Open Tools for Statistical Analysis in Spatial Data Infrastructures

Benedikt Gräler*, Christoph Stasch*, Benjamin Pross*, Olav Peeters**, Simon Jirka*

* 52°North Initiative for Geospatial Open Source Software GmbH
Martin-Luther-King-Weg 24, 48155 Muenster, Germany
Corresponding author: Benedikt Gräler, b.graeler@52north.org

** Belgian Interregional Environment Agency (IRCEL - CELINE)
Kunstlaan 10-12, B-1210 Brussel, Belgium

Abstract Statistical analysis is performed with dedicated tools and software. A prominent example is the R software for statistical computing. We outline an approach for integrating statistical analysis implemented in R into existing spatial data infrastructures using Open Source tools.

Our approach is based on standards defined by the Open Geospatial Consortium (OGC), in particular the OGC Web Processing Service (WPS). The WPS offers an interoperable web service interface for describing and executing processes that transform some input to some output data (e.g. environmental models, GIS operators, etc.). In our approach, users can upload annotated R scripts to web processing services enabling their re-use by any interested party. The R packages SOS4R and sensorweb4R are used to flexibly assess spatio-temporal data from Spatial Data Infrastructures in R. The Uncertainty Model Language (UncertML) is used to integrate uncertainty measures in the results of the statistical processing.

The usage of the software tools is illustrated by an application running at the Belgian Interregional Environment Agency.

1 Introduction

For many statistical indicators, geographic locations play an essential role for the analysis, especially on an international scale. However, the data are provided by various companies, authorities and institutions across different countries posing a need for standardised interfaces to access, exchange and process the raw data as well as statistics derived from it. Different national definitions and conventions underline the importance of clearly communicated semantics of data and statistics [Scheider et al., 2016]. As an example, daily mean temperature can be the average of minimum and maximum temperatures, the average of temperatures at 7:00, 14:00 and twice 21:00, the arithmetic mean of 24 temperature recordings each hour or - in a conceptual version - the integral of the continuous temperature process [Stasch et al., 2014]. The spatial domain additionally needs to address

the presentation of spatial coordinates. While identifiers like the epsg[1] code exist since the mid 1980s, several data sources only implicitly allow to derive their exact coordinate reference system (CRS).

Standardisation bodies like ISO and the Open Geospatial Consortium (OGC) address the issue of common data models and exchange formats for spatial data infrastructures (SDIs). The OGC also standardises the execution of processing and analysis functionalities including a description how the data is processed. But how can we integrate statistical analysis in SDIs? How can we import data from SDIs in our statistical software and also publish statistical indicators in SDIs?

In the following, we present tools that address these issues by integrating statistical analysis in R into SDIs using OGC web processing services. In addition, we also show how to import spatio-temporal data from SDIs into R, analyse it and making the analysis accessible and executable for others in the web, e.g. through browser-based clients. The approach allows for automatisation of worfklows computing statistical indicators and estimates, for sharing them with others and for describing how the indicators or estimates have been derived.

2 Standards for Spatial Data Infrastructures and Statistics

Standards for spatial data infrastructures defined by ISO and OGC are already well established. The European INSPIRE directive is relying on these standards and fosters standardised exchange of spatial information between the European member states. The core standards for spatial information in SDIs are as follows: The OGC Web Map Service (WMS) [de la Beaujardiere, 2006] allows to access rendered maps as images, e.g. for background maps. The Web Feature Service (WFS) Vretanos [2014] provides access to spatial vector data, e.g. boundaries of administrative regions. The Web Coverage Service (WCS) [Baumann, 2012] allows to access raster data, e.g. satellite images. To standardise web-based discovery, access and tasking of different kinds of environmental sensors and to ease their integration in spatial data infrastructures, the OGC has introduced the Sensor Web Enablement (SWE) initiative [Botts et al., 2007, Bröring et al., 2011]. The Sensor Observation Service [Bröring et al., 2012] allows to attach time series of different phenomena to spatial features. The standard for modelling and encoding such time series data with spatial features is Observations&Measurements (O&M) [**?**]. The core idea in O&M is that a spatial feature (feature of interest) has observed properties who may be observed by sensors or humans. Observations are providing property values at different time stamps. For easing the implementation of web-based map clients and analysis tools, GeoJSON is worth

[1] `http://www.epsg.org/`

to mention as another base standard. Based on GeoJSON various encodings of OGC/ISO data models have been defined including a JSON encoding for O&M.

While these standards address the standardised exchange of spatial information, they lack the integration of processing facilities like, for example, aggregating point data to administrative boundaries. Thus, the OGC has specified the Web Processing Service (WPS) [Mueller and Pross, 2015], a web service interface for integrating such processing facilities in SDIs. In this context, we follow the definition of Hofer [2015], where online geoprocessing refers to the manipulation of geospatial input data to generate novel output data in the web. The WPS can also be used to generate web-based aggregates and statistics of the data, as well as to perform modelling and forecasting tasks. The processes can be executed from web clients running in browsers, from common desktop GIS like ArcGIS or Open Source GIS like Quantum GIS, or can be triggered by any other external application. The WPS interface is relying upon a common process model that also allows to describe the processing functionality in a common way independent of certain programming languages like Java, R or Python.

Similar to the OGC for spatial data, several organisations dealing with statistical data encountered the need to define standard formats for statistical (meta-) data in order to ease the discovery and exchange of statistical datasets. Therefore, they founded the Spatial Data and Metadata eXchange (SDMX) initiative[2] that has released several standards, also as ISO standards. These standards include the definition of a common information model for statistical (meta-)data as well as an XML encoding and guidelines for utilizing the data in web services. While time is explicit, the specification of spatial data is not further specified. We hence think that combining SDMX with standards for spatial information sharing is essential in order to combine statistical analysis and spatial data infrastructures. Though we do not utilise SDMX explicitly, our approach may be considered as a first step towards combining these standard frameworks, since we are integrating SDIs and statistical analysis tools and are also considering metadata about statistics.

3 Approach

Our approach focuses on two aspects: On the one hand, we provide a method and corresponding Open Source tools for integrating statistical analysis written in R into SDIs. On the other hand, we show how to utilise spatio-temporal data from SDIs in R scripts. Figure 1 illustrates the conceptual differences between the two set-ups.

[2] http://sdmx.org/

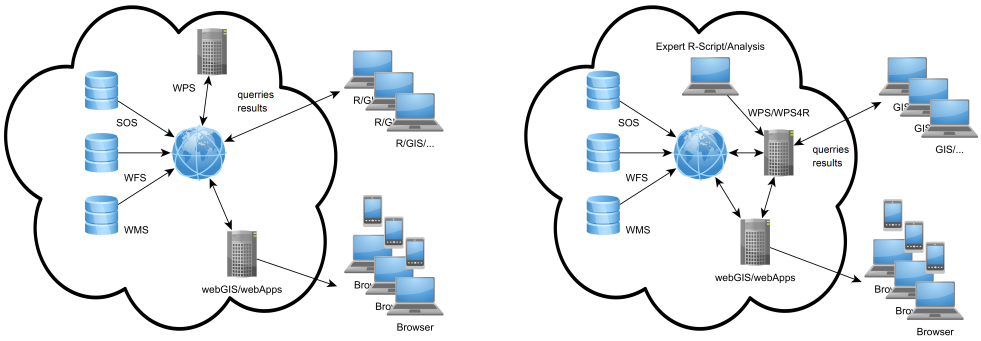

Figure 1: Integrating SDI into local analysis vs. integrating analysis in the SDI.

3.1 Integrating Statistical Analysis in SDIs

Based on the SOS standard or a lightweight REST API, spatio-temporal queries can be made in a standardised form to retrieve time series data at georeferenced locations or spatial fields at certain timestamps from SDIs. Different frontends can be used to retrieve the data ranging from lightweight web clients, over feature-rich GIS tools (e.g. GRASS, ArcGIS) to sophisticated statistical software like R [R Core Team, 2017].

Besides querying the spatio-temporal data, the SOS standard also defines the retrieval of metadata about the data generation procedure such as sensor information, sampling rate, transfer function and units of measure described in the Sensor Model Language (SensorML). SensorML also allows for virtual sensors that represent a statistic that is calculated on a regular basis such as a daily mean, min or max of a time series with finer temporal resolution. This virtual sensor can then be accompanied with standardised metadata on how the statistics have been calculated based on, for instance, linked open data vocabularies. Daily summary statistics already constitutes a simple automation but these statistics can also be wrapped into a exchangeable WPS to perform more complex predefined analyses. During the UncertWeb project [Bastin et al., 2013], this workflow has already been supplemented with standardised exchange of uncertainties going along with model outputs.

Definitions relevant to spatial statistics still need to be extended. Assumptions, design and parameters of a model need to be made explicit to guarantee a meaningful exchange of models. An attempt to describe spatio-temporal random fields via the variogram in a standardised fashion, has been presented in Gräler and Stasch [2012]. This approach poses the limitation of treating a Gaussian random field that follows the dependence structure induced by the variogram function.

Data access can be organised in a client prior to execution, but also carried out automatically by the WPS adding e.g. a spatial buffer to the point of interest or pulling predefined areas and/or timespans. The data source is not limited to a

single data storage, but can be a distributed set of different data providers. This also allows to generate views on the primary data with additional information generating new insights. It also allows to run statistical analysis on data directly in the cloud without having to access chunks of data and running it locally.

3.2 Integrating Data from SDIs in local statistical analysis

R is not only well suited for plain statistical analysis, but also provides packages to handle and analyse spatial data [Bivand et al., 2013]. Therefore, spatial data infrastructures provide large variety of quality assured spatial data. Packages like rgdal [Bivand et al., 2017] allow to easily query data sets from feature services and web coverage services. Integration of spatio-temporal data from the sensor web can be achieved by packages like sos4R [Nüst et al., 2011] and sensorweb4R [Nüst and Autermann, 2015]. Subsequently, the data can be analysed in R with tools for spatial statistics such as RandomFields [Schlather et al., 2015], gstat [Gräler et al., 2016, Pebesma, 2004] or R-INLA [Lindgren and Rue, 2015]. Results can be published as measurements of virtual sensors in transactional sensor observation services or web feature services.

4 Tool set

At 52°North, we develop and implement SOS and WPS standards for the use in productive systems under open source licences. Several extensions have been designed and made available from various research projects and in close cooperation with universities, research institutions, agencies and companies. A conceptual Figure 2 illustrates the tool set integrating geospatial statistical analysis into spatial data infrastructures.

52N SOS

The SOS is the most widely used sensor web standard and defines a web service interface for managing observation data and sensor metadata [Tamayo et al., 2011, Bröring et al., 2012]. It provides information about the observations offered in its service capabilities, e.g. spatio-temporal extent, and offers additional operations for sensor metadata (DescribeSensor) and observation retrieval (GetObservation). New sensors and observations can be inserted using a transactional interface. For modelling and encoding the data, the SOS relies upon the O&M standard and the Sensor Model Language (SensorML) standard for sensor metadata [Botts and Robin, 2014]. The 52°North SOS implementation[3] is based on Java and is fully compliant with the SOS standard. It supports various data sources at the backend including databases (e.g. PostGIS, MySQL), NetCDF files, or custom Web

[3] http://52north.org/sos

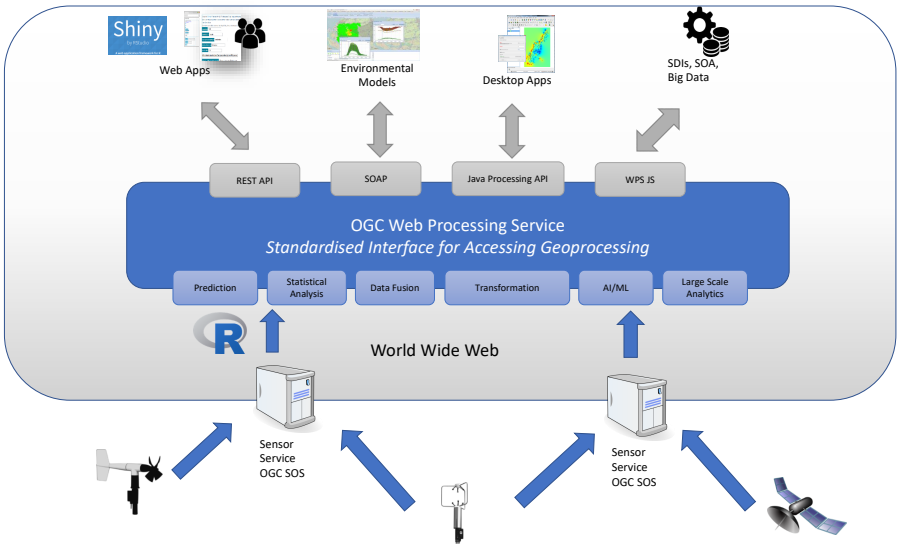

Figure 2: Overview of the tool set interacting with the 52N WPS.

services. Furthermore, the SOS is also available in a bundle with Helgoland, a browser-based client for time series analysis and map views, and with the sensor web REST API, a lightweight REST API that eases client-based access and offers some common preprocessing capabilities to simplify client development.

52N WPS

While most of the OGC standards focus on data models and interfaces for exchanging spatial information, the OGC WPS specifies a common interface for online geoprocessing facilitating an integration of processing facilities in spatial data infrastructures. It may involve simple operations like buffering or intersections, but also complex environmental models. The core operations of WPS consist of operations to obtain service metadata (GetCapabilities), to retrieve process metadata (DescribeProcess) and to run processes (Execute). 52°North has implemented a WPS suite in Java that provides support for common Geoprocessing frameworks such as GRASS GIS, Geotools or Sextante[4]. Custom Java processes can easily be implemented allowing also to publish executables as WPS processes. This also allows to connect existing statistical and environmental models to the web and make them accessible via the standardised WPS interface.

WPS4R

The WPS4R [Hinz et al., 2013] is a generic WPS based on the 52N WPS implementation that wraps annotated R scripts as WPS instances. The annotations

[4] `http://52north.org/wps`

link the in- and outputs of the WPS to the R-script. This allows to make already present analysis R scripts available as a dedicated tool (optionally taking a set of parameters) in the web. As any WPS, this service can then easily be integrated into different systems (Desktop GIS, web GIS, custom tailored web clients) and model chains connecting out- and input from various WPS. This fosters reproducibility and re-usability across organisations and data sets. WPS4R allows to largely simplify an analysis while maintaining a huge degree of expert knowledge. An elaborate R script developed by a statistician can be boiled down to a widget generating a predefined view on the data for different stakeholders. Furthermore, the simple wrapping of an R script allows to easily update and maintain the script in a single location ensuring that all bodies use the same up-to-date method via the WPS4R endpoint. However, this automation also requires an explicit knowledge about the semantics underlying the data and the analysis as not to unintended apply e.g. tools designed for continuous data to discrete outcome.

SOS4R/sensorweb4R

An object oriented SOS-API implementation for R is the sos4R package [Nüst et al., 2011]. It allows to interact with specified endpoints, query meta information as well as to retrieve data. Easy conversions to widely adopted spatial data structures are available easing the application of statistical tools. The R package sensorweb4R [Nüst and Autermann, 2015] utilises the REST endpoint of a 52N SOS featuring a lightweight interaction. The REST-API is not explicitly compliant with any OGC standard, but was developed as a lightweight alternative to the SOAP interface. Long time series load sufficiently fast for more powerful analyses, e.g. using other R packages.

sensorweby/Shiny

Exploiting the Shiny package [Chang et al., 2017] and combining it with the 52N-sensorweb-client leads to a lightweight web GIS platform with extended statistical analysis capabilities. The Shiny frame work allows to wrap a statistical analysis into a web-based GUI by merely using R code. The GUI eases the illustration and access of results of a study and enables the user to freely parametrise and explore the underlying analysis.

UncertML 2.0

UncertML 2.0 has been developed within the Intamap and UncertWeb projects [Bastin et al., 2013]. It is a standardised markup language based on a solid conceptual model for representing probablistic uncertainties. The conceptual model for UncertML is rooted in a basic abstract uncertainty type that is specialised to create distributions (probability distribution functions, including mixture models for multi-modal distributions), statistics (summary statistics, such as moments),

and samples (realisations of random variables). UncertML consists of a dictionary to precisely define the semantics of the uncertainty elements, and can encode both univariate and multivariate random quantities. UncertML focusses on the encoding of uncertainties and is designed to be used with other standards such as O&M, that are used to define the variables being considered, the sampling or model output locations etc.. APIs in Java and JavaScript exist that allow to read and write UncertML.

Uncertainty information can be added to an O&M document in two ways [Stasch et al., 2012]: (1) as additional quality information to the result, or (2) the result itself can be encoded as an uncertain value. In both cases, UncertML is used to model and encode the uncertainties. While O&M is well-suited to observations with spatial vector geometries, grid-based observations are more efficiently encoded using the Network Common Data Format (NetCDF), an established format for exchanging multi-dimensional gridded environmental data. Thus an uncertainty-enabled NetCDF profile (NetCDF-U) has also been developed within UncertWeb.

5 Example: Air quality in Belgium (IRCEL – CELINE)

At the Belgian Interregional Environment Agency (IRCEL – CELINE)[5] the sensorweb4R and sensorweby packages are used for performing advanced statistical analysis making use of decentralised sensorweb data endpoints[6]. Based on sensorweb4R data of external partners is incorporated where a direct database connection would not be possible, e.g. due to security considerations.

The openair[7] package is a feature rich, well documented R-package tailored for air quality professionals. With the combination of sensorweb4R to interact with the REST-API and sensorweby, which integrates a lightweight JavaScript-client[8] to select time series via an interactive map or via a list selector within the Shiny framework, it is possible to build web applications (see Figure 3 for an implementation of the map based selection in Figure 3a, pollutionRose in Figure 3b, scatterPlot in Figure 3c and the schematic design of the web application in Figure 3d). These web applications can be used by domain experts (also without any R-skills) to interactively explore a dataset. Data from different (in- and external) data endpoints can be added to the analysis.

[5] `http://www.irceline.be/en`

[6] More information about the set-up can be found here: `http://blog.52north.org/2015/04/22/advanced-time-series-analysis-on-the-web-with-r/`

[7] `http://www.openair-project.org/`

[8] `https://github.com/52North/js-sensorweb-client`

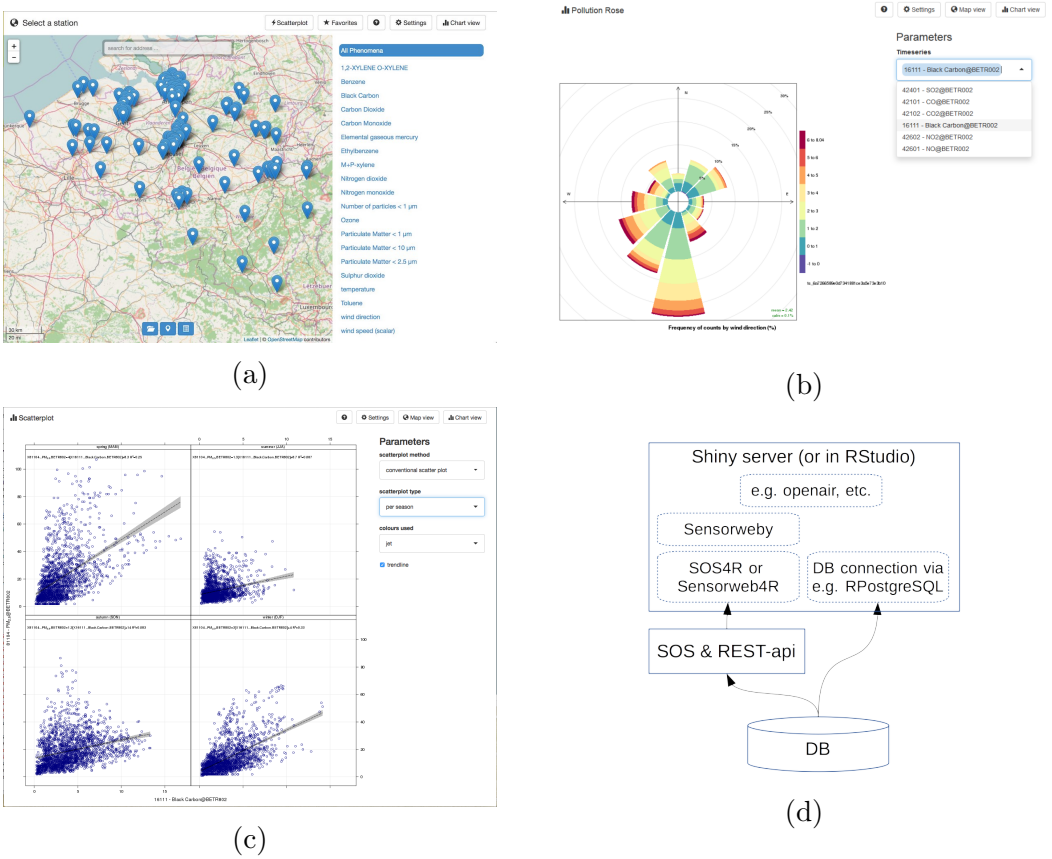

Figure 3: Illustration of different components of the Shiny web application: (a) Map based selection of time series, (b) openair pollutionRose of the selected time series, (c) openair scatterPlot function of the selected time series, (d) schematic design of the entire system.

6 Conclusion: challenges & outlook

While the skeletons to integrate geospatial data and statistics have been presented, the broad integration into daily business still lags behind. It is an important issue to bridge this gap in order to improve the automatic exchange and processing of geospatial data in a meaningful way - an issue that we address in our research projects. The standardisation and integration of spatio-temporal data and metadata about derived statistics is an essential prerequirement. Several challenges have to be resolved subsequently. Scalability, i.e. cloud-readiness, is an issue when services are frequently addressed and suffer from a huge processing load. This goes along with the access control not only of users looking into data, but also of WPSs processing data and sharing results. First attempts to implement a security concept in analysis workflows and model chains are currently explored by the OGC.

Well defined models exists describing basic statistical properties such as a mean value of a series, or how to define probability distributions. However, keeping track of spatial statistics lacks standardised specifications. For instance, the interpolation of point values to a dense grid can rely on very different model assumptions and methods (IDW, kriging - with various variogram models, copula based models, random processes, ...). To communicate and to keep track of these is an open issue of research and standardisation. One way of at least allowing for reproducibility are persistent WPSs that encapsulate the model annotated with metadata.

As different communities are in parallel developing standards and best practices, a mapping between these is necessary to bridge the gap between disciplines for an automatic exchange of data. At least subsets of the SDMX standard can easily be mapped between a subset of the SOS specification. Furthermore, we envision the inclusion of SDMX in an analogous manner as UncertML [Stasch et al., 2012]. A foundation can be provided by developing a common statistical dictionary that is formalised and available on the web allowing an automatic mapping between different concepts of metadata. Bridging these gaps, harmonising concepts and the extension of semantic meta data will largely contribute to ease the exchange of spatio-temporal data and statistical analysis also reflecting the genesis of data and statistics.

References

Lucy Bastin, Dan Cornford, Richard Jones, Gerard BM Heuvelink, Edzer Pebesma, Christoph Stasch, Stefano Nativi, Paolo Mazzetti, and Matthew Williams. Managing uncertainty in integrated environmental modelling: The UncertWeb framework. *Environmental Modelling & Software*, 39:116–134, 2013.

Peter Baumann, editor. *OpenGIS WCS 2.0 Interface Standard - Core, version 2.0.1. OGC 09-110r4.* Open Geospatial Consortium Inc., 2012. URL `http://portal.opengeospatial.org/files/?artifact_id=21273`. Accessed 10 November 2016.

Roger Bivand, Tim Keitt, and Barry Rowlingson. *rgdal: Bindings for the Geospatial Data Abstraction Library*, 2017. URL `https://CRAN.R-project.org/package=rgdal`. R package version 1.2-7.

Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. *Applied spatial data analysis with R, Second edition.* Springer, NY, 2013. URL `http://www.asdar-book.org/`.

Mike Botts and Alexandre Robin, editors. *OGC SensorML: Model and XML Encoding Standard. OGC 12-000.* Open Geospatial Consortium Inc.,

2014. URL http://portal.opengeospatial.org/files/?artifact_id=21273. Accessed 10 November 2016.

Mike Botts, George Percivall, Carl Reed, and John Davidson, editors. *OGC Sensor Web Enablement: Overview And High Level Architecture. OGC 07-165.* Open Geospatial Consortium Inc., 2007. URL http://portal.opengeospatial.org/files/?artifact_id=25562. Accessed 20 March 2011.

Arne Bröring, Johannes Echterhoff, Simon Jirka, Ingo Simonis, Thomas Everding, Christoph Stasch, Steve Liang, and Rob Lemmens. New generation sensor web enablement. *Sensors*, 11(3):2652–2699, 2011. ISSN 1424-8220. doi: 10.3390/s110302652. URL http://www.mdpi.com/1424-8220/11/3/2652/.

Arne Bröring, Christoph Stasch, and Johannes Echterhoff, editors. *OGC Sensor Observation Service Interface Standard. OGC 12-006.* Open Geospatial Consortium Inc., 2012. URL https://portal.opengeospatial.org/files/?artifact_id=47599.

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2017. URL https://CRAN.R-project.org/package=shiny. R package version 1.0.3.

J. de la Beaujardiere, editor. *OpenGIS Web Map Server Implementation Specification. OGC 06-042.* Open Geospatial Consortium Inc., 2006. URL http://portal.opengeospatial.org/files/?artifact_id=21273. Accessed 10 November 2016.

Benedikt Gräler and Christoph Stasch. Flexible representation of spatio-temporal random fields in the model web. In *EGU General Assembly Conference Abstracts*, volume 14, page 4617, 2012.

Benedikt Gräler, Edzer Pebesma, and Gerard Heuvelink. Spatio-temporal interpolation using gstat. *The R Journal*, 8:204–218, 2016. URL https://journal.r-project.org/archive/2016-1/na-pebesma-heuvelink.pdf.

Matthias Hinz, Daniel Nüst, Benjamin Proß, and Edzer Pebesma. Spatial statistics on the geospatial web. In *The 16th AGILE International Conference on Geographic Information Science, Short Papers*, 2013.

Barbara Hofer. Uses of online geoprocessing technology in analyses and case studies: a systematic analysis of literature. *International Journal of Digital Earth*, 8(11):901–917, 2015. doi: 10.1080/17538947.2014.962632. URL http://dx.doi.org/10.1080/17538947.2014.962632.

Finn Lindgren and Håvard Rue. Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63(19):1–25, 2015. URL http://www.jstatsoft.org/v63/i19/.

Matthias Mueller and Benjamin Pross, editors. *OpenGIS WPS 2.0 Interface Standard. OGC 14-065* . Open Geospatial Consortium Inc., 2015. URL `http://portal.opengeospatial.org/files/?artifact_id=21273`. Accessed 10 November 2016.

Daniel Nüst and Christian Autermann. *sensorweb4R: Connect R to the Sensor Web Client API for downloading timeseries data.*, 2015. URL `https://github.com/52North/sensorweb4R`. R package version 0.1.

Daniel Nüst, Christoph Stasch, and Edzer Pebesma. Connecting R to the sensor web. In *Advancing Geoinformation Science for a Changing World*, pages 227–246. Springer, 2011.

Edzer J. Pebesma. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30:683–691, 2004.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

Simon Scheider, Benedikt Gräler, Edzer Pebesma, and Christoph Stasch. Modeling spatiotemporal information generation. *International Journal of Geographical Information Science*, 30(10):1980–2008, 2016.

Martin Schlather, Alexander Malinowski, Peter J. Menck, Marco Oesting, and Kirstin Strokorb. Analysis, simulation and prediction of multivariate random fields with package RandomFields. *Journal of Statistical Software*, 63(8):1–25, 2015. URL `http://www.jstatsoft.org/v63/i08/`.

Christoph Stasch, Richard Jones, Dan Cornford, Martin Kiesow, Matthew Williams, and Edzer Pebesma. Representing uncertainties in the sensor web. In *Proceedings of workshop Sensing a Changing World*, volume 2, 2012.

Christoph Stasch, Simon Scheider, Edzer Pebesma, and Werner Kuhn. Meaningful spatial prediction and aggregation. *Environmental Modelling & Software*, 51:149–165, 2014.

Alain Tamayo, Pablo Viciano, Carlos Granell, and Joaquín Huerta. *Empirical Study of Sensor Observation Services Server Instances*, pages 185–209. Springer Berlin Heidelberg, 2011. URL `http://dx.doi.org/10.1007/978-3-642-19789-5_10`.

Panagiotis (Peter) A. Vretanos, editor. *OpenGIS Web Feature Service 2.0 Interface Standard. OGC 09-025r2.* Open Geospatial Consortium Inc., 2014. URL `http://portal.opengeospatial.org/files/?artifact_id=21273`. Accessed 10 November 2016.