Statistics for spatio-temporal data: an introduction

Edzer Pebesma

ifgi
Institute for Geoinformatics
University of Münster

Geostat Summer School, Bergen, 15-21 Jun 2014

SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

FOR SEVEN YEARS A COMMITTEE of the British Association for the Advancement of Science debated the problem of measurement. Appointed in 1932 to represent Section A (Mathematical and Physical Sciences) and Section J (Psychology), the committee was instructed to consider and report upon the possibility of "quantitative estimates of sensory events"—meaning simply: Is it possible to measure human sensation? Deliberation led only to disagreement, mainly about what is meant by the term measurement. An interim report in 1938 found one member complaining that his colleagues "came out by that same door as they went in," and in order to have another try at agreement, the committee begged to be continued for another year.

For its final report (1940) the committee chose a by the formal (mathematical) properties of the scales. Furthermore—and this is of great concern to several of the sciences—the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered.

A Classification of Scales of Measurement

Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement. The problem then becomes that of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties

J. R. Statist. Soc. A (1996)
159, *Part* 3, *pp.* 445–492

Statistics and the Theory of Measurement

By D. J. HAND†

The Open University, Milton Keynes, UK

[*Read before* The Royal Statistical Society *on Wednesday, March 20th, 1996, the President,* Professor A. F. M. Smith, *in the Chair*]

SUMMARY

Just as there are different interpretations of probability, leading to different kinds of inferential statements and different conclusions about statistical models and questions, so there are different theories of measurement, which in turn may lead to different kinds of statistical model and possibly different conclusions. This has led to much confusion and a long running debate about when different classes of statistical methods may legitimately be applied. This paper outlines the major theories of measurement and their relationships and describes the different kinds of models and hypotheses which may be formulated within each theory. One general conclusion is that the domains of applicability of the two major theories are typically different, and it is this which helps apparent contradictions to be avoided in most practical applications.

Keywords: CLASSICAL MEASUREMENT; MEASUREMENT THEORY; OPERATIONAL MEASUREMENT; REPRESENTATIONAL MEASUREMENT; STATISTICAL MODELS;

Beyond Stevens: A revised approach to measurement for geographic information

Nicholas R. Chrisman
CHRISMAN@u.washington.edu
Department of Geography DP 10, University of Washington
Seattle, Washington 98195 USA

ABSTRACT

Measurement is commonly divided into nominal, ordinal, interval and ratio 'scales' in both geography and cartography. These scales have been accepted unquestioned from research in psychology that had a particular scientific agenda. These four scales do not cover all the kinds of measurements common in a geographic information system. The idea of a simple list of measurement scales may not serve the purpose of prescribing appropriate techniques. Informed use of tools does not depend on the nature of the numbers, but of the whole 'measurement framework', the system of objects, relationships and axioms implied by a given system of representation.

Introduction

The approach to measurement in certain social sciences is still strongly

INT. J. GEOGRAPHICAL INFORMATION SCIENCE,
VOL. 17, NO. 5, JULY–AUGUST 2003, 405–409

Taylor & Francis
Taylor & Francis Group

Guest Editorial

Semantic reference systems

WERNER KUHN
Institute for Geoinformatics, University of Münster, Robert-Koch-Str 26–28, 48149 Münster, Germany
e-mail: kuhn@ifgi.uni-muenster.de

(*Received 12 November 2002; accepted 5 February 2003*)

Four centuries after René Descartes watched a fly walk across his ceiling and wondered how to capture its position (Gribbin 2002), we use Cartesian coordinates routinely to describe locations. We identify the positions of entities in the real world, transform their GIS representations from one coordinate system to another, and integrate spatially referenced data across multiple coordinate systems. A theory of *spatial reference systems* standardises the notions of geodetic datum, map projections, and coordinate transformations (ISO 2002). Similarly, our temporal data refer unambiguously to temporal reference systems, such as calendars, and can be transformed

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
Spatial Statistics

ABSTRACT

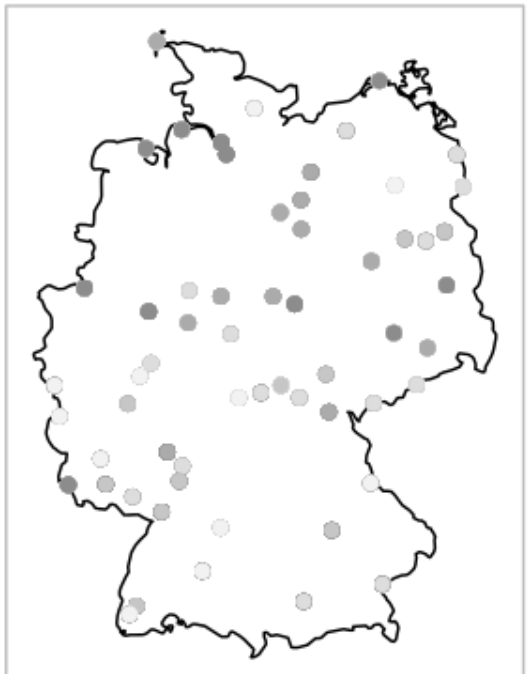
The appropriateness of spatial prediction methods such as Kriging, or aggregation methods such as summing observation values over an area, is currently judged by domain experts using their knowledge and expertise. In order to provide support from information systems for automatically discouraging or 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*

CO_2 emissions of power plants

Sum of CO_2 emissions

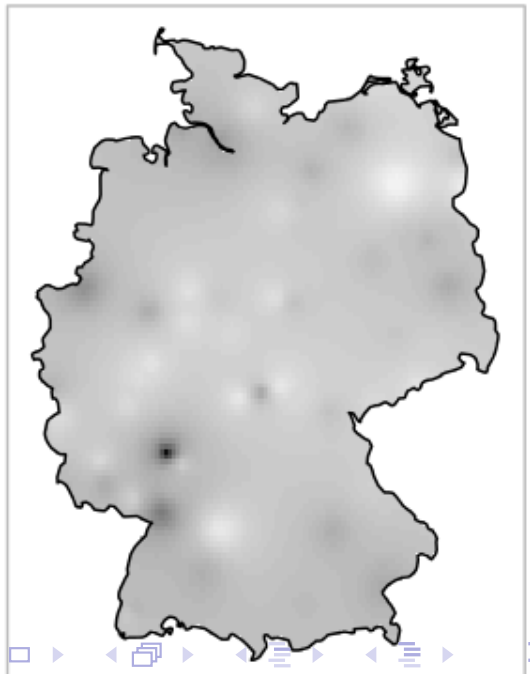
Interpolated CO_2 emissions

PM_{10} measurements



Sum of PM_{10} measurements

Interpolated PM_{10} measurements



All data are spatio-temporal

1. There are no pure-spatial data. Maps reflect either
   - a snapshot in time (remote sensing image)
   - an aggregate over a time period (e.g., interpolated *yearly average* temperature, or yearly aggregated daily interpolations)
   - something that is constant over a period of time (political boundary)
   - a seemingly non-changing phenomenon (geology)
2. There are no pure-temporal data. Time series reflect either
   - spatially aggregated values (global temperature curves)
   - a single spatial location (air quality sensor `DEUB032`, at 8.191934E,50.93033N)
   - vaguely located, or universal aggregates (world market prices, stock quotes)

Functions

We can write function $y = f(x)$ as:

$$f : X \to Y$$

which means that for *any* X, we have a corresponding Y.

$$X \times Y$$

is the *Carthesian product*, the collection of all ordered pairs (x, y)
(Wikipedia): "A function f from X to Y is a subset of the Cartesian product $X \times Y$ subject to the following condition: every element of X is the first component of one and only one ordered pair in the subset. In other words, for every x in X there is exactly one element y such that the ordered pair (x, y) is contained in the subset defining the function f."
X is called the *domain*, Y the *codomain* or *range*

Functions

We can write function $y = f(x)$ as:

$$f : X \to Y$$

which means that for *any* X, we have a corresponding Y.

$$X \times Y$$

is the *Carthesian product*, the collection of all ordered pairs (x, y)
(Wikipedia): "A function f from X to Y is a subset of the Cartesian product $X \times Y$ subject to the following condition: every element of X is the first component of one and only one ordered pair in the subset. In other words, for every x in X there is exactly one element y such that the ordered pair (x, y) is contained in the subset defining the function f."
X is called the *domain*, Y the *codomain* or *range*

Functions

We can write function $y = f(x)$ as:

$$f : X \to Y$$

which means that for *any* X, we have a corresponding Y.

$$X \times Y$$

is the *Carthesian product*, the collection of all ordered pairs (x, y) (Wikipedia): "A function f from X to Y is a subset of the Cartesian product $X \times Y$ subject to the following condition: every element of X is the first component of one and only one ordered pair in the subset. In other words, for every x in X there is exactly one element y such that the ordered pair (x, y) is contained in the subset defining the function f."

X is called the *domain*, Y the *codomain* or *range*

Functions

We can write function $y = f(x)$ as:

$$f : X \to Y$$

which means that for *any* X, we have a corresponding Y.

$$X \times Y$$

is the *Carthesian product*, the collection of all ordered pairs (x, y) (Wikipedia): "A function f from X to Y is a subset of the Cartesian product $X \times Y$ subject to the following condition: every element of X is the first component of one and only one ordered pair in the subset. In other words, for every x in X there is exactly one element y such that the ordered pair (x, y) is contained in the subset defining the function f."
X is called the *domain*, Y the *codomain* or *range*

Inverse functions

for a set of values B in the range,

$$f^{-1}(B) = x \in X : f(x) \in B$$

for a single value b in the range,

$$f^{-1}(b) = x \in X : f(x) = b$$

the resulting set may contain any number of elements.
Example: $f : X \to X^2$, the range (Y) value 4 has corresponding domain values $\{-2, 2\}$.

Reference systems

Reference systems are *conventions* that encode the shared understanding of information. Examples are

- spatial (coordinate) reference systems (where is (52,8)?)
- temporal reference systems (what does

  ```
  > Sys.time()
  ```

  ```
  [1] "2014-06-18 08:57:55 CEST"
  ```

  mean?
- attribute reference systems (e.g., UCUM, Unified Code for Units of Measure)
- semantic reference systems (vocabularies, ontologies, R function index)

Space, Time, Attribute, Identity

We will look at the following four reference system *domains*:

S	space	1,2,3-dimensional, e.g. 2D degrees in WGS84, $\mathbb{R}^2$ or $\mathbb{R}^3$, continuous
T	time	1-dimensional or cyclic, $\mathbb{R}$, sometimes 2-dimensional, continuous
Q	quality	1-dimensional (UCUM), higher-dimensional: functional, multivariate, also possibly nominal, ordinal, interval (Stevens' 1946)
D	discrete	indicating distinct entities (objects, events); $\mathbb{N}$, IDs, primary key in RDBMS, row number in `data.frame`

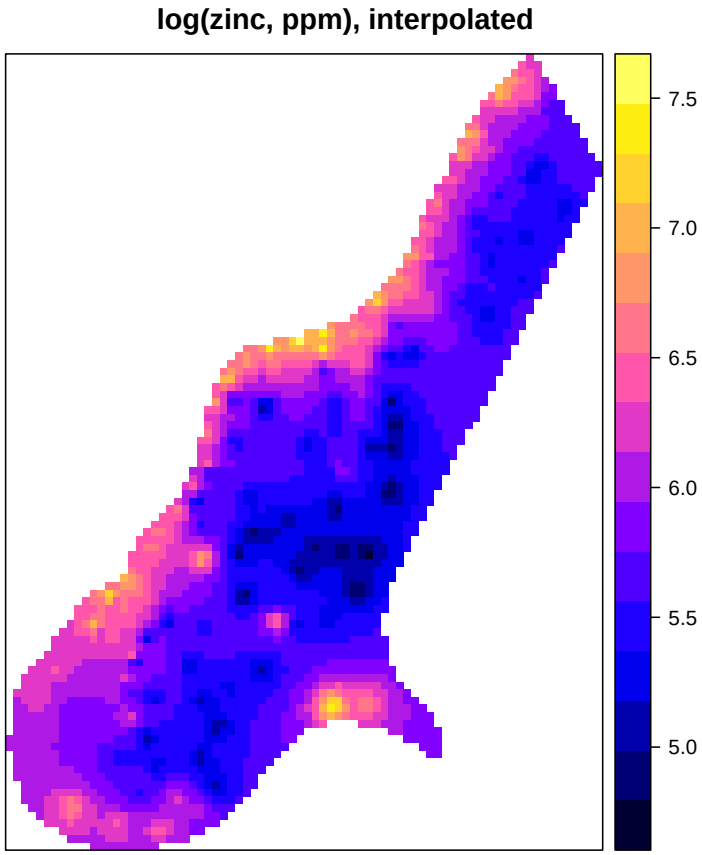
Fields

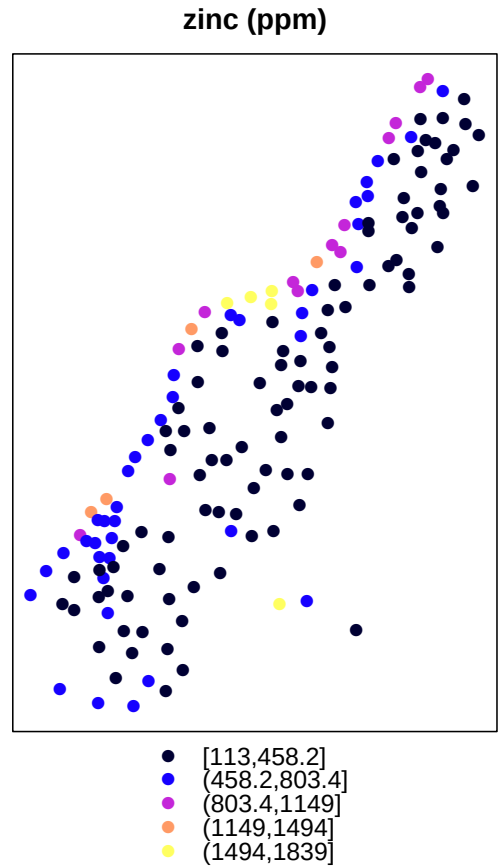
functional form:

$$(S \times T) \to Q$$

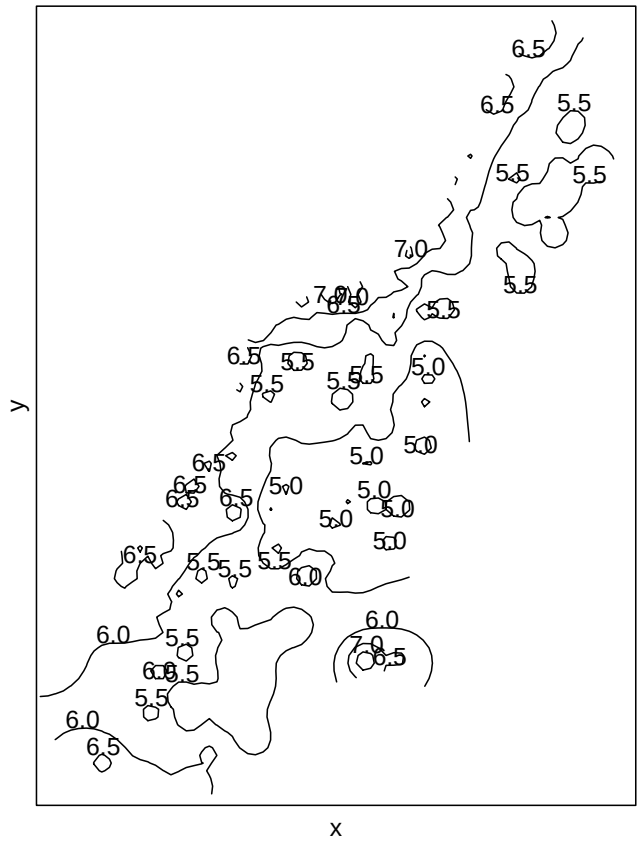
- Answers: "what is then and there?"
- Inverting answers: "when/where was that?"
- Specialisations: $S \to Q$, $T \to Q$
- Incarnations: points (sampled field: meuse), contour lines, coverage

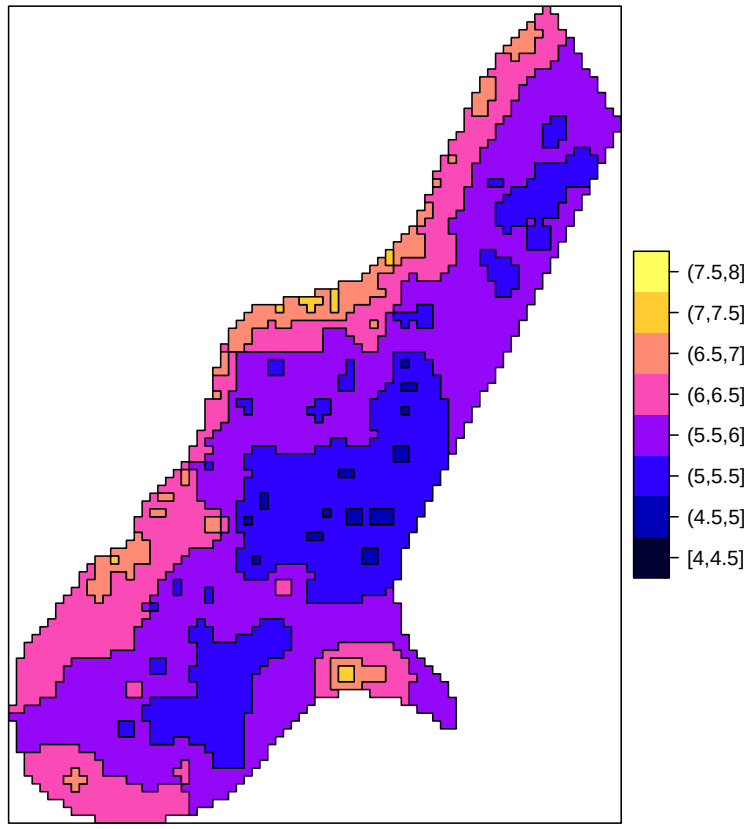
Field examples: grid, points



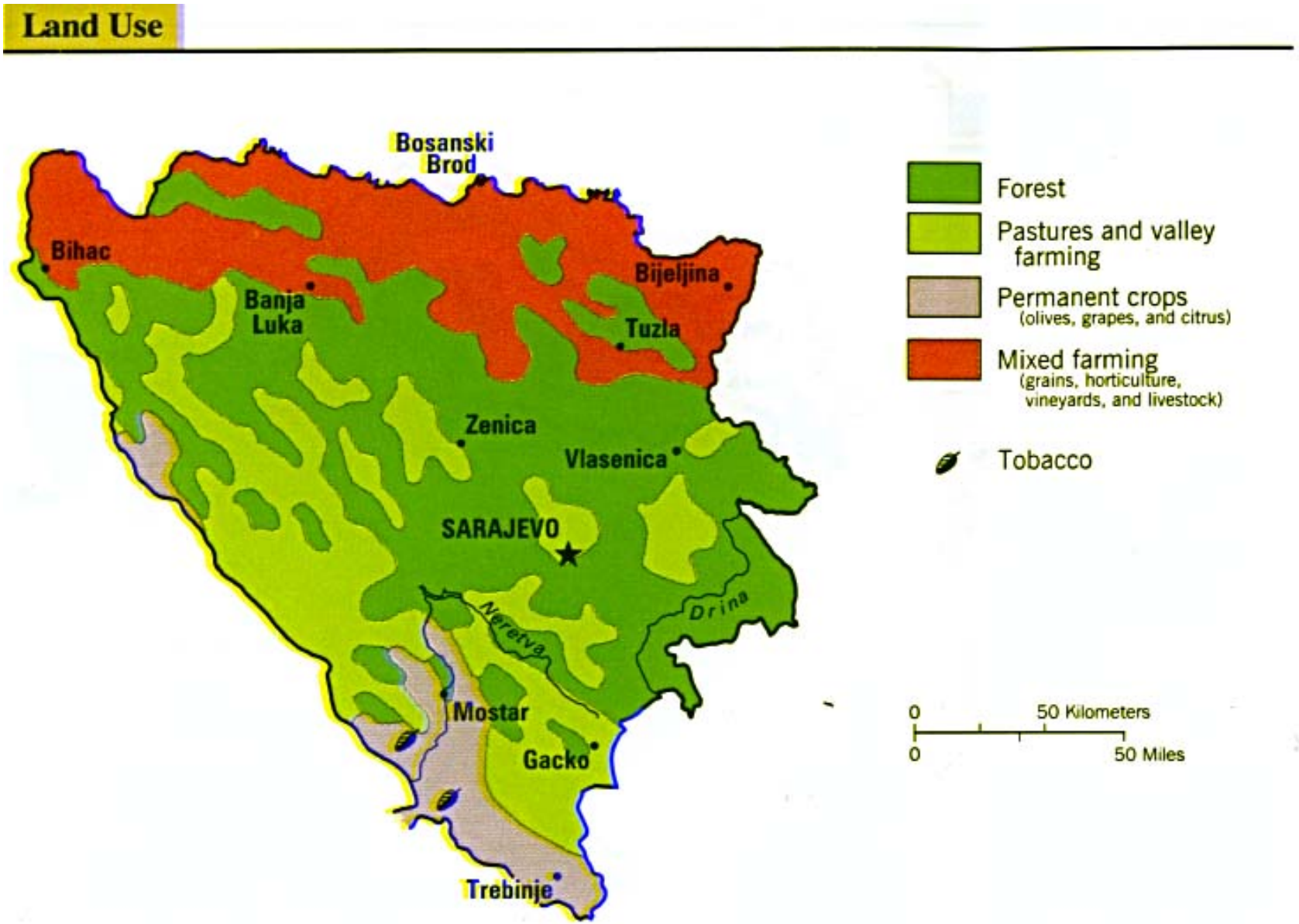


Field examples: lines, polygons

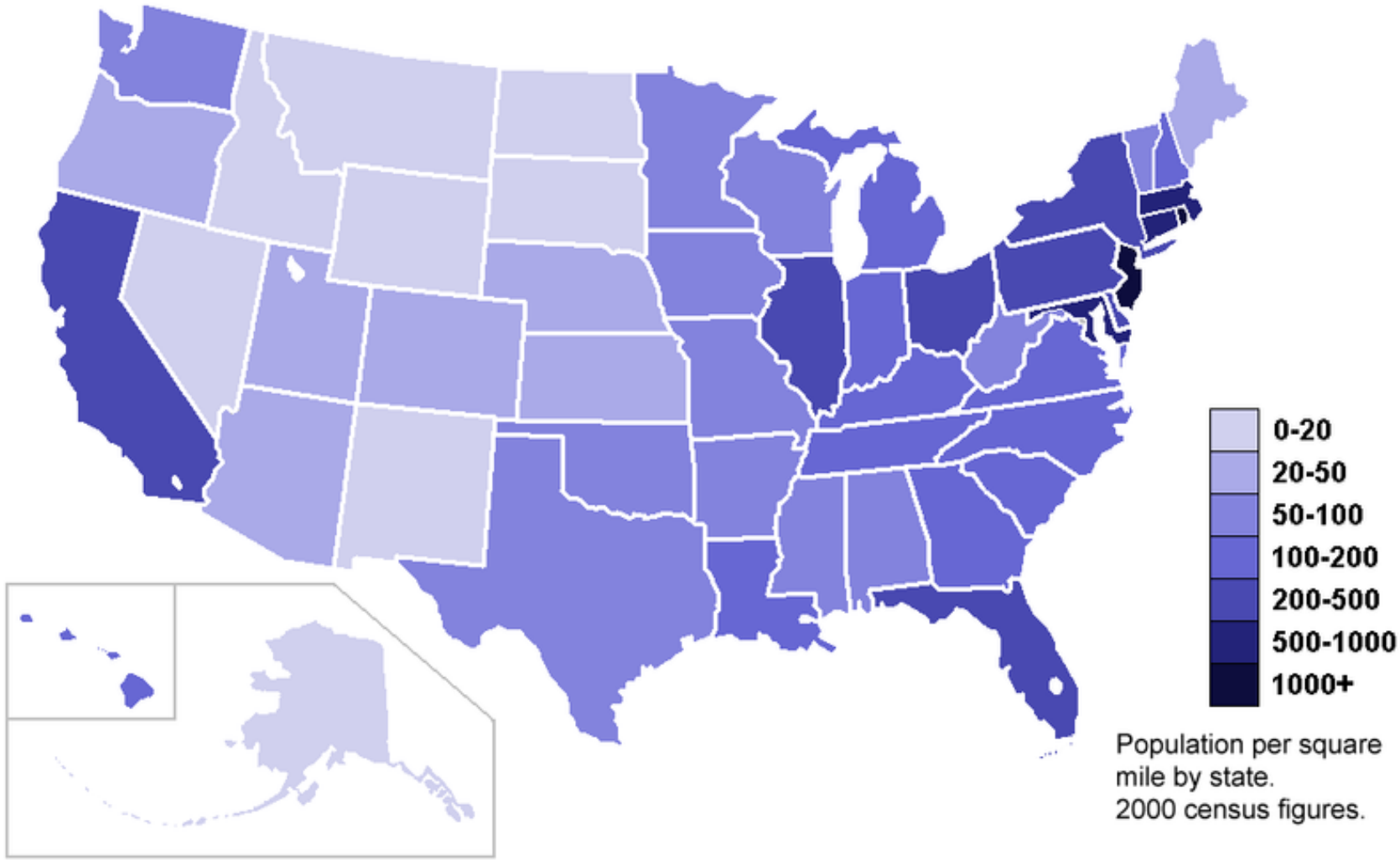




Field: categorical coverage



Non-Field: choropleth, aggregation



Non-moving Entities (objects, events)

functional form:

$$D \to (S \times T \times Q)$$

(for objects without properties, take $Q \equiv 1$)

- Specialisations:
  - $D \to (S \times Q)$: spatial point pattern,
  - $D \to (T \times Q)$: temporal point pattern

Moving entities (objects, events)

functional form:

$$D \to T \to (S \times Q)$$

(for objects without properties, take $Q \equiv 1$)

- generalization of $D \to (S \times T \times Q)$
- specialisations: $D \to T \to Q$, $D \to S \to Q$

Support and aggregation

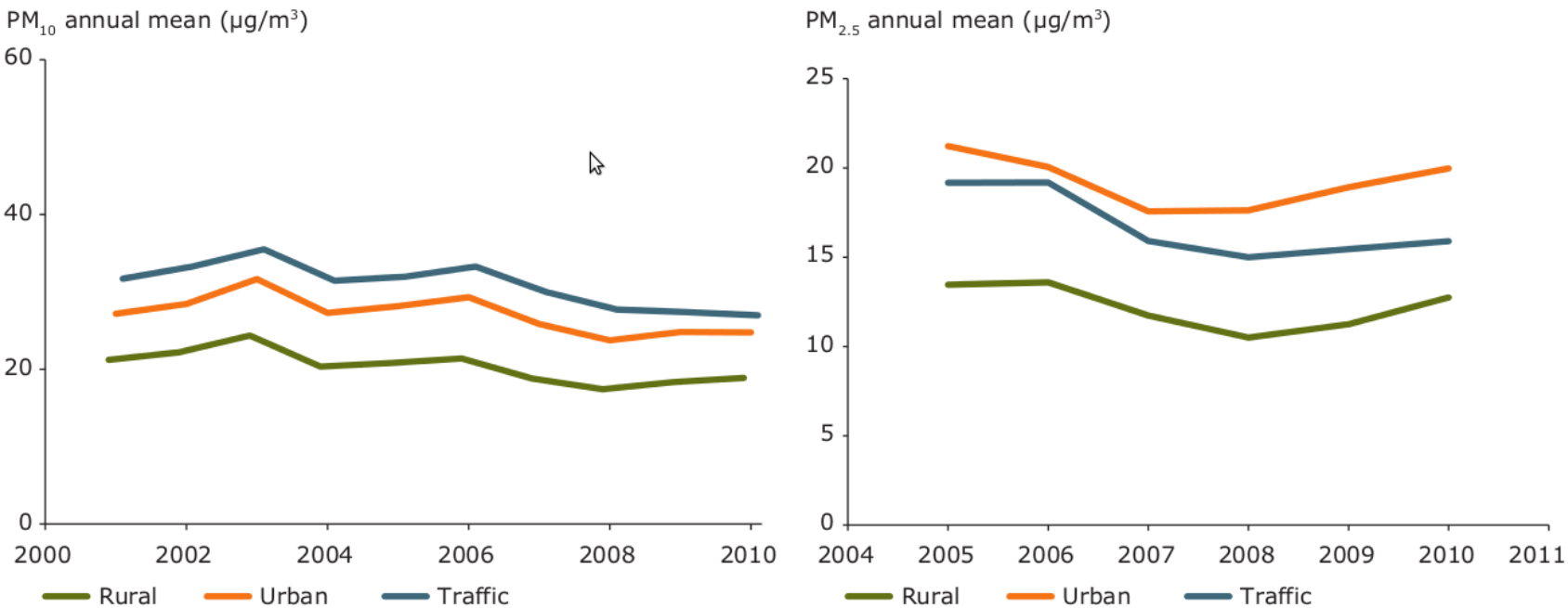
1. we cannot make observations of zero duration, or zero spatial size; the actual size and duration are the measurement *support* (footprint). Think: soil samples, RS cells.
2. often, we want to estimate or compute *aggregated* values, e.g. over periods over areas.
3. even more often, the data we get were aggregated, for convenience (size), or privacy concerns (health data).

Air quality in Europe — 2012 report

ISSN 1725-9177

European Environment Agency

Particulate matter time series, averaged over station type



More complications ...

- "intermediate" phenomena: air quality in street canions ("traffic")
- true "hybrid", 1: time events, spatial fields
  - $D \to ((S \to Q) \times T)$
  - example: election maps
- true "hybrid", 2: spatial events, time fields
  - $D \to (S \times (T \to Q))$
  - example: emission from power plants

How to represent, and then store fields?

1. as functions! Interpolation functions return values at arbitrary times, moments (`gstat::idw` in space, `zoo::na.approx` in time)
2. as evaluated (or observed) functions, at
   - discretized space, regular `raster::raster` or irregular `sp::SpatialPoints`, or
   - time, regular: `stats::ts`, or irregular: `zoo::zoo`, `xts::xts`
3. natural would be to use an index that relates to space and/or time, and records with arbitrarily typed fields $\rightarrow$ *arrays*
4. `netcdf`, `HDF5`;
5. R: array (and raster?) do not support fields of mixed type
6. R for time: zoo, xts do not support fields of mixed type
7. R for space: `sp::SpatialGridDataFrame` do
8. R for space/time: spacetime does too,
9. big data array processing engine: SciDB

How to store objects/events?

Tables are one-dimensional arrays; The `Spatial*` objects in `sp` "behave" like tables (`data.frame`).
Subsetting like `x[3,"zinc"]` works for all, except for `SpatialGridDataFrame`.

I will assume you understand this:

```
> a = data.frame(varA = c(1,1.5,2),
+  varB = c("a", "a", "b"))
> a[1,]

  varA varB
1    1    a

> a[1, drop=FALSE]

  varA
1  1.0
2  1.5
3  2.0

> a[,1]

[1] 1.0 1.5 2.0

> a[1]

  varA
1  1.0
2  1.5
3  2.0

> a[[1]]

[1] 1.0 1.5 2.0
```

```
> a["varA"]

  varA
1  1.0
2  1.5
3  2.0

> a[c("varA", "varB")]

  varA varB
1  1.0    a
2  1.5    a
3  2.0    b

> a$varA

[1] 1.0 1.5 2.0

> a$varA <- 3:1
> a

  varA varB
1    3    a
2    2    a
3    1    b
```

Functional programming

- do it: learn `apply`, `lapply`, `do.call`,
- program generically, e.g. `aggregate`

Time, Time Series Data

1. `POSIXt`, `Date`, `yearmon`, `yearqtr`
2. `zoo`, `xts`, `?aggregate`
3. `forecast`, ...
4. see Task View

Space, Spatial Data

1. `Spatial*`, `raster`,
2. `rgdal`, `rgeos`
3. see Task View
4. selecting records, variables
5. selecting based on spatial match
6. `sp::aggregate`
7. `vignette("over")` (or see CRAN page)
8. `edit(vignette("over"))`, run, modify, run

Space-time, Spatiotemporal Data

1. `spacetime`, `ST*`, also `raster`,
2. back ends: PostGIS, TGRASS, SciDB
3. combines `sp` and `xts`
4. selection, aggregation
5. go through `spacetime` vignettes
6. see Task View