Mixture kriging on granular data

A new approach for imputing building's energy efficiency

Marc Grossouvre, data-scientist at U.R.B.S. [a]

Didier Rullière, researcher at Mines Saint-Etienne[b].

[a] Marc Grossouvre, U.R.B.S. SAS, Saint-Etienne, France, marcgrossouvre@urbs.fr , website `www.urbs.fr`.

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

Abstract

This work deals with three related problems in a geostatistical context.

- Some data are available for given areas of the space, rather than for some specific locations, that's **granular data**.
- Data uncertainty rely both on the input locations and on measured quantities at these locations.
- Multidimensional outputs can be observed, with sometimes missing data.

These problems are addressed simultaneously by considering **mixtures of multivariate random fields**, and by adapting standard **Kriging methodology** to this context. While the usual Gaussian setting is lost, we show that conditional mean, variance and covariances can be derived from this specific setting.

Context

French government recently released a database inventoring the energy efficiency of dwellings that have been diagnosed in the last 10 years. It has therefore become possible to do supervised learning to try and predict the energy efficiency of any dwelling in France and possibly assess whether those dwellings have undergone energy efficient retrofit. A major issue in this project is that most models for energy efficiency are engineering models that require physical visits to collect parameters. From U.R.B.S. company point of view, the question is therefore: **How to predict buildings energy efficiency without visiting them?** From geostatistics point of view, we can assume that granular data such as census data could help us. The question becomes: **How to modelize such granular data in order to include it in a supervised learning model?**

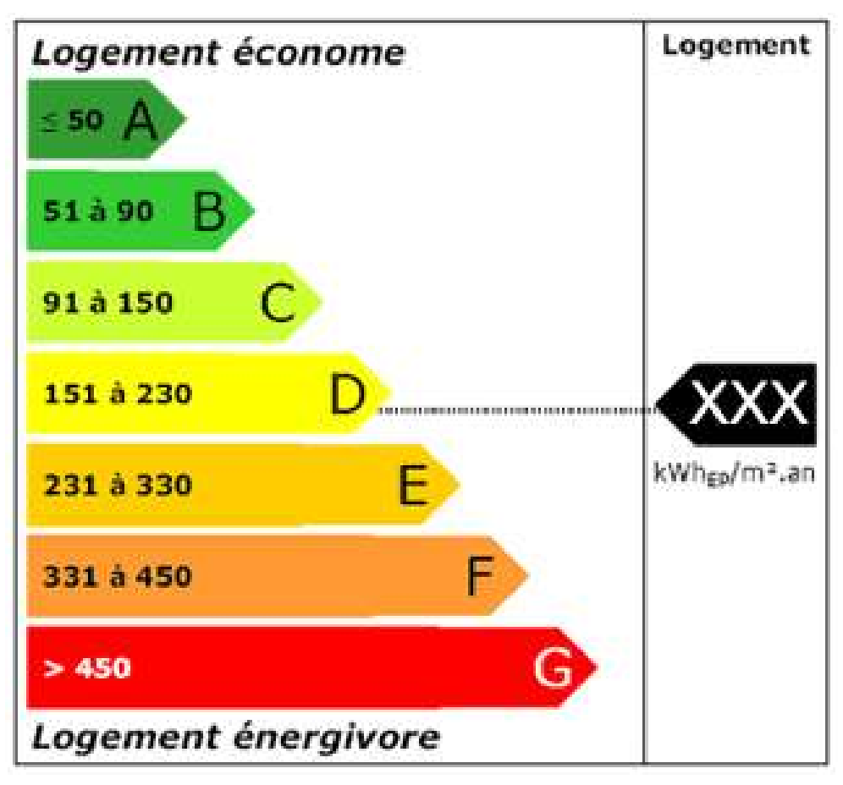

Figure 1: Buildings labelling according to their annual energy consumption per square meter.

Contribution to achieving sustainable development objectives

- *Objectif 11 : Faire en sorte que les villes et les établissements humains soient ouverts à tous, sûrs, résilients et durables.* Act so as to make cities and human buildings open to all, safe, resilient and sustainable.
- *Objectif 12 : Établir des modes de consommation et de production durables.* Adopt sustainable methods of consumption and production.
- *Objectif 13 : Prendre d'urgence des mesures pour lutter contre les changements climatiques et leurs répercussions.* Take emergency measures to fight climatic changes and their impact.

1 Modelize granular data as mixtures rather than averages

Consider output variables such as building's insulation level, households income, square meter price, building's construction date... as a random vector field $\mathbf{Y}(x)$defined at each dwelling x of a district χ. **How to define $\mathbf{Y}$ at a census tract level, denoted $\mathbf{Y}(g)$, g standing for "grain", seen as a set of dwellings?** $\mathbf{Y}(g)$ is commonly seen as the average Y over the census tract. However, averaging reduces the dispersion of the variable as the scale grows. And the application of any highly convex function h would induce a large bias, as $\mathbb{E}\left[h(\mathbf{Y}(g))\right] \neq h\left(\mathbb{E}\left[\mathbf{Y}(g)\right]\right)$ (see Figure 1). Underestimating the dispersion of an output random variable is an adverse effect when we plan to feed a machine learning algorithm with this data. We therefore prefer the mixture approach:

Definition 1 (Outputs). *Outputs are defined on each **point** x of a **territory** χ and each **grain** g of a set of grains $\mathcal{G}$, called a **granularity**.*

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

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

- *For $g \in \mathcal{G}$, denote $\mathbf{Y}(g)$ a p-dimensional real random vector that is $\mathbf{Y}$'s value 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 $\left\{X_g\ :\ g \in \mathcal{G}\right\}$ 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 the random field $\mathbf{Y}$.

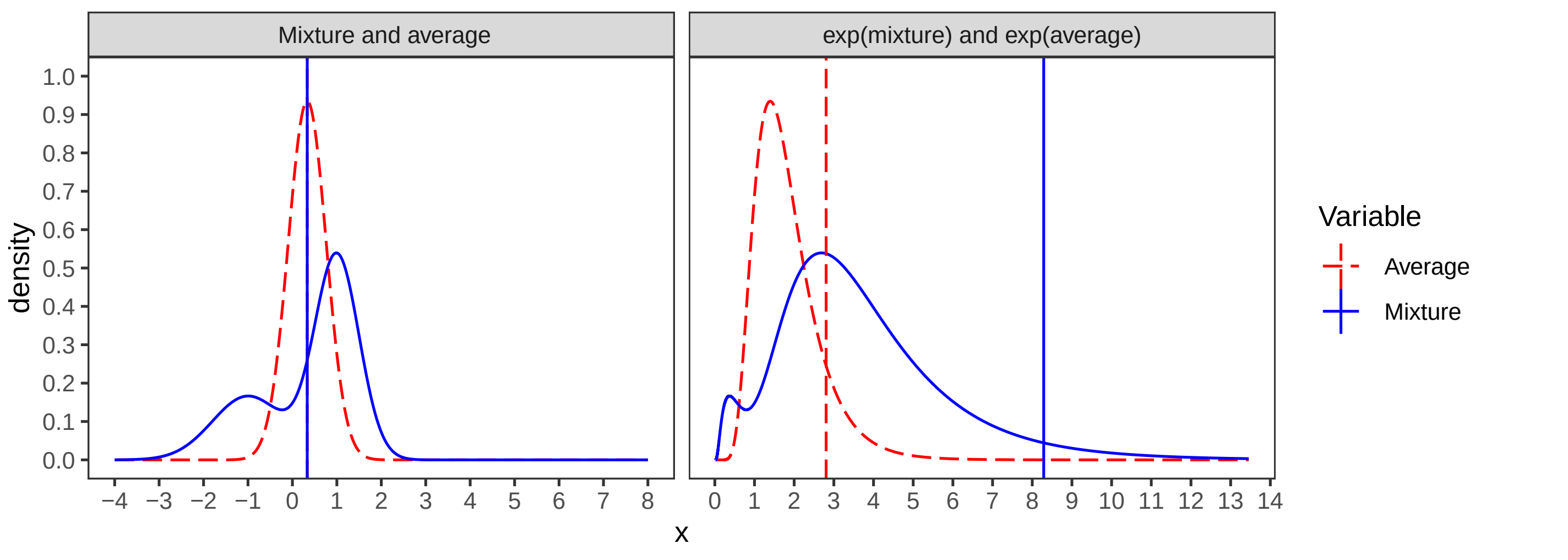

Figure 2: *Compared densities of mixture and average.* Given $Y_a \sim N(-1, 0.8)$ and $Y_b \sim N(1, 0.5)$.
Left: Mixture Y_{mixture} picking Y_a or Y_b with probabilities $\frac{1}{3}$ and $\frac{2}{3}$ (blue plain line). Average $Y_{\text{average}} = \frac{1}{3}Y_a + \frac{2}{3}Y_b$ (red dashed line). Distributions'common mean appears as a vertical line. The mixture is not Gaussian and has a higher dispersion.
Right: $\exp(Y_{\text{mixture}})$ (blue plain line). $\exp(Y_{\text{average}})$ (red dashed line). Differing means appear as vertical lines. Dispersion of the mixture's exponential is higher than the other, and convexity induces a the difference between means.

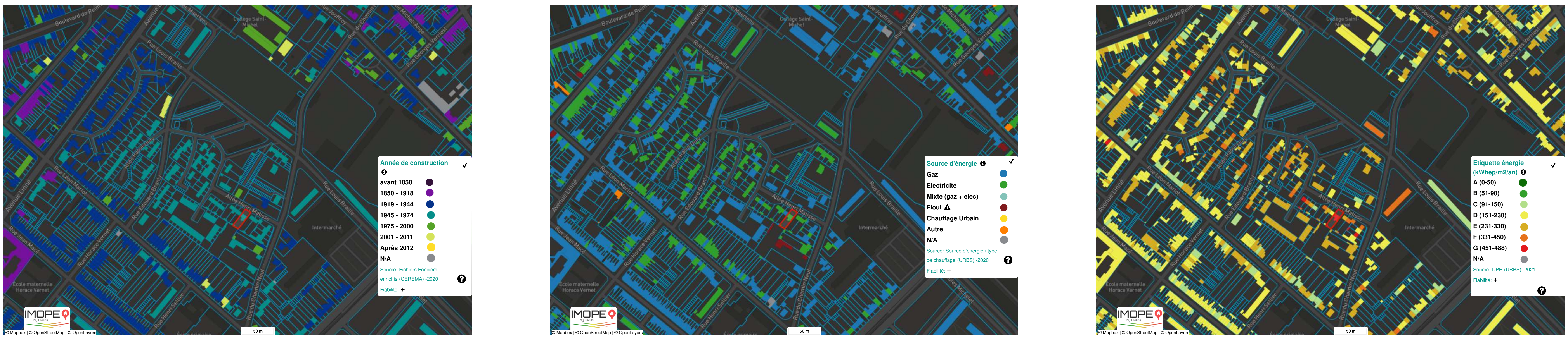

Figure 3: For a neighborhood of Roubaix city, we observe from left to right: building's period of contruction, energy type used for heating the building, building's energy efficiency. The building lying on the red squared landplot has been build in the 50's-60's, is heated with electricity and has a very poor energy efficiency. These images are print outs of IMOPE software, developped by U.R.B.S. company in Saint-Étienne, France.

2 From a mixture model, derive a Best Linear Unbiased Predictor

To define **learning data** we now assume that the output is partially known on a granularity $\mathcal{G}$, i.e. for $(i_1, \ldots, i_n) \in [\![1, p]\!]^n$ and $g_1, \ldots, g_n \in \mathcal{G}$, we know n random variables :

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

For some $g \subset \chi$ and some $i \in [\![1, p]\!]$, we want to **predict** $Y_i(g)$ from a **learning set** $\underline{\mathbf{Y}}$.
We denote :

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

Proposition 1. *If the expectations of $Y_i(x)$ and covariances between $Y_i(x)$ and $Y_j(x')$ are known for all $i, j \in [\![1, p]\!]$, $x, x' \in \chi$, as in usual Kriging assumptions, then $\underline{\boldsymbol{\mu}}$, $\mathbf{K}$ and $\mathbf{h}_i(g)$ can be computed.*

We assume that $\mathbf{K}$ is invertible. We optimize weights $\boldsymbol{\alpha}_i(g) = \left(\alpha_i^j(g)\right)_{j \in [\![1,n]\!]} \in \mathbb{R}^n$ to get a linear unbiased predictor $M_i(g)$ of $Y_i(g)$:

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

minimizing the quadratic error:

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

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

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

Given the first two moments of random variables $\left\{X_g\ :\ g \in \mathcal{G}\right\}$, we get the following result:

Proposition 2 (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 $i \in [\![1, p]\!]$, 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*

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

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

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