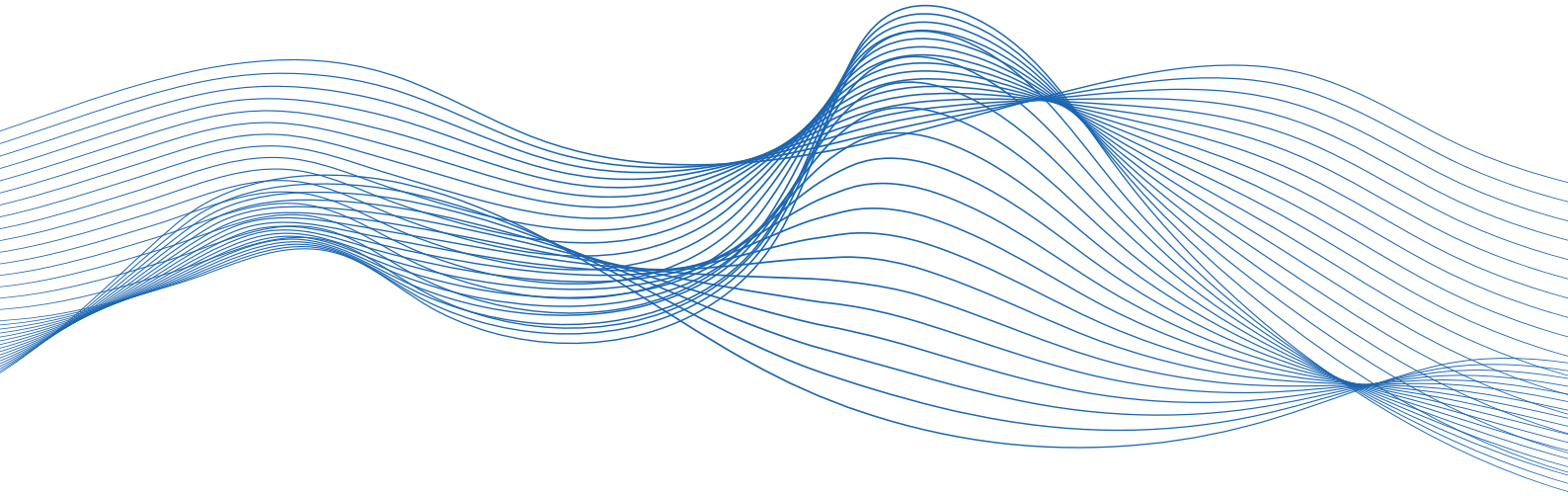
Ensemble Kalman Filter for Pollution Source Characterization in Water Supply Systems

13TH INTERNATIONAL CONFERENCE ON GEOSTATISTICS FOR ENVIRONMENTAL APPLICATIONS

geoENV 2020 PARMA

JUNE 18, 2021



PROCEEDINGS OF geoENV2020

Andrea Zanini & Marco D'Oria, Editors

UNIVERSITÀ DI PARMA

Andrea Zanini & Marco D'Oria
Editors

13TH INTERNATIONAL CONFERENCE ON GEOSTATISTICS FOR ENVIRONMENTAL APPLICATIONS

First Edition 2021

13th International Conference on Geostatistics for Environmental Applications: geoENV2020

Editors: Andrea Zanini & Marco D'Oria

ISBN: 979-12-20341-59-2

handle: https://hdl.handle.net/1889/4373

Preface

The 13th International Conference on Geostatistics for Environmental Applications (geoENV2020) was scheduled in Parma, Italy on July 2020. The international health crisis affected the conference, which was initially postponed to June 2021 and eventually replaced by a one-day virtual event on June 18, 2021 with the presentations of the keynote lecturers.

This book contains the abstracts and extended abstracts submitted to the conference and focusing on geostatistics applied to different fields such as: climate change, ecology, natural resources, forestry, agriculture, geostatistical theory and new methodologies, health, epidemiology, ecotoxicology, inverse modeling, multiple point geostatistics, remote sensing, soil applications, spatio-temporal processes and surface and subsurface hydrology. The Scientific Committee initially selected about 100 abstracts and 68 contributions were confirmed to be published in these proceedings.

The next geoENV conference (geoENV2022) will be held in Parma, Italy on June 2022. We expect more colleagues from all over the world to join this international event next year.

Organizing Committee

Andrea Zanini (Co-Chair)

Marco D'Oria (Co-Chair)

Maria Giovanna Tanda

Valeria Todaro

Jaime Gómez-Hernández

Philippe Renard

Scientific Committee

Teresa Albuquerque – IPCB ICT CERNAS – Portugal
Denis Allard – INRA – France
Peter Atkinson – Lancaster University – United Kingdom
Leonardo Azevedo – CERENA, Instituto Superior Técnico – Portugal
Patrick Bogaert – Université catholique de Louvain – Belgium
Peter Bossew – German Federal Office for Radiation Protection (BfS) – Germany
Ilaria Butera – Politecnico di Torino – DIATI – Italy
Eduardo Cassiraga – Universitat Politècnica de València – Spain
Alessandro Comunian – Università degli Studi di Milano – Italy
Nadim Copty – Bogazici University – Turkey
Sandra De Iaco – University of Salento – Italy
Dimitri D'Or – Ephesia Consult – Belgium
Aldo Fiori – Roma Tre University – Italy
Chantal de Fouquet – Mines Paris Tech – France
Harrie-Jan Hendricks-Franssen – Forschungszentrum Julich GmbH – Germany
Jaime Gómez-Hernández – Universitat Politècnica de València – Spain
Pierre Goovaerts – BioMedware, Inc. – United States
Dario Grana – University of Wyoming – United States
Alberto Guadagnini – Politecnico di Milano – Italy
Claus Haslauer – University of Stuttgart – Germany
George P. Karatzas – Technical University of Crete – Greece
Gregoire Mariethoz – University of Lausanne – Switzerland
Jennifer McKinley – Queen's University Belfast – United Kingdom
Julian Ortiz – Queen's University – Canada
Monica Palma – University of Salento – Italy
Edzer Pebesma – University of Muenster – Germany
Maria João Pereira – CERENA, Instituto SuperiorTécnico – Portugal
Pierre Petitgas – IFREMER – France
Philippe Renard – University of Neuchâtel – Switzerland
Javier Rodrigo Ilarri – Universitat Politècnica València – Spain
Thomas Romary – Mines Paris Tech – France
Xavier Sanchez-Vila – Universitat Politecnica de Catalunya – Spain
Amilcar Soares – CERENA, Instituto SuperiorTécnico – Portugal
Emmanouil Varouchakis – Technical University of Crete – Greece
Hans Wackernagel – Mines Paris Tech, PSL University – France

Keynote lectures

Peter Atkinson, Lancaster University

Implications of the Point Spread Function for downscaling and data fusion in remote sensing

Alessandra Menafoglio, Politecnico di Milano

An object oriented approach to the analysis of spatial complex data

Paula Moraga, King Abdullah University of Science and Technology

How geostatistics can help with decision making in global health – case studies in tropical disease mapping

* Extended abstract

* Extended abstract

* Extended abstract

* Extended abstract

DOWNSTREAM PROPAGATION OF GEOCHEMICAL FOOTPRINTS IN THE TIBER RIVER CATCHMENT (CENTRAL ITALY) ASSESSED THROUGH A CODA APPROACH

Caterina Gozzi (1)* - Antonella Buccianti (1) - Gerd Rantitsch (2) - Orlando Vaselli (1) - Barbara Nisi (3)

University of Florence, Department of Earth Sciences (1) - Montanuniversität Leoben, Geology and Economic Geology (2) - Cnr-igg Institute of Geosciences and Earth Resources (3)

** Corresponding author: caterina.gozzi@unifi.it*

Abstract

Climate change scenarios project an exacerbation of spatiotemporal variations in water cycle dynamics. These changing conditions also affect the Mediterranean catchments, which are suffering drier climate and declining water resources. Within complex and dynamically inter-connected structures of drainage systems, these variations also have a great influence on the inner dynamics of the sediment routing system. Sediment particles and dissolved solids from erosional source regions rarely present a smooth, uniform and continuous pattern from sources to sinks, but rather exhibit a non-linear behavior. In fact, forcing effects such as climate change, anthropic impacts and topographic gradients are able to strongly influence these efflux mechanisms. Large alluvial systems such as the Tiber River catchment, the largest in central Italy (17,156 km^2), have the ability to homogenize or even radically transform incoming geochemical signals during downstream propagation due to the joint contribution of heterogeneous geological-topographical environments and multiple anthropic pressures (Gozzi et al., 2019). In the present study, the chemical composition of stream sediments collected in 2018 from the Tiber River and its main tributaries is analyzed and interpreted in the light of a wide dataset of hydrochemical data. The research aims to investigate physicochemical weathering and transport processes from the up- to the down-reaches of the basin. In order to achieve this goal, advanced statistical methods and graphical-numerical elaborations based on a Compositional Data Analysis approach were applied to process the acquired data. Changes in data variability and pairwise robust Mahalanobis distances were calculated in a compositional context to investigate the transmission of the chemical footprints from the source to the sink and its resilience to changing environmental conditions. In this framework, Compositional Data Analysis appears to be the adequate tool to capture all these features, enabling the detection of potential pollution events or climate-induced modifications.

Gozzi, C., Filzmoser, P., Buccianti, A., Vaselli, O., & Nisi, B., 2019. Statistical methods for the geochemical characterisation of surface waters: The case study of the Tiber River basin (Central Italy). Computer & Geosciences 131, 80-88.

THE EFFECT OF GLOBAL WARMING ON THE MEDITERRANEAN REGION USING A TYPICAL SPECIES (ARBUTUS UNEDO L.)

Maria Margarida Ribeiro (1)* - Alice Maria Almeida (2) - Maria João Martins (3) - Manuel Lameiras Campagnolo (3) - Saki Gerassis (4) - Paulo Fernandez (5) - Teresa Albuquerque (6) - José Carlos Gonçalves (1)

Research Centre for Natural Resources, Environment and Society (cernas) - Instituto Politécnico de Castelo Branco (1) - C4 — Centro de Competências em Cloud Computing (c4-ubi), Universidade da Beira Interior (2) - Forest Research Centre, School of Agriculture, University of Lisbon (3) - University of Vigo, Department of Natural Resources and Environmental Engineering (4) - Instituto Politécnico de Castelo Branco, Escola Superior Agrária (5) - Research Centre for Natural Resources, Environment and Society (cernas) - Instituto Politécnico de Castelo Branco and Ict-university of Évora (6)

** Corresponding author: mataide@ipcb.pt*

Abstract

The Mediterranean region experience now temperatures ~1.3°C higher than during 1880-1920, compared with an increase of ~0.85°C worldwide. Impacts on the Mediterranean forest due to climate change suggest a trend in species migration from south to north and inland to coastal areas. In addition, under this threat, forests may disappear from drier areas. The risk of forest fires will increase a warmer and drier climate and can be further increased by the accumulation of highly flammable biomass in summer. The impact on the forest economy can be extremely severe. It is expected a reduction in productivity, an increase in fire risk, and in the risk of pests and diseases, making forest investment unattractive, resulting in increased forest abandonment. Our aim is to model, from an ecological point of view, a typically Mediterranean species widely distributed in the Mediterranean region, the strawberry tree (Arbutus unedo L.). Through niche modelling, it is possible to reveal the impact of environmental factors on the distribution of strawberry tree habitats using contrasting global warming scenarios. Two different approaches were used aiming at modelling the species' environmental suitability: maximum entropy (MaxEnt) and the convex-hull with Tukey depth approach (CH-Tukey). Both methods were applied to 11487 species presence points, together with a matrix of environmental covariates (bioclimatic and physiographic attributes). Bayesian networks were used in a GIS-based decision-making system to infer the value of the most relevant environmental covariates conditioning strawberry tree environmental suitability. Current and future climate data were obtained from WorldClim. The strawberry tree's vulnerability to the effects of global climate change was examined using two emission scenarios (RCP 4.5 and 8.5), leading to the prediction of species' spatial distribution in 2050 and 2070.

Forest management policy should reflect the impact of climate change on usable areas for forestry, considering species adapted to Mediterranean regions and forest fires such as the strawberry tree.

Acknowledgements: This work was supported by operation Centro-01-0145-FEDER-000019 - C4 - Centro de Competências em Cloud Computing, cofinanced by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio à Investigação Científica e Tecnológica - Programas Integrados de IC&DT; and by Fundação para a Ciência e a Tecnologia I.P. (FCT) through the Forest Research Centre, Portugal, project UID/AGR/00239/2019 and through the CERNAS, Portugal, project UID/AMB/00681/2019.

IDENTIFICATION OF CLIMATE IMPACTS ON WATERSHEDS USING UNSUPERVISED MACHINE LEARNING

Velimir V. Vesselinov (1)*

Los Alamos National Laboratory, Computational Earth Science (1)
** Corresponding author: vvv@lanl.gov*

Abstract

Watersheds are complex systems in which various physical processes impact their state and behavior. The model representation of watersheds is challenging and typically demands computationally intensive multi-physics, high-resolution numerical simulations. The calibration of watershed models, as well as model prediction of the future conditions in the watersheds, is very challenging as well. Here we present the application of advanced unsupervised machine learning (ML) methods to understand watershed behavior better. ML is applied to extract dominant features present in the model outputs which characterize import physics processes.

Unsupervised Machine Learning (ML) methods are powerful data-analytics tools capable of extracting important features hidden in large datasets without any prior information. Recently, we have developed a series of algorithms for unsupervised ML are based on matrix and tensor decomposition. Our ML codes (called NMFk/NTFk/NTNk) are on GitHub (https://github.com/orgs/TensorDecompositions). The web site includes are also example and test problems how our ML codes can be applied to solve a diverse set of problems.

A series of watersheds in Southwestern U.S. has been simulated using a set of different models. The model outputs have been analyzed to characterize model predictions related to attributes such as the evaporation, precipitation, stream flow, soil moisture, etc. A series of common temporal features have been identified across the watersheds, which allow us to group the basins with similar temporal behavior. Differences in the temporal watershed behavior are classified as well. The analyses allow for efficient transfer information between watershed to predict unknown feature states and behavior.

FIRST GEOSTATISTICAL MAPPING OF INDOOR RADON CONCENTRATIONS DATA IN FRANCE

Jean-Michel Metivier (1) - Claire Greau (1)

Irsn (1)

** Corresponding author: jean-michel.metivier@irsn.fr*

Abstract

Radon is a colorless and odorless radioactive gas, naturally present in soils, in greater quantities in granite, volcanic massifs, some shales and sandstones. The health risk is mainly due to the presence of radon in the indoor air of houses in which it can accumulate, depending on their location, design and ventilation. Radon has been classified by the International Agency for Research on Cancer as "certain pulmonary carcinogen" since 1987; it is the second leading cause of lung cancer, after tobacco.

An IRSN mapping of geogenic radon potential has been carried out at the scale 1:1 000 000. This map is only based on geological data and reflects the capacity of geological units to produce radon and to facilitate its transport to the atmosphere. From this map, French counties are classified into 3 categories.

The study proposed here carries out a geostatistical study on the scale of the French territory from more than 30,000 measured values. An ordinary kriging type analysis was performed and a first mapping was achieved.

With a high spatial variability, by calculating the excess percentage of reference values, it is already possible to discriminate areas for which the radon concentrations in the houses appear higher.

Several geostatistical approaches are in prospect: cokriging, conditional simulations.

WIND ENERGY POTENTIAL ESTIMATION USING MACHINE LEARNING: FEATURE ENGINEERING AND SELECTION

Mikhail Kanevski (1)* - Fabian Guignard (1) - Federico Amato (1)

University of Lausanne, Idyst (1)
** Corresponding author: Mikhail.Kanevski@unil.ch*

Abstract

Nowadays, analysis and assessment of renewable energy potential is of great importance. Wind fields are nonlinear and highly variable phenomena at different spatial and temporal scales. If we consider a wind energy potential modelling in a complex mountainous region like Switzerland, the problem becomes very challenging. Therefore, assessment of wind energy potential is often considered in high dimensional input feature spaces (IFS) using different machine learning (ML) algorithms. Usually, IFS are constructed by means of expert knowledge and feature engineering. In the present research, a spatial distribution of wind speed is estimated within the framework of a generic methodology of environmental data driven modelling, which covers a wide range of tasks – from data collection via models calibration and testing to the communication and interpretation of the results. The current paper concentrates on two major topics: 1) feature engineering - construction of IFS; 2) feature selection - a selection of the relevant input features/variables. Feature engineering was performed applying high resolution digital elevation model and GIS tools, simulating new redundant and irrelevant features, and by transforming and shuffling of the raw features. A variety of unsupervised and supervised ML algorithms and tools were adapted and applied to study the problem of feature selection, aiming to improve the modelling results, for example, to reduce a testing error. Filters (methods independent on modelling tool) were used at the pre-processing step, while wrappers and embedding methods were applied at the modelling process. The performance and efficiency of a variety of the algorithms - k-nearest neighbors, Multilayer Perceptrons, General Regression Neural Networks, Random Forest, and Gaussian Processes, were compared. It is shown, that feature selection and variables importance are critical techniques improving data modelling and interpretability of the results. The real data case study consists of the original measurements carried out by MeteoSuisse network, composed of more than one hundred stations distributed over the Switzerland.

NON-STATIONARY MULTIVARIATE CONDITIONAL SIMULATION OF OLYMPIC DAM DEPOSIT

Minniakhmetov Ilnur (1)*

Bhp, Technical Centre of Excellence and Legacy Assets (1)
** Corresponding author: ilnur.minniakhmetov@bhp.com*

Abstract

Conditional simulation models are critical for uncertainty quantification of resources for major capital investment decisions. Olympic Dam deposit is the fourth largest copper deposit and the largest known single deposit of uranium in the world. The main challenges for simulation process are: hundred millions of grid blocks, diffusive nature of grades, non-stationary and spatially non-linear grade distribution, critical Cu:S relationship for smelter performance, non-linear correlation between economical variables. The combination of different techniques have been chosen to address those challenges. First, mineralization zones have been modelled using implicit boundary simulation method. Next, the Projection Pursuit Multivariate Transform has been applied to decorrelate attributes of interests: Au, Ba adjusted S, Cu, Sg, U3O8. Finally, the spectral simulation of decorrelated attributes has been implemented with account of local varying mean, local varying variance, and local varying anisotropy. The models have been validated using histograms, grade-tonnage curves, cross-plots, variogram reproduction. Resulting conditional simulations have been utilized in several decision making frameworks: optimal drill hole spacing analysis, impact of cut-offs and stope footprint sizes to metal revenue and mined tons, and capital investment decision.

THE AGROMET PROJECT: A VIRTUAL WEATHER STATION NETWORK FOR AGRICULTURAL DECISION SUPPORT SYSTEMS

Damien Rosillon (1)* - Jean Pierre Huart (1) - Michel Journée (2) - Viviane Planchon (1)

Cra-w, Productions in Agriculture Department (1) - Royal Meteorological Institute of Belgium (2)
** Corresponding author: d.rosillon@cra.wallonie.be*

Abstract

Objective

Weather-based forecasting models play a major role in agricultural decision support systems but warnings are usually computed at regional level due to a limited amount of automatic weather stations (AWS). Farmers have to refer to the nearest AWS but recommendations are not always adapted to their situation.

The Agromet project aims to set up an operational web-platform designed for real-time agro-meteorological data dissemination at high spatial (1 km x 1 km grid) and temporal (hourly) resolution in Wallonia, southern part of Belgium.

Usually, meteorological data interpolation is performed on low temporal resolution data (eg monthly or yearly) or on climatic data. Interpolate hourly or daily data is much more uncommon and is a real challenge.

Material and methods

Two datasets of meteorological data are used in this study: a first dataset comes from the Pameseb network from the Walloon Agricultural Research Centre CRA-W (28 selected AWS) and a second dataset comes from the Royal Meteorological Institute network (8 selected AWS).

Five learners (or algorithms) are tested: multilinear regression (MultiReg), inverse distance weighted, one nearest neighbor, ordinary kriging and kriging with external drift.

Data analysis is conducted with R software based on mlr package (Machine Learning in R). This package provides a unified interface to more than 160 basic learners. It provides all required interpolation algorithms except kriging. For the purpose of our study, we integrated gstat functions to mlr.

A huge amount of possibilities can be tested in machine learning based on a combination of a learner, one or several explanatory variables, a defined dataset, … To give a structure to our analysis, we define several "explorative constructions" (EC). One EC is a unique combination of a learner, hyper-parameters (if required for the learner e.g. semi-variogram parameters for kriging), one or several explanatory variables (if relevant for the learner) and a dataset.

Each EC is tested by conducting a benchmark. Models are trained on a 2 years of hourly and daily measurement dataset. Training period is from 01/01/2016 00h UTC+2 to 31/12/2017 23h UTC+2. Quality of the prediction models is assessed by a leave-one-out cross validation. Two quality indicators are computed: Root mean square error (RSME) and predicted residuals.

First conclusions

The poster will present the first conclusions of our ongoing project. So far, only the air temperature at hourly and daily step was interpolated and only five learners were explored. However, we can see that even at high temporal resolution of one hour, interpolate data with geostatistical analysis increase the quality of field level air temperature prediction and is better than taking the nearest automatic weather station. Multilinear

regression is the best method for both hourly and daily air. Increasing the dataset from 28 to 36 observations points slightly increases the quality of interpolation.

In the next steps, we plan to go further in results analysis and to focus on extreme deviations. We also plan to interpolate relative humidity and leaf wetness at hourly and daily steps.

GEOSTATISTICAL ANALYSIS IN MYCOTOXIN STUDIES

Ruth Kerry (1)* - Ben Ingram (2) - Esther Garcia-Cela (3) - Brenda Ortiz (4) - Naresh Magan (3)

Brigham Young University, Geography (1) - Talca University (2) - Cranfield University (3) - Auburn University (4)
** Corresponding author: ruth_kerry@byu.edu*

Abstract

Mycotoxins are produced by fungi that can contaminate staple crops. Legislative limits exist for the levels allowed in grain for human/animal consumption because they can cause serious health problems. Mycotoxins are measured post-harvest in stored grain. Crops are accepted or rejected based on average concentrations in grain with no consideration of spatial variability. Factors influencing concentrations have been well-investigated, but there are few studies of spatial variation. Case studies are used to illustrate the need for spatial analysis of mycotoxins at different scales such as 2D and 3D variogram cross-variogram analysis, kriging and Local Moran's I analysis. Insights from spatial analysis will be discussed.

Some mycotoxins develop in field whereas others develop in storage. The collocation of clusters of both types of toxin in stored grain and the smaller size of clusters for those developing in storage suggest their development from foci of toxins that develop in the field. 3D analysis of stored grain shows greater contamination towards the base and outer-surface of the grain pile in moister more aerophillic locations. Aflatoxin variation within fields showed that risk at this scale is associated with soil type and topography. Different risk zones can be managed, harvested and stored separately to reduce wasted grain. Aflatoxin contamination risk of different counties investigated with profile regression is associated with maximum temperatures above, and precipitation levels below, 30-year normals. Future climate change scenarios suggest increased risk of aflatoxin contamination and the need for more irrigation, planting of resistant varieties, shifts in zones where corn is grown, or shifts in growing season scheduling.

THE USE OF CITIZEN OBSERVATIONS FOR BETTER PRECIPITATION ESTIMATION AND INTERPOLATION

Abbas El Hachem (1)* - András Bárdossy (1) - Jochen Seidel (1)

University of Stuttgart, Department of Hydrology and Geohydrology (1)
** Corresponding author: abbas.el-hachem@iws.uni-stuttgart.de*

Abstract

The number of private meteorological stations with data available online through the internet is increasing gradually in many parts of the world. The purpose of this study is to investigate the applicability of these data for the spatial interpolation of precipitation for high intensity events of different durations. Due to unknown biases of the observations, rainfall amounts of the secondary network are not considered directly. Instead, only their temporal order is assumed to be correct. The crucial step is to find the stations with informative measurements. This is done in two steps, first by selecting the locations using time series of indicators of high precipitation amounts. The remaining stations are checked whether they fit into the spatial pattern of the other stations. Thus, it is assumed that the percentiles at the secondary network accurate. These percentiles are then translated to precipitation amounts using the distribution functions which were interpolated using the weather service data only. The suggested procedure was tested for the State of Baden-Württemberg in Germany. A detailed cross validation of the interpolation was carried out for aggregated precipitation amounts of 1, 3, 6, 12 and 24 hours. For each aggregations nearly 200 intense events were evaluated. The results show that filtering the secondary observations is necessary, the interpolation error after filtering and data transformation decreases significantly. The biggest improvement is achieved for the shortest time aggregations.

CONDITIONAL SIMULATION OF CHANNELIZED MEANDERING RESERVOIRS USING PARTICLE FILTERING

Alan Troncoso (1)* - Xavier Freulon (1) - Christian Lantuéjoul (1) - Fabien Ors (1) - Jacques Rivoirard (1)

PSL University - MINES ParisTech, Centre of Geosciences (1)
** Corresponding author: alan.troncoso@mines-paristech.fr*

Abstract

This presentation deals with the characterization of geological reservoirs. Besides its specific interest in oil and gas industries, it can help to address issues and challenges encountered in the environmental sciences, namely in geothermal energy, water management and CO_2 storage. The reservoirs considered here are formed by channelized meandering systems. They are stochastically modeled using a process-based approach that mimics three interacting sedimentary processes (migration, aggradation and avulsion) in order to reproduce the evolution of the reservoir along geological time. In practice, it is important that the simulated reservoirs respect the available field information, such as well facies data or seismic data that provide sand proportions. To achieve this goal, a conditional simulation technique based on particle filtering has been developed. Particle filtering is a statistical technique based on the generation and the selection of so-called particles (in the present context, each particle is a reservoir simulation). The generation of the particles is made stepwise. At each step, the construction of each particle is resumed by stacking a layer of constant thickness made of sediments simulated unconditionally. Then, only the particles that respect at best the conditioning data are selected and replicated, so that the total number of particles remains the same. The simulation process terminates as soon as the top of the reservoir has been reached. One particle among all those produced is then randomly selected to serve as a conditional simulation. This novel approach has been implemented in FLUMY, a software developed by MINES-ParisTech to perform reservoir simulations. It will be presented in detail and illustrated using a case study.

Keywords: Process-based models, Sequential Simulation, FLUMY

1. Introduction

Modeling heterogeneous reservoir is a topic of interest for the geosciences. Many approaches have been developed for this task. Geostatistical techniques utilize variograms as the main tool to estimate or simulate variables either categorical or continuous (Beucher and Renard, 2016). While their capacity to integrate field information is strong, the results in terms of heterogeneity are often viewed as unrealistic. Object-based approach integrates objects in the fields that represent the geological features according to some rules (Deutsch and Wang, 1996). Depending on the complexity of these rules, the capacity to honor the field data and the heterogeneity is accomplished. Finally, a process-based approach that utilizes laws of physics to build reservoir models, honors by construction their features (Cojan et al., 2005), but the conditional step has long been a challenge.

In this work, we considered FLUMY, a stochastic and process-based reservoir model developed by MINES-ParisTech for channelized meandering systems in fluvial and turbidite environments. Here, only the fluvial environment has been considered. This model reproduces the sedimentological evolution of the meandering system and records the associated deposits at the scale of the reservoir. The conditional step is currently

performed dynamically by constantly adapting the simulated processes in order to match the field data (Bubnova, 2018). Here, we propose a different sequential approach using a particle filtering. The adaptation of this technique, widely used in signal and image processing, is applied here to obtain reservoir models that honor fairly well the observations and, at the same time, respect the sedimentological processes, hence the heterogeneities and sand body arrangements.

The paper is organized as follows. Section 2 introduces the FLUMY model. Section 3 introduces the particle filter used for a sequential simulation of the model. Section 4 presents the results. Finally, conclusions are drawn in Section 5.

2. FLUMY: A Process-based Model

2.1. Description of the model

FLUMY reproduces the reservoir sedimentation in time through three main processes: migration of a channel, aggradation, and avulsion (Figure 1). When migrating, the channel develops its meanders, eroding its outer banks, and depositing sandy point-bars within the loops. Meanders are occasionally shortened by cutoffs. From time to time, overbank floods occur, which are responsible for the construction of silty levees and for the deposition of coarse-grain sand at the bottom of the channel and fine-grain alluvium on the floodplain. This corresponds to the aggradation process, which is the cause of the vertical sedimentation through time. Finally, a levee breach can result in a new path for the channel, which is the avulsion process. These three processes interact in competition while building the reservoir. The higher the migration rate, the greater is the reworking of the previous deposits (which increases the resulting net to gross and the lateral sand connectivity). This is the opposite for high aggradation rates which tend to reduce the net to gross and improve the vertical sand connectivity. Whatever the avulsion frequency, this does not modify the resulting net to gross. On the other hand, a high avulsion frequency favors disconnected sand bodies.

The output is a three-dimensional numerical model, with information about lithofacies, grain size and age for each deposition unit. For this work, the 13 different facies have been grouped in two facies, namely sand and shale.

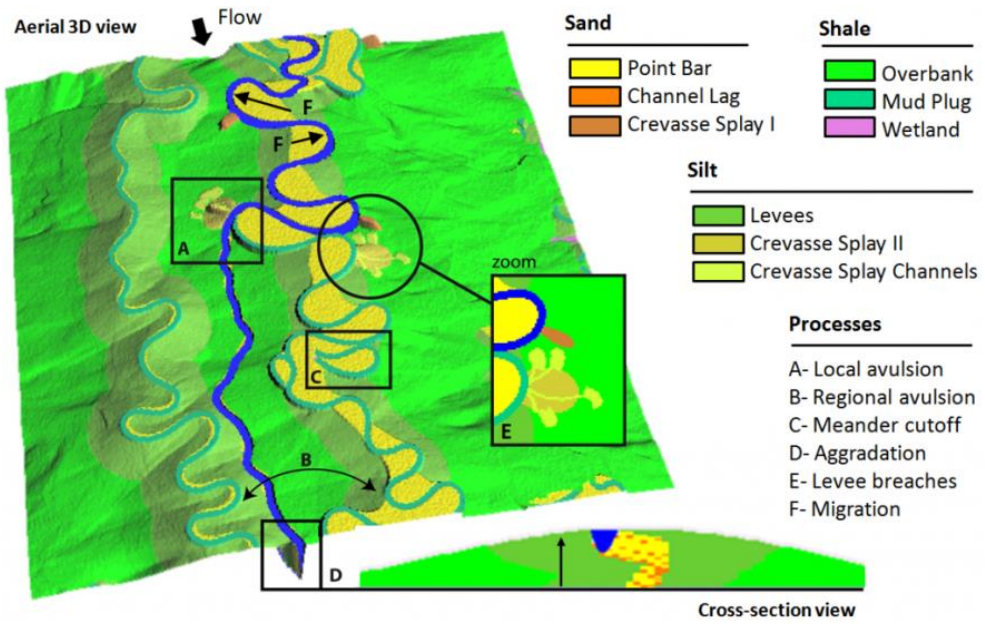


Figure 1 - FLUMY Schematic Image with all its Facies and Processes.

The simulation depends on a few parameters, that rule the extension of the output domain, the processes, and so the type of sedimentation (e.g. the degree of meandering, the net to gross ratio). In the current version of FLUMY software, a "dynamic" conditioning is used to honor well data. It consists in modifying the main processes while running the simulation to attract the channel close to neighboring sand data, or to keep the channel distant from shale data (Bubnova, 2018). However, this requires a detailed knowledge of the processes and may result in distortions affecting the deposits. By contrast, the sequential conditional algorithm presented in this paper avoids modifying the sedimentary processes but in return requires the execution of a large number of simulations.

2.1 Use of the model

The goal of the present paper is to study how sequential simulation technique can be used to condition a process-based model on well data. Here four conditioning vertical wells will be used, along which facies data are known (as either sand or shale).

Such wells and their data are extracted from a previous non-conditional simulation. Both conditional and non-conditional simulations use the same parameters. In particular this guarantees that well data are consistent with the parameters used for the conditional simulation. The simulation domain is 2510m x 2510m x 10m, and the channel depth H = 3m. To ensure that the simulation up to 10m is representative of the model and cannot be modified by further avulsions and migration, processes are run until the simulated topography exceeds 13m everywhere. We will not detail the signification and value of the other parameters, which correspond to the default median scenario proposed by FLUMY.

The non-conditional simulation used here is presented in Figure 2. The location of the 4 wells used for conditioning is shown on the same image. The sand and shale facies are respectively depicted in yellow and green. The proportion of sand in the simulation is about 60%.

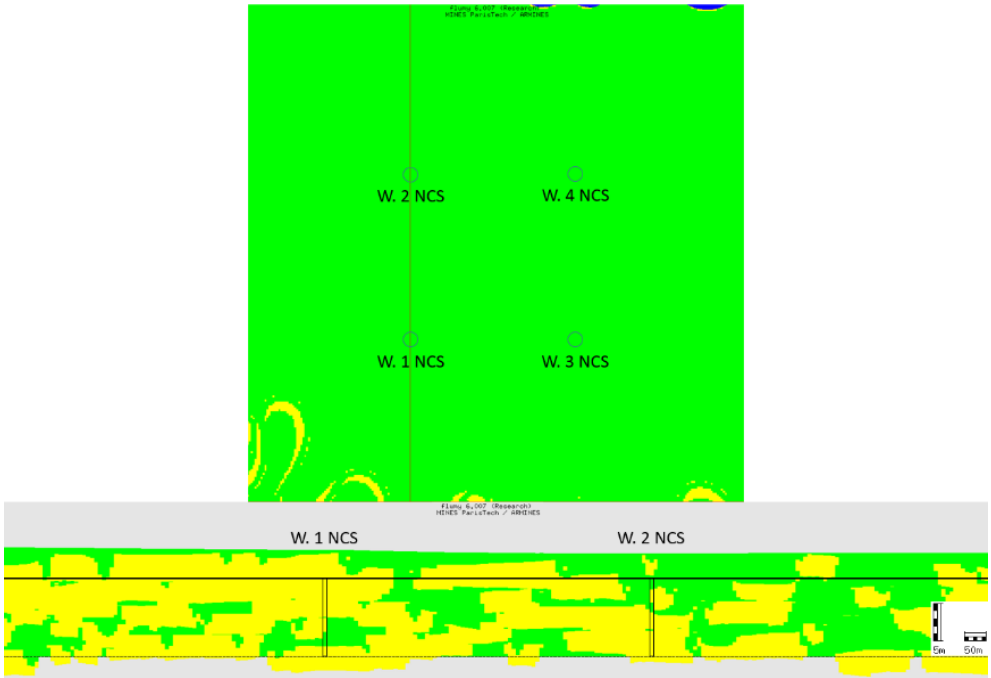


Figure 2 - (top) Aerial view of the non-conditional simulation with the location of the 4 wells and of the cross-section (channel flowing from East to West), (bottom) the North-South vertical cross-section going through two wells. In yellow sand and in green shale.

3. Sequential Monte-Carlo simulation of a process-based model

3.1. SMC in a nutshell

The Sequential Monte-Carlo (SMC) method, also named particle filtering, is a popular method to assimilate sequentially observations of a dynamical system. It has been developed in order to overcome limitations of linear methods such as the Kalman filter, when the system is modeled by non-linear or non-Gaussian processes (Doucet et al., 2001). It may be viewed as the combination of two main elements: i) a hidden Markov chain used to model the dynamic of the system, and ii) a set of weighted models, or particles, defining a discrete approximation of the conditional model. Within this framework, integrating sequentially a new piece of information is performed into three steps. Firstly, particles evolves from the previous state to the time of the new observation according to the *a priori* dynamic, they constitutes the proposals; then, the weight of each proposal is updated according to its compatibility with the new observation; finally, the particles are re-sampled according to the updated weights. These new particles are exchangeable, hence uniformly weighted. As an ensemble technique, SMC describes the evolution of the conditional distribution rather than the evolution of a specific model, and estimates are derived as statistics on the particles. In addition, each particle is a realization of the conditional distribution (aka a conditional simulation).

Regarding the resampling step, Douc et al. (2005) review different strategy, from a plain multinomial resampling scheme to more intricate algorithms such as stratified or residual approaches. Next section details the adaptation of this methodology to the particular case of the process-based model such as FLUMY, first with the split of the continuous depositional process into a sequence of layers, and then the definition of the weights for each layer.

3.2. Layers and non consolidated zones

As presented in Section 2, FLUMY fills the reservoir model with lithofacies, mimicking the main depositional processes in a fluvial context. From a geometrical point of view, this continuous process from the base to the top of the reservoir can be split into N layers with a fixed thickness (the ith layer is defined by elevations between z_i, and z_{i+1}). For a fixed iteration, the upper part of the model, just below the current topography can be later reworked due to the avulsions and the lateral migration of the channel (see Figure 3). Inversely, the content of a layer cannot be further modified on as soon as the bottom of the active channel is definitely above the upper limit of the layer. Hence the complete model is split into a series of depositional sequences, each one achieving the filling of a given layer: for the i^{th} sequence the model is restarted from the end of previous sequence (the initial sequence starts from the bottom of the model) and run until the difference between the topography and the upper limit of the layer exceeds everywhere the channel depth. At this time, the content of the i^{th} layer is definitely defined, and it is called the *consolidated layer* i; the content of the zone between the upper limit z_{i+1} and the current topography can be modified later on and it is called the *unconsolidated zone i*. Then FLUMY can be restarted from the unconsolidated zone only and the position of the channel when it was previously stopped.

These sequences define a Markov process: the consolidated layer belongs to the past and can be saved for the reconstruction of the final model; the current state is the unconsolidated zone and the channel position from which the depositional process can proceed.

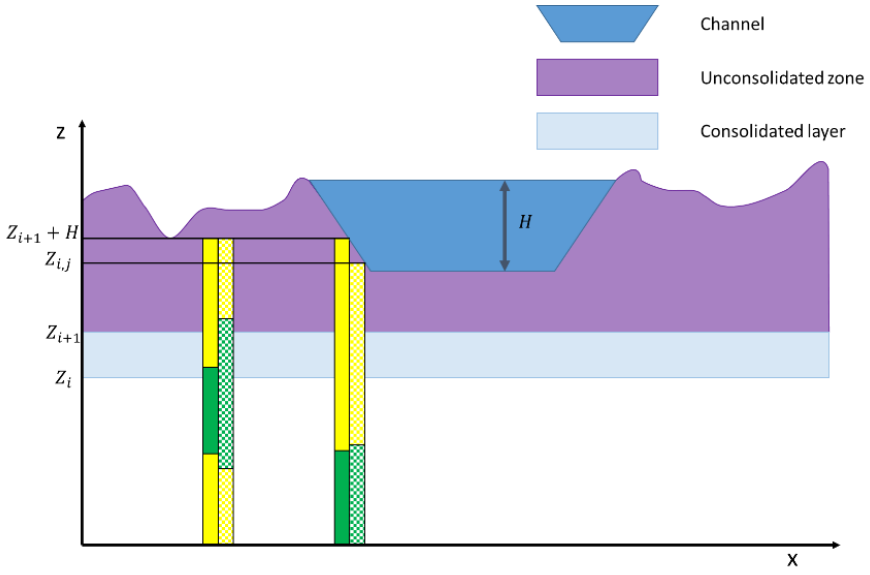


Figure 3 - Definition of the consolidated layer and the unconsolidated zone. Two pairs of wells are represented: solid colors represent the field data and light colors represent a particle.

3.3. Resampling weights for the particles

To implement the particle filter using the Markov process defined above, weights should be defined for the resampling stage. These weights reflect the compatibility between the observations and the simulated values. Observations are facies along the wells: let $F_j(z)$ be the facies of well j at elevation z, either sand or shale. They are compared with the simulated facies of the K particles, $F_j^{(k)}(z)$ for $k \in \{1, \dots, K\}$, both in the consolidated layer and in the unconsolidated zone, defining two types of indicators.

For each stage i, a first indicator is defined as the proportion of match in the consolidated layer along the well j and for each particle k:

$$PC_{i,j}^{(k)} = \frac{\int_{z_i}^{z_{i+1}} 1_{F_j(z)=F_j^{(k)}(z)} dz}{\int_{z_i}^{z_{i+1}} dz}$$

A second indicator is defined as the weighted proportion of match in the unconsolidated zone, taken up to z_{i+1} + H, along the well j and for each particle k:

$$PNC_{i,j}^{(k)} = \frac{\int_{z_{i+1}}^{z_{i,j}} e^{-\mu * R * (z - z_{i+1})} 1_{F_j(z)=F_j^{(k)}(z)} dz}{\int_{z_{i+1}}^{z_{i+1}+H} e^{-\mu * R * (z - z_{i+1})} dz},$$

where

- $z_{i,j}$ is the minimum between $z_{i+1} + H$, and the topography at the end of sequence i along well j;
- R is the ratio between the forecasts for migration and aggradation. This value depends on the simulation parameters and is dimensionless;
- μ is a multiplicative coefficient.

The exponential shape of the weight, $e^{-\mu * R * (z - z_{i+1})}$, has been chosen so as to assign to the upper part of the unconsolidated zone up to $z_{i+1} + H$, a lower weight, as it has a larger chance to be eroded in the following sequences. The values μ and R control the rate for this function to decrease.

Since a single weight is needed for each particle, these indicators of match are linearly combined according to the following formula

$$\omega_i^{(k)} \propto (1-u) * \sum_j^4 PC_{i,j}^{(k)} + u * \sum_j^4 PNC_{i,j}^{(k)}$$

Given these weights, the resampling for stage i is performed using the residual scheme as it ensures the reproduction of the particles with the highest weights (Douc, Cappé, & Moulines, 2005).

4. Results

Sequential conditional simulations have been performed to condition FLUMY by the 4 wells extracted from the non-conditional simulation of Section 2. The thickness of the layers used for conditioning have been taken as $0.3m$, i.e., $1/10$ of the channel depth. It goes without saying that the smaller the layers, the better the quality of results since the comparison with field data is done using a smaller scale. However, the computation time of the full process is slower, due to the higher quantity of resampling stage for the particle filtering. It is known that the result quality of the particle filtering depends on the quantity of the particles K. However, the effect of the unconsolidated zone for the conditioning simulation, measured with the variables u and μ, is unknown. Hence, a sensibility study is performed varying $K = \{10,50,100,250\}$, $u = \{0.00, 0.20, 0.33, 0.43, 0.50, 0.67, 0.71, 0.77, 0.83, 0.91, 1.00\}$ and $\mu = \{1,50,100\}$. 25 independent realizations are computed to control the statistic fluctuations. For each realization, the mean match is calculated for the 4 wells for each particle and, the one with the highest value (best particle) and one randomly (random particle) are taken. Finally, the mean is calculated from the 25 realizations to assign a value for each variable. The figure 4 on the left and in the middle summarises the results. From this sensibility analysis, it is determined that:

1) The higher the number of particles, the better the match.
2) Taking into account the unconsolidated zone has a minor effect on the results.

Analyzing the values from the best particles (Figure 4 in the left), the best result comes from the "optimal parameters" $K = 250$, $u = 0.67$ and $\mu = 50$ with a mean match of 82.3%. A plot showing the influence of the number of particles while fixing $u = 0.67$ and $\mu = 50$ is shown in the figure 4 in the right. The particle filtering taking a random particle in each realization (Figure 4 in the middle) shows a mean match around 70%, even for 10 particles. The effect of the particle filtering conditioning can be seen by comparing this result with the match that would be expected using a non-conditional simulation (which is $p^2 + (1-p)^2 = 52\%$, for a proportion of sand $p = 60\%$).

A single particle with the optimal parameters is taken to show the conditional simulation in the figure 5. A comparison with the field data is shown with the match of each well. While wells are not reproduced exactly, the effect of the conditioning is very clear.

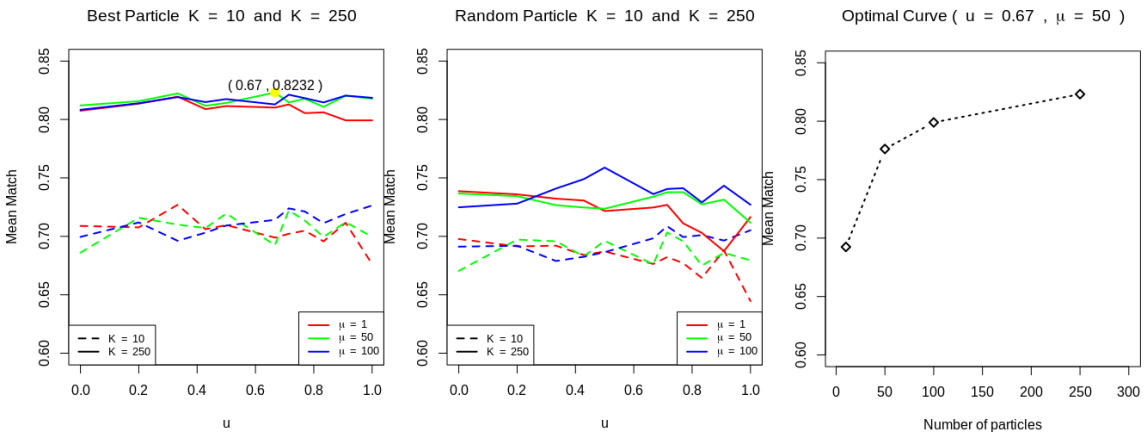


Figure 4 - (left) Mean Match Value over the best particle for each of 25 independent copies of the Particle Filter with K={10,250} particles, (middle) Mean Match Value over a random particle for each of 25 independent copies of the Particle Filter with K={10,250} particles; (right) Mean Match Value for the best particle of each realization increasing for K={10,50,100,250} using u=0.67 and μ=50.

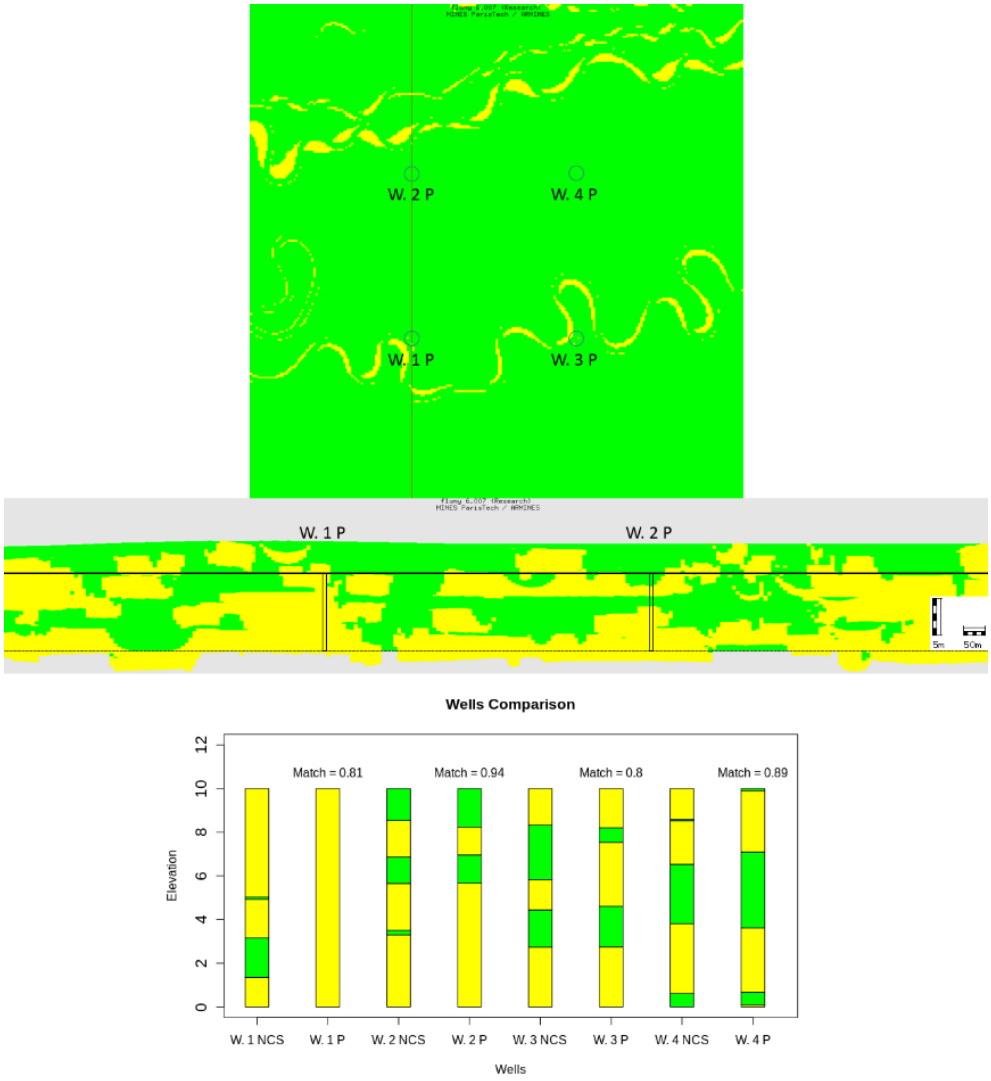


Figure 5 - (top) Aerial view of a conditional simulation from a particle, with the location of the 4 wells and of the cross-section (channel flowing from East to West), (middle) the North-South vertical cross-section going through two wells. (bottom) the wells from the NCS and the particle. Over the particle wells, the facies match. In yellow sand and in green shale.

5. Conclusions

Sequential conditional simulation has been adapted to condition the sedimentary model FLUMY on well data, by proceeding layer by layer. At each stage of the sedimentation, the simulation can be conditioned both from the consolidated layer, which does not change in the sequential evolution, and from the unconsolidated zone, which rests above the previous one and may be subject to further changes. In fact, taking into account the unconsolidated zone does not make a significant difference. The key factor is the number of "particles"(simulations), that must be high enough to ensure a good matching. With 250 particles, it is possible to get a mean match close to 82% by selecting the best particle. Although it is not a 100% match, facies at wells are pretty well reproduced, and the geological evolution is honored. In consequence, no distortions or artifacts are formed. It would be interesting to make an extension of the method to more than two facies or to a conditioning by facies proportions besides a perfect facies match.

Acknowledgements

The first author acknowledges the financial support of Chile National Agency for Research and Development (ANID)/Scholarship Program/DOCTORADO BECAS CHILE/2018 - 72190309.

References

Beucher, H., Renard, D. (2016). Truncated Gaussian and derived methods. Mathematical Geology, 510-519.

Bubnova, A. (2018). Sur le conditionnement des modèles génétiques de réservoirs chenalisés méandriformes à des données de puits. Thèse de doctorat en Géosciences et géoingénierie.

Cojan, I., Fouché, O., Lopez, S., Rivoirard, J. (2005). Process-based Reservoir Modelling in the Example of Meandering Channel. In Geostatistics Banff (pp. 611-619).

Deutsch, C., Wang, L. (1996). Hierarchical object-based stochastic modeling of fluvial reservoirs. Mathematical Geology, 857-880.

Douc, R., Cappé, O., Moulines, E. (2005). Comparison of resampling schemes for particle filtering. 4th International Symposium on Image and Signal Processing and Analysis. Zagreb, Croatia.

Doucet, A., De Freitas, N., Gordon, N. (2001). Sequential Monte Carlo Methods in Practice. New York: Springer.

Liu, J., Chen, R. (1998). Sequential Monte Carlo Methods for Dynamic Systems. Journal of the American Statistical Association, 93.

Makela, A., Landsberg, J., Ek, A. R., Burk, T. E., Ter-Mikaelian, M., Agren, G. I., . . . Puttonen, P. (2000). Process-based models for forest ecosystem management: current state of the art and challenges for practical implementation. Tree Physiology, 289-298.

LAND SURFACE MODELING BY SIMPLE KRIGING WITH LOCALLY VARYING MEAN (SKLVM) WITH VEGETATION ELEVATION AS SECONDARY VARIABLE

Alini Vieira Mancio dos Santos (1)* - João Felipe Coimbra Leite Costa (1) - Camilla Zacche (1,2)

Universidade Federal do Rio Grande do Sul - UFRGS, Departamento de Engenharia de Minas - DEMIN (1) CCGU Alberta (2)

** Corresponding author: alini.mancio@zenithbrasil.com*

Abstract

Areas densely covered by vegetation present significant challenges when reconstituting terrain surface morphology based on data acquired by aerophotogrammetry only. Sparsely distributed terrain sampling points in areas with such characteristics make interpolation by conventional methods difficult, leading to distortions and over-smoothing, which fail to represent the true geomorphology.

The present study proposes the use of elevation data from the tops of trees as a secondary variable to estimate surface elevations based on a multivariate geostatistical framework, including SKVLM and Cokriging.

The results proved that this working approach produced models which correctly represented the terrain surface in areas where surface sampled points are scarce, reducing estimation error. The analysis of the Mean Squared Error of the estimated surfaces in relation to the data from the control points revealed that the incorporation of the dimensions of the top (canopy) of the vegetation in the estimates through SKVLM significantly reduces such error in comparison with conventional methods, where the MSE obtained by the proposed technique (SKVLM) was above 2.75 meters, and the MSE of the most commonly used method for this purpose (PhotoScan algorithm) was 5.90 meters.

Keywords: Simple Kriging local varying mean, cokriging, digital elevation models

1. Introduction

Digital elevation prototypes are a consequence of the interpolation of existing information, that is, the quality of a model is not only related to the quantity and quality of the data used. The model generated is also strongly influenced by the interpolator methods used. Li and Heap (2008) describe 42 interpolation methods commonly used in several areas in their book "Review of Spatial Interpolation Methods for Environmental Scientists": 12 non-geostatistical methods, 22 geostatistical methods, eight combined methods and 14 sub-methods. Of these 42 methods described by Li et. al (2008), only six non-geostatistical interpolators are widely used in the generation of DEM (DEM): Spline, Topo-Grid (topo-raster), Triangulation, Inverse Square of Distance (IQD), Nearest Neighbor and Natural Neighbor. It is worth noting that conventional methods use only data from the field as a database.

Thick vegetation covering imposes difficulties to survey the topography; the average height of such covering can be calculated in order to infer the natural features of the terrain, such as valleys and ridges, in places where the surface is not exposed (Figure 1).

Figure 1 - Landscape photography of dense forest.

A related situation occurs when the data source is the aerophotogrammetric reconstitution of areas of dense forest, in which the terrain surface points applied are scarce or non-existent. The inference of surfaces by interpolating the information of interest only (the elevation of the surface of the terrain) causes distortions and/or smoothings that do not represent the physical reality. Figure 2 shows a schematic profile of points surveyed by aerophotogrammetry (terrain points in brown and vegetation points in green) and the estimated surface projection line (blue line) interpolating only the elevation points of the terrain. In this profile, note in some cases, the surface of the estimated model has a higher elevation than the top level of the vegetation, representing a distortion of the physical reality.

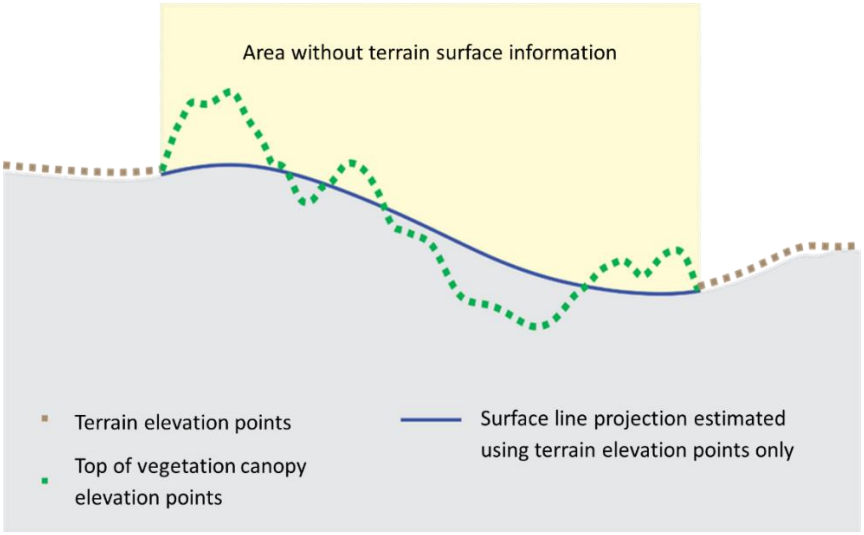


Figure 2 - Schematic profile containing the elevation points of the terrain and vegetation canopy and the surface projection line of a model estimated using terrain elevation data only.

2. Materials and Methods

This study presents, as an alternative, the incorporation of vegetation elevation data as a secondary variable for the estimation of the terrain surface elevation using a multivariate geostatistical method, Simple Kriging with Varying Local Means (SKLVM), as described by Goovaerts (1997). According to the aforementioned author: "Generally, estimates are closer to reality when we add more correlated information, especially if there is little information about the variable of interest." Figure 3 shows an example of the application of this

technique, where in areas where the terrain elevation data are scarce or absent due to the dense vegetation cover, information on the vegetation present was used as a marker to estimate the shapes of the terrain.

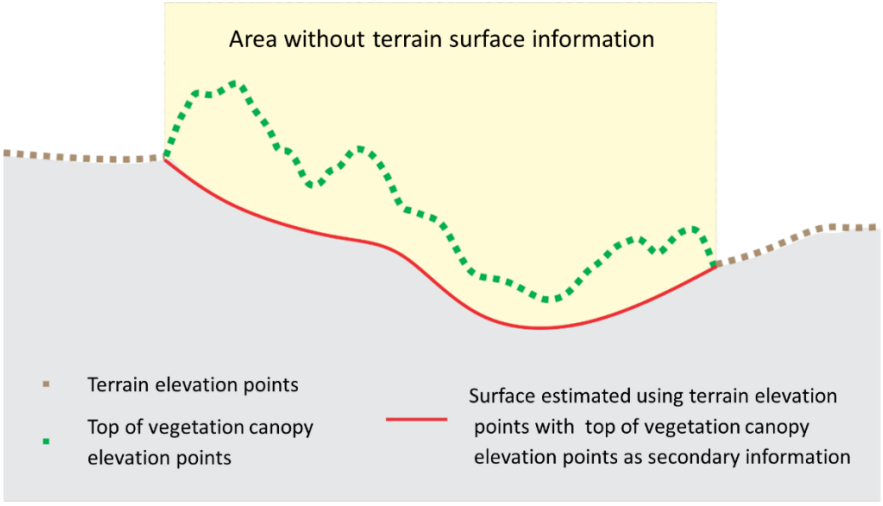


Figure 3 - Schematic profile containing the elevation points of the terrain, the vegetation canopy and the surface projection line of a model estimated using terrain elevation points with top of vegetation canopy elevation points as secondary information.

The methodology applied in this study will be divided into four topics. Firstly, the study area is considered, with a brief summary of the location, the characteristics of the physical environment and the vegetation data. The second subject addressed is the data that served as the basis for this study. There follows the description of the creation of a reference model, generated by triangulation with additional terrain points, six surface models generated by conventional methods, and a model interpolated by simple kriging with varying local means, using the canopy as a secondary variable. Finally, a comparative analysis is made among the surfaces interpolated by the different methods proposed and the control data left unused and used for comparison.

2.1. Study area

The pilot area of this study covers about 20 hectares and is located in the urban area of the municipality of Araxá, in the Alto Parnaíba region, Minas Gerais-Br. Within the limits of the area, the relief comprises part of a hill with two small streams embedded in valleys where the steepness does not exceed a slope of 15% and the elevation differences between the floor and top of the valley are less than 200 meters.

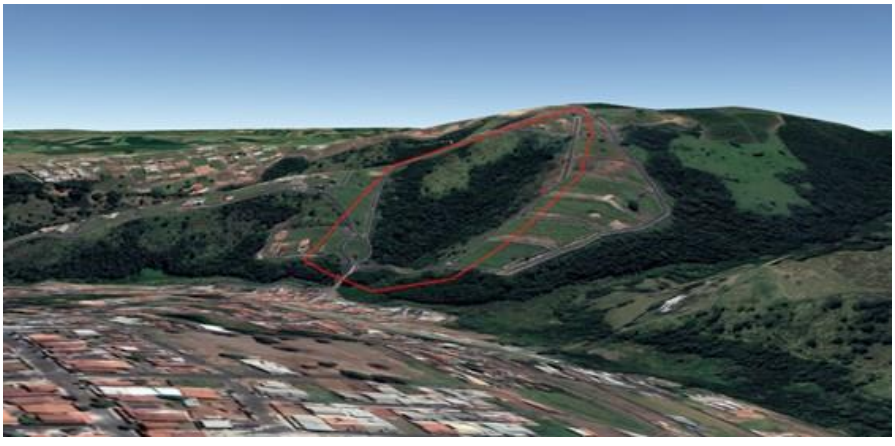


Figure 4 - 3D Google Earth image showing the limits of the test area of the study (red polygon).

With regard to the size and density of vegetation in the study area, uniform behavior and an average individual tree height for this type of vegetation of between 9.2 ± 3.4 meters were observed. The interior of the forest in the test area reveals small differences between individuals, with, in most cases, intertwined canopies, sharing the same space. This last characteristic directly impacts the density of the soil data sampled by aerial drone surveys in areas with dense vegetation.

2.2. Database

The database used is a point cloud obtained by three-dimensional reconstitution of aerial images. The scenes were captured by a DJI Inspire One drone with an embedded RGB Zenmuse X5 30 mm sensor. The photogrammetric reconstitution was performed in the Agisfot MetaShape software package, generating a cloud of 38.3 million points (approximately 170 points/m2), comprising a file containing the information of x, y, z and RGB coordinates. The elevation data are distributed between 975 and 1060 meters with an average of 1027 m and a variance of 386 m. Figure 5 shows the distribution of the data containing terrain elevation data, emphasizing the lack of information along the valleys where the vegetation is dense.

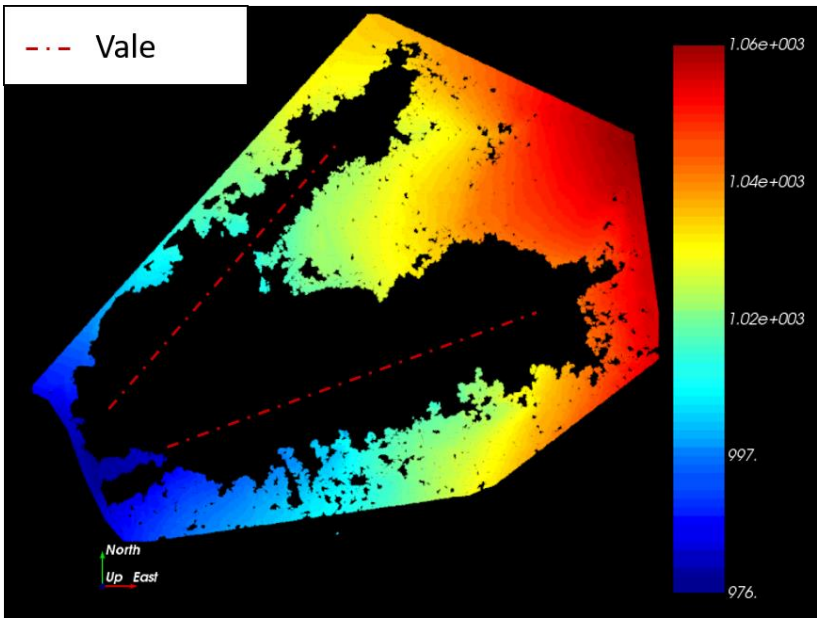


Figure 5 - Spatial distribution of terrain elevation data, with the notable absence of data in the valleys, where the vegetation is dense.

2.3. Reference Model

To measure the adherence of the interpolation methods to the reality in the field, a reference surface interpolated by linear triangulation was modelled. The reference interpolation approach was chosen as it is a commonly used method, and due to the adherence of the modelled surface to the sample data. Three types of data were used, as listed below (Figure 6):

i. 6.9 million points categorized and validated as terrain elevation information (brown points);

ii. 54 points collected in the field with high-precision topographic equipment, described in the data acquisition section in this chapter (blue dots);

iii. 162 points collected from the cloud, carefully selected and checked (orange points).

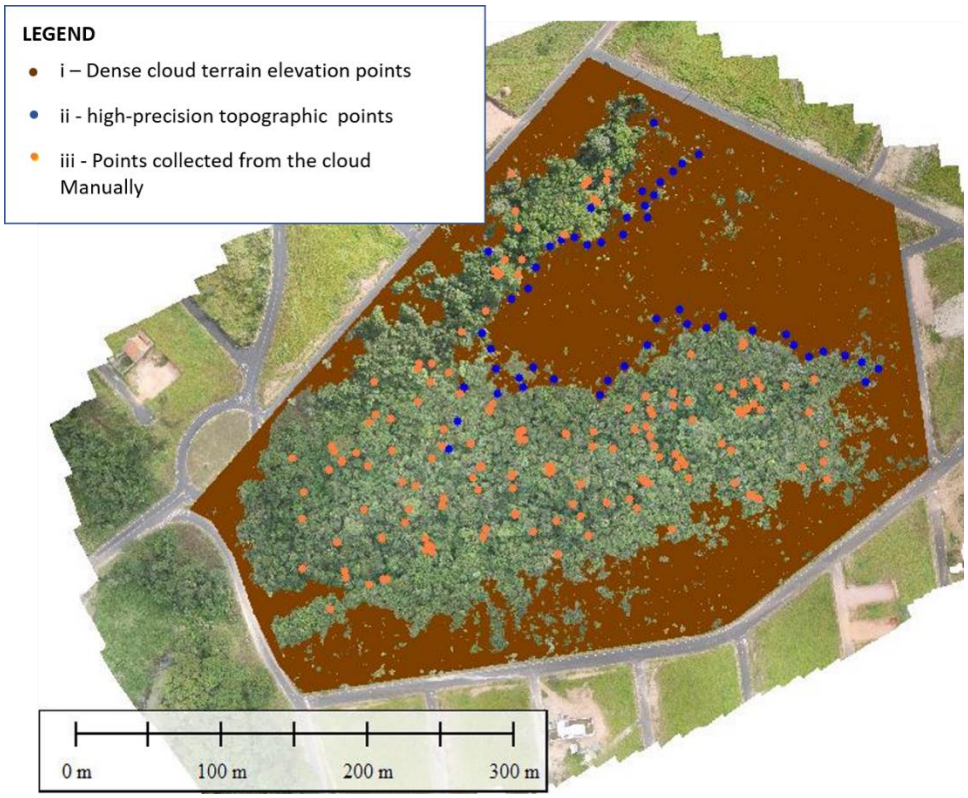


Figure 6 - Types of terrain points according to the acquisition method.

2.4. Terrain Models Generated by Conventional Methods

In order to compare the proposed methods with the conventional methods used, six models were generated, using only the points acquired by aerial survey, and validated with the terrain points.

The six interpolation methods used to generate the digital elevation models of the study area were: Simple Kriging, Delauney Triangulation (Delauney, 1934), Nearest Neighbor, Inverse distance weighting (IDW) (Donald, 1968), Minimum Curvature or Spline and the interpolator used by the PhotoScan software package, which is a modified version of the Delauney Triangulation method. In Figure 7, it can be seen that some surfaces, such as that interpolated by minimum curvature, were modelled with higher elevations than that of the canopy.

2.5. Model interpolated by SKVLM

Simple Kriging with Varying Local Means is a geostatistical interpolating method that incorporates data from a secondary variable in the estimate, in this case information on the canopy elevation. The notable points relating to the application of this technique in the study area are: the primary database, the composition of the secondary variable, the regression equation of the secondary variable in the primary variable and the nuances of the regression residue of this information.

The primary database used is the same as that used for surface modelling by the conventional deterministic interpolators. Simple Kriging with Varying Local Means (Goovaerts, 1997) requires that the information of the secondary variable is available throughout the area where the primary variable will be estimated, that is, the secondary variable should be exhaustive. Therefore, the secondary information used was a 0.5 by 0.5 m grid generated by the interpolation of data points extracted from the original point cloud, obtaining the maximum elevation data within a five by five meter grid. The interpolator used was Delaunay Triangulation.

With the data properly prepared, the surface model generated by SKVLM was prepared in SGEMS (Remy et al., 2009) open and free geostatistical modeling software.

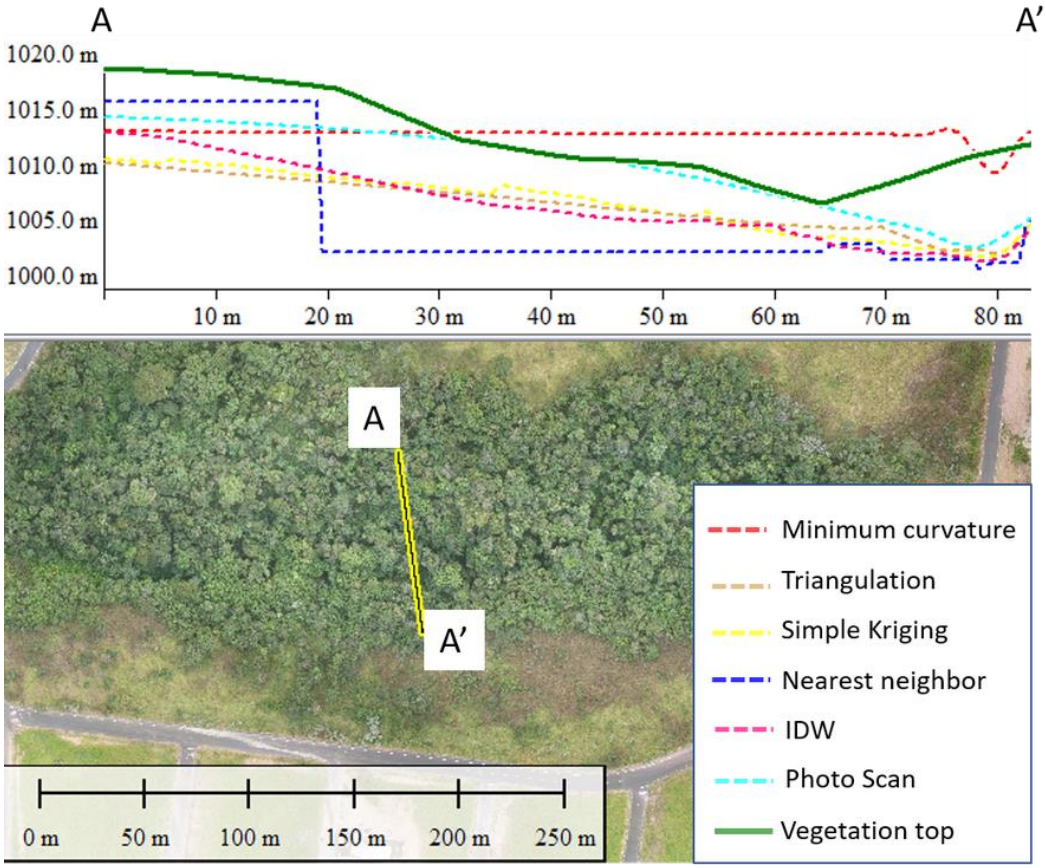


Figure 7 - Cross-sectional profile of surfaces generated by conventional interpolators.

After the database was created, the next step was to determine the correlation between variables by plotting them on a scatter plot. Following the expected trend, the correlation between canopy and terrain elevations is approximately 99%, justifying the use of this method to generate terrain surface models in dense forest areas with sparse primary data. Based on the correlation between the variables, the regression equation is as follows:

$$y = 0.931792\ x + 68.2889$$

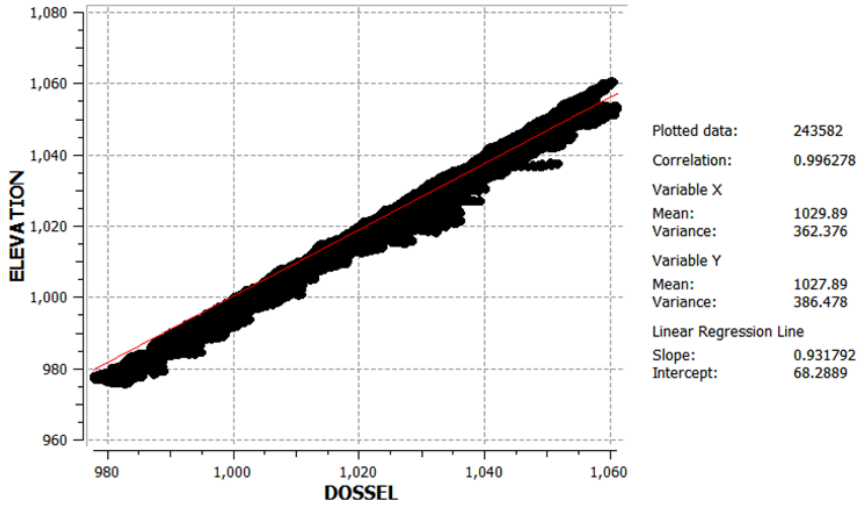


Figure 8 - Scatter plot of top vegetation vs terrain elevation.

The canopy elevation data were regressed to a mean value of the primary variable (terrain elevation) at the grid nodes and at the primary data points. This regression provides the local average elevation of the terrain

which is used to obtain the residues at data locations subtracting the primary data from the local mean. The mean residue value was -0.003 meters with a variance of 0.74 (follow Figure 9).

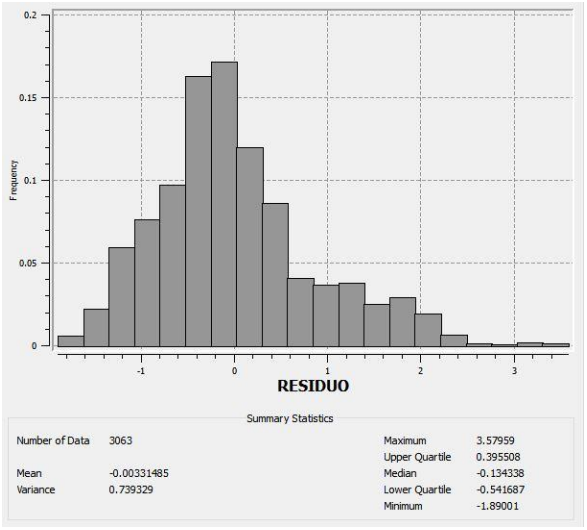


Figure 9 - Histogram of residue.

Once the residue was determined at all sampled points, these data were variographed and a maximum correlation range of approximately 183 meters along 45° direction was observed.

Once the variographic model of the residuals has been defined, the interpolation of the information by SKVLM can be carried out. For this, the standard SGEMS (Remy et al., 2009) algorithm was used, applying the mean local terrain elevation values obtained by regressing the canopy elevation information and the estimated residuals of this regression.

3. Results Obtained

In order to measure the ability of the conventional and proposed interpolator methods (SKVLM) to reproduce the field reality in situations of low sample density, the present study adopted three validation approaches: visual, by indicators and graphical analysis.

3.1. Visual Analysis

The geometry of the estimated surfaces was compared with the geometry of the reference surface (item 2.3 - Reference Model) by vertical sections chosen at random in the study area.

The visual analysis of the estimated surface sections (Figure 10) demonstrates that the surface generated by SKVLM, using the canopy dimensions, most closely adheres to the reference surface. Especially in situations of valleys or ridges covered by dense vegetation. The models interpolated by minimum curvature and nearest neighbor exhibited the least similarity with the benchmark surface.

3.2. Estimate of Error

To compare the differences between the modeled surfaces and the 162 reference points (Figure 6, orange points), two error measurements were used: Root Mean Square Error (RMSE - Root Mean Squared Error) and Pearson's correlation coefficient, according to the methodology proposed in Hallak and Pereira Filho (2011).

The RMSE is defined mathematically by:

$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(Z_s - Z_o)^2\right]^{\frac{1}{2}}$, where n is the number of samples, Z_s is the estimated elevation and Z_o is the sampled elevation.

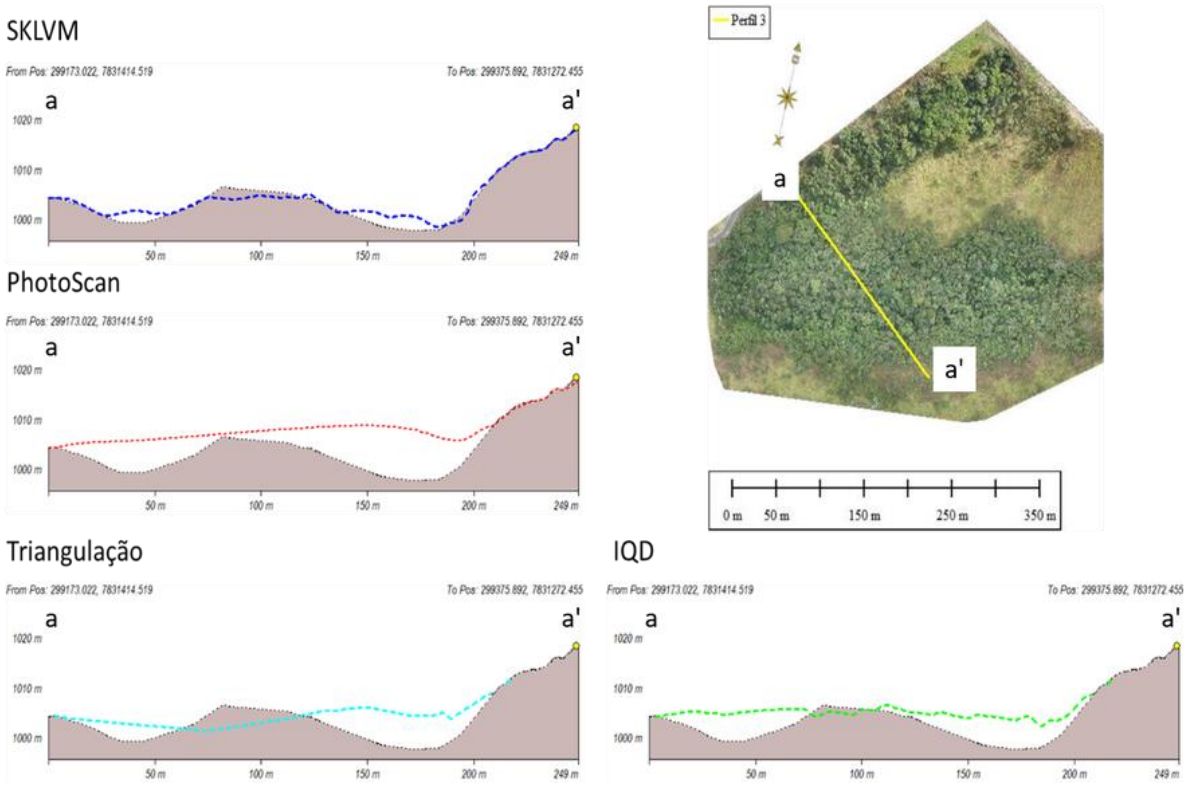


Figure 10 - Topographic sections representing the surfaces with greatest adherence to the reference surface.

Pearson's correlation coefficient is a measure of the existence of a correlation between two variables and its intensity. It is calculated according to the following formula with the definitions presented above:

$$\rho = \frac{\sum_{i=1}^{n}(Z_s - \bar{Z}_s)(Z_o - \bar{Z}_o)}{\sqrt{\sum_{i=1}^{n}(Z_s - \bar{Z}_s)^2} \cdot \sqrt{\sum_{i=1}^{n}(Z_o - \bar{Z}_o)^2}}$$

Table 1 shows the estimated RMSE and Pearson correlation coefficient for each surface modeled with the different interpolators used. Both the RMSE and the correlation coefficient demonstrate that the incorporation of canopy elevation data in the estimation of terrain elevations in cases where it is impossible to densify the sample grid significantly improves the adherence of the results to the sampled reality.

Table 1 - Analysis of error and correlation of the models vs control points.

Interpolator	RMSE (m)	Pearson
SKLVM	2.75	0.995
Triangulation	4.52	0.985
SK	4.76	0.986
IQD	5.04	0.984
NN	6.70	0.967
PhotoScan	5.90	0.982
Minimum Curvature	7.72	0.958

The graphic analysis of the effectiveness of the interpolator methods used two categories of diagrams: scatter plot (Galton, 1886) and box-plot (Tukey, 1969). In the scatter plot (or dispersion graph), the estimated data ("x" axis) and the sampled data ("y" axis) were plotted, to visualize the intensity of the relationship between the estimated models and the sample data (Figure 11). The observation of these graphs confirms the superior quality of the estimate by SKVLM, where the estimated and sampled data exhibit the least dispersion.

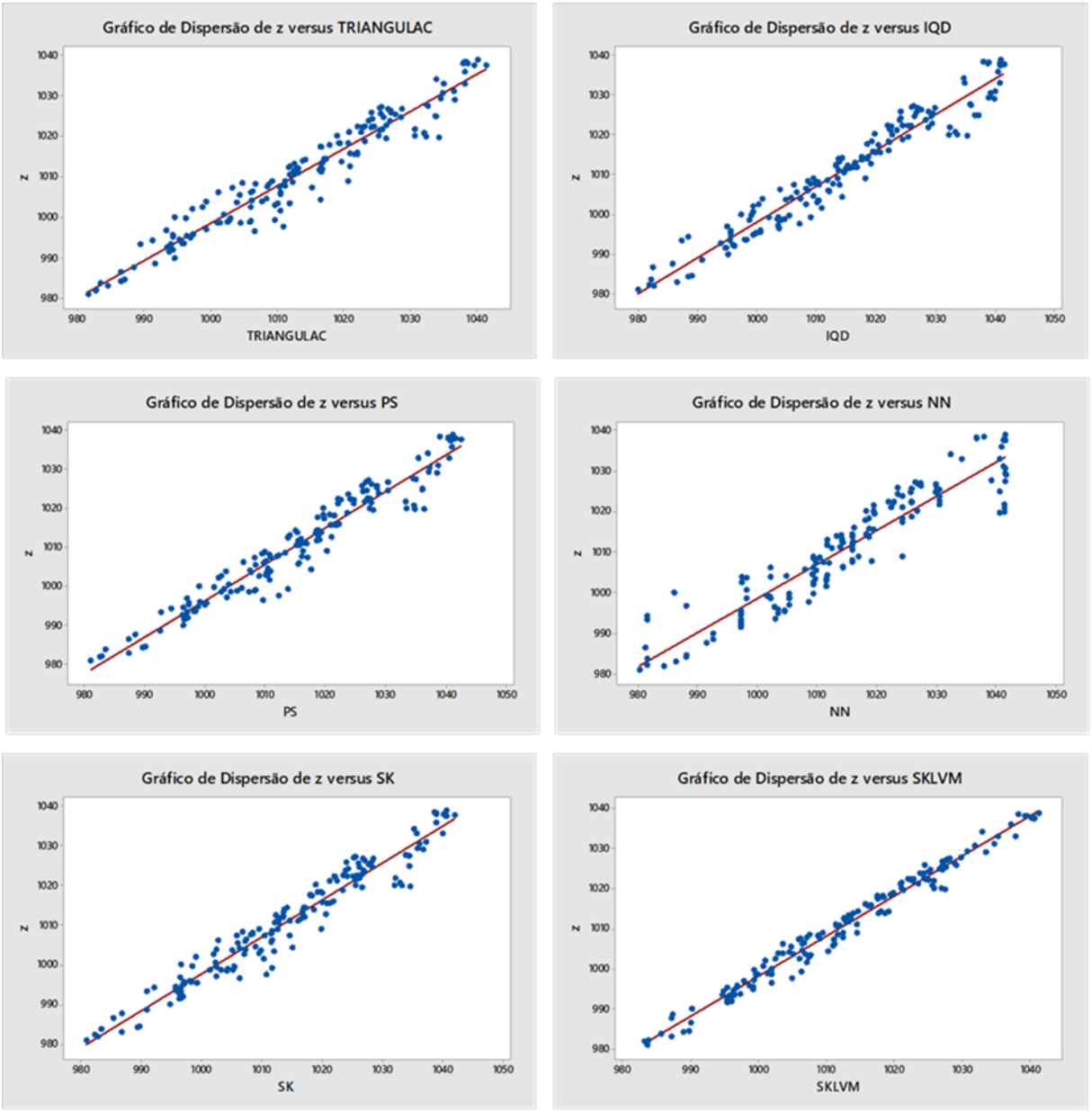


Figure 11 - Scatter plot of estimated values vs sampled values.

In the box-plot, the differences between the estimated data and the sampled data were plotted in order to assess the distribution of errors of the estimate (Figure 12). It was noted that the amplitude of the difference between the data estimated by SKVLM and the sampled data was smaller than for the data estimated by other methods.

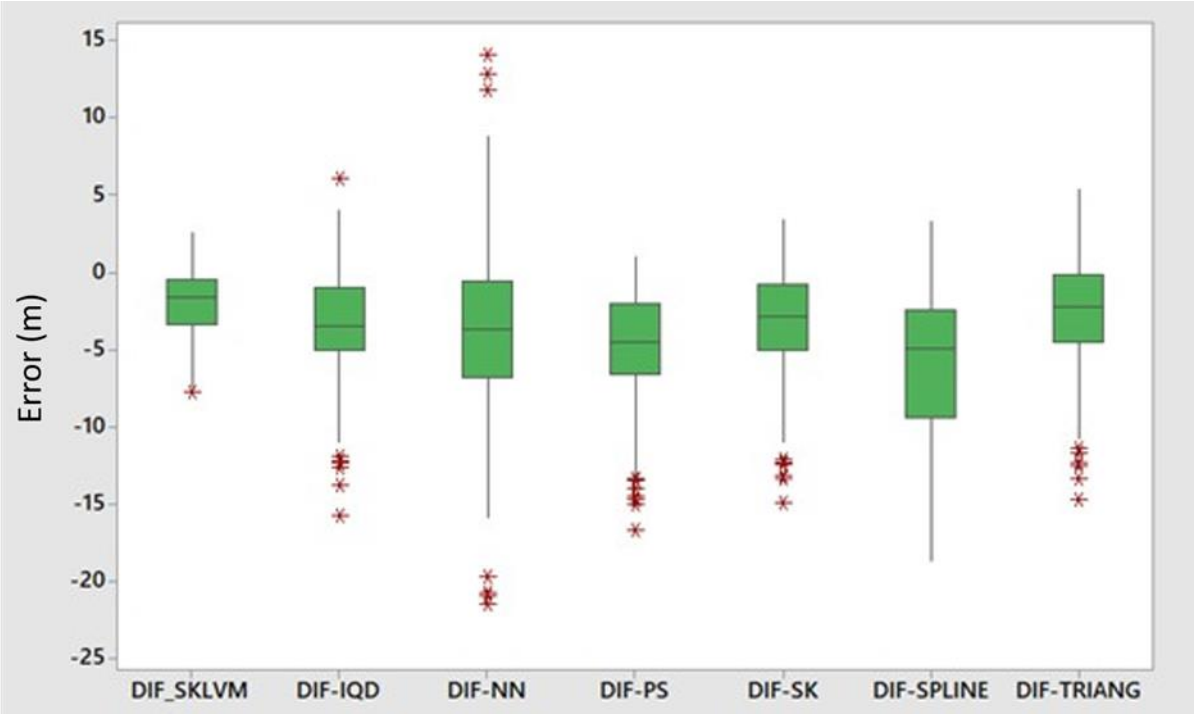


Figure 12 - Box-plot of the differences between estimated vs sampled data.

In summary, the aforementioned graphs show that the estimate of the terrain surface by SKVLM has the highest correlation with the reference data and the lowest extreme error values, emphasizing the greater reliability of the results of the estimates of this interpolator.

4. Conclusions and Recommendations

The use of surface data acquisition technologies to generate digital models of the terrain is increasing every day. Such methods include drones, laser scanners and orbital sensors. The increasing use is due to the low costs involved and facility in obtaining information used across the industry.

At the same time that the advancement of these technologies speeds up the acquisition of data, generating thousands or even millions of geo-referenced information units, there is also a need to improve the tools and techniques used in data processing, transforming such data into useful, high quality information that is fit for purpose. Although several commercial tools exist, one major problem faced by them all is the interpolation of surface data in regions of dense vegetation where the limited amount of terrain information generates distortions in the generated surfaces.

There is therefore a need to improve interpolation techniques that can be used in these regions to reduce errors related to the low density of terrain sampling points. As a result, the present study proposed a methodology for incorporating vegetation elevation data as a secondary variable to estimate terrain surface elevation.

The objective of this study was to investigate how the generation of Digital Elevation Models (DEM) can be more faithful to the true shape of the terrain in locations with a dense vegetation covering, interpolating topographic data from the soil and the top layers of the vegetation canopy as a secondary variable by Simple Kriging with Varying Local Means (SKVLM). For this purpose, the technique proposed in this study, simple kriging with varying local means using information of the top of the vegetation as a secondary variable, demonstrated efficiency and robustness when compared to the interpolating methods conventionally used as triangulation, including the inverse of the square of the distance.

While the proposed method presents satisfactory results in surface modeling in areas of dense vegetation, it is recommended that tests are carried out in other areas with different vegetation types and morphology, to assess its replicability in more diverse situations.

To continue such studies, the implementation of a computational algorithm to automate the data processing of the proposed method in software designed for surface modeling with point clouds, increasing flexibility and reducing manual processes when obtaining results.

References

B. Delaunay: Sur la sphère vide, Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk, 7:793–800, 1934

Galton, F., Okamoto, S., Galton, F. and Pearson, K., 1885. Regression Towards Mediocrity In Hereditary Stature. London: Harrison and Sons.

Goovaerts, P., 1997. Geostatistics For Natural Resources Evaluation. Oxford: Oxford University Press.

Hallak, R. and Pereira Filho, A., 2011. Metodologia para análise de desempenho de simulações de sistemas convectivos na região metropolitana de São Paulo com o modelo ARPS: sensibilidade a variações com os esquemas de advecção e assimilação de dados. Revista Brasileira de Meteorologia, 26(4)(0102-7786), pp.591-608.

Remy, N., Boucher, A. and Wu, J., 2009. Applied Geostatistics With Sgems. Cambridge, UK: Cambridge University Press.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? American Psychologist, 24(2), 83–91. https://doi.org/10.1037/h0027108

Shepard, Donald (1968). «A two-dimensional interpolation function for irregularly-spaced data». Proceedings of the 1968 ACM National Conference. pp. 517–524. doi:10.1145/800186.810616

DOES MORE INFORMATION INCLUDED IN SPATIALLY DISTRIBUTED FIELDS LEAD TO AN IMPROVED MATCH TO OBSERVED DEPENDENT VARIABLES?

Bo Xiao (1) - Claus Haslauer (2)* - Geoff Bohling (3) - András Bárdossy (4)

University of Tübingen, Center for Applied Geoscience (ZAG) (1) - University of Stuttgart, Institute for Modelling Hydraulic and Environmental Systems, Vegas (2) - Kansas Geological Survey (3) - University of Stuttgart, Institute for Modelling Hydraulic and Environmental Systems, LHG (4)

** Corresponding author: claus.haslauer@iws.uni-stuttgart.de*

Abstract

The incentive of this presentation is the age-old quest of stochastic hydrogeology: Are we able to better match observed long-tailed breakthrough curves by an improved description of the spatial dependence of saturated hydraulic conductivity (K)?

This contribution considers two innovations: We include more information than usual by incorporating multiple types of observations at non-collocated locations (data fusion), and we extract more information than usual from the available measurements by analysing statistical properties that go further than typical second-order moments-based analyses (non-Gaussian geostatistics).

The evaluation of these innovations in geostatistical simulation methodologies of spatially distributed fields of K is performed against real-world tracer-tests that were performed at the site of the K measurements. The hypothesis is that fields that contain the most information match the observed solute spreading best.

The spatially distributed K-fields were geostatistically simulated using the multi-objective phase annealing (PA) method. To accelerate the asymmetry updating during the PA iterations, a Fourier transform based algorithm is integrated into the three dimensional PA method. Multiple types of objective functions are included to match the value and/or the order of observations as well as the degree of the "non-Gausianness" (asymmetry). Additionally, "censored measurements" (e.g., high-K measurements above the sensitivity of the device that measures K) are considered.

The MAcroDispersion Experiment (MADE) site is considered the holy grail of stochastic hydrogeology as among the well instrumented sites in the world, the variance of the hydraulic conductivity measurements at the MADE site is fairly large and detailed observations of solute spreading are available. In addition to the classic K-measurements obtained via 2611 flowmeter measurements, recently a large set of 31123 K-measurements obtained via direct push injection logging (DPIL), are available, although not at the same locations where the flowmeter measurements were taken.

The influence of including different types of information on the simulated spatially-distributed fields of K are evaluated by analyzing the ensemble spatial moments and the dispersivity of numerical conservative solute tracer tests performed using particle tracking and the numerical groundwater flow and solute transport model HydroGeoSphere. The improved dependence structure of K with all of the above knowledge contains more information than fields simulated by traditional geostatistical algorithms and expected as a more realistic realization of K at the MADE site and at many other sites where such data-fusion approaches are necessary.

A CONDITIONAL RANDOM FIELD APPROACH TO GEOSTATISTICAL MODELLING

Colin Daly (1)*

Schlumberger Ltd. (1)
** Corresponding author: cdaly@slb.com*

Abstract

One of the most common requirements in Geostatistical modelling is to produce models of one or more primary variables when one or more secondary variables and/or trends are available to provide guidance to the modelling process, for example, through use of co-variograms and co-kriging. A second, equally common requirement is to produce simulations of the type of heterogeneity that is expected of the target variable(s). This latter requirement is particularly important in cases where the geostatistical model is to be used later in processes which depend non-linearly on the heterogeneity, such as fluid flow. This paper considers a method which combines a machine learning algorithm with standard spatial modelling tools. The method for integration of the two methods is by embedding the model as an oracle that makes predictions into the training process by a form of cross validation. A conditional random field approach is used to estimate distributions at each target location from which estimates and realizations are built. The method is applied to a synthetic model with two target variables showing mixed discrete/continuous behavior and it is compared with two standard geostatistical procedures on that model.

Keywords: Embedded Models, Geostatistics, Machine Learning, Ember

1. Introduction

To make a rather gross approximation, one might argue that the great success that Deep Learning has made in classification problems for spatial data is due to the marriage of a machine learning technique, in that case Neural Networks, with pre-existing ideas from pattern recognition such as hand designed filters that produce feature sets for classification. The major step forward was that the convolution net allowed for the filters themselves to be learned LeCun et al. (1998). In spatial estimation, the preexisting technology are spatial estimators such as kriging. In this paper, we take a small step to bringing such estimates together with machine learning techniques for non-linear regression which are quite powerful at working with many variables.

For the problem of spatial estimation or stochastic simulation, the observed data is generally one of 3 types, direct observations of the variable of interest - or target variable, secondary measurements which are only indirectly related to the target variable but are more widely observed and environmental variables which provide information about the location and factors that may affect the target variable. The direct target variable data is generally known at irregular locations and the set of target locations at which estimates, or realizations are required may also be at arbitrary data locations or on a grid. As an example, in the mining or oil industries, the target data is often known at borehole locations only. The secondary variables may be seismic or electromagnetic observations. The environmental variables are to do with location in space but may be more general than simple spatial coordinates. For example, if porosity is a target variable, it may be influenced by such environmental variables as physical depth below the surface, relative depth within a stratum, distance to a fault, zone/region of the field or thickness of the strata bed. This shows that there may be many known variables which are related to the target variable and which could provide information for an estimate. In a

classical geostatistical model, the secondary variables are often used for cokriging and the environmental variables are trends. Well known problems for the classical method are that a) it doesn't generally scale well to estimates with many secondary or environmental variables especially if they are non-linearly related to the target variable or to one another and b) when passing to simulation, assumptions of stationarity are generally required. Either the variables themselves are considered stationary, perhaps after subtraction of a trend (generally an environmental variable) or the relationship between the variables is considered stationary.

The irregularity of the observed data prevents a simple application of a convolution approach for the target variables. The alternative considered here is to use a machine learning algorithm to estimate conditional distributions at each target location. Many off-the-shelf machine learning tools can be used to estimate the conditional distributions if they depend only on the secondary and environmental variables but this will mean that they are not able to exploit the spatial relationship between observations of the target variable, except if it is directly related to the spatial relationships within the secondary variables. For example, the porosity variable may be highly spatially variable with short range correlations dominating behavior, but a secondary seismic attribute may miss the shorter-range variability and only pick up low frequency information. The solution here, motivated by the solution in ConvNets is to try and get the learning algorithm to leverage pre-existing tools for spatial estimation. This is done by embedding the classical spatial estimator into the ML method.

In the general case we embed an oracle into the machine learning estimator. An oracle, when given some observed data and a target location makes a prediction of the value of the target variable at that location. Different calls to the oracle will generally use different sets of observed data, chosen according to some random procedure and so the oracle is a random variable. In this paper, the oracle is kriging which gives different predictions at a location depending on what data is used for the call. The set of estimated distributions, one per target location, is called the envelope in this paper. It can be thought of as a generalization of a trend. Instead of having just a single value at each location, there is now a distribution. A simulation is a sample from the envelope. It is conditional if it matches the observed value at the data locations.

2. Method

Conditional Random Fields (CRF) (Lafferty et. al 2001) avoid construction of the multivariate law. The advantage in direct estimation of each conditional distribution in the envelope compared to a generative Bayesian model is that no effort is expended on establishing relations between the numerous predictor variables. In a full spatial model these involve stringent hypothesis such as the stationarity of the property of interest (perhaps coupled with some simple model of trend) and the stationarity of the relationship between the target variables and the explanatory variables (e.g. the hypothesis that the relationship between porosity and seismic attributes do not change spatially). The principle impact of stationarity in the classic model is seen in stochastic realizations which need to invoke the full multivariate distribution and therefore lean heavily on the hypotheses. This can be greatly reduced in the current proposal.

The form of CRF that we use here to calculate the envelope accommodates and embeds existing spatial models using a Markov type approximation. Let $Z(x)$ be a target variable of interest at the location x, and let $\boldsymbol{Y}(x)$ be a vector of secondary or auxiliary variables observed at x. Let $\{Z_i, \boldsymbol{Y}_i\}$ be observations of the target and secondary variables observed in the field, i.e. Z_i denotes the value of the target variable $Z(x_i)$ at training location x_i. Finally let $\boldsymbol{M}(x) = \boldsymbol{f}(\{Z_i, \boldsymbol{Y}_i\})$ be a vector of pre-existing estimators of $Z(x)$. Then the form of CRF that we require is that the conditional distribution of Z(x) given all available data $\hat{F}(z|\boldsymbol{Y}(\boldsymbol{x}), \{Z_i, \boldsymbol{Y}_i\})$ satisfies,

$$\hat{F}(z|\boldsymbol{Y}(\boldsymbol{x}), \{Z_i, \boldsymbol{Y}_i\}) = E\left[\mathbb{I}_{Z(x)<z}|\boldsymbol{Y}(x), \{Z_i, \boldsymbol{Y}_i\}\right] \approx E\left[\mathbb{I}_{Z(x)<z}|\boldsymbol{Y}(x), \boldsymbol{M}(x)\right] \qquad (1)$$

This hypothesis states that the conditional distribution of $Z(x)$ given all the secondary values observed at x and given all the remote observations of $\{Z_i, \boldsymbol{Y}_i\}$ can be reduced to the far simpler conditional distribution of $Z(x)$ given all the secondary values observed at x and the vector of model predictions at x. The focus is now on trying to estimate the right-hand side of equation 1 at each location. Notice that equation 1 is not an exact Markov hypothesis. There is a loss of information. In particular, using $\boldsymbol{M}(x)$ merely as an oracle that makes predictions at x means that the estimated conditional distribution does not collapse to a singularity at the data locations. The envelope resembles a trend and conditional simulations will require a conditional sampling from it. In this paper the method is extended to more than one target variable in a simple way. A model for each target variable is embedded and the right side of equation 1 is decomposed so that varaiables are modelled sequentially.

The method used to estimate the conditional distribution is a Quantile Random Forest (Meinshausen, 2006) modified to handle embedded variables (e.g kriging) (Daly, 2020). The result is an estimate of the distribution at each location. Stochastic modelling is developed by using that $U(x) = F(Z(x)|Y(x) = y)$ is a uniform random variable. Unlike in the typical usage of a Cloud transform (Kolbjørnsen and Abrahamsen, 2005) F is explicitly calculated at each x, so it is possible to, and shown in the paper how to, write down a relationship for the variogram required to generate U(x) if the uniform field is assumed to be generated by a Gaussian Random field. When F is nearly Gaussian itself, this variogram is just the variogram of scaled residuals of the Ember estimate making it simple to calculate. Concisely, with Ember the distribution F at each location explicitly depends on the variable to be modelled at neighbouring locations through the embedded variable and the simulation is modelling the residuals that are not already captured using a variogram which can be explicitly calculated and modelled. Contrasting that, with Cloud transforms, the distribution at any target location is not made explicit. Any two locations with the same secondary variables have the same distribution. The dependence on location and on the target variable comes through the distribution of uniform values that sample from them. Thus, all information relating to the target variable is relegated to the Uniform Random Field. In Ember, just the residuals are modelled by the uniform random field.

3. Results

To demonstrate the method, a synthetic example is used. This is partly based on a real case study that the author has applied the method to. For this example, the focus is on exploring the method, so the synthetic model used is constructed with this in mind rather than being geologically reasonable in all aspects. A common approach to modelling petrophysical property such as porosity is to do so in two steps. Firstly, a net-to-gross (NTG) property is constructed assigning a value to each cell on a grid. This is the proportion of reservoir sand in that grid cell. In a second step, the porosity of the net sand is estimated (NET). The actual porosity in the cell is the product of the two, PORO = NET x NTG. The reason for this is that many properties, such as fluid saturation, that will be derived from the model depend in a nonlinear way on porosity, so that, averaging over a cell, $f(PORO) \neq NTG \times f(NET)$, so having a NTG property allows for a meaningful calculation of saturations as well as other important engineering variables like relative permeability. Figure 1 shows the 'truth', namely the synthetic model, for the NTG and the NET variables as well as PORO in a typical layer of the model.

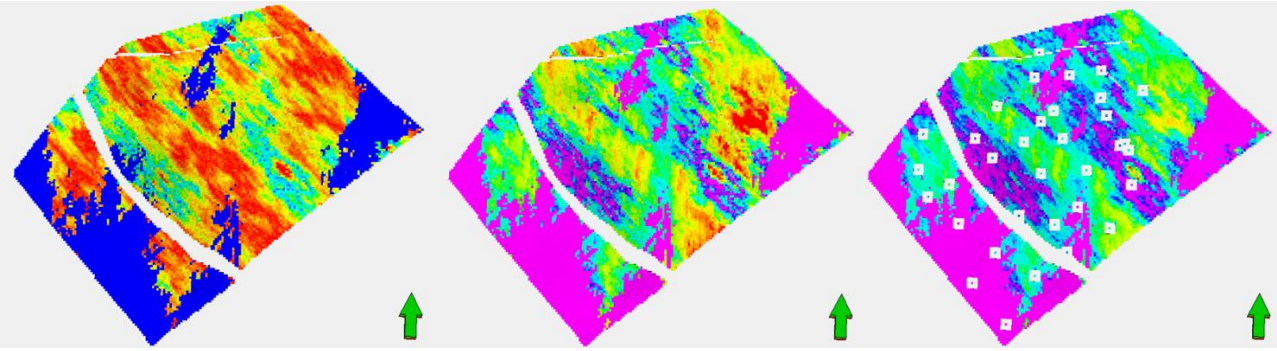


Figure 1 – Synthetic 'truth'. Left to right, NTG, NET and PORO for a representative layer in the model. The locations of the well samples are superimposed on the PORO image.

The 'truth' cases were constructed in a relatively complicated, discontinuous and non-linear way from 3 Gaussian Random Functions. For a first example, we will assume 2 of the Gaussian RF are known as additional, or secondary variables. For this example, these provide more information than would be available in a normal Petroleum or Mining study, but it tests the method when constraints are high. The information available will then be reduced by using only one of the additional variable, and then by using no additional variable. Figure 2 shows the additional variables, called *IndepOfSecondary* and *Secondary* respectively, as well as a 3d crossplot of the two variables with NTG on the z axis which has been calculated empirically using only the well data. The compact nature of the crossplot suggests that a simpler algorithm than Ember, such as a Cloud Transform will give a good result with this quality of secondary data. This will also be considered, as well as a mixed SIS/Gaussian co-simulation to explicitly handle the fact that the data is not drawn from a continuous distribution but is mixed discrete/continuous.

It is clear from figure 1 and the crossplot that the target variables are multimodal. In particular NTG has 'spikes' at 0% and 100% and a continuous distribution of values between. Figure 3 shows the results of Ember estimation of the NTG cases including histograms (fig 3 (iv)) where the spikes can be seen clearly to be well preserved in the simulations. Figure 3 (i) shows the estimate taken by returning the P50 of the envelope at each location. For this study, it give a significantly better estimate in terms of MSE than the mean of the envelope, fig 3 (iv). Measures of uncertainty, such as the Inter Quartile Range (IQR = P75-P25) can be immediately read from the envelope estimate. The IQR is shown in 3 (iii) and can be compared to the actual errors which for the P50 case are shown in 3 (v) and for the mean case in fig 3 (vi). Notice that the IQR looks quite different to the distribution of NTG, but is clearly related to the 'geometry' of the secondary variables which are the dominant source of information and moreover it reflects the actual errors made in estimation very well.

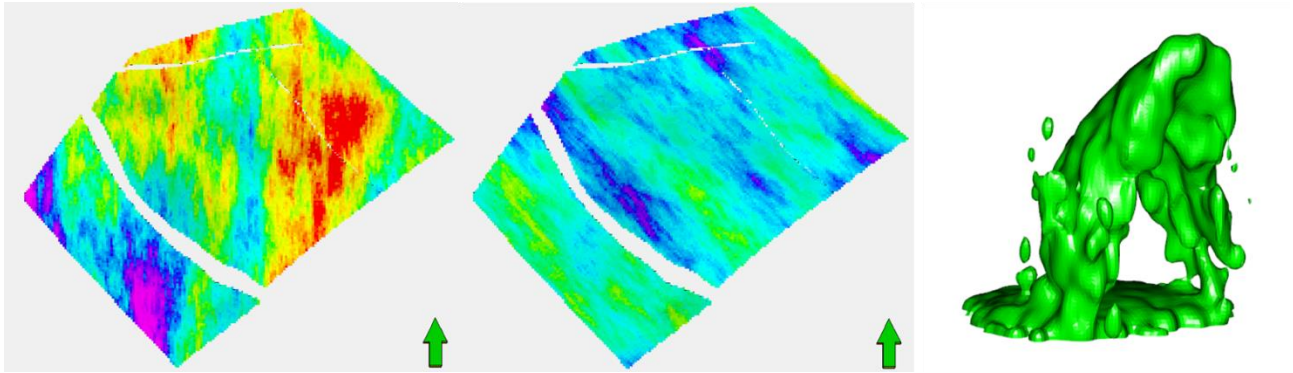


Figure 2 – Secondary variables, 'IndepOfSecondary' on the left, 'Secondary' in the center and a 3d density cross plot of the vector (IndepOfSecondary, Secondary, NET) on the right.

As noted, it is rare in the extractive industries to have secondary data of the quality used so far. To explore how the estimates change when less data is available, consider two situations 1) The variable *IndepOfSecondary* is no longer available but *Secondary* is still available, and 2) Neither of these variables are available and the model is dependent only on the embedded kriging as well as geometric trends in X,Y, and Z directions (which are quite weak for this model). Figure 4 shows the P50, IQR and actual errors made in modelling for these two cases. The first case roughly corresponds to the level of information that one might expect with very good seismic (though slightly more complex in this case because of the U shape of the relationship, see figure 3 (ix)). The second case is the situation when there is no seismic or it is worthless at discriminating reservoir properties. The colours have been kept at the same scale, so it is possible to see that the P50 estimate becomes increasingly washed out and the IQR becomes increasingly more dark (higher values) as less information is used in the modelling. This is the sort of situation where the modeler resorts to stochastic simulation to solve certain problems which depend on non-linear functions of modelled values or on preservation of the extremes, such as fluid flow.

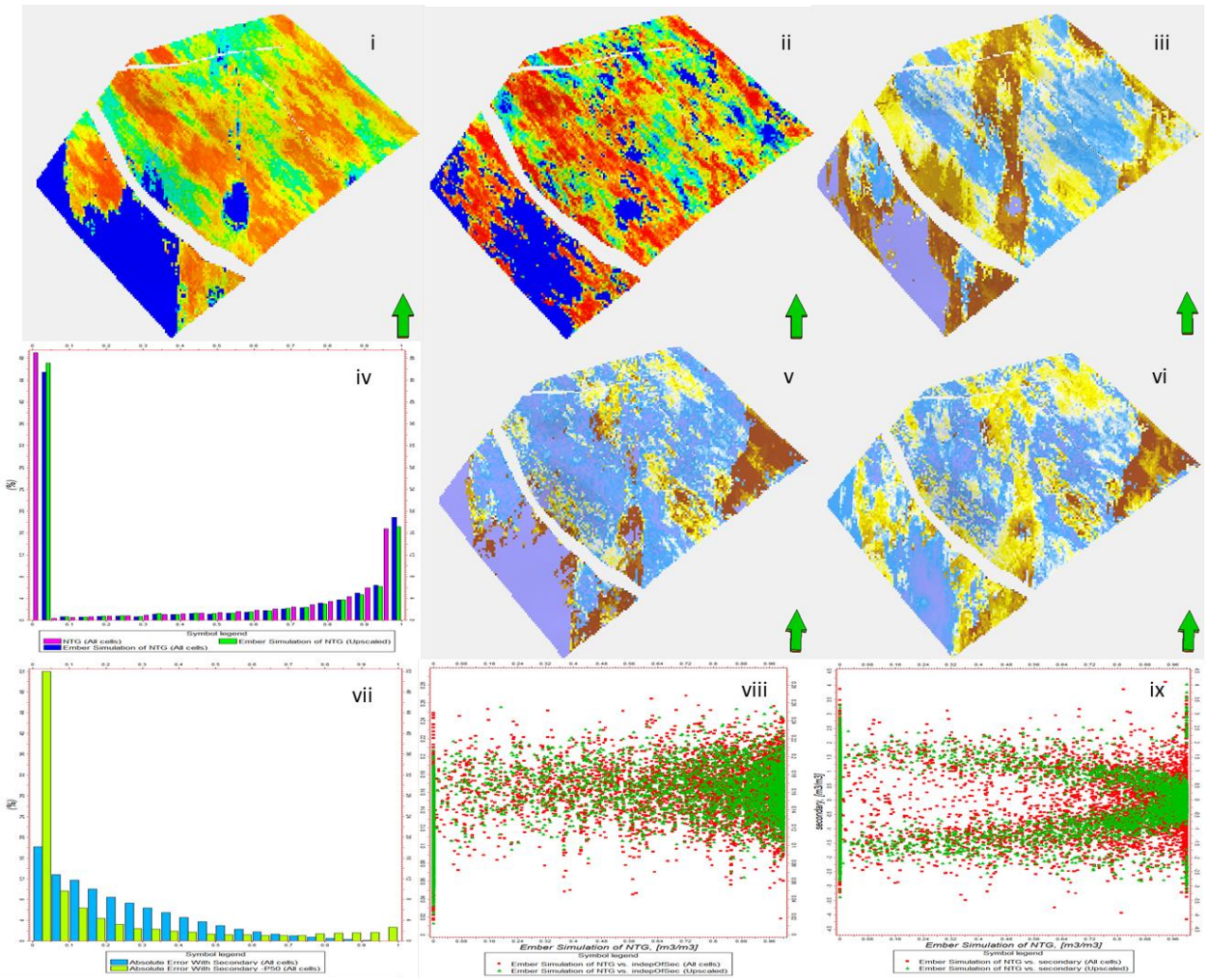


Figure 3. Results of Ember model. (i) P50 Estimate, (ii) Simulation, (iii) Inter Quartile Range of envelope, (iv) NTG histogram showing all data, at wells only and of the simulation, (v) Actual estimation errors from P50 of envelope, (vi) Actual estimation errors from Mean of envelope, (vii) Histogram of actual errors for P50 and Mean estimators, (viii) Cross plot of Ember simulation with 'IndepOfSecondary', (ix) Cross plot of Ember simulation with 'Secondary'

To look at Ember simulation but also to compare with two other methods, consider the simulation of the second variable NET. This variable should be constructed at the same time as NTG to ensure that any further derived variables are consistent. The two other methods used are the Cloud transform and a standard

SIS/Gaussian simulation. The Cloud transform makes explicit use of the pointwise multivariate relationships between variables. In a first step, the NTG is simulated using a P-field to sample from the multivariate relationship shown on the right in figure 2. NET is subsequently generated using full multivariate relationship between NTG, NET and the two additional variables. The implementation of Cloud transform used here is actually just a simplification of Ember – basically it is the Ember algorithm without using embedded models or environmental variables. As such, it does benefit from an estimate of variograms and so gives better results than some commercial applications of the Cloud transform. For the SIS/Gaussian algorithm, the well data is first split into 3 'facies', one with 0% NTG, another with >95% NTG and a third with values between. These facies are modelled with SIS using a facies probability function to condition to the additional variables and then a Gaussian simulation is performed for each facies. The NET variable is made using cosimulation with NTG (correlation is about 0.8). The resulting 3 simulations are shown in figure 5. All models have been constructed with the relevant variograms, trends etc. fitted to the data. The same uniform random field is used for Ember and Cloud transform which makes direct comparison easy, but this was not possible for the SIS model. Histograms of the absolute values of the errors of simulation for the 3 methods are shown in figure 6. Ember gives the best result, as seen in table 1 but the Cloud transform is nearly as good because the two additional variables are by far the dominant variables. The SIS/Gaussian model is considerably worse. It does not appear to leverage the information from the additional variables as well as the other two techniques. While not shown for reasons of space, this is also evident in the reproduction of the cross plots between the simulated result and the additional variables, where the SIS result does not do a good job in respecting the input information compared to the other two methods. While other types of facies models, such as object modelling, were not considered in this short study, it seems likely that results would not be significantly better as SIS usually compares favourably to many other types of facies modelling in terms of conditioning to data.

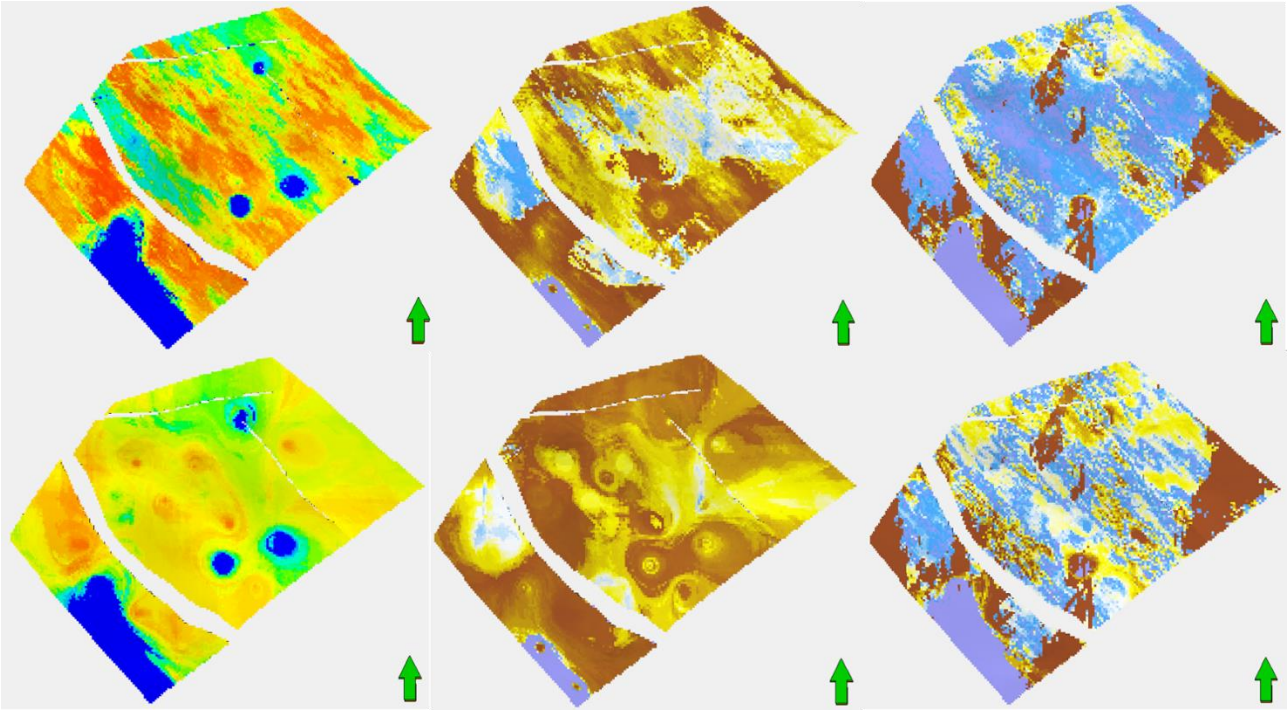


Figure 4. Top Row are Ember results removing the 'IndepOfSecondary' variable. Bottom row uses neither of the additional variables and so only uses embedded kriging as well as geometric variables. Left is P50, center is the IQR and the actual errors in estimation are on the right.

Like the Ember case, the results deteriorate in quality when less additional data is available for the Cloud transform model. The deterioration, as shown by the histogram on the right in figure 6 is somewhat larger that is the case for Ember particularly where the second mode is more pronounced. This second mode is due to values of non-net sand (0% NTG) being 'misclassified' as net sand. When using 2 additional variables the percentage of such misclassifications is 15.2% for Ember (and a similar result for Cloud transform). With only one variable, the misclassifications are 24.2% and 32.4% for the two types of model respectively. The SIS/Gaussian model had 37.2% misclassification already when using 2 additional variables and does not degenerate with less data suggesting that it did not make efficient use of the additional variables in this example.

Finally, returning to one of the original question, can derived variables such as PORO, or fluid saturation be handled correctly? If the relationship between the variables in simulations has been handled correctly, then this should work. For example, the histogram of a model of PORO produced as a product of the two modeled variables is shown in figure 7. The histogram of PORO values at all modelled locations is shown beside the true observed ones and is seen to match well. A crossplot of PORO with one of the additional variables also shows good agreement between the modelled locations and the observed locations. Visually, the simulation looks plausible. And the errors in estimation/simulation are comparable in scale to those made in modelling of the primary variables NTG and NET.

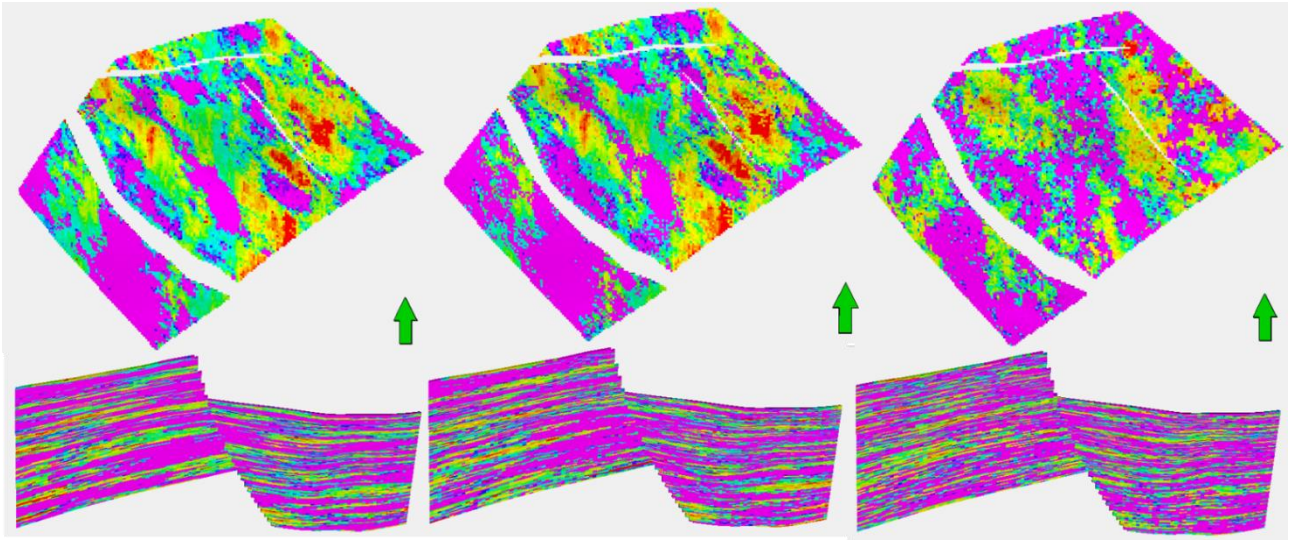


Figure 5. Simulations when both additional variables are used. On the left is Ember, center Cloud transform and on the right is the SIS/Gaussian simulation.

Table 1 – Errors for the three types of model. Standardized so that the mean absolute error of estimation with P50 for Ember is 1. The 2nd column refers to number of additional variables used in model.

Method	Num Addition	P50 Error	Mean Error	Sim Error
Ember	2	1.0	1.23	1.32
Cloud	2	1.24	1.23	1.45
SIS	2	-	-	2.09
Ember	1	1.41	1.65	1.83
Cloud	1	-	-	2.40
Ember	0	1.50	1.72	1.93
SIS	0	-	-	2.12

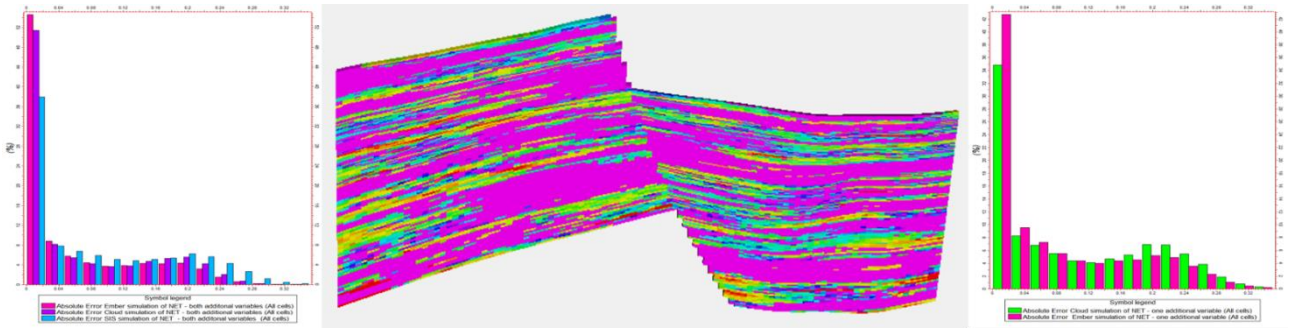


Figure 6. Left. Absolute Errors of simulation when both additional variables are used. SIS, in blue, performs worse than the other two methods. Centre: The cross section of the 'True' NET data. Right: Simulation errors with only one additional variable. SIS result not included. Ember errors in red, Cloud in green.

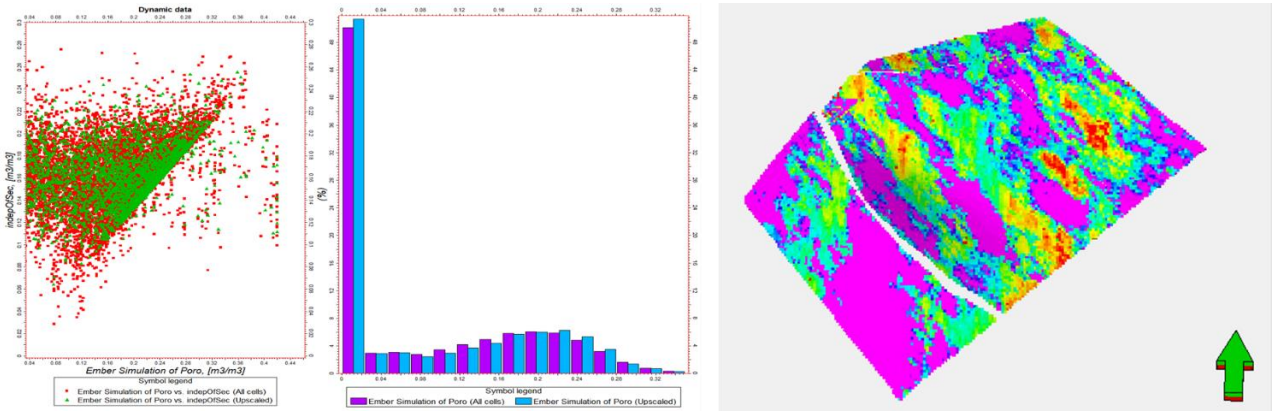


Figure 7. Left. Plot of derived PORO variable with 'Secondary' variable. Green are observed, red are modelled. Centre: Histogram of PORO. Model matched data well. Right: Simulation of PORO.

4. Discussion and Conclusions

The rapid development of machine learning algorithms has led to significant changes in classification type problems of spatial modelling. Progress has been less rapid for regression type spatial problems. The Ember algorithm combines ML with classic geostatistical modelling by a cross validation inspired embedding of models into a well-known algorithm which is stable, is not prone to overtraining and is known to converge to conditional distributions at target locations. The algorithm was run on a synthetic model and compared favorably to two classic geostatistical methods. One of the main reasons for creation of this algorithm was its simplicity to use compared to standard methods, allowing novice users to easily produce results comparable to, or better than existing methods. Runtime is typically 2-4 times longer than running a Gaussian simulation, so given that very little time is spend in data analysis compared to the standard approach, it accelerates the workflow and reduces the potential for human error. For uncertainty/scenario models, users may well still wish to construct many different types of model to explore possibilities, but this method may prove to be a useful addition to modeler's toolkit especially in getting a quick result. While this paper has looked at the most simple problem of direct modelling of heterogeneity without any physics based modelling (distinct for seismic inversion for example), a possible extension to the work here would be to use this type of algorithm in such cases where it would provide a more constrained prior than the Gaussian models that are often used in such cases.

References

Daly, C (2020) An Embedded Model Estimator for Non-Stationary Random Functions using Multiple Secondary Variables. arXiv:2011.04116 [stat.ME]

Kolbjørnsen, O and Abrahamsen, P (2004) Theory of the Cloud Transform for Applications. Geostatistics Banff, Springer, Dordrecht, 2005, pp45-54

Lafferty et al., (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings ICM-2001. Vol. 86 Issue 11, pp2278-2324.

LeCun et al. (1998) Gradient-Based Learning Applied to Document Recognition. Proc. Of the IEEE 1998.

Meinshausen, N (2006) Quantile Random Forests. J. Machine Learning Research 7, 983-999

GEOSTATISTICAL APPROACH TO ESTIMATE THE LOCAL SEISMIC HAZARD IN MUNICIPALITIES OF ANTIOQUIA, COLOMBIA

Lilian Posada (1)* - Luis Sánchez (1) - Gabriel Rosado (1) - Jorge Medina (1)

Universidad Nacional de Colombia, Sede Medellín, Facultad de Minas (1)
** Corresponding author: lposada@unal.edu.co*

Abstract

The last earthquake activity in the state of Antioquia, northwestern Colombia, occurred on October 17th and 18th of 1992, with 6.6 Mw and 7.1 Mw magnitudes, respectively. These events represented significant damage in the region, where 32 municipalities were affected, 17 of which showed high damages in their constructions (3,500 homes, 840 km of roads, 141 educational centers, 15 police inspections, and 7 health centers were affected and 8 deaths were reported). It also caused ecological and social damages, for soil liquefaction, landslides, reactivation of the Cacahual volcano activity and the fully evacuation of Murindó municipality. In the city of Medellín (the most important municipality of the state, with 2.5 million inhabitants), the insured losses were estimated at US$ 11 million.

The Antioquia's engineers propose the development of large engineering projects (eg. The Ituango hydroelectric plant with an investment of 9,8 billion Colombian pesos in December 2018), so it is important to assess the macro-seismic threat for better planning of its territory. A geostatistical analysis of local magnitude, Ml, reported by the Colombian Geological Service for the state of Antioquia, is showing an important surface microseismic activity. The variographic study reflects an anisotropic behavior of geometric type for the data; Y10 (h) = 0.42 Ml2 + Sphere (0.10 Ml2, 171 km) and Y100 (h) = 0.42 Ml2 + Sphere (0.10 Ml2, 60 km). For this to be right, the cortical failure with the greatest impact in the region is of NNE tendency and will serve as a basis for the design of future local seismological networks.

The geostatistical technique of the Polygonal Kriging was used as a global estimation method for each of the 125 municipalities that make up the Antioquia region. The Kriging standard deviation relative to the average of the estimated micro-seismic was used to define a geostatistical module that quantifies the expected macro-seismic activity for the evaluation in earthquake risk management. The municipalities with the highest internal seismic hazard reported with this methodology was Ituango, Dabeiba, Uramita, Urrao, Cañasgordas and Murindó. The Aburrá Valley, with an estimated population of more than 4 million inhabitants, was considered to have a high external seismic hazard due to the Atrato-Urabá fault system (Sutura Dabeiba-Pueblo Rico).

GEOSTATISTICAL INTERPRETATION OF SPATIAL DISTRIBUTIONS OF POROUS MEDIA ATTRIBUTES THROUGH GENERALIZED SUB-GAUSSIAN MODELS

Martina Siena (1)* - Monica Riva (1) - Alberto Guadagnini (1)

Politecnico di Milano, Dipartimento di Ingegneria Civile e Ambientale (1)
** Corresponding author: martina.siena@polimi.it*

Abstract

Geostatistics has allowed quantification of spatial variability patterns displayed by quantities such as rock mineral content, permeability, or porosity. Traditional approaches consider such quantities as multivariate Gaussian (correlated) random fields. There are clear evidences that several hydrologic (and other) quantities exhibit non-Gaussian behavior over a multiplicity of spatial scales. As an example, quantities such as log hydraulic conductivity and the ensuing spatial increments are generally non-Gaussian. Such a behavior is manifest through patterns including distributions of increments (at all separation scales (or lags)) displaying sharp peaks and heavy tails that appear to decay as lag increases. These and other aspects of statistical scaling are manifest in porous as well as fractured media characterized by either one or a hierarchy of spatial correlation scales. We present an extended formulation of the Generalized sub-Gaussian (GSG) model proposed by Riva et al. (2015), according to which the quantity of interest is modeled as the product of a Gaussian random field, G(x), x being a position vector, and a subordinator, U(x), independent of G. Our extension is designed to include multiple subordinator types, thus allowing for increased flexibility and model uncertainty. We consider (i) Log-normal; (ii) Pareto; and (iii) Gamma distributional forms of U, each being characterized by two parameters, respectively controlling the shape and spreading of the probability density function (pdf) of the distribution. We provide the theoretical formulation of the GSG process, Y', and associated spatial increments, ΔY, evaluated for a generic subordinator, as well as for the three selected subordinator forms. The ensuing collection of GSG models is then employed to analyze increments and parent variables associated with two datasets: (i) a two-dimensional distribution of surface-roughness data, collected on a (millimeter-scale) calcite sample resulting from induced mineral dissolution (Dataset 1); and (ii) a one-dimensional field-scale spatial distribution of neutron porosities, collected along a km-scale vertical borehole in a reservoir (Dataset 2). For both datasets, all GSG models are in better agreement with the target sample pdfs than the Normal distribution, with a marked non-Gaussian signature detected in Dataset 1. The degree of similarity between sample and analytical pdfs of the parent variables and their increments, quantified through the Kullback-Leibler divergence, indicates the GSG model with Pareto and Gamma subordinators as the best for the interpretation of Dataset 1 and Dataset 2, respectively. These results suggest that the implementation of multiple subordinators within the GSG framework can enhance the flexibility of the model and improve the accuracy of the interpretation of the (scale-dependent) statistics of a given dataset.

Riva, M., S.P Neuman, and A. Guadagnini (2015). New scaling model for variables and increments with heavy-tailed distributions. Water Resour. Res. 51(6), 4623-4634.

THE KRI-TERRES PROJECT: COMBINING GEOPHYSICS, HYDROGEOLOGICAL MODELLING AND GEOSTATISTICS FOR BETTER CHARACTERIZING CONTAMINATED SOILS

Mathieu Le Coz (1)* - Léa Pannecoucke (2) - Clémence Houzé (3) - Albane Saintenoy (3) - Xavier Freulon (2) - Charlotte Cazala (1) - Chantal De Fouquet (2)

Institute for Radiological Protection and Nuclear Safety (irsn), Pse-env/sedre (1) - Mines Paristech Psl University, Centre de Géosciences (2) - Paris-sud University, Geops (3)
** Corresponding author: mathieu.lecoz@irsn.fr*

Abstract

The Kri-Terres research project, launched in 2017, aims at improving strategies of characterization of contaminated soils by combining geophysics, hydrogeological modelling and geostatistics. The project consists of three main interconnected stages.

The stage 1 aims at developing a method to infer hydraulic parameters that govern flow processes in soil based on monitoring of water infiltration tests with a Ground Penetrating Radar (GPR). Indeed, radargrams show strong reflections coming from the infiltration bulb evolution, and arrival times can be used within a numerical model framework for inversing three hydraulic parameters related to the Mualem-van Genuchten formalism, namely the saturated hydraulic conductivity Ks, the parameter α, inversely proportional to the air-entry value, and the parameter n, related to the pore-size distribution. The method was applied on Fontainebleau Sand outcrops and results showed that arrival times measured in-field and retention curves, obtained by suspended column laboratory experiments performed on samples, are ranked similarly.

The stage 2 aims at quantifying the influence of variations in soil hydraulic parameters on solute transport resulting from a localized source of contaminant. To this end, sets of random fields accounting for spatial variability of Ks, α and n are generated using lognormal distributions with variances, computed through the analysis of a global soil database. These random fields are then used as inputs to an unsaturated flow-and-transport model to simulate contaminant plume migration. By comparison with the homogeneous medium, Ks-, α- and n- random fields respectively result in (i) 25 (variable Ks), 20 (variable α) and 65% (variable n) increase in plume size; (ii) 0.8, 1 and 1.8 m horizontal offsets of the plume center; and (iii) 20, 30 and 50% decrease in plume circularity.

The stage 3 aims at developing a new geostatistical approach that deals with variability in soil hydraulic parameters for better characterizing contaminated soils. This approach called Kriging with Numerical Variogram (KNV) consists in computing the parameters of the geostatistical model from a set of physically-based flow-and-transport simulations rather than from the measurements. The KNV is assessed on a two-dimensional synthetic reference test case reproducing the migration of a tritium plume within an unsaturated soil with hydraulic properties highly variable in space. The results show that the mean absolute error in estimated activities is 50% to 75% lower with KNV compared to classical geostatistical approaches, depending on the sampling scenario (especially, the performance of KNV increases when the number of measurements decreases).

The Kri-Terres project is supported by the French National Radioactive Waste Management Agency (Andra) under the "Investments for the Future" national program.

BAYESIAN ANALYSIS OF SPATIAL DATA WITH MISSING VALUES

Mohsen Mohammadzadeh (1)* - Samira Zahmatkesh (1)

Tarbiat Modares University, Department of Statistics (1)
** Corresponding author: mohsen_m@modares.ac.ir*

Abstract

When dealing with spatially dependent data in various sciences such as geology, meteorology, oceanography, and other Environmental sciences there is often a notable amount of missing values. Some of the factors affecting measurements, such as environmental and atmospheric conditions, sample unit locations, or the time of collecting observations, make missing data inevitable. Due to dependency between spatial observations, missing values that are located at the spatial or temporal neighbourhoods can include useful information that the retrieval of this lost data can increase the accuracy of the data analysis. Missing data mechanisms are classified into three types: missing completely at random, missing at random and missing not at random. When the first two have occurred, missingness is ignorable and when the last one has occurred, missingness is non-ignorable. Estimating model parameters based on the complete cases will yield unbiased estimates even under missing at random assumption. Also, valid inferences are extracted by Likelihood-based techniques as long as the distribution of the response variable is correctly specified and missing at random assumption is hold. There are situations that there is doubt that missingness is non-ignorable, thus methods that are effective in at ignorable case may introduce bias inferences and it is needed to apply methods that consider the model of missingness process into the inferences. In this talk, we develop a method for analysing spatial data with non-ignorable missing values. A statistically principled approach is to build a joint model, which combines information in the observed data with assumptions about the missing value mechanism. The joint modelling of the data along with the missingness process is proposed by using a shared parameter model technique. To model data and the missingness process, a spatial generalized linear mixed model is employed, and to make an inference, a Bayesian approach is used. Traditionally the MCMC methods are used making inference for Bayesian latent Gaussian models. To overcome the heavy computational costs of operations required for model fitting, estimations and spatial predictions that arise when facing with large spatial data set, an approximated Bayesian approach via Integrated Nested Laplace Approximation is applied that is faster than MCMC while their results are equivalent. Then, the presented models, are evaluated and numerically compared in a simulation study, and their application in a real data example is shown.

TOWARDS THE DEVELOPMENT OF A SUSTAINABLE LAND PRICE-SUBSIDENCE SPATIAL MODEL: A REVIEW

Muhammad Akmal Hakim Bin Hishammuddin (1)* - Wang Jianxiu (1)

Tongji University, College of Civil Engineering (1)
** Corresponding author: akmal_hkm@yahoo.com*

Abstract

This paper presents the research findings for a multidisciplinary research of geotechnical (groundwater) civil engineering, urban and regional planning aspects specifically on the land subsidence control, urban underground space (UUS) exploration, urban development, spatial planning, land price and sustainable development in Shanghai megacity. Since 1960s, land subsidence due to groundwater pumping in megacity Shanghai has been successfully controlled by the government to a minimum level. However, starting in 1990s, the land subsidence rate has deteriorate again even though the net withdrawn volume (NWV) of groundwater has remained unchanged since 1980. This is due to rapid urbanisation, building and population load, as well as underground space construction: tunneling for metro. Many researches have been conducted to understand these UUS induced-land subsidence relations in Shanghai using meso-scale methods of geotechnical analysis. However, limited research has been conducted on the impact of UUS induced land subsidence to the urban economic, especially the land price and modelling it spatially. The aim of this research is to understand the relation of UUS induced-land subsidence and land price by spatial modelling using geographic information system (GIS) at megacity scale of Shanghai. It has been determined in the first research question the extent impact of the factors, however this paper focuses on the second research question of determining spatially and adjust the land price and UUS exploration induced-land subsidence to realise the harmony and sustainable urban development of megacity Shanghai. Most of data points are gathered and searched from online search engines and databases for online scientific journals such as Google, Google Scholar, Baidu, Research Gates and Elsevier. Spatiotemporal analysis has been conducted to analyse the land use change over space and time, the land subsidence rate, UUS exploration and land price. There are more than 100 prominent researches especially journal articles have been retrieved to gather the secondary datasets. The time-frame ranges as early as 1960s-90s, 2000s till 2020. There are five (5) main findings which are (1) The spatial situation in the central business district (CBD) area of Shanghai are still under controlled due to proper underground tunneling technologies and control, (2) The development of Shanghai (North and Southern) new area must be controlled both vertical and horizontal due to the current negative correlation of UUS induced-subsidence with the land price due to existing soil condition, (3) Any existing and new development area in Shanghai must be controlled and monitored to avoid further land subsidence and land price negative changes (4) Further validation is needed of the spatial model proposed for Shanghai especially by analyzing the current situation and predict for future sustainability and probability of efficient spatial and land uses. As conclusion, these findings open insights on a complex and economic impact spatial that will give benefits in terms of future planning, geohazards field such as land subsidence and UUS construction. It serves as assistance to future urban planners, geotechnical engineering in the developed megacity like Shanghai in tackling the underground exploration, space and UUS tunneling-induced land subsidence with the land price.

MIXING PGS AND SPDE FRAMEWORKS IN ORDER TO COKRIGE CONTIUOUS AND CATEGORICAL VARIABLES: A FISHERY APPLICATION

Nicolas Bez (1)* - Thibault Cariou (2) - Didier Renard (3) - Camille Vogel (2) - Laurent Dubroca (2)

Ird (1) - Ifremer (2) - Mines Paristech Université Psl (3)
** Corresponding author: nicolas.bez@ird.fr*

Abstract

In this work, we develop a multivariate model designed to combine continuous and discrete variables in a coherent spatial manner. The context of habitat modelling in ecology in general and in fisheries in particular, often leads to cross one or several continuous variables (typically fish densities) and some continuous variables (e.g. sea temperature) but also some qualitative variables describing the environment (e.g. seabed type/facies). Here we consider two different fish species whose spatial distributions are considered to be linked to two different sets of environmental variables. The application concerns the nursery of two key fish species off the Seine estuary (France). Variables describing the environment are pooled into physical abiotic variables (e.g. seabed type, bathymetry, temperature) and biotic variables (e.g. class of benthic species).

We apply a Pluri-Gaussian framework to transform these two sets of environmental variables into two latent Gaussian fields by conditional Pluri-Gaussian simulations (PGS). A linear model of co-regionalisation is then properly fitted to the data. As PGS makes it possible to generate several latent fields conditionally to the data, we produce a large number of such outcomes. The inspection of the model characteristics allows quantifying the impact of each set of environmental variables on the spatial distribution of each one of the two species. Using the SPDE framework, the cokriging also allows estimating the spatial distributions of the two species by simultaneously taking into account their cross and mutual relationships with the controlling environmental variables and the non-stationary anisotropies shaped by bathymetry.

INFERENCE OF NON-STATIONARY SPDE BASED MODELS

Nicolas Desassis (1) - Mike Pereira (2) - Didier Renard (1) - Xavier Freulon (1) - Thomas Romary (1)* - Denis Allard (3)

Mines Paristech, Geosciences (1) - Chalmers University of Technology, Mathematical Sciences (2) - Inrae, Biosp (3)
** Corresponding author: thomas.romary@mines-paristech.fr*

Abstract

When dealing with spatial data sets, the assumption of stationarity is often ill-adapted. Indeed, in most applications with a large number of observations, a spatial structure that varies across the domain can be highlighted. The Stochastic Partial Differential Equation approach (SPDE) allows to easily model such non stationarities. However, non-stationarity is difficult to estimate outside the Matérn model. In this work, we propose an extension of the SPDE framework to other covariance models through generalized random fields. In the stationary case, it allows to build covariance and precision matrices from the discretization of the differential operator involved in the corresponding SPDE. Even if the resulting matrices are not necessarily sparse, the kriging estimator and conditional simulations can be obtained with efficient algorithms in large scale problems, using polynomial approximations. On the other hand, working on Riemannian manifolds allows to generalize this approach to account for local anisotropies. An accelerated version of the EM algorithm is used to estimate the varying parameters controlling the anisotropies. Two cases are presented. In the first one, the anisotropies only depends on the coordinates. In the second one, the anisotropies are controlled by covariates and the parameters of this relationship are also estimated.

FACIES MODELING USING UNSTRUCTURED GRID, A GROUNDWATER FIELD CASE: THE ROUSSILLON COASTAL AQUIFER

Pierre Biver (1)* - Valentin Dall'alba (2) - Francis Morandini (3) - Philippe Renard (2) - Yvan Caballero (4)

Total Sa, Dg/ep/dso/gis/mms/cmg/gi (1) - Université de Neuchâtel, Chyn (2) - Total Sa, Dg/ep/explo/gts/cig (3) - Brgm (4)

** Corresponding author: pierre.biver@total.com*

Abstract

Unstructured grids are useful because they allow adapting locally the model resolution to ensure solving the physical problem under consideration accurately while being parsimonious in the number of cells and saving computing time. Such grids are frequently used in hydrogeology, but the simulation of petrophysical properties on these grids while accounting for the support effect is rarely done.

In a pair of previous publications, a new geomodelling workflow has been proposed to populate directly unstructured grids with lithologies (facies or rock-types) and petrophysical attributes (namely porosity and permeabilities) (Biver et al., 2019; Mourlanette et al., 2020).

In this paper, we illustrate the effectiveness and applicability of the workflow for facies modeling in an environmental application. The study site is the Roussillon coastal aquifer located in the south of France (Dall'Alba et al., 2020). The groundwater model of this aquifer requires a fine resolution around the pumping wells and along the rivers where active groundwater exchanges are occurring. A coarse resolution is sufficient far away from the wells and from the present coastline where the salinity boundary conditions can be expressed more globally. A dedicated unstructured Voronoï grid has been constructed based on these requirements

Subsequently, the facies model is built using all the available well data. Two methods are presented: the Pluri-Gaussian simulation and an object-based technique. In both cases, azimuth and input proportions trends are used. Facies are simulated at integration points inside the cells. For small cells, a single point is used; for larger cells, larger number of points are simulated and averaged (proportions of each facies and most likely facies are computed). With this procedure, the support size effect is handled, and facies mixing is allowed in large cells. The results demonstrate the applicability and efficiency of the method.

Biver, P., Fuet, S., & Allard, D. (2019, September). Direct Geostatistical Simulation on Unstructured Grids I: Recent Improvements for Additive Variables. In Petroleum Geostatistics 2019 (Vol. 2019, No. 1, pp. 1-5). European Association of Geoscientists & Engineers.

Mourlanette, P., Biver, P., Renard, P., Noetinger, B., Caumon, G., & Perrier, Y. A. (2020). Direct simulation of non-additive properties on unstructured grids. Advances in Water Resources, 143, 103665.

Dall'Alba, V., Renard, P., Straubhaar, J., Issautier, B., Duvail, C., & Caballero, Y. (2020). 3D Multiple-Point Statistics simulations of the Roussillon Continental Pliocene aquifer using DeeSse. Hydrology and Earth System Sciences, 24(10), 4997-5013.

HOW GEOSTATISTICS CAN HELP YOU FIND LEAD AND GALVANIZED SERVICE LINES IN PUBLIC WATER SYSTEMS: A COMPOSITIONAL APPROACH

Pierre Goovaerts (1)*

Biomedware, Inc. (1)
** Corresponding author: goovaerts@biomedware.com*

Abstract

In the aftermath of Flint drinking water crisis, most US cities have been scrambling to locate all lead service lines (LSLs) in their water supply systems. This information, which is most often inaccurate or lacking, is critical to assess compliance with the Lead and Copper Rule and to plan the replacement of lead and galvanized service lines (GSLs) as currently under way in Flint. Lack of accurate records is forcing public water systems and State agencies to rely on expensive survey by licensed plumbers to identify the location of lead and galvanized service lines. One should however expect neighboring houses to have similar types of service line, as they were likely built around the same time period and might have undergone similar upgrades to the water supply pipe network (e.g., replacement of lead service lines).

There have been only a few applications of geostatistics to predicting service line composition. In particular, this author used residual indicator kriging to map within the City of Flint the likelihood that a home has a LSL or GSL based on neighboring field data (i.e., 3254 house inspections available) and secondary information (i.e., construction year and city records). This approach had two limitations: 1) kriging was applied separately to each type of material, which does not guarantee that the prediction is coherent (i.e., probabilities can be negative and do not sum to 1), and 2) Euclidean distance was used as measure of proximity, which does not capture potential non-stationarity in spatial autocorrelation.

In this paper, a compositional approach (i.e., simplicial indicator cokriging) is implemented whereby vectors of hard indicator data (presence/absence of 4 types of SL material: lead, copper, galvanized, others) and soft indicator data (probabilities estimated from secondary data by multinomial regression) were converted into two sets of isometric logratios. Differences between the two sets generated residual vectors which were then interpolated by cokriging. Local trends were added to residual estimates, and results were back-transformed to yield for each tax parcel in Flint a coherent vector of probabilities of occurence of the four types of material.

Spatial non-stationarity was tackled by first defining a measure of proximity combining three main types of metrics (Euclidean distances, spatial dependence structures quantified using non-stationary variogram kernel estimators, proximity in feature and covariate spaces). An Euclidean space was then created through a deformation of the original domain by the application of weighted non-metric multidimensional scaling to the matrix of spatial proximities, and transformed coordinates were derived for all tax parcels using thin-plate spline radial basis functions. Simplicial indicator cokriging was conducted in this new transformed space.

Cross-validation analysis using Receiver Operating Characteristic (ROC) Curves was conducted to quantify the benefit of this approach over residual indicator kriging for different sampling densities.

GENERALIZED VARIOGRAMS OF K-ORDER: APPLICATION TO THE SPATIAL VARIABILITY ANALYSIS OF SATELLITE IMAGES

Roberto Bruno (1)* - Sara Kasmaeeyazdi (1) - Francesco Tinti (1)

University of Bologna, Department of Civil, Chemical, Environmental and Materials Engineering (1)
** Corresponding author: roberto.bruno@unibo.it*

Abstract

The Generalized Variogram of k-order allows a non-parametric structural analysis of an IRF-k, currently when data are available on a grid, which is the case of satellite images. This paper applies the GV structural analysis to several satellite images (Copernicus Sentinel 2) for characterizing the spatial variability of some bands. The spatial variability modeling is very important for exploiting satellite images with reference to the correlation study of the indirect information available (the bands) with the actual ReV of interest (grades, …) and to specific problems solution (e.g. selection). Several images from one case study are presented, and the superiority of IRF-k direct analysis for non-stationary regionalization modeling is stressed, both over the current parametric analysis and the classical dichotomic approach "drift + residuals". Many insights are shown, as the recognition and modelling of spatial anisotropies, the inclusion of Generalized Covariances models with sill, and an efficient approach for identifying the k-order. The simplicity of operations is stressed.

NONSTATIONARY NEAREST NEIGHBOR GAUSSIAN PROCESS: HIERARCHICAL MODEL ARCHITECTURE AND MCMC SAMPLING

Sébastien Coube-Sisqueille (1)* - Sudipto Banerjee (2) - Benoît Liquet (3)

Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques et de leurs Applications (1) - University of California, Los Angeles, Biostatistics (2) - Université de Pau et des Pays de l'Adour, Laboratoires de Mathématiques et de leurs Applications (3)

** Corresponding author: sebastien.coube@univ-pau.fr*

Abstract

Hierarchical Bayesian space-time models have proven to be invaluable tools to study variables that are observed through point measurements. While stationary Gaussian Process models can capture quite a large variety of behaviors, nonstationary models are an attractive extension. In relevant cases, a nonstationary model should improve smoothing and prediction and give richer, more informative covariance structures, while a stationary model should underfit.

However, this exciting approach is often hampered by several problems that stem from the complexity of nonstationary models. The first problem is their lack of scalability. Another problem is finding a model that will be complicated enough for the data but simple enough for the user: in the case of a large model with many hierarchical layers, the interpretability of the parameters is critical. Eventually, the high number of potential covariance structures makes it difficult to choose a model and may cause identification problems. Given those problems, we propose a nonstationary Nearest Neighbor Gaussian Process model with an original hierarchical architecture and ad hoc MCMC algorithms.

This model extends the recent development of Nearest Neighbor Gaussian Processes (NNGP). The first aspect of our work is to precise the properties of NNGP with nonstationary covariance from Paciorek (2003).

We embed this nonstationary NNGP in an interpretable hierarchical architecture stemming from the time series model of Heinonen et al. (2016). Our main contribution is to go spatial and define a frugal and consistent prior for the elliptic covariance parameters of Paciorek (2003). We also use NNGP to make the architecture affordable for large spatial data sets.

A large part of our effort is to propose scalable MCMC to sample nonstationary covariance parameters. We try two approaches. The first is a Metropolis Adjusted Langevin Algorithm inspired from the Hamiltonian Monte Carlo of Heinonen et al. (2016). We make several contributions from this basis. In the case of NNGP, usual matrix differentiation formulas cannot be used. Finding the gradients is tedious in the theoretical side, but ends up being computationally efficient. We also extend their response model to full data augmentation, more accurate and allowing non-Gaussian responses, using the interweaving of Yu and Meng (2011) for efficient implementation. The second approach is chromatic sampling, permitted by our model architecture. The rest of our architecture differs little from Coube and Liquet (2020).

Our model's properties are tested on synthetic datasets and we analyze lead contamination measurements in mainland US.

Coube, S. and B. Liquet (2020). Improving performances of mcmc for nearest neighbor gaussian process models with full data augmentation. arXiv preprint arXiv:2010.00896.

Heinonen, M., H. Mannerström, J. Rousu, S. Kaski, and H. Lähdesmäki (2016). Nonstationary gaussian process regression with hamiltonian monte carlo. In Artificial Intelligence and Statistics, pp. 732–740.

Paciorek, C. J. (2003). Nonstationary Gaussian processes for regression and spatial modelling. Ph. D. thesis.

Yu, Y. and X.-L. Meng (2011). To center or not to center: That is not the question—an ancillarity–sufficiency interweaving strategy (asis) for boosting mcmc efficiency. Journal of Computational and Graphical Statistics 20 (3), 531–570.

A CHANGE OF SUPPORT MODEL OPTIMIZATION FOR ENVIRONMENTAL MONITORING

Serena Berretta (1)* - Daniela Cabiddu (2) - Simone Pittaluga (2) - Michela Mortara (2) - Marino Vetuschi Zuccolini (3)

University of Genoa, Department of Mathematics (1) CNR-Imati (2) - University of Genoa, Department for the Earth, Environment and Life Sciences (DiSTAV) (3)
** Corresponding author: serena.berretta@ge.imati.cnr.it*

Abstract

Geometric representations of spatial domains in environmental applications are used to structure, access and render surveyed data; the survey domain is often represented as a discrete regular grid with a fixed spatial resolution, because of its conceptual and implementation simplicity. Nevertheless, the region of interest (ROI) can be delimited by physically-defined criteria leading to boundaries with complex geometries. Such constraints force a representation of free-form domains by means of regular grids with very high-resolution models made of millions or even billions of small elements, requiring high-performance computers to be handled efficiently. Unstructured grids are more suitable to represent free-form domains and flexible to faithfully represent complex geometries.

Moreover, samples used to fill the ROI with interpolation algorithms are often collected with a point support, $z(u_i), i = 1, \dots, N$, while the estimates are instead required over volumes, $Z(v_k), k = 1, \dots, N_b$. **Change-of-support** techniques for additive variables have been previously published in the literature to fill the gap and make unstructured grids suitable for environmental modelling (Chiles and Delfiner, 2009). Among the existing change-of-support techniques, the Discrete Gaussian Model (DGM) is one of the most used techniques for environmental applications where the discrete units are represented by irregular shapes.

Traditional approach of DGM define a coefficient, r, as the correlation between a point value and a volumetric one, so that we can switch from one support to another, without a bias. This coefficient is related to the block-to-block covariance and therefore in a multi-support context of volumes this coefficient will be different from one volume to another, $r_k,\ k = 1, \dots, N_b$. The greater the size of a volume v_k the smaller the coefficient r_k (it will tend to zero); on the other hand, when the volume will tend to the point size, the coefficient will tend to one (maximum correlation). The use of different discretization schemes of quasi-random points for the calculation of block-to-block covariance can affect the results. We implemented an extension of Sobol' sequences to 3D unstructured problems to compute the change-of-support coefficients more efficiently.

We implement the aforementioned theoretical aspects in a specific code applied to a simulation framework that corresponds to an adaptive real-time environmental monitoring for the evaluation of geochemistry in marine waters. We represent the harbor water body by an unstructured grid to describe the highly irregular geometry at finer level: smaller grid cells are generated near the coast to better capture the complex geometries of the seashore; larger ones elsewhere. The first results demonstrate that taking into account the support dependency, enables the optimization in terms of accuracy and computational cost the analysis when the support of available data is different by the supports of the unstructured grid for which statements are required for the estimation. Subsequently, this method will be validated on the harbor waters model.

Keywords: SRF, Support, Simulations

1. Introduction

Environmental monitoring is crucial for investigating the condition of soil, air and water, describing natural phenomena, and promptly reacting to, and even preventing, accidents. In standard environmental monitoring surveys, the number of samples is fixed and their locations are predetermined either on a regular spacing basis or are defined by a-priori knowledge. Once collected, samples are subject to laboratory analysis to provide accurate measurements at point locations, then used to generate geostatistical maps representing an estimation of the continuous distribution of the environmental variables over the domain. The analysis of the distribution might then call for further sampling (e.g., on a portion of the global domain) to reach the desired reliability. This methodology requires very long times and high costs. Nowadays, thanks to the technological improvements, times and costs can be significantly reduced. Indeed, new dynamic positioning systems, lighter and accurate sensors are now available. Having a real-time point support measurement of an environmental variable opens the door to new on-the-fly sampling schema design. However, this requires innovative computational solutions to make spatial data analysis precise and fast.

In Berretta et al. (2018) an adaptive sampling strategy is proposed capable of monitoring inner waters in harbor. This approach exploits geostatistics tools in order to determine in real time the next best location to be sampled. In short, after initializing the system with few randomized sampling points, an iterative routine predicts the variable distribution from the data sampled so far and suggests the next sample to be acquired in order to optimize the data quality by taking advantage of the uncertainty of the estimates. At every iteration, a new waypoint is acquired, and the variable distribution map is refined, along with the uncertainty map related to that distribution. In this work a 2D squared survey area was used for testing the procedure, but in several real cases, the region of interest can be a 3D domain delimited by physically-defined constraints leading to boundaries with very complex geometries. Especially in a geological framework, structured grids lack the ability to model complex reservoir geometries such as external boundaries and internal body heterogeneities. Unstructured grids are composed of several blocks of various shapes and sizes and they are more suitable to represent free-form domains and allow multi-resolution representations: it is becoming more frequent to have a great variety of grid cells of different size and shape to enable fine-scale modelling close to some important locations and coarse-scale modeling in less important regions. Although the measurements are usually made on a point support, one is often interested in volumetric supports (blocks), thus, the support of the collected data is smaller than the support for which estimates are required. Change of support models are necessary to handle such geostatistical applications.

In a preliminary phase, the geometric representation of the domain was supposed to be a structured grid, where cells are equi-volumetric parallelepipeds as shown in Figure 1.

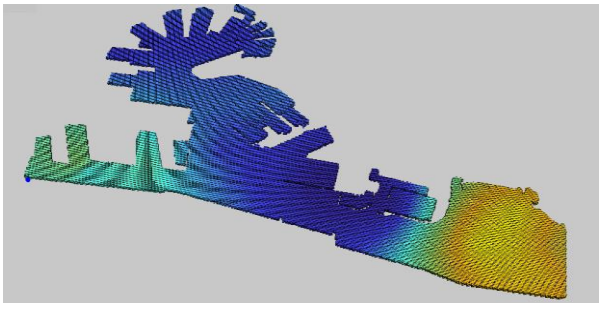


Figure 1 – Geometric model with structured grid (voxels of equal size) where the distribution of a synthetic variable is simulated.

In this work, we generalize the previous approach to support volumetric domains represented as unstructured grids (composed by tetrahedra). Our method keeps into account the volume difference among tetrahedra and

exploits the change of support models to achieve the goal. Since our reference application is intended to be real time, we optimize an existing change of support model to guarantee computational speed. Preliminary results show both efficiency and accuracy of our method.

2. Related Works

2.1. Geostatistics on Unstructured Grids

An unstructured grid, D, is defined as a grid composed by a finite number N_b of no-overlapping blocks:

$$D = \bigcup_{p=1}^{N_b} v_p \ , \qquad v_i \cap v_j = \emptyset, \ \ \forall i \neq j$$

where the volume of v_i, $|v_i|$, is in general different from that of v_j, $|v_j|$, for all $i \neq j$.

If the domain in which the estimates of some environmental variables are required is represented with an unstructured grid, traditional geostatistics methods for the estimation of a variable distribution have not a straightforward application.

The straightforward approach for simulating on unstructured grids is performed by using an auxiliary structured fine-scale grid, the geostatistical simulation methods can be directly applied. The result is then upscaled to the target unstructured grid (Chiles and Delfiner, 2009). This approach has the advantage of avoiding assumptions about i) the change of support law for the random field $Z(u)$, and ii) the spatial distribution of the block average values $Z(v)$. On the other hand, the creation and storage of an auxiliary fine-scale grid increases the number of locations where the random field should be simulated, at high costs in terms of memory usage and computational time. In addition, the auxiliary grid resolution (i.e., the size of blocks) must be decided, as a compromise between accuracy and cost (but at least as small as the smallest block in the target unstructured grid).

Being able to directly use unstructured grids for simulation processes would optimize the pipeline, both in terms of computational time and memory usage. Zaytsev (2016) introduced the change of support models for unstructured grids in a simulation framework. We apply such a model in a real-time environmental adaptive sampling pipeline where the domain has a very complex shape represented by an unstructured grid. In a real-time analysis framework, execution time is a key feature and our proposal speeds up the procedure.

2.2. Global Discrete Gaussian Model

A change of support model is needed when the volumetric support of the data is smaller than the support for which the estimation is required. The most used change of support model in geostatistical analysis is the Discrete Gaussian Model (DGM) (Chiles and Delfiner, 2009).

Consider a stationary random field (SRF) $Z(u)$ that can be expressed as the transformation of an SRF $Y(u)$ with standard Normal marginal distribution, $Z(u) = \varphi(Y(u))$. Consider a block v and a uniform random point x within v, the random variable can be expressed as $Z(x) = \varphi(Y(x))$. Similarly, the mean grade $Z(v)$ of the block v is of the form $Z_v = \varphi_v(Y_v)$, where Y_v is a standard Normal and φ_v the block transformation function to determine. The crucial assumption of the DGM is that the bivariate distribution of the $(Y(x), Y_v)$ pair is Gaussian with a correlation coefficient r assumed positive. In order to determine r, two options are available: the first one (referred to as DGM-1), proposed by Matheron (1963), is based on the fact that the variance defined by the block distribution must be consistent with the block variance σ_v^2 deriving from the covariance $C(h)$ of the SRF $Z(u)$. The second option (DGM-2) was proposed by Emery (2007a) where he investigates the properties of change of support models and provides a simplified method for deriving the change of support coefficient and covariance between the transformed variables, but in a context of structured grids.

In the next section both the DGM-1 and its DGM-2 extension to unstructured grids are described as in Zaytsev (2016).

2.3. DGM for Unstructured Grids

Consider a domain D described by an unstructured grid. Consider $v_i \in D$ and let x a uniform random point within v_i. The mean grade $Z(v_i)$ of the block v_i is of the form: $Z_{v_i} = \varphi_{v_i}(Y_{v_i})$; where Y_{v_i} is a standard Normal variable and φ_{v_i} is the block transformation function of v_i. For each element of the unstructured grid the block transformation function must be determined. The crucial assumption of the DGM becomes that the bivariate distribution of the $\left(Y_{v_i}, Y(x)\right)$ pair is Gaussian with a correlation coefficient r_i. Moreover, Cartier's relation (Chiles & Delfiner (2009)) becomes: $E[Y_{v_i}] = \varphi_{v_i}(Y_{v_i})$.

The calculation of the block transformation functions, $\{\varphi_{v_i}(Y_{v_i}), i = 1, \dots, N\_b\}$, will be carried out using Hermite polynomials: for each block v_i,

$$\varphi_{v_i}(y) = \sum_{p=0}^{\infty} \varphi_{v_i p} H_p(y)$$

The coefficients $\varphi_{v_i p}$ are to be determined. Using the property of Hermite polynomials and the Cartier's relation we found $\varphi_{v_i p} = \varphi_p \, r_i^p$. In conclusion, the result is the following:

$$\varphi_{v_i}(y) = \sum_{p=0}^{\infty} \varphi_p \, r_i^p H_p(y) \qquad (1)$$

The correlation coefficient r_i is selected for each block v_i so that the distribution defined by $\varphi_{v_i}(.)$ has the variance given by $\sigma_{v_i}^2 = \frac{1}{|v_i|^2} \int_{v_i} \int_{v_i} C(x - x') dx dx'$. The variance of $\varphi_{v_i}(Y_{v_i})$, taken as a function of r_i, is:

$$Var[\varphi_{v_i}(Y_{v_i})] = \sum_{p=0}^{\infty} (\varphi_p \, r_i^p)^2$$

and r_i is the solution of

$$\sum_{p=0}^{\infty} (\varphi_p \, r_i^p)^2 = \sigma_{v_i}^2 = \frac{1}{|v_i|^2} \int_{v_i} \int_{v_i} C(x - x') dx dx' \qquad (2)$$

In practice, to solve the polynomial equation on r_i, we use the L-BFGS optimization algorithm in the family of quasi-Newton methods (Matthies, 1979).

DGM-2 is a generalization of the DGM-1. At the cost of a further, more restrictive, assumption, it provides a simpler approach for computing the change of support coefficients. The additional assumption becomes: "for any block v_i and two independent randomized locations ◎ x and x' within v_i, the bivariate distribution of $Y(x)$ and $Y(x')$ is Gaussian". From this assumption, it can be derived the following relation between Y_{v_i} and $Y(v_i) = \frac{1}{v_i} \int_{v_i} Y(x) dx$ for every block (Chiles and Delfiner, 2009)

$$Y(v_i) = r_i Y_{v_i}$$

which provides a simple formula for computing the change-of-support coefficient r_i. In that case, r_i^2 is the block variance of the SRF $Y(v_i)$:

$$r_i^2 = \frac{1}{|v_i|^2} \int_{v_i} \int_{v_i} \rho(x - x') dx dx' \qquad (3)$$

Where $\rho(h)$ is the covariance of the SRF $Y(u)$.

2.4. Change of support coefficients

One of the crucial problems when performing geostatistics on unstructured grids is the problem of computing the block-to-block covariance. Let be v_i and v_j two blocks with volume $|v_i|$ and $|v_j|$ respectively. The covariance between these two blocks is defined as

$$C(v_i, v_j) = \frac{1}{|v_i||v_j|}\int_{v_i}\int_{v_j} C(x - x')dxdx'$$

The computation of the block-to-block covariance is essential for determining the change of support coefficients for each volume of the domain in DGMs. Computation of multidimensional integrals can be effectively performed with several methods including Monte Carlo integration techniques, where the covariance $C(v_i, v_j)$ can be estimated with a number N of pairs of points $(X_k, X'_k) \in v_i \times v_j, k = 1, \dots, N$ sampled from the uniform distribution within blocks:

$$\hat{C}(v_i, v_j) = \frac{1}{N}\sum_{k=1}^{N} C(X_k, X'_k)$$

How to select these pairs of points could be relevant for the computation of block-to-block covariance. In numerical analysis, the quasi-Monte Carlo method is for numerical integration and solving some other problems using low-discrepancy sequences (also called quasi-random sequences). The difference between quasi-Monte Carlo and Monte Carlo method is the way the points on the domain $D \subset R^d$ are chosen. Quasi-Monte Carlo uses a low-discrepancy sequence, whereas Monte Carlo uses a pseudo-random sequence. The main advantage of using low-discrepancy sequences is a faster rate of convergence. The error of the approximation by the quasi-Monte Carlo method is $O(\frac{(log log\, N)^d}{N})$, whereas the Monte Carlo method has a probabilistic error of $O(\frac{1}{\sqrt{N}})$. Hence, the quasi-random sequence reaches the convergence value faster than the pseudo-random sequence.

Several methods could be used to generate quasi-random numbers. In our approach Sobol' sequences are preferred (Sobol', 1967; Antonov and Saleev, 1979), but we have extended them to 3D unstructured grid problems. Indeed, usually Sobol' algorithm generates points in a unitary cube $[0,1]^3$, but in case of using unstructured grid composed by tetrahedra an extension is needed to generate the sequence of points directly inside the tetrahedron avoiding the waste of generating a point in the cube and then to check if the point is inside of the tetrahedron (and if not, discard it or throw it away). Another open problem is the number N of points to generate. This should be a compromise between the quality of the estimate of $\hat{C}(v_i, v_j)$ and the computational time since the unstructured grid could be composed of a very large number of tetrahedra.

Once block-to-block variances computation is completed, the coefficients for the change of support are derived using either the DGM-1 or the DGM-2 (*Equation (2), Equation (3)*).

3. Test and Results

To test the performance of the change of support implementation and its contribution to the quality of the estimation map, we designed the following experiment. We generate a cubic domain 100x100x100 meters discretized into several tetrahedral cells of different size. This unstructured grid is referred as the coarse model (CM) and represents a coarse-scale representation made by $K = 75$ elements. Starting from CM, a fine-scale grid (FM = fine model) has been generated by applying several splitting operations (volume, face and edge splits on their middle points) subdividing each tetrahedron into several smaller tetrahedra. The fine-scale grid (FM = fine model) has $M = 172800$ elements ($K << M$). Note that a mapping between the fine and the coarse

scale model is defined by grid construction: for each tetrahedron in CM the set of tetrahedra in FM belonging to it is known and determined by a finite and integer number of units.

The FM is used to represent a synthetic reality, used as a reference by the sampling procedure. Thus, we associate to FM a synthetic field simulated using an isotropic exponential covariance with range of 45 and nugget of 0.05 (Figure 2, left). Such a scalar field represents a chemical-physical parameter (i.e. dissolved oxygen). Now, in order to assign a value to a tetrahedron of the CM we average all the synthetic values of tetrahedra of the FM that compose it and we obtain an averaged synthetic field associated with the coarse geometric model (Figure 2, right) to use in estimation procedure and as a reference for computations.

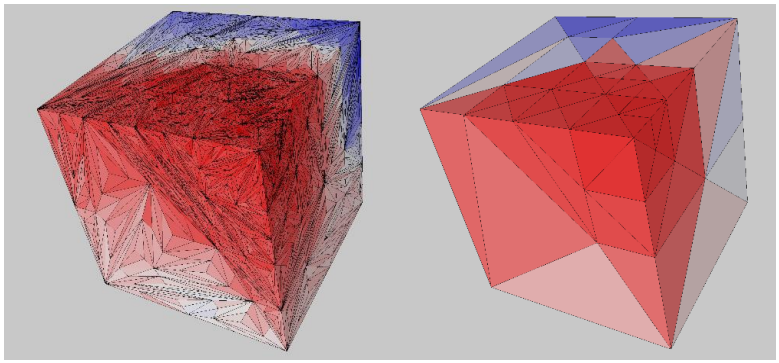


Figure 2 – Cubic geometric model with unstructured grid (tetrahedra) of a cube 100x100x100 meters. On the left: the synthetic reality on the fine-scale unstructured grid; on the right: the averaged synthetic reality on the coarse-scale unstructured grid

To test the performance of the change of support model, we select randomly 25 positions inside the cube and the corresponding values are sampled on the synthetic field in FM. This set of data allows, using Sequential Gaussian Simulations, to provide the estimation map on the CM. The result is then back-transformed either using the change of support model (Figure 4, left) or without considering it (i.e. using only normal score transformation) (Figure 4, right).

The values of change of support coefficients computed with DGM-2 are shown in Figure 3 with respect to the volumes of tetrahedra. The smaller the volume of a tetrahedron, the more the coefficient associated to it tends to 1. In some cases, a very stretched shape of the tetrahedra could give values that are not in accordance with the decreasing trend.

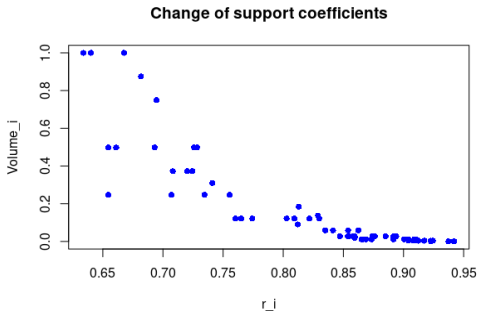


Figure 3 – Plot of change of support coefficients versus normalized volumes of tetrahedra.

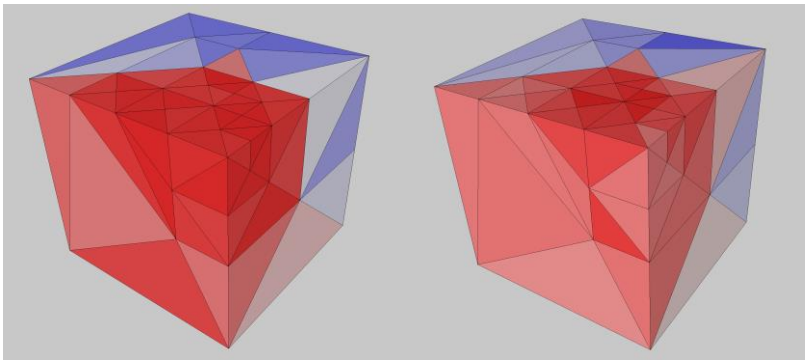


Figure 4 – Estimation map from SGS. On the left: using change of support model in the pipeline; on the right: without considering change of support model in the pipeline.

Once obtained the two maps, it is possible to check which estimates are closer to the averaged synthetic field (Figure 2, right). Mean square error (MSE) is used to verify the improvement in performance. In this case, the estimation using the change of support model has better results (MSE = 28.56) with respect to the estimation without considering it (MSE = 39.75).

4. Application: Adaptive sampling on Unstructured Grids

The use of unstructured grids is crucial in environmental monitoring since the complex geometry of the survey domain. For this, the theoretical aspects described in the previous sections are implemented within a simulation framework that plays an adaptive real-time sampling survey for the evaluation of general geochemistry of the waters in Genoa harbor monitoring, for example the dissolved oxygen. The digital geometric representation of the survey domain is built as a volumetric model of tetrahedra with different volumes in order to achieve a finer detail where needed, e.g., in the narrow regions between close peer structures, to fit the boundaries more accurately, and represent less important areas at a coarser resolution. In Figure 5 is shown the geometric model of the survey domain, that is a section of harbor of Genoa.

To start the procedure an initial set of random samples is necessary for the preliminary estimates. Then the iterative procedure begins. At each iteration, the method selects the best next point to be sampled and updates the known information with new acquired data. The optimization criterion to select new samples is based on the uncertainty of the estimates derived from the simulation method (in this implementation Sequential Gaussian Simulations produce an uncertainty map), following the idea that more samples are necessary to improve the results in areas where uncertainty about estimates is higher.

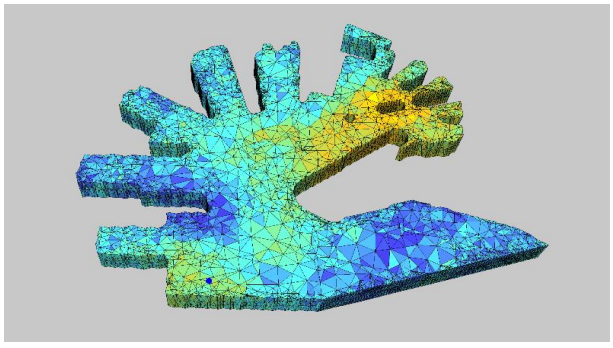


Figure 5 – Geometric model with unstructured grid (tetrahedra) where a synthetic variable is simulated.

Once the estimations at the centers of gravity of each cell are obtained by the simulation process, they are back-transformed using the block transformation functions (Equation (1)) and using the change of support coefficients derived as in Equation (3).

The procedure iterates by updating the prediction and uncertainty maps, until the uncertainty goes below a desired threshold, selected by the expert. Figure 6 (left) shows the estimation map after 7 iterations (7 points plus 8 randomly chosen points for the initialization). In order to evaluate the adaptive sampling approach, we compare the estimation map produced with the same number of samples (15), but all measured at random positions (Figure 6, right). The results highlight that the estimation map given by the adaptive procedure is closer to the synthetic field (Figure 5) than the one derived by random sampling.

5. Discussion and Conclusions

In environmental monitoring often the survey domain to evaluate is very complex and a simple regular grid is not able to well fit its boundaries. In these cases, unstructured grids with tetrahedra are more suitable, but their use in geostatistics analysis must be managed with the change of support model extended to unstructured grids. In this work, the interplay of the estimating geostatistical techniques and change of support models on unstructured grids has been applied in an environmental monitoring procedure, where a real-time adaptive sampling strategy is applied. Furthermore, the computation of the block-to-block covariance was made faster by the generation of Sobol' sequence directly on tetrahedron. This improvement can be crucial in a real-time sampling strategy where the computational time is a challenging issue.

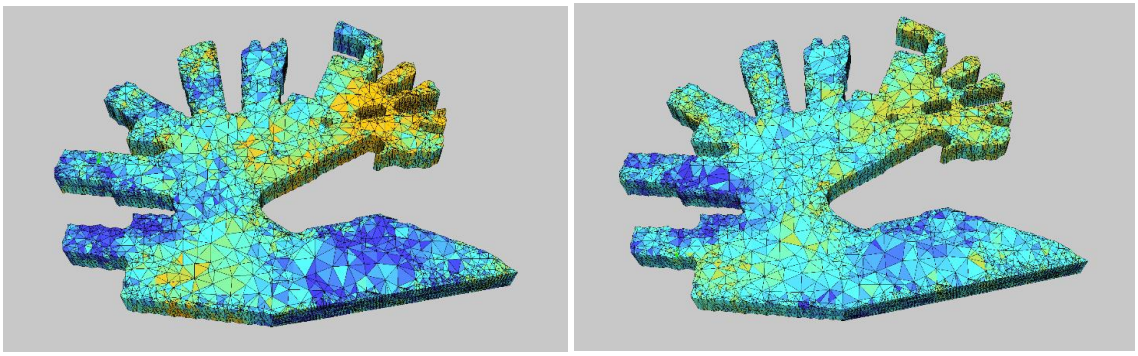


Figure 6 – Estimation map. On the left: using 15 sampled points (8 initial random points + 7 adaptive points); on the right: using 15 random points. Both using change of support model in the estimation process.

In order to check the effectiveness of change of support model for unstructured grids, a simulation experiment on a simpler test domain has been designed: especially when the size of the volumes among tetrahedra is very heterogeneous, the change of support model affects the performance and improve the estimation results.

In future work our environmental monitoring strategy for unstructured grids will be tested in the real setting of the harbor of Genoa focusing on water quality monitoring.

References

Antonov, I. A., & Saleev, V. M. (1979). An economic method of computing LPτ-sequences. USSR Computational Mathematics and Mathematical Physics, 19(1), 252-256.

Berretta, S., Cabiddu, D., Pittaluga, S., Mortara, M., Spagnuolo, M., & Zuccolini, M. V. (2018). Adaptive environmental sampling: The interplay between geostatistics and geometry. In Smart Tools and Apps for Graphics-Eurographics Italian Chapter Conference. The Eurographics Association.

Chiles, J. P., & Delfiner, P. (2009). Geostatistics: modeling spatial uncertainty (Vol. 497). John Wiley & Sons.

Emery, X. (2007). On some consistency conditions for geostatistical change-of-support models. Mathematical geology, 39(2), 205-223.

Matheron, G. (1963). Principles of geostatistics. Economic geology, 58(8), 1246-1266.

Matthies, H., & Strang, G. (1979). The solution of nonlinear finite element equations. International journal for numerical methods in engineering, 14(11), 1613-1626.

Sobol', I. Y. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 7(4), 784-802.

Rivoirard, J. (1994). Introduction to disjunctive kriging and non-linear geostatistics (No. 551.021 R626i). Clarendon Press.

Zaytsev, V., Biver, P., Wackernagel, H., & Allard, D. (2016). Change-of-support models on irregular grids for geostatistical simulation. Mathematical Geosciences, 48(4), 353-369.

ASSESSING LOCAL AND SPATIAL UNCERTAINTY WITH NON-PARAMETRIC GEOSTATISTICS

Stephanie Thiesen (1)* - Uwe Ehret (1)

Karlsruher Institut für Technologie, Institute for Water and River Basin Management - Hydrology (1)
** Corresponding author: stephanie.thiesen@gmail.com*

Abstract

Uncertainty quantification is an important topic for many environmental studies. Our previously proposed Histogram via entropy reduction (HER, Thiesen et al., 2020) combines statistical learning and Information Theory in a geostatistical framework for overcoming parameterization with functions and uncertainty trade-offs present in many traditional methods. It has been shown that, by construction, the method provides a proper framework for uncertainty estimation which accounts for both spatial configuration and data values, while allowing one to introduce or infer properties of the field through the aggregation method. In this study, we explore HER method in the light of uncertainty analysis. In general, uncertainty at any particular unsampled location (local uncertainty) is frequently assessed by nonlinear interpolators such as indicator and multi-gaussian kriging. HER has shown to be a unique approach for dealing with uncertainty estimation in a fine resolution without the need of modeling multiple indicator semivariograms, order-relation violations, interpolation/extrapolation of conditional cumulative distribution functions, or stronger hypotheses of data distribution. In this work, this nonparametric geostatistical framework is adapted to address local and spatial uncertainty in the context of risk mapping. We investigate HER for handling estimations of threshold-exceeding probabilities to map the risk of soil contamination by lead in the well-known dataset of the region of Swiss Jura. Finally, HER method is extended to assess spatial uncertainty through sequential simulation. Its results are compared to indicator kriging and benchmark models available in the literature generated for this particular dataset.

Thiesen S, Vieira DM, Mälicke M, Loritz R, Wellmann JF, Ehret U (2020) Histogram via entropy reduction (HER): an information-theoretic alternative for geostatistics. Hydrol Earth Syst Sci 24:4523–4540. https://doi.org/https://doi.org/10.5194/hess-24-4523-2020

MODELING OF DHS SURVEY DATA AT SUB-NATIONAL ADMINISTRATIVE LEVEL 2

Benjamin K Mayala (1)* - Trinadh Dontamsetti (1) - Tom Fish (1) - Trevor Croft (1)

Icf International, Dhs (1)
** Corresponding author: bmayala2@gmail.com*

Abstract

Over the last several years and within the framework of the Sustainable Development Goals, there has been a need to improve the measurement and understanding of local geographic patterns to support more decentralized decision-making and more efficient program implementation. This requires more disaggregated data that are not currently available in a nationally representative household survey. The spatial modeling techniques that leverage existing survey data, spatial relationships between survey clusters, and relationships with geospatial covariates have become increasingly popular for mapping key development indicators at high spatial resolution. This study explores the potential of model-based geostatistics methodology to model DHS survey indicators. We implement a stacked ensemble modeling approach that combines multiple model algorithmic methods to increase predictive validity relative to a single modeling. The approach captures potentially complex interactions and non-linear effects among the geospatial covariates. Three submodels are fitted to six DHS indicator survey data using the geospatial covariates as exploratory predictors. The model prediction surfaces generated from the submodels are used as covariates in the final Bayesian geostatistical model, which is implemented through a stochastic partial differential equation approach in the integrated nested Laplace approximations. To explore the ability of our modeling approach to estimate indicators below the first subnational level, pixel-level estimates generated from the Bayesian model were aggregated to the second subnational level by using the population-weighted average within the administrative boundary. Results of the individual submodels vary spatially, which is explained by the uncertainties in the individual model algorithm. The use of an ensemble model approach seems more adequate than relying on predictions from any single modeling method. We demonstrate the predictive ability of the model at the second administrative level using cross-validation. The results indicate good predictive performance. The proposed approach can help to inform the allocation of resources and program implementation in areas that need more attention. Countries can use this approach to model other DHS survey indicators at much smaller spatial scales.

EXPLORING THE EFFECTS OF ENVIRONMENTAL TOXINS FROM AIR POLLUTION ON CHRONIC KIDNEY DISEASE

Jennifer Mckinley (1)* - Ute Mueller (2) - Peter Atkinson (3) - Ulrich Ofterdinger (1) - Siobhan Cox (1) - Rory Doherty (1) - Damian Fogarty (4) - Juan Jose Egozcue (5) - Vera Pawlowsky-Glahn (6)

Queen's University Belfast, School of Natural and Built Environment (1) - Edith Cowan University, School of Science (2) - Lancaster University, Lancaster Environment Centre (3) - Belfast Health Trust (4) - U. Politécnica de Cataluña (UPC), Dept. Civil and Environmental Engineering (5) - University of Girona, Dep. Computer Sciences, Applied Mathematics, and Statistics (6)

** Corresponding author: j.mckinley@qub.ac.uk*

Abstract

Recent reviews of the impact of air pollution on human health have shown scientific evidence for the detrimental effects of air pollutants, including environmental toxins which may become blood-borne and translocate to tissues such as the liver, brain and kidney. Atmospheric pollution deposition from traffic and brake wear emissions have been discovered to be important potential sources of toxic metals including arsenic (As), cadmium (Cd), iron (Fe), molybdenum (Mo), lead (Pb), tin (Sn), antimony (Sb), Uranium (U) and Zinc (Zn). Chronic kidney disease (CKD), a collective term for many causes of progressive renal failure, is increasing worldwide due to ageing and a general increase in obesity and diabetes. CKD attributed to unknown aetiology (termed CKDu) is an increasing issue globally with the occurrence of geographic clusters appearing to suggest potential underlying environmental causes of CKDu. This study uses data from the UK Renal Registry including Chronic Kidney Disease of uncertain aetiology (CKDu) to investigate the impact of environmental toxins including air pollution data on human health. Using an urban soil geochemistry database of total element concentrations of potentially toxic elements (PTEs), we examine the spatial statistical relationship between Standardised Incidence Rates (SIRs) of CKDu with environmental toxins and air pollution data. A compositional data analysis approach is used with the use of balances (a special class of log contrasts) to find an elemental balance associated with CKD and CKDu. Using a compositional data analysis approach, informed by the selected PTEs and air pollution balance approach, regression analysis (using glm with log link) reveal a statistically significant correlation between CKD for all SIRs for the age group >16 years and the identified balance of Mo/Zn (significance level of 0.001) and the MDM domains of employment, income and health (significance levels of 0.001, 0.01 and 0.05 respectively). Results from the compositional balance approach indicate an association with the air pollutants SO_2, CO, Benzene, PM10 and PM2.5. However, the relationship between CKD for all SIRs >16 years and these air pollutants was not found to be statistically significant. The findings from this work allow a greater understanding of the link between human health and environmental toxins from anthropogenic sources including air pollution.

Keywords: anthropogenic toxins, toxic metals, atmospheric pollution, compositional data analysis, progressive renal disease

1. Introduction

Studies have shown that chronic exposure to ambient fine particulate matter (PM2.5) is a risk factor for cardiovascular-related morbidity and mortality (Brook et al. 2010). Further work has hypothesised that renal function impairment may be a mediating factor of the cardiovascular effects of long-term PM2.5 exposure (Lue et al., 2013). Experimental evidence on the effects of particle exposure on the kidney is limited (Thomson et al., 2013; Nemmar et al., 2010) but several studies support the premise that atmospheric pollution deposition including traffic pollution (Lue et al., 2013) and long-term PM2.5 exposure negatively affects renal function (Mehta et al., 2016). The detrimental effects of traffic pollution (Carrero et al., 2013; Afsar et al., 2019) and air pollutants, including toxic metals (cadmium (Cd), lead (Pb), uranium (U), arsenic (As), molybdenum (Mo), tin (Sn) and antimony (Sb) have all been linked to potential kidney damage. Brake lining and brake wear emissions have also been shown to be potentially important sources of iron (Fe), copper (Cu), Zinc (Zn), Pb, Sb and Mo (review by Grigoratos and Martini, 2015). The underlying mechanisms for kidney decline are not fully known but studies have shown that ultrafine particles of environmental toxins may become blood-borne and translocate to other tissues such as the liver, brain and kidney (Geiser and Kreyling, 1999; Oberdörster et al., 2005). Soils can be used as tracers for these environmental toxins through identifying both anthropogenic and geogenic signatures. Chronic kidney disease (CKD), a collective term for many causes of progressive renal failure, is associated with a natural decline in renal function over time, with a more rapid decline and resultant impact on life expectancy for individuals who have end-stage kidney disease (ESKD) (Lindeman et al. 1985; Musso and Oreopoulos, 2011). CKD attributed to unknown aetiology (termed CKDu) is an increasing issue globally with the occurrence of geographic clusters appearing to suggest potential underlying environmental causes of CKDu. CKDu is a global issue and while the underlying causes have been linked to environmental factors they are not well understood. Previous research highlights the need for further research into the relationship between the impact of environmental toxins (McIlwaine et al., 2017) and atmospheric pollution on human health (McKinley et al., 2020a; 2020b). Using data from the UK Renal Registry (UKRR) and an urban soil geochemistry database of total element concentrations of potentially toxic elements (PTEs), we examine the statistical relationship between Standardised Incidence Rates (SIRs) of CKD and CKDu with environmental toxins, air pollution data and social deprivation.

2. Material and Methods

Chronic Kidney Disease data

The UKRR collects data on all patients with advanced CKD on dialysis or with a kidney transplant (RRT) across the UK and reports data by age group on primary renal disease (UKRR 2019). For this research the UKRR provided SIRs for patients starting RRT between 2006 and 2016, by Super Output Areas (SOA) which are the smallest administrative wards in Northern Ireland (NI). Data were provided in age brackets, 16-39, 40-64 and 65+, all ages >16 and for uncertain aetiology (CKDu) between 2006 and 2016. A SIR for a SOA is a measure that quantifies the relationship between actual incidence in the SOA and the expected incidence based on that of Northern Ireland as a whole. SIRs of exactly 1 indicate that a SOA's incidence for RRT is equal to that expected based on Northern Ireland's average age-specific incidence rates. SIRs above 1 indicate that the incidence of RRT for a SOA is greater than expected based on the NI's average age-specific incidence rates. For this study, we focus on the Belfast urban area, the capital of NI, UK and the CKD SIR data between 2006 and 2016 for all ages >16 and for uncertain aetiology (CKDu) (Fig. 1).

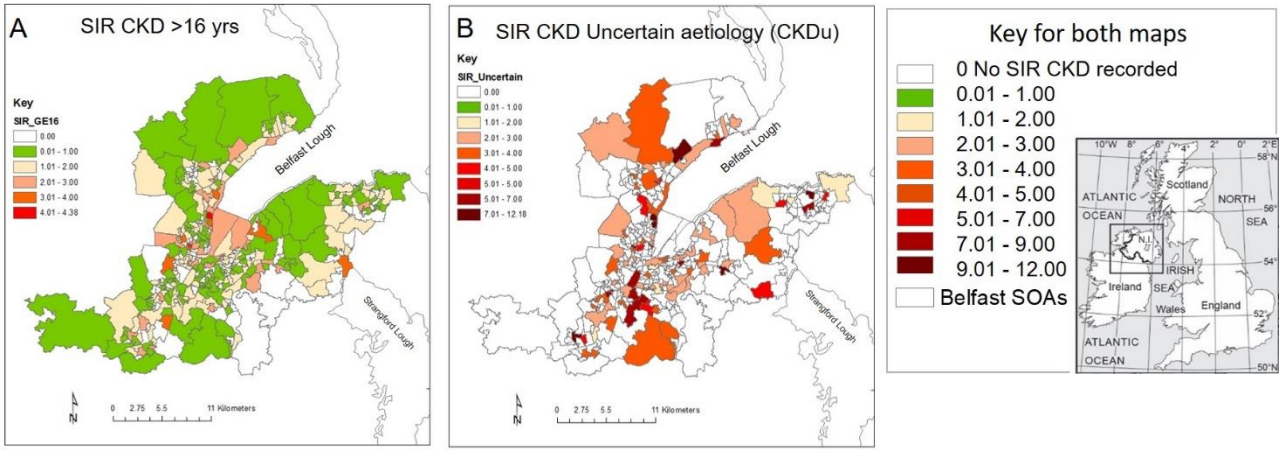


Figure 1 - Standardised Incidence Rates (SIRs) provided by the UKRR, for the Belfast urban area by Super Output Areas (SOA) for all ages >16 and for uncertain aetiology (CKDu).

Environmental factors

For this research we used an urban geochemical database generated as part of the Tellus Survey collected between 2005-2006 (Young and Donald 2013). The database consists of 1164 soil samples collected across the greater Belfast urban area of Northern Ireland, UK and analysed with XRF elemental analysis. Informed by the literature on the impact of anthropogenic contamination on urban areas (McIlwaine et al. 2017) and the potential environmental links with renal disease (Carrero et al., 2013; Grigoratos and Martini, 2015; Afsar et al., 2019; McKinley et al., 2020a and 2020b), 10 geochemical potentially toxic elements (PTEs) (Co, V, Cr, Ni, Zn, Sn, Pb, Sb, As and Mo) were selected for this study. Geochemical data were imputed using the detection limits provided in Young and Donald (2013). A geometric mean value for each geochemical PTE was calculated for each of the 265 SOAs within the greater Belfast urban area.

Air pollution data

Air pollution data were downloaded for Northern Ireland from the Department for Environment Food & Rural Affairs (DEFRA 2014). Annual air pollution data are available, data for 2006 were used for this research to coincide with the Tellus survey and UKRR data. The variables Benzene, CO, NOx, PM2.5, PM10 and SO_2 were used as air pollution covariates. For the pollutants where there were SOAs with missing values ordinary kriging with cubic models was used for imputation.

Measurement of social deprivation

Social deprivation was measured using Multiple Deprivation Measures provided by the Northern Ireland Statistics and Research Agency (NISRA, 2017). The Northern Ireland Multiple Deprivation Measures (MDM) for 2017 provided information on seven individual domains of deprivation (income; employment; health deprivation and disability; education, skills and training; access to services; living environment; and crime and disorder) across the greater Belfast area and an overall MDM ranking. The ranking scale was from 1 which represents the most deprived to 890 for the least deprived. The individual domains of deprivation and the overall MDM rankings were used to examine the relative deprivation of each SOA and explore any observed association with CKD SIRs and air pollution data.

Methods

A statistical approach was used to explore the relationship between the SIRs of CKD and CKDu with social deprivation measures, environmental factors and air pollution. A compositional approach using balances was used to account for the compositional nature of the geochemical data and air pollution data. A forward-

selection balance method using the *selbal* algorithm (discussed in more detail in Rivera-Pinto et al. (2018) and McKinley et al. (2020a)) with an *n*-fold cross-validation (CV) procedure was used to identify the set of balances for MDMs and geochemical PTEs and also for MDMs and air pollution data. The *selbal* approach was explored to identify components (MDMs, geochemical PTEs and air pollution data) whose relative abundance is associated with elevated incidences of CKD and CKDu. Using the elemental balances (MDMs, geochemical PTEs and air pollution data) with the strongest association with CKD and CKDu identified by the *selbal* approach, a generalised linear regression model (*glm* with log link) was used to examine the statistical significance of any observed relationship. To account for spatial autocorrelation the Moran's I statistic and a spatial lag model, using *spatialreg* R package, were used to test the residuals computed from the regression models. Where the Moran's I for the residuals was found to be significantly different from random, the GLM regression results where compared with a spatial lag model and the model fit compared using an Akaike Information Criterion (AIC).

3. Results

Using the forward-selection method using the *selbal* algorithm, the PTEs most frequently identified in the CV procedure, as the most associated with CKD for all SIRs >16 years, are Zn appearing 84% and Mo appearing 66% of the time, respectively. In addition to the global balance of Mo/Zn which appears in 64% of trials, the balances of Sn/Zn and Sn/Sb are also identified in the CV procedure (Table 1; frequencies 12% and 8% respectively). The results for CKDu identify the PTEs of Ni and Cr (Ni appearing 78% and Cr appearing 72% of the time) and the balances of Cr/Ni (Table 1; global balance, frequency 72%), As/Mo and As/Zn as the most frequent in the CV procedure.

The air pollution variables most frequently identified as most associated with CKD for all SIRs >16 years, are SO_2, appearing 96% and CO appearing 66% of the time, respectively. In addition to the balances of SO_2/CO, SO_2/Benzene and SO_2/PM10 are identified in the CV procedure (Table 1 global balance of SO_2/CO appears in 64% of the trials). Investigating the association between CKDu with air pollution data, the results indicate that CO appears in all balances, i.e. in 100% of trials, while PM2.5 appears in 68%, PM10 in 22% and SO_2 in 10% of trials. In the CV procedure, the balance of PM2.5/CO appears in 68% of the iterations, PM10/CO in 22% and SO_2/CO in 10% of trials (Table 1).

Table 1 Results of the forward-selection method using the selbal algorithm for Belfast Urban area (265 SOAs) SIRs of CKD for all ages > 16 years and Belfast Urban area (92 SOAs) SIRs of CKDu with soil PTEs and air pollutants, six individual domains of deprivation domains were used as covariates, in each case the three most common balances (with frequencies shown in brackets).

	CKD>16					CKDu			
	soil PTE	(f)	Air pollutants	(f)		soil PTE	(f)	Air pollutants	(f)
Balance 1	$\ln\left(\frac{Mo}{Zn}\right)$	(0.64)	$\ln\left(\frac{SO_2}{CO}\right)$	(0.64)		$\ln\left(\frac{Cr}{Ni}\right)$	(0.72)	$\ln\left(\frac{PM2.5}{CO}\right)$	(0.68)
Balance 2	$\ln\left(\frac{Sn}{Zn}\right)$	(0.12)	$\ln\left(\frac{SO_2}{Benzene}\right)$	(0.28)		$\ln\left(\frac{Mo}{As}\right)$	(0.14)	$\ln\left(\frac{PM10}{CO}\right)$	(0.22)
Balance 3	$\ln\left(\frac{Sn}{Sb}\right)$	(0.08)	$\ln\left(\frac{SO_2}{PM10}\right)$	(0.04)		$\ln\left(\frac{As}{Zn}\right)$	(0.04)	$\ln\left(\frac{SO_2}{CO}\right)$	(0.10)

Table 2 – Summary of regression results for Belfast Urban area SIRs of CKD for all ages > 16 years informed by the selected PTE and air pollution balances from selbal (using glm with log link).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.343659	2.550254	0.919	0.35898
PTE balance Mo/Zn	0.425762	0.152842	2.786	0.00575
PTE balance Sn/Zn	-0.01776	0.133955	-0.133	0.89462
PTE balance Sn/Sb	0.200353	0.13318	1.504	0.13374
Air pollution balance SO_2/CO	0.594122	0.84384	0.704	0.48204
Air pollution balance SO_2/Benzene	0.432155	0.440733	0.981	0.32777
Air pollution balance SO_2/PM10	-0.45626	0.634027	-0.72	0.47243
mdm$Income	0.000461	0.000215	2.145	0.03295
mdm$Employment	-0.00198	0.000623	-3.174	0.00169
mdm$Health	0.001184	0.000666	1.777	0.07683
mdm$Education	-0.00031	0.000348	-0.9	0.36913
mdm$Service	-0.00041	0.000268	-1.549	0.12268
mdm$Living	0.000135	0.000144	0.943	0.34639

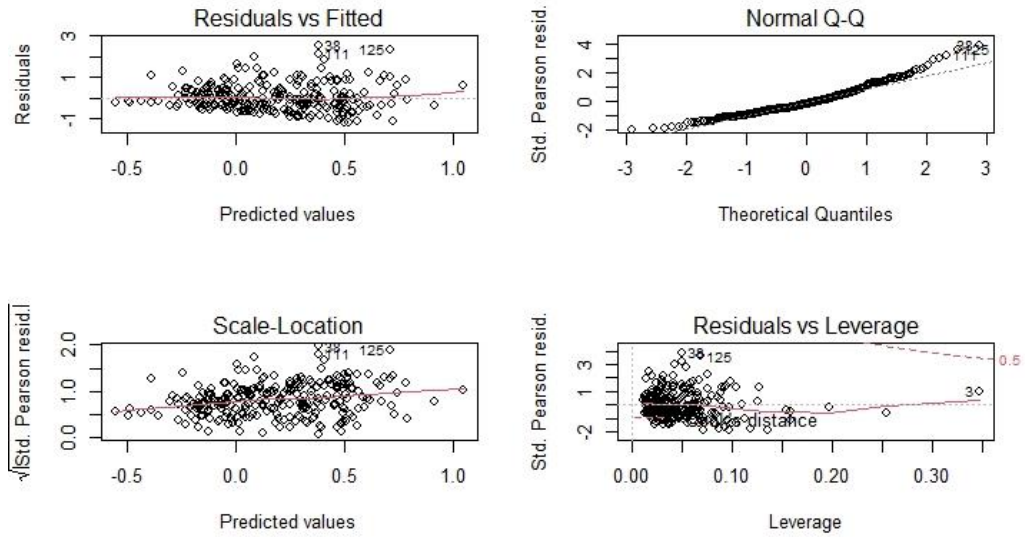


Figure 3 - Graphs for regression results for Belfast Urban area SIRs of CKD for all ages > 16 years informed by the selected PTE and air pollution balances from selbal (using glm with log link) (coefficients shown in Table 2)

The GLM regression results exploring the relationship between the CKD with social deprivation measures, environmental factors and air pollution, indicate a statistically significant correlation between CKD SIR>16 the identified balance of Mo/Zn (Table 2; significance level of 0.001) and the MDM domains of employment, income and health (Table 2; significance levels of 0.001, 0.01 and 0.05 respectively). GLM regression results for

CKDu with social deprivation, environmental factors and air pollution indicated that none of the balances have coefficients that are statistically significant.

To assess the impact of spatial autocorrelation on the regression models, Moran's I test statistic was used to test the residuals computed from the regression models (Fig. 2). The Moran's Index for the residuals for the GLM regression analysis for CKD SIRs >16, with identified balances for PTEs and air pollutants with social derivation measures, was found to be not significantly different from random (Moran's I statistic 0.0304, p-value = 0.3139). Referring to the Akaike Information Criterion (AIC) we find that the GLM has an AIC of 557.59 whereas the spatial lag model using log (CKD SIRs>16) has an AIC of 451.95, indicating that the regression model provided a better fit when a spatial lag was included. The coefficients for the spatial lag model using log (CKD SIRs>16) were not found to be statistically significant.

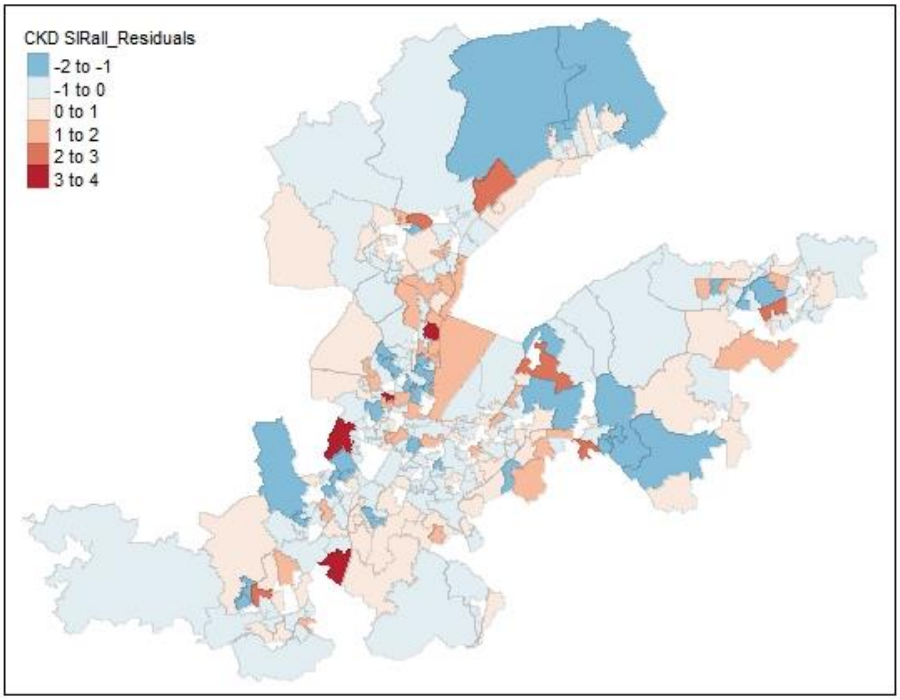


Figure 5 – Residuals shown for the regression results for Belfast Urban area SIRs of CKD for all ages > 16 years informed by the selected PTE and air pollution balances from selbal (using glm with log link) (coefficients shown in Table 2).

4. Discussion and Conclusions

Urbanisation through industrialisation, atmospheric pollution deposition, traffic pollution and brake wear emissions has been linked to harmful impacts on kidney function. (Carrero et al., 2013; Grigoratos and Martini, 2015). Atmospheric pollution deposition, traffic and brake wear emissions have been cited as sources for the PTEs of Zn, Sn, Pb, Sb, As and Mo, with a blood-borne pathway of ultrafine particles of these PTEs which may translocate to the kidney. Using an urban soil geochemistry database of total element concentrations, this study explored the statistical relationship between Standardised Incidence Rates (SIRs) of CKD and CKDu with social deprivation measures and environmental factors including air pollution. The findings the GLM regression (with log link), reveal a statistically significant correlation between CKD for all SIRs for the age group >16 years and the identified balance of Mo/Zn (significance level of 0.001) and the MDM domains of employment, income and health (significance levels of 0.001, 0.01 and 0.05 respectively). However, when a spatial lag was included in the regression model, the coefficients were not found to be statistically significant.

This research also shows that the air pollutants most frequently identified as being most associated with CKD for all SIRs >16 years, are SO_2 appearing 96% and CO appearing 66% of the time, respectively. In addition the

results for CKDu indicate an association with the air pollutants of CO and PM2.5 (appears 100% and 68% of the trials). The association between CKD for all SIRs >16 years and the air pollutants was not found to be statistically significant. However, these preliminary findings do support the argument that atmospheric pollution in the form of SO_2, CO, Benzene, PM10 and PM2.5 exposure deposition and associated toxic metals may negatively affect renal function. Further research is required to fully examine the impact of atmospheric pollutants and chronic kidney disease.

References

Afsar, B., Afsar, R.E., Kanbay, A., Covic, A., Ortiz, A & Kanbay, M. (2019), Air pollution and kidney disease: review of current evidence. Clinical Kidney Journal, vol. 12, no. 1, 19–32 doi: 10.1093/ckj/sfy111.

Brook RD, Rajagopalan S, Pope CA, Brook JR, Bhatnagar A, Diez-Roux AV, al. (2010), Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association.Circulation, 121:2331-237820458016.

Carrero, J.A., Arrizabalaga, I., Bustamante, J., Goienaga, N., Arana, G., Madariaga, J.M., (2013), Diagnosing the traffic impact on roadside soils through a multianalytical data analysis of the concentration profiles of traffic-related elements. Science of the Total Environment. 458e460, 427e434. http://dx.doi.org/10.1016/j.scitotenv.2013.04.047.

Department for Environment Food & Rural Affairs (DEFRA) (2014). Guide to UK Air Pollution Information

Resources, June 2014 https://uk-air.defra.gov.uk/air-pollution/.

Geiser M. Kreyling WG. (1999), Deposition and biokinetics of inhaled nanoparticles. Part Fiber Toxicol 7(2).

Grigoratos T, Martini, G. (2015), Brake wear particle emissions: a review. Environ Sci Pollut Res 22:2491–2504 DOI 10.1007/s11356-014-3696-8.

Lindeman, R., Tobin, J. & Shock, N.W. (1985), Longitudinal studies on the rate of decline in renal function with age. J. Am. Geriatr. Soc., 33, 278–285.

Lue SH, Wellenius GA, Wilker EH, Mostofsky E, Mittleman MA. (2013), Residential proximity to major roadways and renal function. J Epidemiol Community Health 67:629–634.

Mehta AJ, Zanobetti A, Bind MC, Kloog I, Koutrakis P, Sparrow D, Vokonas PS, Schwartz JD. (2016), Long-term exposure to ambient fine particulate matter and renal function in older men: the VA Normative Aging Study. Environ Health Perspect 124:1353–1360; http://dx.doi.org/10.1289/ehp.1510269

McIlwaine, R., Doherty, R., Cox S. & Cave, M. (2017), The relationship between historical development and potentially toxic element concentrations in urban soils. Environmental Pollution Vol. 220, Part B, January 2017, Pages 1036-1049 https://doi.org/10.1016/j.envpol.2016.11.040.

McKinley, J. M., Mueller, U., Atkinson, P. M., Ofterdinger, U., Jackson, C., Cox, S. F., et al. (2020a), Investigating the influence of environmental factors on the incidence of renal disease with compositional data analysis using balances. Applied Computing and Geosciences, 6, 100024. https:// doi.org/10.1016/j.acags.2020.100024.

McKinley, J.M, Mueller, U., Atkinson, P.M., Ofterdinger, U., Cox, S F., Doherty, R., Fogarty, D., Egozcue, J.J., Pawlowsky-Glahn, V. (2020b), Chronic kidney disease of unknown origin is associated with social deprivation and environmental urbanisation in Belfast, UK., Environ Geochem Health (2020). https://doi.org/10.1007/s10653-020-00618-y.

Musso, C.& Oreopoulos, D. (2011), Aging and physiological changes of the kidneys including changes in glomerular filtration rate. Nephron Physiol., 119, p1–p5.

Nemmar A, Al-Salam S, Zia S, Yasin J, Al Husseni I, Ali BH. (2010), Diesel exhaust particles in the lung aggravate experimental acute renal failure. Toxicol Sci 113:267–277.

Northern Ireland Statistics and Research Agency (NISRA (2017), NI Multiple Deprivation Measures 2017 – Summary Booklet 28p.

Oberdörster, G., Oberdörster, E., & Oberdörster, J. (2005), Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. Environ Health Perspect 113:823–839.

Rivera-Pinto, J., Egozcue, J.J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M. & Calle, M.L. (2018), Balances: a new perspective for microbiome analysis. mSystems https://doi.org/10.1101/219386.

Thomson EM, Vladisavljevic D, Mohottalage S, Kumarathasan P, Vincent R. (2013), Mapping acute systemic effects of inhaled particulate matter and ozone: multiorgan gene expression and glucocorticoid activity. Toxicol Sci 135:169–181.

UK Renal Registry (UKRR) (2019). UK Renal registry 21st annual report—data to 31/12/2017. Bristol, UK. Retrieved October 30, 2019 from https://www.renalreg.org/ publications-reports/.

Young, M. E. & Donald, A. W. (2013), A guide to the Tellus data. Geological Survey of Northern Ireland (GSNI), Belfast. http://nora.nerc.ac.uk/509171/. Accessed October 11th 2019.

A COMBINED APPROACH TO EVALUATE LOCAL IMPACTS OF AIR POLLUTION EXPOSURE ON HEALTH USING SYNTHETIC DATA

Manuel Castro Ribeiro (1)* - Maria João Pereira (1)

Universidade de Lisboa, Cerena, Decivil, Instituto Superior Técnico (1)
** Corresponding author: manuel.ribeiro@tecnico.ulisboa.pt*

Abstract

In epidemiologic research, the assessment of the impacts of exposure to ambient air pollution on disease incidence or mortality are usually based on the analysis of misaligned data collected at air-quality monitoring stations not coinciding with health data locations. In such cases, geostatistics does provide suitable methods to map exposure in a continuous spatial domain and to assess exposure uncertainty at health data locations. In turn these results can be combined with regression models to estimate regression coefficients and draw a measure of spatial uncertainty of associations between ambient air pollution exposure and disease incidence or mortality.

However, the impacts of air pollution may vary geographically: for example, the impacts of particulate matter on health may vary from local to local as its composition and toxicity may vary with different source contributions. In these cases, it is important to consider additional local adjustments in exposure regression coefficients, which can be achieved using geographically weighted regression models as they are expressed as functions of spatial locations. Furthermore, local adjustments should be extended to incorporate spatial uncertainty of exposure impacts, since they may vary throughout the spatial domain. If exposures are predicted with a geostatistical interpolator, standard errors associated to exposure coefficients provide a measure of exposure uncertainty, but do not take into account the spatial uncertainty of predicted values as they are subject to prediction error. In this work we propose a solution based on the combination of geographical weighted regression models with geostatistical simulation algorithms incorporating therefore spatial uncertainty of exposure in estimation of local varying regression coefficients.

We used synthetic data differing in sample geometry, sample size and spatial covariance structure of exposure data, to compare the performance of two statistical approaches assessing local uncertainty of associations between fine particles and birth weight: one approach combined geographically weighted regression models with geostatistical interpolation models, the other combined geographically weighted regression with geostatistical sequential simulation algorithms. The performance measures considered were bias and empirical standard errors of the exposure regression coefficient.

Results indicate that the approach combining geographically weighted regression with geostatistical sequential simulation models is the best choice, especially when exposure data have larger sample sizes and exhibit stronger spatial covariance structures. Nevertheless both methods capture local trends of associations between ambient fine particles exposure and birth weight and are sensitive to different levels of spatial uncertainty arising from those trends.

By combining geographically weighted regression models with geostatistical sequential simulation algorithms, we were able to incorporate local uncertainty on exposure regression coefficients estimates with more efficiency (precision), providing an additional tool to health analysts in assessment of the impacts of place in health. All computations were performed using the R software environment, version 3.6.0. The R script for reproducibility of the analysis is available upon request.

EFFECT OF GLYPHOSATE AND PARAQUAT ON SEEDS GERMINATION AND SEEDLINGS OF SORGHUM VULGARE, PHASEOLUS VULGARIS AND VICIA FABA

Mohamed Maldani (1)* - Fatima Zahra Aliyat (1) - Simone Cappello (2) - Marina Morabito (3) - Maria Genovese (2) - Santina Santisi (2) - Filippo Giarratana (4) - Laila Nassiri (1) - Jamal Ibijbijen (1)

University Moulay Ismail, Faculty of Sciences (1) - National Research Council, Institute of Biological Resources and Marine Biotechnology (2) - University of Messina, Department of Chemical, Biological, Pharmaceutical and Environmental Sciences (3) - University of Messina, Polo Universitario dell'Annunziata, Department Of Veterinary Science (4)

** Corresponding author: mohamed.maldani@gmail.com*

Abstract

Agriculture is today faced with the need for a profound change to meet the current challenges that are environmental, climatic, food and social or economic. Among the solutions to meet these challenges, agriculture makes significant use of pesticides that can affect the production of any species and pose a risk, as most can be harmful to health and the environment.

Farmers aim to treat fields with different pesticides against many crop pests without taking into consideration their detrimental effects on the germination and growth of their plants.

In order to determine the effect of pesticides on the germination and emergence of three species of importance in agricultural production in Morocco: Sorghum vulgare, Phaseolus vulgaris and Vicia faba. An experiment is shown in this direction to determine the effect of two pesticides paraquat and glyphosate how the most are used in the Meknes region, which is a farming region in Morocco.

The seeds placed in petri dishes were kept in the dark at a constant temperature of 28 ± 1 °C and irrigated with three solutions of paraquat and glyphosate at different concentrations (0, 1, 2 and 4 g / l) for paraquat and (0, 1, 2 and 5.4 g / l) for glyphosate. Seeds germination was counted at the end of the experiment to calculate the rate and quality of germination. Seedlings are grown in pots that contain peat to determine the effect of pesticides after seed germination.

The results of the experiment showed that the types and concentrations of pesticides reduced the rate, the quality of germination and affected seed growth after germination.

In general, all the tested pesticides stopped the growth of all germinated seeds of the three species after germination when applied at a higher concentration than the recommended dose, but at lower doses, the pesticides had negative effects on growth versus control.

AN O2S2 ANALYSIS OF THE IMPACT ON TOTAL MORTALITY OF THE COVID-19 PANDEMIC IN ITALIAN MUNICIPALITIES

Riccardo Scimone (1)* - Alessandra Menafoglio (1) - Piercesare Secchi (1)

Politecnico di Milano, Dipartimento di Matematica (1)
** Corresponding author: riccardo.scimone@polimi.it*

Abstract

We propose a geostatistical analysis of ISTAT deaths data in Italy, collected daily, for different age classes and in each Italian municipality, during the time period from 2015 to 2020. Such data allows for the exploration of the daily and spatial variability of the death process, in different years and for different age groups, and permits to appreciate the impact of COVID 19 on total mortality in the Country. The yearly sequences of daily death data, indexed by the municipality they are referred to, can be modelled as a sample of spatially dependent functional data and analysed using tools from Object Oriented Spatial Statistics (O2S2). For this talk, the main tools for the analysis will be Kriging in the Bayes space B^2 of probability density functions coupled with Spatial Downscaling for constrained functional data. The analysis will allow us to evaluate the effect of the COVID 19 pandemic on total mortality at the very granular scale of municipalities.

CONTAMINANT RELEASE HISTORY IDENTIFICATION THROUGH SIMULATION-OPTIMIZATION METHOD AND SURROGATE TRANSPORT MODEL

Azade Jamshidi (1)* - Jamal Mohammad Vali Samani (1) - Maria Giovanna Tanda (2) - Andrea Zanini (2)

Tarbiat Modares University (1) - University of Parma (2)
** Corresponding author: azade_jamshidi@modares.ac.ir*

Abstract

Identification of the contaminant release history has received considerable attention in the literature over the past several decades. From a mathematical point of view, this kind of question belongs to the class of ill-posed problems, whose solutions are not characterized by the usual existence, uniqueness and/or stability properties. Among the solution approaches, optimization is one of the most widely used solution approaches to solve the problem. The optimization method consists of the integration of both simulation and optimization models. The main purpose of simulation models is to solve the governing flow and transport equations for given initial and boundary conditions. However, these models are not capable to estimate the inverse solution. Therefore, they have to be integrated with optimization models, which aim to identify the input of the forward model that best fit the observed data. The optimization procedure is based on an iterative process that requires many forward model evaluations. The computational tractability of such a simulation-optimization approach could be enhanced by improving the efficiency of both the simulation model and the optimization method. Several approaches such as surrogate modeling are available to improve the simulation model efficiency by reducing the forward model computation time. In this paper, transfer function theory is applied as surrogate model in the simulation process and, then, it is integrated with an optimization algorithm to estimate the contaminant release history in groundwater. Transfer function is a dynamic system theory, which has been applied to model the groundwater transport process as an input-output system. For this purpose, a literature study case that consists in a 2-dimensional heterogeneous aquifer with two contaminant sources and seven monitoring wells is considered. MODFLOW and MT3DMS codes were applied to compute the transfer functions, which provide information between the sources and the monitoring points. Concentration time series observed at monitoring wells were considered as input data for the inverse problem. The results showed that the present methodology can identify source release fluxes with the same accuracy of the literature study but with only one run of the complete simulation model and in much less computation time.

LOGISTIC GAUSSIAN FIELDS FOR INVERSION BASED ON STOCHASTIC RESPONSES

Athénaïs Gautier (1,2)* - David Ginsbourger (1,2) - Guillaume Pirot (3)

Idiap Research Institute (1) - University of Bern (2) - University of Western Australia (3)
** Corresponding author: athenais.gautier@idiap.ch*

Abstract

When tackling inverse problems involving complex systems with stochastic responses departing from moderate-dimensional Euclidean settings, it is common to appeal to methods such as Approximate Bayesian Computation (ABC) that consists in identifying parameter values generating responses close to observations.

The employed notion of closeness or similarity is typically defined with the help of chosen summary statistics, and one aspect that makes such ABC methods challenging despite aforementioned simplifications is that for any instance of parameters, the obtained dissimilarity inherits from the response's randomness.

Our main aim here is to ease the derivation of approximate posterior distributions by appealing to random distribution field modelling and prediction.

From the geoscientific side, our contribution is motivated by the following problem: localizing the source of a contaminant when the physics is known but the geological structure is not. Our reference observations consist in contaminant concentration curves over time at several observation wells. In order to account for the uncertainty over the geology, we introduce a distribution over plausible geological realizations. In the spirit of ABC, we introduce a measure of dissimilarity over the response space (concentration curves derived from geological realizations) and use it to seek for approximate posterior distributions of the contaminant source location. A means to do so is indeed to model and predict the field of misfit distributions indexed by the source.

While existing methods such as distributional kriging are well known in the geostatistical community, they turned out to be far from straightforward to adapt to the considered specific case where density fields must be modelled and predicted based on samples of heterogeneous sizes across the source space. Our contributions build upon a non-parametric Bayesian approach to modeling fields of probability distributions, and in particular to a spatial extension of the logistic Gaussian model.

We demonstrate the applicability of this model on the aforementioned contaminant localization problem. We further introduce an inversion approach based on it, that delivers a probabilistic prediction of the posterior distribution of the source and enables us to draw realizations of the random posterior induced, hence allowing us to deliver assessments of the associated inversion uncertainty, with a view towards sampling strategies for parsimonious uncertainty reduction.

BAYESIAN TIME-LAPSE INVERSION OF GEOPHYSICAL DATA FOR WATER SATURATION CHANGES DURING SNOWPACK MELTING IN MOUNTAIN WATERSHEDS

Dario Grana (1)* - Natalie Smeltz (1) - Mohit Ayani (1) - Andrew Parsekian (1)

University of Wyoming, Geology and Geophysics (1)
** Corresponding author: dgrana@uwyo.edu*

Abstract

Snow precipitation on mountain hillslopes infiltrates the subsurface and recharges groundwater aquifers. The spatial distribution of the water volume depends on the porosity of the rocks and the fluid saturation. These properties, as well as the temporal changes of water saturation, can be monitored using surface geophysical data, such as seismic refraction and time-lapse electrical resistivity tomography. In this work, we propose a Bayesian joint inversion of geophysical data to predict the spatial distribution of porosity and water saturation in the near surface. Porosity changes laterally and with depth, but we assume that it is constant in time at each location. Water saturation is a function of space and time, since the water volume changes during time when snow melts and infiltrates into the subsurface. The relation between the properties of interest and the measured data is a multivariate rock physics model based on Hertz-Mindlin contact theory for the elastic component and Simandoux equation for the electrical component. The model accounts for pressure and mineralogy changes in depth. Because the properties of interest are bounded between 0 and 1, and the rock physics model is not linear, Bayesian Gaussian-linear inverse methods cannot be applied. We model the joint distribution of model and data variables according to non-parametric probability density functions, approximated using kernel density estimation methods. Because direct measurements of the properties of interest are limited, we generate a training dataset using a Monte Carlo simulation method: we sample porosity and water saturation from a bivariate uniform distribution, we apply the joint elastic-electric rock physics model, and we calculate the joint distribution of model properties and data variables with kernel density estimation. The Bayesian inversion method is first applied to the base survey and then to the repeated time-lapse surveys assuming constant porosity in time. The result of the inversion is a set of pointwise marginal probability density functions of porosity and time-dependent water saturation in a discretized grid of the surface at each time geophysical data are available. We tested the methodology on a geophysical dataset acquired along a 60 m section of mountain hillslope near Laramie, Wyoming, USA. The results are validated using nuclear magnetic resonance observations of water volume in a nearby borehole. The so-obtained porosity and water saturation maps can be used to reduce the uncertainty in hydrological models. Our future research aims to combine geophysical and hydrological models to predict the spatial distribution of the subsurface water produced from snowpack melting that flows and is stored in mountain watersheds in order to make more informed decisions on the water management.

COMBINING 2D GROUNDWATER PARAMETER INVERSION AND TRANSITION PROBABILITY GEOSTATISTICS TO CONSTRUCT A 3D AQUIFER MODEL

Dimitri Rambourg (1)* - Olivier Bildstein (2) - Philippe Ackerer (3)

Lhyges, Université de Strasbourg / Lmte, Cea Cadarache (1) - Lmte, Cea Cadarache (2) - Lhyges, Université de Strasbourg (3)

** Corresponding author: d.rambourg@unistra.fr*

Abstract

The identification of parameter - i.e. specific yield and hydraulic conductivity - over the entirety of an aquifer domain is crucial to properly model groundwater flow and contamination. Most of the time, only discrete data are available, in time and space. Also, measurement of those parameters, through pumping tests for example, are difficult and costly to acquire, especially over large areas. Inversion techniques and geostatistical procedures can be used to overcome these limitations, allowing to unravel complex systems heterogeneity with limited data.

The methodology aims at parametrizing a 3D aquifer model and proceeds in three stages. First, head water measurements (piezometers) are used to calibrate a 2D flow model, averaged over the vertical. The inversion procedure is regularized thanks to a Zonal Adaptive Multi-Scale Triangulation (ZAMT), using a parameter grid dissociated from the calculation mesh, able of local refinement and zoned in order to integrate prior geological knowledge about the nature and distribution of large-scale heterogeneity. In parallel, the 3D distribution of heterogeneity is inferred from spatial analysis of borehole geological data, applying transition probabilities and Markov Chain processes (T-PROGS, S. Carle, 1999) with four classes of hydrofacies (from very low to high permeability). Then, the numerical value of each facies hydraulic conductivity is estimated through a Levenberg-Marquard algorithm, minimizing the discrepancy between the 2D transmissivity, given by the inversion of the 2D flow model, and the computed 3D transmissivity based on the hydrofacies. The resulting parametrized 3D model subsequently allows to run contamination simulations (TRACES, H. Hoteit et al., 2002).

The study site is an alluvial (unconfined) aquifer of 7.6 km², situated in the southern, Mediterranean part of France. The inversion runs with head water time series from 196 piezometers over 7 years (2012-2019), at a decadal time step, and with a convergence threshold ensuring a mean error less than 40 cm. The spatial analysis for constructing the reservoir geometry relies on 639 boreholes descriptions. T-PROGS interpolation well preserves the global lithofacies proportions and produces good continuity of geological layers, consistent with the deposition pattern expected for an alluvial aquifer.

GRAVITY FORWARD MODELLING WITH GECCO TOOLS AND 3D GRAVITY INVERSION APPLIED TO STUDY GEOLOGICAL SUBSURFACE STRUCTURES WITHIN THE URBAN AREAS IN SOUTHERN FINLAND

Eevaliisa Laine (1)* - Hilding Linden (2) - Taija Huotari (1) - Heidi Laxström (1) - Ilkka Suppala (3) - Jan Westerholm (2)

Geological Survey of Finland, Construction and Energy Solutions (1) - Åbo Akademi University, Faculty of Science and Engineering (2) - Geological Survey of Finland, Geophysical Solutions (3)
** Corresponding author: eevaliisa.laine@gtk.fi*

Abstract

The project GECCO combined expertise in high performance computing and geomodelling, and developed tools for faster geological common earth model (CEM) modelling in a powerful computing environment. Geological structures and lithological heterogeneity is modeled by using traditional geological modelling and geostatistical simulations constrained by drill core data and geological cross sections. In the project, the tools were developed for forward electromagnetic, gravity and magnetic modelling. Geophysical field data is coming from specialized geophysical software and included into the GECCO workflow as csv files, either as profile data or interpolated into a regular grid using kriging estimation. A specific interface called GECCOGRAM for gravity and magnetic forward modelling and 3D visualization was created for LINUX operating system either in personal or super computers. The calculated responses are exported as vtk files and visualized using Paraview, which is an open-source, multi-platform data analysis and visualization application. In the GECCOGRAM simple geometric modifications can be done, for example, rotation, translation and scaling of the geological objects.The GECCO workflow, especially the GECCOGRAM and gravity inversion, will be demonstrated using gravity data from southern Finland. The bedrock of southern Finland is mainly composed of Precambrian granites, granodiorites and mica gneisses and is characteristically migmatic. The specific emphasis is laid on the postorogenic granitic intrusions and their relation to brittle structures near urban areas.

PARAMETERIZING SPATIALLY COMPLEX CONCEPTUAL MODELS FOR BAYESIAN OPTIMIZATION

Guillaume Pirot (1)* - Ranee Joshi (1) - Mark Jessell (1) - Mark Lindsay (1)

University of Western Australia, Centre for Exploration Targeting (1)
** Corresponding author: guillaume.pirot@uwa.edu.au*

Abstract

Geological models are useful for the exploration of natural resources such as groundwater, minerals or geothermal energy among other applications. Out of convenience and because of perception bias, geological models often rely on a single geological interpretation, regardless of the purposes they serve or how they are built. However, data and knowledge available for subsurface characterization are always limited and prone to some errors, which should lead modellers to consider a large ensemble of geobody geometries and a wide range of values for subsurface properties. In addition, if we were to ask for a geological interpretation from a hundred different geologists, we would get a hundred different variations. Thus, to minimize the risk of biased and over-confident predictions, modellers need to consider an ensemble of plausible geological interpretations also called conceptual models, in addition to what is classically performed as parametric uncertainty propagation.

Several model selection techniques are available to reduce conceptual uncertainty among a finite set of initial scenarios. However, by considering a discrete and finite set of conceptual models for the inversion of real data – as opposed to synthetic reference data, all initial scenarios might be rejected or the probability of keeping one of them might be very low. Indeed, the lack of knowledge and data combined with human perception, interpretation and bias might lead to a set of erroneous or incomplete initial scenarios. Here, we propose an approach to build a continuous parameter space representation of conceptual models from a discrete and finite set of initial geological scenarios. It allows us to explore additional locations of the conceptual model parameter space to build scenarios that are compatible with collected geological data and it allows us to quantify uncertainty around the initial scenarios.

To illustrate our approach, we consider geological data from the Yalgoo-Singleton area and various geological history scenarios to define several plausible conceptual models. The scenarios are defined as the product of expert knowledge and combinatorial exploration of geological events. The resulting conceptual models are compared in terms of topology, connectivity and geostatistics. Due to the low number of initial scenarios the conceptual model parameter is explored in a low-dimension representation. We use a Bayesian optimization approach relying on the Expected Improvement criteria to localize optimal scenarios in the reduced parameter space. To assess the relevance of a scenario (i.e. a set of model parameters) with respect to geological observations, the objective function to minimize is defined as a distance between summary metrics referring to connectivity, dispersion and multiple-point statistics characteristics of the subsurface property fields.

Acknowledgement: We acknowledge the support from the ARC-funded Loop: Enabling Stochastic 3D Geological Modelling consortia (LP170100985) and DECRA (DE190100431). The work has also been supported by the Mineral Exploration Cooperative Research Centre whose activities are funded by the Australian Government's Cooperative Research Centre Programme. This is MinEx CRC Document 2019/14.

ENSEMBLE KALMAN FILTER FOR POLLUTION SOURCE CHARACTERIZATION IN WATER SUPPLY SYSTEMS

Ilaria Butera (1)* - J. Jaime Gómez-Hernández (2) - Silvia Nicotra (1)

Politecnico di Torino, Diati (1) - Universitat Politècnica de València, Institute for Water and Environmental Engineering (2)
** Corresponding author: ilaria.butera@polito.it*

Abstract

Water distribution systems are a core infrastructure in people lives. Intentional or accidental contaminations can threaten their health and have to be detected in the shortest possible period to reduce damages. Early warning systems should be put in place to detect both the source location and the release intensity.

How to identify a contaminant source from concentration observations at monitoring locations can be cast as an inverse problem for which different approaches are available. In this work, the ensemble Kalman filter (EnKF) is chosen.

The EnKF is demonstrated in the Anytown network, which is a benchmark in water supply system analysis. The network is subject to a time variable demand in which a contaminant is introduced. The contaminant source is determined from concentration observations made in time at different frequencies. Measurement errors on concentration and estimation errors on the base demand are included to make the test case more realistic.

The case study deals with a release with uniform intensity that is originated from a source located in a node of the network. The sensors of the network register concentration values in time with a certain frequency. The scheme adopted for concentration sampling considers a malfunctioning of the sensor network, which introduces observation errors, and it is also assumed that sampling starts some time after the release has occurred and the contaminant has already spread through the pipeline systems.

Different locations of the source, frequency sampling and acquisition data period have been considered.

The results of the tests are very satisfactory for all the examined cases, in spite of the limited number of the ensemble members (48 realizations) and the non-stationarity of the concentration field, due to the intrinsic functioning of the network.

Results show that for the Anytown network, an early detection of solute concentration (within 60 minutes from the release beginning) together with a sampling frequency of 30 minutes is sufficient to accurately detect the source parameters in a short time. If the monitoring starts later, e.g. 3 hours after the beginning, the identification takes a longer time.

ACCOUNTING FOR PETROPHYSICAL UNCERTAINTY IN HYDROGEOPHYSICAL INVERSION WITH THE CORRELATED PSEUDO-MARGINAL METHOD

Lea Friedli (1)* - Niklas Linde (1) - Arnaud Doucet (2) - David Ginsbourger (3)

University of Lausanne, Institute of Earth Sciences (1) - University of Oxford, Statistics (2) - Idiap, Uncertainty Quantification and Optimal Design (3)
** Corresponding author: lea.friedli@unil.ch*

Abstract

Hydrogeophysical investigations aim at obtaining information about hydrogeological properties or processes from indirect geophysical data. The petrophysical relationship describing the link between the hydrogeological and the geophysical properties is generally non-linear and includes significant scatter. We consider the inversion problem, in which hydrogeological parameters are inferred from geophysical data and the intermediate geophysical properties are treated as latent variables. Instead of solving the inverse problem in a traditional two-step approach by first inferring for the geophysical properties and afterwards for the hydrogeological parameters while accounting for petrophysical uncertainty, we use a Metropolis-Hastings scheme to infer directly the hydrogeological parameters from the geophysical data. In doing so, we need to estimate the intractable likelihood of observing the geophysical data given the proposed hydrogeological parameters. The Pseudo-Marginal method relies on an unbiased approximation of this likelihood based on Monte-Carlo averaging over samples from the petrophysical relationship, thereby ensuring that the scattered nature of the petrophysical relationship is taken into account. To increase the efficiency of the resulting Metropolis-Hastings scheme, we lower the variance of the likelihood ratio at each Metropolis step by correlating the samples of the petrophysical relationship used in the proposed and current steps of the Markov chain. We assess the performance of this Correlated Pseudo-Marginal method with two test cases: A synthetic example in which we invert for a porosity field and a field study in which we invert for hydraulic conductivity, both using crosshole ground-penetrating radar (GPR) travel times as geophysical data. The Correlated Pseudo-Marginal method is further compared with an approach in which the petrophysical uncertainty is accounted for by using only one brute-force Monte Carlo sample at each Metropolis step (so-called lithological tomography). A drawback of the latter approach is the need of small MCMC step sizes as a result of the peaky likelihood function of the geophysical data given the geophysical properties. By increasing the number of samples from the petrophysical relationship drawn in each step, we expect that the MCMC algorithm can take much larger step, thereby, alleviating the costs of multiple Monte Carlo samples at each Metropolis step. Furthermore, since brute-force Monte Carlo sampling is an embarrassingly parallel problem, the actual run-time on a computer cluster will be drastically reduced. We further expect that the Correlated Pseudo-Marginal method will outperform the Pseudo-Marginal method by enhancing the efficiency and reduce the computational cost of the algorithm.

DECONVOLUTION OF GAMMA-RAY SPECTROMETRIC MEASUREMENTS FOR RADIOLOGICAL SITE CHARACTERIZATION

Md Moudud Hasan (1,2)* - Tim Vidmar (1) - Bart Rogiers (1) - Eric Laloy (1) - Jos Rutten (1) - Johan Camps (1) - Marijke Huysmans (2)

Belgian Nuclear Research Centre SCK CEN (1) - Vrije Universiteit Brussel (VBU), Department of Hydrology and Hydraulic Engineering (2)
** Corresponding author: mhasan@sckcen.be*

Abstract

A site contaminated by radioactive material needs to be characterized accurately. In situ gamma spectrometry is often used to estimate the level of radioactive contamination in the soil. However, the field of view of a gamma detector can be tens of meters depending on the height of the detector above the ground surface, the source energy and the detector properties, etc. If one works with the assumption that the measurements represent an average over a certain support volume, a local underestimation could result. The contribution of the surrounding area should be disentangled to estimate the level of contamination more correctly. In our experiment, we took gamma spectrometry measurements in a regular grid using a portable gamma detector setup. Calibration radioactive source pads of two different radionuclides i.e., Ba-133 and Cs-137, were used. The detector response as a function of distance is required to de-convolute the measurements. This was simulated using the Monte Carlo N-Particle (MCNP) transport code. For the deconvolution, a least-squares optimization based inversion method was used. Results show that the method is a convenient approach to deconvolute the contribution of nearby areas and increase the resolution of radioactive contamination mapping.

Keywords: Inversion, Gamma spectrometry, In situ, radioactivity, MCNP

1. Introduction

A site contaminated by radioactive material needs to be characterized accurately. An over-or under-estimation of the contamination level may affect the impact assessment and the remediation options. The conventional approach of measuring the radioactivity is to collect samples from the site and measuring them in a laboratory-based gamma-ray spectrometer and/or other radionuclide analytical techniques (IAEA, 2017; Tyler, 2008). This approach provides precise information about the activity level of different radionuclides with a low level of minimum detectable activity (MDA) (IAEA, 2017). However, this approach can be expensive and time-consuming to collect, prepare and measure individual samples. Hence, spatial coverage can be limited and spatial heterogeneities may not be captured through this approach (Tyler et al., 1996; Varley et al., 2017). On the other hand, in situ measurements are likely to be more representative of the measurement area. Furthermore, in situ measurements are relatively low cost and fast, as significant counts can be obtained in a short time because of a large support volume (IAEA, 2017). In situ gamma spectrometry is often used to estimate the level of radioactive contamination in a site (Berens, 2016; Duarte et al., 2011; Guérin and Mercier, 2012; Mikami et al., 2015; Tyler, 2008; Zhukouski et al., 2018). During in situ measurements, a gamma detector's field of view can be tens of meters depending on the height of the detector above the surface, the source energy, the density of soil and air and the detector properties (Androulakaki et al., 2016; Rostron et al., 2014; Zhang et

al., 2015). Therefore, the detector at a particular location will register gamma counts coming from surrounding areas. A heavy shield can be used to avoid such contribution from surrounding areas and limit the investigated area to a well-defined geometry. Such heavy shielding can reduce the portability of the detector and increase the measurement time. On the other hand, assumption that the measurements represent an average over a certain volume can induce local underestimation of the contamination levels. Hence, it is required to disentangle the contribution of the adjacent areas to calculate the concentration of a radionuclide at a particular location. An inversion approach can be used to deconvolute the contribution of the nearby areas when several measurements are available, possibly on a regular grid. However, in such cases, a simple inversion method does not work properly due to measurement errors (Ogawa et al., 2017). A proper inversion method is required for deconvolution of in situ gamma spectrometric measurements. To our knowledge, very few studies have been conducted in this field (Druker, 2017; Ogawa et al., 2017). Moreover, detector's efficiency calibration for in situ measurements can be challenging and is often less accurate than the laboratory-based setup. In this case, using Monte Carlo based numerical simulations of the detector's response provides a good alternative for efficiency calibration (Cinelli et al., 2016; Gutiérrez-Villanueva et al., 2008; Varley et al., 2016).

In this study, we used a portable gamma spectrometry setup for recording gamma spectra on a regular grid. A least-squares optimization based inversion method was used to deconvolute the contribution of nearby grids. Two different radionuclides i.e., calibration source pads of Ba-133 and Cs-137, were used as a radioactive source in this study. The efficiency of the gamma detector was calculated using Monte Carlo N-Particle (MCNP) transport code (Goorley et al., 2013).

2. Material and Methods

2.1. Inversion method

Assuming the gamma-radiation of a specific energy emitted in the decay of a radionuclide is coming from n number of grids in a measurement location, then the recorded gamma-ray count rate (N_j) in grid j can be expressed as,

$$N_j = \sum_{i=1}^{n} \varepsilon_{i,j} P_\gamma A_i \qquad (1)$$

Here, i is the grid number, A_i is the activity of a radioactive nuclide in grid i (Bq), P_γ is the emission probability of the photon energy of a radionuclide and $\varepsilon_{i,j}$ is the detection efficiency of the gamma detector located in grid j, for the considered photon energy in grid i (assuming that the radioactivity is homogeneously distributed over grid i). In equation (1), all the factors are known except the activity (A_i) and equation (1) can be written in a matrix form,

$$\begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1n} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nn} \end{bmatrix} \cdot P_\gamma \cdot \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{bmatrix} = \begin{bmatrix} N_1 \\ N_2 \\ \vdots \\ N_n \end{bmatrix} \qquad (2)$$

Equation (2) is a set of linear equations, which can be written as:

$$N = GA \qquad (3)$$

The efficiency matrix G is the convolution of ε and P_γ. In theory, the activity can be obtained by solving equation (3) for A. However, this problem can be ill posed due to observation errors. Therefore, G is not invertible which then makes it impossible to compute $A = NG^{-1}$, and an ordinary inversion can produce negative A (Ogawa et al., 2017). We used a least-square optimization approach to solve equation (3).

In this study, we used an algorithm to solve this linear least-squares problem (equation (3)) with bounds on the independent variable i.e., the activity of a radionuclide (A). As the radioactivity can not be negative, so the bound was zero to positive infinity. A function ("optimize.lsq_linear") of the *Scipy* module in the python programming language (Virtanen et al., 2020) was used for this purpose. In this function, the Trust Region Reflective algorithm ("trf") (Branch et al., 1999) adapted for a linear least-squares problem was used as a minimization method. In our calculation, the G matrix is a sparse matrix with a larger value in the diagonal and a few low magnitude values (≈0, relative to diagonal values) in the off-diagonal locations. An iterative algorithm for sparse least-squares problems is the LSMR algorithm (Fong and Saunders, 2011), and it is suitable for problems with sparse and large Jacobian matrices. Therefore, the LSMR algorithm was used as a solver to find a solution to the problem. For a given m-by-n design matrix G and a target vector N with m elements, this function solves the following optimization problem:

$$min \;\; 0.5||GA - N||^2 \; subject \; to \;\; lb \leq A \geq ub \qquad (4)$$

Where lb and ub are the lower and upper bounds for A, respectively. This optimization problem is convex; hence a global minimum can be found if iterations have converged.

2.2. Measurement setup

A 38.1 mm × 38.1 mm cylindrical $LaBr_3(Ce)$ detector (Model 38S38/1.5/HV), produced by Saint-Gobain (Saint-Gobain, 2020), was used in this study. The resolution of the detector is ≤ 3.5% at 662 keV. Figure 1shows the measurement setup with a portable radioisotope identification system. In this system, the $LaBr_3(Ce)$ detector was placed in a plastic pipe. This pipe also helped to control the vertical position of the detector. The diameter of the plastic pipe was 5 cm, and the detector was positioned 25 cm above the bottom of the pipe. A portable radio-isotope identification system, named SAM 940 (BNC, 2007), was used to operate the detector.

In this study, 25 radioactive source pads were used. Each pad was 63 cm in length and 59.4 cm in width. These surface radioactive sources were a mixture of two radionuclides (Ba-137 and Cs-137) that was uniformly distributed over each pad. These radioactive source pads were spread over a concrete slab. First, a measurement was taken using all 25 pads (5 by 5) and positioning the detector in the middle of it. This measurement was used to verify the numerical model of the detector setup used in the Monte Carlo simulation. For this purpose, the experimental full energy peak efficiency (FEPE) of 356 keV (Ba-137) and 662 keV (Cs-137) photon energy was calculated.

For inversion analysis, 13 out of 25 pads were taken out randomly (Figure 1). Hence, there were 12 grids (each grid is 63 cm × 59.4 cm area) with radioactive sources and 13 without any sources. Measurements were taken on each grid by positioning the detector set up in the middle of each grid. The measurement time was 5 minutes for each spectrum to attain a reasonable counting statistic. The measured spectra were analyzed later, and the net peak count rate for the 365 and 662 keV gamma line of Ba-133 and Cs-137, respectively, was calculated. The location and the ROI of the gamma lines were estimated using the MultiSpect Analysis software (MultiSpect, 2018) of Kromek Limited.

2.3. Detector efficiency calculation

Monte Carlo N-Particle (MCNP 6.1 cloud version) (Goorley et al., 2013) was used in this study for simulating the detector response. A numerical model of the detector was previously optimized and verified using different point, extended and surface sources (Hasan et al., 2021). This verified detector model was used in this study. The numerical model of the detector setup was further verified in this study using the experimental FEPEs of 356 and 661 keV photon energy using 25 radioactive source pads as described in section 2.2.

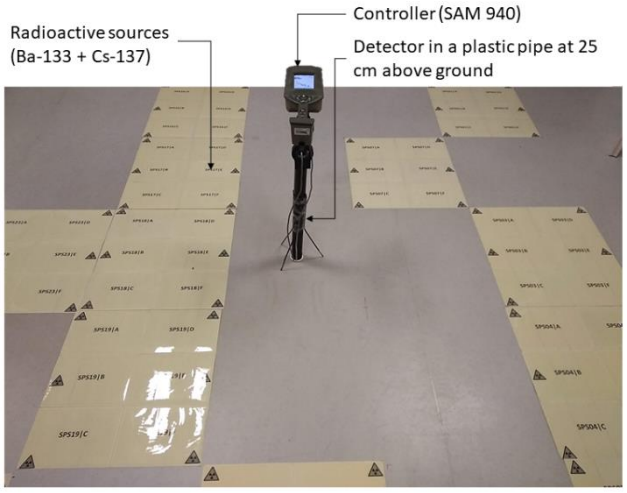


Figure 1- Measurements of gamma spectra on a grid using a portable system.

Detector's full energy peak efficiency (FEPE) of 356 and 662 keV photon energy was calculated using the same geometry described in section 2.2 and Figure 1. In the simulation, the photon source was uniformly distributed on a concrete surface of a particular grid (63 cm × 59.4 cm). The detector efficiency was simulated for 25 grids where the detector was in the first grid in all instances. Therefore, we create a library of detector efficiencies as a function of grid distance. The density of the concrete was 2.4 g/cm^3. The atoms per molecule of the different elements in the simulated concrete slab were H-0.169, C-0.00142, O-0.56252, Na-0.01184, Mg-0.0014, Al-0.02135, Si-0.20411, K-0.00566, Ca-0.01867 and Fe-0.00426.

3. Results

In the Monte Carlo simulation, the numerical model of the detector setup performed well in calculating the FEPE of 356 and 662 keV photon energy using all the 25 radioactive pads. The ratio of simulated to experimental FEPE were 1.06±0.05 and 1.02±0.02 for 356 and 662 keV photon energy, respectively. The uncertainty expansion of the experimental FEPEs covered the simulated ones. Hence, the detector model can be reliably used for the efficiency calculation of similar setups.

Figure 2 shows the results of the inversion analysis of the Cs-137 activity for the 12 slab distributed geometry. The measured net peak count of 662 keV photon energy was obtained in every grid (Figure 2 (b)) though only half of the grid had radioactive sources (Figure 2 (a)). This is because of the expected contribution of radioactive sources in nearby grids. The proposed inversion method was applied to calculate the activity of Cs-137 in each grid from the measured net peak count rate. Figure 2 (c) shows the calculated activity of Cs-137 using the inversion method. The inversion method performed well in calculating the actual radioactivity distributions in the measured area. A good agreement was observed between calculated and known radioactivity (Figure 2 (a) and (c)). In most of the grids, the absolute difference between calculated and known activity was low compared to the known radioactivity (Figure 2 (d)).

Inversion analysis results of Ba-133 are shown in Figure 3. It was observed by comparing Figure 3(a) and Figure 3(c) that the calculated activity was found in six grids (although relatively low) where radioactive source was not existing during the measurements i.e., the known activity was zero in such grids. A relatively higher residual error was also obtained for the Ba-133 than for the Cs-137 case. Due to the lower activity of Ba-133 and the low emission probability of its 356 keV gamma line, the measured net peak count rate had more than 10% uncertainty. The measured peak count rate map of Ba-133 (Figure 3 (b)) was expected to be like the one obtained for Cs-137 (Figure 2(b)) as the activity distribution in the grids was the same in both cases.

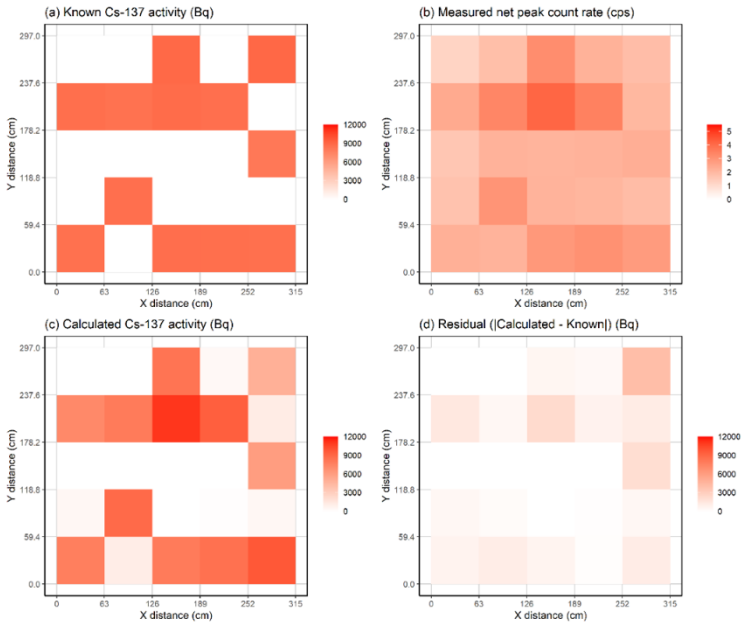


Figure 2 – (a) Known radioactivity of Cs-137 in different grids. (b) Measured net peak count rate of 662 keV photon energy of Cs-137. (c) Calculated radioactivity of Cs-137 from measured net peak count using the inversions method. (d) The absolute difference between calculated and known activity.

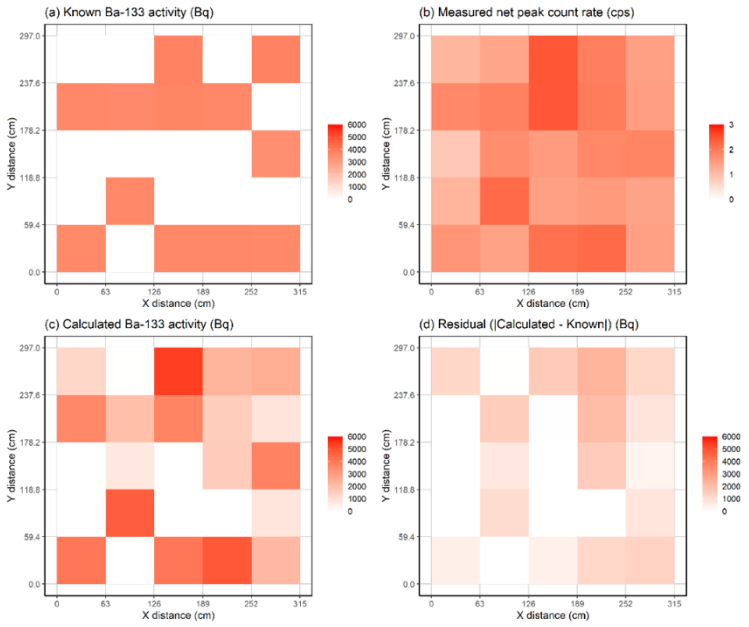


Figure 3 – (a) Known radioactivity of Ba-133 in different grids. (b) Measured net peak count rate of 356 keV photon energy of Ba-133. (c) Calculated radioactivity of Ba-133 from measured net peak count using the inversions method. (d) The absolute difference between calculated and known activity.

4. Discussion and Conclusions

The proposed inversion method performed well in recovering the spatial radioactivity distribution in a contaminated area. Results indicate that it is possible to disentangle the contribution of nearby grids and obtain a better estimation of the spatial activity distribution. Moreover, our approach does not require any heavy shielding of the detector, and the system used in this study was a portable one that can be moved easily.

Therefore, this type of detector setup and inversion method can be used to characterize a contaminated site using less time and other resources. However, the quality of the inversion depends on the measured data, similarly as for any other radiological measurement method. The measurement time and spatial coverage should be carefully selected depending on the target radionuclide. A longer measurement time can increase the quality of the measured count rate, which could in turn increase the quality of the inversion results. Uncertainty analysis of the calculated activity was not reported in this study. In a follow-up experiment, we will use uncertainty analysis in such calculation, and this method will be applied to a contaminated site as well.

References

Androulakaki, E.G., Kokkoris, M., Tsabaris, C., Eleftheriou, G., Patiris, D.L., Pappa, F.K., Vlastou, R., 2016. In situ γ-ray spectrometry in the marine environment using full spectrum analysis for natural radionuclides. Appl. Radiat. Isot. 114, 76–86. https://doi.org/10.1016/j.apradiso.2016.05.008

Berens, A.S., 2016. The use of in situ gamma radiation measurements as a method of determining radon potential in urban environments.

BNC, 2007. Instruction Manual Model 940 SAM Eagle ™.

Branch, M.A., Coleman, T.F., Li, Y., 1999. A Subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. SIAM J. Sci. Comput. 21, 1–23. https://doi.org/10.1137/S1064827595289108

Chirosca, A., Suvaila, R., Sima, O., 2013. Monte Carlo simulation by GEANT 4 and GESPECOR of in situ gamma-ray spectrometry measurements. Appl. Radiat. Isot. 81, 87–91. https://doi.org/10.1016/j.apradiso.2013.03.015

Cinelli, G., Tositti, L., Mostacci, D., Baré, J., 2016. Calibration with MCNP of NaI detector for the determination of natural radioactivity levels in the field. J. Environ. Radioact. 155–156, 31–37. https://doi.org/10.1016/j.jenvrad.2016.02.009

Druker, E., 2017. Airborne gamma-ray spectrometry data processing using 1.5D inversion. J. Environ. Radioact. 177, 13–23. https://doi.org/10.1016/j.jenvrad.2017.05.006

Duarte, P., Mateus, A., Paiva, I., Trindade, R., Santos, P., 2011. Usefulness of systematic in situ gamma-ray surveys in the radiometric characterization of natural systems with poorly contrasting geological features (examples from NE of Portugal). Appl. Radiat. Isot. 69, 463–474. https://doi.org/10.1016/J.APRADISO.2010.10.002

Fong, D.C.L., Saunders, M., 2011. LSMR: An iterative algorithm for sparse least-squares problems. SIAM J. Sci. Comput. 33, 2950–2971. https://doi.org/10.1137/10079687X

Goorley, J., James, M., Booth, T., Brown, F., Bull, J., Cox, L., Durkee, J., Elson, J., Fensin, M., Forster, R., Hendricks, J., Hughes, H., Johns, R., Kiedrowski, B., Mashnik, S., 2013. MCNP6 User's Manual, Version 1.0, LA-CP-13-00634, Los Alamos National Laboratory.

Guérin, G., Mercier, N., 2012. Field gamma spectrometry, monte carlo simulations and potential of non-invasive measurements. Geochronometria 39, 40–47. https://doi.org/10.2478/s13386-011-0056-z

Gutiérrez-Villanueva, J.L., Martín-Martín, A., Peña, V., Iniguez, M.P., de Celis, B., 2008. Calibration of a portable HPGe detector using MCNP code for the determination of 137Cs in soils. J. Environ. Radioact. 99, 1520–1524. https://doi.org/10.1016/j.jenvrad.2007.12.016

Hasan, M.M., Vidmar, T., Rutten, J., Verheyen, L., Camps, J., Huysmans, M., 2021. Optimization and validation of a LaBr 3 (Ce) detector model for use in Monte Carlo simulations. Appl. Radiat. Isot. Under review.

IAEA, 2017. In Situ Analytical Characterization of Contaminated Sites Using Nuclear Spectrometry Techniques.

Mikami, S., Maeyama, T., Hoshide, Y., Sakamoto, R., Sato, S., Okuda, N., Demongeot, S., Gurriaran, R., Uwamino, Y., Kato, H., Fujiwara, M., Sato, T., Takemiya, H., Saito, K., 2015. Spatial distributions of radionuclides deposited onto ground soil around the Fukushima Dai-ichi Nuclear Power Plant and their temporal change until December 2012. J. Environ. Radioact. 139, 320–343. https://doi.org/10.1016/j.jenvrad.2014.09.010

MultiSpect, 2018. MultiSpect Analysis.

Ogawa, H., Minami, K., Kawamoto, T., Kanai, R., Ishikawa, K., Kamimura, R., 2017. Inversion analysis on vertical radiocesium distribution in pond sediment from γ-ray count measurement. J. Environ. Radioact. 175–176, 158–163. https://doi.org/10.1016/j.jenvrad.2017.05.011

Rostron, P.D., Heathcote, J.A., Ramsey, M.H., 2014. Comparison between in situ and ex situ gamma measurements on land areas within a decommissioning nuclear site: A case study at Dounreay. J. Radiol. Prot. 34, 495–508. https://doi.org/10.1088/0952-4746/34/3/495

Saint-Gobain, 2020. Standard Scintillation Product List | Saint-Gobain Crystals [WWW Document]. URL https://www.crystals.saint-gobain.com/products/radiation-detection-products/standard-scintillation-product-list (accessed 11.8.20).

Tyler, A.N., 2008. In situ and airborne gamma-ray spectrometry. Radioact. Environ. 11, 407–448. https://doi.org/10.1016/S1569-4860(07)11013-5

Tyler, A.N., Sanderson, D.C.W., Scott, E.M., Allyson, J.D., 1996. Accounting for spatial variability and fields of view in environmental gamma ray spectrometry. J. Environ. Radioact. 33, 213–235. https://doi.org/10.1016/0265-931X(95)00097-T

Varley, A., Tyler, A., Dowdall, M., Bondar, Y., Zabrotski, V., 2017. An in situ method for the high resolution mapping of 137Cs and estimation of vertical depth penetration in a highly contaminated environment. Sci. Total Environ. 605–606, 957–966. https://doi.org/10.1016/j.scitotenv.2017.06.067

Varley, A., Tyler, A., Smith, L., Dale, P., Davies, M., 2016. Mapping the spatial distribution and activity of 226Ra at legacy sites through Machine Learning interpretation of gamma-ray spectrometry data. Sci. Total Environ. 545–546, 654–661. https://doi.org/10.1016/j.scitotenv.2015.10.112

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., et al., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods. https://doi.org/10.1038/s41592-019-0686-2

Xhixha, G., 2012. Advanced gamma-ray spectrometry for environmental radioactivity monitoring 147.

Zhang, Yingying, Li, C., Liu, D., Zhang, Ying, Liu, Y., 2015. Monte Carlo simulation of a NaI(Tl) detector for in situ radioactivity measurements in the marine environment. Appl. Radiat. Isot. 98, 44–48. https://doi.org/10.1016/j.apradiso.2015.01.009

Zhukouski, A., Anshakou, O., Kutsen, S., 2018. In situ measurement of radioactive contamination of bottom sediments. Appl. Radiat. Isot. 139, 114–120. https://doi.org/10.1016/j.apradiso.2018.04.036

ACCOUNTING FOR MODEL ERRORS USING DEEP NEURAL NETWORKS WITHIN A MARKOV CHAIN MONTE CARLO INVERSION FRAMEWORK

Shiran Levy (1)* - Jürg Hunziker (2) - Eric Laloy (3) - James Irving (1) - Niklas Linde (1)

University of Laussane, Institute of Earth Sciences (1) - Electromagnetic Geoservices Asa (2) - Institute for Environment, Health and Safety, Belgian Nuclear Research Centre (3)
** Corresponding author: she.run.levy@gmail.com*

Abstract

Most geophysical inverse problems are non-linear and rely upon forward solvers based on discretized differential operators and/or simplified representations of physical processes. As a result, model errors are inevitable and they become particularly acute when using Markov chain Monte Carlo (MCMC) inverse methods, in which large number of forward simulations are involved, indicating the need for simplified forward solvers. Because of inherent complexity of addressing model errors, they are often simply ignored or accounted for using Gaussian approximations, which can lead to strongly biased and over-confident inversion results. One way to deal with these errors might be to learn a lower-dimensional probabilistic representation of the discrepancy between high-fidelity and low-fidelity forward solvers, and then to use this parameterization to probabilistically infer the model error, along with the physical property field, using the computationally efficient low-fidelity forward solver. Here, we capitalize on recent advancements in deep learning and combine a MCMC inversion algorithm with a convolutional neural network of the spatial generative adversarial network (SGAN) type, the latter of which has been trained to represent the model error for the problem at hand. At each MCMC step, the low-fidelity forward response is corrected using an estimated model-error realization. The SGAN transformation connecting the model-error space with a low-dimensional latent space is formed by a series of convolution operations and offers a substantial reduction of the number of parameters describing the model errors. The network was successfully trained on images describing discrepancies between curved-ray (high fidelity) and straight-ray (low fidelity) simulations of crosshole ground-penetrating radar (GPR) travel times, and will soon be trained with finite difference full-waveform simulations replacing the curved-ray modeling. Our preliminary results demonstrate that the SGAN can effectively represent the model error for this problem, which paves the way for fast and unbiased MCMC inference. Our methodology may perform better than existing approaches when the multi-Gaussian assumption of model errors is inappropriate. Improvements will be assessed against such state-of-the art approaches.

HANDLING NON-STATIONARITY IN MULTIPLE-POINT STATISTIC SIMULATION WITH A HIERARCHICAL APPROACH

Alessandro Comunian (1)* - Edoardo Consonni (1) - Chiara Zuffetti (1) - Riccardo Bersezio (1) - Mauro Giudici (1)

Università degli Studi di Milano, Dipartimento di Scienze della Terra (1)
** Corresponding author: alessandro.comunian@unimi.it*

Abstract

Many approaches have been proposed to tackle the challenges of non-stationarity in multiple-point statistics (MPS) simulations, including the usage of "auxiliary variables" (AVs) maps. However, obtaining the additional information required to draw these AV maps can be challenging, and in many cases these maps are drawn with subjective ad hoc procedures. Recently, some authors proposed a hierarchical simulation procedure based on a tree-like frame of binary sequential indicator simulations (SIS), with a simulation tree-like frame based on the textural hierarchy of facies. In this work a similar approach is proposed by using MPS instead of SIS; in addition, this work explores the possibility of using a different tree-like frame based on stratigraphic hierarchy and relative chronology.

The proposed approach is demonstrated by using outcrops of alluvial sediments to reconstruct a three-dimensional (3D) volume. First, the outcrops are analyzed to extract a tree-like frame describing the hierarchy of facies. Then, the frame is used to decompose the outcrop into multiple bi-dimensional (2D) training images (TIs), each of which represents the spatial distribution of a simplified interpretation of the outcrop, based on the given hierarchy of facies. Depending on the criteria used to build the tree-like frame, these 2D TIs are composed of a relatively low number of facies; it is therefore straightforward to use a sequence of 2D conditional simulations (s2Dcd approach) to build 3D TIs for each branch of the frame. Finally, the obtained 3D TIs are used to perform a sequence of MPS simulations, nested accordingly to the aforementioned tree-like frame, resulting in a 3D reconstruction of the spatial distribution of the alluvial sediments considered.

On 2D test cases, the results obtained with the proposed approach are comparable with the results obtained by handling non-stationarity using AVs, with the advantage that the proposed approach does not require an AV map. In addition, the decomposition of the simulation problem into smaller groups of facies, allowed to have more control on the low-level reconstructions made with the s2Dcd approach to obtain the 3D TIs, and consequently to improve the final 3D reconstruction.

In conclusion, with the additional effort required to conceptualize a hierarchy of facies, the proposed approach appears as a reliable alternative to obtain non-stationary MPS simulations without the need of additional information, as for example the one required by the use of AVs.

APPLIED MULTI-POINT GEOSTATISTICS FOR TAILINGS CHARACTERIZATION AT KING RIVER DELTA, AUSTRALIA

Sangga Rima Roman Selia (1)* - Raimon Tolosana-delgado (1) - Sibele C. Nascimento (2) - Anita Parbhakar-fox (3) - K. Gerald Van Den Boogaart (1) - Helmut Schaeben (4)

Helmholtz Institute Freiberg for Resource Technology (1) - University of Tasmania (2) - University of Queensland (3) - Tu Bergakademie Freiberg (4)

** Corresponding author: s.selia@hzdr.de*

Abstract

The King River received around 97 million tonnes of pyritic tailings from the mining activities of the Mount Lyell mine within the period of 1916 - 1994. The tailings was dumped into the river's main tributary, the Queen River, and transported downstream into the King River delta. This has been creating environmental problems such as acid drainage which is damaging the ecology of the delta. Within the framework of comprehensive efforts to understand the status of the system and eventually undergo appropriate ecological remediation measures, 3D characterization of the deltaic deposit is required to resolve its internal structure that provides the input for environmental and economic assessment of such actions.

We applied multi-point geostatistics (MPS) to characterize the deposit. Stratigraphic forward modelling is used to generate some training images for the MPS. Conditional data obtained from the field and the training images are fed into the MPS to produce a 3D spatial distribution of relevant minerals in the deposit. This approach can provide high resolution variability of the deposit and is very useful alongside geophysical methods that face challenges in capturing tailings internal structure.

ENVIRONMENTAL RISK ASSESSMENT OF CHINA'S OBOR (BRI) PROJECT IN KAZAKHSTAN – AN EVALUATION OF THE APPEARANCE AND DISAPPEARANCE OF OASIS FARMLAND

Kazuki Seno (1,2) - Christopher McCarthy (3) - Maira Kussainova (4) - Sabir Nurtazin (5) - Buho Hoshino (2)* - James Banfill (6)

Hokkaido University, Graduate School of Environmental Science (1) - Rakuno Gakuen University, Department of Environmental Sciences, College of Agriculture, Food and Environment Sciences (2) - Johns Hopkins University, Zanvyl Krieger School of Arts &sciences (3) - Kazakh National Agrarian University, Department of Soil Science and Agricultural Chemistry (4) - al-Farabi Kazakh National University, Faculty of Biology and Biotechnology (5) - Kyungnam University, Institute of Far Eastern Studies (6)
** Corresponding author: aosier@rakuno.ac.jp*

Abstract

Central Asia is an arid region highly vulnerable to water scarcity. The region's oasis agriculture is one of the most vulnerable landscapes to climate change and human activities. Previous research of land and water use in the region has focused on improving water-use efficiency, soil management and identify. This has a significant impact on the availability of oasis farm production. These include improving water use efficiency and changing the cropping patterns that have a high potential to decrease the exposure and sensitivity of rural communities to the risk of the disappearance of oases farmland. In addition, changes in land use, such as the afforestation of degraded croplands, and the introduction of resource-smart cultivation practices, such as conservation agriculture, may strengthen the capacity of farmers and institutions to respond to land degradation and soil salinization challenges. However, despite the rapid expansion of the outlook for oasis farmland and the imminent risk of the disappearance of oases farms, vulnerabilities to farmland abandonment related to China's One Belt One Road (OBOR or Belt and Road Initiative, BRI) project are rarely considered. The study area of the Zharkent (Panfilov) Region, located on the border between Kazakhstan and China, is a semi-arid area and that has cultivated corn by irrigation from the Ili River and the Usek River for many years. Therefore, there are many abandoned agricultural lands as a result of salinization. However, this area is the start point of the railway and highway network of China's OBOR project to Europe, and corn production and exports are expected to flourish. In this study, we focus on the vulnerability of oasis agriculture and extract changes in agricultural land for about 30 years from 1989 to the present using Landsat series and Sentinel series combined with RGB color combined visualization. The results show that agricultural land has disappeared or desertified at the Ili river basin and at the foot of the Zhongar-Alatau Mountain and that there are several years of fallow even in areas where agriculture is active and show that even areas of currently active agriculture have experienced periods of fallow in the past three decades. This study using the Zharkent region in the irrigated alluvial fan of Zhongar-Alatau Mountain of eastern Kazakhstan as an example, we classify the farm field changing based on Landsat TM and Sentilel-2 satellite imagery and identify the vulnerability of oases farmland.

Keywords: Oases farmland changes, OBOR (BRI), Kazakhstan, Remote sensing

1. Introduction

Located in Central Asia, Kazakhstan is characterized as a semi-arid region, which includes dry steppe land in the south. Agriculture carried out in this area is typically oasis farmland in which water is taken from local rivers and used for irrigation. During the former Soviet Union, irrigation projects were widely carried out to expand agricultural land, and large-scale irrigation projects were created in several areas. However, many of these areas were abandoned since the collapse of the Soviet regime in 1991. In recent years, agricultural reforms have been carried out in Kazakhstan, and the privatization and fragmentation of agricultural land has progressed due to the free transfer of agricultural land use rights. Especially in southern Kazakhstan, most farmers are small and medium-sized farmers who manage their farmland on a family basis (Hamidov et al., 2016). Kazakhstan is an important global producer and exporter of high-quality wheat. Average annual production is about 13 million tons, and between 2 and 8 million tons is exported annually, mainly to destinations in Europe (including Russia and Ukraine), northern Africa, and China, but output is highly dependent on weather. Southern Kazakhstan also produces around 2 million tons of barley, corn, oats, rice and cotton. Historically, Kazakhstan grain production suffers from serious drought two out of every five crop seasons. As a result, yield and production are marked by frequent and sharp year-to-year fluctuations (USDA, 2009; FAO, 2017).

This study evaluates the appearance and disappearance of oasis farmland. An oasis is an area made fertile by a source of freshwater in an otherwise dry and arid region. The Usek River Oases are irrigated by natural snow meltwater from the Zhongar-Altai Mountains. Irrigated corn farmland is widespread in between the Usek River and Ili River oases. However, the fluctuations in oasis agriculture are so severe that farmland is being repeatedly discarded and reclaimed, often due to soil salt accumulation or urbanization (Thevs, et al., 2017; Pueppke, et al., 2018).

Downward infiltration of soil water is unlikely to occur in dry areas where precipitation exceeds evaporation. Also, when the soil is moistened, the transfer of soluble salts occurs. Evapotranspiration of the ground surface causes salts to move to the surface layer and remain in the soil layer, resulting in salt accumulation (Sakai, et al., 2020). Large scale irrigation on farmland with inadequate drainage facilities provide a constant supply of water within or on the soil surface (Anna and Tatiana, 2007). Also, in the same area, the degree of progression varies depending on the soil quality. Soil containing a large number of fine particles, such as clay, is said to have a high-water retention capacity, so salts are likely to accumulate. On the other hand, sandy soil is said to be less prone to salting due to its harsh impression (Yoda et al., 2012). Removal of soluble salts from salinized soil is possible by leaching (Nurtazin et al., 2019).

In 2013, China proposed two economic initiatives that would form the basis of the "One Belt One Road (OBOR)". China has funded the creation of new ports throughout Asia and Africa and infrastructure, such as railways, highways, and gas / crude oil pipelines. It is unpredictable to what extent the changes in the natural and social environments will occur due to rapid large-scale development. In addition, OBOR aims to promote not only economic integration, but also closer political and military ties. For that reason, it is attracting attention from various fields (Hong, 2016; Foggin, 2018; Martin, 2019), especially within the food-water-energy nexus. Central Asia, including Kazakhstan, has many arid and semi-arid areas, making it difficult to secure water and food. Troy et al. (2020, 2017) pointed out that China may seek to dominate the region through the control of food and water resources.

Vulnerable oasis agriculture is greatly affected not only by the natural environment but also by the social environment. In recent years, the changes in the environment that have occurred are not limited to those caused by climate change, but also include changes in the natural environment due to OBOR, changes in supply and demand due to the development of transportation networks, and changes in population and

agricultural land area. In this study, satellite analysis, field surveys, and interviews were used to clarify the vulnerability of oasis agriculture and the rapid change caused by OBOR. Provide a baseline for monitoring future environmental changes (Troy et al., 2020).

2. Material and Methods

2.1. The study area

The study area was the Zharkent region in the Panfilov district of the Republic of Kazakhstan. The Panfilov District is an administrative district located in the southeastern part of Kazakhstan with an area of 1,058,252 ha and a population of about 130,000. The region shares a border with China and is the starting point of OBOR leading to Europe.

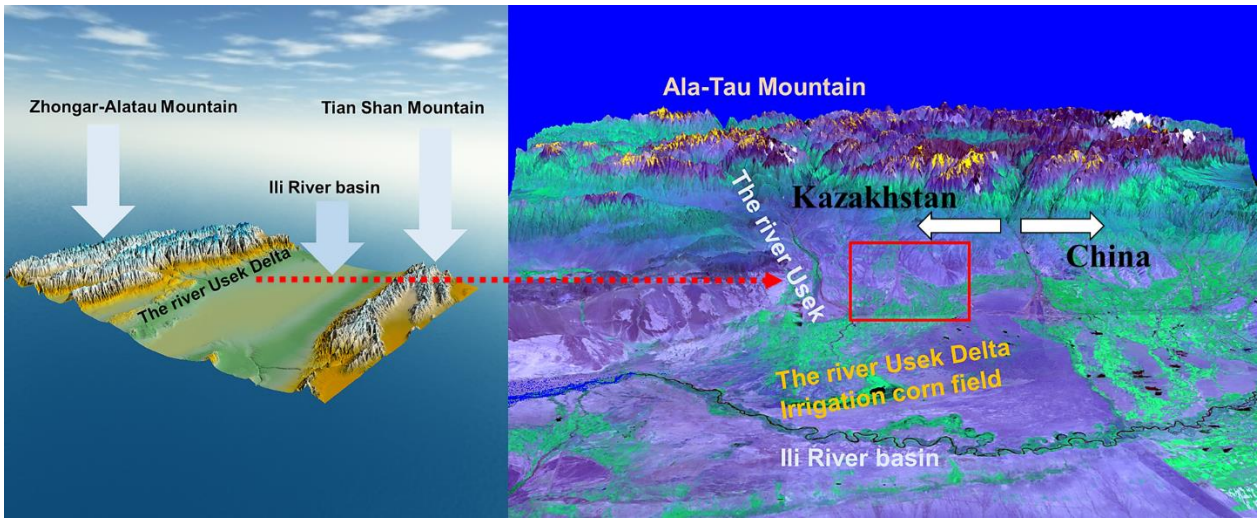


Figure 1 - The study area in the Usek River delta in the southeastern part of Kazakhstan.

2.2. Field survey

The first field survey was conducted in September 2019, including an irrigation water survey, river water quality survey, soil survey and agricultural land survey. A second field survey was conducted in September 2020 and November 2020. An interview survey of local farmers regarding the use of agricultural land and changes due to OBOR was also conducted. Twenty-five responses were received in the first round and 26 responses in the second round.

2.3. Satellite image analysis

Land use change was determined using satellite data for 30 years from 1989 to 2019. The data used are as follows: Using Landsat TM (from USGS): August 22, 1989; August 11 and October 30, 1994; September 6 and October 31, 2006; August 10, 2008; September 11 and October 22, 2011. And used Landsat ETM+ (from USGS): August 26 and October 30, 1999. Landsat 8 (from USGS): September 6 and November 9, 2015; August 22 and October 19, 2019.

Calculation of time series NDVI was based on time series Landsat and Sentinel 2 satellite data. NDVI (Normalized Difference Vegetation Index) was used to extract the vegetation performance of the farmland. NDVI is an index showing the presence or absence of vegetation and activity by utilizing the difference in reflectance depending on the substance, and is calculated by the following formula.

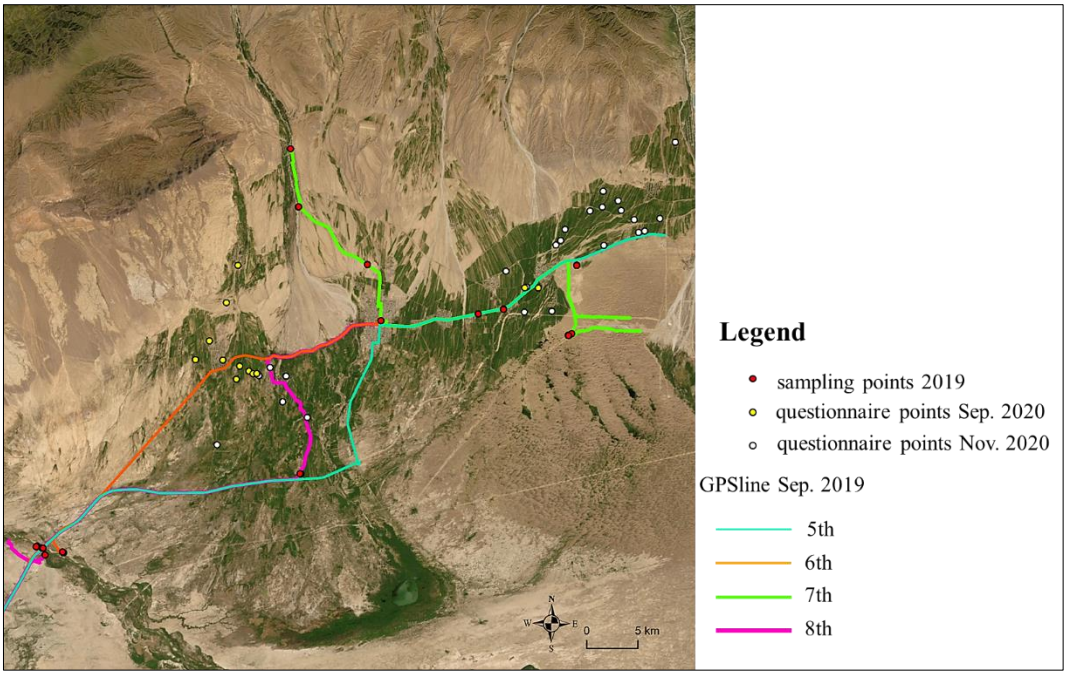


Figure 2 - Distribution of farmers who answered the questionnaire.

NDVI= (NIR-R)/ (NIR+R)

Where, NIR indicates the reflectance in the near-infrared channel, and R in the visible red channel.

In semi-arid areas, the amount and range of vegetation changes greatly depending on the amount of precipitation. Therefore, it was necessary to distinguish between agricultural land and other natural vegetation. In this study, we utilized differences in vegetation before and after harvest, which is an artificial anthropogenic management practice. Two images, summer and autumn, were used per year, and the point where the negative change in the NDVI value was large was classified as agricultural land (Morteza et al., 2017).

3. Results

3.1. Image classification results

Analysis of NDVI time series Landsat imagery provides an overview of agricultural change from 1989 to 2019. The area of agricultural land was largest in 1989 (during the era of the Soviet Union), with nearly 34% of agricultural land abandoned by 1994 (Fig. 2, Fig. 3 and Table 1). After the collapse of the Soviet Union, the area of agricultural land has continued to decrease and began expanding with the OBOR in China. In addition, there was a difference between the change in the agricultural land area within the Panfilov district and the change in the extracted analysis range. According to local government statistics, the area of agricultural land, which had been decreasing since 1990, increased sharply from 2001 and peaking in 2005 across the entire Panfilov district. Investment in the agricultural sector from China was increasing even before the OBOR project officially started. However, such a tendency was not seen in the study area (i.e. the Zharkent region) during this time based on satellite imagery analyzed.

Table 1 - Estimation of agriculture farmland area based on Landsat satellite time series data.

Year	1989	1994	1999	2006	2011	2015	2019
Area (ha)	21663.4	14324.5	14308.7	12984.8	14259.5	15931.3	15385.9

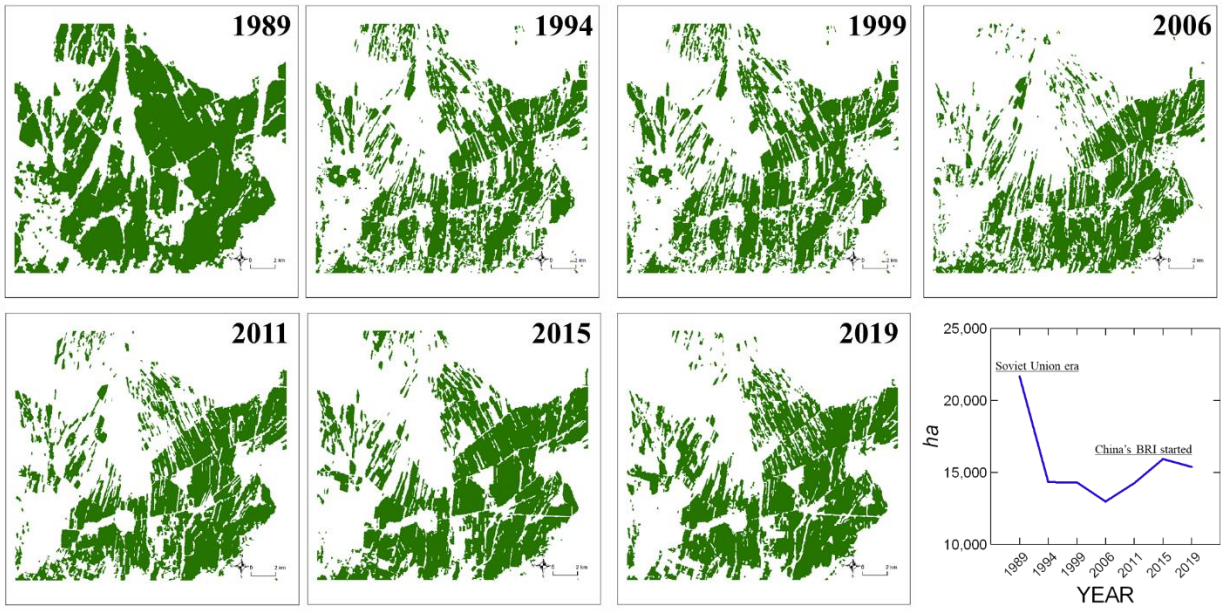


Figure 3 - Extraction of farmland area (green colors) based on Landsat time series data.

3.2. Change detection of Oases farmland using RGB color combination method

RGB image color composite method is useful to detect land use and land cover changes (Suriga et al., 2012; Pujiono et al., 2013). Figure 4, below, shows visual changes in agriculture land from 1989, during the former Soviet era, 2006, when the area of the entire district increased due to Chinese-funded investment in Kazakhstan agriculture, and in 2019, for which the colors were assigned to red (R), green (G), and blue (B), respectively. Agricultural land in all three years of the study is shown in white. In the Khorgas region, the majority of the farmland has remained in production, while in the southern Zharkent and Akaral, some land has also been used continuously. Most of the area in the center of the image is farmland that has been used again although at one time had stopped. On the other hand, in the northern part of Zharkent on the left side of the center of the image, land that has not been used since the collapse of the former Soviet Union, stands out (blue color).

Focusing on the northern region of Zharkent (Fig. 5), where change in land use was prominent, we classified the changes into three patterns using 7-year (1989, 1994, 1999, 2006, 2011, 2015 and 2019) NDVI agricultural land extraction images. The results show a clear division in the southern part where agricultural land is continuously used and the northern part where abandoned cultivated land is conspicuous. It is thought that the long distance from the town and lack of access to major roads is one of the main factors that make it difficult for continuous agriculture production.

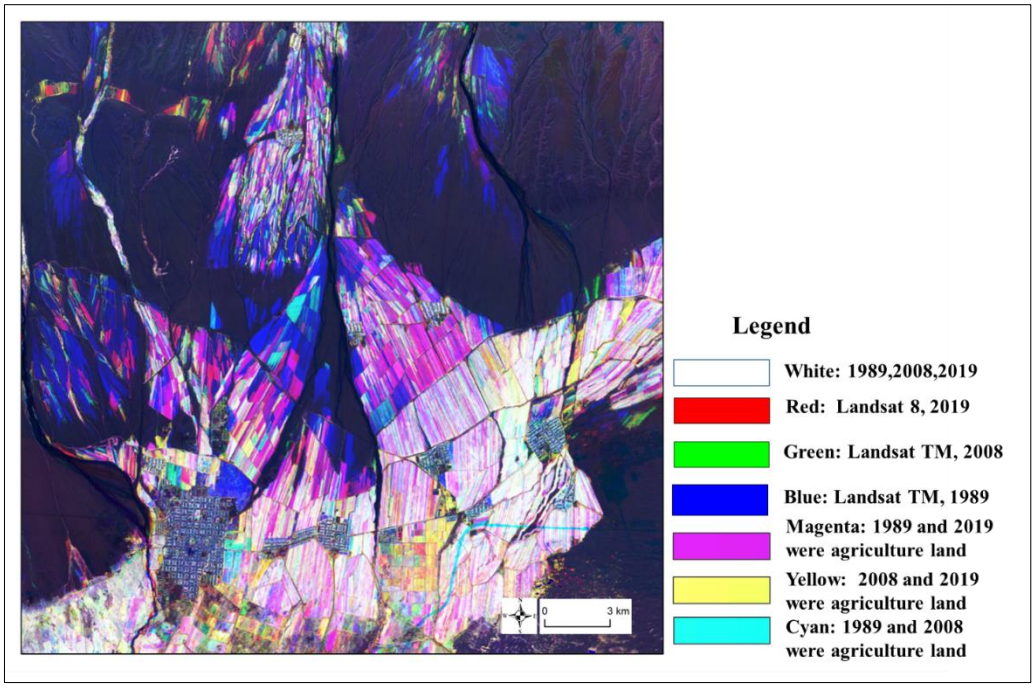


Figure 4 - Change detection of oases farmland using RGB color combination method (Magenta color is noticeable in the northern area, which indicates farmland that was abandoned around 2008).

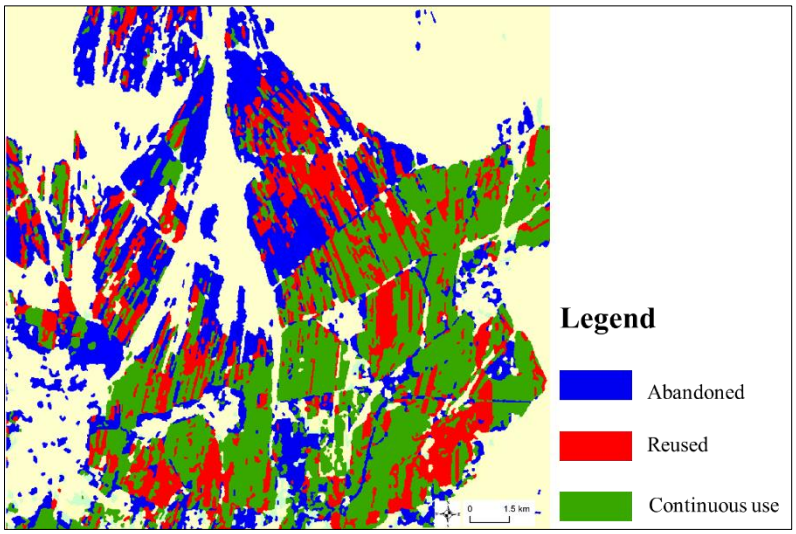


Figure 5 - Image classification result.

3.3. Results of questionnaire to local famers

3.3.1. About agricultural farm

All 51 farmers surveyed used land that was classified as collective farmland during the former Soviet Union era, and no farmer used newly reclaimed farmland. In addition, none of the farms were fallow during the Soviet era. The main crop is maize, and many farmers use alfalfa for crop rotation. Wheat was also cultivated during the former Soviet Union, but it is rarely cultivated today because it is not a suitable crop for this land. Some smallholders used to consume or trade crops with individuals. Many farmers sell their yield to agricultural organizations such as Zharkent starch plant, Asia-Agro-Food Ltd, and Treate Plant LLP in Today.

All the farmers surveyed used gutter irrigation developed during the former Soviet Union. Two large farms with an area of more than 1000 ha also used drip irrigation (see Fig. 6(a,b)).

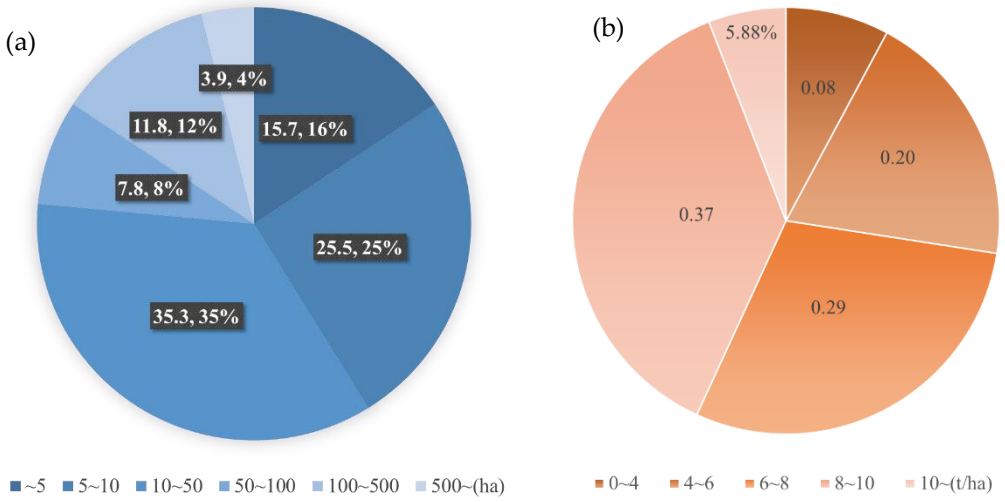


Figure 6 - (a) Percentage of the agriculture land area (ha) owned by farmers around Zharkent and (b) percentage of average crop yield of farmers around Zharkent region.

3.3.2. The Influence of OBOR

The first interview survey provided detailed answers regarding OBOR. Positive opinions were that the opening of roads increased opportunities for communication with other regions and expanded the range of crop transportation. In addition, the opening of the railway has the advantages for expanding transportation opportunities and increasing the profits of the agricultural organization, such as simplifying the introduction of fertilizer. On the other hand, farmers in areas far from the major transportation lines did not notice any change. Farmers who have farmland around the highway reported negative effects, such as flood damage due to poor drainage function and irrigation and groundwater division due to the division of farmland.

3.3.3. Problems and anxiety

In the second interview survey, the answers focused mainly on agricultural land use. None of the farmers felt that yields were reduced or the field was deteriorated due to salinization. Many farmers cited water and funding shortages as anxiety factors. While farmers desire to irrigate three times a season, due to the lack of water resources, this often can be done only once. In addition, there is a wait list for irrigation in early spring in order to share scarce water resources among multiple farmers. In general, the more time between sowing and irrigation, the lower the yield for the year. In addition, compost is mainly used in this area to ensure continuous use of agricultural land and yield. However, compost requires a large amount of chemical fertilizer and is expensive. There were many farmers who wanted to use compost but could not due to lack of funds (see Table 2).

Table 2 - Problems and anxiety from questionnaire samples.

Contents of problems and anxieties	Number of responses
Water shortage	6
Land degradation	2
Disadvantages of OBOR highways	6
Lack of funds	7

4. Discussion and Conclusions

4.1. Factors of abandonment of agriculture land

In northern Zharkent, there was a clear separation between land that was continuously used as agricultural land and land that was not. There may be some unsuitable conditions for farmland in the north. As an example, if the slope of the farmland is large, nutrients will flow off the farm and it will not be suitable for agriculture. One reason why there are many abandoned cultivated farms in the northern part of Zharkent may be because of the mountain range to the north and the large sloping terrain. If that is the case, it can be used again as agricultural land by scraping the ground surface and flattening it. From our result, it can be said that the abandoned agricultural land has a slope of less than 1% and is rather flat. Another possible reason is the difference in soil quality. Mountain ranges extend to the north of the Panfilov region, and the area around Zharkent is an alluvial fan. The soil quality in this area was shifting sand. This type of soil is formed by the influx of large amounts of sand from floods that occur with the melting of snow in the mountains. The upper part is coarse and the soil is renewed quickly. It is difficult to run long-term agriculture because of the young soil that is not conducive to growing. This is thought to be the reason why continuous agricultural land use is not possible in the northern region of Zharkent.

4.2. Reduced yield due to heterogeneous farmland

Among the farmers who visited the site, there was a large difference in yield within the field. Analysis of the farmland using high-resolution satellite images revealed that the NDVI values were not uniform. It is expected that the cause is the slope of the terrain that results in an uneven distribution of nutrients and water throughout the field. In addition, in the second interview survey, five farmers, 19%, stated that they did not know the condition of their fields (see Fig. 6). Many farmers wanted specialized knowledge in order to obtain as much yield as possible in a sustainable manner. We believe it is important for farmers to have an understanding of the factors influencing the quality of their farmland and identify this as an important area for improvement.

4.3. OBOR influence

Population and agricultural land area are increasing due to OBOR. Also, as can be seen from Fig. 7), agricultural output and agricultural income have increased sharply in recent years. For this reason, the demand for agriculture in the region is increasing, and it is expected that it will continue to increase in the future. At this point, fertilizer from China is now easily accessible and may be able to accommodate this change. However, farmland is limited, and all the land that was used for agriculture production during the Soviet era has been reclaimed. Therefore, it is considered that there is almost no land that can be developed for agriculture in the future. In recent years, the reclamation of agricultural land in the north has been remarkable, but there is a possibility that more agricultural land in the region will be reused. It can be said that it is indispensable for future development to seek a method that can be continuously used as agricultural land and to formulate a concrete solution in this area where the risk of salt accumulation is high and water resources are limited. Troy et al., (2020) find many of these rails and roads traverse important agricultural and water zones, creating undetermined risks and opportunities. Land use change was examined within a 10-km buffer around OBOR roads and rails from 2008 to 2018. Railways increased by 23% during this time, yet irrigated and rainfed agriculture decreased whilst urban areas markedly expanded. Contextual research identifies how Chinese policies may encourage agribusiness investment for food exports as possible disruptions to national and regional food supply. However, to date Central Asia provides <1% of Chinese agricultural imports. Evaluating infrastructure change is essential to understand OBOR impacts on environments and societies, with the food-water nexus a particular concern in Central Asia including Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan and Uzbekistan. Limited Chinese imports of Central Asian

agriculture suggests the region's food security will not be significantly altered by the Belt and Road Initiative (Troy et al., 2020).

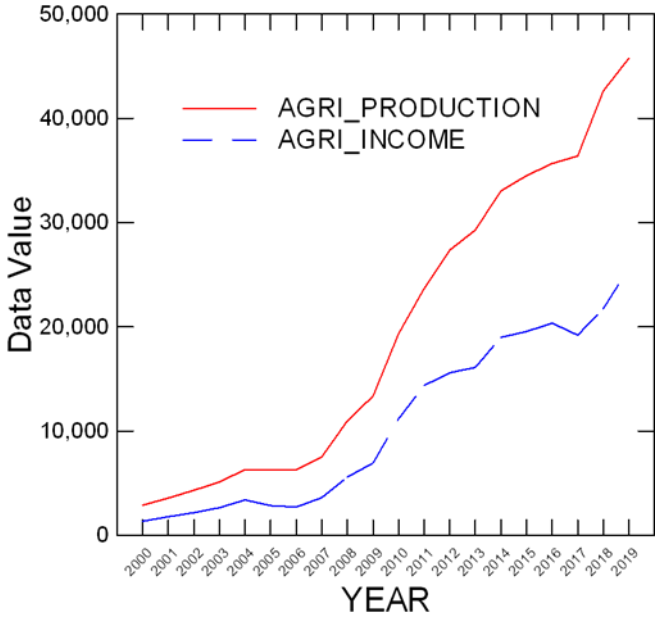


Figure 7 - The government statistics data for agriculture production and agriculture income in Panfilov district of Kazakhstan during 2000 to 2019.

4.4. Conclusions

It was found that agriculture in the area around Zharkent has been more influenced by the social environment than by the natural environment. OBOR has expanded its economic sphere more than ever before and is affecting the areas along the railway lines in various ways. Agricultural land abuse due to increased demand carries a high risk of making continuous use impossible in fragile oasis agriculture. In addition, the soil quality of fluid dunes, where the soil is renewed in a short period of time, hinders the prosperity of agriculture in the northern region located above the alluvial fan. Not only is there a problem with soil quality, but water resources are also limited, so the crops that can be cultivated are also limited. Understanding the conditions of farmland is essential to meet the ever-increasing demand in these circumstances. In addition, "lack of funds" and "water shortage" are very serious factors limiting agriculture production in this area. This could give China the opportunity to take control of the region through funding associated with OBOR. However, the lack of cooperation of local farmers and experts may prevent such a situation.

Author Contributions

Conceptualization, K.S. M.K., S.N. and B. H.; data curation, K.S., C.M. and B.H.; writing—original draft preparation, B.H.; writing—review and editing, K.S., B.H., J.B. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by JSPS KAKENHI Grant Numbers (JP) 19H04362 (Risk assessment of the regional impact of the China "One-Belt-One-Road" (OBOR) project).

Acknowledgments

The authors would like to thanks Dr. Kayo Matsui from Kyoto University and Prof. Satoru Hobara, Prof. Nobutake Nakatani and Dr. Mikoto Kaneko from the Rakuno Gakuen University was participated in the field survey as a key member of the OBOR project. Dr. Ruslan Salmurzauli at Al-Farabi Kazakh National University joint and support field work.

References

Anna and Tatiana Spitsyna (2007), Preliminary Sustainability Assessment of water resources management in the LLi-Balkhash of Central Asia, Master of Science Thesis Stockholm, 35-36. (ISSN 1402-7615) http://www.diva-portal.se/smash/get/diva2:411801/FULLTEXT01.pdf

Eko Pujiono, Doo-Ahn Kwak, Woo-Kyun Lee, Sulistyanto, So-Ra Kim et al., (2013), RGB-NDVI color composites for monitoring the change in mangrove area at the Maubesi Nature Reserve, Indonesia, Forest Science and Technology, 9(4) : 171-179. DOI: 10.1080/21580103.2013.842327

FAO (2017), The future of food and agriculture Trends and challenges, Food and Agriculture Organization of the United Nations. Rome. pp 80-131. http://www.fao.org/3/i6583e/i6583e.pdf

Hamidov, A., Helming, K. & Balla, D. (2016), Impact of agricultural land use in Central Asia: a review. Agron. Sustain. Dev. 36, 6. DOI: 10.1007/s13593-015-0337-7.

Hideaki Matsumoto, Hirotoshi Motoda (2012), Aluminum toxicity recovery processes in root apices. Possible association with oxidative stress. Plant Science, 185(1), DOI: 10.1016/j.plantsci.2011.07.019

Hong Chen Teo, Alex Mark Lechner, Grant W. Walton, Faith Ka Shun Chan, et al., (2019), Environmental Impacts of Infrastructure Development under the Belt and Road Initiative, Environments, 72. DOI: 10.3390/environments6060072

J. Marc Foggin (2018), Environmental Conservation in the Tibetan Plateau Region: Lessons for China's Belt and Road Initiative in the Mountains of Central Asia, Land, 7(2). DOI: 10.3390/land7020052

Kiyotsugu Yoda, Mohamed A.M. Abd Elbasit, Buho Hoshino, et al., (2012), Root System Development of Prosopis Seedlings under Different Soil Moisture Conditions, Journal of Arid Land Studies, 22(1).

Martin Russell (2019), Connectivity in Central Asia Reconnecting the Silk Road, EPRS | European Parliamentary Research Service: 637.891.

Morteza Sadeghi, Ebrahim Babaeian, Markus Tuller, & Scott B. Jones, (2017), The optical trapezoid model: A novel approach to remote sensing of soil moisture applied to Sentinel-2 and Landsat-8 observations, Remote Sensing of Environment 198, 52–68. DOI: 10.1016/j.rse.2017.05.041

Niels Thevs, Volker Beckmann, Altyn Akimalieva, Jan Felix Köbbing, Sabir Nurtazin, et al., (2017) Assessment of ecosystem services of the wetlands in the Ili River Delta, Kazakhstan, Environ Earth Sci; 76(30). DOI: 10.1007/s12665-016-6346-2

Pueppke SG, Zhang Q, Nurtazin ST. (2018), Irrigation in the Ili River Basin of Central Asia: From Ditches to Dams and Diversion. Water, 10(11). DOI: 10.3390/w10111650

Sabir Nurtazin, Niels Thevs, Margulan Iklasov, Norman Graham, Ruslan Salmurzauli, and Steven Pueppke (2019), Challenges to the sustainable use of water resources in the Ili River basin of Central Asia, E3S Web of Conferences: 81. DOI: 10.1051/e3sconf/20198101009

Sternberg, Troy, Ahearn, A. and McConnell, F. (2017), Central Asian 'Characteristics' on China's New Silk Road: The Role of Landscape and the Politics of Infrastructure. Land, 6 (55), DOI: 10.3390/land6030055

Suriga Suriga, Miki Hashimoto, Buho Hoshino, Saixialt[3], Sumiya Ganzorig (2012), Change detection method for pasture degradation using RGB color composite image of multitemporal Landsat TM - A case study of the Inner Mongolian settlement region, IEEE IGARSS 2012(1). DOI: 10.1109/IGARSS.2012.6352691

Thevs N, Nurtazin S, Beckmann V, Salmyrzauli R, Khalil A. (2017), Water Consumption of Agriculture and Natural Ecosystems along the Ili River in China and Kazakhstan. Water, 9(3), DOI: 10.3390/w9030207

Troy Sternberg, Chris McCarthy, Buho Hoshino (2020), Does China's Belt and Road Initiative Threaten Food Security in Central Asia?, Water, 12, 2690. DOI: 10.3390/w12102690

Yuji Sakai, Chie Shimizu, Hironori Murata, Hitomi Seto, Ryosuke Fukushima, Takashi Koga, Chang Wang (2020), Changes in Soil Physicochemical Properties and Maize Production Following Improvement of Salt-Affected Soils Using Coal Bio-Briquette Ash in Northeast China. Agronomy 10:3, pages 348. DOI: 10.3390/agronomy10030348

USDA (2009), Kazakhstan Agricultural Overview, Commodity Intelligence Report (https://ipad.fas.usda.gov/highlights/2010/01/kaz_19jan2010/)

TIME SERIES ANALYSIS OF VIIRS-DNB NIGHTTIME LIGHTS IMAGERY FOR CHANGE DETECTION IN URBAN AREAS: A CASE STUDY OF DEVASTATION IN PUERTO RICO FROM HURRICANES IRMA AND MARIA

Guofeng Cao (1)* - Naizhuo Zhao (2)

Texas Tech University, Geosciences (1) - Mcgill University (2)
** Corresponding author: guofeng.cao@ttu.edu*

Abstract

Brightness of nighttime lights (NTL) collected by the Suomi National Polar-orbiting Partnership (S-NPP) Visible Infrared Imaging Radiometer Suite (VIIRS) is compatible across different times of images thanks to the on-board calibration system. However, the NTL radiance observed by the S-NPP VIIRS shows clear seasonality corresponding to the seasonal changes in the albedo of land surface. Additionally, the existence of many uncertain factors (e.g. complex atmospheric conditions) renders it inappropriate to directly use the NTL radiances to derive changes on the ground. In this study, we adopt a statistical procedure of time series analysis, namely seasonal and trend decomposition using Loess (STL), to model the time series observations of NTL brightness by decomposing the observations into three separable time series components (i.e. trend, seasonality, and remainder). Based on the time series model, forecast can be made for short-term future with confidence measure, and by comparing the model forecast with observed NTL brightness, significant changes can then be detected at pixel levels. We applied this method to the Puerto Rico area to detect and assess the damages caused by Hurricanes Irma and Maria, and to monitor the recovery after the disaster. Our results show that the proposed method successfully captures the changes of NTL brightness due to the damage of the hurricanes and general economic decline. Moreover, we also find that after removing the seasonal and remainder components, the time series of NTL image can more accurately reflect the temporal trends of economic status in Puerto Rico.

MAPPING VANADIUM IN THE BAUXITE TAILINGS WITH THE INTEGRATION OF REMOTE SENSING AND GEOSTATISTICAL APPROACHES

Sara Kasmaeeyazdi (1)* - Emanuele Mandanici (1) - Efthymios Balomenos (2) - Francesco Tinti (1) - Stefano Bonduà (1) - Roberto Bruno (1)

University of Bologna, Department of Civil, Chemical, Environmental and Materials Engineering (1) - Mytilineos S.a. (2)

** Corresponding author: sara.kasmaeeyazdi2@unibo.it*

Abstract

In remote sensing analysis, bands information and indices are used to map different regionalized variables, for different applications of earth and environmental sciences. Moreover, the classifications methods can be used to differentiate the areas, and then with kriging tools to estimate the target variable values and variances. Often, these analyses are enriched by the validation of the obtained estimation maps using values from in-situ samples. On the other hand, to get effective and reliable maps, there is the need of high amount of data. In this research, remote sensing studies (statistical studies, spectrum view and unsupervised classifications) applied to Copernicus Sentinel-2 images have been combined with advanced geostatistical approaches (Gaussian simulation using Turning Bands (TBs) algorithm) to map the distribution of one critical raw material (Vanadium element-V2O5). The approach has been applied to a Bauxite tailings case study, for a fixed-time sequence (4 months).

Simulation results have been obtained for the Vanadium grade variability maps in the Bauxite tailings for 1000 realizations using 60 samples as direct and Sentinel-2 images as collocated variable. To test the simulation results, the reproduced experimental variograms of the realizations are compared with the selected variogram model of the Vanadium concentration and a coherent convergence has appeared. Hence, despite the lack of band-ratio existence for Vanadium identification in remote sensing analysis and, on the other hand, the limited number of initial sampling of data for geostatistical analysis, the integration of both approaches has generated appropriate maps of Vanadium grade distribution, within the Bauxite tailings case study.

SAMPLING HILLSIDES OR FLOODPLAINS TO DETERMINE GEOCHEMICAL BACKGROUNDS FOR SOILS? A CRITICAL ANALYSIS THROUGH GEOSTATISTICAL AND MACHINE LEARNING APPROACHES

Carlos Boente (1)* - Saki Gerassis (2) - Teresa Albuquerque (3) - Margarida Ribeiro (3) - Susana Fernández (4) - Arturo Colina (1) - José Luis Rodríguez Gallego (1)

University of Oviedo, Indurot (1) - University of Vigo, Department of Natural Resources and Environmental Engineering (2) - Instituto Politécnico de Castelo Branco, Research Centre for Natural Resources, Environment and Society (cernas) (3) - University of Oviedo, Department of Geology (4)
** Corresponding author: carboente@gmail.com*

Abstract

The determination of soil screening levels (SSLs) for contaminant elements is a key factor to calculate the Risk-Based Soil Screening Levels (RBSSL), which mark the threshold of ecotoxicological risk for human health. Despite of their importance in health studies, obtaining these values (SSLs) is often complex since they depend on multiple factors such as geology, geomorphology, mining/industrial legacy, among others.

Multielemental soil contents are intrinsically related to the location. Thus, the aim of this research is to assess the differences between sampling in hillsides and/or floodplains for soil's geochemical backgrounds determination. On one hand, samples in floodplains across valleys are presumably closer to cities, agriculture, industry and, generally, to all the main anthropogenic activities, whereas, on the other hand, mountain hillsides are more prone to be influenced by intrinsic characteristics such as geological and geomorphological features.

The research work considered a full geochemical database of 334 soil samples that were specifically used for the determination of the official RBSSLs of Asturias (NW Spain), currently in force since 2014. The design of the sampling campaign implied in similar proportions both hillsides and floodplains. The Potentially Toxic Elements (PTEs) under study were Ag, As, Cd, Co, Cr, Cu, Hg, Mn, Mo, Ni, Pb, Sb, Tl, V and Zn, which constitute prominent pollutants in multiple studies involving Asturian soils. For operational purposes, the dataset was streamlined in three blocks (full data, hillsides, foodplains) and studied independently.

The adopted methodology is formed by a series of mathematical computation combining classical statistical and geostatistical approaches together with novel machine learning techniques. The entire process is divided into the following steps: 1. The elimination of outliers through the Mahalanobis' Distance (MD) algorithm; 2. Mathematical determination of Soil Screening Levels; 3. Principal Components Analysis (PCA) aiming to ascertain the associations among PTEs in each dataset; 4. A supervised Bayesian learning as a solution to determine the PTEs conditioning the most SSLs values; 5. Stochastic Simulation aiming the computation of a mean image for PTEs spatial distribution as well as the definition of clusters of high-low spatial uncertainty. Furthermore, this work intended to overlay the obtained spatial uncertainty patterns for both hillside and floodplain datasets, delineating the best fit regarding the definition of a robust geochemical background.

The findings revealed that floodplains provide higher SSLs for the majority of the PTEs, and that SSLs showed higher spatial uncertainty when considering hillsides. These results strongly suggest a higher association to anthropogenic activities in the case of floodplains and a close dependence between natural geological enrichments and hillsides.

In addition, all the results were compared with the SSLs obtained for the municipality of Langreo study case. Langreo is a territorial division belonging Asturias that presented a very intense industrial/mining activity in the past, but it is currently decaying. This gives rise to a similar case to the one presented here but on a minor area, in such a way the effect of the scale may be also addressed.

MAPPING THE GEOGENIC RADON POTENTIAL FOR GERMANY BY MACHINE LEARNING

Eric Petermann (1)* - Hanna Meyer (2) - Madlene Nussbaum (3) - Peter Bossew (1)

Federal Office for Radiation Protection (bfs), Radon and Norm (1) - Westfälische Wilhelms-universität Münster, Institute of Landscape Ecology (2) - Berne University of Applied Sciences (bfh), School of Agricultural, Forest and Food Sciences (3)

** Corresponding author: epetermann@bfs.de*

Abstract

The radioactive gas radon (Rn) is considered as an indoor air pollutant due to its detrimental effects on human health. Radon is known as the second most important cause for lung cancer after tobacco smoking. The dominant source of indoor Rn is the ground beneath the building in most cases. Following the European Basic Safety Standards, all EU Member States are required to delineate Rn priority areas, i.e. areas with increased risk of high indoor radon concentrations. One possibility to this end is using the "geogenic Rn potential" (GRP), which quantifies the availability of geogenic Rn for infiltration into buildings. The GRP is defined as a function of Rn concentration in soil gas and soil gas permeability.

In this study we used > 4,000 point measurements across Germany in combination with ~50 environmental co-variables (predictors). We fitted machine learning regression models to the target variables Rn concentration in soil and soil gas permeability. Subsequently, the GRP is calculated from both quantities. We compared the performance of three algorithms: Multivariate Adaptive Regression Splines (MARS), Random Forest (RF) and Support Vector Machines (SVM). Potential candidate predictors are geological, hydrogeological and soil landscape units, soil physical properties, soil chemical properties, soil hydraulic properties, climatic data, tectonic fault data, and geomorphological parameters.

The identification of informative predictors, tuning the model hyperparameters and estimation of the model performance was conducted using a spatial 10-fold cross-validation, where the folds were split by spatial blocks of 40×40 km. This procedure counteracts spatial autocorrelation of predictor and response data and is expected to ensure independence of training and test data. MARS, RF and SVM were evaluated in terms of its prediction accuracy and prediction variance. The results revealed that RF provided the most accurate predictions so far. The effect of the selected predictors on the final map was assessed in a quantitative way using partial dependence plots and spatial dependence maps. The RF model included 8 and 14 informative predictors for radon and permeability, respectively. The most important predictors in the RF model were geological and hydrogeological units as well as field capacity for radon and soil landscape, geological and hydrogeological units for soil gas permeability.

IMPACT OF DIFFERENT VARIOGRAM MODELS OF TOTAL ORGANIC CARBON ON SAMPLING SCHEME OPTIMIZATION AND POTENTIALITY OF COVARIATE INFORMATION IN THE PRECISION AGRICULTURE FRAMEWORK

Giuseppe Pappagallo (1)* - Emanuele Barca (1) - Daniela De Benedetto (2) - Anna Maria Stellacci (3)

National Research Council (cnr), Water Research Institute (irsa) (1) - Council for Agricultural Research and Economics (crea), Research Centre for Agriculture and Environment (crea-aa) (2) - University of Bari, Department of Soil, Plant and Food Sciences (di.s.s.p.a.) (3)

** Corresponding author: giuseppe.pappagallo@ba.irsa.cnr.it*

Abstract

Assessment at field scale of soil organic carbon (TOC) is of primary interest for agronomic management, particularly, in the precision agriculture framework. The knowledge about the spatial distribution of TOC is invaluable to implement strategies for improving the crop yield. However, the assessment of TOC spatial distribution requires the collection and the analysis of a large number of samples that is a costly and time-consuming activity.

In the present work, a strategy is proposed to optimize a sampling scheme of the considered soil property by means of an indirect auxiliary variable coming from the proximal geophysical sensing survey (GPR data). This variable has a greater spatial continuity than the soil organic carbon, then can be straightforwardly modelled in the geostatistical fashion. This allows to apply the spatial simulated annealing as an efficient mean for reducing optimally the sampling scheme. In addition, two different variogram models are derived by an automatic method. Such models are compared for assessing i) which is more suited as a descriptor of the indirect variable spatial behaviour and ii) allows the efficient reduction of the sampling scheme discarding all the redundant sampling points and saving those truly informative.

GEOSTATISTICAL INVERSION OF ELECTROMAGNETIC INDUCTION DATA FOR MODELLING WASTE DEPOSITS

João Narciso (1)* - Leonardo Azevedo (1) - Ellen Van De Vijver (2) - Marc Van Meirvenne (2)

Instituto Superior Técnico, Cerena (1) - Ghent University, Faculty of Bioscience Engineering, Department of Environment (2)

** Corresponding author: joao.narciso@tecnico.ulisboa.pt*

Abstract

The consumption of natural resources and the management of environmental issues arising from the associated production of waste has received increased attention in the recent decades. In many European countries, a paradigm shift is observed towards a smarter and more sustainable exploitation and management of resources, including considering the recovery of materials and/or energy from old landfills. This sparked new interest in innovative techniques that can support the characterization and modelling of deposits of urban and industrial wastes, including mine tailings. As waste deposits typically show a very complex and spatially heterogeneous composition, characterization that relies exclusively on direct observations from sampling is expensive and time-consuming, and, maybe even more problematically, may provide results that are not representative of the entire area of interest. Non-invasive geophysical methods have been proven effective alternatives to investigate the spatial distribution of the subsurface properties of landfill sites. Electromagnetic induction (EMI) surveys have been successfully applied to the qualitative characterization of landfills through imaging variations in the subsurface electrical conductivity (EC) and magnetic susceptibility (MS) which can be related to changes in waste composition and conditions (e.g. moisture content). However, due to the non-unique electromagnetic signature of different waste materials and the complex three-dimensional sensitivity of EMI measurements, the use of quantitative interpretation techniques such as inversion is still limited. Furthermore, most geophysical inversions are done under a deterministic framework. These approaches have two main limitations: (1) they do not allow quantification of the uncertainties about the inversion results, and (2) they generally produce overly smoothed versions of an actually much more heterogeneous subsurface reality. Therefore, geostatistical inversion of EMI data emerges as a powerful tool to improve the landfill modelling from geophysical data.

This work presents a new iterative geostatistical EMI inversion method where ensembles of near-subsurface petrophysical models, expressed in terms of EC and MS, are generated with stochastic sequential simulation and co-simulation. Resulting models are conditioned locally by existing borehole data of these properties and a spatial continuity pattern as defined by a variogram model. For each model, the synthetic instrument response including both the in-phase and quadrature-phase components of the secondary magnetic field is calculated using a 1-D forward model to link the EMI data with the petrophysical domain. The misfit between observed and synthetic FDEM data and the sensitivity to each petrophysical property in function of depth is used to drive the generation of a new set of models in the subsequent iteration, in which a global multi-objective optimizer is used to converge the inversion from iteration-to-iteration. We apply this method both to a synthetic landfill data set, which were created based on real observations made at a Portuguese mine tailing, and to a real landfill data set located in Belgium. Furthermore, this new methodology provides a flexible framework for data integration and spatial prediction of subsurface EC and MS along with its spatial uncertainty.

EXTRAPOLATION OF A LEGACY SOIL MAP TO SURROUNDING AREAS BY MACHINE LEARNING BASED MODEL AVERAGING

Madlene Nussbaum (1)* - Stéphane Burgos (1)

Berne University of Applied Sciences (bfh), School of Agricultural, Forest and Food Sciences (hafl) (1)
** Corresponding author: madlene.nussbaum@bfh.ch*

Abstract

Spatial information on soil is crucial for many applications such as spatial planning, erosion reduction, agricultural management or climate mitigation. In Switzerland, political pressure has recently risen to improve the basis for soil related decision making. The administration of the Swiss Canton of Schaffhausen aims to map agricultural soils (15000 hectares) with high resolution to allow for decisions relevant to landownership. One third of the area has been mapped by a conventional approach in the early 1990s at a scale of 1:5000.

To reduce the future soil sampling effort to complete the soil information for Schaffhausen the legacy soil map was fully exploited by a non-parametric bootstrap approach. During the survey in the 1990s the observed values of the soil samples have only been recorded in aggregated form as soil polygons. Besides the missing spatial coordinates the observed soil properties have been grouped into classes. We generated a dense set of "virtual soil samples" from the legacy map. Class width and multiple legend units per soil map polygon were considered by repeated generation of virtual samples and random assignment of response values (e. g. percent of topsoil clay or soil depth). We applied a model weighted averaging approach to these responses combining seven models that could be built automatically. The inverse of the mean squared error of the cross-validation was used as weights. Models were fitted with a large set of environmental covariates by 1) model selection for linear models through grouped lasso, 2) robust external drift kriging (georob), 3) geoadditive models selecting penalized smoothing spline terms by componentwise gradient boosting (geoGAM), and three different tree-based methods 4) boosted regression trees, 5) random forest, 6) rule-based linear regression (Cubist) and 7) support vector machines with non-linear basis functions.

The predictions of the seven models were averaged at the nodes of a 10 m-grid. To avoid extrapolation into areas with different soil forming factors we have carefully chosen the target area for prediction based on a similarity analysis. The non-parametric bootstrap allowed to create predictive distributions and map the uncertainty with 90 %-prediction intervals. The predicted maps have been successfully validated with 211 legacy soil profiles and 350 new soil samples chosen by a stratified random design.

BAYESIAN MODELING OF SPATIO-TEMPORAL TRENDS IN SOIL PROPERTIES USING INLA AND SPDE

Nicolas Saby (1)* - Thomas Opitz (2) - Bifeng Hu (3) - Hocine Bourennane (3) - Blandine Lemercier (4)

Inrae, Infosol (1) - Inrae, Biosp (2) - Inrae, Ur Sols (3) - Agrocampus Ouest, Umr Sas (4)
** Corresponding author: nicolas.saby@inrae.fr*

Abstract

The assumption of spatial and temporal stationarity does not hold for many ecological and environmental processes. This is particularly the case for many soil processes, often driven by factors such as biological dynamics, climate change and anthropogenic influences. For better understanding and predicting such phenomena, we develop a Bayesian inference framework that combines the integrated nested Laplace approximation (INLA) with the stochastic partial differential equation approach (SPDE). We put focus on modeling complex temporal trends varying through space with an accurate assessment of uncertainties, and on spatio-temporal mapping of processes that are only partially observed through soil variables measurements.

We model observed data through a latent (i.e., unobserved) smooth process whose additive components are endowed with Gaussian process priors. We use the SPDE approach to implement flexible sparse-matrix approximations of the Matérn covariance for spatial fields. The separate specification of the spatially varying linear trend allows us to conduct component-specific statistical inferences (range and variance estimates, standard errors, confidence bounds). For observed data following a Gaussian distribution, we add independent measurement errors. We also include in our model covariate information on parent material, climate and seasonality. Finally, in this approach we propose an analysis of the estimated model to summarize salient features the spatio-temporal predictions using a multivariate analysis.

In this work, we used this approach to study possible trends in space and time of several agronomic properties of agricultural fields in France. We used a large dataset, comprising more than 2 million values collected over the French territory over a 25 years period (1990-2014). This database, called the 'French Soil tests database', gathers soil test results produced by soil test laboratories on the request of French farmers.

For example, we studied the spatio-temporal variation of soil pH values, which can rapidly change within several years due to the effects of agricultural practices. We also explored spatio-temporal variation of carbon content and nutriments (P, K, Mg). Our models reveal significant temporal trends with strong spatial heterogeneity. We show that soil pH increased in a large part of agricultural soils monitored in the database. This finding suggests that the soils became less acidic across the country during the survey period. Conversely, decreases were almost never detected. When significant, the pH increase averages 0.25 over a 15-yr period. We assume that part of the increase can be explained by changes in nitrogen inputs to soil, especially associated with animal breeding systems, and possible decreases in acid deposition from the atmosphere. One limitation in proving the significance of the changes is the number of samples collected in certain part of the study area. Despite its limitations, the soil test database coupled with advanced statistical model appears to be a useful tool for monitoring such overall changes at a national level. Overall, our results demonstrate the high potential for the use of historical data stored in such large databases.

GEOCARE, DEVELOPMENT OF GEOPHYSICAL METHODS FOR CHARACTERIZATION AND REHABILITATION OF CONTAMINATED SITES

Theo Declerq (1)* - Laurent Thannberger (1) - Abderrahim Jardani (2)

Valgo (1) - University of Rouen Mont Saint Aignan, M2c (2)
** Corresponding author: theo.declercq@valgo.com*

Abstract

The rehabilitation of polluted sites and soils is a major issue that responds both to increasing land pressure and to environmental objectives in order to limit the impact on the health of individuals, on water resources and on biodiversity. The current means used for characterizing a polluted site and from which management measures are defined are generally based on one-off investigations (soil surveys, piezometers, piezairs, etc.). This creates uncertainty about the characterization of pollution depending intimately on the number and position of the piezometers.

The question of the spread of these measurements in the space-time in order to obtain an overview of the parameters needed for designing the management plan is essential for cost-effectiveness optimization. This work is often complex due to the heterogeneity of environmental properties and pollution. In this context, geophysical and hydro-geophysical methods, inherently spatialized and for some of them non-intrusive, offer an alternative to conventional investigative techniques for characterizing a hydrocarbon pollution type.

After 3D hydraulic characterization of the aquifer based on cross pumping and 3D electrical resistivity for heterogeneities determination, we tried to understand the LNAPL distribution in the soil in 3 dimensions though geophysical methods. The geophysical methods are geostatistically based with variograms and kriging used to treat and select the most representative data.

In this context, the objective of the GEOCARE project is to assess the potential of these methods to characterize a LNAPL source zone (oil cuts) and to follow up on the pumping/skimming treatment technique.

Several geophysical methods were tested in cubitainers and then on a pilot site previously characterized by innovative hydro-geophysical cross-pumping methods based on geostatistics. These maps, obtained after stochastic and deterministic inversions of data from different wells tracked over several pumping, identify wells directly related to the strongest LNAPL quantities for more efficient pumping.

In addition, in order to correlate geophysical measurements to field observations, coupled pressure sensors were used to continuously monitor the water table level to define a unique hydrological behaviour in relation to the lunar cycle and the mobile LNAPL thickness located in the well as an anti-correlation varying with the tides and the lunar cycle.

The geophysical method in the proposed presentation is the electrical resistance tomography one with the peculiarity of having electrodes arranged on the surface in 2D but also vertically: either in the wells or directly in contact with the soil; This characterization method allows to highlight greater electrical resistances in correlation with the amount of hydrocarbons observed in wells.

The work carried out has allowed different approaches to be compared to the pilot area. Investigations and interpretations are underway to assess the relevance of the various tools used for the diagnosis and follow-up of treatment pilots.

SPATIO-TEMPORAL GEOSTATISTICAL ANALYSIS AND PREDICTION FOR FINANCIAL DATA

Sandra De Iaco (1) - Monica Palma (1) - Daniela Pellegrino (1)*

University of Salento (1)
** Corresponding author: daniela.pellegrino@unisalento.it*

Abstract

Recently, the interest on financial risk management has increased and the analysis of the joint spatial and temporal profiles of the associated financial variables provides appreciable basis for interpreting their behavior.

Analysis over space and time might highlight some characteristics, which are typical of local areas and/or temporal spans, due to specific political actions on different social, economic and financial backgrounds. Thus, predictions may offer significant hints to policy makers, investors and market operators in planning strategies for the economic growth and macroeconomic stability.

A Geostatistical approach in this context is thus suitable to model jointly the spatial and temporal variability exhibited by the data, to predict financial variables as well as to realize probability maps related to different levels of financial risk.

In particular, after introducing the available data regarding financial variables observed in the Italian provinces, for different time points, a space-time geostatistical analysis for modeling and prediction purposes will be discussed. Finally, probability financial risk maps will be produced.

DYNAMIC RAINFALL MODELLING USING SPATIOTEMPORAL GEOSTATISTICS: BLENDING SATELLITE AND GROUND OBSERVATIONS

Emmanouil A Varouchakis (1)* - Dionissios T Hristopulos (2) - George P Karatzas (1) - Gerald Corzo (3)

Technical University of Crete, Environmental Engineering (1) - Technical University of Crete, Mineral Resources Engineering (2) - Ihe Delft Institute for Water Education, Integrated Water Systems and Governance (3)
** Corresponding author: varuhaki@mred.tuc.gr*

Abstract

Spatiotemporal precipitation monitoring and analysis is useful for water resources management studies. However, precipitation is usually measured at a limited number of locations. In particular, in areas of complex terrain, where topography plays a key role in the precipitation process, precipitation stations are usually sparse. Satellite precipitation data are an attractive alternative to precipitation observations. Usually though, they present inconsistencies due to the complexity of the retrieval algorithms and/or failure of infrared observation capability. In addition, estimations are available in spatial resolutions that occasionally miss significant terrain characteristics. This work presents a methodology that combines satellite and ground precipitation observations for the improvement of spatiotemporal mapping and analysis. The methodology is based on a geostatistical framework using Space-time Residual Kriging approach involving precipitation satellite data and elevation as covariates. Such an approach is applied to approximate non-stationarity of data and to suppress outliers. Separable and non-separable variogram functions are assessed to identify the optimal function to determine the spatiotemporal interdependence of fluctuations including the sum-metric and the Spartan variograms. The case study is the island of Crete, Greece, and the available data consist of ground observations from 54 stations and satellite observations at annual scale during the period 2010–2018.

MODELING MULTIVARIATE SPACE-TIME ANISOTROPIC COVARIANCE FUNCTION

Sandra De Iaco (1) - Monica Palma (1)* - Donato Posa (1)

University of Salento, Dept. of Management and Economics (Sect. Mathematics and Statistics) (1)
** Corresponding author: monica.palma@unisalento.it*

Abstract

In multivariate context, it is common to adopt the linear coregionalization model (LCM) based on isotropic independent hidden components underlying the phenomenon of interest. In this paper, a spatio-temporal LCM which takes into account the presence of possible spatial anisotropies is proposed and practical aspects in fitting and modeling are faced. A case study concerning daily averages of climatic variables, such as minimum and maximum temperature, and 10-centimeters soil temperature, recorded at some stations of the Irish Meteorological Service for 20-year span, is discussed. Thus, after establishing the possible presence of spatial anisotropy, the independent latent components which jointly describe the direct and cross-correlation among the variables under study, are identified; then the spatio-temporal anisotropic LCM is fitted.

Key words: covariance matrix, spatial direction, linear coregionalization model, fitting procedure

1. Introduction

The LCM is one of the most used models in multivariate geostatistical applications (Bevilacqua et al.,2015; Emery, 2010; Genton and Kleiber, 2015) and, in the last decade (De Iaco et al., 2012, 2013), it has been successfully proposed in the spatio-temporal context thanks to its flexibility and versatility in modelling the direct and cross-correlation detected for the variables at hand. Moreover, if the researcher has spotted a specific spatial direction for the variables over the study area, a spatio-temporal LCM with spatial anisotropy can be easily defined even by applying the fitting procedure proposed in De Iaco et al. (2019a).

In this paper, after a brief review on the main concepts of the multivariate Geostatistics in space-time (Section 2), the procedure to define a suitable spatio-temporal anisotropic LCM is presented (Section 3) and then applied at three climatic correlated variables (Section 4).

2. A brief review on space-time multivariate Geostatistics

In multivariate geostatistical analyses, the measurements of the $m \geq 2$ variables under study are considered as a realization of a multivariate space-time random field (MSTRF) $\{\mathbf{X}(\mathbf{s},t), (\mathbf{s},t) \in D \times T \subseteq \mathbb{R}^d \times \mathbb{R}\}$, with $\mathbf{X}(\mathbf{s},t) = [X_1(\mathbf{s},t), \dots, X_m(\mathbf{s},t)]^T$, where $(\mathbf{s},t)$ is a point in the spatio-temporal domain $D \times T$.

Assuming second-order stationarity, the covariance matrix $\mathbf{C}$ for the random vectors $\mathbf{X}(\mathbf{s},t)$ and $\mathbf{X}(\mathbf{s}',t')$ exists and depends on the space-time separation vector $\mathbf{h} = (\mathbf{h}_s, h_t)$, with $\mathbf{h}_s = (\mathbf{s} - \mathbf{s}')$ and $h_t = (t - t')$, i.e.

$$\mathbf{C}(\mathbf{X}(\mathbf{s},t), \mathbf{X}(\mathbf{s}',t')) = \mathbf{C}(\mathbf{h}_s, h_t) = [C_{ij}(\mathbf{h}_s, h_t)], \tag{1}$$

where

$$C_{ij}(\mathbf{h}_s, h_t) = E[(X_i(\mathbf{s} + \mathbf{h}_s, t + h_t) \cdot X_j(\mathbf{s},t))] - \mu_i\,\mu_j, \qquad i,j = 1, \dots, m,$$

are the cross-covariance functions, when $i \neq j$, and the direct covariance functions, when $i = j$; while μ_i and μ_j are the expected values of X_i and X_j, respectively.

In the spatio-temporal LCM, the covariance matrix $\mathbf{C}$ is modeled as

$$\mathbf{C}(\mathbf{h}_s, h_t) = \sum_{l=1}^{L} \mathbf{B}_l \; c_l(\mathbf{h}_s, h_t), \tag{2}$$

where $c_l(\mathbf{h}_s, h_t)$ are the covariances of the spatio-temporal latent variables describing the MSTRF $\mathbf{X}$ and $\mathbf{B}_l = [b_{ij}^l], l = 1, \dots, L$, with $L \leq m$, are $(m \times m)$ positive definite matrices, called in the literature coregionalization matrices (Chiles and Delfiner, 2012).

Two main issues have to be considered during the fitting procedure of model (2), i.e. the right detection of the basic latent components, as well as the choice of appropriate classes of covariance models for the selected components (De Iaco et al., 2019a). In order to make this process easier, a complete procedure has been recently developed by Cappello et al. (2021) which is structured in the following sub-procedures:

A. identification of the latent basic components;
B. modeling of the basic components by choosing the most appropriate models on the basis of test statistics' results concerning full symmetry, separability and type of separability of each latent component;
C. computation of the coregionalization matrices.

3. A fitting procedure for a spatio-temporal anisotropic LCM

As underlined in De Iaco et al. (2019b), the definition of isotropy has no meaning for spatio-temporal random fields; however, it is common to refer to the presence of spatial isotropy in space-time, when the the spatio-temporal covariance function depends on the modulus of the spatial lag and the temporal lag. Thus, for the LCM in (2), it is often assumed that $c_l(\mathbf{h}_s, h_t) = c_l(\|\mathbf{h}_s\|, |h_t|)$. However, in some applications, it is convenient to introduce some forms of spatial anisotropy in the spatio-temporal model used to describe the correlation structure of the basic components.

It is well-known that a geometric anisotropic covariance model in a 2D space can be seen as an isotropic model in a new system of coordinates obtained by a linear transform (rotation followed by re-scaling) of the original vector coordinates, fixed the anisotropy factor τ (i.e. the ratio between the minor range and the major range) and the angle α that the vertical axis forms (clockwise) with the axis corresponding to the direction of maximum continuity. Alternatively, an anisotropic covariance model can be described through a zonal model.

On the basis of these considerations, at the first stage of the above fitting procedure (Section 2), the structural analysis for the variables under study shall be also aimed at establishing possible spatial anisotropies. Note that, if the variables are also characterized by a systematic behavior in time, then at each spatial point of the area of interest the periodic components of the analyzed variables need to be estimated and then removed. Successively, the structural analysis will be carried out on the deseasonalized data by computing the direct and cross-covariance surfaces at different spatial directions in order to detect possible spatial anisotropies and consequentially, fit a spatio-temporal anisotropic LCM, as it will be shown in the following case study.

4. A case study

Climate issues represent a fundamental component in ecological researches; for this reason ecologists and agronomists tend to analyze the relationships among soil conditions and atmospheric variables. Soil temperature, namely the temperature measured at some centimeters soil depth, is one of the most important environmental variables which greatly affects the eco-system. This variable is, in turn, significantly influenced from other climatic variables, such as temperature, rainfall, solar radiation (Jungqvist et al., 2014; Islam et al., 2015; Pogacar et al., 2018).

The analyzed data set concerns 3 variables, that is 10-cm Soil Temperature, Minimum and Maximum Air Temperature, whose values (expressed in $^{\circ}C$) have been daily recorded from 1999 to 2018 at 14 stations (Fig. 1) of the Ireland's environmental monitoring network.

The exploratory data analysis has revealed the following features of the variables under study:

a) the lowest mean values occur in the North of Ireland, while in the Southern part of the island there are the greatest mean values, for all variables, as the posting maps in Fig. 1-a) show;

b) at low (high) minimum and maximum temperature, usually correspond low (high) soil temperature;

c) the values of the analyzed variables jointly increase in summer and decrease in winter, as it is evident in Fig. 1-b);

d) symmetry assumption between the variables is reasonable and is confirmed by the cross-correlation computed between the variables at different stations and time points, i.e. between the daily data recorded for one of the variable at $(\mathbf{s}, t)$ and the data recorded for another variable at $(\mathbf{s} + \mathbf{h}_s, t + h_t)$, and vice versa; as an example, in Fig. 2 the scatter plots of the daily data recorded at two different couples of stations during 2018 are shown: the values of the correlation coefficients and the shapes of the scatter plots highlight the linear relationships between the variables.

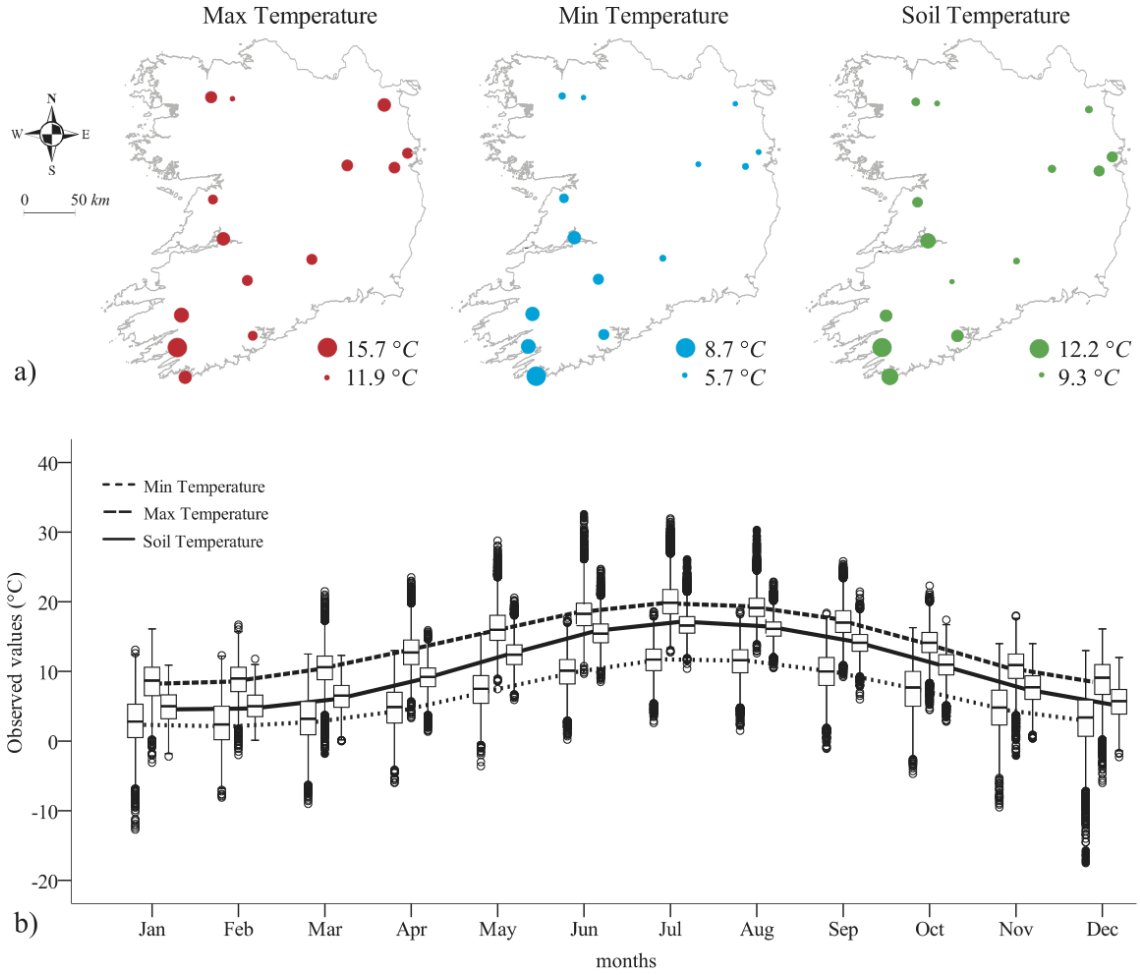


Figure 1: a) Posting maps of the mean values observed at the sample points, b) box plots of the daily observed data, grouped by month.

Then, the seasonal component shown by each variable has been estimated through the monthly averages and successively removed from the observed values. Therefore the deseasonalized data have been retained to compute the direct and cross-correlation in space-time. In particular, for a selection of 8 spatial lags and 21

temporal lags, direct and cross-covariance surfaces have been estimated in four spatial directions (North, East, $N45E$, $S65E$, with a tolerance degree equal to $\pm 22.5°$ in each direction). In this way, a geometric anisotropy has been identified for the deseasonalized variables with maximum spatial continuity in the North direction and minimum spatial continuity in the East direction.

At this point, the fitting procedure of the spatio-temporal LCM has gone on step-by-step as described in De Iaco et al. (2019a), for different spatial directions. In the following the results obtained for the East direction have been presented.

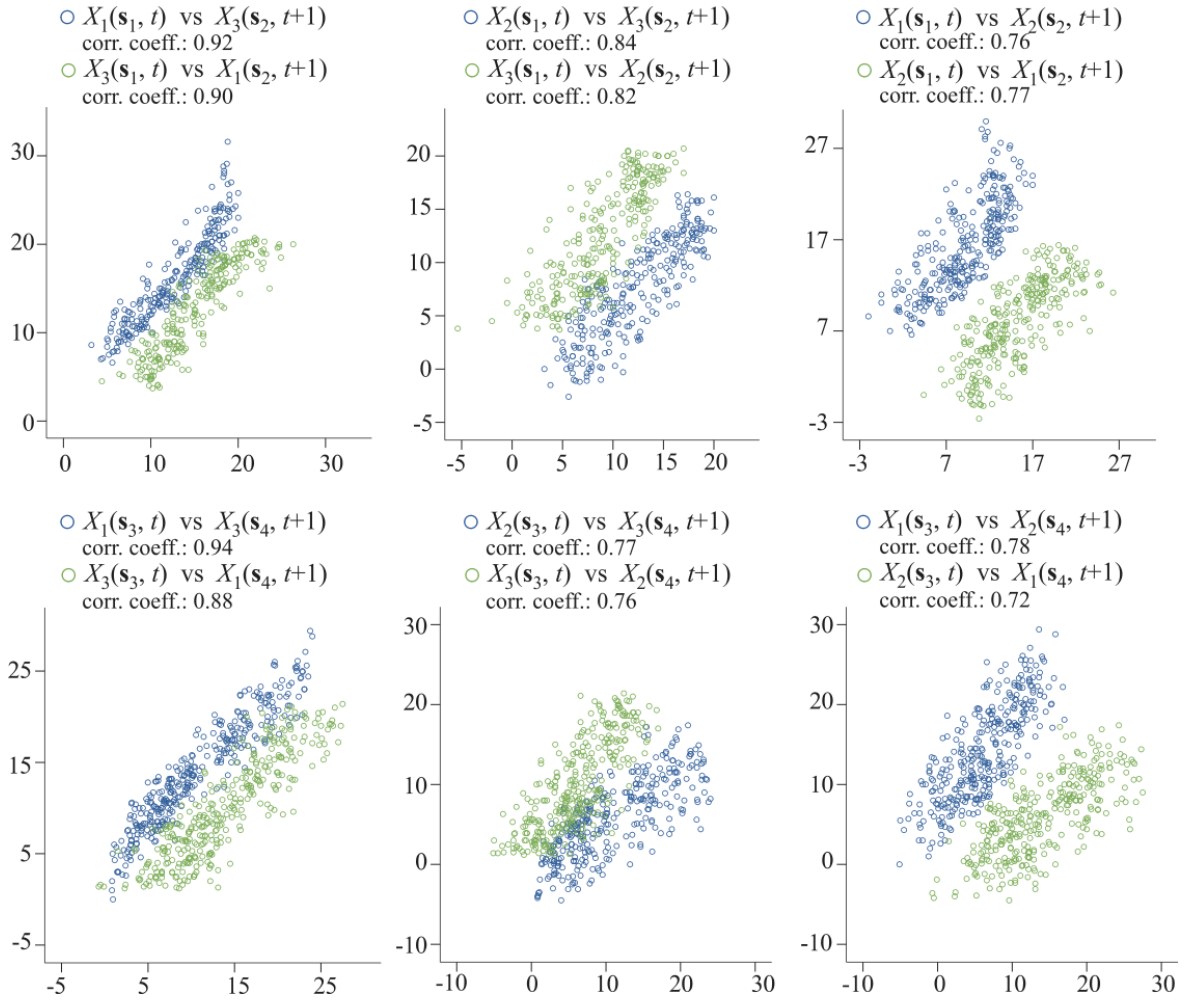


Figure 2: Scatter plots of the daily data observed in 2018 at two different couples of stations for the variables X_1,X_2 and X_3 corresponding to Max Temperature, Min Temperature and Soil Temperature, respectively.

First of all, the (3×3) matrices of sample directional covariances computed for 8 spatial and 21 temporal lags (in total 168 square matrices) have been simultaneously diagonalized through the following orthonormal matrix

$$\mathbf{\Phi} = \begin{bmatrix} 0.7482 & 0.5805 & -0.3213 \\ -0.6623 & 0.6243 & -0.4143 \\ -0.0399 & 0.5227 & 0.8516 \end{bmatrix}$$

As presented in De Iaco et al. (2019a), simultaneously nearly diagonalized of several symmetric and square matrices can be easily obtained by the R package Jade (Miettinen et al., 2017).

Then the 3 independent latent components have been calculated through the product between the data matrix and the above orthonormal matrix $\mathbf{\Phi}$. Among the three latent components, solely two of them have exhibited different variability scales in space-time (Fig. 3), namely

a) 90 *km* and 5 days, which is considered as the short-scale component (SSC);

b) 120 *km* and 10 days, which represents the long-scale component (LSC).

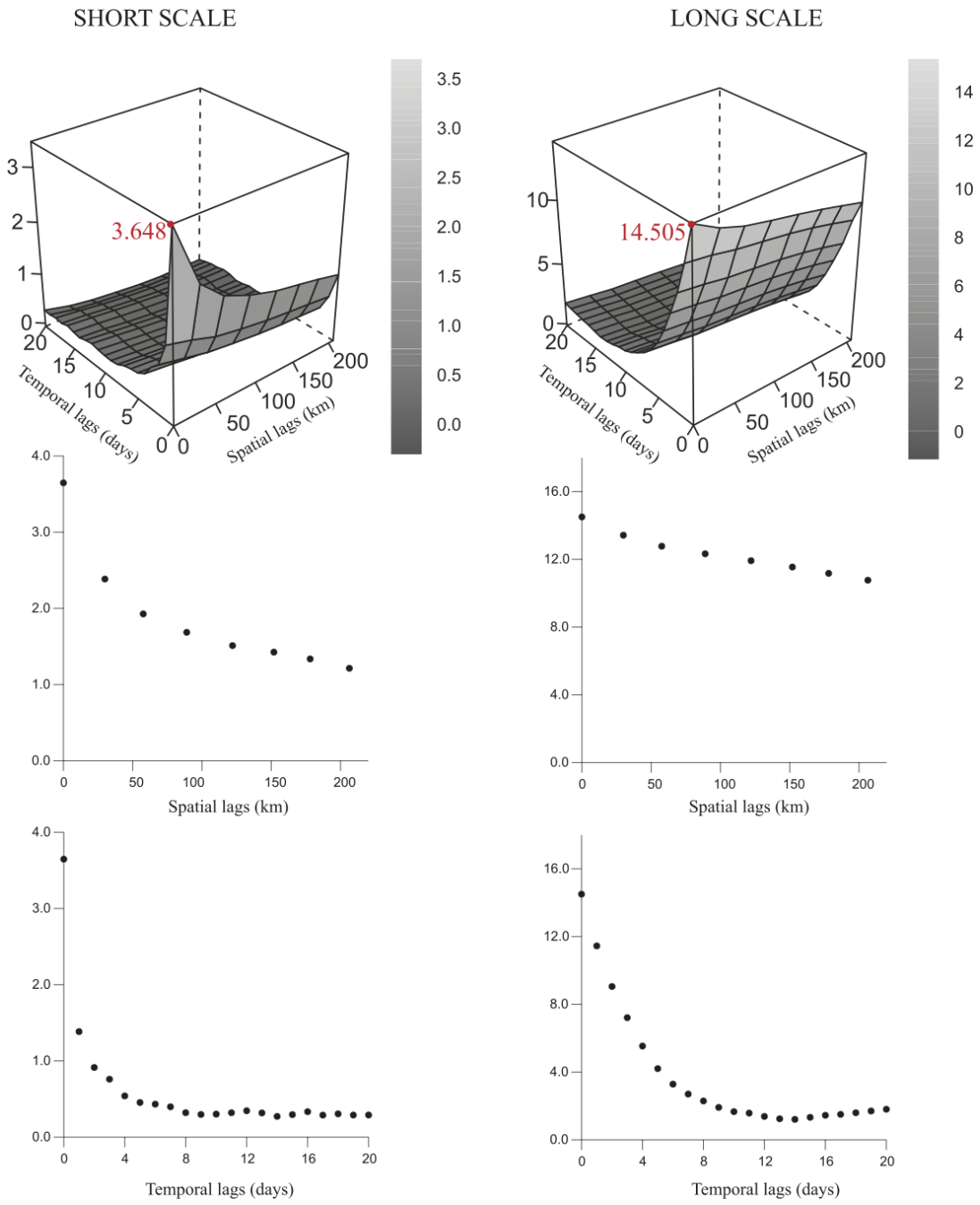


Figure 3: Space-time surfaces of SSC and LSC computed in the East direction, and their marginals in space and time.

The other unused component has been discarded due to the fact that its spatio-temporal variability was similar to one of the selected components. The next stage of the fitting procedure has regarded a series of tests in order to check the main features of the selected latent components. Full symmetry, separability and type of separability have been tested for the two basic components; details about these testing procedures can be found in De Iaco et al. (2016, 2019a).

Test results have driven the analyst in choosing the most properly classes of covariance models for the selected latent components. In particular, the tests' results have highlighted that, for the SSC, a fully symmetric space-

time covariance function which is positive for some lags and negative non-separable otherwise can be used, while for the LSC, a fully symmetric and negative non-separable covariance function can be considered.

Hence, the fitted space-time anisotropic LCM is

$$\mathbf{C}(\mathbf{h}_s, h_t) = \mathbf{B}_1 \ {}_{IPS}c_1(\mathbf{h}_s, h_t) + \mathbf{B}_2 \ {}_{PS}c_2(\mathbf{h}_s, h_t), \tag{3}$$

where, according to the above mentioned tests' results, the basic structures are modeled by using

- for the SSC, the integrated product-sum covariance models

$$ {}_{IPS}c_1(\mathbf{h}_s, h_t) = k_{11}\frac{1}{\frac{h_s'}{b_1}+\frac{|h_t|}{c_1}+1} + k_{21}\frac{1}{\frac{h_s'}{b_1}+1} + k_{31}\frac{1}{\frac{|h_t|}{c_1}+1}, \tag{4}$$

  where $k_{11} > 0, k_{21} \geq 0, k_{31} \geq 0$, while $b_1 > 0$ is the scaling parameter in space, $c_1 > 0$ is the scaling parameter in time and h_s' is the spatial distance in the new system of coordinates obtained by considering the anisotropy factor equal to 0.33 and the angle equal to 90°;

- for the LSC, the product-sum covariance model

$$ {}_{PS}c_2(\mathbf{h}_s, h_t) = k_{12}Exp(h_s'; b_2)Exp(|h_t|; c_2) + k_{22}Exp(h_s'; b_2) + k_{32}Exp(|h_t|; c_2), \tag{5}$$

  where $k_{12} > 0, k_{22} \geq 0, k_{32} \geq 0$, $Exp(\cdot\,; r)$ denotes the exponential covariance model, with practical ranges $r = b_2$ for the spatial marginal and $r = c_2$ for the temporal one, and h_s' is the spatial distance in the new system of coordinates obtained by considering the anisotropy factor equal to 0.8 and the angle equal to 90°.

By the constraint non-linear regression method implemented in the SPSS package, the parameters of models (4) and (5) have been estimated and their values have been reported in Tab. 1.

Table 1: Estimations of the covariance models parameters

Short-scale component ($l = 1$)	Long-scale component ($l = 2$)
Parameters of model (4)	Parameters of model (5)
$k_{11} = 1.890$	$k_{12} = 1.227$
$k_{21} = 0.042$	$k_{22} = 1.283$
$k_{31} = 1.715$	$k_{32} = 11.995$
$b_1 = 0.050$	$b_2 = 120$
$c_1 = 1.207$	$c_2 = 10$

To completely specify the space-time LCM, the two coregionalization matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ have to be computed (last step of the fitting procedure). Hence, by using the contributions of the SSC and LSC, the corresponding values of $c_1(\mathbf{0}, 0)$ and $c_2(\mathbf{0}, 0)$ (red values indicated in Fig. 3), as well as the empirical values of the direct and cross directional covariances of the variables nearby the first and the second variability scales (De Iaco et al., 2019a), the following matrices have been determined

$$\mathbf{B}_1 = \begin{bmatrix} 2.006 & 0.847 & 0.857 \\ 0.847 & 1.411 & 0.626 \\ 0.857 & 0.626 & 0.841 \end{bmatrix}, \mathbf{B}_2 = \begin{bmatrix} 0.073 & 0.066 & 0.068 \\ 0.066 & 0.088 & 0.075 \\ 0.068 & 0.075 & 0.070 \end{bmatrix}, \tag{6}$$

which are positive definite as confirmed by the results obtained through the spectral decomposition of $\mathbf{B}_1$ and $\mathbf{B}_2$.

5. Conclusions

In this paper, the use of a space-time anisotropic LCM is introduced and the fitting procedure, which allows the analyst to detect the most proper classes of covariance models to be used for the latent components of the LCM, is discussed. In the case study, the spatio-temporal correlation modeling among three climatic variables

was presented and the LCM characterized by the presence of spatial anisotropic latent components was fitted. It is worth pointing out that further advances might regard the use of the obtained space-time LCM for spatio-temporal predictions of the variables under study, where the spatial direction of maximum continuity is not neglected.

References

Bevilacqua M, Hering AS, Porcu E (2015), On the flexibility of multivariate covariance models: comment on the paper by Genton and Kleiber. *Stat. Sci.* **30** (2), 167–169.

Cappello C, De Iaco S, Palma M, Pellegrino D (2021), Spatio-temporal modeling of an environmental trivariate vector combining air and soil measurements from Ireland. *Spatial Statistics* **42** 100455.

Chilès JP, Delfiner P (2012), *Geostatistics. Modeling spatial uncertainty.* Second Edition. New York: Wiley.

De Iaco S, Maggio S, Palma M, Posa D (2012), Towards an automatic procedure for modeling multivariate space-time data. *Comput. Geosci.* **41**: 1–11.

De Iaco S, Myers DE, Palma M, Posa D (2013), Using Simultaneous Diagonalization to Identify a Space-Time Linear Coregionalization Model. *Math. Geosc.* **45**: 69–86.

De Iaco S, Palma M, Posa D (2016), A general procedure for selecting a class of fully symmetric space-time covariance functions. *Environmentrics* **27**(4): 212–224.

De Iaco S, Palma M, Posa D (2019), Choosing suitable linear coregionalization models for spatio-temporal data. *Stoch. Environ. Res. and Risk Assess.* **33**: 1419–1434.

De Iaco S, Posa D, Cappello C, Maggio S (2019), Isotropy, symmetry, separability and strict positive definiteness for covariance functions: a critical review. *Spat. Stat.* **29**: 89–108.

Emery X (2010), Interactive algorithms for fitting a linear model of coregionalization. Comput. Geosci. **36**(9): 1150–1160.

Genton MG, Kleiber W (2015), Cross-covariance functions for multivariate geostatistics. Stat. Sci. 30 (2), 147–163.

Islam K, Khan A, Islam T (2015), Correlation between Atmospheric Temperature and Soil Temperature: A Case Study for Dhaka, Bangladesh. Atmospheric and Climate Sciences, 5, 200–208.

Jungqvist G, Oni SK, Teutschbein C, Futter MN (2014), Effect of climate change on soil temperature in Swedish boreal forests. PLoS ONE, **9**(4), 1–12.

Miettinen J, Nordhausen K, Taskinen S (2017), Blind Source Separation Based on Joint Diagonalization in R: The Packages JADE and BSSasymp. *J. of Stat. Software*, **76**: 1–31.

Pogacar T, Zupanc V, Bogatai LK, CrepinSek Z (2018), Soil temperature analysis for various locations in Slovenia, *Ital. J. Agrometeorol.* **1**, 25–34.

ATMOSPHERIC CONDITIONS AT A WILDFIRE START: SPATIOTEMPORAL GEOSTATISTICS APPROACH

Eduardo Henrique de Moraes Takafuji (1)* - Marcelo Monteiro da Rocha (1) - Rodrigo Lilla Manzione (2)

Geoscience Institute, University of São Paulo (1) - Biosystems Engineering Department, School of Sciences and Engineering, São Paulo State University (2)
** Corresponding author: eduardo.takafuji@alumni.usp.br*

Abstract

Wildfires happen every year around the globe, but their frequency had increased as could be seen in United States, Australia, Brazil, Portugal, Indonesia, and Canada, just to cite a few places. Independently from its location, all fire triangles are composed of fuel, heat, and oxygen and they can be started naturally or by anthropogenic reasons. However, the fire needs the right condition in order to increase its status from a flame to a wildfire, that is, it should have enough fuel and proper atmospheric conditions. This study shows the atmospheric conditions (temperature, humidity, dew point, and wind speed) at three of the major wildfires ever registered in British Columbia (Canada). Every summer British Columbia forests present a wildfire risk and according to the Statistics & Geospatial Data of the British Columba Wildfire Service three of the largest wildfires ever recorded occurred in 2017. The fires at Plateau Fire and Hanceville Riske Creek were discovered on July 7th, 2017 and they burned 521,012 ha and 239,298 ha, respectively. Moreover, another huge wildfire at Elephant Hill was discovered on July 6th, 2017 and burned 191,865 ha. It is intuitive that in July 2017 British Columbia has the fuel and the atmospheric conditions to maintain wildfires. In those wildfire proximities, 34 weather stations are measuring several variables (*i.e.* min and max temperatures, wind conditions, precipitation, and others) daily. This study aimed to estimate the atmospheric condition at the specific fire region and the dates before and after its discovery by spatiotemporal kriging. In order to estimate these circumstances, the temporal range of the sampling was 31 days (15 days pre and 15 days post the event). Furthermore, the knowledge of the circumstances that a wildfire started, and spread can be helpful to prevent the worldwide forests from wildfires. From the results was possible to verify the relation between the weather and the wildfires in the region.

Keywords: spatio-temporal geostatistics, atmospheric conditions, wildfire

1. Introduction

According to the Government of British Columbia (2021), the largest wildfires registered between 2012 and 2017 were in Plateau Fire (521,012 ha) and Hanceville Riske Creek (239,298 ha), they were discovered on July 7th. The third-largest occurred in Elephant Hill (191,865 ha) was discovered on July 6th of 2017. It is evident that those days the weather was perfect to start a wildfire. Thus, we decided to investigate the atmospheric parameters that started these huge wildfires. Over the last few years, there are increasing interest in the wildfires subject in scientific production. Some studies about wildfire propagation are Khakzad (2019), Egorova et al. (2020), Cruz et al. (2020), Grasso and Innocente (2020), and Liu et al. (2021). Moreover, Parente et al. (2019) showed the relation between drought and wildfires in Portugal, Sayad et al. (2019) used machine learning modeling with air temperature, wild and soil moisture and Adhikari et al. (2021) used a data-driven wildfire simulation model to assess wildfire risk.

Several parameters influence wildfires spread (*e.g.* topography, soil moisture, vegetation, wind, temperature, and others) and this study is focused on atmospheric parameters measured by weather stations. Coen (2015) explains that weather affects fires from global and seasonal patterns to small-scale, where there are rapidly changing conditions in the fire's environment by altering the fluid dynamics of the air.

This study uses spatiotemporal kriging to investigate the atmospheric condition of the locals and the time of the wildfires' start. Spatiotemporal geostatistics is a fast way to estimate the atmospheric parameters without the huge time and processing demand of numerical climate/weather models. Geostatistics is a data-driven estimation method that is used to predict values at unsampled points. It uses spatial variability (variogram) to model the natural variability of the desired variable. In order to predict an unobserved location at any desired time, spatiotemporal geostatistics uses the space-time correlation obtained in the dataset (Montero et al., 2015). Several authors develop and explain spatiotemporal geostatistics (Kyriakidis and Journel, 1999, Sherman, 2011, Cressie and Wilke, 2011 and Montero et al., 2015). Moreover, spatiotemporal geostatistics can be applied in several areas such as meteorological variables (Spadavecchia and Williams, 2009, Fernández-Cortés et al., 2006 and Takafuji et al., 2020), short-term deforestation prediction in Amazonia (Sales et al., 2017), solar irradiance forecast in California (Jamaly and Kleissl, 2017), soil properties (Gasch et al, 2015), air pollution (Montero-Lorenzo et al. 2013, Menezes et al. 2016, Monteiro et al. 2017, and Hu et al., 2019) and groundwater (Snepvangers et al., 2003 and Manzione et al., 2019).

2. Material and Methods

In order to estimate the atmospheric conditions at the wildfire starts, this study used the dataset Global Summary of The Day GSOD (2018) from weather stations in British Columbia (Canada). This dataset contains dew point, sea level pressure, temperature, and wind speed. Coen (2015) explains that the parameter that directly affects combustion rates, or the air-fluid dynamics are: atmospheric temperature, relative humidity, and wind. Thus, the relative humidity was estimated through temperature and dew point.

All processing was done in the R project (R Core Team, 2021), 'gstat' (Pebesma, 2004 and Gräler et al., 2016), and 'spacetime' (Pebesma, 2012 and Bivand et al., 2013) R packages. All maps are considered as raster images and treated with 'raster' (Hijmans, 2020) and shapefiles as 'rgdal' (Bivand et al., 2021). The shapefiles for the Canadian provinces were retrieved from the website Global Administrative Areas (GADM, 2018). The shapefiles of the wildfires were retrieved from the Government of British Columbia (2021) website.

3. Results

In order to generate the map of each parameter (dew point, sea level pressure, temperature, and wind speed) on July 6th and 7th, the spatiotemporal variograms were modeled (Figure 1). The experimental variogram (left) measures the spatiotemporal variance at different space and time lags. The variogram model represents the spatiotemporal continuity to estimate the values at a specific point in space and time. The difference (right) shows how well the model fits the experimental variogram.

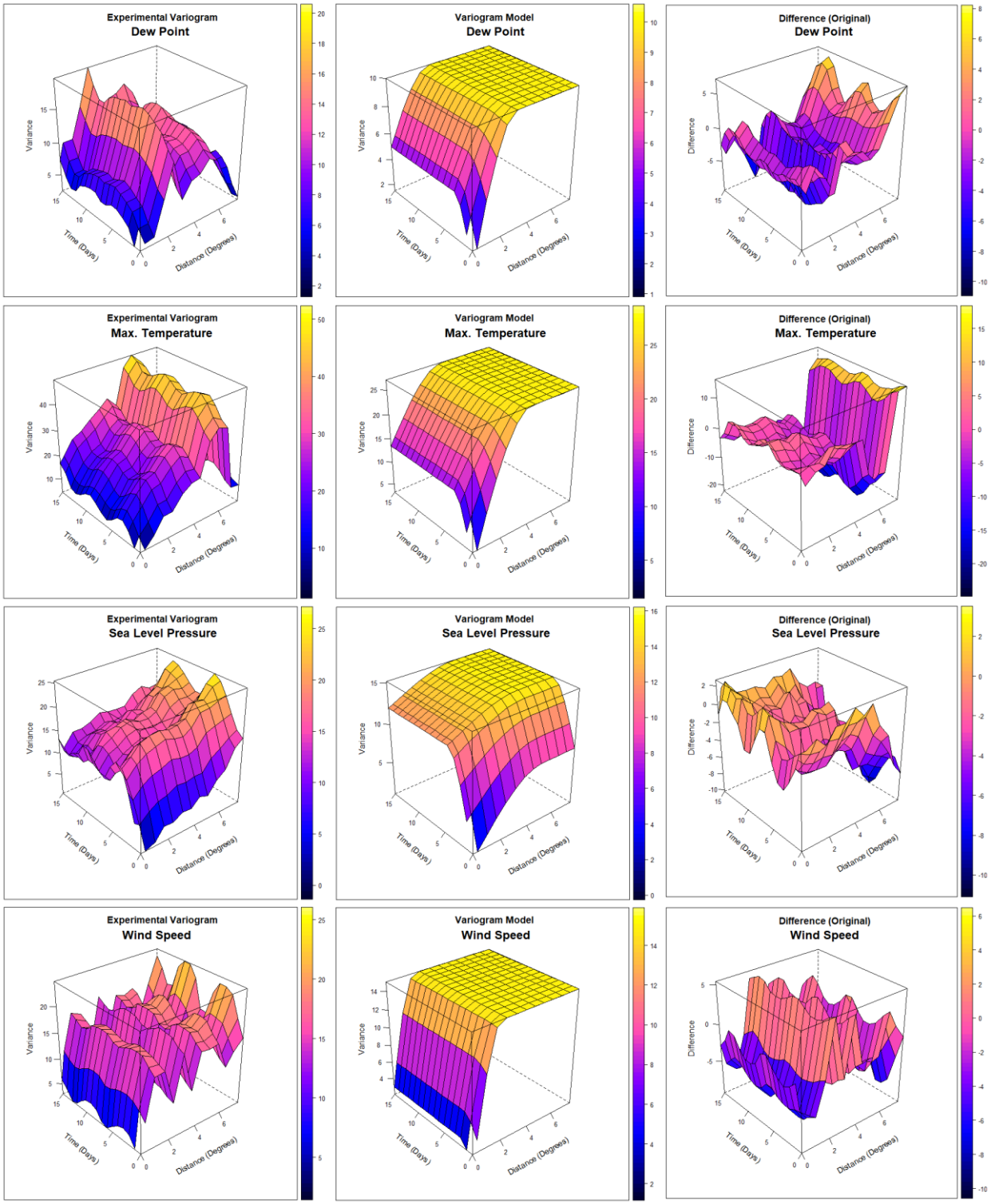


Figure 2 - Variogram (experimental, model, and difference) of each atmospheric parameter.

Then, the ordinary kriging was applied to generate the maps through a spatiotemporal dataset and variogram. Figures 2 and 3 show, respectively, the maps of four atmospheric parameters on July 6th and 7th. The blue stars the position of the weather stations, the red dot is where the fire start, and the shapefiles are the burned area. The burned area started at Elephant Hill (Lat: 121.297600W, Long: 50.699783N) on July 6th, 2017 and started at Plateau Fire (Lat: 123.084717W, Long: 52.503383N) and Hanceville Riske Creek (Lat: 123.168633W, Long: 51.944000N) on July 7th, 2017.

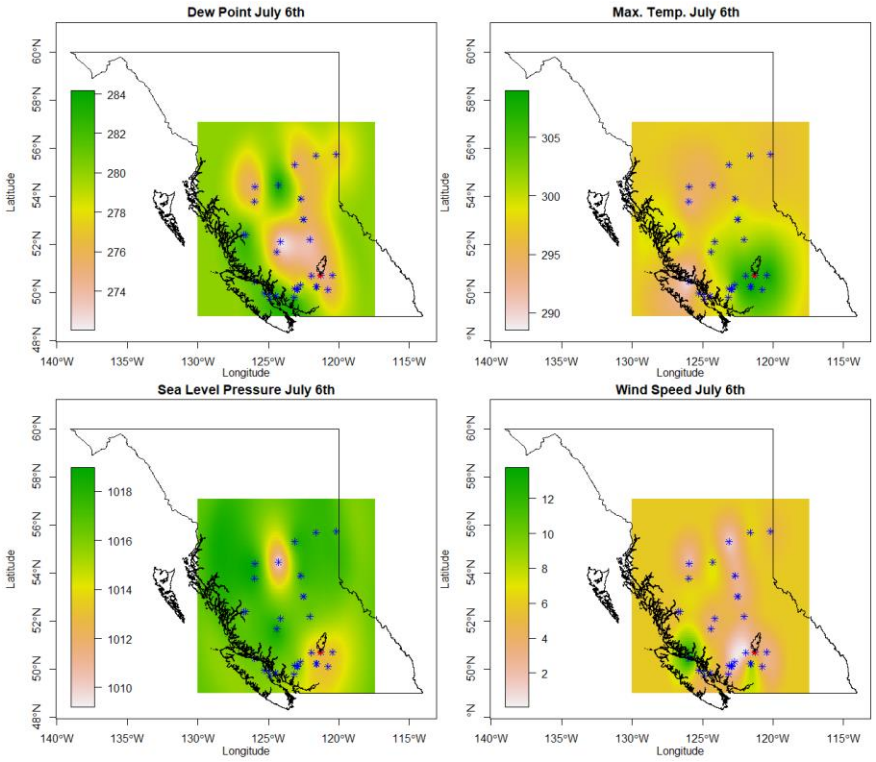


Figure 2 - British Columbia map with meteorological stations (blue stars), wildfire start point (red dot), wildfire area burned (black shape), and kriged wind speed (colored legend).

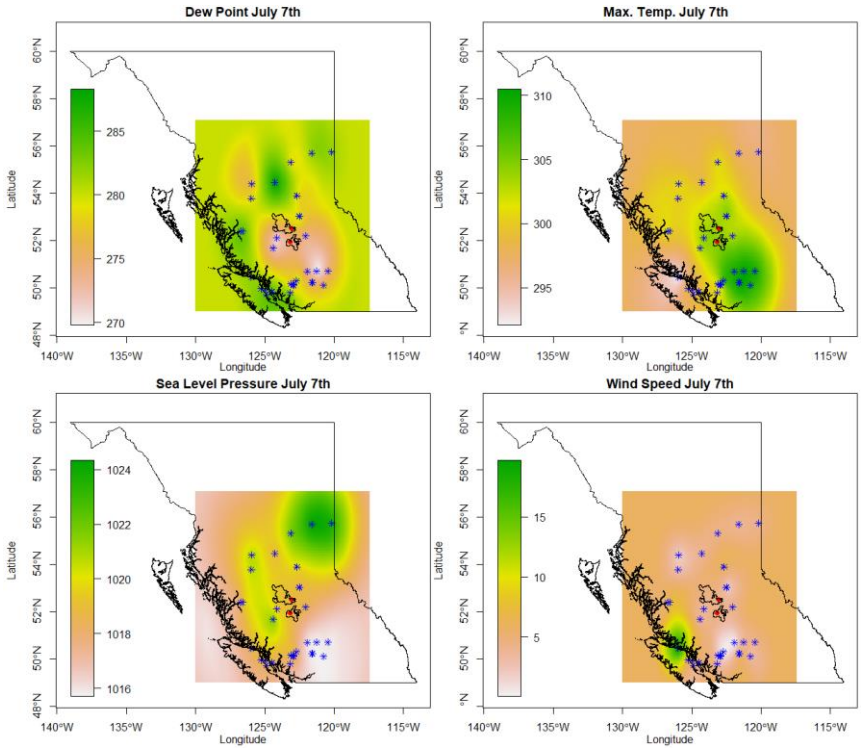


Figure 3 - British Columbia map with meteorological stations (blue stars), wildfire start point (red dot), wildfire area burned (black shape), and kriged wind speed (colored legend).

Figure 4 shows the time series of the atmospheric parameters at each wildfire origin. These values were estimated with spatiotemporal kriging.

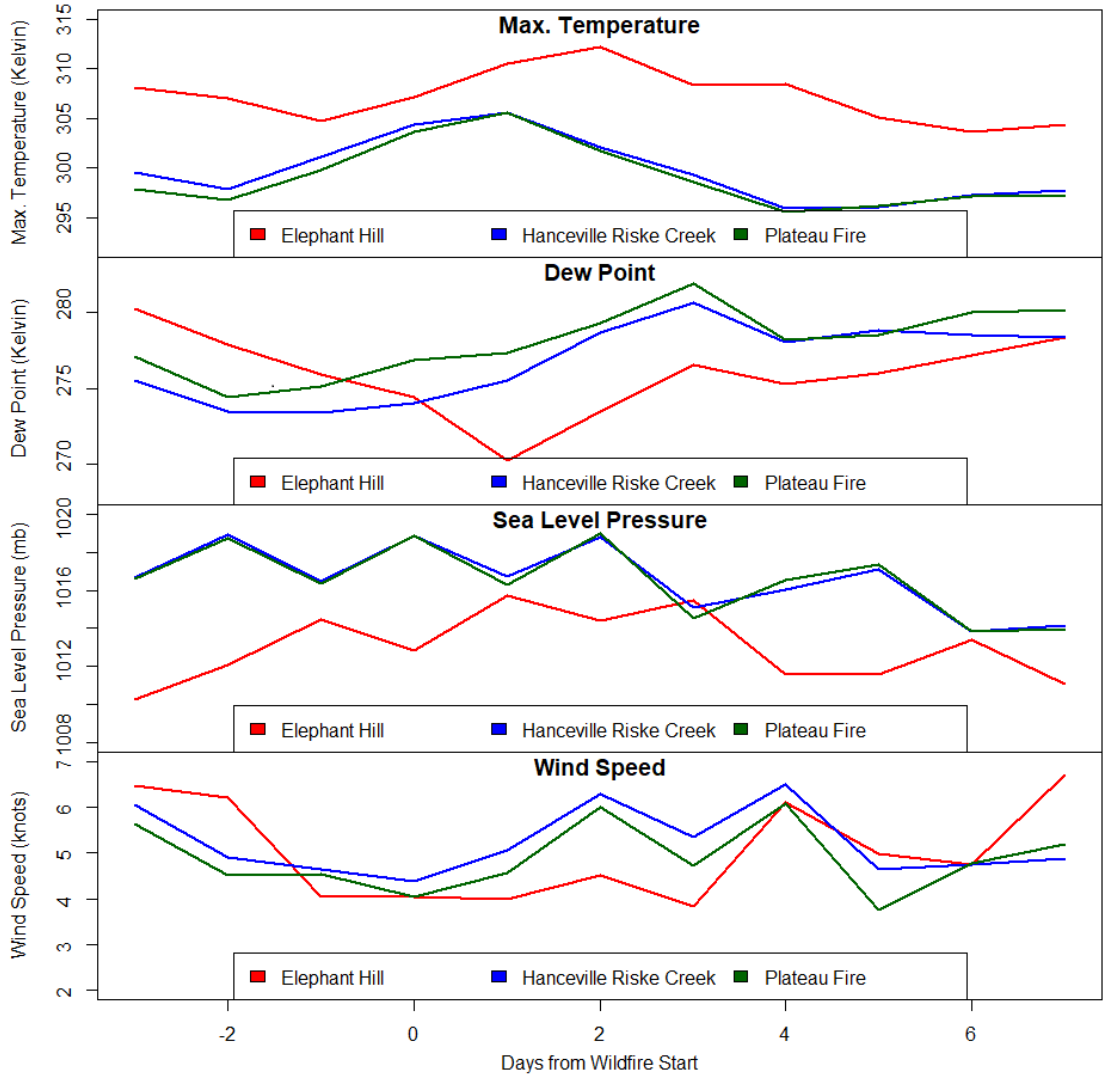


Figure 4 - Time series of atmospheric parameters at the origin of each wildfire. Day 0 is the day when the wildfire started.

Another important parameter is the relative humidity. It was computed by a simple conversion presented by Lawrence (2005) where it is demonstrated that:

$$RH \sim 100 - 5(t - t_d) \qquad \text{(eq. 1)}$$

where t is temperature and t_d is the dew point. The relative humidity approximation is used for moist air. But, in this study, it is used as a general rule to show how low is the relative humidity. Figure 5 is the map algebra (eq.1) of the dew point and maximum temperature maps.

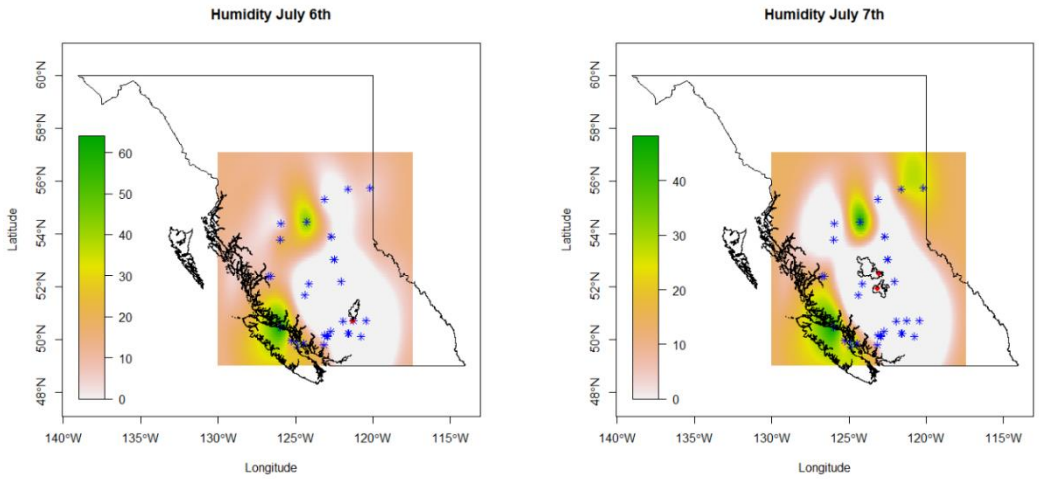


Figure 5 - British Columbia map with meteorological stations (blue stars), wildfire start point (red dot), wildfire area burned (black shape), and kriged humidity (colored legend).

4. Discussion and conclusions

The results (Figures 2, 3, and 4) show that the temperature is high and the dew point is low. The combination of these parameters indicates that the relative humidity is extremely low, and it is notable that its value is near zero around the registered origin of the wildfires (Figure 5). Moreover, the high temperature and low wind speed support the fire start and maintenance. Coen (2015) explains that, in summary, fires benefit from sustained high temperatures, low RH (single digits), dry fuels, strong winds, no RH recovery overnight, dry lightning, passing dry thunderstorms, and gusty winds. The time series (Figure 4) shows that the temperature was rising in all locations on the day the wildfire started. The dew point tends to decrease in the days before the fire. This combination created the extremely low relative humidity that is essential to start and maintain the wildfire. Moreover, the wind speed was relatively low on the day of the wildfire start and increased in the next days, which may help its spread. When observed spatially, the sea level pressure is known to guide the speed and direction of the winds, however, no correlation with the wildfire day was noticed in a single point.

This study shows how spatiotemporal geostatistics can estimate the atmospheric parameters at an unobserved local and time. This technique can be used for atmospheric parameters without the complexity and time-demanding of climate/weather numerical models. Thus, it is noticeable the relation between the weather and the start and maintenance of wildfires. As a suggestion of upcoming work, a complete study of the wildfires may show the weather patterns needed to generate machine learning that warns when there is a possible wildfire.

Acknowledgments

The authors thank everyone involved in collecting, processing, storing, and distributing the data used in this study. Data from weather stations were obtained from the National Oceanic and Atmospheric Administration (NOAA) National Center for Environmental Information (NCEI). Besides, the author is also grateful to those involved in the R and GADM projects and all collaborators of the R packages.

References

Adhikari B., Xu C., Hodza P., Minckley T. (2021), Developing a geospatial data-driven solution for rapid natural wildfire risk assessment. Applied Geography 126, doi.org/10.1016/j.apgeog.2020.102382.

Bivand R., Keitt T., Rowlingson B. (2021), rgdal: Bindings for the 'Geospatial' Data Abstraction Library, R package version 1.5-23, https://CRAN.R-project.org/package=rgdal, Accessed 20 March 2021.

Bivand R.S., Pebesma E., Gomez-Rubio V. (2013), Applied spatial data analysis with R (Second edition), Springer, New York.

Coen J. (2015), Weather Forecasting | Wildfire Weather, in: North G.R., Pyle J., ZhangF., Encyclopedia of Atmospheric Sciences (2nd Edition), vol 6, doi.org/10.1016/B978-0-12-382225-3.00481-3.

Cressie N., Wikle C.K. (2011), Statistics for Spatio-Temporal Data, John Wiley & Sons, Hoboken

Cruz M.G., Alexander M.E., Fernandes P.M. Kilinc M., Sil A. (2020), Evaluating the 10% wind speed rule of thumb for estimating a wildfire's forward rate of spread against an extensive independent set of observations, Environmental Modeling and Software 133, doi.org/10.1016/j.envsoft.2020.104818.

Egorova V.N., Trucchia A., Pagnini G. (2020), Fire-spotting generated fires. Part I: The role of atmospheric stability, Applied Mathematical Modelling 84:590-609, doi.org/10.1016/j.apm.2019.02.010.

Fernández-Cortés A., Calaforra J. M., Jiménez-Espinosa R., Sánchez-Martos F. (2006), Geostatistical spatiotemporal analysis of air temperature as an aid to delineating thermal stability zones in a potential show case: Implications for environmental management, Journal of Environmental Management 81:371-383, doi.org/10.1016/j.jenvman.2005.11.011.

Gasch C.K., Hengl T., Graler B., Meyer H., Magney T.S., Brown D.J. (2015), Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D + T: The Cook Agronomy Farm data set, Spatial Statistics, doi.org/10.1016/j.spasta.2015.04.001.

Global Administrative Areas (2018), GADM database of Global Administrative Areas (version 2.0), www.gadm.org, Accessed 16 July 2018.

Government of British Columbia (2021), https://www2.gov.bc.ca/gov/content/safety/wildfire-status/about-bcws/wildfire-statistics, Accessed 20 March 2021.

Gräler B., Pebesma E., Heuvelink G.B.M. (2016), Spatio-Temporal Interpolation using gstat, The R Journal, doi.org/10.32614/RJ-2016-014.

Grasso P., Innocente M.S. (2020), Physics-based model wildfire propagation towards faster-than-real-time simulations, Computers and Mathematics with Applications 80: 790-808, doi.org/10.1016/j.camwa.2020.05.009.

GSOD (2018), Climate Data Online (CDO), https://gis.ncdc.noaa.gov/maps/ncei/cdo/daily, Accessed 10 February 2018.

Hijmans R.J. (2020), raster: Geographic Data Analysis and Modeling, R package version 3.4-5, https://CRAN.R-project.org/package=raster, Accessed 20 March 2021.

Hu H., Hu Z., Zhong K., Xu J., Zhang F., Zhao Y., Wu P. (2019), Satellite-based high-resolution mapping of ground-level PM 2.5 concentrations over East China using a spatiotemporal regression kriging model, Science of the Total Environment 672: 479-490, doi.org/10.1016/j.scitotenv.2019.03.480.

Jamaly M., Kleissl J. (2017), Spatiotemporal interpolation and forecast of irradiance data using Kriging, Solar Energy 158: 407-423, doi.org/10.1016/j.solener.2017.09.057.

Khakzad N. (2019), Modeling wildfire spread in wildland-industrial interfaces using dynamic Bayesian network, Reliability Engineering and System Safety 189:165-176, doi.org/10.1016/j.ress.2019.04.006.

Kyriakidis P., Journel A.G. (1999), Geostatistical Space-Time Models: A Review, Mathematical Geology, doi.org/10.1023/A:1007528426688.

Lawrence M.G. (2005), The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications, Bulletin of the American Meteorological Society 86: 225-234, doi.org/10.1175/BAMS-86-2-225.

Liu N., Lei J., Gao W., Chen H., Xie X. (2021), Combustion dynamics of large-scale wildfires, Proceedings of the Combustion Institute, doi.org/10.1016/j.proci.2020.11.006.

Manzione, R.L., Takafuji, E.H.M., De Iaco, S., Cappello, C., Rocha, M.M. (2019), Spatio-temporal Kriging to Predict Water Table Depths from Monitoring Data in a Conservation Area at São Paulo State, Brazil, Geoinformatics & Geostatistics: An Overview, v. 7, doi.org/10.4172/2327-4581.1000205.

Menezes R., Piairo H., Garcia-Soidán P., Sousa I. (2016), Spatial–temporal modellization of the NO2 concentration data through geostatistical tools, Statistical Methods & Applications, doi.org/10.1007/s10260-015-0346-3.

Monteiro A., Menezes R., Silva M.E. (2017), Modelling spatio-temporal data with multiple seasonalities: The NO2 Portuguese case, Spatial Statistics, doi.org/10.1016/j.spasta.2017.04.005.

Montero J-M., Fernández-Avilés G., Mateu J. (2015), Spatial and Spatio-Temporal Geostatistical Modeling and Kriging, John Wiley & Sons, Chichester.

Montero-Lorenzo J-M., Fernández-Áviles G., Mondéjar-Jiménez J. (2013), A spatio-temporal geostatistical approach to predicting pollution levels: The case of mono-nitrogen oxides in Madrid, Computers, Environment and Urban Systems, doi.org/10.1016/j.compenvurbsys.2012.06.005.

Parente J., Amraoui M., Menezes I., Pereira M.G. (2019), Drought in Portugal: Current regime, comparison of indices and impacts on extreme wildfires, Science of the Total Environment 685: 150-173, doi.org/10.1016/j.scitotenv.2019.05.298.

Pebesma E. (2012), spacetime: Spatio-Temporal Data in R, Journal of Statistical Software, doi.org/10.18637/jss.v051.i07.

Pebesma E.J. (2004), Multivariable geostatistics in S: the gstat package, Computers & Geosciences, doi.org/10.1016/j.cageo.2004.03.012.

R Core Team (2021), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Sales M., Bruin S., Herold M., Kyriakidis P., Souza Jr C. (2017), A spatiotemporal geostatistical hurdle model approach for short-term deforestation prediction, Spatial Statistics 21: 304-318, doi.org/10.1016/j.spasta.2017.06.003.

Sayad Y.O., Mousannif H., Moatassime H.A. (2019), Predictive modeling of wildfires: A new dataset and machine learning approach, Fire Safety Journal 104:130-146, doi.org/10.1016/j.firesaf.2019.01.006.

Sherman M. (2011), Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties, John Wiley & Sons, Chichester.

Snepvangers J.J.J.C., Heuvelink G.B.M., Huisman J.A. (2003), Soil water content interpolation using spatio-temporal kriging with external drift, Geoderma, doi.org/10.1016/S0016-7061(02)00310-5.

Spadavecchia L., Williams M. (2009), Can spatio-temporal geostatistical methods improve high resolution regionalization of meteorological variables? Agricultural and Forest Meteorology 149:1105-1117, doi.org/10.1016/j.agrformet.2009.01.008.

Takafuji, E.H.M., Rocha, M.M., Manzione, R.L. (2020), Spatiotemporal forecast with local temporal drift applied to weather patterns in Patagonia, Springer Nature Applied Sciences, v. 2, doi.org/10.1007/s42452-020-2814-0.

FUZZY LOGIC AND SPACE-TIME INTERACTION PARAMETER IN COVARIANCE MODEL

Sandra De Iaco (1) - Monica Palma (1)* - Claudia Cappello (1)

University of Salento, Dept. of Management and Economics (Sect. Mathematics and Statistics) (1)
** Corresponding author: monica.palma@unisalento.it*

Abstract

In the literature, several efforts have been made in order to define a comprehensive procedure for the identification of the most appropriate class of space-time covariance functions for the observations at hand. Several statistical tests have been proposed in order to check the relevant aspects presented by the variable under study, which are particularly significant to select an appropriate class of space-time covariance functions. Among these aspects, it is crucial the identification of the space-time interaction parameter, which discriminates between separable and non-separable covariance models.

Space-time interaction produces effects in the behavior of the covariance/variogram surface. In particular, it can be observed that the spatio-temporal correlation is stronger (weaker) than the theoretical separable correlation (or correlation without space-time interaction). This can happen for all lags or just some of them.

However, the knowledge of the interaction parameter is imprecise, thus it might be useful to recall the fuzzy logic. In the literature, the fuzzy logic was applied in order to address imprecise information on variogram parameters (sill, nugget, range) in spatial kriging, while in the context of spatio-temporal geostatistics fuzzy theory was used to treat fuzzy data in very few cases.

In this paper, a novel approach based on the fuzzy logic is proposed to assess the space-time interaction parameter of the covariance function. Hence, the vector of possible parameters of the spatio-temporal covariance is considered to be a fuzzy set and each parameter vector is assigned a membership value in this set. A case study on environmental variable is thoroughly discussed.

GEOSTATISTICAL DOWNSCALING OF OFFSHORE WIND SPEED DATA DERIVED FROM NUMERICAL WEATHER PREDICTION MODELS USING HIGHER SPATIAL RESOLUTION SATELLITE PRODUCTS

Stylianos Hadjipetrou (1)* - Stelios Liodakis (1) - Anastasia Sykioti (1) - Phaedon Kyriakidis (1) - No-Wook Park (2)

Cyprus University of Technology, Department of Civil Engineering and Geomatics (1) - Inha University, Department of Geoinformatic Engineering (2)

** Corresponding author: sk.hadjipetrou@edu.cut.ac.cy*

Abstract

Regional offshore wind assessment studies typically rely on forecasts from Numerical Weather Prediction (NWP) models. NWP products are typically available at fine temporal resolutions (e.g., on an hourly basis) but relatively coarse spatial resolutions (e.g., on the order of several kilometers) to be used directly for more detailed local assessments. Satellite data, e.g., SAR (Synthetic Aperture Radar) data, on the other hand have been widely used in the literature to reveal high spatial resolution wind fields along with their variations but are available only at a few instances within a month's period. The C-Band SAR instrument onboard the Sentinel-1 platform, in particular, provides wind speed data at 10 m above sea surface with a repeat frequency of 6 days since 2016.

Statistical downscaling techniques are often employed to obtain finer spatial resolution products from coarse NWP products for use in finely resolved impact assessment studies. This study investigates the application of a novel geostatistical approach for downscaling Regional Reanalysis wind speed data using SAR data in order to spatially enhance information captured by the former. The data used comprise Sentinel-1A and 1B VV-polarized SAR wind field measurements and Uncertainties in Ensembles of Regional Reanalyses (UERRA) data, both bias-corrected using in-situ data from local meteorological coastal stations. The reference data used for bias correction are generated via spatial interpolation and aggregation (upscaling) of the local meteorological station wind speed values within the closest Sentinel pixel (1 km) and UERRA cell (11 km).

Prior to the downscaling procedure, Weibull distribution models are fitted to the wind speed time-series both at the coarse and fine spatial resolutions. Downscaled UERRA Weibull distributions parameters (scale (a) and shape (b)) are then generated via Area-To-Point Kriging with External Drift (ATPKED), whereby Weibull parameter values are computed at a finer spatial resolution as a weighted linear combination of neighboring coarse resolution attribute values. The fine resolution parameters are used as auxiliary variables. ATPKED is mass preserving, in that the average of the downscaled Weibull parameter values within a coarse cell reproduce the bias-corrected UERRA value at that cell. Once the fine scale parameters are estimated, the wind speed distribution at the pixel level can be extracted. Statistical comparison indicated that more than half of the wind speed variability in Sentinel images can be explained by the contemporaneous downscaled estimates. Geostatistical simulation is also employed to assess the uncertainty in the fine resolution values.

As an illustration of the methodology, offshore wind speed values are estimated at a spatial resolution of 1km for the coastal areas of the Republic of Cyprus at a 6-hour interval over a period of 1 year. The results imply that the downscaled products could furnish a basis for a more spatially resolved offshore wind power assessment for the region, provided the above procedure is generalized for a longer time period.

Keywords: Weibull distribution, Area-to-Point Kriging, External Drift, Simulation

1. Introduction

Wind exploitation offers new opportunities and a promising future towards renewable energy production. The offshore wind global market is growing big surpassing 35 GW of offshore wind capacity, which alone represents the 4.8% of the total global cumulative wind capacity, according to the latest Global Wind Energy Council report (Council, 2020). Weather model outputs provided at global or regional scales, however, entail uncertainties which do not allow for local scale wind resource assessment. To that end, finely resolved wind field information is vital towards assessing the offshore wind resource potential (Shin et al., 2018).

In-situ wind measurements, and especially offshore, are typically sparse and not easily accessible. In the absence of such data, researchers rely on coarse spatial resolution climate model outputs e.g., Numerical Weather Predictions (NWP). Local scale wind assessments, however, calls for fine-scale wind field data in order to drive a plethora of applications ranging from wind farm layout optimization (Pillai et al., 2017) to offshore wind energy storage and distribution (Katsaprakakis, 2016). Despite the readily available satellite data providing more detailed information regarding the spatial variability of offshore wind resource, the local assessment is still limited by the satellite revisit frequency and thus the few instances of actual wind data provided in time.

Dynamical and statistical (or empirical) techniques are being routinely employed to downscale the coarse spatial resolution climate information (Hewitson and Crane 1996). The former embed atmospheric mesoscale models nested within the large-scale forecast models (Al-Yahyai et al., 2012; Beaucage et al., 2014) while the latter rely on statistical relationships between coarse scale patterns of climate variables (predictor) and local scale products (predictand) to infer high-resolution probability density functions (pdf) (González-Aparicio et al., 2017; Winstral et al., 2017). Computation efficiency and complexity (Javad Alizadeh et al. n.d.) are, among other issues (Hong and Kanamitsu, 2014), the main drawbacks of the dynamical downscaling technique that favor the use of the statistical approaches. Temporal statistical models have also been developed to downscale wind-related climate variables. Kumar et. al. (2012) employed a neural network to downscale meteorological variables while Guo et. al. (2016) accounted for the diurnal patterns of wind speed to obtain hourly data.

In this study, we propose a geostatistical approach to refine the coarse spatial resolution wind information obtained from UERRA via integration with the finer scale Sentinel-1 SAR Level-2 OCN wind speed estimates. To overcome the temporal inconsistency between the two data sources, we fit Weibull distributions to their respective time-series at each location in order to obtain scale and shape parameters which are subsequently fused to derive wind speed distributions at a finer scale. The downscaling is achieved on the residuals of a regression model developed between the corresponding Weibull parameters of UERRA and Sentinel-1 via Area-To-Point Regression Kriging (ATPRK) where the Sentinel-1 information is used as external drift. Lastly, uncertainty assessment is achieved by means of stochastic simulation where alternative realizations amount to generating deviates from each local distribution of the downscaled/fused Weibull parameters, while accounting for the spatial auto-correlation and the (lag-0) cross-correlation in these parameters.

2. Material and Methods

2.1. Study area and data acquisition

The European Centre for Medium-Range Weather Forecasts (ECMWF) archive includes among other products UERRA regional reanalysis gridded data. UERRA HARMONIE/V1 model outputs, available at the spatial resolution of 11km and at a 6- hour interval, were acquired for the period of January 2014 to July 2019. Sentinel-1 Level-2 Ocean (OCN) products, on the other hand, are spatially finer (1km spatial resolution) but only

available from 2017 onwards. Moreover, Sentinel 1A and 1B satellites are recording tiles in the broader offshore Cyprus area approximately at 3:45 Coordinated Universal Time (UTC) and 15:45 UTC, leading to a (spatially partial) coverage of 1 to 2 scenes per day within a 4-day run, leaving 3 days in between without a scene. Ground range gridded estimates of wind speed at the 10m height above the sea surface, provided from the Ocean Wind Fields (OWI) geophysical component data from 456 Sentinel-1 Level-2 Ocean (OCN) products, were obtained for a time frame from June 1, 2017 until May 31, 2019. These were acquired in Interferometric Wide (IW) Swath mode under Vertical-Vertical (VV) + Vertical Horizontal (VH) dual polarisation operation.

In-situ wind speed observations from 5 Cyprus coastal meteorological stations located at Limnitis, Pafos, Akrotiri, Larnaca and Famagusta areas were retrieved online from the NCEI GIS Map Portal available at: https://gis.ncdc.noaa.gov/maps/ncei/cdo/hourly. The aerodrome routine meteorological reporting format is coded as FM-15, while observations coming from a fixed land station, either manned or automatic, as FM-12, and are widely known as METAR and SYNOP, respectively. To form a consistent wind time-series, only the METAR weather information (FM-15) was used. This led to a 2-year time-series of in-situ measurements identical to the Sentinel data time frame. The Sentinel tiles used along with the location of the 5 coastal weather monitoring stations are depicted in Figure 1.

Nearest neighbour resampling was used for both UERRA and Sentinel datasets, in order to bring all the information at a common basis; a regular square grid. A maximum distance of 1 pixel (~1km) was set for the resampling process to prevent long distance allocation of Sentinel-1 pixel values to the grid. The regular grid bounding box is shown with white outline in Figure 1.

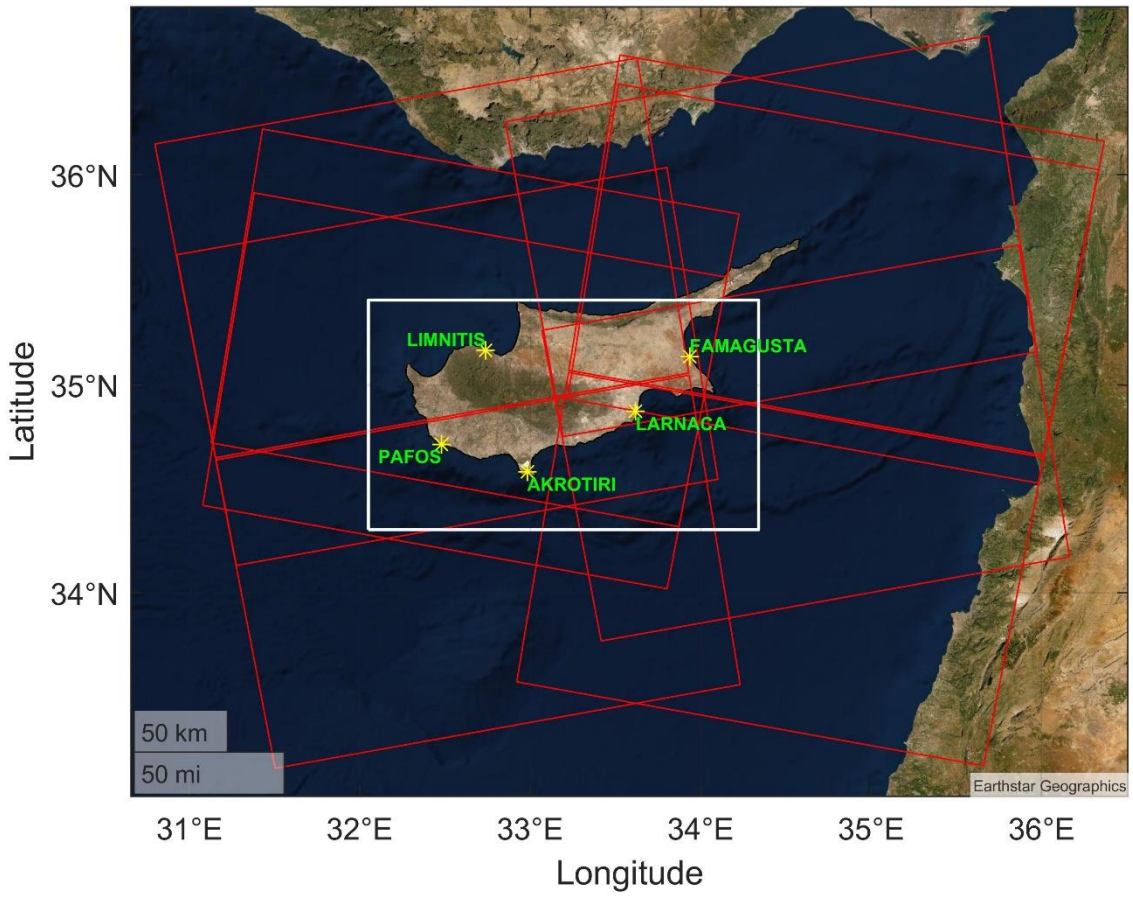


Figure 1: Outline of the study area (white polygon), Sentinel tiles (red polygons) and weather monitoring stations

2.2. Weibull distribution fitting

In the context of preliminary wind resource assessments, one is mainly interested on annual averages (i.e. wind power density annual average) rather than knowing the wind speed values at particular time instances. In order to derive such statistics, empirical distribution functions are typically fitted to the time-series of wind speed data available at each grid cell or pixel. The Weibull probability density function (PDF) has been widely used to fit wind speed distributions for wind energy applications as it appears to be related to the nature of the wind in certain conditions. The theoretical PDF is fitted to the sample wind speed data by estimating the parameters, scale (a) and shape (b), of the Weibull distribution, so that some measure of agreement between the model-derived and the sample statistics (or quantiles and/or probabilities) is maximized. Weibull distribution fitting was conducted via the method of maximum likelihood using all the available coarse (UERRA) and fine (Sentinel) spatial resolution samples over the period of interest for each grid cell/pixel. Once the Weibull distribution is fitted to the data and the corresponding parameters have been calculated, a measure of uncertainty can also be obtained by using bootstrap to resample with replacement from each cell's/pixel's Weibull distribution. The corresponding standard deviation maps of for each parameter obtained via bootstrap reveal that the variability around Weibull scale and shape parameter values is spatially varying across the area of interest. In addition, that variability is larger for the Sentinel than for the UERRA time-series data due to the much smaller data extent of the former.

Weibull distribution functions were also fitted to the coastal meteorological stations' time-series to allow for comparison with the distributions of the closest UERRA node and Sentinel pixels lying at most 10km from each station. The gridded estimates were also adjusted to the station height prior to the analysis. The first step of the comparison workflow includes interpolating the Weibull parameters from stations to 1km x 1km cell with 100m discretization and 11km x 11km cell with 1 km discretization along each dimension, both using the correlogram of the corresponding Sentinel scale and shape Weibull parameters. The interpolated Weibull parameters within the discretized cells are then averaged to obtain the area-adjusted parameters around the station locations. Following that, the bias for the UERRA and Sentinel scale and shape Weibull parameters is computed as a weighted average of the differences between the area-adjusted station and the nearest UERRA and Sentinel parameters. The weights are calculated in terms of the proportion of the number of data available at the meteorological stations, UERRA cells and Sentinel pixels. The spatial distribution of the Weibull-derived parameters after the bias correction is depicted in Figure 2.

3. Methodology

3.1. Downscaling UERRA data using Sentinel data and Area-to-Point Kriging with External Drift

To generate thematic information at a finer scale resolution from coarse scale datasets, various downscaling algorithms have been proposed and applied to spatial downscaling of various environmental parameters such as precipitation (Immerzeel et al., 2008; Jia et al., 2011; Park, 2013; Park et al., 2016) land surface temperature (Hutengs and Vohland, 2016; Yoo et al., 2017) and evapotranspiration (Ke et al., 2016). Particularly, among spatial downscaling algorithms, Kyriakidis (2004) proposed area-to-point (ATP) kriging interpolation, a modification of conventional univariate kriging, which can properly account for variations within coarse scale blocks. The objective of area-to-point spatial interpolation is to predict any unknown point value using the areal data available on a coarse level. The prediction locations are arbitrary, that is, they need not comprise a regular grid and they can lie inside or outside any areal support. (Park, 2013)) extended ATP kriging to a multivariate framework by combining regression-based trend estimates with ATP kriging-based residual correction.

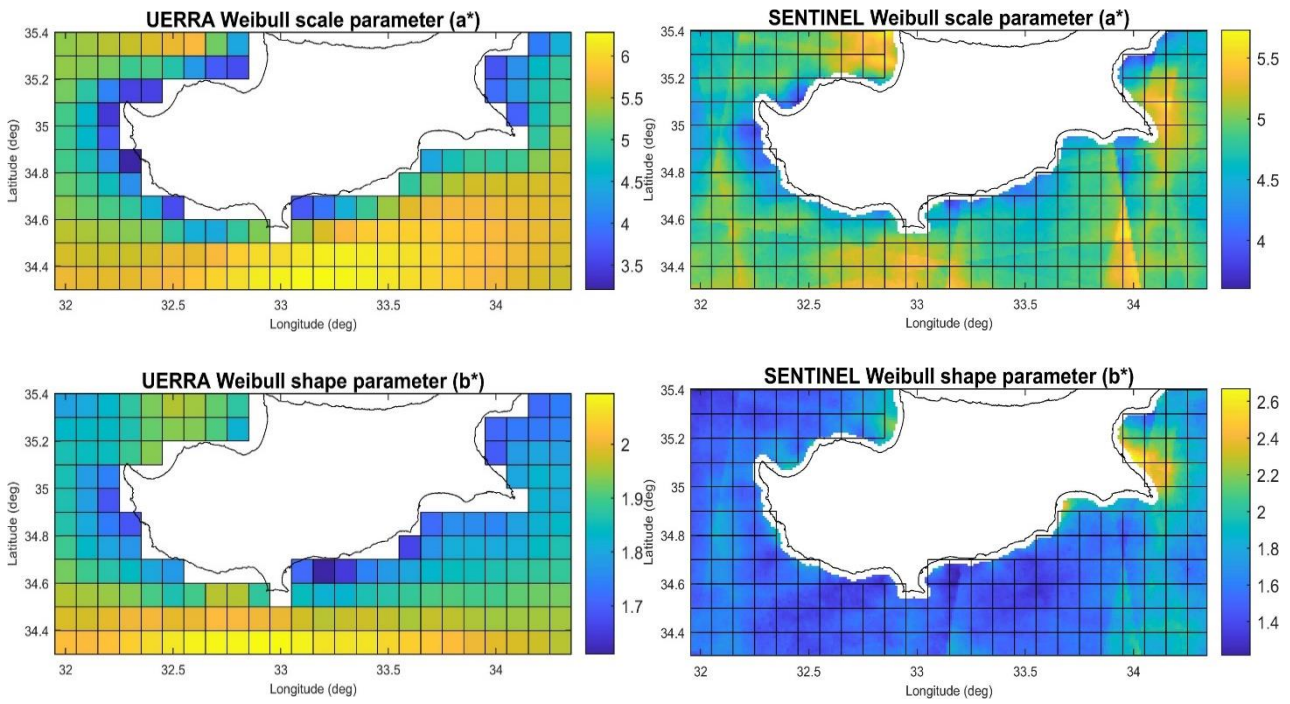


Figure 2: UERRA and Sentinel Weibull scale (a) and shape (b) parameters after bias correction

In this work an extension of ATP downscaling is adopted, namely Area-To-Point Kriging with External Drift (ATPKED), which is a downscaling algorithm theoretically based on the component decomposition of the attribute of interest into deterministic trend and stochastic residual components (Immerzeel et al., 2008). The deterministic trend components are estimated using regression modeling based on the statistical relationships between input coarse scale data and auxiliary variables available at a finer scale. The stochastic residual components that cannot be explained by the auxiliary variables are estimated using ATP kriging. The final downscaling result is obtained by adding the two components; the results satisfy the coherence property and reproduce the original coarse scale block values when upscaled.

The coarse spatial resolution (UERRA) Weibull parameters are downscaled using the fine spatial resolution (Sentinel-1) Weibull parameters as auxiliary variables. This is achieved by substituting the actual wind speed values at the coarse and fine spatial resolution level with the UERRA and Sentinel Weibull scale and shape parameters. It should be noted, however, that both Weibull a and b parameters are used as explanatory variables in the regression models in this case. This is proved to lead to statistically significant results in contrast to using only a single parameter each time.

To perform ATPKED downscaling, Sentinel fitted a, and b Weibull parameters are initially upscaled to the coarse spatial resolution (11km) of UERRA reanalysis dataset. A regression model, with dependent variables the UERRA a, b Weibull parameters and independent variables the upscaled Sentinel a, b Weibull parameters is fitted, and the residuals are calculated. The variogram of a, b Weibull residuals is deconvolved or de-regularized, to estimate residual variograms of a, b Weibull parameters at the fine spatial resolution (1km) of Sentinel images. Lastly, Area-to-Point Kriging with External Drift (ATPKED) is performed within local neighborhoods, to compute downscaled a, b Weibull parameters and their prediction standard errors at the fine spatial resolution (1km). The applied ATPKED method ensures that the area-average of downscaled Weibull parameters within UERRA cells reproduce the coarse corresponding resolution UERRA Weibull parameters up to a measure of uncertainty of the latter (here derived via bootstrap). The flowchart presented in Figure3 is depicting the above described steps of the ATPKED methodology.

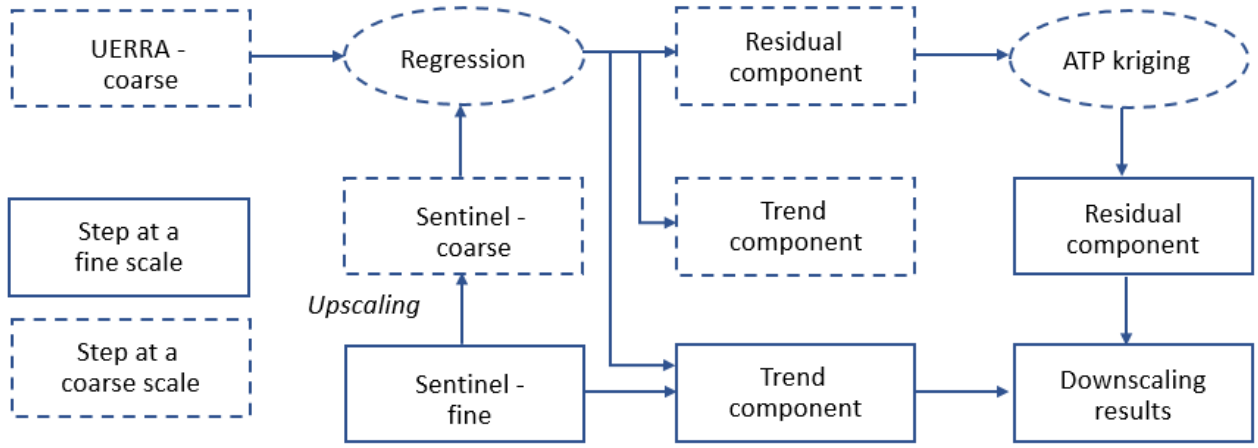


Figure 3: Flowchart of ATP Regression Kriging (ATPKED) Spatial Downscaling Algorithm

Lastly, the wind power density (WPD) is calculated using the ATPKED downscaled parameters. The variability of Sentinel-1 wind speed data provides some added value regarding the offshore wind power estimates and the current approach could lay the groundwork for a spatially finer resolved offshore wind assessment.

3.2. Uncertainty propagation of the downscaling model results

The uncertainty in wind power density (WPD) is additionally examined within a Monte Carlo context via the use of geostatistical simulation. Consensus fusion (Deutsch and Zanon, 2004; Doyen et al., 1996) is adopted to estimate a local distribution at each 1km pixel. The mean of that local distribution is a weighted combination of UERRA a, b Weibull parameters downscaled using Area-to-Point Ordinary Kriging (ATPOK) and Sentinel a, b Weibull parameters. The variance of that local distribution is a combination of error-variances attached to each source of information (i.e., ATPOK variance for downscaled UERRA, bootstrap variance for Sentinel). Standard Gaussian unconditional realizations are simulated (via Cholesky decomposition) for each Weibull parameter, using the corresponding Sentinel-derived correlogram/variogram model. Each pair of simulated Gaussian deviates at each 1km x 1km pixel is transformed into (cross)correlated realizations to account for the fact that Weibull a and b parameters exhibit significant (cross)correlation.

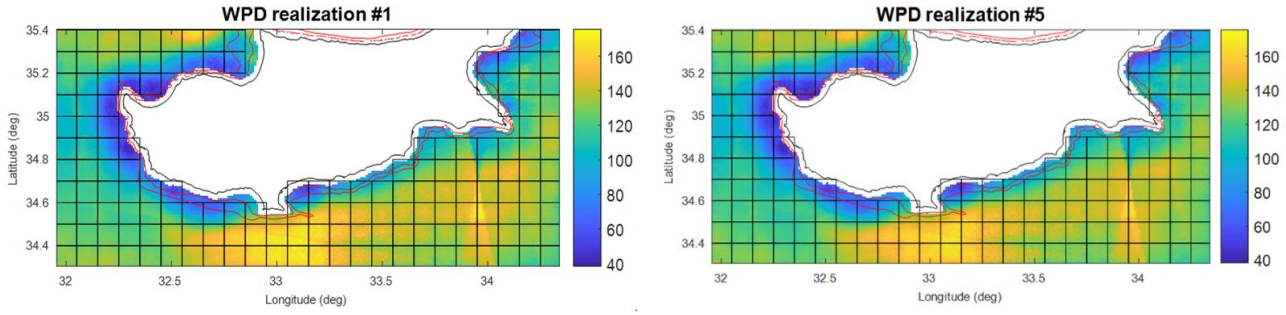


Figure 4. Two WPD realizations computed by the respectful simulated Weibull a and b parameters

Weibull a and b parameter values are simulated at each location by multiplying each simulated Gaussian deviate with the corresponding local standard deviation and adding the corresponding local mean; spatial correlation in simulated Gaussian deviates induces spatial correlation in simulated Weibull parameters. Finally, the simulated WPD is computed at each 1km x 1km pixel using pairs of simulated Weibull a and b parameter values; Figure 4 depicts two realizations of WPD.

4. Results

The inclusion of Sentinel-1 fine resolution information in the downscaling method employed in this work furnishes important local variability in the downscaling endeavor that is otherwise missing from downscaling UERRA data alone. The downscaled a, and b Weibull parameter predictions computed by ATPKED are depicted in Figure 5.

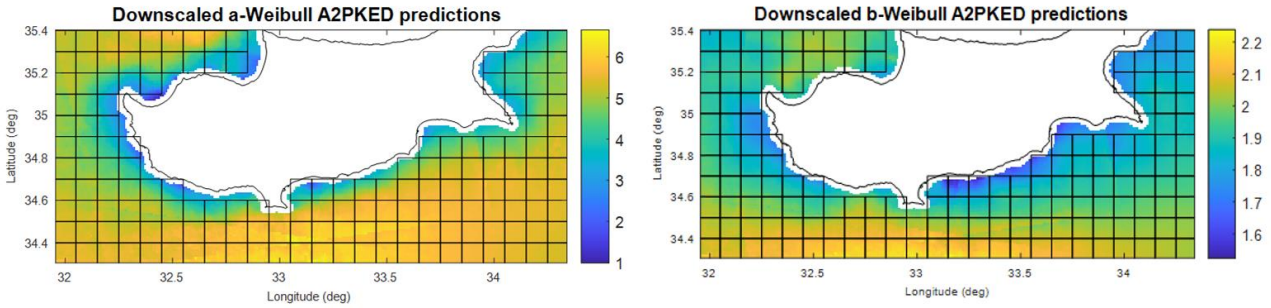


Figure 5. Downscaled a, and b-Weibull parameter estimation computed by ATPKED

Similarly, wind power density spatial patterns derived by the above described ATPKED downscaling method tend to be less smooth compared to downscaling methods only incorporating UERRA coarse wind data. Moreover, the uncertainty analysis conducted via the simulation of Weibull a, and b parameters results in an ensemble of simulated wind power density realizations; Figure 6 depicts the ensemble WPD average and standard deviations of the simulated WPD values at each 1km grid node.

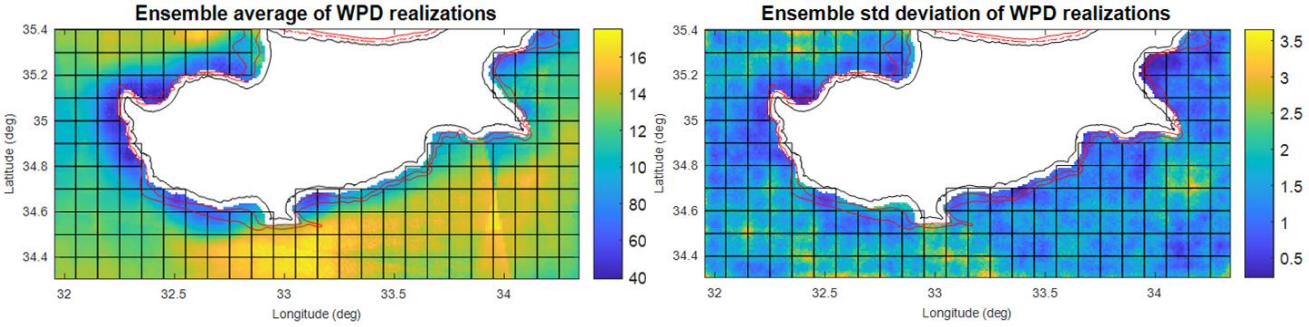


Figure 6. Ensemble average (left) and std deviation (right) of the WPD realizations

A measure of uncertainty is obtained via the downscaled estimates allowing to report the deviation around the wind power density values which does not seem to be particularly high in this case (ranges between 0-3.5 W/m^2). The proposed geostatistical uncertainty analysis may lead to risk conscious policy making for a potential future wind farm allocation scheme for the offshore area of Cyprus.

5. Discussion and Conclusions

The analysis and modeling based on satellite products depend heavily on spatial and temporal resolutions of such products. Using satellite products (Sentinel-1 wind data in this project) with appropriate spatial resolutions is crucial for both global/regional and local analyses. The revisiting time or temporal resolution is also important in the proper selection of satellite products for periodic monitoring. Periodic environmental monitoring and time-series analysis are possible using reanalysis products (UERRA wind data in this work) owing to their high temporal resolutions. However, the spatial resolution of reanalysis products is too coarse to perform analysis at a local scale.

Spatial downscaling is often applied to coarse scale satellite products with high temporal resolution for environmental monitoring at a finer scale. This work describes the implementation of area-to-point Kriging with External Drift (ATPKED) algorithm for the downscaling of Weibull a, and b parameters derived by UERRA coarse scale data, also accounting for the Sentinel data available at a fine scale resolution. The ATPKED algorithm combines regression modeling and residual correction with area-to-point kriging. The employed geostatistical downscaling of coarse scale Weibull parameters demonstrates that ATPKED algorithm generates downscaling results in which overall variations/patterns in input coarse scale data are preserved and local details are also well captured. Although the applicability of ATPRK was tested on regular raster data such as satellite Sentinel products, the tool is flexible and can be applied to spatial data with irregular shapes because any objects with irregular shapes are discretized by internal points.

Additionally, downscaled UERRA and Sentinel a, b Weibull parameters are employed in a Monte Carlo framework for assessing the uncertainty in spatial distribution of Wind Power Density (WPD) derived from the downscaled Weibull distributions. In this approach, the integration two datasets is achieved via the use of consensus fusion; the local distribution is described by the weighted average of the two levels of information and the variance by the combination by the Kriging variance for downscaled UERRA, and the bootstrap variance for Sentinel. Such simulated realizations can be used to assess the uncertainty in WPD values linked to a potential future wind farm allocation scheme in the offshore areas of Cyprus.

References

Al-Yahyai, S., Charabi, Y., Al-Badi, A., and Gastli, A. (2012). "Nested ensemble NWP approach for wind energy assessment." Renewable Energy, 37(1), 150–160.

Beaucage, P., Brower, M. C., and Tensen, J. (2014). "Evaluation of four numerical wind flow models for wind resource mapping." Wind Energy, John Wiley & Sons, Ltd, 17(2), 197–208.

Council, G. W. E. (2020). GWEC Global Wind Report 2021. Wind Energy Technology.

Deutsch, C. V., and Zanon, S. D. (2004). "Direct prediction of reservoir performance with Bayesian updating under a multivariate Gaussian model." Canadian International Petroleum Conference 2004, CIPC 2004, Petroleum Society of Canada (PETSOC).

Doyen, P. M., den Boer, L. D., and Pillet, W. R. (1996). "Seismic Porosity Mapping in the Ekofisk Field Using a New Form of Collocated Cokriging." All Days, SPE.

González-Aparicio, I., Monforti, F., Volker, P., Zucker, A., Careri, F., Huld, T., and Badger, J. (2017). "Simulating European wind power generation applying statistical downscaling to reanalysis data." Applied Energy, Elsevier Ltd, 199, 155–168.

Guo, Z., Chang, C., and Wang, R. (2016). "A novel method to downscale daily wind statistics to hourly wind data for wind erosion modelling." Communications in Computer and Information Science, Springer Verlag, 611–619.

Hewitson, B. C., and Crane, R. G. (1996). "Climate downscaling: Techniques and application." Climate Research, Inter-Research, 7(2), 85–95.

Hong, S.-Y., and Kanamitsu, M. (2014). "Dynamical Downscaling: Fundamental Issues from an NWP Point of View and Recommendations." J. Atmos. Sci, 50(1), 83–104.

Hutengs, C., and Vohland, M. (2016). "Downscaling land surface temperatures at regional scales with random forest regression." Remote Sensing of Environment, Elsevier Inc., 178, 127–141.

Immerzeel, W. W., Rutten, M. M., and Droogers, P. (2008). "Spatial downscaling of TRMM precipitation using vegetative response on the Iberian Peninsula." Remote Sensing of Environment, 113, 362–370.

Javad Alizadeh, M., Reza Kavianpour, M., Kamranzad, B., and Etemad-Shahidi, A. (n.d.). "A Weibull Distribution Based Technique for Downscaling of Climatic Wind Field."

Jia, S., Zhu, W., Lu, A., and Yan, T. (2011). "A statistical spatial downscaling algorithm of TRMM precipitation based on NDVI and DEM in the Qaidam Basin of China." Remote Sensing of Environment, 115(12), 3069–3079.

Katsaprakakis, D. A. (2016). "Energy storage for offshore wind farms." Offshore Wind Farms: Technologies, Design and Operation, Elsevier Inc., 459–493.

Ke, Y., Im, J., Park, S., and Gong, H. (2016). "Downscaling of MODIS One Kilometer Evapotranspiration Using Landsat-8 Data and Machine Learning Approaches." Remote Sensing, MDPI AG, 8(3), 215.

Kumar, J., Brooks, B.-G. J., Thornton, P. E., and Dietze, M. C. (2012). "Sub-daily Statistical Downscaling of Meteorological Variables Using Neural Networks." Procedia Computer Science, 9, 887–896.

Kyriakidis, P. C. (2004). "A Geostatistical Framework for Area-to-Point Spatial Interpolation." Geographical Analysis, Wiley, 36(3), 259–289.

Park, N. W. (2013). "Spatial downscaling of TRMM precipitation using geostatistics and fine scale environmental variables." Advances in Meteorology, 2013.

Park, N. W., Hong, S., Kyriakidis, P. C., Lee, W., and Lyu, S. J. (2016). "Geostatistical downscaling of AMSR2 precipitation with COMS infrared observations." International Journal of Remote Sensing, Taylor and Francis Ltd., 37(16), 3858–3869.

Pillai, A. C., Chick, J., Khorasanchi, M., Barbouchi, S., and Johanning, L. (2017). "Application of an offshore wind farm layout optimization methodology at Middelgrunden wind farm." Ocean Engineering, Elsevier Ltd, 139, 287–297.

Shin, J.-Y., Jeong, C., and Heo, J.-H. (2018). "A Novel Statistical Method to Temporally Downscale Wind Speed Weibull Distribution Using Scaling Property." Energies, MDPI AG, 11(3), 633.

Winstral, A., Jonas, T., and Helbig, N. (2017). "Statistical downscaling of gridded wind speed data using local topography." Journal of Hydrometeorology, American Meteorological Society, 18(2), 335–348.

Yoo, C., Im, J., Park, S., and Cho, D. (2017). "Thermal Characteristics of Daegu using Land Cover Data and Satellite-derived Surface Temperature Downscaled Based on Machine Learning." 대한원격탐사학회지, The Korean Society of Remote Sensing, 33(6), 1101–1118.

CAN RADIOMETRIC DATA IMPROVE LITHOLOGY MAPPING AND GEOLOGICAL UNDERSTANDING THROUGH UNSUPERVISED CLASSIFICATION?

Zeinab Smillie (1)* - Vasily Demyanov (1) - Jennifer Mckinley (2) - Mark Cooper (3)

Heriot-watt University, Institute of Geoenergy Engineering (1) - Queen's University, School of Natural and Built Environment (2) - British Geological Survey (3)
** Corresponding author: z.smillie@hw.ac.uk*

Abstract

Pattern classification algorithms can enhance the pattern recognition and prediction of large multivariate data sets, that otherwise would be difficult to detect. These techniques can be used to visualise the contribution or role of various features in shaping the patterns of a large data set.

Self-organising map (SOM) is an unsupervised classification tool that is trained by competitive learning. The method is useful in analysing and visualising high-dimensional data, based on principles of vector quantification of similarities and clustering in a high-dimensional space. The method can be used to perform prediction, estimation, pattern recognition of large data sets.

One main advantage of the SOM is that it can be applied to categorical and continuous variables making the tool ideal for analysing a complex combination of geological feature such as rock classifications, ages, geochemical composition, terrain elevations, etc.

We here employ the tool to predict geological features using geophysical data, mainly the airborne geophysical data acquired through the Tellus project 2011/12. Tellus radiometric data present a high-resolution data set (Line spacing of 200 m and point spacing of 60 m). The data characterise the K, U and Th distribution associated with the natural geological features in Northern Ireland.

The SOM of the radiometric data displayed patterns that are evidently associated with both bedrock and superficial geology. However, the addition of other natural features, such as terrain elevations modifies the clarity of the clusters and contribute to the prediction of geological formations.

The SOM enhances the visualisation and recognition of the signals of geochemical variations within the bedrocks, although now concealed with superficial deposits. These advantages of SOM, combined with the high-resolution nature of the radiometric data input, presents an efficient tool to improve or complement conventional geological mapping techniques especially for "hard to recognise" stages of igneous rock emplacements, rock mass zonation and alteration/contact zones and also provides fundamental attempt toward understanding geological processes.

ON THE USE OF ARTIFICIAL NEURAL NETWORKS TO IDENTIFY RELATIONSHIPS AMONG NEARBY RAINFALL STATIONS TO INFER PAST RAINFALL DATA

Ioannis Trichakis (1)* - Nikolaos Kyriakou (1) - George Karatzas (1)

Technical University of Crete, Environmental Engineering (1)
** Corresponding author: itrichakis@isc.tuc.gr*

Abstract

With the advancement of the meteorological sensors, many more weather stations are now installed in a far denser network than in the past. This creates the question at some points, if it is possible to infer past values of currently installed stations, before their installation in the field. Artificial neural networks (ANNs) have been deployed to search for relationships among nearby rainfall stations in the island of Crete, Greece. During the ANN training, observed data from a newly installed station served as target data for the ANN, while observed data from nearby stations constituted the input data for the network. The results show that the ANN is able to find complex relationships among many different stations and a comparison proves its ability to outperform the simple regression models. The calibrated ANN has the added value of being able to infer missing data values even in case the new station has a failure for some days.

HELP: THE SANDBOX HAS BECOME CONTAMINATED

J. Jaime Gómez-Hernández (1)* - Zi Chen (1) - Teng Xu (1) - Andrea Zanini (2)

Universitat Politècnica de València, Institute for Water and Environmental Engineering (1) – University of Parma, Department of Engineering and Architecture (2)
** Corresponding author: jgomez@upv.es*

Abstract

The identification of a pollutant source together with the underlying distribution of hydraulic conductivity from state variable observations in time is a difficult task. It has been solved in synthetic aquifers under strictly controlled conditions using the restart normal-score ensemble Kalman filter. In this study, we present an application of this technique to a laboratory sandbox experiment. Our expectation was that the methodology would work smoothly and both source and conductivity heterogeneity would be easily identified, several problems arose related with the not-fully-controlled analog set up. The problems were circumvented, and the technique was able to perform its purpose. The specifics of the problems encountered, and their solution will be discussed.

COMBINING FLOW AND TRANSPORT NUMERICAL MODELING AND GEOSTATISTICS TO IMPROVE THE ASSESSMENT OF GROUNDWATER CONTAMINATION: AN APPLICATION TO THE CHERNOBYL SITE

Léa Pannecoucke (1) - Mathieu Le Coz (2)* - Xavier Freulon (1) - Chantal De Fouquet (1)

Mines Paristech - Psl University, Centre de Géosciences (1) - Irsn, Pse-env Sedre/leli (2)
** Corresponding author: mathieu.lecoz@irsn.fr*

Abstract

Following the Chernobyl nuclear power plant accident in 1986, the "Red Forest" (irradiated forest surrounding the plant) was bulldozed and buried in several trenches dug in the unsaturated part of a sandy aquifer. In 1999, a pilot site was built to monitor the migration of radionuclides from the T22 trench to the underlying saturated zone. Piezometers enable measurements on groundwater once or twice a year, showing a plume of strontium-90 that extends to more than 20m downstream from the T22. However, the estimation of the amount of strontium-90 released to the groundwater remains a complex issue, especially due to the high variability and non-stationarity in the measured activity in strontium-90 both in time and space. The aim of this study is to estimate the activity in strontium-90 at several dates on a 2D cross-section of the aquifer, based on multi-level measurements from six piezometers.

The inference of geostatistical non-stationary models generally requires a lot of observations, which are not available here. However, physical processes governing strontium-90 migration may be characterized and modeled. Therefore, the lack of observations is compensated by numerical modeling of flow and transport. Two non-stationary methods are implemented and compared to stationary ordinary kriging (OK): kriging with an external drift (KED), using average simulations as a drift, and kriging with numerical variograms (KNV) computed from flow and transport models (Pannecoucke et al., 2020), which does not rely on any stationarity assumption. In order to implement KED and KNV, a set of about 200 flow and transport simulations taking into account the uncertainties regarding the source term and the hydraulic parameters are run. On the one hand, the mean of the simulation outputs is computed and used as an external drift in KED. On the other hand, non-stationary empirical variograms of the activity are computed and used in a kriging procedure. The performances of the three approaches are compared by cross-validation.

The two approaches combining geostatistics and numerical modeling of flow and transport give more accurate estimates of activity especially when the number of available measurements decreases. For example, when the measurements from 3 out of 6 piezometers are kept, the mean absolute error is reduced by 1.6 when using KNV instead of OK and by 1.4 when using KED instead of OK. When the measurements of 5 piezometers are kept, the errors are only reduced of 1.1 for KNV compared to OK, and are the same for KED and OK. Those results underline the advantage of combining numerical modeling and geostatistics. A more precise comparison between KED and KNV shows that KED is more robust when less than a hundred simulations are available. Finally, the methods developed in this work are very general and can be used in various context, such as nuclear sites monitoring and dismantlement, but also for other types of pollution, as soon as a numerical code that simulates the studied phenomenon is available.

This work is part of Kri-Terres project, supported by the French National Radioactive Waste Management Agency under the "Investments for the Future" program.

USING GEOSTATISTICAL METHODS TO HELP OPTIMIZING AN EXISTING GROUNDWATER MONITORING NETWORK

Nathalie Courtois (1)*

French Alternative Energies and Atomic Energy Commission (cea), Laboratory of Modelling of Transfers in Environment (lmte) (1)

** Corresponding author: nathalie.courtois@cea.fr*

Abstract

The studied site is a research center, located in the South of France. Three superimposed aquifers are in presence in, from surface to the bottom, Quaternary, Miocene and Cretaceous Formations. Since its creation in the 60's, the center started to constitute a groundwater monitoring network dedicated to the survey of both groundwater levels and quality which has continuously evolved as a function of different needs: survey of new facilities, knowledge of flow rate and directions, improvement of the 3D hydrogeological model used for flow and transport simulations, etc. The monitoring network is now composed of about 400 wells distributed in the three superimposed aquifers. Geostatistical methods are used to help optimizing this network in terms of number and spatial distribution of the wells.

An original and specific geostatistical methodology is developed. First, variograms are calculated on hydraulic heads surveys at different dates, covering a large panel of hydrological conditions. Corresponding head distributions are then constructed by kriging. For some aquifers, as hydraulic heads and elevations are correlated, a smoothed digital elevation model is used as external drift. Then, trajectories starting from specific zones (facilities, buildings, etc.) are calculated, in order to highlight the downstream positions. Finally, a network optimization is conducted in two parts: (i) sequential addition of new wells, allowing to decrease the uncertainty on hydraulic head in zones with few information, (ii) sequential removal of existing wells, on a criteria of geometrical redundancy. During the calculation process, several constraints are imposed such as a minimal thickness of geological formation to add a new well and a minimal distance to existing or added wells. This sequential and automated process allows testing different configurations (number of additions/removals, minimal distance between wells, etc.).

As a result, the study leads to an optimized list of new wells to add and existing wells to remove for each of the three aquifers. This list is a precious base for optimization that has to be further discussed, taking into account other criteria that cannot be included in the geostatistical analysis, e.g. presence of faults, available space, access conditions for drilling machines, etc.

THE PLAN.T.E PROJECT: AN AFRICAN MISSING LINK TO FIGHT DESERTIFICATION

Patrick Pierron (1)* - Pascal Bernasconi (1)

Geo-csp Sas (1)
** Corresponding author: pp@geo-csp.fr*

Abstract

Facing desertification is a pregnant priority for humankind. Since the early 30's its spreading mechanism has been unveiled by Pr. Henri Erhart who theorized pedogenesis as a geological phenomenon ("Biostasie" et "rhexistasie": Esquisse d'une théorie sur le rôle de la pédogénèse en tant que phénomène géologique /Paris, 1955). When considering the Bio-Rhexistasy cycle, it is noticeable that a strong unbalance prevails between a slow colonisation rate of soil creating vegetation, and abrupt reversing conditions of soil erosion when the vegetation hold is lost. When seeking for remediation in reforesting and re-cultivating arid lands, the crucial question is how to restore the water retention capacity of surface formations, essential for plants to survive. In most desert landscapes very altered soil conditions prevail; in fact former elaborated soils have generally been washed away by scarce torrential rains, leaving lateritic cuirass or barren rock grounds. The basics of the PLAN.T.E project are to remediate the consequent lack of water retention capacity by creating local systems of small reservoirs able to collect a significant part of the impluvium, directly and through local run off concentration, then retain appreciable amounts of shallow depth water out of reach of intense evaporation, but still available for plants. PLAN.T.E is an imperative that can be read as "PLANt Trees with Explosives". The method is known as "soft cratering", implementing low energy explosives (Anfo) classified for agricultural use and allows to produce from appropriate blasting in 5-7 m deep drilled holes a series of cones of crushed rock with a high fracture porosity. When applied to impervious rocks like laterites, marls or shales, individual reservoirs with up to 25 m^3 capacity are made available. Success in planting trees of selected resilient stocks in these locations is then to be achieved through appropriate fertilisation with organic compost and adapted watering till self-sufficiency, when roots gain access to the eventual water reserve. This application of a mining method to fighting desertification is innovative, and thought to provide to most of the fragile ecosystems, that currently vanish in arid regions due to the double penalty of adverse climatic drift and increasing human pressure, the necessary impulse for a medium term recovery and long term prosperity. The method is also applicable to enhance infiltration in clogged recent sedimentary floodplains, with a double benefit in aquifer recharge in terms of pumpable volumes and lowered salinity. The first pilots are being scheduled in southern Tunisia with applications in creating low cost reservoirs, enhanced deep aquifer recharge from flood spreading racks, and of course direct trees plantation. Besides the technical presentation of the method, our contribution mostly focuses on the framing conditions of the project PLAN.T.E, through specific studies and partnerships that back the first projected developments and illustrate our collaborative strategy to rally expertise and funding.

A GEOSTATISTICAL DATA FUSION APPROACH FOR PROBABILISTIC ASSESSMENT OF WATER TABLE DEPTH RISKS USING MULTI SOURCE DATA

Rodrigo Lilla Manzione (1)* - Annamaria Castrignanò (2)

Unesp - São Paulo State University, School of Sciences and Engineering (1) - Cra - Council for Agricultural Research and Economics (2)

** Corresponding author: rlmanzione@gmail.com*

Abstract

Extreme water table depths can affect plants water supply, agricultural crops development, field machinery and alter soil-water relationships causing environmental changes. In general, water table depths risks are estimated from monitoring networks that provide scarce and irregular data. The use of exhaustive data derived from satellites images can not only improve these estimatives but also incorporate addition physical knowledge about the process in progress. When combined, environmental, agricultural and geotechnical variables can describe possible states of a certain phenomenon, which are interpreted as spatial random variables. Multiple stochastic realizations of spatial variables can form the basis of risk assessment. Treating these realizations as possible realities, risk assessment consists essentially in calculating the frequency (probability) with which specified criteria are exceeded or fail to be met. The aim of this paper is to describe a geostatistical methodology for data fusion in order to estimate risks of water table depths and exemplify with a case study in a Cerrado conservation area in Águas de Santa Barbara/SP-Brazil. The risk of water table depth was defined using critical thresholds. Mean water table depths observed at 56 wells during 2016-17 hydrological year and exhaustive TanDEM-X elevation data with 90m resolution were utilized and combined with prior information from homogeneous zones delineated for each risk category. Using sequential indicator simulation, the realizations were post processed calculating the probability of each category and corrected for local probabilities. The final maps present the most probable corrected category and were evaluated by entropy calculation. Cerrado restoration strategies from water table depths risks at this area were discussed, showing that areas nearby watersheds divisors and in the north part of the region present low risk of shallow water levels which could potentially influence plants adaptation.

HIGH FREQUENCY OXYGEN DATA ASSIMILATION IN WATER QUALITY ASSESSMENT

Thomas Romary (1)* - Shuaitao Wang (1) - Nicolas Flipo (1)

Mines Paristech, Geosciences (1)
** Corresponding author: thomas.romary@mines-paristech.fr*

Abstract

The coupling of high frequency data of water quality with physically based models of river systems is of great importance for the management of urban socio-ecosystems. The hydro-biogeochemical model Prose has been under development for several decades in Mines ParisTech center of geosciences. It consists of three modules: hydrodynamics, transport and biogeochemistry and simulates the metabolism of river systems. In the meantime, a high frequency sensors network measuring the dissolved oxygen (DO) concentration in the Seine River has been deployed. The calibration of the Prose model to the DO data is made difficult by the large number of, possibly dynamic, parameters involved. These parameters are both physical and related to micro-organisms (reaeration coefficient, photosynthetic parameters, growth rates, respiration rates and optimal temperature). Focusing on the most influential parameters, selected through a sensitivity analysis of the biogeochemical module C-Rive, we develop a sequential data assimilation approach based on a particle filtering algorithm. It is able to reproduce the observed DO concentrations while characterizing the dynamics of the parameters governing the Seine river metabolism. In particular, we show that the physical, bacterial and phytoplanktonic parameters can be retrieved properly.

SPATIAL DISPERSION OF A FIELD IN AN AREA IN DEPENDENCE OF ITS SIZE

Peter Bossew (1)*

German Federal Office for Radiation Protection (1)
** Corresponding author: pbossew@bfs.de*

Abstract

Knowing the expected spatial dispersion of a field within an area is important for planning surveys and sampling campaigns, because the sample size required to determine the mean (or other statistics) with given precision depends on dispersion. The common question of a survey planner is thus: Which sample size do we need? If dispersion is not known, usually a pilot survey is performed to get an estimate of that quantity. In other cases, such as in radon surveys, it is relatively well known for certain size of spatial unit. However, what is required, is a function that describes dispersion in dependence of area size.

According to Tobler's First Law of Geography ("everything is related to everything else, but near things are more related than distant things"), one must expect that dispersion of a field increases, in average, with size of a spatial unit, or more precisely, with mean separation between random locations within the unit. This is reflected by the shape of variograms.

In this paper, we discuss how variograms and areal dispersion are related, including a short excursion about distribution and expected length of random segments within an area. As the main measure of dispersion, the geometrical standard deviation is used, but also other quantities are available. For empirical fields, dispersion-area functions derived from variograms are compared with ones extracted from data directly by computing dispersions within windows of varying size. As another example, for the European Indoor Radon Map, which consists of a dataset for measurements aggregated into 10 km x 10 km cells, the function is recovered by re-aggregation of cells and re-calculating dispersion.

Keywords: Spatial dispersion, survey design, ambient radon´

1. Introduction

Among tasks in environmental sciences is surveying areas for quantities indicating pollution, hazard or mineral resources. Often one is interested in the spatial distribution or pattern, but equally commonly in the mean value or other statistics which are compared to reference levels in order to decide about certain action. The straight forward way is sampling. Through statistical procedures the wanted statistic is computed together with its uncertainty, usually expressed as confidence intervals (or their Bayesian cousins).

There are two different approaches. In a *design-based approach*, one generates a sample from which the wanted statistic can be directly estimated, such as the expectation (true mean) as the arithmetical mean. The sampling design must be such that pre-set scores of accuracy and precision are met. In essence, the former is a condition on representativeness of a sample, the latter on the sample size. Assuring representativeness may be a difficult task in real-world cases, as it means generating a sample whose statistical distribution equals the (unknown) true one; commonly achieved by drawing randomly from the population; very difficult to succeed e.g. in radon surveys.

The alternative is a *model-based approach*, which is more relaxed in respect of spatial sample design, i.e. representativeness of the draw. Applying a statistical model, possibly exploiting the autocorrelation structure of the sample allows estimating the spatial pattern, i.e. producing a map, from which the wanted statistic can be derived. The drawback against the design-based approach is that the model induces additional model-related uncertainty components, that its practical implementation may be nontrivial, as may be estimation of uncertainty of derived statistics.

If previous knowledge is available, e.g. through pilot surveys, optimal sample designs can be generated in a Bayesian spirit by iteratively setting additional sample points or relocating them until the required precision is achieved. The standard textbook is Müller (2007).

However, here we focus on the design-based approach because it is so frequently used for its intuitive simplicity. Assuming that representativeness can reasonably be achieved, the remaining task is determining the sample size. Clearly it depends on the statistical dispersion of data: the higher dispersion, the larger a sample necessary to achieve a required precision, e.g. of the mean. The question is therefore, how to estimate it from previous generic knowledge.

2. Theory

2.1. Minimum sample size and dispersion

Simple reasoning leads to the so-called Hale formula: Assume a log-normal process, Z~LN(μ,σ). Then Y:=ln(Z)~N(μ,σ). Let s the standard deviation (SD) of a sample $\{y_1,...,y_n\}$, size n, of Y, m its arithmetical mean. The SD of the mean, or standard error (SE) equals $s/\sqrt{n}$. As approximation, substitute s by an anticipated σ. The approximate confidence limits of the mean are $CL(low,high)=(m\mp x_\alpha\sigma/\sqrt{n})$. As a constraint on the precision of μ, the quantity $x_\alpha\sigma/\sqrt{n}$ shall be below a tolerance or a maximum deviation δ between true and observed logarithmic mean μ. Since μ=ln(GM), GM the geometrical mean, δ=ln[GM(obs)]–ln[GM(true)] = ln[GM(obs)/GM(true)]=ln[(GM(obs)-GM(true))/GM(true) + 1]. [(GM(obs)-GM(true)]/GM(true) is the maximum tolerated relative deviation of the GM; call this the required precision PREC. Inverting $x_\alpha\sigma/\sqrt{n}=\delta$, σ=ln(GSD), GSD – the geometrical standard deviation, and inserting results in

$$n_{\min} = \left(\frac{x_\alpha \ln(GSD)}{\ln(PREC+1)} \right)^2 \qquad (1)$$

This is a variant of a formula shown by Hale (1972), often used because of its simplicity. (In the original, it is given for finite population; in the limit population $\to\infty$ one finds formula (1)).

However, the precision constraint PREC is defined for the geometrical, but not for the arithmetic mean AM, which is the more important quantity as unbiased estimate of the mean.

A further approximation which enters via the CLT in the above statement about the SE consists in assuming statistical *independence* of the sample elements (i.e. the physical samples) z_i or y_i. In general, this is not true, as the sample is drawn from an autocorrelated field Z. However, the resulting bias is generally not known in real-world experiments, because it depends on the locations of the sample elements in the sampled area and the degree of autocorrelation of Z (its true variogram) is normally not known beforehand. For an approximate estimate of the minimal sample size n_{min} one will therefore live with the simplification.

If one prefers to derive n_{min} from the precision of the AM, one may proceed as follows. Confidence intervals of the AM under LN are complicated, but approximations are available. An easily tractable one is the modified Cox approximation (Olsson 2005),

$$\left(M_{upper}, M_{lower}\right) \approx \exp\left[m + \frac{s^2}{2} \pm t_{p;n-1}\, s \sqrt{\frac{1}{n} + \frac{s^2}{2(n-1)}} \right] \qquad (2a)$$

m and s – the sample estimates of μ and σ, M=exp(m+s²/2) the AM under LN hypothesis und p the significance. (See also Parkin et al. 1990, who give (n+1) in the denominator instead of (n-1).) The targeted precision, in analogy to the one defined above for the Hale formula, equals

$$\mathrm{PREC} = (M_{upper} - M_{lower})/(2M) \qquad (2b)$$

PREC does not contain m anymore. The minimal sample size is determined such that PREC remains below a set threshold. For known GSD (and s) and given p, (2a,b) can be solved for n by iteration, e.g. with Excel solver. For PREC=0.2, p=0.95 and GSD=2 (about realistic for Rn fields within 100 km² areas), one finds n_{min}=39 and 43 from the Hale formula (1) and from (2), respectively.

2.2. Tobler's law, proportional effect

The problem, which is the focus of this paper, is estimation of GSD as an input into (1) or (2). Tobler's *First Law of Geography* states that *"everything is related to everything else, but near things are more related than distant things"* (Tobler 1970). This law is the very condition of any spatial analysis and interpolation. It follows that variability or dispersion of a field increase, in average, with size of a spatial unit, or more precisely, with mean separation between random locations within the unit; more formally, in average disp(B_1)>disp(B_2) if $|B_1|>|B_1|$, in particular, if $B_1 \supset B_2$. B are spatial units, $|B|$ their areas, disp(B), measures of dispersion of the field within B, such as the coefficient of variation (CV), the geometrical standard deviation (GSD), the relative median absolute deviation (MAD/MED), p-quantile deviation $(Q_{1-p}-Q_p)/(Q_{1-p}+Q_p)$, and others. This fact is reflected by the shape of variograms.

Additionally, fields of positive definite physical quantities seem to have the general property that local (within a neighborhood) dispersion increases with their local level. This is called proportional effect (e.g., Manchuk et al. 2006, 2009) and can cause troubles in geostatistics. While for variables ~N(μ,σ), SD=σ and AM=μ are independent, for LN(μ,σ), SD= AM$\sqrt{}$(exp(σ²)-1), i.e. they are proportional. Assuming local log-normality (or "permanence", e.g. Agterberg, 1984; which ideally holds for LN multifractals, section 2.4), the proportional effect follows. Environmental radon can often be well described as LN (among many other, Bossew, 2010) and higher variability in high-Rn areas is indeed empirically observed (e.g., Bossew et al., 2008; Dubois et al., 2010). Reversely, assume a power-type relationship between the local SD and the local AM (which seems to be realistic), SD=a AM^b. Then CV = SD/AM ~ AM^{b-1}, i.e. the CV is slightly spatially variable. In the LN case and with some algebra, one finds,

$$GSD = \exp\sqrt{\ln\left(a^2\ AM^{\,2(b-1)} + 1\right)}\,.$$

A functional dependency between GSD or σ and GM and μ exists too, as a consequence, but it cannot be written analytically. For b=1, this becomes the "pure" proportional effect and CV (=a, in this case), GSD are spatially constant. Further studying the realistic case b>1, or "superproportional effect" would be worth the effort. - However, this paper will focus on the "Tobler effect", but not treat the proportional effect further.

2.3. Variogram and variance within an area

An increasing variogram is a consequence of the „Tobler effect" (or vice versa). The variance of process Z in area B equals $\mathrm{Var}_B(Z)=\mathrm{mean}_{h\in B}\ \gamma(h) = \int_{h\in B} \gamma(h)\, f_B(h)\, dh$, with γ the semivariance or variogram, $f_B(h)$ the distribution density of random segments of length h in B (figure 1). In general, this can only be computed numerically.

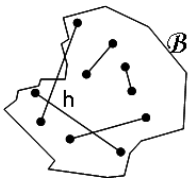


Figure 1 - Random segments of length h within area B.

As a rough approximation, set $Var_B(Z) \approx \gamma(\langle h \rangle)$ and if γ is the semivariance of the ln-transformed data,

$$GSD \approx \exp[\sqrt{\gamma(\langle h \rangle)}] \qquad (3)$$

where ⟨h⟩ denotes the mean length of a random segment within *B*. Due to the non-linear functions involved, (3) is clearly a biased estimate of $\exp(\sqrt{Var_B})$ (whose sampling distribution has not yet been studied to my knowledge), apart from the approximation. For *B* a square with unit side length, one finds ⟨h⟩≈0.521. For *B* a circle with radius 1, ⟨h⟩≈0.905. See the annex for details about random segments. Thus, for a square- and circle-shaped area, ⟨h⟩=0.521√area and 0.511√area, respectively. The approximation, though first-order in f_B only, performs relatively well, as will be shown in section 3. However, one may develop better approximations by including higher moments of h within *B*.

2.4 Power model and multifractality

Environmental fields can often be described as multifractals. In particular, this can be empirically shown for geogenic radon (as most important predictor of indoor radon, in most cases; for the Austrian Friedmann-type radon potential, see section 3, Bossew et al. 2008 and Bossew 2020). Its physical cause is geochemical distribution whose spatial distribution (represented by geology) can be characterized as outcome of a multiplicative cascade (e.g., Turcotte 1997). These lead to asymptotically (with cascade generation) LN fields and power variograms for low lags, $\gamma(h) \sim h^{\beta}$, with β=2H, H – the Hurst exponent, H=3-D, D – the fractal dimension of the surface (e.g., Bölviken et al., 1992; Herzfeld and Overbeck, 1999). (On the other hand, Cheng and Agterberg (1996), Cheng (1999) derived approximately log-shaped variograms for 1-dim de Wijs cascades.) In consequence, dispersion within an area can be expected to follow an approximate power law with its size. Again, this does not honor the proportional effect.

For a multiplicative de Wijs cascade (De Wijs, 1951; Agterberg, 2007a,b, 2015; and other), one finds

$\sigma^2 = \frac{k}{4}\left(\ln\left(\frac{1+\upsilon}{1-\upsilon}\right)\right)^2$, k – the cascade generation and υ the splitting factor. See also Bossew (2020).

3. Data

As examples we use data from radon (Rn) surveys: the first Austrian indoor Rn survey; the German indoor Rn database and the German database of the geogenic Rn potential (GRP); and the database of the European indoor Rn map.

The *Austrian indoor survey* is described in Friedmann (2005). We use a sample of size 25,160, about demographically representative. The measurements have been transformed to a standard situation called Friedmann radon potential, representing hypothetical Rn concentration in a standard room (ground floor, etc.).

The *German indoor Rn dataset* comprises 39,810 measurements, similarly transformed to a standard situation. Methodology is shortly described in Petermann and Bossew (2021) (Annex A). The sample is neither demographically nor geographically representative.

The *European dataset* consists of about 1 million measurements of indoor Rn concentration in ground floor living rooms across Europe, aggregated into 20,328 10 km × 10 km cells, of which AM, SD, AM and SD of ln-data, minimum, maximum, median and number of data are available (Cinelli et al., 2019; European Commission, 2019). Original data are not available due to data protection. However, these statistics allow numerical re-aggregation into differently sized units for the purpose of this study.

4. Results

GSD over 100 km² areas estimated by different methods for the Austrian, German and European datasets are summarized in table 1. For Austria and Germany, the area B is a circle of 100 km² area or 5.642 km radius. This corresponds to a theoretical mean distance of data points, ⟨h⟩=5.117 km. The empirical estimate has been found by calculating the GSD in 5,000 random circles of same area and averaging.

For the European dataset, the area is the original 10 km × 10 km cell. The variograms are shown in figure 2. The one of the European set is not directly comparable with the former ones, because its data are the means over cells; original data are not available. Therefore, the variogram-based method to estimate the GSD is not applicable here.

The physical reason for the difference between the estimates for Austria and Germany is probably their different definition.

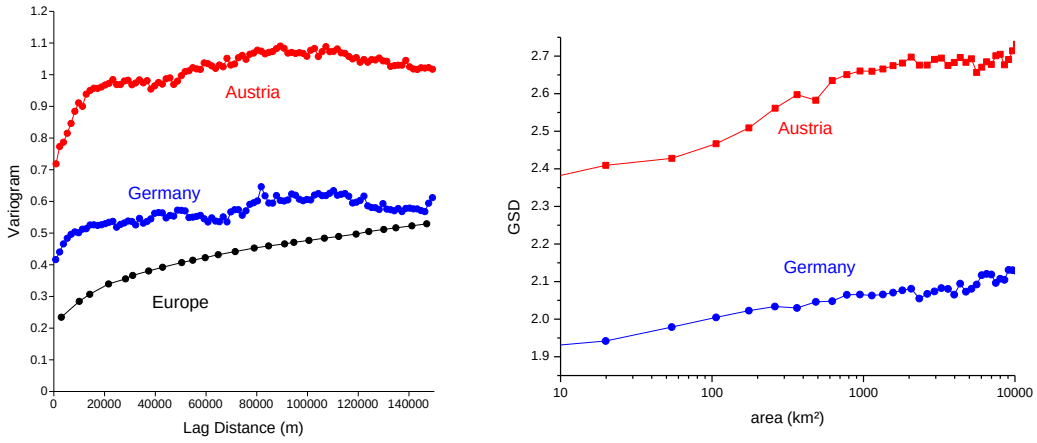


Figure 2 - Left: Unidirectional variograms of the Austrian, German and European ln-transformed indoor Rn data; right: GSD estimated from the variograms.

Empirical evaluation of the power model $GSD(h) \sim \alpha h^{\beta}$, (GSD in quadratic random windows, h – empirical mean separation of data points in the window, km) by regression leads to:

Germany: $\alpha=1.880\pm0.056$, $\beta=0.0758\pm0.0127$; Austria: $\alpha=2.110\pm0.039$, $\beta=0.0894\pm0.0086$ (see figure 3).

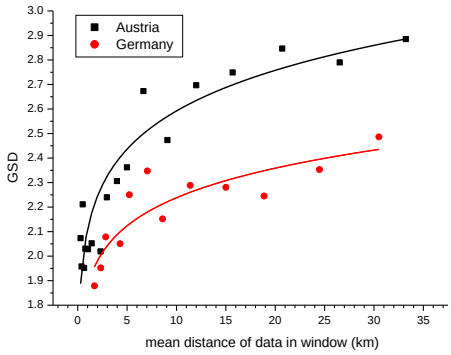


Figure 3 - Power model dependence of the GSD on data separation distance for two datasets (see text).

(Computational details: create circular random window of radius r; if it contains at least 8 data, compute mean data distance and GSD; aggregate into distance classes by computing median of the mean distances per window and median GSD; perform non-linear regression as shown in the figure.)

The power model overestimates GSD for larger windows (above ca. 200 km²). The reason is that the power function is unbounded, which does not reflect the stationarity property of the fields. The total GSD of German indoor Rn data, referring to an area of 358,000 km², equals 2.18 only. For Austria, the power model appears more realistic; GSD=3.0 for 84,000 km². It may be that the data points in figure 3 are affected by strong data clustering, which renders window statistics problematic. For Austria, a power-law dependence has been found for the CV for window radiuses below about 60 km, CV=0.46 $r^{0.343}$, r in km, Bossew et al. (2008).

For the European dataset, the empirical dependence of the GSD of indoor Rn concentration within cells on their areas is shown in figure 4.

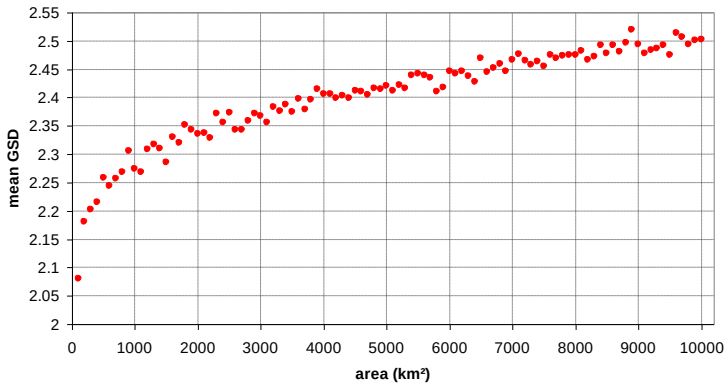


Figure 4 - Dependence of GSD on area, European indoor Rn data.

(Computational details: 1000 rectangular, where possible square shaped arrays of given size of cells were generated at random locations and the GSD recalculated from the individual ones in the 10 km × 10 km source cells; the data points in the graph represent the AMs per array size. SDs, not shown in the graph for better readability, are between 1.3 and 1.7.)

For the power model, one finds, if 'area' is taken as independent variable:

α=1.758±0.010, β=0.0378±0.0007; choosing a linear dimension d as independent, since d~√area, β=0.0756.

For d=⟨h⟩ of a theoretical square of given area, one finds α=1.842±0.009. This result is very similar to the German one, reported above.

Table 2 – Estimated GSD within 100 km².

Dataset	area	empirical	eq. (3)	power
Austria	100 km² circle	2.50	2.47	2.44
Germany	100 km² circle	2.02	2.01	2.13
Europe	100 km² square	2.08	n.a.	2.09

5. Conclusions

As it could be expected, data dispersion within areas increases with its size; here it has been shown for three datasets of indoor radon concentration. For small areas, i.e. typically the size of small municipalities, models and empirical findings agree well.

At current state of knowledge, for planning radon surveys, I recommend using the GSD(area) dependence estimated from the European indoor radon map (figure 4). For small areas, the power model, as derived from German data, may be a good option, otherwise it overestimates the GSD. If the variogram is known, estimating from it according to section 2.3 (formula (3)) may be a good alternative, as long as the shape of the area, for which the sample size shall be determined, does not deviate too strongly from elliptic or rectangular, and is not topologically awkward (holes inside or disconnected areas). The minimum sample size is most easily calculated from the Hale formula (1), or somewhat more complicated, by iterative solution of formula (2).

Tobler's First Law is very profound in spite of its seeming intuitive simplicity. Its consequences are far reaching and further theoretical discussion appears worthwhile, as does studying them on real-world examples of environmental fields.

References

Agterberg, F.P. (1984): Use of spatial analysis in mineral resource evaluation. Journal of the International Association for Mathematical Geology, 16(6), 565–589. doi:10.1007/bf01029317

Agterberg F.P. 2007a: Mixtures of multiplicative cascade model in geochemistry. Nonlin. Processes Geophys. 14, 201–209; DOI:10.5194/npg-14-201-2007

Agterberg F.P. 2007b: New applications of the model of de Wijs in regional geochemistry. Math. Geol. 39 (1), 1–25; https://doi.org/10.1007/s11004-006-9063-7

Agterberg F. (2015): Self-similarity and multiplicative cascade models. The Journal of The Southern African Institute of Mining and Metallurgy 115, 1–11; https://www.saimm.co.za/Journal/v115n01p001.pdf (accessed 9 April 2021)

Bölviken, B., Stokke, P.R., Feder, J., Jössang, T. (1992): The fractal nature of geochemical landscapes. Journal of Geochemical Exploration, 43(2), 91–109. doi:10.1016/0375-6742(92)90001-o

Bossew P., Dubois G., Tollefsen T., De Cort M. (2008): Spatial analysis of radon concentration at very short scales I. 33th IGC, Oslo, 12-14 Aug 2018.

Bossew P. (2010): Radon: Exploring the Log-normal Mystery. J. Environm. Radioactivity 101 (10), 826-834,. http://dx.doi.org/10.1016/j.jenvrad.2010.05.005

Bossew P. (2020): Log-log linearity of the asymptotic distribution – a valid indicator of multi-fractality? EGU General Assembly 2020, Online, 4–8 May 2020, Vienna https://meetingorganizer.copernicus.org/EGU2020/EGU2020-6447.html

Cheng, Q. and Agterberg, F.P. (1996): Multifractal modeling and spatial statistics. Mathematical Geology, 28(1), 1–16; doi:10.1007/bf02273520

Cheng, Q. (1999): Multifractality and spatial statistics. Computers & Geosciences, 25(9), 949–961. doi:10.1016/s0098-3004(99)00060-6

Cinelli G, Tollefsen T, Bossew P, Gruber V, Bogucarskis K, De Felice L, De Cort M. (2019): Digital version of the European Atlas of natural radiation, Journal of Environmental Radioactivity 196: 240–252. https://doi.org/10.1016/j.jenvrad.2018.02.008

De Wijs H.J. (1951): Statistics for ore distributions part 1. Geologie en Mijnbouw 13 (11), 365–375

Dubois G., Bossew P., Tollefsen T., De Cort M. (2010): First steps towards a European Atlas of Natural Radiation: Status of the European indoor radon map. JER 101 (10), 786–798. http://dx.doi.org/10.1016/j.jenvrad.2010.03.007

European Commission, Joint Research Centre – Cinelli G, De Cort M & Tollefsen T (Eds.), European Atlas of Natural Radiation, Publication Office of the European Union, Luxembourg, 2019; Printed version: ISBN 978-92-76-08259-0; doi:10.2760/520053 ; Catalogue number KJ-02-19-425-EN-C; Online version: ISBN 978-92-76-08258-3; doi:10.2760/46388 ; Catalogue number KJ-02-19-425-EN-N; https://remon.jrc.ec.europa.eu/About/Atlas-of-Natural-Radiation/Download-page

Friedmann H. (2005): Final results of the Austrian radon project. Health Physics 89 (4): 339–348. DOI: 10.1097/01.hp.0000167228.18113.27

Ghosh B. (1943): On the distribution of random distances in a rectangle. Science and Culture 8 (9), 388.

Hale W.E. (1972): Sample size determination for the log-normal distribution. Atmospheric environment 6, 419–422; doi:10.1016/0004-6981(72)90138-2

Hammersley J.M. (1950): The distribution of distances in a hypersphere. Ann. Math. Stat. 21 (3), 447–452; DOI: 10.1214/aoms/1177729805

Herzfeld, U.C. and Overbeck, C. (1999): Analysis and simulation of scale-dependent fractal surfaces with application to seafloor morphology. Computers & Geosciences, 25(9), 979–1007. doi:10.1016/s0098-3004(99)00062-x

Manchuk J. (2006): The Proportional Effect: What it is and how do we model it? http://www.ccgalberta.com/ccgresources/report08/2006-109-proportional_effect.pdf (accessed 1 April 2021)

Manchuk, J.G., Leuangthong, O. & Deutsch, C.V. (2009): The Proportional Effect. Math Geosci 41, 799–816. https://doi.org/10.1007/s11004-008-9195-z

Mathai, A. M., Moschopoulos, P., & Pederzoli, G. (1999). Random points associated with rectangles. Rendiconti Del Circolo Matematico Di Palermo, 48(1), 163–190. doi:10.1007/bf02844387

Müller W. G. (2007): Collecting spatial data - Optimum Design of Experiments for Random Fields. Springer Berlin-Heidelberg, DOI 10.1007/978-3-540-31175-1

Olsson J. (2005): Confidence Intervals for the Mean of a Log-Normal Distribution. Journal of Statistics Education Volume 13, Number 1 (2005), ww2.amstat.org/publications/jse/v13n1/olsson.html (accessed 1 April 2021)

Parkin T.B., Chester S.T., Robinson J.A. (1990): Calculating Confidence Intervals for the Mean of a Lognormally Distributed Variable. Soil. Sci. Am. J. 54, 321–326; https://pubag.nal.usda.gov/download/49166/PDF (accessed 11 April 2021)

Philip, J., (2007): The Probability Distribution of the Distance between Two Random Points in a Box. www.math.kth.se/wjohanph/habc.pdf (accessed 10 April 2021)

Petermann E., Bossew P. (2021): Mapping indoor radon hazard in Germany: The geogenic component. Science of the Total Environment 780, 146601; https://doi.org/10.1016/j.scitotenv.2021.146601

Rosenberg E. (2004): The expected length of a random line segment in a rectangle. Operations research letters 32, 99-102; doi:10.1016/S0167-6377(03)00072-5

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46 (Suppl. l), 234–240 (The famous quote is on the bottom of p.236, right column). www.jstor.org/stable/143141

Turcotte D.J. (1997): Fractals and Chaos in Geology and Geophysics (2nd ed.). Cambridge University Press. ISBN 0521-56733-5.

Annex: Random segments

Random segments are straight connections between two random points within an area. The distribution of their lengths can be calculated in cases of simple shaped areas, such as rectangles and circles, but is more easily found by simulation.

The distribution of segments within a sphere, where x:=(segment length)/(circle diameter), is $f(x)=\frac{16}{\pi}x\left(\arccos x - x\sqrt{1-x^2}\right)$ (e.g., Hammersley 1950) with expectation of the length within a circle of radius=1, $\langle d\rangle=128/(45\,\pi)\approx 0.9054$. The formula for the distribution within a rectangle is more complicated; e.g. shown in Ghosh (1943), Mathai et al. (1999) or Philip (2007). For the expectation, see the same or Rosenberg (2004). For the unit square, it is $\langle d\rangle=(1/15)(2+\sqrt{2}+5\ln(\sqrt{2}+1))\approx 0.5214$.

Mean lengths of random segments computed numerically for differently shaped areas, all of size 1, are shown in figure 5. The shape parameter is denoted eps:=b/a. For rectangles these are the side lengths, for ellipses the long and short half axes (numerical eccentricity=$\sqrt{(1-\text{eps}^2)}$), for rectangular triangles the catheti and for the orthogonal L-shapes the legs; for the latter, additional parameters α and β measure the "cut-out" part, whose side lengths are (1-α)a and (1-β)b. For these non-convex shapes, also segments running across the outside are counted. For shape L1, α=β=0.5 (a "bold" L), for shape L2, α=β=0.2 (a "thin" L). As an example of a topological more complex area, the "donut" has been chosen, i.e. a circular ring. In this case, the shape parameter eps denotes 1-inner/outer radius. The results are based on 10 million simulations each. Convergence is low, but calculation fast, few minutes at most for each graph with an average notebook. Accuracy has been checked with theoretical values for square and circle; error is about 0.1%. $\langle d\rangle$ scales with √area.

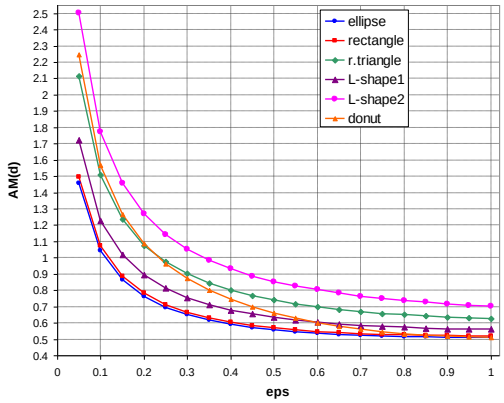


Figure 5 - Mean lengths of random segments in differently shaped areas of size 1.

One notices that the lengths ⟨d⟩ do not differ very much for high ε, i.e. as long as the shape is not too far from circular or quadratic. For ellipses and rectangles of same ε, ⟨d⟩ is very similar. However, deviation from these basic shapes (triangles and L-shapes) result in quite different mean segment lengths for same area. The donut example shows (as was to be expected) that topological more complicated areas lead to different ⟨d⟩.

R

S

T

V

W

X

Z

Proceedings of

geoENV2020

13TH INTERNATIONAL CONFERENCE ON GEOSTATISTICS FOR ENVIRONMENTAL APPLICATIONS

PARMA, ITALY JUNE 18, 2021

Edited by **ANDREA ZANINI & MARCO D'ORIA**

The 13th International Conference on Geostatistics for Environmental Applications (geoENV2020) was scheduled in Parma, Italy on July 2020. The international health crisis affected the conference, which was initially postponed to June 2021 and eventually replaced by a one-day virtual event on June 18, 2021 with the presentations of the keynote lecturers.

This book contains the abstracts and extended abstracts submitted to the conference and focusing on geostatistics applied to different fields such as: climate change, ecology, natural resources, forestry, agriculture, geostatistical theory and new methodologies, health, epidemiology, ecotoxicology, inverse modeling, multiple point geostatistics, remote sensing, soil applications, spatio-temporal processes and surface and subsurface hydrology. The Scientific Committee initially selected about 100 abstracts and 68 contributions were confirmed to be published in these proceedings.

The next geoENV conference (geoENV2022) will be held in Parma, Italy on June 2022. We expect more colleagues from all over the world to join this international event next year.

UNIVERSITÀ DI PARMA