Support of observations and predictions in spatial and temporal statistics: practical aspects and software challenges.

Edzer Pebesma

ifgi
Institute for Geoinformatics
University of Münster

DAGStat Tagung 2016, 14.3. - 18.3.2016, Göttingen

Overview

1. “Change of Support”
2. Examples on spatial data
3. Examples on time series
4. Modelling spatiotemporal information generation
5. Software challenges

Change of support (COS)

"Support" is the physical size and temporal duration of that, where a measurement or prediction refers to.
All approaches to spatial and spatiotemporal data adopt some kind of stationary model for the data, e.g.

$$Z(s) = \mu + e(s), \quad Z(s) \sim \mathcal{N}(\mu, \Sigma)$$

with $\Sigma_{ij} = \text{Cov}(Z(s_i), Z(s_j))$, leading to the simple kriging / BLP equations

$$\hat{Z}(s_0) = \mu + \Sigma_0 \Sigma^{-1}(Z(s) - \mu)$$

$$\text{Var}(\hat{Z}(s_0) - Z(s_0)) = \sigma_Z^2 - \Sigma_0 \Sigma^{-1} \Sigma_0$$

where element i of Σ_0 equals $\text{Cov}(Z(s_i), Z(s_0))$.

Change of support (COS) - 2

Block kriging estimates the "block" mean value

$$Z(B_0) = |B|^{-1} \int_B Z(u)\,du$$

by

$$\hat{Z}(B_0) = \mu + \Sigma_0 \Sigma^{-1}(Z(s) - \mu)$$

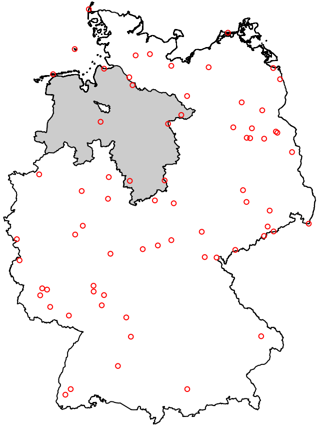
$$\mathrm{Var}(\hat{Z}(B_0) - Z(B_0)) = \sigma^2_{Z(B)} - \Sigma_0 \Sigma^{-1} \Sigma_0$$

when replacing

- $\mathrm{Cov}(Z(s_i), Z(s_0))$ with $\mathrm{Cov}(Z(s_i), Z(B_0)) = |B|^{-1} \int_B \mathrm{Cov}(Z(s_i), Z(u))\,du$
- σ^2_Z with $\sigma^2_{Z(B)} = |B|^{-2} \int_B \int_B \mathrm{Cov}(Z(u), Z(v))\,du\,dv$

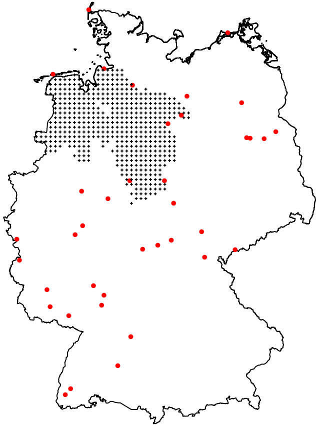
COS: What is it?

```
> library(sp)
> library(spacetime)
> data(air) # loads stations, dates, air, DE_NUTS1
> rural = STFDF(stations, dates,
+  data.frame(PM10 = as.vector(air)))
> utm32N = CRS("+proj=utm +zone=32 +north +datum=WGS84")
> rural = spTransform(rural, utm32N)
> DE_NUTS1 = spTransform(DE_NUTS1, utm32N)
> library(rgeos)
> DE = gUnionCascaded(DE_NUTS1)
> plot(DE)
> Niedersachsen = DE_NUTS1["Niedersachsen",]
> plot(Niedersachsen, col = grey(.8), add = TRUE)
> points(as(rural, "Spatial"), col = 'red')
> r = rural[ , "2009-01-10"]
> (sample_mean = as.data.frame(
+  aggregate(r, Niedersachsen, FUN = mean, na.rm = TRUE)))

                PM10
Niedersachsen 21.677
```



COS: What is it? – 2

```
> r = r[!is.na(r$PM10),]
> library(gstat)
> v = variogram(PM10~1, r)
> (f = fit.variogram(v, vgm("Exp")))

  model  psill range
1   Nug   0.00     0
2   Exp 352.75 92427

> plot(v, f)
> pts = spsample(Niedersachsen, 500, "regular",
+  offset = c(.5,.5))
> k1 = krige(PM10~1, r, pts, f) # 500 points

[using ordinary kriging]

> c(mean(k1$var1.pred), mean(k1$var1.var), var(k1$var1.pred))

[1]  22.558 203.103  46.955

> k2 = krige(PM10~1, r, Niedersachsen, f) # 1 block

[using ordinary kriging]

> as.data.frame(k2)[,3:4]

              var1.pred var1.var
Niedersachsen    22.558   35.383

> sample_mean

                 PM10
Niedersachsen 21.677
```



COS: history

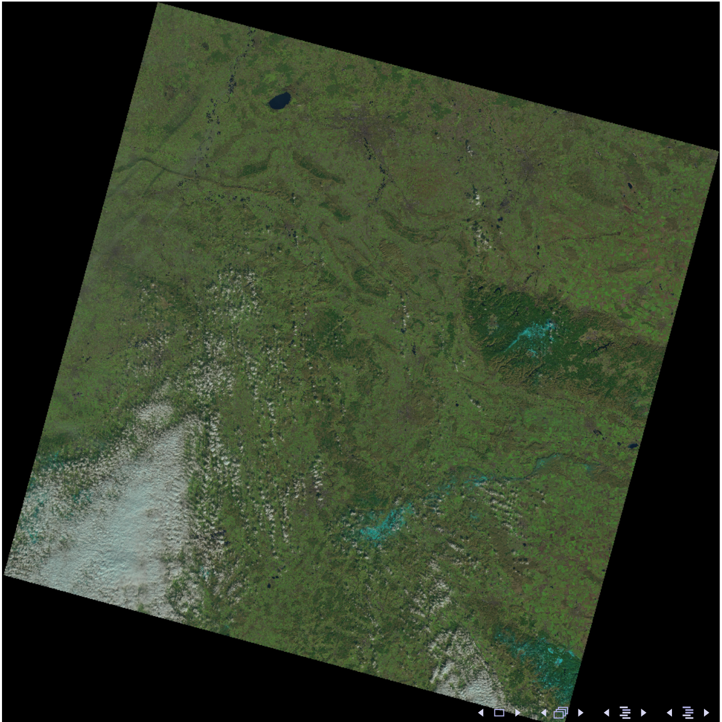
- 1960's: mining industry, D. Krige, G. Matheron
- motivation: measurements are cores, 'minable units' are blocks.
- other "mining remains":
  - "kriging" – Danie G. Krige was a South-African mining engineer
  - "nugget effect": sudden, dramatic variations over short distances
  - dominant use of the (semi)variogram, rather then the covariogram
- observed "blocks" data: socio-economic, population, satellite
- generated "blocks" data: GCM's, weather models
- Cressie: for non-linear $g(\cdot)$, $\int g(Z(s)) \neq g(\int(Z(s)))$
- in "point data", what does the word "point" mean?
- Spatial Statistics: "areal or lattice data"
- ecological regression: build models from (spatially) aggregated data

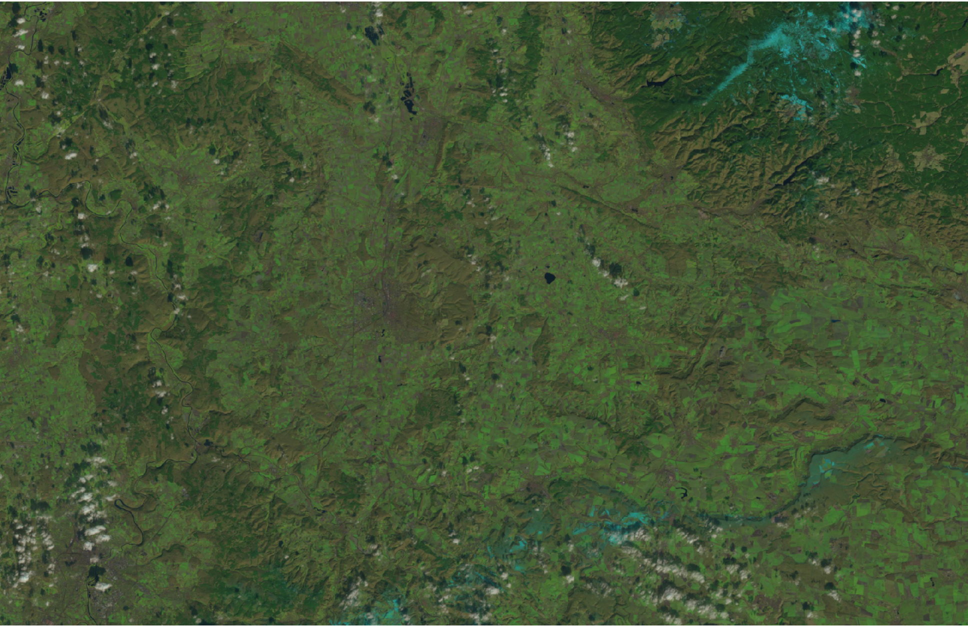
Spatial Data

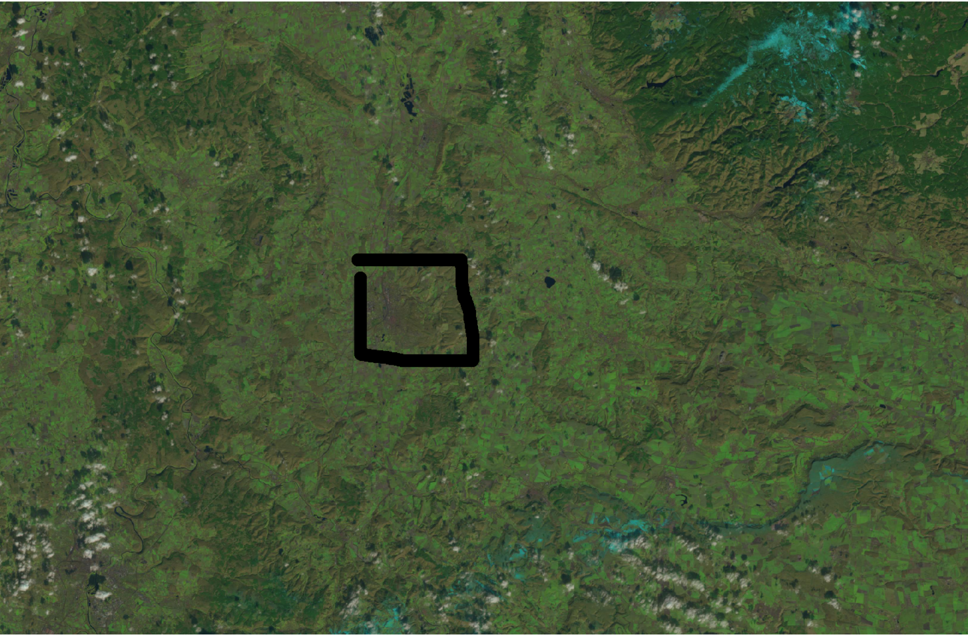
Landsat 8 data

- Landsat 8, Göttingen area, 9-3-16
- http://earthexplorer.usgs.gov/ ; free registration, download trivial
- format: georeferenced jpeg; Coordinates in UTM
- 30 m $\times$ 30 m pixels;
- extent: $\pm$ 8000 $\times$ 8000 pixels, $\pm$ 240 km $\times$ 240 km
- 7 spectral bands; took (default) RGB composite;
- "scene" imported and plotted by R:

```
> library(rgdal)
> r = readGDAL("LC81950242016069LGN00.jpg")

LC81950242016069LGN00.jpg has GDAL driver JPEG
and has 7991 rows and 7881 columns

> image(r, red = "band1", green = "band2", blue = "band3")
```







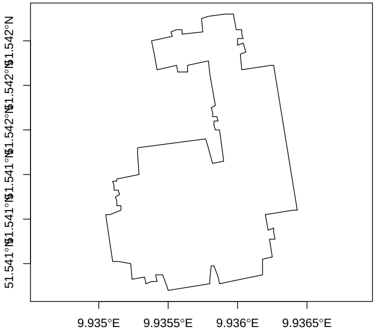
Where are we?

- OpenStreetMap is a data set with roads, buildings, and other many things
- after searching for "Göttingen shapefile", I found that Bike friends had cut it in pieces convenient for tourists

  http:

  //download.bbbike.org/osm/bbbike/Goettingen/
- I downloaded the Göttingen area as a shapefile, and identified the ZHG building (ID 115376791) using Quantum GIS (an open source, interactive GIS).

```
> g = readOGR("Goettingen-shp/shape", "buildings")

OGR data source with driver: ESRI Shapefile
Source: "Goettingen-shp/shape", layer: "buildings"
with 105184 features
It has 3 fields

> proj4string(g)

[1] "+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +t

> ZHG = subset(g, osm_id == 115376791) # ZHG
> plot(ZHG, axes = TRUE)
```

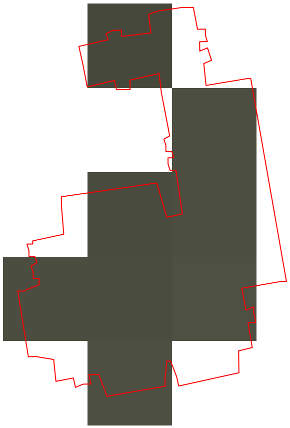


... in the “context” of Landsat 8?

```
> proj4string(r)

[1] "+proj=utm +zone=32 +datum=WGS84 +units=m +no_defs +ellps=WG

> ZHG = spTransform(ZHG, CRS(proj4string(r)))
> bbr = bbox(r)
> bbz = bbox(ZHG)
> # y, rows:
> (bbr[2,2] - (bbz[2,1] + 30 * 20))/30

[1] 4514.7

> # x, cols:
> ((bbz[1,1] - 30 * 20) - bbr[1,1])/30

[1] 4449.2

> r0 = r[4514:4554, 4449:4490]
> par(mar = c(0,0,1,0))
> image(r0, red = "band1", green = "band2",
+     blue = "band3")
> plot(ZHG, border = 'red', add=TRUE)
```

… and what is the color of our roof?

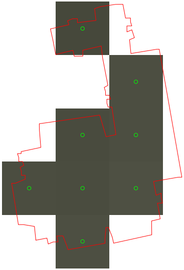
```
> fullgrid(r0) = FALSE
> image(r0[ZHG,,drop=TRUE],
+  red = "band1", green = "band2", blue = "band3")
> plot(ZHG, border = 'red', add=TRUE)
```



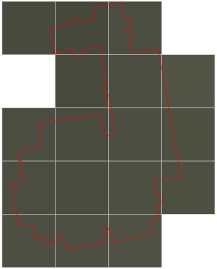
Roof color: ...

Compute mean of:

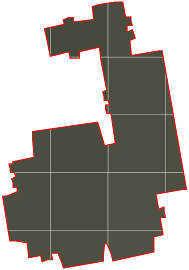
Intersecting centres:



Intersecting cells:



Intersection, area weighted:

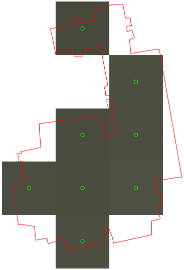


all answers are `FALSE`

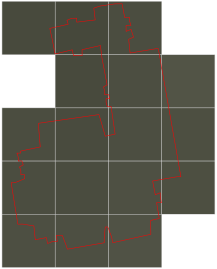
Roof color: ...

Compute mean of:

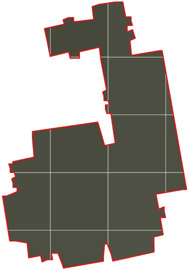
Intersecting centres:



Intersecting cells:



Intersection, area weighted:



all answers are `FALSE`

Generalizing block kriging

When data are blocks, (how) can we estimate (i) arbitrary blocks and (ii) point values (disaggregation)?

$$Z(B) = \mu + e(B), \quad Z(B) \sim \mathcal{N}(\mu, \Sigma)$$

with $\Sigma_{ij} = \text{Cov}(Z(B_i), Z(B_j))$, which equals

$$|B_i|^{-1}|B_j|^{-1} \int_{Bi} \int_{Bj} \text{Cov}(Z(u), Z(v))\, du dv$$

and

$$\hat{Z}(B_0) = \mu + \Sigma_0 \Sigma^{-1}(Z(B) - \mu)$$

and where element i of Σ_0 equals $\text{Cov}(Z(B_i), Z(B_0))$.

This still needs the "point-to-point" covariance. How to infer this from block-only data?

$\Rightarrow$ what does a point covariance mean, when the process is discrete (e.g. population counts)?

Generalizing block kriging

When data are blocks, (how) can we estimate (i) arbitrary blocks and (ii) point values (disaggregation)?

$$Z(B) = \mu + e(B), \quad Z(B) \sim \mathcal{N}(\mu, \Sigma)$$

with $\Sigma_{ij} = \mathrm{Cov}(Z(B_i), Z(B_j))$, which equals

$$|B_i|^{-1}|B_j|^{-1} \int_{Bi} \int_{Bj} \mathrm{Cov}(Z(u), Z(v))\, du dv$$

and

$$\hat{Z}(B_0) = \mu + \Sigma_0 \Sigma^{-1}(Z(B) - \mu)$$

and where element i of Σ_0 equals $\mathrm{Cov}(Z(B_i), Z(B_0))$.
This still needs the "point-to-point" covariance. How to infer this from block-only data?
$\Rightarrow$ what does a point covariance mean, when the process is discrete (e.g. population counts)?

Generalizing block kriging

When data are blocks, (how) can we estimate (i) arbitrary blocks and (ii) point values (disaggregation)?

$$Z(B) = \mu + e(B), \quad Z(B) \sim \mathcal{N}(\mu, \Sigma)$$

with $\Sigma_{ij} = \mathrm{Cov}(Z(B_i), Z(B_j))$, which equals

$$|B_i|^{-1}|B_j|^{-1} \int_{Bi} \int_{Bj} \mathrm{Cov}(Z(u), Z(v))\, dudv$$

and

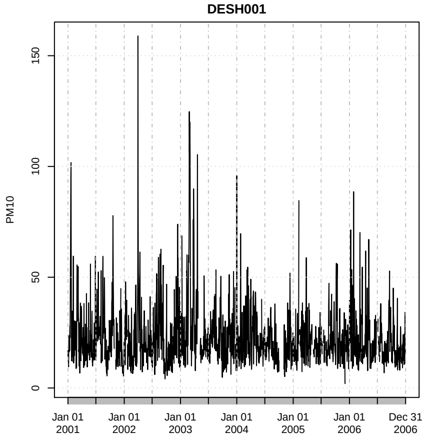
$$\hat{Z}(B_0) = \mu + \Sigma_0 \Sigma^{-1}(Z(B) - \mu)$$

and where element i of Σ_0 equals $\mathrm{Cov}(Z(B_i), Z(B_0))$.
This still needs the "point-to-point" covariance. How to infer this from block-only data?
$\Rightarrow$ what does a point covariance mean, when the process is discrete (e.g. population counts)?

Temporal Data

Time aggregation: PM10 data

```
> library(spacetime)
> data(air)
> rural = STFDF(stations, dates,
+  data.frame(PM10 = as.vector(air)))
> class(rural)

[1] "STFDF"
attr(,"package")
[1] "spacetime"

> pm10 = rural[1,"2001::2006"][,1]
> class(pm10)

[1] "xts" "zoo"

> station = row.names(rural[,1])[1]
> class(index(pm10))

[1] "Date"

> plot(pm10, main = station, ylab = "PM10")
```



Time aggregation

```
> yr = with(as.POSIXlt(index(pm10)), 1900 + year)
> pm10.yr = aggregate(pm10, yr, na.rm = TRUE)
> class(pm10.yr)

[1] "zoo"

> pm10.yr

2001 8270.0
2002 8389.4
2003 9236.4
2004 7066.1
2005 7059.2
2006 7348.2
```

… ehm …

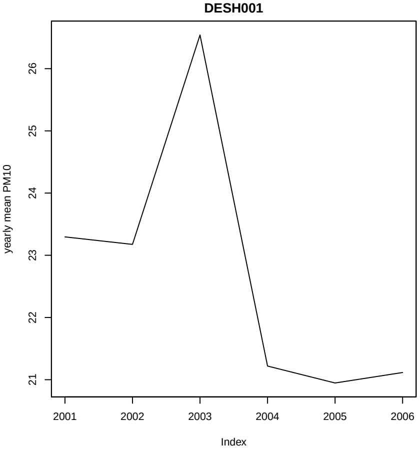
Time aggregation

```
> yr = with(as.POSIXlt(index(pm10)), 1900 + year)
> pm10.yr = aggregate(pm10, yr, na.rm = TRUE)
> class(pm10.yr)

[1] "zoo"

> pm10.yr

2001 8270.0
2002 8389.4
2003 9236.4
2004 7066.1
2005 7059.2
2006 7348.2
```

... ehm ...

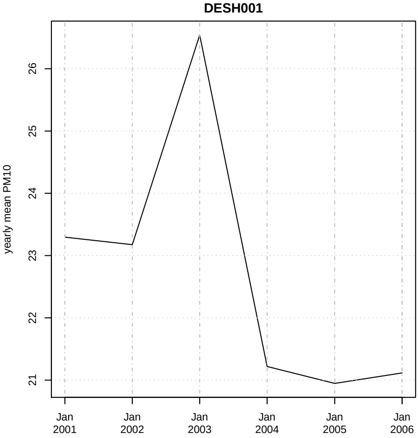
Time aggregation .. 2

```
> yr = with(as.POSIXlt(index(pm10)), 1900 + year)
> pm10.yr = aggregate(pm10, yr, FUN = mean, na.rm = TRUE)
> plot(pm10.yr, main = station, ylab = "PM10")
> pm10.yr

2001 23.296
2002 23.175
2003 26.541
2004 21.220
2005 20.947
2006 21.116

> class(pm10.yr)

[1] "zoo"

> class(index(pm10.yr))

[1] "numeric"
```

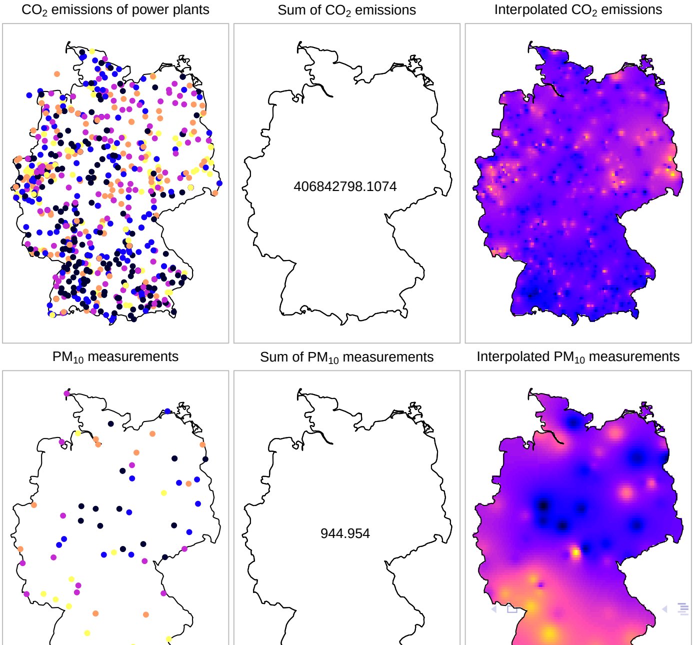


Time aggregation .. 3

```
> yr.POSIXct = strptime(paste0(index(pm10.yr), "-01-01"),
+  format = "%Y-%m-%d", tz = "UTC")
> library(xts)
> pm10.yr = xts(pm10.yr, yr.POSIXct)
> plot(pm10.yr, main = station, ylab = "PM10")
> pm10.yr

                x
2001-01-01 23.296
2002-01-01 23.175
2003-01-01 26.541
2004-01-01 21.220
2005-01-01 20.947
2006-01-01 21.116

> class(pm10.yr)

[1] "xts" "zoo"

> class(index(pm10.yr))

[1] "POSIXlt" "POSIXt"
```



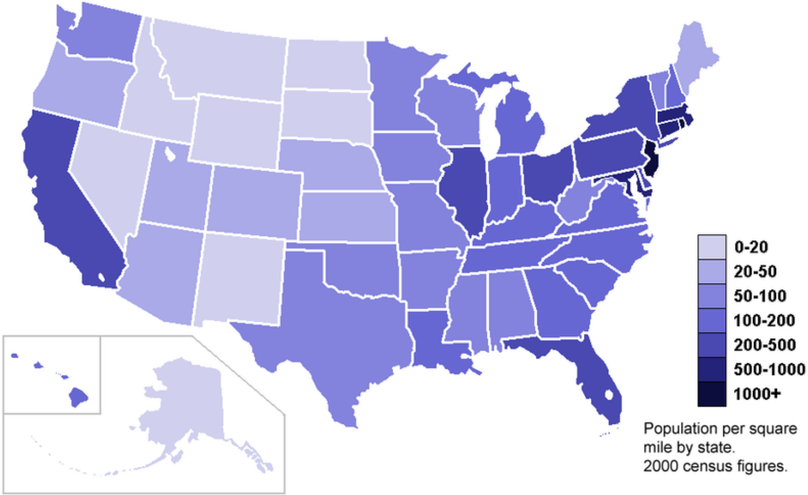
Meaningful spatial statistics



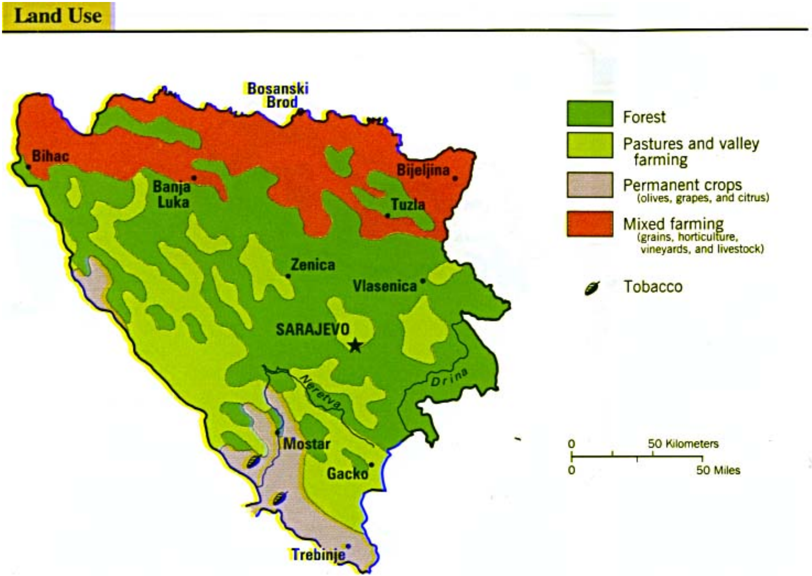
Spatial data bases: PostGIS view

```
user=# select * from co2 limit 3;
 pk | plant_id |     name     | carbon_2007 |        location
----+----------+--------------+-------------+------------------------
  1 |    20075 | JANSCHWALDE  |    27400000 | POINT(14.45305 51.83248)
  2 |    14153 | FRIMMERSDORF |    24100000 | POINT(6.575827 51.0547)
  3 |    31142 | NIEDERAUSSEM |    30400000 | POINT(6.668831 50.99228)
(3 rows)

user=# select * from pm10 limit 3;
 pk | station |    time    | pm10 |         location
----+---------+------------+------+-------------------------
  1 | AT0ENK1 | 2005-06-01 |   14 | POINT(13.67111 48.39167)
  2 | AT30202 | 2005-06-01 |  9.7 | POINT(15.91944 48.10611)
  3 | AT4S108 | 2005-06-01 |  7.8 | POINT(14.57472 48.53111)
(3 rows)

user=# select * from geometry_columns;
 f_table_name | f_geometry_column | dim  | srid  | type
--------------+-------------------+------+-------+------
 pm10         | location          |    2 |  4326 | POINT
 co2          | location          |    2 |  4326 | POINT
```

Choropleth: aggregate values per polygon



Coverage: “every” point is mapped



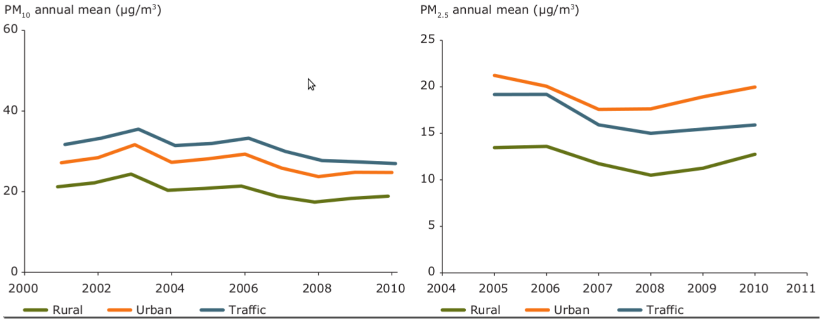
EEA Report | No 4/2012

Air quality in Europe — 2012 report

ISSN 1725-9177

European Environment Agency

Particulate matter time series, averaged over station type



Modelling spatiotemporal information generation

- Scientists create a lot of data, but how do we discover data they created, and how do we advertise data we create ourselves?
- Jim Frew's laws of metadata: (i) scientists don't write metadata, (ii) scientists can be forced to write bad metadata.
- Much of data description focuses *when, where* and *what* questions (semantics), less so on *how* and *why* (pragmatics)
- We developed an algebra for information generation (i.e., the how), using functions composed of reference systems.
- We hope this can help solve the discovery problem.

INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **25**: 1965–1978 (2005)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/joc.1276

VERY HIGH RESOLUTION INTERPOLATED CLIMATE SURFACES FOR GLOBAL LAND AREAS

ROBERT J. HIJMANS,[a,*] SUSAN E. CAMERON,[a,b] JUAN L. PARRA,[a] PETER G. JONES[c] and ANDY JARVIS[c,d]

[a] *Museum of Vertebrate Zoology, University of California, 3101 Valley Life Sciences Building, Berkeley, CA, USA*
[b] *Department of Environmental Science and Policy, University of California, Davis, CA, USA; and Rainforest Cooperative Research Centre, University of Queensland, Australia*
[c] *International Center for Tropical Agriculture, Cali, Colombia*
[d] *International Plant Genetic Resources Institute, Cali, Colombia*

Received 18 November 2004
Revised 25 May 2005
Accepted 6 September 2005

ABSTRACT

We developed interpolated climate surfaces for global land areas (excluding Antarctica) at a spatial resolution of 30 arc s (often referred to as 1-km spatial resolution). The climate elements considered were monthly precipitation and mean, minimum, and maximum temperature. Input data were gathered from a variety of sources and, where possible, were restricted to records from the 1950–2000 period. We used the thin-plate smoothing spline algorithm implemented in the ANUSPLIN package for interpolation, using latitude, longitude, and elevation as independent variables. We quantified uncertainty arising from the input data and the interpolation by mapping weather station density, elevation bias in the weather stations, and elevation variation within grid cells and through data partitioning and cross validation. Elevation bias tended to be negative (stations lower than expected) at high latitudes but positive in the tropics. Uncertainty is highest in mountainous and in poorly sampled areas. Data partitioning showed high uncertainty of the surfaces on isolated islands, e.g. in the Pacific. Aggregating the elevation and climate data to 10 arc min resolution showed an

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2016
http://dx.doi.org/10.1080/13658816.2016.1151520

Taylor & Francis
Taylor & Francis Group

Modeling spatiotemporal information generation

Simon Scheider[a,d], Benedikt Gräler[b], Edzer Pebesma[b] and Christoph Stasch[b,c]

[a]Departement Bau, Umwelt und Geomatik, Institut für Kartographie und Geoinformation, ETH Zürich, Zürich, Switzerland; [b]Fachbereich Geowissenschaften, Institute for Geoinformatics, University of Münster, Münster, Germany; [c]52°North Initiative for Geospatial Open Source Software GmbH, Münster, Germany; [d]Human Geography and Spatial Planning, Universiteit Utrecht, Utrecht, The Netherlands

ABSTRACT

Maintaining knowledge about the provenance of datasets, that is, about how they were obtained, is crucial for their further use. Contrary to what the overused metaphors of 'data mining' and 'big data' are implying, it is hardly possible to use data in a meaningful way if information about sources and types of conversions is discarded in the process of data gathering. A generative model of spatiotemporal information could not only help automating the description of derivation processes but also assessing the scope of a dataset's future use by exploring possible transformations. Even

ARTICLE HISTORY
Received 1 September 2015
Accepted 2 February 2016

KEYWORDS
Spatiotemporal data types; data generation; provenance model; algebra

Basic types

Basic reference system types and simple derivations thereof. Each type needs to go along with its reference system (RS).

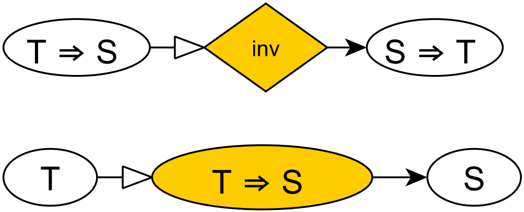
$\mathcal{P}$ denotes the power set (set of all subsets).

Symbol	Definition	Meaning	Description
S		$\mathbb{R}^3$	Set of possible spatial locations with RS.
T		$\mathbb{R}$	Set of possible moments in time with RS.
D		$\mathbb{N}$	Set of possible discrete entity identifier with RS.
Q		$\mathbb{R}$	Set of possible observed values with RS.
R	S set	$\mathcal{P}(S)$	Set of regions: bounded by polygons, or collection of isolated locations and combinations thereof.
I	T set	$\mathcal{P}(T)$	Set of collections of moments in time: continuous intervals or a set of moments in time or combinations thereof.
D set	D set	$\mathcal{P}(D)$	Sets of object identifiers
Q set	Q set	$\mathcal{P}(Q)$	Sets of quality values.
bool		$\{\mathrm{T}, \mathrm{F}\}$	Boolean, also used to express predicates for selection
Extent	$R \times I$	$R \times I$	set of spatio-temporal extent as the orthogonal product of the spatial and temporal projections
Occurs	$(S \times T)$ set	$\mathcal{P}(S \times T)$	set of spatio-temporal subsets, occurrences of events and objects, but also of certain values or conditions in a field; footprint, support

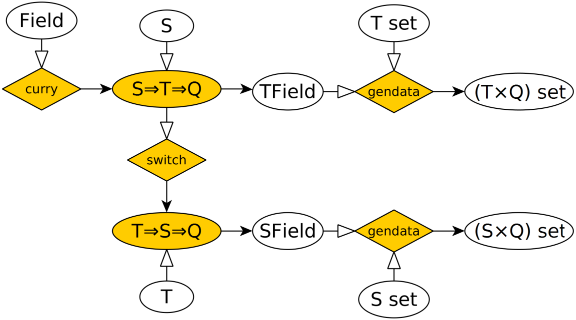
Data Generation Types

Symbol	Type definition	Description
Field	$S \times T \Rightarrow Q$	spatio-temporal field
Lattice	$R \Rightarrow I \Rightarrow Q$	spatio-temporal lattice
Event	$D \Rightarrow S \times T$	spatio-temporal events
Trajectory	$T \Rightarrow S$	trajectory
Objects	$D \Rightarrow T \Rightarrow S$	objects in time and space
LatticeT	$S \Rightarrow I \Rightarrow Q$	spatial temporal lattice
BlockEvent	$D \Rightarrow \text{Extent}$	events affecting a set of locations and lasting for son
RegionalTrajectory	$T \Rightarrow R$	trajectory of regions
BlockObjects	$D \Rightarrow I \Rightarrow R$	objects in space and time defined over regions and c

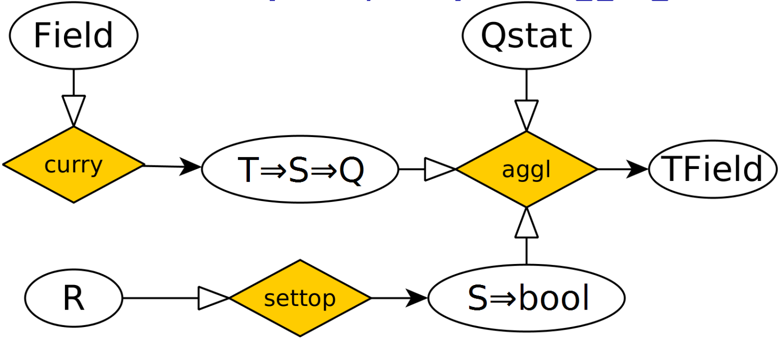
Data derivation

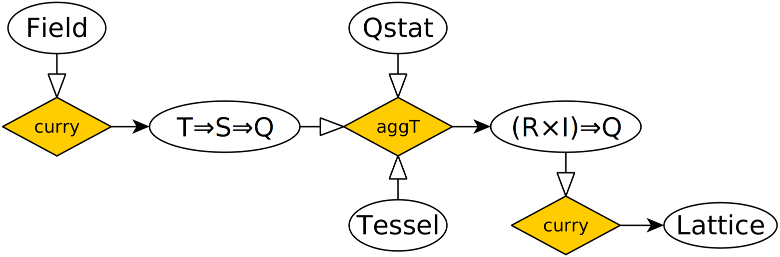


Data derivation: generating field data

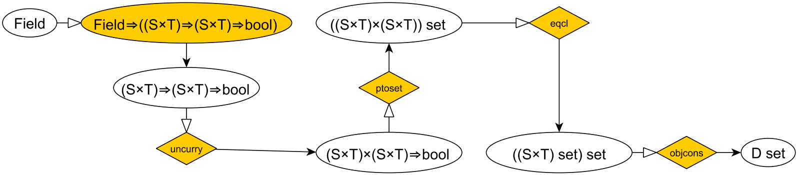


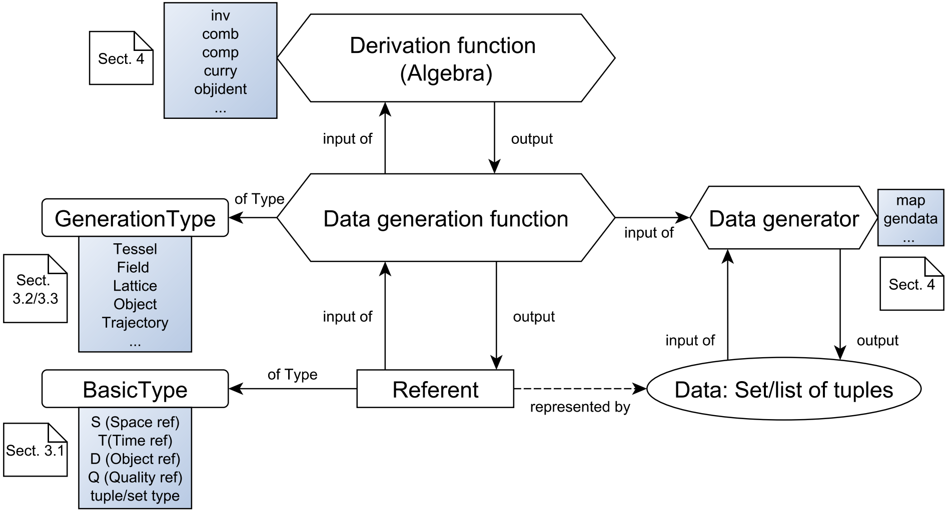
Data derivation: spatial/temporal aggregation





Data derivation: deriving objects from fields





Sect. 4
inv
comb
comp
curry
objident
...
Derivation function
(Algebra)
input of
output
GenerationType
of Type
Data generation function
input of
Data generator
map
gendata
...
Tessel
Field
Lattice
Object
Trajectory
...
Sect.
3.2/3.3
Sect. 4
input of
output
input of
output
BasicType
of Type
Referent
represented by
Data: Set/list of tuples
S (Space ref)
T(Time ref)
D (Object ref)
Q (Quality ref)
tuple/set type
Sect. 3.1

How smart is R?

- R does have `factor` and `ordered` for nominal and ordinal variables, but does not support interval or ratio variables.
- R has no support for measurement units.
- R aggregate functions can't check whether its variable is extensive (sum) or intensive (mean)
- R supports time (`Date`, `POSIXt` and time zones), but not time intervals
- Package `lubridate` does this, but does not support time series data, similar to `zoo` or `xts` do.
- `spacetime` compensates (somewhat) for this
- `sp` and `rgdal` support coordinate reference systems, interoperably
- `zoo`, `sp`, `spacetime` let you aggregate data over time and/or space, but do not annotate returned objects that they are the result of aggregation.

Software challenges (Discussion/Conclusions)

How smart should software be?

- COS is everywhere, but it's not registered with our data.
- How can I find datasets generated using procedure y?
- Which analysis could I apply to dataset x, or avoid?
- R scripts convey syntax and numerical manipulation, only implicit semantics
- Many R functions could trivially annotate returned objects with meaningful bits
- Instead of points/lines/grids/polygons, we need field/lattice/event/trajectory/object
- For meaningful discovery, R should (optionally and automatically) write metadata describing data provenance
- We will next try to implement some of the concepts mentioned above in R (possibly using CXXR, Silles and Runnalls)

Thank you!

Software challenges (Discussion/Conclusions)

How smart should software be?

- COS is everywhere, but it's not registered with our data.
- How can I find datasets generated using procedure y?
- Which analysis could I apply to dataset x, or avoid?
- R scripts convey syntax and numerical manipulation, only implicit semantics
- Many R functions could trivially annotate returned objects with meaningful bits
- Instead of points/lines/grids/polygons, we need field/lattice/event/trajectory/object
- For meaningful discovery, R should (optionally and automatically) write metadata describing data provenance
- We will next try to implement some of the concepts mentioned above in R (possibly using CXXR, Silles and Runnalls)

Thank you!

References

- S. Scheider, B. Gräler, E. Pebesma, C. Stasch, 2016. Modelling spatio-temporal information generation. Int J of Geographic Information Science, published online (pdf).
- Stasch, C., S. Scheider, E. Pebesma, W. Kuhn, 2014. Meaningful Spatial Prediction and Aggregation. Environmental Modelling & Software, 51, (149–165, open access).
- Sinton, David. "The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study." Harvard papers on geographic information systems 6 (1978): 1-17.
- Ferreira, Karine Reis, Gilberto Camara, and Antônio Miguel Vieira Monteiro. "An algebra for spatiotemporal data: From observations to events." Transactions in GIS 18.2 (2014): 253-269.
- Camara, Gilberto, et al. "Fields as a generic data type for big spatial data." Geographic Information Science. Springer International Publishing, 2014. 159-172.
- Goodchild, Michael F., May Yuan, and Thomas J. Cova. "Towards a general theory of geographic representation in GIS." International journal of geographical information science 21.3 (2007): 239-260.
- Galton, Antony. "Fields and objects in space, time, and space-time." Spatial cognition and computation 4.1 (2004): 39-68.
- Cressie, N. Statistics for Spatial Data. Wiley, 1993.