Are current spatial data bases useful for meaningful analysis?

Edzer Pebesma, Christoph Stasch,
Simon Scheider, Werner Kuhn

ifgi
Institute for Geoinformatics
University of Münster

52north
exploring horizons

Oliver Schmitz' defence symposium, UU, May 8, 2014

Motivation

- more data becomes available from an increasing number of sources;
- (interdisciplinary) research tries to integrate more, and more different types of data (satellite imagery, emission statistics, air quality sensor data and model predictions, human trajectories);
- the distance between researchers and the persons who understand the observation process increases

Motivation

- more data becomes available from an increasing number of sources;
- (interdisciplinary) research tries to integrate more, and more different types of data (satellite imagery, emission statistics, air quality sensor data and model predictions, human trajectories);
- the distance between researchers and the persons who understand the observation process increases

$\Rightarrow$ the risk of inappropriate or meaningless analysis increases

Motivation

- more data becomes available from an increasing number of sources;
- (interdisciplinary) research tries to integrate more, and more different types of data (satellite imagery, emission statistics, air quality sensor data and model predictions, human trajectories);
- the distance between researchers and the persons who understand the observation process increases

$\Rightarrow$ the risk of inappropriate or meaningless analysis increases
(How) can we design software such that it warns against this?

What does meaningful mean?

```
> f = factor(c("yellow", "yellow", "red", "blue"))
> f

[1] yellow yellow red    blue
Levels: blue red yellow

> f[1] < f[3]

[1] NA
Warning in Ops.factor(f[1], f[2]) : < not meaningful for factors

> mean(f)

[1] NA
Warning message:
In mean.default(f) : argument is not numeric or logical: returning NA

> x = factor(c("Small", "Large"), ordered = TRUE, levels = c("Small", "Large"))
> x

[1] Small Large
Levels: Small < Large

> x[1] < x[2]

[1] TRUE
```

What does meaningful mean?

```
> f = factor(c("yellow", "yellow", "red", "blue"))
> f

[1] yellow yellow red    blue
Levels: blue red yellow

> f[1] < f[3]

[1] NA
Warning in Ops.factor(f[1], f[2]) : < not meaningful for factors

> mean(f)

[1] NA
Warning message:
In mean.default(f) : argument is not numeric or logical: returning NA

> x = factor(c("Small", "Large"), ordered = TRUE, levels = c("Small", "Large"))
> x

[1] Small Large
Levels: Small < Large

> x[1] < x[2]

[1] TRUE
```

`factor` variables represent categorical (nominal or ordinal) data; for these, it is meaningless to compute means and variances.

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

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

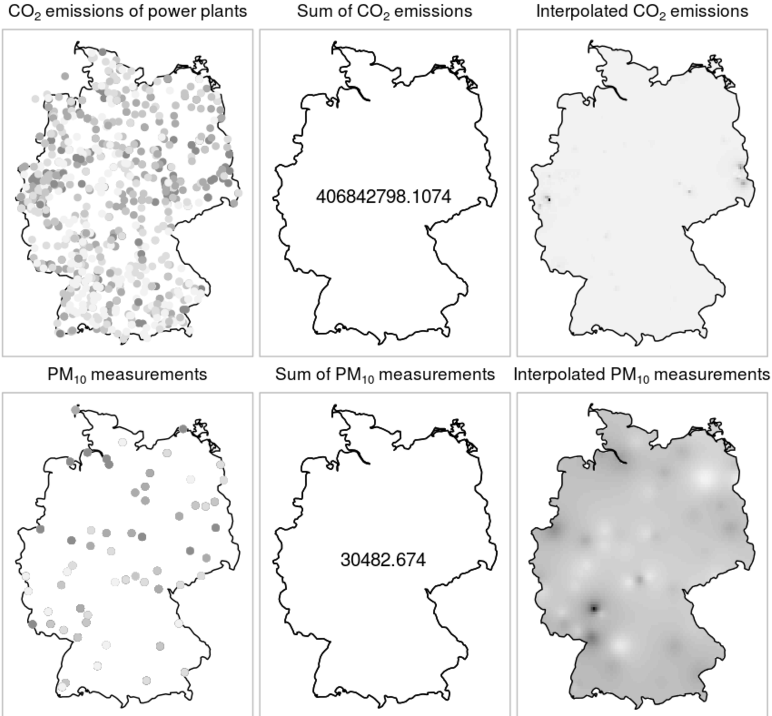
A R T I C L E I N F O

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

A B S T R A C T

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*


CO2 emissions of power plants
Sum of CO2 emissions
Interpolated CO2 emissions
406842798.1074
PM10 measurements
Sum of PM10 measurements
Interpolated PM10 measurements
30482.674

Sinton (1978)

Sinton said that we have location, theme, and time, and from these we can fix one, control a second, and measure the third.[1]

we fix	control	measure	example
time	location	theme	land use map, satellite image
time	theme	location	where are we (now)?
location	time	theme	temperature time series, stock market quotes
location	theme	time	arrival times, earth quakes
theme	location	time	phenology
theme	time	location	tracer experiments, epidemic

[1] *The inherent structure of information as a constraint to analysis: mapped thematic data as a case study. In: Dutton G (ed.) Harvard Papers on Geographic Information Systems, Vol. 6. Addison-Wesley, Reading MA*

Sinton (1978)

Sinton said that we have location, theme, and time, and from these we can fix one, control a second, and measure the third.[1]

we fix	control	measure	example
time	location	theme	land use map, satellite image
time	theme	location	where are we (now)?
location	time	theme	temperature time series, stock market quotes
location	theme	time	arrival times, earth quakes
theme	location	time	phenology
theme	time	location	tracer experiments, epidemic

But: Sinton's *theme* does not distinguish discrete entities from continuous phenomena.

[1] *The inherent structure of information as a constraint to analysis: mapped thematic data as a case study. In: Dutton G (ed.) Harvard Papers on Geographic Information Systems, Vol. 6. Addison-Wesley, Reading MA*

Why reference systems?

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

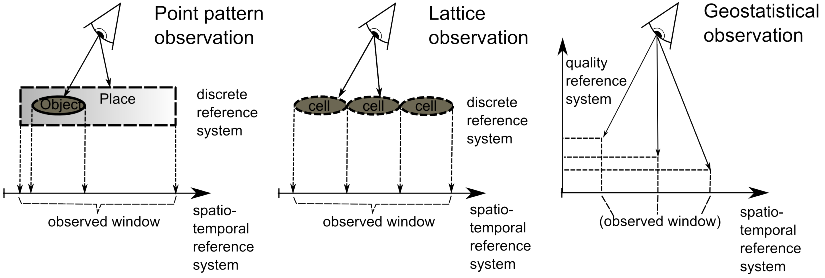
(*Received 12 November 2002; accepted 5 February 2003*)

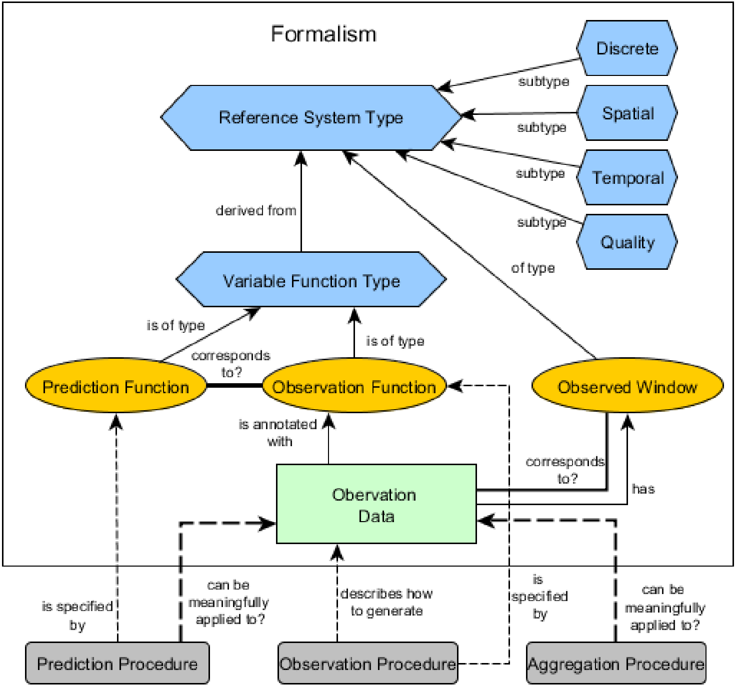
Four centuries after René Descartes watched a fly walk across his ceiling and wondered how to capture its position (Gribbin 2002), we use Cartesian coordinates routinely to describe locations. We identify the positions of entities in the real world, transform their GIS representations from one coordinate 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

Types of Reference System Domains.

Reference Domain	Type	Description	Example
Spatial	D_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	D_t	All possible times defined in a temporal reference system	POSIX time (seconds from 1970-01-01 UTC) with $D_t \subset \mathbb{Q}$
Quality	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)
Discrete Entities	D_d	Set of discrete objects or events.	Set of coal power plants in Germany in 2010

Spatial statistics data




Formalism
Discrete
subtype
Reference System Type
Spatial
subtype
Temporal
subtype
Quality
subtype
derived from
of type
Variable Function Type
is of type
is of type
corresponds to?
Prediction Function
Observation Function
Observed Window
is annotated with
corresponds to?
has
Obervation Data
is specified by
can be meaningfully applied to?
describes how to generate
is specified by
can be meaningfully applied to?
Prediction Procedure
Observation Procedure
Aggregation Procedure

Point Pattern Variables

Functional type	Example
$D_d \Rightarrow D_s$	Locations of longleaf pines in 4 ha of a natural forest in Thomas County, Georgia
$D_d \Rightarrow (D_s \times D_q)$	Locations of longleaf pines in 4 ha of a natural forest in Thomas County, Georgia with diameters-at-breast-heights (DBH)
$D_d \Rightarrow D_t$	Occurrence of earthquakes in an american county
$D_d \Rightarrow (D_t \times D_q)$	Occurrence of earthquakes in an american county with magnitudes
$D_d \Rightarrow (D_t \times D_s)$	Occurrence of earthquakes at particular locations in space and time
$D_d \Rightarrow (D_t \times D_s \times D_q)$	Occurrence of earthquakes at particular locations in space and time with magnitudes

Geostatistical and Lattice Variables

Variable type	Functional type	Example
Geostatistical Variable	$(D_s \times D_t) \Rightarrow D_q$	PM_{10} concentrations across Germany
Lattice Variable	$({}^r D_s \times D_t) \Rightarrow D_q$	Number of inhabitants per cantons of the Midi-Pyrenees

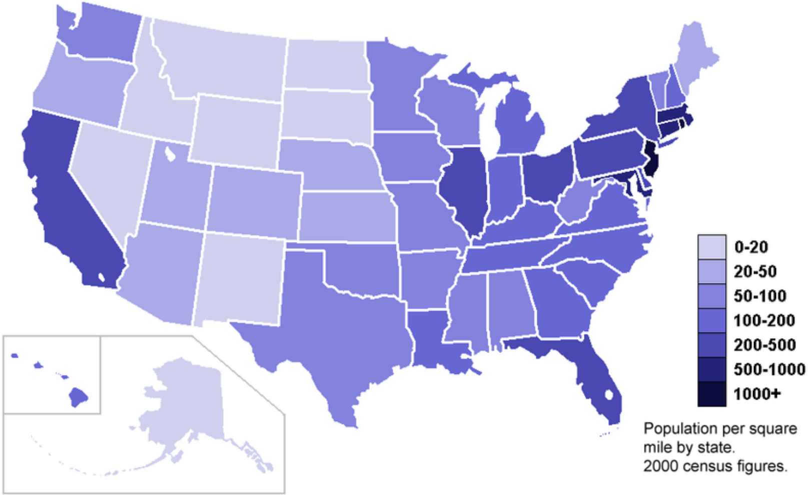
Trajectory Variable Type in Spatial Statistics

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

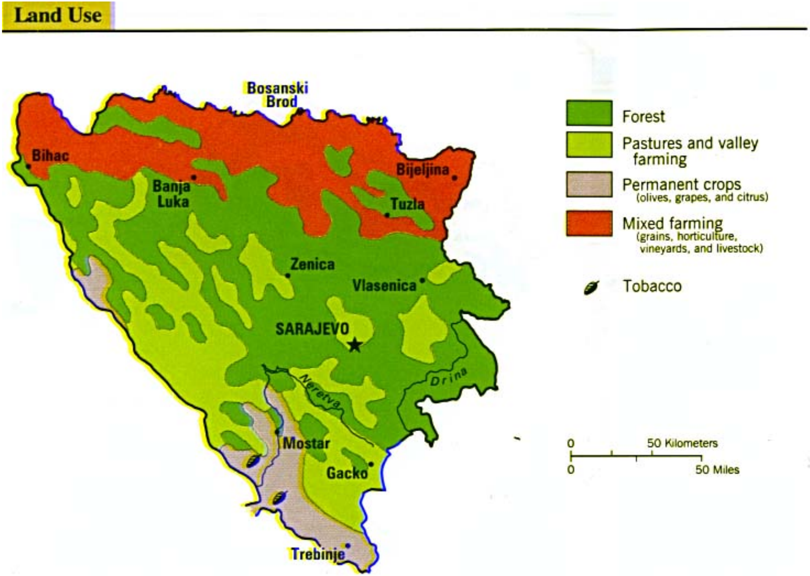
Spatial data bases: PostGIS view

```
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



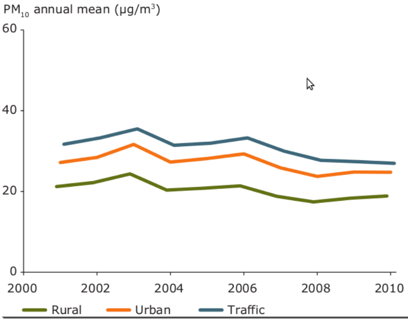
EEA Report | No 4/2012

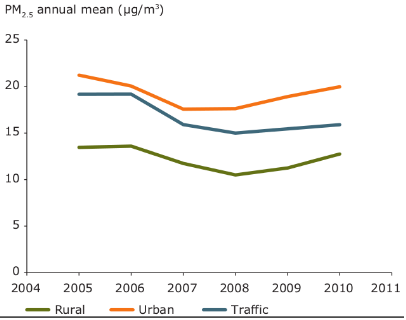
Air quality in Europe — 2012 report

ISSN 1725-9177

European Environment Agency

Particulate matter time series, averaged over station type





- not one, but multiple (discrete) indexes form “primary key”
- dimensions: space (2 or 3), time, spectral, ...
- records may contain attributes of different types
- built for Tb-Pb scale, scientific data
- supports *sparse arrays*

When representing

```
> R = 6378.1
> C = 2 * pi * R * 1e6 # earth's circumference, in mm
> C / 2^64 # resolution, in mm, if earth covered

[1] 2.172458e-09
```

Summarizing

- “points” may represent discrete entities, or measurements taken on a continuous field
- meaningfulness of interpolation or aggregation (sum) depends on whether we have the one, or the other
- polygons (or lines) attributes may represent aggregates over varying values, or coverages (constant values)
- sampling (downscaling) is only meaningful for coverages
- GIS and (relational) data bases do not tell us one from the other
- function types, constructed from 4 reference system domains, can do so

Summarizing

- "points" may represent discrete entities, or measurements taken on a continuous field
- meaningfulness of interpolation or aggregation (sum) depends on whether we have the one, or the other
- polygons (or lines) attributes may represent aggregates over varying values, or coverages (constant values)
- sampling (downscaling) is only meaningful for coverages
- GIS and (relational) data bases do not tell us one from the other
- function types, constructed from 4 reference system domains, can do so
- array data bases can meaningfully represent continuous phenomena
- "taming" them for spatiotemporal data, and integrating them with other types of spatiotemporal data remains a challenge

Hybrid types