Towards meaningful spatial statistics

Edzer Pebesma, Christoph Stasch,
Simon Scheider, Werner Kuhn

ifgi
Institute for Geoinformatics
University of Münster

52north
exploring horizons

Spatial Statistics, Columbus OH
Jun 5-7, 2013

Motivation

- more data becomes available from an increasing number of sources
- (interdisciplinary) research tries to integrate more different types of data
- the distance between researcher and the act of observation increases

Motivation

- more data becomes available from an increasing number of sources
- (interdisciplinary) research tries to integrate more different types of data
- the distance between researcher and the act of observation increases

⇒ the risk of inappropriate or meaningless analysis increases

What does **meaningful** mean?

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

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

> mean(f)

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

> as.numeric(f)

[1] 3 3 2 1

> mean(as.numeric(f))

[1] 2.25
```

What does meaningful mean?

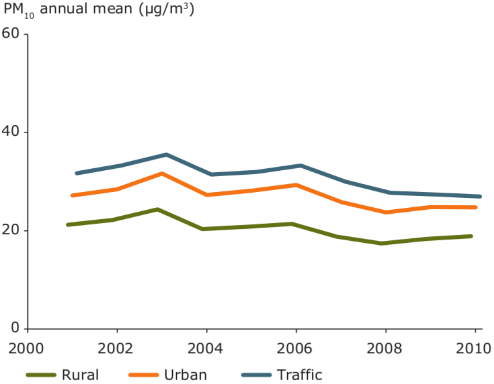
```
> f = factor(c("yellow", "yellow", "red", "blue"))
> f

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

> mean(f)

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

> as.numeric(f)

[1] 3 3 2 1

> mean(as.numeric(f))

[1] 2.25
```

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

Stevens, S.S., 1946. *On the Theory of Scales and Measurement.* Science 103, 677–680.

Air quality in Europe: EEA report 4/2012

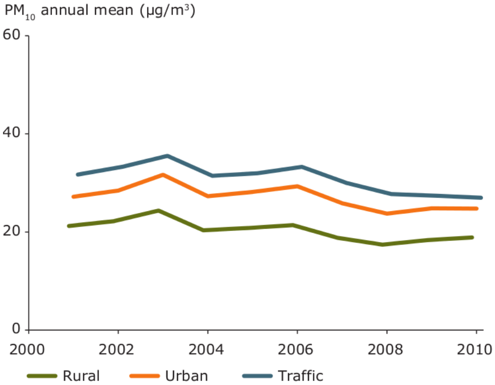
Trends in PM_{10} ($\mu g/m^3$), 2001-2010, per station type



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

Air quality in Europe: EEA report 4/2012

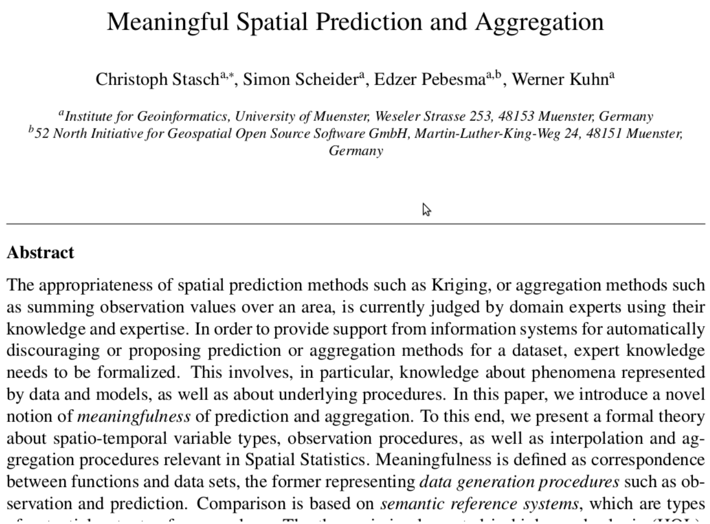
Trends in PM_{10} ($\mu g/m^3$), 2001-2010, per station type



"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)

Why "Towards"?



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
```

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
```

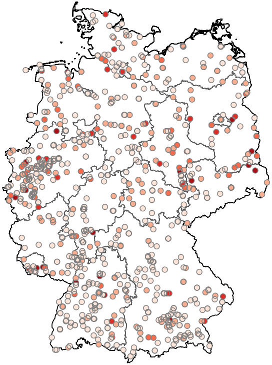
these data look similar, but can we do similar things with them?

Example: sum of top soil zinc concentrations

```
> sum(meuse$zinc)

[1] 72806

> coordinates(meuse) = ~x+y
> data(meuse.grid)
> gridded(meuse.grid) = ~x+y
> library(rgeos)
> area = gUnionCascaded(as(meuse.grid, "SpatialPolygons"))
> aggregate(meuse["zinc"], area, sum)[[1]]

[1] 72806

> sum(meuse[area,][["zinc"]])

[1] 72806
```

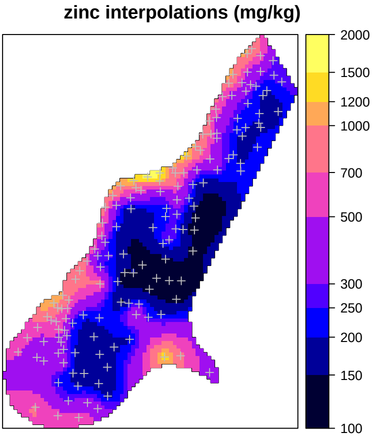
Example: sum of coal power plant CO_2 emissions

```
> sum(co2$carbon_2007)

[1] 408032358

> coordinates(co2) = ~longitude+latitude
> library(spacetime)
> data(air)
> DE = gUnionCascaded(DE_NUTS1)
> proj4string(co2) = proj4string(DE)
> sum(co2[DE,][["carbon_2007"]])

[1] 407203574

> as(aggregate(co2["carbon_2007"], DE_NUTS1, sum), "data.frame")

                        carbon_2007
Baden-Wurttemberg        29911306.2
Bayern                   12524147.5
Berlin                   13049336.8
Brandenburg              47260944.6
Bremen                    7328829.7
Hamburg                   1927826.4
Hessen                   11758650.6
Mecklenburg-Vorpommern    5188286.9
Niedersachsen            22296016.3
Nordrhein-Westfalen     184510334.0
Rheinland-Pfalz           5545028.7
Saarland                 18568351.3
Sachsen                  34103366.0
Sachsen-Anhalt           10268063.7
Schleswig-Holstein        2433435.1
Thuringen                  529650.3
```



Example: interpolating top soil zinc concentration

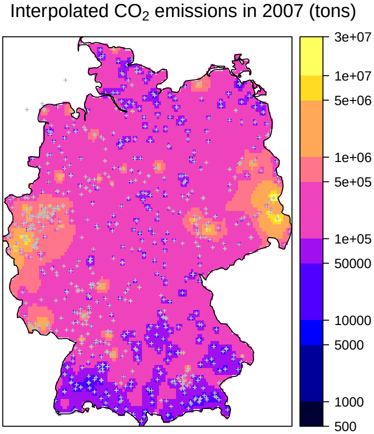
```
> v = variogram(log(zinc)~1, meuse)
> v.fit = fit.variogram(v, vgm(1, "Sph", 900, 1))
> m.kr = krige(log(zinc)~1, meuse, meuse.grid, v.fit)

[using ordinary kriging]

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

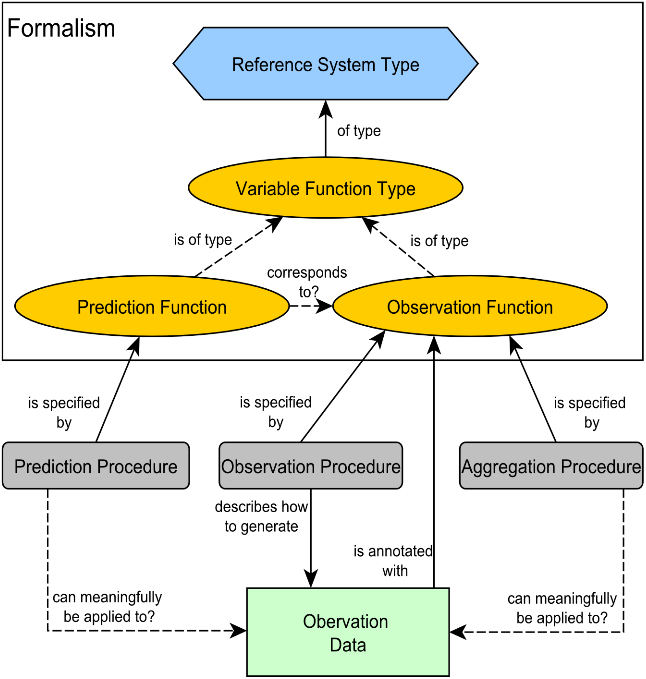


Interpolating power plant CO_2 emissions

```
> # create interpolation grid:
> grd = spsample(DE, 10000, "regular", offset = c(0,0))
> gridded(grd) = TRUE
> # interpolate, idw:
> co2_interpolated <- krige(carbon_2007~1, co2, grd)
```

[inverse distance weighted interpolation]


Interpolated CO_2 emissions in 2007 (tons)


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 Discrete Entities	D_d	Set of discrete objects or events.	Set of coal power plants in Germany in 2010

Variable Types in Spatial Statistics

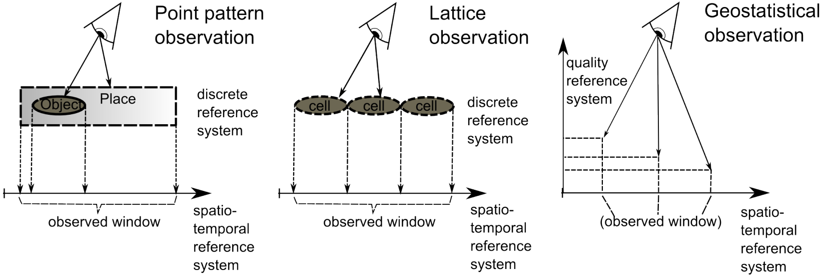
- **Point patterns**: (unmarked, marked) × (spatial, temporal, spatio-temporal)
  $MSTPP = ''D_d \Rightarrow (D_t \times D_s \times D_q)''$
- **Geostatistical variables**: (spatial, temporal, spatio-temporal)
  $GEOST = ''(D_s \times D_t) \Rightarrow D_q''$
- **Lattice variables**: (spatial, temporal, spatio-temporal)
  $LAT = ''(^r D_s \times D_t) \Rightarrow D_q''$
- **Trajectories**: (unmarked, marked)
  $MTRAJECT = ''D_d \Rightarrow D_t \Rightarrow (D_s \times D_q)''$

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.

Observation windows


Point pattern observation
Object
Place
discrete reference system
observed window
spatio-temporal reference system
Lattice observation
cell
cell
cell
discrete reference system
observed window
spatio-temporal reference system
Geostatistical observation
quality reference system
(observed window)
spatio-temporal reference system

What we would like to see:

```
> options(warn = 1)
> library(mss)
> meuse = as(meuse, "GeostatisticalDataFrame")
> m = aggregate(meuse["zinc"], area, sum)

Warning in aggregate.GeostatisticalDataFrame(meuse["zinc"], area, sum) :
  aggregation using a sum function is not considered meaningful for Geostatistical data

> co2 = as(co2, "PointPatternDataFrame")
> co2_interpolated = krige(carbon_2007~1, co2, grd)

Warning in krige(carbon_2007 ~ 1, co2, grd) :
  interpolating point patterns is not considered meaningful
[inverse distance weighted interpolation]

> options(warn = 2)
> co2_interpolated <- krige(carbon_2007~1, co2, grd)

Error in krige(carbon_2007 ~ 1, co2, grd) :
  (converted from warning) interpolating point patterns is not considered meaningful
```

What we would like to see:

```
> options(warn = 1)
> library(mss)
> meuse = as(meuse, "GeostatisticalDataFrame")
> m = aggregate(meuse["zinc"], area, sum)

Warning in aggregate.GeostatisticalDataFrame(meuse["zinc"], area, sum) :
  aggregation using a sum function is not considered meaningful for Geostatistical data

> co2 = as(co2, "PointPatternDataFrame")
> co2_interpolated = krige(carbon_2007~1, co2, grd)

Warning in krige(carbon_2007 ~ 1, co2, grd) :
  interpolating point patterns is not considered meaningful
[inverse distance weighted interpolation]

> options(warn = 2)
> co2_interpolated <- krige(carbon_2007~1, co2, grd)

Error in krige(carbon_2007 ~ 1, co2, grd) :
  (converted from warning) interpolating point patterns is not considered meaningful
```

... but should this require additional coding?

What we would like to see:

```
> options(warn = 1)
> library(mss)
> meuse = as(meuse, "GeostatisticalDataFrame")
> m = aggregate(meuse["zinc"], area, sum)


Warning in aggregate.GeostatisticalDataFrame(meuse["zinc"], area, sum) :
  aggregation using a sum function is not considered meaningful for Geostatistical data

> co2 = as(co2, "PointPatternDataFrame")
> co2_interpolated = krige(carbon_2007~1, co2, grd)


Warning in krige(carbon_2007 ~ 1, co2, grd) :
  interpolating point patterns is not considered meaningful
[inverse distance weighted interpolation]

> options(warn = 2)
> co2_interpolated <- krige(carbon_2007~1, co2, grd)


Error in krige(carbon_2007 ~ 1, co2, grd) :
  (converted from warning) interpolating point patterns is not considered meaningful
```

... but should this require additional coding?

⇒ at the time of data *import* the variable type should be set.

`http://meaningfulspatialstatistics.org`

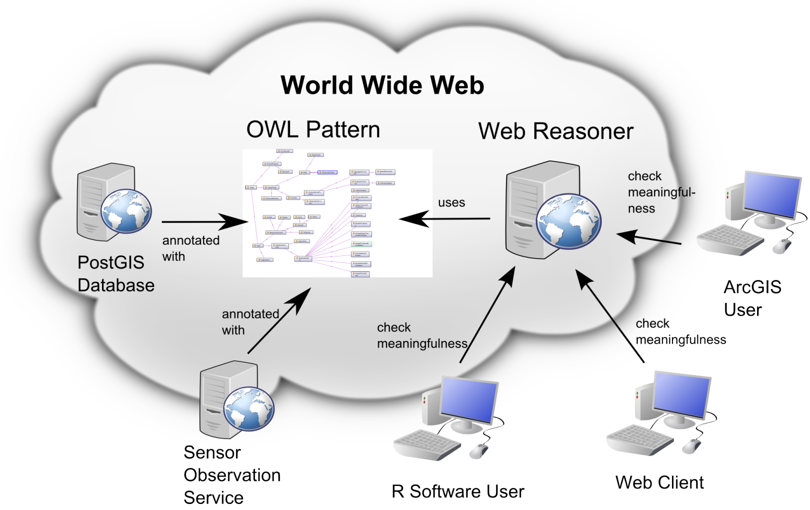
... serves our OWL (web ontology language) pattern, which

- provides the classes defined, and their relationships
- allows semantic annotation of data sets, and procedures

... provides a meaningful spatial statistics data portal, which

- is based on linked data, storing RDF triples, e.g.
  DataSet_XYZ isOfType MarkedSpatialPointPattern
  with appropriate links (URIs) to resources and definitions
- allows retrieving and uploading annotations about data *types*
- allows e.g. retrieving collections of a particular type,
- supports SPARQL directly, or from R using package `RSPARQL`

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

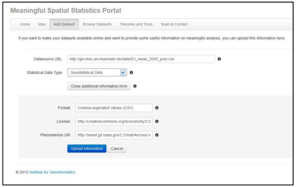
Conclusions, discussion:

- Semantic annotation of data can help prevent meaningless operations
- Automated reasoning is still challenging (HOL implementation of formalism)
- the “what is a data set?” question: DOI, fields, etc.
- Filling the annotation data portal by scraping CRAN, `jstatsoft.org`, and other published analyses scripts
- Handling ambiguities in annotations
- (user management and authority in the annotations data base)

The limits of crisp classes

- are traffic station air quality measurements geostatistical variables? (stationarity assumption)
- what if a car measures engine temperature *and* outside air temperature?
- what does the sum of observed bird counts mean, in particular in case of volunteered information?
- time series of power plant CO_2 emmissions: hybrid classes?
- if not meaningful as predictions, what *do* interpolated point pattern marks tell us?

Web Portal



Meaningful Spatial Statistics Portal

If you want to make your datasets available online and want to provide some useful information on meaningful analysis, you can upload this information here.

Datasource URL: http://giv-mss.uni-muenster.de/data/EU_meas_2005_june.csv

Statistical Data Type: Geostatistical Data

Format: Comma-seperated Values (CSV)

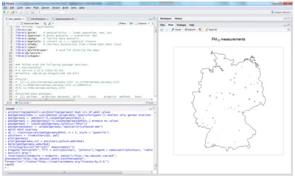
License: http://creativecommons.org/licenses/by/3.0

Phenomenon URI: http://sweet.jpl.nasa.gov/2.2/matrAerosol.o

© 2012 Institute for Geoinformatics

annotate datasets / browse annotations

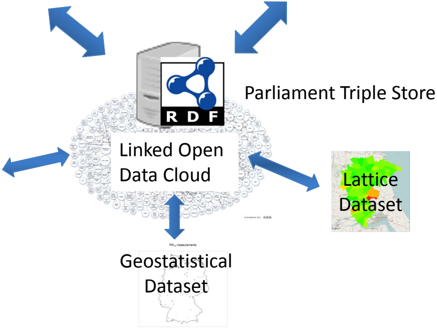
Meaningful Analysis in R



```
### SECTION: requirements
library(sp)
library(gstat)     # geostatistics -- loads spacetime, zoo, xts
library(spatstat) # point patterns -- overwrites idw!
library(spdep)     # lattice data analysis
library(maptools) # convert sp <--> spatstat classes
library(SPARQL)    # retrieve annotations from Linked Open Data cloud
library(rgeos)
library(RColorBrewer)      # used for plotting the maps
library(classInt)
library(cshapes)


### TESTED with the following package versions:
# > sessionInfo()
# R version 2.15.0 (2012-03-30)
#Platform: x86_64-pc-mingw32/x64 (64-bit)
#
#locale:
#  [1] LC_COLLATE=German_Germany.1252  LC_CTYPE=German_Germany.1252
#[3] LC_MONETARY=German_Germany.1252 LC_NUMERIC=C
#[5] LC_TIME=German_Germany.1252
#
#attached base packages:
#  [1] splines   grDevices datasets  utils     stats     graphics  methods   base
```

```
> proj4string(pm10sel)<-proj4string(germany) #set crs of pm10 values
> pm10germanyIndex <- over(pm10sel,as(germany,"SpatialPolygons")) #select only german stations
> pm10germany <- pm10sel[!is.na(pm10germanyIndex),]
> pm10germany <- pm10germany[!is.na(pm10germany$PM10),] #remove NA values
> pm10germany <- subset(pm10germany,select=c("PM10"))
> pm10germanyGeost <- as(pm10germany,"GeostatisticalDataFrame")
> #plot PM10 stations
> q5 <- classIntervals(pm10germany$PM10, n = 5, style = "quantile")
> q5Colours <- findColours(q5, pal)
> plot(germany)
> plot(pm10germany,col = q5Colours,pch=19,add=TRUE)
> #plot(pm10germany,add=TRUE)
> title(expression("PM"[10]*" measurements"))
> #legend("bottomleft", fill = attr(q5Colours, "palette"),legend = names(attr(q5Colours, "table"
> box(col='grey')
> insertGeosts(endpoint = endpoint, dataUrl="http://my.dataset.com/ds6",
phenomenon="http://my.dataset.phens/testPhenomenon",
format="csv",license="http://creativecommons.org/licenses/by/3.0/")
raw(0)
>
```

retrieve/upload annotations + meaningful analysis



Parliament Triple Store

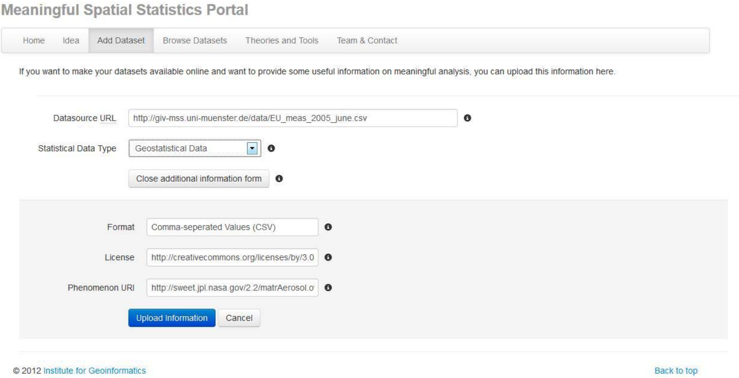
Linked Open Data Cloud

Point Pattern Dataset

Lattice Dataset

Geostatistical Dataset

Add annotations



If you want to make your datasets available online and want to provide some useful information on meaningful analysis, you can upload this information here.

Datasource URL: http://giv-mss.uni-muenster.de/data/EU_meas_2005_june.csv

Statistical Data Type: Geostatistical Data

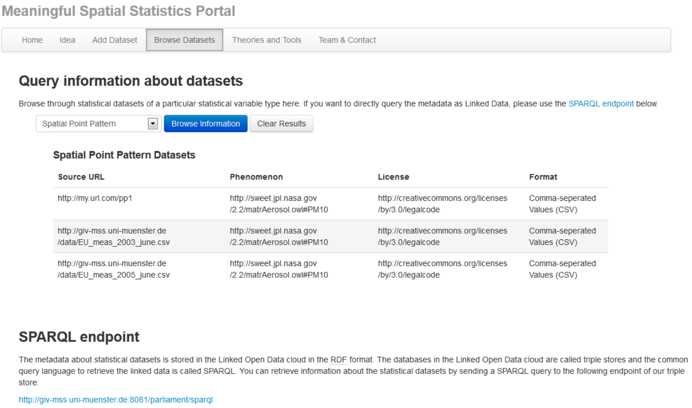
Format: Comma-seperated Values (CSV)

License: http://creativecommons.org/licenses/by/3.0

Phenomenon URI: http://sweet.jpl.nasa.gov/2.2/matrAerosol.o

© 2012 Institute for Geoinformatics

Browse annotations



Query information about datasets

Browse through statistical datasets of a particular statistical variable type here. If you want to directly query the metadata as Linked Data, please use the SPARQL endpoint below.

Spatial Point Pattern Datasets

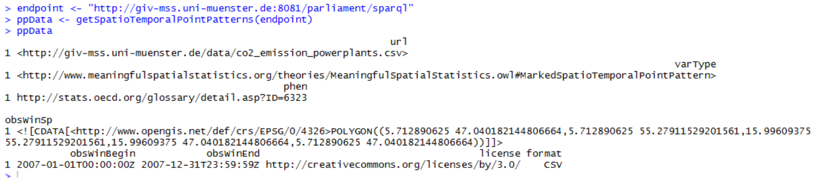
Source URL	Phenomenon	License	Format
http://my.url.com/pp1	http://sweet.jpl.nasa.gov /2.2/matrAerosol.owl#PM10	http://creativecommons.org/licenses /by/3.0/legalcode	Comma-seperated Values (CSV)
http://giv-mss.uni-muenster.de /data/EU_meas_2003_june.csv	http://sweet.jpl.nasa.gov /2.2/matrAerosol.owl#PM10	http://creativecommons.org/licenses /by/3.0/legalcode	Comma-seperated Values (CSV)
http://giv-mss.uni-muenster.de /data/EU_meas_2005_june.csv	http://sweet.jpl.nasa.gov /2.2/matrAerosol.owl#PM10	http://creativecommons.org/licenses /by/3.0/legalcode	Comma-seperated Values (CSV)

SPARQL endpoint

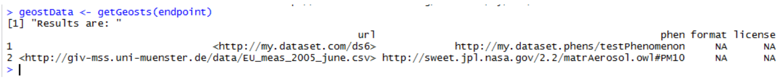
The metadata about statistical datasets is stored in the Linked Open Data cloud in the RDF format. The databases in the Linked Open Data cloud are called triple stores and the common query language to retrieve the linked data is called SPARQL. You can retrieve information about the statistical datasets by sending a SPARQL query to the following endpoint of our triple store:

http://giv-mss.uni-muenster.de:8081/parliament/sparql

Retrieve Annotations in R

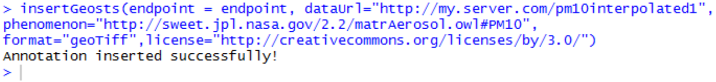


```
> endpoint <- "http://giv-mss.uni-muenster.de:8081/parliament/sparql"
> ppData <- getSpatioTemporalPointPatterns(endpoint)
> ppData
                                                                 url
1 <http://giv-mss.uni-muenster.de/data/co2_emission_powerplants.csv>
                                                                                                                 varType
1 <http://www.meaningfulspatialstatistics.org/theories/MeaningfulSpatialStatistics.owl#MarkedSpatioTemporalPointPattern>
                                               phen
1 http://stats.oecd.org/glossary/detail.asp?ID=6323

obsWinSp
1 <![CDATA[<http://www.opengis.net/def/crs/EPSG/0/4326>POLYGON((5.712890625 47.040182144806664,5.712890625 55.27911529201561,15.99609375
55.27911529201561,15.99609375 47.040182144806664,5.712890625 47.040182144806664))]]>
           obsWinBegin            obsWinEnd                                     license format
1 2007-01-01T00:00:00Z 2007-12-31T23:59:59Z http://creativecommons.org/licenses/by/3.0/    csv
> 
```



```
> geostData <- getGeosts(endpoint)
[1] "Results are: "
                                                          url                                               phen format license
1                                <http://my.dataset.com/ds6>             http://my.dataset.phens/testPhenomenon     NA      NA
2 <http://giv-mss.uni-muenster.de/data/EU_meas_2005_june.csv> http://sweet.jpl.nasa.gov/2.2/matrAerosol.owl#PM10     NA      NA
> 
```

Upload Annotations



```
> insertGeosts(endpoint = endpoint, dataUrl="http://my.server.com/pm10interpolated1",
phenomenon="http://sweet.jpl.nasa.gov/2.2/matrAerosol.owl#PM10",
format="geoTiff",license="http://creativecommons.org/licenses/by/3.0/")
Annotation inserted successfully!
> |
```