R Spatial Concerns

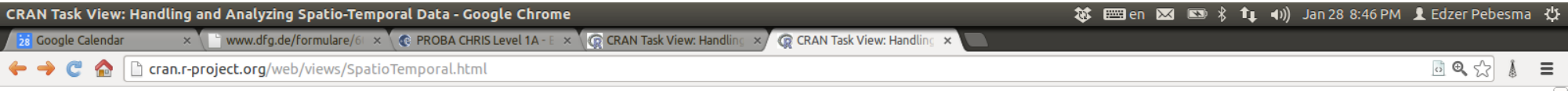
Edzer Pebesma

ifgi
Institute for Geoinformatics
University of Münster

Bay Area R User Group meeting, Dec 17, 2014

Overview

- what is spatial/spatio-temporal statistics concerned about?
- which challenges are we working on?
- why do we do this?



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.

Environmental Modelling & Software 51 (2014) 149–165

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

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, Heisenbergstr. 2, 48149 Muenster, Germany
[b] 52 North Initiative for Geospatial Open Source Software GmbH, Martin-Luther-King-Weg 24, 48151 Muenster, Germany

ARTICLE INFO

Article history:
Received 23 December 2012
Received in revised form
16 September 2013
Accepted 16 September 2013
Available online 22 October 2013

Keywords:
Meaningfulness
Knowledge-based environmental modelling

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

Spatial and Spatio-Temporal Statistics

Space and time are different from other feature dimensions because

- they "span" the world we live in
- prediction in space & forecasting in time are economic activities
- real-world processes are usually correlated in space and/or time
- (random sampling might take care of this, but then disables prediction and forecasting)

Spatial statistics: usual book chapters

1. Geostatistics
2. Point patterns
3. Areal (lattice) data

But, given a data set, which chapter should we start?

Data sets usually come from files...

... with the metadata in the file name:

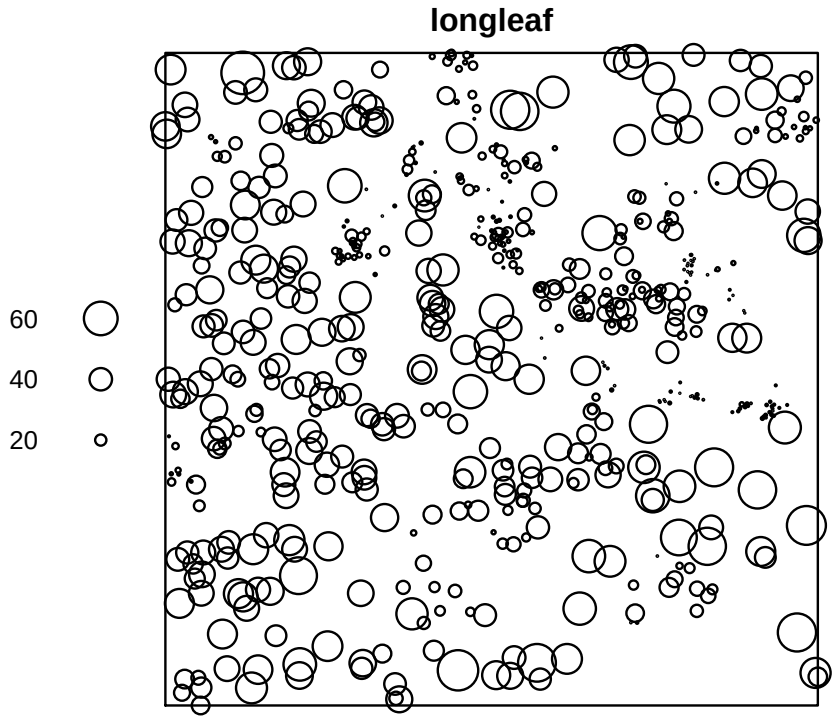
```
> co2 = read.csv("co2_emission_powerplants.csv")
> co2[1:5, c("longitude", "latitude", "carbon_2007")]

  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

> cc = c("factor", "Date", "Date", "factor", "factor",
+         "numeric", "numeric", "numeric")
> pm10.tab <- read.table("EU_meas_2005_june.dat", header = TRUE, colClasses = cc)
> pm10.tab[1:5, c("x", "y", "time", "PM10")]

         x        y       time PM10
1 13.67111 48.39167 2005-06-01 14.0
2 15.91945 48.10611 2005-06-01  9.7
3 14.57473 48.53111 2005-06-01  7.8
4 15.04668 48.87861 2005-06-01 21.9
5 16.43333 48.08611 2005-06-01 11.2
```

similar, but do they afford the same analyses?

Point patterns: data on longleaf pines

from `?longleaf`: *This is a marked point pattern; the mark associated with a tree is its diameter at breast height ('dbh'), a convenient measure of its size.*[1]

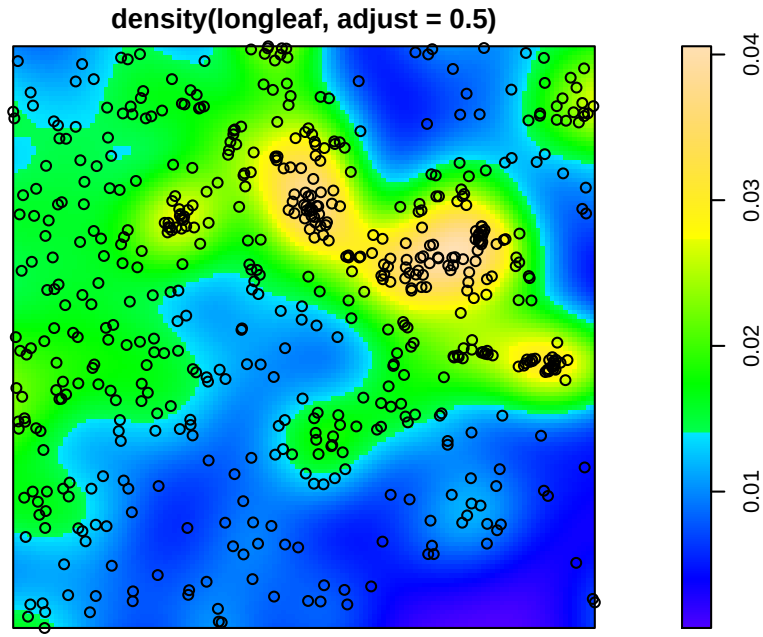
```
> library(spatstat)
> data(longleaf)
> plot(longleaf)
```



[1] it doesn't actually *say* that *dbh* is in cm

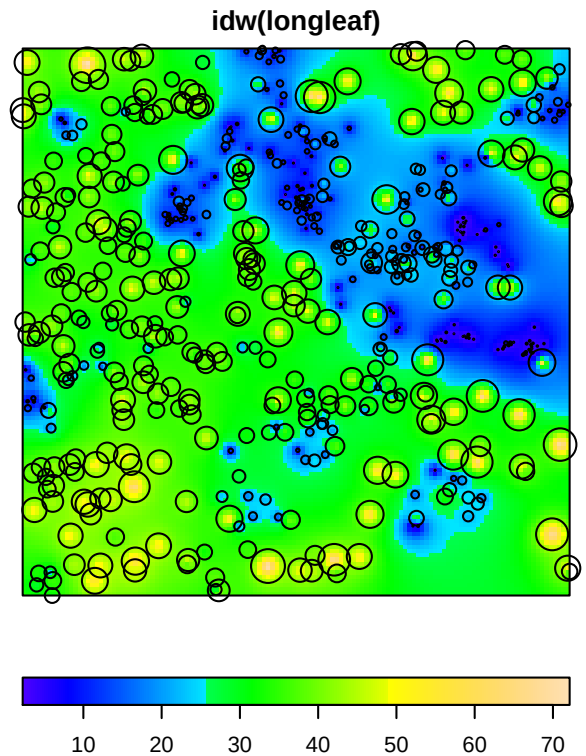
Point patterns: densities

```
> library(spatstat)
> data(longleaf)
> plot(density(longleaf, adjust = 0.5))
> points(longleaf)
```

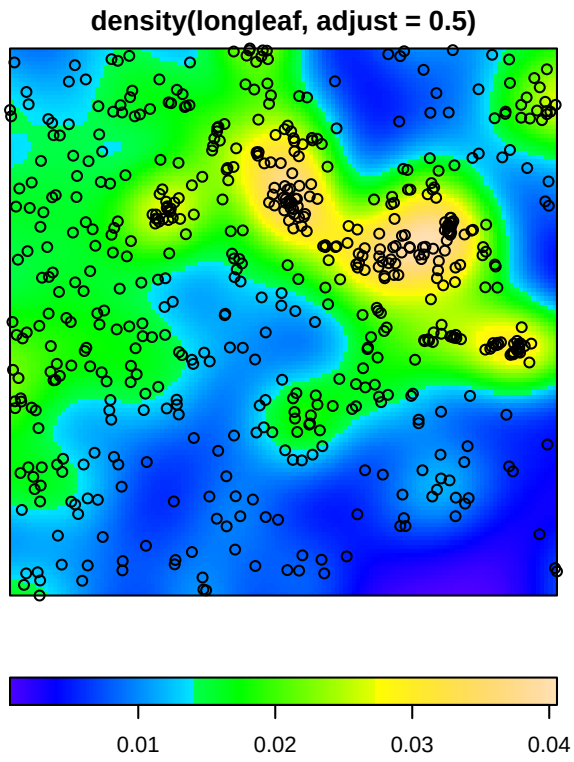


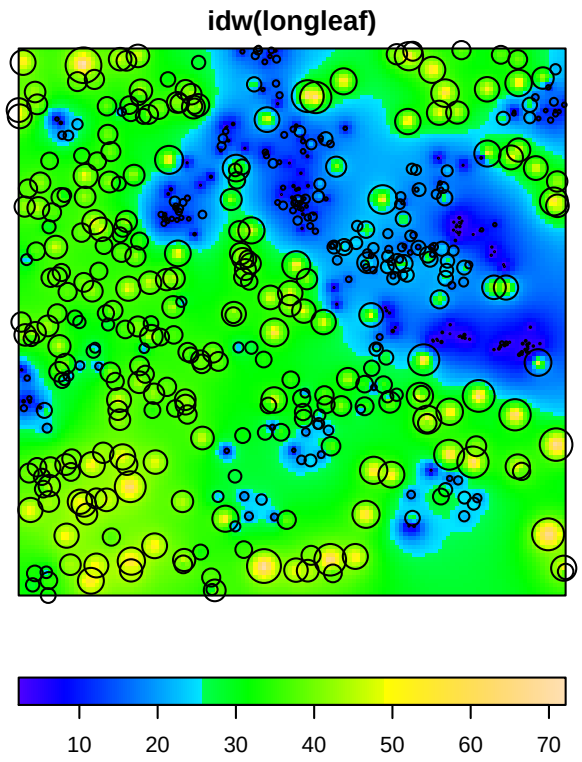
Spatial points: interpolation

```
> plot(idw(longleaf), ribside="bottom")
> plot(longleaf, add = TRUE)
```

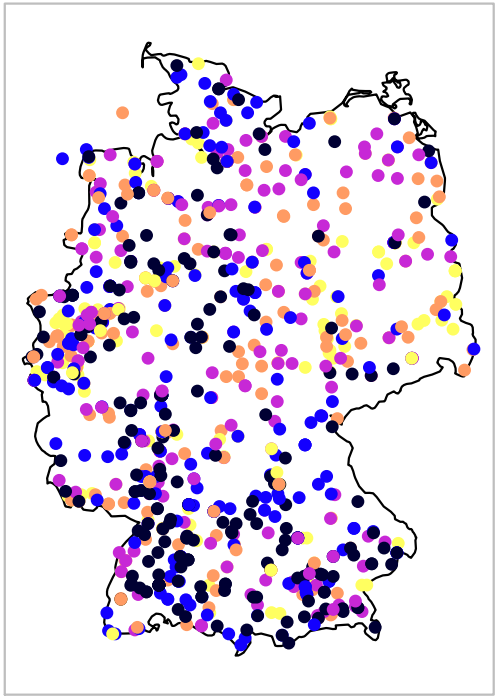


Densities vs. interpolation



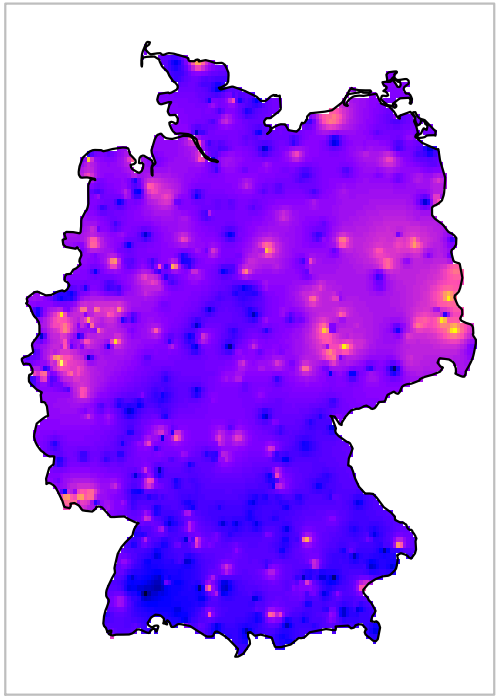


CO_2 emissions of power plants

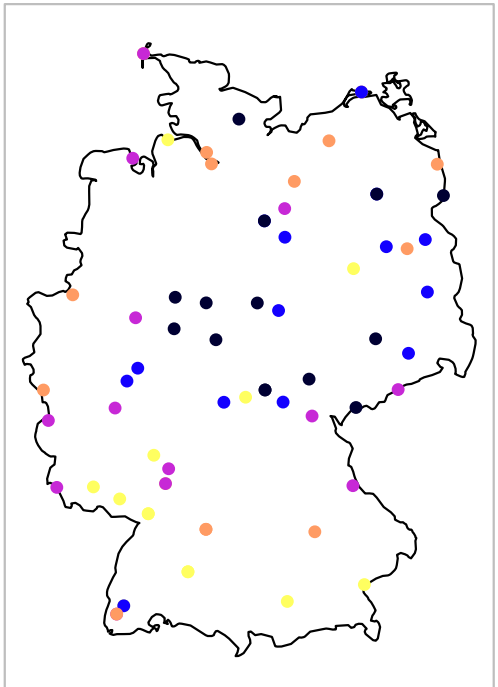


Sum of CO_2 emissions

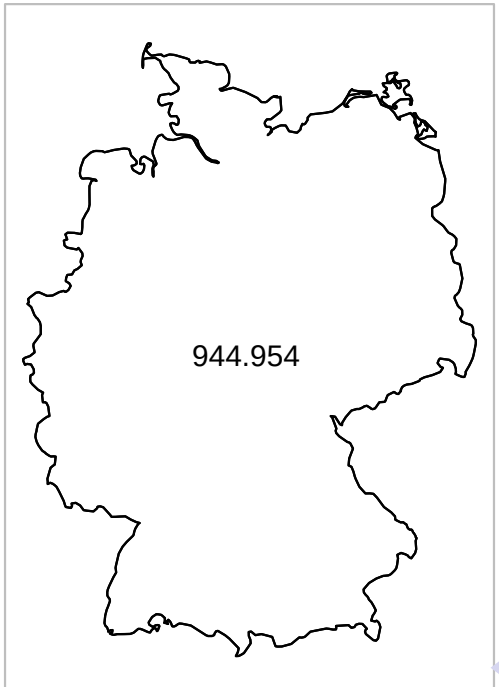
Interpolated CO_2 emissions



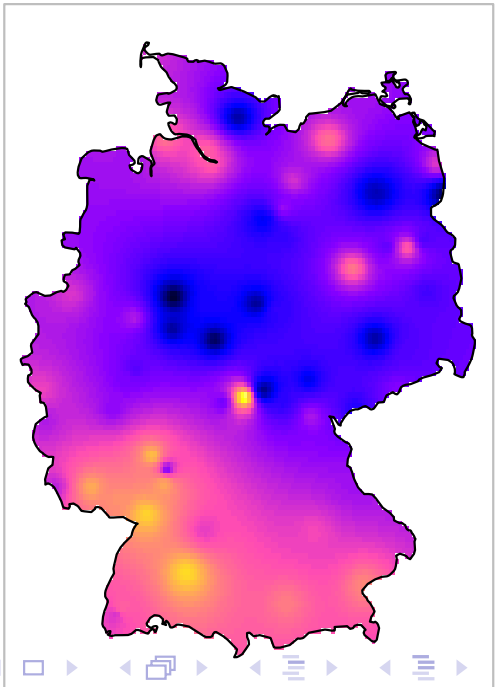
PM_{10} measurements



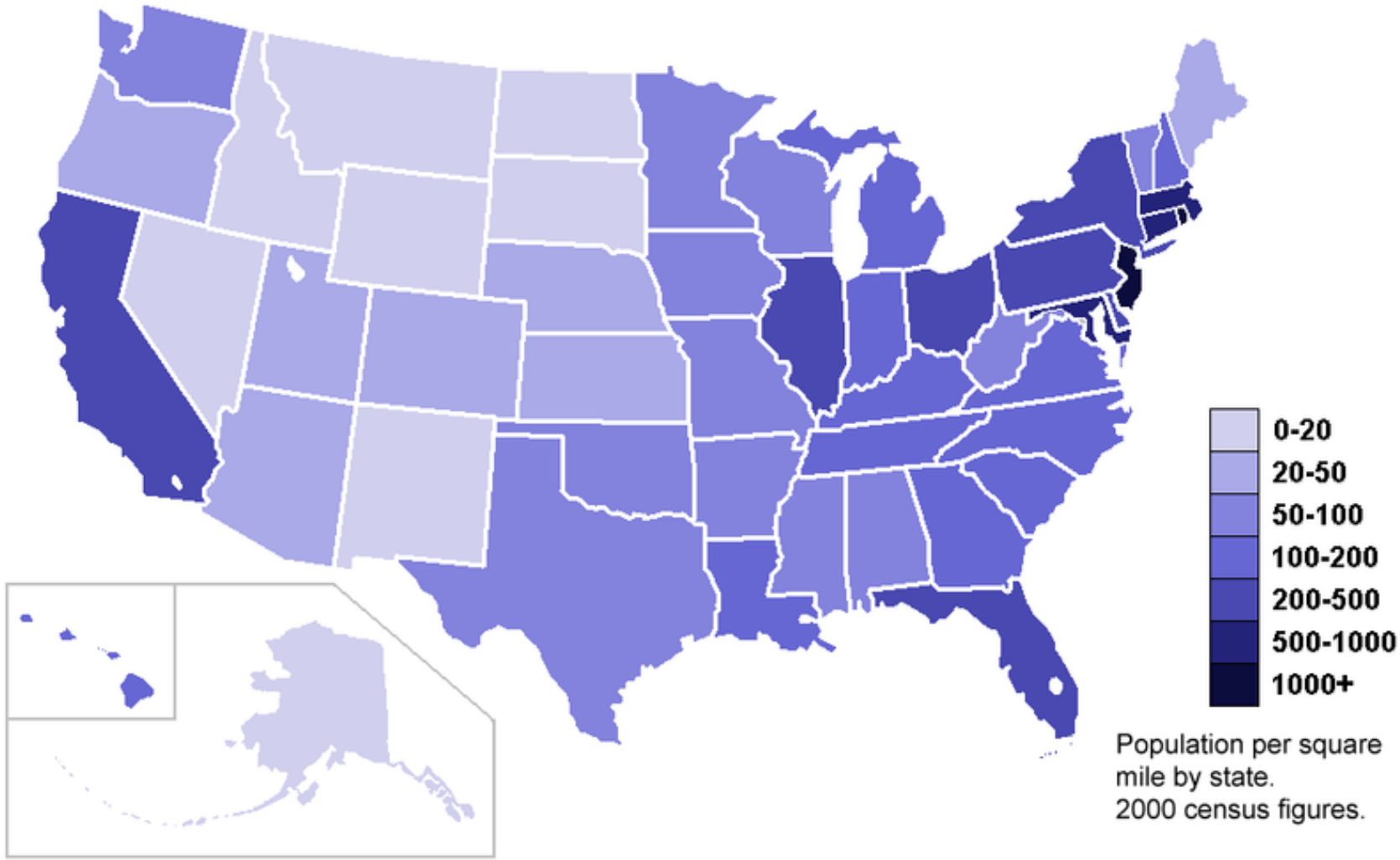
Sum of PM_{10} measurements



Interpolated PM_{10} measurements



Choropleth: aggregate values per polygon



Coverage: “every” point is mapped

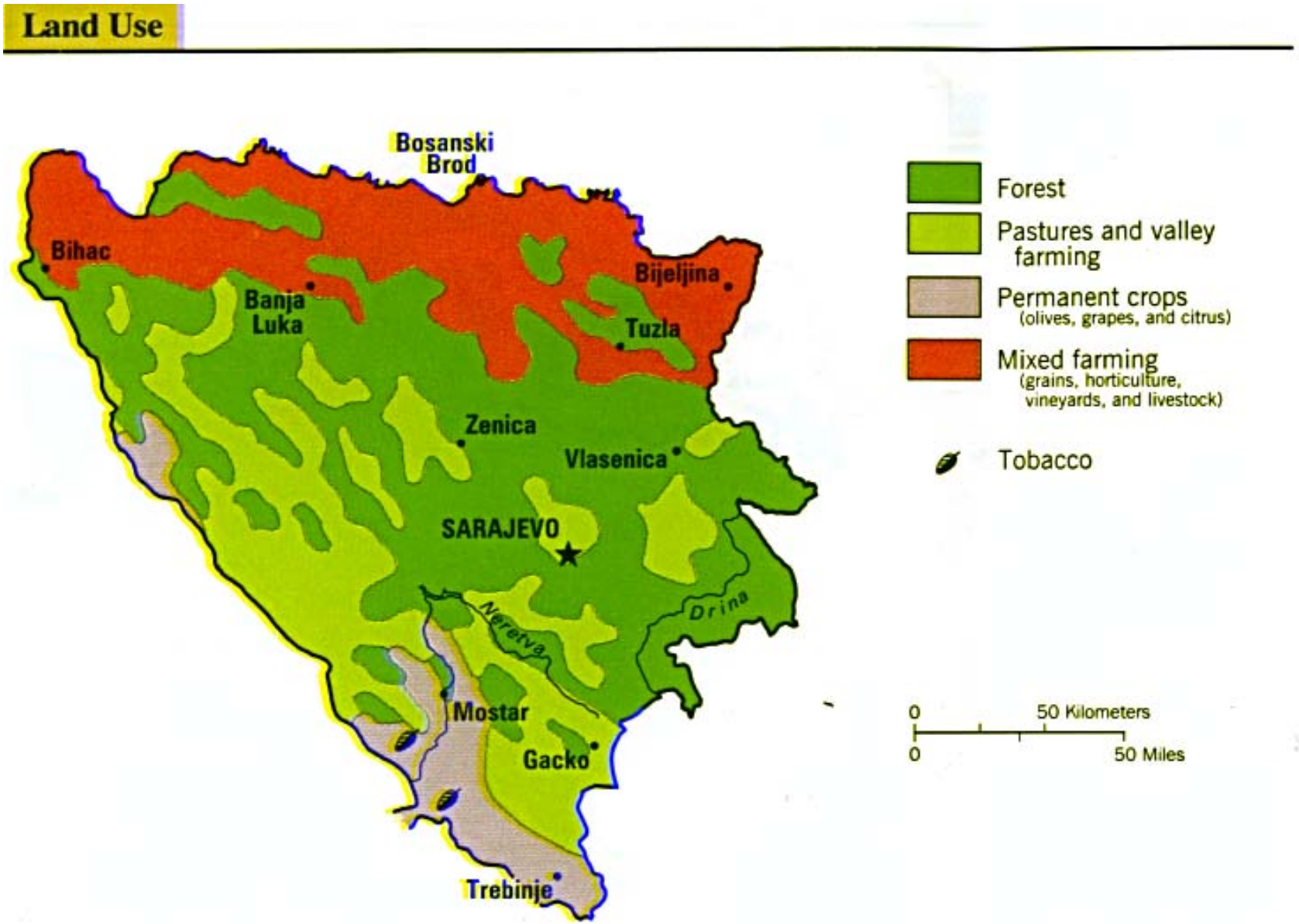
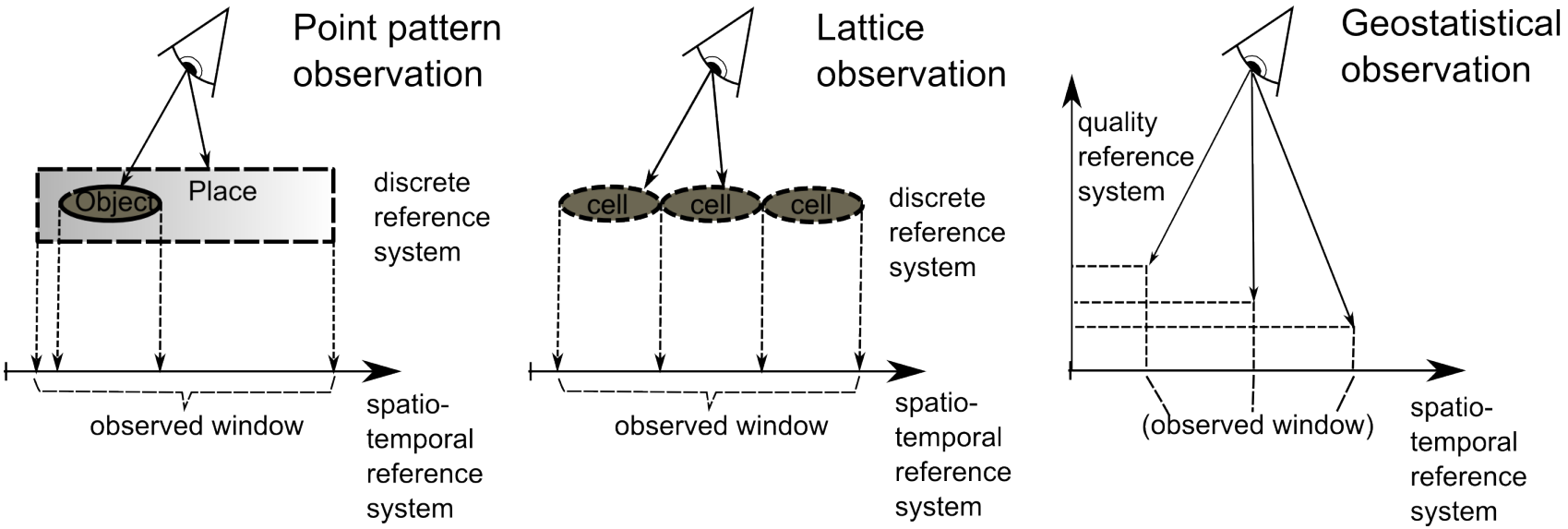


Image: colors, still with no meaning

Observed windows



Meaningfulness

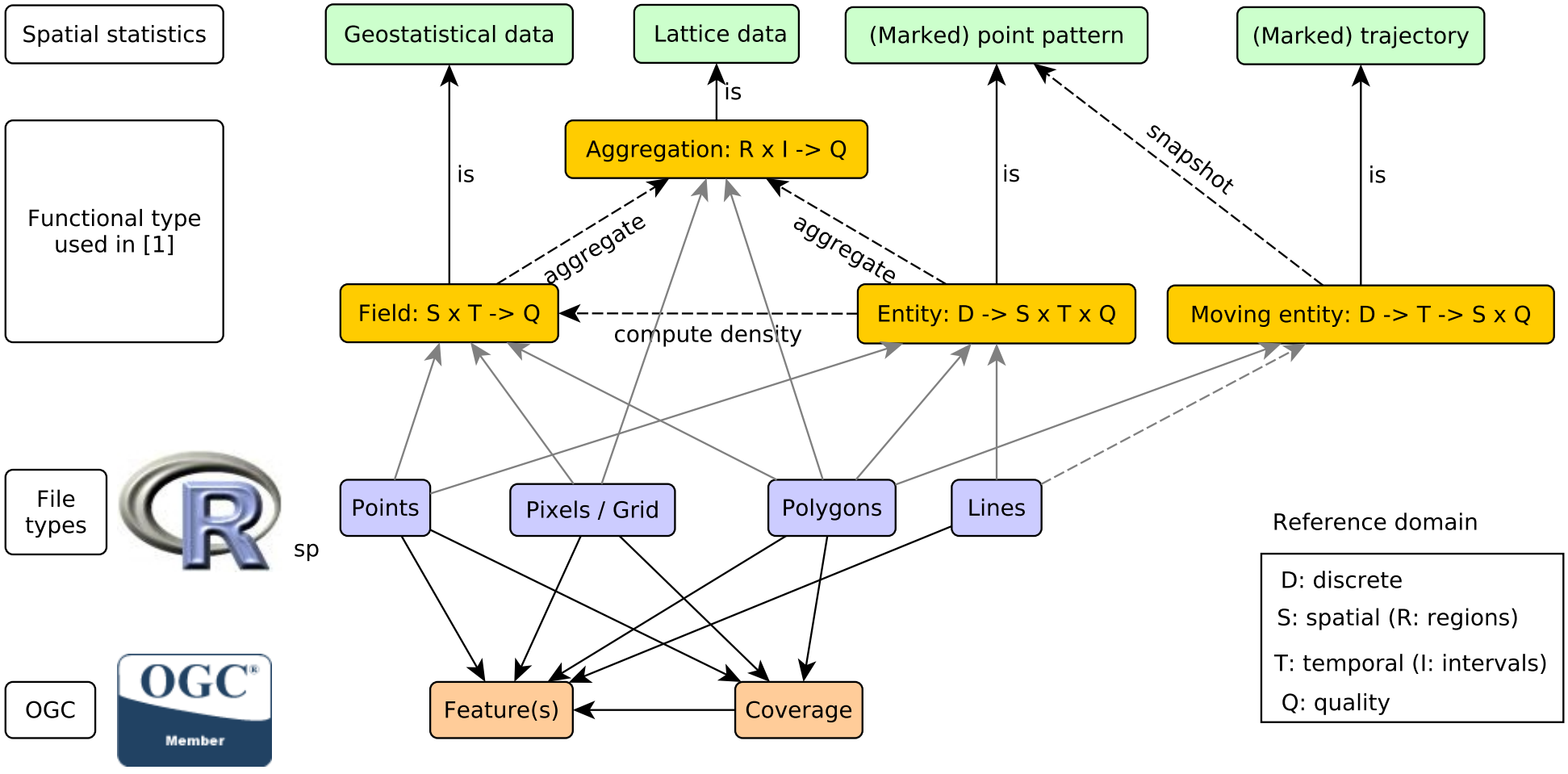
Meaningfulness checks are implemented in our formalism as correspondence checks:

- 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.

To check correspondence, we need types.
We build types from reference system domains.

Types of Reference System Domains.

Reference Domain		Description	Example
Spatial	S	All possible locations that are defined in a spatial reference system	$([-90, 90] \times [-180, 180]) \subset \mathbb{R}^2$ defined in WGS84
Temporal	T	All possible times defined in a temporal reference system	POSIX time (seconds since Jan 1, 1970 UTC) with $T \subset \mathbb{Q}$
Quality	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)
Discrete Entities	D	Set of discrete objects or events.	Set of coal power plants in Germany in 2007

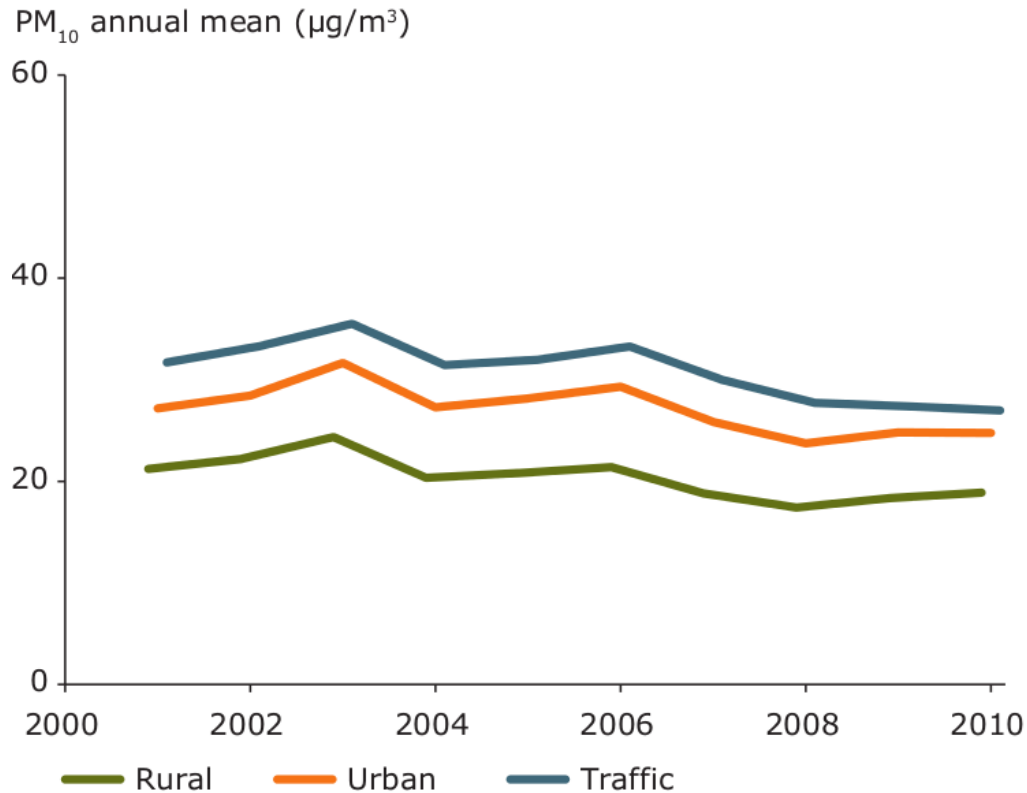

Spatial statistics
Functional type used in [1]
File types
sp
OGC
Member
Geostatistical data
Lattice data
(Marked) point pattern
(Marked) trajectory
is
Aggregation: R x I -> Q
aggregate
snapshot
Field: S x T -> Q
compute density
Entity: D -> S x T x Q
Moving entity: D -> T -> S x Q
Points
Pixels / Grid
Polygons
Lines
Feature(s)
Coverage
Reference domain
D: discrete
S: spatial (R: regions)
T: temporal (I: intervals)
Q: quality

Air quality in Europe — 2012 report

ISSN 1725-9177

European Environment Agency

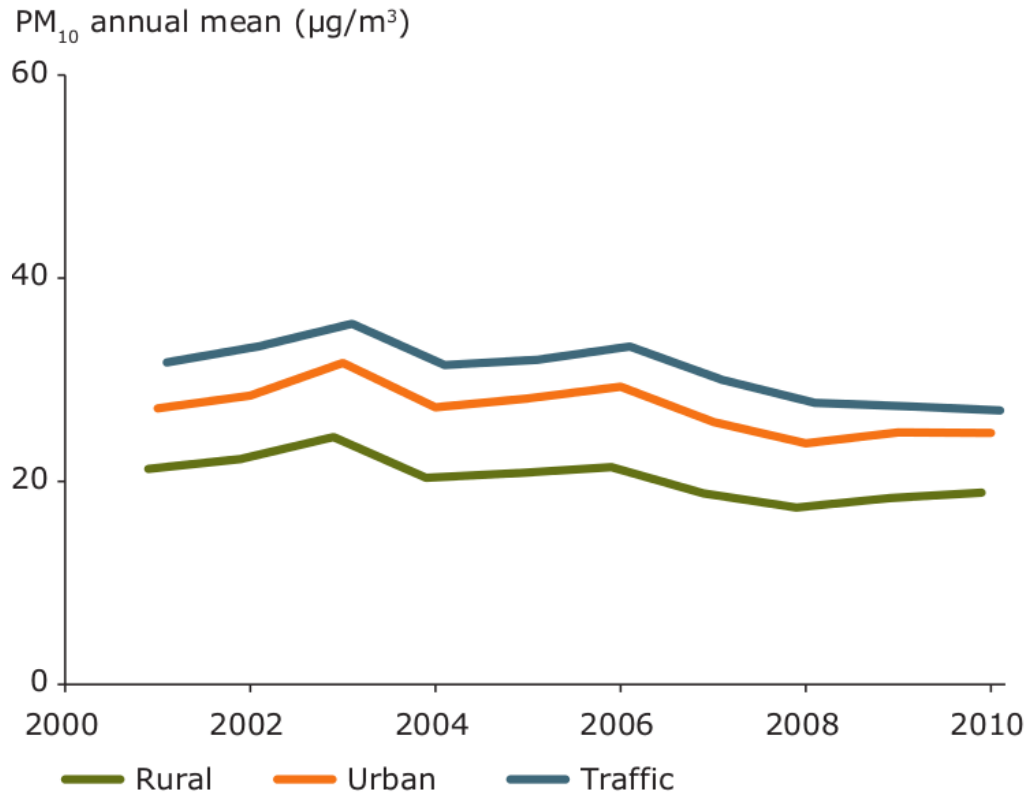
Air quality in Europe: EEA report 4/2012



"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



"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)

R-spatial: Further challenges

- Space-Time, integrated, non-conditional analysis
- Spatial Data+: multivariate, functional data, integrated analysis
- Meaningful spatial statistics, semantic web
- Working directly on/in (relational or array) data bases (`rgdal2`, `SciDB-R`)

The bigger picture: global change

- if we want to reach concensus on global change and its causes, we need to open up our workflows
- R is a language for sharing data analysis workflows that is taught in universities
- our data (EO: observed, GCMs: modelled) has become too large to download
- there is a need for open, scalable and shared compute infrastructure

Why does Google Earth Engine work, but do GEOSS[2], and maybe EarthCube, not work?

[2] global earth observation system of systems

Conclusions and final remarks

1. working from files, spatial statistics can easily lead to non-meaningful analyses; our current data infrastructure does not deal well with this (e.g., observation window)
2. in analogy with Stevens' (1946) measurement scales, we propose types, built from reference domains (S, T, D, Q)
3. with increasing inter-disciplinarity, the risk of non-meaningful analysis increases
4. studying global change asks for inter-disciplinarity, and sharing of open workflows
5. with Today's data sizes, we need shared processing facilities ("bring user to the data")
6. R could be a key component in this, because it works, and is widely understood.

Thank you!

Conclusions and final remarks

1. working from files, spatial statistics can easily lead to non-meaningful analyses; our current data infrastructure does not deal well with this (e.g., observation window)
2. in analogy with Stevens' (1946) measurement scales, we propose types, built from reference domains (S, T, D, Q)
3. with increasing inter-disciplinarity, the risk of non-meaningful analysis increases
4. studying global change asks for inter-disciplinarity, and sharing of open workflows
5. with Today's data sizes, we need shared processing facilities ("bring user to the data")
6. R could be a key component in this, because it works, and is widely understood.

Thank you!

Conclusions and final remarks

1. working from files, spatial statistics can easily lead to non-meaningful analyses; our current data infrastructure does not deal well with this (e.g., observation window)
2. in analogy with Stevens' (1946) measurement scales, we propose types, built from reference domains (S, T, D, Q)
3. with increasing inter-disciplinarity, the risk of non-meaningful analysis increases
4. studying global change asks for inter-disciplinarity, and sharing of open workflows
5. with Today's data sizes, we need shared processing facilities ("bring user to the data")
6. R could be a key component in this, because it works, and is widely understood.

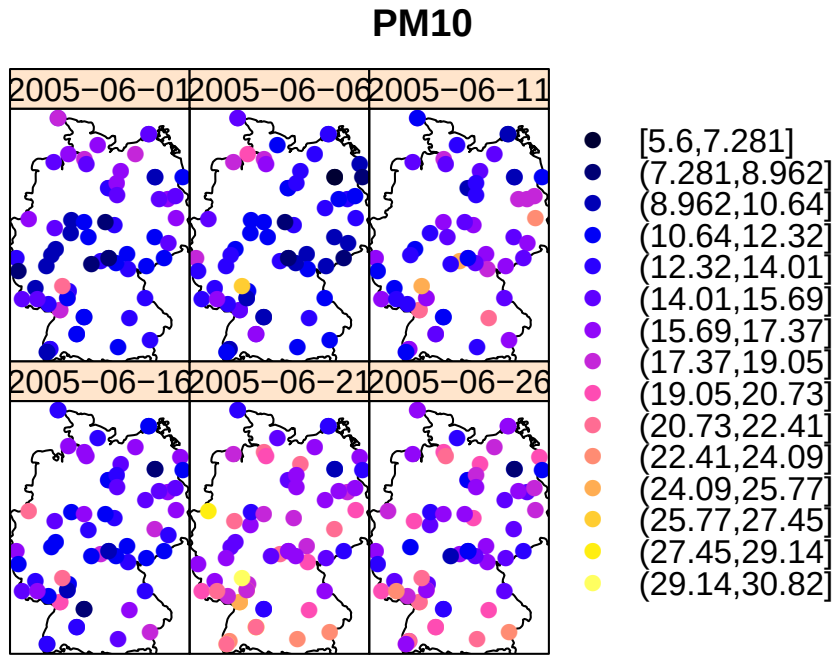
Thank you!

The limits of classes

- how to deal with hybrid classes (space-field, time-point pattern OR space-point pattern, time-field)?
- 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 marks tell us?

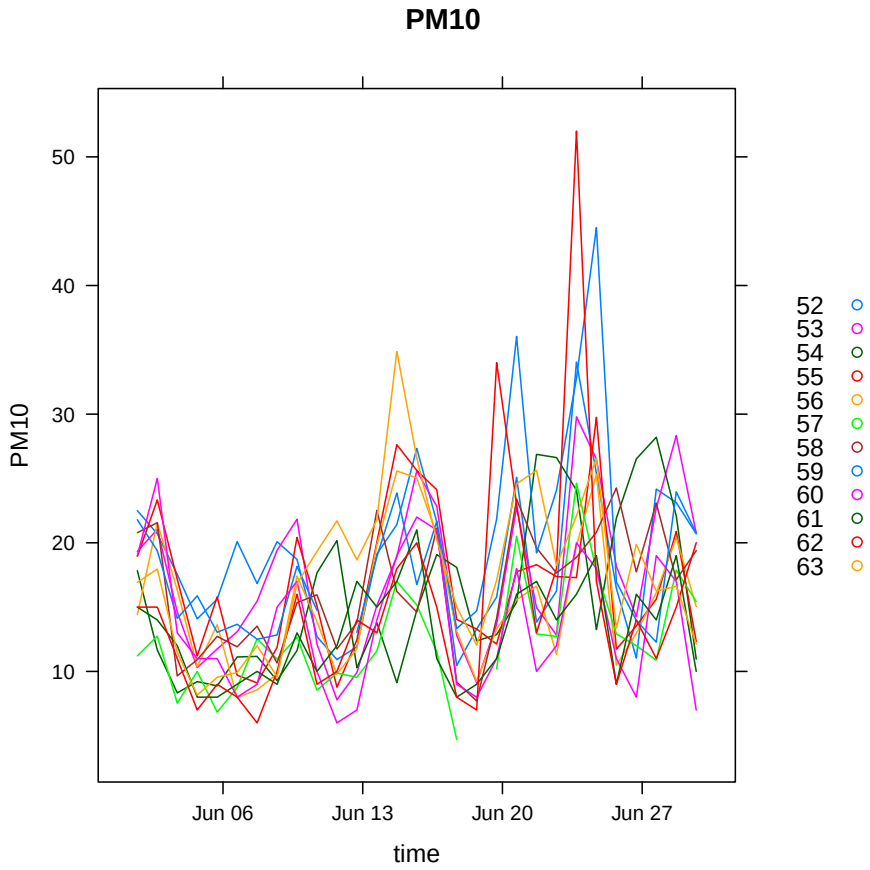
Space-time syntax and plots

```
> library(sp)
> library(spacetime)
> WGS84 = CRS("+proj=longlat +ellps=WGS84")
> pm10all = STIDF(
+  SpatialPoints(pm10.tab[c("x", "y")], WGS84),
+  as.POSIXct(pm10.tab$time),
+  pm10.tab)
> pm10all = as(pm10all, "STFDF")
> load("germany.rda")
> class(germany)

[1] "SpatialPolygonsDataFrame"
attr(,"package")
[1] "sp"

> pm10sel <- pm10all[germany, "2005-06", "PM10"]
> stplot(aggregate(pm10sel, "5 days", mean),
+  sp.layout=list("sp.polygons", germany))
```

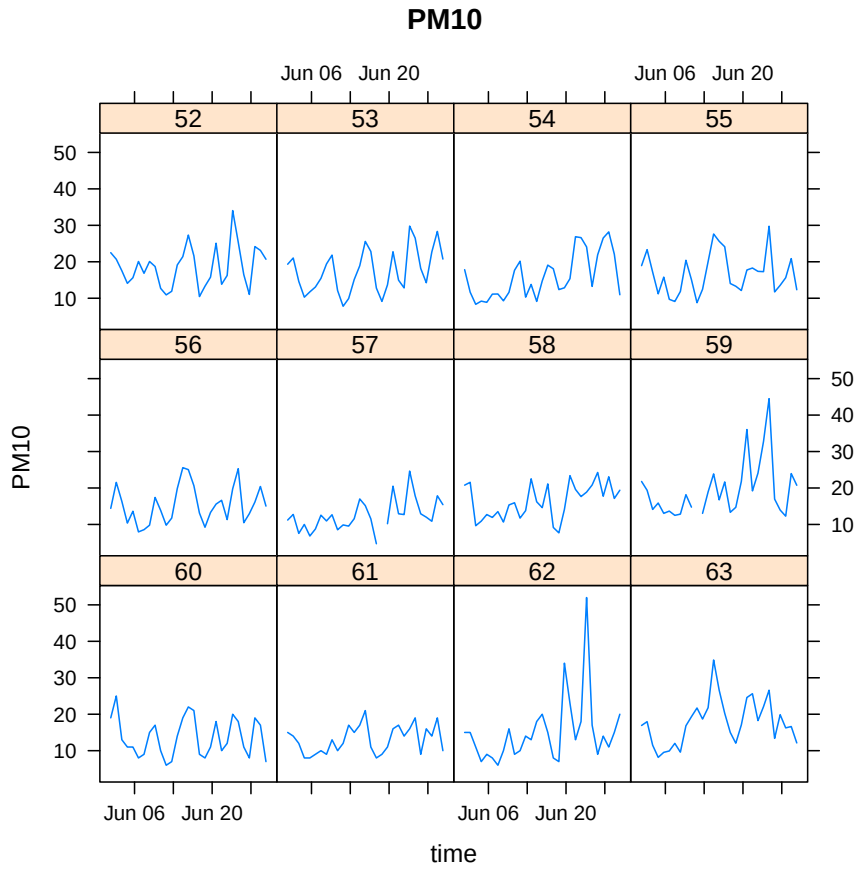


Space-time syntax and plots

```
> stplot(pm10sel[1:12, ], mode = "ts")
```

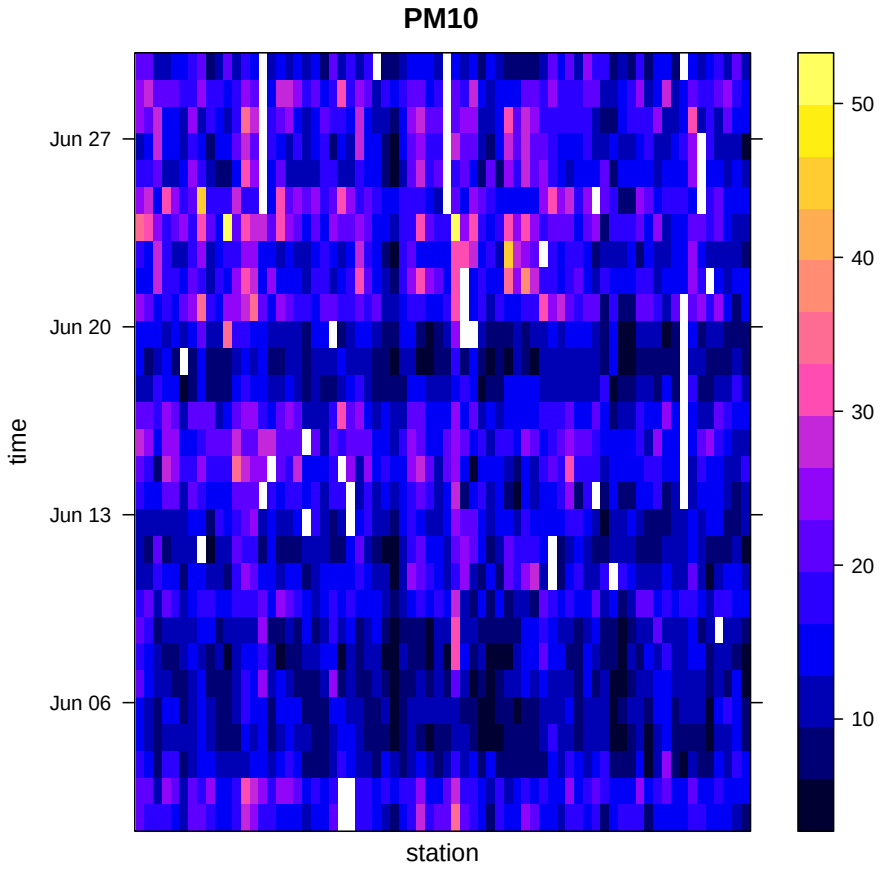



```
> stplot(pm10sel[1:12, ], mode = "tp")
```



Space-time syntax and plots

```
> bpy = bpy.colors()
> stplot(pm10sel, mode = "xt", col.regions = bpy,
+   xlab = "station",
+   scales = list(x = list(draw = FALSE)))
```




```
> m = aggregate(pm10sel, "time", mean, na.rm = TRUE)
> o = order(m[[1]])
> stplot(pm10sel[o,], mode = "xt", col.regions = bpy,
+   xlab = "station",
+   scales = list(x = list(draw = FALSE)))
```