Credit: © F. Cormon/Auvergne-Rhône-Alpes Tourisme

MASCOT-NUM 2022
Clermont Ferrand, France
7-9 June 2022

Mixture kriging on granular data

M. Grossouvre
Mines Saint-Etienne

Supervisor(s): D. Rullière (Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS), J. Villot (Mines Saint-Etienne, Univ Lyon, CNRS, UMR 5600 Environnement, Ville, Société)

PhD expected duration: Aug. 2021 - Jul. 2024

Address: U.R.B.S.
Bâtiment des Hautes Technologies
20, rue du Professeur Benoît Lauras
42000 Saint-Étienne

E-mail: marcgrossouvre@urbs.fr

Abstract:

We consider an input space over which is defined a field of multidimensional random output variables. The specificity we introduce is that outputs may be defined and observed not only for points of the input space but also for some regions of this same input space: for instance, imagine that some sociological variables (salaries, expenses, etc.) are available for different geographical areas: cities, regions, countries, etc. In this context, we refer to **granular data** and **grains** for these areas of the input space. For such data, one is interested in defining a suitable data model that is able to predict output variables for given inputs, be it points or grains. The underlying assumption in this work is that there is some form of dependence between outputs based on the relative positions of the associated inputs.

A possible application of this model is in the field of geographic information to handle data that is released in open format by institutions. Say for instance that a government releases the distribution of inhabitants salaries at municipality level. A private company may try to use this data to estimate the distribution of salaries at a smaller scale, say for a district in a city. This company may be willing to include in its model both this institutional data and some known salaries at specific locations of the territory.

To handle this problem, we expose here a general Kriging approach that generalizes the usual Simple or Ordinary (Co)Kriging techniques.

Different ways to predict using areal data have been proposed and a review of methods has been published by Gotway and Young [2]. This review shows a constant approach: the observed output over an areal unit is an average of random variables, so that aggregation means averaging values. In meteorological studies, change of support problems have been studied for time series [1] with a fully Bayesian approach. Mathematically, a milestone has been set by Kyriakidis [3] with a complete Kriging model including area-to-point and sketching area-to-area prediction. This work has been cited a large number of times. The main identified problem of these approaches is the variance shrinkage of the average compared to the initial random variable. Even if averaging has turned out to be useful for suitable data such as satellite imaging ([4]), it is usually considered as a source of concerns [2]. Another issue is that covariance between blocks has been commonly computed by reducing a block to its centroid, this distorts the covariance especially when blocks are close to each other. Those approaches have also in common the fact that they consider ”block” (areas) and points as intrinsically different objects. Moreover blocks are considered as connected surface areas in $\mathbb{R}^2$ that have to be ”discretized”.

We propose an approach that does not make distinction between points and grains in observations and prediction. It also that makes no difference between grains containing continuous or discrete sets of points. **The originality of this paper is that it considers mixture random variables rather than averaged random variables over areas.** We define a simple and general framework where **points** lie in a **territory** $\chi \subset \mathbb{R}^d$ and **grains** are any non-empty subsets of χ. A vector of output variables $\mathbf{Y}(x) = (Y_1(x), ..., Y_d(x))^\top$ is defined for any point x of χ. Output variables $\mathbf{Y}(g)$ on a grain g

are defined to be the output variables at a random location X_g of g (mixture distribution). Moreover, we assume that for any point x in χ and any i in $[\![1,d]\!]$, $\mu_i(x) := \mathbb{E}\left[Y_i(x)\right]$ is known and for any 2 points of χ, any i,j in $[\![1,d]\!]$, we also know $k_{i,j}(x,x') := \operatorname{Cov}\left[Y_i(x), Y_j(x')\right]$.

With this framework, we have:

$$k_{i,j}(g,g') := \operatorname{Cov}\left[Y_i(g), Y_j(g')\right] = \mathbb{E}\left[k_{i,j}(X_g, X_{g'})\right] + \operatorname{Cov}\left[\mu_i(X_g), \mu_j(X_{g'})\right] \tag{1}$$

and we can prove that the variance of a grain is always greater than the variance we would obtain replacing the mixture at grain level by an average. In this framework, there is no more distorsion when grains are close to each other or even overlapping since no approximation is made.

We prove that given a set of observations on grains, the best linear unbiased predictor of the output on a new grain is the Kriging predictor. If grains are singletons then we retrieve the usual Simple Kriging and Ordinary Kriging predictors so that Mixture Kriging can be seen as a generalization of usual Kriging interpolation. Mixtures of gaussian random variables are not gaussian so that gaussian interpretation of Kriging is lost. And Mixture Kriging in not interpolating. If grains are not singletons then repeated observations are supported.

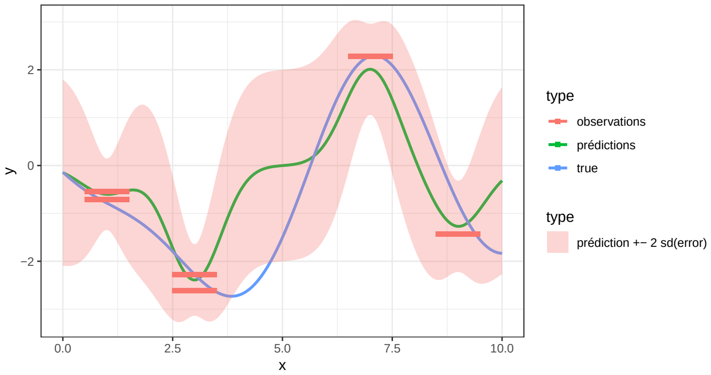

Figure 1: We consider a gaussian random field. Assume that $Y(x)$ is observed with a noise on x "round to the nearest integer". We modelize this noise saying that we observe Y on grains of length 1. For instance the first 2 red lines on the left show 2 observation of the grain $[0.5, 1.5[$. We have 6 observations from which we learn lengthscale and amplitude. After optimizing RMSE, noise is controlled and variations of predictions are ample.

References

[1] Alan E. Gelfand, Li Zhu, and Bradley P. Carlin. On the change of support problem for spatio-temporal data. *Biostatistics*, 2(1):31–45, March 2001. Publisher: Oxford Academic.

[2] Carol Gotway and Linda Young. Combining Incompatible Spatial Data. *Journal of the American Statistical Association*, 97:632–648, February 2002.

[3] Phaedon Kyriakidis. A Geostatistical Framework For Area-To-Point Spatial Interpolation. *Geographical Analysis*, 36, August 2004.

[4] Qunming Wang, Wenzhong Shi, and Peter M. Atkinson. Area-to-point regression kriging for pan-sharpening. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:151–165, April 2016.

Short biography – After studying engineering, applied mathematics and history of sciences, Marc Grossouvre has been teaching Mathematics in high school for more than 10 years. Much involved in the ecological transition, he retrained and became a data-scientist. After studying at Data-Science Tech Institute (DSTI, Sophia-Antipolis), he was hired in U.R.B.S. start-up company as a PhD student to work on building's enery efficiency. This work is also partially funded by ANRT.