# About spatial interpolation using mixture distributions for predicting Energy Performance Certificate

Marc Grossouvre, Didier Rullière, Jonathan Villot

# About spatial interpolation using mixture distributions for predicting Energy Performance Certificate

Marc Grossouvre [1] & Didier Rullière [2] & Jonathan Villot

[1] *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`.
[2] *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.*
[3] *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.*

**Résumé.** Pour planifier la transition énergétique, les décideurs ont besoin d'une connaissance approfondie de leur territoire. Pour cela, des données sont collectées de sources multiples, à plusieurs échelles, avec des contraintes comme les politiques de confidentialité. De telles données fournissent des informations sur des zones spatiales mais sans localisation spécifique. C'est le cas du Diagnostic de Performance Energétique (DPE). Les bases de données des DPE réalisés sont publiées sous des contraintes spécifiques : anonymisation, géolocalisation avec adresse postale, détails manquants. Ce document montre que l'apprentissage des DPE observés pour prédire les DPE manquants peut être considéré comme un problème d'interpolation spatiale. Il présente une manière de traiter le DPE en tant qu'information géolocalisée et de prédire sa valeur au niveau du bâtiment. La méthodologie du krigeage est appliquée à des champs aléatoires observés à des emplacements aléatoires pour trouver le meilleur prédicteur linéaire non biaisé (BLUP). Ce nouveau modèle est appelé krigeage de mixtures. Bien que le cadre gaussien habituel soit perdu, nous montrons que la moyenne conditionnelle, la variance et la covariance peuvent être calculées. Ce nouveau modèle donne des résultats intéressants dans la prédiction du DPE au niveau du bâtiment, ce qui est une condition préalable pour que les décideurs ciblent les efforts de rénovation. Le cas spécifique d'une ville française est pris comme exemple. Le modèle présenté inclut également le co-krigeage de mixtures de sorte que les covariables puissent être utilisées pour améliorer le résultat. Il est également suggéré que le krigeage de mixtures puisse être utilement mis en oeuvre pour contrôler la propagation de l'incertitude. Nous présentons des applications en ce sens sur des données simulées.

**Mots-clés.** Processus multi-échelle, régression de surface à points, données surfaciques, krigeage par blocs, changement d'échelle, transition énergétique.

**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 maker to target renovation efforts. The specific case of a city in France is taken as an example. The presented model includes Mixture co-Kriging so that covariates 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

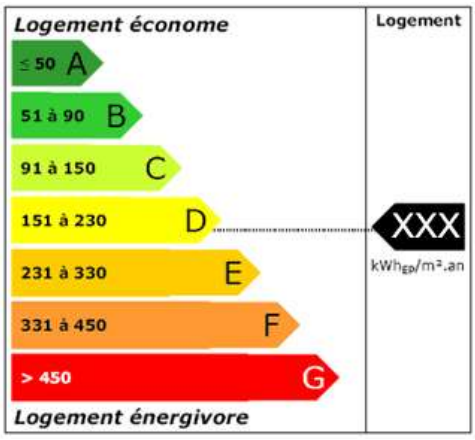


Figure 1: Prescribed vignette appearing on the French energy certificate up to 2021. Top left: efficient dwelling; Top right: dwelling; Bottom: energy intensive dwelling.

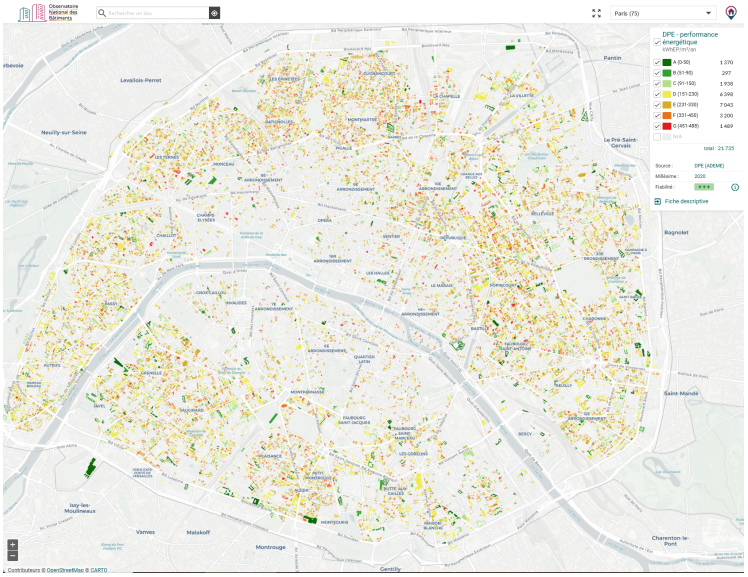

Figure 2: Map of inventoried EPCs in Paris. Screen capture of the French National Observatory of Buildings (Observatoire National des Bâtiments)

An Energy Performance Certificate (EPC) is given in France as an energy consumption associated with a qualitative labelling letter ranging from A to G as shown in Figure 1 . The labels are inventoried in a database along with the addresses of the dwellings and can be matched with a land plot in a second database. However, the exact location of each dwelling

on the land plot is uncertain. Decision-makers want to infer the energy consumption and label of buildings that have not been observed to identify targets for energy retrofit incentives. We call this problem the EPC prediction problem. The smallest unit of information for a table with one EPC per row is a part of a building, which is not clearly defined as an object in a 3D space but has features that describe it.

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.

Spatial interpolation is a technique that uses known geographical point samples to estimate values at unknown points. It mostly relies on the assumption that points close to each other in the input space are more likely to have similar output values. Gaussian Process Regression, also known as Kriging, is a major spatial interpolation approach based on a linear weighted combination of observation values that produces the Best Linear Unbiased Predictor (BLUP) for point spatial interpolation. However, the EPC prediction problem does not deal with points but with areal observations that involve the transformation of data from one set of boundaries to another. Areal interpolation research assumes that areas, also known as blocks, that are close to each other in the input space are more likely to have similar output values. Block Kriging is a derivative of Kriging that is designed for handling areal data. It assumes that feature at block level is an average value over the block. This averaging heavily influences the correlations between output variables in areal Kriging models, causing a family of problems described by Gotway and Young (2002), called the change of support problems among which is the Modifiable Areal Unit Problem (MAUP) and the ecological inference problem. Despite its limitations, the averaging method has proven to be effective for interpolating areal data, as demonstrated by several successful applications.

A new data model has been proposed by Godoy et al. (2022) to address change of support problems by defining a Gaussian random field on the class of closed subsets of a domain using the Hausdorff distance and a Matérn kernel. However, this model lacks interpretability and consistency between the output at the areal level and the point level. Additionally, it does not solve the input data uncertainty problem in the EPC prediction problem.

This paper proposes a new model that can handle both aggregated and point support data, introducing an object category called grain to express this approach. The model addresses issues related to determining a consistent covariance model for points and blocks, and proposes a method of incorporating a mixture distribution to account for stochastic dependence between blocks resulting from uncertainty on the input values. The approach effectively manages input uncertainty, but mixtures of Gaussian random variables are generally not Gaussian, so the usual Gaussian process interpretations and conditioning will no longer hold. The present document is a summary of a more detailed article of the same authors Grossouvre and Rullière (2023). It focuses on the problem of EPC prediction.

# 2 Optimal linear interpolation of mixture distributions

This approach assumes that the output variables, such as sociological variables, can be defined and 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 now 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 $\chi$ 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 $\chi$ is a finite set of grains of $\chi$ .*

**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 $\chi$ denoted: $\forall x \in \chi,\ \mathbf{Y}(x) := (Y_1(x), \ldots, Y_p(x))^\top \in \mathbb{R}^p$*
- *For each $g \in \mathcal{G}$, a $p$-dimensional real random vector $\mathbf{Y}(g)$ is defined to be the value of $\mathbf{Y}$ at a random location $X_g \in g$: $\forall g \in \mathcal{G},\ \mathbf{Y}(g) := \mathbf{Y}(X_g) \in \mathbb{R}^p$*

*For a given granularity $\mathcal{G}$, we assume that the set of random variables $\{X_g : g \in \mathcal{G}\}$, is defined and known, and that the dependence structure between those random variables is also known. We assume furthermore that these random variables are independent from $\mathbf{Y}$.*

We assume that the output is partially known on a set of grains: For $(i_1, \ldots, i_n) \in \{1, \ldots, p\}^n$ and $g_1, \ldots, g_n \in \mathcal{G}$ we know $n$ random variables:

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

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

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

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.

## 2.2 Best unbiased linear predictor

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. Now, 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 those moments.

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

*(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:*

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

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

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

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

*Proof.* (i) is a direct application of the conditional expectation formula where $Y_i(g)$ is the result of conditioning $Y_i(x)$ with $X_g$. And (ii) derives 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 any one of the grains $g$ 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. 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.

Let $\underline{\mathbf{Y}}$ be the vector of observations forming the learning set, $g \subset \chi$ a grain such that for some $i \in \{1, \dots, p\}$, $Y_i(g)$ is to be predicted. For a given set of weights $\boldsymbol{\alpha}(g) = (\alpha^1(g), \dots, \alpha^n(g)) \in \mathbb{R}^n$, we define a linear predictor $M_{\boldsymbol{\alpha}(g)}$:

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

Provided that those weights exist and are unique, the optimal weights $\boldsymbol{\alpha}_i(g)$ are defined

to be those minimizing a quadratic error, $\boldsymbol{\alpha}_i(g) \in \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \mathbb{E}\left[\left(Y_i(g) - \boldsymbol{\alpha}^\top \underline{\mathbf{Y}}\right)^2\right]$, over all unbiased linear predictors. The following proposition gives an optimal predictor that can be computed under the minimal assumptions of Proposition 1. It 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 $\mathbf{Y}(g)$: given the set of $p$ best predictors, we define the matrix $A(g) = (\alpha_i^j(g))_{i\in\{1,\dots,p\},\, j\in\{1,\dots,n\}}$ so that the predictor $M_{A(g)} = A\underline{\mathbf{Y}}$ is the best linear unbiased predictor of $\mathbf{Y}(g) = (Y_1(g)\dots Y_p(g))^\top$ for the total quadratic error $\mathbb{E}\left[||\mathbf{Y}(g) - A\underline{\mathbf{Y}}||_2^2\right]$.

**Proposition 2** (Simple Mixture Kriging prediction)**.** *Given a set of observations $\underline{\mathbf{Y}}$, for any $g \subset \chi$, and in particular for a single point $g = \{x\}$, for any $i \in \{1,\dots,p\}$, 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: If $\underline{\boldsymbol{\mu}} = (0,\dots,0)^\top$ and $\mu_i(g) = 0$ then*

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

*Where $\underline{\boldsymbol{\mu}} := \mathbb{E}\left[\underline{\mathbf{Y}}\right] \in \mathbb{R}^n$, $\mathbf{K} := \left(\mathrm{Cov}\left[Y^j, Y^{j'}\right]\right)_{j,j'\in\{1,\dots,n\}} \in \mathcal{S}_n^+(\mathbb{R})$ (semi-definite positive $n \times n$ matrix), $\mathbf{h}_i(g) := \left(\mathrm{Cov}\left[Y^j, Y_i(g)\right]\right)_{j\in\{1,\dots,n\}} \in \mathbb{R}^n$, $\epsilon_i(g) := Y_i(g) - M_i(g)$, $v_i(g) := c_{i,i}(g,g)$ and $c_{i,j}(g,g') := \mathbb{E}\left[\epsilon_i(g)\,\epsilon_j(g')\right]$.*

$\mathbf{K}$ is assumed to be invertible. Note that if the expectations of $Y_i(x)$ and covariances between $Y_i(x)$ and $Y_j(x')$ are known for all $i,j \in \{1,\dots,p\}$, $x, x' \in \chi$, $\underline{\boldsymbol{\mu}}$, $\mathbf{K}$ and $\mathbf{h}_i(g)$ can be computed using Proposition 1. Similar results are available for Ordinary Mixture Kriging in Grossouvre and Rullière (2023).

# 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 on the one hand we observe EPC with an uncertainty on the location of the observed dwelling on the land plot where it lies. The observed dwelling can not be identified among all the known 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 dwellings 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 a 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.

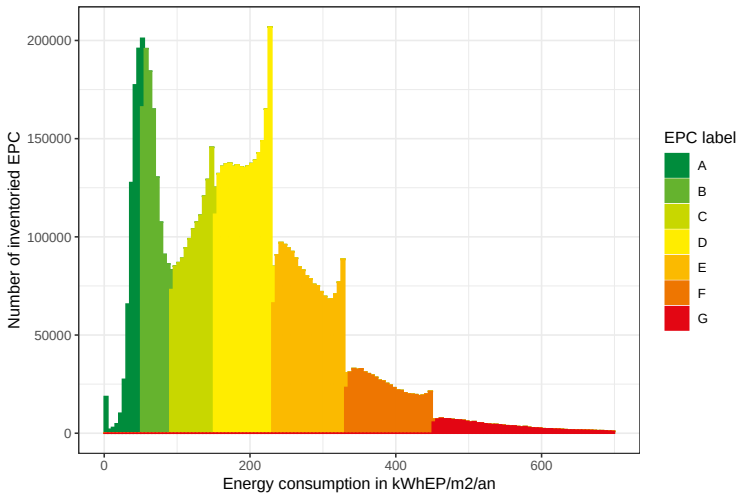


Figure 3: Bar plot of EPC labels frequencies among all EPCs collected in France between 20214 and 2021. Classes are highly heterogeneous.

Denoting $T_L$ the number of true observations with label $L$ and $TP_L$ the number of good predictions with label $L$, the balanced accuracy is given by the formula:

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

Consider the following model $M_1$:

- $\chi$ is the territory of an urban area in Angers city in France in a 3 dimensional space where coordinates represent construction year, latitude and longitude.
- A random field $Y(x)$ is defined on $\chi$. 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 point in a 3 dimensional space $\chi$. 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 $\mathcal{G}$. Variables are normalized based on standard normal Gaussian quantiles associated with ranks so that $\chi = \mathbb{R}^3$.
- 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 $\underline{\mathbf{Y}}$.

Note that grains seem to be disjoint on Figure 4 but they are not due to overlaps on the 3$^{\text{rd}}$ dimension. Moreover, by construction, $Y$ is centred. 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

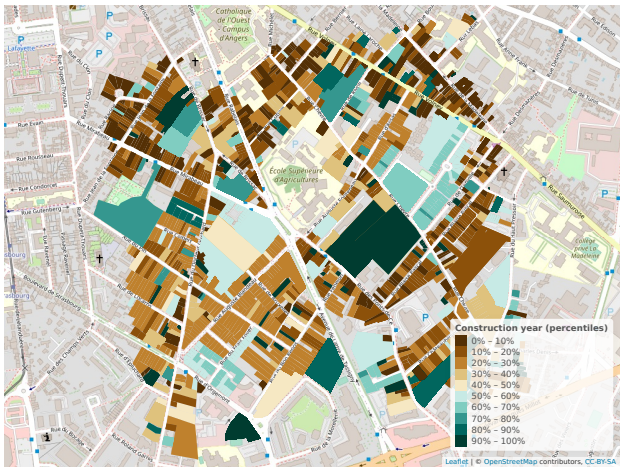


Figure 4: Construction year of an urban area in Angers. The side of the square is 1km. Construction years range from 1340 (first percentile) to 2019 (last percentile).

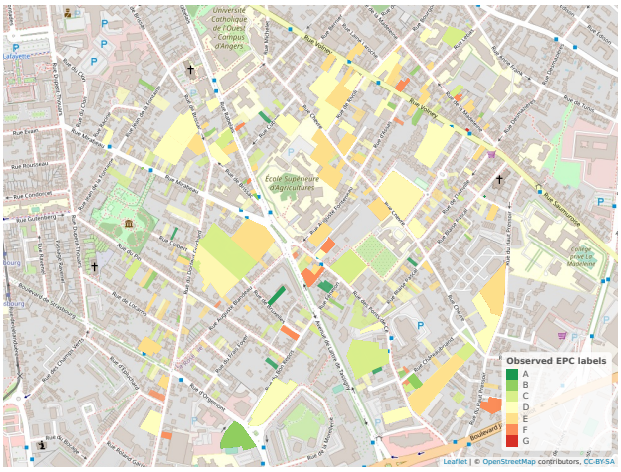


Figure 5: Map of the 365 observations. Each colour represents a label associated with the numeric value. See Figure 1 .

| | | EPC int. | | EPC num. | | |
|---|---|---|---|---|---|---|
| Model | BA | MAE | RMSE | MAE | RMSE | Variance |
| 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.

Table 1: Optimal performances achieved by $M1$ and $M2$ and respective overall variance of numeric predictions.

Matérn $3/2$ model. $\sigma^2$ 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 is used to predict energy consumption a plot level. Without nugget effect the mean prediction, in the case of a one dimensional output, does not depend on $\sigma^2$. $\sigma^2$ is therefore set to 1. $\Theta$ is chosen so as to maximize the BA criterion of the predicted labels derived from the predicted energy consumptions. BA is computed using leave-one-out cross validation.

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.

| True labels | Predicted labels | | | | | | |
|---|---|---|---|---|---|---|---|
| | 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 2: Confusion matrix of $M1$ predictions

| True labels | Predicted labels | | | | | | |
|---|---|---|---|---|---|---|---|
| | 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 3: Confusion matrix of $M2$ predictions

- A nugget effect has to be introduced so as to avoid oscillation since Kriging predictor is otherwise interpolating: $\mathbb{V}[Y(x)] = \sigma^2 + \epsilon^2$ 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. The percentage of true labels A and B that are predicted as A or B is 25% with $M1$ (Mixture Kriging) and 10% with $M2$ (Kriging). 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 and yields a better BA as compared with Kriging which requires the introduction of a nugget effect.

# 4 Discussion and conclusion

Mixture Kriging is a new model that is consistent with Kriging, but it generates mean predictions that are not impacted by the size or shape of the grains. This means that there is no prediction shrinkage due to these factors. Mixture Kriging also has no measurable Modifiable Areal Unit Problem (MAUP) effect, and its predictions are smooth without introducing a nugget effect, which tends to shrink the mean predictions in Kriging models. Mixture Kriging is designed to handle data with uncertainty on input values without introducing a nugget effect. The main difference between block-to-block Kriging and Mixture Kriging is in the method of computing the observations variance and covariance between covariates associated with the same grain. This results in a higher diagonal value in the

observations covariance matrix than what is found with Kriging, which makes the model predictions smoother, which is the sought effect of introducing a nugget effect in Kriging. However, Mixture Kriging has a higher computational cost than Kriging, which increases with the squared value of the density of points in the grains.

This new approach opens the way for feeding 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 numeric 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

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`).

# Bibliographie

Godoy, L. da C., Oliveira Prates, M. and Yan J. (2022), An unified framework for point-level, areal, and mixed spatial data: the Hausdorff-Gaussian Process, `http://arxiv.org/abs/2208.07900` (preprint).

Gotway, C. and Young L. (2002), Combining Incompatible Saptial Sata, *Journal of the American Statistical Association*, 97, pp. 632-648.

Gösgens, M., Zhiyanov, A., Tikhonov, A. and Prokhorenkova, L. (2021), Good Classification Measures and How to Find Them, *Advances in Neural Information Processing Systems, 34, pp. 17136-17147.*

Grossouvre, M. and Rullière, D. (2023), Spatial interpolation using mixture distributions: A Best Linear Unbiased Predictor, `https://hal.science/hal-03276127` (preprint).