O2S2: a new venue for computational geostatistics

Alessandra Menafoglio[a,*], Piercesare Secchi[a,b]

[a] *MOX, Department of Mathematics, Politecnico di Milano*
[b] *Center for Analysis Decision and Society, Human Technopole, Milano*

Abstract

Applied sciences have witnessed an explosion of georeferenced data. Object oriented spatial statistics (O2S2) is a recent system of ideas that provides a solid framework where the new challenges posed by the *GeoData revolution* can be faced, by grounding the analysis on a powerful geometrical and topological approach. We shall present a perspective on O2S2, as a fruitful ground where novel computational approaches to geosciences can be developed, at the very interface among varied fields of applied sciences – including mathematics, statistics, computer science and engineering.

Keywords: Object Oriented Data Analysis; Georeferenced data; Kriging; Computational statistics
2010 MSC: 62M30, 62H11, 62P12

1. Introduction: O2S2 for Modern Applied Geosciences

The availability of large amounts of data is shaping a new era for applied geosciences. Nowadays, field studies may not rely on small-scale datasets of scalar variables only, but rather on a multitude of complex datasets from different sources, which provide direct and indirect observations of the phenomenon under investigation. For instance, seismic monitoring relies on dense networks of measurement instruments *in-situ*, which typically record signals at high-frequency in time (i.e., functional data). Here, additional sources of information are represented by remote sensing data (e.g., satellite images), and soft data, such as those provided by the resident population via crowdsourced platforms (smartphone applications, social networks or online surveys, see, e.g., [1]).

Data analyses in these settings cannot ignore the data heterogeneity and complexity. Data streams, functional data, images, tensors, networks, and texts are few paradigmatic examples of the different types of data objects that may represent the core of the geostatistical analysis. In the GeoData deluge, the context where classical geostatistics was developed is rapidly disappearing, opening a new frontier for *GeoData Science*.

*Corresponding author
Email address: alessandra.menafoglio@polimi.it (Alessandra Menafoglio)

This framework is fostering a compelling need for innovative paradigms of analysis. Object oriented spatial statistics (O2S2, [2]) is a recent system of ideas that provides a solid framework where these new and revolutionary challenges can be faced, by grounding the analysis on a powerful geometrical and topological approach. O2S2 embraces the philosophy of object oriented data analysis (OODA, [3]), and is rooted in the interpretation of the data point (e.g., the curve, image, or network) as the *atom* of the statistical analysis. The data points (also called *data objects*) are thus modeled as points in a mathematical space – named *feature space* – that should properly represent the data characteristics, particularly their dimensionality and constraints.

In the context of O2S2, new challenging problems can be formulated and tackled, and classical paradigms of analysis reinterpreted (e.g., those based on variography, kriging and stochastic simulation), opening new venues for computational geostatistics. The intent of this paper is not to provide an exhaustive review or deep mathematical treatment, but rather a perspective on O2S2 as a paradigm for the development of new computational approaches to geosciences. We shall also open views on the varied contexts that are being challenged by the complexity of modern GeoData-driven problems, in areas of applied sciences well beyond the classical fields of application of geostatistics. The focus will be posed not only on the complexity of the data, but also on the potential complexity of the study domain (e.g., for its size or shape), with reference to the computational methods and software developed in O2S2.

2. A key role for the feature space

Among the pillars of O2S2, the feature space plays a key role. For instance, the operations (sum, product by a constant) defined in the feature space are key to the definition of linear predictors such as kriging [2]. As an example, we consider a set of data objects (e.g., functional data), collected at locations $s_1, ..., s_n$ in a spatial domain D, and denoted by $\mathcal{X}_{s_1}, ..., \mathcal{X}_{s_n}$. We represent the data as elements of a Hilbert feature space $\mathcal{F}$ (e.g., the space L^2), with operations $(+, \cdot)$, inner product $\langle \cdot, \cdot \rangle$, and associated norm $\| \cdot \|$. Loosely stated, in O2S2 for Hilbert data, the kriging predictor is defined as the *linear combination* of the data $\sum_{i=1}^{n} \lambda_i^* \cdot \mathcal{X}_{s_i}$ with 'optimal' scalar weights $\lambda_1^*, ..., \lambda_n^*$ (see [2]). Here, the form taken by the *linear combination* is precisely determined by the operations $(+, \cdot)$. On the other hand, the metric induced by the inner product in $\mathcal{F}$ implies a notion of similarity between data objects observed at nearby locations ($\|\mathcal{X}_{s_i} - \mathcal{X}_{s_j}\|^2$, $i, j = 1, ..., n$), which is instrumental in defining the variogram and the associated notions of stationarity (see [4, 2]). As a matter of fact, the feature space $\mathcal{F}$ should be selected as to properly represent the data characteristics that one is willing to account for in the analysis. In this perspective, a feature space selected for unconstrained functional data (e.g., the space L^2 or a Sobolev space), most likely will be inappropriate to represent tensor data or distributional data, such as probability density functions (PDFs). These latter types of data are not uncommon in the geosciences. For instance, particle-size

fractions (PSFs) and particle-size densities (PSDs, i.e., the continuous counterpart) are routinely used in hydrogeological studies and in all the areas of applied science where flow and transport phenomena are to be modeled, being related to the porosity and permeability of the medium. The analysis of PSFs (PSDs) requires to account for the fact that they are *closed data*, i.e., their sum (integral) is one. It has been widely recognized [5, 6, 7] that an approach which neglects this aspect and treats each component of a PSF separately (or fixed quantiles of a PSD) is affected by spurious correlations. Furthermore, it leads to biased results and inappropriate estimates. For instance, embedding PSDs in L^2 and building kriging predictors through its geometry most likely leads to negative kriged densities or totals different from one (e.g., [8], and references therein). All these issues arise because an Euclidean space is not the appropriate feature space for the analysis, as the data belong to a simplex. In the perspective of O2S2, the analyst should first focus on these geometrical properties of the data (positivity and closeness), and, on this basis, select an appropriate feature space – a possible choice being, for PSFs, the Aitchison geometry for compositional data in the simplex [CoDa 6] and, for PSDs, its continuous counterpart [9, 8]. In this vein, the feature space may not necessarily be finite-dimensional and Euclidean — the working assumptions of geostatistics — but could be an infinite-dimensional Hilbert space, a Riemannian manifold or a Banach space, if better representative of the data objects [4, 10].

3. O2S2 in action

The areas of potential application of O2S2 are varied. O2S2 has been used for the spatial prediction of particle-size distribution in heterogenous aquifers [8, 11], to model and forecast gas rate production curves in shale reservoirs [12], and, more recently, for the analysis, prediction and simulation of shaking fields generated from earthquakes events [13]. In fact, O2S2 can be used naturally in all those settings where compositional, symbolic and functional data analysis [6, 14, 15] approaches were already successfully introduced, such as geochemistry (see [7] and references therein), climatology [16], oceanography [17], water quality [18]. A short case study on water quality will be presented at the end of Sect. 5.

O2S2 also allows for kriging meta-modeling [19], enabling one to perform efficient uncertainty assessment in numerical models where the response is a complex object (e.g., a function of time or a field in space). This has found application in models for fluid flow in reservoirs [20, 21], but also in diffusion-reaction PDE models [22]. A similar approach is currently used to provide a full uncertainty assessment on a mathematical model for sediment transport in a mountain basin, within the SMART-SED project [23].

In all these contexts, calibrating the model inherently requires to take advantage of the rich but heterogenous set of information available at different sources, integrating the data collected *in situ* with those given at other open data repositories (e.g., on region geology, soil composition, land use). *Data fusion* – i.e., the process of combining information from multiple data sources

based on sound statistical models – is still one of the most challenging yet compelling issues to bring O2S2 into further action. A critical topic in this regard is definitely the *change of support* for the data, particularly the problem of downscaling (i.e., of bringing the data support to a smaller spatial scale). Developing effective downscaling methods in O2S2 will be the key to further broaden its potential in modern applied geosciences.

4. A GeoData revolution beyond classical applications

The advent of modern low-cost technologies for data collection and storage is fostering the GeoData revolution well-beyond Earth sciences. *Smart cities* are equipped with huge networks of sensors, which provide real-time information on various aspects of life in urban areas. Hot research topics in this field are those related to *urban mobility*, particularly for the development and optimization of shared approaches. In this context, urban dynamics of vehicles and people can be then inferred from the integrated analysis of large amounts of georeferenced digital 'contrails' and weak signals left by the users, such as mobile phones traffic data, social networks activity, or GPS locations collected from smartphone applications [24].

In our view, the GeoData revolution represents an incredible opportunity for knowledge dissemination across very disparate areas of science and engineering. For instance, GeoData are also widespread in the context of Industry 4.0, where data-rich environments are feeding the 4th industrial revolution. Production plants are becoming highly sensored, to allow for a real-time quality monitoring of the produced parts. In this broad context, additive manufacturing (i.e., 3D printing) is leading the industrial and statistical research, at the very frontier of statistical process control. Monitoring of parts is interpreted in this context as (real-time) analysis of data objects represented by complex shapes, often described by manifold geometries [25]. The challenge to take on in this framework is to allow for data-driven semi-automatic product and process monitoring, based on streams of high-frequency signals (e.g., videos [26]) or tomographic reconstructions [25], these data being naturally subject to spatial dependence (i.e., GeoData). Here, O2S2 has the clear potential to be successfully employed for modeling the spatial structure of complex data, in the same varied industrial settings were kriging and multi-fidelity paradigms have been already successfully introduced (see, e.g., [27] in free-form surface monitoring).

5. Complex data or complex domains?

Historically, the analysis of spatial data has been dominated by the use of global approaches to the field modeling, mostly based on the assumption that the generating process is stationary (or mildly non-stationary) and distributed over a Euclidean domain. However, in vast areas of geosciences, the proximity between data locations is naturally expressed through the shortest path (i.e., the geodesic) induced by the physics of the phenomenon, which may well be non-Euclidean. For instance, while measuring aquatic variables in a stream network

system, the closeness among monitoring sites should be represented through a *water distance* – i.e., the shortest path *within* water, or a distance accounting for the fluid flow in the system – rather than the Euclidean shortest path, which may pass through land.

The extensive availability of GeoData defined at large spatial scales, calls for innovative methodological and computational approaches able to deal not only with massive and complex data, but also with data distributed over general types of study domains. This area is the focus of active research in geostatistics. Indeed, whenever the metric on the spatial domain is non-Euclidean, widely-used parametric covariance families may no longer be valid [28]. In special cases, flexible classes of valid covariance models have been developed; these include the case of stream networks (e.g., [29, 30]) and spherical domains (e.g.,[31, 32, 33]), which naturally arise when dealing with global climate data (see, e.g., [34]). However, strategies based on the development of *ad hoc* valid models for the specific metric at hand seem hardly applicable in general contexts. Recent literature has shown that overcoming the issue is possible by using different modeling or computational approaches in the analysis. Relevant contributions in this sense are those encoding the spatial dependence precisely through the physics of the phenomenon, described via partial differential equations (PDEs, see, e.g., [35, 36]) or stochastic PDEs (SPDEs, see, e.g.,[37]). Although these approaches are yet to be developed for general types of object data, their modeling perspective is naturally suited to take full advantage of the possible prior knowledge on the laws governing the phenomenon under study.

In the context of O2S2, we recently proposed [18] a computational approach able to deal jointly with the data and the domain complexities, by following a *divide-et-impera* strategy, in a bagging framework [38]. The methodology is based on iterated random partitions of the study domain (random domain decompositions, RDD, [18, 39]), that allow performing an ensemble of locally stationary and Euclidean *weak* analyses – each conditioned to a realization of the RDD – to be then aggregated into a final *strong* result. Natural fields of application of the approach are those of environmental monitoring within large estuarine systems, where sensible data analysis should properly account for the complex topology of the spatial domain. For instance, in [18] we used Kriging via RDDs to predict the PDFs of dissolved oxygen (DO) within the Chesapeake Bay (US), a large estuarine system which is regularly monitored to assess the impact of human activities on aquatic variables deemed critical for its ecosystem. The feature space for the O2S2 analysis was set to the Bayes space of [9], to properly account for the closed nature of PDF data (see Sect. 2). The kriged PDFs are shown in Fig. 1a. Note that kriging the entire PDFs, instead of, e.g., their summary statistics (mean, variance, or selected quantiles), allows projecting the full information content embedded in PDFs to unsampled locations in the system. In the context of our study, the kriged PDFs were then used to support the identification of the so-called *dead zones*, which are areas of the estuary where the presence of oxygen in water is below 2 mg/l hindering the life of most marine species. Figure 1b shows the predicted map of probability $\mathbb{P}(DO < 2\text{mg/l})$, obtained from the kriged PDFs. The spatial

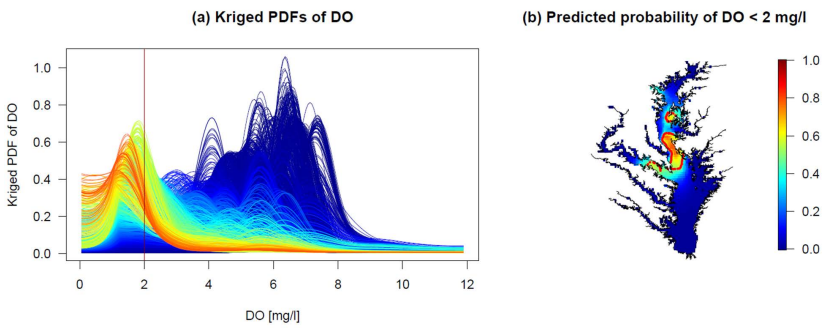

Figure 1: O2S2 prediction results for the distribution of DO in the Chesapeake Bay. (a) Kriged PDFs obtained via RDD; (b) map of probability of being a Dead Zone ($\mathbb{P}(DO < 2\text{mg/l})$), obtained from the kriged PDFs (modified from [18]). To enhance interpretation, the contour line of level $\mathbb{P}(DO < 2\text{mg/l}) = 0.5$ is marked with a thick red line in panel (b). Colors in panels (a) and (b) are given consistently.

patterns clearly follows the water dynamics within the system, and are indeed insightful for the assessment of its critical areas. The correct identification of these latter areas is key to plan effective restoration and protection programs for the Bay. More generally, developing sound mathematical frameworks for modern computational geosciences ultimately means providing valid decision making support to the stakeholder (national and local agencies, administrators, final users), with potential impacts on economy, environment and human health.

6. Computational challenges and software

Statistical methods taking on the challenge of the GeoData revolution cannot neglect the computational feasibility of developed algorithms. GeoData scientists will definitely need to take advantage of state-of-the-art numerical methods and IT technologies. For instance, recent methods based on PDEs or SPDEs [35, 37] rely on advanced techniques of numerical analysis and on statistical approximations (INLA [40]), leading to highly sparse matrices. Localization through RDD leads to embarrassingly parallel computer schemes (i.e., the structure of the algorithms is naturally suited to code parallelization), allowing for highly efficient implementations. Hardware acceleration can provide further technological support to address the challenges of computational geosciences, allowing to achieve higher degrees of efficiency by using hardware components on the system to perform pre-defined types of tasks (see, e.g., [41]). In all these cases, the availability of open, efficient and effective software packages will be crucial to knowledge dissemination. A few R packages are already available for O2S2, allowing for spatial modeling and kriging of Hilbert data (`fdagstat` [42]) and of manifold data (`Manifoldgstat` [43]). These software packages take advantage of scalable routines allowing for fast computations on relatively large

datasets, or for the use of bagging algorithms – whose backbone is the iterative repetition of model estimates and kriging predictions.

The ultimate key to moving forward the frontier of GeoData Science will definitely be a strong interplay among varied areas of applied sciences and engineering – including mathematics, statistics, engineering and computer science.

References

[1] F. Finazzi, The earthquake network project: Toward a crowdsourced smartphone-based earthquake early warning system, Bulletin of the Seismological Society of America 106 (3) (2016) 1088–1099.

[2] A. Menafoglio, P. Secchi, Statistical analysis of complex and spatially dependent data: A review of object oriented spatial statistics, European Journal of Operational Research 258 (2) (2017) 401 – 410.

[3] J. S. Marron, A. M. Alonso, Overview of object oriented data analysis, Biometrical Journal 56 (5) (2014) 732–753.

[4] A. Menafoglio, G. Petris, Kriging for Hilbert-space valued random fields: The operatorial point of view, Journal of Multivariate Analysis 146 (2016) 84–94.

[5] J. Aitchison, The statistical analysis of compositional data, Journal of the Royal Statistical Society. Series B (Methodological) 44 (2) (1982) 139–177.

[6] V. Pawlowsky-Glahn, J. Egozcue, R. Tolosana-Delgado, Modelling and analysis of compositional data, John Wiley & Sons, Ltd, 2015.

[7] A. Buccianti, E. Grunsky, Compositional data analysis in geochemistry: Are we sure to see what really occurs during natural processes?, Journal of Geochemical Exploration 141 (2014) 1 – 5.

[8] A. Menafoglio, A. Guadagnini, P. Secchi, A Kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers, Stochastic Environmental Research and Risk Assessment 28 (7) (2014) 1835–1851.

[9] K. G. van den Boogaart, J. Egozcue, V. Pawlowsky-Glahn, Bayes Hilbert spaces, Australian & New Zealand Journal of Statistics 56 (2014) 171–194.

[10] D. Pigoli, A. Menafoglio, P. Secchi, Kriging prediction for manifold-valued random field, Journal of Multivariate Analysis 145 (2016) 117–131.

[11] A. Menafoglio, A. Guadagnini, P. Secchi, Stochastic Simulation of Soil Particle-Size Curves in Heterogeneous Aquifer Systems through a Bayes space approach, Water Resources Research 52 (2016) 5708–5726.

[12] A. Menafoglio, O. Grujic, J. Caers, Universal kriging of functional data: trace-variography vs cross-variography? Application to forecasting in unconventional shales, Spatial Statistics 15 (2016) 39–55.

[13] F. Lentoni, A class of geostatistical methods to predict and simulate seismic ground motion fields: from a univariate to a functional approach, Master's thesis, Politecnico di Milano (2018).

[14] L. Billard, E. Diday, Symbolic Data Analysis: Conceptual Statistics and Data Mining, Wiley, 2007.

[15] J. Ramsay, B. Silverman, Functional data analysis, 2nd Edition, Springer, New York, 2005.

[16] P. Delicado, R. Giraldo, C. Comas, J. Mateu, Statistics for spatial functional data, Environmetrics 21 (3-4) (2010) 224–239.

[17] D. Nerini, P. Monestiez, C. Manté, Cokriging for spatial functional data, Journal of Multivariate Analysis 101 (2) (2010) 409–418.

[18] A. Menafoglio, G. Gaetani, P. Secchi, Random domain decompositions for object-oriented kriging over complex domains, Stochastic Environmental Research and Risk Assessment.

[19] J. P. Kleijnen, Kriging metamodeling in simulation: A review, European Journal of Operational Research 192 (3) (2009) 707–716.

[20] F. Bottazzi, E. D. Rossa, A functional data analysis approach to surrogate modeling in reservoir and geomechanics uncertainty quantification, Mathematical Geosciences 49 (4) (2017) 517–540.

[21] O. Grujic, A. Menafoglio, G. Yang, J. Caers, Cokriging for multivariate Hilbert space valued random fields: application to multi-fidelity computer code emulation, Stochastic Environmental Research and Risk Assessment 32 (7) (2018) 1955–1971.

[22] S. Pagani, A. Manzoni, A. Quarteroni, Efficient State/Parameter Estimation in Nonlinear Unsteady PDEs by a Reduced Basis Ensemble Kalman Filter, SIAM/ASA J. Uncertainty Quantification 5 (1) (2017) 890–921.

[23] D. Brambilla, M. Papini, L. Longoni, Temporal and Spatial Variability of Sediment Transport in a Mountain River: A Preliminary Investigation of the Caldone River, Italy, Geosciences 8 (5) (2018) 163.

[24] P. Secchi, S. Vantini, V. Vitelli, Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan (with discussion), Statistical Methods & Applications 24 (2) (2015) 279–300.

[25] X. Zhao, E. del Castillo, An intrinsic geometrical approach for statistical process control of surface and manifold data, Tech. rep. (2019).

[26] B. Colosimo, M. Grasso, Spatially weighted pca for monitoring video image data with application to additive manufacturing, Journal of Quality Technology 50 (4) (2018) 391–417.

[27] E. Del Castillo, B. Colosimo, S. Tajbakhsh, Geodesic Gaussian processes for the parametric reconstruction of a free-form surface, Technometrics 57 (1) (2015) 87–99.

[28] F. Curriero, On the use of non-Euclidean distance measures in geostatistics, Mathematical Geology 38 (2006) 907–926.

[29] P. Asadi, A. Davison, S. Engelke, Extremes on river networks, The Annals of Applied Statistics 9 (4) (2015) 2023–2050.

[30] J. Ver Hoef, E. Peterson, A moving average approach for spatial statistical models of stream networks, Journal of the American Statistical Association 105 (489) (2010) 6–18.

[31] C. Huang, H. Zhang, S. M. Robeson, On the validity of commonly used covariance and variogram functions on the sphere, Mathematical Geosciences 43 (6) (2011) 721–733.

[32] E. Porcu, D. Daley, M. Buhmann, M. Bevilacqua, Radial basis functions with compact support for multivariate geostatistics, Stochastic Environmental Research and Risk Assessment 27 (4) (2013) 909–922.

[33] M. Jun, M. L. Stein, Nonstationary covariance models for global data, Ann. Appl. Stat. 2 (4) (2008) 1271–1289.

[34] S. Castruccio, M. Genton, Principles for statistical inference on big spatio-temporal data from climate models, Statistics & Probability Letters 136 (2018) 92 – 96.

[35] L. M. Sangalli, J. O. Ramsay, T. O. Ramsay, Spatial spline regression models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75 (4) (2013) 681–703.

[36] M. S. Bernardi, M. Carey, J. O. Ramsay, L. M. Sangalli, Modeling spatial anisotropy via regression with partial differential regularization, Journal of Multivariate Analysis 167 (2018) 15 – 30.

[37] F. Lindgren, H. Rue, J. Lindström, An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (4) (2011) 423–498.

[38] L. Breiman, Bagging predictors, Mach Learn 24 (1996) 123–140.

[39] P. Secchi, S. Vantini, V. Vitelli, Bagging Voronoi classifiers for clustering spatial functional data, International Journal of Applied Earth Observation and Geoinformation 22 (2013) 53 – 64.

[40] H. Rue, S. Martino, N. Chopin, Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (2) (2009) 319–392.

[41] Y. Zhang, X. Zheng, Z. Wang, G. Ai, Q. Huang, Implementation of a parallel GPU-based space-time kriging framework, ISPRS International Journal of Geo-Information 7 (5) (2018) 193.

[42] O. Grujic, A. Menafoglio, fdagstat, an R package (2017).
URL `https://github.com/ogru/fdagstat`

[43] I. Sartori, L. Torriani, Manifoldgstat, an R package (2019).
URL `https://github.com/LucaTorriani/KrigingManifoldData`