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
Spatial prediction
Spatiotemporal aggregation

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 of potential outputs of a procedure. The theory is implemented in higher-order logic (HOL), and theorems about meaningfulness are proved in the semi-automated prover Isabelle. The type system of our theory is available as a Web Ontology Language (OWL) pattern for use in the Semantic Web. In addition, we show how to implement a data-model recommender system in the statistics tool environment R. We consider our theory groundwork to automate semantic interoperability of data and models.

1. Introduction

Summing temperature measurements or interpolating point source emissions is not meaningful. This paper formalises meaningfulness of applying prediction and aggregation procedures to data. With ever increasing data volumes (Bell et al., 2009) of diverse origin and nature (Parsons et al., 2011), we observe an increase in importance of *information semantics* to the application of Spatial Statistics and environmental modelling (Villa et al., 2009). Although we do have access to more and more data, the distance between those who collect the data and those who analyse it has become larger. Also, in interdisciplinary settings, data from heterogeneous sources are combined by researchers without specific domain knowledge, increasing the risk of *inappropriate analysis* (Pebesma et al., 2011). Making sense of these large data volumes exceeds the limits and competence of a particular group of scientists (Weinberger, 2011), and thus there is a need for *semantic metadata* that can bridge the gap of knowledge which exists between groups (Gray et al., 2005).

NASA has recently argued that model reuse and *data-model interoperability* has a significant added value, as about *60%* of the time of NASA scientists is spent on making data and models compatible (National Aeronautics and Space Administration, 2012). In order to achieve integrated environmental modelling, standards that allow to describe and publish data and models in an automated fashion need to be developed (Laniak et al., 2013). When integrating environmental models, it is crucial to avoid “constructs that are perfectly valid as software, but ugly or even useless as models” (Voinov and Shugart, 2013, p.149).

Observations form the basis of empirical and physical sciences. They provide samples for a process of interest, enabling us to infer knowledge about this process and to evaluate assumptions and hypotheses. In order to infer knowledge or test hypotheses about a process, *statistical models and procedures* can be applied to observations. The syntactical integration of observations in statistical modelling frameworks is not an issue (R Development Core Team, 2011). However, the semantic integration of observations in such

☆ This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

* Corresponding author. Current address: Institute for Geoinformatics, University of Muenster, Heisenbergstr. 2, 48149 Muenster, Germany. Tel.: +49 251 8339760; fax: +49 251 8339763.

E-mail addresses: staschc@uni-muenster.de (C. Stasch), s_sche30@uni-muenster.de (S. Scheider), e.pebesma@uni-muenster.de, pebesma@52north.org (E. Pebesma), kuhn@uni-muenster.de (W. Kuhn).

1364-8152/$ – see front matter © 2013 The Authors. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.09.006

systems still forms a major challenge (Sheth et al., 2008), since not all syntactically possible applications are meaningful. Although there are already ontologies for describing observable properties and sensing devices such as the NASA SWEET ontologies[1] or the W3C SSN ontology (Compton et al., 2012), a formalization of analysis procedures is missing. In this paper, we address the challenge of meaningful interpolation of a set of environmental observations and of meaningful aggregation in space and time. While there are sophisticated methods for interpolation and aggregation, such as Kriging (Journel and Huijbregts, 1978), determining *which* method is appropriate for *which* kind of data in an automated fashion is an open research question.

To illustrate the problem, consider two real datasets from the atmospheric domain shown in Fig. 1: (i) total emissions of CO_2 for the year 2007 from power plants in Germany[2] and (ii) daily mean concentrations of fine dust (PM_{10}) measured at air quality stations in Germany.[3] Both datasets have an indistinguishable data structure: records of scalar values indexed by points in space and time. Hence, the datasets are often treated as the same in spatial analysis. However, though both datasets can be easily interpolated spatially (right column of Fig. 1), interpolation of points is only *meaningful* for PM_{10} concentration and not for total emissions of CO_2 from power plants. Users unaware of the data semantics may apply inappropriate procedures because they do not distinguish between datasets with equivalent structure representing incommensurable phenomena. A comparable problem concerns the application of *statistical measures* in spatio-temporal aggregation, such as summing up observation values within spatial regions. Computing the sum of the total CO_2 emissions of power plants over Germany may be meaningful, while the sum of PM_{10} concentrations over an area may not.

Though basic prediction and aggregation functionality for spatial data is often available in Geographical Information System (GIS) or statistical software in an adhoc fashion, the choice of a particular method is usually up to the user and its appropriateness is not checked by the system. Furthermore, while measurement scales (Stevens, 1946; Suppes and Zinnes, 1967; Chrisman, 1995) are well established, in many cases, allowable operations are unknown for a dataset. It is, e.g., not meaningful to compute the mean value of a numerical ordinal variable, although it is possible from a computational viewpoint.

We argue that the problem of meaningful prediction and aggregation requires knowledge about the *meaning of data*, i.e., *semantic knowledge*, in machine readable form to help users determine which prediction or aggregation method can be applied to which dataset. In this paper, we suggest a way how the notion of meaningfulness can be operationalized:

1. Formal specifications make some of the knowledge underlying meaningful statistics more explicit and readable for machines.
2. In a rough approximation, a statistical prediction or aggregation method can be said to be meaningfully applicable to a data set, if it is *semantically interpretable* in the observation context of the data. This context can be captured, to a significant degree, by *(semantic) reference systems* (Kuhn, 2003).
3. On this basis, the well-known conceptual distinction between *(marked) point pattern*, *geostatistical variables* and *lattice data* (Illian et al., 2008; Burrough and Mcdonnell, 1998; Cressie and Wikle, 2011), can be made formally explicit.
4. Meaningfulness of prediction can then be checked by testing whether *prediction functions* underlying statistical models *formally correspond* to *observation functions* underlying data, where both are typed by semantic reference systems.
5. Meaningfulness of summation can be checked by testing whether regions over which the data is aggregated formally correspond to the *observed window* of the data to be aggregated.
6. This shows a way to design a recommender tool in Spatial Statistics,[4] in which data and variables can be linked to interpolation and aggregation methods.

The contribution of this paper is a formalization of meaningfulness of spatial prediction and aggregation with respect to datasets. Out of scope are the semantic description of observable properties, of statistical models, and of application problems. We make the case for our notion of meaningfulness based on the two scenarios from the atmospheric domain introduced above (air quality and CO_2 emissions). We test our theory and prove meaningfulness in Isabelle/HOL, a higher-order theorem prover. A preliminary Web Ontology Language (OWL) pattern, which can be used by statistical applications on the Web, and a prototypical implementation in R illustrate some of its potential.

The remainder of this paper is structured as follows. The next section introduces the background of our work including overviews on Spatial Statistics, on spatio-temporal aggregation, on meaningfulness in measurement theory, and on semantic reference systems. Afterwards our functional approach to formalize Spatial Statistical knowledge is described in detail. Then, a description of a prototypical implementation, an R package that extends the package sp, is introduced. Finally, after discussion of our approach, conclusions and directions for future research are presented.

2. Background

This section provides background and a preliminary discussion of the problem of meaningfulness. First, basic variable types are introduced which are relevant for meaningful Spatial Statistics. Then, we describe our notion of spatio-temporal aggregation. Afterwards, we discuss the definition of meaningfulness in measurement theory. Finally, semantic reference systems for space, time, as well as for thematic domains are described, and it is argued why they are useful for making the necessary distinctions.

2.1. Spatial Statistics

Spatial Statistics (Ripley, 1981; Cressie, 1993; Cressie and Wikle, 2011) is a branch of Statistics that deals with spatial and spatio-temporal processes. Although all observations are taken under circumstances that can be characterized by a location and time, in many cases location and time do play a minor role, for instance where controlled experiments in lab conditions eliminate the role of space and time. In case of medical experiments, the subject's identity and age may form the major reference. However, when observations are taken outside a lab, non-controllable factors typically cause them to be correlated in space and/or over time. Spatial Statistical models address such correlations, allow inferences, and are used for the prediction of phenomena in space and time.

For spatio-temporal processes, Cressie and Wikle (2011) use the following notation:

[1] The SWEET ontologies are accessible at http://sweet.jpl.nasa.gov/ontology/.

[2] The data is accessible for free at http://www.carma.org.

[3] The data is available at http://www.eea.europa.eu/themes/air/airbase.

[4] In this paper, we follow the distinction introduced by Cressie and Wikle (2011), where statistics refers to summaries of data and Statistics to the Statistical Science. The same is applied to Spatial Statistics and Geostatistics.