Spatial interpolation using mixture distributions: A Best Linear Unbiased Predictor

Marc Grossouvre, Didier Rullière, Jonathan Villot

Spatial interpolation using mixture distributions: A Best Linear Unbiased Predictor

Marc Grossouvre [*†] Didier Rullière [‡] Jonathan Villot [§]

Abstract

Planning the energy transition requires decision makers to have an in-depth knowledge about a given territory. To achieve this, data is collected from multiple sources, at multiple scales, with constraints such as privacy policies. Resulting data informs about given areas of space without a specific point location. Such is the case of Energy Performance Certificate (EPC). EPC databases are released under specific constraints: anonymization, geo-localization with postal address, missing details. This paper shows that learning the observed EPCs to predict missing ones can also be seen as a spatial interpolation problem. It presents a way to treat EPC as a geo-localized information and predict its value at building level.

Kriging methodology is applied to random fields observed at random locations to find a Best Linear Unbiased Predictor (BLUP). This new model is referred to as Mixture Kriging. While the usual Gaussian setting is lost, we show that conditional mean, variance and covariance can be derived. This new model gives interesting results in EPC prediction at building level which is a prerequisite for decision makers to target renovation efforts. The specific case of a city in France is taken as an example. The presented model includes Mixture coKriging so that covariates, even with missing observations, can be used to improve the result. It is also suggested that Mixture Kriging can be usefully implemented to control uncertainty propagation. We present potential applications on simulated data.

Keywords— multi-scale processes, area-to-point regression, areal data, block Kriging, change of support, energy transition

1 Introduction

1.1 Classifying the EPC prediction problem in research

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

[*]Marc Grossouvre, U.R.B.S. SAS, Bâtiment des Hautes Technologie, 20 Rue Professeur Benoît LAURAS, 42000 Saint-Etienne, France, marcgrossouvre@urbs.fr, website `www.urbs.fr`.

[†]Marc Grossouvre, Mines Saint-Etienne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, Departement GMI, Espace Fauriel, 29 rue Ponchardier, F - 42023 Saint-Etienne, France, marc.grossouvre@emse.fr

[‡]Didier Rullière, Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, Departement GMI, Espace Fauriel, 29 rue Ponchardier, F - 42023 Saint-Etienne, France. didier.rulliere@emse.fr

[§]Jonathan Villot, Mines Saint-Etienne, Univ Lyon, CNRS, Univ Jean Monnet, Univ Lumière Lyon 2, Univ Lyon 3 Jean Moulin, ENS Lyon, ENTPE, ENSA Lyon, UMR 5600 EVS, Institut Henri Fayol, F - 42023 Saint-Etienne France. jonathan.villot@emse.fr

along the present paper.

In literature, this problem can be approached from an engineering perspective, from a data management one and from a geostatistics points 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. To work at a larger scale, they 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 building level. Some features 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.

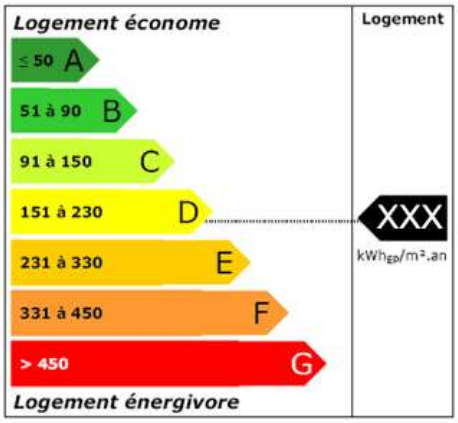

Figure 1: Prescribed vignette appearing on the French energy certificate up to 2021. The meaning of the legend appearing at the top left is efficient dwelling; top right: dwelling; bottom: energy intensive dwelling.

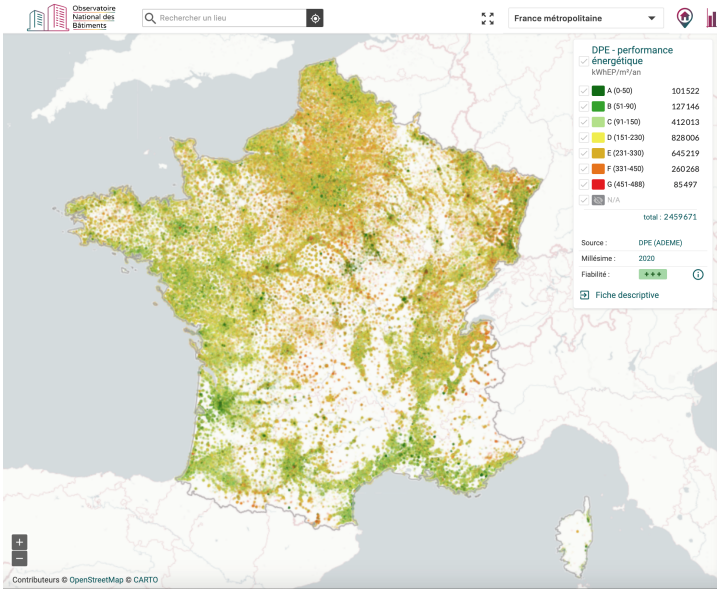

Figure 2: Map of French inventoried EPCs. This image is a screen capture of the French National Observatory of Buildings (Observatoire National des Bâtiments - ONB), released with the consent of the rights holders U.R.B.S. SAS.

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 of buildings since many governments lack a detailed map of 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. 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.

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 dimensional space but it has features that describe it. And to predict EPC of buildings, one also has to define buildings. Data fusion requires to define the same way smallest units of information, also known as a granules for each dataset[1] . The field of study that focuses on representing, constructing and processing these information granules is called Granular Computing

[1] "Informally, a granule of a variable X is a clump of values of X which are drawn together by indistinguishability, equivalence, similarity, proximity or functionality. For example, an interval is a granule." Zadeh [2005]

(Pedrycz [2013]). Assuming that dwellings, EPC observations and complete buildings are represented in a same data model, meaning that an appropriate data fusion process is implemented, a relevant predictive model should now be constructed. Granular computing is multidisciplinary but since we are dealing with geo-localized information, the natural field of research is geostatistics which have been defined as "dealing with spatial processes indexed over continuous space" (Cressie [1993], p7).

From a geostatistics perspective, among other issues, the irreducible uncertainty about granules' positions (dwellings, buildings...) in their underlying space restricts the use of traditional spatial interpolation models such as Kriging. 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 are already identified and that several solutions have been proposed with their benefits and shortcomings.

1.2 The limits of systematic averaging for spatial interpolation

As defined by Comber and Zeng [2019], spatial interpolation "is a technique which uses sample values of known geographical points (or areal units) to estimate (or predict) values at other unknown points (or area units)". The same article presents a summary table of the major spatial interpolation approaches among which is Gaussian Process Regression (Williams and Rasmussen [1996]), also known as Kriging. Kriging theory was first published by Matheron [1963] based on Daniel Krige's master thesis. It relies on the general assumption that points close to each other in the input space are more likely to have similar output values. The original article states that Kriging is a "weighted combination" (linear combination) of observation values that "leads to achieve the best possible estimation" making it the Best Linear Unbiased Predictor (BLUP) in the least squares sense for point spatial interpolation. Kriging has been first defined to interpolate point observations. But 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. Spatial or areal interpolation research is based on a the following assumption: granules that are close to each other in the input space are more likely to have similar features (output values). This is reasonably understandable for temperatures which are continuously defined over the 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 include: 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 presented 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. This technique inherited from mining activities assumes that feature at block (granule) level is the average of 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 estimating the variogram, that is to say of 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 output variables 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-Redón [2022]). While the MAUP refers to correlation between output variables, the ecological inference problem is a result of the correlations at individual level being different from the correlations of the averaged outputs 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: "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 MAUP and 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 output variable is observed at multiple scales, including point level. Indeed systematic averaging makes points and areas different objects with different variability, different correlation structure and therefore different predictors. The classification of problems such as "area-to-point" or "area-to-area" reflects this categorization. Moraga has built a Bayesian framework that can be iterated both with point observations and block observations, based on averaging at areal level for output variables continuously defined over the territory (Moraga et al. [2017]). This model, like other models derived from Kriging, considers blocks to be connected surface areas in $\mathbb{R}^2$ that need to be "discretized" (Goovaerts [2008]) which can distort reality for outputs that are not continuously defined over the space such as populations that are often 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 outputs at areal level do not require systematic averaging. In this regard, Godoy et al. [2022] defined a Gaussian random field on the class $\mathcal{B}_D$ of closed subsets of a certain domain $D \in \mathbb{R}^n$. Distances between elements of $\mathcal{B}_D$ are measured with the Hausdorff distance and the correlation structure between outputs 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 output at areal level and the output at 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, Hausdorff-Gaussian process does not solve the problem of input data 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. MAUP is related to determining a covariance model for points from which are derived the covariance between blocks and the covariance between points and blocks. The standard aggregation approach is a weighted average. These weights 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 reduction in variance, whereas mixing a large number of random variable does not tend to reduce the variance. We will show that this approach effectively manages input 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 means and covariances of output variables. 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 input values 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. Finally, in Section 4, we discuss the pros and cons of the new model.

2 Optimal linear interpolation of mixture distributions

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. Let us consider an input space over which is defined a field of multidimensional random output variables. The output variables, such as sociological variables, are 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 for new inputs, whether they be points or grains, based on the assumption that there is dependence between outputs based 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 $g \subseteq \chi$.*
- *A **granularity** $\mathcal{G} = \{g_1, g_2, ...\}$ of a territory χ is a finite set of grains of χ.*

It is common in some application fields to use a different terminology to talk about grains: blocks, pixels, 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. 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 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 disjoints and those authors are not interested in the family itself (see for instance Kyriakidis [2004]). The reader may find in Appendix C some considerations about those families that arise when relaxing the disjunction constraint.

Granularities are defined in order to work with families of grains. When dealing with geographic data, a granularity is usually the minimum scale at which information is available. For instance, granularities 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 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 where exactly on this tract. Then dwellings' surface area is neither continuous nor clearly geo-localized. Definition 2 below unifies outputs that are continuously defined over a territory and outputs that are not.

Definition 2 (Outputs)**.** *Let $\mathcal{G}$ be a granularity. Outputs are defined over points and grains of $\mathcal{G}$ as follows:*

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

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

- *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 $X_g \in g$:*

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

  For a given granularity $\mathcal{G}$, 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 $\mathbf{Y}$.

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

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

As an example, if k observations of the whole random vector $\mathbf{Y}(g_j)$ are conducted for $j \in \{1, \dots, k\}$, then $n = k \cdot p$ and the vector of observations is:

$$\underline{\mathbf{Y}} = (Y_1(X_{g_1}), \dots, Y_p(X_{g_1}), \dots, Y_1(X_{g_j}), \dots, Y_p(X_{g_j}), \dots, Y_1(X_{g_k}), \dots, Y_p(X_{g_k}))^\top . \quad (1)$$

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 output observations.

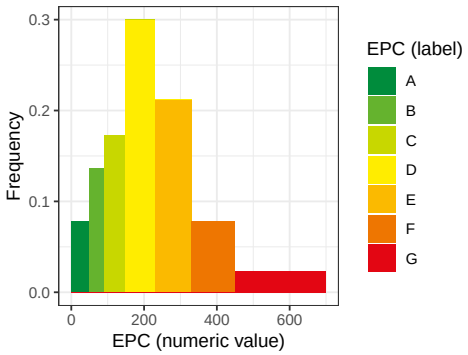

Figure 3: Bar plot of EPC labels frequencies among all EPCs collected in France between 2014 and 2021. Classes are highly heterogeneous.

Example 1 (Buildings energy efficiency)**.** *Keeping in mind that an Energy Performance Certificate (EPC) is given as an energy consumption in $kWh/m^2/year$ (see Figure 3), one can consider a model for which χ is a city viewed as a 2 dimensional space with latitude and longitude as coordinates, $\mathcal{G}$ is the set of plots and a point in χ is associated with a given square meter of a building on the plot. $Y(x)$ is the energy consumption associated with the square meter of building x. Then an EPC in the database is the observed energy efficiency rating associated with one unknown point among those located on the plot pointed by the address. Therefore for a certain plot g, this EPC is an observation of $Y(X_g)$. This model is further developed in subsection 3.3.*

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 equal to $\mathbf{Y}(X_g)$, 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 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 covariances 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 $\mu_i(x) := \mathbb{E}[Y_i(x)]$, we have:

$$\mu_i(g) := \mathbb{E}[Y_i(g)] = \mathbb{E}[\mu_i(X_g)] \tag{2}$$

(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:

$$k_{i,j}(g, g') := \mathrm{Cov}[Y_i(g), Y_j(g')] = \mathbb{E}[k_{i,j}(X_g, X_{g'})] + \mathrm{Cov}[\mu_i(X_g), \mu_j(X_{g'})] \tag{3}$$

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 where $Y_i(g)$ is the result of conditioning $Y_i(x)$ with X_g. And (ii) is derived from the conditional covariance (variance) formula, after conditioning by the joint random vector $(X_g, X_{g'})$ (random variable X_g). □

Note that $\mathrm{Cov}[\mu_i(X_g), \mu_j(X_{g'})] = 0$ in the case where $\mu_i(x)$ is constant over g or g' or in the case where 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 of this grain is the same as the output of the underlying point.

Example 2. *For two distinct and finite grains g and g' of cardinalities $[g], [g']$, assuming that X_g and $X_{g'}$ are independent uniform random variables, we get:*

$$\begin{aligned}
\mu_i(g) &= \frac{1}{[g]} \sum_{x \in g} \mu_i(x) \\
k_{i,j}(g, g') &= \frac{1}{[g]\,[g']} \sum_{(x,x') \in g \times g'} \mathrm{Cov}[Y_i(x), Y_j(x')] \\
k_{i,j}(g, g) &= \frac{1}{[g]} \sum_{x \in g} \mathrm{Cov}[Y_i(x), Y_j(x)]
\end{aligned}$$

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

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

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

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

However

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

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

2.3 Best unbiased linear predictor

In this section, it is proved that there exists a best linear predictor to predict the output associated with a new grain $g \subset \chi$ given a learning set of observations. The problem amounts to predicting any component of the output.

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:

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

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

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

$$M_{\boldsymbol{\alpha}(g)} = \sum_{j=1}^{n} \alpha^j(g) Y^j = \boldsymbol{\alpha}(g)^\top \underline{\mathbf{Y}}. \tag{4}$$

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:

$$\boldsymbol{\alpha}_i(g) \in \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \mathbb{E}\left[\left(Y_i(g) - \boldsymbol{\alpha}^\top \underline{\mathbf{Y}}\right)^2\right] \tag{5}$$

Given the optimal predictor $M_i(g)$, the prediction errors are denoted:

$$\begin{aligned}
\epsilon_i(g) &:= Y_i(g) - M_i(g) && (6) \\
c_{i,j}(g, g') &:= \mathbb{E}\left[\epsilon_i(g)\, \epsilon_j(g')\right] && (7) \\
v_i(g) &:= c_{i,i}(g, g) && (8)
\end{aligned}$$

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*

$$\begin{aligned}
\boldsymbol{\alpha}_i(g) &= \mathbf{K}^{-1} \mathbf{h}_i(g) && (9) \\
c_{i,j}(g, g') &= k_{i,j}(g, g') - \mathbf{h}_i(g)^\top \mathbf{K}^{-1} \mathbf{h}_j(g') && (10)
\end{aligned}$$

(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*

$$\begin{aligned}
\boldsymbol{\alpha}_i(g) &= \mathbf{K}^{-1}\left(\mathbf{h}_i(g) + \lambda_i(g)\underline{\boldsymbol{\mu}}\right) \quad \text{where} \quad \lambda_i(g) = \frac{\mu_i(g) - \underline{\boldsymbol{\mu}}^\top \mathbf{K}^{-1} \mathbf{h}_i(g)}{\underline{\boldsymbol{\mu}}^\top \mathbf{K}^{-1} \underline{\boldsymbol{\mu}}} && (11) \\
c_{i,j}(g, g') &= k_{i,j}(g, g') - \mathbf{h}_i(g)^\top \mathbf{K}^{-1} \mathbf{h}_j(g') + \lambda_i(g)\lambda_j(g) \underline{\boldsymbol{\mu}}^\top \mathbf{K}^{-1} \underline{\boldsymbol{\mu}} && (12)
\end{aligned}$$

Proof of Proposition 2 is given in Appendix A.

The above Proposition 2 is valid to predict a single component $Y_i(g)$ of the output $\mathbf{Y}(g)$, but it can be extended to the prediction of $\mathbf{Y}(g)$: the best linear unbiased predictor of $\mathbf{Y}(g) = (Y_1(g) \ldots Y_p(g))^\top$ 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, three important particular cases are explored. The first one considers the Ordinary Mixture Kriging situation where the output 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. The last particular case shows that Kriging is a special case of Mixture Kriging.

Particular case 1 ($\underline{\boldsymbol{\mu}} = \mu_0(1,\dots,1)^\top$)**.** *Regarding ordinary mixture Kriging, assuming that all random variables $Y_i(g)$ have the same unknown expectation μ_0, setting $\mathbf{1}_n = (1,\dots,1)^\top$, Equation 9 simplifies into:*

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

and setting

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

$M_i(g)$ 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 outputs, observations and covariances:*

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

Then 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 observations covariance matrix writes:

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

And covariance vector between observations and a new grain writes:

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

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 $\mathbf{K}_\epsilon$ is a diagonal matrix and $\mathbf{h}_{\epsilon,i}(g)$ is null as long as g is not among the observed grains.

Particular case 3 (Gaussian Singleton)**.** *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.*

It is also to be noticed that 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 Appendix B.

3 Illustration

3.1 Unidimensional case: rounded inputs

A common issue when feeding statistical models with real data is the precision of input data and its impact on a model's performance. Usual applications of Kriging take this uncertainty into account increasing output variances by a value that is known as 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 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 input values that are rounded to the nearest unit i.e. for 2 input values of $x_1, x_2 \in]0.5, 1.5]$, the observer sees the same value $\tilde{x}_1 = \tilde{x}_2 = 1$. For a Kriging model, these are multiple observations of a 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. Therefore, it 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 `R` language, we simulate a 1-dimensional random field realization with Gaussian covariance kernel which parameters are detailed in Table 1 . x is discretized between 0 and 10 with step 0.05. We pick 8 points for observations as listed in Table 2 . These observations are plotted on Figure 4 . Observations $\{o1, o2, o6\}$ form the learning set, observations $\{o4, o5, o7\}$ form the test set and observations $\{o3, o8\}$ form the validation set.

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 1: Parameters and performances of fitted models in the case of observations with rounded input. Note that 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 $[0, 10]$.

		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 2: Observations of the simulated Gaussian random field.

Kriging model (Figure 4 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 $(10^{-i})_{i \in \{1,\dots,10\}}$. 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 same kernel and the previously selected nugget effect. This model is applied to compute a validation MSE and a total MSE computed on all points in $[0, 10]$. The variance of the prediction error is also predicted using formula given in Proposition 2 .

Regarding Mixture Kriging (Figure 4 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. 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 learning set and test set are used to fit a model with the same kernel parameters as for simulation and without nugget effect. Validation MSE and total MSE are computed for comparison with Kriging.

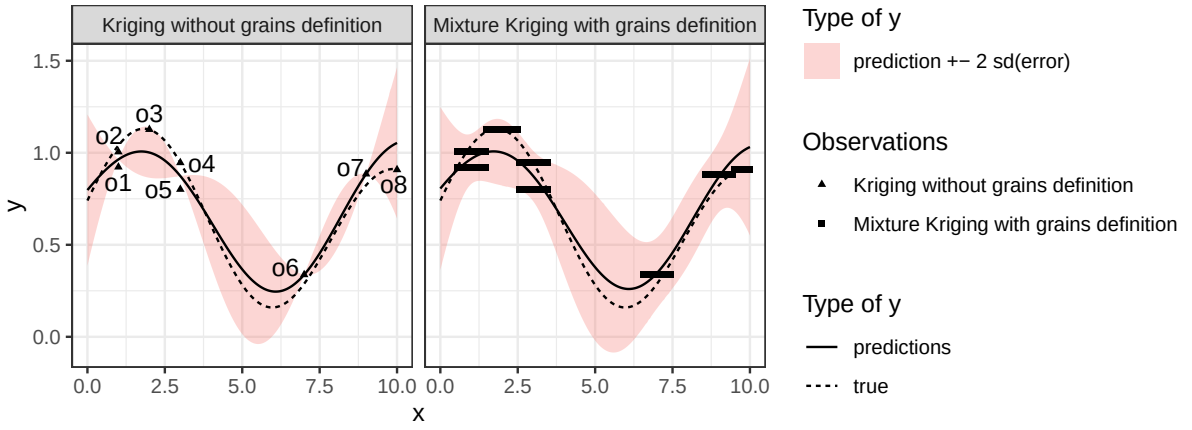

Figure 4: *Rounded inputs.* **Left and right** Dashed line labelled "true" shows a simulated uniform random field. Solid line labelled "predictions" shows fitted model mean prediction (see Table 1). Ribbon shows an interval of radius twice the root square of the estimated error variance. **Left**: *Kriging model.* Triangular dots show observations. **Right** *Mixture Kriging.* Horizontal line segments show observations. (See Table 2 for more details about observations).

In this case the mean prediction is almost the same for both models. But the predicted error variance (visible on the ribbons in Figure 4) differs. By construction, Kriging is supposed to interpolate observations exactly resulting in a very small error variance near observations. However, Mixture Kriging takes into account the input uncertainty and predicts a significantly positive error variance even near observations. If one increases the nugget effect on the Kriging model, the predicted error variance increases but there remains a "bottle neck" effect near the observations and predictions are shrunk towards 0.

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 x, it belongs to a set of permissible values χ 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 on the manufactured object induces some uncertainty on the performance. 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 $x_1, \dots, x_n$, the industry observes performances $\mathbf{Y}(g_1), \dots, \mathbf{Y}(g_n)$. 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 that 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 $\{\mathbf{y}(x_i) = \sin(x_i^2) \,:\, i \in \{1, \dots, n\}\}$. 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 : Manufactured object is produced exactly according to the design, precision interval is restricted to a point.
- P_2 : The precision is the same for all designs, the associated interval is of fixed measure.

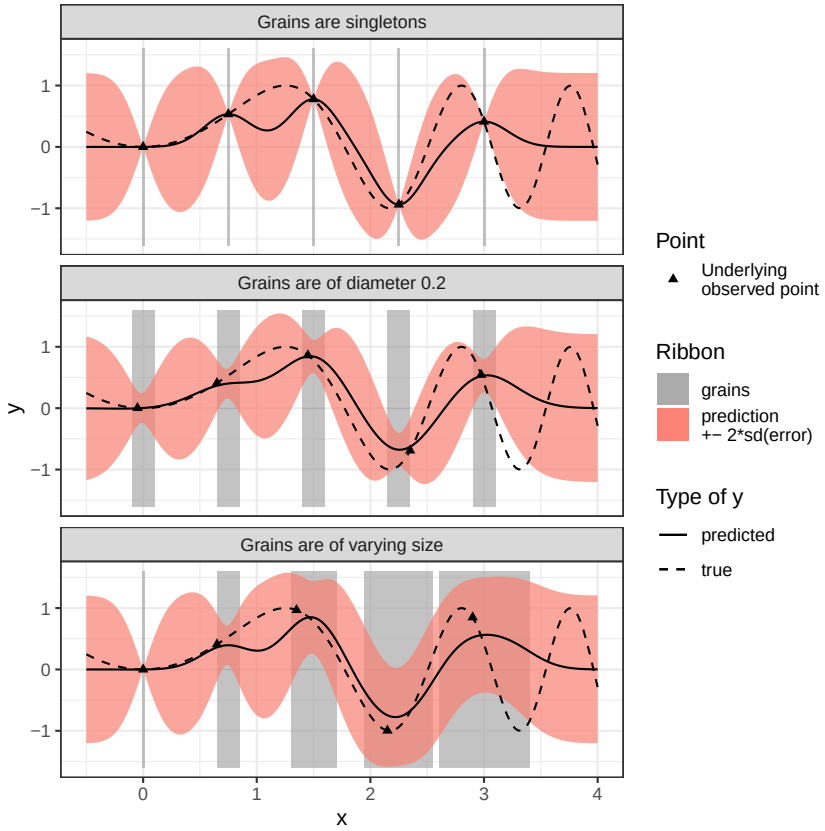

Figure 5: *Mixture Kriging and grain sizes.*
All: Dashed line represents $\mathbf{y}(x)$. Solid line is the mean prediction. Ribbon shows an interval centred on the mean prediction, of radius twice the square root of the predicted error variance. Vertical columns show the grains as x intervals. Black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

- P_3: The larger is x, the larger is the uncertainty on 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 with the overall variance of $\mathbf{y}$ on $\chi = [0, 4]$ as variance parameter. Range parameter is optimized minimizing mean squared error between $\mathbf{y}$ and point prediction on χ. 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 same size (Figure 5 middle), predicted values are not interpolating any more, predicted error is not null on the grains but is also smaller than above far from the grains. In the bottom figure, the greater is x, the greater the uncertainty on the manufactured object as compared to design. Predicted error (ribbon) is increasing with the grains diameter.

	Model properties			
Granularity	Variance	Nugget	Range	Exact interpolation
Grains are singletons	0.36	0	0.3	Yes
Grains are of equal measure	0.36	0	0.4	No
Grains are of increasing measure	0.36	0	0.3	No

Table 3: Compare models quality for different types of granularities.

Overall, it is important to note that mixture Kriging model accounts for the randomness of input values without any nugget effect, therefore preserving variability of predicted values.

3.3 Energy Performance Certificate (EPC) prediction

Let us now address the EPC prediction problem in a more detailed manner. EPC is given as a numeric energy consumption per square meter and per year which is associated with a letter ranging from A to G. A and B label the most energy saving dwellings (less that $90kWh/m^2/year$). 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 an 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. And we want to predict an EPC at the whole land plot level, that is to say for the set of buildings it contains.

As can be seen in Figure 3, observations are strongly unbalanced, meaning that labels A, B, F, G are rarely observed while labels C, D, E are very common. As a result, labels A, B, F, 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 T_L the number of observations with label L and TP_L the number of good predictions with label L, the balanced accuracy is given by the formula:

$$BA = \frac{1}{7} \sum_{L \in \{A,\dots,G\}} \frac{TP_L}{T_L}$$

Let us first consider the following variable normalization: given a real random variable X, F_X its cdf, supposed to be invertible and $F_{\mathcal{N}}$ the standard Gaussian distribution cdf, we denote $H(X)$ the random variable given by $F_{\mathcal{N}}^{-1}(F_X(X))$. It follows a standard Gaussian distribution. Moreover H has an inverse since $X = F_X^{-1}(F_{\mathcal{N}}(H(X)))$.

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 construction year, latitude and 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. The set of all grains form the granularity $\mathcal{G}$.
- 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 location of observations. On g, the output variable is defined as: $Y(g) = Y(X_g)$.
- A vector of observations of n distinct grains is given and denoted $\underline{\mathbf{Y}}$.

Moreover, by construction, Y is centred.

The granularity $\mathcal{G}$ is mapped in Figure 6. Note that the grains seem to be disjoint but they are not due to overlaps on the 3rd dimension. The set of observations is represented in Figure 7.

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:

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

σ^2 is called the variance coefficient and $\Theta = (\theta_1, \theta_2, \theta_3)$ 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.

Mixture Kriging predictor described in subsection 2.3 is used to predict energy consumption at plot level. It can be proved that without nugget effect the mean prediction, in the case of a one 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

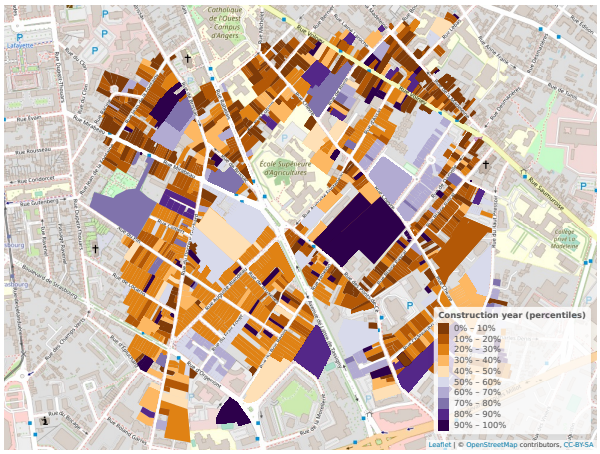

Figure 6: An urban area in Angers: latitude is the vertical dimension, longitude is the horizontal dimension, construction year is given by the colour. The side of the square is 1km. Construction years range from 1340 (first percentile) to 2019 (last percentile).

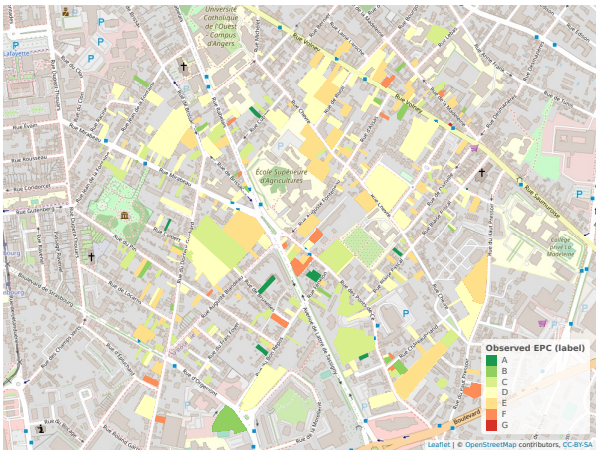

Figure 7: Map of the 365 observations. Each colour represents a label associated with a numeric value. See also Figure 1.

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 R language to implement Mixture Kriging.

Let us now consider a Kriging model $M2$ to compare performances with Mixture Kriging model $M1$. $M2$ has same properties as M_1 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 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 eldest part of a building is usually also the largest one.
- A nugget effect has to be introduced so as to have a smooth predictor:

$$\mathbb{V}[Y(x)] = \sigma^2 + \epsilon^2 \text{ where } \epsilon^2 \in [0, 1] .$$

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

Both models $M1$ and $M2$ are optimized with a 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 to default.

With regards to the optimal parameters in Table 4, 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 5, $M1$ reaches a larger BA than $M2$ by 37%. However $M1$ has lower performances on other indicators with a difference of approximately 10%. The variance of all 365 predictions with $M1$ is 150% larger than with $M2$. These figures are better understood by examining the confusion matrices in Tables 6 and 7. Indeed, the percentage of large errors (represented by the red area) is 3% with model $M1$ and 0.5% with $M2$. We know that large errors have an important impact on 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-intensive dwellings.

These results suggest that Mixture Kriging ($M1$) predictions have an improved variability compared to Kriging ($M2$). Despite having fewer parameters, Mixture Kriging significantly improves the BA although it leads to more frequent large errors. Kriging accounts for uncertainty in the input data eliminating the need to add uncertainty to the output. In this example it avoids grouping predictions near the mean value (shrinkage) and yields a better BA as compared with Kriging which requires the introduction of a nugget effect.

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

*: These parameters are treated as constant parameters.

Table 4: Optimal parameters for $M1$ and $M2$.

	EPC int.			EPC num.		
Model	BA	MAE	RMSE	MAE	RMSE	Prediction range
Mixture Kriging ($M1$)	**0.26**	0.93	1.37	78.93	106.16	**6,66**
Kriging ($M2$)	0.19	**0.85**	**1.22**	**72.22**	**92.98**	2,59

EPC int.: Energy Performance Certificate treated as an integer: 1 for label A, ..., 7 for G.

EPC num.: energy consumption expressed in $kWh/m^2/year$.

BA: Balanced Accuracy; MAE: Mean Absolute Error; RMSE: Root Mean Squared Error.

Prediction range: Variance of the vector of predictions.

Table 5: Optimal performances achieved by $M1$ and $M2$ with 3 input variables and no output covariate.

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

Table 6: Confusion matrix of $M1$ predictions

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

Table 7: Confusion matrix of $M2$ predictions

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 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 Modifiable Areal Unit Problem (MAUP), ecological inference and variance reduction problems can become challenging to solve. Over the past few decades researchers have been developing methods to assess and/or correct MAUP effect (Briz-Redón [2022]). Modifying territory partitioning (Li et al. [2009]) is also an effective solution for addressing variance reduction problem, but 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, hereby simulating the addition of a white noise to the outputs. This transformation smooths predicted values but also shrinks them: the range between minimal and maximal predicted values is reduced, thus degrading the prediction of values that are particularly large or particularly small.

The availability of new datasets with uncertainty on the inputs and where averaging is not a meaningful interpretation has driven us to seek a novel method of linear interpolation. We have introduced a new element in the model that is a random position (input value) associated with the output at areal level. 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 points are acceptable. The terms "grains" and "granularity" have 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 prediction shrinkage due to this factor. If the output variance is the same everywhere at point level, then it is also the same as the output 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 point level, then it will also be the same as the covariance at grain level regardless of the grain's shape. This implies that this model has no measurable MAUP effect.

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 a 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. 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. In practical applications, Mixture Kriging is therefore designed to handle data with uncertainty on the input values without introducing nugget effect.

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. In particular, datasets that inform about granules that are uncertainly defined such as dwellings, buildings, streets, human persons, households. It can also be used for datasets informing about granules which should have deterministic shapes or position in the input space but come with a 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 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 test the benefits of using covariates. We also study the possibility to develop 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`).

References

Walter Jeremy Koch Aldworth. *Spatial prediction, spatial sampling, and measurement error*. Doctor of Philosophy, Iowa State University, Digital Repository, Ames, 1998. URL `https://lib.dr.iastate.edu/rtd/11842/`. Pages: 6510332.

Usman Ali, Mohammad Haris Shamsi, Mark Bohacek, Cathal Hoare, Karl Purcell, Eleni Mangina, and James O'Donnell. A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings. *Applied Energy*, 267:114861, June 2020. ISSN 0306-2619. doi: 10.1016/j.apenergy.2020.114861. URL `https://www.sciencedirect.com/science/article/pii/S0306261920303731`.

Ilaria Ballarini, Vincenzo Corrado, Francesco Madonna, Simona Paduos, and Franco Ravasio. Energy refurbishment of the Italian residential building stock: energy and cost analysis through the application

of the building typology. *Energy Policy*, 105:148–160, June 2017. ISSN 0301-4215. doi: 10.1016/j.enpol.2017.02.026. URL http://www.sciencedirect.com/science/article/pii/S0301421517301015.

Álvaro Briz-Redón. A Bayesian shared-effects modeling framework to quantify the modifiable areal unit problem. *Spatial Statistics*, 51:100689, October 2022. ISSN 2211-6753. doi: 10.1016/j.spasta.2022.100689. URL https://www.sciencedirect.com/science/article/pii/S2211675322000537.

Alexis Comber and Wen Zeng. Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations. *Geography Compass*, 13(10):e12465, 2019. ISSN 1749-8198. doi: 10.1111/gec3.12465. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/gec3.12465. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gec3.12465.

Noel A.C. Cressie. *Statistics for spatial data revised edition*. Wiley series in probability and mathematical statistics. Applied probability and statistics., john wiley & sons inc. edition, 1993. ISBN 0-471-00255-0. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781119115151.

Lucas da Cunha Godoy, Marcos Oliveira Prates, and Jun Yan. An unified framework for point-level, areal, and mixed spatial data: the Hausdorff-Gaussian Process, August 2022. URL http://arxiv.org/abs/2208.07900. arXiv:2208.07900 [stat].

Pierre Goovaerts. Kriging and Semivariogram Deconvolution in the Presence of Irregular Geographical Units. *Mathematical Geology*, 40(1):101–128, 2008. ISSN 0882-8121.

Carol Gotway and Linda Young. Combining Incompatible Spatial Data. *Journal of the American Statistical Association*, 97:632–648, February 2002. doi: 10.1198/016214502760047140.

Martijn Gösgens, Anton Zhiyanov, Aleksey Tikhonov, and Liudmila Prokhorenkova. Good Classification Measures and How to Find Them. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17136–17147. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/8e489b4966fe8f703b5be647f1cbae63-Paper.pdf.

Edward H. Isaaks and R. Mohan Srivastava. *An Introduction to Applied Geostatistics*. Oxford University Press, 1989. ISBN 978-1-61583-082-4. URL https://www.sciencedirect.com/science/article/pii/009830049190055I.

Yan Jin, Yong Ge, Jianghao Wang, Gerard Heuvelink, and Le Wang. Geographically Weighted Area-to-Point Regression Kriging for Spatial Downscaling in Remote Sensing. *Remote Sensing*, 10:579, April 2018. doi: 10.3390/rs10040579.

Ruth Kerry, Pierre Goovaerts, Izak P.J. Smit, and Ben R. Ingram. 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, 2013. ISSN 1365-8816. doi: 10.1080/13658816.2012.663917. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341904/.

Phaedon Kyriakidis. A Geostatistical Framework For Area-To-Point Spatial Interpolation. *Geographical Analysis*, 36, August 2004. doi: 10.1353/geo.2004.0009.

Nina Siu-Ngan Lam. Spatial Interpolation Methods: A Review. *The American Cartographer*, 10(2):129–150, January 1983. ISSN 0094-1689. doi: 10.1559/152304083783914958. URL https://doi.org/10.1559/152304083783914958. Publisher: Taylor & Francis _eprint: https://doi.org/10.1559/152304083783914958.

Changjiang Li, Zhiming Lu, Tuhua Ma, and Xingsheng Zhu. A simple kriging method incorporating multiscale measurements in geochemical survey. *Journal of Geochemical Exploration*, 101(2):147–154, May 2009. ISSN 0375-6742. doi: 10.1016/j.gexplo.2008.06.003. URL http://www.sciencedirect.com/science/article/pii/S0375674208000666.

Yu-Pin Lin, Bai-You Cheng, Guey-Shin Shyu, and Tsun-Kuo Chang. 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, January 2010. ISSN 0269-7491. doi: 10.1016/j.envpol.2009.07.015. URL https://www.sciencedirect.com/science/article/pii/S0269749109003534.

Georges Matheron. Principles of geostatistics. *Economic Geology*, 58(8):1246–1266, December 1963. ISSN 0361-0128. doi: 10.2113/gsecongeo.58.8.1246. URL https://doi.org/10.2113/gsecongeo.58.8.1246.

Paula Moraga, Susanna M. Cramb, Kerrie L. Mengersen, and Marcello Pagano. A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. *Spatial Statistics*, 21:27–41, August 2017. ISSN 2211-6753. doi: 10.1016/j.spasta.2017.04.006. URL https://www.sciencedirect.com/science/article/pii/S2211675317301318.

Witold Pedrycz. *Granular computing : analysis and design of intelligent systems*. Industrial electronics series. Taylor & Francis, 2013. ISBN 978-1-4398-8681-6. URL https://doi.org/10.1201/9781315216737.

Osvaldo José Ribeiro Pereira, Adolpho José Melfi, Célia Regina Montes, and Yves Lucas. Downscaling of ASTER Thermal Images Based on Geographically Weighted Regression Kriging. *Remote Sensing*, 10(4):633, April 2018. doi: 10.3390/rs10040633. URL https://www.mdpi.com/2072-4292/10/4/633. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Laura Poggio and Alessandro Gimona. Downscaling and correction of regional climate models outputs with a hybrid geostatistical approach. *Spatial Statistics*, 14:4–21, November 2015. ISSN 2211-6753. doi: 10.1016/j.spasta.2015.04.006. URL https://www.sciencedirect.com/science/article/pii/S2211675315000305.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass, 2006. ISBN 978-0-262-18253-9. OCLC: ocm61285753.

Marc Rocas, Alberto García-González, Sergio Zlotnik, Xabier Larráyoz, and Pedro Díez. Nonintrusive uncertainty quantification for automotive crash problems with VPS/Pamcrash. *Finite Elements in Analysis and Design*, 193:103556, October 2021. ISSN 0168-874X. doi: 10.1016/j.finel.2021.103556. URL https://www.sciencedirect.com/science/article/pii/S0168874X21000408.

Pascal Schetelat, Lucie Lefort, and Nicolas Delgado. Urban data imputation using multi-output multi-class classification. *Building to Buildings: Urban and Community Energy Modelling*, November 2020.

Brian J. Smith, Jun Yan, and Mary Kathryn Cowles. Unified Geostatistical Modeling for Data Fusion and Spatial Heteroskedasticity with R Package ramps. *Journal of Statistical Software*, 25:1–21, April 2008. ISSN 1548-7660. doi: 10.18637/jss.v025.i10. URL https://doi.org/10.18637/jss.v025.i10.

Phuong Truong and Gerard Heuvelink. Bayesian Area-to-Point Kriging using Expert Knowledge as Informative Priors. *International Journal of Applied Earth Observation and Geoinformation*, 30:2291, April 2013. doi: 10.1016/j.jag.2014.01.019.

Qunming Wang, Wenzhong Shi, and Peter M. Atkinson. Area-to-point regression kriging for pan-sharpening. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:151–165, April 2016. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2016.02.006. URL http://adsabs.harvard.edu/abs/2016JPRS..114..151W.

Christopher Williams and Carl Rasmussen. Gaussian Processes for Regression. *Advances in Neural Information Processing Systems 8*, 8, March 1996.

Simon N. Wood. *Generalized Additive Models: An Introduction with R, Second Edition*. CRC Press, May 2017. ISBN 978-1-4987-2834-8.

E.-H. Yoo and P. C. Kyriakidis. Area-to-point Kriging with inequality-type data. *Journal of Geographical Systems*, 8(4):357–390, November 2006. ISSN 1435-5930, 1435-5949. doi: 10.1007/s10109-006-0036-7. URL http://link.springer.com/10.1007/s10109-006-0036-7.

Lotfi A. Zadeh. Toward a generalized theory of uncertainty (GTU)—an outline. *Information Sciences*, 172(1):1–40, June 2005. ISSN 0020-0255. doi: 10.1016/j.ins.2005.01.017. URL https://www.sciencedirect.com/science/article/pii/S002002550500054X.

Xiaohu Zhang, Wenjun Zuo, Shengli Zhao, Li Jiang, Linhai Chen, and Yan Zhu. 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):33, January 2018. doi: 10.3390/ijgi7010033. URL `https://www.mdpi.com/2220-9964/7/1/33`. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

A Proof of Proposition 2

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:

$$\begin{aligned} F_i(g) &:= \left\{\boldsymbol{\alpha}^\top \underline{\mathbf{Y}} : \mu_i(g) = \boldsymbol{\alpha}^\top \underline{\boldsymbol{\mu}}\right\} \\ G_i(g) &:= \{\alpha Y_i(g) : \alpha \in \mathbb{R}\} \end{aligned}$$

And similarly, we denote:

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

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 8). 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 8.

Proof. For given $i \in \{1, \ldots, p\}$ and $g \subseteq \chi$, let $M_{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^\top \underline{\mathbf{Y}}$ be a linear predictor of $Y_i(g)$, where $\boldsymbol{\alpha} = (\alpha^1, \ldots, \alpha^n)$ 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 $\underline{\boldsymbol{\mu}} = (0, \ldots, 0)^\top$ 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 $\boldsymbol{\alpha}$,

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

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.

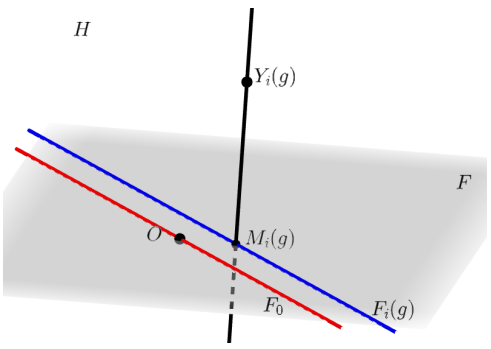

Figure 8: *Geometrical interpretation of the prediction process.*

(ii) If $\underline{\boldsymbol{\mu}} \neq (0, \ldots, 0)^\top$ 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:

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

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

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

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

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

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

Let us consider now the cross-errors:

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

Due to unbiasedness condition, it means that:

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

Which rewrites:

$$c_{i,j}(g, g') = k_{i,j}(g, g') - \boldsymbol{\alpha}_j(g')^\top \mathbf{h}_i(g) - \boldsymbol{\alpha}_i(g)^\top \mathbf{h}_j(g') + \boldsymbol{\alpha}_i(g)^\top \mathbf{K} \boldsymbol{\alpha}_j(g')\ . \tag{13}$$

Note that equation (13) 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:

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

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

B Cross-errors and conditional covariances

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

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

This is for example the case when $\{\mathbf{Y}(x) : x \in \chi\}$ *is a vector-valued Gaussian random field and when each* X_g *is Dirac distributed (see Remark 3). In this setting, under assumption* (A)*, one can show that cross errors for both Simple Mixture Kriging and Ordinary Mixture Kriging are*

$$c_{i,j}(g,g') = \mathbb{E}\left[\mathrm{Cov}\left[Y_i(g), Y_j(g')|\underline{\mathbf{Y}}\right]\right]. \tag{14}$$

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 simplifies:* $\mathbb{E}\left[\mathrm{Cov}\left[Y_i(g), Y_j(g')|\underline{\mathbf{Y}}\right]\right] = \mathrm{Cov}\left[Y_i(g), Y_j(g')|\underline{\mathbf{Y}}\right]$.

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 $\mathbb{E}\left[\epsilon_i(g)\right] = 0$ and $\mathbb{E}\left[\delta\right] = 0$:

  $$\forall \delta \in F_0,\ \forall i \in \{1,\dots,p\},\ \forall g \in \chi,\quad \mathbb{E}\left[\epsilon_i(g)\,\delta\right] = \mathrm{Cov}\left[\epsilon_i(g), \delta\right] = 0. \tag{15}$$

  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 (15), we have:

  $$\mathrm{Cov}\left[\epsilon_i(g) + \delta, \epsilon_j(g') + \delta'\right] = c_{i,j}(g,g') + 0 + 0 + \mathrm{Cov}\left[\delta,\, \delta'\right] \tag{16}$$
- On the other hand, using the conditional covariance formula, we have:

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

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

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

  Furthermore, we have assumed in Assumption (A) that $M_i(g) = \mathbb{E}\left[Y_i(g)|\underline{\mathbf{Y}}\right]$ and $M_j(g') = \mathbb{E}\left[Y_j(g')|\underline{\mathbf{Y}}\right]$, therefore $\mathbb{E}\left[\epsilon_i(g)|\underline{\mathbf{Y}}\right] = \mathbb{E}\left[\epsilon_j(g')|\underline{\mathbf{Y}}\right] = 0$ and:

  $$\mathrm{Cov}\left[\epsilon_i(g) + \delta, \epsilon_j(g') + \delta'\right] = \mathbb{E}\left[\mathrm{Cov}\left[Y_i(g), Y_j(g') \mid \underline{\mathbf{Y}}\right]\right] + \mathrm{Cov}\left[\delta, \delta'\right] \tag{17}$$

  Identifying the equations (16) and (17), we get the expected result.

□

C Operations on granularities, overlapping granularities

In the course of our research, we started studying some granularities available in our databases and their relations/classifications: e.g. what is the relation between the set of land plots and the set of census tracts? We also thought about ways to build non-overlapping granularities from existing granularities. This lead us to the definitions of the following concepts.

Definition 3 (Non-overlapping granularity)**.** *A granularity $\mathcal{G}$ is said to be **non-overlapping** when all intersections of grains are empty: $\forall g, g' \in \mathcal{G},\ g \cap g' = \emptyset$.*

Definition 4 (Granularity order)**.** *The **granularity order** $\mathcal{G} \leq \mathcal{H}$, or equivalently $\mathcal{H} \geq \mathcal{G}$, holds for two granularities $\mathcal{G}$ and $\mathcal{H}$ under the following condition:*

$$\mathcal{G} \leq \mathcal{H} \quad \Leftrightarrow \quad \forall g \in \mathcal{G}, \begin{cases} g \in \bigcup\limits_{h \in \mathcal{H}} h \\ \text{and} \quad \forall h \in \mathcal{H},\ g \cap h \in \{\emptyset, g\} \end{cases}$$

$\mathcal{G}$ is said to be thinner than $\mathcal{H}$, or equivalently $\mathcal{H}$ coarser than $\mathcal{G}$. In particular, $\mathcal{G} \leq \mathcal{H}$ implies that any grain of $\mathcal{G}$ is a subset of at least one grain in $\mathcal{H}$, but it also implies that a grain of $\mathcal{G}$ does not partly overlap a grain of $\mathcal{H}$.

Relation $\leq$ is transitive on the set of granularities defined on χ. It defines of partial order on this set.

Proposition 4 (Non-overlapping granularities)**.** *Define an **insertion operator** $\oplus$, for any non-overlapping granularity $\mathcal{G}$ and any grain h by:*

$$\mathcal{G} \oplus \{h\} := \left\{ g_0 \,:\, g_0 \neq \emptyset \text{ and } \ g_0 \in \{g \cap h \,:\, g \in \mathcal{G}\} \cup \{g \setminus h \,:\, g \in \mathcal{G}\} \cup \left\{ h \setminus \bigcup_{g \in \mathcal{G}} g \right\} \right\}.$$

This operator $\oplus$ adds a partition of the grain h to the non-overlapping granularity $\mathcal{G}$, while ensuring that $\mathcal{G} \oplus \{h\}$ is non-overlapping and has the same union of grains as $h \cup \bigcup\limits_{g \in \mathcal{G}} g$.

Then we have:

(i) For any non-overlapping granularity $\mathcal{G}$ and grain h, the resulting granularity is thinner than $\mathcal{G} \cup \{h\}$:

$$\mathcal{G} \oplus \{h\} \leq \mathcal{G} \cup \{h\}\,.$$

(ii) For any non-overlapping granularity $\mathcal{G}$ and grains h, h', the insertion order does not matter:

$$(\mathcal{G} \oplus \{h\}) \oplus \{h'\} = (\mathcal{G} \oplus \{h'\}) \oplus \{h\}\,.$$

*(iii) Among the granularities that are thinner than a finite granularity $\mathcal{G} = \{g_1, \ldots, g_n\}$, there is a unique **maximal non-overlapping granularity** $\mathcal{G}^{\oplus}$ and we can construct it iteratively with the insertion operator.*

$$\mathcal{G}^{\oplus} := \{g_1\} \oplus \ldots \oplus \{g_n\}\,. \tag{18}$$

This granularity is a non-overlapping granularity such that $\mathcal{G}^{\oplus} \leq \mathcal{G}$, and it is maximal, in the sense that any other non-overlapping $\mathcal{G}'$ that is thinner than $\mathcal{G}$ is also thinner than $\mathcal{G}^{\oplus}$: $\mathcal{G}' \leq \mathcal{G} \Rightarrow \mathcal{G}' \leq \mathcal{G}^{\oplus}$.

Proof. • Let us prove the item (i)
Let us prove that $\mathcal{G} \oplus \{h\} \leq \mathcal{G} \cup \{h\}$. Let $g_+ \in \mathcal{G} \oplus \{h\}$ and $g' \in \mathcal{G} \cup \{h\}$. It is clear by construction that $g_+ \in \bigcup\limits_{g \in \mathcal{G} \cup \{h\}} g$. Moreover:

$$g_+ = g \cap h \text{ or } \ g_+ = g \setminus h \text{ or } \ g_+ = h \setminus \bigcup_{g \in \mathcal{G}} \quad \text{AND} \quad g' \in \mathcal{G} \text{ or } \ g' = h$$

One can prove that in all 6 different combined cases, either $g_+ \cap g' = g_+$ or $g_+ \cap g' = \emptyset$.
As a consequence, $\mathcal{G} \oplus \{h\} \leq \mathcal{G} \cup \{h\}$.

- Let us prove the item (ii).
  Let $g_2 \in (\mathcal{G} \oplus \{h\}) \oplus \{h'\}$ then:

  $$\text{(A)}\ \ \exists g_1 \in \mathcal{G} \oplus \{h\},\ g_2 = g_1 \cap h' \quad \text{or} \quad \text{(B)}\ \ \exists g_1 \in \mathcal{G} \oplus \{h\},\ g_2 = g_1 \setminus h' \quad \text{or} \quad \text{(C)}\ \ g_2 = h' \setminus \bigcap_{g \in \mathcal{G} \oplus \{h\}} g$$

  Let $g_1 \in \mathcal{G} \oplus \{h\}$ then:

  $$\text{(a)}\ \ \exists g_0 \in \mathcal{G}, g_1 = g_0 \cap h \quad \text{or} \quad \text{(b)}\ \ \exists g_0 \in \mathcal{G}, g_1 = g_0 \setminus h \quad \text{or} \quad \text{(c)}\ \ g_1 = h \setminus \bigcup_{g \in \mathcal{G}} g$$

  $$\begin{array}{lll}
  \text{(Aa)}\ \ g_2 = g_0 \cap h \cap h' & = g_0 \cap h' \cap h & \in (\mathcal{G} \oplus \{h'\}) \oplus \{h\},\ \text{see case (Aa)} \\
  \text{(Ab)}\ \ g_2 = (g_0 \setminus h) \cap h' & = (g_0 \cap h') \setminus h & \in (\mathcal{G} \oplus \{h'\}) \oplus \{h\},\ \text{see case (Ba)} \\
  \text{(Ac)}\ \ g_2 = (h \setminus \bigcup_{g \in \mathcal{G}} g) \cap h' & = (h' \setminus \bigcup_{g \in \mathcal{G}} g) \cap h & \in (\mathcal{G} \oplus \{h'\}) \oplus \{h\},\ \text{see case (Ac)} \\
  \text{(Ba)}\ \ g_2 = (g_0 \cap h) \setminus h' & = (g_0 \setminus h') \cap h & \in (\mathcal{G} \oplus \{h'\}) \oplus \{h\},\ \text{see case (Ab)} \\
  \text{(Bb)}\ \ g_2 = (g_0 \setminus h) \setminus h' & = (g_0 \setminus h') \setminus h & \in (\mathcal{G} \oplus \{h'\}) \oplus \{h\},\ \text{see case (Bb)} \\
  \text{(Bc)}\ \ g_2 = (h \setminus \bigcup_{g \in \mathcal{G}} g) \setminus h' & = h \setminus \bigcup_{g \in \mathcal{G} \oplus \{h'\}} g & \in (\mathcal{G} \oplus \{h'\}) \oplus \{h\},\ \text{see case (C)} \\
  \text{(C)}\ \ g_2 = h' \setminus \bigcup_{g \in \mathcal{G} \oplus \{h\}} g & = (h' \setminus \bigcup_{g \in \mathcal{G}} g) \setminus h & \in (\mathcal{G} \oplus \{h'\}) \oplus \{h\},\ \text{see case (Bc)}
  \end{array}$$

  For cases (Bc) and (C), we used the fact that $\bigcup_{g \in \mathcal{G} \oplus \{h\}} g = h \cup \bigcup_{g \in \mathcal{G}} g$.

- Let us prove the item (iii)
  Note that due to item (ii), $\mathcal{G}^{\oplus}$ does not depend on the indexing order of the grains composing $\mathcal{G}$.
  Moreover, due to item (i), $\{g_1\} \oplus \{g_2\} \leq \{g_1, g_2\}$ and by recurrence, $\mathcal{G}^{\oplus} \leq \mathcal{G}$.
  Now let us prove that for any non-overlapping granularity $\mathcal{H}$, any granularity $\mathcal{G}$, any grain g_0:

  $$\mathcal{G} \leq \mathcal{H} \cup \{g_0\} \ \Rightarrow\ \mathcal{G} \leq \mathcal{H} \oplus \{g_0\}$$

  Suppose $\mathcal{G} \leq \mathcal{H} \cup \{g_0\}$. Let $g \in \mathcal{G}$ and $g_+ \in \mathcal{H} \oplus \{g_0\}$, taking into account that $\mathcal{H}$ is non-overlapping:

  $$\text{(A)}\ \ \exists h \in \mathcal{H} : g \subset h \cap g_0 \quad \text{or} \quad \text{(B)}\ \ \exists h \in \mathcal{H} : g \subset h \setminus g_0 \quad \text{or} \quad \text{(C)}\ \ g \subset g_0 \setminus \bigcup_{h \in \mathcal{H}} h$$
  $$\text{and}\ \ \text{(a)}\ \ \exists h' \in \mathcal{H} : g_+ = h' \cap g_0 \quad \text{or} \quad \text{(b)}\ \ \exists h' \in \mathcal{H} : g_+ = h' \setminus g_0 \quad \text{or} \quad \text{(c)}\ \ g_+ = g_0 \setminus \bigcup_{h \in \mathcal{H}} h$$

  In cases Ab, Ac, Ba, Bc, Ca, Cb, we have $g \cap g_+ = \emptyset$. In cases Aa and Bb, if $h = h'$ then $g \cap g_+ = g$, otherwise $g \cap g_+ = \emptyset$. In case Cc, $g \cap g_+ = g$. Therefore in either case, $g \cap g_+ \in \{g, \emptyset\}$ and $\mathcal{G} \leq \mathcal{H} \oplus \{g_0\}$.

□

When a non-overlapping granularity is needed, one can thus use Proposition 4 and build $\mathcal{G}^{\oplus}$ directly from any finite granularity $\mathcal{G}$, possibly overlapping. However, we will see in the rest of the paper that the proposed model is also suited for overlapping granularities.

When two data sources are available, relying on two granularities $\mathcal{G}$ and $\mathcal{H}$ it can also be convenient to define $\mathcal{G} \oplus \mathcal{H} := (\mathcal{G} \cup \mathcal{H})^{\oplus}$ to get a non-overlapping resulting granularity allowing to work with both data sources. As an example, if an information is given at the level of a grid reference system $\mathcal{G}$, and also at a level of urban areas $\mathcal{H}$, it may be convenient to build all intersection areas by this way. The Proposition 4 gives a simple way to do so, even in more complicated situations where both $\mathcal{G}$ and $\mathcal{H}$ are overlapping granularities.

In the Example 3 below, one investigates the impact of overlapping granularities. In many cases, the overlaps impact is limited. In situations where this impact can be important, one can use the construction of non-overlapping granularity presented in Proposition 4 (see Appendix C).

Example 3 (Overlapping granularity)**.** *Consider two overlapping grains g and g', with non-empty intersection $g_0 = g \cap g'$. We want to compare the situation where X_g is dependent on $X_{g'}$ with a situation of independence.*

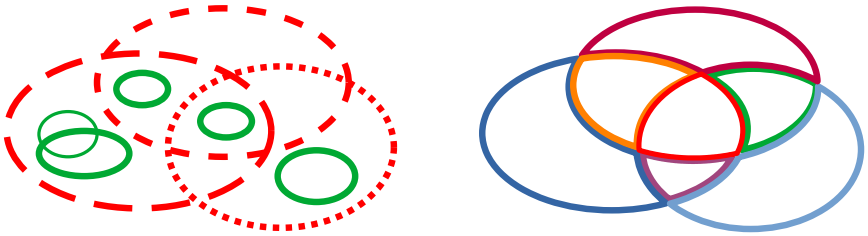

Figure 9: *Thinner granularity and maximal thinner non-overlapping granularity.* **Left:** The granularity comprising the 5 green grains (solid lines) is thinner than the granularity comprising the 3 red grains (dashed lines). **Right:** The granularity comprising 7 non overlapping grains is the maximal non-overlapping granularity that is thinner than the red granularity on the left.

- *Case of dependence. We define random locations* X_{g_0} , $X_{g\setminus g_0}$, $X_{g'\setminus g_0}$ *and two Bernoulli random variables* B *and* B'. *We assume that those five random variables are mutually independent. Let:*

$$\begin{cases} X_g &= BX_{g_0} + (1-B)X_{g\setminus g_0} \\ X_{g'} &= B'X_{g_0} + (1-B')X_{g'\setminus g_0} \end{cases} \tag{19}$$

- *Case of independence. We introduce here* $X_{g_0}^{\perp}$ *an independent copy of* X_{g_0} , *independent from* X_{g_0} , $X_{g\setminus g_0}$, $X_{g'\setminus g_0}$, B *and* B'. *Let:*

$$\begin{cases} X_g &= BX_{g_0} + (1-B)X_{g\setminus g_0} \\ X_{g'}^{\perp} &= B'X_{g_0}^{\perp} + (1-B')X_{g'\setminus g_0} \end{cases} \tag{20}$$

Let Δ *be the covariance difference due to the dependence structure of* X_g *and* $X_{g'}$,

$$\Delta := \mathrm{Cov}\left[Y_i(X_g), Y_j(X_{g'}^{\perp})\right] - \mathrm{Cov}\left[Y_i(X_g), Y_j(X_{g'})\right]. \tag{21}$$

Then setting $\rho_{\max} = \sup\left\{|k_{i,j}(x,x) - k_{i,j}(x,x')| \,:\, x \in g_0,\, x' \in g_0\right\}$, *assuming that*

$$\forall x \in g \cup g', \quad \begin{cases} \mu_i(x) = \mu_i(g) = \mu_i(g') \\ \mu_j(x) = \mu_j(g) = \mu_j(g') \end{cases}$$

one can show that:

$$|\Delta| \leq \mathbb{P}\left[B = B' = 1\right]\mathbb{P}\left[X_{g_0} \neq X_{g_0}^{\perp}\right]\rho_{\max}\,. \tag{22}$$

The variation due to the common dependence structure on the overlap can be significant if all of the three factors are not negligible. This shows in particular that overlapping grains are not too problematic, when means are identical, if the probability of selecting the intersection g_0 *for both grain is small, or if the probability of selecting different points in the intersection is small.*

Proof of the results in Example 3. Under given assumptions on the means μ_i and μ_j, Applying the total covariance formula on $\mathrm{Cov}\left[Y_i(X_g), Y_j(X_{g'}^{\perp})\right]$ and $\mathrm{Cov}\left[Y_i(X_g), Y_j(X_{g'})\right]$, we get

$$\Delta = \mathbb{E}\left[\mathrm{Cov}\left[Y_i(X_g), Y_j(X_{g'}^{\perp})|(B,B')\right]\right] - \mathbb{E}\left[\mathrm{Cov}\left[Y_i(X_g), Y_j(X_{g'})|(B,B')\right]\right],$$

and the difference is non zero in the only case where $B = B' = 1$, so that using independence,

$$\Delta = \mathbb{P}\left[B = B' = 1\right]\left(\mathbb{E}\left[\mathrm{Cov}\left[Y_i(X_{g_0}), Y_j(X_{g_0})\right]\right] - \mathbb{E}\left[\mathrm{Cov}\left[Y_i(X_{g_0}), Y_j(X_{g_0}^{\perp})\right]\right]\right)$$

The parenthesis vanishes in any conditional cases where $X_{g_0}^{\perp} = X_{g_0}$, and in other cases, the conditional difference is bounded by $\rho_{\max}$, hence the result. □

Contents