Predicting missing Energy Performance Certificates: Spatial interpolation of mixture distributions

Marc Grossouvre, Didier Rullière, Jonathan Villot

Predicting missing Energy Performance Certificates: Spatial interpolation of mixture distributions

Marc Grossouvre[a,b], Didier Rullière[c], Jonathan Villot[d]

[a]*Mines Saint-Etienne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, Departement GMI, Espace Fauriel, 29 rue Ponchardier, Saint-Etienne, 42023, France;*
[b]*U.R.B.S. SAS, Bâtiment des Hautes Technologie, 20 Rue Professeur Benoit LAURAS, Saint-Etienne, 42000, France*
[c]*Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, Espace Fauriel, 29 rue Ponchardier, Saint-Etienne, 42023, France*
[d]*Mines Saint-Etienne, Univ Lyon, CNRS, Univ Jean Monnet, Univ Lumiere Lyon 2, Univ Lyon 3 Jean Moulin, ENS Lyon, ENTPE, ENSA Lyon, UMR 5600 EVS, Institut Henri Fayol, Saint-Etienne, 42023, France*

Abstract

Mass renovation goals aimed at energy savings on a national scale require a significant level of public financial commitment. To identify target buildings, decision-makers need a thorough understanding of energy performance. Energy Performance Certificates (EPC) provide information about areas of space, such as land plots or a building's footprint, without specifying exact locations. They cover only a fraction of dwellings. This paper demonstrates that learning from observed EPCs to predict missing ones at the building level can be viewed as a spatial interpolation problem with uncertainty both on input and output variables. The Kriging methodology is applied to random fields observed at random locations to determine the Best Linear Unbiased Predictor (BLUP). Although the Gaussian setting is lost, conditional moments can still be derived. Covariates are admissible, even with missing observations. We present applications using both simulated and real data, with a specific case study of a city in France serving as an example.

Keywords: climate governance, energy efficiency, multi-scale processes, areal data, change of support

Email addresses: marcgrossouvre@urbs.fr (Marc Grossouvre), drulliere@emse.fr (Didier Rullière), jonathan.villot@emse.fr (Jonathan Villot)

1. Introduction

1.1. Classifying the EPC prediction problem in research

Energy Performance Certificates are delivered in many countries around the world to assess the energy efficiency of buildings. Various approaches can be found in the European Union, Turkey, the UK, the USA... An Energy Performance Certificate (EPC) is defined in France as an energy consumption associated with a qualitative labelling letter ranging from A to G as shown in Figure 1. Energy consumptions associated with dwellings, identified by their addresses, are inventoried in a database released in open access and mapped in Figure 2. A second database matches each address with a land plot. Finally, a third database gives the living area of every dwelling, be it house or apartment, together with the land plot where they are located, and a few other technical specifications. However, the exact location of these dwellings on each land plot is not certain. From these datasets, decision-makers such as municipalities, would like to infer the EPC (energy consumption and label) of buildings that have not been observed in order to identify targets for energy retrofit incentives. This problem is referred to as the EPC prediction problem in the present paper.

In the literature, this problem can be approached from an engineering perspective, from a data management perspective or from a geostatistics point of view.

From an engineering perspective, heat engineers have physical models that compute an energy balance in order to find a given building's energy consumption. These models require a large number of technical features and may be used to design a refurbishment (improvement) strategy (Baker et al., 2021). To work at a larger scale, heat engineers define typologies of buildings, compute a distribution of these types on a given territory, and therefore infer a distribution of EPC labels. This approach has proven to be efficient (Ballarini et al., 2017). However, the lack of knowledge about the detailed technical features of each building is a strong limitation for a prediction at the building level. Some feature reduction efforts have been made (Ali et al., 2020), but the remaining features are still problematic to infer and require extra efforts (Schetelat et al., 2020). The present work considers an alternative approach wherein detailed technical knowledge of each building is relinquished, and instead leverages the geolocated nature of EPC information.

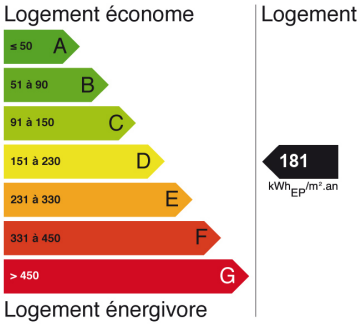

Figure 1: Prescribed vignette appearing on the French energy performance certificate up to 2021. Label A refers to energy-efficient dwellings, and label G refers to energy-intensive dwellings.

From a data management perspective, the EPC prediction problem requires a process to combine datasets from multiple sources available at multiple scales, which is known as data fusion (Smith et al., 2008). These types of problems are becoming increasingly complex due to the growing amount of data available, whether it be ecological, social, or institutional. These datasets relate to space units of varying shapes, dimensions, and cardinality. And in some cases, it may be difficult to determine the exact position of an observed object. This is the case with buildings, since many governments lack a detailed map of the building stock in their country. Property tax is typically based on intrinsic factors such as surface area and number of bedrooms, but not extrinsic factors such as the floor number or window orientation (see Table 1). As a result of this uncertainty, large-scale studies on housing stock have to rely on an abstract concept of dwelling. This idea of dwelling can refer to a house or an apartment; it is not clearly delimited but it is described by a set of features such as an area or a number of bedrooms. These features are gathered in a table with one dwelling per row, meaning that the dwelling is the smallest unit of information.

dwelling ID	address	area (m^2)	bedrooms	...	land plot ID	
024830065432	161 rue du Chateau 02089 BILLY	83	3	...	024830000C0057	

Table 1: Structure of the dwellings table from the French Ministry of Finance. The actual table comprises 118 features. Geographic position is identified by a land plot ID.

Similarly, the smallest unit of information for a table with one EPC per row is a part of a building. It is not clearly defined as an object in a 3-

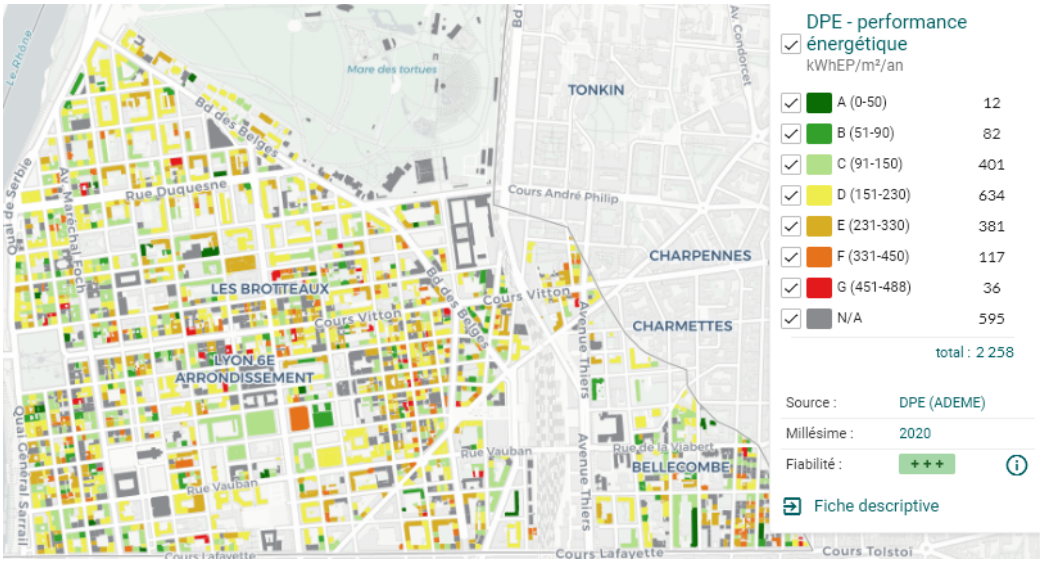

Figure 2: Map of French inventoried EPCs over a neighborhood of Lyon city. This image is a screen capture of the French National Observatory of Buildings (Observatoire National des Bâtiments - ONB), released with the consent of the rights holders U.R.B.S. SAS.

address	area (m^2)	walls	...	energy consumption	EPC
161 rue du Chateau 02089 BILLY	83	bricks	...	210	D

Table 2: Structure of the observed EPCs table. Geographic positions are indicated by addresses.

dimensional space, but it has features that describe it (see Table 2). And to predict the EPC of buildings, one also has to define buildings. In the same way, data fusion requires defining the smallest units of information, also known as granules for each dataset: "Informally, a granule of a variable X is a clump of values of X that are drawn together by indistinguishability, equivalence, similarity, proximity, or functionality. For example, an interval is a granule." Zadeh (2005). The field of study that focuses on representing, constructing, and processing these information granules is called Granular Computing (Pedrycz, 2013). Assuming that an appropriate data fusion process is implemented, dwellings, EPC observations, and complete buildings are represented in the same data model. It remains to define a relevant predictive model. Granular computing is multidisciplinary, but since we are dealing with geo-localized information, the natural field of research is geo-statistics, which has been defined as "dealing with spatial processes indexed

over continuous space" (Cressie, 1993, p7).

From a geostatistics perspective, the irreducible uncertainty about granules' positions (dwellings, buildings, etc.) in their underlying space restricts the use of traditional spatial interpolation models such as Kriging as well as more recent models such as those proposed by Roksvåg et al. (2021), although the latter efficiently combines point and areal observations. This work aims to overcome the latter limitation and develop a comprehensive framework capable of handling data with uncertainty about the position of observed objects while still allowing for the definition of an optimal linear predictor for spatial interpolation of EPC values. As is first presented below, the literature shows that the problems to solve have already been identified and that several solutions have been proposed with their benefits and shortcomings.

1.2. The limits of systematic averaging for spatial interpolation

Gaussian Process Regression (Williams and Rasmussen, 1996), also known as Kriging, is one of the major spatial interpolation approaches (Comber and Zeng, 2019). Kriging theory relies on the assumption that points close to each other are more likely to have similar features. It achieves the Best Linear Unbiased Predictor (BLUP) in the least squares sense for point spatial interpolation. However, the EPC prediction problem deals with observations that are not point observations but areal observations. Areal interpolation, as defined by Lam (1983), involves "the transformation of data from one set of boundaries to another". Lam also used the terms source zone and target zone. For the EPC prediction problem, source zones are dwellings and buildings' parts that are observed, while target zones are whole buildings, including those for which no part has been observed. Spatial or areal interpolation research is based on the assumption that granules close to each other are more likely to have similar features. This is reasonably understandable for temperatures that are continuously defined over space, but it may be more challenging to observe and model when dealing with areal data where granules can be of various sizes and shapes, sometimes uncertainly defined. Gotway and Young (2002) highlighted the terms used to describe areal interpolation and its challenges; this terminology includes block Kriging, multi-scale and multi-resolution modelling, the ecological inference problem, the modifiable areal unit problem (MAUP), the scaling problem, the change of support problem, and the reduction of variance problem. Below are the aspects of this work that are more relevant for solving the EPC prediction problem.

Block Kriging is a derivative of Kriging designed for handling areal data. It distinguishes point-to-area, area-to-point, and area-to-area predictions. It assumes that a feature at block (granule) level is the average of the block's point features. Point-to-area prediction produces an estimate "identical to that obtained by averaging the point estimates produced by [Kriging]" (Isaaks and Srivastava, 1989; Cressie, 1993). Kyriakidis (2004) described a complete Kriging model for area-to-point prediction, proved that it is an optimal predictor, and sketched area-to-area prediction. Goovaerts (2008) studied in depth the problem of estimating the variogram, that is to say, measuring the similarity between 2 points at different distances, for block Kriging. He showed that averaging reduces the sill of the variogram and tried to tackle this bias. Moreover, while point estimates obtained by Kriging are optimal, area-to-area Kriging may not be the optimal predictor for the average value over the block.

A known issue resulting from systematic averaging in areal Kriging models arises in scenarios such as analysing crop yields, where the set of agricultural fields to aggregate for a certain type of crop varies from year to year. It states that correlations between features at areal level are heavily dependent on the aggregation process, making it difficult to compare correlations between different years. This is the Modifiable Areal Unit Problem (MAUP) for which a measuring approach has been recently proposed (Briz-Redon, 2022). While the MAUP refers to the correlation between aggregated features, the ecological inference problem is a result of the correlations at the individual level being different from the correlations of the averaged features at the ecological (group) level. A lack of information about the individuals' positions leads to a bias when the averaged information about individuals distributed into areal units is cross-classified by other individual (point-level) variables (sex, race). According to Gotway and Young (2002), "The smoothing effect that results from averaging is the underlying cause of both the scale problem in the MAUP and aggregation bias in ecological studies." Apart from correlations, the variance itself is affected by systematic averaging. Indeed, the average of identical random variables has a smaller variance than the variance of the individuals themselves. The specific issue of variance reduction at the block level was partially addressed in Li et al. (2009) using rectangular blocks at multiple scales.

Despite its limitations, the averaging method has proven to be effective for interpolating areal data. For example, Poggio and Gimona (2015) downscaled climate models and predicted soil wetness using Kriging on the residuals

of a generalized additive model (Wood, 2017). Area-to-point Kriging, also called disaggregation, has also been implemented by Kerry et al. (2013); Truong and Heuvelink (2013); Yoo and Kyriakidis (2006). Additionally, area-to-area Kriging (block Kriging) has been used effectively by Zhang et al. (2018) and has been apply to downscaling by Jin et al. (2018) as well as Pereira et al. (2018). The satellite imaging field has also notably benefited from this framework, as in the pan-sharpening process, which is "a technique to combine the fine spatial resolution panchromatic (PAN) band with the coarse spatial resolution multispectral bands of the same satellite to create a fine spatial resolution multispectral image" Wang et al. (2016). In this process, points are weighted according to their distance from the centroid of the satellite pixel when computing the average value.

Both the MAUP and the ecological inference problem belong to a family of problems related to the combination of different types of granules in the same model, e.g. observing dwellings and predicting buildings. These problems are gathered in the change of support problems family. Another particular change of support problem known as spatial misalignment arises when a given feature is observed at multiple scales, including point level. Systematic averaging makes points and areas different objects with different different correlation structures and therefore different predictors. The classification of problems such as "area-to-point" or "area-to-area" reflects this categorization. To address spatial misalignment, a Bayesian framework that can be iterated both with point observations and block observations has been proposed by Moraga et al. (2017). However, this model is still based on averaging at areal level for features that are continuously defined over the territory. Like other models derived from Kriging, it considers blocks to be connected surface areas in $\mathbb{R}^2$ that need to be discretized (Goovaerts, 2008), which can distort reality for features that are not continuously defined over the space. Such is the case of populations of individuals that are discrete points heterogeneously located within a block, such as a county or census tract.

1.3. Beyond systematic averaging

A way to try and overcome change of support problems is to define a new data model for which features at areal level do not require systematic averaging. In this regard, Godoy et al. (2022) defined a Gaussian random field on the class $\mathscr{B}_D$ of closed subsets of a certain domain $D \in \mathbb{R}^n$. Distances

between elements of $\mathscr{B}_D$ are measured with the Hausdorff distance, and the correlation structure between features is based on this distance together with a Matérn kernel. Eventually, a Bayesian framework is used to fit the model with respiratory cancer data, yielding encouraging results. This model seems very general and will probably find other fields of application. However, it is not interpretable in the sense that there is no obvious link between the feature at the areal level and the feature at the point level, therefore eluding the question of consistency. In other words, it is not known whether the aggregation of cancer incidence predictions at a small scale would give the prediction of cancer incidence at a larger scale. Beside this limitation, the Hausdorff-Gaussian process does not solve the problem of position uncertainty that is found in the EPC prediction problem.

In this paper, a new model is proposed where learning and prediction can be made from both aggregated and point support data. An object category called grain is introduced to express this new approach, consistent with research realities where it may be desirable to complete large aggregated open datasets with local observations and predict at various scales. Grains containing a continuous or discrete set of points are treated identically. As is detailed above, a weighted average is the standard aggregation approach. In this respect, the MAUP is related to determining a covariance model for points from which the covariance between blocks and the covariance between points and blocks are derived. Weights for averaging are assumed to be fully determined for a given block; they are not regarded as a probability distribution for a block, thereby ignoring some related statistics and other potential sources of stochastic dependence between blocks. The present paper proposes a method of incorporating a mixture distribution to address this issue. Kriging has already been developed for features that are mixtures at the point level (Lin et al., 2010), but Lin et al. make no assumption about the distribution of features at the areal level. Instead, we assume the aggregation of information at the areal level to be a mixture. Averaging a large number of random variables results in a variance reduction, whereas mixing a large number of random variables does not tend to reduce the variance. We will show that this approach effectively manages position uncertainty. However, one drawback is that mixtures of Gaussian random variables are generally not Gaussian, which means that the usual Gaussian process interpretations and conditioning will no longer hold.

The present study proposes a new model for processing granular data,

as detailed in Section 2. In Subsection 2.1, a suitable data model is established, while in Subsection 2.2, we define the feature variables' means and covariances. Moreover, a Best Linear Unbiased Predictor is derived in Subsection 2.3. We illustrate the model with examples in Section 3, starting with simulated rounded positions in Subsection 3.1, followed by simulated areal data with varying area sizes in Subsection 3.2. Subsection 3.3 focuses on presenting the EPC prediction problem with real data. Finally, in Section 4, we discuss the pros and cons of the new model.

2. Prediction model

This work is motivated by the will to handle data that is released in open format by public or private institutions. The goal is to use institutional data, such as the distribution of salaries at the municipality level, to estimate the distribution of salaries at a smaller scale, such as a district in a city, while also including known salaries at specific locations. To achieve this, we propose here a general Kriging approach that extends the traditional Simple or Ordinary Kriging and coKriging techniques. The model will explain some variables (such as the energy consumption, the salary, etc.) using some explanatory variables (such as the location, the construction year, etc.). The former will be referred to as output variables and the latter as input variables. Let us consider a space (input space, sometimes known as study space) over which is defined a field of multidimensional random variables (output variables, features of interest) such as sociological variables, assumed to be defined and potentially observed for both points in the input space and for geographic areas, such as cities, regions, or countries. These areas are referred to as "grains". The model predicts output variables at unobserved points or grains, based on the assumption that the dependence between outputs depends on the relative positions of the inputs. No assumption is made regarding the shape of the grains, which can even overlap partially or completely.

2.1. Data model

Let us define the structure of the input space.

Definition 1 (Inputs)**.** *Let d be a positive integer. A territory and grains inside this territory are defined as follows:*

- *A* ***territory*** *is a subset χ of $\mathbb{R}^d$;*

- *A **point** is any element $x \in \chi$;*
- *A **grain** is any non-empty subset $g \subseteq \chi$.*

It is common in some application fields to use a different terminology to talk about grains: blocks, pixels, or areas for instance. In the above definition, there is no constraint on grains, contrary to pixels that are usually forming a regular grid known as a raster. A set of grains does not have to cover the whole territory, and its elements might overlap. Moreover, a grain is not necessarily a connected set, contrary to blocks. And an area is usually seen as associated with a surface area (a set of strictly positive measure) whereas a grain may be a finite set of points.

For instance, suppose that the points are represented as pairs of latitude and longitude coordinates in an appropriate coordinate reference system. In this case, χ could be defined as the set of all latitude-longitude pairs that fall within a specific country, yielding $d = 2$ and $\chi \subset \mathbb{R}^2$. A grain may correspond, for example, to a specific city, to a specific land plot, or to a specific building's footprint. Previous Kriging models refer to blocks or areas for sets of points that are disjoint (see, for instance, Kyriakidis, 2004).

When dealing with geographic data, a set of grains is usually the minimum scale at which information is available; that is to say, the data granularity. For instance, it may be the set of land plots, the set of cities, the set of buildings' footprints, etc. However, considered grains may have non-empty intersections and may come from different datasets, at different scales, such as land plots and census tracts. Definition 1 is general enough to include such sets of grains. Data that describe population or buildings are not continuously defined over a territory, as opposed to temperature or pollutant concentration. Census data are anonymized at the census tract level before being released. For instance, in a census table describing dwellings, a row describes a dwelling that exists on a certain census tract, but we don't know exactly where it is on this tract. Then dwellings' surface area is neither continuous nor clearly geo-localized. Definition 2 below unifies output features that are continuously defined over a territory and output features that are not.

An originality of this work is to consider a set of random locations that model the uncertainty of explanatory variables over each considered grain. Let $\{X_g, g \in \mathcal{G}\}$ be a given sequence of random variables that are random locations, or more generally, random explanatory variables. Their joint distri-

bution is assumed to be known. As an example, for non-overlapping grains, one can choose a sequence of independent uniform random variables over each grain, but any other joint distribution, possibly dependent, can be chosen. Definition 2 associates output variables with these random explanatory variables.

Definition 2 (Outputs)**.** *Let $\mathcal{G}$ be a set of grains. We assume that for each grain $g \in \mathcal{G}$, there is a random variable X_g with values the points of g. Output variables are defined over points and grains of $\mathcal{G}$ as follows:*

- $\mathbf{Y}$ *is a p-dimensional multivariate random field over χ such that:*

$$\forall x \in \chi,\ \mathbf{Y}(x) := (Y_1(x), \ldots, Y_p(x))^\top \in \mathbb{R}^p$$

- *For each $g \in \mathcal{G}$, a p-dimensional real random vector $\mathbf{Y}(g)$ is defined to be the value of $\mathbf{Y}$ at a random location $X_g \in g$:*

$$\forall g \in \mathcal{G},\ \mathbf{Y}(g) := \mathbf{Y}(X_g) \in \mathbb{R}^p$$

Defined accordingly, $\mathbf{Y}(g)$ is a mixture distribution.

For a given set of grains $\mathcal{G}$, the set of random variables $\{X_g : g \in \mathcal{G}\}$, is assumed to be defined and known, and the dependence structure between those random variables is supposed to be known. Furthermore, these random variables are assumed to be independent from the random field $\mathbf{Y}$.

Example 1. *The importance of X_g should be stressed here. For instance, if one studies the distribution of capital owned by citizens of a given municipality, $P(X_g = x)$ gives the probability of a citizen x to be observed. $P(Y(X_g) = y)$ is the probability to observe y when a citizen picked randomly according to X_g unveils his capital:*

$$P(Y(g) = A) = P(Y(X_g) = A) = \sum_{x \in g} P(X_g = x) P(Y(x) = A)\ .$$

It is clear that individuals are not distributed regularly (along a grid for instance) in the grain. However, in this example, it makes sense to consider that $\forall x \in g,\ P(X_g = x) = 1/[g]$ where $[g]$ is the cardinality of g. This means that the contribution of all citizens are equally valued in $Y(g)$.

Let us now suppose that the outputs are partially known on a set of grains. For $(i_1, \ldots, i_n) \in \{1, \ldots, p\}^n$ and $g_1, \ldots, g_n \in \mathcal{G}$ the following n random variables are known:

$$\underline{\mathbf{Y}} = (Y^1, \ldots, Y^n)^\top \text{ with } Y^j = Y_{i_j}(g_j) \text{ for } j \in \{1, \ldots, n\}$$

As an example, if ℓ observations of the whole random vector $\mathbf{Y}(g_h)$ are conducted for $h \in \{1, \ldots, \ell\}$, then $n = \ell \cdot p$ and the vector of observations is:

$$\underline{\mathbf{Y}} = (Y_1(X_{g_1}), \ldots, Y_p(X_{g_1}), \ldots, Y_1(X_{g_\ell}), \ldots, Y_p(X_{g_\ell}))^\top . \tag{1}$$

If some observations are incomplete, that is to say some components of $\mathbf{Y}_{g_j}$ are missing for some j, then $\underline{\mathbf{Y}}$ will be a subvector of $\underline{\mathbf{Y}}$ given in Equation (1). It means that there may be missing data in the outputs' observations.

2.2. Mean and covariances of output variables

The originality of the present work is that for a grain g, $\mathbf{Y}(g)$ is defined to be $\mathbf{Y}(X_g)$, the value of $\mathbf{Y}$ at a random location $X_g \in g$. If the random field $\{\mathbf{Y}(x) : x \in \chi\}$ and the joint distribution of $\{X_g \in \chi : g \in \mathcal{G}\}$ are known, then the joint distribution of $\{\mathbf{Y}(g) : g \in \mathcal{G}\}$ can be deduced. And, if one only knows the moments of order one and cross moments of order two of $\{Y(x) : x \in \chi\}$ together with the joint distribution of $\{X_g \in \chi : g \in \mathcal{G}\}$, then one can expect to be able to deduce expectation and cross covariances of $\{\mathbf{Y}(g) : g \in \mathcal{G}\}$.

In the rest of the paper, we assume that the first two moments of $\{\mathbf{Y}(x) : x \in \chi\}$, $\{X_g \in \chi : g \in \mathcal{G}\}$ and $\{\mathbf{Y}(g) : g \in \mathcal{G}\}$ exist. In the following proposition, we show that we can indeed deduce the moments of grains' outputs.

Proposition 1 (Mean and covariance of $\mathbf{Y}(g)$)**.** *From Definition 2, we derive the following results:*

(i) For any grain $g \in \mathcal{G}$ and any index $i \in \{1, \dots, p\}$, assuming that for all $x \in g$ we know $\mu_i(x) := \mathbb{E}[Y_i(x)] = \mathbb{E}[Y_i(g)|X_g = x]$, we have:

$$\mu_i(g) := \mathbb{E}[Y_i(g)] = \mathbb{E}[\mu_i(X_g)]$$

(ii) For any two grains g, g' in $\mathcal{G}$ and any two indices $i, j \in \{1, \dots, p\}$, assuming that for all $x \in g$, $x' \in g'$ we know $k_{i,j}(x, x') := \mathrm{Cov}[Y_i(x), Y_j(x')]$, we have:

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

In particular,

$$k_{i,i}(g, g) = \mathrm{Cov}[Y_i(g), Y_i(g)] = \mathbb{V}[Y_i(g)] = \mathbb{E}[k_{i,i}(X_g, X_g)] + \mathbb{V}[\mu_i(X_g)].$$

Proof. (i) is a direct application of the conditional expectation formula $\mathbb{E}[V] = \mathbb{E}[\mathbb{E}[V|U]]$ where $Y_i(x)$ is the result of conditioning $Y_i(g)$ with X_g.

(ii) is derived from the conditional covariance formula:

$$\mathrm{Cov}[U, V] = \mathbb{E}[\mathrm{Cov}[U, V|W]] + \mathrm{Cov}[\mathbb{E}[U|W], \mathbb{E}[V|W]]$$

after conditioning by the joint random vector $(X_g, X_{g'})$ (random variable X_g). □

Note that $\mathrm{Cov}[\mu_i(X_g), \mu_j(X_{g'})] = 0$ when $\mu_i(x)$ is constant over g or g' or when X_g and $X_{g'}$ are independent. Also note that this framework yields the expected result that if a grain is restricted to a point, then the output variables associated with this grain are the same as those associated with the underlying point.

Example 2. *For two distinct and finite grains g and g' of cardinalities $[g], [g']$, assuming in this example that X_g and $X_{g'}$ are independent uniform*

346 *random variables, we get:*

$$\mu_i(g) = \frac{1}{[g]} \sum_{x \in g} \mu_i(x)$$

$$k_{i,j}(g, g') = \frac{1}{[g]\,[g']} \sum_{(x,x') \in g \times g'} \mathrm{Cov}\left[Y_i(x), Y_j(x')\right]$$

$$k_{i,j}(g, g) = \frac{1}{[g]} \sum_{x \in g} \mathrm{Cov}\left[Y_i(x), Y_j(x)\right]$$

Remark 1 (Comparison with average – block-to-block covariances)**.** *Previous models using the concept of blocks define* $\bar{Y}_i(g) := \mathbb{E}\left[Y_i(X_g)|\{Y_i(x), x \in g\}\right] = \int_g Y_i(x) dF_g(x)$*, with* F_g *the cumulative distribution function (cdf) of the, possibly discrete, random variable* X_g*,* $i \in \{1, \ldots, p\}$*. One can check that with this setting, the mean of the mixture* $Y_i(g)$ *and the average* $\bar{Y}_i(g)$ *are identical:*

$$\mathbb{E}\left[Y_i(g)\right] = \bar{Y}_i(g)\,.$$

Regarding the covariances, when X_g *and* $X_{g'}$ *are two independent random variables, one can check that*

$$\mathbb{E}\left[k_{i,j}(X_g, X_{g'})\right] = \mathrm{Cov}\left[\bar{Y}_i(g), \bar{Y}_j(g')\right]$$

However,

$$\mathbb{E}\left[k_{i,j}(X_g, X_g)\right] \neq \mathrm{Cov}\left[\bar{Y}_i(g), \bar{Y}_j(g)\right]$$

347 *because the independence assumption does not hold any more. As a*
348 *consequence,* $\mathbb{V}\left[Y_i(g)\right] \neq \mathbb{V}\left[\bar{Y}_i(g)\right]$*, even in the specific case where*
349 $\forall i, j, g, g'$, $\mathrm{Cov}\left[\mu_i(X_g), \mu_j(X_{g'})\right] = 0$*. The difference between a mixture and*
350 *an average is retrieved here:* $\mathbb{V}\left[Y_i(g)\right] \geq \mathbb{V}\left[\bar{Y}_i(g)\right]$*.*

351 *2.3. Best unbiased linear predictor*

352 A Gaussian Process is a collection of random output variables indexed
353 over points in the input space of explanatory variables, typically denoted
354 as $Y(.)$. An observation is therefore a random variable $Y(x)$ evaluated at
355 a given point x, and the covariance between $Y(x)$ and $Y(x')$ is a function
356 of (x, x'). But we rather consider here an uncertainty on the explanatory
357 variable, meaning that an observation is modelled as a random field $Y(.)$

evaluated at a random location X_g over a given gain g. Thus, one observes a mixture of Gaussian random variables that are not Gaussian any more. Moreover the covariance between $Y(X_g)$ and $Y(X_{g'})$ depends on the joint random variables $(X_g, X_{g'})$. In the previous subsection, some assumptions have been made that are sufficient to be able to compute the covariance between two observations. In the present subsection, it is proved that, given the above defined framework and a learning set of observations, a best linear predictor can be inferred to predict the output features associated with a grain $g \subset \chi$ that has not been observed, given a learning set of observations. Note that the problem amounts to predicting any component of the output variable and that the specific covariance structure resulting from the uncertainty on the explanatory variable requires the development of a new software package, as usual packages such as `DiceKriging` can not fit such a model.

Let $\underline{\mathbf{Y}}$ be the vector of observations forming the learning set, and let $g \subset \chi$ be a grain such that for some $i \in \{1, \ldots, p\}$, $Y_i(g)$ is to be predicted.

Denote:

$$\begin{aligned} \underline{\boldsymbol{\mu}} &:= \mathbb{E}\left[\underline{\mathbf{Y}}\right] && \in \mathbb{R}^n \\ \mathbf{K} &:= \left(\mathrm{Cov}\left[Y^j, Y^{j'}\right]\right)_{j,j' \in \{1,\ldots,n\}} && \in \mathcal{S}_n^+(\mathbb{R}) \\ \mathbf{h}_i(g) &:= \left(\mathrm{Cov}\left[Y^j, Y_i(g)\right]\right)_{j \in \{1,\ldots,n\}} && \in \mathbb{R}^n \end{aligned}$$

where $\mathcal{S}_n^+(\mathbb{R})$ is the set of semi-definite positive, $n \times n$, real matrices.

In the following, $\mathbf{K}$ is assumed to be invertible.

With a given set of weights $\boldsymbol{\alpha}(g) = (\alpha^1(g), \ldots, \alpha^n(g)) \in \mathbb{R}^n$, is associated a linear predictor $M_{\boldsymbol{\alpha}(g)}$:

$$M_{\boldsymbol{\alpha}(g)} = \sum_{j=1}^{n} \alpha^j(g) Y^j = \boldsymbol{\alpha}(g)^\top \underline{\mathbf{Y}}\,.$$

The optimal weights $\boldsymbol{\alpha}_i(g)$, provided that they exist and are unique, are defined to be those minimizing a quadratic error over all unbiased linear predictors:

$$\boldsymbol{\alpha}_i(g) \in \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \mathbb{E}\left[\left(Y_i(g) - \boldsymbol{\alpha}^\top \underline{\mathbf{Y}}\right)^2\right]$$

Given the optimal predictor $M_i(g)$, the prediction error and the Kriging

(co)variance are denoted as:

$$\begin{aligned} \epsilon_i(g) &:= Y_i(g) - M_i(g) \\ c_{i,j}(g,g') &:= \mathbb{E}\left[\epsilon_i(g)\,\epsilon_j(g')\right] \quad (2) \\ v_i(g) &:= c_{i,i}(g,g) \quad (3) \end{aligned}$$

The following proposition gives an optimal predictor that can be computed under the minimal assumptions of Proposition 1. Given the first two moments of random variables $\{X_g : g \in \mathcal{G}\}$, all components of $\underline{\boldsymbol{\mu}}$, $\mathbf{K}$, and $\mathbf{h}_i(x)$ can be computed.

Proposition 2 (Mixture Kriging prediction)**.** *Given a set of observations* $\underline{\mathbf{Y}}$*, for any* $g, g' \subset \chi$*, and in particular for a single point* $g = \{x\}$*, for any* $i \in \{1, \ldots, p\}$*, the weights* $\boldsymbol{\alpha}_i(g)$ *yielding the best linear unbiased predictor (BLUP) of* $Y_i(g)$ *and the associated cross errors are as follows:*

(i) ***Simple Mixture Kriging.*** *If* $\underline{\boldsymbol{\mu}} = (0, \ldots, 0)^\top$ *and* $\mu_i(g) = 0$ *then*

$$\begin{aligned} \boldsymbol{\alpha}_i(g) &= \mathbf{K}^{-1}\mathbf{h}_i(g) \quad (4) \\ c_{i,j}(g,g') &= k_{i,j}(g,g') - \mathbf{h}_i(g)^\top \mathbf{K}^{-1}\mathbf{h}_j(g') \end{aligned}$$

(ii) ***Ordinary Mixture Kriging.*** *If* $\underline{\boldsymbol{\mu}} \neq (0, \ldots, 0)^\top$ *then the condition for unbiasedness writes* $\mu_i(g) = \boldsymbol{\alpha}_i(g)^\top \underline{\boldsymbol{\mu}}$ *and*

$$\begin{aligned} \boldsymbol{\alpha}_i(g) &= \mathbf{K}^{-1}\left(\mathbf{h}_i(g) + \lambda_i(g)\underline{\boldsymbol{\mu}}\right) \quad (5) \\ \text{where } \lambda_i(g) &= \frac{\mu_i(g) - \underline{\boldsymbol{\mu}}^\top \mathbf{K}^{-1}\mathbf{h}_i(g)}{\underline{\boldsymbol{\mu}}^\top \mathbf{K}^{-1}\underline{\boldsymbol{\mu}}} \\ c_{i,j}(g,g') &= k_{i,j}(g,g') - \mathbf{h}_i(g)^\top \mathbf{K}^{-1}\mathbf{h}_j(g') + \lambda_i(g)\lambda_j(g)\underline{\boldsymbol{\mu}}^\top \mathbf{K}^{-1}\underline{\boldsymbol{\mu}} \end{aligned}$$

Proof of Proposition 2 is given in Supplementary material Appendix A.

Proposition 2 is presented as an algorithm in pseudo-language for Simple Mixture Kriging in Algorithm 1.

Assume that $\{\mathbf{Y}(x) : x \in \chi\}$ is a vector-valued Gaussian random field and that each X_g is Dirac distributed for all grains. This last condition holds in particular when each grain is restricted to one singleton point. In this Gaussian case, one retrieves Simple Kriging and Ordinary Kriging predictors, as defined, for example, in Rasmussen and Williams (2006). In this sense,

the Mixture Kriging results presented here can be seen as a generalization of the Kriging interpolation.

The above Proposition 2 is valid to predict a single component $Y_i(g)$ of the output variable $\mathbf{Y}(g)$, but it can be extended to the prediction of $\mathbf{Y}(g)$: the best linear unbiased predictor of $\mathbf{Y}(g) = (Y_1(g) \dots Y_p(g))^\top$ for the quadratic error $\mathbb{E}\left[||\mathbf{Y}(g) - \mathbf{A}\underline{\mathbf{Y}}||_2^2\right]$ is $M_{\mathbf{A}(g)} = \mathbf{A}(g)\underline{\mathbf{Y}}$ where $\mathbf{A}(g)$ is the matrix of which the i-th row is equal to $\boldsymbol{\alpha}_i(g)^\top$ given by Proposition 2.

2.4. Particular cases

In this subsection, two important particular cases are explored. The first one considers the Ordinary Mixture Kriging situation, where the output variable's expectation is the same everywhere. An estimator of this constant expectation is derived. The second particular case considers Mixture Kriging with noisy observations and shows that a nugget effect can be introduced the same way as for Kriging.

Particular case 1 (Constant mean $\underline{\boldsymbol{\mu}} = \mu_0(1, \dots, 1)^\top$)**.** *Regarding Ordinary Mixture Kriging, assuming that all random variables $Y_i(g)$ have the same unknown expectation μ_0, and setting $\mathbf{1}_n = (1, \dots, 1)^\top$, Equation* (5) *simplifies into:*

$$\boldsymbol{\alpha}_i(g) = \mathbf{K}^{-1}\left(\mathbf{h}_i(g) + \frac{1 - {\mathbf{1}_n}^\top \mathbf{K}^{-1}\mathbf{h}_i(g)}{{\mathbf{1}_n}^\top \mathbf{K}^{-1}\mathbf{1}_n}\mathbf{1}_n\right),$$

$$\text{and setting } \hat{m}(g) := \frac{{\mathbf{1}_n}^\top \mathbf{K}^{-1}\underline{\mathbf{Y}}}{{\mathbf{1}_n}^\top \mathbf{K}^{-1}\mathbf{1}_n},$$

$$M_i(g) \text{ becomes: } M_i(g) = \hat{m}(g) + \mathbf{h}_i(g)^\top \mathbf{K}^{-1}(\underline{\mathbf{Y}} - \mathbf{1}_n\hat{m}(g)),$$

therefore $\hat{m}(g)$ is an unbiased estimator of μ_0. $\hat{m}$ can be compared with usual sample mean for independent observations $\overline{\underline{\mathbf{Y}}} = \frac{{\mathbf{1}_n}^\top \underline{\mathbf{Y}}}{{\mathbf{1}_n}^\top \mathbf{1}_n}$.

Particular case 2 (Noisy observations)**.** *Let us consider the case where, for a given $x \in \chi$, we can only observe $\tilde{Y}_i(x) = Y_i(x) + \epsilon_i(x)$ where $\epsilon_i(x)$ is independent from any $Y_j(x')$. We denote the resulting noisy output variables, observations and covariances as follows:*

$$\begin{aligned}
\tilde{Y}_i(g) &:= \tilde{Y}_i(X_g) = Y_i(g) + \epsilon_i(g) \\
\tilde{Y}^j &:= \tilde{Y}_{i_j}(X_{g_j}) = Y^j + \epsilon^j \\
\eta_{i,j}(x, x') &:= \mathrm{Cov}\left[\epsilon_i(x), \epsilon_j(x')\right]
\end{aligned}$$

Algorithm 1: Simple Mixture Kriging predictor presented as an algorithm in pseudo-language.

Data:

It is assumed that all grains are discretized and that for any grain g, X_g is uniform. A single output random field $Y(.)$ is observed.

$\mathcal{G}$: A list of n observed grains $\mathcal{G}_i, i \in \{1, \ldots, n\}$, each grain being a table with its points coordinates.

$\underline{\mathbf{Y}}$: Observed values $Y^i, i \in \{1, \ldots, n\}$, a numeric vector of same length as $\mathcal{G}$.

$k(.,.)$: Covariance kernel, a function that takes 2 points and returns a positive real number.

g_0 : An unobserved grain to be predicted i.e. a table with its points coordinates.

ϵ : A positive real number giving the nugget effect.

Result:

Optimal weights $\boldsymbol{\alpha}$ for g_0.

Conditional expectation M of $Y(g_0)$.

Kriging variance $v = \mathbb{V}\left[Y(g_0) - M\right]$ (variance of the prediction error).

begin

Fill $\mathbf{K}$*:*

for $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, n\}, i \geq j$ **do**

if $i = j$ **then**

$\mathbf{K}_{i,i} = \frac{1}{[\mathcal{G}_i]} \sum_{x \in \mathcal{G}_i} k(x, x) + \epsilon$

else $\mathbf{K}_{i,j} = \mathbf{K}_{j,i} = \frac{1}{[\mathcal{G}_i] \times [\mathcal{G}_j]} \sum_{x \in \mathcal{G}_i, x' \in \mathcal{G}_j} k(x, x')$

Fill $\mathbf{h}$*:*

for $i \in \{1, \ldots, n\}$ **do**

$\mathbf{h}_i = \frac{1}{[\mathcal{G}_i]} \sum_{x \in \mathcal{G}_i, x' \in g_0} k(x, x')$

Get results:

$\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{h}$

$M = \boldsymbol{\alpha}^\top \underline{\mathbf{Y}}$

$v = \sum_{x \in g_0} k(x, x) - \mathbf{h}^\top \mathbf{K}^{-1} \mathbf{h}$

Then the covariance between 2 grains' outputs is:

$$\tilde{k}_{i,j}(g,g') := \mathrm{Cov}\left[\tilde{Y}_i(g), \tilde{Y}_j(g')\right] = k_{i,j}(g,g') + \mathbb{E}\left[\eta_{i,j}(X_g, X_{g'})\right]$$

Therefore the observations covariance matrix writes:

$$\begin{aligned}
\tilde{\mathbf{K}} &:= \left(\mathrm{Cov}\left[\tilde{Y}^j, \tilde{Y}^{j'}\right]\right)_{j,j'\in\{1,\dots,n\}} \\
\tilde{\mathbf{K}} &= \mathbf{K} + \left(\mathrm{Cov}\left[\boldsymbol{e}^j, \boldsymbol{e}^{j'}\right]\right)_{j,j'\in\{1,\dots,n\}} \\
\tilde{\mathbf{K}} &= \mathbf{K} + \mathbf{K}_e
\end{aligned}$$

And the covariance vector between the observations and a new grain writes:

$$\begin{aligned}
\tilde{\mathbf{h}}_i(g) &:= \left(\mathrm{Cov}\left[Y^j + \boldsymbol{e}^j, Y_i(g) + \boldsymbol{e}_i(g)\right]\right)_{j\in\{1,\dots,n\}} \\
\tilde{\mathbf{h}}_i(g) &= \mathbf{h}_i(g) + \left(\mathbb{E}\left[\eta_{i_j,i}(X_{g_j}, X_g)\right]\right)_{j\in\{1,\dots,n\}} \\
\tilde{\mathbf{h}}_i(g) &= \mathbf{h}_i(g) + \mathbf{h}_{e,i}(g)
\end{aligned}$$

Typically, we can assume that $\mathbb{E}\left[\eta_{i,j}(X_g, X_{g'})\right] = \mathbb{1}_{\{i=j\}}\mathbb{1}_{\{g=g'\}}\eta_{i,i}(g,g)$. *In which case* $\mathbf{K}_e$ *is a diagonal matrix and* $\mathbf{h}_{e,i}(g)$ *is null as long as* g *is not among the observed grains.*

Contrary to Gaussian Process Regression, the prediction cross error $c_{i,j}(g,g')$ defined in Equation (2) is usually not equal to the conditional covariance $\mathbb{E}\left[\mathrm{Cov}\left[Y_i(g), Y_j(g')|\underline{\mathbf{Y}}\right]\right]$. However, under certain assumptions, one can prove that if $M_i(g) = \mathbb{E}\left[Y_i(g)|\underline{\mathbf{Y}}\right]$, then the cross error can also be viewed as a conditional expectation: $c_{i,j}(g,g') = \mathbb{E}\left[\mathrm{Cov}\left[Y_i(g), Y_j(g')|\underline{\mathbf{Y}}\right]\right]$. Details are given in Supplementary material Appendix B.

3. Illustration

3.1. Unidimensional case: rounded inputs

A common issue when feeding geo-statistical models with real data is the precision of the input data and its impact on a model's performance. Usual applications of Kriging take this uncertainty into account when increasing output variables' variances by a value that is known as the nugget effect (e.g. Rocas et al., 2021). Precision being a typical case of input data uncertainty, the example below simulates the effect of rounding input values (coordinates

in the study space) to the nearest units. Let us consider a one-dimensional, centred Gaussian random field $Y(x),\ x \in [1, 10]$ of constant variance. Let us assume that this field is observed at some points for which coordinates are rounded to the nearest unit, i.e., for 2 input values $x_1, x_2 \in]0.5, 1.5]$, the observer sees the same value $\tilde{x}_1 = \tilde{x}_2 = 1$. For a Kriging model, these are multiple observations of the same point, and it is necessary to introduce a nugget effect in the model for the observations' covariance matrix to be invertible. This nugget effect simulates an uncertainty on the output values, while the uncertainty is really on the input values. It rather makes sense to describe those input values as random positions $\tilde{x}_{1,g}$ and $\tilde{x}_{2,g}$ in $g =]0.5, 1.5]$ instead of deterministic $\tilde{x}_1 = \tilde{x}_2 = 1$. Then, we can model the observed objects as mixture distributions and fit a Mixture Kriging model. Let us compare both approaches.

Using the `geoR` package in the `R` language, we simulate a 1-dimensional random field realization with a Gaussian covariance kernel. The specific parameters are detailed in Table 3. x is discretized between 0 and 10 with step 0.05. We pick 8 points for observations as listed in Table 4. These observations are plotted on Figure 3. Observations $\{o1, o2, o6\}$ form the learning set, observations $\{o4, o5, o7\}$ form the validation set, and observations $\{o3, o8\}$ form the test set. These sets of observations are deliberately very small so as to represent the Mixture Kriging's behavior in a readable graphic.

Underlying field		Model properties				Validation	Total
Variance	Range	Model	Variance	Nugget	Range	MSE	MSE
1	4	Kriging	1	10^{-9}	4	0.037	1.14
1	4	Mixture Kriging	1	0	4	0.027	1.18

Table 3: Parameters and performances of fitted models in the case of observations with rounded inputs. Note that the nugget effect for Kriging is the result of an optimization process. For Mixture Kriging, nugget is null by design. Validation MSE: Mean Squared Error on validation set. Total MSE: Mean Squared Error on the complete interval $[0, 10]$.

The Kriging model (Figure 3 left) has repeated observations for $x = 1$ and $x = 3$. The learning set is used to fit a family of models with the same kernel parameters as those used for simulation plus a nugget effect among $(10^{-i})_{i\in\{1,\dots,10\}}$. The nugget effect yielding the smallest mean squared error (MSE) on the test set is selected. A new model is fitted with both learning and test sets using the same kernel and the previously selected nugget effect. This model is applied to compute a validation MSE and a total MSE

		Input			Output
Set	Label	Underlying x (true value)	Rounded x (for Kriging)	Grain (for Mixture Kriging)	y
Learning	$o1$	0.55	1	$g_1 =]0.5, 1.5]$	0.923
Learning	$o2$	0.85	1	$g_2 =]0.5, 1.5]$	1.005
Validation	$o3$	1.65	2	$g_3 =]1.5, 2.5]$	1.127
Test	$o4$	3.00	3	$g_4 =]2.5, 3.5]$	0.946
Test	$o5$	3.45	3	$g_5 =]2.5, 3.5]$	0.801
Learning	$o6$	7.20	7	$g_6 =]6.5, 7.5]$	0.337
Test	$o7$	9.40	9	$g_7 =]8.5, 9.5]$	0.884
Validation	$o8$	9.70	10	$g_8 =]9.5, 10]$	0.908

Table 4: Observations of the simulated Gaussian random field.

computed on all points in $[0, 10]$. The variance of the prediction error is also predicted using the formula given in Proposition 2.

Regarding Mixture Kriging (Figure 3 right), grains $g_1 = [0.5, 1.5[$ and $g_3 = [2.5, 3.5[$ are observed twice each while the other grains are observed once each. The Mixture Kriging model can handle repeated observations by design. Uncertainty on the input is resulting from the random position that generates the observation. The grain covariances are computed from the point covariances as detailed in Proposition 1. The random positions $(X_{g_i})_{i\in\{1,\dots,8\}}$ are assumed to be uniform on the points of the associated grains. Both the learning set and the test set are used to fit a model with the same kernel parameters as for simulation and with no nugget effect. Validation MSE and total MSE are computed for comparison with Kriging.

In this case, the mean prediction is almost the same for both models. But Kriging variance (visible on the ribbons in Figure 3) differs. By construction, Simple Kriging is supposed to interpolate observations exactly, resulting in a very small Kriging variance near observations. Too many observations may be outside the confidence band. If one increases the nugget effect on the Simple Kriging model, mean predictions move towards 0 and their range is reduced. Therefore, with a large nugget effect, one ends up with a nearly constant mean prediction and a large Kriging variance. Mixture Kriging takes into account the input uncertainty and predicts a significant Kriging variance even near observations improving the coverage without any nugget effect.

In this very simple example, the reader may be surprised that both Kriging and Mixture Kriging yield remarkably good predictions. However,

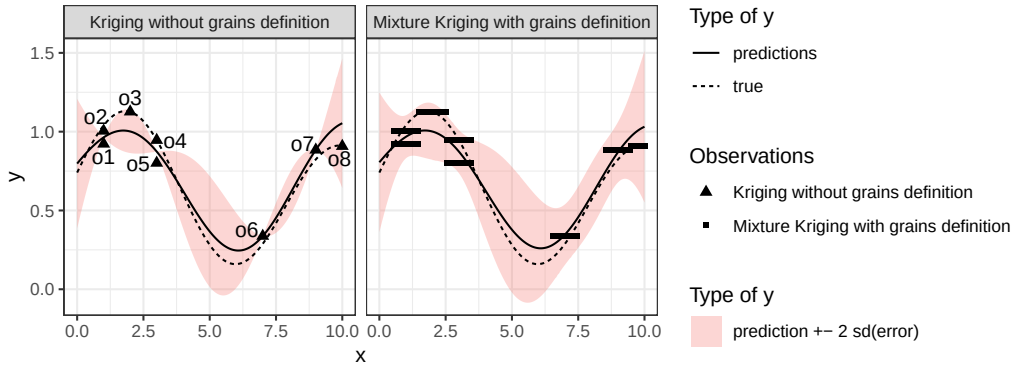

Figure 3: *Rounded inputs.* **Left and right**: The dashed line shows the underlying simulated random field. The solid line labelled "predictions" shows the fitted model mean prediction (see Table 3). The ribbon shows an interval of radius twice the root square of the estimated prediction error variance. **Left**: *Kriging model.* Triangular points show observations. **Right**: *Mixture Kriging.* Horizontal line segments show observations. See Table 4 for more details about observations.

the prediction error value represented by the ribbon's height is important as compared with the predicted values. This means that if the underlying output is also noisy, error can quickly increase. This is the reason why, in real life, one needs much more observations to learn from, see Subsection 3.3.

3.2. Unidimensional case: grains of varying size

Imagine a company that wants to measure some performance indicator for manufactured objects that are produced according to certain design specifications. The design is denoted as x; it belongs to a set χ of permissible values, and $\mathbf{Y}(x)$ is the performance indicator. For instance, $\mathbf{Y}$ can measure the lift of an aircraft wing depending on some shape parameter x. Because of some unavoidable manufacturing precision issues, the manufactured object's characteristics do not match the design's specifications exactly. This uncertainty about the manufactured object induces some uncertainty about the performance indicator. Thus, the constructed design can be viewed as a random vector X_{g_x}, taking values in some tolerance set $g_x \subset \chi$ around the design $x \in \chi$. When testing some designs $x_1, \ldots, x_n$, the industry observes performances $\mathbf{Y}(g_1), \ldots, \mathbf{Y}(g_n)$. Measuring both the expectation and the variance of $\mathbf{Y}(x)$ for each permissible design $x \in \chi$ is one method to find the best design, but this can be costly, so fitting an interpolation model with the set of k observations is preferable. In this setting, for the sake of simplicity, we assume that $\mathbf{Y}(x)$ is conditioned by observations $\{\mathbf{y}(x_i) = \sin(x_i^2) \ : \ i \in \{1, \ldots, n\}\}$.

In this case, we assume that the precision associated with a design x_i is an interval centred on x. The real characteristic of the object having performance $\mathbf{y}(x_i)$ is a random value in this grain, which is assumed to be uniform on all points of the grain.

We compare 3 models:

- P_1: The manufactured object is produced exactly according to the design, the precision interval is restricted to a point.
- P_2: The precision is the same for all designs, the associated interval is of fixed measure.
- P_3: The larger is x, the larger is the uncertainty about the manufactured object, which means that intervals' measures are growing with the design x.

All three models have a null nugget effect and a Gaussian kernel having for variance parameter the overall variance of $\mathbf{y}$ on $\chi = [0, 4]$. The range parameter is optimized by minimizing the mean squared error between $\mathbf{y}$ and point predictions on χ (see Table 4). When grains are restricted to points (Figure 5 top), we get the usual results on Simple Kriging, in particular, predicted values are exactly interpolating observations. When grains are intervals of the same size (Figure 5 middle), predicted values are not interpolating any more; predicted error is not null on the grains but far from the grains, it is smaller than in the previous case. In the bottom figure, the greater is x, the greater the uncertainty about the manufactured object as compared to design. The predicted error (ribbon) is increasing with the grain diameter. Overall, it is important to note that the Mixture Kriging model accounts for the randomness of input values without any nugget effect. This eliminates the adverse consequences of a nugget effect that could otherwise shrink mean predictions towards zero.

In the previous example, a very small set of observations was enough to make very good predictions. In the present one, the situation is different because observations are not drawn from a Gaussian random field but from a deterministic function. This underlying function is modelled as a noisy random field. Therefore, the Kriging error is greater than in the previous case. A potential extension of this illustration would be to optimize both the range and the nugget effect, but the purpose here is to visualize the effect of the uncertainty on the input and not the output.

	Model properties			
Set of grains	Variance	Nugget	Range	Exact interpolation
P_1: Grains are singletons	0.36	0	0.3	Yes
P_2: Grains are of equal measure	0.36	0	0.4	No
P_3: Grains are of increasing measure	0.36	0	0.3	No

Figure 4: Properties of models P_1, P_2, P_3. Range is an optimal value so as to minimize mean squared error.

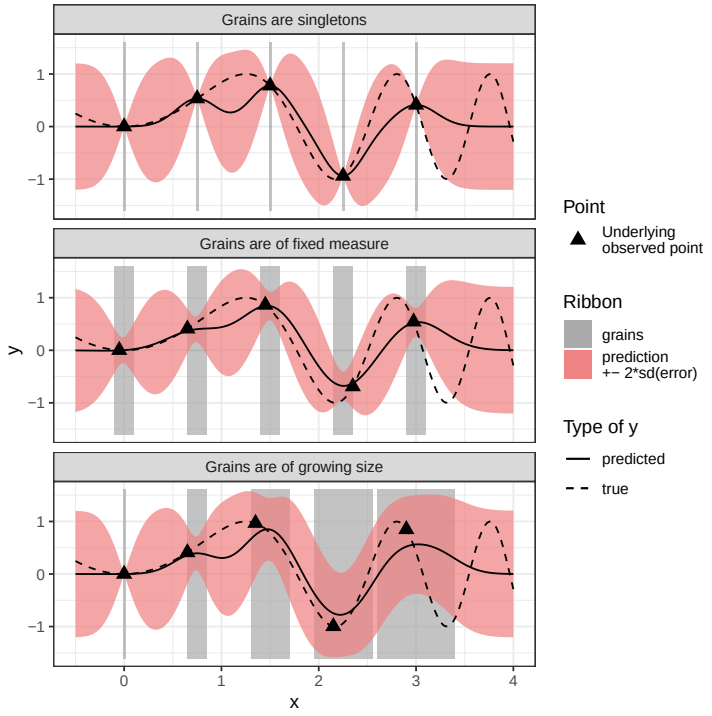

Figure 5: *Mixture Kriging and grain sizes.* The dashed line represents $\mathbf{y}(x)$. The solid line is the mean prediction. The ribbon shows an interval centred on the mean prediction, of radius twice the square root of the predicted error variance. Grey vertical columns show the grains as x intervals. Black triangles show the underlying observed point (observed X_g and associated output value).

3.3. Energy Performance Certificate (EPC) prediction

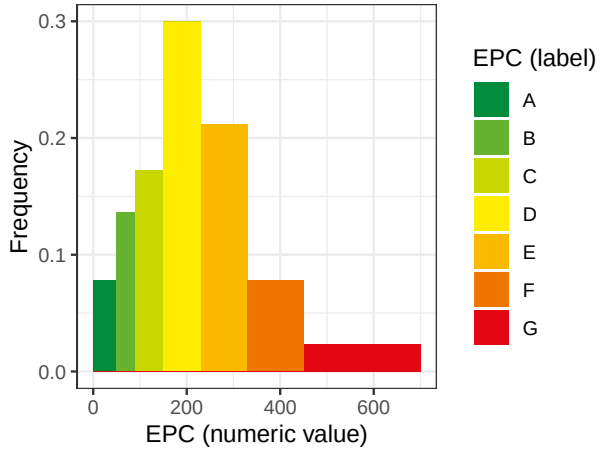

Figure 6: Bar plot of EPC labels frequencies among all EPCs collected in France between 2014 and 2021. Classes A, B, F, G are rare while classes C, D and E are frequent.

Let us now address the EPC prediction problem, keeping in mind that an Energy Performance Certificate (EPC) is given as an energy consumption in $kWh/m^2/year$. The observed distribution of this energy consumption is provided in Figure 6. One considers a model for which χ is a city viewed as a 2-dimensional space with latitude and longitude as coordinates after proper projection, $\mathcal{G}$ is the set of plots, and a point in χ is associated with a given floor square meter of a building on the plot. A floor square meter is regarded here as a granule and not as a set of points in χ. This would not make sense since, for a multi-storey building, there are more floor square meters than the building's footprint area. $x \in \chi$ is therefore a reference point for this floor square meter the same way a point would be used to locate a citizen in a city (see Example 1). $Y(x)$ is the areal energy consumption in x, typically the EPC of the dwelling to which belongs the floor square meter represented by x. Then an EPC in the database is the observed energy efficiency rating associated with one unknown point among those located on the plot indicated by the address. Therefore, for a certain plot g, this EPC is an observation of $Y(X_g)$.

EPC is given as a numeric energy consumption per square meter and per year. This energy consumption is associated with a letter ranging from A to G. A and B label the most energy-saving dwellings (less than $90kWh/m^2/year$). F and G label the most consuming dwellings (more than

$330kWh/m^2/year$). We want to model a situation where we observe EPC with uncertainty on the location of the observed dwelling on the land plot, where it lies, and where the observed dwelling can not be distinguished among all the dwellings of this land plot. We also want to predict an EPC numeric value at the whole land plot level, that is, for the set of dwellings it contains.

As can be seen in Figure 6, observations are strongly unbalanced, meaning that labels A, B, F, and G are rarely observed while labels C, D, and E are very common. As a result, labels A, B, F, and G are difficult to predict, although they are more interesting for decision-makers. Therefore, we introduce the Balanced Accuracy (BA) criterion. It is an asymmetric performance measure that focuses on good results (Gösgens et al., 2021) and it gives the same weight to each class. Denoting n_ℓ the number of observations with label ℓ and $n_{\hat{\ell},\ell}$ the number of predictions $\hat{\ell}$ with true label ℓ (true predictions of label ℓ), the balanced accuracy is given by the formula:

$$BA = \frac{1}{7} \sum_{\ell \in \{A,\dots,G\}} \frac{n_{\hat{\ell},\ell}}{n_\ell}$$

Given a real random variable X and F_X its cdf, supposed to be invertible. Let $H(X) := F_{\mathcal{N}}^{-1} \circ F_X(X)$ where $F_{\mathcal{N}}$ is the standard Gaussian distribution cdf. H is invertible, and $H(X)$ follows a standard Gaussian distribution by the probability integral transform theorem. Using H we normalize input and output variables.

Let us consider the model M_1 such that:

- χ is the territory of an urban area in the French city of Angers in a 3-dimensional space where coordinates represent the image through H of the construction year, the latitude, and the longitude.
- A random field $Y(x)$ is defined on χ. It represents the image through H of the energy consumption per square meter and per year at x.
- A grain g is defined as a set of points in a 3-dimensional space χ. A grain represents a land plot. Each point represents a square meter of living area. It has 3 coordinates. $\mathcal{G}$ denotes the set of all grains.
- For any grain $g \in \mathcal{G}$, the random variable X_g is the uniform law on the points of g. It represents the uncertainty on the observations' location. On g, the output variable is defined as: $Y(g) = Y(X_g)$. By construction, Y is centred.

- A vector of observations of n distinct grains is given and denoted as $\underline{\mathbf{Y}}$.

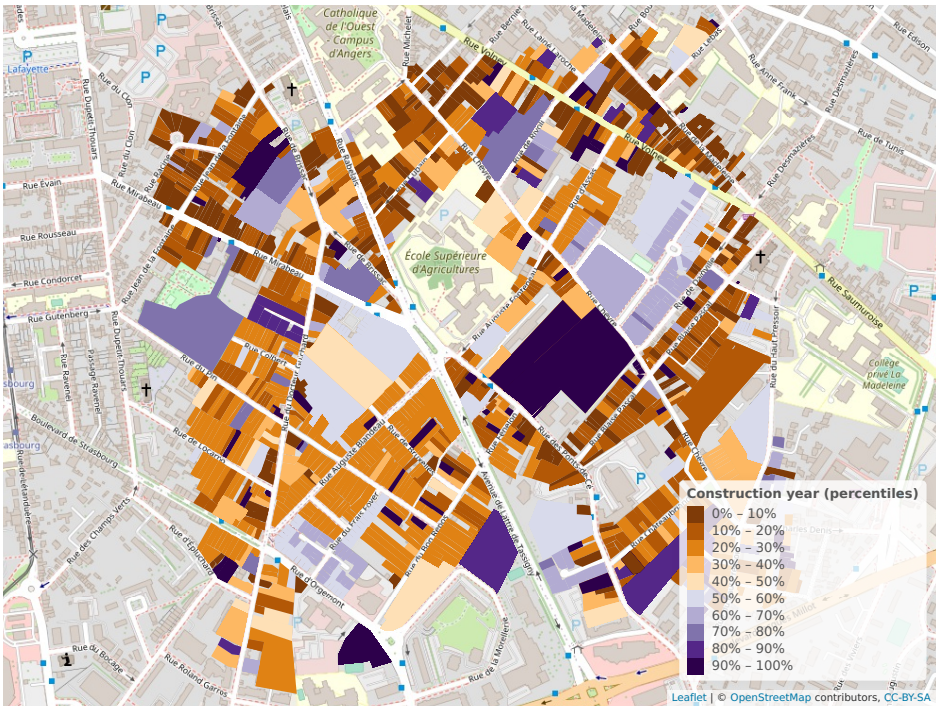

Figure 7: An urban area in Angers: latitude is the vertical dimension, longitude is the horizontal dimension, and construction year is given by the colour. The side of the square is 1km. Construction years range from 1340 (first percentile) to 2019 (last percentile).

$\mathcal{G}$ is mapped in Figure 7. Note that the grains seem to be disjoint, but they are not due to overlaps in the construction year dimension. The set of observations is represented in Figure 8.

For this model, the following assumptions are made:

- For any two distinct grains g, g', random variable X_g is independent from $X_{g'}$.
- For any two points x, x', the covariance between $Y(x)$ and $Y(x')$ is following a Matérn $3/2$ model:

$$\text{Cov}\left[Y(x), Y(x')\right] = \sigma^2 \left(1 + \sum_{i=1}^{3} \frac{|x_i - x_i'|}{\theta_i}\right) \exp\left(-\sum_{i=1}^{3} \frac{|x_i - x_i'|}{\theta_i}\right)$$
$$\text{where } U = (\sigma^2, \theta_1, \theta_2, \theta_3) \in]0, 1] \times]0, +\infty[^3$$

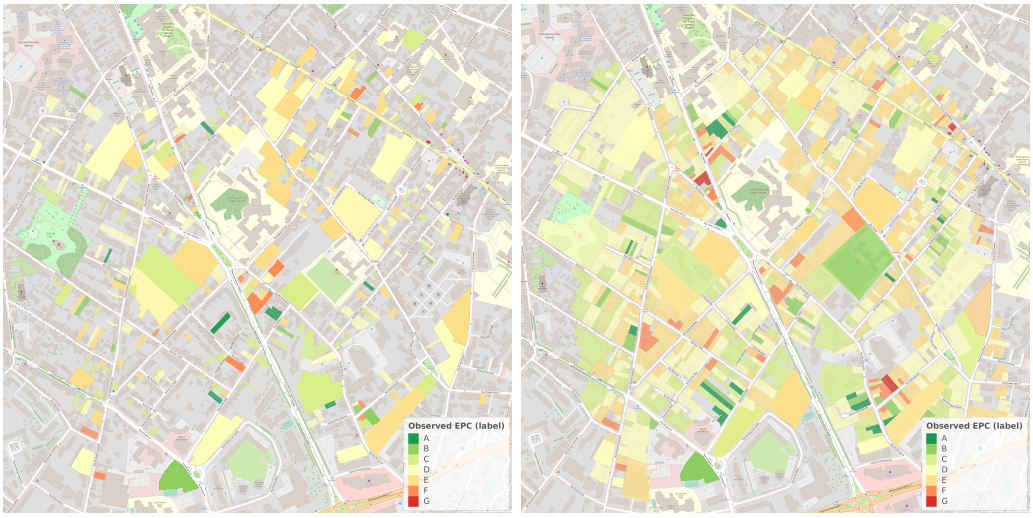

Figure 8: Left: Map of the 365 observations. Right: Map of all predicted values (labels derived from Mixture Kriging means). Each colour represents a label associated with a numeric value. See also Figure 1.

σ^2 is called the variance coefficient, and $\Theta = (\theta_1, \theta_2, \theta_3)$ are the length scale coefficients. Note that no nugget effect is required because the model takes into account the spatial uncertainty of the input by construction.

The Mixture Kriging predictor described in subsection 2.3 is used to predict energy consumption at the plot level. It can be proved that, without the nugget effect, the mean prediction, in the case of a 1-dimensional output, does not depend on σ^2 (the proof is simply deduced from the fact that for an invertible matrix A, we have $(\lambda A)^{-1} = \lambda^{-1} A^{-1}$). σ^2 is therefore set to 1. Θ is chosen so as to maximize the BA criterion of the predicted labels derived from the predicted energy consumptions. BA is computed using leave-one-out cross-validation. Note that the leave-one-out cross-validation predictor that is derived from Proposition 2 is also linear and optimal for quadratic error. A code has been developed in the R language to implement Mixture Kriging.

So as to assess the effect of balanced accuracy on the optimum, we also consider a model $M1'$, which is the same as $M1$ except that parameters are assessed optimizing the accuracy. The accuracy is the total number of labels correctly predicted divided by the number of predictions.

Let us now consider the Kriging model $M2$ to compare performances with the Mixture Kriging model $M1$. $M2$ has the same properties as $M1$ presented above, except that:

- Grains are singletons. A grain $g = \{x^1, ..., x^q\}$ is replaced by a point x of coordinates the minimum construction year and the mean latitude and longitude values. Note that it is assumed that the year of construction of the eldest building portion is the most meaningful information for prediction. This makes sense, especially because the oldest part of a building is usually also the largest one.

- A nugget effect σ_ϵ^2 has to be introduced so as to have a smooth predictor:

$$\mathbb{V}\left[Y(x)\right] = \sigma^2 + \sigma_\epsilon^2 \ .$$

For $M2$, the Kriging predictor is used. $V = (\sigma^2, \theta_1, \theta_2, \theta_3, \sigma_\epsilon^2)$ is chosen so as to maximize BA, the same way as for M_1. The standard R package `DiceKriging` is used for prediction.

There are 365 observations on the given territory. The best parameters are estimated by optimizing the performance indicator, Balanced Accuracy or Accuracy, computed by leave-one-out cross validation. All models $M1, M1'$ and $M2$ are optimized with the genetic algorithm provided by R package `ga` parametrized with population size 50, elitism 5, maximum number of iterations 100, maximum number of iterations without improvement 100. Other parameters are left as default.

With regards to the optimal parameters in Table 5, length scale parameters are smaller in $M1$ than in $M2$, meaning that $M1$ prediction is influenced by fewer neighbours than $M2$. The nugget effect found for $M2$ is small. As for the optimal performances in Table 6, $M1$ reaches a larger BA than $M2$ by 37%. However, $M1$ has lower performances on other indicators with a difference of approximately 10%. The range of all 365 mean predictions with $M1$ is 150% larger than with $M2$. These figures are better understood by examining the confusion matrices in Tables 7 and 8. Indeed, the percentage of large errors (represented by the red area) is 3% with model $M1$ and 0.5% with model $M2$. We know that large errors have an important impact on

Model	ϵ^2	σ^2	θ_1	θ_2	θ_3
Mixture Kriging ($M1$)	*0.00**	*1.00**	0.28	0.44	1.22
Mixture Kriging ($M1'$)	*0.00**	*1.00**	0.93	0.78	0.91
Kriging ($M2$)	0.02	0.53	0.98	0.82	1.49

*: These parameters are treated as constant parameters.

Table 5: Optimal parameters for $M1$ and $M2$.

			EPC int.		EPC num.		
Model	BA	Acc.	MAE	RMSE	MAE	RMSE	Range
Mixture Kriging $M1$	**0.26**	0.40	0.93	1.37	78.93	106.16	**6.66**
Mixture Kriging $M1'$	0.21	**0.42**	0.93	1.38	79.46	108.55	6.54
Kriging $M2$	0.19	0.38	**0.85**	**1.22**	**72.22**	**92.98**	2.59

EPC int.: Energy Performance Certificate treated as an integer: 1 for A, ..., 7 for G.
EPC num.: Energy consumption expressed in $kWh/m^2/year$.
BA: Balanced Accuracy. Acc.: Accuracy.
MAE: Mean Absolute Error. Range: Variance of the predicted values ($\times 10^3$)
RMSE: Root Mean Squared Error. viewed as a measure of the predictions' range.

Table 6: Optimal performances achieved by $M1$, M_1' and $M2$. For each indicator, best value is indicated in bold font.

MAE and RMSE. However, the percentage of true labels A and B that are predicted as A or B is 25% with $M1$ and 10% with $M2$. For labels F and G, these figures are 16% and 0% respectively. This information is valuable for decision-makers seeking to identify energy-efficient and/or energy-intensive dwellings.

These results suggest that Mixture Kriging ($M1$, $M1'$) predictions have an improved range as compared to Kriging ($M2$): the range of mean predictions by Mixture Kriging is greater than by Kriging. This allows better predictions for extreme labels A, B, F, and G. Despite having fewer parameters (ϵ^2 and σ^2 are regarded as constants), Mixture Kriging improves the BA, although it also leads to more frequent large errors. Mixture Kriging accounts for uncertainty in the input data, eliminating the need to add a nugget effect. In this example, it avoids grouping predictions near the mean value (shrinkage) and yields a better BA as compared to Kriging that requires the introduction of a nugget effect.

Among Mixture Kriging models, as expected, $M1$ has a better Balanced Accuracy than $M1'$, and $M1'$ has a better Accuracy than $M1$. Other indi-

True values	Predicted values						
	A	B	C	D	E	F	G
A	2	1	3	2	2	0	0
B	1	3	1	9	2	2	0
C	1	3	25	26	15	4	0
D	3	5	21	80	33	5	1
E	4	2	12	36	36	5	1
F	0	3	2	4	5	3	0
G	0	0	0	1	1	0	0

Table 7: Confusion matrix of $M1$ predictions.

True values	Predicted values						
	A	B	C	D	E	F	G
A	1	0	3	5	1	0	0
B	0	2	1	11	4	0	0
C	0	1	13	48	12	0	0
D	2	1	19	94	32	0	0
E	0	1	9	56	30	0	0
F	1	0	2	11	3	0	0
G	0	0	1	1	0	0	0

Table 8: Confusion matrix of $M2$ predictions.

True values	Predicted values						
	A	B	C	D	E	F	G
A	0	2	2	5	1	0	0
B	1	0	3	9	3	2	0
C	2	2	23	29	14	4	0
D	1	6	17	91	28	2	3
E	1	6	14	31	36	6	2
F	1	0	3	8	2	3	0
G	0	0	1	0	1	0	0

Table 9: Confusion matrix of $M1'$ predictions

	True	Predicted
A	10	11
B	18	17
C	74	64
D	148	158
E	96	94
F	17	19
G	2	2

Table 10: Distribution of labels in $M1$

cators are very similar, let alone the smaller variance of $M1'$'s predictions. Optimizing parameters based on Balanced Accuracy forces the model to predict more often labels A, B, F, and G so that the distribution of predicted labels is very close to the distribution of observed labels as can be seen in Table 10. In our case, the confusion matrices show that this effect is positive for labels A and B, as more true A or B are predicted as A or B. But the effect of balanced accuracy does not bring benefits for labels F and G, on the contrary, it has a tendency to predict more F and G where the true label is D or E. A possible explanation for this moderate benefit of introducing the balanced accuracy is that we are missing some information. The moderate size of observations (365 individuals) makes it difficult for a model to discriminate between rare labels and frequent labels. For instance, there are only 2 observed G labels. One can expect a model learning from a larger number of observations to perform better. Moreover, in an area where buildings are

old for instance, our model cannot distinguish a building that has never been renovated from the others. It may be useful in further studies to introduce more variables, such as a comfort level. However, as discussed below, the proposed model is quite heavy in terms of computation resources; therefore, scaling up or adding variable has an important computational cost.

4. Discussion and conclusion

Since the discovery of Kriging, the issue of learning from and predicting areal data has been a concern. Proposed models have mainly assumed that the output variable at the areal level is the mean of the point outputs, which has proven helpful in various fields such as mining, climatology, or satellite imaging, where averaging makes sense for interpretation and where blocks tend to have similar shapes and sizes. However, in other fields such as agriculture or social studies, blocks can have varying shapes or sizes, and averaging is not always the most meaningful interpretation. In these cases, problems like the Modifiable Areal Unit Problem (MAUP), the ecological inference problem, and the variance reduction due to averaging can become challenging to solve. Over the past few decades, researchers have been developing methods to assess and/or correct the MAUP effect (Briz-Redon, 2022). Modifying territory partitioning (Li et al., 2009) is also an effective solution for addressing variance reduction problems, but it is not always possible. Both Kriging and block-Kriging incorporate uncertainties on input and/or output values through the addition of a nugget effect to variances, thereby simulating the addition of a white noise to the output variables. This transformation smooths predicted values but also shrinks them; the range between minimal and maximal predicted values is reduced, thus degrading the prediction quality of values that are particularly large or particularly small.

The availability of new datasets with uncertainty on the inputs (uncertain positions) and where averaging is not a meaningful interpretation has driven us to seek a novel method of spatial interpolation. We have introduced a new element in the model that is a random input value. It has been found that resulting mixture distributions can be interpolated optimally, and the resulting Best Linear Unbiased Predictor (BLUP) requires only the first 2 moments of the prior random field and a spatial covariance function. This model can learn from and predict outputs associated with grains of any shape, size, or cardinality. Even single points are acceptable. The term "grain" has been introduced to describe these objects.

The new model called Mixture Kriging is still consistent with Kriging in the sense that Kriging is a special case of Mixture Kriging where grains are restricted to singletons. However, Mixture Kriging generates a mean prediction range that is not impacted by the grain's shape or size under usual conditions. As a consequence, there is no reduction in the mean prediction's range due to this factor. If the output variable's variance is the same everywhere at point level, then it is also the same as the output variable's variance at grain level, meaning that there is no variance reduction either. Similarly, if the covariance between the output variable of interest and another output variable is the same everywhere at the point level, then it will also be the same as the covariance at the grain level, regardless of the grain's shape. This implies that this model has no measurable MAUP effect in the sense of Briz-Redon (2022).

Without any MAUP effect, the Mixture Kriging approach is able to handle multi-scale data. We hope that this can help handling datasets coming from multiple sources in the same model. This model can potentially be used to fit ecological data or social data. For instance, on a global scale, the Intergovernmental Panel on Climate Change (IPCC) studies planetary boundaries on water based on gridded data, but for the study of specific territories, studies are commonly done on watersheds. The Mixture Kriging model has the ability to combine these two scales of study in the same model to benefit from both global and local studies.

The main computational distinction between block-to-block Kriging and Mixture Kriging lies in the method of computing the observations variance and the covariance between covariates associated with the same grain. This results mainly in the diagonal of the observations covariance matrix being greater than what is found with Kriging. This is precisely the sought effect when introducing supplementary noise on the outputs (nugget effect) in Kriging for smoothing predictions. This explains why Mixture Kriging has smooth predictions but with limited shrinkage, hence a good performance with Balanced Accuracy. In practical applications, Mixture Kriging is therefore designed to handle data with uncertainty on the input values without introducing the nugget effect.

Regarding computational differences, it should also be noted that Mixture Kriging (like block-to-block Kriging) has a higher computational cost than Kriging, this cost is growing like the squared value of the density of points in the grains. This is an important limitation of the model. For instance, in the models $M1$, $M1'$ and $M2$ presented in Subsection 3.3, there are 395

observations. The Kriging model $M2$ requires $365 \times 366/2 = 66,795$ covariances to be computed. But the Mixture Kriging models $M1$ and $M1'$ require to compute $3,770,500,618$ point-to-point covariance in order to compute the $66,795$ covariances between grains. Scaling up the model may, therefore, be difficult. This computational complexity is highly dependent on the definition of the random position X_g for each grain and on its discretization. For the above models, X_g is supposed to be uniform for all grains, and the number of discretized points is the number of square meters of living space on the grain. But any new model based on Mixture Kriging requires an appropriate definition of these random variables, depending both on the grains' geometries and on the studied output variable(s). Another limitation of the model is the difficulty of assessing its parameters, especially the range. It is difficult to compute a variogram because there is no natural definition of a distance between grains. Estimating the range is also possible by minimizing an error measure, but this process requires computing numerous different models, which is costly, as mentioned above.

Keeping in mind its limitations, this new approach opens the way for implementing Mixture Kriging models with new datasets that have been impossible to fit in the usual Kriging framework or with usual Kriging software packages. In particular, datasets that inform about granules that are uncertainly defined, such as dwellings, buildings, streets, human persons, and households. It can also be used for datasets informing about granules, which should have deterministic shapes or positions in the input space, but come with numerical uncertainty such as measure precision, rounding effect, observations' aggregations, or observations' anonymization. Moreover, the model can handle multivariate outputs, even if some output components are missing in the observations. Encouraging results have been found when studying the prediction of Energy Performance Certificates (EPC). Results show that Mixture Kriging can be useful to improve the prediction of values far from the average and, in our case, to improve the detection of energy-saving homes. Future studies should test the upscaling feasibility of the already developed model and the benefits of using covariates. We also study the possibility of developing a similar model with Universal Kriging.

Acknowledgements

The authors acknowledge support from the URBS enterprise, www.urbs.fr. They thank in particular Maximilien Brossard for careful reading and constructive comments.

This research was jointly supported by Mines Saint-Etienne graduate engineering school and research institute (https://www.mines-stetienne.fr/en/), URBS enterprise (https://www.imope.fr/) and French National Agency for Research and Technology (https://www.anrt.asso.fr/fr).

Bibliography

Aldworth, W. J. K. (1998). *Spatial prediction, spatial sampling, and measurement error*. Doctor of Philosophy, Iowa State University.

Ali, U., Shamsi, M. H., Bohacek, M., Hoare, C., Purcell, K., Mangina, E., and O'Donnell, J. (2020). A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings. *Applied Energy*, 267.

Baker, E., Challenor, P., and Eames, M. (2021). Future Proofing a Building Design using History Matching Inspired Level-set Techniques. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70(2):335–350.

Ballarini, I., Corrado, V., Madonna, F., Paduos, S., and Ravasio, F. (2017). Energy refurbishment of the Italian residential building stock: energy and cost analysis through the application of the building typology. *Energy Policy*, 105:148–160.

Briz-Redon, A. (2022). A Bayesian shared-effects modeling framework to quantify the modifiable areal unit problem. *Spatial Statistics*, 51.

Comber, A. and Zeng, W. (2019). Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations. *Geography Compass*, 13(10). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gec3.12465.

Cressie, N. A. (1993). *Statistics for spatial data revised edition*. Wiley series in probability and mathematical statistics. Applied probability and statistics., New York, john wiley & sons inc. edition.

Godoy, L. d. C., Prates, M. O., and Yan, J. (2022). An unified framework for point-level, areal, and mixed spatial data: the Hausdorff-Gaussian Process.

Goovaerts, P. (2008). Kriging and Semivariogram Deconvolution in the Presence of Irregular Geographical Units. *Mathematical Geology*, 40(1):101–128.

Gösgens, M., Zhiyanov, A., Tikhonov, A., and Prokhorenkova, L. (2021). Good Classification Measures and How to Find Them. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17136–17147, New York. Curran Associates, Inc.

Gotway, C. and Young, L. (2002). Combining Incompatible Spatial Data. *Journal of the American Statistical Association*, 97:632–648.

Isaaks, E. H. and Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*. Oxford University Press, New York.

Jin, Y., Ge, Y., Wang, J., Heuvelink, G., and Wang, L. (2018). Geographically Weighted Area-to-Point Regression Kriging for Spatial Downscaling in Remote Sensing. *Remote Sensing*, 10.

Kerry, R., Goovaerts, P., Smit, I. P., and Ingram, B. R. (2013). A comparison of multiple indicator kriging and area-to-point Poisson kriging for mapping patterns of herbivore species abundance in Kruger National Park, South Africa. *International journal of geographical information science (IJGIS)*, 27(1):47–67.

Kyriakidis, P. (2004). A Geostatistical Framework For Area-To-Point Spatial Interpolation. *Geographical Analysis*, 36.

Lam, N. S.-N. (1983). Spatial Interpolation Methods: A Review. *The American Cartographer*, 10(2):129–150.

Li, C., Lu, Z., Ma, T., and Zhu, X. (2009). A simple kriging method incorporating multiscale measurements in geochemical survey. *Journal of Geochemical Exploration*, 101(2):147–154.

Lin, Y.-P., Cheng, B.-Y., Shyu, G.-S., and Chang, T.-K. (2010). Combining a finite mixture distribution model with indicator kriging to delineate and map the spatial patterns of soil heavy metal pollution in Chunghua County, central Taiwan. *Environmental Pollution*, 158(1):235–244.

Moraga, P., Cramb, S. M., Mengersen, K. L., and Pagano, M. (2017). A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. *Spatial Statistics*, 21:27–41.

Pedrycz, W. (2013). *Granular computing : analysis and design of intelligent systems*. Industrial electronics series. Taylor & Francis, Boca Raton.

Pereira, O. J. R., Melfi, A. J., Montes, C. R., and Lucas, Y. (2018). Downscaling of ASTER Thermal Images Based on Geographically Weighted Regression Kriging. *Remote Sensing*, 10(4):633.

Poggio, L. and Gimona, A. (2015). Downscaling and correction of regional climate models outputs with a hybrid geostatistical approach. *Spatial Statistics*, 14:4–21.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass.

Rocas, M., García-González, A., Zlotnik, S., Larráyoz, X., and Díez, P. (2021). Nonintrusive uncertainty quantification for automotive crash problems with VPS/Pamcrash. *Finite Elements in Analysis and Design*, 193:103556.

Roksvåg, T., Steinsland, I., and Engeland, K. (2021). A Two-Field Geostatistical Model Combining Point and Areal Observations—A Case Study of Annual Runoff Predictions in the Voss Area. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70(4):934–960.

Schetelat, P., Lefort, L., and Delgado, N. (2020). Urban data imputation using multi-output multi-class classification. *Building to Buildings: Urban and Community Energy Modelling*.

Smith, B. J., Yan, J., and Cowles, M. K. (2008). Unified Geostatistical Modeling for Data Fusion and Spatial Heteroskedasticity with R Package ramps. *Journal of Statistical Software*, 25:1–21.

Truong, P. and Heuvelink, G. (2013). Bayesian Area-to-Point Kriging using Expert Knowledge as Informative Priors. *International Journal of Applied Earth Observation and Geoinformation*, 30:2291.

Wang, Q., Shi, W., and Atkinson, P. M. (2016). Area-to-point regression kriging for pan-sharpening. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:151–165.

Williams, C. and Rasmussen, C. (1996). Gaussian Processes for Regression. In *Proceedings of the 1995 Conference*, Cambridge, Mass. The MIT Press.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. CRC Press, Boca Raton.

Yoo, E.-H. and Kyriakidis, P. C. (2006). Area-to-point Kriging with inequality-type data. *Journal of Geographical Systems*, 8(4):357–390.

Zadeh, L. A. (2005). Toward a generalized theory of uncertainty (GTU)—an outline. *Information Sciences*, 172(1):1–40.

Zhang, X., Zuo, W., Zhao, S., Jiang, L., Chen, L., and Zhu, Y. (2018). Uncertainty in Upscaling In Situ Soil Moisture Observations to Multiscale Pixel Estimations with Kriging at the Field Level. *ISPRS International Journal of Geo-Information*, 7(1).

Appendix A. Proof of Proposition 2 (supplementary material)

This proof employs a classical statistical approach to compute a Best Linear Unbiased Predictor using a family of observed random variables that are not necessarily Gaussian but have known first and second moments.

It is interesting for the understanding of the problem to give it a geometrical approach. Let us denote $F_i(g)$ the set of linear unbiased predictors of $Y_i(g)$ given an observation vector $\underline{\mathbf{Y}}$. With previous notations, it means that:

$$F_i(g) \;:=\; \left\{\boldsymbol{\alpha}^\top \underline{\mathbf{Y}} \,:\, \mu_i(g) = \boldsymbol{\alpha}^\top \underline{\boldsymbol{\mu}}\right\}$$

And similarly, we denote:

$$\begin{aligned}
G_i(g) &:= \{\alpha Y_i(g) \,:\, \alpha \in \mathbb{R}\} \\
F &:= \left\{\boldsymbol{\alpha}^\top \underline{\mathbf{Y}} \,:\, \boldsymbol{\alpha} \in \mathbb{R}^n\right\} \text{ (the feature space generated by observations)} \\
F_0 &:= \left\{\boldsymbol{\alpha}^\top \underline{\mathbf{Y}} \,:\, \boldsymbol{\alpha}^\top \underline{\boldsymbol{\mu}} = 0\right\} \\
H &:= F \times G_i(g)
\end{aligned}$$

One can note that F_0 is a subspace of F of dimension $\dim(F) - 1$. Moreover $F_0 + F_i(g) = F_i(g)$, meaning that $F_i(g)$ is an affine subspace of F having F_0 for underlying vector space (see Figure A.9). But it also means that the sets of unbiased linear predictors for each output variable are parallel:

$$\forall i, j \in \{1, \ldots, p\},\ \forall g, g' \in \chi,\ F_i(g) \parallel F_j(g')$$

Now, given that we are minimizing the quadratic error between $Y_i(g)$ and $M_i(g)$, which can be seen as the distance between $Y_i(g)$ and $M_i(g)$ in H, the optimization process is geometrically a projection of $Y_i(g)$ on $F_i(g)$. This approach is illustrated in Figure A.9.

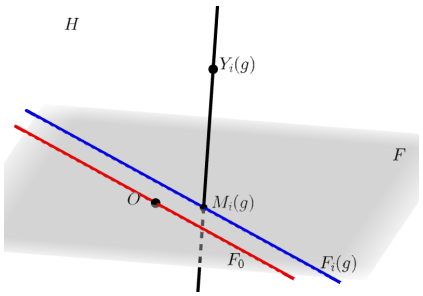

Figure A.9: *Geometrical interpretation of the prediction process.*

Proof. For given $i \in \{1, \dots, p\}$ and $g \subseteq \chi$, let $M_{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^\top \underline{\mathbf{Y}}$ be a linear predictor of $Y_i(g)$, where $\boldsymbol{\alpha} = (\alpha^1, \dots, \alpha^n)$ is a vector of weights, and denote the associated error $v_i(g, \boldsymbol{\alpha}) := \mathbb{E}\left[(Y_i(g) - M_{\boldsymbol{\alpha}})^2\right]$, then:

$$\begin{aligned} v_i(g, \boldsymbol{\alpha}) &= \mathbb{E}\left[\left(\boldsymbol{\alpha}^\top \underline{\mathbf{Y}} - Y_i(g)\right)^2\right] \\ &= \mathbb{E}\left[\boldsymbol{\alpha}^\top \underline{\mathbf{Y}}\, \underline{\mathbf{Y}}^\top \boldsymbol{\alpha} - 2Y_i(g)\boldsymbol{\alpha}^\top \underline{\mathbf{Y}} + Y_i(g)^2\right] \\ &= \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \underline{\boldsymbol{\mu}}\, \underline{\boldsymbol{\mu}}^\top \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top \left(\mathbf{h}_i(g) + \underline{\boldsymbol{\mu}} \mu_i(g)\right) + \mathbb{V}\left[Y_i(g)\right] + \mu_i(g)^2 . \end{aligned}$$

(i) If $\underline{\boldsymbol{\mu}} = (0, \dots, 0)^\top$ and $\mu_i(g) = 0$ then

$$v_i(g, \boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top \mathbf{h}_i(g) + \mathbb{V}\left[Y_i(g)\right].$$

By differentiation over each component of $\boldsymbol{\alpha}$,

$$\frac{\partial v_i(g, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} := \left(\frac{\partial v_i(g, \boldsymbol{\alpha})}{\partial \alpha^j}\right)_{j \in \{1, \dots, p\}} = 2\mathbf{K}\boldsymbol{\alpha} - 2\mathbf{h}_i(g) .$$

Without constraints, this value should be null at any extremum, and thus the optimal vector of weights is

$$\boldsymbol{\alpha}_i(g) = \mathbf{K}^{-1} \mathbf{h}_i(g) .$$

Since $\mathbf{K}$ is symmetric positive, this only extremum is a minimum.

(ii) If $\underline{\boldsymbol{\mu}} \neq (0, \dots, 0)^\top$ then the condition for unbiasedness writes $\mu_i(g) = \boldsymbol{\alpha}^\top \underline{\boldsymbol{\mu}}$ by linearity of expectation.

$v_i(g, \boldsymbol{\alpha})$ rewrites again:

$$v_i(g, \boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top \mathbf{h}_i(g) + \mathbb{V}\left[Y_i(g)\right].$$

We introduce the Lagrangian operator:

$$\mathscr{L}(\boldsymbol{\alpha}, \lambda) = v_i(g, \boldsymbol{\alpha}) - 2\lambda(\boldsymbol{\alpha}^\top \underline{\boldsymbol{\mu}} - \mu_i(g)) \ .$$

We are minimizing a quadratic function over a single affine equality constraint. A necessary optimality condition is:

$$\frac{\partial \mathscr{L}}{\partial \boldsymbol{\alpha}}(\boldsymbol{\alpha}, \lambda) = 0 \, ,$$

that is to say:

$$2\mathbf{K}\boldsymbol{\alpha} - 2\mathbf{h}_i(g) - 2\lambda\underline{\boldsymbol{\mu}} = 0 \, ,$$

and therefore the optimal weights are

$$\boldsymbol{\alpha}_i(g) = \mathbf{K}^{-1}(\mathbf{h}_i(g) + \lambda\underline{\boldsymbol{\mu}}) \, .$$

The unbiasedness condition is:

$$\underline{\boldsymbol{\mu}}^\top(\mathbf{K}^{-1}(\mathbf{h}_i(g) + \lambda\underline{\boldsymbol{\mu}})) = \mu_i(g) \, ,$$

so that

$$\lambda_i(g) = \frac{\mu_i(g) - \underline{\boldsymbol{\mu}}^\top \mathbf{K}^{-1}\mathbf{h}_i(g)}{\underline{\boldsymbol{\mu}}^\top \mathbf{K}^{-1}\underline{\boldsymbol{\mu}}} \, .$$

933 Therefore this only solution is a minimum of $v_i(g, \boldsymbol{\alpha})$.

Let us consider now the cross-errors:

$$c_{i,j}(g, g') = \mathbb{E}\left[(Y_i(g) - M_i(g))\,(Y_j(g') - M_j(g'))\right] .$$

934 Due to unbiasedness condition, it means that:

$$\begin{aligned}
c_{i,j}(g, g') &= \mathrm{Cov}\left[Y_i(g) - M_i(g), Y_j(g') - M_j(g')\right] \\
&= \mathrm{Cov}\left[Y_i(g), Y_j(g')\right] - \mathrm{Cov}\left[Y_i(g), M_j(g')\right] \\
&\quad - \mathrm{Cov}\left[M_i(g), Y_j(g')\right] + \mathrm{Cov}\left[M_i(g), M_j(g')\right] \\
&= \mathrm{Cov}\left[Y_i(g), Y_j(g')\right] - \mathrm{Cov}\left[Y_i(g), \boldsymbol{\alpha}_j(g')^\top \underline{\mathbf{Y}}\right] - \mathrm{Cov}\left[\boldsymbol{\alpha}_i(g)^\top \underline{\mathbf{Y}}, Y_j(g')\right] \\
&\quad + \mathrm{Cov}\left[\boldsymbol{\alpha}_i(g)^\top \underline{\mathbf{Y}}, \boldsymbol{\alpha}_j(g')^\top \underline{\mathbf{Y}}\right] .
\end{aligned}$$

935 Which rewrites:

$$c_{i,j}(g, g') = k_{i,j}(g, g') - \boldsymbol{\alpha}_j(g')^\top \mathbf{h}_i(g) - \boldsymbol{\alpha}_i(g)^\top \mathbf{h}_j(g') + \boldsymbol{\alpha}_i(g)^\top \mathbf{K}\boldsymbol{\alpha}_j(g') \ . \quad \text{(A.1)}$$

936 Note that equation (A.1) is true for any linear unbiased predictor. Which, in the case of simple Mixture Kriging, simplifies into:

$$c_{i,j}(g,g') = k_{i,j}(g,g') - \mathbf{h}_i(g)^\top \mathbf{K}^{-1}\mathbf{h}_j(g')\,.$$

And in the case of ordinary Mixture Kriging:

$$c_{i,j}(g,g') = k_{i,j}(g,g') - \mathbf{h}_i(g)^\top \mathbf{K}^{-1}\mathbf{h}_j(g') + \lambda_i(g)\lambda_j(g)\underline{\boldsymbol{\mu}}^\top \mathbf{K}^{-1}\underline{\boldsymbol{\mu}}\,.$$

937 The expressions of $v_i(g) = c_{i,i}(g,g)$ in both cases follow immediately. □

Appendix B. Cross-errors and conditional covariances (supplementary material)

It is well known that the best predictor of $Y_i(g)$ is also the conditional variable $\mathbb{E}\left[Y_i(g)|\underline{\mathbf{Y}}\right]$. However, this best predictor is not necessarily linear, especially in non Gaussian cases. The following proposition proves that if the Best Linear Unbiased Predictor is the best overall predictor then the error covariances can also be seen as conditional covariances.

Proposition 3 (Cross-errors and conditional covariances)**.** *Consider the assumption*

$$(\mathrm{A}): \quad \forall i \in \{1,\ldots,p\},\ \forall g \in \mathcal{G},\ M_i(g) = \mathbb{E}\left[Y_i(g)|\underline{\mathbf{Y}}\right].$$

Under assumption (A)*, cross errors for both Simple Mixture Kriging and Ordinary Mixture Kriging are:*

$$c_{i,j}(g,g') = \mathbb{E}\left[\mathrm{Cov}\left[Y_i(g), Y_j(g')|\underline{\mathbf{Y}}\right]\right]. \tag{B.1}$$

Moreover, if $\mathrm{Cov}\left[Y_i(g), Y_j(g')|\underline{\mathbf{Y}}\right]$ *does not depend on* $\underline{\mathbf{Y}}$*, as it is the case for conditional Gaussian vectors, Equation* (B.1) *simplifies:* $\mathbb{E}\left[\mathrm{Cov}\left[Y_i(g), Y_j(g')|\underline{\mathbf{Y}}\right]\right] = \mathrm{Cov}\left[Y_i(g), Y_j(g')|\underline{\mathbf{Y}}\right]$.

Assumption (A) holds for example when $\{\mathbf{Y}(x) : x \in \chi\}$ is a vector-valued Gaussian random field and when each X_g is Dirac distributed.

Proof. The proof uses a classical approach on orthogonality of Best Linear Unbiased Predictors. It is presented here in three steps. The proof can be simplified in the Simple Mixture Kriging setting.

- First, given the notations introduced in Appendix A, let $\delta \in F_0$ be a non-zero vector and β a real number.

  Let $M_i^{\beta}(g) := M_i(g) + \beta\,\delta \in F_i(g)$. Recall that $\epsilon_i(g) := Y_i(g) - M_i(g)$ and $v_i(g) := \mathbb{E}\left[(\epsilon_i(g))^2\right]$.

  We have:

  $$\mathbb{E}\left[\left(Y_i(g) - M_i^{\beta}(g)\right)^2\right] = v_i(g) - 2\beta\mathbb{E}\left[\epsilon_i(g)\,\delta\right] + \beta^2\mathbb{E}\left[\delta^2\right].$$

  The minimum value of this polynomial expression is reached for:

  $$\beta_0 = \frac{\mathbb{E}\left[\epsilon_i(g)\,\delta\right]}{\mathbb{E}\left[\delta^2\right]}.$$

Since the only optimal point is $M_i(g)$, $M_i^{\beta_0}(g) = M_i(g)$ and therefore $\beta_0 = 0$. As a consequence, as both $\mathbb{E}[\epsilon_i(g)] = 0$ and $\mathbb{E}[\delta] = 0$:

$$\forall \delta \in F_0,\ \forall i \in \{1, \ldots, p\},\ \forall g \in \chi,\ \ \mathbb{E}[\epsilon_i(g)\,\delta] = \mathrm{Cov}[\epsilon_i(g), \delta] = 0\,. \tag{B.2}$$

From a geometrical point of view it is equivalent to say that the inner product of the error and any vector of F_0, such as the difference of any linear unbiased predictors of $Y_j(g')$, is null. This approach can be found for example in Aldworth (1998), section 4.5.1. page 122, in the case of ordinary Kriging on a stationary process.

- Now, let δ and δ' be any two vectors of F_0. As a consequence of the previous result in Equation (B.2), we have:

$$\mathrm{Cov}[\epsilon_i(g) + \delta, \epsilon_j(g') + \delta'] = c_{i,j}(g, g') + 0 + 0 + \mathrm{Cov}[\delta,\, \delta'] \tag{B.3}$$

- On the other hand, using the conditional covariance formula, we have:

$$\begin{aligned}\mathrm{Cov}[\epsilon_i(g) + \delta, \epsilon_j(g') + \delta'] = {} & \mathbb{E}[\mathrm{Cov}[\epsilon_i(g) + \delta, \epsilon_j(g') + \delta' \mid \underline{\mathbf{Y}}]] \\ & + \mathrm{Cov}[\mathbb{E}[\epsilon_i(g) + \delta \,|\, \underline{\mathbf{Y}}], \mathbb{E}[\epsilon_j(g') + \delta' \,|\, \underline{\mathbf{Y}}]]\end{aligned}$$

Given a $\underline{\mathbf{Y}}$, the random variables $\delta,\ \delta',\ M_i(g)$ and $M_j(g')$ are constant, so that the first term is

$$\mathbb{E}[\mathrm{Cov}[\epsilon_i(g) + \delta, \epsilon_j(g') + \delta' \mid \underline{\mathbf{Y}}]] = \mathbb{E}[\mathrm{Cov}[Y_i(g), Y_j(g') \mid \underline{\mathbf{Y}}]].$$

Furthermore, we have assumed in Assumption (A) that

$$M_i(g) = \mathbb{E}[Y_i(g)|\underline{\mathbf{Y}}] \text{ and } M_j(g') = \mathbb{E}[Y_j(g')|\underline{\mathbf{Y}}]\ ,$$

therefore

$$\begin{gathered}\mathbb{E}[\epsilon_i(g)|\underline{\mathbf{Y}}] = \mathbb{E}[\epsilon_j(g')|\underline{\mathbf{Y}}] = 0 \\ \text{and } \mathrm{Cov}[\epsilon_i(g) + \delta, \epsilon_j(g') + \delta'] = \mathbb{E}[\mathrm{Cov}[Y_i(g), Y_j(g') \mid \underline{\mathbf{Y}}]] + \mathrm{Cov}[\delta, \delta']\end{gathered} \tag{B.4}$$

Identifying the equations (B.3) and (B.4), we get the expected result.

□

Contents