21. Geoscience and Geoinformation – From data acquisition to modelling and visualisation

Nils Oesterling, Massimiliano Cannata, Michael Sinreich, Elmar Brockmann

Swiss Geological Survey
Swiss Geodetic Commission
Swiss Geophysical Commission
Swiss Hydrogeological Society

TALKS:

21.1 Araya D., Kumi M., Podgorski J., Berg M.: Exploring geographically weighted logistic regression to predict groundwater fluoride contamination in Ghana

21.2 Cannata M., Antonovic M., Oesterling N., Brodhag S.: Borehole Data Management System v1.0


21.4 Oriani F., Mariethoz G.: Missing data imputation for incomplete multisite time series: a data pattern-based approach

21.5 Podgorski J., Berg M.: Second-generation global risk map of groundwater arsenic contamination

POSTER:

P 21.1 Zakeri F., Mariethoz G.: Multiple-Point Geostatistical Simulation Models in Remote Sensing Applications
21.1 Exploring geographically weighted logistic regression to predict groundwater fluoride contamination in Ghana

Dahyann Araya1, Michael Kumi2, Joel Podgorski1, Michael Berg1

1 Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department Water Resources and Drinking Water, 8600 Dübendorf, Switzerland (dahyann.araya@eawag.ch).
2 CSIR, Water Research Institute, Environmental Chemistry Division, Box695, Tamale, Ghana.

About 25% of Ghana’s population lacks access to essential drinking water services, and 73% have no access to reliable and clean water sources (Safe Water Network 2017). The population, especially in rural areas, is dependent on groundwater for drinking. However, dental and skeletal fluorosis related to groundwater fluoride contamination is well documented across the north of the country (Figure 1), where the majority of the population lives in rural areas (Apambire et al. 1997). The World Health Organization (WHO) has set a maximum concentration guideline of 1.5 mg/L for drinking water. Different geological, topographical and climatic characteristics throughout Ghana are responsible for the distribution of fluoride contamination in groundwater.

Logistic regression (LR) has been used to study the relationship between geogenic groundwater contamination and various environmental variables. However, this type of modelling method applies a global approach to regression, i.e., it assumes that relationships between variables do not vary across space. Therefore, LR runs the risk of not addressing the heterogeneous and complex characteristics of aquifer environments.

The first law of geography states that closer things tend to be more related than things that are farther apart, thus recognizing spatial dependence and self-correlation. Geographically weighted regression (GWR) attributes weights to models according to location and distance (Brunsdon et al. 1996). It adapts the structure of the model over space to address the local statistical associations among environmental variables. Therefore, the closer a point is to an observation location, the higher the probability to be related to this observation. Hence, GWRs allow heterogeneity in spatial relationships to be addressed through local rather than global regression models.

Figure 1. High concentration of fluoride in groundwater in Ghana, the red dots show a concentration greater than 1 mg/L.

The present study explores GWLR for probability modelling of groundwater fluoride contamination throughout Ghana. Data of 3150 groundwater wells were collected to map and model the occurrence of fluoride exceeding 1.0 and 1.5 mg/L. The GWLR and LR models were created using geospatial predictor variables of topography, geology, soil, hydrology, climate,
and ecology. Preliminary model validation results indicate that GWLR performs better than LR. Fluoride hotspots were identified in the Upper East and Northern Region with over 95% confidence using Getis Ord Gi* statistic. Geological and soil variables were identified as positive predictors of groundwater fluoride contamination. The derived prediction map of groundwater fluoride contamination throughout Ghana highlights high-risk areas. This map could be used to raise awareness and understanding, and to inform actions of local scope in order to avoid or mitigate fluoride water contamination-related risks.

This study is part of the Groundwater Assessment Project GAP where other geospatial statistical methods such as multivariate logistic regression (Rodríguez-Lado et al. 2013) and machine learning (Podgorski and Berg 2020) have been used to assess and predict geogenic groundwater contamination.

REFERENCES

21.2
Borehole Data Management System: feature

Massimiliano Cannata¹, Milan Antonovic¹, Nils Oesterling², and Sabine Brodhag²

¹ SUPSI, Istituto scienze della Terra, DACD, Canobbio, Switzerland (massimiliano.cannata@supsi.ch)
² Federal Office of Topography swisstopo, Wabern, Switzerland

Most of the time boreholes data, particularly those collected in the past, are in the form of static data reports that describe the stratigraphy and the related characteristics; these data types are generally available as paper documents, or static files like .pdf of images (.ai). While very informative, these documents are not searchable, not interoperable nor easily reusable, since they require a non negligible time for data integration. Sometime, data are archived into database. This certainly improve the find-ability of the data and its accessibility but still do not address the interoperability requirement and therefore, combining data from different sources remain a problematic task. To enable FAIR borehole data and facilitate the different entities (public or private) management Swisstopo (www.swisstopo.ch) has funded the development of a Web application named Borehole Data Management System (BDMS) [1] that adopt the borehole data model () [2] implemented by the Swiss Geological Survey. From the first beta release (2019) several improvements to the platform has been implemented leading to the first official release of the platform (v1.0) officially available on www.swissforages.ch. The latest released features includes:

- Borehole document storage
- Interface customization
- Improved access & authorization management
- External WMS/WMTS background map support
- Added user feedbacks form
- Handling of personalized and versioned terms of service
- Enhanced bulk data import
- Minor enhancements and bug fixes

Figure 1. Screenshoot of Web interface of the swissforages platform with a test data input.

REFERENCES
21.3 
**Challenges for Integrating Web-EGIS Functionalities in the Environmental Research Data Portal EnviDat**

Ionuț Iosifescu Enescu, David Hanimann, Dirk Nikolaus Karger, Gian-Kasper Plattner, Dominik Haas-Artho, Rebecca Kurup Buchholz, Lucia Espona, Niklaus E. Zimmermann, Loïc Pellissier & Martin Hägeli

EnviDat is the environmental data portal of the Swiss Federal Research Institute WSL covering datasets from forest, landscape, biodiversity, natural hazards, and snow and ice research. EnviDat offers a wide range of services that support Research Data Management (RDM) and Open Science. For example, EnviDat actively implements the FAIR (Findability, Accessibility, Interoperability and Reusability) principles by offering formal publication of research data with proper citation information and Document Object Identifiers (DOIs) [Iosifescu et al. 2018, 2019].

Recently EnviDat evolved beyond the set of core features expected from a research data management portal with a new, built-in repository. This evolution is driven by the diversity of researchers’ requirements for a specialized environmental data portal. Examples include (i) immediate access to data collected by automatic measurements stations and (ii) environmental data visualization on charts and maps, with geoservices for large geodatasets. The latter is highly relevant for the topic “Geoscience and Geoinformation – From data acquisition to modelling and visualization”.

The environmental research data published in EnviDat is georeferenced and thus inherently spatial. Consequently, there are several user requirements related to mapping and visualization of heterogeneous environmental data in vector and raster form directly in the publishing portal. In this context we first review the concept of a Web-based Environmental Geospatial Information Systems (Web-EGIS). A Web-EGIS is a special case of Web-GIS, which is designed to integrate the functionalities of geospatial information systems with “environmental geo-referenced data and services” based on a generic hypercube-based data organization and visualization [Iosifescu et al. 2015]. This concept can guide the extension of the existing EnviDat portal with geoportal components, such as the online mapping of geospatial data.

The mapping requirements also pose, however, several important challenges. First, a hypercube-based architecture cannot be applied directly because it requires a common geospatial database with a unified schema for all data sets. In EnviDat, the data deposited by researchers is heterogeneous in file formats which were produced with highly different schemata, workflows and software. Although geospatial databases and technologies such as PostgreSQL and PostGIS are already an integral part of EnviDat’s architecture, harmonizing and inserting user-uploaded data into a single geospatial database is not possible at this time. Migrating the data into a common database would require both a significant amount of labor and a harmonized schema that allows to organize the data from various environmental research fields in a unified manner. Consequently, the hypercube-based concept needs to be simplified and adapted to the realities of an existing data repository.

Second, the envisioned mapping capabilities must be supported by an extensible Spatial Data Infrastructure (SDI). Implementing and managing an SDI requires substantial computing and storage resources that are significantly larger than what is available for running a research data management repository such as EnviDat. Beside dedicated servers for providing the necessary mapping and geoprocessing services, additional storage is needed for optimizing data for GIS display. According to our tests, the additional pyramids and multi-band raster GeoTIFFs increase the total data storage size from 1.3 to 2.5 times, despite using raster compression algorithms such as LZW or Packbits for the data files, and the JPEG lossy compression for the associated pyramids.

Third, some environmental data sets are quite large and can contain a complex directory structure. Relevant examples in this context are the “High resolution climate data for Europe” (https://www.envidat.ch/#/metadata/eur11) or the “Climatologies at high resolution for the Earth's land surface areas (CHELSA)” (https://www.envidat.ch/#/metadata/chelsa-climatologies). Terabytes of georeferenced raster data are produced by climate models, and their size is expected to increase in the petabytes range with a complex directory structure containing a large number of individual files. Obviously, raster time-series with up to 10'000 layers for a single data set pose quite a performance challenge: such vast number of layers can not be configured and served interactively by traditional SDI map services. Subsetting or clipping smaller spatial areas from a global dataset would also take a significant amount of time with standard geoprocessing services, which is no longer adequate for real-time interactions with many users. Moreover, well-known mapping frameworks are not adapted for providing user-friendly temporal navigation for a huge number of layers.

The above challenges were uncovered during an early proof-of-concept for the integration of selected Web-EGIS functionalities for raster data in EnviDat and shown in Figure 1. Nevertheless, this proof-of-concept demonstrates some of the possible features that could become part of EnviDat in the future, once these challenges are resolved. Some features are (i) automatic creation of the necessary geoservices based on a user-provided configuration file, (ii) automatic display of the data in the data-set map (as a 2D map or on a globe), (iii) temporal navigation for a large number of layers of a time
series, (iv) split view and difference map for visually comparing layers at different times, and (v) querying and displaying point data on charts for the entire time series. The mock-up implementation of these features in the proof-of-concept allows us to precisely test and refine the generic user requirements, thus improving the estimates of the necessary implementation effort and overall costs. These estimates also help us taking better informed decisions about which features to keep or drop in any future prototype and release phases.

The many challenges mentioned above also represent opportunities for further improving the exchange of scientific information in the environmental domain. Geospatial technologies have the potential to become a central element for any specialized environmental data portal, triggering the convergence between publishing repositories and geoportals. Ultimately these new requirements demonstrate the increased expectations that institutions and researchers have towards the future capabilities of research data portals and repositories in the environmental domain. With EnviDat, we are ready to take up these challenges over the years to come.

The many challenges mentioned above also represent opportunities for further improving the exchange of scientific information in the environmental domain. Geospatial technologies have the potential to become a central element for any specialized environmental data portal, triggering the convergence between publishing repositories and geoportals. Ultimately these new requirements demonstrate the increased expectations that institutions and researchers have towards the future capabilities of research data portals and repositories in the environmental domain. With EnviDat, we are ready to take up these challenges over the years to come.

Figure 1. Proof-of-concept for integrating Web-EGIS functionalities in EnviDat

REFERENCES
21.4  
**Missing data imputation for incomplete multisite time series: a data pattern-based approach**

Fabio Oriani¹, Gregoire Mariethoz¹

¹ Faculty of Geosciences and Environment, University of Lausanne, Geopolis, CH-1015 Lausanne  
(fabio.oriani@protonmail.com)

In the era of big data, missing data imputation (gap filling) remains a task of primary importance to obtain homogeneously informed and representative datasets for environmental studies, monitoring, and numerical modelling. We propose here a data reconstruction algorithm, called Vector Sampling (VS), for missing data imputation in multisite measurement networks.

Based on a multiple-point resampling scheme, VS first retrieves the pattern of data present for a given day from multiple stations. Then it looks for the \( k \) most similar patterns in the historical record from the same stations. Finally, it draws an estimate of the missing data by applying a weighted mean on the \( k \) historical data for the target stations. The process is repeated for all missing days to complete all the dataset. The VS algorithm presents three main advantages with respect to usual imputation approaches:

I) Unlike traditional \( k \)-nearest neighbor (\( k \)-nn) algorithms, it can handle incomplete training datasets robustly: this means it can use the available dataset as both historical (training) and target dataset. The user can simply input an incomplete dataset to obtain the complete version, with the estimations based on its own data patterns.

II) It does not require building statistical models, only having the parameter \( k \) which can be manually setup, taking typical values between 10 and 20.

III) Unlike most machine-learning techniques, it does not require computationally intensive training procedures, drawing on-the-fly estimations based on simple matrix operations.

VS has been applied to daily rainfall networks from Denmark, Australia, and Switzerland, and compared with inverse-distance, kriging, and \( k \)-nn interpolations [1]. On flat terrains, with spatially uniform rain events, kriging interpolation tends to minimize the error, while, in mountainous regions with non-stationary rainfall statistics, VS can better recover the complex rainfall patterns. VS turns out to be a convenient option for routine or automatic applications, which can be applied to continuous environmental variables with the only requirement of a representative historical dataset.

VS for Python3 is open-source and freely available at: https://bitbucket.org/orianif/vs/src/master/

REFERENCES

21.5
Second-generation global risk map of groundwater arsenic contamination

Joel Podgorski\textsuperscript{1,2} & Michael Berg\textsuperscript{1,3}

\textsuperscript{1} Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department Water Resources and Drinking Water, 8600 Dübendorf, Switzerland (joel.podgorski@eawag.ch, michael.berg@eawag.ch)
\textsuperscript{2} University of Manchester, Department of Earth and Environmental Sciences, Manchester, United Kingdom
\textsuperscript{3} UNESCO Chair on Groundwater Arsenic within the 2030 Agenda for Sustainable Development, School of Civil Engineering and Surveying, University of Southern Queensland, 4350 QLD, Australia

Naturally occurring arsenic in groundwater affects millions of people around the world. Odorless and tasteless, arsenic can present significant hazards to human health in the concentrations frequently found in nature. Due to generally not being measured in groundwater quality analyses and the fact that its health effects can be similar to those stemming from other causes, its presence can go undetected for a long time. In order to assess the extent of this problem, we have used machine learning to create an up-to-date global prediction map of groundwater arsenic concentrations exceeding the WHO drinking water guideline of 10 µg/L.

Over 200,000 measurements of arsenic concentration in groundwater were compiled from a wide variety of sources while excluding measurements known to have originated from a depth greater than 100 m. These were aggregated into 58,555 data points by taking the geometric mean of concentrations falling within 1-km square pixels (Fig. 1), which corresponds to the resolution of the predictor variables used. A collection of 52 spatially continuous predictor variables with global coverage representing various climatic, geologic, soil and other parameters related to the dissolution and accumulation of arsenic in groundwater was assembled. Recursive feature elimination was employed to identify a subset of 11 variables, which were then used in a random forest model grown with 10,001 trees.

The model was then validated against a test dataset. Despite a prevalence of high values (>10 µg/L) of only 0.22 in the data, the model performs well in predicting both high values (sensitivity: 0.79) and low values (specificity: 0.85) at a probability cutoff of 0.50. Likewise, the model’s Area Under the Curve (AUC), which considers the full range of possible cutoffs and generally ranges from 0 to 1, has the very high value of 0.89 with the test dataset

The resulting model predicts areas with high arsenic concentrations in groundwater on all continents. Known areas of groundwater arsenic contamination are identified as are new areas of potential geogenic arsenic contamination, including large sections of Central Asia, the Sahel region and Okavango Delta in Africa, and parts of the Arctic. Combining the global arsenic prediction model with household groundwater-usage statistics, we estimate that 94-220 million people are potentially exposed to high arsenic concentrations in groundwater. As groundwater is increasingly utilized to support a growing population and buffer against increasing water scarcity due to a changing climate, this model will help raise awareness, identify suitable areas for safe wells and guide where testing for arsenic should be prioritized.
Multiple-Point Geostatistical Simulation Models in Remote Sensing Applications

Fatemeh Zakeri¹, Gregoire Mariethoz¹

¹ Faculty of Geosciences and Environment, University of Lausanne, Geopolis, CH-1015 Lausanne
(fatemeh.zakeri, gregoire.mariethoz@unil.ch)

Multiple-point geostatistical simulation (MPS) aims at generating realizations of a spatial phenomenon based on training images (TIs) while preserving complex patterns [1]. These methods have been increasingly used for analyzing remotely sensed data over the past few decades. This presentation will review some uses of MPS to remote sensing and summarize the main concepts. Common applications include downscaling, improving land use/land cover classifications, and gap-filling.

The resolution of satellite imagery is often insufficient in geoscience applications. The sub-pixel information is typically not uniquely determined, making a case for using geostatistical approaches to quantify uncertainty [2]. The most widely used MPS methods in downscaling are Direct Sampling (DS), Single Normal Equation Simulation (SNESIM), and Filter-based Simulation (FILTERSIM). For instance, DS, which is suited for both categorical and continuous variables, performs downscaling by sampling small-scale patterns based on a pair of high-resolution and low-resolution TIs that provides a correspondence between resolutions.

One of the main issues affecting classification accuracy is “salt-and-pepper” noise (i.e., isolated pixels) or spatially discontinuous appearance, which usually occurs in classification results. MPS models have been used to improve the accuracy of classification maps, often as a post-processing step. These methods usually incorporate the spatial and spectral information or apply the MPS to the obtained classified images directly to improve the accuracy of them.

Gap-free satellite data are often unavailable due to obstructions, such as clouds or clouds shadows [3]. MPS methods have been used to fill these gaps with realistic patterns. The general approach is to use a training image presenting a similar structure as the gaps to be filled, such as the same scene taken at a different date, or an image of another texturally similar area. Using multivariate MPS simulation allows the inclusion of covariates in the gap-filling process, such as elevation or land cover classes.

Our review shows that MPS approaches have a strong potential for applications that need to derive spatial information. However, more research is required to apply different MPS methods in new domains, such as detecting changes or information fusion. Moreover, future studies need to tailor the existing MPS algorithms to specific remote sensing applications.

REFERENCES