Meanfully Integrating Big Earth Science Data

Edzer Pebesma¹, Christoph Stasch², Benedikt Gräler¹, Simon Scheider³

¹ Institute for Geoinformatics, University of Münster, DE; ² 52°North GmbH; ³ ETH Zürich; IN33A-3757

**Objectives**

We propose a description of data at the level of their meaning that allows for notifying data users when meaningless operations are being executed, and present a prototypical implementation in R.

**Introduction**

When more data is open and problems are more interdisciplinary, data analysts need to work more with data they do not understand well, increasing the risk of analysis that is non-meaningful [1, 2].

The following figure shows summing over a region and spatial interpolation of:

- CO₂ emissions from coal power plants in Germany (top)
- PM₁₀ concentrations for rural background stations in Germany (bottom)

**What Is Meaningful?**

Based on matching variable types, we proof meaningfulness using HOL as follows [1]:

A prediction is meaningful, if it provides an estimate for a potential observation.

- Summing up values over an area is meaningful, if the observed window corresponds to the target geometry (grouping predicate) of the aggregation.

Further assertions:

- Polygon data may reflect constant values over areas (geology) or means (population density), this matters for (further) aggregation or downsampling.
- The same is true for grid cells.
- Point pattern data often come without observation window, over which aggregation is meaningful.
- Field measurements often come without a notion for which locations interpolations based on them make sense.
- File types do not inform on meaningfulness.

**Implementation: R**

```r
> install_github("mss", "edzer")
> library(mss)
> co2 = as(pm10, "PointPatternDataFrame")
> pm10 = as(pm10, "PointPatternDataFrame")
```

**Open Challenges**

- How can we set these types automatically? By (i) annotating data [2], (ii) annotating phenomena, (iii) scraping R scripts, or (iv) by teaching students?
- How can we incorporate quality and uncertainty criteria?
- How do we represent hybrid spatio-temporal variable types, e.g. {point pattern in space, field in time}?
- How can we link our approach to existing Web portals/catalogues/brokers?

**References**


**Contact Information**

- [http://www.meanfulspatialstatistics.org](http://www.meanfulspatialstatistics.org)
- Email: edzer.pebesma@uni-muenster.de
- Phone: +49 251 83 33081