Where do spatial statistics and geoinformatics meet?

Edzer Pebesma

ifgi
Institute for Geoinformatics
University of Münster

Geodätischen Kolloquium der
Leibniz Universität Hannover

Jan 29, 2013

Where do spatial statistics and geoinformatics meet?

My answer will be partial (and egocentric).
It will address the question at three levels

- engineering level:
  - Spatial and spatio-temporal data analysis in the R project
- societal level:
  - Spatial statistics, reproducible research
- scientific level:
  - Meaningful spatial prediction and aggregation

Engineering level: INTAMAP, UncertWeb

- INTAMAP: interoperability and automated mapping (2006-2009)
- UncertWeb: the uncertainty-enabled model web (2010-2013)
- both focus on interoperability, the model web, uncertainty, and web services
- outcomes:
  - UncertML, a markup language for probability distributions
  - greenland, a visualisation client for probabilistic, spatio-temporal data
  - R packages WPS4R, spacetime

To which extent is, for instance, R a model web?

Engineering level: INTAMAP, UncertWeb

- INTAMAP: interoperability and automated mapping (2006-2009)
- UncertWeb: the uncertainty-enabled model web (2010-2013)
- both focus on interoperability, the model web, uncertainty, and web services
- outcomes:
  - UncertML, a markup language for probability distributions
  - greenland, a visualisation client for probabilistic, spatio-temporal data
  - R packages WPS4R, spacetime

To which extent is, for instance, R a model web?

Engineering level: INTAMAP, UncertWeb

- INTAMAP: interoperability and automated mapping (2006-2009)
- UncertWeb: the uncertainty-enabled model web (2010-2013)
- both focus on interoperability, the model web, uncertainty, and web services
- outcomes:
  - UncertML, a markup language for probability distributions
  - greenland, a visualisation client for probabilistic, spatio-temporal data
  - R packages WPS4R, spacetime

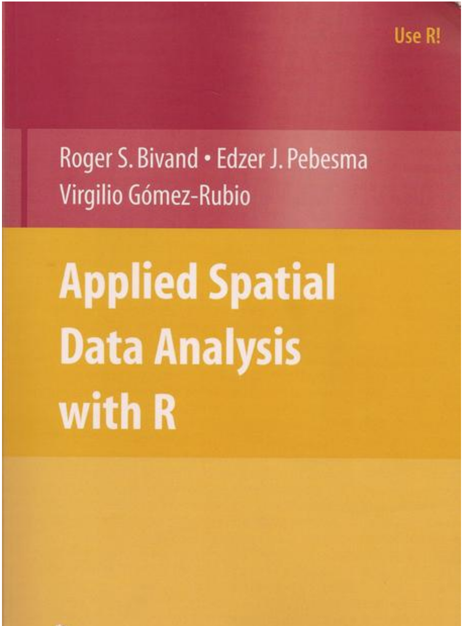
To which extent is, for instance, R a model web?

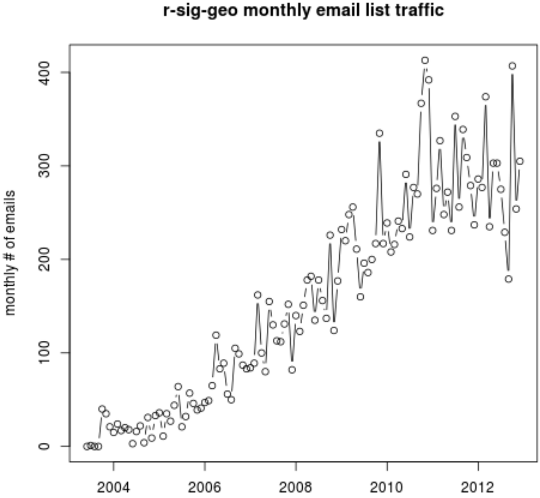
Engineering level: INTAMAP, UncertWeb

- INTAMAP: interoperability and automated mapping (2006-2009)
- UncertWeb: the uncertainty-enabled model web (2010-2013)
- both focus on interoperability, the model web, uncertainty, and web services
- outcomes:
  - UncertML, a markup language for probability distributions
  - greenland, a visualisation client for probabilistic, spatio-temporal data
  - R packages WPS4R, spacetime

To which extent is, for instance, R a model web?

Engineering level: 1.




r-sig-geo monthly email list traffic
monthly # of emails
0
100
200
300
400
2004
2006
2008
2010
2012

CRAN Task View: Handling and Analyzing Spatio-Temporal Data

Maintainer:	Edzer Pebesma
Contact:	edzer.pebesma at uni-muenster.de
Version:	2013-01-28

This task view aims at presenting the useful R packages for the analysis of spatio-temporal data.

Please let the maintainer know if something is inaccurate or missing.

Although one could argue that all data are spatio-temporal, as they must have been taken somewhere and at some point in time, in many cases the spatial locations or times of observation are not registered, and irrelevant to the purpose of the study. Here, we will address the cases where both location *and* time of observation are registered, and relevant for the analysis of the data. The Spatial and TimeSeries task views shed light on spatial, and temporal data handling and analysis, individually.

Representing data

- **In long tables:** In some cases, spatio-temporal data can be held in tables (`data.frame` objects), with longitude, latitude and time as three of the columns, or an identifier for a location or region and time as columns. For instance, data sets in package plm for linear panel models have repeated observations for observational units, where these units often refer to spatial areas (countries, states) by an index. This index (a name, or number) can be matched to the spatial coordinates (polygons) of the corresponding area, an example of this is given by Pebesma (2012, Journal of Statistical Software) . As these data sets usually contain more than one attribute, to hold the data in a two-dimensional table a *long table* form is chosen, where each record contains the index of the observational unit, observation time, and all attributes.

Journal of Statistical Software
November 2012, Volume 51, Issue 7. http://www.jstatsoft.org/

spacetime: Spatio-Temporal Data in R

Edzer Pebesma
University of Münster

Abstract

This document describes classes and methods designed to deal with different types of spatio-temporal data in R implemented in the R package **spacetime**, and provides examples for analyzing them. It builds upon the classes and methods for spatial data from package **sp**, and for time series data from package **xts**. The goal is to cover a number of useful representations for spatio-temporal sensor data, and results from predicting (spatial and/or temporal interpolation or smoothing), aggregating, or subsetting them, and to represent trajectories. The goals of this paper is to explore how spatio-temporal data can be sensibly represented in classes, and to find out which analysis and visualisation methods are useful and feasible. We discuss the time series convention of representing time intervals by their starting time only. This document is the main reference for the R package **spacetime**, and is available (in updated form) as a vignette in this package.

ww.Systat.com

Year	Impact Factor (IF)	Total Articles	Total Cites
[illegible]	4.01	95	1795
2010	2.647	60	1039
2009	2.32	42	714
2008	1.033	43	423

ISI Web of Knowledge℠

Journal Citation Reports®

WELCOME HELP

2011 JCR Science Edition

Journal Summary List

Journal Title Changes

Journals from: **subject categories STATISTICS & PROBABILITY** VIEW CATEGORY SUMMARY LIST

Sorted by: Impact Factor SORT AGAIN

Journals 1 - 20 (of 116)

[1 | 2 | 3 | 4 | 5 | 6]

Page 1 of 6

MARK ALL UPDATE MARKED LIST

Ranking is based on your journal and sort selections.

Mark	Rank	Abbreviated Journal Title *(linked to journal information)*	ISSN	JCR Data						*Eigenfactor®* Metrics	
				Total Cites	Impact Factor	5-Year Impact Factor	Immediacy Index	Articles	Cited Half-life	*Eigenfactor®* Score	*Article Influence®* Score
	1	J STAT SOFTW	1548-7660	1795	4.010	4.791	1.537	95	4.3	0.01176	2.729
	2	J R STAT SOC B	1369-7412	12345	3.645	5.281	0.793	29	>10.0	0.02073	5.382
	3	STAT SCI	0883-4237	3054	3.035	4.205	0.259	27	>10.0	0.00886	3.659
	4	ANN STAT	0090-5364	11722	3.030	3.700	0.423	104	>10.0	0.04315	4.157
	5	ECONOMETRICA	0012-9682	19659	2.976	4.700	0.688	48	>10.0	0.04393	8.648
	6	STAT METHODS MED RES	0962-2802	1835	2.443	2.988	0.500	36	>10.0	0.00570	1.930
	7	STATA J	1536-867X	1250	2.222	3.063	0.147	34	6.4	0.00604	1.871
	8	BIOSTATISTICS	1465-4644	2225	2.145	3.162	0.519	54	7.3	0.01154	2.313
	9	J R STAT SOC A STAT	0964-1998	1685	2.110	2.275	0.327	49	>10.0	0.00631	1.685
	10	PHARM STAT	1539-1604	422	2.067	2.160	0.463	67	4.1	0.00254	1.056
	11	J AM STAT ASSOC	0162-1459	21348	1.992	3.310	0.240	121	>10.0	0.03691	3.115

scholar.google.com:edzer pebesma

Change photo

Edzer Pebesma Edit

Professor of geoinformatics, University of Muenster Edit

spatial statistics - geostatistics - interoperability - reproducible research - R Edit

Verified email at uni-muenster.de Edit

My profile is public Edit Link Homepage Edit

Citation indices

	All	Since 2008
Citations	2531	1733
h-index	22	19
i10-index	31	24

Citations to my articles

Select: All, None Actions ▾ Show: 20 ▾ **1-20** Next >

Title / Author	Cited by	Year
Multivariable geostatistics in S: the gstat package EJ Pebesma Computers & Geosciences 30 (7), 683-691	448	2004
Applied spatial data analysis with R RS Bivand, EJ Pebesma, V Gómez-Rubio Springer	386	2008
Gstat: a program for geostatistical modelling, prediction and simulation EJ Pebesma, CG Wesseling Computers & Geosciences 24 (1), 17-31	356	1998
Spatial aggregation and soil process modelling G Heuvelink, EJ Pebesma Geoderma 89 (1), 47-65	137	1999
Classes and methods for spatial data in R EJ Pebesma, RS Bivand R News 5 (2), 9-13	99	2005

Societal level: OSS and reproducible research

Eos, Vol. 93, No. 16, 17 April 2012

FORUM

The R Software Environment in Reproducible Geoscientific Research

PAGE 163

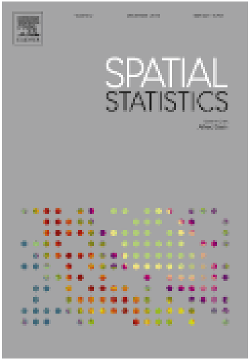
Reproducibility is an important aspect of scientific research, because the credibility of science is at stake when research is not reproducible. Like science, the development of good, reliable scientific software is a social process. A mature and growing community relies on the R software environment for carrying out geoscientific research. Here we describe why people use R and how it helps in communicating and reproducing research.

R is a multiplatform open-source software environment [*R Development Core Team*, 2012] that implements S, a language designed for data analysis. It pro-

relieving developers of the burden of multiplatform support. Package verification mechanisms make CRAN a reliable and powerful resource for users and developers.

When writing R packages, one assumes that R works in a certain way and will continue doing so. When the working of R changes, there is a chance that this change will break a package, i.e., stop an extension package from working, especially when the package was using R syntax sloppily. Although R core uses the 3500 extension packages on CRAN to verify the impact of planned changes, improvements that break some packages are at times needed. In such a case, maintainers of packages affected are notified in a timely way so that they can take

Societal level: spatial statistics @Elsevier



Scientific level:

Meaningful Spatial Prediction and Aggregation

Christoph Stasch[a,*], Simon Scheider[a], Edzer Pebesma[a,b], Werner Kuhn[a]

[a] *Institute for Geoinformatics, University of Muenster, Weseler Strasse 253, 48153 Muenster, Germany*
[b] *52 North Initiative for Geospatial Open Source Software GmbH, Martin-Luther-King-Weg 24, 48151 Muenster, Germany*

Abstract

The appropriateness of spatial prediction methods such as Kriging, or aggregation methods such as summing observation values over an area, is currently judged by domain experts using their knowledge and expertise. In order to provide support from information systems for automatically discouraging or proposing prediction or aggregation methods for a dataset, expert knowledge needs to be formalized. This involves, in particular, knowledge about phenomena represented by data and models, as well as about underlying procedures. In this paper, we introduce a novel notion of *meaningfulness* of prediction and aggregation. To this end, we present a formal theory about spatio-temporal variable types, observation procedures, as well as interpolation and aggregation procedures relevant in Spatial Statistics. Meaningfulness is defined as correspondence between functions and data sets, the former representing *data generation procedures* such as observation and prediction. Comparison is based on *semantic reference systems*, which are types

How do point data look?

```
> library(gstat)
> data(meuse)
> meuse[1:5, c("x","y","zinc")]

       x      y zinc
1 181072 333611 1022
2 181025 333558 1141
3 181165 333537  640
4 181298 333484  257
5 181307 333330  269

> co2 = read.csv("co2_emission_powerplants.csv")
> co2[1:5, c("longitude", "latitude", "carbon_2007")]

  longitude latitude carbon_2007
1 14.453050 51.83248    27400000
2  6.575827 51.05470    24100000
3  6.668831 50.99228    30400000
4  6.615766 51.03780    22200000
5  6.313576 50.83805    22000000
```

similar, but can we do similar things with them?

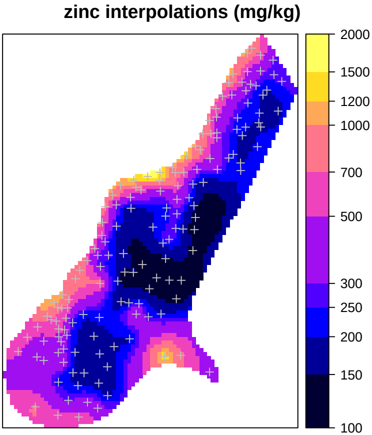
Interpolating heavy metal concentration in soil

following Burrough & McDonell, 1998:

```
> coordinates(meuse) = ~x+y
> v = variogram(log(zinc)~1, meuse)
> v.fit = fit.variogram(v, vgm(1, "Sph", 900, 1))
> data(meuse.grid)
> gridded(meuse.grid) = ~x+y
> m.kr = krige(log(zinc)~1, meuse, meuse.grid, v.fit)
```

```
[using ordinary kriging]
```

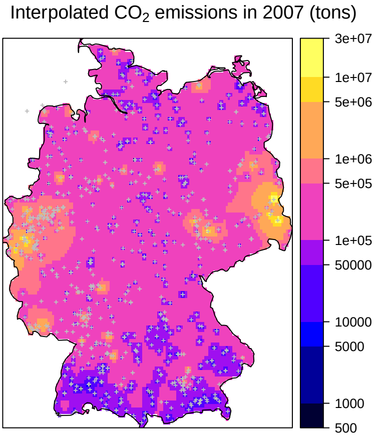
```
> spplot(m.kr["var1.pred"])
```



Interpolating power plant CO_2 emissions

```
> # load the country border of Germany:
> library(cshapes)
> cntr <- cshp(date=as.Date("2008-06-30"))
> germany <- cntr[cntr$CNTRY_NAME == "Germany",]
> # clean co2 data:
> co2 <- co2[co2$latitude != 0 & co2$carbon_2007 != 0,
+  c("latitude","longitude","carbon_2007")]
> # convert table to Spatial:
> coordinates(co2) = ~longitude+latitude
> proj4string(co2) = proj4string(germany)
> # create interpolation grid:
> grd = spsample(germany, 10000, "regular", offset = c(0,0))
> gridded(grd) = TRUE
> # interpolate, idw:
> co2_interpolated <- krige(carbon_2007~1, co2, grd)

[inverse distance weighted interpolation]
```



Which one is meaningful?

At an unobserved location, what does

a predicted zinc concentration in top soil mean?

b predicted coal power plant CO_2 emission mean?

what does *unobserved location* mean?

a a location, with soil, with "similar" conditions?

b a site with

- a power plant with unknown emission, or
- an arbitrary site with no power plant?

Which one is meaningful?

At an unobserved location, what does

- **a** predicted zinc concentration in top soil mean?
- **b** predicted coal power plant CO_2 emission mean?

what does *unobserved location* mean?

- **a** a location, with soil, with "similar" conditions?
- **b** a site with
  - a power plant with unknown emission, or
  - an arbitrary site with no power plant?

Which one is meaningful?

At an unobserved location, what does

a predicted zinc concentration in top soil mean?

b predicted coal power plant CO_2 emission mean?

what does *unobserved location* mean?

a a location, with soil, with "similar" conditions?

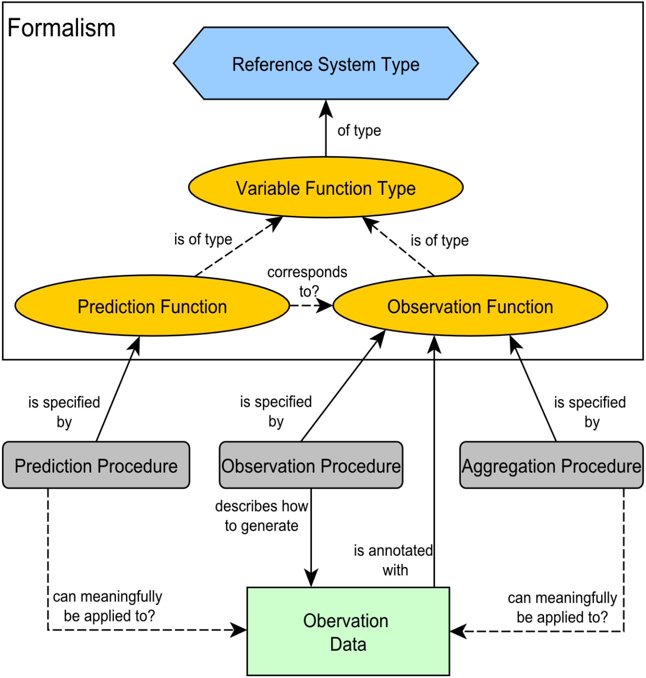
b a site with

- a power plant with unknown emission, or
- an arbitrary site with no power plant?

When is summing meaningful?

```
> with(as.data.frame(meuse), sum(zinc))
[1] 72806
> with(as.data.frame(co2), sum(carbon_2007))
[1] 407225925
```

Which of these two sums is meaningful?


Formalism
Reference System Type
of type
Variable Function Type
is of type
is of type
corresponds
to?
Prediction Function
Observation Function
is specified
by
is specified
by
is specified
by
Prediction Procedure
Observation Procedure
Aggregation Procedure
describes how
to generate
is annotated
with
can meaningfully
be applied to?
can meaningfully
be applied to?
Obervation
Data

Types of Reference System Domains.

Reference Domain	Type	Description	Example
Domain of a Spatial Reference System	D_s	All possible locations that are defined in a spatial reference system; we restrict D_s to $D_s \subset \mathbb{R}^2$	$([-90, 90] \times [-180, 180]) \subset \mathbb{R}^2$ defined in WGS84
Domain of a Temporal Reference System	D_t	All possible times defined in a temporal reference system	POSIX time (seconds from 1st January 1970 UTC) with $D_t \subset \mathbb{Q}$
Domain of a Quality Reference System	D_q	Set of all values that a quality might take	$[0, 10^6] \subset \mathbb{R}$ with unit ppm as defined in Unified Code for Units of Measure (UCUM)
Domain of a Discrete Entities	D_d	Set of discrete objects or events.	Set of coal power plants in Germany in 2010

Types of Point Pattern Variables in Spatial Statistics

Variable type	**Functional type**	**Example**
Spatial Point Pattern	$SPP = "D_d \Rightarrow D_s"$	Locations of longleaf pines in 4 ha of a natural forest in Thomas County, Georgia
Marked Spatial Point Pattern	$MSPP = "D_d \Rightarrow (D_s \times D_q)"$	Locations of longleaf pines in 4 ha of a natural forest in Thomas County, Georgia with diameters-at-breast-heights (DBH)
Temporal Point Pattern	$TPP = "D_d \Rightarrow D_t"$	Occurrence of earthquakes in an american county
Marked Temporal Point Pattern	$MTPP = "D_d \Rightarrow (D_t \times D_q)"$	Occurrence of earthquakes in an american county with magnitudes
Spatio-temporal Point Pattern	$STPP = "D_d \Rightarrow (D_t \times D_s)"$	Occurrence of earthquakes at particular locations in space and time

Types of Geostatistical and Lattice Variables in Spatial Statistics

Variable type	Functional type	Example
Geostatistical Variable	$GEOST = "D_s \Rightarrow D_t \Rightarrow D_q"$	PM_{10} concentrations across Germany
Lattice Variable	$LAT = "\, ^rD_s \Rightarrow D_t \Rightarrow D_q"$	Number of doctor-prescriptions per consultation in cantons of the Midi-Pyrenees

Trajectory Variable Type in Spatial Statistics

Variable type	Functional type	Example
Trajectory	$TRAJECT =$ "$D_d \Rightarrow D_t \Rightarrow D_s$"	paths of tracked animals
Marked Trajectory	$MTRAJECT =$ "$D_d \Rightarrow D_t \Rightarrow (D_s \times D_q)$"	paths of tracked animals with measurements of body temperature

Functions for Basic Observation Procedures

Observation procedure	Observation function	Example
Object localization procedure	$obs_{loc} :: D_d \Rightarrow D_t \Rightarrow D_s$	Location of a coal power plant by centroid (any spatial point pattern)
Object property observation procedure	$obs_{prop} :: D_d \Rightarrow D_t \Rightarrow D_q$	CO_2 emission rate of a power plant at a series of times
Continuous phenomenon observation procedure	$obs_{cphen} :: D_s \Rightarrow D_t \Rightarrow D_q$	Observation of PM_{10} concentrations across Germany

Function for an ordinary Kriging procedure

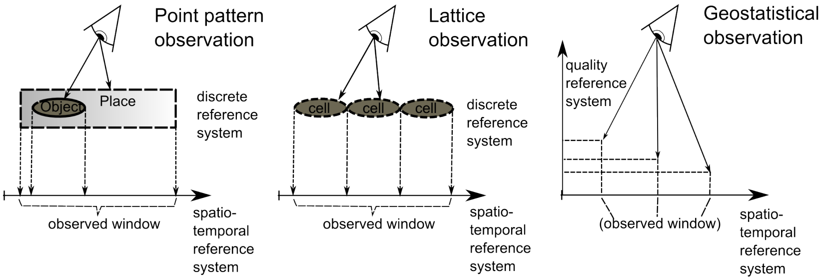
Prediction procedure	**Prediction Function**	**Example**
Ordinary kriging procedure	$pred_{geost} :: GEOST$	Spatial interpolation of PM_{10} measurements using ordinary kriging.

Types of Measurement Scales and Permissible statistics

(after Stevens, 1946); statistics permissible for lower scales are also permissible for higher scale variables, but not vice-versa.

Scale Type	**Permissible Statistics**
Nominal	Count (number of cases), Mode, Contingency
Ordinal	Median, Percentiles
Interval	Mean, Standard Deviation, rank-order correlation, product-moment correlation
Ratio	Coefficient of variation

Observation windows

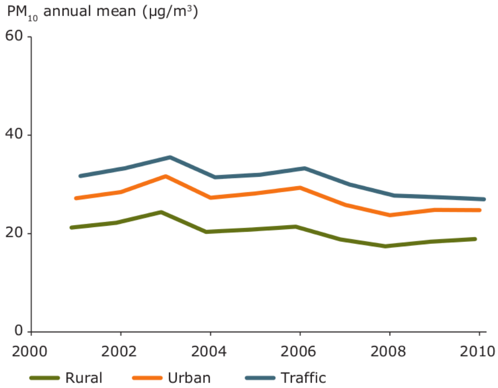


Meaningfulness

- Meaningfulness checks are implemented in our formalism as correspondence checks:
- Meaningful prediction is introduced based on a correspondence check between observation functions and prediction functions that ensures that there is a possible observation for each prediction.
- Meaningful aggregation is based on checking whether an observed window corresponds to the target regions of an aggregation, hence testing the condition, that the target region needs to be observed completely in case of using the sum as an aggregation function.

Air quality in Europe: EEA report 4/2012

Trends in PM_{10} ($\mu g/m^3$), 2001-2010, per station type

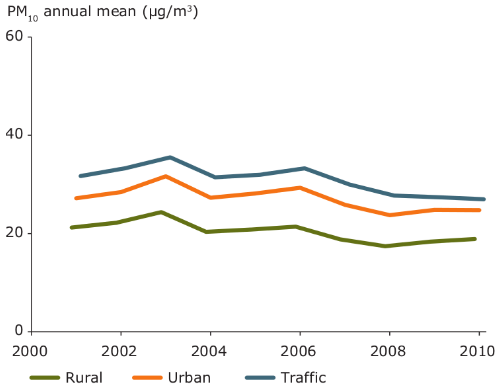


"in the diagrams a geographical bias exists towards central Europe where there is a higher density of stations"

to obtain aggregate values for Europe, one needs to aggregate predictions over Europe (block kriging)

Air quality in Europe: EEA report 4/2012

Trends in PM$_{10}$ ($\mu g/m^3$), 2001-2010, per station type



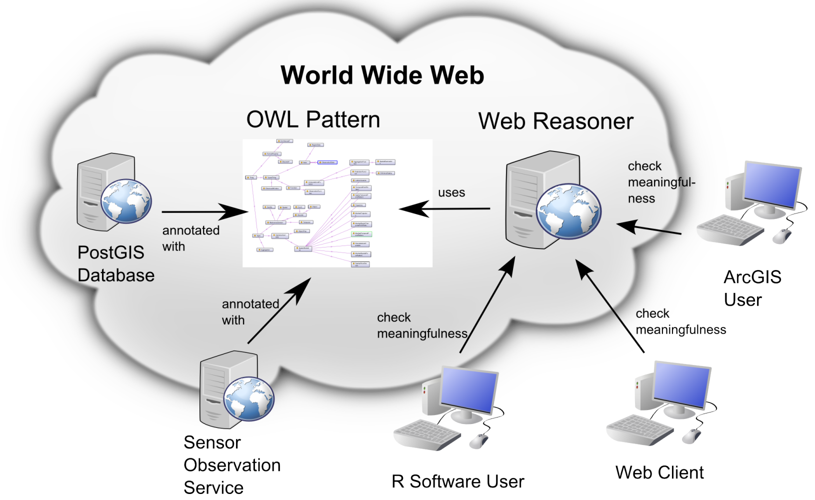
"in the diagrams a geographical bias exists towards central Europe where there is a higher density of stations"

to obtain aggregate values for Europe, one needs to aggregate predictions over Europe (block kriging)

The limits of classes

- are traffic station air quality measurements geostatistical variables? (stationarity assumption)
- what does the sum of observed bird counts mean, in particular in case of volunteered information?
- if not meaningful, what *do* interpolated point pattern mark tell us?

OWL pattern


World Wide Web
OWL Pattern
Web Reasoner
PostGIS
Database
annotated
with
uses
check
meaningful-
ness
ArcGIS
User
annotated
with
check
meaningfulness
check
meaningfulness
Sensor
Observation
Service
R Software User
Web Client

Outlook

- this paper constrained to meaningfulness of spatio-temporal prediction and aggregation, and addressed geostatistical and point pattern data.
- lattice data, or trajectory data were not addressed
- rasters, imagery has not been addressed
- "deeper" statistical problems include model estimation, and model selection (evaluating assumptions)
- what can we do to make our theory work (help) in practice?
- how can random variables, or variables with uncertainty, be represented in the formalism?

Conclusions

- The fringe zone between geoinformatics and spatial statistics offers several exciting challenges at the engineering, societal and scientific level, even that of information theory.
- I mentioned a few, related to activities in my group, that addressed R, interoperability, model web, and semantic reference systems (and ignored data analysis, air quality and exposure modelling, monitoring network design, land use change in Brazil, 52°North, and citizen science)
- Most of this work is completely in the open, meaning that everyone is welcome to participate!