14TH INTERNATIONAL CONFERENCE ON GEOSTATISTICS FOR ENVIRONMENTAL APPLICATIONS

geoENV

JUNE 22-24
PARMA
2022

PROCEEDINGS OF geoENV2022

Andrea Zanini & Marco D'Oria, Editors

UNIVERSITÀ
DI PARMA

Andrea Zanini & Marco D’Oria

Editors

14TH INTERNATIONAL CONFERENCE ON GEOSTATISTICS FOR ENVIRONMENTAL APPLICATIONS

UNIVERSITÀ DI PARMA

First Edition 2022

14th International Conference on Geostatistics for Environmental Applications: geoENV2022

Editors: Andrea Zanini & Marco D’Oria

handle: https://hdl.handle.net/1889/4778

Preface

The 14th International Conference on Geostatistics for Environmental Applications (geoENV2022) was held in Italy, at the Campus of the University of Parma. From June 22 to June 24, 2022, over 80 experts on geostatistics gathered to discuss about environmental applications of this discipline.

This book contains the abstracts and extended abstracts submitted to the conference and focusing on geostatistics applied to different fields such as: ecology, natural resources, environmental pollution and risk assessment, 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.

Organizing Committee

Andrea Zanini (Co-Chair)

Marco D'Oria (Co-Chair)

Maria Giovanna Tanda

Valeria Todaro

Jaime Gómez-Hernández (President)

Philippe Renard (Secretary)

Scientific Committee

Teresa Albuquerque – IPCB ICT CERNAS – Portugal

Denis Allard – BioSP, INRAE – France

Peter Atkinson – Lancaster University – United Kingdom

Leonardo Azevedo – CERENA, Instituto Superior Técnico, Universidade de Lisboa – Portugal

Patrick Bogaert – Université catholique de Louvain – Belgium

Peter Bossew – German Federal Office for Radiation Protection (BfS) – Germany

Ilaria Butera – Politecnico di Torino – Italy

Guofeng Cao – University of Colorado Boulder – United States

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

Aldo Fiori – Roma Tre University – Italy

Chantal de Fouquet – MINES ParisTech, Universitè PSL – France

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

Harrie-Jan Hendricks-Franssen – Forschungszentrum Julich – Germany

George P. Karatzas – Technical University of Crete – Greece

Sara Kasmaeeyazdi – University of Bologna – Italy

Liangping Li – South Dakota School of Mines – United States

Gregoire Mariethoz – University of Lausanne – Switzerland

Jennifer McKinley – Queen's University Belfast – United Kingdom

Alessandra Menafoglio – Politecnico di Milano – Italy

Paula Moraga – King Abdullah University of Science and Technology – Saudi Arabia

Julian Ortiz – Queen's University – Canada

Monica Palma – University of Salento – Italy

Edzer Pebesma – University of Muenster – Germany

Maria João Pereira – CERENA, Instituto Superior Técnico, Universidade de Lisboa – Portugal

Pierre Petitgas – IFREMER – France

Javier Rodrigo Ilarri – Universitat Politècnica València – Spain

Thomas Romary – MINES ParisTech, Universitè PSL – France

Xavier Sanchez-Vila – Universitat Politecnica de Catalunya – Spain

Amilcar Soares – Instituto SuperiorTécnico, Universidade de Lisboa – Portugal

Francesco Tinti – University of Bologna – Italy

Emmanouil Varouchakis – Technical University of Crete – Greece

Hans Wackernagel – Mines Paris Tech, Universitè PSL – France

Keynote lectures

Wednesday, June 22, 2022
Carolina Guardiola Albert, Geological Survey of Spain (IGME-CSIC)
Exploitation of InSAR ground movement measurements through geostatistics

Thursday, June 23, 2022
Leonardo Azevedo, CERENA, Instituto Superior Técnico, Universidade de Lisboa, Portugal
Modelling the ocean with acoustic waves: a geostatistical inversion approach

Friday, June 24, 2022
Laura Poggio, ISRIC – World Soil Information, the Netherlands
Geostatistics, machine learning and spatial patterns

BAYESIAN U-NET FOR ORE TYPE UNCERTAINTY MODELING IN COMPLEX GEOLOGICAL ENVIRONMENTS....1

A GEOSTATISTICAL APPROACH FOR MERCURY SPATIAL PATTERNS ASSESSMENT IN SEDIMENTS IN AN OLD MINING REGION -THE CAVEIRA MINE CASE STUDY, PORTUGAL....2

STREAM SEDIMENTS POLLUTION: A COMPOSITIONAL BASELINE ASSESSMENT AT THE CAVEIRA MINE, PORTUGAL....4

THE GEOGRAPHICAL PATTERN OF LOCAL STATISTICAL DISPERSION OF ENVIRONMENTAL RADON IN EUROPE* 6

SPATIOTEMPORAL PREDICTIONS OF MULTIPLE AIR POLLUTION DATA: A CASE STUDY 15

GEOSPATIAL MODEL OF COMPOSITION OF WATER SERVICE LINES IN FLINT, MI: VALIDATION USING EXCAVATION DATA AND A NEW COMPOSITIONAL GEOSTATISTICAL APPROACH....17

GEOSTATISTICAL MAPPINGS OF INDOOR RADON CONCENTRATIONS DATA IN FRANCE 18

INTERPOLATED SURFACES AS MAPPING UNITS FOR BINARY CLASSIFICATION OF RADON PRONE AREAS: A CASE-STUDY FROM CENTRAL PORTUGAL....19

A SPATIO-TEMPORAL MULTILEVEL MULTIVARIATE MODEL TO EVALUATE THE DETERMINANTS OF AIR POLLUTION IN APULIA REGION 20

OPERATIONAL SPATIALISATION OF HOURLY AND DAILY WEATHER DATA (AIR TEMPERATURE AND RELATIVE HUMIDITY) FOR AGRICULTURAL DECISION SUPPORT SYSTEMS 21

USING OF SEQUENTIAL INDICATOR SIMULATION TO MODEL NON-STATIONARY GEOLOGICAL DOMAINS COMBINING WITH A MACHINE LEARNING ALGORITHM* 23

AN OBJECT-ORIENTED APPROACH TO THE ANALYSIS OF SPATIAL FUNCTIONAL DATA OVER STREAM-NETWORK DOMAINS 31

IMPLICIT GEOLOGICAL DOMAIN MODELING BY USING GAUSSIAN PROCESSES ALGORITHM* 32

TWO APPROACHES TO SCENARIO MODELLING USING THE EMBEDDED ESTIMATOR METHOD 40

ANALYSIS OF TWO PRECIPITATION GAP-FILLING METHODS IN A STUDY AREA OF NORTHERN ITALY....41

A NEW APPROACH TO SIMULATE CONDITIONAL RANDOM FIELDS WITH CORRELATION TO AN EXTERNAL VARIABLE USING FFTMA AND P-FIELD SIMULATION 42

COMBINED ANALYSIS OF SPATIALLY MISALIGNED DATA USING GAUSSIAN FIELDS AND THE STOCHASTIC PARTIAL DIFFERENTIAL EQUATION APPROACH....43

BIVARIATE DEEPKRIGING FOR LARGE-SCALE SPATIAL INTERPOLATION OF WIND FIELD....44

GEOSTATISTICS FOR COMPOSITIONAL DATA: COMPARISON OF LOGARITHMIC TRANSFORMATION METHODS IN AN IRON DEPOSIT*....45

A RIEMANNIAN TOOL FOR CLUSTERING OF GEO-SPATIAL MULTIVARIATE DATA....52

USING GEOSTATISTICAL APPROACHES TO ESTIMATE THE RGB VALUES UNDER A CLOUD COVERED AREA WITHIN THE SENTINEL-2 IMAGES 53

* Extended abstract

SURROGATE MODELS AS MANAGEMENT TOOLS FOR THE REQUENA-UTIEL AND CABRILLAS-MALACARA AQUIFERS........54

APPLICATION OF GEOSTATISTICS AND SELF-ORGANIZING MAPS FOR ESTIMATION OF GROUNDWATER LEVEL SPATIAL DISTRIBUTION IN COMPLEX HYDROGEOLOGICAL SYSTEMS 55

A COMPARISON BETWEEN BAYESIAN AND ORDINARY KRIGING BASED ON VALIDATION CRITERIA: APPLICATION TO RADIOLOGICAL CHARACTERISATION*........56

DOES MORE INFORMATION INCLUDED IN SPATIALLY DISTRIBUTED FIELDS LEAD TO AN IMPROVED MATCH TO OBSERVED DEPENDENT VARIABLES?........63

A STOCHASTIC MODEL OF AN ALERT SYSTEM FOR DETECTING LOCAL ANOMALOUS INCIDENCE VALUES OF COVID-19........64

SPATIALEPISIM: AN R SHINY APP FOR TRACKING INFECTIOUS DISEASES IN LOW- AND MIDDLE-INCOME COUNTRIES (LMIC)........65

UNPACKING OCCUPATIONAL HEALTH DATA IN THE TERTIARY SECTOR. FROM SPATIAL CLUSTERING TO BAYESIAN DECISION MAKING*........66

FUNCTIONAL DATA ANALYSES TO MODEL COVID-19 WAVES........74

SPATIO-TEMPORAL POINT PROCESS COMPARTMENT MODELING FOR INFECTIOUS DISEASES......75

GEOSPATIAL MODELING OF NATIONAL HEALTH SERVICE DELIVERY SURVEY DATA........76

GEOSTATISTICAL COVID-19 RISK AND UNCERTAINTY IN A SINGLE MAP WITH R OPEN-SOURCE CODE........78

MACHINE LEARNING-BASED INVERSE MODELING FOR THE IDENTIFICATION OF HYDRAULIC CONDUCTIVITY........79

SEISMIC OCEANOGRAPHY IMAGING AND INVERSION OF THE MADEIRA ABYSSAL PLAIN.............80

SOLVING GEOPHYSICAL RANDOM EFFECT MODELS WITH INTRACTABLE LIKELIHOODS: LINEARIZED GAUSSIAN APPROXIMATIONS VERSUS THE CORRELATED PSEUDO-MARGINAL METHOD........81

GEOSTATISTICAL INVERSION FOR ROCK PHYSICS PROPERTIES IN THE CRITICAL ZONE..................82

ON A CONSTRUCTIVE SPECTRAL METHOD FOR CONDITIONING PLURIGAUSIAN SIMULATIONS TO BOREHOLES OBSERVATIONS AND INDIRECT DATA. APPLICATION TO AQUIFER MODELS* 83

EFFICIENT INVERSION WITH COMPLEX GEOSTATISTICAL PRIORS USING NORMALIZING FLOWS AND VARIATIONAL INFERENCE........92

GEOSTATISTICAL ELECTRICAL RESISTIVITY TOMOGRAPHY INVERSION FOR GROUNDWATER CHARACTERIZATION........93

PERMEABILITY PREDICTION WITH GEOSTATISTICAL SEISMIC INVERSION CONSTRAINED BY ROCK PHYSICS........94

GROUNDWATER CONTAMINANT SOURCE CHARACTERIZATION THROUGH ARTIFICIAL NEURAL NETWORKS........95

MODELLING THE COMPLEXITY BENEATH OUR FEET: A JOINT INVERSION FDEM AND ERT TECHNIQUE........96

* Extended abstract

IDENTIFICATION OF CONTAMINANT SOURCES AND PLUMES AFFECTED BY BIODEGRADATION AND SORPTION PROCESSES BY ENSEMBLE KALMAN FILTERS....97

ASSESSMENT OF HYDRAULIC CONDUCTIVITY FIELD USING LABORATORY SANDBOX TRACER TEST DATA AND AN ENSEMBLE KALMAN FILTER METHOD....98

HANDLING NON-STATIONARITY IN MULTIPLE-POINT STATISTIC SIMULATION WITH A HIERARCHICAL APPROACH....99

PIXEL-BASED MULTIPLE-POINT-STATISTICS PARAMETRIZATION BASED ONLY ON TRAINING IMAGES....100

GEOSTATISTICAL SIMULATION FOR OFFSHORE WIND SPEED SPATIO-TEMPORAL GAP-FILLING 101

APPLYING MPS TO POINT DATA MERGING USING PATTERN-TO-POINT (P2P) CATALOGS....102

A GEOSTATISTICAL POINT OF VIEW ON HETEROSUPPORT AND HETEROTOPIC CO-REGIONALIZATION OF REMOTE SENSED INFORMATION....103

REMOTE SENSING CHANGE DETECTION FOR THE APPEARANCE AND DISAPPEARANCE OF OASES FARMLAND IN CENTER ASIA....104

MAPPING OF CRITICAL RAW MATERIALS IN BAUXITE MINING RESIDUES USING GEOSTATISTICS AND REMOTE SENSING....105

DETECTION OF GROUND DEFORMATION EVENTS: A METHODOLOGY FOR THE STATISTICAL ANALYSIS OF INSAR TIME SERIES....106

DERIVING HIGH SPATIAL RESOLUTION DAILY VEGETATION INDEX IMAGES FROM SENTINEL AND MODIS DATA: A GEOSTATISTICAL APPROACH....107

A NOVEL METHODOLOGY FOR THE IDENTIFICATION OF ERRORS AND SIGNIFICANT ACCELERATION EVENTS FROM AUTOMATIC INCLINOMETERS TIME SERIES RELATED TO SLOW AND VERY SLOW-MOVING LANDSLIDES....108

A NONPARAMETRIC SPATIO-TEMPORAL APPROACH TO EVALUATE UNCERTAINTY IN GEOSTATISTICS....110

COVARIANCE MODELING FOR SPATIO-TEMPORAL COMPLEX-VALUED RANDOM FIELDS....111

VARIABLE KERNEL ESTIMATES OF FIRST-ORDER SUMMARY DESCRIPTOR OF SPATIO-TEMPORAL POINT PROCESSES....112

GEOSTATISTICAL ANALYSIS OF EXTREME PRECIPITATION RECORDS OVER NORTH-WEST ITALY 113

A MARKED POINT-PROCESS SPATIO TEMPORAL MODEL TO UNDERSTAND FOREST FIRES IN THE MEDITERRANEAN BASIN....114

AN ENSEMBLE-BASED APPROACH FOR THE ANALYSIS OF SPATIALLY MISALIGNED DATA....116

ROBUST STATISTICAL PROCESSING OF LONG-TIME DATA SERIES ON SOIL-ATMOSPHERE INTERACTION: PRELIMINARY RESULTS....117

USING GEOSTATISTICAL METHODS TO HELP OPTIMIZING AN EXISTING GROUNDWATER MONITORING NETWORK....118

* Extended abstract

CRITICAL COMPARISON OF THREE COMPOSITIONAL INDICES TO TRACE GEOCHEMICAL CHANGES DOWNRIVER 119

A SINGLE SPATIOTEMPORAL FRAMEWORK FOR MONTHLY LOW FLOW ESTIMATION IN AUSTRIA – INCLUDING EMPIRICAL ORTHOGONAL FUNCTIONS, VARIABLE SELECTION, AND FUNCTIONAL CLUSTERING 120

COUPLED SURFACE-SUBSURFACE HYDROLOGICAL MODEL FOR THE ESTIMATION OF NET RECHARGE OF THE KONYA CLOSED BASIN, TURKEY 121

IMPORTANCE OF MULTI-PARAMETER APPROACHES IN THE DEVELOPMENT OF MACHINE LEARNING ALGORITHMS FOR LANDSLIDE DISPLACEMENT FORECASTING 122

MODELING DENSITY DATA ACROSS THE IBERIAN MASSIF (PORTUGAL): RELIABILITY ASSESSMENT FOR ENVIRONMENTAL APPLICATIONS 123

ENVIRONMENTAL GEOSTATISTICS OF HEAVY METALS IN FINE ACTIVE SEDIMENTS OF OESTE ANTIOQUEÑO SUBREGION ANTIOQUIA - COLOMBIA. 124

CONDITIONED SIMULATION FOR GEOSTATISTICALTREATMENT OF SEISMIC THREAT IN ANTIOQUIIA STATE, NORTHWESTERN COLOMBIA. 125

ANALYSIS OF SURFACE TEXTURE USING SPATIAL CONTINUITY INDICES: INTERPRETABILITY, FLEXIBILITY, AND ROBUSTNESS 126

* Extended abstract

BAYESIAN U-NET FOR ORE TYPE UNCERTAINTY MODELING IN COMPLEX GEOLOGICAL ENVIRONMENTS

Helga Jordão (1)* - Amilcar Soares (1)

CERENA - Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal (1)
** Corresponding author: helgajordao@tecnico.ulisboa.pt*

Abstract

Deep learning had a substantial growth in the geosciences research community. There are several successfully applications examples such as in remote sensing, seismic interpretation, geomorphology and geomodelling. But still there is a slow uptake of these approaches in real applications and one reason for that is the lack of interpretability of deep learning approaches. Deep learning models are deterministic, producing a single estimate, or prediction without providing a quantitative assessment of uncertainty of the prediction made. Thus, when one use deep learning for a model prediction, it is important to quantify the attached uncertainty (a measure of the confidence of the predictions), since all models are subject to noise and model inference errors. This is critical in decision making and risk assessment applications.

Several techniques for quantifying model uncertainty in deep learning have recently been proposed. Inside Bayesian Deep Learning methods, one of the most used is Monte Carlo dropout. Dropout is a regularization method used to reduce overfitting and improve generalization error in deep neural networks but it can also be used for providing uncertainty of model prediction. Dropout randomly drops some units of the neural network providing some randomness to the system. If used at inference, leads to multiple different parameter settings, and creates a probabilistic Bayesian Neural Network. By passing several times the same input we generate multiple realizations which can provide an estimate of the model uncertainty.

In this paper, we introduce a Bayesian Deep Learning approach (Monte Carlo dropout) on a real case problem of the mining industry, ore type morphology modeling in complex geological environments, where uncertainty quantification of the boundaries of different ore types, plays a pivotal role in resources evaluation and uncertainty assessment. One of the main sources of risk in mining resources and reserves evaluation is the heterogeneity of the orebody. Therefore, is crucial to quantify the uncertainty of orebody boundaries, since the confidence in a feasibility study and investment or operational decisions must be based on relevant and reliable resources predictions. We implemented a probabilistic Bayesian U-Net for automatic delimiting the geological domains of an orebody, conditioned on drill-hole data, but also producing spatial uncertainty maps of those domains.

A GEOSTATISTICAL APPROACH FOR MERCURY SPATIAL PATTERNS ASSESSMENT IN SEDIMENTS IN AN OLD MINING REGION -THE CAVEIRA MINE CASE STUDY, PORTUGAL

Natália Mota (1) - Rita Fonseca (1) - Joana Araújo (1) - Margarida Antunes (1) - Teresa Valente (2) - Ana Barroso (2) - Alexandre Araújo (1) - Teresa Albuquerque (3)*

ICT, University of Évora, Geosciences, Évora, Portugal (1) – ICT, University of Minho, Geosciences, Braga, Portugal (2) - Polytechnique Institute of Castelo Branco (CERNAS) and ICT, University of Évora, Civil Engineering, Castelo Branco, Portugal (3)
** Corresponding author: teresal@ipcb.pt*

Abstract

Mercury pollution is significant in many former mining communities worldwide, including in developing countries. Anthropic contributions to environmental Hg pollution are mostly connected to fuel fossil emissions, industrial and mining activities. Among mining operations, gold exploration contributes to the highest Hg contamination rates, given the processes, widely used in the past, of mixing Hg with the gold-containing ore, to separate this metal from the bulk impurities.

This study, as part of the GeoMaTre project, an ongoing collaborative network (2021-2024) between the Polytechnic Institute of Castelo Branco and the University of Évora, Portugal, aimed to evaluate the potential risk of mercury pollution in stream sediments in the Caveira area, an abandoned Cu, Pb, Zn, Ag, and Au mine, included in the Iberian Pyrite Belt, at South Portugal. This mine corresponds to a Gossan developed on pyrite mineralization, with high gold and silver content at the official beginning of its exploitation, in 1863, having exhausted the reserves in these precious metals in the 1920s. Until the date of its abandonment (1966) the exploitation focused on the remaining metals (Cu, Pb, Zn) and S. Currently, the surrounding area of Caveira mine is essentially composed of areas of waste accumulation, from mining activity, with little or no vegetation.

Thirty-three sediment samples were collected from within 0 to 10 cm depth, in a grid of 1Km x 1Km. Hg was determined in samples preserved at about 4ºC at the time of collection, through a mercury analyzer (NIC MA-3000) based on thermal decomposition, gold amalgamation, and cold vapor atomic absorption spectroscopy detection.

A multivariate preliminary study was conducted to evaluate the spatial distribution of Hg at the mine area and to determine the spatial clusters of Hg concentration. Analysis showed very high values (50-130μgg^{-1}), in the sediments deposited in the mainstream crossing the mine heaps, with concentrations reaching 340 μgg^{-1} in the meeting with the major waterway of the region. In the latter, near the confluence zone, there is an attenuation of Hg levels, although still above the reference values for sediments, 0.3μgg^{-1}, according to the Netherlands Regulation (2009), followed by many European countries. Since this is a complex mining area with diffuse distribution of the water system, levels significantly higher than reference values were also found in other small streams in the vicinity of the mine heaps. According to the Hg limits established by this regulation, mitigation measures are required when Hg is greater than 36μgg^{-1}. Therefore, to identify spatial patterns of the Hg concentration distribution, geostatistical modeling was used throughout conventional variography followed by the Sequential Gaussian Simulation (SGS). The Mean Image of the one hundred performed simulations followed by local G clustering allowed the definition of the significant hotspots for contamination risk. The probability maps of exceeding, respectively, the 0.3$\mu gg^{-1}$and the 36μgg^{-1} thresholds were computed and acted as a measurement of the obtained clusters' robustness, thus providing a faster and more intuitive way to verify whether the previously detected problematic zones are true of concern and in need of mitigation.

Keywords: Caveira mine; Mercury; Sequential Gaussian Simulation; G clustering; Probability map.

STREAM SEDIMENTS POLLUTION: A COMPOSITIONAL BASELINE ASSESSMENT AT THE CAVEIRA MINE, PORTUGAL

Araújo Joana (1) - Rita Fonseca (1) - Natália Mota (1) - Alexandre Araújo (1) - Margarida Antunes (2) - Teresa Valente (2) - Ana Barroso (3) - Teresa Albuquerque (4)*

ICT, University of Évora, Geosciences, Évora, Portugal (1) – ICT, University of Minho, Geosciences, Braga, Portugal (2) - ICT, University of Minho, Geosciences, Castelo Branco, Portugal (3) - Polytechnique Institute of Castelo Branco (CERNAS) and ICT, University of Évora, Civil Engineering, Castelo Branco, Portugal (4)
** Corresponding author: teresal@ipcb.pt*

Abstract

A high concentration of Potentially Toxic Elements (PTE) can affect ecosystem health. It is therefore essential that spatial trends of pollutants are assessed and controlled. River sediment pollution is widespread in mining communities around the world, including in developing countries. This study, as part of the GeoMaTre project, restoration of water bodies impacted by mine drainage, an ongoing collaborative project (2021-2024) between the Polytechnic Institute of Castelo Branco and the University of Évora, Portugal, aimed to evaluate the potential risk of PTEs pollution in stream sediments under the direct influence of Caveira mine, a Cu-Pb-Zn-Ag-Au old mine included in the Iberian Pyrite Belt, South Portugal. Quantifying pollution implies first the understanding of pollution-free stream sediment. Often, this background, or pollution baseline, is undefined or only partially known. Given that the concentration of chemical elements is compositional, as the attributes vary together, a compositional approach was used aiming to find a compositional balance, based on Compositional Data (CoDa) principles. A dataset of 33 samples was collected from within 0 to 10 cm depth, in a grid of 1Km x 1Km and thirteen chemical elements, including PTEs of variable toxicity (As, Cd, Co, Cr, Cu, Hg, Mn, Ni, Pb, Zn, V) and major elements from lithogenic sources (Fe, Al), were analyzed in preserved samples at about 4°C. The most extractable forms of metals (except for Hg) were obtained by partial digestion with aqua regia (HCl and HNO_3) in a high-pressure microwave digestion unit, followed by ICP-OES analysis. Hg was analyzed determined by a mercury analyzer based on thermal decomposition, gold amalgamation, and cold vapor atomic absorption spectroscopy detection.

Very high levels of Pb, As and Hg were found in sediments in a stream closer to the mine tailings pile and in an accumulation area where it flows into a larger waterway. These concentrations, reaching 3.8% Pb, 750μgg^{-1} As and 340 μgg^{-1} Hg, are well beyond the intervention values imposed by the Netherlands legislation (2009), one of the only European legislation that includes reference values for freshwater sediments. In some of these locations, Zn contents above the legislated reference values (120 μgg^{-1}) corroborate the nature of the ore previously exploited, pyrite mineralization enriched in Cu, Pb, Zn, Au, and Ag.

The methodological approach implied the geological background selection in terms of a trimmed subsample that can be assumed as non-pollutant (Al and Fe) and the selection of a list of pollutants based on the based on expert knowledge and previous studies (As, Zn, Pb, and Hg); Identifying a compositional balance, including pollutant and non-pollutant elements, with sparsity and simplicity as properties, is crucial for the construction of the novel Compositional Pollution Indicator (CPI).

A sequential stochastic Gaussian Simulation was performed on the new CPI. The results of the 100 computed simulations are summarized through mean image maps and probability maps of exceeding a given statistical

threshold, thus, allowing the characterization of the spatial distribution and variability of the CPI. A better understanding of the trends of relative enrichment and PTEs fate is discussed.

Keywords: Caveira mine; Stream sediment; Compositional Indicator; Sequential Gaussian Simulation; Probability map.

THE GEOGRAPHICAL PATTERN OF LOCAL STATISTICAL DISPERSION OF ENVIRONMENTAL RADON IN EUROPE

Peter Bossew (1)*

Federal Office for Radiation Protection (BfS), Dpt. Ur-2, Berlin, Germany (1)
** Corresponding author: peter.bossew@reflex.at*

Abstract

Acknowledged as a significant health hazard, increasing attention has been given to indoor and geogenic radon for some 20 years. One part of the efforts is surveying in order to assess the geographical extent of the hazard. Results acquired in surveys serve to support decisions in radon abatement policy, aimed to reduce exposure and consisting in prevention, mitigation and remediation measures. One particular element is delineation of so-called radon priority areas, or areas in which the hazard is such that abatement measures should be implemented with priority. These areas are estimated from radon measurements supported by different modelling methods.

Methods which are based on estimating local probabilities that radon concentration exceeds a reference level often rely on measures of local variability, expressed e.g. by the geometric standard deviation or the coefficient of variation, because these describe the shape of the distribution whose tail areas are the sought-after probabilities. Evidently, delineation of radon priority areas thus depends, apart from mean concentration, on dispersion within the area whose priority status shall be assessed.

A second use of spatial variability measures of radon is survey planning, because the sample size necessary to estimate a mean with given precision depends on the dispersion of the quantity to be assessed. It is estimated through pilot surveys or derived from general knowledge.

The large radon databases accumulated for years allow more detailed insight into spatial properties of dispersion, some of which are discussed in this papers, in the first place the relation between local dispersion and mean and sampling density. Not least, they also grant insight – so far mostly speculative – about the process, understood as a stochastic process, which generates the spatial dynamic.

Keywords: Environmental radon; Spatial dispersion; Mapping

1. Introduction

Radon (Rn; here restricted to the isotope 222Rn) is an important hazard to human health, believed to cause 100,000s of lung cancer fatalities world-wide annually through its equally radioactive short-lived progeny (e.g., Zeeb and Shannoun, 2009; Gaskin et al. 2018). One element on Rn abatement strategies is mapping. In Europe, an indoor Rn map based on over one million measurements has been generated (Cinelli et al 2019), consisting of statistics within 10 km × 10 km cells. The statistics include mean and standard deviation of the values and their logarithms.

Spatial distribution of Rn within given units is distributed approximately log-normal (among many other, Bossew 2010). This is used to estimate exceedance probabilities (i.e. that a reference level is exceeded), which serve as decision base in Rn abatement policy. Another use is planning of surveys. In both cases, together with mean, dispersion within a spatial unit must be known, measured e.g. as geometric standard deviation - in the first case, for estimating the distribution tail, in the second, to estimate the sample size necessary to estimate the local mean with required precision (Bossew 2021).

The dependence of dispersion on the size of areal unit has been investigated previously (Bossew 2021). In addition, dependence of dispersion per areal unit on the mean has been identified (Dubois et al. 2010), which exceeds the familiar proportional effect.

The physical origin of this "super-proportional" effect has not yet been explained. In this paper, the effect is shown along empirical data as well as its consequences for estimation of local Rn hazard through calculation of exceedance probability. Possible reasons are discussed, stemming (1) from data realm (spatial observation density, data uncertainty which may inflate sample dispersion) or (2) from the fundamental principle that generates the about-lognormal distribution (e.g. multiplicative cascades related to the multi-fractal nature of generating natural processes). In any case, the effect should be taken into account for modelling.

2. Material and Methods

Data analyzed here are taken (1) from the database underlying the European Indoor Radon Map (EIRM), part of the European Atlas of Natural Radiation, Cinelli et al. (2019). The database consists of statistics within 10 km × 10 km cells aligned to the European LAEA (Lambert azimuthal equal area) system. Each case (cell) contains the following statistics: n (number of original data), AM (arithmetical mean), SD (standard deviation), AML and SDL (AM and SD of ln-transformed data), minimum, median and maximum. The statistics are built by national competent authorities per country; the original data are not available to the authors of the EIRM for privacy and data protection reasons which are particularly sensitive for indoor Rn data. The version of the database used (3/2019) is based on about 1.2 million original data. It is not publicly available. (2) Additionally, a newer dataset of Germany only (2021) is investigated, based on about 58,000 original data (also not public).

3. Results

3.1. Log-normality

Under assumption of log-normality (LN), we have $m = \exp(\mu+\sigma^2/2)$ and $s^2=\exp(2\mu+\sigma^2)(\exp(\sigma^2)-1)$, where m and s – the mean and standard deviation, sample estimates AM and SD, μ and σ - the analogues for the ln-transformed population, sample estimates AML and SDL. Under LN, AM and AM':= exp(AML+SDL^2/2) must be equal, as must be SD^2 and SD'^{2} according to the formula above.

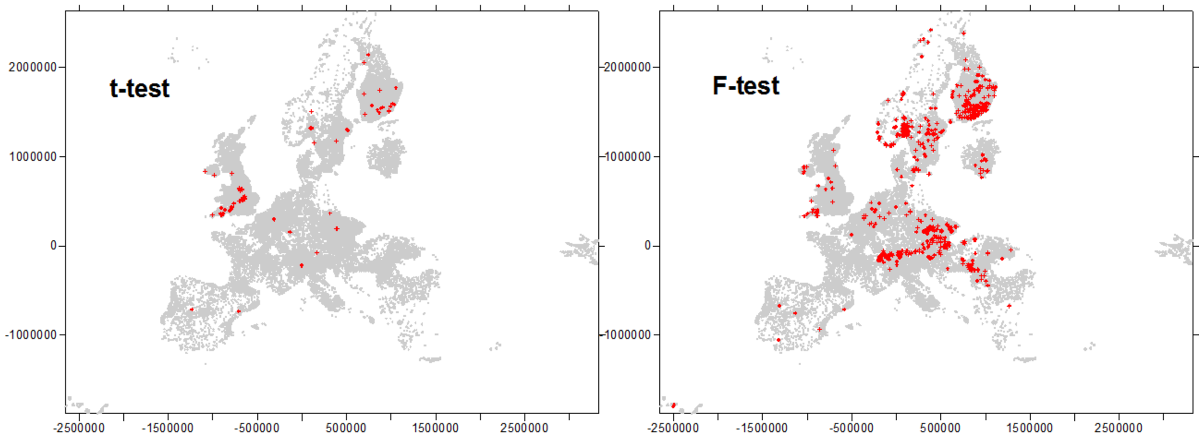


Figure 1 – Maps of cells (marked red) deviating from LN (see text). In this and following maps: North is up, axis units: m.

This can be tested by t-test (unequal variances) for the means AM and AM', and by F-test for the variances. $p>0.05$ are necessary conditions for LN, while $p<0.05$, sufficient for deviation of LN. For the European data, one finds $p(t)<0.05$ and $p(F)<0.05$ for 0.36% and 2.7% of all cells, respectively, which can be considered a sufficiently low rate of deviation from LN, to call Rn LN within 10km×10km cells in acceptable approximation. The geographical distribution of cells deviating from LN are shown in Figure 1. The pattern of deviation which appears in particular for $p(F)<0.05$ has not yet been further investigated. For Germany, the rate of deviating cells is even lower, 0.12% and 0.56%, respectively.

3.2. Bias of standard deviation

The sample standard deviation is biased, which under normality – assumed for the log data - can be corrected, SDLcorr = SDL $\sqrt{}((n-1)/2)\ \Gamma((n-1)/2)/\Gamma(n/2)$ (derivation e.g. in stackexchange 2012). An acceptable, frequently used approximation consists in replacing the common Bessel correction (n-1) in the sample SDL by (n-3/2) (e.g., Brugger 1969). Gurland (1971) proposed the correction factor $1+1/(4(n-1))$. For n=5, the exact correction factor equals 1.064, simple approximation 1.069, Gurland: 1.063. The raw variance SDL^2 is unbiased. The bias problem of SD is more serious for autocorrelated samples, which is actually the normal case; no easy solution for the 2-dim (i.e. spatial) case is known to me.

However, here we deal mostly with the variance and the geometric standard deviation GSD. In analogy to the geometrical mean, GM:=exp(μ), it is defined GSD:=exp(σ) or exp(SDL) for the sample estimate. Under LN, the sample GSD can be shown by simulation (I am not aware of an analytical treatment) to be almost unbiased if based on the common Bessel-corrected SDL.

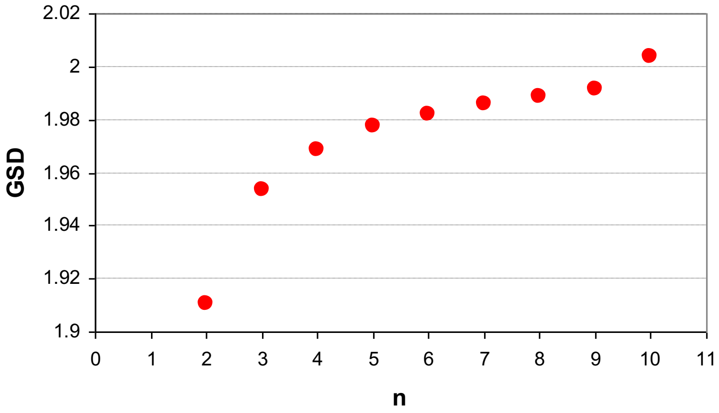


Figure 2 – Dependence of the empirical GSD on sample size. True population GSD=2 assumed.

The bias problem appears therefore alleviated, but still present. Fig. 2 shows the sampling GSD in dependence of sample size (n). For the simulation, true GM=100 Bq/m^3 (which does not matter) and GSD=2 was assumed and the empirical GSD of n values was computed. The points in the graph are means over 10,000 realizations. For low populated cells, the GSD is therefore underestimated up to about 5%. However, the bias depends on GSD: For GSD> about 2.5, the bias is positive.

For the real-world datasets, the relations of mean GSD per sample size are shown in Fig. 3. It seems that for low n (up to about 10 or 20) the empirical graphs reflect the bias effect. This can however not explain the shape for higher n.

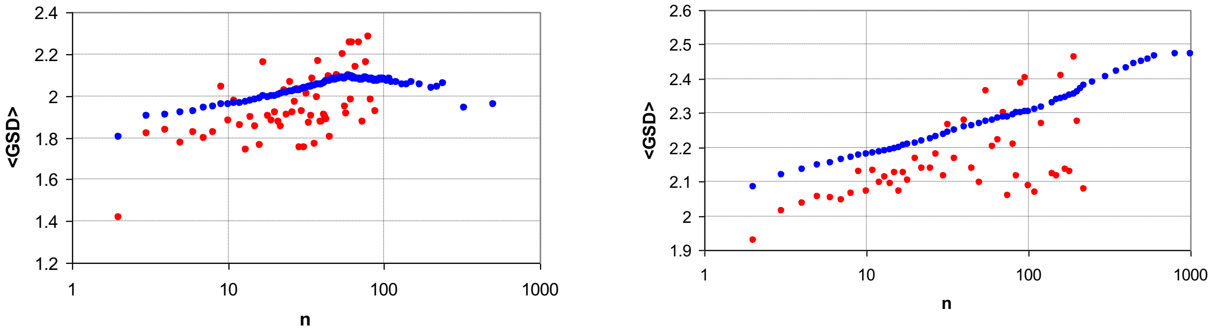


Figure 3 – Dependence of the mean empirical GSD, ⟨GSD⟩, within cells on sample size (n) per cell. Left: German dataset, Right: European dataset. Red: Mean GSD of cells with n measurements; blue: mean GSD of cells with ≥ n measurements.

3.2 Geographical distribution of the geometric standard deviation

The European dataset consists of 20,328 cells with n≥2 measurements. The AM(GSD)=2.08, the median equals 1.95. For the new German dataset, there are 2,457 cells (n≥2). The AM equals 1.81, the median, 1.72. The reason for the difference may be data definition: All Rn concentrations refer to residential rooms on ground floor; in Germany, buildings with full basement are defined as additional standard condition. Presence of basement is an important controlling factor and the additional constraint therefore reduces dispersion. German data are standardized per model, but it seems that model uncertainty introduces less variability than is removed by standardization.

Figures 4a and 4b show number of data n (i.e. observation density), GSD and AM per 10km×10km cell. One can visually notice the association between the statistics. The association is less clear in low populated cells, which is no surprise because few observations per 100km^2 are hardly representative, given the high spatial variability of Rn also in small-scale. (A less noisy, interpolated map taking advantage of geology as trend predictor has been shown by Elío et al. 2019).

3.3. Dependence of GSD on GM and logarithmic proportional effect

(The text of this par. is partly taken from Bossew 2021.) 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√(exp(σ^2)-1), i.e. they are proportional. Assuming local log-normality (or "permanence", e.g. Agterberg 1984; which ideally holds for LN multifractals, section 4.3), the proportional effect follows. Reversely, assume a power-type relationship between the local SD and the local AM (which seems to be realistic), SD=a AMb. Then CV = SD/AM ~ AMb-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. Here we are concerned with the apparently realistic case b>1, or "superproportional effect". Higher variability in high-Rn areas has indeed been empirically observed (e.g., Bossew et al. 2008, Dubois et al. 2010).

Define w:=ln(exp(SDL^2)-1). According to the above model, w is linearly related to ln(AM), w=a*+b*ln(AM) with a*=2 ln(a) and b*=2(b-1). Fig. 5 shows the scatter plots (w, ln(AM)) for the European and German datasets (for $n \geq n_{min}=10$ to reduce noise). For the "normal" case, $b=1$, $b^*=0$, no slope would be visible. (Scatter plots (GSD, ln(AM)) and (CV,AM) look similar.) For Europe, one finds $a \approx 0.38$ and $b \approx 1.18$, for Germany, $a \approx 0.13$, $b \approx 1.43$ (the parameters depend slightly on n_{min}.). The reason for the differences is unclear; either the model is insufficient or its parameters are themselves subject to regional variability, for whatever reason. It seems that this heuristic model makes sense until a theoretically grounded one has been found.

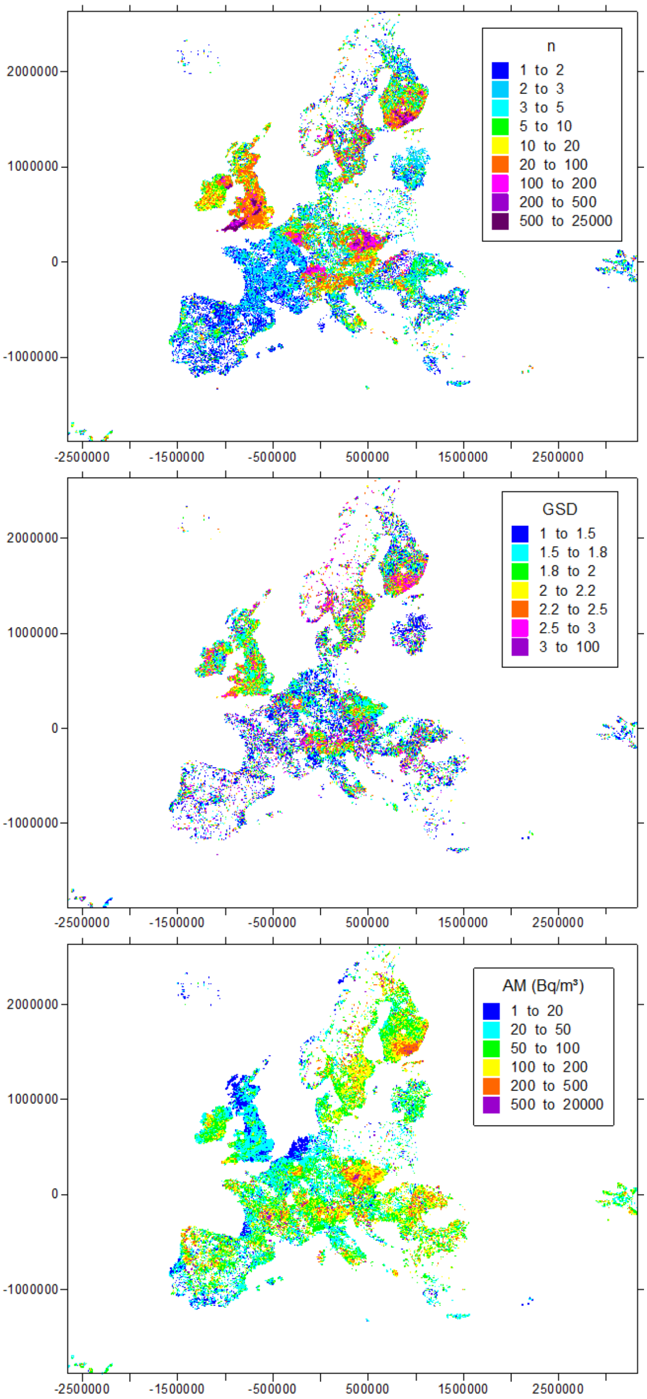


Figure 4a – European indoor Rn dataset (version 2019): n – number of observations per cell; GSD – geometrical standard deviation, AM – arithmetical mean Rn concentration per 10km×10km cell.

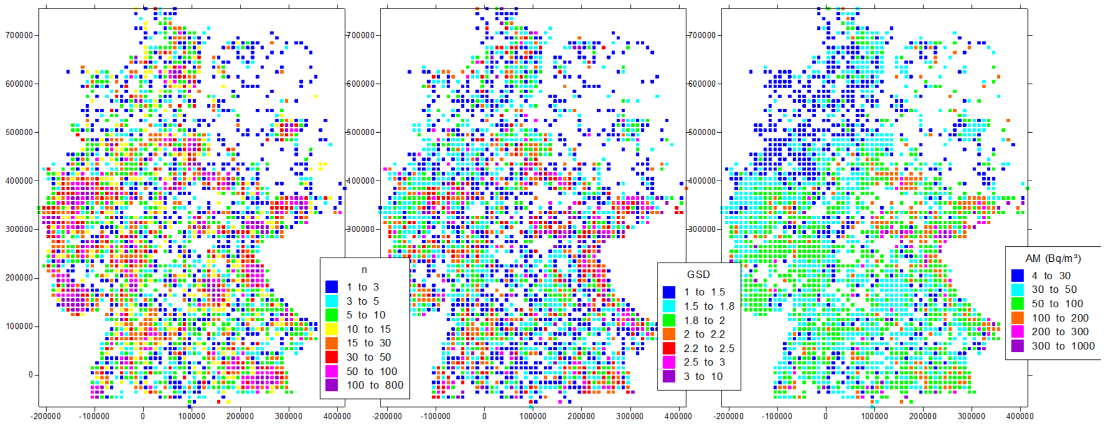


Figure 4b – German indoor Rn dataset (version 2021): n – number of observations per cell; GSD – geometrical standard deviation, AM - arithmetical mean Rn concentration per 10km×10km cell.

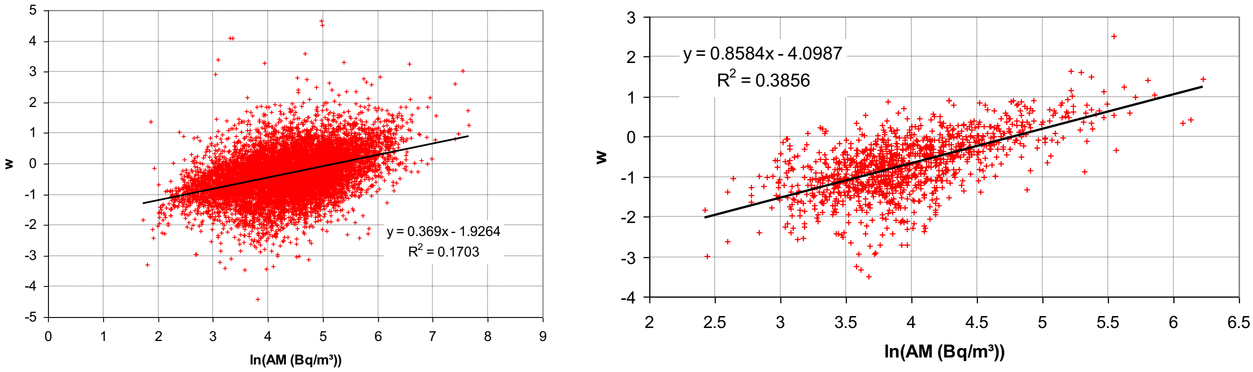


Figure 5 – Dependence of w (transformed empirical GSD) on cell AM. Left: European, right: German dataset.

3.4. Exceedance probability

Under LN(μ,σ²), the probability that Z exceeds a threshold or reference level z, equals p(z):=prob(Z>z)= (1/2)(1+erf((μ-ln z)/(σ√2)))=1-Φ((ln z-μ)/(σ√2)). Therefore, estimation of exceedance probability from the LN model depends on the choice of σ: from individual data; assuming a constant σ; or a modelled σ. For unknown μ and σ, this has to be estimated from the data based on the sampling distributions of p(z). Following Liebermann & Resnikoff (1955) and especially for Rn, Murphy & Organo (2008) (apparently with some mistakes in their eq. 13), p(z)=Beta(u_b,β), Beta – the cumulative symmetric beta distribution with β=(n-2)/2 and u_b=max[0,(1-(ln RL-AML)/(SDL √(n-1))]. Beard (1960) suggests p(z)=t_{n-1}[u_t √(n/(n+1))],

u_t=(ln RL-AML)/SDL; used in Bossew et al. (2015) and also in the following.

A common definition of Rn priority areas (RPA) is p(300 Bq/m³)≥0.1 (e.g., Bossew 2018). In Fig. 6, RPAs are mapped in a part of Central Europe (chosen because over entire Europe the differences are more difficult to recognize) according this definition; in p(I), estimated with GSD=1.95 (median over Europe); p(II): with empirical SDL (Fig. 4a, SDL=ln GSD); p(III): modelled GSD according section 3.3, with a=0.41 and b=1.16.

One notices that the gross patterns are similar, but that locally there are significant differences. In the future, the analysis may be refined by Bayesian reasoning, e.g. using p(III) as prior to estimate p(II).

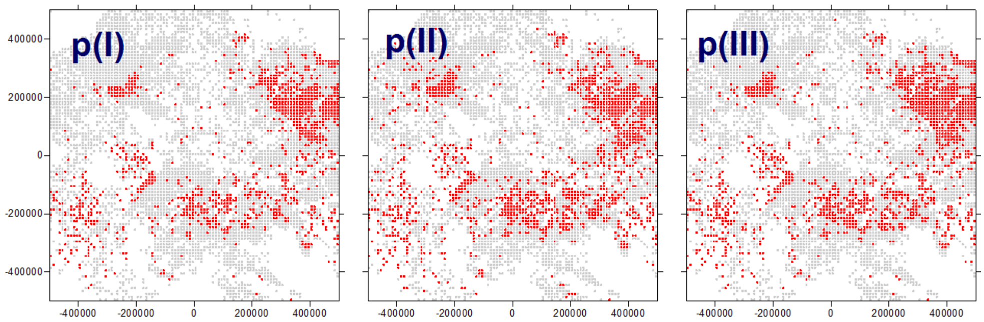


Figure 6 – Radon priority areas (red) in Central Europe, defined as prob(Rn>300 Bq/m³)≥0.1, calculated with 3 different models. Blank: cells with less than 2 measurements.

4. Discussion

The observed dependence of dispersion (GSD or CV) per cell on its mean may have different physical or statistical reasons. The following have been identified as possible so far: Data uncertainty and generating process as physical and estimation bias as statistical cause. Further may emerge in future investigation.

4.1. Data uncertainty

Relative uncertainty (unc) is higher for lower Rn concentrations. No commonly accepted unc model for usual indoor Rn measurements (by alpha track-etch detectors) exists; after all, different detector brands as used across Europe have different uncertainty. However, a model of decreasing unc leads to GSD decreasing with mean, as shown with the following artificial, but about plausible model: $unc(z)=u_1+(u_2-u_1)(z_1/z)^{\alpha}$, with $u_1=0.1$, $u_2=0.3$, $z_1=10$ Bq/m³, $\alpha=0.6523$ such that unc(50 Bq/m³)=0.17 and $unc \rightarrow u_1$ for $z \rightarrow \infty$.

Rn concentration was sampled ~LN(μ,σ^2), exp(σ)=GSD=2 assumed true spatial dispersion; then $z' := z(1+u)$, $u \sim N(unc(z), s_{unc})$, $s_{unc}=1$ assumed (in fact unknown); from several thousand realizations of z', which is the "observed" concentration, the GSD was computed and plotted in Figure 7 (error bars over replicates of 1000 realizations). As expected, this uncertainty model leads to dispersion inflation above the theoretical population dispersion σ (=ln(2)) or GSD=2, but cannot explain the increase of GSD with mean as observed in the real world.

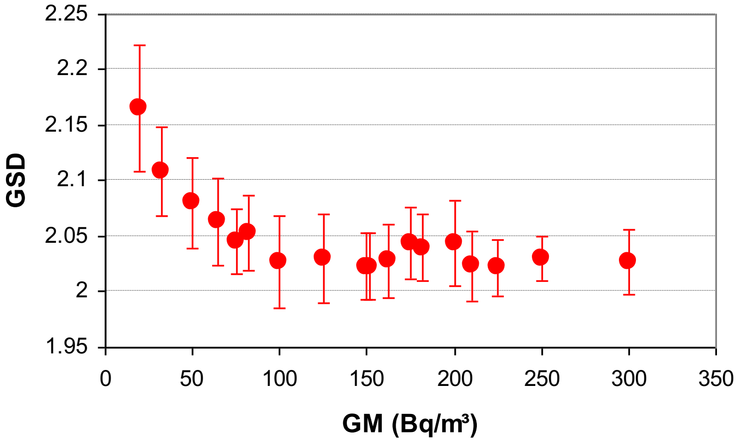


Figure 7 – Dependence of GSD on GM for an about realistic uncertainty model. True population GSD=2 assumed.

4.2. Sampling effect

The empirical GSD is very slightly biased. Often in environmental surveys, areas which are known or suspected for higher levels of the investigated quantity are sampled preferentially. In areas with higher levels sample density would therefore be higher (i.e. more measurements per cell), and consequently, due to the bias, the observed GSD higher. However, as shown in section 3.2, this could explain the effect only for low sample densities.

4.3 Generating process

The spatial variability of radon (geogenic Rn and indoor Rn which is largely controlled by the former) depends on the one of geogenic reality, represented by geology. Since about 1950, there has been much discussion about conceiving geochemical variability as result of multiplicative cascade, leading to multi-fractality and local LN (depending on the cascade model), not to be discussed further here (see e.g. literature quoted in Bossew 2021, section 2.4). However, the regional variability of the dispersion cannot be explained with a simple de Wijs model, which leads to constant GSD. One may hypothesize that the splitting factors are not geographically constant, or that the cascades are not equally developed in all geological regions. It seems that the first option can indeed give rise to the observed effect (to be further investigated elsewhere).

5. Conclusions

The availability of large Rn datasets – generated as a consequence of radon abatement policy which includes Rn surveys - allows discovering and investigating statistical effects which have so far been unknown. One key quantity in survey planning and in estimation of local exceedance probability and in consequence, status as radon priority area, is local dispersion. It has been shown that it is not spatially constant but slightly depends on the local mean. Exceedance probability can be estimated by geostatistical methods, typically indicator kriging or conditional simulation. If, on the other hand, one attempts to estimate it from data, as done here, much care has to be taken to estimate the dispersion (GSD) as correctly as possible. Applying a mean GSD (model p(I)) is probably a too rough approximation.

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.

Beard, L. R. (1960): Probability estimates based on small normal-distribution samples. J. Geophys. Res. 65(7), 2143–2148; https://doi.org/10.1029/JZ065i007p02143.

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. Environ. Radioactivity 101 (10), 826 - 834,. http://dx.doi.org/10.1016/j.jenvrad.2010.05.005.

Bossew P., Tollefsen T., Cinelli G., Gruber V. and De Cort M. (2015): Status of the European Atlas of Natural Radiation. RPD 167 (1-3): 29 - 36 doi:10.1093/rpd/ncv216.

Bossew P. (2018): Radon priority areas – definition, estimation and uncertainty. Nuclear Technology & Radiation Protection 33 (3), 286 - 292; http://doi.org/10.2298/NTRP180515011B.

Bossew P. (2021): Spatial dispersion of a field in an area in dependence of its size. GeoENV 2021, www.repository.unipr.it/handle/1889/4373 (accessed 30 March 2022).

Brugger R.M. (1969): A Note on Unbiased Estimation of the Standard Deviation. The American Statistician, 23 (4), 32. No doi available.

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. J. Environ. Radioactivity 101 (10), 786 - 798. http://dx.doi.org/10.1016/j.jenvrad.2010.03.007.

Elío J., Cinelli G., Bossew P., Gutierrez-Villanueva JL., Tollefsen T., De Cort M., Nogarotto A., Braga R. (2019): First steps towards an All-European Indoor Radon Map. Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2019-102.

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. doi: 10.2760/520053.

Gaskin J., Coyle D., Whyte J., Krewski D. (2018): Global Estimate of Lung Cancer Mortality Attributable to Residential Radon. Environ Health Perspect. 126(5):057009. doi: 10.1289/EHP2503.

Gurland, J., & Tripathi, R. C. (1971): A Simple Approximation for Unbiased Estimation of the Standard Deviation. The American Statistician, 25(4), 30. doi: 10.2307/2682923.

Lieberman, G. J., & Resnikoff, G. J. (1955): Sampling Plans for Inspection by Variables. Journal of the American Statistical Association, 50(270), 457. doi: 10.2307/2280972.

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 30 March 2022).

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.

Murphy, P., & Organo, C. (2008): A comparative study of lognormal, gamma and beta modelling in radon mapping with recommendations regarding bias, sample sizes and the treatment of outliers. Journal of Radiological Protection, 28(3), 293–302. doi:10.1088/0952-4746/28/3/001.

stackexchange (2012): Why is sample standard deviation a biased estimator of σ, URL (version: 2012-05-08): https://stats.stackexchange.com/q/27984 (accessed 15 Mar 2022).

Zeeb H., Shannoun, F. & World Health Organization. (2009): WHO handbook on indoor radon: a public health perspective. World Health Organization. https://apps.who.int/iris/handle/10665/44149 (accessed 30 March 2022).

SPATIOTEMPORAL PREDICTIONS OF MULTIPLE AIR POLLUTION DATA: A CASE STUDY

Claudia Cappello (1) - Sandra De Iaco (1) - Monica Palma (1)*

University of Salento, Dept. of Economic Sciences, Lecce, Italy (1)
** Corresponding author: monica.palma@unisalento.it*

Abstract

Air quality is one of the most serious issues for numerous world's major cities. Due to the dangerous effects of outdoor air pollution to human health, the International Agency for Research on Cancer has classified air pollution as carcinogenic to humans. Several studies have established a cause-and-effect relationship between high concentrations of common air pollutants, such as nitrogen dioxide (NO2), carbon monoxide (CO), ground-level ozone (O3), and particulate matter (PM2.5 and PM10) and the development of respiratory diseases. In the modern society, the advanced systems for the continuous monitoring of pollutants' concentrations have become fundamental to keep control of the air quality and support the policy decision makers to promptly act to prevent environmental risks.

In Europe, the European Environment Agency (EEA) collects and maintains European air quality database, consisting of multi-annual time series of air quality measurements for the air pollutants recorded at the environmental stations of the monitoring networks of the EU Member States. Such large datasets can be adequately analyzed by multivariate spatiotemporal geostatistical methods for modelling and prediction, in order to obtain further information about the air quality over the area of interest. In particular,

- spatiotemporal variability scales,
- relationships in space-time among two or more air pollutants,
- possible air pollution levels at unobserved points,
- probabilities of exceeding a limit of attention,

represent additional information which can be of interest for both researchers and environmental decision-makers.

In the literature, the spatiotemporal linear coregionalization model represents the most common model used to describe the correlation in space-time which characterizes the multiple variables under study. Thanks to its computational flexibility, the above model has been used in several studies and recently some computational advances have been proposed with the twofold aim of choosing the most appropriate basic models (covariances) at the different spatiotemporal variability scales shown by the data, as well as simplifying the modeling stage. The fitted model is then used to obtain stochastic predictions through cokriging: in this stage of the analysis, the space-time linear coregionalization model with different basic covariance models allows the data scientists to obtain reliable predictions for the pollutant of interest, over the spatiotemporal domain.

In this paper, the above-mentioned tools have been used to deeply study air pollution in Germany, where nitrogen dioxide and particulate matter pollution are still issues in metropolitan German cities. On the basis of air quality EEA database, the hourly measurements of PM10, PM2.5 and NO2 recorded during 2021 at several monitoring stations over Germany have been analyzed. In particular, a linear coregionalization model based on mixture models, related to different scales of spatio-temporal variability, has been fitted to the empirical multivariate correlation matrix. Then, stochastic predictions for each analyzed pollutant have been

obtained by spatiotemporal cokriging interpolation and finally, through nonparametric estimation methods applied in the multivariate case, risk maps of the probability of exceeding some fixed pollution thresholds have been realized for the pollutants under study.

GEOSPATIAL MODEL OF COMPOSITION OF WATER SERVICE LINES IN FLINT, MI: VALIDATION USING EXCAVATION DATA AND A NEW COMPOSITIONAL GEOSTATISTICAL APPROACH

Pierre Goovaerts (1)*

Biomedware, Inc., Ann Arbor, United States (1)
** Corresponding author: goovaerts@biomedware.com*

Abstract

In the aftermath of Flint drinking water crisis, a service line replacement program was implemented to identify lead and galvanized service lines (SL) connecting residences to Flint's water system and replace them. This program led to the excavation and inspection over a 3-year period (2018-2020) of a total of 26,750 lines, representing close to 50% of all tax parcels in the city of Flint. These recent data were used to validate an earlier geospatial model created by residual indicator kriging to predict the likelihood that a home has a lead, galvanized or copper SL based on neighboring field data (i.e., house inspection conducted in 2017 at 3,254 homes) and secondary information (i.e., construction year and city records). Receiver Operating Characteristic Curves indicated an average frequency of detection (i.e., Area Under the Curve (AUC)) of 0.9 for copper and galvanized service lines, and 0.6 for lead SLs. Predicting the composition of SL at unmonitored residences by indicator kriging, however, can result in negative probabilities of occurrence and probabilities that do not sum to 1. These limitations were overcome by adopting simplicial indicator kriging whereby data undergo a log-ratio transform before the geospatial analysis and mapping. This first application of a compositional approach to service line data improved the detection of lead service lines (AUC= 0.74 vs 0.6) while providing coherent predictions. As for the traditional (i.e., non-compositional approach), better predictions are obtained when incorporating secondary information and there is no benefit in using cokriging to account for multiple SL data at each location as it is an equally-sampled or isotopic case.

GEOSTATISTICAL MAPPINGS OF INDOOR RADON CONCENTRATIONS DATA IN FRANCE

Jean-Michel Metivier (1)* - Claire Greau (1) - Nahla Mansouri (1)

IRSN, Environment, Fontenay-aux-Roses, France (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.

The study proposed here carries out a geostatistical study on the scale of the French territory from more than 30,000 measured values.

For geolocated data, an OK was performed. The geocoded data at the centroid of the municipality were also taken into account and an OK with change of support (deconvolution) was carried out.

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.

INTERPOLATED SURFACES AS MAPPING UNITS FOR BINARY CLASSIFICATION OF RADON PRONE AREAS: A CASE-STUDY FROM CENTRAL PORTUGAL

Gustavo Luís (1)* - Alcides Pereira (1)

CITEUC, Department of Earth Sciences, University of Coimbra, Coimbra, Portugal (1)
** Corresponding author: gustavo.psl96@gmail.com*

Abstract

Radon is a naturally occurring radioactive gas that is formed from the decay of Radium present on rocks and soils. This gas accumulates into the dwellings up to indoor radon concentration (IRCs) that could be a major health concern. It exposes the lungs to ionizing radiation, increasing the risk of lung cancer. The European Council Directive 2013/59/EURATOM establishes that Member States should identify Radon Prone Areas (RPA), where a significant percentage of dwellings have IRC above the national reference level. How to map these areas is still a topic under debate. The definition of the basic mapping unit is an important pillar of the methodologies and are per se one of the divergences. It could be a grid cell, a lithological unit or an administrative unit. Map scale is another important characteristic, since data can refer to a regional scale or to a smaller one (e.g., municipal scale). Inside a RPA identified at a regional scale, as the detail and variability increase at smaller scales, it's possible to differentiate smaller RPAs and non-RPAs, or statistically differentiate areas with higher and lower priority, according to the percentage of dwellings above the RL, for mitigation purposes.

Dwellings' locations are spatially biased, because they agglomerate into towns. This bias, that intrinsically extends to IRC distribution, is not well suited for the most common methods of geostatistics. Descriptive or inference statistic are frequently used on IRC or other proxy variables. The concept of RPA or non-RPA is binary and thus a binary classification system should be optimized. At a municipal scale, the mapping units can greatly impact the results. Reducing the grid cell size would increase the number of grid cells without data. Increasing it, the detail vanishes and the map loose spatial variability. A cost-benefit relation should be considered. The use of lithological limits at that scale would suffer the same lack of intern variability.

Our work aims to find an optimal classifier limit on interpolated surfaces of a proxy variable of IRC, namely Total Gamma Radiation (TGR), through Receiver Operating Characteristic (ROC) curve analysis, using a region of Central Portugal as a case study. TGR data can be spatially unbiased and so is suited for the most common geostatistical methods like ordinary kriging. The most promising advantage of the use of interpolated surfaces is the variability detail preserved in the ROC curve analysis and passed to the map of RPAs, with TGR isolines as the mapping units. The Matthews Correlation Coefficient values obtained (0.33 and 0.25 for the two studied areas) indicate a good correlation between the observed and predicted binary classifications, comparable to results obtained in other studies that use grid cells as mapping units. The resulting binary classifications using interpolated surfaces are region-specific and should not be extrapolated to other areas. Future work is aiming to study the uncertainty of the RPA map associated to the interpolations. Geostatistical simulation techniques can generate multiple interpolated surfaces, each used to establish an optimal classifier limit whose originate multiple RPA maps.

A SPATIO-TEMPORAL MULTILEVEL MULTIVARIATE MODEL TO EVALUATE THE DETERMINANTS OF AIR POLLUTION IN APULIA REGION

Sabrina Maggio (1)* - Claudia Cappello (1) - Sandra De Iaco (1)

University of Salento, Dept. of Economic Sciences, Lecce, Italy (1)
** Corresponding author: sabrina.maggio@unisalento.it*

Abstract

One of the most critical risk factors to human health and environment is represented by air pollution, since it can cause cardiovascular diseases, lung cancer, chronic, acute respiratory diseases, as well as damages to the ecosystem. The study of air pollution concentration has to consider that it is vehiculated in the atmosphere and that it can considerably change over time and space.

The purpose of this work is to apply a three-level multivariate model to detect the key spatio-temporal determinants influencing air quality in Apulia region. For this reason, the dataset of the concentrations of the most harmful air pollutants (such as particulate matter, ground-level ozone and nitrogen dioxide), collected by the ARPA (Regional Agency for the Protection of the Environment) through the monitoring network of air quality stations is considered together with the observations related to some meteorological variables (such as atmospheric pressure, rainfall, atmospheric temperature, wind velocity), as well as the information on the type of area where the monitoring stations are located (traffic/city center, residential, rural and industrial sites). In particular, the pollutants and the meteorological variables are measured in 23 monitoring stations distributed in Apulia region, during the year 2019.

The model is fitted in order to assess the dependence structure of the three above mentioned pollutants (PM, O_3 and NO_2) from the meteorological variables (such as Atmospheric Pressure, Rainfall, Atmospheric Temperature, Wind Velocity) and the type of area where the monitoring stations are located (traffic/city center, residential, rural and industrial sites). This model is also exploited in combination with the geostatistical techniques, for prediction purposes, over space and time.

Acknowledgments

This contribution is supported by the Project "Rigenerazione Sostenibile dell'agricoltura nei territori colpiti da Xylella fastidiosa".

OPERATIONAL SPATIALISATION OF HOURLY AND DAILY WEATHER DATA (AIR TEMPERATURE AND RELATIVE HUMIDITY) FOR AGRICULTURAL DECISION SUPPORT SYSTEMS

Damien Rosillon (1)* - Alban Jago (1) - Jean Pierre Huart (1) - Viviane Planchon (1) - Michel Journée (2) - Patrick Bogaert (3)

Walloon Agricultural Research Centre, Agriculture, Territory and Technologies Integration, Gembloux, Belgium (1) - Royal Meteorological Institute of Belgium, Weather and Climate Information Service, Brussels, Belgium (2) - Uclouvain, Earth and Life Institute, Louvain-la-Neuve, Belgium (3)
** Corresponding author: d.rosillon@cra.wallonie.be*

Abstract

Weather-based forecasting models play a major role in agricultural decision support systems (DSS) 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. Spatialization could be a solution to estimate weather conditions in farmer's field but it must meet specific operational constraints: near real-time spatialization, high temporal resolution and robust methods.

This is the goal of the Agromet platform (www.agromet.be) an operational web-platform designed for real-time agro-meteorological data dissemination at high spatial (1 km x 1 km grid) and temporal (hourly and daily) resolution in Wallonia, southern part of Belgium. Two meteorological parameters are interpolated: air temperature and relative humidity.

The poster will present the methodology used to choose the best spatialization models, the main conclusions and some research perspectives for the Agromet platform.

Material and method

Two datasets of meteorological observations are used: a first dataset comes from the Pameseb network of the Walloon Agricultural Research Centre CRA-W (28 selected AWS) and a second one comes from the Royal Meteorological Institute network (8 selected AWS). Five algorithms of spatialization are tested: nearest neighbor, inverse distance weighted, multilinear regression, ordinary kriging and kriging with external drift. Four explanatory variables are tested: longitude, latitude, elevation (all three static variables) and gridded weather forecasts (a dynamic variable).

Models are trained using two years of hourly and daily air temperature and relative humidity measurements. Quality of the prediction is assessed by a leave-one-out cross validation. The mean absolute error is used as performance indicator. In addition to the overall performance scores, a more in-depth analysis of "worst cases" is carried out to understand the reliability of the scenarios.

Main conclusions

Kriging with elevation as external drift is the scenario with the best score in all cases i.e. for both temperature and humidity and for hourly and daily steps. Integrating weather forecast as a dynamic explanatory variable seems not to improve the quality of the spatialization but this requires further studies. Enriching the training dataset by increasing the number of AWS (from 28 to 36) does not dramatically improve the overall performance score but the issue is complex: the positive or negative impact of the integration of additional stations differs both by scenario and by station.

This first operational version of the Agromet platform is a solid foundation on which to base future developments. However, under certain atmospheric conditions like temperature inversion, local weather is not well modelized and the impact on DSS simulations must be assessed.

Research perspectives

Future research will focus on a better simulation of local weather conditions at mesoclimate scale suitable for agronomic models (potato late blight, orange wheat blossom midge and wheat phenology model).

Possibilities for improvement are: adding new explanatory variables (meteorological satellite images, weather forecasting model outputs), integration of farmers AWS to increase the size of the training dataset or using new prediction methods from machine learning (random forests, neural networks).

USING OF SEQUENTIAL INDICATOR SIMULATION TO MODEL NON-STATIONARY GEOLOGICAL DOMAINS COMBINING WITH A MACHINE LEARNING ALGORITHM

Almas Amirzhan (1)* - Nasser Madani (1)

School of Mining and Geosciences, Nazarbayev University, Nur-Sultan, Kazakhstan (1)
** Corresponding author: almas.amirzhan@nu.edu.kz*

Abstract

Sequential indicator simulation is a widely used method for simulation of variables in 3D geological modeling, which can be utilized not only for continuous variable but also for categorical variables. This method is designed to aid in the characterization of uncertainty on the structure or behavior of natural geological systems. However, there are legitimate criticisms against the sequential indicator simulation technique. Among others, the traditional algorithm is not able to deal with the non-stationary assumption of geological domains, as it is conventionally built upon the stationary properties of such variables. This paper addresses the problem of sequential indicator simulation for modeling the non-stationary geological domains. A machine learning algorithm called logistic regression is proposed to model the spatial probabilistic variability of the non-stationary geological domain by using only the conditioning data. Logistic regression is a machine learning classification algorithm used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable containing data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts a dependent variable as a function of one independent variable. In this paper, we deal with three categories, meaning that there is more than one possible categorical outcome. As a consequence in this case, we consider it as a "Multinomial Logistic Regression". This machine learning method allows the creation of a probability map for each categorical variable, which in our proposed method; they are considered as geological domains. This map can then be used as soft information (secondary data) in sequential indicator simulation to generate the non-stationary geological domains. To test the algorithm, a synthetic map is simulated, where the geological domains show a strict heterogeneity. We used this as the reference map to evaluate the performance of the algorithm. Then, the map is sampled randomly to obtain a set of conditioning data. The conditioning data is split into two sub-datasets - test and train values. The 20% of the conditioning data is considered for the test target points and the rest of 80% is employed as the train dataset. Using Multinomial Logistic Regression method, a machine learning system is trained and tested using the corresponding datasets. The obtained regression model is then used to predict the probability of each geologic domain at each single target location of the reference map. These maps are then used as soft information in one of the variants of sequential indicator simulation algorithm that uses non-stationary simple kriging with residuals from the locally varying mean probabilities. The proposed algorithm is applied to compare the results with the traditional sequential indicator simulation algorithm, which does not use soft information. This comparison showed that the Multinomial Logistic Regression algorithm is properly able to model the non-stationary geologic domain in the region and one can use it for modeling the heterogeneity.

Keywords: Sequential indicator simulation; Logistic Regression; Categorical variable; Non-stationary geologic domain.

1. Introduction

For categorical data, indicator approaches are widely utilized. For each rock type, different indicator variograms are being used. Secondary data may readily be coded as soft indicator or soft probabilistic data that can be suitable for modeling the non-stationary phenomena. Many people utilize indicator kriging (IK) (Cressie, 1989) and sequential indicator simulation (SIS) for modeling the geological domains. The SIS algorithm is a widely used tool for geostatistical modelling (Journel and Isaaks, 1984; Journel, 1983). This stochastic paradigm simulates the geological domains at the target block locations. The traditional SIS method, commonly used in industry, shows acceptable results when working in stationary structures. However, this method becomes problematic in the case of dealing with non-stationary geological domains. The reason is that traditional SIS is developed based on stationary assumptions. In heterogeneous characteristics of rock variabilities, the realization produced by traditional sequential indicator simulation may appear very patchy and unstructured. The simulated categories can be observed all over the deposit (pretending homogeneity), making this method extremely unreliable. To circumvent this problem, using soft data in SIS algorithm can be of great help as this information instruct the algorithm to produce the non-stationary geological domains. The soft data can be obtained from geophysical data and geological interpretation (Deutsch, 2006). The former is rarely available in mining industry. Use of the latter is quite subjective since it is very time consuming to produce a reliable model. Another issue concerning such deterministic models is how to convert them to local probabilities so that it can be used as soft data in the SIS algorithm. This step is not also quite straightforward. An algorithm is presented in the study that uses a machine-learning algorithm to produce such soft information. The method is very quick in identification of local probabilities and needs the minimum intervention of the practitioner. In the following, a theoretical background of the method proposed is discussed. The applicability of the method is also tested over a synthetic case study.

2. Material and Methods

2.1. Traditional sequential indicator simulation (SIS)

The SIS algorithm uses IK to estimate the probability density function (pdf) of categorical variable Z for a categorical variable (u). It uses a combination of indicator formalism and the sequential paradigm to mimic a non-parametric distribution (Remy et al. 2008). A sequence of alternative, equally likely realizations of an indicator variable z(u) distribution are generated using stochastic simulation. For example, if z(u) from category k is simulated at spatial site u, the predicted pdf is as follows:

$$\mathrm{Prob}\{\mathrm{I(u) = 1|(n)}\} = \mathrm{E\{I(u)|(n)\}} \tag{1}$$

Assume that $i(u;\ zk)$ is category z_k. It turns to 1 if u belongs to z_k, otherwise 0. Mutual exclusion must meet the following criteria:

$$\mathrm{i_k(u) i_{k'}} = 0, = \ \forall \mathrm{k} \ \neq \mathrm{k'} \text{ and} \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{i(u; z_k)} = 1 \tag{2}$$

In reality, meeting those two requirements will be mutually exclusive and exhaustive. For example, when simple kriging is used to estimate the probability of variables z_k on location u, $\mathrm{i}^*_{SK}(\mathrm{u;k})$, the following results are obtained:

$$\mathrm{i}^*_{SK}(\mathrm{u;k}) = \ \mathrm{p_k} + \sum_{\alpha=1}^{\mathrm{n}} \lambda_\alpha [\mathrm{I(u; z_k)} - \ \mathrm{p_k}] \tag{3}$$

where $\mathrm{p_k}$ is global proportion of category z_k and λα is the weight of sample in kriging system. The SIS procedure is as following: (1) Create a route that passes through all of the simulated places. (2) For each u along the route: (a) Get the data for neighboring category conditioning: $\mathrm{z(u_\alpha)}, \alpha = 1, \ldots, \mathrm{N}$. (b) Solve a kriging system to estimate the indicator random variable I(u; $\mathrm{z_k}$) for each of the K categories (Eq. 3). (c) After correcting order relation problems, estimate values of $\mathrm{i^*(u;\ z_k) = Prob^*(Z(u) =\ z_k)}$, define an estimate of the discrete conditional probability density function (cpdf) of the categorical variable Z(u), and draw a realization, by using Monte Carlo simulation from cpdf and assign it as a datum at position u. After then, the preceding simulated values can be utilized as conditioning data for the subsequent unsampled site. (d) Repeat until all of the location have been visited. (3) To make new realizations, repeat the preceding stages with different random number in Monte Carlo simulation (Deutsch and Journel 1998).

2.3. Multinomial Logistic Regression (MLR)

Most multivariate analysis techniques require the basic assumptions of normality and continuous variables, involving independent and/or dependent variables as aforementioned. MLR exists to handle the case of dependents with more classes. This is referred to as the multivariate case. Thus, it is expected that the multinomial logistic regression approach would do better when there is evidence of substantial departures from multivariate normality, as is the case where there are some dichotomous or zero/one variables or where distributions are highly skewed or heavy-tailed, especially in dynamic settings. Tabachnick et al. (2007) argued that the multinomial logistic regression technique has several significant advantages as a summary to the discussion above: (1) it is more robust to violations of assumptions of multivariate normality and equal variance-covariance matrices across groups; and (2) it is similar to linear regression, but more easily interpretable diagnostic statistics. Widely use MLR as a problem-solving tool, particularly in medicine, psychology, mathematical finance, and engineering, due to the above advantages listed. This listed relevance attracted the present author's attention to the study case described in this paper. Multinomial Logistic Regression algorithm allows predicting each point of categorical variable. The most important thing is that only MLR can calculate the probability of belonging those points to the first, second or third category and determines the probability of category to which most likely the categorical variable will belong. This data is vital for residual calculation, which will be implemented in our proposed approach.

2.4. Non-stationary sequential indicator simulation

Non-stationary sequential indicator simulation offers a reliable algorithm to model the categorical data with heterogeneous characteristics. This algorithm uses a non-stationary simple kriging with residuals from the locally varying mean probabilities (Deutsch, 2006):

$$\mathrm{i}^*_{LM}(\mathrm{u;k}) = \mathrm{p_k(u)} + \sum_{\alpha=1}^{\mathrm{n}} \lambda_\alpha[\mathrm{I(u;z_k)} - \mathrm{p_k}(u_\alpha)] \tag{4}$$

In this formula, one is working with the residuals $\mathbf{I(u;z_k)} - \mathbf{p_k}(\boldsymbol{u_\alpha})$ over the sample points that are obtained from using a regression function. Therefore, a regression function should be fitted over the sample points to derive the residuals; then the same regression function should be used to predict the values at target location to derive $\mathbf{p_k(u)}$. Variogram analysis should be implemented over the residuals at sample points. Since $\boldsymbol{p_k(u)}$ and $\boldsymbol{p_k(u_\alpha)}$ in Eq. (4) are probabilities, one needs to use a regression functions flexible to produce those local probabilities in this algorithm. In this study, we propose to use MLR to infer such probabilities over the sample points and target grids. As far as the probability $\boldsymbol{i^*_{LM}(u;k)}$ estimated at target location, the further processing steps are similar to traditional sequential indicator simulation. Hereafter, SIS-lm and SIS-trad refer to our proposed and traditional sequential indicator simulation approaches, respectively.

3. Results

The proposed sequential indicator simulation using local probability means tested on a synthetic data set. This map obtained by using plurigaussian simulation (Madani, 2021) with an anisotropy with maximum continuity along North direction (Fig. 1). As can be seen, the produced map shows a very strict heterogeneous characteristic where the category 1 (blue) is located on the left, category 2 (green) is located on the center, and category three (red) is located on the right hand side of the grid. Then, 50 and 100 samples randomly selected from this map to supply the conditioning data in the simulation algorithms. As mentioned before, MLR is an appropriate method when data is not binary; in this case, the model has to predict values over three categories. Procedure performed several times by the different number of random values for assessing the quality of the Logit algorithm. To do so, 20% test and 80% train split those points' values. The classification report shows that the estimated accuracy of the MLR model is 82%. Attempts to increase accuracy by tuning parameters and changing test/train ratio did not significantly affect the absolute accuracy for both cases.

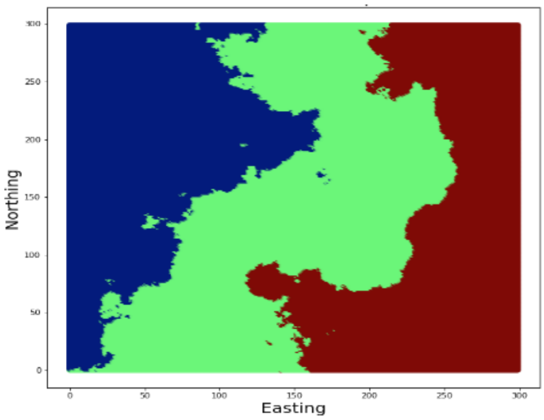


Figure 1 – Reference map; blue: category 1, green: category 2, and red: category 3.

It should be mentioned that before implementing the SIS-lm, it is necessary to codify the data into the proper category (0 or 1) and calculate residuals for each category: $Residuals = probability - indicator$. Therefore, the experimental variogram has been calculated for each indicator (for SIS-trad) and their residuals (for SIS-lm) along several directions (multidirectional) with a minimal tolerance to quantify the underlying anisotropy for each category separately. The intuitive results showed significant continuity along Northing as it was expected. Behind the experimental analysis of variogram, it is crucial to fit a justified theoretical variogram to the experimental ones. As a result, spherical theoretical variogram model was fitted manually. In order to provide unbiased results for this case study, 100 realizations were produced using the proposed algorithms with local means (SIS-lm) and traditional SIS (SIS-trad). The grid dimension of target block was considered 300× 300×1 identical to the reference map. As it was expected, SIS-trad shows an unstructured and patchy results in both cases where using 50 points and 100 points (Fig. 2, Fig. 3). However, the results produced by SIS-lm showed that the realizations bear resemblance to the reference map. Here, to illustrate the results, realization #20 selected randomly for 50 points and 100 points (Fig. 2, Fig. 3).

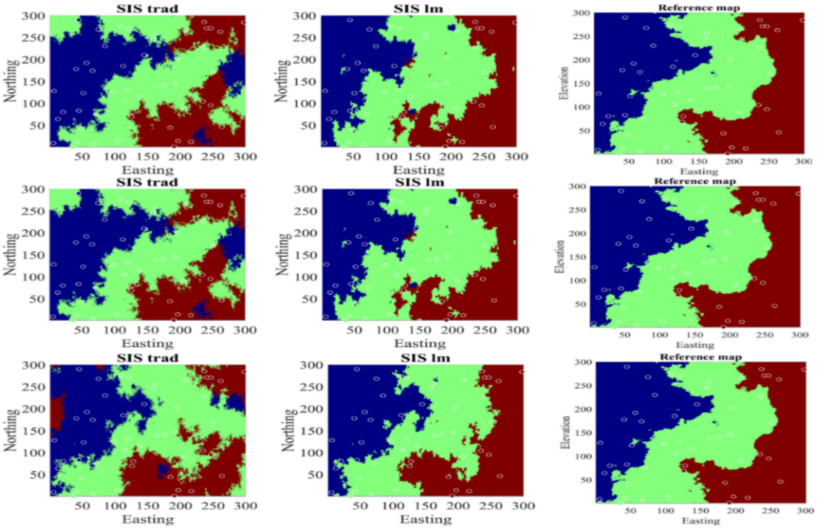


Figure 2 – Comparison of realizations obtained by different techniques for 50 points; blue: category 1, green: category 2, and red: category 3.

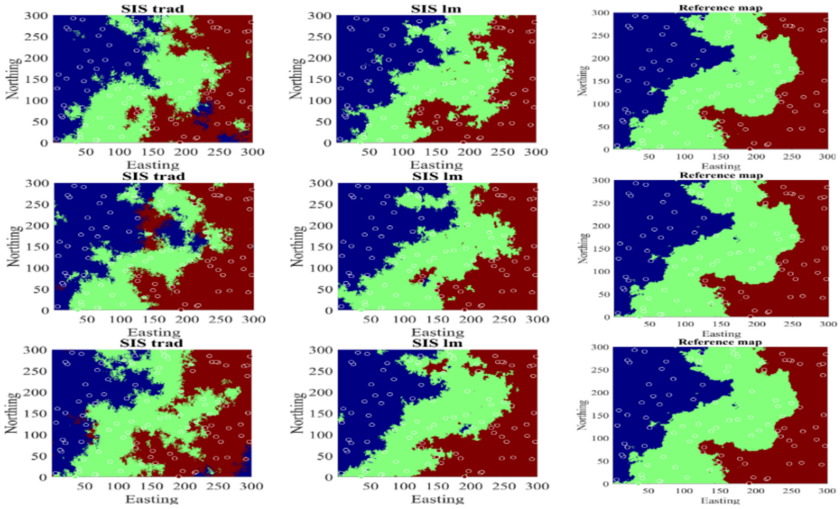


Figure 3 – Comparison of realizations obtained by different techniques for 100 points; blue: category 1, green: category 2, and red: category 3.

Geological uncertainty is critical in orebody evaluation, as it can be proposed in the layout of the boundaries (Emery 2007). This uncertainty can be presented by probabilistic modelling of each categorical domain (conditional simulation). Probability maps are assessed at a local scale for each categorical domain to quantify the uncertainty. The maps are constructed by calculating, for each block, the frequency of occurrence of each rock unit over the 100 conditional realizations. They show the risk of finding a mineralized zone different from others. The sectors with little uncertainty are those associated with a high probability for a given rock unit, indicating that there is little risk of not finding this rock unit, or those associated with a very low probability, indicating that one is pretty sure of not finding this unit. In contrast, the other sectors (painted in light blue, green, or yellow are more uncertain. As can be seen from Figures 4 & 5, the proposed approach produced the more strong certainty of the presence of categories over the expected areas that conditioning dataset might be scarce.

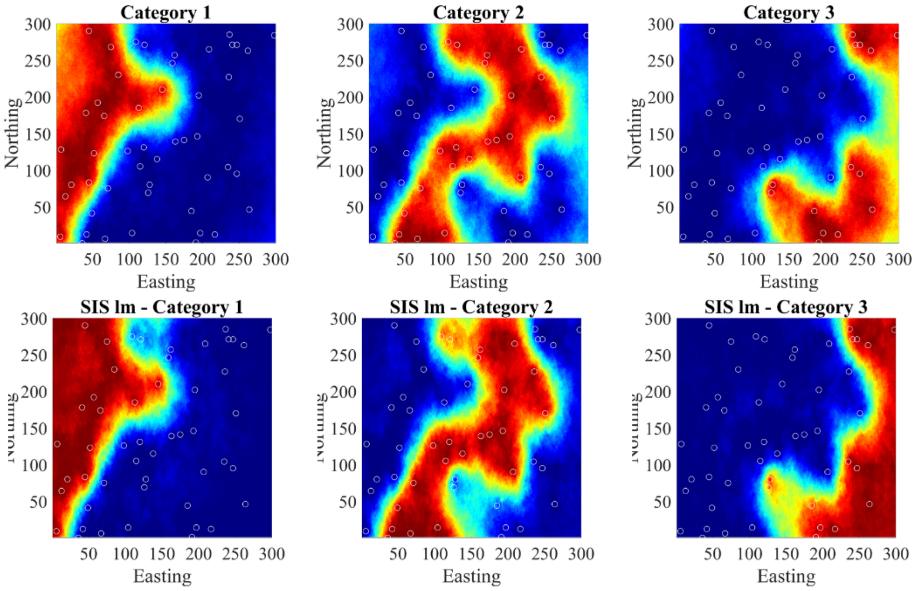


Figure 4 – Comparison of probability maps of each categories obtained by different techniques for 50 points.

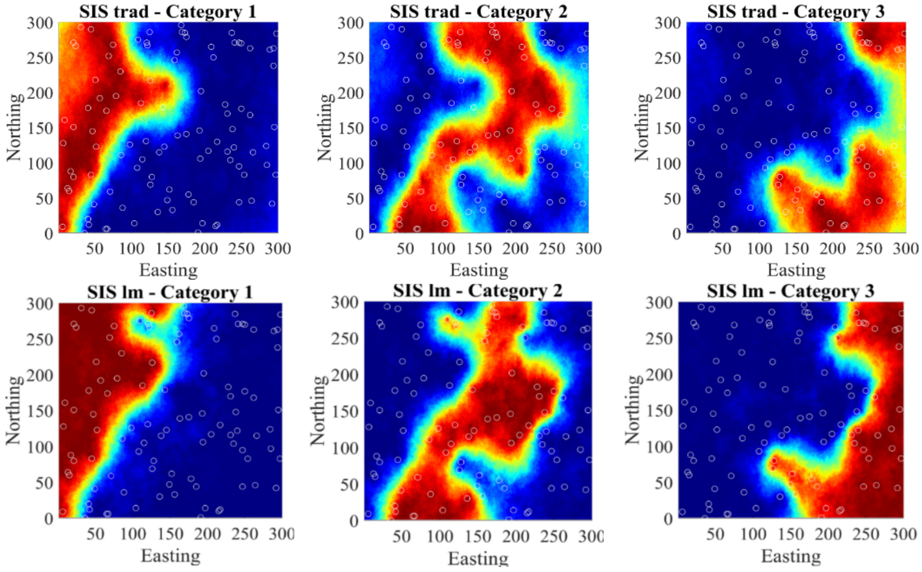


Figure 5 – Comparison of probability maps of each categories obtained by different techniques for 100 points.

In order to validate the realizations, it is necessary to calculate the frequency of each category along with the simulated results. This measure of global uncertainty provides an intuitive tool to compare with the properties of the experimental sampling data. It can also show how the proportion of each category is reproduced over each realization. As can be seen from Table 1, the average global proportions are presented for the categories evaluated by different approaches over 100 realizations. There is a slight difference between the global proportions of each category for the realizations obtained by proposed approaches (Table 1). Indeed, both methods produced the global proportions properly.

Another method for uncertainty evaluation is to calculate the relative error (RE) between reproduced proportions and original proportions. Table (2) compares RE of SIS-trad and SIS-lm for 50 and 100 points. As can be seen, the relative error is much less when using SIS with local mean as proposed in this study.

Table 1 – Comparison of global proportions reproduced by SIS-trad and SIS-lm with an original proportion.

	Category 1	Category 2	Category 3
Original proportion (Reference map)	0.296	0.401	0.302
SIS- lm (50 points)	0.318	0.411	0.270
SIS-trad (50 points)	0.297	0.422	0.279
SIS- lm (100 points)	0.324	0.373	0.301
SIS-trad (100 points)	0.303	0.342	0.354

Table 2 – Comparison of relative errors evaluated by SIS-trad and SIS-lm with original proportion.

	Category 1	Category 2	Category 3	Sum of errors
SIS- lm (50 points)	0.074	0.025	-0.106	-0.006
SIS-trad (50 points)	0.006	0.052	-0.076	-0.016
SIS- lm (100 points)	0.096	-0.06	-0.004	0.024
SIS-trad (100 points)	0.024	-0.147	0.171	0.048

4. Discussion and Conclusions

The use of sequential indicator simulation with local mean probabilities and residuals calculated from utilizing Multinomial Logistic Regression is a viable approach for building realizations of non-stationary geological domains. The resulting algorithm correctly represents each geological domain compared with traditional SIS. Moreover, realizations and probability maps produced by the proposed algorithm result in an increased accuracy and reduced the error when comparing with original map and, therefore, they have the potential to provide improved support for engineering decisions.

5. Acknowledgments

Both authors are grateful to Nazarbayev University for funding this work via Faculty Development Competitive Research Grants for 2021–2023 under Contract No. 021220FD4951.

References

Cressie, N., & Chan, N. H. (1989). Spatial modeling of regional variables. Journal of the American Statistical Association, 84(406), 393-401.

Deutsch, C. V. (2006). A sequential indicator simulation program for categorical variables with point and block data: BlockSIS. Computers & Geosciences, 32(10), 1669-1681.

Emery, X. (2007). Simulation of geological domains using the plurigaussian model: new developments and computer programs. Computers & geosciences, 33(9), 1189-1201.

Goovaerts, P. (1994). Comparative performance of indicator algorithms for modeling conditional probability distribution functions. Mathematical Geology, 26(3), 389-411.

Gómez-Hernández, J. J., & Srivastava, R. M. (1990). ISIM3D: An ANSI-C three-dimensional multiple indicator conditional simulation program. Computers & Geosciences, 16(4), 395-440.

Journel, A. G. (1983). Nonparametric estimation of spatial distributions. Journal of the International Association for Mathematical Geology, 15(3), 445-468.

Journel, A. G., & Isaaks, E. H. (1984). Conditional indicator simulation: application to a Saskatchewan uranium deposit. Journal of the International Association for Mathematical Geology, 16(7), 685-718.

Kupfersberger, H., Deutsch, C. V., & Journel, A. G. (1998). Deriving constraints on small-scale variograms due to variograms of large-scale data. Mathematical geology, 30(7), 837-852.

Madani N. (2021) Plurigaussian Simulations. In: Daya Sagar B., Cheng Q., McKinley J., Agterberg F. (eds) Encyclopedia of Mathematical Geosciences. Encyclopedia of Earth Sciences Series. Springer, Cham. https://doi.org/10.1007/978-3-030-26050-7_251-1.

Remy, É., Ruet, P., & Thieffry, D. (2008). Graphic requirements for multistability and attractive cycles in a Boolean dynamical framework. Advances in Applied Mathematics, 41(3), 335-350.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). Using multivariate statistics (Vol. 5, pp. 481-498). Boston, MA: pearson.

AN OBJECT-ORIENTED APPROACH TO THE ANALYSIS OF SPATIAL FUNCTIONAL DATA OVER STREAM-NETWORK DOMAINS

Chiara Barbi(1) - Alessandra Menafoglio (1)* - Piercesare Secchi (1)

Politecnico di Milano, MOX, Department of Mathematics, Milan, Italy (1)
** Corresponding author: alessandra.menafoglio@polimi.it*

Abstract

The large availability of complex spatial data (curves, images, distributions) has posed a new and revolutionary challenge to spatial statistics. Object Oriented Spatial Statistics (O2S2) is a system of ideas which takes on the challenge, founding on a geometrical viewpoint to the data analysis. The foundational idea behind O2S2 is that the data object is considered as an indivisible unit (a.k.a. the *atom* of the analysis) rather than a collection of features, and it is analyzed via a mathematical embedding into an appropriate space (e.g., a Hilbert space or a Riemannian manifold, depending on the characteristics of the data).

We here focus on the problem of spatial prediction for functional data, when their embedding space can be assumed to be a Hilbert space, and their spatial domain of observation is a river network. Here, the peculiar reticular nature of the domain requires to use geostatistical methods based on the concept of *stream distance*, which captures the spatial connectivity of the points in the river induced by the network branching.

Extending to Hilbert data the pioneering ideas of Ver Hoef and Petersen (2010), we develop a class of functional moving average models based on the stream distance, allowing one to account for both the geometry of the data (as determined by their embedding space) and that of the spatial domain (as from the binary tree representing the stream network). Within this broad class of models, we shall focus on the so-called *purely tail-up* and *purely tail-down* models, and give a consistent definition of their covariance structure. The associated estimators, allowing one to assess the spatial structure of the data, will also be illustrated. This will eventually enable us to discuss on kriging prediction methods over stream networks, for both stationary and non-stationary object data.

We will illustrate our proposal on water temperature profiles in the Middle Fork River, USA.

IMPLICIT GEOLOGICAL DOMAIN MODELING BY USING GAUSSIAN PROCESSES ALGORITHM

Talgatbek Bazarbekov (1)* - Nasser Madani (1) - Timur Merembayev (2)

Nazarbayev University, School of Mining and Geosciences, Nur-Sultan, Kazakhstan (1) - International Information Technology University, Almaty, Kazakhstan (2)
** Corresponding author: talgatbek.bazarbekov@nu.edu.kz*

Abstract

With the increase of computational power in the digital electronic machines, more and more mathematical computations, in particular machine learning (ML) algorithms, rapidly are replacing the traditional routine time-consuming jobs. This context has also impacted the process of geological modeling either. In ore body evaluation, the classical framework is first orebody modeling by, for instance, wireframing the geological domains and then carrying out the resource estimation. Currently, most of the industrial practitioners are using explicit modeling for the first step -geological modeling, which is very time consuming and requires trained specialists. Even though they find ones, the result of the model is substantially subjective to the opinion of the geologist.

In contrary to the explicit geological modeling, implicit modeling uses mathematical algorithms and computational power to construct automatically the 2D/3D geological domains using drillhole data. In recent years, several interpolation algorithms have been proposed to improve the implicit modeling techniques. For implicit modeling, there is not much man-machine interaction needed, and due to high automation and quick partial update, it is being more and more popular in the field of geological modeling of an orebody.

There are plenty of implicit modeling techniques coming from mathematical science. For instance, in traditional geostatistics, indicator kriging (IK) is frequently used as a trivial implicit approach for geological domain modeling. However, IK suffers from some problems such as order relation problem, negative weights in the kriging system, and support effect, which makes the algorithm computationally intensive and sometimes untrustworthy for large number of blocks and complex geological domains.

In this paper, we discuss an alternative and powerful tool, widely used in ML, the Gaussian processes algorithm, which can solve the problem of IK in implicit modeling of geological domains. We treated the domain modeling as the classification problem in classical ML. The main hyperparameter to tune here is the kernel used for the interpolation. This arises from the fact that, by changing only the kernel function, one can obtain different results. In this paper, we compared the results of nearest neighbor method and indicator kriging with GP, where different kernels are used. In both cases, squared exponential kernel, a type of radial basis functions (RBF), was applied. The comparison of the algorithms was performed on synthetic dataset and the results showed that the performance of the traditional approach and GP is alike. We can also impose the anisotropy in the kernel function derived from the variogram analysis.

The procedure of modeling with GP has more advantages compared to traditional IK approach, although having almost the same result as an output. Among the advantages is that computation of GP is much faster than of IK and the process itself also requires less human interaction. The problem of order relation deviations and negative weights are not actually obstacles in the GP algorithm.

Keywords: Machine learning; Indicator kriging, Domain modeling; Gaussian processes; Implicit modeling.

1. Introduction

The main goal of the geological modeling is to identify the geological features, i.e. lithology, mineralization, geochemistry etc., to understand and visualize the picture of the sub-soil in a better way. Geologists draw cross-sections upon their knowledge and expertise and construct 3D models. Sometimes, the geological models are quite complex, and in the process of constructing such models, geologists try to decrease the level of complexity by reducing the number of geo-domains or connecting the geo-domains with the straight lines. To solve this issue, geoscientists try to adapt the mathematical algorithms to automate the process, and here comes the implicit geological modeling.

Implicit geological modeling is two-fold, scientific community is searching new approaches to alleviate the process of constructing the geological model using very quick process with a click of a button, whereas geologists tend to use their expertise where they can construct the model they think it is more compatible with the geological setting. Nevertheless, this work is dedicated to one of the numerical methods from machine learning, which in turn can be applied to solve the problem of geological domain modeling.

Implicit modeling uses mathematical algorithms and computational power to automatically construct 3D volumes using drillhole data. A lot of interpolation algorithms were tested and being tested to improve implicit modeling. For implicit modeling, there is not much man-machine interaction needed, and due to high automation and quick partial update, it is being more and more popular in the field of three-dimensional modeling of orebodies. Among others, the interpolation methods applicable for such implicit geological modeling can be: triangulation with linear interpolation (Dyn, Levin, and Rippa 1990), nearest neighbor (Olivier and Hanqiang 2012), inverse distance weights (Lu and Wong 2008), linear interpolation (Powell 1994), local polynomial (Baker and Pixley 1975), radial basis functions (RBF) interpolation (Morse et al. 2001), spline interpolation (R. and de Boor 1980), indicator Kriging, discrete smooth interpolation (Frank, Tertois, and Mallet 2007), moving least squares (MLS) (Fleishman, Cohen-Or, and Silva 2005), support vector machine (SVM) (Smirnoff, Boisvert, and Paradis 2006; 2008), potential field method with Gaussian Processes (Gonçalves, Kumaira, and Guadagnin 2017). The most used method nowadays in some commercial geological modeling software programs (Micromine, Leapfrog, Datamine, Gocad) is RBF-based interpolator (Wang et al. 2018; Zhong et al. 2019).

Among these algorithms, the nearest neighbors and indicator Kriging are one of the simplest, classical, and therefore the most used in the industry. In this paper, we will focus on the machine learning algorithm called Gaussian Processes which we will compare to the results of classical approaches.

2. Methodology

2.1. Indicator Kriging

Indicator Kriging is a classical geostatistical approach for modeling spatial data. There are different applications of indicator Kriging in the modeling, it can be applied to categorical variables as well as to continuous variables. In this paper, we will focus on the categorical variables modeling with two categories. The routine of indicator Kriging begins with processing of the dataset. We have to assign the indicators to each category as shown in the Table 1.

After we assign the indicators, we compute indicator semi-variograms for each category and define variogram models. From the variogram model we calculate the weights λ_{OK} and apply ordinary Kriging for each category. After that for each block we choose the maximum probability category.

Kriging is the best linear unbiased estimator (BLUE), which means it is exact interpolation method, but it is computationally intensive, because for each location, we have to compute large system of equations.

Table 1 – Data processing before indicator Kriging.

u Sample location	*Z(u)* Category (lithology)	*i(u, z_{clay})* Indicators (Clay)	*i(u, z_{sand})* Indicators (Sand)
u_1	Clay	1	0
u_2	Sand	0	1
u_3	Sand	0	1
…			

2.2. Machine learning approach

Machine learning algorithms are becoming more and more popular in all fields. There are several approaches in machine learning, based on the problem set one is aiming to solve, such as supervised learning and unsupervised learning, clustering, classification, regression.

For the case of geological domaining, the most appropriate approach is supervised learning binary classification algorithms. The advantage of the machine learning algorithms is that it can take several input parameters. The input parameters for the proposed model (Gaussian process), we chose (x, y) coordinates. For binary classification problem we don't apply the indicator approach, we just convert the classes into numerical values.

Table 2 – Data processing before binary classification algorithm.

u Sample location	*Z(u)* Category (lithology)	*i(u, z_{clay})* Indicators (Clay)
u_1	Clay	1
u_2	Sand	2
u_3	Sand	2
…		

Another significant difference between Machine learning algorithm and traditional geostatistics like Kriging is the application of weights. In machine learning, the weight are applied to input values, whereas in Kriging, the weights are applied to the known output values as shown in the Figure 1.

2.3. Gaussian Processes

Gaussian Processes algorithm belongs to the machine learning family. It is a generalization of the Gaussian probability distribution. Gaussian probability distribution functions summarize the distribution of random variables, whereas Gaussian processes summarize the properties of the functions, e.g. the parameters of the functions. As such, one may think of Gaussian processes as one level of abstraction or indirection above Gaussian functions. (Rasmussen and Williams 2018)

Similar to Kriging, one of the crucial parameters for Gaussian Processes is the covariance function, for which it is used as a kernel function for prediction. Covariance function gives the algorithm, the assumption of the prediction model, and depending on the chosen kernel function, one can obtain different models using the same input data.

For domain modeling as a kernel, we decided to use a radial basis function – squared exponential kernel.

$$k\left(x_i, x_j\right) = \exp\,(-\frac{d\left(x_i, x_j\right)^2}{2l^2})$$

where: l – length scale of the kernel; $d(x,y)$ – Euclidean distance to conditioning data.

So, basically our model is controlled by the length scale of the kernel and a kernel multiplier, to sharpen the change between two classes.

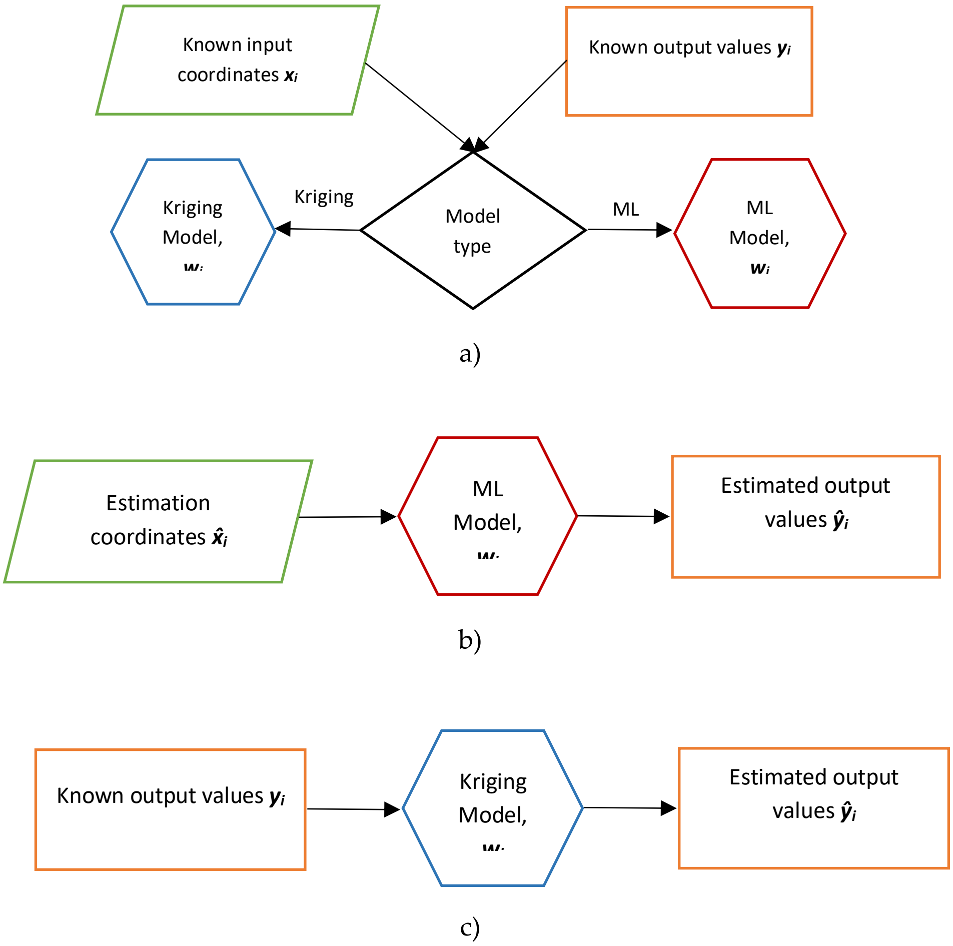

Known input coordinates x_i
Known output values y_i
Kriging Model, w_i
Kriging
Model type
ML
ML Model, w_i
a)
Estimation coordinates $\hat{x}_i$
ML Model, w_i
Estimated output values $\hat{y}_i$
b)
Known output values y_i
Kriging Model, w_i
Estimated output values $\hat{y}_i$
c)

Figure 1 – Machine learning approach vs kriging approach: a) model construction to obtain weights w_i; b) estimation with ML algorithm; c) estimation with kriging algorithm.

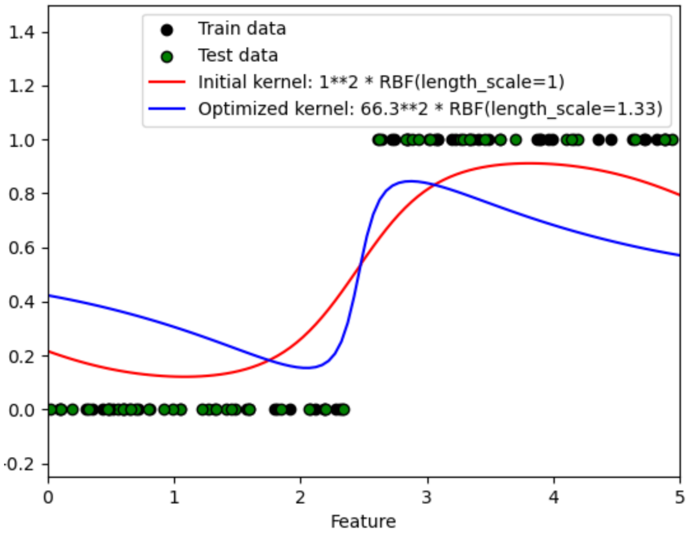


Figure 2 – The effect of kernel multiplier (*https://scikit-learn.org/*).

2.4. Computer program

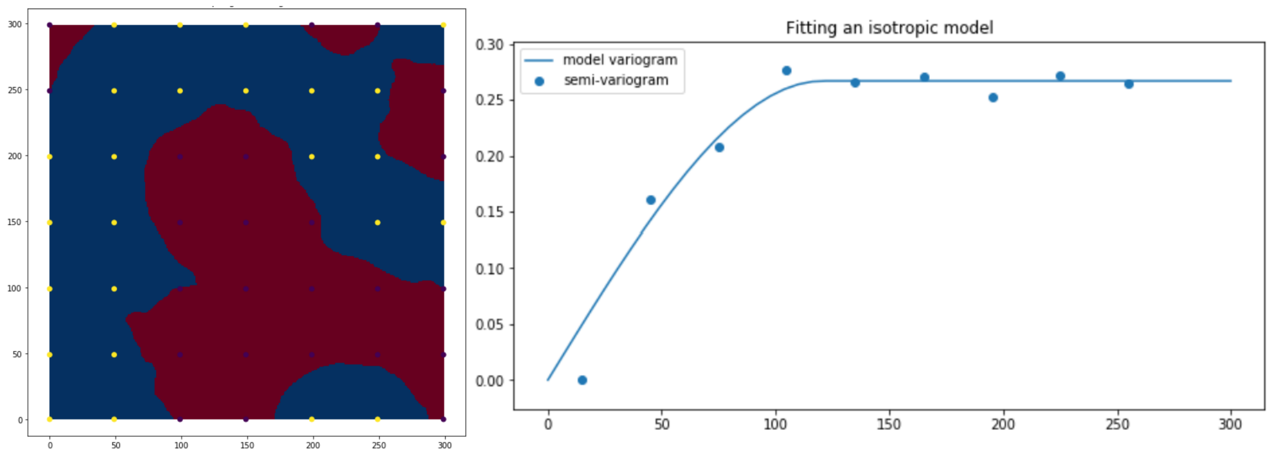
A computer program for indicator Kriging, Gaussian Processes was written in Python. For Kriging part, we used *gstools* and *pykrige* libraries from *geostat-framework* (https://geostat-framework.org/). For Gaussian Processes part, we used *scikit-learn* package (https://scikit-learn.org/).

3. Results

In this work, we compared the Gaussian Processes algorithm with classical geostatistical approach on the synthetic dataset. We have two reference maps of a size of 300x300m, generated from plurisim software, one isotropic and the other anisotropic.

3.1. Isotropic model

We generated a simple reference map of two categories with isotropic distribution (Figure 2a). Then we regularly sampled our map with 50x50m grid mesh and computed indicator variogram (Figure 2b).



a) b)

Figure 3 – Reference map with regular sampling.

The model for variogram we obtained was Spherical (Sill=0.267, Range=121, Nugget=0).

The kernel parameter: length scale = 50, kernel multiplier = 200

The decision about the length scale of the kernel depends on several factors. Changing the length scale will give us different models. So, in this work, to obtain a model similar to indicator Kriging, our recommendation is to take around ½ of the Range obtained from the variogram or around the distance between samples.

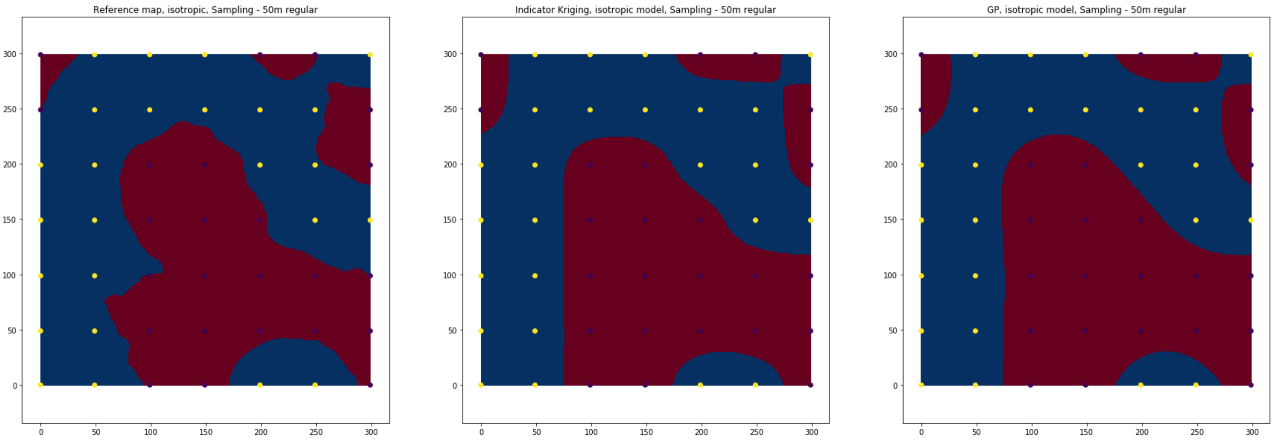

Reference map, isotropic, Sampling - 50m regular
Indicator Kriging, isotropic model, Sampling - 50m regular
GP, isotropic model, Sampling - 50m regular
0
50
100
150
200
250
300

Figure 4 – Comparing the obtained results for IK and GP.

3.2. Anisotropic model

To test the performance of the Gaussian process algorithm on anisotropic dataset, we generated an anisotropic map with an anisotropy with maximum continuity along 45 degrees. As in previous case, firstly we run the traditional approach with indicator kriging. In this case we took finer mesh for sampling, 25x25m and tried to run the algorithm. For running the indicator kriging, two cases are considered with isotropic and anisotropic variograms. In Figure 5, we observe our reference map (left) and comparing with the obtained models by indicator Kriging with isotropic variogram model (middle) and anisotropic variogram model (right).

The indicator variogram parameters were the following:

Isotropic variogram: Spherical (Sill=0.26, Range=70, Nugget=0);

Anisotropic variogram: Spherical (Sill=0.26, Range=[90,45], Nugget=0, Angle=45deg).

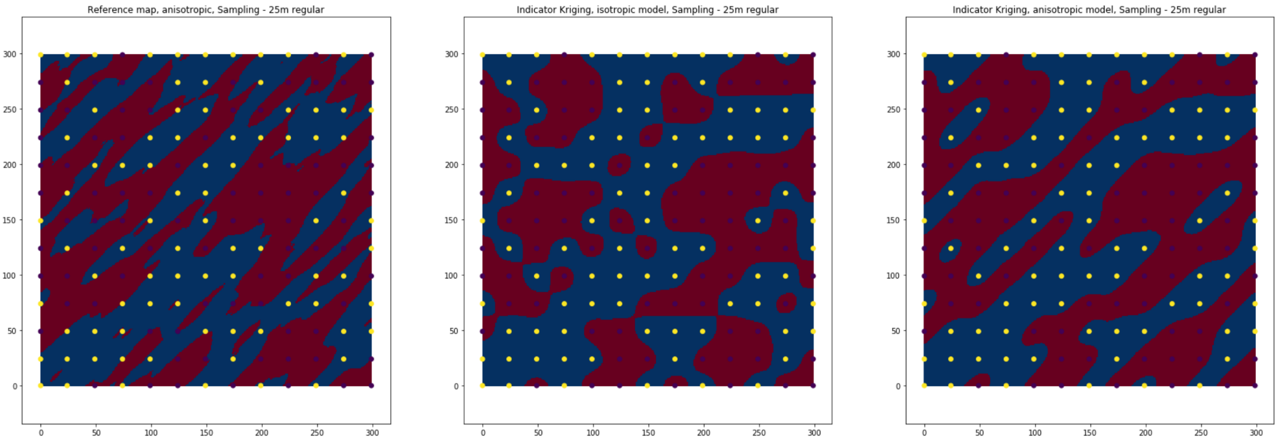

Reference map, anisotropic, Sampling - 25m regular
Indicator Kriging, isotropic model, Sampling - 25m regular
Indicator Kriging, anisotropic model, Sampling - 25m regular
0
50
100
150
200
250
300

Figure 5 – Comparing the results from isotropic IK and anisotropic IK with the reference map.

From Figure 5 we observe, that anisotropic model looks better that isotropic model, because it does not regard the directional variogram.

Now let's compare the results obtained from Gaussian Processes algorithm. As hyperparameters, we again apply only the isotropic kernel length scale and kernel multiplier. The length scale = 25, the kernel multiplier = 200.

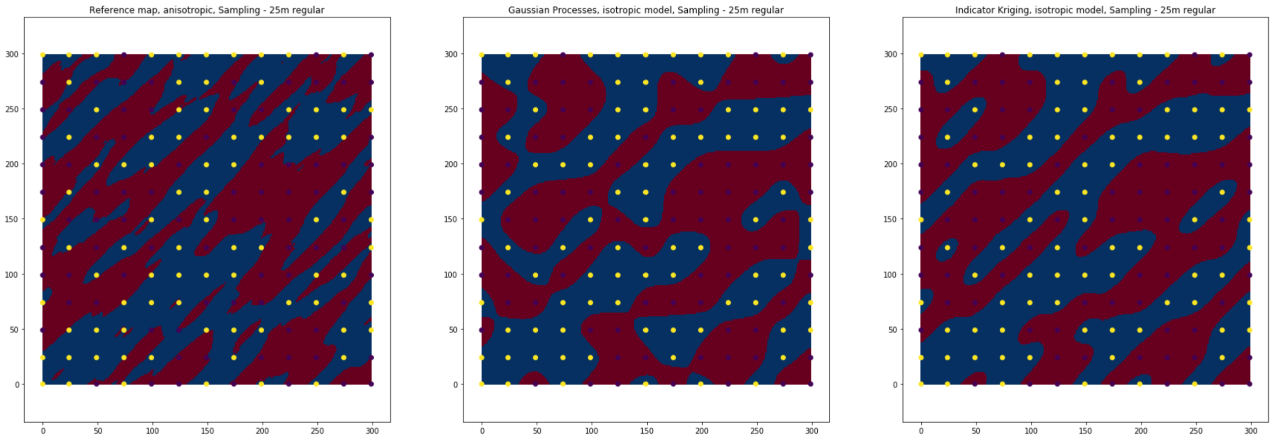

Reference map, anisotropic, Sampling - 25m regular
Gaussian Processes, isotropic model, Sampling - 25m regular
Indicator Kriging, isotropic model, Sampling - 25m regular
0
50
100
150
200
250
300

Figure 6 – Comparing the results from Gaussian Processes and anisotropic IK with the reference map.

In Figure 6, the results obtained with isotropic kernel for GP shows that in some regions, it found the direction of distribution and is obviously showed better performance than isotropic IK, but a bit lower performance than anisotropic IK.

3.3. The computational performance of the algorithms.

In this work we used small model with size 300x300m with block size 1x1m, so we had 90000 blocks. The computer program was written in Python. The time for computing anisotropic model for each algorithm is shown in the table below.

Algorithm	Time, ms
Indicator Kriging	1420
Gaussian Processes	360

4. Discussion and Conclusions

The aim of the paper is not to show that the performance of the Gaussian Processes algorithm is much better than of Indicator Kriging. One of the main ideas is to demonstrate an alternative method for implicit numerical geological domain modeling. This paper showed only the application in 2D case, but it can be easily converted to 3D volume construction. Saying about the advantages of the proposed method is that its performance is much faster than of kriging, in our case it showed 3 times faster computation, this might be even larger for big models. One of the main reasons is demonstrated in Figure 1, the working principle of ML approaches. Another advantage is that there are not much hyperparameters to tune, and the performance is very similar to what the IK gives.

But of course, there is some work to do, we need to be able to specify the directional anisotropy, if we clearly observe it from a directional variogram. In this work, we showed the performance of the squared exponential kernel, for further research other kernels might be tested, and depending on the variogram analysis Gaussian processes algorithm can also be applied for estimation of continuous variables.

References

Baker, Kirby A., and Alden F. Pixley. 1975. "Polynomial Interpolation and the Chinese Remainder Theorem for Algebraic Systems." Mathematische Zeitschrift 143 (2). https://doi.org/10.1007/BF01187059.

Dyn, Nira, David Levin, and Samuel Rippa. 1990. "Data Dependent Triangulations for Piecewise Linear Interpolation." IMA Journal of Numerical Analysis 10 (1). https://doi.org/10.1093/imanum/10.1.137.

Fleishman, Shachar, Daniel Cohen-Or, and Cláudio T. Silva. 2005. "Robust Moving Least-Squares Fitting with Sharp Features." In ACM Transactions on Graphics. Vol. 24. https://doi.org/10.1145/1073204.1073227.

Frank, Tobias, Anne Laure Tertois, and Jean Laurent Mallet. 2007. "3D-Reconstruction of Complex Geological Interfaces from Irregularly Distributed and Noisy Point Data." Computers and Geosciences 33 (7). https://doi.org/10.1016/j.cageo.2006.11.014.

Gonçalves, Ítalo Gomes, Sissa Kumaira, and Felipe Guadagnin. 2017. "A Machine Learning Approach to the Potential-Field Method for Implicit Modeling of Geological Structures." Computers and Geosciences 103. https://doi.org/10.1016/j.cageo.2017.03.015.

Lu, George Y., and David W. Wong. 2008. "An Adaptive Inverse-Distance Weighting Spatial Interpolation Technique." Computers and Geosciences 34 (9). https://doi.org/10.1016/j.cageo.2007.07.010.

Morse, Bryan S., Terry S. Yoo, Penny Rheingans, David T. Chen, and K. R. Subramanian. 2001. "Interpolating Implicit Surfaces from Scattered Surface Data Using Compactly Supported Radial Basis Functions." In Proceedings - International Conference on Shape Modeling and Applications, SMI 2001, 89–98. https://doi.org/10.1109/SMA.2001.923379.

Olivier, Rukundo, and Cao Hanqiang. 2012. "Nearest Neighbor Value Interpolation." International Journal of Advanced Computer Science and Applications 3 (4). https://doi.org/10.14569/ijacsa.2012.030405.

Powell, M. J. D. 1994. "A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation." In Advances in Optimization and Numerical Analysis. https://doi.org/10.1007/978-94-015-8330-5_4.

R., J., and Carl de Boor. 1980. “A Practical Guide to Splines.” Mathematics of Computation 34 (149). https://doi.org/10.2307/2006241.

Rasmussen, Carl Edward, and Christopher K. I. Williams. 2018. Gaussian Processes for Machine Learning. Gaussian Processes for Machine Learning. https://doi.org/10.7551/mitpress/3206.001.0001.

Smirnoff, Alex, Eric Boisvert, and Serge J Paradis. 2006. “3D Geological Modeling: Solving as a Classification Problem with the Support Vector Machine”.

Smirnoff, Alex, Eric Boisvert, and Serge J. Paradis. 2008. “Support Vector Machine for 3D Modelling from Sparse Geological Information of Various Origins.” Computers and Geosciences 34 (2): 127–43. https://doi.org/10.1016/j.cageo.2006.12.008.

Wang, Jinmiao, Hui Zhao, Lin Bi, and Liguan Wang. 2018. “Implicit 3D Modeling of Ore Body from Geological Boreholes Data Using Hermite Radial Basis Functions.” Minerals 8 (10). https://doi.org/10.3390/min8100443.

Zhong, De yun, Li guan Wang, Lin Bi, and Ming tao Jia. 2019. “Implicit Modeling of Complex Orebody with Constraints of Geological Rules.” Transactions of Nonferrous Metals Society of China (English Edition) 29 (11). https://doi.org/10.1016/S1003-6326(19)65145-9.

TWO APPROACHES TO SCENARIO MODELLING USING THE EMBEDDED ESTIMATOR METHOD

Colin Daly (1)*

Schlumberger, Modelling, Abingdon, United Kingdom (1)
** Corresponding author: cdaly@slb.com*

Abstract

A recently proposed Conditional Random Field method of producing geostatistical models of one or more target variables with many covariables of mixed discrete/continuous type works in two steps. Firstly, it finds approximate conditional distributions at the model target locations using a relatively low dimensional machine learning regression leveraging embedding spatial estimators, such as kriging. The idea of embedding simpler models gives the method its name, Ember. Its advantage is that no explicit random function is required to produce the approximate conditional distribution. In a second stage, when stochastic simulations are required, the residual variation is sampled from the previously estimated set of conditional distributions using a uniformly distributed random field. By construction the marginal distribution of the target variable, bivariate marginals and even higher order marginals of the target with any of covariates are reproduced in the simulated model. In the applications considered so far this has been simulated via a Gaussian Random field (i.e a Gaussian copula is used). However, sometimes this simulation approach can be too restrictive and a way of relaxing this is required.

Many applications of Geostatistics require an ensemble of stochastically generated models. One example, amongst many, would be to understand the uncertainty in the flow of fluids in an aquifer. Geoscientists will often want to do more than just sample realizations from a fixed parameter stochastic model. It may be necessary to vary parameters such as the global mean or to create scenarios based on geological hypothesis which are consistent with, but not determined by the data. For example, a geologist might draw one or more possible maps of facies fairway based on domain knowledge. This type of scenario variable may only have a weak linear correlation with the target data but have a textural relevance difficult to capture with limited data. Scenario modelling is about ensuring that these possibilities are considered. With most current forms of geological modelling, it can be difficult to produce scenarios and yet continue to honour the relevant marginals.

This presentation, after a quick introduction to the Ember method, looks at ways that it can be extended to produce scenarios. One straightforward method is to modify the envelope of approximate conditional distributions to incorporate the specific scenario. This approach changes the marginal distributions although the differences are often small enough to remain acceptable. A second method is to modify the sampling strategy. It is necessary to ensure that the sampling random variable remains uniform if the marginals are to be conserved. One method which achieves that is considered here by modelling the sampling RF as a Substitution RF with the scenario variable as directing function, allowing the texture of the scenario to be caught.

ANALYSIS OF TWO PRECIPITATION GAP-FILLING METHODS IN A STUDY AREA OF NORTHERN ITALY

Camilla Fagandini (1)* - Valeria Todaro (1) - Maria Giovanna Tanda (1) - Andrea Zanini (1)

University of Parma, Department of Engineering and Architecture, Parma, Italy (1)
** Corresponding author: camilla.fagandini@unipr.it*

Abstract

Missing data are a frequent problem in meteorological and hydrological observation datasets. They are caused by many circumstances, e.g., sensor malfunction, errors in measurements, faults in data acquisition from the operators, etc. Finding efficient methods to deal with this problem is an important issue because it is necessary to have complete time series to carry out reliable hydrological analyses. There are numerous gap-filling procedures, usually specific to the nature of the variable under study. In this research, two approaches – FAO-based linear regression and Kriging interpolation – were compared in terms of their ability to fill missing data. In addition, some general criteria proposed by the World Meteorological Organization (WMO) were considered. The FAO procedure fills the gaps using data collected at the gap time in other stations. Aiming at filling the missing data of a certain monitoring station, the approach proceeds as follows: first, the correlation coefficients with the other existing stations, having an appropriate common recording period, are computed; then, the station with the highest correlation is considered to estimate the parameters of a regression equation, which is used to obtain the missing value of the interested station. Instead of considering time series at a specific location, the Kriging approach analyses the spatial distribution of hydrometeorological data at a certain time. As a result, to fill each gap at a monitoring station, the approach requires recognition of the stations with available data, computation of the variogram and interpolation of the missing data at a specific location through the Kriging approach. For testing purposes, complete-time series were selected from several rain gauges in a large area of Northern Italy. To evaluate the efficacy of the approaches, few data were removed from one station. Then, the two methods were used to estimate the precipitation in the missing periods. The FAO method is suggested in case of a small number of gaps and requires little computational effort. Whereas Kriging can manage more intensive processes but it involves the use of a large number of monitoring stations. The pros and cons of the two gap-filling approaches were discussed by measuring the goodness of integration with shared metrics.

A NEW APPROACH TO SIMULATE CONDITIONAL RANDOM FIELDS WITH CORRELATION TO AN EXTERNAL VARIABLE USING FFTMA AND P-FIELD SIMULATION

Sebastian Hörning (1)* - András Bárdossy (2)

The University of Queensland, Centre for Natural Gas, Brisbane, Australia (1) – University of Stuttgart, Institute for Modelling Hydraulic and Environmental Systems, Dept. of Hydrology and Geohydrology, Stuttgart, Germany (2)
** Corresponding author: s.hoerning@uq.edu.au*

Abstract

The simulation of conditional spatial random fields with correlation to an external variable is omnipresent in environmental sciences. In hydrology, for example, soil water content exhibits a relationship with precipitation while precipitation itself is often correlated to the topography. Considering these relationships can help to improve the estimation of the variable of interest at unsampled locations. Traditionally, external drift Kriging is used to incorporate a linear relationship to an external variable. This method however is mainly used for estimation often leads to values outside the admissible range of the variables.

We propose a new approach to simulate conditional random fields with correlation to an external variable using FFTMA and p-field simulation. Using p-field simulation combined with FFTMA, one can simulate conditional realizations of the variable of interest. The p-field represents the conditional cumulative distribution functions (ccdf) at each sampled and unsampled location. These ccdfs can for example be determined using Kriging or copula interpolation. FFTMA can then be used to simulate an unconditional spatial random field with a given spatial covariance which is estimated from the variable of interest. This field is subsequently transformed into a conditional field via the p-field’s ccdfs. In order to introduce the correlation to an external variable, one can use inverse FFTMA. Assuming that the external variable represents a random field, one can determine the underlying Gaussian random numbers by inverting the FFTMA equations. These Gaussian random numbers can then be used to generate correlated Gaussian random numbers for the simulation of the unconditional field using FFTMA, where the correlation is obtained from the relationship between the variable of interest and the external variable. Thus, one obtains an unconditional spatial random field which exhibits the desired correlation. Using the p-field, this unconditional field can then be transformed into a conditional field with correlation to the external variable.

We demonstrate this new approach using rainfall data from Baden-Wuerttemberg. Rainfall in BW exhibits significant correlation with the state’s topography. A digital elevation model will be used to determine this correlation and will serve as external variable for the simulations.

COMBINED ANALYSIS OF SPATIALLY MISALIGNED DATA USING GAUSSIAN FIELDS AND THE STOCHASTIC PARTIAL DIFFERENTIAL EQUATION APPROACH.

Paula Moraga (1)*

King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia (1)
** Corresponding author: paula.moraga@kaust.edu.sa*

Abstract

Spatially misaligned data are becoming increasingly common due to advances in both data collection and management in a wide range of scientific disciplines including the epidemiological, ecological and environmental fields. Here, we present a Bayesian geostatistical model for fusion of data obtained at point and areal resolutions. The model assumes that underlying all observations there is a spatially continuous variable that can be modeled using a Gaussian random field process. The model is fitted using the integrated nested Laplace approximation (INLA) and the stochastic partial differential equation (SPDE) approaches. In the SPDE approach, a continuously indexed Gaussian random field is represented as a discretely indexed Gaussian Markov random field (GMRF) by means of a finite basis function defined on a triangulation of the region of study. In order to allow the combination of point and areal data, a new projection matrix for mapping the GMRF from the observation locations to the triangulation nodes is proposed which takes into account the types of data to be combined. The performance of the model is examined via simulation when it is fitted to (i) point, (ii) areal, and (iii) point and areal data combined to predict several simulated surfaces that can appear in real settings. The model is also applied to predict the concentration of fine particulate matter (PM2.5) in Los Angeles and Ventura counties, USA. The results show that the combination of point and areal data provides better predictions than if the method is applied to just one type of data, and this is consistent over both simulated and real data. We conclude the approach presented may be a helpful advance in the area of spatial statistics by providing a useful tool that is applicable in a wide range of situations where information at different spatial resolutions needs to be combined.

BIVARIATE DEEPKRIGING FOR LARGE-SCALE SPATIAL INTERPOLATION OF WIND FIELD

Pratik Nag (1)* - Ying Sun (1) - Brian Reich (2)

Environmental Statistics Department, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia (1) - Department of Statistics, North Carolina State University, Raleigh, United States (2)

** Corresponding author: pratik.nag@kaust.edu.sa*

Abstract

High spatial resolution wind data are essential for a wide range of applications in climate, oceanographic and meteorological studies. Large-scale spatial interpolation or downscaling of bivariate wind fields is a challenging task because wind data tend to be non-Gaussian with high spatial variability and heterogeneity. In spatial statistics, cokriging is commonly used for predicting bivariate spatial fields. However, the cokriging predictor is not optimal except for Gaussian processes. Additionally, cokriging is computationally prohibitive for large datasets. In this paper, we propose a method, called bivariate DeepKriging, which is a spatially dependent deep neural network (DNN) with an embedding layer constructed by spatial Radial basis functions for bivariate spatial data prediction. We then develop a distribution-free uncertainty quantification method based on bootstrap and ensemble DNN. Our proposed approach outperforms the traditional cokriging predictor with commonly used covariance functions, such as the linear model of co-regionalization and flexible bivariate Mat\'ern covariance. We show that the proposed DNN model is computationally efficient and scalable, with twenty times faster computations on average. We apply the bivariate DeepKriging method to the wind data over the Middle East region at 506771 locations. The prediction performance of the proposed method is superior over the cokriging predictors and dramatically reduces the time of computation and the large-scale computational complexity.

GEOSTATISTICS FOR COMPOSITIONAL DATA: COMPARISON OF LOGARITHMIC TRANSFORMATION METHODS IN AN IRON DEPOSIT

Milena Nasretdinova (1)* - Nasser Madani (1)

School of Mining and Geosciences, Nazarbayev University, Nur-Sultan City, Kazakhstan (1)
** Corresponding author: milena.nasretdinova@nu.edu.kz*

Abstract

The modeling of mineral resources is one of the important stages in a mining project. Based on regionalized variables corresponding to the type of deposit, the initial data can be evaluated and modeled to produce some informative maps that can be obtained based on either geostatistical estimation or simulation techniques. Regarding this, compositional variable, e.g., grade of ore, the mineralogical composition of rocks, etc. often is used in geostatistical modeling. The main issue for this type of data is closure problem, which lead to spurious linear cross-correlation between the pair of such variables, impacting the mineral resource evaluation of the deposit. Therefore, the correlation of compositional data cannot represent a true linear relationship between the components, and the generally accepted interpretation of the deduced correlation may be interpreted incorrectly. In this case, using standard multivariate geostatistical methods for modeling compositional data without pre-processing step can lead to inconsistencies in the final results. This crucial step can be implemented using advanced data analysis techniques based on, for instance, log-ratio transformations. This methodology allows freeing compositional data from subordination to a constant sum, which in turn makes it possible to use standard geostatistical methods in real space for modeling of such complex variables. After this step, a post-processing step is needed to restitute the simulation results to the original scale of the data. The purpose of this research is to compare two methods of logarithmic ratio transformation to evaluate the mineral resources in order to find out the most reliable transformation technique. To do so, the closure problem of compositional nature of the data is solved by additive logarithmic ratio (alr) and centered logarithmic ratio (clr) transformations of the borehole data. The algorithm is illustrated over the Carajas Iron ore deposit in Brazil, where five geochemical components (Fe, Al2O3, Mn, P and SiO2) are required to be considered for mineral resource evaluation in this deposit. A filler variable is also introduced. This variable is a collection of undefined components that have not been analyzed through each sample. After pre-processing step, the transformed variables are subjected into the variogram analysis to infer the linear model of coregionalization. Then, using the direct and cross-variogram models, the transformed variables are co-simulated over the target grid cells entire the deposit. The realizations are then post-processed to back-transfer the simulation results to the original scale of the five geochemical elements. Furthermore, a mineral resource evaluation is taken into account to quantify the recovery functions (tonnage, mean grade, and metal quantities) for the whole deposit to compare the results obtained from both log-ratio transformations. A cross-validation technique is also performed. At the end, a proper discussion is provided so that one can evaluate the consistency of these two log-ratio transformation techniques for a proper mineral resource evaluation.

Keywords: Compositional data; Geostatistical analysis; Log-ratio transformation; Carajas Iron mine.

1. Introduction

Compositional data is one of the most common data types in geostatistical modeling (Pawlowsky-Glahn & Egozcue, 2016). Compositional data usually includes the grades of ore, mineralogical composition of rocks, concentrations of chemical elements in soils and rocks (Pawlowsky, 1984). By definition, compositional data (CoDa) is two or more variables, which together characterize the relative weight of each variable in relation to the whole (Pawlowsky-Glahn et al. 2015). In other words, CoDa consists of certain parts, each of which represents a fraction of the whole (Pawlowsky-Glahn & Olea, 2004). Since the compositional data is only positive numbers, and their sum is always 100%, changing one value within the sum entails changing others to maintain a constant amount. For instance, when determining the proportion of chemical elements, a specific element corresponds to its percentage of content, but when the proportion of an element changes (for example, from 15% to 20%), the sum of the remaining concentrations cannot be automatically more than 80%. In addition, correlations in this case might be spurious due to the limitation of a constant sum. Thus, the "constant sum" problem, which is also called as the "closure effect", can limit the application of standard statistical methods in relation to such data and eventually, leads to erroneous statistical results (Grunsky & Caritat, 2019). Thereby, for a relatively accurate evaluation of mineral resources and ore reserves, it is necessary to take into account the compositional nature of the data and implement advanced data analysis techniques based, for instance, on logarithmic ratio transformations basis. In this paper, different methods of log-ratio transformations are compared. In the following, a theoretical background is given mostly about such transformations and their geostatistical modeling, and then an Iron deposit case study is used to compare different approaches.

2. Material and Methods

2.1. Compositional data analysis (CoAn)

The concept of compositional data analysis (CoAn) historically refers to Aitchison method (Aitchison, 1986). This technique is based on log-ratio transformations, in which data is transformed using appropriate transformations that preserve the geometry of compositional data on the simplex and the support space of compositional data. In other words, this method includes pre-processing of data before their analysis and modeling, allowing to free compositional data from subordination to the closure effect that represents a positive constant *Δ* on the scale of the underlying random vector (Aitchison, 1986). According to Aitchison (1986) a vector of δ components Z(x) = {Z_1(x); Z_2(x); …; Z_δ(x)} is a composition of:

$$\forall Z_i(x) > 0 \qquad \sum_{i=1}^{\delta} Z_i(x) = \Delta \tag{1}$$

In this case, the closure problem of compositional nature of the data is solved by additive log-ratio (alr) and centered log-ratio (clr) transformations of the borehole data. These types of transformations, corresponding to compositional data, are briefly described below.

2.1.1. Additive log-ratio transformation (alr)

The additive log-ratio transformation converts original data into log-ratios as (Aitchison, 1986):

$$F(x) = (ln\frac{Z_1(x)}{Z_\delta(x)}; ln\frac{Z_2(x)}{Z_\delta(x)}; \dots; ln\frac{Z_{\delta-1}(x)}{Z_\delta(x)}) \tag{2}$$

where the numerator is the original composites Z(x), and one of the variables is chosen as the denominator. The distinctive features of the denominator are that, firstly, it must be strictly positive, and secondly, the same denominator must be applied to all variables. For this study, a filler variable is introduced, which is a collection of undefined components that have not been analyzed through each sample. In addition, for alr, the filler was chosen as the denominator $Z_\delta(x)$, since it facilitates the analysis by reducing the number of transformations by one compared to the initial number of variables, and the choice of this denominator does not affect the results of forward and back-transformations (Job, 2010). After obtaining the transformed variables, geostatistical algorithms can be applied over the log-ratio transformed regionalized variables (F(x)). But the result of modelling should be backward transformed into the compositional space (Pawlowsky-Glahn & Egozcue, 2016). For the back-transformation, the equation has the following form (Aitchison, 1986):

$$B(x) = (\frac{\exp(F_1)}{\sum_{i=1}^{\delta-1}\exp(F_i)+1}; \frac{\exp(F_2)}{\sum_{i=1}^{\delta-1}\exp(F_i)+1}; \dots; \frac{\exp(F_{\delta-1})}{\sum_{i=1}^{\delta-1}\exp(F_i)+1}) \times \Delta \qquad (3)$$

2.1.2. Centered log-ratio transformation (clr)

The centered log-ratio transformation is able to transfer the compositions to the real sample space for further application of statistical analysis methods (Aitchison, 1986):

$$F(x) = (ln\frac{Z_1(x)}{\sqrt[\delta]{Z_1(x)\times Z_2(x)\times\dots\times Z_\delta(x)}}; ln\frac{Z_2(x)}{\sqrt[\delta]{Z_1(x)\times Z_2(x)\times\dots\times Z_\delta(x)}}; \dots; ln\frac{Z_{\delta-1}(x)}{\sqrt[\delta]{Z_1(x)\times Z_2(x)\times\dots\times Z_\delta(x)}}) \qquad (4)$$

where the numerator are the composites, and the denominator is the geometric mean of those composites over the sample points. The main advantage of clr transformation over alr is that clr can display composition isometrically. The back-transformation clr is represented as (Aitchison, 1986):

$$B(x) = (\frac{\exp(F_1)}{\sum_{i=1}^{\delta-1}\exp(F_i)}; \frac{\exp(F_2)}{\sum_{i=1}^{\delta-1}\exp(F_i)}; \dots; \frac{\exp(F_{\delta-1})}{\sum_{i=1}^{\delta-1}\exp(F_i)}) \times \Delta \qquad (5)$$

After pre-processing step, the transformed variables are subjected into the normal score transformation and then into the variogram analysis to infer the model of coregionalization. Then, using direct and cross-variogram models, the transformed variables are simulated and co-simulated over the target grid cells entire the region using any Gaussian simulation approaches such as turning bands (co)-simulation that we used in this study.

2.2. Turning Bands (Co)-Simulation

The turning bands simulation (TBSIM) is designed to generate a realization from a normal distribution with zero mean and a specified covariance structure (Matheron, 1973). The concept of the TBSIM is to convert of a two- or three-dimensional problem to a series of one-dimensional problems, which implies the creation of a series of one-dimensional random processes along lines radiating from a coordinate origin and their subsequent projection and combination at arbitrary points in space, yielding discrete values or realizations of the field (Journel & Huijbregts, 1978).

Having the covariance model fitted to the primary de-clustered normal score variable, the covariance function is derived from one-dimensional random fields. TBSIM provides a non-conditional multi-dimensional random field compatible with the target covariance model, in which the simulated values are practically standard Gaussian (Emery & Lantuéjoul, 2006).

TBCOSIM is an extension of TBSIM, indicating an approximate algorithm based on the multi-Gaussian distribution assumption of the underlying random field as first introduced by Matheron (1973). The existence of cross-correlation among certain variables motivates one to use Gaussian co-simulation approaches rather than independent simulation (Wackernagel, 2003). The reason relates to considering the inter-dependency characteristic among certain variables, for which it leads to generate results that reproduce the local and global multivariate statistical parameters of data. In TBCOSIM, it is of interest to simulate stochastically the cross-correlated variables. In this regard, the cross-covariance function is needed to construct such one- and multi-dimensional Gaussian random fields in the region.

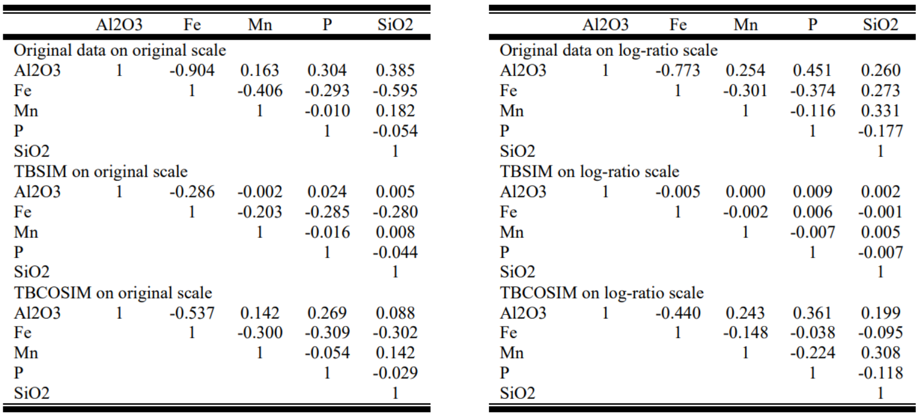
2.3. Case study

The proposed algorithm is based both on TBSIM and TBCOSIM using both simple and ordinary kriging/co-kriging methods, illustrated over the Carajas Iron ore deposit in Brazil, where five geochemical components (Fe, Al2O3, Mn, P and SiO2) are considered for mineral resource evaluation in this deposit. The main stages of the algorithm is briefly described below. The closure problem of the data is solved by additive log-ratio (alr) and centered log-ratio (clr) transformations of the borehole data after introducing a filler variable. The transformed variables are subjected into the variogram analysis to infer the linear model of coregionalization after normal score transformation of alr- and clr-transformed data. Using the direct and cross-variogram models, the transformed variables are simulated and co-simulated over the target grid cells entire the deposit (for this, simple and ordinary kriging/co-kriging are used). The realizations are post-processed to back-transfer the simulation results to the original scale of these five geochemical elements. A mineral resource evaluation is then considered to quantify the recovery functions (tonnage, mean grade, and metal quantities). To save the space, the output of Fe and Al2O3 are only presented to compare the simulation results obtained from both log-ratio transformations. A cross-validation technique is also performed.

3. Results

In order to validate the simulation results, Mueller et al. (2014) illustrated that it is necessary that the simulation results either reproduce the global statistical parameters in original scale or in log-ratio transformed values. For this purpose, the results of simulation and co-simulation using simple kriging/co-kriging for alr-transformed data were reflected in this paper because the statistical parameters calculated for the simulation and co-simulation using simple/ordinary kriging/co-kriging for clr-transformed data and for the simulation and co-simulation using ordinary kriging/co-kriging for alr-transformed data produced not very satisfying results, particularly in reproduction of original global distributions. The correlation coefficient matrix is provided in Table 1 for the simulation results in simplex space and original space. As can be seen, in both cases, the TBCOSIM is superior in reproduction of correlation coefficients in both original scale and log-ratio scales. This difference can be explained by the fact that compared to TBCOSIM, TBSIM does not consider the cross-dependency between variables, which leads to poor reproduction of the cross-correlation between the modeled variables.

Table 1 – Correlation coefficients on original scale and log-ratio scale of original data, TBSIM and TBCOSIM obtained from (alr) transformation.



	Al2O3	Fe	Mn	P	SiO2
Original data on original scale					
Al2O3	1	-0.904	0.163	0.304	0.385
Fe		1	-0.406	-0.293	-0.595
Mn			1	-0.010	0.182
P				1	-0.054
SiO2					1
TBSIM on original scale					
Al2O3	1	-0.286	-0.002	0.024	0.005
Fe		1	-0.203	-0.285	-0.280
Mn			1	-0.016	0.008
P				1	-0.044
SiO2					1
TBCOSIM on original scale					
Al2O3	1	-0.537	0.142	0.269	0.088
Fe		1	-0.300	-0.309	-0.302
Mn			1	-0.054	0.142
P				1	-0.029
SiO2					1

	Al2O3	Fe	Mn	P	SiO2
Original data on log-ratio scale					
Al2O3	1	-0.773	0.254	0.451	0.260
Fe		1	-0.301	-0.374	0.273
Mn			1	-0.116	0.331
P				1	-0.177
SiO2					1
TBSIM on log-ratio scale					
Al2O3	1	-0.005	0.000	0.009	0.002
Fe		1	-0.002	0.006	-0.001
Mn			1	-0.007	0.005
P				1	-0.007
SiO2					1
TBCOSIM on log-ratio scale					
Al2O3	1	-0.440	0.243	0.361	0.199
Fe		1	-0.148	-0.038	-0.095
Mn			1	-0.224	0.308
P				1	-0.118
SiO2					1

Next step is to quantify the reproduction of original distribution. For this, QQ-plots are drawn in Figure 1. According to the QQ-plots, co-simulation with simple co-kriging gives better results for both Fe and Al2O3 compared to simulation method with simple kriging, and this significantly is better for Al2O3 (Fig. 1). The clr-transformed simulation results are all failed to reproduce the original distributions.

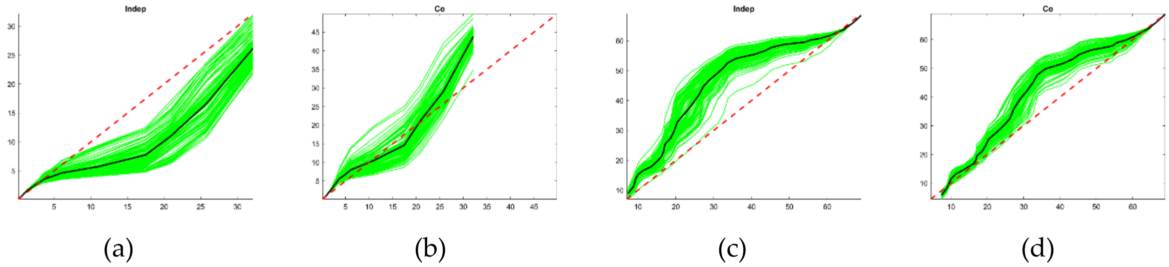


Figure 1 – QQ-plots of Al2O3 for (a) TBSIM and (b) TBCOSIM and of Fe for (c) TBSIM and (d) TBCOSIM obtained from (alr) transformation. Green line: individual realizations; black: average of distributions.

In order to identify the dependence relationship regardless of measure of correlation coefficient for comparison of two algorithms, the scatter plots between pairs of Al2O3 and Fe for simulation and co-simulation are depicted (Fig. 2). As can be seen, co-simulation with simple co-kriging over alr-transformed data gives better results compared to simulation method. In TBCOSIM, the reproduction of bivariate relation between the variables, compared to TBSIM, demonstrates that, not only the reproduction of bivariate relations is improved, but also the bivariate relation is roughly in agreement with the original data.

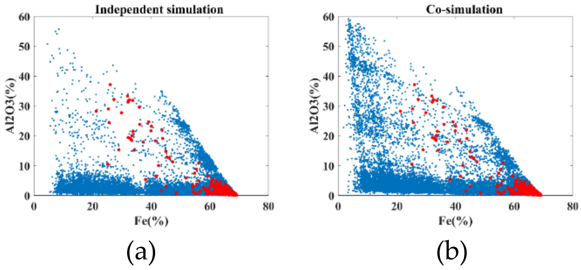
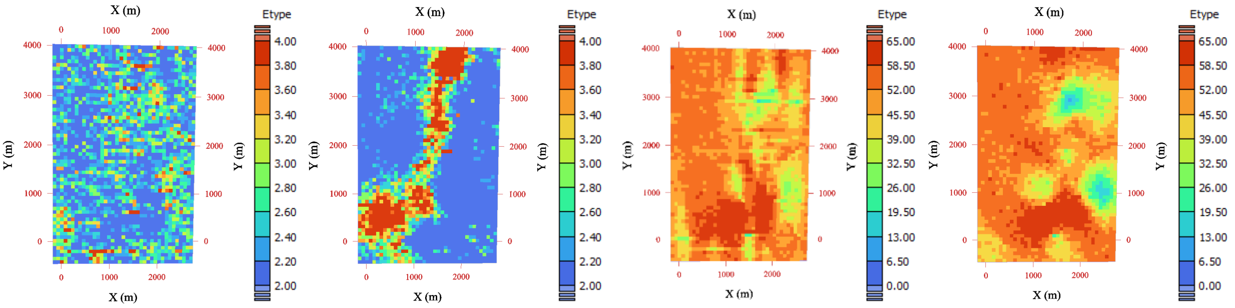


Figure 2 – Scatter plots of Al2O3 and of Fe for (a) TBSIM and (b) TBCOSIM for realization No.1 obtained from (alr) transformation. Blue points: simulation results; red points: original data.

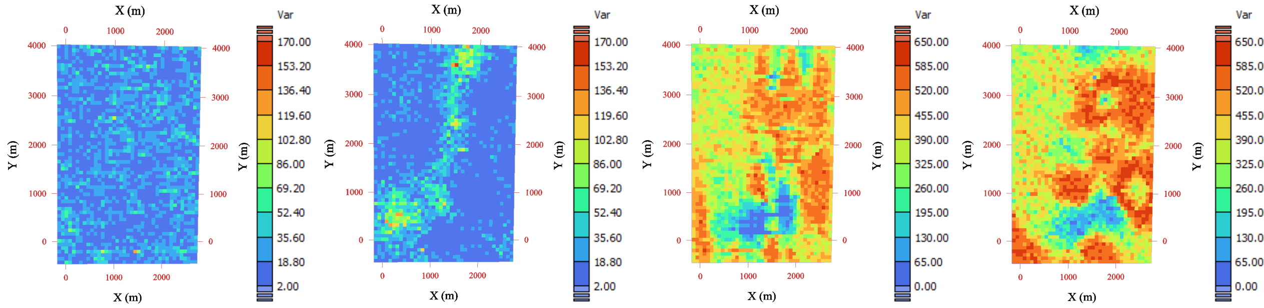
E-type maps, obtained by averaging the original scaled simulated results across 100 realizations of TBSIM and TBCOSIM per block, are represented in Figure 3. As can be seen, independent simulation for Al2O3 generated very noisy and unstructured results. However, for Fe, this revealed specific outlines that are mainly concentrated in the lower part of the region and then decrease to the north-east. Compared to simulation, co-simulation showed more reliable results from a visual inspection. It can be noted that for Al2O3, its spatial variability mainly lies in the direction from the southwest to the northeast. In addition, in the case of Fe, co-simulation has a similar pattern of high values distribution, since the highest values also lie in the southern part of the region, and the lowest values in the east. This good illustration of high and low values in TBCOSIM is related to the influence of co-variates in the process of modeling.

To assess the uncertainty, the variance maps were also obtained for 100 realizations both for independent simulation and co-simulation (Fig. 4). A distinctive feature of simulation and co-simulation at this stage is that in the case of simulation, the produced conditional variance map is very noisy. However, co-simulation for Al2O3 shows high uncertainty where high ore values are located, which may be due to the proportional effect of Al2O3 variability in this deposit (Fig. 3). For Fe, both methods reproduce a similar result and low uncertainty corresponds to places with a high iron content.



(a) *(b)* *(c)* *(d)*

Figure 3 – E-type maps of Al2O3 for (a) TBSIM and (b) TBCOSIM and of Fe for (c) TBSIM and (d) TBCOSIM obtained from (alr) transformation.



(a) *(b)* *(c)* *(d)*

Figure 4 – Conditional variance maps of Al2O3 for (a) TBSIM and (b) TBCOSIM and of Fe for (c) TBSIM and (d) TBCOSIM obtained from (alr) transformation.

As a result of the analysis of grade-tonnage curves, it was found that the co-simulation for Al2O3 produces a much higher value of the metal content compared to the simulation. However, for the fraction of tonnage, the results of both methods are approximately the same. In the case of Fe, both TBSIM and TBCOSIM reproduced very similar results (Fig. 5). This highly affects the mine planning process, since Al2O3 is a deleterious element in this deposit and proper evaluation of this component lead to better evaluation of a mine plan and Net Present Value of a project.

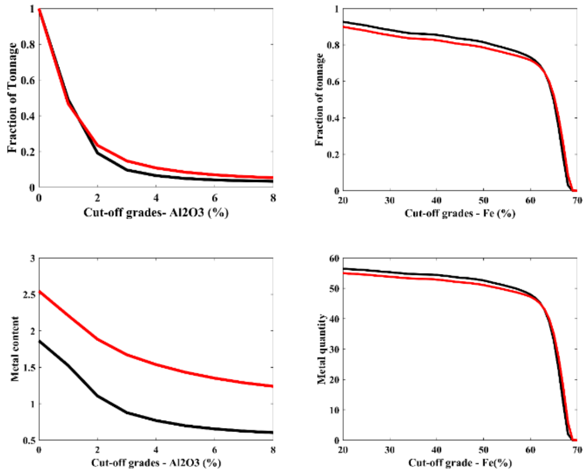


Figure 5 – Grade-Tonnage Curves for TBSIM (black) and TBCOSIM (red) obtained from (alr) transformation.

4. Discussion and Conclusions

In an iron deposit, the clr- and alr-transformation techniques are used to model five geochemical components in this deposit using independent simulation and co-simulation techniques. After analysis of reproduction of original distribution, it was revealed that all clr-transformation techniques are failed to reproduce the original distribution of variables. In addition, using ordinary kriging and co-kriging in the simulation paradigms for alr-transformed data also produced biased results for reproduction of original distributions. Therefore, we decided to show only the more or less acceptable results of simulation and co-simulation over the alr-transformed data when using simple kriging and co-kriging. The results showed that co-simulation outperforms the simulation in terms of reproduction of original bivariate relations, original distribution and conditional variance. The final resource estimation also showed better results for co-simulation.

References

Aitchison, J. (1986). The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44(2), 139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x.

Emery, X., & Lantuéjoul, C. (2006). TBSIM: A computer program for conditional simulation of three-dimensional Gaussian random fields via the turning bands method. Computer & Geoscience, 32(10), 1615–1628. https://doi.org/10.1016/j.cageo.2006.03.001.

Grunsky, E. C., & Caritat, P. D. (2019). State-of-the-art analysis of geochemical data for mineral exploration. Geochemistry: Exploration, Environment, Analysis, 20(2), 217–232. https://doi.org/10.1144/geochem2019-031.

Job, M. (2010). Application of logratios for compositional data. Centre for Computational Geostatistics Report. University of Alberta, Canada.

Journel, A. B., & Huijbregts, C. J. (1978). Mining geostatistics. Mineralogical Magazine, 43(328), 563–564. https://doi.org/10.1180/minmag.1979.043.328.34.

Matheron, G. (1973). The intrinsic random functions and their applications. Advances in Applied Probability, 5(3), 439–468. https://doi.org/10.2307/1425829.

Mueller, U., Tolosana-Delgado, R., & van den Boogaart, K. G. (2014). Simulation of compositional data: A nickel laterite case study. Advances in orebody modelling and strategic mine planning. Melbourne: AusIMM.

Pawlowsky, V. (1984). On spurious spatial covariance between variables of constant sum. Sciences de la terre. Informatique géologique, 21, 107–113. http://pascal- https://futur.upc.edu/1657692.

Pawlowsky-Glahn, V., & Egozcue, J. J. (2016). Spatial analysis of compositional data: A historical review. Journal of Geochemical Exploration, 164, 28–32. https://doi.org/10.1016/j.gexplo.2015.12.010.

Pawlowsky-Glahn, V., & Olea, R. A. (2004). Geostatistical analysis of compositional data. Oxford University Press.

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). Modeling and analysis of compositional data. John Wiley & Sons.

Wackernagel, H. (2003). Multivariate geostatistics: An introduction with applications. Berlin: Springer.

A RIEMANNIAN TOOL FOR CLUSTERING OF GEO-SPATIAL MULTIVARIATE DATA.

Alvaro Riquelme (1)* - Julian Ortiz (1)

Queen's University, Department of Mining, Kingston, Canada (1)
** Corresponding author: alvaro.riquelme@queensu.ca*

Abstract

Geological modeling is required for the correct characterization of natural phenomena and can be done in two steps: (1) clustering the data into consistent groups and (2) modeling the extent of the domains in space honoring the labels defined in the previous step. The clustering step can be based on the information of continuous variables in space, instead of relying on the geological logging of the data. Methodologies able to deal with this problem are applied in the definition of stationary spatial domains, where the assumption of a constant mean within a given spatial domain is critical for resource estimation.

In this work, we propose a method to cluster the data that can then be used for domaining when multiple geochemical variables are available. The method assumes that changes in the local correlation between these attributes can be used to characterize the domains. The method looks at the local (linear) correlations between variables at sample locations, inferred in a local neighborhood. These local correlations defined at sample locations can be mapped into the space of correlation matrices which form a Riemannian manifold, where the Euclidean distance is no longer a suitable metric. The correlation matrices are then clustered by adapting a k-means algorithm to the manifold context. The main challenge is to find a suitable metric able to cluster data on the correlation matrix space, with the purpose of solving the computation of distances between correlation matrices and determine the centroid associated to each cluster. This is addressed by using tools from Riemannian Geometry. An application of the procedure is shown in a real case study to illustrate and capture the essential steps of the methodology. This example demonstrates how the clustering methodology proposed honors the spatial configuration of data delivering continuous clusters and agrees with previous domaining attributes.

USING GEOSTATISTICAL APPROACHES TO ESTIMATE THE RGB VALUES UNDER A CLOUD COVERED AREA WITHIN THE SENTINEL-2 IMAGES

Ashkan Tayebi Gholamzadeh (1)* - Roberto Bruno (1) - Sara Kasmaeeyazdi (1) - Francesco Tinti (1)

University of Bologna, Department of Civil, Chemical, Environmental, and Materials Engineering, Bologna, Italy (1)
** Corresponding author: ashkan.tayebi@gmail.com*

Abstract

Remote Sensing tools and approaches are now widely used in many different fields such as mining engineering, soil sciences, environmental monitoring, etc. Satellite data is useful within many geosciences because it provides a large amount of time-space continuous data which is easily and quickly accessible. Clouds and shadows within satellite images, however, are an important challenge as they obscure surface features. In many cases, the image cannot be used or the target area cannot be studied because of cloud cover. To counteract this, Geostatistical tools can be used to estimate the missing data in the target area.

In this study, a Sentinel-2 image from Copernicus data of land cover in Emilia Romagna (Italy) is used and the spectrum bands are resampled into 10-m resolution. The objective of the study is to estimate the RGB-values of a cloud-covered area within the image. The statistical parameters and the spatial variability of nearby pixels are studied by testing different neighborhoods to analyze the possibility of interpolation within the cloud-covered area. Then the estimator properties are controlled and the values of the cloud-covered pixels are estimated by assuming different parameters such as the size and shape of the neighborhood in consideration, number of pixels used for the estimation and the pixels distribution in the image. The reliability of results is then evaluated by analyzing the estimation variance of each estimated pixel and mapping them. The analysis has been carried out using mainly MATLAB and VBA programming.

To validate the results, an image taken at a similar a timeframe has been used for comparison. Results show the advantages and disadvantages of different ways to estimate the cloud-covered area and reliability of the estimations are compared. In addition, by comparing the results with the original values, the effect of the cloud cover has been evaluated on different remote sensed band values.

SURROGATE MODELS AS MANAGEMENT TOOLS FOR THE REQUENA-UTIEL AND CABRILLAS-MALACARA AQUIFERS

Janire Uribe-Asarta (1)* - Vanessa A. Godoy (1) - J. Jaime Gómez-Hernández (1)

Universitat Politecnica de Valencia, Valencia, Spain (1)
** Corresponding author: j.uribeasarta@gmail.com*

Abstract

Water authorities, legislators and groundwater users have commonly used groundwater numerical models as tools for aquifer management, as they provide crucial information to make decisions. However, these models are complex and computationally expensive. Moreover, numerical models require specialized personnel who can operate them and interpret the results. An approach to overcoming these disadvantages is to replace the numerical model with a data-driven surrogate. Surrogate models are built with machine learning methods, and they are trained from a number of scenarios considering possible ranges of variations in the inputs or outputs of the numerical model, such as, the extracted or injected flow, recharge or evapotranspiration.

This study has assessed different machine learning methods for the construction of groundwater flow surrogate models for the Requena-Utiel and Cabrillas-Malacara aquifers, in Valencia, Spain. The goal is to provide a fast and precise tool that allows evaluating the impact of possible changes in external variables (pumping, precipitation) on the decrease or increase in piezometric heads. To facilitate the use of the model and allow anyone to make a query, it is freely accessible from the internet.

The surrogate models have resulted in very precise approximations for the input and output ranges for which the training data was generated. Likewise, the computational reductions are remarkable. We conclude that the surrogate models are fast, easy-to-use and powerful tools to assist in aquifer management.

This research was developed under the scope of the InTheMED project. InTheMED is part of the PRIMA Program supported by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No 1923.

APPLICATION OF GEOSTATISTICS AND SELF-ORGANIZING MAPS FOR ESTIMATION OF GROUNDWATER LEVEL SPATIAL DISTRIBUTION IN COMPLEX HYDROGEOLOGICAL SYSTEMS

Emmanouil Varouchakis (1)* - George P. Karatzas (2) - Ioannis Trichakis (3)

Mineral Resources Engineering, Technical University of Crete, Chania, Greece (1) - Chemical and Environmental Engineering, Technical University of Crete, Chania, Greece (2) - European Commission, Joint Research Centre, Ispra, Italy (3)
** Corresponding author: evarouchakis@isc.tuc.gr*

Abstract

Water scarcity is a major global problem and expected to become even more significant in the near future. Overexploitation of groundwater, from direct and indirect activities, mainly due to intensive agricultural activity, combined with projected climatic change, has a great impact on the hydrological/hydrogeological conditions of the Mediterranean region. This generates concerns over the sustainability of groundwater resources.

In this work, a geostatistical analysis approach, in combination with a machine-learning algorithm, i.e. Self-Organizing Maps (SOM), was applied with three main goals. a) to develop reliable spatial maps of groundwater level variability in a large-scale hydrogeological system of complex aquifers, and to identify groundwater level zones, b) to process efficiently a large dataset of 2524 wells using geostatistical analysis and c) to accurately map groundwater level spatial distribution in a local scale. As a first step, the algorithm applied Self-Organizing Maps to identify locally similar input data, and then used those identified clusters of data, by means of Ordinary Kriging to estimate the spatial distribution of groundwater level and produce maps in large and local scales.

The proposed method provides a complementary tool to physically based models that require extensive data and hydrogeological details to produce reliable large-scale groundwater level fields. Such maps can be helpful to organize management scenarios for the sustainability of groundwater resources in large hydrogeological districts and to assess the climate change effect of hydrometeorological variables on the groundwater resources.

This research was developed under the scope of the InTheMED project. InTheMED is part of the PRIMA Program supported by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No 1923.

A COMPARISON BETWEEN BAYESIAN AND ORDINARY KRIGING BASED ON VALIDATION CRITERIA: APPLICATION TO RADIOLOGICAL CHARACTERISATION

Martin Wieskotten (1,3)* - Marielle Crozet (1) - Bertrand Iooss (2,5) - Céline Lacaux (3) - Amandine Marrel (4,5)

CEA, DES, ISEC, DMRC, Univ. Montpellier, Marcoule, Lma Université D'Avignon, Avignon, France (1) EDF R&D, Chatou, France (2) - LMA Université D'Avignon, Avignon, France (3) - CEA, DES, IRESNE, DER, Cadarache, Saint-Paul-Lez-Durance (4) - Institut de Mathématiques de Toulouse, Toulouse, France (5)

** Corresponding author: martin.wieskotten@gmail.com*

Abstract

Radiological characterisation is one of the main challenges of the decommissioning and dismantling projects of nuclear facilities. This is an important step in decommissioning projects as it aims to estimate the quantity and spatial distribution of different radionuclides. To carry out the estimation, measurements are performed on site to obtain preliminary information and spatial interpolation, for example using the kriging tool, which allows to predict the value of interest for the contamination (radionuclide concentration, radioactivity, etc.) at unobserved positions. A strong assumption made when applying ordinary kriging is that the variance and range parameters are known, which is rarely the case. Furthermore, the estimation error made when these parameters are estimated from the data is never taken into account, although this can lead to biased kriging predictions and overoptimistic prediction variances. This problem is emphasised when only a few observations are available (a quite common case in decommissioning projects), since the variance of the parameters' estimators becomes larger. To address this issue, we propose to use Bayesian kriging where the model parameters are considered as random variables, which allows to take into account their uncertainties. The use of prior specifications in Bayesian kriging also allows for more robust parameter estimate when only a few observations are available. As such, the present work focused on assessing the usefulness of Bayesian kriging whilst comparing its performance to that of ordinary kriging. First, in order to make a relevant comparison, a simulated data set with known parameters is initially studied and several cross-validation criteria, such as the predictivity coefficient (Q^2), the Predictive Variance Adequacy (PVA), and the α-CI plot, are estimated for varying data sizes to quantify the performances of both kriging methods. A new criterion, the Predictive Interval Adequacy (PIA), is also introduced and studied. Then, the same comparison is applied on a real data set from the decommissioning project of the G3 reactor at the CEA's Marcoule site.

Keywords: Geostatistics; Bayesian kriging; Ordinary kriging; Validation criterion.

1. **Introduction**

Radiological characterization is one of the main challenges encountered in the nuclear industry for the decommissioning and dismantling (D&D) of old infrastructures. Its main goal is to evaluate the quantity and spatial distribution of radionuclides. As such, measurements are made to constitute a data set and obtain preliminary information. While measurements are made, many problems can arise. The radioactivity present on site can be dangerous for operators and does not allow for many measurements. In some extreme cases, drones and robots have to be used CEA DEN (2017), making measurements more expensive and reducing data set's sizes. It is therefore quite common in nuclear D&D characterisation to have only a small number of data available. A balance has to be found between information and costs, and statistical tools make it possible to optimise the information extracted, within a rigorous mathematical framework giving associate confidence intervals.

More precisely, spatial statistics or geostatistics are used to predict the variable of interest at unobserved position (prediction of the expected value), with an indication of the expected error in prediction (the prediction variance). The methodology is often based on two steps: first the construction of a statistical model with the estimation of its parameters, followed by prediction with the linear interpolator based on kriging. The classical kriging model also receives a common critic: its predictions do not take into account the uncertainty in the estimation of the model parameters. The variances of the predictions are too optimistic and the neglected model uncertainties can have a significant impact. This problem is made worse for smaller data sets, which can be common in D&D projects. The work of Desnoyers (2010) is one of the first example of application of kriging to radiological characterisation. In his work a practical case study was analysed, but was based on many measurements, which is not realistic in the case of industrial nuclear D&D projects.

To overcome this, a Bayesian approach was first proposed by Kitanidis (1986). Its main goal was to take into account uncertainties in the scale and mean parameters of the model. The work of Handcock and Stein (1993) then completed the full Bayesian approach which considers all the parameters of the model as unknown. More recently, a slightly different approach was presented by Krivoruchko and Gribov (2019) and is called empirical Bayesian kriging. While the equations are similar to the ones of regular Bayesian kriging, the prior choices are obtained through unconstrained simulations of the random field. This approach was adapted to allow for multi-fidelity applications, where Bayesian theory is used to update the initial data with new, more accurate data (classically used with cokriging if correlations between old and new data exist). Some examples can be found in meteorology with Gupta et al. (2017) or for oil extraction in Al-Mudhafar (2019). Note that a more complete description of Bayesian kriging with an extension to generalised linear model is presented in Diggle and Ribeiro (2007).

The following Section describes the different kriging models and the model validation criteria that are used. Section 3 provides our models' comparison results on several numerical tests and on our real application case study. Section 4 gives some conclusions.

2. Material and Methods

2.1. Spatial Model and Predictions

The model considered is the following random field:

$$\{Z(x), x \in \mathbb{R}^2\},$$

which is constrained to $D \subset \mathbb{R}^2$. The random field $Z(.)$ is isotropic and stationary, meaning:

$$\forall\, x \in\ D, E[Z(x)] = \beta,$$
$$\forall\, x, y\ \in\ D, Cov\big(Z(x), Z(y)\big) = \sigma^2 C_\phi(|x - y|),$$

where β, σ^2 and ϕ are the mean variance and range parameters, respectively. The term C_ϕ corresponds to a semidefinite positive function. Moreover, the random field is considered Gaussian. Thus for n observations at positions $\{x_1, \dots, x_n\}$, we obtain the Gaussian random vector $\boldsymbol{Z} = (Z(x_1), \dots, Z(x_n))'$. We then have:

$$\boldsymbol{Z}|\beta, \sigma^2, \phi \sim \mathcal{N}(\beta \mathbf{1}_n, \sigma^2 \boldsymbol{K}_\phi),$$

with $\mathbf{1}_n = (1, \dots 1)'$, and $\sigma^2\, \boldsymbol{K}_\phi = (Cov(Z(x_i), Z(x_j)))_{1 \le i,j \le n}$ the covariance matrix. The observation sample of $\boldsymbol{Z}$ is written $\boldsymbol{z} = (z(x_1), \dots, z(x_n))'$.

The model is therefore specified by 3 different parameters: the trend parameter $\beta \in D_\beta$, the scale (or variance) parameter $\sigma^2 \in D_{\sigma^2}$ and the range parameter $\phi \in D_\phi$. The first step of the geostatistical methodology is to estimate these parameters. Two main procedures exist: variographic analysis and maximum likelihood estimation. An extensive literature is available about parameter estimation with variographic analysis, such as Chilès and Delfiner (2012), Webster and Oliver (2007). In this work, we use maximum likelihood estimation to avoid automatic fitting of variograms since our numerical tests will require the estimation of parameters for numerous data sets generated by simulations. Automatic fitting of variograms is strongly discouraged in most of the literature (Chilès and Delfiner, 2012; Webster and Oliver, 2007), so we avoid this method here.

The kriging predictor is a linear interpolator which expressions are derived from supplementary conditions, such as minimizing the prediction variance. For a detailed description of kriging and its construction, the reader can refer to the reference books of Chilès and Delfiner (2012), Cressie (1993) for geostatistics, but also Rasmussen and Williams (2006) for computer code surrogate models. Let x_0 be an unobserved position at which we wish to predict the expected value and the variance of $Z(x_0)|\sigma^2, \phi, \boldsymbol{Z} = \boldsymbol{z}$ (the mean is considered unknown), the ordinary kriging equations are:

$$E[Z(x_0)|\sigma^2, \phi, \boldsymbol{Z}] = \left(\boldsymbol{k} + \boldsymbol{1}_n \frac{1 - \boldsymbol{1}_n' \boldsymbol{K}_\phi^{-1} \boldsymbol{k}}{\boldsymbol{1}_n' \boldsymbol{K}_\phi^{-1} \boldsymbol{1}_n}\right)' \boldsymbol{K}_\phi^{-1} \boldsymbol{Z},$$

$$Var[Z(x_0)|\sigma^2, \phi, \boldsymbol{Z}] = \sigma^2 \left(\boldsymbol{1} - \boldsymbol{k}' \boldsymbol{K}_\phi^{-1} \boldsymbol{k} + \frac{\left(1 - \boldsymbol{1}_n' \boldsymbol{K}_\phi^{-1} \boldsymbol{k}\right)^2}{\boldsymbol{1}_n' \boldsymbol{K}_\phi^{-1} \boldsymbol{1}_n}\right),$$

with $\sigma^2 \boldsymbol{k} = (Cov(Z(x_0), Z(x_j)))_{1 \le j \le n}$. A major concern for applications of these equations is that they are conditional on the knowledge of the variance and correlation length parameters, which is mostly unrealistic since they are estimated. This assumption yields overoptimistic prediction variances and narrower confidence intervals. This problem is made worse in case of a small data set where parameter estimation is sensible to each observation. To address this issue, more robust methods exist such as cross-validation estimation (Bachoc (2013a)). Another solution is to consider the parameters as random variables. Bayesian approach seems natural in this case and leads to Bayesian kriging.

Indeed, Bayesian kriging deals simultaneously with estimation and predictions by considering the parameters as random variables that must be predicted conditionally to the observed data. We use here the approach described by Diggle and Ribeiro (2002). For ease of notation, densities will be denoted as $p(.)$ and the conditioning to parameters will be simplified from $Z|\beta = \tilde{\beta}\ to\ Z|\beta$. Bayesian kriging predictions are derived from the predictive distribution as follows:

$$\begin{aligned} p(Z(x_0)|\boldsymbol{Z} = z) &= \int_{D_\beta \times D_{\sigma^2} \times D_\phi} p(Z(x_0), \beta, \sigma^2, \phi | \boldsymbol{Z} = \boldsymbol{z}) d\beta d\sigma^2 d\phi \\ &= \int_{D_\beta \times D_{\sigma^2} \times D_\phi} p(Z(x_0), \beta, \sigma^2 | \phi, \boldsymbol{Z} = \boldsymbol{z}) p(\phi | \boldsymbol{Z} = \boldsymbol{z}) d\beta d\sigma^2 d\phi \\ &= \int_{D_\phi} p(Z(x_0) | \phi, \boldsymbol{Z} = \boldsymbol{z}) p(\phi | \boldsymbol{Z} = \boldsymbol{z}) d\phi\,. \end{aligned}$$

As per usual in Bayesian framework, we choose a joint prior distribution for β, σ^2:

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

For the correlation length, the prior is reduced to a uniform law between the minimum and the maximum distance allowed by the data set:

$$\phi \sim U(D_{\phi,min}, D_{\phi,max}).$$

2.2 Validation criteria

Different validation criteria have been recently deeply studied by Demay et al. (2022), to compare and choose different covariance models for geostatistical predictions. These criteria aim to assess the quality of both the predictions of the model and the associated prediction variances. Their expressions are given here in their leave-one-out cross-validation form, but can be extended to K-fold cross-validation or to validation sets cases. These criteria with some new adaptations are given below.

Predictivity coefficient (Q^2)

The main goal of this coefficient is to evaluate the predictive accuracy of the model by normalising the errors, allowing a direct interpretation in terms of explained variance. Its definition is the following:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(z(x_i) - \hat{z}_{-i})^2}{\sum_{i=1}^{n}(z(x_i) - \hat{\mu})^2},$$

where $\hat{z}_{-i}$ is the value predicted at location x_i by the model built without the i-th observation and $\hat{\mu}$ is the empirical mean of the data set. This coefficient measures the quality of the predictions and how near they are to the observed values. It is similar to the coefficient of determination used for regression (with independent observations), but estimated here by cross-validation. The closer its value is to 1, the better the predictions are (relatively to the observations). As a rule of thumb, if the Q^2 is smaller than 0.5 (i.e. less than 50% of output variance explained), the model is not considered valid.

Predictive variance adequacy (PVA)

This second criterion aims to quantify the quality of the prediction variances given by the model and kriging. Finely studied in Bachoc (2013b), it is defined by the following equation:

$$PVA = \left| log\left(\frac{1}{n}\sum_{i=1}^{n}\frac{(z(x_i) - \hat{z}_{-i})^2}{\hat{s}_{-i}^2}\right)\right|,$$

where $\hat{s}_{-i}^2$ is the prediction variance (at location x_i) of the model built without the i-th observation. It estimates the average ratio between the squared observed prediction error and the prediction variance. It therefore gives an indication of how much a prediction variance is bigger or smaller than the one expected. The closer the PVA is to 0, the better the prediction variances are. For example, a $PVA \approx 0.7$ indicates prediction variances that are on average two times bigger or smaller than the squared errors.

Predictive interval adequacy (PIA)

The PVA is a criterion of variance adequacy but does not take into account a possible skewness in the predictive distribution. In the Gaussian case (like ordinary kriging), mean and variance completely characterise the distribution. But in the case of Bayesian kriging where the predictive distribution is no longer Gaussian, the Q^2 and PVA are not sufficient to evaluate the quality of the model and its prediction. As such, we propose a new complementary geometrical criterion called the predictive interval adequacy (PIA) and defined as follows:

$$PIA = \left| log\left(\frac{1}{n}\sum_{i=1}^{n}\frac{(z(x_i) - \hat{z}_{-i})^2}{\left(\hat{q}_{0.31,-i} - \hat{q}_{0.69,-i}\right)^2}\right)\right|,$$

where $\hat{q}_{0.31,-i}$ (respectively $\hat{q}_{0.69,-i}$) is the estimation of the quantile of order 0.31 (respectively 0.69) of the predictive distribution (at location x_i) without the i-th observation. Note that it has been defined to be identical to the PVA for a Gaussian distribution, but rather than comparing squared errors to the predictive variance, it compares the width of prediction intervals with the squared errors. Another main difference is that the intervals considered by the PIA are centered on the median while those of the PVA are centered around the mean. Finally, an estimation of the predictive distribution is necessary to compute in practice this criterion, whereas the PVA only requires the computation of predictive mean and variance.

α-CI plot and Mean Squared Error α ($MSE\alpha$)

The Gaussian process model allows to build prediction intervals of any level $\alpha \in]0,1[$:

$$CI_\alpha\big(z(x_i)\big) = \left[\hat{z}_{-i} - \hat{s}_{-i} q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}; \hat{z}_{-i} + \hat{s}_{-i} q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}\right],$$

where $q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}$ is the quantile of order $1-\frac{\alpha}{2}$ of the standard normal distribution. This expression is only valid if all parameters are known. For example, if the scale parameter is incorrectly estimated, the width of the predicted confidence intervals will not reflect what we might observe. But how can we validate a confidence interval without prior knowledge of the model parameters? The idea behind this criterion is to evaluate empirically the number of observations falling into the predicted confidence intervals and to compare this empirical estimation to the theoretical ones expected:

$$\Delta_\alpha = \frac{1}{n}\sum_{i=1}^{n}\delta_i \text{ where } \delta_i = \begin{cases} 1 & \text{if } z(x_i) \in CI_\alpha(z(x_i) \\ 0 & \text{else} \end{cases}.$$

This value can be computed for varying α, and can then be visualised against the theoretical values, yielding what Demay et al. (2022) call the α-CI plot.

Similarly to the PIA, the α-CI plot must be adapted to the Bayesian kriging since the posterior distribution is not Gaussian. We therefore introduce a slightly different criterion based on the quantiles of the predictive distribution. More precisely, the α-CI plot relies now on credible intervals defined as:

$$\widetilde{CI}_\alpha(z(x_i)) = \left[\hat{q}_{\frac{1-\alpha}{2},-i}\,; \hat{q}_{\frac{1+\alpha}{2},-i}\right],$$

where $\hat{q}_{\frac{1-\alpha}{2},-i}$ (respectively $\hat{q}_{\frac{1+\alpha}{2},-i}$) is the estimation of the quantile of order $\frac{1-\alpha}{2}$ (respectively $\frac{1+\alpha}{2}$) of the predictive distribution (at location x_i) of the model built without the i-th observation.

Once again, we obtain a criterion that is identical for both methods when the predictive distribution is Gaussian.

Illustrations of α-CI plot can be found in Demay et al. (2022). To summarise the α-CI plot, we also introduce a quantitative criterion called the Mean Squared Error α defined as follows:

$$MSE\alpha = \sum_{j=1}^{n_\alpha}(\Delta_{\alpha_j} - \alpha_j)^2,$$

where n_α is the number of widths considered for prediction intervals and α_j the width of the j-th confidence interval $\widetilde{CI}_{\alpha_j}$ (in practice a regular discretization of α on $]0,1[$ will be considered to compute $MSE\alpha$). The closer this criterion is to 0, the better the confidence/credible intervals are in average.

The different aforementioned criteria provide complementary information to evaluate the prediction quality of the kriging model, either in terms of mean, variance or confidence/credible intervals. They will be used in the following to compare the performance of ordinary and Bayesian kriging.

3. Numerical tests and results

3.1. Analytical example

Our goal is to compare Bayesian and ordinary kriging (the latter being the more commonly used kriging method). To do so, we will compute the different criteria mentioned above on data sets of different sizes.

First, we consider data sets simulated from an analytical Gaussian process model with known parameters. More precisely, the data sets are simulated in the input space $[0,10]^2$ from a Gaussian process with an exponential covariance and the following parameters:

$$\beta = 0.5, \sigma^2 = 0.1, \phi = 4.5.$$

We simulate data sets of different sizes, varying from 16 and 81 observations, sampled in a square grid on the input space. For each size, the process is repeated 100 times with independent random Gaussian process simulations.

For each data set, Bayesian and ordinary kriging models are estimated and the different validation criteria are computed by cross-validation. Results are given in Figure 1 with boxplots w.r.t. the data set sizes.

The results indicate that Bayesian kriging performs better in terms of both mean and prediction variance for small sample sizes. More precisely, Bayesian kriging outperforms ordinary kriging on all the four criteria for data sets with less than 40 observations. This result is especially visible for the PVA and PIA and shows that the main difference between both kriging methods still lies in the predictive variance estimation. This is mainly because the Bayesian kriging accounts for more uncertainty of the estimates of Gaussian process parameters than ordinary kriging. Bayesian kriging therefore yields larger and more accurate prediction intervals, and as a result better PVA, PIA, and $MSE\alpha$ criteria.

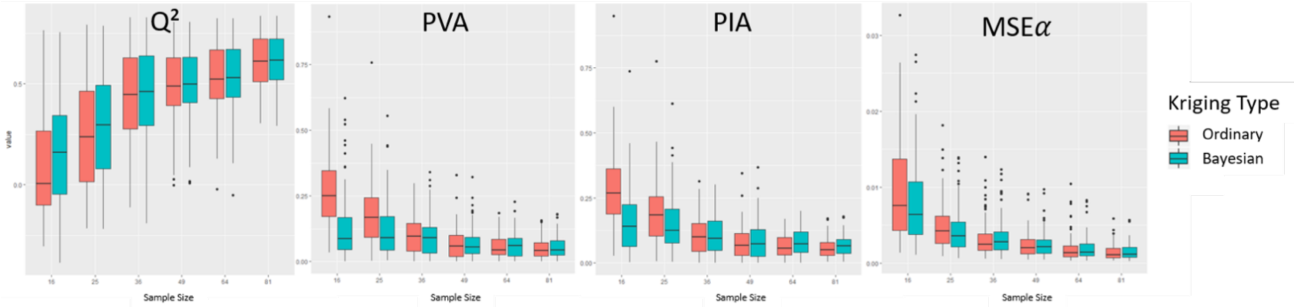


Figure 1 – Distribution of validation criteria (Q², PVA, PIA, and $MSE\alpha$) w.r.t. the size of data sets, for simulated data.

It can also be noted that for larger data sets, Bayesian and ordinary kriging yield similar results. This observation was to be expected, since Bayesian and inferential methodology coincide for larger data sets. It can be therefore argued that Bayesian kriging becomes less advantageous and relevant for data set with more than 40 observations, since its computational cost is higher than that of ordinary kriging. Note that Q^2 values are also extremely low for 49 observations or fewer, but again this is to be expected for very small data sets.

3.2. Real application case: G3's data set

We apply a similar protocol to the real data set of the G3 reactor in CEA Marcoule. This data set is made of 70 observations of radioactivity measurements sampled in the input domain [0,10] × [0,7]. To generate multiple data sets, we resampled without replacement data sets of various sizes 20, 30, 40, 50, 60 and 70 observations, with the last one being the real size of the original data set. Once again, the process is repeated 100 times for each sample size (except for 70 observations) and for each sample a cross-validation is applied to estimate the validation criteria.

The obtained results are given in Figure 2. They are similar to the ones obtained for the simulated data sets. We can remark that the variance of each validation criterion is reduced as the data sets size grows. This is both explained by the larger data sets, but also by our protocol, where observations are randomly drawn without replacement among the original 70 observations, so that as the data set sizes increases, the samples differ less and less. It can be noted that ordinary kriging seems to be slightly better than Bayesian kriging for larger data sets, reinforcing our precedent argument that Bayesian kriging should be reserved for smaller data sets for which the uncertainty in parameter estimation is high.

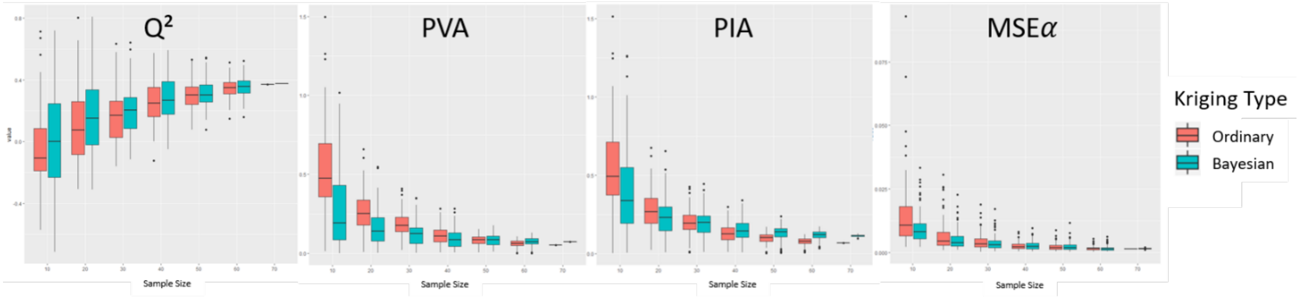


Figure 2 – Distribution of validation criteria (Q², PVA, PIA, and $MSE\alpha$) w.r.t. the size of data sets, for the G3 data set.

4. Discussion and Conclusions

In conclusion, the use of Bayesian kriging for spatial interpolation of data sets in support of decommissioning and dismantling projects shows promising results. It is particularly true for small data sets for which it outperforms the ordinary kriging in terms of accuracy of predictive mean, variance and predictive intervals. This advantage becomes less important as the sample size increases: ordinary kriging, less computationally expensive, is then preferable for large data sets. Bayesian kriging has also the drawback of requiring a prior specification, which is often difficult to choose and can strongly influence the predictions. Therefore, the use of Bayesian kriging should be restricted to smaller data sets or cases in which prior information on parameters is well known. Our future work will focus on better modelling of measurement uncertainty in Bayesian kriging, particularly through the use of heteroscedastic models (Ng and Yin (2012)).

References

Al-Mudhafar, W.J. (2019). Bayesian kriging for reproducing reservoir heterogeneity in a tidal depositional environment of a sandstone formation. Journal of Applied Geophysics 160, 84–102. https://doi.org/10.1016/j.jappgeo.2018.11.007.

Bachoc, F. (2013a). Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. Computational Statistics & Data Analysis 66, 55–69. https://doi.org/10.1016/j.csda.2013.03.016

Bachoc, F. (2013b). Estimation paramétrique de la fonction de covariance dans le modèle de krigeage par processus gaussiens: application à la quantification des incertitudes en simulation numérique, PhD Thesis of University Paris VII.

CEA DEN (2017). L'assainissement-démantèlement des installations nucléaires, Le Moniteur. ed, Monographie CEA.

Chilès, J.-P., Delfiner, P. (2012). Geostatistics : Modeling Spatial Uncertainty, Second Edition. ed, Wiley Series In Probability and Statistics. Wiley.

Cressie, N. (1993). Statistics for spatial data. John Wiley & Sons.

Demay, C., Iooss, B., Le Gratiet, L., Marrel, A. (2022). Model selection based on validation criteria for Gaussian process regression: An application with highlights on the predictive variance. Quality and Reliability Engineering International, 38:1482-1500. https://doi.org/10.1002/qre.2973.

Desnoyers, Y. (2010). Approche méthodologique pour la caractérisation géostatistique des contaminations radiologiques dans les installations nucléaires, PhD Thesis of Ecole des Mines de Paris.

Diggle, P.J., Ribeiro, P.J. (2007). Model-based Geostatistics, Springer Series in Statistics. Springer.

Diggle, P.J., Ribeiro, P.J. (2002). Bayesian Inference in Gaussian Model-based Geostatistics. Geographical and Environmental Modelling 6, 129–146. https://doi.org/10.1080/1361593022000029467.

Gupta, A., Kamble, T., Machiwal, D. (2017). Comparison of ordinary and Bayesian kriging techniques in depicting rainfall variability in arid and semi-arid regions of north-west India. Environ Earth Sci 76, 512. https://doi.org/10.1007/s12665-017-6814-3.

Kitanidis, P.K. (1986). Parameter Uncertainty in Estimation of Spatial Functions: Bayesian Analysis. Water Resources Research 22, 499–507. https://doi.org/10.1029/WR022i004p00499.

Krivoruchko, K., Gribov, A. (2019). Evaluation of empirical Bayesian kriging. Spatial Statistics 32, 100368. https://doi.org/10.1016/j.spasta.2019.100368.

Ng, S.H., Yin, J. (2012). Bayesian Kriging Analysis and Design for Stochastic Simulations. ACM Trans. Model. Comput. Simul. 22, 17:1-17:26. https://doi.org/10.1145/2331140.2331145.

Rasmussen, C.E., Williams, C.K.I. (2006). Gaussian Processes for Machine Learning. MIT Press.

Webster, R., Oliver, M.A. (2007). Geostatistics for environmental scientists. John Wiley & Sons.

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, Zag, Tübingen, Germany (1) - University of Stuttgart, Vegas, Stuttgart, Germany (2) - Kansas Geological Survey, The University of Kansas, Lawrence, United States (3) - University of Stuttgart, Lhg, Stuttgart, Germany (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.

Various hypotheses regarding the representation of the vertical non-stationarity are evaluated. The most complex model involves 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 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 STOCHASTIC MODEL OF AN ALERT SYSTEM FOR DETECTING LOCAL ANOMALOUS INCIDENCE VALUES OF COVID-19

Ana Filipa Duarte (1)* - Leonardo Azevedo (1) - Amilcar Soares (1)

CERENA, DECivil, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal (1)
** Corresponding author: filipadamasoduarte@tecnico.ulisboa.pt*

Abstract

At the beginning of the COVID-19 pandemic, the uncertainty about the dynamics of the phenomenon was so high that one of the most important alert tools for managing the pandemic was the real-time characterization of the cumulative incidence rates (i.e., the infection risks) in wide geographic domains. Thus, geostatistical models have been used to characterize a daily risk map of COVID-19 incidence and the associated uncertainty, like the COVID-19 Risk Map for Portugal (Azevedo et al., 2019) which was adopted by the Portuguese Directorate-General for Health as a tool for controlling the incidence and managing the mitigation measures.

After the different stages of confinement, risk control and management concerns moved to a more local scale. Hence, one of the major challenges of public health management to face in a new pandemic outbreak of COVID-19 is to detect local anomalous daily values, in order to take local pandemic containment measures promptly. In this context, an anomalous incidence value of COVID-19 is the one that exceeds the predicted or expected value at one spatial location, above a certain threshold. An anomalous value differs from neighbourhood values, as it is a spatial discontinuity and has different temporal behaviour.

In this study, we propose a method for predicting local distributions of incidence values and detect the anomalous values based on the local predicted probability distribution functions, by accounting for the spatiotemporal evolution of COVID-19 cumulative incidence rates of a given region, namely a set of municipalities in the metropolitan region of Lisbon used to validate the model.

For this purpose, the space and time components of the prediction model are treated separately. In the first step, we consider historical data of each municipality to forecast incidence values of COVID-19 by training a Machine Learning algorithm (Genetic Programming). The predicted and observed incidence values allow to build the corresponding predicted probability distribution function (PDF) of cumulative incidence rates for each municipality centroid location and for the required day.

In a second step, the spatial component is added by characterising the local predicted PDFs, using geostatistical sequential simulation. With the simulated realizations, we obtain a space-time predicted model of local PDFs for the "next day". Comparing the real observed incidence values of the "next day", at a specific location, we can classify this value as anomalous if it exceeds one extreme percentile of the predicted PDF at that location and also has a different behaviour from the neighbourhood value.

SPATIALEPISIM: AN R SHINY APP FOR TRACKING INFECTIOUS DISEASES IN LOW- AND MIDDLE-INCOME COUNTRIES (LMIC)

Ashok Krishnamurthy (1)* - Crystal Wai (1) - Gursimran Dhaliwal (1) - Jake Doody (2)

Mount Royal University, Mathematics and Computing, Calgary, Canada (1) - University of Maryland Baltimore County, Department of Mathematics and Statistics, Baltimore, United States (2)
** Corresponding author: akrishnamurthy@mtroyal.ca*

Abstract

It is essential to understand what future epidemic trends will be, as well as the effectiveness and potential impact of public health intervention measures. The goal of this research is to provide insight that would support public health officials towards informed, data-driven decision making. We present spatialEpisim, an R Shiny app that integrates mathematical modelling and open-source tools for tracking the spatial spread of infectious diseases in low- and middle-income countries (LMIC). Our app uses open-source GIS tools and freely available population count data downloaded as a gridded raster map at the 30-arc second resolution from WorldPop (www.worldpop.org) to assess the geographical spread of an epidemic. With this app we can visualize how an infectious disease spreads across a large geographical area. The rate of spread of the disease is influenced by changing the model parameters and human mobility patterns. We present spatial compartmental models of epidemiology (ex: SEIR, SEIRD, SVEIRD) to capture the transmission dynamics of the spread of COVID-19. The rate of spread of the disease is influenced by changing the model parameters and human mobility patterns. First, we run the spatial simulations under the worst-case scenario, in which there are no major public health interventions. Next, we account for mitigation efforts including strict mask wearing and social distancing mandates, targeted lockdowns, and widespread vaccine rollout to vaccinate priority groups. As a test case Nigeria is selected and the projected number of newly infected and death cases are estimated and presented. Projections for disease prevalence with and without mitigation efforts are presented via time-series graphs for the epidemic compartments. Predicting the transmission dynamics of COVID-19 is challenging and comes with a lot of uncertainty. In this research we seek primarily to clarify epidemiological and mathematical ideas, rather than to offer definitive medical answers. Our analyses may shed light more broadly on how an infectious disease spreads in a large geographical area with places where no empirical data is recorded or observed.

UNPACKING OCCUPATIONAL HEALTH DATA IN THE TERTIARY SECTOR. FROM SPATIAL CLUSTERING TO BAYESIAN DECISION MAKING

María Pazo (1)* - Carlos Boente (2) - Teresa Albuquerque (3, 4, 5) - Natália Roque (3, 4) - Saki Gerassis (1) - Javier Taboada (1)

GESSMin Research Group, Department of Natural Resources and Environmental Engineering, University of Vigo, Vigo, Spain (1) - CIQSO-Center for Research in Sustainable Chemistry, Associate Unit CSIC-University of Huelva "Atmospheric Pollution", Huelva, Spain (2) - Instituto Politécnico de Castelo Branco, Castelo Branco, Portugal (3) - Centro de Estudos de Recursos Naturais, Ambiente e Sociedade (CERNAS), Instituto Politécnico de Castelo Branco, Portugal (4) – ICT, Universidade de Évora, Évora, Portugal (5)
** Corresponding author: sakis@uvigo.es*

Abstract

The health status of the service sector workforce is a great unknown for medical geography. Despite the advances carried out by spatial epidemiology to predict spatial patterns of disease incidence, there are important challenges unsolved. In particular, the main issue resides in the ability to effectively simplify and visually represent the problem domain, given the need to cover very different service activities and, at the same time, consider the impact of numerous emerging risk factors such as those stemming from bioclimatic and socioeconomic variables. This article proposes a new approach that allows to consider, simplify, prioritise and visualise multiple occupational health risk factors giving rise to not healthy workers. For that, it is used a twofold approach based on an innovative synergy between Bayesian machine learning and geostatistics, to analyse up to 74.401 occupational health surveillance tests gathered between 2012-2016 in Spain. This solution allows to extract relevant patterns over those risk factors that cannot be further discriminated in the Bayesian network, such as *spine* or *limbs observations*, depicting distribution maps of key differentiating variables computed by an ordinary kriging approach.

Keywords: Health data; Information theory; Ordinary kriging; Target analysis.

1. Introduction

The service sector, generally known as the tertiary sector of the economy, consists of the provision of services to other businesses, including end consumers. Services generate approximately 70% of the European Union's Gross domestic product (GDP) and employment (Eurostat, 2022). Some of the most common areas of the service sector are tourism (e.g., accommodation, travel agents), catering, education, real state, transport, and financial-related services. The variety of possible activities within this sector makes extremely complex the estimation of the health status of their workforce.

This aspect has been accentuated by the impact of the COVID-19 pandemic (Chang et al., 2021). Overall, men's and women's work tasks are in many cases considerably different, triggering occupational health risks for each gender. Despite the advances carried out by spatial epidemiology methods to predict spatial patterns of disease incidence, the abundance and accuracy of occupational health risk maps are still very limited (Gerassis et al., 2021). This is due in part to the multiple variables to represent without a clear approach to simplify the problem domain, and even more, to find out those differentiating variables. To this situation, it must be added the already undeniable impact on the health of climate change (Orlov et al., 2020). The rise in temperatures is expected to open the door for an increasing number of pathologies whose effects can worse at work.

In this respect, given the outcomes of numerous investigations related to the effect that the climate change has on morbidity, reduced productivity of people, and increased sick leaves (Ebi et al., 2021; Wondmagegn et al., 2021) is necessary to bring a new perspective that addresses these challenges. For that, this study aims to develop an innovative approach, based on a methodological decision-to-visualization process that bears into consideration the impact of bioclimatic and socioeconomic variables as any other medical variable as part of the decision making process of a worker health status and associated occupational risks.

In practice, this research aims to improve the projections of occupational health risk factors and the characterization of the health status, which is the target node of the model, exploiting the combination of Bayesian machine learning and spatial techniques. In that manner, the added value is the possibility to identify and characterise those variables that may have a differentiating impact that apparently is not meaningful from a mathematical point of view. All in all, the results of this research work are expected to be one more contribution towards the medical services of the future, where the patient health status will not be any more subject to only a series of traditional medical tests and underlying medical conditions (Awotunde et al., 2021).

2. Material and Methods

2.1. Data characterization

A total of 74,401 occupational health surveillance tests gathered from workers belonging to the service sector in the period between 2012-2016 throughout the Spain territory were used as a medical data source for this study. More specifically, the workers for this research database carried out activities related to administrative and auxiliary services (31,894), financial and insurance services (12,958), education (13,938), and hostelry (15,611). Each clinical examination was undertaken according to the Spanish occupational health legislation (Ley 31/1995). Relevant occupational health organizations and hospital services conducted the medical tests gathering major information about the state of workers' health defining the main health risk factors causing pathologies, including the main physical conditions and health habits.

This study goes a step beyond traditional occupational health surveillance analyses, adding to the medical record of each worker a cross prediction with climatic and socioeconomic factors as an instrument to better characterize and predict those factors disrupting the health status. For that, *Maximum Temperature (BIO5)* or *Annual Rainfall (BIO12)*, and *Unemployment Rate or GDP* are examples of the bioclimatic and socioeconomic variables used respectively. Procedurally, this research is conducted in four levels. First, from the 37 initial variables considered, a total of 26 were finally taken for modeling purposes after reducing the problem dimension (Level 1). In the second level of analysis, these reduced variables were used to characterize the four main groups of service activities (Level 2). Later, for each activity group, the health status acts as a target node for which the relevant patterns are ascertained (Level 3). Figure 1 provides a scheme of this methodological process applied. These three levels presented correspond to the development of a Bayesian methodology that is complemented with a geostatistical analysis (Level 4) for those parts of the network where further clarity is needed.

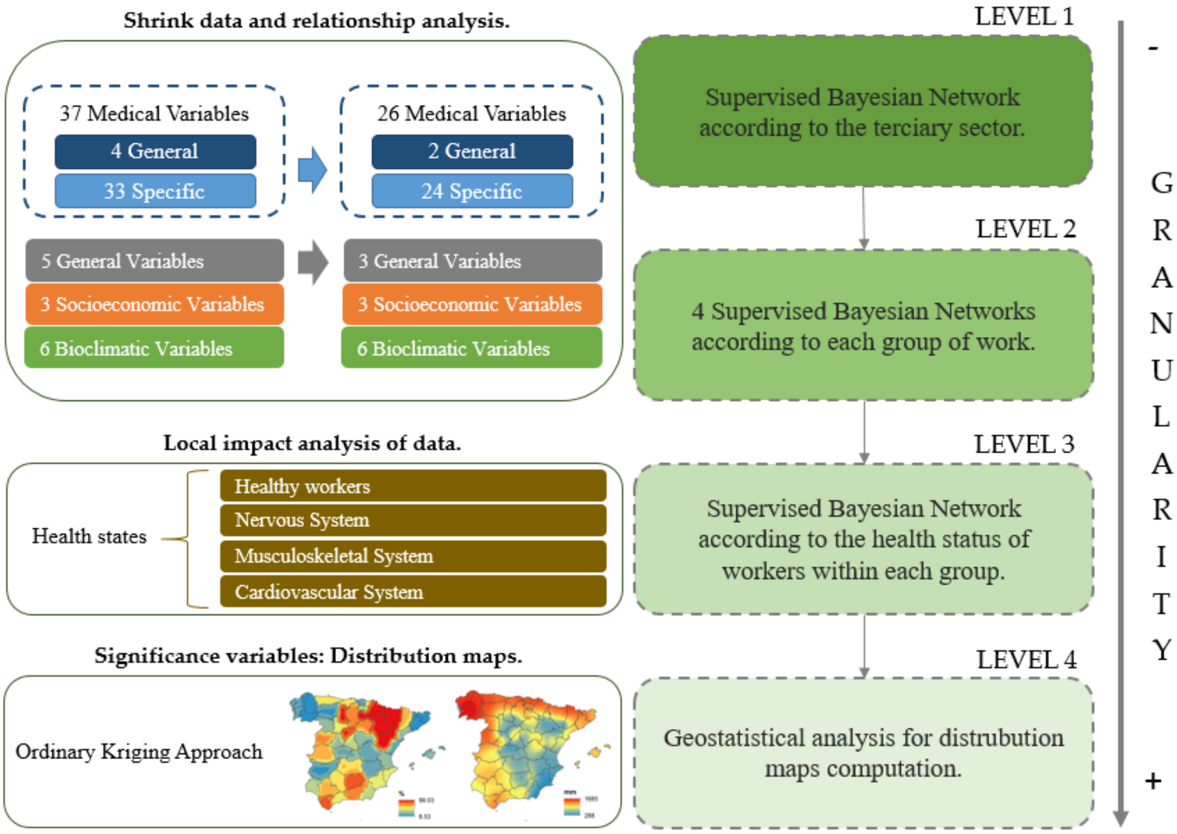

Shrink data and relationship analysis.
37 Medical Variables
4 General
33 Specific
26 Medical Variables
2 General
24 Specific
5 General Variables
3 Socioeconomic Variables
6 Bioclimatic Variables
3 General Variables
3 Socioeconomic Variables
6 Bioclimatic Variables
Local impact analysis of data.
Health states
Healthy workers
Nervous System
Musculoskeletal System
Cardiovascular System
Significance variables: Distribution maps.
Ordinary Kriging Approach
LEVEL 1
Supervised Bayesian Network according to the terciary sector.
LEVEL 2
4 Supervised Bayesian Networks according to each group of work.
LEVEL 3
Supervised Bayesian Network according to the health status of workers within each group.
LEVEL 4
Geostatistical analysis for distrubution maps computation.
-
GRANULARITY
+

Figure 1 – The implemented methodological process with four levels of analysis.

As anticipated, for those occupational health variables that cannot be further discriminated in the network, a higher level of granularity in the analysis is provided by carrying out a geostatistical ordinary kriging approach as an effective solution to extract relevant patterns and produce reliable health risk maps. Ordinary kriging is the most widespread method of kriging. It serves to estimate a value at a point of a region for which a variogram is known, using neighboring data to the estimation of an unknown location (Goovaerts, 1997). This approach allows focusing the spatial representation on those variables with a more differentiating nature across the four different groups of service activities under analysis. Spatial interpolations were carried out by means of Geostatistical Wizard module in ArcGIS v-10.2.2. Semivariograms were manually adjusted assuming spatial isotropy in the search of preferential directions.

2.2. Supervised machine learning techniques for target characterization

Recent advances in computer science offer the possibility to couple machine learning with traditional statistical methods such as Bayesian networks (Benavoli et al., 2017). Bayesian networks have shown their potential in problem domains with manifold variables of different typologies, where the medical and occupational health domain is a showcase of their performance. Concretely, information theory in combination with Bayesian networks is used to respond to the different stages of this study, quantifying the reduction of uncertainty brought by each medical variable to the knowledge of the health state.

On this basis, once the Bayesian model is built as a result of the machine learning process aimed to discover significant relationships in the problem space search, the Kullback-Leibler (KL) divergence is used as a measure of strength in the relationship between two nodes that are directly connected by an arc (Conrady et al., 2015). This parameter allows measuring how the probability distribution in each variable drifts away from the state of health (target node). From a mathematical viewpoint, let P and Q represent the distribution of two joint probabilities defined for the same set of variables or X nodes.

$$D_{KL}\big(P(X)||Q(X)\big) = \sum_{X} P(X)log_2 \frac{P(X)}{Q(X)} \quad (1)$$

For target node characterization, the relative weight value is shown as a fraction of the maximum KL Divergence value. Likewise, these weights can be depicted as the global contribution percentage of each arc to the target node quantifying the value between two directly connected nodes D_{KL}(Parent|Child) and the sum of all KL Divergence values across the network. In addition, the independence test G is computed from the KL divergence of the relationship, thus its value is reckoned from the network.

3. Results

Given the need to clarify the understanding of occupational health risks triggering workers' sick leave, this section presents the preliminary results obtained from the application of Bayesian machine learning and geostatistics to the occupational health data for the four service activities under analysis. The results outline the findings obtained based on the four methodological levels summarised in Figure 1.

In the first place, a general Bayesian network was built delving into the statistical association stemming from the state of health and each variable in the model, considering all service activities (administrative and auxiliary services, financial and insurance services, education, and hostelry). From the resulting Bayesian network, a relationship analysis was carried out. The more representative parent-child connections were identified. These relationships are shown in Table 1, where *age* excels by its high impact, followed by the *location* and the *total cholesterol*.

Table 1 – Characterization of the target node (health state). Relationship analysis for the most representative medical, socioeconomic and bioclimatic variables.

Parent	Child	KL(Parent|Child)	Relative weight	Contribution	G Test
Health state	Age	0.0521	1.0000	16.6440%	5,374.1635
Health state	Location	0.0341	0.6552	10.9056%	3,521.2995
Health state	Total Cholesterol	0.0287	0.5504	9.1604%	2,957.8012
Health state	Drug Consumption	0.0213	0.4084	6.7969%	2,194.6465
Health state	Hearing test	0.0169	0.3238	5.3896%	1,740.2371
Health state	Spine Observation	0.0149	0.2852	4.7465%	1,532.5901
Health state	Limbs Observation	0.0145	0.2789	4.6413%	1,498.6285
Health state	Physical Limitations	0.0126	0.2416	4.0214%	1,298.4567
Health state	Minimum Rainfall	0.0100	0.1922	3.1997%	1,033.1395
Health state	Population	0.0074	0.1417	2.3587%	761.5890
Health state	Annual Rainfall	0.0073	0.1410	2.3476%	758.0027
Health state	Sleep Quality	0.0068	0.1312	2.1836%	705.0538
Health state	Maximum Temperature	0.0067	0.1280	2.1306%	687.9423

In the second place, four supervised Bayesian networks were built, corresponding to each of the four defined service activities and whose common target node was the health status of the worker. The application of a Naïve Bayes algorithm allowed to generate a pragmatic network structure for the analysis of the influence of each variable on the health status of the workers (Figure 2). The characterization of the target node revealed that age, location, and total cholesterol, previously identified as the most significant factors in the general network of the service sector, also present a high impact on all the concrete service activities under study. In that context, the authors have considered the need to deepen the understanding of those variables that are a priori not that significant, but which may hold key differentiating aspects within each population group.

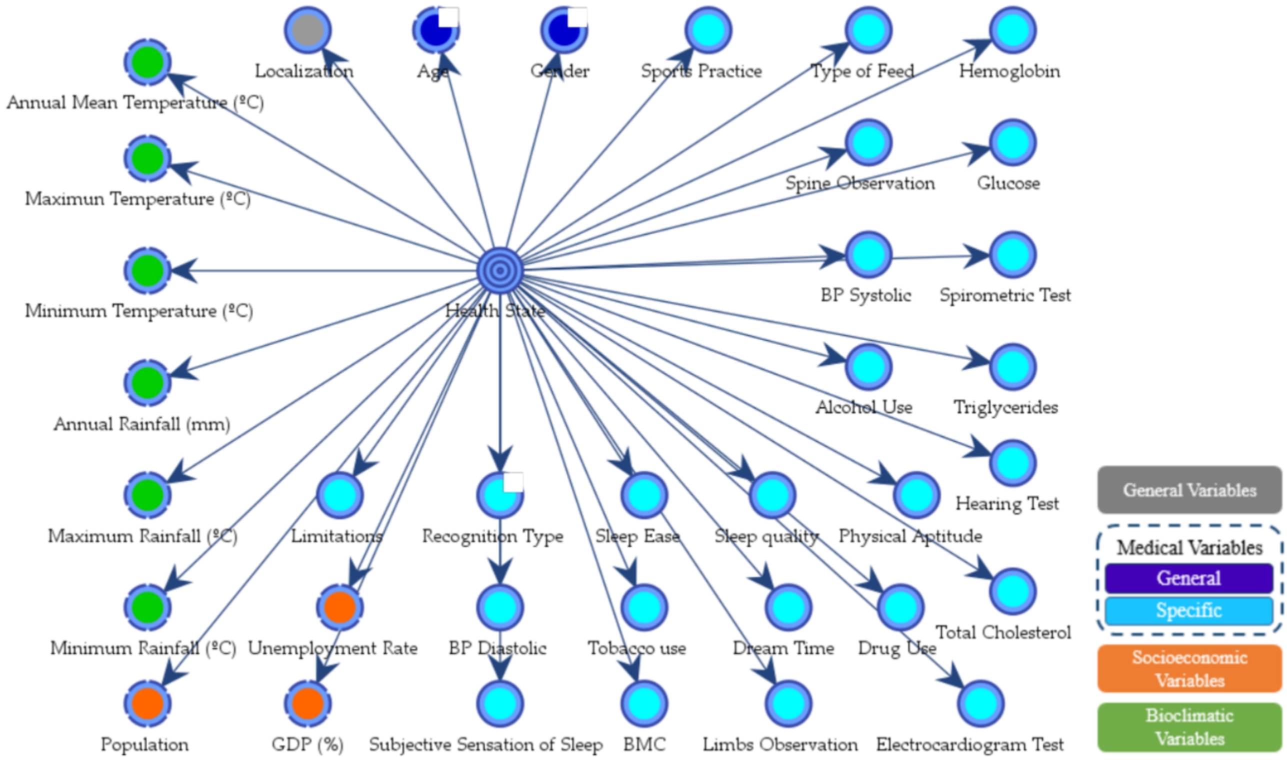


Figure 2 – Supervised Bayesian network built with Naïve Bayes algorithm. The graph presents the administrative and auxiliary services network with the target node (health state) in the center.

When looking at the distribution of contributions of each variable to the characterization of the state of health, it is found that the nervous system (15%-19%) matches to a high extent the characterization of the medical examinations of healthy workers (64%-70%). The most significant medical conditions conditioning these two states are age, total cholesterol, and location, whereas hearing problems and drug use are always reflected as differential variables. As an example, after an inference analysis on patients with high levels of total cholesterol belonging to hostelry services, a greater impact could be seen on elderly workers (>50) belonging to the autonomous community of the Basque Country (38.26 % of registered cases) located in the North of Spain. In contrast, it can be concluded the strong need to provide a higher level of granularity on the musculoskeletal (8%-11%) and cardiovascular (6%-9%) pathologies, as here the differences among possible additional differential variables, even if relevant from a mathematical point of view, they are not meaningful from a policy perspective (Table 2).

The great horizontality of variables such as *age, location,* and *total cholesterol* directed this study towards the need to add value to those differentiating variables of the musculoskeletal and cardiovascular systems. This situation leads to the spatial representation of the variable's *spine observation, annual precipitation (BIO 12), limbs observation, and annual mean temperature (BIO 1)* under an ordinary kriging approach (Figure 3). This approach allows identifying both a spatial distribution of spinal problems and potentially related extremities, as well as two clearly differentiated regions where these problems have a higher impact. Particularly, in the Northeast of Spain, except Catalonia, and the South, with vascular problems such as the presence of varicose veins. As for the Western part of Spain, a higher rate of spinal problems, derived from muscle contractures or other minor discomforts, is identified. Based on Bayesian results, it can be demonstrated that this type of injury is related to a great extent to pathologies of the musculoskeletal system which is potentially present in service activities such as hospitality (31.23%) and administration (32.05%). In addition, it is also seen how these pathologies are also an underlying cause for the appearance of problems in the extremities.

Table 2 – Local impact analysis of data over the target node for the states representing musculoskeletal and cardiovascular systems by service activity.

	Musculoskeletal System		**Cardiovascular System**	
	Relative Binary Mutual Information		Relative Binary Mutual Information	
Group of Work	Spine Observation	Annual Rainfall	Limbs Observation	Annual Temperature
Administrative and auxiliary services	3.4737%	0.64%	1.7552%	1.3914%
Financial and insurance activities	2.2066%	0.5638%	1.5169%	0.9984%
Education	3.6072%	0.7496%	0.9042%	0.4487%
Hostelry	3.2213%	1.3584%	2.1152%	0.0360%

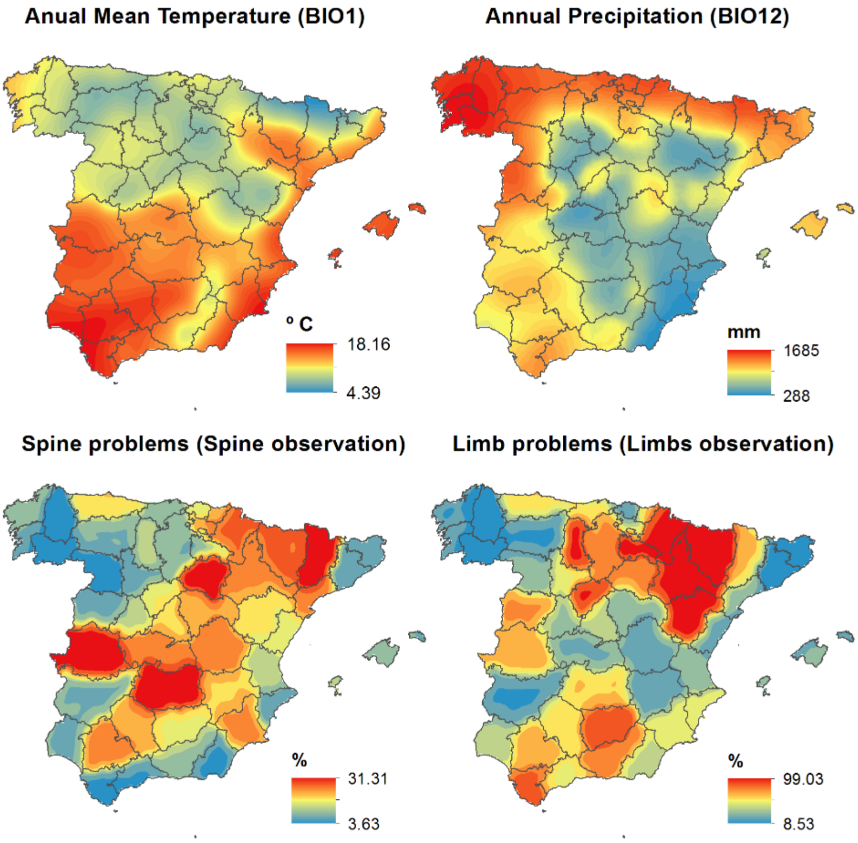

Anual Mean Temperature (BIO1)
Annual Precipitation (BIO12)
° C
18.16
4.39
mm
1685
288
Spine problems (Spine observation)
Limb problems (Limbs observation)
%
31.31
3.63
%
99.03
8.53

Figure 3 – Distribution maps for annual mean temperature (ºC), annual rainfall (mm), and spine and limb observation variables using rate data between 2012 and 2016 interpolated by Ordinary Kriging.

4. Discussion and Conclusions

Given the greater complexity of characterizing the musculoskeletal and cardiovascular systems, and the impossibility of achieving greater conceptual discrimination of the detailed variables with Bayesian networks, a new level of granularity is needed. Here, an ordinary kriging approach enters into play, offering the possibility to differentiate and obtain meaningful policy findings at a regional scale, revealing what are the exact implications, above all, of the bioclimatic variables and how they affect unhealthy workers within the service sector.

As a showcase of the potential of this combined approach, Figure 4 shows the Bayesian results of the inference analysis on the medical variables *spine* and *limbs observation*, and the bioclimatic variables *annual mean temperature (ºC) (BIO1)* and *annual rainfall (mm) (BIO12).* This inference is carried out with two variables of reference that are the service sector activities (group of work) and the state of health (pathology). In general, the results allowed to conclude a greater impact of spinal problems for hostelry activities, as well as a direct relationship of this pathology with limb problems, increasing the cases of workers with adverse vascular conditions by 19.64%. This is an example, which shall be complemented always with a more granular spatial mapping to deepen on the regional variations.

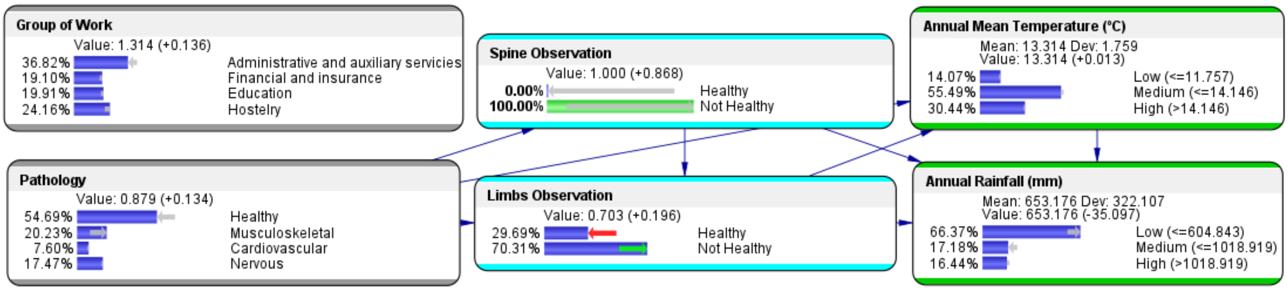


Figure 4 – Inference results for the work groups when the evidence reflects spinal problems.

In conclusion, the results of this study revealed that variables such as age, location, and cholesterol, with contributions to the general network between 9-17%, are generally critical for the characterization of the health status of workers in the service sector. To a second extent, it was possible to identify a series of differentiating variables such as pine observation, annual precipitation (BIO 12), limbs observation, and annual mean temperature (BIO 1) that despite not being extremely significant from a mathematical point of view, they play a key role and show a great impact at regional level.

Likewise, this article exposes the potentialities of the combination of Bayesian machine learning complemented by geostatistics to translate the complex occupational health problem of workers' health status into evident visual findings that can feed medical policy developments across different service activities. At this stage, it is already possible to demonstrate the high influence of bioclimatic and socioeconomic variables within the medical decision making of a worker health state. Looking forward, further analysis is needed to identify more health risk factors that can be derived, for example, from the impact of high temperatures or income level. In this respect, depending on the data available and the scope of the analysis, more sophisticated geostatistical approaches would have also to be explored.

References

A. Benavoli, A., G. Corani, J. Demsar, M. Zaffalom. Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis. Journal of Machine Learning Research. 2017.

A. Orlov, J. Sillmann, K. Aunan, T. Kjellstrom, and A. Aaheim, "Economic costs of heat-induced reductions in worker productivity due to global warming," Glob. Environ. Chang., vol. 63, no. September 2019, p. 102087, 2020, doi: 10.1016/j.gloenvcha.2020.102087.

B. Y. Wondmagegn et al., "Increasing impacts of temperature on hospital admissions, length of stay, and related healthcare costs in the context of climate change in Adelaide, South Australia," Sci. Total Environ., vol. 773, p. 145656, Jun. 2021, doi: 10.1016/J.SCITOTENV.2021.145656.

C.-H. Chang, R. Shao, M. Wang, and N. M. Baker, "Workplace Interventions in Response to COVID-19: an Occupational Health Psychology Perspective," *Occup. Heal. Sci.*, vol. 5, no. 1–2, pp. 1–23, Mar. 2021, doi: 10.1007/S41542-021-00080-X/TABLES/1.

Eurostat, "Contributions of each sector - Institutional sector accounts - Eurostat". European Commission, 2022. https://ec.europa.eu/eurostat/web/sector-accounts/detailed-charts/contributions-sectors

J. B. Awotunde, A. E. Adeniyi, R. O. Ogundokun, G. J. Ajamu, and P. O. Adebayo, "MIoT-Based Big Data Analytics Architecture, Opportunities and Challenges for Enhanced Telemedicine Systems," *Stud. Fuzziness Soft Comput.*, vol. 410, pp. 199–220, 2021, doi: 10.1007/978-3-030-70111-6_10.

K. L. Ebi et al., "Extreme Weather and Climate Change: Population Health and Health System Implications," https://doi.org/10.1146/annurev-publhealth-012420-105026, vol. 42, pp. 293–315, Apr. 2021, doi: 10.1146/ANNUREV-PUBLHEALTH-012420-105026.

P. Goovaerts. Geostatistics for Natural Resources Evaluation. Applied Geostatistics Series. Oxford University Press, New York, NY (USA), 837 483 p., 1997.

S. Conrady, L. Jouffe. Bayesian Networks & BayesiaLab - A Practical Introduction for Researchers. Bayesia USA. 2015. ISBN-10: 0996533303.

S. Gerassis, C. Boente, M.T.D. Albuquerque, M.M. Ribeiro, A. Abad, J. Taboada, "Mapping occupational health risk factors in the primary sector—A novel supervised machine learning and Area-to-Point Poisson kriging approach" *Spatial Statistics*, vol. 42, 100434, 2021.

FUNCTIONAL DATA ANALYSES TO MODEL COVID-19 WAVES

Maria João Pereira (1)* - Leonardo Azevedo (1) - Manuel Ribeiro (1) - Amilcar Soares (1)

CERENA, Instituto Superior Tecnico, Universidade de Lisboa, Lisboa, Portugal (1)
** Corresponding author: maria.pereira@tecnico.ulisboa.pt*

Abstract

Since its outbreak, the SARS-CoV-2 pandemic has been showing complex dynamics in both time and space. Over two years of pandemic different strategies and polices have been adopted by countries to control and mitigate the impacts of the disease propagation. But there is still much to understand and lessons to learn, about how human behaviour, effectiveness of vaccines over time, infection prevention policies, changes of coronavirus itself and the number of people who are vulnerable controlled the several COVID-19 waves in each country. In this work we analysed the spatiotemporal patterns of the 5 COVID-19 waves in Portugal using geostatistical functional data analysis.

The daily number of infection data by municipality reported by the Portuguese Directorate-General for Health are used to build time series of infection since the beginning of the outbreak in Portugal. We divided the time series in 5 sub-sets corresponding to the 5 waves. We employ a dimensionality reduction of these curves using functional principal component analysis. The objective of this step is twofold, detect municipalities with similar temporal evolution and get a small number of coefficients to describe the temporal pattern of the series. The low-dimension coefficients were used cluster municipalities with similar behaviour. Results for the different waves were analysed and compared giving new insights about data and allowing to set up new hypothesis about disease spread.

SPATIO-TEMPORAL POINT PROCESS COMPARTMENT MODELING FOR INFECTIOUS DISEASES

André Victor Ribeiro Amaral (1)* - Jonatan A. González (1) - Paula Moraga (1)

King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia (1)
** Corresponding author: andre.ribeiroamaral@kaust.edu.sa*

Abstract

Infectious diseases such as COVID-19 may have a significant impact on society, overloading health systems and affecting individuals' lives at different levels. For example, they can lead to job losses, increase inequalities and cause mental health problems. Therefore, it is crucial to understand how these diseases may spread and how decision-makers may act to prevent them. In this context, different mathematical models have been proposed to describe how the number of infected cases evolves over time. An often-popular choice is the SIR compartment model, in which individuals are assigned to three different groups, namely Susceptible (S), Infected (I), and Recovered (R), and are modeled according to a system of Ordinary Differential Equations (ODEs). Still, deterministic approaches, especially when the population size is not sufficiently large, may not correctly describe the phenomena of interest, and models with stochastic components may perform better in explaining the disease-spreading dynamics. In this work, we reinterpret the SIR model and write it as a system of Stochastic Differential Equations (SDEs) with respect to time, describing the intensity functions of spatio-temporal point processes for the occurrence of susceptible, infected, and recovered individuals in the studied area. In particular, the introduced randomness aims to account for the uncertainty on the newly infected individuals. For such a system, we propose a numerical solution, and describe the spatial dependence through a Cox process. To conclude, our work approaches a common problem in epidemiology from a different point of view, namely point processes in space and time, which may bring new insights into how infectious diseases can be modeled.

GEOSPATIAL MODELING OF NATIONAL HEALTH SERVICE DELIVERY SURVEY DATA

Eun-Hye Enki Yoo (1) - John E. Roberts (2) - Becky Powell (3) - Tia Palmero (4) - Qiang Pu (1)

The State University of New York at Buffalo, Geography, Buffalo, United States (1) - The State University of New York at Buffalo, Psychology, Buffalo, United States (2) - University of Denver, Geography, Denver, United States (3) - The State University of New York At Buffalo, Epidemiology and Environmental Health, Buffalo, United States (4)
** Corresponding author: eunhye@buffalo.edu*

Abstract

There is an urgent need for ecological approaches to assess accessibility to healthcare services for adolescents in low- and middle-income countries (LMIC; Mark 2013). The Demographic and Health Surveys (DHS) program is a principal source of data on the provision of health services in LMICs, from which ecological studies can be conducted. In particular, the DHS has collected data from a survey on individual households and another on health facilities and service delivery environment in LMICs. The present study focuses on the latter, referred to as Service Provision Assessment (SPA) surveys, which provides the information about the characteristics of health facilities and services available in individual country. In recent years, SPAs also collected geographic information of health facilities that participated the survey, although the spatial coverage of currently available data is limited given that they are based on samples rather than census. To improve data availability, we developed multiple spatial interpolation methods to estimate key information on health service availability and service delivery environments using both geostatistics (e.g., Kriging with external drift) and machine learning algorithms (e.g., stacked ensemble model). Prior to the model development, we identified key geospatial covariates that represent the service delivery environment of health facilities. Here, we explored machine learning algorithms to capture potentially complex interactions and non-linear effects among geospatial covariates, and geostatistical interpolation to account for spatial structure (i.e., spatial correlation) of health service availability in the study area. The model prediction surfaces generated from multiple interpolation methods were further aggregated at subnational administrative level 2 (i.e., small scale administrative boundary). For geostatistical models, the uncertainty associated with estimates of health service availability was quantified. Model performance was evaluated following Yoo et. al. (2021) using two criteria—prediction accuracy and classification error. The performance evaluation of the geostatistical linkage method, demonstrated using information on the general service readiness of sampled health facilities in Tanzania, showed that geostatistical methods and machine learning methods are comparable in terms of both prediction accuracy and classification error. However, we also found that the results of machine learning models vary spatially, which is explained by the uncertainties in the individual model algorithm. Among the machine learning algorithms, the use of an ensemble model approach seems more adequate than relying on predictions from any single modeling method. The proposed geospatial approach minimizes the methodological issues and has potential to be used in various public health research applications where facility and population-based data need to be combined at fine spatial scale. Particularly, we expect that the prediction uncertainty from geostatistical models will be useful to establish reliable linkages between DHS household surveys (i.e., indicators for population, health, and nutrition) and SPA surveys (i.e., service availability and delivery environments).

References

Mark, T. "Adolescence: a second chance to tackle inequities." Lancet Infectious Diseases 382.9904 (2013): 1535.

Yoo, Eun-Hye, Tia Palermo, and Stephen Maluka. "Geostatistical linkage of national demographic and health survey data: a case study of Tanzania." Population health metrics 19.1 (2021): 1-14.

GEOSTATISTICAL COVID-19 RISK AND UNCERTAINTY IN A SINGLE MAP WITH R OPEN-SOURCE CODE

Manuel Ribeiro (1)* - Leonardo Azevedo (1) - Amilcar Soares (1) - Maria João Pereira (1)

Centro de Recursos Naturais e Ambiente, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal (1)
** Corresponding author: manuel.ribeiro@tecnico.ulisboa.pt*

Abstract

With the emergence of COVID-19 pandemic in Portugal, a Block Direct Sequential Simulation (Block DSS) tool has been developed to model COVID-19 spread in mainland Portugal (Azevedo et al., 2020) and support decision-making by health authorities and policy makers. Besides mapping the risk, the model includes the assessment of spatial uncertainty of estimates. Yet uncertainty is difficult to be visualized with the estimated risks and it is usually disregarded as a tool to support decision-making process. This can be misleading since the extent of risk uncertainty varies throughout the spatial domain.

To overcome this problem, an R package performing pixelation (Taylor et al., 2020) has been proposed to visualize uncertainty in maps of disease risk into a single map. The application is illustrated to map 2017 P. falciparum (a parasite causing malaria in humans) incidence in central Africa as proof of concept. The proposed solution provides disease risks maps with varying pixel size such that areas of high average uncertainty have large pixels, while areas with low average uncertainty have small pixels. This means that in areas where uncertainty is higher, the risk is smoothed over a larger area by performing pixelation of uncertainty (i.e., by increasing the pixel size). The pixelated map summarizes effectively in one single map the two key elements required in disease risk mapping and provides policy makers with a decision-support tool capable of rapidly identify high risks in areas with high spatial uncertainty and high risks in areas with low spatial uncertainty preventing fine-scale inference in regions with high-risk uncertainty.

While Block DSS algorithm is implemented as a stand-alone software tool and distributed free of charge, it is not open source, and its use can be cumbersome. Therefore, programming code solutions combining analysis of disease mapping based on the modelling approach proposed by Azevedo and colleagues (Azevedo et al., 2020) with the pixelation approach to visualize uncertainty in maps of disease risk are very limited, require a considerable programming effort and a high level of expertise.

To address this gap, we present a complete open-source R code for the rapid computation and visualization of uncertainty in maps of disease risk in a single map, based on the COVID-19 geostatistical modelling approach proposed by Azevedo and colleagues (Azevedo et al., 2020). The resulting map can be a valuable tool to help policymakers to make informed decisions about COVID-19 pandemic.

Bibliography
Azevedo, L., Pereira, M.J., Ribeiro, M.C., Soares, A., 2020. Geostatistical COVID-19 infection risk maps for Portugal. Int. J. Health Geogr. 19, 1–8. https://doi.org/10.1186/s12942-020-00221-5.
Taylor, A.R., Watson, J.A., Buckee, C.O., 2020. Pixelate to communicate: visualising uncertainty in maps of disease risk and other spatial continua. arXiv:2005.11993 [stat.AP].

MACHINE LEARNING-BASED INVERSE MODELING FOR THE IDENTIFICATION OF HYDRAULIC CONDUCTIVITY

Vanessa A. Godoy (1)* - Gian F. Napa-Garcia (1) - J. Jaime Gómez-Hernández (1)

Universitat Politècnica de València, Valencia, Spain (1)
** Corresponding author: godoyalmeida@gmail.com*

Abstract

Identifying parameters based on state variables is a well-known task in groundwater modeling. In the last decades, many works have focused on the assimilation of observed data (in hydrogeology, the hydraulic head) to identify model parameters heterogeneity (in hydrogeology, the hydraulic conductivity) by using variants of the ensemble Kalman filter (EnKF). In the EnKF, the assimilation is based on a prediction of how the system will progress, followed by a correction of the parameters based on the discrepancy between predictions and observations. The correction is performed based on a linear interpolation that can only capture the linear component of the non-linear relationship between conductivities and piezometric heads. In contrast, machine learning algorithms, besides having permeated all ambits of science and technology today, are known for capturing the relationship between dependent and independent variables, whether linear or not. Although these algorithms have proven their ability to replace process-driven models with data-driven ones to predict piezometric heads or solute concentrations from ancillary variables, they are seldom used for inverse modeling purposes. In this work, we propose to couple machine learning algorithms with the well-established EnKF to perform stochastic inverse modeling. The plan is to take advantage of the ensemble of realizations to train machine learning algorithms and to use it to perform a non-linear correction, which should give better results than the one obtained with the EnKF. The validity of the proposed method was demonstrated by applying it in a synthetical example and, for the sake of completeness, the results were compared to the results obtained by using the EnKF. The proposed method not only has appropriately characterized the patterns of spatial variability of the reference fields and reduced the uncertainty but did it by using a small number of realizations and historical data. Despite some limitations related to the difficulty about the definition of parameters of the machine learning algorithms, the proposed method has proved to be a powerful tool to characterize subsurface heterogeneity.

SEISMIC OCEANOGRAPHY IMAGING AND INVERSION OF THE MADEIRA ABYSSAL PLAIN

Ana Filipa Duarte (1)* - Leonardo Azevedo (1) - Luis Matias (2) - Álvaro Peliz (2) - Renato Mendes (3)

CERENA, DECivil, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal (1) - Instituto Dom Luiz, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal (2) - Collaborative Laboratory, +Atlantic, Matosinhos, Portugal (3)

** Corresponding author: filipadamasoduarte@tecnico.ulisboa.pt*

Abstract

Seismic oceanography is a multidisciplinary research field that provides insights about ocean processes happening at large- to small-scales, leveraging already available common marine multichannel seismic reflection data (MCS) to create high-resolution images of the structure of the ocean (Biescas et al., 2008).

This work exemplifies the potential of seismic oceanography data to study the ocean by processing and inverting a set of three parallel two-dimensional MCS profiles acquired in the Madeira Abyssal Plain (Northeast Atlantic) during June 2006.

The MCS data undergoes a series of processing steps aiming to image: i) the structure of the water column close to the sea surface, by attenuating the direct wave arrivals; ii) and preserving the true seismic amplitudes required to invert the seismic data for the physical properties of the ocean (i.e., ocean temperature and salinity). The processed seismic oceanography profiles clearly identify two layers: the top layer, above approximately 2000 m, has bright and coherent reflection content, while the bottom layer, below this depth, is reflection-free as expected for the homogeneous North Atlantic Deep Waters (Segade et al., 2015). The top layer comprises several features of interest namely, eddies at the expected Mediterranean Outflow Water depths, steeply dipping reflectors, which indicate the possible presence of frontal activity or secondary dipping eddy structures and strong horizontal reflections between intermediate water masses suggestive of double diffuse processes.

While the structural interpretation of the observed seismic reflections provides valuable oceanographic insights, the ability to predict the ocean temperature and salinity allows a deeper understanding of these processes. We show the application of a geostatistical seismic oceanography inversion methodology to these data and its spatial interpolation in three-dimensions. The geostatistical inversion combines information from different sources (i.e., direct and indirect measurements) about the temperature and salinity of the ocean. The inverted models reproduce the overall vertical distribution of both properties, as interpreted from the global ocean models. Also, these models capture the oceanic features of interest at a finer scale, revealing the filamentation structures around the Eddie's core and, specifically, the warm intrusions around the homogenous nucleus.

SOLVING GEOPHYSICAL RANDOM EFFECT MODELS WITH INTRACTABLE LIKELIHOODS: LINEARIZED GAUSSIAN APPROXIMATIONS VERSUS THE CORRELATED PSEUDO-MARGINAL METHOD

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

University of Lausanne, Institute of Earth Sciences, Lausanne, Switzerland (1) - University of Bern, Institute of Mathematical Statistics and Actuarial Science and Oeschger Center for Climate Change Research, Bern, Switzerland (2)

** Corresponding author: lea.friedli@unil.ch*

Abstract

We consider a Bayesian inversion problem in which the posterior distribution of (hydro)geological parameters of interest is inferred from geophysical data. A critical aspect in this setting is the noisy petrophysical relationship linking the hydrogeological target parameters to the geophysical properties sensed by the geophysical measurements. To account for the uncertainty resulting from this noisy petrophysical relationship, the intermediate geophysical properties are treated as latent (unobservable) variables. To perform an inversion in such a random effect model, the intractable likelihood of the (hydro)geological parameters given the geophysical data needs to be estimated. We aim to achieve this by approximating the likelihood with a Gaussian probability density function based on local linearization of the geophysical forward operator. This allows including the effect of the noise in the petrophysical relationship by a corresponding inflation of the data covariance matrix. The new approximate method is compared against the general correlated pseudo-marginal (CPM) method estimating the likelihood by Monte Carlo averaging over samples of the latent variable. We compare the performances of the two methods on synthetic test examples, in which we infer for porosity using crosshole ground-penetrating radar (GPR) first-arrival travel times. By varying the non-linearity and levels of noise in the petrophysical relationship, we elaborate recommendations for the choice of the appropriate method depending on the problem under consideration. The linearized Gaussian approach, while attractive due to its relative computational speed, suffers from a decreasing accuracy with increasing noise in the petrophysical relationship and/or non-linearity. The computationally more expensive CPM method, by comparison, performs very well even for strongly non-linear settings with high amounts of noise in the petrophysical relationship.

GEOSTATISTICAL INVERSION FOR ROCK PHYSICS PROPERTIES IN THE CRITICAL ZONE

Dario Grana (1)* - Andrew Parsekian (1)

University of Wyoming, Laramie, United States (1)
** Corresponding author: dgrana@uwyo.edu*

Abstract

Understanding the physical processes in the critical zone requires accurate predictions of the spatial distribution of rock and fluid properties, such as porosity and fluid (water and air) saturations. On mountain hillslopes, snow precipitation infiltrates the subsurface and recharges groundwater aquifers. The spatial distribution of the water volume depends on the porosity of the rocks and on the fluid saturations in the pore space. These petrophysical properties of porous rocks can be predicted from surface geophysical data, such as seismic refraction and time-lapse electrical resistivity tomography. This modeling problem can be formulated as a geophysical inverse problem. We present a Bayesian inversion approach based on the ensemble smoother algorithm to generate multiple realizations of porosity and water saturation conditioned on geophysical properties, specifically P-wave velocity of seismic waves and electrical resistivity. The model realizations are generated using a geostatistical algorithm, for example the probability field simulation, and updated using the ensemble smoother algorithm. The prior distribution includes a spatial correlation function such that the model realizations mimic the geological spatial continuity. 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 Archie's equation for the electrical component. The model accounts for pressure and mineralogy changes in depth. The result of the inversion includes a set of realizations of porosity and water saturation, that are used to infer most likely model and its uncertainty. The posterior uncertainty is analyzed in a low-dimensional space using multi-dimensional scaling. The results are compared to those of traditional Bayesian inversion methods and the predictions of the proposed method show a higher level of accuracy than traditional inversion algorithms. The proposed approach is tested on synthetic data and applied to a real geophysical dataset measured along a 60 m section of mountain hillslope near Laramie, Wyoming, USA. The results are validated using direct observations of porosity and water saturation from core samples and borehole measurements. The so-obtained porosity and water saturation maps are used to predict the spatial distribution of the subsurface water produced from snowpack melting that flows and is stored in mountain watersheds. These results can reduce the uncertainty in hydrological model predictions and can be used to make more informed decisions on the water management.

ON A CONSTRUCTIVE SPECTRAL METHOD FOR CONDITIONING PLURIGAUSIAN SIMULATIONS TO BOREHOLES OBSERVATIONS AND INDIRECT DATA. APPLICATION TO AQUIFER MODELS

Dany Lauzon (1)* - Denis Marcotte (1)

Polytechnique Montréal, Civil, Geological and Mining Engineering Department, Montréal, Canada (1)
** Corresponding author: dany.lauzon@polymtl.ca*

Abstract

This presentation combines the S-STBM method with the pluriGaussian simulations for the conditioning of indirect data when modeling categorical hydrogeological systems. It emphasizes the advantages of the sequential construction of a facies field using S-STBM as the calibration method. The boreholes conditioning is done quickly by replacing the slow Gibbs sampler method with an approach based on the calibration of latent Gaussian fields subject to inequality constraints. We use a novel approach based on parametrization to reduce the optimization space, generally multivariate with the pluriGaussian simulations, to a unidimensional space. The methodology is exposed, two case studies are presented, and future applications envisaged are mentioned such as the simultaneous calibration of facies units and hydrogeological properties.

Keywords: Inverse problems; PluriGaussian simulations; Constructive calibration; Spectral simulations; Gradual deformation.

1. Introduction

Geostatistical simulations provide non-unique models to illustrate uncertainties and assess environmental risks associated with a geological model (Chilès and Delfiner, 2012). For categorical fields, a commonly used approach covering a wide variety of geological phenomena for groundwater aquifers is the pluriGaussian simulations (PGS) where one or many latent Gaussian random fields (L-GRF) are truncated to assess the categorical field (Armstrong et al., 2011). However, conditioning these models to indirect data (e.g., hydraulic heads, drawdowns, first arrival times between wells, proportion maps, or tracer tests) is a tedious step due to the non-linear relationship between the indirect data, the geological properties, and the discrete nature of the geological facies. Optimization methods are one of the most widely used approaches to best calibrate the response of an aquifer to available indirect observations (Hu et al., 2001; Lauzon and Marcotte, 2022).

A recently developed approach, the sequential spectral turning band method (S-STBM), consists of building from zero the aquifer model constrained to indirect observations like an engineer building a plane while flying it (Lauzon and Marcotte, 2020). The method consists of sequentially adjusting each band to minimize the error between the observed data and the simulated data. The S-STBM has proven itself in both continuous and discrete domains (especially when combined with truncated and pluriGaussian simulations) for the calibration of first arrival travel times between wells, the conditioning of boreholes data, and the calibration of pressure heads (Lauzon and Marcotte, 2020a, 2020b, 2022). For categorical problems, the recent study by Lauzon and Marcotte (2022) showed that S-STBM is better suited to calibrate indirect data on categorical problems than usual calibration methods such as gradual deformation, and iterative spatial resampling.

This article seeks to apply the S-STBM method combined with PGS methods to build efficiently categorical fields calibrated to indirect data applied to aquifer models. We first describe the S-STBM approach, present how to perform fast conditioning of categorical data using S-STBM instead of the well-known Gibbs sampler, and inform about PGS simulations. The novel approach presents in this paper concerns the phase vector parametrization which reduces the n-dimensional optimization process to a one-dimensional problem. This aims to significantly reduce the number of calls to the forward model, a time-consuming step in inverse modeling. Next, we present two case studies to illustrate the calibration of a five-facies model of a confined aquifer to borehole data and pressure heads. Possible future applications in environment and sedimentary deposits are discussed.

2. Theory and Methodology

2.1. The sequential spectral turning bands method (S-STBM)

The S-STBM approach (Lauzon and Marcotte, 2020a) consists of building a GRF by adding cosine functions where each new phase $\varphi_i \epsilon [0,2\pi]$ is optimized to minimize any objective function:

$$Y_i(x) = \sqrt{\frac{i-1}{i}} Y_{i-1}(x) + \sqrt{\frac{1}{i}} \sqrt{2} \cos(\langle \omega_i, x \rangle + \varphi_i) \tag{1}$$

where $Y_i(x)$ is the GRF at iteration i defined at coordinate x in $\mathbb{R}^d$, d is the dimension of the GRF, $\langle \cdot,\cdot \rangle$ is the scalar product and ω_i is a frequency vector randomly oriented over a unit half-sphere of $\mathbb{R}^d$, and sample randomly from the radial spectral density of a covariance function C (Lauzon and Marcotte, 2020a). Note that the coordinate system x is grid free, the computational complexity is $O(n)$, n is the number of points, and the scalar products $\langle \omega_i, x \rangle$ can be computed simultaneously for all points x using parallelization on GPU (Lauzon and Marcotte, 2020a). Typically, at least one hundred phases are required to converge to multivariate Gaussian distribution (Lauzon and Marcotte, 2022).

Lauzon et Marcotte (2020a) have shown that for an equal number of calls to the flow simulator, it is better to shallow optimize the phase and add more cosine functions than the reverse. This characteristic is used in section 2.4 to propose a new optimization strategy when S-STBM is combined with PGS methods.

One advantage of S-STBM over usual calibration methods (e.g., phase annealing, gradual deformation, iterative spatial resampling) is its constructive nature which eliminates the requirement of an initial state (i.e., $Y_0(x)$ is an empty field for S-STBM). This ensures that S-STBM builds the property fields directly without having to adapt to the unfavorable characteristics of the initial state, a reason why the usual algorithms show slow convergences when they are combined with PGS simulations (Lauzon and Marcotte, 2022).

2.2. PluriGaussian simulations (PGS)

The idea behind PGS methods is to combine one or many L-GRFs simulated over the area of interest and assign facies according to the simulated values at each point. The assignation leads to the categorical field, $\mathcal{C}(x)$, and this is done using a truncation rule $T\colon \mathbb{R}^p \to \mathbb{N}$ that transforms the vector formed of the p L-GRFs to a categorical field. The function T takes a coordinate x, forms a vector with the respective values of the p L-GRFs at coordinate x, and returns a categorical variable. The categorical field is obtained when all vectors are transformed (See Fig.1 for an example):

$$T\left(Y_1(x), Y_2(x), \dots, Y_p(x)\right) \to \mathcal{C}(x) \tag{2}$$

A wide variety of geological phenomena may be simulated using PGS methods. One way to simulate heterogeneity in reservoir and aquifer modeling is to generate two or more PGS models that reflect different geological layouts. For example, with Bi-PGS where one represents the sedimentation process and the other the diagenesis (Renard et al., 2008). Another possibility to generate complex phenomena consists of using correlated L-GRFs through the linear model of coregionalization where the GRF are obtained by linear combinations of underlying variables (Armstrong et al., 2011). Its generalization to shift operator, regularization over support, or partial derivatives along orthogonal directions can help to model asymmetric cross-covariances (Marcotte, 2012). For example, cyclic and rhythmic facies architectures observed in sedimentary rocks sequences have been modeled using a shift operator which mimics the spatial asymmetric relationships of a sedimentary deposit (Blévec et al., 2020).

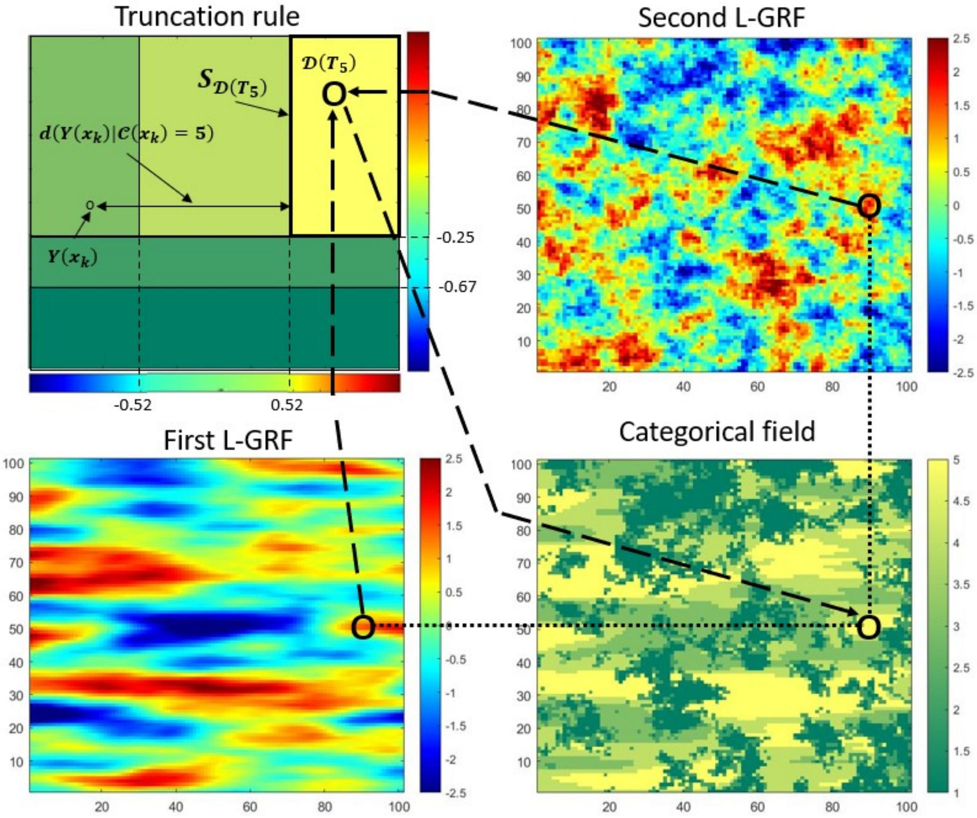


Figure 1 – Example of a pluriGaussian simulation with two uncorrelated L-GRFs. The dotted arrows represent an example of truncation described by Eq. 2. An example of notions of facies domain $\left(\mathcal{D}(T_j)\right)$, facies boundary $\left(S_{\mathcal{D}(T_j)}\right)$ and distance $\left(d_k(Y(x_k)|\mathcal{C}(x_k) = j)\right)$ is shown on the truncation rule for the yellow facies (facies 5) (See Eq.3 and Eq.4).

2.3. Alternative to the Gibbs sampler for boreholes conditioning

The idea behind the Gibbs sampler is to condition GRFs to inequality constraints. This is a Markov chain Monte Carlo algorithm in which each iteration consists of replacing a randomly selected point with a random Gaussian value drawn from a truncated Gaussian distribution derived from simple kriging. Thus, the Gibbs sampler aims primarily to ensure the constraints on the Gaussian values are respected and subsequently introduces the spatial correlation by drawing from the conditional distributions considering all other points, a slow and time-consuming step when several constraints are present (Marcotte et Allard, 2018). Here, we adopt an opposite point of view. We rather sequentially build a field where the spatial correlation is reproduced by construction and progressively enforce the constraints on the Gaussian values by optimizing an objective function.

To constrain the L-GRFs by S-STBM, the categorical observations, $C(x_k)$, are sequentially introduced by the optimization of each phase $\phi_{i,p}$ where $\phi_{i,p}$ refers to the ith phase of the pth L-GRF. At each iteration, p phases must be optimized using multivariate algorithms. As the geological problem is categorical, gradient-based optimization algorithms cannot be applied (Hu et al., 2001). We introduce in section 2.4 a new methodology based on the idea of gradual deformation to explore the space of the phase-vector using a single parameter to optimize.

To define the objective function to be minimized, the notions of facies domain and facies boundary are introduced. The domain $\mathcal{D}(T_j)$ refers to each Gaussian vector $A \in \mathbb{R}^p$ which, when $T(A)$ is applied, gives facies j (see Eq. 3). We note the boundary of a domain $S_{\mathcal{D}(T_j)}$. The distance between a categorical observation $C(x) = j$ and the Gaussian vector $Y(x)$ is 0 when $T(Y(x)) = j$, otherwise it is defined as the minimum Euclidean distance D to the boundary $S_{\mathcal{D}(T_j)}$. The objective function (OF) is the average of these shortest distances (see Eq. 4).

$$\mathcal{D}(T_j) = \{\forall A \in \mathbb{R}^p | T(A) = j\} \tag{3}$$

$$OF(Y) = \frac{1}{M}\sum_{k=1}^{M} d_k\,(Y(x_k)|C(x_k) = j), where \tag{4}$$

$$d_k(Y(x_k)|C(x_k) = j) = \begin{cases} \min\limits_{\forall Y \epsilon S_{\mathcal{D}(T_j)}} \{D(Y(x_k), Y)\} & if\ Y(x_k) \notin \mathcal{D}(T_j) \\ 0 & otherwise \end{cases}$$

where M is the number of categorical observations. Fig. 1 shows an illustration of Eqs 3 and 4 for the facies numbered 5 (the yellow one in Fig. 1). Note that the domain of each facies can be arbitrarily complex. It may be composed of many separate polygons, each one not necessarily convex as in D'or et al. (2017)

2.4. Optimization using gradual deformation

In PGS methods, truncation needs two or more GRFs. This involves several phases to be optimized at each iteration for the purpose to calibrate indirect data. In such conditions, the number of calls to the flow simulator may be higher than the univariate case to find an adequate set of parameters. This is due to the discrete nature of the geological models which proscribes the uses of gradient-based optimization methods.

We propose instead to take advantage of the building nature of S-STBM. As mentioned before, we only need to optimize slightly each phase, $\varphi_{i,p}$, and add several cosine functions to get an adequate calibrated field. The idea is to reduce the optimization process to a univariate one using a parametrization on the phase vector. To achieve this, gradual deformation is used. Two Gaussian white noise vectors of length p, y_1 and y_2, are fused (see Eq. 5). The inverse anamorphosis, Ψ^{-1}, enables to transform the normal vectors towards the theoretical distribution of the initial parameters, here, an independent uniform distribution on $[0,2\pi]$ for each phase.

$$\varphi_{i,n} = \Psi^{-1}\left(y_{1,i,n}\cos(t) + y_{2,i,n}\sin(t)\right) \tag{5}$$

3. Results

The MATLAB Reservoir Simulation Toolbox (Lie, 2019) was used to perform a synthetic example of a confined aquifer made of five geological units model using PGS with two uncorrelated L-GRFs. The truncation rule is the one shown in Fig. 1. The first example shows the conditioning of borehole data and the second one inverts the categorical field from the pressure heads. The same reference is used in both scenarios. The field size is 100 m × 100 m and is discretized on a 101 × 101 grid. The upper and lower sides were set as no-flow boundaries and fixed head boundary conditions of 1 m and 0 m were set on the left and right sides, respectively. The flow simulation was performed in a steady state. The first L-GRF is modeled using an isotropic exponential covariance with an effective range of 15 m. The second L-GRF refers to the cubic covariance with an anisotropic range of ax=50 m and ay=10 m.

3.1. Example 1: Conditioning to borehole data

One hundred samples are randomly selected from the reference for conditioning data (black circles in Fig.2(a-d)). The conditioning is performed using the methodology explained previously. The stopping criteria are the total number of OF evaluations (here 5000 OF calls) or when the OF returns a value of zero, indicating that the conditioning was performed adequately. Fig. 2(a-d) illustrates the reference (a) and three simulations (b-d) calibrated to the one hundred borehole data.

For our case, convergence is obtained quickly in very few iterations (See Fig. 2(e)). We stress that after 1000 iterations, the average distance is lower than 0.001 indicating that the remaining perturbations required to fulfill the constraints are small everywhere (i.e., close to their boundaries). One may end the calibration by S-STBM at this stage and switch to a T-SGS approach as proposed by Lauzon and Marcotte (2020b) to force the conditioning of the remaining points as they are extremely close to the boundary. Finally, Fig. 2. (f-h) presents respectively the variogram for the first L-GRF and the second L-GRF. The match between simulated and theoretical variogram is almost perfect for the two L-GRF.

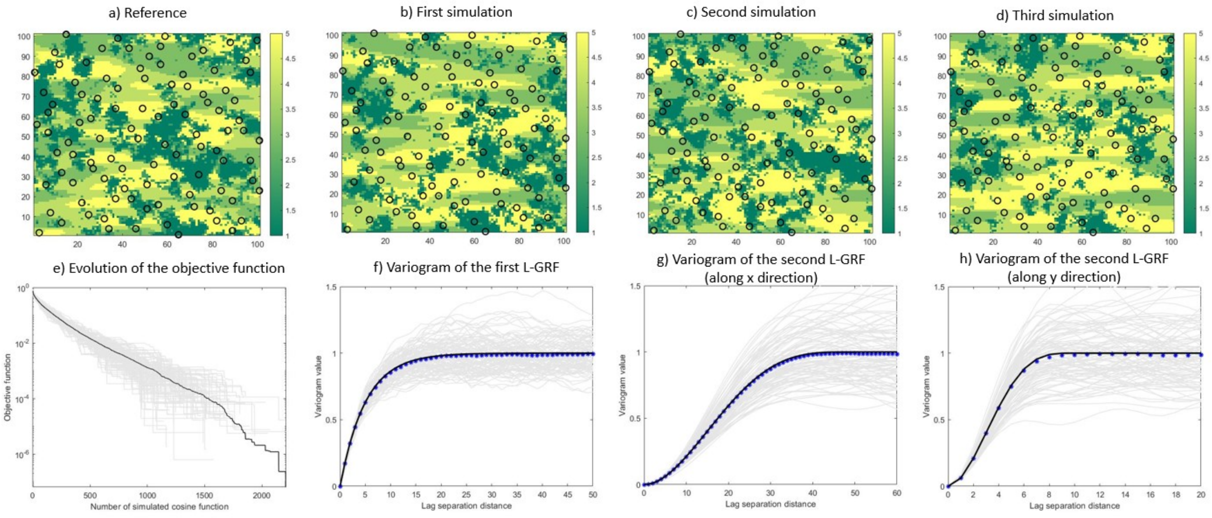


Figure 2 – Reference categorical field followed by three calibrated simulations by S-STBM to borehole data (black circles: localization of the 100-conditioning data). e) Evolution of the objective function (black line: mean objective function over 100 simulations; gray lines: objective function of the 100 calibrated realizations). f) Isotropic exponential variogram of the first L-GRF (effective range of 15m). g-h) Respectively, anisotropic cubic variogram along the x-direction (ax=50) and the y-direction (ay=10) (black line: theoretical model; gray lines: experimental variogram of the 100 calibrated realization; blue markers: mean experimental variogram).

3.2 Example 2: Calibration to pressure heads

The same confined aquifer was used to invert categorical fields to pressure heads (Fig.2 (a) and Fig. 3(a)). A pumping well is located at the center of the field with a pumping rate of 0.003 m^3/s. The 21 pressure heads obtained from a constant rate pumping test (black Xs in Fig. 3(c)) formed the calibration data. Conditioning to borehole data is performed using post-conditioning by kriging (Chilès and Delfiner, 2012). The latent Gaussian values are the ones obtained in the first example. So, the calibrated latent Gaussian at borehole locations values associated with the ith-simulation, in the first example, are used as hard data in the ith-simulation of the second example. The OF computed the MSE between the measured (h^m) and simulated (h^s) pressure heads:

$$OF(\mathcal{C}(x)) = \frac{1}{M}\sqrt{\sum_{i=1}^{M}\left(h_i^s - h_i^m\right)^2}$$

where $M = 21$, the number of head observations. The number of calls to the flow simulator was fixed to 2000. Table 1 indicates the hydrogeological properties used to model the confined aquifer.

Table 1 – Hydrogeological properties of the five units (kx, ky: permeability along x, y directions; φ: porosity).

Unit	k_x (mD)	k_y (mD)	φ (%)
Facies 1 (Shadow green)	1050	1050	0.22
Facies 2 (Medium green)	850	850	0.20
Facies 3 (Light green)	750	750	0.23
Facies 4 (Yellow green)	40	40	0.17
Facies 5 (Yellow)	20	20	0.16

The calibration results of one hundred simulations are shown in Figure 3(a). A two-order reduction amplitude of the objective function is obtained by using 2000 optimized cosine functions. One can note a rapid reduction of OF at the beginning which questions the need to calibrate 2000 cosine functions instead of, say, 100 cosine functions. A practitioner could have decided to stop the calibration process at 100 cosine functions and obtain satisfactory results according to his objectives. Note that the stopping criteria of an inversion depend on the objectives of the calibration, the quality and precision of the indirect data, and the execution time of the forward model (e.g., a flow simulator). Fig. 3. (f-h) presents respectively the variogram for the first L-GRF and the second L-GRF. The match between the average variograms of the hundred simulations is close to the theoretical variogram for both L-GRFs indicating that S-STBM can preserve the spatial correlation.

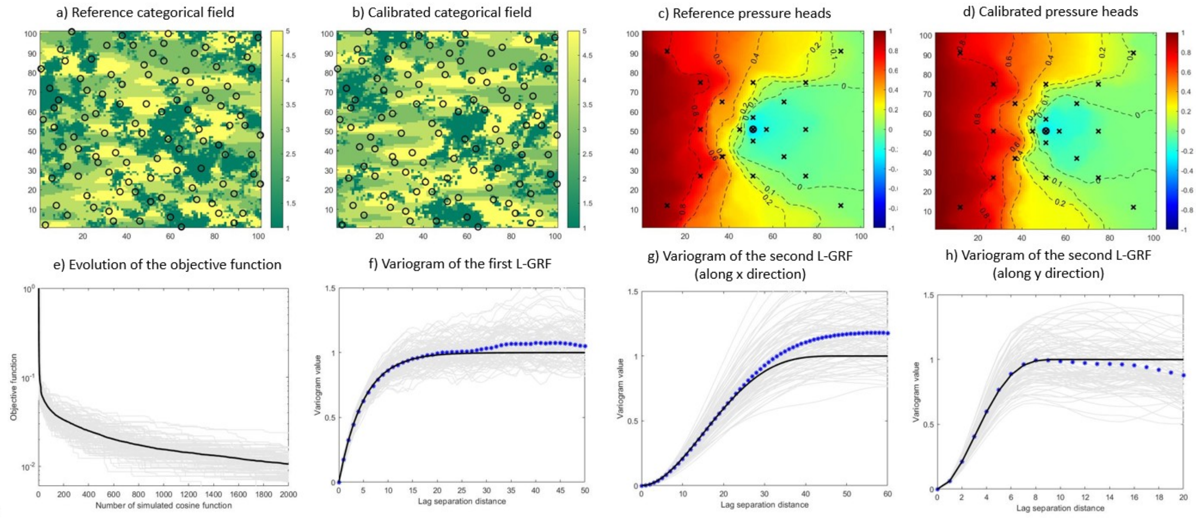


Figure 3 – a-b) Respectively, reference and calibrated categorical field (black circles: localization of the 100-conditioning data). c-d) Respectively, reference and calibrated pressure heads (black Xs: measured pressure heads; black circle: pumping well location). e) Evolution of the objective function (black line: mean objective function over 100 calibrated realizations; gray lines: objective function of the 100 calibrated realizations). f) Isotropic exponential variogram of the first L-GRF (effective range of 15m). g-h) Respectively, anisotropic cubic variogram along the x-direction (ax=50) and the y-direction (ay=10) (black line: theoretical model; gray lines: experimental variogram of the 100 calibrated realizations; blue markers: mean experimental variogram).

4. Discussion and Conclusions

This article presented a novel methodology to apply S-STBM with PGS methods. The optimization, which normally requires a multivariate approach in PGS, is reduced to a univariate optimization process by taking advantage of the properties of the S-STBM method combined with an approach based on gradual deformation. The case studies result demonstrates the effectiveness of S-STBM for borehole conditioning, an alternative to the reputedly slow Gibbs sampler when multiple boreholes are available, and for calibrating a geological model to pressure heads.

Our proposed parameterization using gradual deformation is compared to a multivariate approach and a univariate calibration method based on alternating phase calibration (i.e., calibrating each phase one by one). Fig.4 presents the results of the comparison for example 2. The left image shows the OF evolution based on the number of calls to the flow simulator (fixed at 800 calls) and the right illustrates the OF evolution based on the number of simulated cosine functions for each L-GRF. One can see that the multivariable approach is less efficient (red line) than the two univariate scenarios due to the time required to find an adequate phase vector (10 calls to perform one iteration for the multivariate optimization compared to 2 for the parametrization and 4 for the alternating approach). Note that parametrization using gradual deformation and the alternating approach shows similar calibration results when compared to the numbers of calls to the flow simulator (Fig.4, left one). In a case with more variables to calibrate, it is likely that the parametric approach becomes more efficient than the alternating phase approach. This however is for the moment speculative and remains to be validated.

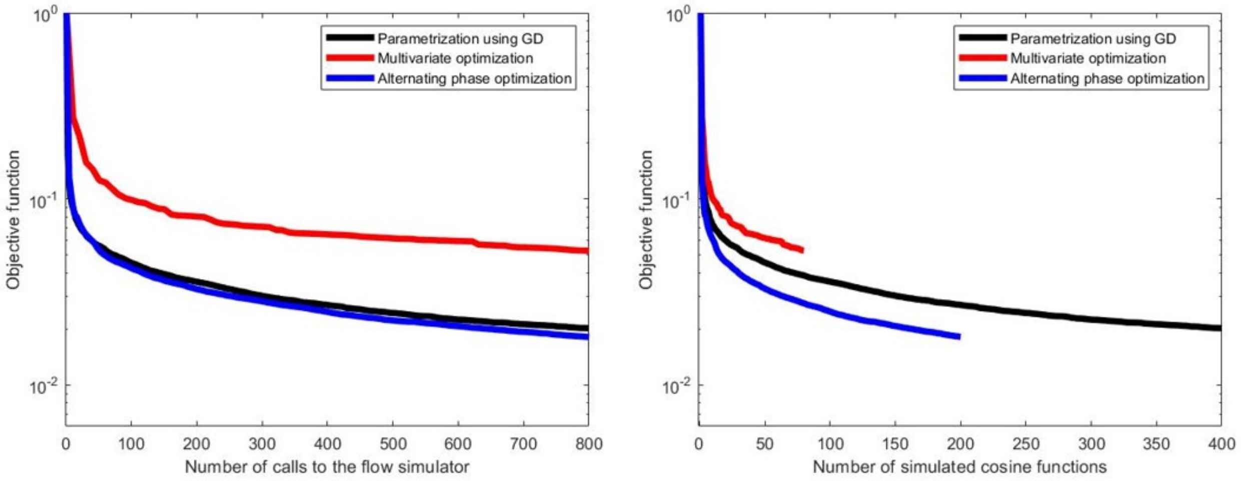


Figure 4 – Comparison of three optimization strategies. Left) based on the number of calls to the flow simulator. Right) based on the numbers of simulated cosine functions. (Black: parametrization using gradual deformation; Red: multivariate optimization using a Nelder–Mead method; Blue: univariate optimization using an alternating scheme with one phase at a time).

Note that a few tens or hundreds of cosine functions allow a reduction of the OF by 1 to 2 orders of magnitude (see. Fig. 2(e) and Fig. 3(e)). This is probably sufficient to stop the calibration accounting for modeling errors and the accuracy of real data or to account for errors in borehole positioning, and geological units identification. Fig. 2(e) and Fig. 3(e) also show that the calibration of borehole data requires more cosine functions than the calibration of pressure heads. This is not an issue because the objective function associated with the borehole conditioning requires only calculation of distances, a fast step compared to the use of a flow simulator. To be mentioned, the hydrogeological problem took more time than conditioning the data, despite the significant difference in the number of cosine functions generated.

The main drawback of PGS methods lies in the determination of variogram models for L-GRFs. Since several sets of parameters may satisfy the field data, it may be difficult to justify the use of a given variogram model. In practice, the geological interpretation by geoscientists or analogous models may guide the variogram model selection.

To check the impact of the calibration of borehole data and pressure heads, the average of the facies proportions for 100 realizations and the percentage of well-simulated facies with respect to the reference facies were computed. In addition, the Kullback-Liebler divergence was computed relatively to the reference facies proportions. Note that the borehole data represents only 1% of the field. Table 2 resumes the results of four cases: uncalibrated (b), calibrated only to borehole data (c) and to pressure heads (d), and calibrated to both borehole data and pressure heads (e). Scenario a) is the reference. Results indicate that the scenario calibrated on heads alone (d) marginally improves the well identified facies (22.6 vs 20.8) and deteriorates the KL-divergence with respect to uncalibrated case (b) (0.45 vs 0.32). The scenario calibrated on only the borehole facies (c) improves significantly the well identified facies (36.7 vs 20.8) and leaves the KL-divergence similar to the uncalibrated case (0.31 vs 0.32). Finally, jointly calibrating to heads and facies (e) gives the best results on proportions of well classified (38.2 vs 20.8) and KL-divergence (0.20 vs 0.32). Hence, although the pressure heads are not very informative about the facies when used alone, combining them with the facies observed in boreholes enables to significantly reduce the KL-divergence and increase slightly the facies identification with respect to using only facies in boreholes. Note that these results are obtained despite facies proportions observed in boreholes were quite different from facies proportions over the whole reference field as indicated by the KL-divergence of 1.68 for the boreholes.

Table 2 – Proportion of facies before and after calibration. a) Reference. Averages of 100 realizations b) uncalibrated, c) calibrated to observed facies, d) calibrated to pressure heads, e) calibrated to observed facies and pressure heads.

Unit	a)	b)	c)	d)	e)	Proportion of observed facies in boreholes
Facies 1 (Shadow green)	21.4	24.3	19.8	24.6	20.5	16
Facies 2 (Medium green)	14.0	14.8	13.7	15.0	13.5	11
Facies 3 (Light green)	18.8	17.9	18.1	16.3	17.9	22
Facies 4 (Yellow green)	26.2	24.4	25.6	25.6	25.9	28
Facies 5 (Yellow)	19.6	18.8	22.8	18.6	22.2	23
Percentage of points with coincident simulated/reference facies (%)	-	20.8	36.7	22.6	38.2	-
KL-Divergence with respect to reference a)	-	0.32	0.31	0.45	0.20	1.68

References

Armstrong, M., Galli, A., Beucher, H., Loc'h, G., Renard, D., Doligez, B., Eschard, R., Geffroy, F. (2011). Plurigaussian simulations in geosciences. Springer Berlin Heidelberg. doi: 10.1007/978-3-642-19607-2.

Blévec, T.L., Dubrule, O., John, C.M., Hampson, G.J. (2020). Geostatistical Earth modeling of cyclic depositional facies and diagenesis. AAPG Bulletin. 104(3): p.711–734. doi: 10.1306/05091918122.

Chilès, J.-P., Delfiner, P. (2012). Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons, Inc., ISBN: 9781118136188. doi: 10.1002/9781118136188.

D'Or, D., David E., Walgenwitz A., Pluyaud P., and Allard D. (2017). Non stationary plurigaussian simulations with auto-adaptative truncation diagrams using the cart algorithm. 79th European Association of Geoscientists and Engineers Conference and Exhibition 2017, p.1-5. doi: 10.3997/2214-4609.201701019.

Hu, L. Y., Le Ravalec, M., & Blanc, G. (2001). Gradual deformation and iterative calibration of truncated Gaussian simulations. Petroleum Geoscience, 7(S), S25-S30. doi: 10.1144/petgeo.7.S.S25.

Lauzon, D., Marcotte, D. (2020a). Calibration of random fields by a sequential spectral turning bands method. Computers & Geosciences. 135, 104390. doi: 10.1016/j.cageo.2019.104390.

Lauzon, D., Marcotte, D. (2020b). The sequential spectral turning band simulator as an alternative to Gibbs sampler in large truncated- or pluri- Gaussian simulations. Stochastic environmental research and risk assessment. 34 (11), 1939–1951. doi: 10.1007/s00477-020-01850-9.

Lauzon, D., Marcotte, D. (2022). Statistical comparison of variogram-based inversion methods for conditioning to indirect data. Computers & Geosciences. 160, 105032. doi: 10.1016/j.cageo.2022.105032.

Lie, K.-A. (2019). An introduction to reservoir simulation using MATLAB/GNU Octave: User guide for the MATLAB reservoir simulation toolbox (MRST). Cambridge University Press, doi: 10.1017/9781108591416.

Marcotte, D. (2012). Revisiting the Linear Model of Coregionalization. In: Abrahamsen P, Hauge R, Kolbjørnsen O (eds) Geostatistics Oslo 2012 Quantitative Geology and Geostatistics, vol 17 Springer, Dordrecht. doi: 10.1007/978-94-007-4153-9_6.

Marcotte, D., Allard, D. (2018). Gibbs sampling on large lattice with GMRF. Computers & Geosciences. 111:190–199. doi: 10.1016/j.cageo.2017.11.012.

Renard, D., Beucher, H., Doligez, B. (2008). Heterotopic Bi-Categorical Variables in PluriGaussian Truncated Simulations. VIII International Geostatistical Congress pp 289–298.

EFFICIENT INVERSION WITH COMPLEX GEOSTATISTICAL PRIORS USING NORMALIZING FLOWS AND VARIATIONAL INFERENCE

Shiran Levy (1)* - Eric Laloy (2) - Niklas Linde (1)

Institute of Earth Sciences, University of Lausanne, Lausanne, Switzerland (1) - Institute for Environment, Health and Safety, Belgian Nuclear Research Centre, Mol, Belgium (2)
** Corresponding author: shiran.levy@unil.ch*

Abstract

We study the combination of inverse autoregressive flows (IAF) and variational inference (VI) within the context of geophysical inverse problems as an efficient alternative to Markov chain Monte Carlo (MCMC) sampling. Variational inference seeks to approximate a target distribution parametrically for a given family of distributions by solving an optimization problem. It provides a computationally-efficient approach that scales well to high-dimensional problems, but the approximation is limited by the chosen parameterized family of distributions. To enable more expressive approximate distributions, we explore the combination of VI with inverse autoregressive flows (IAF), in which a series of neural transport maps transform an initial density of random variables into a target density. In this VI-IAF routine, samples from a normal distribution are pushed forward through a series of invertible transformations onto a variational density approximating the unnormalized posterior. The parameters of the IAF are learned by minimizing the Kullback-Leibler divergence between the variational density and the unnormalized target posterior distribution. In our study, we use a deep generative adversarial network (GAN) to generate complex geostatistical priors described by the low-dimensional, latent space of the GAN. We compare this approach against popular methods for solving geophysical inverse problems such as deterministic gradient-based methods and MCMC sampling. Even if previous attempts to perform gradient-based inversion in combination with GANs of the same architecture were proven unsuccessful, preliminary results with VI-IAF on channelized subsurface models and linear physics suggest that this approach recovers the true model reliably and provides appropriate uncertainty quantification with a relatively low amount of computation. As a next step, we will test this routine on cases where the forward model is nonlinear. As most of the nonlinearity comes from the GAN generator, we expect the results to be similar to those obtained in the linear case.

GEOSTATISTICAL ELECTRICAL RESISTIVITY TOMOGRAPHY INVERSION FOR GROUNDWATER CHARACTERIZATION

João Lino Pereira (1)* - Mafalda Oliveira (2) - Rui Guinote (2) - Emmanouil A. Varouchakis (3) – J. Jaime Gómez-Hernández (4) - Leonardo Azevedo (1)

CERENA, DECivil, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal (1) - Somincor, Neves-Corvo, Portugal (2) - School of Mineral Resources Engineering, Technical University of Crete, Chania, Greece (3) - Institute of Water and Environmental Engineering, Universitat Politècnica de València, Valencia, Spain (4)
** Corresponding author: joao.lino.pereira@tecnico.ulisboa.pt*

Abstract

Electrical Resistivity Tomography (ERT) is a geophysical method used to characterize the spatial distribution of subsurface electrical properties such as electrical resistivity. These geophysical data are acquired by injecting electrical current into the ground via a pair of electrodes and measuring the resulting potential field by a corresponding pair of potential electrodes. As standard practice, the observed measurements are then converted into apparent resistivity, a weighted average of the resistance of earth materials to current flow. These data may be used to map geological structures, fractures, stratigraphic units, groundwater bodies, as they depend on variations in lithology, porosity, permeability, water saturation and fluid conductivity.

To predict the model parameters from the field data we need to solve a geophysical inverse problem. From near-surface characterization studies, this is often accomplished with deterministic resistivity inversion models. Deterministic approaches have three main limitations: the predicted models are smooth representations of the subsurface; the predicted models are highly dependent on the initial solution; and have limited uncertainty assessment capabilities. We propose herein an alternative stochastic resistivity inversion method based on geostatistical simulation and co-simulation as model perturbation technique. A set of electrical resistivity models are generated conditioned to the available resistivity borehole data, and assuming a spatial continuity pattern as revealed by a variogram model retrieved directly from the data. From the set of simulated models, we use a forward model that allows to compute synthetic apparent resistivity models. These are locally compared against the observed one. The portions of the geostatistical realizations that ensure the maximum similarity between predicted and observed apparent resistivity are stored in an auxiliary volume along with the similarity coefficient. Both auxiliary volumes are used as secondary variable in the co-simulation of a new set of models in the subsequent iteration.

We illustrate this methodology by applying it to a set of two-dimensional profiles obtained from an ERT survey carried out at Neves-Corvo mining site (Alentejo region, Portugal). The survey was performed to characterize the spatial distribution of the aquifer located within the mine premises. This methodology was able to predict electrical resistivity models from the data obtained in the ERT survey. All the predicted models are able to generate synthetic geophysical data that match the observed one (around 0.90 similarity coefficient) while reproducing the borehole data and the imposed variogram models. We show the ability of the model to assess uncertainty and compare the results against a conventional deterministic inversion methodology available at a commercial software.

PERMEABILITY PREDICTION WITH GEOSTATISTICAL SEISMIC INVERSION CONSTRAINED BY ROCK PHYSICS

Roberto Miele (1)* - Dario Grana (2) - João Felipe Costa (3) - Paula Bürkle (4) - Luiz Eduardo Varella (4) - Bernardo Viola Barreto (4) - Leonardo Azevedo (1)

CERENA - Civil Engineering Department - Instituto Superior Técnico, Lisbon, Portugal (1) - School of Energy Resources - Wyoming University, Laramie (WY), United States (2) - DEMIN, Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, Brazil (3) - Petrobras, EXP/GEOP/TGEO, Rio de Janeiro, Brazil (4)

** Corresponding author: roberto.miele@tecnico.ulisboa.pt*

Abstract

Rocks' permeability (k) is a property describing the ease with which a given fluid in the subsurface can flow through the pores of a rock sample. The prediction of its spatial distribution is thus fundamental when characterizing a hydrocarbon or CO2 reservoirs. Nonetheless, rocks' permeability is strongly sensible to several geological factors, such as the sedimentary and diagenetic processes, and to the rocks' structures, causing its values to change over several order of magnitude even in a single rock type (Ma, 2019; Yang, 2017). Different approaches have been developed to predict the spatial distribution of k for a given area of interest. Data-driven approaches, such as stochastic sequential simulation (Deutsch, 2002), aim at the reproduction of the experimental data distributions (marginal and joint distributions of k and other rock properties, often ϕ) and their spatial patterns (e.g., a variogram model). On the other hand, rock-physics-modelling-based methods consist in using empirical or heuristic equations (e.g., Kozeny-Carman, RGPZ) (Mavko et al., 2019; Glover et al., 2006) to estimate k form ϕ, given other rocks' parameters such as grain size, cementation and tortuosity.

Iterative geostatistical seismic inversion methods (e.g., Azevedo et al., 2020; Bosch et al., 2010; Grana et al., 2017, 2021; Sen, 2006) use a data-driven approach to predict the distribution of the subsurface' rock physics and its uncertainties by relating k, ϕ and acoustic impedance (I_P) to seismic reflection amplitudes. Hence, the objective function is the minimization of the seismic data misfit. Nonetheless, the strong variability of k can represent a limitation to these methods: since the stochastic models' perturbation lacks any physics constraint, the simulated models can finally fit the observed seismic but lack a physical or geological meaning.

To tackle such issues, we propose an iterative geostatistical seismic inversion algorithm to invert for litho-fluid facies and k, where each model is simulated following a data-driven approach, but its iterative optimization is also conditioned to rock physics constraints. For a post-stack seismic volume, we first simulate facies models by means of 1D first-order Markov chain Monte Carlo method, then we sequentially simulate k and co-simulate ϕ and I_P. We can thus calculate a synthetic seismic volume from the reflectivity model obtained from the I_P model. Through a pre-calibrated facies-dependent rock physics models, we calculate a second volume of k from the simulated I_P and ϕ models. This estimated model represents the expected k given the available rock physics a priori knowledge, for the I_P and synthetic seismic data generated. Thus, the algorithm's objective function is the minimization of both the seismic data misfit and of the two permeability models.

We validated the proposed inversion algorithm through a synthetic, one-dimensional case study application, comparing the results to those of a purely data-driven approach. We also apply the method to a three-dimensional real dataset. The algorithm generated accurate models of permeability, demonstrating to predict better results than those of conventional approaches.

GROUNDWATER CONTAMINANT SOURCE CHARACTERIZATION THROUGH ARTIFICIAL NEURAL NETWORKS

Laura Molino (1) - Daniele Secci (1)* - Andrea Zanini (1)

University of Parma, Department of Engineering and Architecture, Parma, Italy (1)
** Corresponding author: daniele.secci@unipr.it*

Abstract

Water plays a crucial role in human life and in all its activities. For this reason, all water resources and in particular groundwater should be managed in a sustainable way in order to satisfy current needs and without causing environmental consequences. Unfortunately, economies based on intensive agriculture and industrial production lead to unsustainable use of water, the effect of which also includes the contamination of aquifers. In this context, the identification of the location of the contaminant source with its release history has attracted great attention within the scientific community called upon to provide theoretical methods to limit the spread of the contaminant. To identify remediation strategies immediately is essential to have a tool that can provide accurate results in real time. With this aim, surrogate models can become the conceptual models of primary choice being able to study forward and inverse transport problem using a number of observations, which is not much greater than the unknown parameters to be calculated, reducing in this way the computational cost compared with other more complex models. Data-driven surrogate models lead to the field of Artificial Intelligence where neural networks, trained on a finite dataset, are able to estimate the desired output by means of a learning process emulating the behavior of the human brain.

In this work, a feedforward artificial neural network (FFWD-ANN) has been developed to analyze different cases as surrogate model. The investigated domain has been selected from a literature study (Ayvaz, 2010) and the training dataset has been randomly developed by means of the Latin Hypercube Sampling in order to reduce the number of forward simulations. Initially, the network has been trained to solve forward transport problem. In the proposed approach, the ANN well estimates the pollutant concentrations in 7 monitoring wells, at different times, by using as input data the release history at two contaminant sources with known locations. Then, the surrogate model has been trained to deal with inverse transport problem related to different application cases: 1. estimation of the release history at one contaminant source with known location; 2. simultaneous estimation of the release history and location of one contaminant source; 3. estimation of the release history at two contaminant sources with known location; 4. simultaneous estimation of the release history at two contaminant sources with known location and error on observations.

The results have been compared with literature data (Ayvaz, 2010; Jamshidi et al. 2020). Artificial Neural Network seems to be well suited to dealing with this type of forward and inverse problems, preserving the reliability of the results and reducing the computational burden of numerical models.

This research was developed under the scope of the InTheMED project. InTheMED is part of the PRIMA Program supported by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No 1923.

Jamshidi, A., Samani, J.M.V., Samani, H.M.V., Zanini, A., Tanda, M.G., Mazaheri, M., 2020. Solving Inverse Problems of Unknown Contaminant Source in Groundwater-River Integrated Systems Using a Surrogate Transport Model Based Optimization. Water 12, 2415.

Ayvaz, M.T., 2010. A linked simulation–optimization model for solving the unknown groundwater pollution source identification problems. J. Contam. Hydrol. 117, 46–59.

MODELLING THE COMPLEXITY BENEATH OUR FEET: A JOINT INVERSION FDEM AND ERT TECHNIQUE

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

CERENA, Universidade de Lisboa, Instituto Superior Técnico, Lisboa, Portugal (1) - Ghent University, Department of Environment, Gent, Belgium (2)
** Corresponding author: joao.narciso@tecnico.ulisboa.pt*

Abstract

The near-surface is a complex and highly dynamic region of the Earth, resulting from interacting processes of both natural and anthropogenic origin, and characterized by physical properties with small-scale heterogeneity. Due to its importance in many human activities, an accurate characterization of the spatial distribution of near-surface properties is often challenging, yet essential in different activities of socio-economic, mineral resources and environmental fields. The modelling of these systems is often based on discrete direct observations acquired through conventional invasive sampling techniques, such as boreholes and trenches, that have proven competent in one-dimensional modelling but are expensive and impractical to acquire in some sites, cause ecosystem tampering and reveal limitations in capturing the spatial variability of near-surface properties. From the pressing need to make the characterization more detailed and comprehensive, but also with reduced costs, time efficiency and operational flexibility, the characterization of the near-surface through geophysical surveys have been emerging as powerful techniques in modelling the complexity of these systems through indirect, and virtually continuous, measurements of the subsurface physical properties. Within the most common near-surface geophysical techniques, frequency-domain electromagnetic (FDEM) induction and electrical resistivity tomography (ERT) methods have demonstrated their efficiency to characterize heterogeneous subsurface systems due to their simultaneous sensitivity to two key subsurface properties, electrical conductivity (EC) and magnetic susceptibility (MS). Due to differences in the spatial resolution of both methods, these are often acquired jointly, but interpreted and modelling separately.

The prediction of reliable subsurface models from these geophysical data can be solved through a joint geophysical inverse problem. However, handling the differences in the resolution and nature of both methods is not straightforward and prone to uncertainties. On the other hand, the joint inversion leverages the benefits of each method individually.

In this work, we show an iterative geostatistical joint FDEM and ERT inversion technique. A geostatistical framework is used to couple both data domains in a consistent spatial model. The method is illustrated with its application to synthetic and real data sets where the benefits of the joint approach are discussed against the individual inversions.

IDENTIFICATION OF CONTAMINANT SOURCES AND PLUMES AFFECTED BY BIODEGRADATION AND SORPTION PROCESSES BY ENSEMBLE KALMAN FILTERS

Alicia Sanz-Prat (1)* - J. Jaime Gómez-Hernández (1)

Institute of Water and Environmental Engineering, Universitat Politècnica de València, Valencia, Spain (1)
** Corresponding author: asanpra@iiama.upv.es*

Abstract

Contaminant events disrupt the stability and resilience of increasingly vulnerable soil and groundwater resources. Identifying where, when, and how much contaminant spill is released into aquifers is critical for improving remediation techniques and clarifying environmental liability, but commonly troublesome in contaminant sites where sparse observation networks are the unique tool to define the status of soil and aquatic ecosystem. Such ecosystems usually are subject to coupled nonlinear physicochemical processes and dynamic environmental conditions. Despite constrained model assumptions and demanding computational time, hydrogeology stochastic inverse models (SIM) are considered excellent methodologies to extract consistent and valued input-parameter information from non-sampled areas by analysing predictive responses of the system in comparison with actual observed responses. Among the SIM stands out the data assimilation method ensemble Kalman Filter (EnKF) capable of simultaneously estimating model parameters (hydraulic conductivity and porosity), as well as the location of the contaminant source and the evolution of the discharge mass flow from observations of the piezometric head and concentration (state variables). After an initialization of the model parameters, the steps of the EnKF are (i) prediction of the state variables by direct modelling from time k = 0 and (ii) updating the estimated values of the parameters from the deviations between observations and predictions. Parameters and corrected variables serve as input data in the next iteration at time k + 1. It is well known that EnKF performance has been favourable when modelling conservative transport. The main objective of this study is to move forward into reactive transport. For that purpose, this study applies the ensemble Kalman Filter (EnKF) data assimilation for transport inverse modelling when biodegradation and isotherm sorption processes are present. We test spurious effects of aquifer heterogeneity, multicomponent and multiparameter reactive cases, as well as the influence of initial/boundary conditions in synthetic scenarios.

ASSESSMENT OF HYDRAULIC CONDUCTIVITY FIELD USING LABORATORY SANDBOX TRACER TEST DATA AND AN ENSEMBLE KALMAN FILTER METHOD

Valeria Todaro (1)* - Andrea Zanini (1) - Marco D'Oria (1) - J. Jaime Gómez-Hernández (2) - Maria Giovanna Tanda (1)

University of Parma, Department of Engineering and Architecture, Parma, Italy (1) - Universitat Politècnica de València, Valencia, Spain (2)
** Corresponding author: valeria.todaro@unipr.it*

Abstract

Knowledge of hydraulic and transport characteristics of natural aquifers is fundamental for planning several engineering applications as groundwater extraction or recharge systems and prediction of spatio-temporal evolution of subsurface contamination. Tracer tests are often used to characterize aquifers. In this work, we apply an ensemble Kalman filter technique to infer the hydraulic conductivity field from concentration data obtained by means of tracer test. The methodology is validated using data collected in a laboratory sandbox that reproduces a vertical cross-section of an unconfined aquifer, in which the flow is driven by constant gradient head at the upstream and downstream boundary. The porous medium consists of glass beads of different diameters that reproduce a heterogeneous field and fluorescein sodium salt is used as tracer. The plexiglass walls of the sandbox allow to collect concentration values by means of an image technique process. Breakthrough curves recorded at different monitoring points are used as observations in an inverse problem aimed at estimating the hydraulic conductivity field. Here, we propose to apply the Ensemble Smoother with Multiple Data Assimilation (ES-MDA) to solve this type of inverse problem. ES-MDA is an iterative data assimilation method that updates the unknown parameters based on the knowledge of observed measurements and a numerical model that describes the forward process. In this case, the groundwater flow and transport processes are modeled with MODFLOW and MT3DMS software, respectively. Two different approaches are tested to estimate the heterogeneous hydraulic conductivity field. In the first test, the conductivity field is estimated using the pilot points method: the hydraulic conductivity is estimated in a finite number of points, which are then interpolated using an Ordinary Kriging to obtain the solution over the whole domain. This reduces the number of parameters to be estimated and consequently the computational burden. On the other hand, the pilot point approach can lead to over-smoothed solutions.

In the second test, a fully parameterized approach is adopted. The hydraulic conductivity is estimated at each cell of the discretized domain leading to a large number of unknown parameters. The second approach can better characterize the true heterogeneity but requires more computation time than the first one. To reduce the computational burden, which is related to the ensemble size used to perform ES-MDA, some corrections on the algorithm, such as covariance localization and covariance inflation, are applied. This leads to promising results for both approaches showing the capability of Ensemble Kalman filter methods to handle tracer test data with the aim to characterize the aquifer conductivity.

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)

Dipartimento di Scienze della Terra "A.Desio", Università degli Studi di Milano, Milan, Italy (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.

PIXEL-BASED MULTIPLE-POINT-STATISTICS PARAMETRIZATION BASED ONLY ON TRAINING IMAGES

Mathieu Gravey (1)* - Gregoire Mariethoz (2)

Department of Physical Geography, Utrecht University, Utrecht, Netherlands (1) - IDYST, University of Lausanne, Lausanne, Switzerland (2)
** Corresponding author: conf@mgravey.com*

Abstract

Developments in multiple-point-statistics (MPS) algorithms over the last decade have made the approach more and more viable for practical applications. Nowadays, MPS can reproduce structures better than ever before, opening opportunities for increasingly complex simulations. However, such good results can only be achieved with a good algorithmic parametrization. Often, finding an appropriated parametrization for MPS simulation is more an art than a science, which requires training and practice.

While it is generally possible to find an acceptable parametrization with a try-and-error approach for univariate or sometimes bivariate problems, this solution is impractical when the number of variable increases or with complex parameters. A classical solution, adopted in previous work, is to find an optimal set of parameters with optimization approaches using an objective function over simulations. These approaches require a significant number of simulations, and therefore an important computation time.

Here, we propose a novel approach that uses exclusively the training image to find an optimal set of parameters. The main advantage of our approach is to remove the risk of over-fitting the objective function. At the same time, we don’t use an optimization approach, which means that we find a set of parameters in a predictable time. Our approach is based on the understanding of how the simulation algorithms works. We developed it for QuickSampling (QS), but it can be easily adaptable to any other pixel-based MPS algorithm.

GEOSTATISTICAL SIMULATION FOR OFFSHORE WIND SPEED SPATIO-TEMPORAL GAP-FILLING

Stylianos Hadjipetrou (1)* - Gregoire Mariethoz (2) - Phaedon Kyriakidis (1)

Cyprus University of Technology, Department of Civil Engineering and Geomatics, Limassol, Cyprus (1) - University of Lausanne, Institute of Earth Surface Dynamics, Lausanne, Switzerland (2)
** Corresponding author: sk.hadjipetrou@edu.cut.ac.cy*

Abstract

Wind resource assessment in the context of renewable energy feasibility studies is of utmost importance for the urgent need of decarbonizing the energy sector. Researchers from multiple disciplines typically rely on long term, albeit spatially coarsely resolved datasets, e.g., Numerical Weather Prediction (NWP) model outputs, to conduct local wind energy/resource studies. Attempts have been made to downscale these products, via dynamical or statistical downscaling, which turned to be computationally expensive without being able to reproduce the spatiotemporal variability inherent in wind speed complex patterns. While finer spatial resolution datasets have been made available e.g., Synthetic Aperture Radar (SAR) data, the revisit frequency of the satellites carrying relevant sensors leads to temporal gaps not allowing for a complete assessment of the wind resource potential.

In this study, we employ a geostatistical framework, namely Multiple-Point Statistics (MPS) to simulate wind speed patterns in the offshore area around Cyprus, aiming to fill Sentinel-1 information gaps where they exist. The data used comprise Sentinel-1A and 1B SAR gridded estimates of the surface wind speed derived from Interferometric Wide (IW) Swath beam mode under Vertical-Vertical (VV) + Vertical Horizontal (VH) dual polarization operation and Uncertainties in Ensembles of Regional Reanalyses (UERRA) data after being resampled to the satellite's spatial resolution (1km), both validated against in-situ measurements from local meteorological coastal stations. More precisely, pairs of UERRA and Sentinel-1 co-registered information are used as Training Images (TIs) from which wind patterns are eventually inferred. The selection of TIs used to create each TI set is based on Root Mean Square Error (RMSE) minimization while the recently developed Quick Sampling (QS) algorithm was used to generate the MPS simulations. Multiple realizations are generated to provide an uncertainty estimate of the final output while synthetic image time-series were evaluated via cross-validation as well as by statistical comparison against Sentinel reference data.

As an illustration of the methodology, offshore wind speed images are simulated at a spatial resolution of 1km for the offshore areas of Cyprus over a 2-years period. Results imply that the proposed methodology could form a consistent time series of high spatial resolution wind speed images leading to temporally finer resolved offshore wind power estimates for the region, provided the above procedure is generalized for a longer time-period.

APPLYING MPS TO POINT DATA MERGING USING PATTERN-TO-POINT (P2P) CATALOGS

Fabio Oriani (1,2)* - Gregoire Mariethoz (1)

Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland (1) Agroscope - Swiss Centre of Excellence for Agricultural Research, Zürich, Switzerland (2)
** Corresponding author: fabio.oriani@protonmail.com*

Abstract

Data merging is the common geostatistical practice of interpolating point data using gridded images as predictive variables (e.g. remote sensing imagery, modeled rasters, or reanalysis datasets) to obtain continuous representations for spatial analysis and numerical modeling. Multiple-point-statistics (MPS) algorithms [1,3,2], which preserve complex spatial data patterns for realistic representations, are generally not applicable in this case, since they can only operate on gridded images instead of sparse data points. In this study, we propose a strategy to extend the MPS concept to the merging of any point data with gridded images.

The novel approach, called pattern-to-point (P2P) catalogs, retrieves from the gridded image the pixel values close to any data point. This way, a catalog of local pixel patterns associated to point data is obtained. The P2P catalog, built over one or multiple images, constitutes a training dataset that can be used to project the point-data information over the whole grid, where point data are not present.

We present here a preliminary test where the P2P technique is used to bias-correct a weather radar image for the estimation of 10-min rainfall intensity over Switzerland (MeteoSwiss). The national rain gauge dataset is used as ground-truth point data source to merge with the radar images. The bias-corrected value for every pixel is obtained by retrieving the pixel pattern in its neighborhood and looking for similar patterns in the P2P catalog, built using 10-min images from 2 days of radar activity. The found ensemble of similar patterns carries the associated point data, whose values are averaged to estimate the bias-corrected value for the target pixel. The process, repeated for every pixel, generates a corrected image revealing small-scale structures, which reflect the latent information in the P2P catalog.

The P2P strategy presents as a promising approach that can be used for different applications, including: 1) deriving space-varied probability distributions of a point variable, 2) realistic gridded estimations based on the projection of historical data, and 3) assessing the uncertainty on each grid value where the scale jump between point and grid is taken into account.

Bibliography

[1] Guardiano, F., and Srivastava R., Multivariate geostatistics: beyond bivariate moments Geostatistics, Troia, Kluwer Academic Publications, Dordrecht, 1993, 1, 133-144

[2] Mariethoz, G.; Renard, P. & Straubhaar, J. The Direct Sampling method to perform multiple-point geostatistical simulationsWater Resources Research, Amer Geophysical Union, 2010, 46, W11536

[3] Strebelle, S., Conditional simulation of complex geological structures using multiple-point statistics, Mathematical Geology, Springer, 2002, 34, 1-21

A GEOSTATISTICAL POINT OF VIEW ON HETEROSUPPORT AND HETEROTOPIC CO-REGIONALIZATION OF REMOTE SENSED INFORMATION

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

DICAM, University of Bologna, Bologna, Italy (1)
** Corresponding author: roberto.bruno@unibo.it*

Abstract

The growing use of remote sensing data (typically multi / hyperspectral satellite / drone images) has pushed towards the exploitation of indirect information known throughout the field of study. The use of remote sensing as indirect information, together with a limited number of direct and indirect field data, should improve the characterization of the target variable. The main challenge is to seek and quantify the possible correlations between the available information.

The starting point is the regularized nature of the information associated with the pixels. This prevents the identification and quantification of existing correlations, if at a scale below resolution.

An intuitive correlation study must be handled carefully for many reasons, including the position model of each terrestrial data within the corresponding pixel resolution surface, being the sample positions isotopic / heterotopic with the pixel or randomly distributed.

Often, comparing different images ignores some important issues. A typical problem is that images derived from different satellites / drones / equipment have different pixel resolution, therefore different support (heterosupport). Finally, the image data is generally heterotopic, even in the case of the same scene shot by the same satellite at different times.

Grasping the meaning of an experimental correlation coefficient becomes a sensitive issue. This contribution focuses on these issues and by a geostatistical approach explains the different meaning of apparently equivalent operations. There is a need to deepen the geostatistical co-regionalization analysis to quantify and overcome the inaccuracies and uncertainties of any experimental correlation study. And the solution tool remains the modelling of the cross covariance for different supports.

REMOTE SENSING CHANGE DETECTION FOR THE APPEARANCE AND DISAPPEARANCE OF OASES FARMLAND IN CENTER ASIA

Buho Hoshino (1)*

Rakuno Gakuen University, Department of Environmental Sciences, College of Agriculture, Food and Environment Sciences, Ebetsu, Japan (1)
** Corresponding author: aosier@rakuno.ac.jp*

Abstract

The earth's surface changes are driving by climate change and human activates. Remote sensing techniques are very useful detecting and analyzing the change on the earth's surface. Change detection captures the spatial changes from multi temporal satellite images due to manmade or natural phenomenon. It is of great importance in remote sensing, monitoring environmental changes and land use –land cover change detection. Remote sensing satellites acquire satellite images at varying resolutions and use these for change detection. This paper focusses 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 and visualized them using RGB color combined techniques. The results show that agricultural land is 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. 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 using Landsat TM and Sentilel-2 satellite imagery and identify of vulnerability to the disappearance of oases farmland. China's investment in agriculture could lead to the depletion of water resources in the region. Because oases agriculture is one of the most vulnerable anthropogenic landscapes to climate change and human activates. Central Asia is one of the arid regions highly vulnerable to water scarcity. 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 with water taken from local rivers 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. Therefore, many irrigated farmlands were abandoned due to the collapse of the former Soviet Union. However, China's investment in Kazakhstan agriculture is cultivating once abandoned agricultural land and developing new oases agricultural land. 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. Locals want to invest in China OBOR, but depletion of water resources could put the region in poverty again in the future.

MAPPING OF CRITICAL RAW MATERIALS IN BAUXITE MINING RESIDUES USING GEOSTATISTICS AND REMOTE SENSING

Sara Kasmaeeyazdi (1)* - Adriana Guatame-Garcia (2) - Francesco Tinti (1) - Mike Buxton (2) - Emanuele Mandanici (3) - Joachim Schick (3) - Francoise Bodenan (4) - Dimitris Sparis (5) - Efthymios Balomenos (6) - Roberto Bruno (1)

DICAM, University of Bologna, Bologna, Italy (1) - Technical University of Delft, Delft, Netherlands (2) - Orano Company, Limoges, France (3) - BRGM, Orleans, France (4) - National Technical University of Athens, Athens, Greece (5) - Mytilineos, Athens, Greece (6)
** 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 in Greece.

Simulation results have been obtained for the Vanadium grade variability maps in the Bauxite tailings for 1000 realizations using infield samples as direct and Sentinel-2 images as an auxiliary 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 they have shown a coherent convergence. 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.

DETECTION OF GROUND DEFORMATION EVENTS: A METHODOLOGY FOR THE STATISTICAL ANALYSIS OF INSAR TIME SERIES

Laura Pedretti (1)* - Massimiliano Bordoni (1) - Valerio Vivaldi (1) - Silvia Figini (2) - Matteo Parnigoni (3) - Alessandra Grossi (3) - Luca Lanteri (4) - Mauro Tararbra (4) - Nicoletta Negro (5) - Claudia Meisina (1)

University of Pavia, Department of Earth and Environmental Sciences, Pavia, Italy (1) - University of Pavia, Department of Political and Social Sciences, Pavia, Italy (2) - Res It S.r.l., Milano, Italy (3) - Arpa Piemonte, Torino, Italy (4) - Regione Piemonte, Torino, Italy (5)

** Corresponding author: laura.pedretti01@universitadipavia.it*

Abstract

The A-InSAR Time Series (TS) interpretation is advantageous to understand the relation between ground movement processes (subsidence, slow-moving landslides) and triggering factors (snow, heavy rainfall), both in areas where it is possible to compare satellite TS with in-situ monitoring systems, and in areas where in situ instruments are scarce or absent. To identify areas of potential interest for significant ground deformations exploiting large datasets of satellite data, a new methodology ("ONtheMOVE" - InterpolatiON of SAR Time series for the dEtection of ground deforMatiOneVEnts) has been developed. The methodology has been tested on Sentinel-1 data available for the period 2016-2020 and covering Piemonte region, in an area prone to slow-moving landslides.

This work aims to classify the trend of TS (uncorrelated, linear, non-linear); to identify breaks in non-linear TS, i.e. those events (heavy rainfall, the melting of snow) when a significant modification of the TS trend is observed; to provide the descriptive parameters (beginning and end of the break, length in days, cumulative displacement, the average rate of displacement) that characterize the magnitude and timing of changes in ground motion.

To determine whether a TS has a trend, the Spearman statistical test is used. It is a non-parametric test which, based on the Spearman correlation index, determines whether two series are related to each other. Only TS that have evidence of any trend are considered in the discrimination between linear or non-linear trends.

To identify the type of the trend, four methodologies, based on two different approaches, are available: a statistical one (Terasvirta test and the White test) and a mathematical-modelling one (Polynomial-Pl and Polynomial Moving Average-PlMa). For the statistical approach, the function that synthesizes the data is composed of a non-linear and a linear part and after having optimized the coefficients, it checks if the coefficients of the non-linear part are zero by means of a score test. For the mathematical-modelling approach, the two methods are based on the interpolation of data with polynomials.

This innovative methodology can be applied to any type of satellite datasets characterized by low or high-temporal resolution of measures, it can be tested in any areas to identify any ground instability (slow-moving landslides, subsidence) at local or regional scales. It provides a supporting and integrated tool with conventional methods for planning and management of the area, both for back analysis and for near real-time monitoring of the territory and it can be helpful as regards the characterization and mapping of the kinematics of the ground instabilities, the assessment of susceptibility, hazard and risk.

DERIVING HIGH SPATIAL RESOLUTION DAILY VEGETATION INDEX IMAGES FROM SENTINEL AND MODIS DATA: A GEOSTATISTICAL APPROACH

Alzira Ramos (1)* - Leonardo Azevedo (2) - Cristina Branquinho (3) - Gregory Duveiller (4) - Maria João Pereira (2)

Instituto Superior Técnico e Faculdade de Ciências da Universidade de Lisboa, Universidade de Lisboa, Lisboa, Portugal (1) - Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal (2) - Faculdade de Ciências da Universidade de Lisboa, Universidade de Lisboa, Lisboa, Portugal (3) - Joint Research Centre (JRC), European Commission, Ispra, Italy (4)
** Corresponding author: alzira.ramos@tecnico.ulisboa.pt*

Abstract

Open oak woodland is an agroforestry system found in drylands where the extreme temperatures and the scarcity of water are becoming more frequent and more extensive, which may affect the ecosystem productivity and the services it provides. The landscape of open oak woodland is characterized by a low density of trees. The decline of open oak woodland in Southern of Portugal is a problem widely accepted by the scientific community. Many factors have been identified as contributing to this process, but a deep understanding that could help setting effective actions to mitigate and/or revert the process is still lacking.

The analysis of vegetation phenology (i.e. the annual cycle of vegetation growing) over time and space using Earth Observation (EO) data can provide important clues to better understand resilience and adaptation of open oak woodland to present climate changes and predict the long-term effects. That is the case of Moderate Resolution Imaging Spectroradiometer data (MODIS), which is acquiring data since 2000 and delivers daily information with ~250m pixel size. However, the relative coarse spatial resolution presents limitations in the space domain for these studies. More recently, the Sentinel 2 A/B, developed under the Copernicus Program, was launched in 2015 and 2017. This system has 2-3 days temporal frequency in mid-latitudes. These images have a high spatial resolution, of about ~10m pixel size, but the time-frequency approximates that of the MODIS.

The proposed methodology simulates daily radiometric vegetation index images, on the spatial resolution of Sentinel imagery using geostatistical stochastic simulation. We assume that the pixel value observed by MODIS can be approximated by applying the instrument's point spread function (PSF) over Sentinel 2 pixels observed at the same date. Then, we use an inverse modelling approach and direct sequential simulation with local means to simulate the daily images on the Sentinel 2 spatial resolution. This is an iterative process which starts by estimating the NDVIs from historical data and characterizing the relationship between MODIS and Sentinel 2 pixels. Then, using the MODIS vegetation index image we derive the map of local means and simulate in the spatial resolution of Sentinel 2. The images are upscaled using the PSF and compared with MODIS pixels. The mismatched between predicted and observed images are used to generate a new set of images at the Sentinel 2 spatial scale. The process iterates until satisfy to a convergence criterium.

A NOVEL METHODOLOGY FOR THE IDENTIFICATION OF ERRORS AND SIGNIFICANT ACCELERATION EVENTS FROM AUTOMATIC INCLINOMETERS TIME SERIES RELATED TO SLOW AND VERY SLOW-MOVING LANDSLIDES

Valerio Vivaldi (1)* - Massimiliano Bordoni (1) - Laura Pedretti (1) - Mauro Tararbra (2) - Luca Lanteri (2) - Matteo Parnigoni (3) - Alessandra Grossi (3) - Silvia Figini (4) - Nicoletta Negro (5) - Claudia Meisina (1)

University of Pavia, DISTA, Pavia, Italy (1) - Dipartimento Tematicogeologia e Dissesto, Arpa Piemonte – Agenzia Regionale per la Protezione Ambientale, Turin, Italy (2) - Res It Srl, Milan, Italy (3) - Department of Political and Social Sciences, University of Pavia, Pavia, Italy (4) - Regione Piemonte, Turin, Italy (5)
** Corresponding author: valerio.vivaldi@unipv.it*

Abstract

In-place automatic inclinometers are typical devices used to monitor displacements of extremely slow to slow-moving landslides.

The landslides monitored by automatic inclinometers in Piemonte region (Northern Italy) are characterized by slow and very slow deformation velocity, with rate of displacement ranging between 1.6 m/yr and 13 m/month and 16 mm/yr and 1.6 m/yr.

This work aimed to develop a novel method of displacement data analysis acquired by automatic inclinometers from the RERCOMF dataset (landslide regional monitoring networking) of ARPA Piemonte.

The methodology was developed in four steps. i) Evaluation of the reliability of the instrument: errors and instrumental anomalies were identified and erased by the time series scanning through a R based algorithm, carrying out the standard deviation of the time series movements standard deviations. At the same time the algorithm analyses potential anomalies based on Azimuth angle displacement from a specific Local Azimuth, assigned to each landslide and related to the main direction of landslide displacement. All values exceeding+/-30° the landslide Local Azimuth are considered anomalies and they are removed from the time series. ii) Individuation of significant moments of acceleration (events) in the rate of landslide displacement, retrieving in the time series a change in the rate of cumulated movements and velocities for a large range of duration. An event was indeed defined as a period, characterized by an increase in the rate of velocity respect to the typical trend of the deformation measured by an inclinometer. iii) The significant acceleration events clustering allowed to classify landslides in terms of total displacement, duration and average velocity. Duration and displacement data of the identified events were then implemented in a Hierarchical Cluster Analysis (HCA) through the software Orange 2.7, aiming to group events with similar kinematic behaviour during active phases corresponding to events.

iv) Reconstruction of thresholds of the triggers which influence the change in the rate of displacement to forecast the possible rate of displacement, according to the values of the main triggers influencing the deformation pattern of a phenomenon, considering the landslide geomorphological features and the complex groundwater hydrodynamics, that generally determine complicated hydro-mechanical relationships between rainfalls, groundwater table depth, snow-melting and the resulting deformation.

This method has shown excellent results in terms of time series errors identification, similar for both landslides characterized by seasonal movements and for landslide with strong response towards intense meteorological

events. The R based algorithm is capable to filter and analyse the time series of automatic inclinometers, starting from the raw data and giving back the landslide kinematics.

A NONPARAMETRIC SPATIO-TEMPORAL APPROACH TO EVALUATE UNCERTAINTY IN GEOSTATISTICS

Claudia Cappello (1)* - Sandra De Iaco (1) - Sabrina Maggio (1) - Monica Palma (1)

University of Salento, Lecce, Italy (1)
** Corresponding author: claudia.cappello@unisalento.it*

Abstract

Improving the methods of determining the spatiotemporal distribution and uncertainty of environmental variables can provide considerable benefits when developing risk assessments and strategic policies. In the nonparametric geostatistical framework, the uncertainty assessment methods include the histogram via entropy reduction.

In particular, in the spatial context, the histogram via entropy reduction method has been recently proposed as a novel nonparametric interpolation approach which can be used for estimating threshold-exceeding probabilities without assuming any underlying distribution for the data under study, and avoiding the modeling stage of the spatial correlation shown by the data. Hence, this approach is just based on the empirical probability distribution to quantify the spatio-temporal dependence and it minimizes entropy of the space-time predictions. The histogram via entropy reduction methodology is in between the geostatistical and statistical learning approach since the estimations are based on the empirical data recalling some fundamental aspects of Geostatistics.

Some advances of this technique already applied in the spatial context, can be further proposed for the spatio-temporal domain and the advantages in using this method are thoroughly pointed out in a case study concerning spatiotemporal measurements of an environmental variable. This includes the following main steps: i) evaluation of spatio-temporal dependence, ii) definition of an aggregation method, and iii) prediction of the target conditional probability distribution. A comparison with indicator kriging results is also discussed.

Acknowledgments

This contribution is supported by the Project "Rigenerazione Sostenibile dell'agricoltura nei territori colpiti da Xylella fastidiosa".

COVARIANCE MODELING FOR SPATIO-TEMPORAL COMPLEX-VALUED RANDOM FIELDS

Sandra De Iaco (1)*

University of Salento, Lecce, Italy (1)
** Corresponding author: sandra.deiaco@unisalento.it*

Abstract

In Geostatistics, the theory of complex-valued random fields is often used to provide an appropriate characterization of vector data with two components. In this context, constructing new classes of complex covariance models represents a goal of particular interest in the scientific community and in many areas of applied sciences, such as in electrical engineering, oceanography or meteorology, since they are used in structural analysis and, then for stochastic interpolation or simulation.

In the literature, there are various contributions focused only on modeling the spatial evolution of vector data with a reasonable representation on a complex domain. However, the temporal perspective is analyzed separately or used to model time-varying complex covariance models. However, it is surely challenging to propose some advances in modeling the joint spatial and temporal behavior of phenomena, whose decomposition in modulus and direction is natural.

After introducing the theoretical background regarding the complex formalism of a spatio-temporal random field, some techniques for building new families of spatio- temporal models are discussed. Then, the spatio-temporal complex modeling is applied to sea current data referred to the US East and Gulf Coast. The results regarding a comparative analysis between different complex-valued covariance models are also presented.

VARIABLE KERNEL ESTIMATES OF FIRST-ORDER SUMMARY DESCRIPTOR OF SPATIO-TEMPORAL POINT PROCESSES

Jonatan A. González (1)* - Paula Moraga (1)

King Abdullah University of Science and Technology (KAUST), Thuwal- Jeddah, Saudi Arabia (1)
** Corresponding author: jonathan.gonzalez@kaust.edu.sa*

Abstract

Spatio-temporal statistical methodologies have been widely applied, developed and demanded in epidemiology. Point process theory offers an appropriate scenario to analyse the spatio-temporal variation of the expected number of disease cases from information collected at the individual level, that is, reports of residential location and onset.

We illustrate an application of point process tools to study deaths caused by COVID-19 disease in the Santander-Colombia metropolitan area. Specifically, we explore the geographical distribution of the number of fatalities per unit area through the estimation by using adaptive kernels (kernels with variable bandwidth).

Adaptive kernels make it possible to capture the local variation in the intensity without the bias problem induced by high clustering and without relying on a single bandwidth: a sensible choice with critical consequences on the estimation quality. Lastly, we address the issue of the efficient computation of the adaptive spatio-temporal estimator.

GEOSTATISTICAL ANALYSIS OF EXTREME PRECIPITATION RECORDS OVER NORTH-WEST ITALY

Paola Mazzoglio (1)* - Ilaria Butera (1) - Pierluigi Claps (1)

Politecnico di Torino, Turin, Italy (1)
** Corresponding author: paola.mazzoglio@polito.it*

Abstract

Regional rainfall frequency analyses are based on rain gauge data that are affected by spatio-temporal discontinuities and gaps that can significantly influence the results of the statistical analyses. Neglecting the shorter series leads to ignore an amount of information that can be essential to correctly understand the spatial variability of the extremes.

In this work we present an application of the patched kriging technique, a year-by-year application of ordinary kriging equations, that overcomes the data inconsistency by considering all the records, independently on the length of the time series. The methodology is applied to short-duration (1 to 24 hours) annual maximum rainfall depths recorded by rain gauges coming from the recently-released Improved Italian – Rainfall Extreme Dataset (I2-RED). The trend with elevation is removed and, for each duration, the sample variogram is evaluated as the mean of the annual variograms weighted on the number of active rain gauges for any year.

The sequential application of the ordinary kriging allows to reconstruct a “rainfall data cube” and a “variance data cube” in the (x, y, t) space. By coring the cube along the t-axis, a complete series of measured and estimated values is obtained at each location. The cored series are then analyzed using the L-moments, weighted on the related series of kriging variance, to consider the different nature of the data (measured and estimated).

To overcome inconsistency of the L-moment statistics, a bias-correction procedure is introduced, that preserves the coefficient of variation from the smoothing effect induced by the spatial interpolation.

The methodology is applied to short-duration annual maximum rainfall depths in the whole North-West of Italy, that includes areas affected by the most severe extremes on record. The dataset used in this study covers the period 1928-2021, including the all-time Italian record events up to now, some of which observed in 2021 (377.8 mm / 3h, 496 mm / 6h, 740.6 mm / 12h).

A MARKED POINT-PROCESS SPATIO TEMPORAL MODEL TO UNDERSTAND FOREST FIRES IN THE MEDITERRANEAN BASIN

Oscar Rodriguez De Rivera Ortega (1)* - Juncal Espinosa Prieto (2) - Javier Madrigal (3) - Marta Blangiardo (4) - Antonio López-Quílez (5)

University of Kent, Canterbury, United Kingdom (1) - Escuela Tecnica Superior de Ingenierias Agrarias, Universidad de Valladolid, Palencia, Spain (2) - Centro Superior de Investigaciones Agrarias, Madrid, Spain (3) - Imperial College of London, Department of Epidemiology and Biostatistics, London, United Kingdom (4) - Universitat de Valencia, Department of Statistics and Operational Researc, Valencia, Spain (5)
** Corresponding author: o.ortega@kent.ac.uk*

Abstract

Many areas across the world have seen a rise in extreme fires in recent years. Those include South America and southern and western Europe. They also include unexpected places above the Arctic Circle, like the fires in Sweden during the summer of 2018. The Mediterranean area is no stranger to these changes and fires have become larger and more frequent.

Wildfires have already been studied with point process approaches to assess how the spatial heterogeneity of wildfires observed over a given time interval depends on the spatial distribution of land use information such as vegetation, urban zones or wetlands. In practice, raw data of environmental covariates is usually only available at very specific spatial and temporal scales, often with very high resolution, and comes in different numerical formats. Appropriate preprocessing of such data is important to obtain good predictive models. The high spatio-temporal dimension of wildfire data (observed occurrences and control cases without occurrences) has often been coped with by separating the data into subsets or by strongly aggregating them, by year or by spatial areas. Some recent approaches have concentrated more strongly on studying the interplay of the spatial and temporal structures, or on the usefulness of a specific Fire Weather Index aggregating weather data.

We here use log-Gaussian Cox process models, which have already been identified as useful models for wildfires since they allow capturing spatio-temporal aggregation structures through random effects.

Fitting spatial point process models to some spatial patterns is computationally intensive due to - amongst other things - the large number of individual points in the data set. Here, we consider a rather different situation. In some applications difficulties arise since point patterns with only a very small number of points can be collected, due to logistic limitations (e.g. for reasons of accessibility). These patterns are sometimes too small to justify the modelling of a single pattern. However, if replicates exist, a joint model of all replicates with a factor that accounts for variability among replicates caused by different conditions on different days may be more suitable.

Bayesian inference for log-Gaussian Cox processes using the integrated nested Laplace approximation (INLA) is now well-established, but remains challenging with the high dimension of our regression model. Instead of simplify the problem building a grid structure to populate the information in our work, we explore a scenario in which complex spatio-temporal structures can be incorporated into the analysis of a deeper understanding of forest fires. We follow the spread of forest fires across the Mediterranean basin, examining the role of multiple random fields in capturing spatial-clustering dynamics in the fires distribution across the Mediterranean between 2003 and 2013. Here, the point pattern of forest fires also reflects the observation

process. Accounting for spatially varying detection probability is a particular strength of inlabru, which was developed specifically for (ecological) datasets with complex observation processes.

AN ENSEMBLE-BASED APPROACH FOR THE ANALYSIS OF SPATIALLY MISALIGNED DATA

Ruiman Zhong (1)* - Paula Moraga (1)

King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia (1)
** Corresponding author: ruiman.zhong@kaust.edu.sa*

Abstract

In recent years, spatial data from satellite imagery, remote sensing, monitoring stations, and surveys can be collected in large quantities and at high spatial resolutions. The analysis of these data is crucial for decision-making in many disciplines such as epidemiology, climatology, and the environment. An ongoing challenge when analyzing this type of data is the spatial data misalignment problem which occurs when data at different spatial scales need to be combined. In this work, we propose a new ensemble-based approach for the analysis of spatially misaligned data that combines multiple statistical and machine learning approaches including Bayesian hierarchical models, geostatistical models, random forest, and gradient boosted trees. Our new approach improves prediction and propagates uncertainty from individual models for better uncertainty quantification. We assess the predictive performance of the ensemble-based approach by conducting a simulation study. Specifically, we generate spatial datasets that can appear in real settings, fit the individual and ensemble approaches, and compute a number of prediction error measures using spatial cross-validation designs. The new approach is also used to predict fine particulate matter emissions (PM2.5) in the UK in 2020 using data obtained from monitoring stations and satellite-derived environmental indicators. Our results show that the new proposed ensemble-based approach combines the strengths and drawbacks of each base model and result in a better final prediction. Moreover, the ensemble approach also shows robustness and generalization and avoids extremely deviant predictions. We believe our approach can enhance the reliability of predictions of outcomes obtained by combining multiple spatially misaligned data and can help decision-making in a wide range of disciplines.

ROBUST STATISTICAL PROCESSING OF LONG-TIME DATA SERIES ON SOIL-ATMOSPHERE INTERACTION: PRELIMINARY RESULTS

Mirko Anello (1) - Marco Riani (1) - Fabrizio Laurini (1) - Marco Bittelli (2) - Massimiliano Bordoni (3) - Claudia Meisina (3) - Roberto Valentino (1)*

University of Parma, Parma, Italy (1) - University of Bologna, Bologna, Italy (2) - University of Pavia, Pavia, Italy (3)
** Corresponding author: roberto.valentino@unipr.it*

Abstract

Large-scale quantitative assessment of water resources, useful in hydrology, hydrogeology, agriculture and other fields, is generally carried out using models that take into account soil-atmosphere interaction and the hydraulic behaviour of the soil. In particular, the shallow part of the soil is normally in unsaturated condition. The water content and the water potential in the soil are the main quantities to be considered in the evaluation of the hydraulic behaviour of unsaturated soil in relation to rainfall events. These quantities are very useful, because they constitute the input data for different types of water balance models. Most of these models are based on the knowledge of Soil Water Characteristic Curves (SWCC) which indicate a "two-way" link between soil water content and water potential. These curves have one or more hysteresis, linked to the drying and wetting cycles that soils undergo under natural conditions. Experimental evidence shows that the "biunivocal" link between the quantities considered does not allow the real behaviour of a partially saturated soil to be adequately reproduced. In the modelling chain, this mismatch between model and physical reality can lead to an inappropriate estimate of water resources in relation to rainfalls.

The aim of this research is providing a detailed description of interaction between soil and atmosphere. In particular, we intend to find a new function linking the quantities involved in the phenomena, namely soil volumetric water content, soil-water potential, air temperature, rainfalls, solar radiation. To achieve this goal, we treat long time series of field experimental data from continuous monitoring over a long period at two experimental sites, representative of two different geological contexts in Oltrepò Pavese. These data are treated in the framework of robust statistics by using the combination of robust parametric and non parametric models (LTS, SSA and SARIMA). We show that the fitted models can capture the relevant features present in the data and therefore can be used for prediction purposes.

USING GEOSTATISTICAL METHODS TO HELP OPTIMIZING AN EXISTING GROUNDWATER MONITORING NETWORK

Nathalie Courtois (1), Claire Faucheux (1)

French Alternative Energies and Atomic Energy Commission (CEA), Laboratory of Modelling of Transfers in the Environment (LMTE), Saint-Paul-Lez-Durance, France (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.

CRITICAL COMPARISON OF THREE COMPOSITIONAL INDICES TO TRACE GEOCHEMICAL CHANGES DOWNRIVER

Caterina Gozzi (1)* - Antonella Buccianti (1)

University of Florence, Department of Earth Sciences, Florence, Italy (1)
** Corresponding author: caterina.gozzi@unifi.it*

Abstract

River water and sediment composition serve as a sentinel to monitor the complex responses of watersheds across time and space. Watersheds exhibit a wide variety of spatial heterogeneity in landscape properties, inherent nonlinearity of many geohydrological processes and vary over time for the effects of climate changes and disturbance by human activities. Progress in developing new methods able to capture process interactions and comprehensive behaviors may favor an enhanced prediction of watershed's response to perturbing factors.

This work presents a critical comparison among three statistical indices developed under the Compositional Data Analysis theory (Aitchison, 1982): i) the cumulative sum of unclosed perturbation factors of each composition (row sum) with respect to a pristine reference, ii) the robust Mahalanobis distance, describing the compositional differences from the same reference and, iii) the geometric mean of the entire composition. All these measures are supposed to monitor the collective evolution of a geochemical composition, keeping record of the interactions among constituents thus allowing to go beyond the analysis of single variables. The performance of these indices was tested to examine source-to-sink compositional changes in the surface water and stream sediment composition of the Tiber River, the third-longest Italian river (Gozzi & Buccianti, 2021).

The results allowed to understand how geochemical footprints propagate downstream and compare the fluctuations with the variations in the drained lithotypes and other external drivers (e.g., soil use, human impact and morphometric parameters). All indices provide consistent results, especially if the chemical species having a high variability are treated separately and low values in the dataset are rare. Under this latter condition, the geometric mean of the composition shows a high correlation with the cumulative sum of unclosed perturbation factors. This evidence suggests that the geometric mean could find potential applications as a simple and effective monitoring index of the complex relationships among the involved constituents. Conversely, the robust Mahalanobis distance occasionally diverges from the other two measures and its application is recommended for larger datasets. These methods could be applied to different river basins worldwide and may facilitate the comprehension of their complex responses to potential hydrogeochemical threats.

Aitchison, J., (1982). The Statistical Analysis of Compositional Data (with discussion). Journal of the Royal Statistical Society Series B, 44(2), 139–177.

Gozzi, C. & Buccianti, A., (2021). Assessing Indices Tracking Changes in River Geochemistry and Implications for Monitoring. Natural Resource Research. Accepted for Publication (NARR-D-21-00982R2).

A SINGLE SPATIOTEMPORAL FRAMEWORK FOR MONTHLY LOW FLOW ESTIMATION IN AUSTRIA – INCLUDING EMPIRICAL ORTHOGONAL FUNCTIONS, VARIABLE SELECTION, AND FUNCTIONAL CLUSTERING

Johannes Laimighofer (1)* - Michael Melcher (2) - Gregor Laaha (1)

Institute of Statistics, University of Natural Resources and Life Sciences (BOKU), Vienna, Austria (1) - Institute of Information Management, Fh Joanneum, University of Applied Sciences, Graz, Austria (2)
** Corresponding author: johannes.laimighofer@boku.ac.at*

Abstract

A range of statistical models have been introduced for spatiotemporal modeling of environmental problems. However, their implementation to streamflow is still rare. The marginal use is likely due to the specific tree-wise structure of river networks, which poses particular challenges. Additionally, catchments vary in several aspects as anthropogenic influences, variability of meteorological conditions and heterogeneity of the catchment characteristics. To our knowledge there exists no study, that tries to model monthly low flow in one spatiotemporal model. Therefore, we propose to adapt an approach originally used for air-pollution modeling (Lindström et al. 2013, Szipro et al. 2010) for modeling monthly low flow in Austria.

The 260 gauging stations used for this study are located in Austria. All these stations are consistently monitored between 1982 and 2018 and they cover about 60% of the national territory of Austria. Out of each daily time series we calculate the 5% monthly quantile – monthly Q95 – and standardize the values by each catchment area (L s-1 km -2). Our approach is based on temporal empirical orthogonal functions (EOFs), that should capture the temporal part of the model. The EOFs are weighted by a universal kriging structure. A seasonal, annual, or overall trend can be included by spatiotemporal covariables, which are based on meteorological data. Variable selection for the universal kriging structure and the spatiotemporal covariables is performed by a linear boosting model and a bootstrapping approach. Finally, the space-time residuals are estimated by kriging. We propose to not solely use geographic coordinates but extend kriging to a physiographic space – using principal components and partial least square components. To avoid one large heterogeneous region in the single model framework, we added a functional clustering approach for smaller homogeneous regions. Cluster membership is estimated by conditional forest.

We found that this single model framework can yield high prediction accuracy of a cross validated R2 of 0.8. The main performance gain can be reached by dividing the study area through functional clustering.

References
Johan Lindström, Adam Szpiro, Paul D Sampson, Silas Bergen, and Lianne Sheppard. Spatiotemporal: An r package for spatio-temporal modelling of air-pollution. Journal of statistical software (in press), http://cran. rproject.org/web/packages/SpatioTemporal/index.html, 2013.
Szpiro, A. A., P. D. Sampson, L. Sheppard, T. Lumley, S. D. Adar, and J. D. Kaufman (2010). Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics 21(6), 606-631.

COUPLED SURFACE-SUBSURFACE HYDROLOGICAL MODEL FOR THE ESTIMATION OF NET RECHARGE OF THE KONYA CLOSED BASIN, TURKEY

Onur Cem Yologlu (1)* - Nadim Copty (1) - Izel Uygur (1) - Mehmet Can Tunca (1) - Elif Bal (1) - Buse Yetisti (2) - Irem Daloglu (1) - Ali Kerem Saysel (1)

Bogazici University, Institute of Environmental Sciences, Istanbul, Turkey (1) - Universitat Politècnica de Catalunya, Barcelona, Spain (2)
** Corresponding author: ncopty@boun.edu.tr*

Abstract

The accurate estimation of aquifer recharge is essential for the sustainable utilization of groundwater resources. This task is particularly challenging for large watersheds where the definition of the surface-subsurface system parameters is associated with high level of uncertainty. In this paper a water flow model, based on the coupling of the HYDRUS and MODFLOW computer programs, is developed for the Konya closed basin, one of the major agricultural regions of Turkey. Groundwater levels in the semiarid basin have experienced rapid decline in recent years due to excessive over exploitation. The problem is exacerbated by the large number of unregulated groundwater extraction wells, estimated to exceed 70,000 in number, rendering the estimation of actual groundwater extraction rates highly uncertain. The regional model, covering an area of about 62,000 km2, simulates unsaturated vertical flow through the vadose zone and groundwater flow through the underlying aquifer system. The model combines data collected at different scales including point groundwater level data, meteorological data and pumping test data, as well as land use, surface topography and water content (GRACE) satellite data. To better define the precipitation spatial distribution, cokriging of precipitation and topography data was used. Groundwater extraction data were estimated based on historical agricultural yields and crop water demands. Historical groundwater level data were used as the main calibration parameter. Results show that the model was able to reproduce the observed groundwater level trends of the past 2 decades. Recharge areas are mostly located in the higher elevation regions of the basin with water flowing towards the central portions of the plain where intensive irrigation is located. The modeling results underscore the impacts of the expansion of irrigated lands and the switch to more water-demanding crops on the basin's overall water deficits.

IMPORTANCE OF MULTI-PARAMETER APPROACHES IN THE DEVELOPMENT OF MACHINE LEARNING ALGORITHMS FOR LANDSLIDE DISPLACEMENT FORECASTING

Marco Conciatori (1)* - Alessandro Valletta (1) - Andrea Segalini (1)

University of Parma, Department of Engineering and Architecture, Parma, Italy (1)
** Corresponding author: marco.conciatori@unipr.it*

Abstract

Landslides are a constant source of danger for people and the built environment. Even with more recent advancements in this field, these natural hazards still represent a serious problem worldwide and the effort to mitigate their impact has driven the development of specific knowledge, technology, and practices.

One of the current challenges involves the ability of new sensors and monitoring techniques to provide a considerable amount of information, which should necessarily be addressed with automatic procedures for their elaboration. A possible approach relies on the introduction of algorithms from the field of Machine Learning applied to the interpretation of the landslide behavior. ML algorithms have key advantages in this context: they do not need explicit knowledge or models of the problem, since they learn directly form the examination of data. They scale with data, so that they get more accurate when presented with large amount of information. Moreover, they can adapt to learn from many different sources like numeric measurements, images, text, or sound.

The methodology here discussed was designed for landslide monitoring and early warning activities based on hydrogeological information, with the objective to predict the monitored site behavior few days in advance. In particular, input data for this process are displacement measurements, water level, and meteorological conditions. The model relies on an Artificial Neural Network that will read the values of these parameters measured over a predefined number of consecutive days, predicting the displacement of the day following the last observed. While the model is training, the inputs are selected from past data, so that the model's prediction can be compared with already available measurements. The error in the prediction is used to adjust the algorithm until it starts to make accurate forecasts. Once the model has reached this stage, it can be shown the measurements of the last days, so that in will predict the development of the slope a few days from now. Moreover, the presence of several sensors on the studied site could give the possibility to assess the landslide behavior in sectors where no monitoring tool is present, thanks to spatial interpolation procedures performed on forecasted displacements.

For these types of algorithms, the addition of irrelevant information can lead to lower accuracy in the results. Theoretically, with a large enough dataset, the model should become able to distinguish between useful and inconsequential inputs, in practice however it is very hard to have the necessary tools to achieve that. Accounting for that, the training process will be performed on models including different subsets of the available data, in order to identify the informational value of cross-correlations between monitored parameters.

MODELING DENSITY DATA ACROSS THE IBERIAN MASSIF (PORTUGAL): RELIABILITY ASSESSMENT FOR ENVIRONMENTAL APPLICATIONS

Filipa Domingos (1)* - Gustavo Luís (2) - Sérgio Sêco (1) - Alcides Pereira (2)

Laboratory of Natural Radioactivity, University of Coimbra, Department of Earth Sciences, Coimbra, Portugal (1) – CITEUC, Centre for Earth and Space Research of the University of Coimbra, University of Coimbra, Coimbra, Portugal (2)
** Corresponding author: lipa_domingos@hotmail.com*

Abstract

The Basic Safety Standards (BSS) for protection against the dangers pertaining to ionizing radiation were established in the Council Directive 2013/59/EURATOM. The main sources of ionizing radiation due to exposure to natural gamma-ray emitters, such as 40K and isotopes from the 238U and 232Th decay series, including radon (222Rn) and thoron (220Rn), indoors are the materials underlying the buildings (soil and/or bedrock) and the building materials. Research efforts have been focused mainly on the underlying bedrock materials. However, several authors have reported a significant contribution to the inhalation dose received by the population by 220Rn, whose main source are the building materials.

To assess the safety of building materials, the BSS requires firstly to estimate an activity concentration index denoted the I index from the activity concentrations of terrestrial radionuclides (226Ra, 40K and 232Th). This index is used as a screening tool to discriminate between "safe" and "unsafe" materials for use in construction. If a building material presents an index I greater than 1, then the gamma radiation dose must be estimated using other data such as the density, the thickness, and the intended use of the material, among other factors. Hence, it is essential to know the density of the materials to ensure reliable dose estimates. The density of the materials can be determined by several methods. Direct measurement of density may be performed using standardized methods whereas indirect assessments of density make use of the chemical and/or mineralogical composition of the materials. Despite the existence of easily applicable methods to determine density accurately in the case of rocks, density data are often drawn from literature according to lithology.

Geochemical databases constructed from local studies on rock geochemistry drawn from literature have been demonstrated effective to estimate the content of terrestrial radionuclides over geological units because they generally lack geospatial information, constraining the use of geostatistical tools. In this work we compare density results obtained using a standardized method (EN 1936:2006) to density results estimated from geochemical analysis results taken from literature using different models. Direct measurements of density ranging from 2100 kg/m3 to 3900 kg/m3 were drawn from a database with over 560 assays carried out in hand specimens of rock samples collected from northern to southern Portugal in the Iberian Massif (of Variscan age). The main objective is to determine whether geochemical results may provide accurate density estimates which may be used to improve dose estimates as well as in other applications, such as exploration by gravimetric methods and resource estimation. The reliability of density data estimated from geochemical data is discussed. Spatial correlations of density between and within geological units are assessed using geostatistical tools and the possibility of using density results estimated from literature data in spite of the lack of geospatial information is investigated.

ENVIRONMENTAL GEOSTATISTICS OF HEAVY METALS IN FINE ACTIVE SEDIMENTS OF OESTE ANTIOQUEÑO SUBREGION ANTIOQUIA - COLOMBIA.

Luis Hernán Sánchez Arredondo (1)* - Jheyson Andrés Bedoya Londoño (2) - Sergio Alejandro Garavito Higuera (2)

Universidad Nacional de Colombia, Materiales y Minerales, Medellín, Colombia (1) - Universidad Nacional de Colombia, Area Curricular de Recursos Minerales, Medellín, Colombia (2)
** Corresponding author: lhsanche@unal.edu.co*

Abstract

The subregion of west of Antioquia is characterized by having a low concentration of rural land ownership and a high process of ecological destruction and desertification in the municipalities of Santa Fe de Antioquia, San Jeronimo, Anza and Sopetran. Tropical dry forest biome, secondary vegetation and growing forests, sub-Andean forest, Andean forest, wetlands and associated forests, paramos and basal forest predominate.

One of the most transcendental environmental problems has been the transformation of natural ecosystems that have become mainly agroecosystems and due to population growth, due to the development of new projects among other anthropic activities. In the subregion, the main global greenhouse effect (GHG) emissions are from deforestation, livestock, agriculture, energy consumption and sanitation. Climate risk of the subregion is medium to low.

Through geostatistical techniques of simulation conditioned by turning bands, different scenarios were assessed for the analysis of environmental contamination of metals Co, Cu, Cr, Ni, Pb and Zn in an area of 7,291 km2, equivalent to the subregion of west from department of Antioquia. Geostatistical cartography prepared for the pessimistic case of each of the metals analyzed, report anomalous values above the standards established for the different types of rocks in the region, which would generate a moderate relative mobility for water contamination, being subjected to surface weathering processes under oxidation conditions in acidic environments. Under reducing conditions, the simulated elements would have a behavior of relative immobility and would offer environmental danger for the soils of the region.

Potentially dangerous anomalous values were estimated for Co ≥ 130 mg/kg, Cu ≥ 1,710 mg/kg, Cr ≥ 2,507 mg/kg, Ni ≥ 138 mg/kg, Pb 165 ≥ mg/kg and Zn ≥ 1,170 mg/kg, which are consistent with volcano-sedimentary sequences and acid to intermediate plutonism present in the subregion. Analysis of the samples were carried out by optical emission spectrography, which although it is a semi-quantitative method, has made it possible to formulate the state of the art of environmental geochemistry and identification of possible vulnerable areas in order to focus future studies, with greater precision and detail.

Regarding mining activity, 20% of the area of the subregion is concessioned for extraction of minerals and 34% is requested as mining concession contracts, whose activities of exploration and exploitation of georesources are concentrated mainly in Au and precious metals. Mine with the highest production is of the gold type and is in in the municipality of Buritica, which intends to extract 9.1 tons of gold each year; however, medium, and small-scale mining represents 93% of mining titles and applications in the area of influence.

If the high concentrations of Co, Cu, Cr, Ni, Pb and Zn in some soils of the subregion of west of Antioquia are validated, their use should be technically oriented to economic activities in the mining sector, instead of agriculture and livestock, due to the negative impacts of heavy metals on human health.

CONDITIONED SIMULATION FOR GEOSTATISTICALTREATMENT OF SEISMIC THREAT IN ANTIOQUIIA STATE, NORTHWESTERN COLOMBIA.

Luis Hernán Sánchez Arredondo (1)* - Lilian Posada Garcia (2) - Adiela Martínez (3)

Universidad Nacional de Colombia, Materiales y Minerales, Medellín, Colombia (1) - Universidad Nacional de Colombia, Geociencias y Medio Ambiente, Medellín, Colombia (2) - Sociedad Colombiana para la Defensa del Patrimonio Geológico y Minero-Metalúrgico, Nn, Ginebra, Suiza (3)
** Corresponding author: lhsanche@unal.edu.co*

Abstract

The Colombian Northwest region is a tectonically active territory located in the Colombian Oceanic Mobile Belt (Cuna and Calima Terrains). The Colombian Seismological Network was created in 1999. Since its creation up to 2018 it has registered around 11,000 earthquakes in the Antioquia state, 3,000 of those have surface activity with average magnitudes of 2 and a maximum of 5.2 on the Richter scale, which means an important surface activity related to active geological faults. A geostatistical analysis focused on the simulation by turning bands was elaborated, whose procedure was to transform available seismic data to Gaussians (Anamorphosis), elaborate a variographic analysis, generate the simulation and make the automatic cartography. The structural geostatistical analysis showed an anisotropic behavior with a strong nugget effect and two preferential directions according to azimuth N10° and N100°, where the first one is representative of the maximum continuity of the seismic activity in and is mostly coupled to the preferential direction of the geological faults in the Antioquia region. Three cases of analysis are presented: the average of 100 simulations, the pessimistic and the optimistic cases. For the pessimistic case, the greatest susceptibility to surface microseismic activity in Antioquia region are identified in Urabá (Turbo, Murindó), the far southwest (Urrao, Salgar- Ciudad Bolívar), the west (Frontino, Cañasgordas), and the north (Ituango, Sabanalarga). Historically, two macro-seisms stand out; the Turbo earthquake of September 7, 1882 at 3:20 a.m. local time of intensity X, which has been associated with the Mutatá Fault located in the area, with unquantified ecological damage. The second one occurred in Panama (former Colombian territory) with the collapse of buildings and loss of human life. The earthquake generated a tsunami that mainly affected the San Blas Archipelago, where several waves about 3 meters high caused numerous damages and 65 death people. Other secondary effects in nature such as ground cracking, liquefaction and landslides were also reported. The Murindó earthquake occurred on October 17th and 18th of 1992, with 6.6 Mw and 7.1 Mw magnitudes, respectively. These events caused significant damage in the region, where 32 municipalities were affected, 17 of which showed high damages in their infrastructure, severe ecological and social damages, for soil liquefaction, landslides, reactivation of the Cacahual volcano activity and the fully evacuation of Murindó municipality.

ANALYSIS OF SURFACE TEXTURE USING SPATIAL CONTINUITY INDICES: INTERPRETABILITY, FLEXIBILITY, AND ROBUSTNESS

Sebastiano Trevisani (1)*

University IUAV of Venice, Venice, Italy (1)
** Corresponding author: strevisani@iuav.it*

Abstract

Despite the long record of applications and the well-known theoretical framework, geostatistical based image/surface texture tools have still not gained a wide diffusion in the context of geomorphometric analysis, even for the evaluation of surface roughness. Depending on the disciplines and authors, roughness can be considered a synonym of surface texture (as in this presentation) or as a specific component of it (e.g., according to surface metrology), referred to short-range spatial variability of surface roughness. Surface roughness, in its general meaning, is a complex multiscale and directional property of topographic surfaces; accordingly, multiple roughness-related metrics can be defined. In this context, geostatistical spatial continuity indices are capable to characterize multiple aspects of surface texture, by means of interpretable metrics. Moreover, the geostatistical indices can be adapted and mixed with other geocomputational approaches, tailoring the algorithms for the geomorphometric analysis. In this presentation, an ad-hoc implementation for the analysis of high resolution digital terrain models is presented, outlining the main characteristics and potentialities of the approach. The considerations are valid also for image analysis, for example in the context of remote sensing and applied geophysics.

References

Atkinson, P.M. and Lewis, P., 2000. Geostatistical classification for remote sensing: An introduction. Computers and Geosciences, 26(4), pp. 361-371.

Balaguer, A., Ruiz, L.A., Hermosilla, T. and Recio, J.A., 2010. Definition of a comprehensive set of texture semivariogram features and their evaluation for object-oriented image classification. Computers and Geosciences, 36(2), pp. 231-240.

Guth, P.L., 2001. Quantifying terrain fabric in digital elevation models. GSA Reviews in Engineering Geology, 14, pp. 13-25.

Herzfeld, U.C. and Higginson, C.A., 1996. Automated geostatistical seafloor classification - Principles, parameters, feature vectors, and discrimination criteria. Computers and Geosciences, 22(1), pp. 35-41.

Trevisani, S., Cavalli, M. and Marchi, L., 2009. Variogram maps from LiDAR data as fingerprints of surface morphology on scree slopes. Natural Hazards and Earth System Science, 9(1), pp. 129-133.

Trevisani, S., Cavalli, M. and Marchi, L., 2012. Surface texture analysis of a high-resolution DTM: Interpreting an alpine basin. Geomorphology, 161-162, pp. 26-39.

Trevisani, S. and Rocca, M., 2015. MAD: Robust image texture analysis for applications in high resolution geomorphometry. Computers and Geosciences, 81, pp. 78-92.

Trevisani, S. and Cavalli, M., 2016. Topography-based flow-directional roughness: Potential and challenges. Earth Surface Dynamics, 4(2), pp. 343-358.

A

A. Godoy Vanessa 54; 79
A. Varouchakis Emmanouil 93
Albuquerque Teresa 2; 4; 66
Amirzhan Almas 23
Anello Mirko 117
Antunes Margarida 2; 4
Araújo Alexandre 2; 4
Araújo Joana 2
Azevedo Leonardo 64; 74; 78; 80; 93; 94; 96; 107

B

Bal Elif 121
Balomenos Efthymios 105
Barbi Chiara 31
Bárdossy András 42; 63
Barreto Bernardo Viola 94
Barroso Ana 2; 4
Bazarbekov Talgatbek 32
Bedoya Londoño Jheyson Andrés 124
Bersezio Riccardo 99
Bittelli Marco 117
Blangiardo Marta 114
Bodenan Francoise 105
Boente Carlos 66
Bogaert Patrick 21
Bohling Geoff 63
Bordoni Massimiliano 106; 108; 117
Bossew Peter 6
Branquinho Cristina 107
Bruno Roberto 103; 105
Buccianti Antonella 119
Bürkle Paula 94
Butera Ilaria 113
Buxton Mike 105

C

Cappello Claudia 15; 20; 110
Claps Pierluigi 113
Comunian Alessandro 99
Conciatori Marco 122
Consonni Edoardo 99
Copty Nadim 121
Costa João Felipe 94
Courtois Nathalie 118
Crozet Marielle 56

D

Daloglu Irem 121
Daly Colin 40
De Iaco Sandra 15; 20; 110; 111
Dhaliwal Gursimran 65
Domingos Filipa 123
Doody Jake 65
D'Oria Marco 98
Duarte Ana Filipa 64; 80
Duveiller Gregory 107

E

Espinosa Prieto Juncal 114

F

Fagandini Camilla 41
Faucheux Claire 118
Figini Silvia 106; 108
Fonseca Rita 2; 4
Friedli Lea 81

G

Garavito Higuera Sergio Alejandro 124
Gerassis Saki 66
Ginsbourger David 81
Giudici Mauro 99
Gómez-Hernández J. Jaime 54; 79; 93; 97; 98
González Jonatan A. 75; 112
Goovaerts Pierre 17
Gozzi Caterina 119
Grana Dario 82; 94
Gravey Mathieu 100
Greau Claire 18
Grossi Alessandra 106; 108
Guatame-Garcia Adriana 105
Guinote Rui 93

H

Hadjipetrou Stylianos 101
Haslauer Claus 63
Hörning Sebastian 42
Hoshino Buho 104
Huart Jean Pierre 21

I

Iooss Bertrand 56

J

Jago Alban 21
Joana Araújo 4
Jordão Helga 1
Journée Michel 21

K

Karatzas George P. 55
Kasmaeeyazdi Sara 103; 105
Krishnamurthy Ashok 65
Kyriakidis Phaedon 101

L

Laaha Gregor 120
Lacaux Céline 56
Laimighofer Johannes 120
Laloy Eric 92
Lanteri Luca 106; 108
Laurini Fabrizio 117
Lauzon Dany 83
Levy Shiran 92
Linde Niklas 81; 92
Lino Pereira João 93
López-Quílez Antonio 114
Luís Gustavo 19; 123

M

Madani Nasser 23; 32; 45
Madrigal Javier 114
Maggio Sabrina 20; 110
Mandanici Emanuele 105
Marcotte Denis 83
Mariethoz Gregoire 100; 101; 102
Marrel Amandine 56
Martínez Adiela 125
Matias Luis 80
Mazzoglio Paola 113
Meisina Claudia 106; 108; 117
Melcher Michael 120
Menafoglio Alessandra 31
Mendes Renato 80
Merembayev Timur 32
Metivier Jean-Michel 18
Miele Roberto 94
Molino Laura 95
Moraga Paula 43; 75; 112; 116
Mota Natália 2; 4

N

Nag Pratik 44
Nahla Mansouri 18
Napa-Garcia Gian F. 79
Narciso João 96
Nasretdinova Milena 45
Natália Roque 66
Negro Nicoletta 106; 108

O

Oliveira Mafalda 93
Oriani Fabio 102
Ortiz Julian 52

P

Palma Monica 15; 110
Palmero Tia 76
Parnigoni Matteo 106; 108
Parsekian Andrew 82
Pazo María 66
Pedretti Laura 106; 108
Peliz Álvaro 80
Pereira Alcides 19; 123
Pereira Maria João 74; 78; 107
Planchon Viviane 21
Posada Garcia Lilian 125
Powell Becky 76
Pu Qiang 76

R

Ramos Alzira 107
Reich Brian 44
Riani Marco 117
Ribeiro Amaral André Victor 75
Ribeiro Manuel 74; 78
Riquelme Alvaro 52
Roberts E. John 76
Rodriguez de Rivera Ortega Oscar 114
Rosillon Damien 21

S

Sánchez Arredondo Luis Hernán 124; 125
Sanz-Prat Alicia 97
Saysel Ali Kerem 121
Schick Joachim 105
Secchi Piercesare 31
Secci Daniele 95
Sêco Sérgio 123
Segalini Andrea 122
Soares Amilcar 1; 64; 74; 78
Sparis Dimitris 105
Sun Ying 44

T

Taboada Javier 66
Tanda Maria Giovanna 41; 98
Tararbra Mauro 106; 108
Tayebi Gholamzadeh Ashkan 53
Tinti Francesco 103; 105
Todaro Valeria 41; 98
Trevisani Sebastiano 126

Trichakis Ioannis......55
Tunca Mehmet Can......121

U

Uribe-Asarta Janire......54
Uygur Izel......121

V

Valente Teresa......2; 4
Valentino Roberto......117
Valletta Alessandro......122
Van De Vijver Ellen......96
Varella Luiz Eduardo......94
Varouchakis Emmanouil......55
Vivaldi Valerio......106; 108

W

Wai Crystal......65
Wieskotten Martin......56

X

Xiao Bo......63

Y

Yetisti Buse......121
Yologlu Onur Cem......121
Yoo Eun-Hye Enki......76

Z

Zanini Andrea......41; 95; 98
Zhong Ruiman......116
Zuffetti Chiara......99

Proceedings of

geoENV2022

14TH INTERNATIONAL CONFERENCE ON GEOSTATISTICS FOR ENVIRONMENTAL APPLICATIONS

PARMA, ITALY JUNE 22-24, 2022

Edited by **ANDREA ZANINI & MARCO D'ORIA**

The 14th International Conference on Geostatistics for Environmental Applications (geoENV2022) was held in Italy, at the Campus of the University of Parma. From June 22 to June 24, 2022, over 80 experts on geostatistics gathered to discuss about environmental applications of this discipline.

This book contains the abstracts and extended abstracts submitted to the conference and focusing on geostatistics applied to different fields such as: ecology, natural resources, environmental pollution and risk assessment, 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.

geoENVia

ALMA · UNIVERSITAS · STUDIORUM · PARMENSIS · A.D. 962
UNIVERSITÀ DI PARMA