing data to providing services and operations on data. The model web takes this idea into the domains of environmental and ecological models, where data and model components are all offered as web services and chained to complex model structures where components and data can easily exchanged. The uncertainty enabled model web will implement a number of these chains, and set out a framework for rigorous quantitative error propagation through such chains. (Benjamin Proß; FP7 UncertWeb)

Driving forces in Amazon deforestation

Deforestation rates in Brazil have decreased over the last five years, as has been confirmed by the DETER and PRODES systems of from National Institute for Space Research (INPE) in Brazil. The yearly spatial patterns of deforestation over the last ten years are mapped in great detail, but the actual emerging spatio-temporal patterns and their driving forces are little understood. Are they a consequence of global markets, of national policies, or of local factors? (Giovana Mira de Espindola, PhD student at INPE)

Geospatial Service Level Agreements (SLAs)

Geospatial computation will rapidly move to the cloud. This means that organizations will have to rely on Cloud Computing for their services, and need to adopt to the economic rules of that game. Which consequences does this have for geospatial services and data? How can service providers utilize Cloud Computing infrastructures to guarantee certain Quality of Service (QoS)? In addition to the usual metrics, how can measures related to geospatial data in particular be used in Service Level Agreements? How much should one pay for a certain level of spatial detail? Or for a particular level of attribute accuracy? (PhD student: Bastian Baranski)

Geostatistics on accelerated hardware

New hardware architectures such as grid computing, cloud computing and graphical processing units require algorithms to be executed in parallel to speed up processing, and have different memory models under which this is realized. Geospatial, geometric and geostatistical algorithms all have different challenges and problems that need to be addressed before this new branch of computing can be fully exploited. (PhD student: Katharina Henneböhl)

ifgicopter

Quadcopters are not only fun to fly with, but are also operationally useful and cost effective for detailed aerial photography, photographing details of objects that are hard to reach such as buildings or wind mills, or sensing air quality parameters in case of fires in chemical installations. The study project ifgicopter looks at the communication to and from a quadcopter, the planning of routes, and the integration of these platforms with the sensor web architecture. (Holger Fritze, Thore Fechner, Matthes Rieke, Phillip Verhoeven, Christian Knoth, Juliane Exeler and many others)

6 Outlook: what we will do

Over the next few years we will demonstrate that ifgi combines fundamental technological and methodological innovation with strategic demonstrations and implemenations in particular domains, in cooperation with the major partners within Europe and abroad.

Geochange: Amazon deforestation

In cooperation with INPE, a research proposal is in preparation for a project that builds GIScience theory, methodology, and technology dynamic environmental processes that include sensor data management and analysis, process modelling, ubiquitous computing, and semantic interoperability. It aims at integrating aspects of environmental change over four fundamental levels: the sensor level, the modelling level, the level of resulting spatial patterns, and the user interface level where results are shared with decisions makers and stakeholders.

Geoinformatics as an interdisciplinarity science

Spatial location and spatial or spatio-temporal relationships play an increasing role in the collection and analysis of data through many disciplines, and often act as a binding agent. The institute for geoinformatics has the potential to extent its role within and outside the university in interdisciplinary research. Interactions may include but are not limited to the fields of medicine, health, psychology, history, archeology, psychology, religion, economy.

Combining open source projects

A major strength in open source projects comes from transporting ideas accross projects and building further connections between successful projects, environments or applications. The geostatistics community of 52°North has this as one of its aims, and will try to connect R with SWE[5] services, and other services for providing and processing geospatial data.

The overall motivation

Figure 3 shows a picture of the overall information flow and feed-back mechanism that motivates my research. Ideally, decisions made by politicians or individuals are supported by available information. This information is usually derived from processing (analysing, modelling) raw sensor information. Raw sensor information comes top-down from authorities (governements, companies, institutes) or bottom-up from individuals carrying mobile phones with sensors, or using mobile devices to pass information they sense themselves. Flying drones may be used for ad-hoc collection of in-situ or remotely sensed data.

Two major questions we now see are (i) whether the modelling and sensor data available are useful and adequate to support decisions, and if not, (ii) how and to which extent additional information needs to be collected.

7 Zum Schluß

I am very grateful to the University of Münster and the faculty of geosciences to offer me this chair, and to Gilberto Câmara for connecting ifgi to me. The institute for geoinformatics has by far been the most exciting place where I've worked so far. Interactions with my colleages, members in my group, and with

[5] sensor web enablement

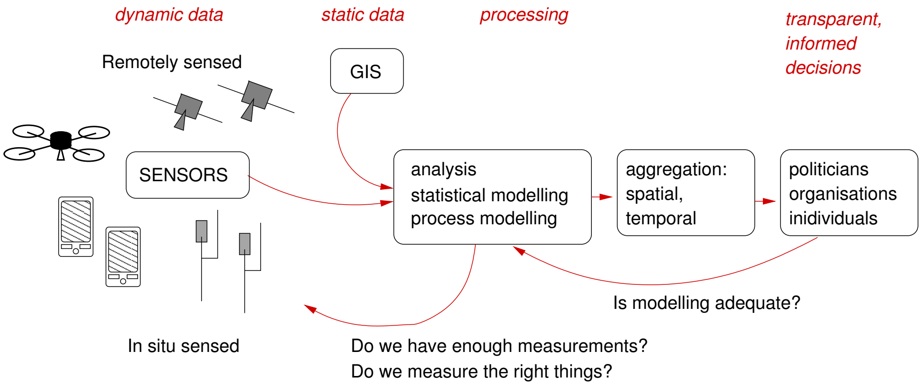


Figure 3: The place of geostatistics and sensor data in integrated, interdisciplinary research

students from all different directions make nearly every day into a new adventure, and something to look forward to. I am very grateful to my family – Ellen, Ulla and Mandus – who had the courage to follow me in order to realize this. Their support and courage during the past few years make me a very happy and proud person.

References

[1] P.J. Diggle and P.J. Ribeiro Jr. *Model-based Geostatistics.* Springer, 2009.

[2] B. Finkenstadt, L. Held, and V Isham. *Statistical Methods for Spatio-Temporal Systems.* Chapman & Hall, 2006.

[3] L S Gandin. *Objective Analysis of Meteorological Fields.* Leningrad, 1963.

[4] H. Kazianka and J. Pilz. Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stochastic Environmental Research and Risk Assessment*, 2010. in press.

[5] Daniel G Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *J. of the Chem., Metal. and Mining Soc. of South Africa*, 52(6):119139, 1951.

[6] B Matern. Spatial variation. *Meddelanden fran Statens Skogsforskningsinsitut*, 49(5), 1960.

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

[8] H. Wackernagel. *Multivariate Geostatistics: An Introduction with Applications.* Springer, 2010.

[9] R Webster and G B M Heuvelink. The kalman filter for the pedologist's tool kit. *European Journal of Soil Science*, 57(6):758 – 773.