



HAL
open science

Assessing the energy performance of the building stock at a fine scale in order to massify the renovation works and fulfill the national commitments

Marc Grossouvre

► To cite this version:

Marc Grossouvre. Assessing the energy performance of the building stock at a fine scale in order to massify the renovation works and fulfill the national commitments. Environmental Sciences. École des Mines de Saint-Etienne, 2024. English. NNT : 2024EMSEM040 . tel-04888233

HAL Id: tel-04888233

<https://hal.science/tel-04888233v1>

Submitted on 15 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



NNT : 2024EMSEM040

THÈSE DE DOCTORAT

de l'Institut Mines-Télécom
École Nationale Supérieure des Mines Saint-Étienne

Ecole Doctorale N°488 (Sciences, Ingénierie, Santé)

Spécialité de doctorat : Sciences et génie de l'environnement

Soutenue publiquement le 8 octobre 2024 par :

Marc Grossouvre

Évaluation de la performance énergétique du parc bâti à fine échelle dans une optique de massification des travaux et d'atteinte des objectifs nationaux

Devant un jury composé de :

Elena DI BERNARDINO	Professeure, Université Côte d'Azur, Nice	Rapporteuse
Robin GIRARD	Professeur, Directeur de Recherche, Mines Paris	Rapporteur
Natacha GONDRAN	Professeure, Mines Saint-Etienne	Examinatrice
Gaël POETTE	Ingénieur-chercheur CEA/Prof. Bordeaux INP	Examineur
Simon ROUCHIER	Maître de conférence, Polytech Annecy Chambéry	Examineur
Didier RULLIERE	Professeur, Mines Saint-Etienne	Directeur de thèse
Jonathan VILLOT	Maître assistant, Mines Saint-Etienne	Co-encadrant

Affidavit

Je soussigné, Marc Grossouvre, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Didier Rullière, le co-encadrement de Jonathan Villot, dans le respect des principes d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect de la charte nationale de déontologie des métiers de la recherche.

Ce travail n'a pas été précédemment soumis dans sa globalité en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Saint-Etienne, le 14 juillet 2024

Ce travail de thèse est une œuvre de l'esprit, protégée par le droit d'auteur, tel que prévu aux articles L111-1 du CPI et suivants disposant que « *L'auteur d'une œuvre de l'esprit jouit sur cette oeuvre, du seul fait de sa création, d'un droit de propriété incorporelle exclusif et opposable à tous. [...]* »

Il est rappelé que par exception au droit d'auteur, la loi française autorise l'utilisation d'une œuvre divulguée, sans autorisation de son auteur, suivant les conditions définies dans l'article L122-5 du CPI disposant que « *Lorsque l'œuvre a été divulguée, l'auteur ne peut interdire [...] la représentation ou la reproduction d'extraits d'œuvres, [...] sous réserve que soient indiqués clairement le nom de l'auteur et la source [...] les analyses et courtes citations justifiées par le caractère critique, polémique, pédagogique, scientifique ou d'information de l'œuvre à laquelle elles sont incorporées [...]* »

To the well being of present and future generations,

Acknowledgements

It took me much more time than most Ph.D. students and many more encounters to finally indulge in writing a thesis. Among the contributors to my relationship with research, I should thank, among others, Professor Velimir Jurdjevic from the University of Toronto and Professor Roshdi Rashed from the University of Paris VII. They were the first ones to show me the pleasures of research.

This Ph.D. project emerged from a maze of converging interactions between benevolent people to whom I am very grateful: Maximilien Brossard and Jonathan Villot, who got the bright idea of creating **U.R.B.S.**; Mines Saint-Etienne, in particular Rodolphe Leriche and Mireille Batton-Hubert, who pointed out the significance of this project; and Jennifer Hyslop, from **DSTI**, who encouraged me on this new adventure all the way.

During those three years, three interns contributed to this research: Ndeye Khady Mbengue, Rafaël Quiblier, and Nathan Seychal. Thank you for supporting me and for your patience. I wish you luck and success in your professional life.

As sources should be named, I must confess that my work for collecting the data manipulated in this work is minimal as compared to the work of my colleagues Safia Raouf and Benoît Génot. Thank you so much for your work and friendship.

And of course, a bundle of thanks goes to Didier Rullière, who properly survived my thesis. Thank you for opening my mind to geostatistics. Thank you for your unwavering support and infinite patience. We shared more than time and work; thank you for that too.

As Didier rightly pointed out regarding journal reviewers, it is important to express gratitude to those who invest their time in reading and commenting on one's work. In the same spirit, I extend my heartfelt thanks to you, esteemed members of the jury, for granting me the honour of your time and attention.

Foreword

Pr. Roshdi Rashed taught me that any research belongs to one or more traditions. It is not possible to understand a research work if it is not connected with the preceding, and sometimes following, works in this tradition. Today, we call this tradition “bibliography” or “state of the art”, and we tend to reduce it to some specific literature, as is done in **Introduction**. But there are other traditions to explore if one wants to embrace the conditions of the production of knowledge. These traditions can be religious, political, social, or stylistic, among others.

Here, what matters to us is the quality of dwellings. A dwelling is both determined by and determinant of the relationship of individuals with their environment, be it material or immaterial. And the quality of a dwelling, even though we try to reduce it to a few quantitative indicators, is resulting from interactions between some walls, a roof, some windows, and the wind, the temperature, the rain, but also the behaviour of humans who move, eat, and sleep in these dwellings. This is the reason why geostatistics, thermal engineering, data fusion, and machine learning are not enough to explain the birth of this research project, which also belongs to the modern tradition of environmental studies.

We could try to draw a simplified causality chain leading to this thesis as follows: The industrial revolution valued fossil energy-powered machines in Europe and ignited climate change. The aftermath of World War II required reconstructing Europe and brought in millions of immigrants, for whom dwellings were hastily constructed. The conjunction of those phenomena brought us to a situation where millions of dwellings are not adequate for their inhabitants to live decently. Moreover, these dwellings consume a lot of energy. Beyond the fact that poor people are packed indignantly in those huge buildings containing hundreds of dwellings, the consequences of energy overconsumption and its impact on the economy and the environment are a matter of concern. At this point, the lawmakers enter the scene. They define scales and thresholds, authorise and forbid, and provide money for supporting renovation projects. The burden of implementing this falls on the shoulders of local institutions, departments, and municipalities. Hearing about it, some guys in Mines Saint-Etienne start developing a solution to help these decision-makers in their task. This question of producing an improved knowledge of the building stock from the energy performance point of view arises.

At the same time, not so far from Mines Saint-Etienne, I am seeking a career transition

in accordance with my wish to humbly contribute to the ecological transition with my best abilities, which are doing mathematics and coding algorithms. I have bookmarked in my navigator this video of Maximilien and Jonathan pitching their project. And after completing my training in data science at [DSTI](#), I tell them that I want to be part of their project. They make the connection between my abilities, energy efficiency, lawmakers, and their new company and offer me to start this Ph.D. project.

And it is only one century after the primary cause of the problem appeared that thermal engineering, geostatics, and others were successfully called for help, showing the interdisciplinary aspect of environmental studies. Although this thesis contains important mathematical developments, the reader may benefit from keeping in mind that the end users of our results are not statisticians but politicians and urbanists.

Overview

Introduction (en français)	1
Introduction	27
I Pre-processing Data	51
II Handling Multi-Scale Uncertainties	65
III Constrained Classification	91
IV Enhancing Prediction's Performances	135
Conclusion and Future Directions	155
Conclusion et perspectives (en français)	163
Bibliography	169
Supplementary Material	181
Supplementary Material for Pre-processing data	181
Supplementary Material for Mixture Kriging	221
Supplementary Material for Joint Kriging	245
Supplementary Material for Fuzzy Classification	263
Supplementary Material for the Conclusion	269
Table of Contents	297
List of Tables.	298
List of Figures	300

Acronyms

This document is incomplete. The external file associated with the glossary ‘acronym’ (which should be called `00_main.acr`) hasn’t been created.

Check the contents of the file `00_main.acn`. If it’s empty, that means you haven’t indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can’t be generated. If the file isn’t empty, the document build process hasn’t been completed.

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`.

For example:

```
\usepackage[automake]{glossaries-extra}
```

- Run the external (Lua) application:

```
makeglossaries-lite.lua "00_main"
```

- Run the external (Perl) application:

```
makeglossaries "00_main"
```

Then rerun \LaTeX on this document.

This message will be removed once the problem has been fixed.

Mathematical Notations

General Notations

Sets

\mathbb{R}	The set of real numbers.
\mathbb{R}^n	The \mathbb{R} -vector space of dimension n .
$\mathcal{M}_{n \times m}(\mathbb{R})$	The \mathbb{R} -vector space of $n \times m$ matrices.
$\mathcal{S}_n^+(\mathbb{R})$	The \mathbb{R} -vector space of symmetric semi-definite positive $n \times n$ matrices.
$\mathcal{S}_n^{+*}(\mathbb{R})$	The \mathbb{R} -vector space of symmetric definite positive $n \times n$ matrices.
\mathbb{N}	The set of natural numbers.
$\{a, b\}$	A set comprising elements a and b .
$\{1, \dots, n\}$	The set of natural numbers from 1 to n .
$[A]$	The number of elements in the finite set A .
$\{x : x \geq 2\}$	The set comprising all elements x such that $x \geq 2$.
$\mathbb{1}_{\{A\}}(x)$	The indicator function of the set A evaluated at point x .
$\mathbf{1}_n, \mathbf{0}_n$	A column vector of n ones, zeros.

Probabilities

$P[A B]$	The probability of A knowing B.
$E[X]$	The expectation of the random variable X .
$\text{Var}[X]$	The variance of the random variable X .
$\text{Cov}[X, Y]$	The covariance between the random variables X and Y .
$\text{Corr}[X, Y]$	The correlation between the random variables X and Y .

Operators

\mathbf{A}^\top	The matrix transposed of matrix \mathbf{A} .
$\text{diag}[\cdot]$	The diagonal of a matrix.
$\mathcal{L}(x, \lambda)$	The Lagrangian function of variable x and Lagrangian multiplier λ .
$:=$	Is defined to be equal to.
$ x $	Absolute value of x .
\cap, \cup	Intersection and union of sets, and probabilistic operators AND and OR.

Specific Variables

Locations, input variables, design points

χ	The set of all locations.
g, h	Grains: subsets of χ .
\mathcal{G}, \mathcal{H}	Granularities: sets of grains.
n, q	The number of observed locations, of prediction locations.
x	Any location.
x_1, \dots, x_n	All observed locations.
x^*	Any prediction location.
x_1^*, \dots, x_q^*	All prediction locations.
X_g	Random position on a grain g .
X^*	A random variable over prediction locations.
π	The $q \times 1$ distribution of X^* given by $(\pi_{x_1^*}, \dots, \pi_{x_q^*})$.

Output variables

p	The number of output variables.
$\mathbf{Y}(x)$	The $p \times 1$ vector of output variables at location x .
$\boldsymbol{\mu}$	The $p \times 1$ mean of $\mathbf{Y}(x)$, when constant over x : $\mathbb{E}[\mathbf{Y}(x)]$.
\mathbf{Y}	The $p \times n$ matrix of observed output variables: $[\mathbf{Y}(x_1), \dots, \mathbf{Y}(x_n)]$.
\mathbf{Y}^*	The $p \times q$ matrix of output variables to predict: $[\mathbf{Y}(x_1^*), \dots, \mathbf{Y}(x_q^*)]$.

Prediction

$\mathbf{M}(x^*)$	A $p \times 1$ predictor of $\mathbf{Y}(x)$
\mathbf{M}	The $p \times q$ matrix of all predictions $[\mathbf{M}(x_1^*), \dots, \mathbf{M}(x_q^*)]$.
$\boldsymbol{\alpha}(x^*)$	The $n \times 1$ linear weights for the prediction in x^* .
\mathbf{A}	The $n \times q$ matrix of weights for all predictions $[(\boldsymbol{\alpha}(x_1^*), \dots, \boldsymbol{\alpha}(x_q^*))]$.
\mathbf{m}	A given constant $p \times 1$ vector of prescribed mean predicted values.
$\Delta(x^*)$	The loss to be minimised for finding $\mathbf{M}(x^*)$.
$\mathcal{C}_{\mathbf{M}, \hat{\mathbf{M}}}$	A confusion matrix comparing true membership degrees with predicted ones.
$\boldsymbol{\lambda}$	A $q \times 1$ vector of Lagrange multipliers (relative to sum of weights).
$\boldsymbol{\lambda}'$	A $p \times 1$ vector of Lagrange multipliers (relative to predicted values).
\mathbf{Z}	An additional $p \times 1$ factor for affine predictions.

Covariances

\mathbf{W}	A given symmetric positive definite matrix for computing norms.
$\mathbf{h}(x^*)$	The $n \times 1$ vector given by $\mathbb{E}[\mathbf{Y}^\top \mathbf{W} \mathbf{Y}(x^*)]$.
\mathbf{H}	The $n \times q$ matrix given by $[\mathbf{h}(x_1^*), \dots, \mathbf{h}(x_q^*)]$.
\mathbf{K}	The $n \times n$ matrix given by $\mathbb{E}[\mathbf{Y}^\top \mathbf{W} \mathbf{Y}]$.
$\tilde{\mathbf{K}}, \tilde{\mathbf{h}}(x^*), \tilde{\mathbf{H}}$	The same as above but using centred expressions (covariances).
\mathbf{P}	The $n \times 1$ covariance vector between \mathbf{Z} and $\mathbf{Y}(x_i)$.
\mathbf{Q}	The $q \times 1$ covariance vector between \mathbf{Z} and $\mathbf{Y}(x_j^*)$.

Introduction (en français)

1	Contexte institutionnel et environnemental	2
2	Vue d'ensemble du projet de recherche	3
2.1	Problématique de recherche	3
2.2	Hypothèses et objectifs.	4
2.3	Champ d'application et limites.	5
2.4	Importance de ce travail	6
3	Méthodologie et structure générale du document	6
4	État des lieux	9
4.1	Qu'est-ce que le diagnostic de performance énergétique ?	9
4.2	À quoi ressemble un DPE ?	12
4.3	Objectifs de rénovation.	14
5	État de l'art	17
5.1	Évaluation de la performance énergétique d'un parc immobilier	17
5.2	Interpolation spatiale	20

Le principal but de cette thèse est de construire un modèle prédictif explicable pour estimer l'efficacité énergétique de chaque bâtiment en France. L'ensemble d'entraînement pour ce travail est fourni par l'ensemble des **DPE** (*Diagnostic de Performance Énergétique – Energy Performance Certificate*) observés. Après une présentation des contextes juridiques et scientifiques actuels dans les sections suivantes 4 et 5, le processus de prétraitement des données et ses défis sont présentés dans le Chapitre I. Le Chapitre II introduit une méthode novatrice pour interpoler les distributions de mélange, appelée Mixture Kriging. Ce modèle tient compte de l'incertitude des emplacements des bâtiments et présente l'important avantage de ne pas avoir de problème d'agrégation spatiale – **MAUP** (**Modifiable Areal Unit Problem**) mesurable. Ce modèle est efficace à petite échelle – une ville, par exemple – mais il s'avère trop lourd pour être étendu au niveau national. Compte tenu des enseignements de ce modèle, le Chapitre III présente un modèle plus léger, appelé Joint Kriging, qui ne tient pas compte de l'incertitude de position des bâtiments et a un effet **MAUP**, mais se révèle être très efficace et facile à étendre à grande échelle. Ce modèle peut être contraint afin d'imposer une distribution globale des prédictions sur un territoire donné, ce qui est particulièrement intéressant pour prédire les **DPE**. Mixture Kriging est un modèle de régression alors que Joint Kriging est un modèle de classification floue. Le Chapitre IV compare les résultats de classification de Joint Kriging avec ceux du modèle actuellement utilisé par la société **U.R.B.S.**. Ce dernier est un modèle **KNN** (**k-Nearest Neighbours**) optimisé. Ces deux modèles sont également comparés avec un classificateur dur, Random Forest. La classification floue avec Joint Kriging se révèle très prometteuse et a été choisie pour remplacer le modèle actuel utilisé à **U.R.B.S.**. Les Chapitres II, III, et IV sont trois articles de recherche qui ont été écrits au cours des trois dernières années. Le premier et le troisième sont déjà publiés; le second est en cours de révision.

1 Contexte institutionnel et environnemental

Ce travail résulte de la conjonction de trois facteurs déterminants : le contexte écologique du dérèglement climatique, la décision de l'Union européenne de tenter d'atteindre la durabilité, et la création de la société **U.R.B.S.** qui accompagne les institutions dans la gestion de la transition du parc immobilier situé sur le territoire qu'elles administrent.

Les politiques européennes en faveur de la durabilité, définies par de multiples directives du Parlement européen [1], [2], incitent les scientifiques, les parties prenantes et les politiques à explorer de nouvelles approches pour réduire la consommation d'énergie et minimiser les émissions de **GES** (**Gaz à effet de serre**). Les pays européens définissent donc des stratégies pour améliorer la performance énergétique des activités anthropiques et relever les défis urgents du changement climatique. Conformément à cette feuille de route, les initiatives visant à renforcer les mesures d'efficacité énergétique dans le secteur du bâtiment sont essentielles. En effet, le secteur du bâtiment est l'un des principaux consommateurs d'énergie au monde, représentant 40 % de la consommation d'énergie

européenne, [3], [4]. Cette consommation contribue de manière significative aux émissions de GES (36 % du total), principalement les émissions de CO₂, altérant ainsi le climat de notre planète. Et elle a connu une tendance générale à la hausse au cours des dernières décennies.

La société **U.R.B.S.** a été fondée en 2019 pour apporter un soutien aux décideurs des politiques publiques urbaines. Une équipe pluridisciplinaire composée d'ingénieurs de la donnée, de développeurs de logiciels et de chercheurs produit une base de données rassemblant les informations disponibles sur les logements. Cette base est enrichie par l'appariement de sources, l'imputation de données manquantes et le calcul d'indicateurs de synthèse. Elle est mise à disposition des utilisateurs via un géo-service personnalisable pour chaque territoire. Il est rapidement apparu à l'entreprise que les institutions locales telles que les **EPCI** (*Établissement Public de Coopération Intercommunale – Public Intermunicipal Cooperation Authority*) ou les départements manquaient de connaissances sur le parc immobilier du territoire qu'ils administrent, ce qui retardait la montée en puissance de la rénovation énergétique dans le contexte précité des politiques européennes de durabilité. Désireux de contribuer à cet effort collectif, **U.R.B.S.** et Mines Saint-Étienne ont lancé, en 2021, un projet de recherche commun pour l'imputation des données manquantes dans les observations des Diagnostics de Performance Énergétique. Cette thèse résulte de ce projet.

“Les États membres prennent les mesures nécessaires pour garantir que des exigences minimales en matière de performance énergétique des bâtiments ou des unités de bâtiment soient fixées en vue de parvenir à des niveaux optimaux en fonction des coûts. [...] Les États membres prennent les mesures nécessaires pour garantir que des exigences minimales en matière de performance énergétiques soient fixées pour les éléments de bâtiment qui font partie de l'enveloppe du bâtiment et qui ont un impact considérable sur la performance énergétique de cette enveloppe lorsqu'ils sont remplacés ou rénovés, en vue de parvenir à des niveaux optimaux en fonction des coûts.” [1]

2 Vue d'ensemble du projet de recherche

2.1 Problématique de recherche

Ce travail de recherche vise à fournir suffisamment d'informations aux décideurs pour qu'ils puissent diagnostiquer le parc immobilier d'un territoire donné, identifier les bâtiments efficaces et inefficaces sur le plan énergétique, et donc mieux cibler les mesures d'incitation. Le principal problème est que la recherche actuelle, comme nous le verrons, décrit le parc immobilier à grande échelle ou analyse des bâtiments particuliers en fonction d'une connaissance technique détaillée de leur structure. Cependant, pour atteindre les objectifs de rénovation, les institutions ont besoin de connaître l'efficacité énergétique de tous les bâtiments sur le territoire qu'elles administrent, à un coût raisonnable,

c'est-à-dire sans les visiter tous physiquement. Le défi est donc d'extraire suffisamment d'informations des données disponibles à l'échelle nationale pour compenser, au moins partiellement, le manque de connaissances techniques dans chaque bâtiment. L'objectif est de prédire, pour chaque bâtiment en France, son efficacité énergétique, en se concentrant sur la détection des bâtiments énergivores, appelés "passoires énergétiques".

2.2 Hypothèses et objectifs

Information suffisante. La première hypothèse de ce travail est que le volume d'informations disponibles sur le parc immobilier français est suffisant pour en déduire des connaissances sur l'efficacité énergétique de chaque bâtiment sans nécessiter de visite physique des installations. Les observations disponibles dans la base de données des **DPE**, résultant de visites physiques des logements, couvrent moins de 20 % du parc immobilier¹ (3.2 millions de **DPE** ont été produits en 2022 et 4.5 millions en 2023) et il n'est pas prévu, à ce stade, de rendre le **DPE** obligatoire pour tous les bâtiments. Par conséquent, cette hypothèse détermine fortement la faisabilité de la résolution du problème de recherche. Les informations disponibles sur les bâtiments comprennent une diversité de caractéristiques telles que leur âge, leur structure générale et le contexte socio-économique des occupants et des propriétaires.

Possibilité d'apprentissage. La deuxième hypothèse, qui découle de la première, consiste à supposer qu'il existe un ou plusieurs algorithmes capables d'effectuer un apprentissage supervisé pour prédire les étiquettes **DPE** de chaque bâtiment français, en utilisant des informations sur les bâtiments comme données d'entrée et la base de données des **DPE** comme observations. Cela signifie que deux bâtiments sont supposés être plus susceptibles d'avoir la même étiquette **DPE** s'ils sont similaires l'un à l'autre d'une manière qui reste à définir. Cette hypothèse est objectivée par un indicateur de performance qui est, dans notre cas, la précision équilibrée (balanced accuracy en anglais).

Les deux hypothèses ci-dessus sont remises en question si des bâtiments, qui semblent similaires du point de vue des données, ont des efficacités énergétiques très différentes. Cela peut-être dû à des facteurs non disponibles, tels qu'une rénovation, qu'il n'est pas obligatoire de déclarer. Au contraire, il est plus facile d'aborder le problème de recherche s'il suffit que deux bâtiments aient le même âge ou soient géographiquement proches l'un de l'autre pour qu'ils aient des efficacités énergétiques similaires.

Notre objectif est de prédire l'efficacité énergétique des bâtiments avec la meilleure performance possible, sur la base de ces hypothèses. En chemin, nous les reformulons, les interprétons, et finalement les validons ou les réfutons dans une certaine mesure. Cet objectif est contraint par les exigences de son application : il devrait être possible de mettre en œuvre les résultats pour rendre les prédictions disponibles dans le logiciel **U.R.B.S.** appelé **ONB** (*Observatoire National des Bâtiments – National Buildings Observatory*). En particulier, nous traitons de grands ensembles de

1. Statistiques de la base de données **IMOPE**, en comptant les adresses comme décrit dans la sous-section 2.3.

données et nous devons être en mesure d'étendre le processus conçu au niveau national.

2.3 Champ d'application et limites

Cette recherche ne concerne que les logements, à l'exclusion de tout autre type de bâtiment tels que des bureaux, des bâtiments industriels ou de stockage. Elle couvre un large territoire, dans notre cas toute la France, prenant ainsi en compte les questions de zones climatiques et d'altitude.

Le processus de fusion des données est décrit dans la partie suivante mais n'est pas abordé sous l'angle de la recherche. Il a été conçu et mis en œuvre conjointement avec l'équipe **U.R.B.S.**. Il est évident qu'il a un fort impact sur notre travail et de nombreuses améliorations ont été apportées au cours des trois années qu'a duré ce projet, afin de se concentrer sur les données pertinentes. Un point essentiel à garder à l'esprit est que nous travaillons à l'échelle de l'adresse. Cela signifie que nous identifions une adresse, telle que le "7 rue Bergson, Saint-Étienne", à un bâtiment. Ce n'est pas tout à fait vrai, car une même adresse peut être utilisée pour plusieurs bâtiments et, plus rarement, un même bâtiment peut avoir plusieurs entrées associées à plusieurs adresses. Mais lorsque cette recherche a commencé, il n'y avait pas d'ensemble de données disponible pour lever l'ambiguïté de ces situations. Ce n'est qu'à la fin de l'année 2023 que la France a commencé à dresser un inventaire officiel des bâtiments physiques. Il reste que statistiquement, selon la base de données **MoF (Ministry of Finances)**, la grande majorité (92 %) des bâtiments ont une seule adresse et la majorité (68 %) des adresses pointent vers un seul bâtiment. Il convient de noter que même si une adresse renvoie à plusieurs bâtiments, il n'y a généralement qu'un seul bâtiment avec des logements.

En raison de la nouveauté du projet, et afin de pouvoir justifier les prédictions auprès d'un client de **U.R.B.S.**, il est demandé de proposer un algorithme explicable. Pour cette raison, les algorithmes de type "boîte noire" ont été exclus de l'étude. En particulier, nous n'avons pas travaillé avec des réseaux de neurones ni avec des forêts aléatoires. Cette contrainte est discutée au Chapitre **IV**, où les performances de Random Forest sont évaluées, et au Chapitre **Conclusion et perspectives (en français)**, où les résultats donnés par le réseau de neurones **tabnet** sont présentés.

Les données présentées dans ce travail ne concernent que la France métropolitaine mais la méthodologie peut être utile pour d'autres pays, en particulier les pays européens. Notre travail a d'ailleurs suscité de l'intérêt lorsque nous l'avons présenté en France, en Belgique et aux États-Unis. Les trois articles présentés ici ont été évalués positivement, deux sont déjà publiés, le troisième est en cours d'évaluation.

Ce travail contribue aux objectifs de développement durable suivants :



11. Rendre les villes et les installations humaines inclusifs, sûrs, résilients et durables : La prédiction de l'efficacité énergétique des bâtiments facilite les programmes de rénovation.



12. Garantir des modes de consommation et de production durables : Encourager la rénovation des bâtiments énergivores permet de réduire leur consommation d'énergie et d'augmenter leur durée de vie, ce qui permet d'économiser des matériaux de construction.

2.4 Importance de ce travail

La principale conclusion de ce travail est qu'il est effectivement possible de tirer des informations sur l'efficacité énergétique d'un bâtiment à partir des bases de données disponibles. Cette recherche montre également qu'il est possible de considérer les **DPE** comme des données géolocalisées, ce qui signifie qu'en plus de comparer les descriptions de deux bâtiments, il est également intéressant de savoir qu'ils sont situés à proximité l'un de l'autre afin d'améliorer la prédiction de l'efficacité énergétique. À notre connaissance, il s'agit d'une nouveauté. Elle suggère qu'au lieu de conserver une approche par bâtiment avec les mêmes caractéristiques pour encourager la rénovation, il peut être intéressant de considérer l'inefficacité énergétique par quartier. Cela n'est pas surprenant pour les personnes familières des questions de développement communautaire, mais c'est certainement surprenant pour les institutions qui examinent habituellement le parc immobilier d'un point de vue très technique. Notre travail suggère également qu'il est plus pertinent de prédire une étiquette **DPE** plutôt qu'une consommation d'énergie standardisée. Cela confirme a posteriori que la nature du **DPE** va au-delà de l'ingénierie thermique. En ce qui concerne la détection des passoires énergétiques, ce travail montre que les passoires énergétiques (étiquettes **DPE F** et **G**) sont plus difficiles à prédire que les autres étiquettes A à E. Pour surmonter cette difficulté, il est avantageux d'envisager une approche de classification floue. En fonction des modèles et de notre niveau d'acceptation des faux positifs, il est désormais possible de détecter 36 % à 66 % des passoires énergétiques sans visiter physiquement les bâtiments, alors que nous ne pouvions pas dépasser 23 % avant le début de ce projet de recherche, et alors qu'aucun des autres laboratoires travaillant sur ce sujet en France n'a encore publié de résultats quantitatifs.

3 Méthodologie et structure générale du document

Pour construire une méthodologie, nous suivons deux lignes directrices. La première est que, d'un point de vue industriel, l'un des principaux attendus de ce projet est l'identification des passoires énergétiques. Par conséquent, au-delà de la connaissance commune des facteurs déterminants pour l'efficacité énergétique, tels que l'âge d'un bâtiment, il est intéressant d'identifier les facteurs qui sont en corrélation avec l'inefficacité énergétique. Nous émettons l'hypothèse que certains facteurs socio-économiques pour-

raient être corrélés de manière significative avec l'inefficacité énergétique. La deuxième ligne directrice est que nous traitons des données géolocalisées. Nous formulons donc une deuxième hypothèse de recherche, selon laquelle il est bénéfique pour notre travail de traiter les **DPE** comme des données géolocalisées au lieu de s'en tenir à l'approche traditionnelle de l'ingénierie thermique.

Le thème central de ce travail est le **DPE**. La définition et la perception de ce qu'est un **DPE** varient selon les acteurs et évoluent dans le temps. D'un objet d'ingénierie thermique, il est devenu une question politique, modifié à plusieurs reprises par l'organe législatif, et il est maintenant accessible en tant que base de données [5]. Afin de trouver un terrain d'entente pour la discussion, dans le cadre de cette thèse, le **DPE** est considéré comme une classification légale des logements et des bâtiments. Cela ne doit pas faire oublier l'ingénierie thermique utilisée pour établir un **DPE**, mais compte tenu de la variété des acteurs qui participent à sa définition, il est beaucoup trop restrictif de le réduire au résultat d'un modèle d'ingénierie. Le **DPE** est une conséquence légale des objectifs de durabilité et, en particulier, de la nécessité d'intensifier l'effort de rénovation. Les multiples programmes nationaux de rénovation ont en commun de ne pas atteindre leurs objectifs. Il est probable qu'une meilleure connaissance du parc immobilier permettrait d'atteindre les objectifs nationaux. La définition du **DPE** et l'état de l'art dans la production de connaissances à grande échelle sur l'efficacité énergétique d'un parc immobilier sont présentés dans les sections suivantes 4 et 5.

Les données rassemblant des informations sur les logements pâtissent de l'incertitude de la géolocalisation. Cet aspect semble être un problème important à résoudre pour pouvoir intégrer les données dans un modèle géostatistique. C'est pourquoi le premier article présenté dans le Chapitre II introduit un modèle géostatistique qui prend nativement en compte les incertitudes de position et s'adapte aux données multi-échelles. Il s'agit d'un modèle de régression qui prédit la consommation d'énergie standardisée des bâtiments en fonction de leur latitude, de leur longitude, de leur âge et éventuellement d'autres variables. Bien que ce modèle soit prometteur à l'échelle d'une ville, il s'avère coûteux à étendre à l'échelle nationale.

S'il est difficile de prendre en compte l'incertitude de la position dans les données d'entrée, il est possible de l'inclure dans les données de sortie, c'est-à-dire dans les valeurs prédites. En suivant cette voie, nous abandonnons l'approche de régression pour nous tourner vers la classification, c'est-à-dire pour prédire l'étiquette **DPE** et travailler sur la classification floue. La classification floue permet non seulement de prédire une classe, mais aussi de donner des informations sur sa probabilité et celle de l'autre classe. Le deuxième article, au Chapitre III, présente un modèle géostatistique de classification appelé Joint Kriging. Il est plus léger que le Mixture Kriging et, par conséquent, en plus de ses bonnes performances, il est moins difficile à mettre à l'échelle.

Parallèlement à cette recherche à mi-parcours, il était nécessaire, en arrière-plan, de développer un modèle à mettre en production à **U.R.B.S.**. Ce modèle est également un modèle de classification, pas spécifiquement pour les données géographiques, et il inclut

de multiples variables socio-économiques. Il est basé sur un algorithme **FKNN** (**Fuzzy k-Nearest Neighbours**) optimisé avec une pseudo-descente de gradient de type **SPSA** (**Simultaneous Perturbation Stochastic Approximation**). Le troisième article, dans le Chapitre **IV**, présenté dans cette thèse, compare les performances de Joint Kriging, de cet algorithme **FKNN**, et de Random Forest.

Aperçu du projet de recherche : L'essentiel

Problème industriel. Acquérir une connaissance fine de l'efficacité énergétique des bâtiments.

Problématique de recherche. Apprendre à partir des **DPE** observés pour prédire l'efficacité énergétique des bâtiments non observés.

Contraintes. Le modèle doit être explicable. Bien que l'objectif principal soit de prédire le **DPE** au niveau du bâtiment, la distribution globale prédite doit aussi correspondre autant que possible à la distribution globale estimée.

Hypothèses de recherche Les informations disponibles pour l'ensemble du parc immobilier sont suffisantes pour calculer l'efficacité énergétique sans visite du site. Et nous pouvons trouver un algorithme qui apprend à partir des **DPE** existants et prédit les **DPE** des bâtiments non observés.

Champ d'application et limites Nous nous intéressons uniquement au parc de logements français. Nous cherchons un modèle explicable.

4 État des lieux

4.1 Qu'est-ce que le diagnostic de performance énergétique ?

Définition légale d'un DPE

Le **DPE** (*Diagnostic de Performance Énergétique – Energy Performance Certificate*) d'un bâtiment ou d'une partie de bâtiment est un document qui comporte la quantité d'énergie effectivement consommée ou estimée, exprimée en énergie primaire et finale, ainsi que les émissions de gaz à effet de serre induites, pour une utilisation standardisée du bâtiment ou d'une partie de bâtiment et une classification en fonction de valeurs de référence permettant de comparer et évaluer sa performance énergétique et sa performance en matière d'émissions de gaz à effet de serre. Il comporte une information sur les conditions d'aération ou de ventilation. Il est accompagné de recommandations destinées à améliorer ces performances et du montant des dépenses théoriques de l'ensemble des usages énumérés dans le diagnostic.

Il est établi par une personne répondant aux conditions prévues par l'article L. 271-6.

Sa durée de validité est fixée par voie réglementaire.

Article L126-26 du Code de la construction et de l'habitation [6]

En France, le **DPE** (*Diagnostic de Performance Énergétique – Energy Performance Certificate*) est réalisé par un diagnostiqueur certifié pour ce travail. Il ou elle recueille des informations qualitatives et quantitatives sur un bâtiment ou un logement et les saisit dans un logiciel qui a été approuvé par l'ADEME (*Agence Française pour la Transition Écologique – French Agency For Ecological Transition*). Sur la base de ces données d'entrée, en prenant des valeurs par défaut pour les données manquantes, le logiciel produit deux chiffres représentant une consommation d'énergie annuelle normalisée par mètre carré et par an et un poids normalisé de dioxyde de carbone émis, CO_2 , par mètre carré et par an. Une combinaison de deux seuils permet d'obtenir une étiquette énergétique, un niveau d'émission de **GES** (*Gaz à effet de serre*) et une étiquette **DPE** globale.

Le lecteur peut d'ores et déjà constater l'importance du contrôle exercé par le gouvernement sur ce système. En effet, la loi française définit non seulement le principe de la classification des bâtiments, mais aussi l'algorithme qui calcule la consommation d'énergie et les émissions de **GES** et elle définit les seuils qui déterminent les étiquettes. Ces étiquettes, à leur tour, sont utilisées pour définir des droits et des devoirs légaux, tels que les normes de construction pour les nouveaux bâtiments ou l'autorisation de louer un logement. L'écosystème législatif en la matière a été réécrit en 2021 à la lumière de nouvelles connaissances scientifiques, de nouveaux objectifs européens et d'un nouvel agenda politique. Nous décrivons ci-après l'état de la législation après 2021, en mentionnant les changements importants par rapport au système précédent. En particulier, la partie **GES** de la définition a été ajoutée en 2021, bien que les informations sur les émissions de **GES** apparaissaient déjà dans les **DPE** établis avant 2021.

Examinons plus en détail la définition juridique d'un **DPE** (voir l'encadré ci-dessus) :

— **“Bâtiment ou partie de bâtiment”**. La même échelle est utilisée pour un lo-

gement spécifique dans un immeuble et pour l'ensemble du bâtiment. Toutefois, la méthode n'est pas la même. Pour un immeuble, le diagnostiqueur visite un échantillon représentatif de bâtiments pour en déduire le **DPE** des autres. Et la consommation d'énergie d'un immeuble n'est pas égale à la somme des consommations des logements car il y a des zones dans un bâtiment qui n'appartiennent à aucun appartement, comme l'escalier, l'ascenseur, éventuellement, ou le hall d'entrée. Notre objectif est de prédire **DPE** au niveau du bâtiment uniquement. Par conséquent, un algorithme d'apprentissage doit trouver un moyen d'estimer **DPE** d'un bâtiment sur la base des observations des logements.

- **“Énergie primaire et finale”, “émissions de GES”**. En fait, l'étiquette **DPE** est déterminée par la consommation d'énergie primaire. Elle est calculée en multipliant l'énergie finale estimée par un facteur de conversion, qui est de 2.3 pour l'électricité (2.58 avant 2021) et de 1 pour les autres sources d'énergie. De même, un facteur de conversion est appliqué pour calculer les émissions de **GES** : 0.079 pour l'électricité, 0.227 pour le gaz de ville, 0.324 pour le fioul domestique, alors qu'elles étaient respectivement de 0.180, 0.234, et 0.300 avant 2021 [7]. Ces facteurs sont en partie politiques : La réduction des facteurs de conversion pour l'électricité vient d'un gouvernement favorable à l'énergie nucléaire, qui lance de nombreux projets de construction de centrales nucléaires.
- **“Utilisation standardisée du bâtiment”**. La loi fait des hypothèses sur le nombre d'occupants d'un logement, la température de consigne du chauffage ou l'utilisation de l'eau sanitaire. Le **DPE** ne se préoccupe pas de la consommation réelle d'énergie. En pratique, cela signifie que les personnes aisées, qui ont tendance à consommer plus d'énergie, sont susceptibles de consommer plus que le standard **DPE**, et que les personnes aux revenus modestes sont susceptibles de consommer moins d'énergie que le standard. Pour plus de détails sur la pauvreté énergétique, voir [8].
- **“Une classification”**. L'objectif principal du **DPE**, comme le suggèrent les règles de l'Union Européenne, est de classer les bâtiments, de les comparer entre eux, de pouvoir détecter les bâtiments énergivores et de les rénover.

Par conséquent, le **DPE** essaie de servir deux objectifs, informer et classer, mais ils ne sont pas totalement compatibles. Il ressort clairement des remarques ci-dessus que l'aspect classification a été privilégié pour induire ce qui est perçu par le gouvernement comme un comportement transitoire, tel que le remplacement d'un système de chauffage au gaz de ville par un système électrique. Il n'y a pas lieu ici de discuter du pour et du contre de ces faits, mais il est plutôt important de garder à l'esprit qu'il n'y a qu'une faible corrélation entre la consommation d'énergie réelle et la consommation d'énergie mesurée par le **DPE**. Cela est régulièrement souligné par de nombreux acteurs [9].

Enfin, il convient de mentionner un autre biais important du **DPE** : la classification est basée sur la consommation d'énergie par an et par mètre carré de surface habitable. Or, la consommation d'énergie est fortement liée de manière linéaire à la surface de

l'enveloppe du bâtiment. Par conséquent, la consommation d'énergie par mètre carré est proportionnelle au quotient de la surface de l'enveloppe divisée par la surface habitable, qui est un indicateur de compacité. Plus elle est petite, plus le bâtiment est compact. Les cubes sont très compacts, ce qui donne un indicateur de compacité de $1/8$. Mais si un bâtiment ou un logement a un plafond bas avec des pièces disposées en ligne, l'indicateur de compacité tend vers $1/2$. Par conséquent, selon la façon dont les logements sont disposés dans un immeuble, l'ensemble du bâtiment peut avoir une bonne efficacité énergétique telle qu'évaluée par l'indicateur **DPE**, alors que certains de ses logements peuvent être mesurés comme des passoires énergétiques. Cet effet est empiré si le logement se trouve dans un coin du bâtiment ou au sommet de celui-ci. De plus, pour les très petits logements, la partie chauffage de la consommation d'énergie est considérablement réduite par rapport à la consommation d'eau chaude, ce qui tend à mettre en évidence des passoires énergétiques même si le logement est économe en énergie d'un point de vue thermique.

Malgré ces limites, le **DPE** reste un indicateur important. C'est le seul indicateur qui résulte d'une étude approfondie du comportement énergétique d'un bâtiment. Les diagnostiqueurs collectent un grand nombre d'indicateurs qui peuvent aider à comprendre la structure du parc immobilier. Et c'est l'indicateur de base pour l'élaboration des réglementations thermiques. C'est pourquoi un **DPE** est obligatoire dans les cas suivants :

- Pour la livraison d'un nouveau bâtiment ou l'extension d'un bâtiment existant. Il est remis à l'acheteur.
- Pour la vente ou la location d'un bâtiment ou d'une partie de bâtiment. Il est remis à l'acheteur ou au locataire potentiel et annexé au contrat de bail. Le fait de ne pas informer l'acheteur ou le locataire est punissable.
- Pour tout immeuble construit avant 2013.

Toutefois, certains bâtiments sont exclus de cette obligation : les maisons dont la surface est inférieure à 50 m^2 , les bâtiments temporaires qui doivent rester moins de deux ans, les bâtiments du patrimoine national, les parties d'un bâtiment qui n'ont pas de système de chauffage ou qui sont occupées moins de 4 mois par an.

Ainsi, des millions de **DPE** sont collectés chaque année. Près de 8 millions de **DPE** ont été collectés entre juin 2021 et décembre 2023 pour des maisons, des appartements ou des immeubles. Mais cet ensemble d'observations n'est pas représentatif du parc immobilier français car certains bâtiments ont fait l'objet de plusieurs diagnostics alors que d'autres n'ont jamais été visités par un diagnostiqueur. En particulier, les maisons sont sous-représentées car les personnes qui possèdent une maison ont tendance à la garder très longtemps (voir Figure 1). Sur l'ensemble du parc immobilier, les maisons constituent 90 % des bâtiments, alors qu'elles ne représentent que 69 % des observations. Par conséquent, les immeubles sont sur-représentés dans les observations.

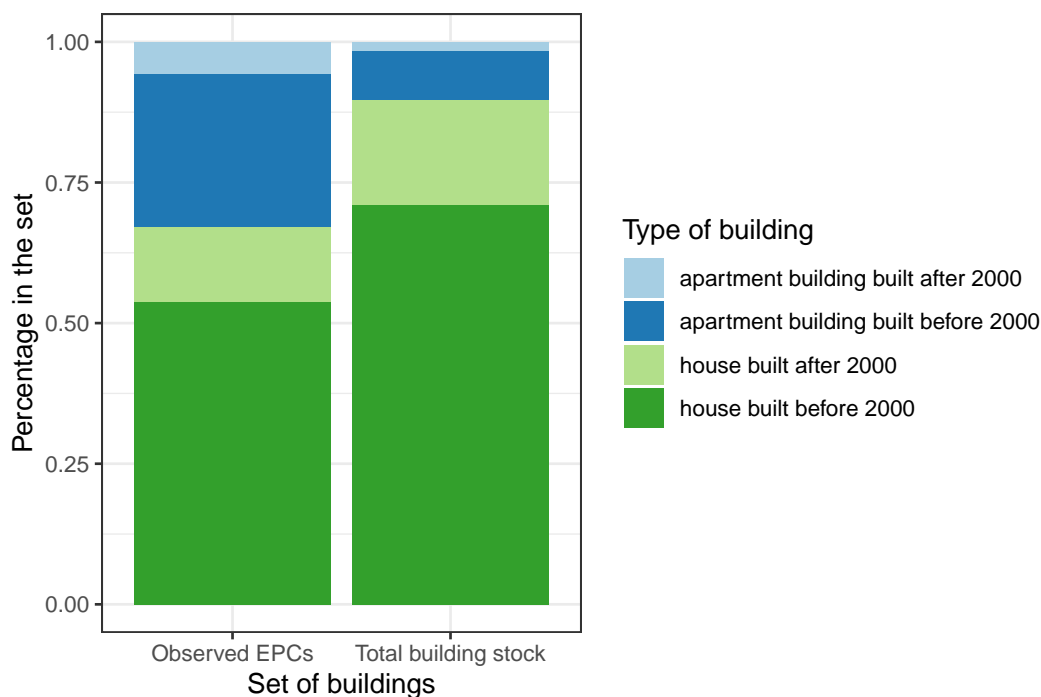


Figure 1 – Comparaison entre le parc immobilier des **DPE** observés et le parc immobilier total en France : proportions des populations construites avant ou après 2000, qui sont des maisons ou des appartements.

4.2 À quoi ressemble un DPE ?

La consommation d'énergie primaire est exprimée en kilowattheures par mètre carré de surface habitable et par an ($\text{kWh}/\text{m}^2/\text{an}$), voir la Figure 2. En particulier, il convient de noter que si la consommation d'énergie dépend linéairement de la surface de l'enveloppe du logement, elle est considérée ici comme une quantité par unité de surface de la zone d'habitation. Cela pose des problèmes de compacité, comme indiqué dans la sous-section 4.3.

Les émissions de **GES** sont exprimées en kilogrammes d'équivalent CO_2 par mètre carré de surface habitable et par an ($\text{kg}_{\text{CO}_2}/\text{m}^2/\text{an}$). Cette unité pose également quelques problèmes de compacité. La conception de cette unité consiste à convertir le poids de tout **GES** en un poids de CO_2 qui aurait le même effet de serre. Son principal avantage est de produire un résultat lisible pour l'utilisateur final, mais son inconvénient est qu'il masque la diversité des **GES** émis. Certains d'entre eux peuvent avoir des effets négatifs autres que l'effet de serre.

Comme le montre la Figure 3, deux séries de seuils définissent une étiquette d'énergie primaire et un seuil d'émissions de **GES**, tous deux compris entre A et G. L'étiquette **DPE** est déterminée par la couleur du point, dont les coordonnées sont les émissions **GES** et la consommation d'énergie primaire. La principale conséquence de ce système est que l'étiquette **DPE** est la plus mauvaise des deux sous-étiquettes. Par exemple, si un logement a des émissions **GES** égales à $80 \text{ kg}_{\text{CO}_2}/\text{m}^2/\text{an}$ et une consommation d'énergie

primaire égale à 300 kWh/m²/an, alors son étiquette **GES** est F et son étiquette énergie est E, donc son étiquette **DPE** est la plus mauvaise de F et E, c'est à dire F.

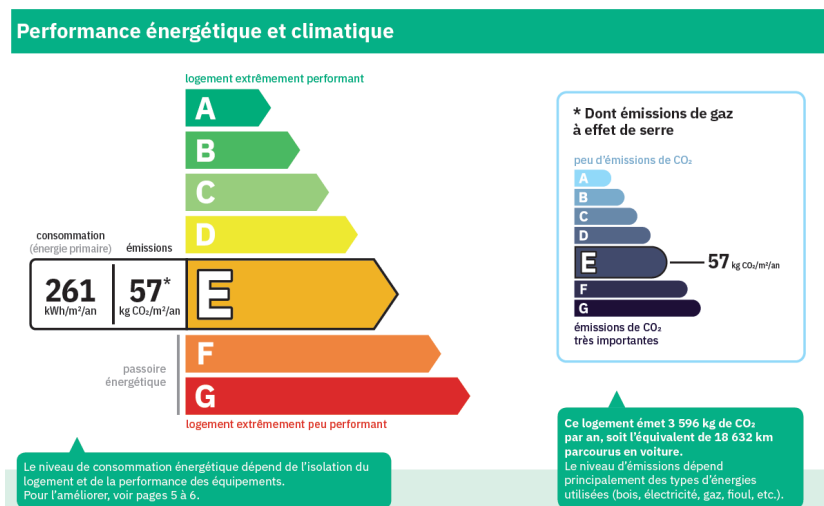


Figure 2 – L'étiquette qui donne le résultat d'un diagnostic de performance énergétique.

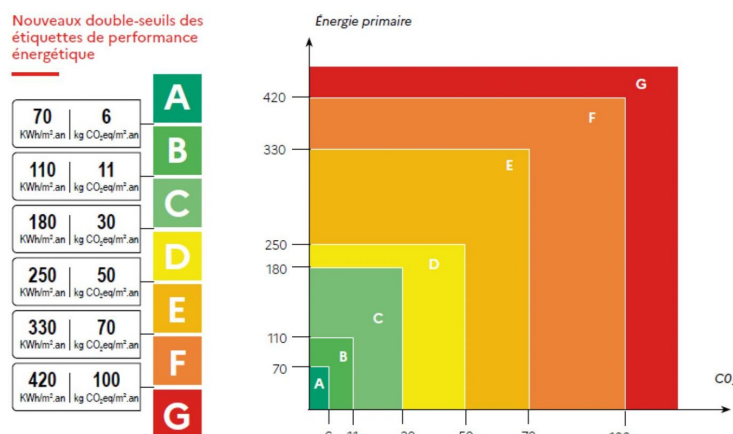


Figure 3 – Le processus de double seuil utilisant la consommation d'énergie primaire et les émissions de **GES** pour le calcul de l'étiquette **DPE** en France depuis 2021.

Les résultats de l'étude technique apparaissent dans une vignette présentée dans la Figure 2. Les résultats qualitatifs (étiquettes) et quantitatifs (consommation/émissions) sont présentés. Mais d'après notre expérience, la plupart des gens ne s'intéressent qu'à l'aspect qualitatif. C'est logique, car le cadre juridique dépend des étiquettes. Mais cela peut masquer la quantité d'efforts nécessaires pour améliorer le **DPE** d'une seule étiquette.

4.3 Objectifs de rénovation

Au 21^{ème} siècle, le premier jalon de l’engagement français en faveur du développement durable a été le “Grenelle de l’Environnement”, qui s’est déroulé de septembre à décembre 2007. Les conclusions de cette réunion ont ouvert la voie à deux séries de lois connues sous le nom de lois Grenelle I et Grenelle II. La première, publiée en 2009, place la France parmi les pays européens les plus ambitieux. L’objectif est de diviser par quatre les émissions de **GES** en 2050 par rapport à 1990 ; voir l’encadré ci-dessous. En ce qui concerne les bâtiments, le parc immobilier existant doit réduire sa consommation d’énergie d’au moins 38 % avant 2020, et un nombre minimal de rénovations annuelles est fixé. Un diagnostic de tous les bâtiments publics est prescrit, avec des objectifs spécifiques de réduction de la consommation d’énergie. Des objectifs encore plus spécifiques sont fixés pour les opérateurs de logements sociaux, pour lesquels des zones prioritaires sont définies. Des incitations financières sont également prévues pour les propriétaires, les banques et les compagnies d’assurance.

“Article 2: La lutte contre le changement climatique est placée au premier rang des priorités. Dans cette perspective, est confirmé l’engagement pris par la France de diviser par quatre ses émissions de gaz à effet de serre entre 1990 et 2050 en réduisant de 3 % par an, en moyenne, les rejets de gaz à effet de serre dans l’atmosphère, afin de ramener à cette échéance ses émissions annuelles de gaz à effet de serre à un niveau inférieur à 140 millions de tonnes équivalent de dioxyde de carbone (140 Mt_{CO₂}).

Article 5: L’État se fixe comme objectif de réduire les consommations d’énergie du parc des bâtiments existants d’au moins 38 % d’ici à 2020. A cette fin, l’État se fixe comme objectif la rénovation complète de 400 000 logements chaque année à compter de 2013.”

Loi Grenelle I, 2009-967, publiée le 3 août 2009 [10].

La loi Grenelle II a été publiée en 2010. Elle a pour objectif d’organiser la mise en œuvre des engagements pris dans le cadre de la loi Grenelle I. Elle introduit la notion de pauvreté énergétique et précise que les efforts de réduction de la consommation d’énergie doivent bénéficier aux ménages en situation de pauvreté énergétique ; voir l’encadré ci-contre. Elle redéfinit le **DPE** pour y inclure les émissions de **GES** et rend le **DPE** obligatoire en cas de vente ou de mise en location du bâtiment. Un diagnostic doit être réalisé dans tous les immeubles équipés de systèmes de chauffage collectif avant 2017. Et tous les **DPE** sont collectés dans une base de données commune.

Est en situation de précarité énergétique au titre de la présente loi une personne qui éprouve dans son logement des difficultés particulières à disposer de la fourniture d’énergie nécessaire à la satisfaction de ses besoins élémentaires en raison de l’inadaptation de ses ressources ou de ses conditions d’habitat.

Loi Grenelle II, 2010-788, publiée le 12 juillet 2010 [11].

Ces lois ont été modifiées presque chaque année depuis leur première publication. Bien qu'il y ait sans aucun doute une inflexion dans l'évolution historique de notre société, il est également vrai que l'un des objectifs de la rénovation a été atteint. La consommation d'énergie des bâtiments de bureaux a augmenté de plus de 29 % entre 1990 et 2019 [12], et les bâtiments résidentiels ont augmenté leur consommation d'énergie de plus de 10 % entre 1990 et 2021, avec un maximum en 2013, lorsque la consommation était 20 % plus élevée qu'en 1990 (données du ministère de la transition écologique). Toutefois, le pays a réduit ses émissions de GES de 371 Mt_{CO₂} en 1990 à 252 Mt_{CO₂} en 2020, y compris l'impact de COVID-19 – il était de 291 Mt_{CO₂} en 2019. Ceci est principalement dû à des efforts dans le secteur de la production d'énergie et des processus industriels et résulte de choix controversés tels que le développement de l'électricité nucléaire.

En 2021, l'ensemble de l'approche du Grenelle a été redéfinie dans une loi appelée *Loi portant lutte contre le dérèglement climatique et renforcement de la résilience face à ses effets* [13]. Les conséquences de cette loi sont résumées dans la Figure 4, page 16. Son principal impact sur les bâtiments est la limitation de l'extension des villes sur les terres agricoles ou naturelles. Cette loi redéfinit le DPE tel que décrit dans la sous-section 4.1. Elle définit des termes tels que la rénovation performante, qui est une rénovation qui amène le bâtiment à une étiquette A ou B, ou la rénovation globale, qui est une rénovation performante en moins de dix-huit mois. Au lieu de se concentrer sur les stratégies de rénovation énergétique, elle introduit des mesures coercitives à l'encontre des propriétaires de bâtiments. En particulier, elle stipule que les logements portant l'étiquette DPE G ne pourront plus être loués à partir de 2025. Il en va de même pour les étiquettes F en 2028 et pour les étiquettes E en 2034. Parallèlement, le modèle thermique qui calcule le DPE a été révisé. Les agents immobiliers, tels que la puissante FNAIM (*Fédération Nationale de l'Immobilier – National Federation of Real Estate*), la plus grande fédération française de l'immobilier, ont rapidement compris que cela allait créer une tempête dans le secteur de la location, en particulier dans les grandes villes comme Paris où les petits bâtiments anciens et énergivores sont faciles à louer, même non rénovés. Les lobbyistes ont réagi en exerçant une forte pression sur le gouvernement. Les fonctionnaires ont déclaré que le nouveau modèle thermique était défectueux et ont ajusté le modèle pour les bâtiments construits avant 1975 afin de produire moins de passoires énergétiques. Une autre préoccupation a été soulevée concernant les petits logements où la consommation d'eau chaude sanitaire devient relativement importante dans la consommation globale d'énergie par rapport aux besoins de chauffage, introduisant un biais dans le DPE. Par conséquent, une autre modification du modèle a également été introduite en 2024 pour minimiser cet effet sur les appartements dont la surface est inférieure à 40 m² par le biais d'une modification de l'indicateur de compacité.



Figure 4 – Un résumé visuel de la loi portant lutte contre le dérèglement climatique.

5 État de l'art

Dans cette section, nous présentons une revue de la littérature scientifique concernant deux aspects importants de la présente recherche. Le premier est l'évaluation de la performance énergétique d'un parc immobilier. Le second concerne l'interpolation de données géolocalisées. Notre travail se situe à la croisée de ces traditions de recherche, qui ont vu le jour principalement au 19^{ème} siècle. L'étude de la performance énergétique a été rendue possible par les travaux de Joule et Kelvin, entre autres en thermodynamique, et a été rendue nécessaire par la révolution industrielle, qui a nécessité de grandes quantités d'énergie. L'interpolation de données géolocalisées est une branche des statistiques spatiales qui est un champ de recherche ancien, notamment en épidémiologie, mais qui a pris un nouveau tournant en bénéficiant des avancées dans la connaissance des maladies contagieuses par des chercheurs comme Pasteur. John Snow, par exemple, est célèbre pour avoir identifié un puits de pompage contaminé responsable de l'épidémie de choléra de 1854 à Soho grâce à une approche géostatistique ; les travaux de J. Snow sont présentés par le Lancet dans [14]. L'interpolation spatiale elle-même a grandement bénéficié des travaux de Danie G. Krige en tant qu'ingénieur des mines.

La bibliographie ci-dessous est volontairement limitée aux aspects qui présentent un intérêt pour les trois articles présentés dans les Chapitres II, III, et IV. Chacun de ces chapitres contient également une revue spécifique de la littérature scientifique.

5.1 Évaluation de la performance énergétique d'un parc immobilier

Dans la littérature scientifique, l'évaluation de la performance énergétique d'un parc immobilier peut être abordée d'un point de vue technique, d'un point de vue de la gestion des données qui inclut la métamodélisation, et d'un point de vue géostatistique, ce qui est le choix de ce travail.

Modèles physiques et typologies de bâtiments. D'un point de vue technique, les ingénieurs thermiciens disposent de modèles physiques qui calculent un bilan énergétique afin de déterminer la consommation d'énergie d'un bâtiment donné. C'est l'approche traditionnelle, et le **DPE** est calculé par cette méthode. Les détails du modèle physique, appelé modèle 3CL, sont disponibles dans [15]. Pour travailler à plus grande échelle, les ingénieurs thermiciens définissent des typologies de bâtiments, calculent une distribution de ces types sur un territoire donné, et en déduisent donc une distribution des étiquettes **DPE** ou des consommations d'énergie. Cette approche s'est avérée efficace [16]. Cependant, le manque de connaissance des caractéristiques techniques détaillées de chaque bâtiment constitue une limitation forte pour une prédiction au niveau du bâtiment. Cela signifie que la consommation d'énergie totale d'une ville peut être correctement prédite, alors que la consommation d'énergie d'un bâtiment particulier dans cette ville est mal prédite. En fait, la consommation d'énergie au niveau du bâtiment peut même ne pas être prédite du tout lorsque l'on dispose de la connaissance d'un échantillon représentatif. C'est le cas, par exemple, de l'étude nationale française dite *Enquête TREMI* [17]. Ces modèles nécessitent de nombreux paramètres qu'il est pra-

tiquement impossible de collecter pour chaque bâtiment d'un territoire. Certains efforts de réduction des caractéristiques ont été réalisés [18] mais les caractéristiques restantes sont toujours problématiques à déduire et nécessitent des efforts supplémentaires [19], ce qui soulève des problèmes de propagation de l'incertitude.

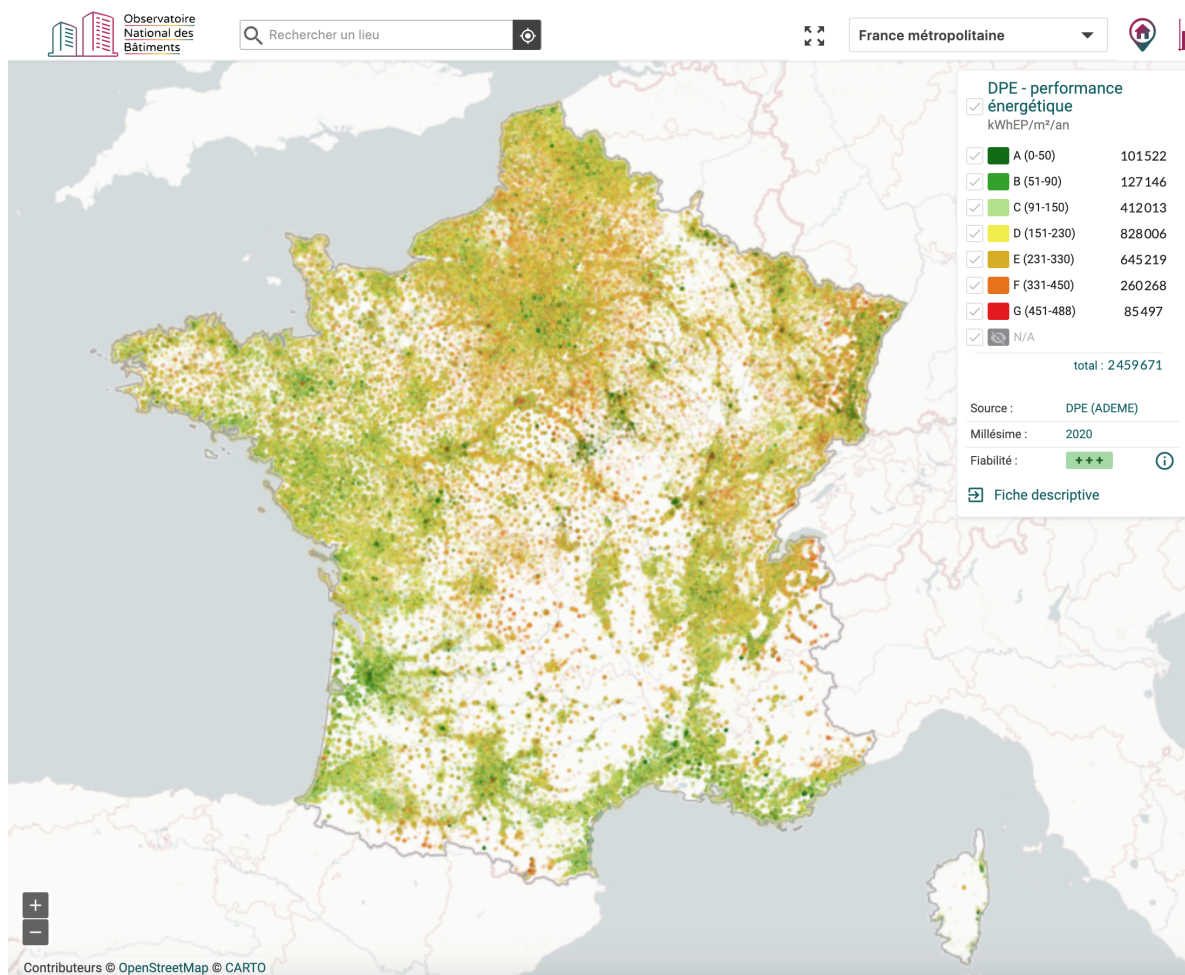


Figure 5 – Carte de France des DPE répertoriés. Cette image est une capture d'écran de l'ONB (*Observatoire National des Bâtiments – National Buildings Observatory*), diffusée avec le consentement de U.R.B.S., détenteur des droits.

Fusion de données. Du point de vue de la gestion des données, la prédiction du DPE nécessite un processus qui combine l'ensemble de données provenant de sources multiples et disponibles à des échelles multiples. Ce processus est connu sous le nom de fusion de données (data fusion en anglais) [20]. Il devient de plus en plus complexe en raison de la quantité croissante de données disponibles, qu'elles soient écologiques, sociales ou institutionnelles. Ces ensembles de données se rapportent à des unités spatiales de formes, de dimensions et de cardinalité variées. Et dans certains cas, il peut être difficile de déterminer la position exacte d'un objet observé. C'est le cas des bâtiments, car de nombreux gouvernements ne disposent pas d'une carte détaillée du parc immobilier de leur pays. L'impôt foncier est généralement basé sur des facteurs intrinsèques tels que

la surface et le nombre de chambres, mais pas sur des facteurs extrinsèques tels que le nombre d'étages ou l'orientation des fenêtres. En raison de cette incertitude, les études à grande échelle sur le parc immobilier doivent s'appuyer sur un concept abstrait de logement. Ce concept de logement peut se référer à une maison ou à un appartement ; il n'est pas clairement délimité mais il est décrit par un ensemble de caractéristiques telles qu'une surface ou un nombre de chambres à coucher. Ces caractéristiques sont rassemblées dans un tableau avec un logement par ligne, ce qui signifie que le logement est la plus petite unité d'information. Le processus de fusion des données, ainsi que le pré-traitement des données, sont présentés au Chapitre I.

Métamodélisation. De même, la plus petite unité d'information d'un tableau comportant un **DPE** par ligne est une partie d'un bâtiment. Elle n'est pas clairement définie comme un objet dans un espace tridimensionnel, mais elle possède des caractéristiques qui la décrivent. Et pour prédire le **DPE** des bâtiments, il faut également définir les bâtiments. La fusion de données nécessite de définir, de la même manière, les plus petites unités d'information, également appelées granules, pour chaque ensemble de données. "De façon informelle, un granule d'une variable X est un groupe de valeurs de X qui sont liées par l'indiscernabilité, l'équivalence, la similarité, la proximité ou la fonctionnalité. Par exemple, un intervalle est une granule." [21]. Le domaine d'étude qui se concentre sur la représentation, la construction et le traitement de ces granules d'information s'appelle le calcul granulaire (Granular Computing) [22]. En supposant que les logements, les observations des **DPE** et les bâtiments complets soient représentés dans le même modèle de données, ce qui signifie qu'un processus approprié de fusion des données est mis en œuvre, un modèle prédictif pertinent, que nous pouvons appeler, dans ce cas, un métamodèle, doit maintenant être construit. Le calcul granulaire est pluridisciplinaire, mais comme nous traitons d'informations géolocalisées, le champ de recherche naturel est la géostatistique, qui a été définie comme "traitant des processus spatiaux indexés sur un espace continu" [23, p. 7].

Géostatistiques. Probablement pour des raisons historiques, la performance énergétique étant généralement un problème pour les ingénieurs du bâtiment, le **DPE** n'est pas traité comme une information géolocalisée dans la littérature. La nouveauté de ce travail est que nous prenons en compte l'aspect géostatistique à chaque étape de notre recherche. Nous essayons de nous affranchir de la question importante des connaissances techniques détaillées à grande échelle. Au lieu de cela, nous privilégions une approche géostatistique pour bénéficier de la nature géolocalisée des informations **DPE**. Cet aspect peut être visualisé dans la Figure 5, qui présente les **DPE** observés sur une carte de France. Il est visible que les **DPE** sont meilleurs à l'ouest qu'à l'est, au sud qu'au nord, et dans les villes qu'à la campagne. Des phénomènes similaires apparaissent à l'échelle de la ville, par exemple. Du point de vue de la géostatistique, l'incertitude irréductible de la position des granules (logements, bâtiments, etc.) dans l'espace sous-jacent influe sur l'utilisation des modèles traditionnels d'interpolation spatiale tels que le krigeage. Dans le Chapitre II, nous proposons un moyen de surmonter cette dernière limitation et

de développer un cadre complet capable de traiter des données avec une incertitude sur la position des objets observés tout en permettant la définition d'un prédicteur linéaire optimal pour l'interpolation spatiale des valeurs numériques des **DPE**. Un autre problème est que le **DPE** est catégoriel alors que, par définition, l'interpolation spatiale est davantage développée pour les variables quantitatives. C'est l'objet du Chapitre III, qui présente une approche de krigeage pour la classification floue. Les performances de ce dernier modèle sont évaluées et comparées à celles d'autres modèles dans le Chapitre IV.

5.2 Interpolation spatiale

Selon la définition de [24], l'interpolation spatiale “est une technique qui utilise des valeurs d'échantillon de points géographiques connus, ou d'unités spatiales², pour estimer (ou prédire) des valeurs en d'autres points (ou unités spatiales) inconnus”. Le même article présente un tableau récapitulatif des principales méthodes d'interpolation spatiale. L'interpolation spatiale repose sur une hypothèse fondamentale qui est similaire à de nombreux modèles de prédiction : Pour une variable aléatoire donnée définie en chaque point d'un territoire, c'est-à-dire un champ aléatoire, on suppose que plus deux points du territoire sont proches l'un de l'autre, plus les variables qui leur sont associées sont similaires. Par exemple, on suppose que la température a plus de chances d'être similaire entre deux points séparés par une distance de 1m qu'entre deux points séparés par une distance de 1km. En termes de statistiques, cela signifie que la corrélation entre deux points augmente à mesure que la distance entre ces points diminue.

Une façon simple mais efficace d'aborder le problème consiste à considérer que la dépendance spatiale est limitée à des unités spatiales indépendantes les unes des autres. Cela conduit à la carte chloroplèthe, où une statistique locale telle que la médiane ou la moyenne est associée à chaque unité géographique. La première carte chloroplèthe produite est présentée dans la Figure 6 page 24. Cette approche est toujours très utile. Les images satellites, par exemple, sont des grilles de pixels dont la surface est de quelques mètres carrés ou centimètres carrés et dont la valeur est la valeur moyenne d'une certaine intensité de longueur d'onde lumineuse. Il est intéressant de noter que la prédiction d'un **DPE** au niveau du bâtiment que nous poursuivons dans cette recherche peut conduire à la production d'une variété de cartes chloroplèthes. C'est le cas, par exemple, de la carte présentée dans la Figure 7 page 25.

Si l'on suppose que la variable aléatoire observée est continue, il serait logique de prédire une quantité continue, en tenant compte des valeurs agrégées observées (les moyennes, par exemple). Cette contrainte a été appelée contrainte pycnophylactique par Tobler en 1979 [26]. Diverses solutions ont été proposées pour résoudre ce problème, comme le lissage par noyau. Cela n'entre pas dans le cadre de ce travail, puisque nous traitons d'observations individuelles plutôt que d'observations de valeurs agrégées. Cependant, il est utile de garder à l'esprit qu'il est difficile de prédire des valeurs

2. Faute de terminologie adaptée en français, nous avons traduit “areal unit” par “unité spatiale”. Pour être précis, on entend généralement par “areal unit” une unité spatiale non réduite à un point.

ponctuelles sous contrainte de valeurs agrégées. C'est ce que nous présentons au Chapitre III.

La régression par processus gaussien [27], également connue sous le nom de krigeage, est devenue un domaine de recherche en soi. La théorie du krigeage a été publiée pour la première fois par Matheron [28] sur la base de la thèse de maîtrise de Daniel Krige. Comme indiqué précédemment, il repose sur l'hypothèse générale selon laquelle les points proches les uns des autres dans l'espace d'entrée sont plus susceptibles d'avoir des valeurs de sortie similaires. L'article original indique que le krigeage est une "combinaison pondérée" (combinaison linéaire) des valeurs d'observation qui "permet d'obtenir la meilleure estimation possible", ce qui en fait le meilleur prédicteur non biaisé – en anglais **BLUP (Best Linear Unbiased Predictor)** – au sens des moindres carrés pour l'interpolation spatiale ponctuelle. Le krigeage a d'abord été défini pour interpoler des observations ponctuelles. Il s'agit intrinsèquement d'un algorithme de régression, mais au Chapitre III, nous proposons un moyen de l'utiliser pour de la classification floue.

L'interpolation d'unités spatiales, telle que définie par Lam [29], concerne "la transformation de données d'un ensemble de limites à un autre". Lam a également utilisé les termes de zone source et de zone cible. Pour le problème de prédiction du **DPE**, les zones sources sont les logements et les parties de bâtiments observées, tandis que les zones cibles sont des bâtiments entiers. La recherche sur l'interpolation spatiale ou d'unités spatiales repose sur l'hypothèse suivante : les granules proches les uns des autres dans l'espace d'entrée sont plus susceptibles d'avoir des caractéristiques similaires (valeurs de sortie). Cette hypothèse est raisonnablement compréhensible pour les températures définies de manière continue dans l'espace, mais elle peut s'avérer plus difficile à observer et à modéliser lorsqu'il s'agit de données spatiales où les granules peuvent être de tailles et de formes diverses, parfois définies de manière incertaine. Gotway and Young [30] a mis en évidence les termes utilisés pour décrire l'interpolation d'unités spatiales et ses défis : le krigeage par blocs, la modélisation multi-échelle et multi-résolution, le problème de l'inférence écologique, le **MAUP (Modifiable Areal Unit Problem)**, le problème de la mise à l'échelle, le problème du changement de support et le problème de la réduction de la variance. Nous présentons ci-dessous les aspects de l'inventaire de Gotway and Young qui sont les plus pertinents pour résoudre le problème de prédiction des **DPE**.

Le krigeage par blocs (block Kriging en anglais) est un dérivé du krigeage conçu pour traiter des unités spatiales non réduites à des points. Il distingue les prédictions point-vers-zone, zone-vers-point et zone-vers-zone. Cette technique, héritée des activités minières, suppose que la caractéristique au niveau du bloc (granule, unité spatiale) est la moyenne des caractéristiques ponctuelles du bloc. La prédiction point-vers-zone produit une estimation "identique à celle obtenue en faisant la moyenne des estimations ponctuelles produites par [Krigeage]" [31], [23]. Kyriakidis [32] a décrit un modèle de Krigeage complet pour la prédiction zone-vers-zone, a prouvé qu'il s'agissait d'un prédicteur optimal et a esquissé la prédiction zone-vers-zone. Goovaerts [33] a étudié en profondeur le problème de l'estimation du variogramme, c'est-à-dire la mesure de la

similarité entre deux points situés à des distances différentes, pour le krigeage par blocs. Il a montré que le calcul de la moyenne réduit le palier (sill en anglais) du variogramme et a tenté de remédier à ce biais. En outre, si les estimations ponctuelles obtenues par le krigeage sont optimales, le krigeage zone-vers-zone peut ne pas être le prédicteur optimal pour la valeur moyenne sur le bloc.

Un problème connu résultant de l'établissement systématique de moyennes dans les modèles de krigeage d'unités spatiales se pose dans des scénarios tels que l'analyse des rendements des cultures, où l'ensemble des champs agricoles à agréger pour un certain type de culture varie d'une année à l'autre. Il consiste en le fait que les corrélations entre les variables de sortie dépendent fortement du processus d'agrégation, ce qui rend difficile la comparaison des corrélations entre différentes années. C'est le **MAUP** pour lequel une méthode de mesure a été récemment proposée par Briz-Redon [34]. Alors que le **MAUP** fait référence à la corrélation entre les variables de sortie, le problème de l'inférence écologique résulte, lui, du fait que les corrélations au niveau individuel sont différentes des corrélations des sorties moyennes au niveau écologique (niveau du groupe) : un manque d'informations sur les positions des individus conduit à un biais lorsque les informations moyennes sur les individus distribués dans des unités spatiales sont croisées avec d'autres variables individuelles (niveau du point : sexe, race). Selon Gotway and Young : "L'effet de lissage qui résulte du calcul de la moyenne est la cause sous-jacente du problème d'échelle dans l'étude du **MAUP** et du biais d'agrégation dans les études écologiques". Outre les corrélations, la variance elle-même est affectée par le calcul systématique de la moyenne. En effet, la moyenne de variables aléatoires identiquement distribuées a une variance plus petite que la variance des individus eux-mêmes. La question spécifique de la réduction de la variance au niveau des blocs a été partiellement abordée dans [35] en utilisant des blocs rectangulaires à plusieurs échelles.

Malgré ses limites, la méthode par calcul de la moyenne s'est avérée efficace pour l'interpolation de données d'unités spatiales. Par exemple, Poggio and Gimona [36] a réduit l'échelle des modèles climatiques et prédit l'humidité des sols en utilisant le krigeage sur les résidus d'un modèle additif généralisé [37]. Le krigeage zone-vers-point, aussi appelé désagrégation, a également été mis en œuvre par [38]–[40]. En outre, le krigeage zone-vers-zone (krigeage par blocs) a été utilisé efficacement par [41] et a été appliqué à la réduction d'échelle par [42] ainsi que par [43]. Le domaine de l'imagerie satellitaire a également bénéficié de ce cadre de travail, comme dans le processus de pan-sharpening, qui est "une technique visant à combiner la bande panchromatique (PAN) à résolution spatiale fine avec les bandes multispectrales à résolution spatiale grossière du même satellite pour créer une image multispectrale à résolution spatiale fine" [44]. Dans ce processus, les points sont pondérés en fonction de leur distance par rapport au centroïde du pixel du satellite lors du calcul de la valeur moyenne.

Le **MAUP** et le problème de l'inférence écologique appartiennent tous deux à une famille de problèmes liés à la combinaison de différents types de granules dans le même modèle, par exemple l'observation des habitations et la prédiction des bâtiments. Ces

difficultés sont regroupées dans la famille des problèmes de changement de support. Un autre problème particulier de changement de support, connu sous le nom de désalignement spatial, survient lorsqu'une variable de sortie donnée est observée à plusieurs échelles, y compris au niveau des points. En effet, le calcul systématique de la moyenne fait des points et des zones des objets différents avec une variabilité différente, des structures de corrélation différentes et donc des prédicteurs différents. La classification des problèmes tels que "zone-vers-point" ou "zone-vers-zone" reflète cette catégorisation. Moraga *et al.* [45] ont construit un cadre bayésien qui peut être itéré à la fois avec des observations ponctuelles et des observations par bloc, sur la base d'une moyenne au niveau du bloc pour les variables de sortie définies de manière continue sur le territoire. Ce modèle, comme d'autres modèles dérivés du krigeage, considère les blocs comme des surfaces connectées dans \mathbb{R}^2 qui doivent être "discrétisées" [33], ce qui peut fausser la réalité pour les sorties qui ne sont pas définies de manière continue sur l'espace, telles que les populations qui sont souvent des points discrets situés de manière hétérogène dans un bloc, comme un comté ou un secteur de recensement.

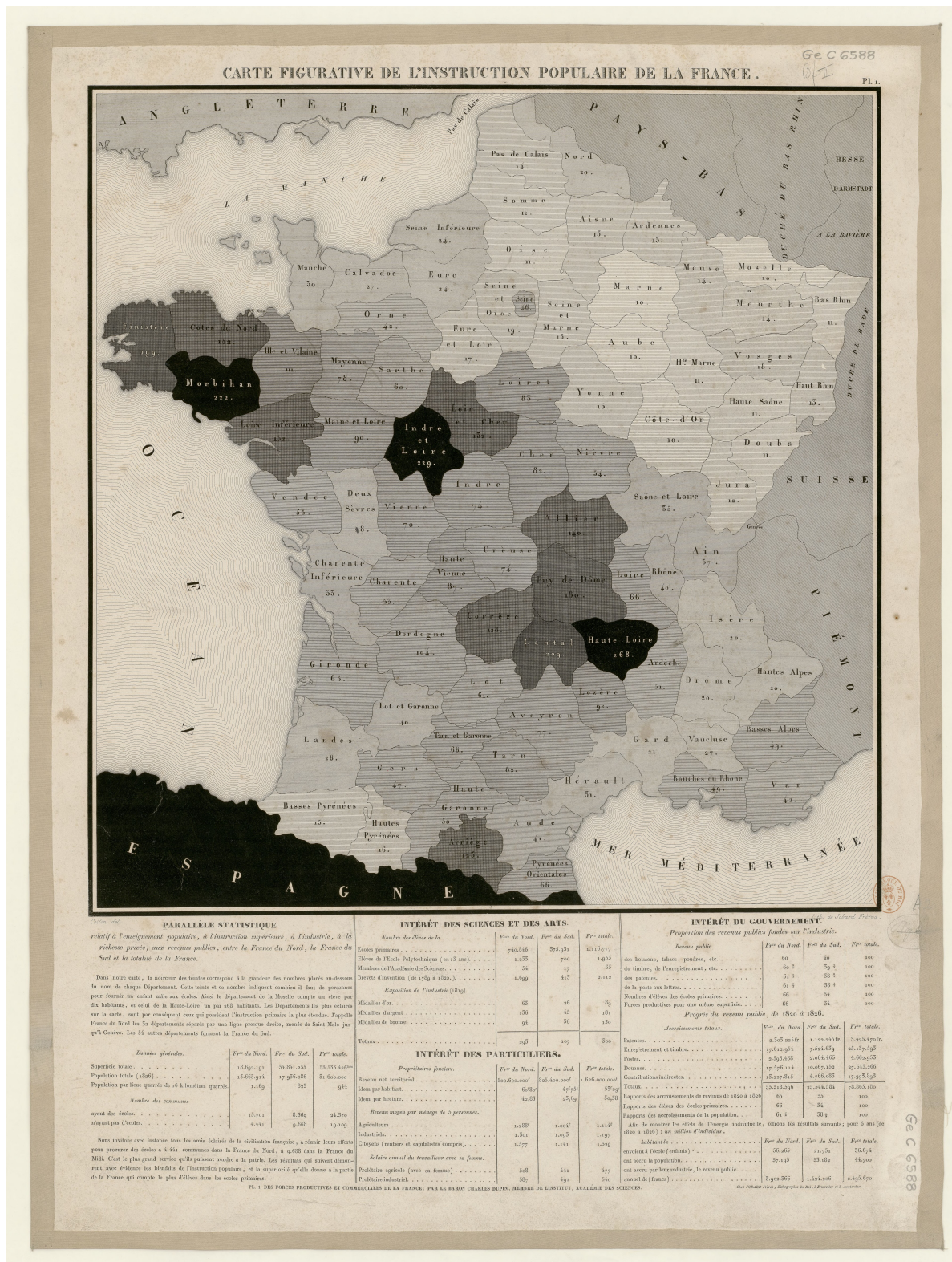


Figure 6 – La première carte choroplèthe réalisée par Charles Dupin en 1826. Il s'agit d'une carte représentant l'alphabétisation en France. Chaque unité de surface est un département. Plus la zone est foncée, plus la valeur associée est grande. Cette valeur représente "le nombre de personnes nécessaires pour fournir un enfant mâle à l'école" (sic). Par exemple, dans le département de la Loire, il y a un garçon scolarisé pour 66 habitants en moyenne.

LA RÉPARTITION GÉOGRAPHIQUE DES PASSOIRS ÉNERGÉTIQUES DANS LA COMMUNAUTÉ URBAINE

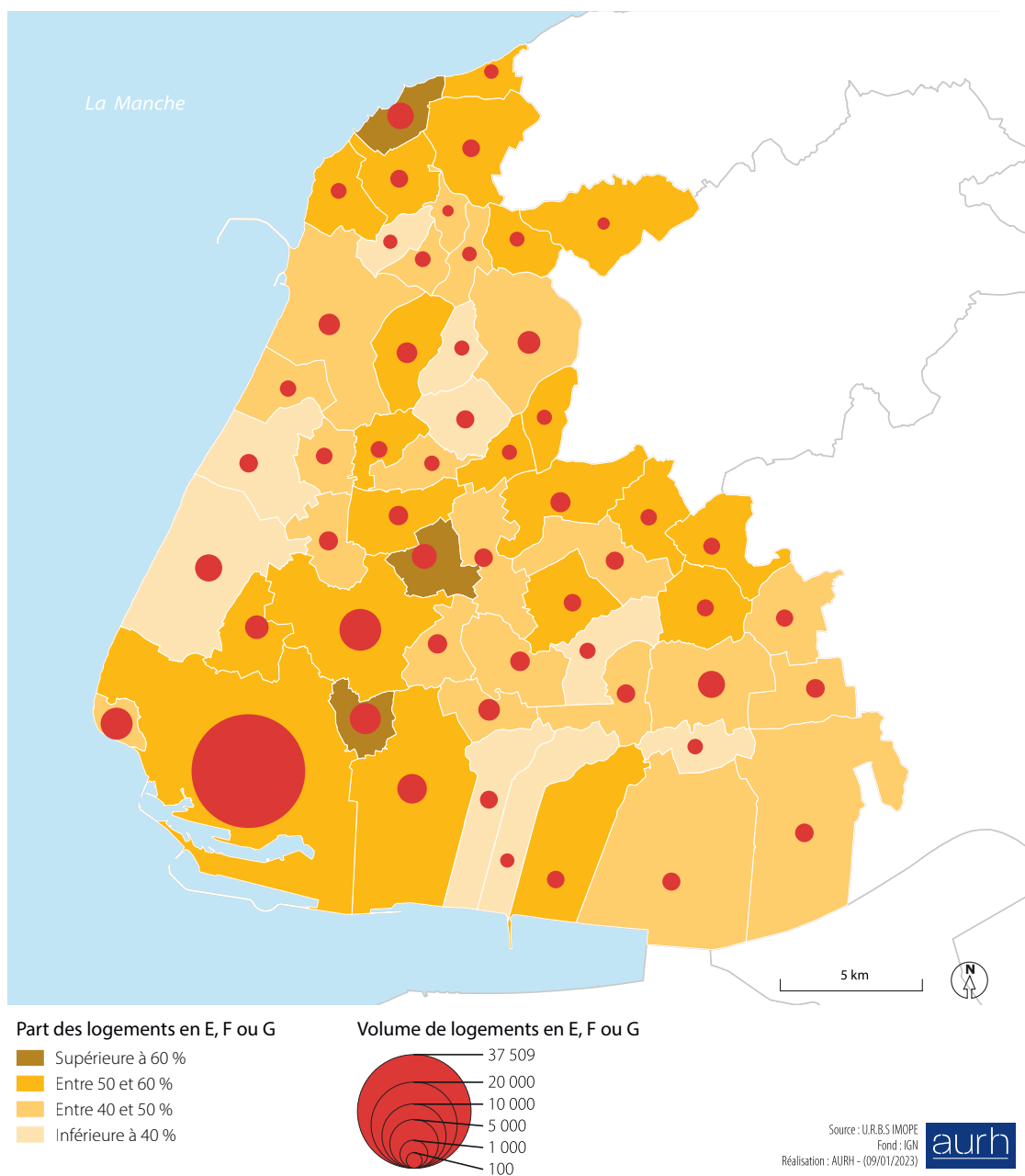


Figure 7 – Carte chloroplèthe représentant le pourcentage prédit de logements pour les étiquettes E, F ou G dans l'agglomération du Havre. Les unités spatiales sont les communes. La couleur de l'unité spatiale représente le pourcentage de logements avec les étiquettes E, F ou G ; plus la couleur est foncée, plus le pourcentage est élevé. Les cercles représentent le nombre de logements étiquetés E, F ou G ; plus le rayon est grand, plus le nombre est élevé [25]. Cette étude utilise les DPE prédits à l'aide de FKNN (Fuzzy k-Nearest Neighbours) tels que présentés dans le Chapitre IV.

Introduction

1	Institutional and Environmental Context28
2	Overview of the Research Project.29
2.1	Research Problem29
2.2	Hypotheses and Objectives.29
2.3	Scope and Limitations30
2.4	Significance31
3	Methodology and Overall Structure.32
4	State of Things34
4.1	What is the Energy Performance Certificate?34
4.2	What Does an EPC Look Like?37
4.3	Renovation Objectives38
5	State of the Art42
5.1	Assessing the Energy Performance of a Building Stock.42
5.2	Spatial Interpolation44

The main purpose of this thesis is to construct an explainable predictive model for estimating the energy efficiency of each and every building in France. The training set for this work is provided by the set of observed **EPC (Energy Performance Certificate)**. After a presentation of the current legal and scientific contexts in the following Sections 4 and 5, the data preprocessing process and its challenges are presented in Chapter I. Chapter II introduces a novel method to interpolate mixture distribution. It is called Mixture Kriging. It takes into account the uncertainty of the buildings' locations and presents the important advantage of having no measurable **MAUP (Modifiable Areal Unit Problem)**. This model is efficient at a small scale – a city, for instance – but turns out to be too heavy to upscale at the national level. Taking into account the teachings of this model, Chapter III presents a lighter model, called Joint Kriging, that does not account for the position uncertainty of the buildings and has a **MAUP** effect, but turns out to be very efficient and easy to upscale. This model can be constrained in order to impose an overall distribution on a given territory, which is especially interesting for predicting **EPCs**. While Mixture Kriging is a regression model, Joint Kriging is a fuzzy classification model. Chapter IV compares the classification results of Joint Kriging with those of the model that is currently in use at **U.R.B.S.** company. The latter is an optimised **KNN (k-Nearest Neighbours)** model. Those two models are also compared with a hard classifier, Random Forest. It turns out that fuzzy classification with Joint Kriging is very promising and has been chosen to replace the current model used at **U.R.B.S.**. Chapters II, III, and IV are three research articles that have been written during the last three years. The first and third ones are already published; the second one is under review.

1 Institutional and Environmental Context

This work arises from the conjunction of three determining factors: the ecological context of the climate changes, the European Union decision to try to reach sustainability, and the creation of the **U.R.B.S.** company that is accompanying institutions in managing the transition of the building stock of the territory they administer.

European policies for sustainability, defined by multiple European Parliament directives [1], [2], draw scientists, stakeholders, and politics to explore novel approaches to reduce energy consumption and minimise **GHG (Greenhouse Gas Emissions)**. Hence, European countries

“Member States shall take the necessary measures to ensure that minimum energy performance requirements for buildings or building units are set with a view to achieving cost-optimal levels. [...]

Member States shall take the necessary measures to ensure that minimum energy performance requirements are set for building elements that form part of the building envelope and that have a significant impact on the energy performance of the building envelope when they are replaced or retrofitted, with a view to achieving cost-optimal levels.”

[1]

are defining strategies to enhance the energy performance of anthropogenic activities and address the urgent challenges of climate change. Pursuant to this roadmap, initiatives aimed at enhancing energy-efficient measures in the building sector are pivotal. Indeed, the building sector is one of the world's key energy consumers, accounting for 40 % of European energy consumption, [3], [4]. It contributes significantly to GHG (36 % of the total), primarily CO₂, thereby altering our planet's climate, and has been experiencing an overall rising trend over the past decades.

The **U.R.B.S.** company was founded in 2019 to provide support for urban public policy's decision-makers. A multidisciplinary team comprising data engineers, software developers, and researchers produces a database gathering available information about dwellings. This database is enriched by pairing sources, imputing missing data, and computing synthetic indicators. It is made available for users through a geo-service that can be customised for each territory. It was quickly brought to the company's attention that local institutions such as *EPCI (Établissement Public de Coopération Intercommunale – Public Intermunicipal Cooperation Authority)* or departments were missing knowledge about the building stock of the territory that they administer, which was delaying upscaling of energy renovation in the above-mentioned context of European policies for sustainability. Willing to contribute to this collective effort, **U.R.B.S.** and Mines Saint-Etienne launched, in 2021, a joint research project for imputing missing data in Energy Performance Certificates observations. This thesis is the result of that project.

2 Overview of the Research Project

2.1 Research Problem

This research is focused on providing sufficient information to decision-makers so that they can diagnose the building stock of a given territory, identify energy-efficient and energy-inefficient buildings, and therefore better target incentives. The main issue is that current research, as will be proved, is either describing the building stock at a large scale or analysing particular buildings given a detailed technical knowledge of their structure. However, to address renovation objectives, institutions need to know the energy efficiency of all buildings on the territory they administer, at a reasonable cost, that is, without physically visiting all of them. The challenge is therefore to extract sufficient information from available data at the national scale so as to compensate, at least partially, for the lack of technical knowledge in each building. The goal is to predict, for each and every building in France, its energy efficiency, focusing on the detection of energy-inefficient buildings, called energy sieves.

2.2 Hypotheses and Objectives

Sufficient information. The first hypothesis for this work is that the volume of available information on the French building stock is enough to derive knowledge about

each building’s energy efficiency without requiring a physical visit to the facilities. The available observations given in the **EPC** database, resulting from physical visits to the dwellings, cover less than 20 % of the building stock³ (3.2 million **EPC**s were produced in 2022 and 4.5 millions in 2023) and there is no plan, at this stage, to render **EPC** mandatory for all buildings. Therefore, this hypothesis strongly determines the feasibility of solving the research problem. Available information about buildings includes a diversity of features such as their age, general structure, and socio-economic context for occupants and owners.

Learnability. The second hypothesis, deriving from the first, is to assume that there exists some algorithm(s) that can perform supervised learning to predict **EPC** labels for every French building, using information about buildings as input data and the database of **EPC**s as observations. It means that two buildings are assumed to be more likely to be of the same **EPC** label if they are similar to each other in a way that remains to be defined. This hypothesis is objectivised by a performance indicator which is, in our case, the balanced accuracy.

The above two hypotheses are challenged for buildings that look similar from a data point of view but have very different energy efficiencies. This is due to some unavailable factors, such as a renovation, that are not mandatory to declare. On the contrary, addressing the research problem is easier if it suffices that two buildings are of the same age or are geographically close to each other for them to have similar energy efficiencies.

Our objective is to predict the buildings’ energy efficiency with the best possible performance, based on these hypotheses. On our way, we reformulate them, construe them, and finally validate or refute them to a certain extent. This objective is constrained by its application requirements: it should be possible to implement the results to make predictions available in the **U.R.B.S.** software called **ONB** (*Observatoire National des Bâtiments – National Buildings Observatory*). In particular, we are dealing with large datasets, and we must be able to upscale the designed process to the country level.

2.3 Scope and Limitations

This research is concerned only with dwellings, excluding any other kind of facility, be it dedicated to offices, industry, or storage for instance. It covers a large territory, in our case all of France, therefore taking into account questions of climate areas and altitude.

The data fusion process is described in the following part but is not discussed from a research aspect. It was jointly designed and implemented by **U.R.B.S.** team. Obviously it has a strong impact on our work and there have been multiple improvements made during the 3 years that this project lasted, to concentrate on relevant data. A key point to keep in mind is that we are working at the address scale. Meaning that we

3. Statistics from the **IMOPE** database, counting addresses as described in Subsection 2.3.

identify an address such as “7 Bergson street, Saint-Etienne” with a building. This is not completely true as a single address may be used for multiple buildings and, rarely, a single building may have multiple entries associated with multiple addresses. But when this research started, there was no available dataset to disambiguate the situations. It is only in the end of 2023 that France started an official inventory of physical buildings. It remains that statistically, according to the **MoF (Ministry of Finances)** database, the vast majority (92%) of buildings have a single address and the majority (68%) of addresses point to as single building. Note that even if an address points to multiple buildings, there is usually only one building with dwellings.

Due to the novelty of the project, and in order to be able to justify the predictions to a client of **U.R.B.S.**, it is requested to propose an explainable algorithm. For that reason, black boxes algorithms have been excluded from the study. In particular, we did not work with neural networks nor with random forests. This constraint is discussed in Chapter **IV**, where the performance of Random Forest is assessed and in the Chapter **Conclusion and Future Directions**, results given by **tabnet** neural network are presented.

Although regarding data, our research is contextualised in mainland France, the methodology can be useful for other countries, especially European ones. And our work was perceived positively when we presented it in France, Belgium and United States. All three articles that are presented here have been positively reviewed, two of them are already published and the third one is being reviewed.

This work is a contribution to the following **SDG (Sustainable Development Goals)**:



11. Make cities and human settlements inclusive, safe, resilient and sustainable: The prediction of buildings’ energy efficiency facilitates the renovation programmes.



12. Ensure sustainable consumption and production patterns: Encouraging the renovation of energy-inefficient buildings reduces their energy consumption and increases their lifespan saving construction material.

2.4 Significance

The main conclusion of this work is that it is indeed possible to draw information about a building’s energy efficiency from available databases. This research also shows that it is possible to look at **EPCs** as geolocated data, meaning that in addition to comparing two buildings’ descriptions, it is also of interest to know that they are located close to each other to improve the prediction of energy efficiency. To the extent of our knowledge, this is new. It suggests that instead of keeping an approach per building with the same characteristics for encouraging renovation, it may also be of interest to consider the energy inefficiency per neighbourhood. This is not surprising to people familiar with community development issues, but it is certainly surprising for institutions that usually look at the building stock from a highly technical point of view. Our work also suggests that predicting an **EPC** label instead of a standardised energy consumption is more

relevant. This is an a posteriori confirmation that the nature of the **EPC** is beyond thermal engineering. And regarding the detection of energy sieves, this work shows that energy sieves (**EPC** labels F and G) are more difficult to predict than other labels A to E. To overcome this difficulty, it is beneficial to consider a fuzzy classification approach. Depending on models and our level of acceptance for false positives, it is now possible to detect 36 % to 66 % of energy sieves without physically visiting the buildings, while we could not overpass 23 % before this research project started, and while none of the other labs working on this topic in France published any quantitative results yet.

3 Methodology and Overall Structure

To design a methodology, we follow two guidelines. The first one is that, from an industrial perspective, a key outcome of this project is the identification of energy sieves. Therefore, beyond the common knowledge of determining factors for energy efficiency, such as the age of a building, it is of interest to identify factors that are correlated with energy inefficiency. We make the research hypothesis that some socio-economic factors could be significantly correlated with energy inefficiency. The second guideline is that we are dealing with geolocated data. So we make a second research hypothesis, saying that it is beneficial for our work to treat **EPCs** as geolocated data instead of sticking to the traditional thermal engineering approach.

A central theme of this work is the **EPC**. The definition and perception of what is an **EPC** vary among the actors and change over time. From a thermal engineering object, it has become a political issue, modified multiple times by the legislative body, and it is now accessible as a database [5]. In order to find a common ground for discussion, along this thesis, the **EPC** is regarded as a legal classification of dwellings and buildings. This should not obliterate the thermal engineering used for establishing an **EPC**, but considering the variety of actors who take part in its definition, it is way too restrictive to reduce it to the result of an engineering model. The **EPC** is a legal outcome of the sustainability objectives and, in particular, of the need for intensifying the renovation effort. The multiple national renovation programmes have in common to fall short of their expectations. It is likely that a better knowledge of the building stock would help achieve national targets. The definition of the **EPC** and the state of the art in producing large-scale knowledge about the energy efficiency of a building stock are presented in the following Sections 4 and 5.

The data gathering information about dwellings suffers from geolocation uncertainty. That aspect appears to be an important issue to tackle to be able to fit the data into a geostatistics model. This is why the first article presented in Chapter II introduces a geostatistics model that natively takes into account position uncertainties and accommodates multi-scale data. It is a regression model that predicts the standardised energy consumption of buildings based on their latitude, longitude, age, and potentially other variables. Although this model is promising at the scale of a city, it turns out to be costly to upscale.

If it is difficult to take into account the position uncertainty in the input data, it may be possible to include it in the output data, meaning in the predicted values. Following this track, we leave the regression approach to turn towards classification, meaning to predict the **EPC** label and work on fuzzy classification. Fuzzy classification not only predicts a class but also gives information about its likelihood and the likelihood of the other class. The second article, in Chapter **III**, presents a geostatistics model of classification called Joint Kriging. It is lighter than Mixture Kriging and, therefore, in addition to its good performances, is less difficult to upscale.

Beside this midterm research, it was necessary, in the background, to develop a model to be implemented in **U.R.B.S.** production. This model is also a classification model, not specifically for geographic data, and it includes multiple socio-economic variables. It is based on a **FKNN (Fuzzy k-Nearest Neighbours)** algorithm optimised with **SPSA (Simultaneous Perturbation Stochastic Approximation)** pseudo-gradient descent. The third article, in Chapter **IV**, presented in this thesis compares the performances of Joint Kriging, this **FKNN** algorithm, and Random Forest.

Overview of the research project: Essentials

Industrial problem. Acquiring a fine knowledge of the buildings' energy efficiency.

Research problem. Learning from the observed **EPCs** to predict the energy efficiency of unobserved buildings.

Constraints. Model should be explainable. While the main objective is to predict the **EPC** at the building level, the overall predicted distribution should match as much as possible the overall estimated distribution.

Research hypotheses The information available for the complete building stock is sufficient to derive energy efficiency without a site visit. And we can find an algorithm that learns from the existing **EPCs** and predicts the **EPCs** of unobserved buildings.

Scope and limitations We are concerned only with the French stock of dwellings. We look for an explainable model.

4 State of Things

4.1 What is the Energy Performance Certificate?

Legal definition of an EPC

The **EPC (Energy Performance Certificate)** for a building or part of a building is a document that includes the quantity of energy actually consumed or estimated, expressed in primary and final energy, as well as the induced greenhouse gas emissions, for a standardised use of the building or part of the building, and a classification based on reference values for comparing and evaluating its energy performance and greenhouse gas emissions performance. It provides information on ventilation or air exchange conditions. It is accompanied by recommendations aimed at improving these performances and the amount of theoretical expenses for all the uses listed in the diagnosis.

It is established by a person meeting the conditions provided for in Article L. 271-6.

Its validity period is determined by regulations.

Article L126-26 of the Building and Housing Code [6], translated from French to English by the author

In France, the **EPC (Energy Performance Certificate)** is called Diagnostic of Energy Performance – **DPE (Diagnostic de Performance Énergétique – Energy Performance Certificate)**, and it is performed by a so-called diagnostician who is certified for this job. He or she collects qualitative and quantitative information about a building or a dwelling and enters it into software that has been approved by the **ADEME (Agence Française pour la Transition Écologique – French Agency For Ecological Transition)**. Based on this input data, taking default values for missing data, the software produces two figures representing a standardised annual energy consumption per square metre and per year and a standardised weight of emitted carbon dioxide, CO_2 , per square metre and per year. By a combination of thresholds, an energy label, a **GHG (Greenhouse Gas Emissions)** level, and a global **EPC** label are derived.

The reader can already notice the strong government control in this system. In fact, the French law not only defines the principle of classifying buildings but also defines the computational algorithm that computes energy consumption and the **GHG** emissions and defines the thresholds that decide labels. These labels, in turn, are used to define legal rights and duties, such as the construction standards for new buildings or the permission to rent out a dwelling. The ecosystem of laws around this was re-written in 2021 in the light of new scientific knowledge, new European targets, and a new political agenda. We describe hereafter the state of the law after 2021, mentioning important changes as compared to the previous system. In particular, the **GHG** part was added in 2021, although the information about **GHG** emissions was already appearing in the **EPCs** established before 2021.

Let us examine the legal definition of an **EPC** (see the text box above) in more detail:

- **“Building or part of a building”**. The same scale is used for a specific dwelling in a block of flats and for the whole building. However, the method is not the same.

For a block of flats, the diagnostician visits a representative sample of buildings to derive the **EPC** of the others. And the energy consumption of a block of flats is not equal to the sum of the dwellings' consumptions because there are areas in a building that do not belong to any flat, such as the staircase, the lift, if any, or the entry hall. Our goal is to predict **EPC** at building level only. Therefore, a learning algorithm somehow has to find a way to estimate a building's **EPC** based on dwelling observations.

- **“Primary and final energy”, “GHG emissions”**. In fact, the **EPC** label is determined by the primary energy consumption. It is computed by multiplying the estimated final energy by a conversion factor, which is 2.3 for electricity (2.58 before 2021) and 1 for the other sources of energy. Similarly, a conversion factor is applied to compute **GHG** emissions: 0.079 for electricity, 0.227 for city gas, 0.324 for domestic fuel oil, while they were respectively 0.180, 0.234, and 0.300 before 2021 [7]. These factors are partly political: Reducing the conversion factors for electricity is the result of a pro-nuclear energy government that is launching multiple nuclear plant construction projects.
- **“Standardised use of the building”**. The law is making assumptions on the number of occupants in a dwelling, the target temperature of heating, or the use of sanitary water. **EPC** is not concerned with real energy consumption. Practically, it means that wealthy people, who have a tendency to consume more energy, are likely to consume more energy than the **EPC** standard, and people with low incomes are likely to consume less. More details about energy poverty can be found in [8].
- **“A classification”**. The main purpose of **EPC**, as suggested by European Union rules, is to classify buildings, to compare them with each other, to be able to detect energy-inefficient buildings and renovate them.

As a result, **EPC** is trying to serve two purposes, informing and classifying, but they are not fully compatible. It is clear from the above remarks that the classification aspect has been privileged to induce what is perceived by the government as a transitional behaviour, such as changing a city gas heating system into an electric one. There is no point here in discussing the pros and cons of these facts, but rather, it is important to keep in mind that there is only a weak correlation between real energy consumption and energy consumption measured by the **EPC**. This is regularly pointed out by multiple actors [9].

Finally, it is worth mentioning another important bias of the **EPC**: the classification is based on the energy consumption per year and per square metre of living area. But energy consumption is highly linearly related to the building's envelope surface area. Therefore, the energy consumption per square metre is proportional to the quotient of the envelope's surface divided by the living surface area, which is a compactness indicator. The smaller it is, the more compact the building. Cubicles are very compact, giving a compactness indicator of $1/8$. But if a building or dwelling has a low ceiling with

rooms arranged in line, the compactness indicator tends towards $1/2$. As a consequence, depending on the way dwellings are arranged in a block of flats, the whole building may have good energy efficiency as evaluated by the **EPC**, while some of its dwellings may be measured as energy sieves. This effect is worth it if the dwelling is at a corner of the building or at the top of it. Moreover, for very small dwellings, the heating part of the energy consumption is considerably reduced compared to the consumption of hot water, tending to point out energy sieves even if the dwelling is energy-efficient.

In spite of these limitations, the **EPC** remains an important indicator. It is the only indicator that results from an in-depth study of a building's energy-related behaviour. Diagnosticians collect a large number of indicators that can help in understanding the structure of the building stock. And it is the base indicator for designing thermal regulations. That is why an **EPC** is mandatory in the following cases:

- For the delivery of a new building or an extension of an existing building. It is given to the buyer.
- For selling or renting out a building or part of a building. It is given to the potential buyer or tenant and attached to the bail contract. Failing to inform the buyer or tenant is punishable.
- Any block of flats that was constructed before 2013.

However, certain buildings are excluded from this requirement: houses with a surface area smaller than 50 m^2 , temporary buildings that should stay less than two years, national heritage buildings, parts of a building that have no heating system or that are occupied less than 4 months a year.

As a result, millions of **EPCs** are collected every year. Almost 8 million **EPCs** have been collected between June 2021 and December 2023 for houses, flats, or blocks of flats. But this set of observations is not representative of the French building stock because some buildings have undergone multiple diagnostics while others have never been visited by a diagnostician. In particular, houses are underrepresented because people who own a house tend to keep it for a very long time (see Figure 8). Among the total building stock, houses constitute 90% of buildings, while they represent only 69% of the observations. Consequently, blocks of flats are overrepresented in the observations.

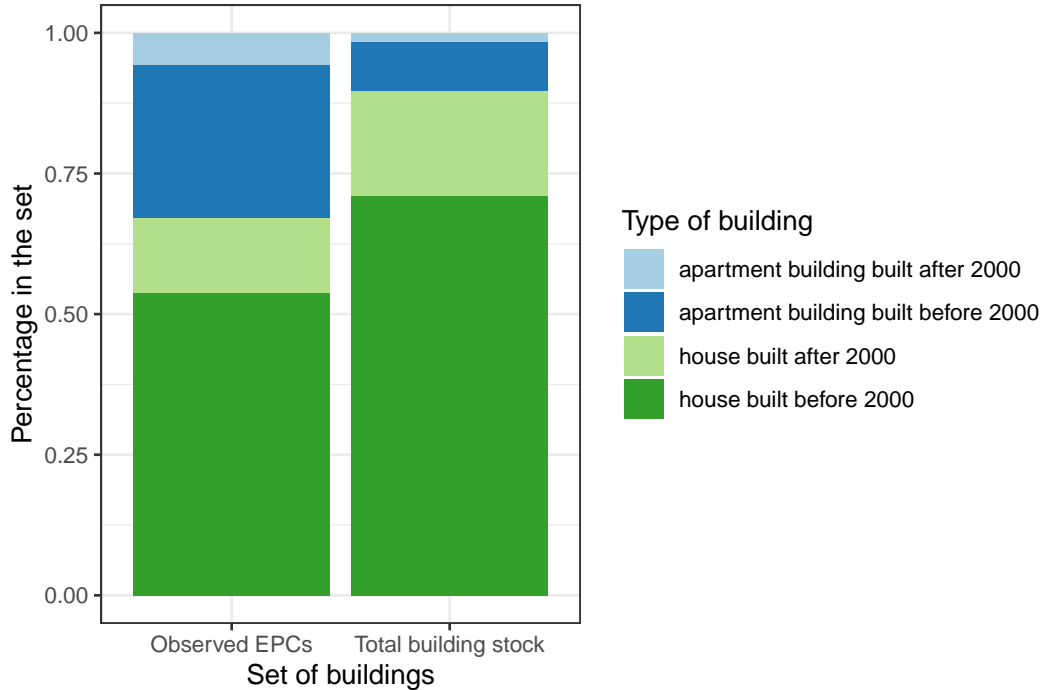


Figure 8 – Observed EPCs vs. total building stock in France: proportions of the populations built before or after 2000, that are houses or flats.

4.2 What Does an EPC Look Like?

The primary energy consumption is expressed in kilowatt-hours per square metre of living area and per year ($\text{kWh}/\text{m}^2/\text{year}$), see Figure 9. In particular, it is worth noting that while energy consumption is linearly dependent on the surface of the dwelling's envelope, it is regarded here as a quantity per surface unit of living area. This raises issues of compactness, as detailed in Subsection 4.3.

The GHG emissions are expressed in kilograms of equivalent CO_2 per square metre of living area and per year ($\text{kg}_{\text{CO}_2}/\text{m}^2/\text{year}$). It also raises some compactness issues. The design of this unit is to convert the weight of any GHG into a weight of CO_2 that would have the same greenhouse effect. Its main advantage is to yield a readable result for the end user, but a drawback is that it hides the diversity of GHG that are emitted. Some of which may have adverse effects other than the greenhouse effect.

As can be seen in Figure 10, two series of thresholds define a primary energy label and a GHG emissions threshold, both of which range from A to G. The EPC label is determined by the colour of the point, whose coordinates are the GHG emissions and the primary energy consumption. The main consequence of this system is that the EPC label is the worst of both sub-labels. For instance, if a dwelling has GHG emissions equal to $80 \text{ kg}_{\text{CO}_2}/\text{m}^2/\text{year}$ and a primary energy consumption equal to $300 \text{ kWh}/\text{m}^2/\text{year}$, then its GHG label is F and its energy label is E, therefore its EPC label is the worst of F and E, i.e. F.

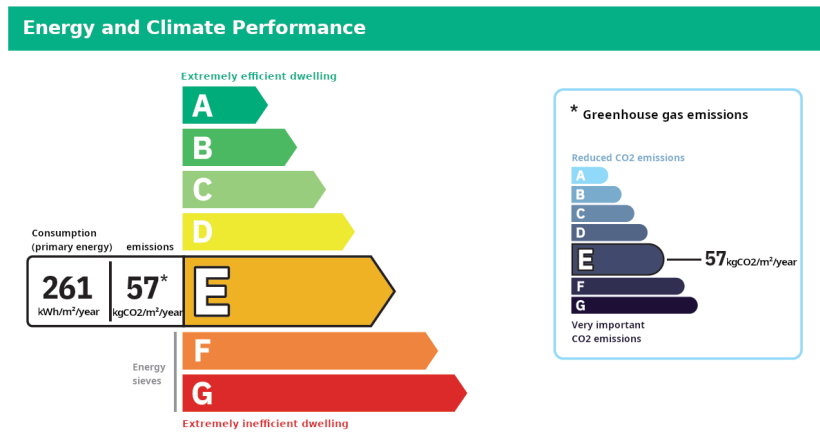


Figure 9 – The French official vignette, which gives the result of the energy efficiency diagnostic.

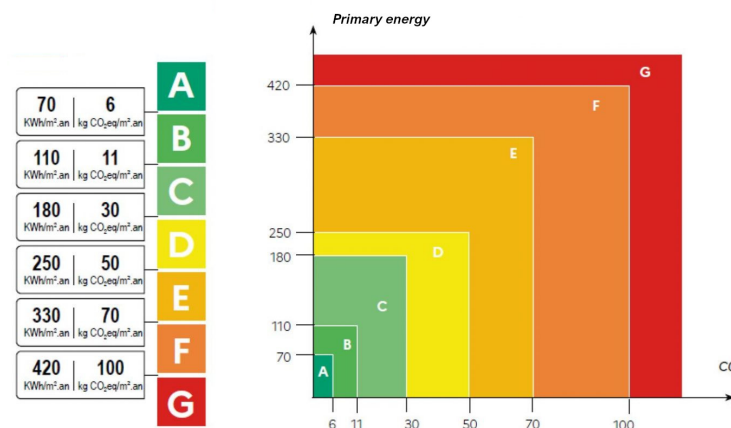


Figure 10 – The double thresholding process involving primary energy consumption and GHG emissions for computing the EPC label in France since 2021.

The results of the technical study appear in a vignette presented in Figure 9. Both qualitative and quantitative results are presented. But in our experience, most people are only interested in the qualitative aspect. It makes sense because the legal framework depends on the labels. But it may hide the amount of effort necessary to improve the EPC by one label.

4.3 Renovation Objectives

In the 21st century, the first landmark for the French commitment to sustainability was the so-called “Grenelle de l’Environnement” which took place from September to December 2007. The conclusions of this meeting paved the way for two series of laws known as Grenelle I and Grenelle II. The first of these, published in 2009, places France among the most ambitious of European countries. The objective is to divide by four

the **GHG** emissions in 2050 as compared to 1990; see the text box below. Regarding buildings, the existing building stock has to reduce its energy consumption by at least 38 % before 2020, and a minimal number of annual renovations are set. A diagnostic of all public buildings is prescribed, with specific targets of energy reduction for them. Even more specific objectives are set for social housing operators, for which priority areas are defined. Some financial incentives are also organised for owners, banks, and insurance companies.

“Article 2: The fight against climate change is placed at the top of the list of priorities. In this perspective, France confirms its commitment to reduce its greenhouse gas emissions by fourfold between 1990 and 2050 by decreasing greenhouse gas emissions into the atmosphere by an average of 3 % per year. The aim is to bring its annual greenhouse gas emissions to a level lower than 140 million tonnes of carbon dioxide (140 Mt_{CO₂}) equivalent by that deadline.

Article 5: The state sets the objective of reducing the energy consumption of existing building stock by at least 38 % by 2020. To achieve this, the state aims to completely renovate 400 000 homes annually starting in 2013.” Grenelle I Law, 2009-967, published on the 3rd of August 2009 [10].

The Grenelle II law was published in 2010. Its goal is to organise the implementation of the commitments made in Grenelle I. It introduces the concept of energy poverty and states that energy reduction efforts should benefit households that are in energy poverty; see the text box below. It redefines the **EPC** to include the **GHG** emissions and renders the **EPC** mandatory in cases where the building is sold or put for rent. A diagnostic has to be performed in all blocks of flats equipped with collective heating systems before 2017. And all **EPCs** are collected in a common database.

These laws have been modified almost every year since their first publication. Although there is, without a doubt, an inflexion in the historical development of our society, it is also true that one of the renovation objectives has been attained. The energy consumption of office buildings has increased by more than 29 % between 1990 and 2019 [12], and residential buildings have increased their energy consumption by more than 10 % between 1990 and 2021, with a maximum in 2013, when the consumption was 20 % more than in 1990 (data from the Ministry of Ecological Transition). However, the country has reduced its **GHG** emissions from 371 Mt_{CO₂} in 1990 to 252 Mt_{CO₂} in 2020 (this includes the impact of COVID-19; it was 291 Mt_{CO₂} in 2019). This is mainly due to efforts in the energy production sector and industrial processes and results from controversial choices such as the development

A person is considered to be in a situation of energy poverty under this law if they experience particular difficulties in obtaining the energy supply necessary to meet their basic needs in their home due to the inadequacy of their resources or living conditions.

Law Grenelle II, 2010-788, published on the 12th of July 2010 [11].

of nuclear electricity.

In 2021, the whole Grenelle approach was redefined into a law called *The law on combating climate change and strengthening resilience to its effects* [13]. The consequences of this law are summarised in Figure 11, page 41 (in French). Its main impact on buildings is the limitation of cities' extension over agricultural or natural land. This law is re-defining the EPC as described in Subsection 4.1. It defines terms such as efficient renovation, which is a renovation that brings the building to a label A or B, or global renovation, which is an efficient renovation in less than eighteen months. Instead of focusing on energy renovation strategies, it introduces coercive measures against building owners. In particular, it states that dwellings with a G EPC label cannot be rented out, starting in 2025. The same happens for F labels in 2028 and for E labels in 2034. At the same time, the thermal model that computes the EPC was revised. Real estate agents, such as the powerful FNAIM (*Fédération Nationale de l'Immobilier – National Federation of Real Estate*), the biggest French real estate federation, quickly realised that this was to create a storm in the renting business, especially in big cities like Paris where small, old, and energy-inefficient buildings are easy to rent. Lobbyists reacted with a surge of pressure on the government. Officials declared that the new thermal model was flawed and adjusted the model for buildings constructed before 1975 in order to produce fewer energy sieves. Another concern was raised regarding small dwellings where hot sanitary water consumption becomes relatively important in the overall energy consumption as compared to the heating needs, introducing a bias in the EPC. Therefore, another modification of the model was also introduced in 2024 to minimise this effect on flats whose surface is less than 40 m² through a modification of the compactness indicator.



Figure 11 – A visual summary of the Law on combating climate change.

5 State of the Art

In this section, we present a review of the scientific literature regarding two important aspects of the present research. The first one is the assessment of the energy performance of a building stock. The second one is the interpolation of geolocated data. Our work is at the crossroads of these research traditions, which started mainly in the 19th century. Studying energy performance was made possible by the works of Joule and Kelvin, among others in thermodynamics, and was made necessary by the industrial revolution, which required a lot of energy. Interpolation of geolocated data is a branch of spatial statistics that is an old field of research, especially in epidemiology, but it took a new turn when it benefited from the advances in the knowledge of contagious diseases by researchers like Pasteur. John Snow, for instance, is famous for having identified a contaminated pump well responsible for the cholera outbreak of 1854 in Soho thanks to a geostatistical approach; the work of J. Snow is presented by the *Lancet* in [14]. Spatial interpolation itself greatly benefited from Danie G. Krige’s work as a mining engineer.

The bibliography below is purposely limited to the aspects that are of interest for all three articles that are presented in Chapters II, III, and IV. Each of these chapters contains a specific review of the scientific literature, too.

5.1 Assessing the Energy Performance of a Building Stock

In the scientific literature, assessing the energy performance of a building stock can be approached from an engineering perspective, from a data management perspective that includes metamodeling, and from a geostatistics point of view, which is the choice of this work.

Physical models and building typologies. From an engineering perspective, heat engineers have physical models that compute an energy balance in order to find a given building’s energy consumption. This is the traditional approach, and the **EPC** is computed by this method. The details of the physical model, called the 3CL model, are available in [15]. To work at a larger scale, thermal engineers define typologies of buildings, compute a distribution of these types on a given territory, and therefore infer a distribution of **EPC** labels or energy consumptions. This approach has proven to be efficient [16]. However, the lack of knowledge about the detailed technical features of each building is a strong limitation for a prediction at the building level. It means that the total energy consumption of a city might be properly predicted, while the energy consumption of a particular building in this city is poorly predicted. In fact, the energy consumption at the building level may even not be predicted at all when the knowledge of a representative sample is available. This is the case, for instance, of the French national study called “Enquête TREMI” [17]. Those models require many parameters that are almost impossible to gather for each and every building in a territory. Some feature reduction efforts have been made [18] but the remaining features are still problematic to infer and require extra efforts [19], raising propagation of uncertainty concerns.

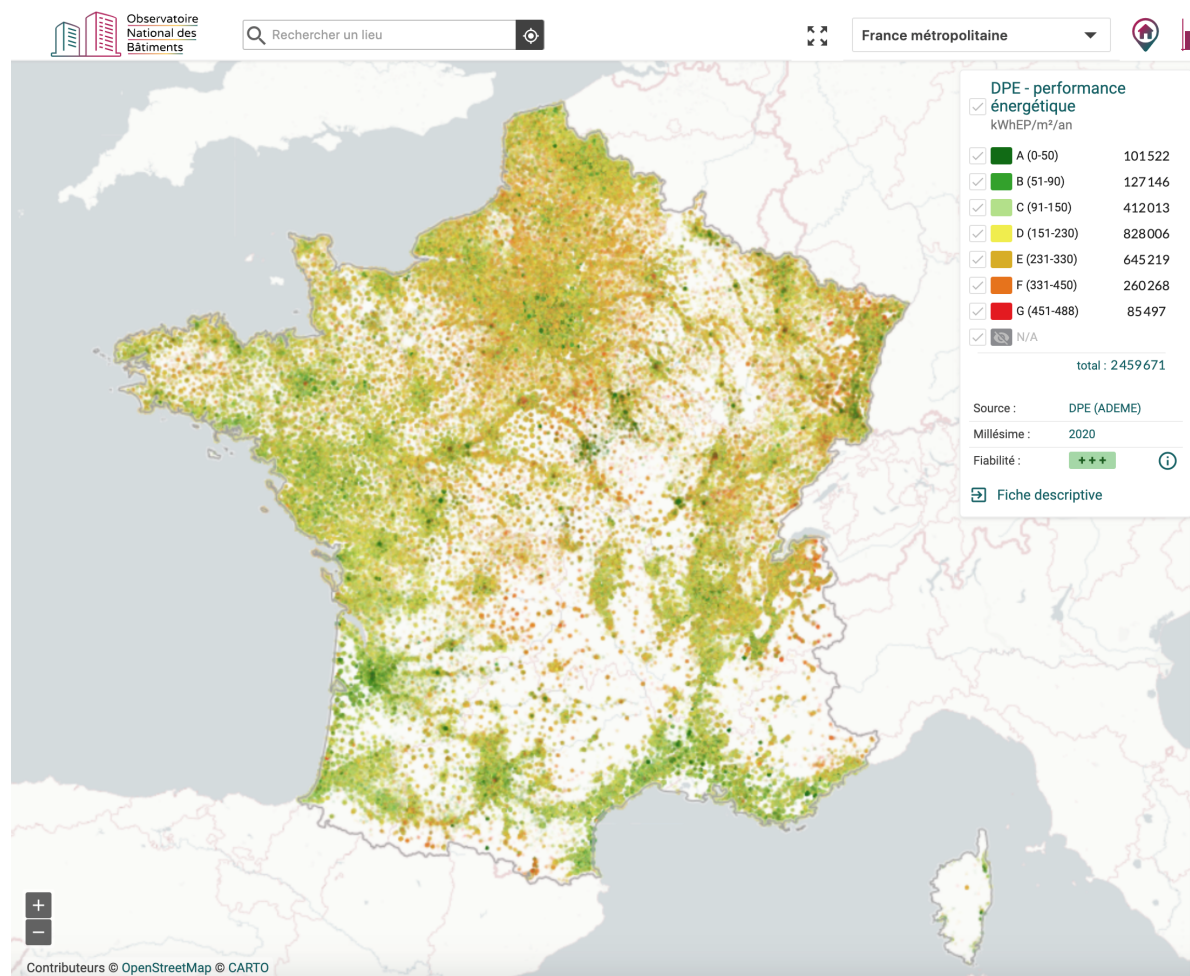


Figure 12 – Map of French inventoried EPCs. This image is a screen capture of the ONB (*Observatoire National des Bâtiments – National Buildings Observatory*), released with the consent of the rights holders U.R.B.S..

Data fusion. 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 [20]. This process is becoming increasingly complex due to the growing amount of data available, whether it be ecological, social, or institutional. These datasets relate to space units of varying shapes, dimensions, and cardinality. And in some cases, it may be difficult to determine the exact position of an observed object. This is the case with buildings, since many governments lack a detailed map of the building stock in their country. Property tax is typically based on intrinsic factors such as surface area and number of bedrooms, but not extrinsic factors such as 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 a flat; 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. The

data fusion process, together with the data preprocessing, are presented in Chapter I.

Metamodeling. 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 defining, in the same way, the smallest units of information, also known as granules, for each dataset. “Informally, a granule of a variable X is a clump of values of X that are drawn together by indistinguishability, equivalence, similarity, proximity, or functionality. For example, an interval is a granule.” [21]. The field of study that focuses on representing, constructing, and processing these information granules is called Granular Computing [22]. Assuming that dwellings, **EPC** observations, and complete buildings are represented in the same data model, meaning that an appropriate data fusion process is implemented, a relevant predictive model, which we can call, in this case, a metamodel, should now be constructed. Granular computing is multidisciplinary, but since we are dealing with geo-localised information, the natural field of research is geostatistics, which has been defined as “dealing with spatial processes indexed over continuous space” [23, p. 7].

Geostatistics. Probably for some historical reasons, as energy performance is usually a problem for building engineers, **EPC** is not treated as geolocated information in the literature. A novelty of this work is that we take into account the geostatistics aspect at every step of our research. We are trying to free ourselves from the important issue of detailed technical knowledge on a large scale. Instead, we favour a geostatistics approach to benefit from the geolocated nature of the **EPC** information. This aspect can be visualised in Figure 12, which presents the observed **EPCs** on a map of France. It is visible that the **EPCs** are better in the west than in the east, in the south than in the north, and in the cities than in the countryside. Similar phenomena appear at the city scale, for instance. From this geostatistics perspective, among other issues, the irreducible uncertainty of the granules’ positions (dwellings, buildings, etc.) in their underlying space impacts the use of traditional spatial interpolation models such as Kriging. In Chapter II, we propose a way to overcome the latter limitation and develop a comprehensive framework capable of handling data with uncertainty in the position of the observed objects while still allowing for the definition of an optimal linear predictor for spatial interpolation of **EPC** values. Another issue is that **EPC** is categorical while, by definition, spatial interpolation is more developed for quantitative variables. This is the object of Chapter III, which introduces a Kriging approach for fuzzy classification. The performance of this last model is assessed and compared with other models in Chapter IV.

5.2 Spatial Interpolation

As defined by [24], spatial interpolation “is a technique that uses sample values of known geographical points (or areal units) to estimate (or predict) values at other unknown points (or areal units)”. The same article presents a summary table of the major

spatial interpolation approaches. Spatial interpolation relies on a fundamental hypothesis that is similar to many prediction models: For a given random variable defined on each point of a territory (that is, a random field), we assume that the closer to each other two points are on the territory, the more similar are their associated variables. For example, we assume that the temperature is more likely to be similar between two points separated by a distance of 1m than between two points separated by a distance of 1km. In terms of statistics, it means that the correlation between two points is increasing as the distance between those points is decreasing.

A simple yet efficient way to approach the problem is to consider that the spatial dependency is limited to areal units that are independent of each other. This leads to the choropleth map, where a local statistic such as the median or the average is associated with each areal unit. The first ever produced choropleth map is given in Figure 13 page 48. This approach is still very useful. Satellite images, for instance, are grids of pixels whose area is a few square metres or square centimetres and whose value is the mean value of a certain light wave length intensity. Interestingly enough, the prediction of an EPC at the building level that we are pursuing in this research may lead to producing a variety of choropleth maps. This is the case, for instance, of the map presented in Figure 14 page 49.

If it is assumed that the observed random variable is continuous, then it would make sense to predict a continuous quantity, taking into account the observed aggregated values (means, for example). This constraint was called the pycnophylactic constraint by Tobler in 1979 [26]. There are a variety of proposed solutions to this problem, such as kernel smoothing. This is out of the scope of this work, since we are dealing with individual observations rather than observations of aggregated values. However, it is important to keep in mind that there is difficulty in predicting point values under constraints on aggregated values. This is presented in Chapter III.

Gaussian Process Regression [27], also known as Kriging, generates a domain of research *per se*. Kriging theory was first published by Matheron [28] based on Daniel Krige's master thesis. As mentioned before, 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 BLUP (Best Linear Unbiased Predictor) in the least squares sense for point spatial interpolation. Kriging has been first defined to interpolate point observations. It is intrinsically a regression algorithm, but in Chapter III, we propose a way to use it for Fuzzy Classification.

Areal interpolation, as defined by Lam [29], 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 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 that are continuously defined over space, but it may be more challenging to observe and model when dealing with areal data where granules can be of various sizes and shapes, sometimes uncertainly defined. Gotway and Young [30] highlighted the terms used to describe areal interpolation and its challenges: block Kriging, multiscale and multi-resolution modelling, the ecological inference problem, the **MAUP (Modifiable Areal Unit Problem)**, the scaling problem, the change of support problem, and the reduction of variance problem. We present below the aspects of Gotway and Young’s inventory 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 the feature at block (granule) level is the average of the block’s point features. Point-to-area prediction produces an estimate “identical to that obtained by averaging the point estimates produced by [Kriging]” [31], [23]. Kyriakidis [32] described a complete Kriging model for area-to-point prediction, proved that it is an optimal predictor, and sketched area-to-area prediction. Goovaerts [33] studied in depth the problem of estimating the variogram, that is, 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 **MAUP** for which a measuring approach has been recently proposed by Briz-Redon [34]. While the **MAUP** refers to correlation between output variables, the ecological inference problem is a result of the correlations at the 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 identically distributed 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 [35] using rectangular blocks at multiple scales.

Despite its limitations, the averaging method has proven to be effective for interpolating areal data. For example, [36] downscaled climate models and predicted soil wetness

using Kriging on the residuals of a generalised additive model [37]. Area-to-point Kriging, also called disaggregation, has also been implemented by [38]–[40]. Additionally, area-to-area Kriging (block Kriging) has been used effectively by [41] and has been applied to downscaling by [42] as well as [43]. 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” [44]. 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 the ecological inference problem belong to a family of problems related to the combination of different types of granules in the same model, e.g., observing dwellings and predicting buildings. These difficulties are gathered in the family of the change of support problems. Another particular change of support problem known as spatial misalignment arises when a given output variable is observed at multiple scales, including the point level. Indeed, systematic averaging makes points and areas different objects with different variability, different correlation structures, and therefore different predictors. The classification of problems such as “area-to-point” or “area-to-area” reflects this categorisation. Moraga *et al.* [45] have 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. This model, like other models derived from Kriging, considers blocks to be connected surface areas in \mathbb{R}^2 that need to be “discretised” [33] 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.

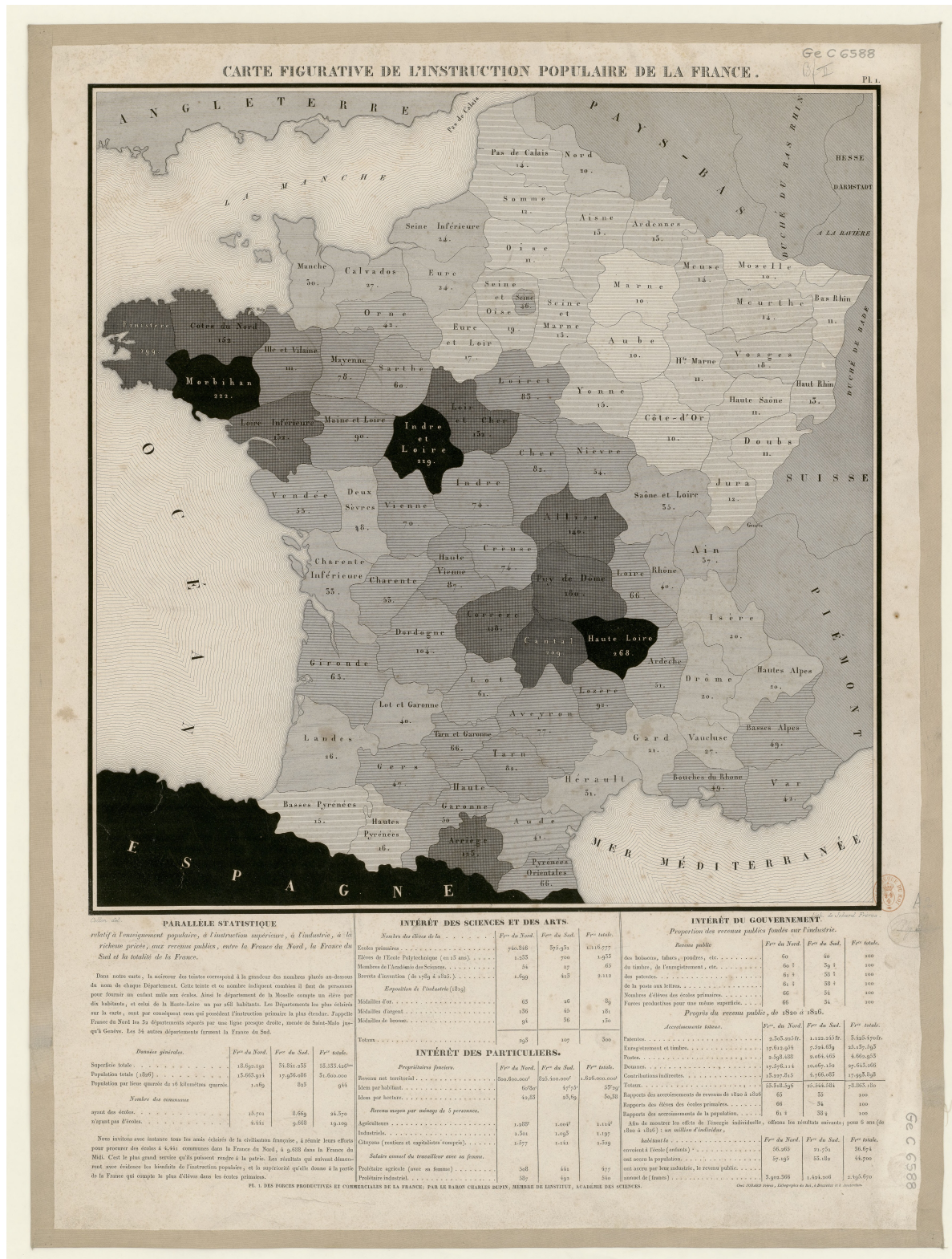


Figure 13 – The first choropleth map produced by Charles Dupin in 1826. It is a map representing literacy in France. Each areal unit is a department. The darker the area, the larger the associated value. This value represents “the number of persons required to provide a male child to the school” (sic). For instance, in the Loire department, there is one boy in school for 66 inhabitants on average.

LA RÉPARTITION GÉOGRAPHIQUE DES PASSOIRS ÉNERGÉTIQUES DANS LA COMMUNAUTÉ URBAINE

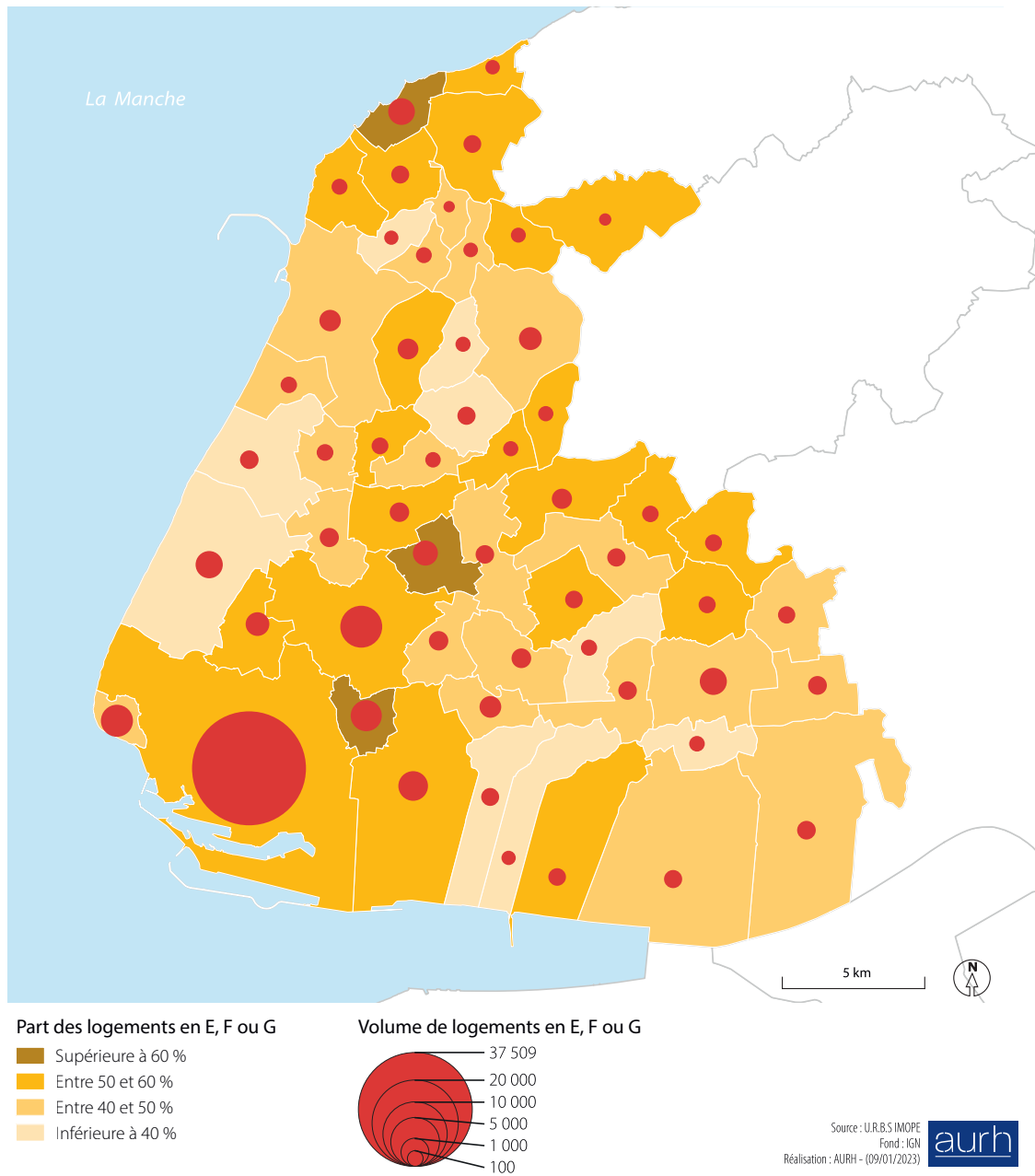


Figure 14 – A choropleth map representing the percentage of dwellings predicted to be of labels E, F, or G in the Le Havre metropolitan area. Areal units are municipalities. The colour of the areal unit represents the percentage of dwellings with labels E, F, or G; the darker the colour, the greater the percentage. The circles represent the number of dwellings with labels E, F, or G; the greater the radius, the greater the count [25]. This study is using the EPCs predicted using FKNN (Fuzzy k-Nearest Neighbours) such as presented in Chapter IV.

CHAPTER I

Pre-processing Data

1	Data Fusion and Uncertainties52
2	Imputation of Missing Values.55
3	From Categories to Numbers.59
4	Ranks and Quantiles60
5	Nearest Neighbours and Distance.61
	Conclusion.61

The starting point of this research is the release in open access of all the **EPC (Energy Performance Certificate)** observations in France, thereby defining a learning set from which we can expect to improve our knowledge of unobserved buildings. However, there are technical issues to be addressed in order to leverage these observations. This brief chapter is dedicated to explaining the nature of these technical problems and describing the solutions that are implemented in order to produce the data we have worked with.

1 Data fusion and Resulting Uncertainties

As explained in Subsection 5.1, data fusion is about combining multiple datasets from multiple sources, available at multiple scales. Each dataset describes small units of information known as granules. As far as the database of **EPCs** is concerned, granules are buildings or parts of buildings, typically one or more dwellings. On the other hand, the main source of information available at a large scale is the database of the **MoF (Ministry of Finances)**, which describes dwellings. And the goal of **U.R.B.S.**, through the **IMOPE (Inventaire Multi-Objets du Parc Existant – Multi-Object Inventory of the Existing Building Stock)** database [46], is to describe addresses, seen as a set of dwellings and assumed to be pointing at one only building or at least a very small number of buildings. Let us enumerate the different cases that may be encountered:

- **Addresses pointing to one only dwelling: houses.** Any house is uniquely described in the **MoF** database, but its **EPC** may not have been observed, or may have been observed one or multiple times. If it has been observed multiple times, only the last observation is kept as the most up-to-date one.
- **Addresses pointing to multiple dwellings: blocks of flats.** Each dwelling in the building is described separately in the **MoF** database. Those dwellings are merged to form a granule in the **IMOPE** database. If the associated address is not found in the **EPCs** database, the building is not observed. If the address is found in the **EPCs** database, multiple situations may occur:
 - At least one **EPC** observation exists for the whole building: the last one of the building’s **EPC** is kept as the most up-to-date one.
 - Only dwellings’ **EPCs** are available for this building: it is difficult to know which one of these observations is the most relevant to describe the building. At this point, the last observation is kept. In particular, we do not try to match an observed dwelling in a block of flats with a specific dwelling in the **MoF** database.

It is important to understand that this fusion process uses postal addresses. The addresses in the **EPCs** database are typed by the technician performing the diagnostic, while the addresses in the **MoF** database use a particular standard. This results in discrepancies that are difficult to straighten out. A specific tool has been developed in **U.R.B.S.** to tackle this difficulty. An example of difficult matching is given in Figure I.1: *28 & 30 Gabriel Vicaire Street \n Prosper Convert Street* (\n indicates a line break) is successively cleaned into *28 & 30 Gabriel Vicaire Street* and split into *28 Gabriel Vicaire*

Street and *30 Gabriel Vicaire Street*. After that, a match is found with *30 GABRIEL VICAIRE STREET* in the **MoF** database. If no direct matching is found, then a second series of tests is made, trying to match the addresses of each database with the **BAN** (**National Database of Addresses**). The resulting process is described in Figure I.2.



Figure I.1 – Uncertainty of the addresses in the EPCs database: An entry in the **EPCs** database writes *28 & 30 Gabriel Vicaire Street \nProsper Convert Street*. It appears that the technician intended to mean *28 et 30 Gabriel Vicaire Street, at the corner of Prosper Convert Street*. The **MoF** database shows a *30 GABRIEL VICAIRE STREET* but no 28. It appears that two houses were merged into a single one, and in this case, **MoF** keeps only one address. (Image: Map data ©2019 Google, annotated by the author.)

Similar problems appear for merging all data sources enumerated in Figure IV.1, page 141. The fusion process described above carries inherent uncertainties. We have tried to identify and classify uncertainties on the observed **EPCs**. This is a necessary step to understand the behaviour of the models we present in this work. I have made a presentation of the **EPCs** uncertainties on the occasion of an interdisciplinary meeting at Institut Fayol in March 2023. The slides are given in Supplementary Material 1, page 181, (in French). Uncertainties can be classified into:

- **Measurement uncertainties.** This is the uncertainty resulting from the level of information accessible to the technician, the quality of this information, and the general behaviour of the technician. It is also a result of the software used to compute the standardised energy consumption. This uncertainty is visible when two diagnostics are performed in the same dwelling and result in different labels, although no significant changes occur between both technicians' visits.

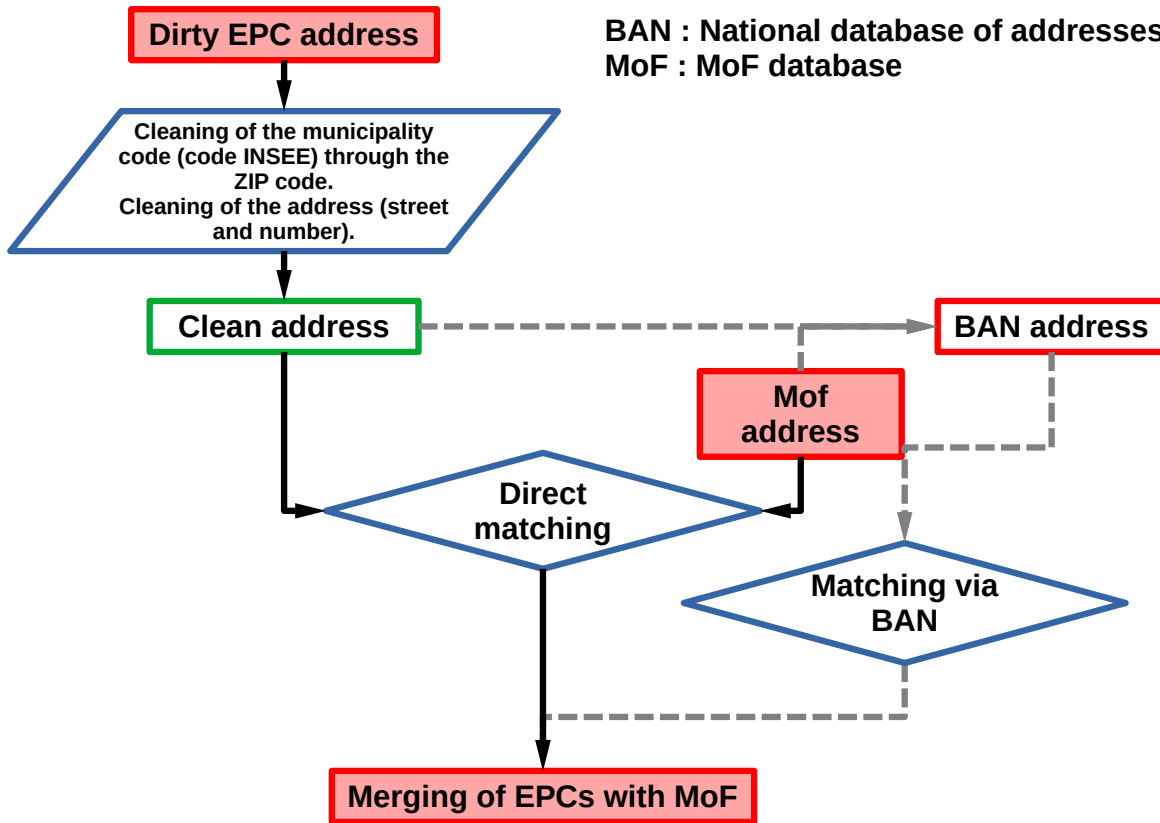


Figure I.2 – Representation of the algorithm used to match EPCs database addresses with MoF addresses.

- **Threshold effects.** A threshold is seen here as a regulatory number bounding the domain associated with an energy label. The existence of this limit has consequences for the observed labels. The threshold effect encompasses all phenomena that occur near this bounding value. Statistically, the threshold often results in an accumulation of observations on one side of the boundary and a deficit of observation on the other side, as shown in Figure IV.4, page 143. It may be interesting to observe the behaviour of the EPCs distribution when a smoothing procedure is applied in order to erase the threshold effects [47]. But the benefits of such a procedure are not clear, because the underlying idea is that some of the observed EPCs are biased and that this bias should be straightened out. But, aside from the fact that there is no actual proof of this bias, it seems really difficult to identify those biased observations and to determine the value of the bias. In this work, we have therefore assumed that the observed EPCs are sincere, and we have not made any assumptions about the technicians' behaviour.
- **Geolocation uncertainty.** It is only recently that the French government started a systematic inventory of the buildings on the national territory. Up until now, a dwelling has been associated with a land plot or a group of land plots. The exact position of the dwelling in a building and of the building in the land plot is only described in natural language in the sales contract, with sentences such as

“the dwelling is located on the 3rd floor of the building, which is in the northern part of the land plot.” This uncertainty is the reason why the **IMOPE** database is centred on addresses and not buildings as such. While it is possible to assert that two dwellings have the same address, it is very difficult to certify that they are in the same building, and it is even more difficult to draw this building on a map. That said, the reader should keep in mind that, from a statistical point of view, the most frequent type of building in France is the house, alone on a land plot. In which case, an address is associated with exactly one building. If there are dependencies on the land plot, such as a separate garage, uncertainty arises. As far as blocks of flats are concerned, in most cases, a land plot contains only one building containing all the dwellings having the same address. Uncertainty arises when there are two different addresses pointing to the same land plot or when two buildings have the same address.

Among these three uncertainties, measurement uncertainty is the easiest to handle, as it is natively included as an underlying assumption in most statistical models. For instance, in the case of Kriging, the concept of the nugget effect represents this uncertainty in the output value. We did not perform any thorough study to measure the uncertainty generated by the threshold effect, and, in any case, it is also an uncertainty on the output of a predictive model. Uncertainty about the position of the observation is more tricky. It amounts to considering that the input values are uncertain, which is not usual for predictive models. This is the problem we try to tackle in Chapter II, with the Mixture Kriging model.

2 Imputation of Missing Values

While we can associate with any address a land plot and, therefore, a geographic location with latitude, longitude, and altitude, most other features have some missing values. See Supplementary Material 11, page 264, for a dictionary of the main variables. For instance, the year of construction is missing for 0.5% of the addresses, which, at the scale of the country, represents more than 120 000 addresses. It was decided not to ignore those addresses and to predict an **EPC** even if some features are missing. Therefore, we designed an algorithm to impute missing values. At first, we imputed missing values in numerical variables with their local median value following the Algorithm 1. The same was done with the mode for categorical variables. The advantages of this algorithm are its stability and its simplicity of implementation. But it does not fulfil any condition of optimality. We hired an intern, Rafaël Quiblier, to specifically work on this issue under my supervision. The following paragraphs present a summary of this specific project [48].

An imputation algorithm is expected to preserve not only some statistical indicators of a variable but also its consistency and its variability [49]. The imputation by the local median surely preserves the median indicator, but it also reduces the variability of the variable by reinforcing the centre of the distribution, and it ignores any relationship with

other variables beside latitude and longitude. An algorithm known to be reasonably well performing on most datasets and fast is `missForest` [50], which is described in Algorithm 2. It does not fulfil any optimality criterion but a stability one, which is an interesting approach. In order to predict optimally and also preserve variables' variability, an algorithm called stochastic regression combines linear (or quadratic) regression with random imputation of some residuals to improve variability, as presented in Algorithm 3. The new algorithm we propose is based on linear regression using a series of models in order to make the best use of all available data. It is described in Algorithm 4. The same process is implemented with logistic regression for boolean variables.

It turns out that for quantitative variables, this iterative linear regression algorithm is performing better than both the local medians and `missForest` on quantitative variables, with a 25 % improvement for the **RMSE**. And the overperformance is very significant for boolean variables because the local median has a tendency to predict only the most frequent value in this case.

Algorithm 1 Iterative Local Imputation of Missing Values

Input: Data matrix D with missing values
Output: Data matrix D with imputed values

for each missing value $x_{i,j}$ in D **do**
 if non-missing values on the same land plot exist **then**
 $x_{i,j} \leftarrow$ median of non-missing values on the same land plot
 else if non-missing values in the neighbourhood exist **then**
 $x_{i,j} \leftarrow$ median of non-missing values in the neighbourhood
 else if non-missing values in the municipality exist **then**
 $x_{i,j} \leftarrow$ median of non-missing values in the municipality
 else if non-missing values in the region exist **then**
 $x_{i,j} \leftarrow$ median of non-missing values in the region
 else
 $x_{i,j}$ remains missing
 end if
end for
return D

Algorithm 2 missForest Imputation

Input: Data matrix D with missing values, where D_j is the column of the variable j .

Output: Data matrix D with imputed values

Initialise the missing values in D with initial guesses (e.g., column means)

$converged \leftarrow False$

$iteration \leftarrow 0$

while not $converged$ and $iteration < maxiter$ **do**

$iteration \leftarrow iteration + 1$

for each variable j in D **do**

$obs_j \leftarrow$ observed values of variable j

$miss_j \leftarrow$ missing values of variable j

Train a random forest RF_j on $(D \setminus D_j) \cup obs_j$

Predict $miss_j$ using RF_j and update D_j

end for

Calculate the difference between the new and previous imputed values

if difference is small enough **then**

$converged \leftarrow True$

end if

end while

return D

Algorithm 3 Stochastic Regression for Imputation

Input: Data matrix D with missing values, where D_j is the column of the variable j .

Output: Data matrix D with imputed values

for each variable j in D **do**

$obs_j \leftarrow$ observed values of variable j

$miss_j \leftarrow$ missing values of variable j

Train a linear model β_j on $(D \setminus D_j) \cup obs_j$

Predict obs_j using β_j and compute the resulting residual η_j

Predict $miss_j$ using β_j

For each predicted value of $miss_j$, add a randomly chosen residual in η_j

Update D_j

end for

return D

Algorithm 4 Iterative Imputation with Linear Regression

Input: Data matrix D with missing values, where D_j is the column of the variable j .

Output: Data matrix D with imputed values

for each variable j in D **do**

$obs_j \leftarrow$ observed values of variable j

$miss_j \leftarrow$ missing values of variable j

Step 1: Variable selection

Select the best variables $D_{\bar{j}}$ from $(D \setminus D_j)$ to predict obs_j by forward selection

Step 2: Iterative imputation

for each variable k in $D_{\bar{j}}$ **do**

Calculate $[miss_{jk}]$, the number of individuals for which both j and k are missing

end for

Order variables in $D_{\bar{j}}$ by increasing value of $[miss_{jk}]$

for each variable $k \in \{1, \dots, K\}$ in $D_{\bar{j}}$ **do**

Train a linear model β_k on $(D_1 \cup \dots \cup D_k) \cup obs_j$

Predict $miss_j \setminus miss_{1\dots k}$ using β_k

\triangleright *Because of the missing values in the predictors, each iteration imputes less missing values in j but predicts them more accurately than the previous iterations.*

end for

end for

return D

3 From Categories to Numbers

The models we are using, be it Kriging or **KNN**, require underlying distance functions between individuals (see also Section 5). Some of the addresses' features we use are categorical, such as the material of the main walls or the type of roof. At first, we started converting those categorical variables into so-called dummy variables: one boolean (0 or 1) variable for each type of wall, one boolean variable for each type of roof, etc. The advantage of this method is that it can refine the variable selection process to identify the most influential materials on the **EPC**. But the drawback is that it further increases the number of variables, which increases the computational complexity.

In the models presented in the following chapters, we convert categorical variables into numerical variables by ordering the different classes following the order of their associated mean energy consumption, as presented in Algorithm 5. This method is simpler to handle than the method using dummy variables because it does not increase the number of variables. And it performs well because we make sure that there is a positive correlation between the numerical variable that is created and the energy consumption.

Algorithm 5 Categorical to Numerical Conversion Based on Energy Consumption

Input: Data matrix D with categorical variable C and numerical energy consumption variable E

Output: Data matrix D with numerical variable C_{num}

Step 1: Calculate the mean energy consumption for each category

Initialise an empty dictionary $mean_energy$

for each unique category c in C **do**

$mean_energy[c] \leftarrow$ mean of E where $C = c$

end for

Step 2: Rank the categories based on mean energy consumption

Sort the categories in $mean_energy$ by their mean energy consumption values

Initialise an empty dictionary $rank_dict$

$rank \leftarrow 1$

for each category c in sorted $mean_energy$ **do**

$rank_dict[c] \leftarrow rank$

$rank \leftarrow rank + 1$

end for

Step 3: Convert categorical variable C to numerical variable C_{num}

for each row i in D **do**

$D[i][C_{num}] \leftarrow rank_dict[D[i][C]]$

end for

return D

4 From Numbers to Ranks and Quantiles

During the optimisation phase of the algorithms that we use, a crucial step is to assign weights to each input variable. In order for the algorithm to work this out properly, it is easier if the orders of magnitude of all variables are the same. People usually refer to this as rescaling or normalisation. The approach that is chosen in this work is a two-step transformation described in Algorithm 6.

Algorithm 6 Variables normalisation

Input: Data matrix D with numerical variables j

Output: Data matrix D with normalised numerical variables j

Let n be the number of individuals in D

for each variable j of D **do**

 Extract the column D_j from D

$sorted_indices \leftarrow \text{argsort}(D_j)$ \triangleright Get indices that would sort D_j

for $i = 0$ to $n - 1$ **do**

$D_j[sorted_indices[i]] \leftarrow \Phi^{-1}\left(\frac{i+1}{n+1}\right)$ \triangleright Update D_j with the Gaussian quantile of the ranks (starting from 1) divided by the number of observations

end for

end for

return D

Example 1. *Simplified example for the normalisation of a vector of dates.*

$$\begin{array}{ccccccc}
 \begin{pmatrix} 1982 \\ 2004 \\ 1953 \\ 1885 \\ 1975 \\ 1968 \end{pmatrix} & \longrightarrow & \begin{pmatrix} 5 \\ 6 \\ 2 \\ 1 \\ 4 \\ 3 \end{pmatrix} & \longrightarrow & \begin{pmatrix} 5/7 \\ 6/7 \\ 2/7 \\ 1/7 \\ 4/7 \\ 3/7 \end{pmatrix} & \longrightarrow & \begin{pmatrix} 0.57 \\ 1.07 \\ -0.57 \\ -1.07 \\ 0.18 \\ -0.18 \end{pmatrix} \\
 \text{raw values} & & \text{ranks} & & \text{pseudo-observations} & & \text{pseudo-quantiles} \\
 & & & & & & \text{(Standard Gaussian} \\
 & & & & & & \text{quantiles)}
 \end{array}$$

This normalisation process is presented in its mathematical form in Supplementary Material 2, page 195. It shows, in particular, that the transformation we propose here is invariant if the variable is transformed through a strictly increasing function. As a result of this normalisation process, all numerical variables have a normal distribution. This algorithm can be adjusted to handle boolean (0/1) variables in order to keep the order of magnitude of the normal distribution. The effect of this transformation can be observed.

For instance, in Figure I.3, page 63, three maps of Saint-Etienne represent the city living space either with true latitude/longitude, on the left, with pseudo-observations, in the middle, or with normalised pseudo-observations on the right. It is clear from those maps that the distribution of distances between two points in the territory is not the same on the three maps.

5 Nearest Neighbours and Distance

In Supplementary Material 4, page 201, we study the effects of the normalisation process on the nearest neighbours' distributions. For a better interpretation of the process, we represent in Figure I.4 the effect of normalisation on the variable `nlogh`, which indicates the number of dwellings in a building. The buildings with `nlogh = 1` are the houses. On the left, the bar plot represents the number of buildings for each value of `nlogh` between 1 and 250. There is the same distance between `nlogh = 0` and `nlogh = 25` as between `nlogh = 25` and `nlogh = 50`. However, there is a huge structural difference between houses and blocks of flats of 25 dwellings, while the difference between two buildings of 25 and 50 dwellings is less obvious. In the middle graphic, the same bars are distributed differently; the first few bars are distant from each other (buildings comprising 1 to 10 dwellings), while the bigger buildings are close to each other. On the right, the distribution has a longer tail but is still isolating houses from the other buildings.

In addition to this variable transformation, it is important, especially for **KNN** models, to choose a proper distance function. In 2014, Hassanat published an article [51] which introduces a new similarity measure, which, he claims, is a distance, and gives very good results with **KNN**. Unfortunately, the proof he gives of the fact that it is a distance is false. The main problem in proving that a measure is a distance is to prove the triangular inequality. We worked on that and finally proved that the Hassanat distance is indeed a distance. The proof is given in Supplementary Material 5, page 209. We tested multiple combinations of variable transformations with distances. It turns out that the Hassanat distance performs better with raw variables, but with normalised variables, a regular Euclidean distance seems to perform as well as a more elaborate distance like the Hassanat distance. This observation is qualitative, as a proper classification of such combinations between variables' transformation and distance would require much more work.

Still, it is important to keep in mind that we finally opted for a variables' normalisation and a Euclidean distance for **KNN**. For Kriging models, the distance is an underlying parameter of a covariance kernel. Covariance kernels are further discussed in the following chapters.

Conclusion

Data fusion, imputation of missing values, trans-typing of variables, and transformation of variables form a large part of the development time necessary to implement

the models presented hereafter. Data fusion is mainly the work of the data team in **U.R.B.S.** However, all the other tasks have been performed by the author.

Cartography of living areas in Saint-Étienne city

One dot represents 30m² of living surface.

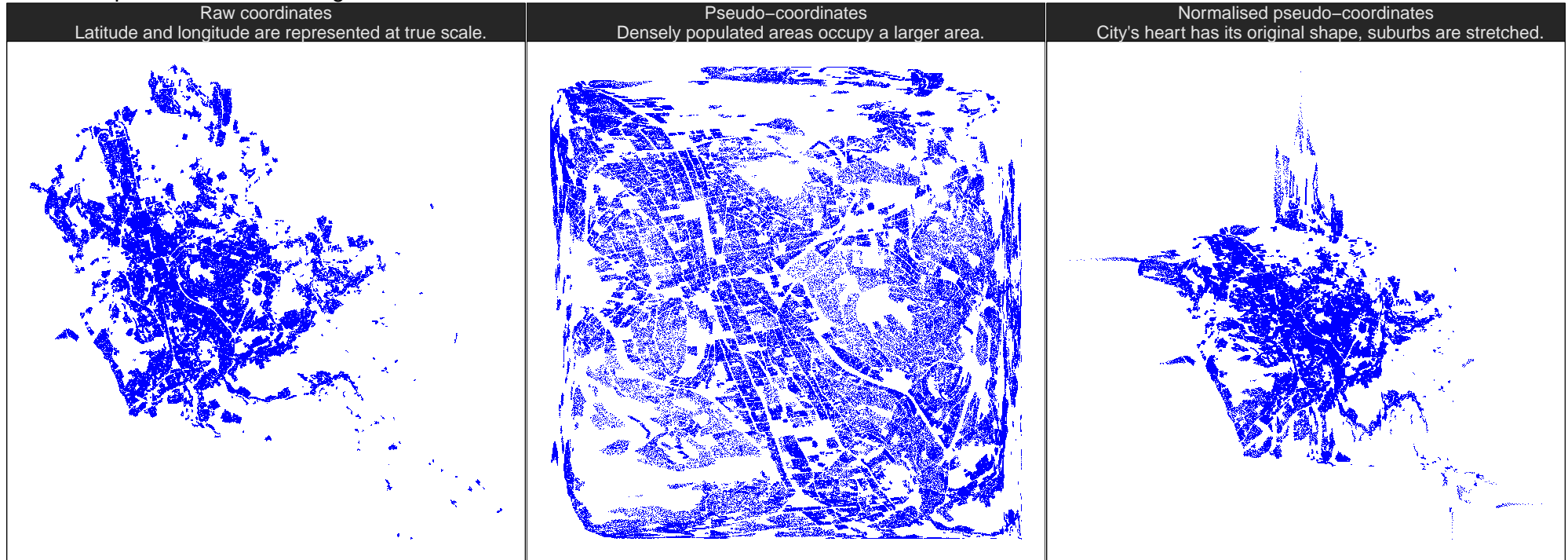


Figure I.3 – The above maps represent the city of Saint-Etienne. Each dot stands for 30m² of living space. On the left, points are located on a latitude-longitude system of axes. In the middle, the coordinates are transformed as pseudo-observations of the coordinates, as described in Example 1: Densely populated areas expand and fill the square, while less densely populated areas tend to shrink. On the right, the pseudo-quantiles are used: the heart of the city is not distorted much, but the extreme values are pushed away. These transformations induce different behaviours of the KNN algorithm or the Kriging algorithm as described in Supplementary Material 4, page 201.

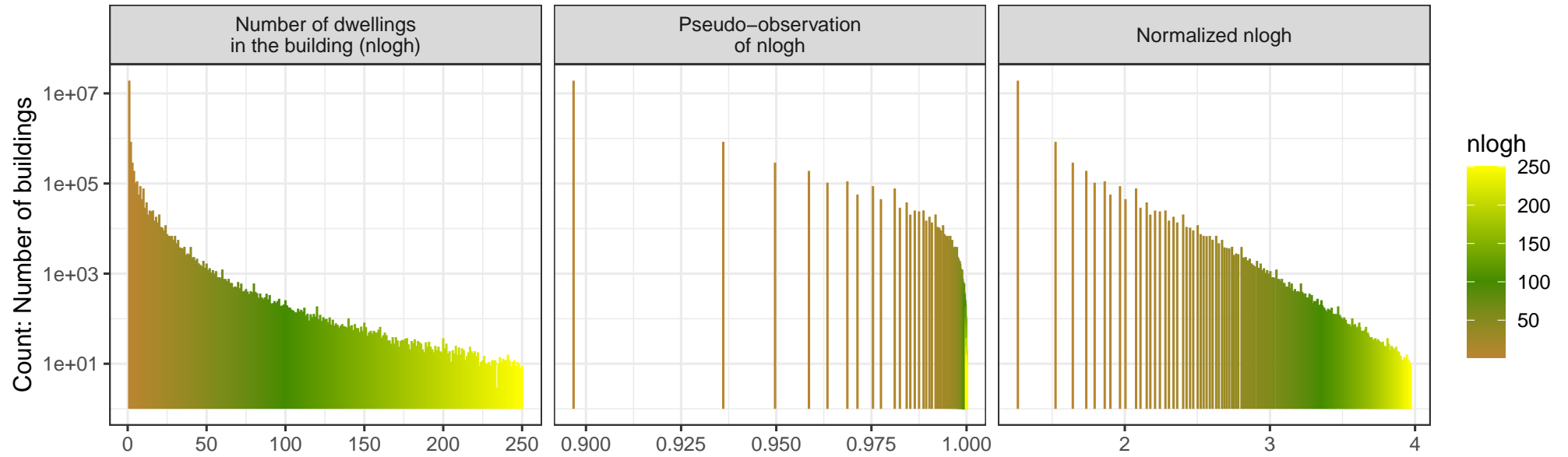


Figure I.4 – Effect of variables normalisation on an integer variable. These bar plots represent the number of buildings per value of the variable `nlogh`, which indicates the number of dwellings in a building.

Handling Multi-Scale Uncertainties

The content of this chapter is published, under the title *Predicting missing Energy Performance Certificates: Spatial interpolation of mixture distributions* in the *Energy and IA* journal, volume 16, May 2024 [52]. It is published in open access under the license Creative Commons Attribution 4.0 International.

	Résumé en français66
	Abstract67
1	EPC Prediction Problem67
2	Optimal Interpolation68
2.1	Data Model69
2.2	Mean and Covariances of Output Variables71
2.3	Best Unbiased Linear Predictor73
2.4	Particular Cases74
3	Illustration76
3.1	Unidimensional Case: Rounded Inputs.76
3.2	Unidimensional Case: Grains of Varying Size78
3.3	EPC Prediction81
	Discussion and Conclusion88

Résumé en français

La transition énergétique nécessite une connaissance approfondie des territoires, pour laquelle les décideurs s'appuient sur des données issues de multiples sources. Ces données, souvent restreintes par des politiques de confidentialité, ne renseignent pas toujours précisément sur des emplacements spécifiques. C'est le cas des bases de données de **DPE** (*Diagnostic de Performance Énergétique – Energy Performance Certificate*), qui sont anonymisées. C'est aussi le cas de covariables potentielles, comme celles issues du recensement, qui sont anonymisées à des échelles géographiques supérieures (**IRIS** ou carreau). Il s'agit donc de construire un modèle d'interpolation spatiale susceptible de rendre compte de cette incertitude de position. De façon générale, nous voulons un modèle qui tienne compte d'une incertitude sur les données d'entrée, liées à des sources multi-échelles, qu'il s'agisse de coordonnées géographiques ou pas.

Dans ce chapitre, nous traitons la prédiction des **DPE** non observés comme un problème d'interpolation spatiale, en utilisant une nouvelle approche nommée Mixture Kriging (krigeage de distributions de mélange). Il s'agit d'un modèle de régression qui prédit la consommation standardisée d'énergie primaire en kWh/m²/year. Cette méthode est adaptée, de façon générale, pour des variables qui sont observées en des lieux aléatoires. Dans notre cas, il prédit "un **DPE** observé quelque part sur une parcelle", sans que l'on sache exactement où sur la parcelle. La méthode permet de définir un prédicteur linéaire optimal non-biaisé. Bien que le cadre gaussien, habituel dans le krigeage, soit perdu, il est possible de calculer l'espérance de la variable à une position non observée et sa variance, c'est à dire une estimation de l'erreur possible.

Nous illustrons cette méthode à travers le cas d'une ville en France, montrant que Mixture Kriging produit des résultats prometteurs pour la prédiction des **DPE** au niveau des bâtiments. Ce résultat est important pour les décideurs qui cherchent à cibler les efforts de rénovation. Le modèle inclut également le co-krigeage, permettant l'utilisation de covariables, même en l'absence d'observations complètes, pour améliorer les prédictions.

De plus, l'implémentation de Mixture Kriging peut servir à contrôler la propagation de l'incertitude, comme la performance d'un produit industriel connaissant les incertitudes sur les paramètres de fabrication. Nous présentons des applications potentielles sur des données simulées, soulignant son utilité pour les processus multi-échelles, la régression d'une observation zonale à une prédiction ponctuelle, et la valorisation des données zonales, comme les données du recensement agrégées à l'**IRIS**, dans le contexte de la transition énergétique.

Abstract

Planning the energy transition requires decision-makers to have 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. The resulting data informs about given areas of space without a specific point location. Such is the case of **EPC (Energy Performance Certificate)**. **EPC** databases are released under specific constraints: anonymisation, geo-localisation with postal address, and missing details. In this chapter, we show 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 geo-localised information and predict its value at the building level.

Kriging methodology is applied to random fields observed at random locations to find a **BLUP (Best Linear Unbiased Predictor)**. 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 the 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 the propagation of uncertainty. We present potential applications based on simulated data.

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

1 Classifying the EPC Prediction Problem

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 9. Energy consumptions associated with dwellings, identified by their addresses, are inventoried in a database released in open access and mapped in Figure 12. A second database matches each address with a land plot. Finally, a third database gives the living area of every dwelling, be it a house or a flat, 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 throughout the present chapter.

As is explained in Section 5, **EPC** prediction usually requires a lot of technical details that may not be available on a large scale. 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.

From a geostatistics perspective, among other issues, the irreducible uncertainty of granules' positions (dwellings, buildings, etc.) 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 have already been identified and that several solutions have been proposed with their benefits and shortcomings.

A way to try and overcome change of support problems is to define a new data model for which outputs at the areal level do not require systematic averaging. In this regard, [53] 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 the areal level and the output at the point level. Therefore, it is eluding the question of consistency. In other words, it is not known whether the aggregation of cancer incidence predictions at a small scale would give the prediction of cancer incidence at a larger scale. Beside this limitation, the Hausdorff-Gaussian process does not solve the problem of uncertainty in the input data, that is found in the **EPC** prediction problem.

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 3.3, 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 a field of multidimensional random output variables is defined. 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** $\mathcal{G} = \{g_1, g_2, \dots\}$ of a territory χ is a finite set of grains of χ .

It is common in some application fields to use different terminology to talk about grains: blocks, pixels, and 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 disjoint, and those authors are not interested in the family itself, such as in [32]. The reader may find in Supplementary Material 8, page 227, 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, the 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 anonymised at the census tract level before being released. For instance, in a census table describing dwellings, a row describes a dwelling that exists on a certain census tract, but we don’t know exactly where it is on this tract. Then dwellings’ surface area is neither contin-

uous nor clearly geo-localised. 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 for 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:

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

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 (II.1). It means that there may be missing data in the output observations.

Example 2 (Buildings energy efficiency). *Keeping in mind that an **EPC** is given as an energy consumption in kWh/m²/year (see Figure II.1), 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 metre of a building on the plot. $Y(x)$ is the energy consumption associated with the square metre 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.*

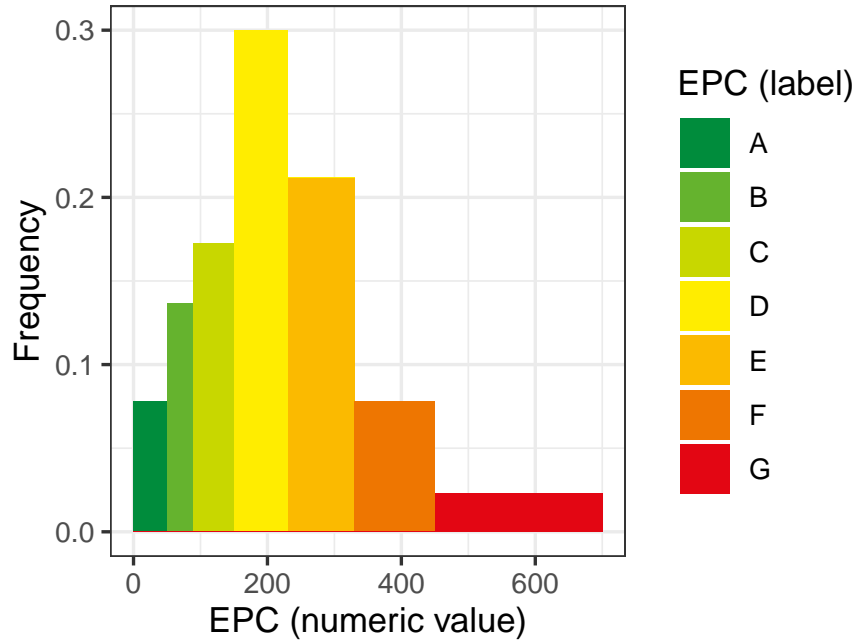


Figure II.1 – Bar plot of EPC labels frequencies among all EPCs collected in France between 2014 and 2021. Classes are highly heterogeneous.

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 $\{\mathbf{Y}(x) : x \in \chi\}$ 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 chapter, we assume that the first two moments of variables $\{\mathbf{Y}(x) : x \in \chi\}$, $\{X_g \in \chi : g \in \mathcal{G}\}$, 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:*

$$\mu_i(g) := \mathbb{E}[Y_i(g)] = \mathbb{E}[\mu_i(X_g)] \quad (\text{II.2})$$

(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 $k_{i,j}(x, x') := \text{Cov}[Y_i(x), Y_j(x')]$, we have:*

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

In particular,

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

Proof. (i) is a direct application of the conditional expectation formula

$$\mathbb{E}[V] = \mathbb{E}[\mathbb{E}[V | U]]$$

where $Y_i(g)$ is the result of conditioning $Y_i(x)$ with X_g . And (ii) is derived from the conditional covariance (variance) formula

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

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

Note that $\text{Cov}[\mu_i(X_g), \mu_j(X_{g'})] = 0$ 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 3. *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:*

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

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 *cdf (cumulative distribution function)* of the, possibly discrete, random variable X_g , for $i \in \{1, \dots, p\}$. 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

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

However

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

because the independence assumption does not hold any more. As a consequence, $\text{Var}[Y_i(g)] \neq \text{Var}[\bar{Y}_i(g)]$, even if $\forall i, j, g, g', \text{Cov}[\mu_i(X_g), \mu_j(X_{g'})] = 0$. The difference between a mixture and an average is retrieved here: $\text{Var}[Y_i(g)] \geq \text{Var}[\bar{Y}_i(g)]$.

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 $i \in \{1, \dots, p\}$, $Y_i(g)$ is to be predicted.

Denote:

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

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

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

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

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

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

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

$$\begin{aligned} \epsilon_i(g) &:= Y_i(g) - M_i(g) \\ c_{i,j}(g, g') &:= \text{Var}[\epsilon_i(g) \epsilon_j(g')] \\ v_i(g) &:= c_{i,i}(g, g) \end{aligned}$$

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 $g, g' \in \mathcal{X}$, and in particular for a single point $g = \{x\}$, for any $i \in \{1, \dots, p\}$, the weights $\boldsymbol{\alpha}_i^*(g)$ yielding the *BLUP* of $Y_i(g)$ and the associated cross errors are as follows:*

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

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

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

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

Proof of Proposition 2 is given in Supplementary Material 6, page 222.

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 $\mathbf{Y}(g) = (Y_1(g) \dots Y_p(g))^\top$ for the quadratic error $\mathbb{E}[\|\mathbf{Y}(g) - \mathbf{A}\underline{\mathbf{Y}}\|_2^2]$ 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, and 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 (II.4) 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:

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

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

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:

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

Then covariance between 2 grains outputs is:

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

Therefore, observations covariance matrix writes:

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

And covariance vector between observations and a new grain writes:

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

Typically, we can assume that $\mathbb{E}[\eta_{i,j}(X_g, X_{g'})] = \mathbf{1}_{\{i=j\}} \mathbf{1}_{\{g=g'\}} \eta_{i,i}(g, g')$. In which case \mathbf{K}_ϵ 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 [54]. In this sense, the Mixture Kriging results presented here can be seen as a generalisation of the Kriging interpolation.

It is also to be noticed that under certain assumptions, one can prove that if $M_i(g) = \mathbb{E}[Y_i(g) | \mathbf{Y}]$ then the cross error can also be viewed as a conditional expectation: $c_{i,j}(g, g') = \mathbb{E}[\text{Cov}[Y_i(g), Y_j(g') | \mathbf{Y}]]$. Details are given in Supplementary Material 7, page 225.

3 Illustration

3.1 Unidimensional Case: Rounded Inputs

A common issue when feeding statistical models with real data is the precision of the input data and its impact on a model’s performance. Usual applications of Kriging take this uncertainty into account, increasing output variances by a value that is known as the nugget effect [e.g., 55]. 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 the same point, and it is necessary to introduce a nugget effect in the model for the observations’ covariance matrix to be invertible. This nugget effect simulates 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 the R language, we simulate a 1-dimensional random field realisation with a Gaussian covariance kernel, whose parameters are detailed in Table II.1. x is discretised between 0 and 10 with step 0.05. We pick 8 points for observations as listed in Table II.2. These observations are plotted in Figure II.2. 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	1×10^{-9}	4	0.037	1.14
1	4	Mixture Kriging	1	0	4	0.027	1.18

Table II.1 – Parameters and performances of fitted models in the case of observations with rounded input. Note that the nugget effect for Kriging is the result of an optimisation process. For Mixture Kriging, the nugget is null by design.

Validation MSE: Mean Squared Error on validation set.

Total MSE: Mean Squared Error on the complete interval $[0, 10]$.

The Kriging model (Figure II.2 left) has repeated observations for $x = 1$ and $x = 3$. The learning set is used to fit a family of models with the same kernel parameters as those used for simulation plus a nugget effect among $(10^{-i})_{i \in \{1, \dots, 10\}}$. The nugget effect yielding the smallest **MSE (Mean Squared Error)** on the test set is selected. A new model is fitted with both learning and test sets using the same kernel and the

		Input			Output
Set	Label	Underlying x (True value)	Rounded x (for Kriging)	Grain (for Mixture Kriging)	y
Learning	o_1	0.55	1	$g_1 =]0.5, 1.5]$	0.923
Learning	o_2	0.85	1	$g_2 =]0.5, 1.5]$	1.005
Validation	o_3	1.65	2	$g_3 =]1.5, 2.5]$	1.127
Test	o_4	3.00	3	$g_4 =]2.5, 3.5]$	0.946
Test	o_5	3.45	3	$g_5 =]2.5, 3.5]$	0.801
Learning	o_6	7.20	7	$g_6 =]6.5, 7.5]$	0.337
Test	o_7	9.40	9	$g_7 =]8.5, 9.5]$	0.884
Validation	o_8	9.70	10	$g_8 =]9.5, 10]$	0.908

Table II.2 – Observations of the simulated Gaussian random field.

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 the formula given in Proposition 2.

Regarding Mixture Kriging (Figure II.2 right), grains $g_1 = [0.5, 1.5[$ and $g_3 = [2.5, 3.5[$ are observed twice each while the other grains are observed once each. The Mixture Kriging model can handle repeated observations by design. The uncertainty of the input results from the random position that generates the observation. The grain covariances are computed from the point covariances, as detailed in Proposition 1. The random positions $(X_{g_i})_{i \in \{1, \dots, 8\}}$ are assumed to be uniform on the points of the associated grains. Both the learning set and the test set are used to fit a model with the same kernel parameters as for simulation and without a nugget effect. Validation **MSE** and total **MSE** are computed for comparison with Kriging.

In this case, the mean prediction is almost the same for both models. But the predicted error variance (visible on the ribbons in Figure II.2) 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 zero.

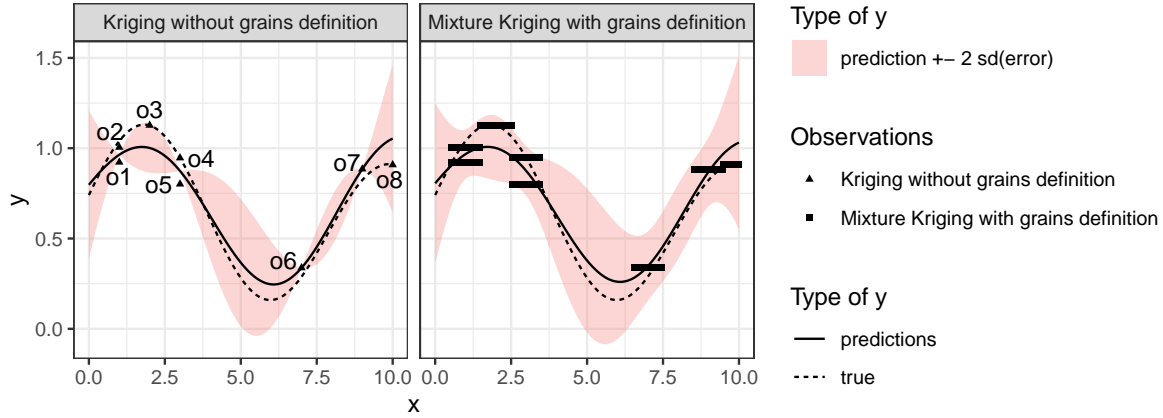


Figure II.2 – Rounded inputs. Left and right The dashed line labelled “true” shows a simulated uniform random field. The solid line labelled “predictions” shows the fitted model mean prediction (see Table II.1). The ribbon shows an interval of radius twice the root square of the estimated error variance.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

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 about the manufactured object induces some uncertainty about 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 : The manufactured object is produced exactly according to the design; the precision interval is restricted to a point.
- P_2 : The precision is the same for all designs; the associated interval is of fixed measure.
- P_3 : The larger is x , the larger is the uncertainty about the manufactured object,

which means that intervals' measures are growing with the design x .

All three models have a null nugget effect and a Gaussian kernel with the overall variance of \mathbf{y} on $\chi = [0, 4]$ as a variance parameter. The range parameter is optimised by minimising the mean squared error between \mathbf{y} and point prediction on χ . When grains are restricted to points (Figure II.3 top), we get the usual results on simple Kriging; in particular, predicted values are exactly interpolating observations. When grains are intervals of the same size (Figure II.3 middle), predicted values are not interpolating any more; the predicted error is not null on the grains but is also smaller than the value above far from the grains. In the bottom figure, the greater x , the greater the uncertainty on the manufactured object as compared to the design. The predicted error (ribbon) is increasing with the grains' diameter.

Granularity	Model properties			
	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 II.3 – Comparison of models quality for different types of granularities.

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

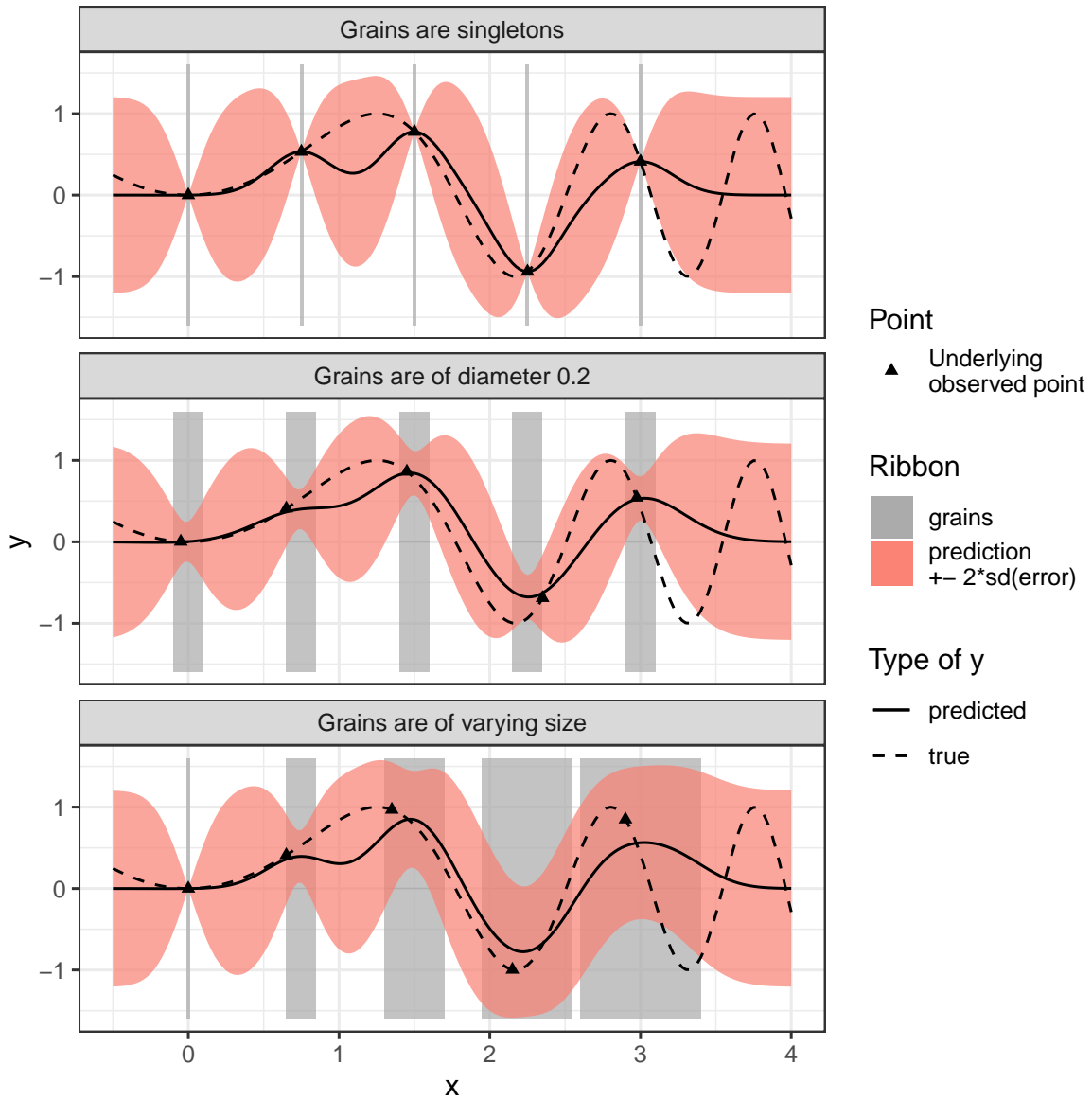


Figure II.3 – *Mixture Kriging and varying grain sizes.* **All:** The dashed line represents $y(x)$. The solid line is the mean prediction. The ribbon shows an interval centred on the mean prediction, of radius twice the square root of the predicted error variance. Vertical columns show the grains as x intervals. Black triangles show the underlying observed point (observed X_g and associated output).

3.3 EPC Prediction

The details of the practical implementation of Mixture Kriging are presented in Supplementary Material 9, page 231.

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

EPC is given as a numeric energy consumption per square metre and per year, which is associated with a letter ranging from A to G. A and B label the most energy-saving dwellings (less than $90 kWh/m^2/year$). F and G label the most consuming dwellings (more than $330 kWh/m^2/year$). We want to model a situation where we observe **EPC** with an uncertainty in the location of the observed dwelling on the land plot where it lies and where the observed dwelling cannot 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 II.1, observations are strongly unbalanced, meaning that labels A, B, F, and G are rarely observed while labels C, D, and E are very common. As a result, labels A, B, F, and G are difficult to predict, although they are more interesting for decision-makers. Therefore, we introduce the BA criterion. It is an asymmetric performance measure that focuses on good results [56] and it gives the same weight to each class. Denoting n_ℓ the number of observations with label ℓ and $n_{\hat{\ell},\ell}$ the number of predictions $\hat{\ell}$ with true label ℓ (true predictions of label ℓ), the balanced accuracy is given by the formula:

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

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

Let us consider the model M_1 such that:

- χ is the territory of an urban area in the French city of Angers in a 3 dimensional space where coordinates represent 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 metre 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 metre of living area. It has 3 coordinates. The set of all grains forms 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 of the observations' locations. On g , the output variable is defined as: $Y(g) = Y(X_g)$. By construction, Y is centred.
- A vector of observations of n distinct grains is given and denoted \underline{Y} .

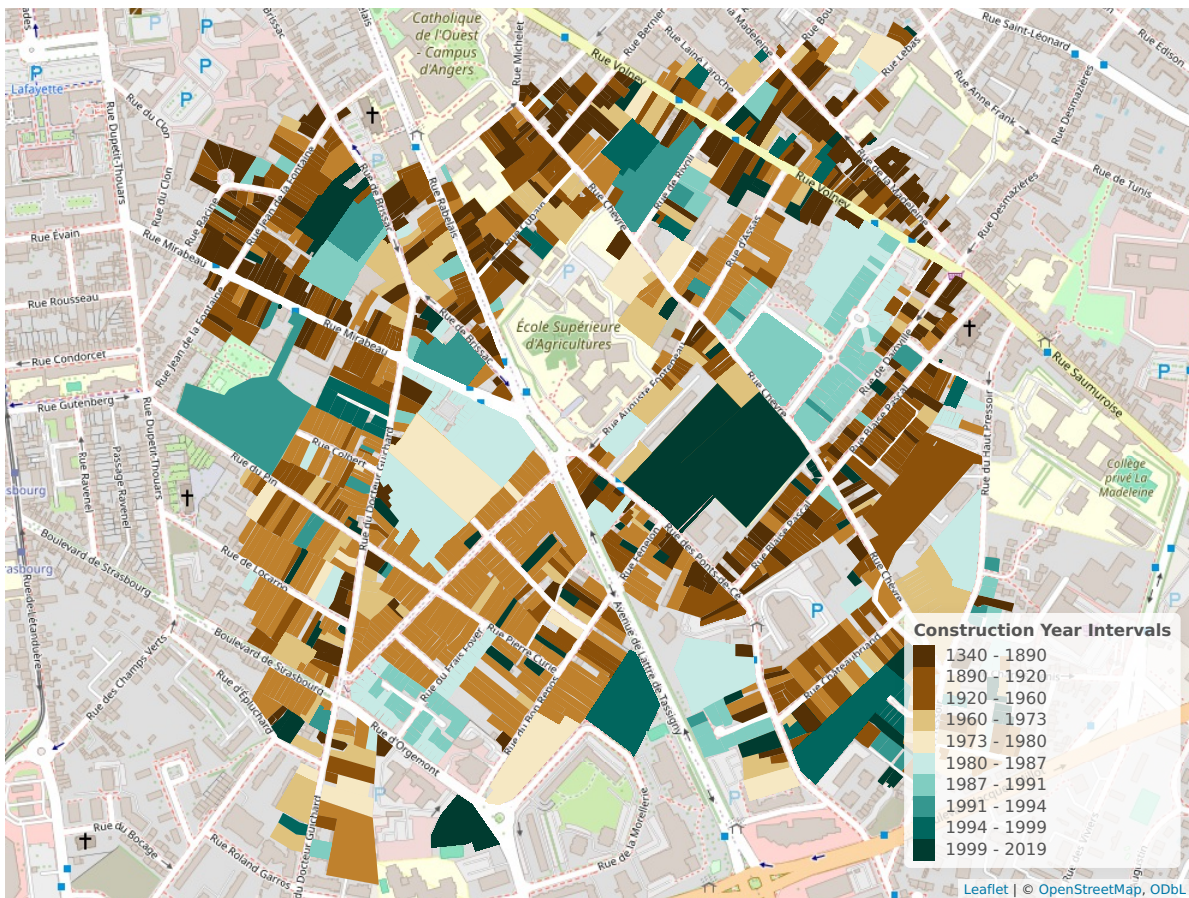


Figure II.4 – An urban area in Angers: latitude is the vertical dimension, longitude is the horizontal dimension, and construction year is given by the colour. The side of the square is 1 km. Construction years range from 1340 to 2019. The boundaries of the years' intervals are the years' deciles.

The granularity \mathcal{G} is mapped in Figure II.4. 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 II.5.

For this model, the following assumptions are made:

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

$$\text{Cov}[Y(x), Y(x')] = \sigma^2 \left(1 + \sum_{i=1}^3 \frac{|x_i - x'_i|}{\theta_i} \right) \exp \left(- \sum_{i=1}^3 \frac{|x_i - x'_i|}{\theta_i} \right)$$

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 $\Theta = (\theta_1, \theta_2, \theta_3)$ is the vector of length scale coefficients. Note that no nugget effect is required because the model takes into account the spatial uncertainty of the input by construction.

The Mixture Kriging predictor described in Subsection 2.3 is used to predict energy consumption at the plot level. It can be proved that without the nugget effect, the mean prediction, in the case of a 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 maximise the BA criterion of the predicted labels derived from the predicted energy consumptions. BA is computed using leave-one-out cross validation. Note that the leave-one-out cross-validation predictor that is derived from Proposition 2 is also linear and optimal for quadratic error. A code has been developed in **R** language to implement Mixture Kriging.

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

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

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

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

For $M2$, the Kriging predictor is used. $V = (\sigma^2, \theta_1, \theta_2, \theta_3, \epsilon^2)$ is chosen so as to maximise BA, the same way as for $M1$. The standard **R** package `DiceKriging` is used for prediction.

There are 365 observations on the given territory. The best parameters are estimated by optimising the performance indicator, Balanced Accuracy or Accuracy, computed by leave-one-out cross validation. All models $M1, M1'$ and $M2$ are optimised with

the genetic algorithm provided by **R** package `ga` parametrised with population size 50, elitism 5, maximum number of iterations 100, and maximum number of iterations without improvement 100. Other parameters are left as defaults.

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

*: These parameters are treated as constant parameters.

Table II.4 – Optimal parameters for $M1$ and $M2$.

Model	EPC int.		EPC num.				
	BA	Acc.	MAE	RMSE	MAE	RMSE	Range
Mixture Kriging ($M1$)	0.26	0.40	0.93	1.37	78.93	106.16	6.66
Mixture Kriging ($M1'$)	0.21	0.42	0.93	1.38	79.46	108.556	6.54
Kriging ($M2$)	0.19	0.38	0.85	1.22	72.22	92.98	2.59

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

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

BA: Balanced Accuracy.

Acc.: Accuracy.

MAE: Mean Absolute Error.

Range: Variance of the predicted values ($\times 10^3$)

RMSE: Root Mean Squared Error.

viewed as a measure of the predictions' range.

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

When looking at the models' performances, one should keep in mind that classes are heterogeneous, in the sense that the labels C, D, and E are much more frequent than labels A, B, F, and G. In particular, a model predicting only D labels would have a fairly good Accuracy of 0.40. But its Balanced Accuracy would be $1/7 = 0.14$, and it would be of no use for identifying energy sieves. Therefore, it is important to find a model that predicts a distribution of labels as close as possible to the expected distribution of labels.

With regards to the optimal parameters in Table II.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 II.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 II.6 and II.7. Indeed, the percentage of large errors (represented by the red area) is 3% with model $M1$ and 0.5% with model $M2$. We know that large errors have an important impact on MAE (Mean Absolute Error) 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.

True values	Predicted values							
	A	B	C	D	E	F	G	
A	2	1	3	2	2	0	0	10
B	1	3	1	9	2	2	0	18
C	1	3	25	26	15	4	0	74
D	3	5	21	80	33	5	1	148
E	4	2	12	36	36	5	1	96
F	0	3	2	4	5	3	0	17
G	0	0	0	1	1	0	0	2
	11	17	64	158	94	19	2	365

Table II.6 – Confusion matrix of $M1$ predictions.

True values	Predicted values							
	A	B	C	D	E	F	G	
A	1	0	3	5	1	0	0	10
B	0	2	1	11	4	0	0	18
C	0	1	13	48	12	0	0	74
D	2	1	19	94	32	0	0	148
E	0	1	9	56	30	0	0	96
F	1	0	2	11	3	0	0	17
G	0	0	1	1	0	0	0	2
	4	5	48	226	82	0	0	365

Table II.7 – Confusion matrix of $M2$ predictions.

True values	Predicted values							
	A	B	C	D	E	F	G	
A	0	2	2	5	1	0	0	10
B	1	0	3	9	3	2	0	18
C	2	2	23	29	14	4	0	74
D	1	6	17	91	28	2	3	148
E	1	6	14	31	36	6	2	96
F	1	0	3	8	2	3	0	17
G	0	0	1	0	1	0	0	2
	6	16	63	173	85	17	5	365

Table II.8 – Confusion matrix of $M1'$ predictions

These results suggest that Mixture Kriging ($M1, M1'$) predictions have an improved range as compared to Kriging ($M2$): the range of mean predictions by Mixture Kriging is greater than by Kriging. This allows better predictions for extreme labels A, B, F, and G. Despite having fewer parameters (ϵ^2 and σ^2 are regarded as constants), Mixture Kriging improves the BA, although it also leads to more frequent large errors. Mixture

Kriging accounts for uncertainty in the input data, eliminating the need to add a nugget effect. In this example, it avoids grouping predictions near the mean value (shrinkage) and yields a better BA as compared to Kriging, which requires the introduction of a nugget effect.

Among the Mixture Kriging models, as expected, $M1$ has a better Balanced Accuracy than $M1'$, and $M1'$ has a better Accuracy than $M1$. Other indicators are very similar, let alone the smaller variance of $M1'$'s predictions. Optimising parameters based on Balanced Accuracy forces the model to predict more often labels A, B, F, and G so that the distribution of predicted labels is very close to the distribution of observed labels, as can be seen in Table II.6. In our case, the confusion matrices show that this effect is positive for labels A and B, as more true A or B are predicted as A or B. But the effect of balanced accuracy does not bring benefits for labels F and G. On the contrary, it has a tendency to predict more F and G where the true label is D or E. A possible explanation for this moderate benefit of introducing the balanced accuracy is that we are missing some information. The moderate size of observations (365 individuals) makes it difficult for a model to discriminate between rare labels and frequent labels. For instance, there are only two observed G labels. One can expect a model learning from a larger number of observations to perform better. Moreover, in an area where buildings are old, for instance, our model cannot distinguish a building that has never been renovated from the others. It may be useful in further studies to introduce more variables, such as a comfort level. However, as discussed below, the proposed model is quite heavy in terms of computation resources; therefore, scaling up or adding variables has an important computational cost.

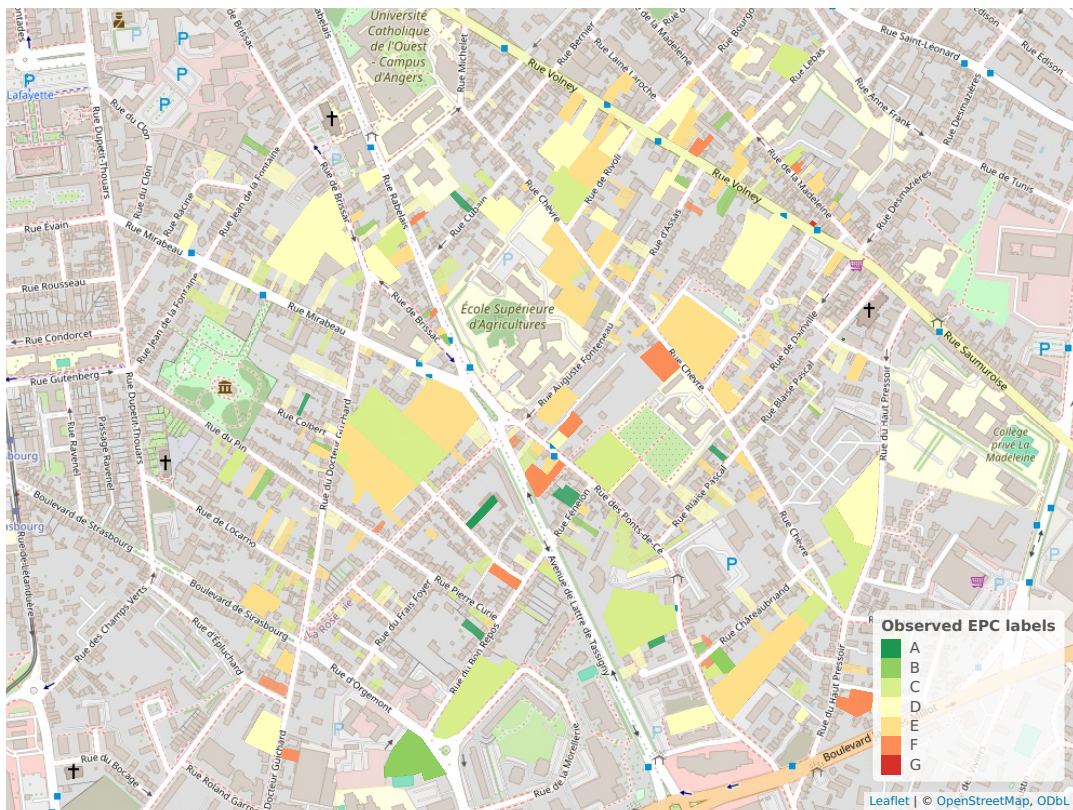


Figure II.5 – Map of the 365 observations. Each colour represents a label associated with a numeric value.

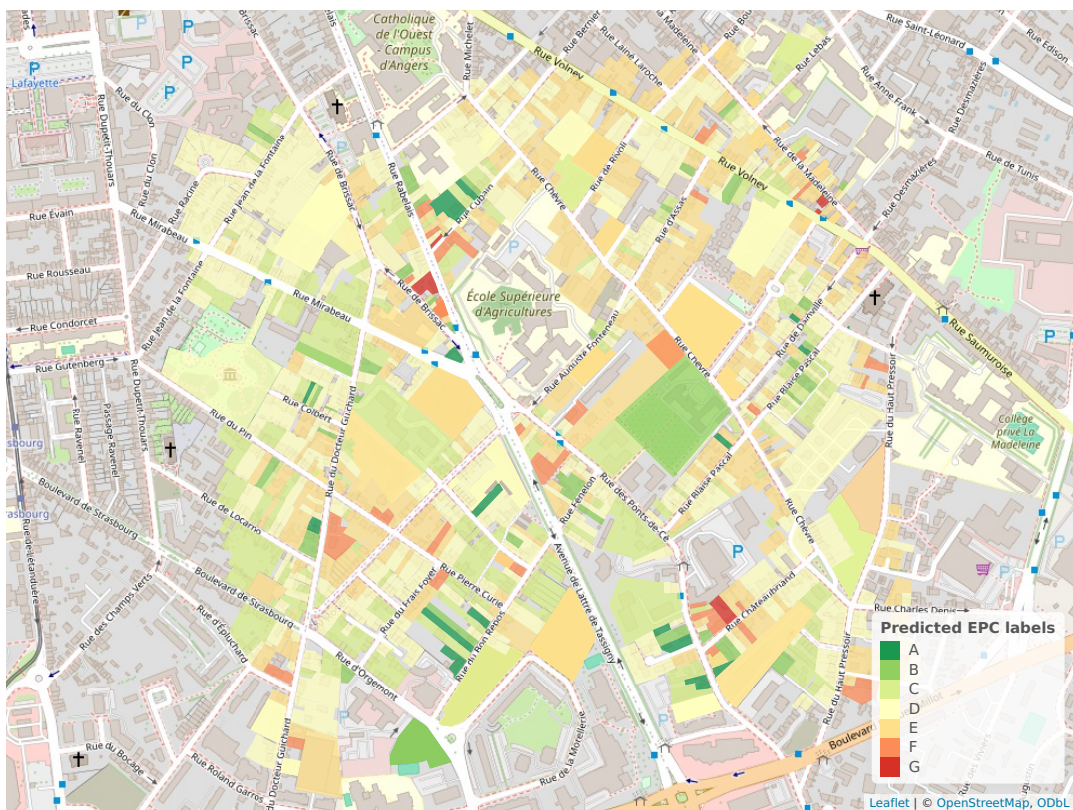


Figure II.6 – Map of all predicted labels derived from Mixture Kriging means.

Discussion and Conclusion

The issue of learning from and predicting areal data has long been a concern. The models that have been proposed mainly assume 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 **MAUP** (**Modifiable Areal Unit Problem**), 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 the **MAUP** effect [34]. Modifying territory partitioning [35] is also an effective solution for addressing variance reduction problems, but it is not always possible. Both Kriging and block-Kriging incorporate uncertainties on input and/or output values through the addition of a nugget effect to variances, thereby simulating the addition of 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 in 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 the areal level. It has been found that the resulting mixture distributions can be interpolated optimally, and the resulting **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 areas (grains) of any shape, size, or cardinality. Even points are acceptable. The terms “grains” and “granularity” have been introduced to describe these objects. The Mixture Kriging model has the ability to handle indifferently overlapping grains of even multiple granularities defining objects at multiple scales: land plots and census tracts, for instance.

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 shrinkage of the prediction 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 the point level, then it will also be the same as the covariance at the grain level, regardless of the grain’s shape. This implies that this model has no measurable **MAUP** effect.

The main computational distinction between block-to-block Kriging and Mixture

Kriging lies in the method of computing the observations' variance and the covariance between covariates associated with the same grain. This results mainly in the diagonal of the observations' covariance matrix being greater than what is found with Kriging. This is precisely the sought effect when introducing supplementary noise on the outputs (nugget effect) in Kriging for smoothing predictions. This explains why Mixture Kriging has smooth predictions but with limited shrinkage, hence a good performance with Balanced Accuracy. 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 in the input values without introducing nugget effect.

Regarding computational differences, it should also be noted that Mixture Kriging (like block-to-block Kriging) has a higher computational cost than Kriging, this cost is growing like the squared value of the density of points in the grains. This is an important limitation of the model. For instance, in the models $M1$, $M1'$ and $M2$ presented in Subsection 3.3, there are 395 observations. The Kriging model $M2$ requires $365 \times 366/2 = 66,795$ covariances to be computed. But the Mixture Kriging models $M1$ and $M1'$ require to compute 3,770,500,618 point-to-point covariance in order to compute the 66,795 covariances between grains. Scaling up the model may, therefore, be difficult. This computational complexity is highly dependent on the definition of the random position X_g for each grain and on its discretisation. For the above models, X_g is supposed to be uniform for all grains, and the number of discretised points is the number of square metres of living space on the grain. But any new model based on Mixture Kriging requires an appropriate definition of these random variables, depending both on the grains' geometries and on the studied output variable(s). Another limitation of the model is the difficulty of assessing its parameters, especially the range. It is difficult to compute a variogram because there is no natural definition of the distance between grains. Estimating the range is also possible by minimising an error measure, but this process requires computing numerous different models, which is costly, as mentioned above.

Despite its limitations, this new approach opens the way for implementing Mixture Kriging models with new datasets that have been impossible to fit in the usual Kriging framework. In particular, datasets that inform about granules that are uncertainly defined, such as dwellings, buildings, streets, human persons, and households. It can also be used for datasets informing about granules, which should have deterministic shapes or positions in the input space but come with a numerical uncertainty such as measure precision, rounding effect, observations' aggregations, or observations' anonymisation. 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 EPC. Results show that Mixture Kriging can be useful to improve the prediction of values far from the average and, in our case, the detection of energy-saving

homes. Future studies should test the upscaling feasibility of the already-developed model and the benefits of using covariates. We also study the possibility of developing a similar model with Universal Kriging.

Acknowledgements

The authors acknowledge support from the **U.R.B.S.** 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/>), **U.R.B.S.** enterprise (<https://www.imope.fr/>) and French National Agency for Research and Technology (<https://www.anrt.asso.fr/fr>).

CHAPTER III

Constrained Classification

The content of this chapter is published as a preprint, under the title *A Joint Kriging Model with Application to Constrained Classification* in the HAL repository [57]. It was submitted to and reviewed by the Statistics and Computing Journal. We have reviewed it and re-submitted it. The preprint is in open access under the license Creative Commons Attribution 4.0 International.

	Résumé en français94
1	Introduction95
2	Joint Kriging Model100
2.1	Optimal Weights Without Constraints101
2.2	Optimal Weights Summing to One102
2.3	Optimal Weights With Constraint on Predictions103
2.4	Optimal Weights With Affine Extension106
2.5	Joint Kriging Mean and Variance108
3	Constrained Classification110
3.1	Prescribed Constraints110
3.2	Application of the Joint Kriging Model111
3.3	Positivity Requirement113
4	Filling Cross-Covariances114
5	Numerical Illustrations117
5.1	A Simplified Toy Example117
5.2	A Multi-Output Time Series Example120
5.3	A Constrained Classification Example125
6	Conclusion132

Transition from Chapter II to Chapter III

EPC is a geolocated indicator The implementation we made of the Mixture Kriging model informs us that it is possible to treat **EPC (Energy Performance Certificate)** as geolocated information, meaning that latitude and longitude are informative for learning **EPC**. That is a major breakthrough that opens the way for multiple spatial statistics studies. Additionally, the Mixture Kriging model can handle any other input variable, such as the construction year, socio-economic indicators, or structural information about a building. Any research is embedded in a tradition, and energy efficiency studies have traditionally been part of thermal engineering studies for obvious reasons. Intersectional approaches proved that there is an overlapping of poverty, substandard housing, and overcrowded dwellings, which can be spatially described [58]. It suggests that buildings' energy efficiency should also follow this rule, but for some reason, up until now, the consequences of these observations have not been fully acknowledged. This research has emerged at the crossroads of spatial interpolation and thermal engineering. We gathered multiple reasons to believe that we ought to try spatial interpolation with **EPCs**: most French cities are built as a sequence of rings strongly associated with social criteria, the climate and the altitude are important factors determining the **EPC**, and the architecture is historically regionalised.

Strengths and limitations of Mixture Kriging Theoretically, Mixture Kriging is a powerful model. It has no **MAUP (Modifiable Areal Unit Problem)** effect. It includes the uncertainty of the buildings' positions. It can potentially take advantage of covariates, even if they are areal, such as census data, and even if they have missing observations. And it gives encouraging results. However, practically, the upscaling of the Mixture Kriging model is challenging due to its computational complexity. We thought we might be able to make simplification hypotheses to alleviate the computational burden. But approximations in Kriging can jeopardize the model because the covariance matrix must be symmetric and positive, while its eigenvalues may be quite sensitive to minor changes in the matrix. Even with a code in C++, the model was taking more than 6 hours to run on a part of Strasbourg city with 48 cores in parallel. And the more we simplified the model, the less performance we obtained.

From regression to fuzzy classification More generally, we found that regression might not be the best approach for determining the **EPC**. We had started the study with **EPC** observations generated before the legal framework was defined in 2021. At this time, **EPC** labels were defined through thresholding of the energy consumption only and the **GHG** emissions. After 2021, the problem was increasing in complexity since we had to define a regression model both for the energy consumption and the **GHG** emissions and then derive an **EPC** label. Such a process was triggering some error propagation issues as well as problems with the aggregation of models. Another important aspect to keep in mind is that the thresholds applied on the energy consumption and the **GHG** emission to define their respective labels are not linearly distributed. Therefore, the

same quantitative error can result in different behaviours as far as the label is concerned: an important error for energy-efficient dwellings, a minor error for energy sieves. Due to this, it is very difficult to control the errors on specific labels. These observations convinced us that we should look for a classification approach.

Concomitantly with the development of Mixture Kriging, I was also developing a production model for **U.R.B.S.**, which was a regression model based on **KNN**, and which raised similar issues. For the same reasons as mentioned above, we converted it to a classification model, which slightly improved the prediction performance. Working on recent bibliography on this topic, we found a seemingly efficient approach for fuzzy classification with **KNN** [59]. We implemented it and found a real benefit. We thought that if we could find a way to combine the benefits of fuzzy classification with the advantages of spatial interpolation, we might surpass all previous models. The outcome is presented in this chapter.

Résumé en français

Ce chapitre explore les techniques avancées d'interpolation de données, sous contraintes, par krigeage appliquées à la prédiction simultanée (jointe) de plusieurs variables. Le problème de Krigage abordé ici consiste à prédire les valeurs de plusieurs variables d'intérêt dans le contexte, par exemple, d'expériences coûteuses ou chronophages comme les diagnostics immobiliers, tout en respectant des contraintes spécifiques sur les valeurs prédites. Le modèle ouvre la porte à des applications en classification floue où l'on prédit un degré d'appartenance pour chaque modalité d'une même variable catégorielle, ce qui revient à prédire autant de variables que de modalités.

La méthode présentée, nommée Joint Kriging, utilise des poids de combinaison linéaires communs pour à toutes les variables de sortie simultanément, ce qui réduit considérablement le nombre de paramètres nécessaires par rapport aux approches traditionnelles où chaque variable de sortie est traitée séparément. On définit une seule combinaison linéaire qui est appliquée aux observations de la première variable pour prédire la première variable, aux observations de la deuxième variable pour prédire la deuxième variable, et ainsi de suite. Cette approche unifiée permet une meilleure cohérence et une augmentation de l'efficacité de l'interpolation, particulièrement lorsque les observations sont rares ou que les variables sont fortement corrélées, ce qui est le cas de la classification floue. Joint Kriging inclut le cas gaussien sans s'y limiter.

La prédiction jointe de plusieurs variables, décrite ci-dessus, permet l'intégration de deux types de contraintes. La première contrainte assure que la somme des valeurs de sortie prédites à un point donné soit toujours égale à 1. Cela permet de faire de la classification floue. La deuxième contrainte prescrit la moyenne de chaque variable de sortie pour une prédiction sur un ensemble de points donné. Cette contrainte permet d'intégrer des informations extérieures au modèle, par exemple, ou de faire des prédictions dans des scénarios favorables ou défavorables. Cela nous intéresse particulièrement dans le cas des **DPE** dont nous pouvons estimer la distribution sur un territoire donné.

En complément, une approche est aussi proposée pour l'interpolation affine, et non plus linéaire, dans le cas où l'on connaît un comportement des variables par défaut, en l'absence d'observations. La détermination des fonctions de corrélation est discutée y compris dans le cas complexe de coordonnées sur une sphère. Le modèle donne de très bon résultats en classification pour la prédiction de la magnitude des tremblements de terre, par rapport à près de 70 modèles testés.¹

Ce chapitre propose donc une approche robuste pour le krigeage simultané de plusieurs variables. Les hypothèses de simplification des approches classiques de co-krigeage permettent de prendre en compte d'autres informations comme la distribution globale des variables, sous forme de contraintes. Cette double approche de simplification d'une part, et prise en compte d'une plus grande complexité d'autre part, permet d'avoir un modèle efficace et performant.

1. La prédiction des DPE avec Joint Kriging est présentée dans le Chapitre **IV**.

Abstract

Interpolating or predicting data is of utmost importance in machine learning, and Gaussian Process Regression is one of the numerous techniques that are often used in practice. In this chapter, we consider the case of multi-input and multi-output data. A simple *Joint* Kriging model is proposed, where common combination weights are applied to all output variables at the same time. This dramatically reduces the number of hyperparameters to be optimised while keeping nice interpolating properties. An original constraint on predicted values is also introduced, useful for considering external information or adverse scenarios. Finally, it is shown that, when applied to membership degrees, the model is especially helpful for fuzzy classification problems. In particular, the model allows for prescribed average percentages of each class in predictions. Numerical illustrations are provided for both simulated and real data and show the importance of the constraint on predicted values. The method also competes with the 69 other models of an open real-world benchmark.

Keywords – Multi-output Kriging, Cokriging, Constrained classification, Spatial Prediction, multi-task Gaussian Process regression.

1 Introduction

Interpolating data is widely used in many fields of computer experiments. It is especially useful to predict the values of one or several variables of interest in the context of time-consuming or costly experiments. One considers here a Kriging interpolation problem on several output variables, with specific constraints on predicted values, so that applications to classification are possible. Let us detail the need to deal with such a problem.

Kriging on several outputs. Kriging, or Gaussian Process Regression is a method of interpolation, especially suited when there are only a few observations that have to be interpolated. It is widely used in many fields of Machine Learning, originally for geostatistical studies and spatial interpolation, but also for computer experiments in many domains (finance, industry, environment, etc.).

The most basic Kriging theory aims at predicting a single real-valued quantity of interest, the *output* (for instance, gold concentration in the ground), depending on some explanatory variables that are referred to as *input values* or *locations* (for instance, latitude, longitude, and depth). From a statistical point of view, the Kriging method is based on the best linear unbiased combination of observed outputs, with the assumption that observations are random variables whose correlation depends on locations. From a Gaussian random field point of view, in a Gaussian setting, the interpolation is the mean of a conditional Gaussian random field, with confidence bands derived from the variance of the conditional random field. An in-depth review of Gaussian Processes can be found in [54].

The method has several advantages. First, it is interpretable: the prediction is a

weighted average of observations, with quite a logical behaviour of the weights. Second, the method fully interpolates the data, that is, predicts exactly an observed output if one uses the same input values. And third, it not only gives a prediction but also confidence intervals for this prediction. Among limitations of the method and proposed extensions in the literature, one can cite the difficulty to handle numerous observations, see e.g. [60]–[62], and references therein, the difficulty to specify the covariance model and to estimate its hyperparameters [63], the difficulty to treat multivalued outputs [64].

In this chapter, we mainly consider this multivalued output problem, which is clearly of practical interest. One originality of this work is that this kind of multivalued interpolation is also applied to membership degrees in a classification setting. Moreover, a proposed simplification of the model is especially useful since it keeps the property of membership degrees summing to one. At last, another novelty is to consider a specific constraint on predicted values. As detailed below, it will allow for proportion constraints in a classification setting.

Constraints on predicted values. Such constraints can be useful when having external information, for adverse modelling, or for homogenisation needs, as illustrated below. The constraint we consider, for the model presented in this chapter, focuses on the average of predicted values.

It can be very helpful to prescribe a specific value for the average of predicted values. Let us instantiate some examples: Due to an industrial accident, one wishes to measure the pollution in the soil for different chemical products. Measures are done at some spatial places, but the number of measures is limited. One would like to infer the quantity of all chemical products everywhere in the soil. Knowing stockpiles of products before the accident, the total quantity of lost chemicals may be known for every chemical product. While Gaussian Process Regression is especially suited to predicting one product dosage in the soil, it has difficulty handling jointly a lot of products as it needs to model many cross-covariances. Furthermore, it cannot handle any constraint at all, like prescribing the sum of predicted values to be equal to the known quantity of spilled product. Another example is the case where one needs to build a prediction under an adverse scenario: even if the total quantity of lost chemicals is unknown, it can be useful to get an idea of the distribution of pollutants in an adverse case of massive loss.

In other investigations, there may be external knowledge to consider. For example, a regional study might want to be in line with some given national statistics if there is no reason that the regional statistics differ on average. One can observe data due to an exceptional situation (e.g., COVID), and one may want to use it knowing that the situation has returned to normal. Or one might want predictions over different years or over different regions to coincide, at least on average. For instance, one may want that some disease incidence prediction does not differ, on average, over different medical centres. Another situation is the following: imagine that one knows, under an arbitrage-free setting, that some predicted stock returns must be zero on average, or

imagine that the regulator wants to force a prediction under specific shock scenarios. Fairness constraints can also be introduced to limit unfairness in algorithmic decision-making [65]. Therefore, prescribing the average value of predictions is useful in multiple contexts, be it external information (known quantity of chemical, national statistic, etc.), adverse modelling (regulation, simulation under specific scenarios, etc.), or the need to homogenise results (over different regions, observed years, fairness constraints, etc.).

In the context of **EPC (Energy Performance Certificate)** prediction, it is possible to derive from a set of observations the overall distribution of **EPCs** over a given territory by drawing a representative sample of the territory. The ability to constrain a model to be consistent with this known distribution over a territory can be seen as a benefit for prediction reliability and as a way to balance out tendencies to predict too often or too rarely certain labels.

Constrained Kriging and fuzzy classification. Fuzzy classification is useful when an individual may simultaneously belong to multiple classes of a categorical variable, or when one is trying to predict a distribution of the probabilities for an individual to belong to each class of a categorical variable. In both cases, one usually builds a model to predict a quantitative variable associated with each class. The larger the quantitative variable, the more likely an individual is to belong to the associated class; see Example 4. These quantitative variables are called membership degrees. If those membership degrees are positive and sum to 1, they can be assumed to be probabilities.

Example 4. *Let us give a simple example of fuzzy classification. Assume that one wants to classify apples according to their colour. We assume that the only possible colours are green, red, and yellow. We define 3 numerical variables G , R , and Y associated with those colours. For a given apple, the triplet $(G = 0; R = 0.3; Y = 0.7)$ can be read in either of the following ways, depending on the chosen model:*

- *The apple is $3/10$ red and $7/10$ yellow.*
- *The probability for the apple to be green is 0, to be red is 0.3, and to be yellow is 0.7.*

The degree of membership of the apple to the class “yellow” is 0.7. The usual terminology is “membership degree”.

Applying multi-output Kriging on membership degrees has several advantages for fuzzy classification. One advantage is that the interpolation property can be preserved, which is not necessarily the case for other classification or clustering techniques like **KNN**: even at a location very close to a given observation, **KNN** can predict another class than the observed one. Another advantage of using a multi-output Kriging model on membership degrees is to get an estimation of the uncertainty of the prediction. For instance, at a specific location, one may predict 10% of the chance that the class is one, but one can also give a confidence interval for this quantity.

Applied to classification, a constraint on average predicted membership degrees is also

useful for the same reasons as above-mentioned. A specific simplification of multi-output Kriging will nevertheless be required to fulfil all considered constraints on predicted membership degrees.

Literature. The proposal here is to use multi-output Kriging with classification, under specific constraints on predicted values.

Regarding Multi-output Kriging, there is a huge amount of literature available. Reference books can be found on the topic, such as [66] and [67]. The modelling of cross-covariance functions is detailed in several papers, as in [68], [69]. Recent papers are dealing with inference and prediction using multitask Gaussian Processes [70], [71]. Co-Kriging techniques are built to treat several outputs, but there is usually one main output, and others are used to improve the prediction of the main considered output. Furthermore, all cross-covariances between outputs at different locations have to be modelled, which creates $O(p^2)$ covariance models, where p is the number of outputs [72], [73], [64]. While highly parametrised models are useful in many situations, the prediction quality relies on the proper specification of the model and on the estimation of its parameters. A fine model with the wrong parameters can sometimes be less efficient than a simpler model with more control over a few parameters [74]. One considers, here, a model where Kriging is applied to multivalued outputs in \mathbb{R}^p , but with a specific simplification leading to a single covariance function to tune instead of $O(p^2)$.

Some works can be found in the literature about clustering or classification under constraints. A survey on constrained classification can be found in [75, and references therein]. Some research works treat size constraints for clustering [76]–[78], while others treat the problem of fuzzy clustering with weights (membership degrees), as in the present work, see for example [79]. Fairness constraints are also considered in [65].

Regarding Kriging and classification, some works on classification using Gaussian settings can be found in a dedicated Chapter 3 in the book [54]. In particular, for binary classification, membership probabilities can be approximated by a sigmoid transformation of some latent Gaussian Process. The approach can be generalised to multi-class problems, and Bayesian inference can be conducted using analytic approximations of integrals, or solutions based on Monte Carlo sampling [80], [54, and references therein]. Other recent approaches involving Multi-task Gaussian processes, using several latent Gaussian processes, and Bayesian inference with approximations or sampling can be found in [81], [82].

Among works closer to what is proposed in the present work, Indicator Kriging aims at determining the **cdf (cumulative distribution function)** of an underlying random field at an unknown location, as a weighted average of indicators. It uses linear combinations of transformed observations too, but relies on a direct link between indicators and the underlying random field using thresholds. Hence, it does not seem to be directly suited to classify non-ordinal data (without any hierarchy between classes). It also requires the observation of the latent process that generates the indicators [83]–[86]. Extensions

like indicator co-Kriging require a large number of cross-covariances [87]. In the present paper, the proposed method can be applied to non-ordinal data, and does not require a specific model or thresholds between indicators of membership and an underlying real random field; furthermore, in a simplified setting, the whole method can also rely on a single covariance function.

Proposal To the best of our knowledge, the use of Kriging on several outputs with application to classification under constraints on predicted values has not been developed yet. We present, in this work, such a model. It involves a reduction of the number of hyperparameters. And it includes the possibility of considering specific constraints. The present approach directly yields closed-form formulas without the need for conditional density approximations or sampling. Such original constraints are not typically addressed by classical multi-output Kriging or Gaussian Process regression.

It seems to us that using multi-output Kriging on classification offers many modelling perspectives as well as practical results and performance. We will see that the proposed model competes with the best available methods on an open data set, among the 69 competitors of an open benchmark.

Structure The chapter is structured as follows. In Section 2, we define a simplified Kriging model that is suited for multivalued outputs. The model is detailed in three cases:

- with no specific constraint, similarly to Simple Kriging;
- with weights summing to 1, similarly to Ordinary Kriging;
- with constraints on weights summing to 1 and on average predicted values.

In each case, we derive optimal weights together with the prediction mean and variance. An extension using an affine prediction is also developed. In Section 3, the proposed interpolation technique is applied to membership degrees, and it is shown that it preserves useful basic properties for the prediction. Section 4 details strategies to fill the required covariance matrices and hyperparameters. In Section 5, numerical applications of the proposed interpolation technique are given. One considers in particular a minimal application on a toy example, an illustration on a multivalued time series on a real data set, and a more detailed real-world application on a classification problem. A conclusion closes the chapter.

Supplementary Material For more readability, all proofs are gathered in Supplementary Material 10, page 246. A list of symbols and notations is given in Chapter **Mathematical Notations**. All illustrations are generated with notebooks that are available as online supplementary material², in modifiable and executable format `.Rmd` and in already executed directly readable `.html` format [88]. Hence, the results are fully reproducible, and all specifications for drawing figures are easy to retrieve.

2. at <https://gitlab.emse.fr/marc.grossouvre/jointkrigingsupplementary/>

2 Joint Kriging Model

Let us consider a multivalued random field $\mathbf{Y}(x) := (Y_1(x), \dots, Y_p(x))^\top \in \mathbb{R}^p$, $x \in \chi$ where χ is a metric set of input points, typically $\chi = \mathbb{R}^d$. For the sake of clarity and using analogy with geostatistics, we will refer to x as *locations*, but χ may contain any explanatory variable. The components $Y_1(\cdot), \dots, Y_p(\cdot)$ will be referred to as the p considered *output variables*. Components of the random field $\mathbf{Y}(x)$ can be dependent. Furthermore, \mathbf{Y} or its components are not necessarily Gaussian. However, one assumes that first- and second-order moments exist. One considers here that $\mathbf{Y}(x) \in \mathbb{R}^p$ and $\chi = \mathbb{R}^d$, but other metric spaces would be possible as soon as expectation and covariances between $\mathbf{Y}(x)$ and $\mathbf{Y}(x')$ can be derived.

Given n observations of $\mathbf{Y}(x_1), \dots, \mathbf{Y}(x_n)$, we aim at predicting the values of the random field at some unobserved locations x_1^*, \dots, x_q^* , i.e., we aim at giving a predictor of $\mathbf{Y}(x_1^*), \dots, \mathbf{Y}(x_q^*)$. At an unobserved location x^* , we define the **Joint Kriging predictor** as a predictor $\mathbf{M}(x) = (M_1(x), \dots, M_p(x))^\top$ depending linearly on observations, where real coefficients apply jointly to all components of the observations:

$$\mathbf{M}(x^*) := \sum_{i=1}^n \alpha_i(x^*) \mathbf{Y}(x_i) \text{ where } \forall i \in \{1, \dots, n\}, \alpha_i(x^*) \in \mathbb{R}. \quad (\text{III.1})$$

These weights $\boldsymbol{\alpha}(x^*) := (\alpha_1(x^*), \dots, \alpha_n(x^*))^\top$ are optimised in order to minimise some error that we will detail later on, under various possible constraints. Now, defining the $p \times n$ matrix $\mathbf{Y} := [\mathbf{Y}(x_1), \dots, \mathbf{Y}(x_n)]$, Equation (III.1) also writes in a compact way:

$$\mathbf{M}(x^*) = \mathbf{Y} \boldsymbol{\alpha}(x^*). \quad (\text{III.2})$$

The main assumption here is that the weights are impacting all components the same way: the first component $M_1(x^*)$ is a linear combination of the observed first components, namely $Y_1(x_1), \dots, Y_1(x_n)$; the second component $M_2(x^*)$ is the same linear combination of the observed second components $Y_2(x_1), \dots, Y_2(x_n)$, etc. In other words, the weights affect jointly, or simultaneously, all the components of observed $\mathbf{Y}(x_i)$, $i = 1, \dots, n$, hence the chosen name of **Joint Kriging model**. We will see in Section 3 that this key **simplifying assumption** is especially useful for classification under constraints. It would be technically possible to release this assumption, e.g., by replacing weights $\alpha_i(x)$ by some $p \times p$ matrix for $i = 1, \dots, n$; one would get closer to some general co-Kriging model with $O(p^2)$ covariance models, but that is not the purpose of the present work.

Let us define the prediction error associated with a vector of weights $\boldsymbol{\alpha}(x^*)$, at a prediction location x^* . This loss is defined as the scalar value:

$$\Delta(x^*) := \mathbb{E} \left[\|\mathbf{M}(x^*) - \mathbf{Y}(x^*)\|_{\mathbf{W}}^2 \right], \quad (\text{III.3})$$

where $\|\mathbf{v}\|_{\mathbf{W}}^2 := \mathbf{v}^\top \mathbf{W} \mathbf{v}$ is a squared norm,

with $\mathbf{W} \in \mathcal{S}_n^{+*}(\mathbb{R})$.

For instance, if one changes the unit of the first output variable, say multiply it by 100, then it sounds logical that the resulting norm be unchanged. Thus, some weights' matrix may seem reasonable: an inverse covariance matrix as in the Mahalanobis distance, or a diagonal matrix of inverse variances, etc. For simplicity, the reader may imagine that all p output variables are already scaled and that \mathbf{W} is the $p \times p$ identity matrix.

The main difficulty is to derive the optimal weights $\boldsymbol{\alpha}(x^*)$ under the various constraints one would like to consider. At all prediction locations x_1^*, \dots, x_q^* , one thus aims at determining the optimal weights, gathered in a $n \times q$ matrix:

$$\mathbf{A} := [\boldsymbol{\alpha}(x_1^*), \dots, \boldsymbol{\alpha}(x_q^*)].$$

This is performed in the three following subsections under different constraints.

2.1 Optimal Weights Without Constraints

In this subsection, we define optimal weights that minimise the prediction error without supplementary constraints.

The following proposition expresses the weights such that $\mathbf{M}(x^*)$ is a **BLUP (Best Linear Unbiased Predictor)** of $\mathbf{Y}(x^*)$, in the sense of minimising the loss (III.3). The result looks exactly the same as in the simple Kriging model, but the components in the symmetric positive semidefinite matrix \mathbf{K} and in the vector $\mathbf{h}(x^*)$ here aggregate the values of all p observed output variables. One retrieves the usual Simple Kriging equations in the case where $p = 1$ and \mathbf{W} is the identity matrix.

Proposition 3 (Simple Joint Kriging weights). *The optimal weights $\boldsymbol{\alpha}(x^*)$ minimising the loss of Equation (III.3) are given by:*

$$\boldsymbol{\alpha}(x^*) = \mathbf{K}^{-1}\mathbf{h}(x^*) \in \mathbb{R}^n, \quad (\text{III.4})$$

or equivalently, using a matrix expression to predict simultaneously over the q locations,

$$\mathbf{A} = \mathbf{K}^{-1}\mathbf{H} \in \mathcal{M}_{n \times q}(\mathbb{R}), \quad (\text{III.5})$$

where

$$\begin{aligned} \mathbf{K} &:= \mathbb{E} [\mathbf{Y}^\top \mathbf{W} \mathbf{Y}] \in \mathcal{S}_n^+(\mathbb{R}) \text{ is assumed to be invertible,} \\ \mathbf{h}(x^*) &:= \mathbb{E} [\mathbf{Y}^\top \mathbf{W} \mathbf{Y}(x^*)] \in \mathbb{R}^n, \\ \mathbf{H} &:= [\mathbf{h}(x_1^*), \dots, \mathbf{h}(x_q^*)] \in \mathcal{M}_{n \times q}(\mathbb{R}) \end{aligned}$$

If furthermore for all output variables $j = 1, \dots, p$, for all location $x \in \chi$, $\mathbb{E}[Y_j(x)] = 0$, then $\mathbf{M}(x^*)$ is unbiased.

Proof. The proof is given in Supplementary Material 10.1, page 246. \square

Note that the matrix \mathbf{K} is necessarily a covariance matrix since it is symmetric positive semidefinite.

Here, we have weights applied jointly to all components, which leads to a simplified predictor. We operate with general assumptions that are the finite moments of orders 1 and 2. We do not require stationarity or a Gaussian setup. The assumptions we make are about the covariance function of a tunable weighted sum of components. The predictor can still take into account dependencies between those components and non-stationarities, even if the final results are approximated, because of its simplicity.

We will see that this simplified predictor is required to handle specific constraints, such as “higher scale constraints” in Section 3.1. More details on covariances in \mathbf{K} and \mathbf{H} will be given in a dedicated Section 4, where links with specific cross-covariance models of the literature are also presented.

2.2 Optimal Weights Summing to One

In this section, one considers an additional constraint. This constraint raises naturally when the random variables $Y_i(x)$ are not centred, and it leads to weights summing to one, as in Ordinary Kriging [89], namely for all x^* ,

$$\boldsymbol{\alpha}^\top(x^*)\mathbf{1}_n = 1 \quad (\text{III.6})$$

where $\mathbf{1}_n$ is a $n \times 1$ vector of ones.

The above constraint implies that the prediction is a weighted average of observations. Therefore, in the case where output variables’ expectation is constant over the territory χ and equal to $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^\top$, then the expectation of the prediction is also $\boldsymbol{\mu}$:

$$\begin{aligned} &\text{If (III.6)} \\ &\text{and } \forall x \in \chi, \mathbb{E}[\mathbf{Y}(x)] = \boldsymbol{\mu} \\ &\text{then } \mathbb{E}[\mathbf{M}(x^*)] = \boldsymbol{\mu} \\ &\text{and therefore } \mathbb{E}[\mathbf{M}(x^*)] = \mathbb{E}[\mathbf{Y}(x^*)]. \end{aligned}$$

Conversely,

$$\begin{aligned} &\text{If } \forall x \in \chi, \mathbb{E}[\mathbf{Y}(x)] = \boldsymbol{\mu} \\ &\text{and } \forall i \in \{1, \dots, p\}, \mu_i \neq 0 \\ &\text{and } \mathbb{E}[\mathbf{M}(x^*)] = \boldsymbol{\mu} \\ &\text{then (III.6)}. \end{aligned}$$

Hence, (III.6) is a very natural constraint. It does not imply, however, that $\mathbf{M}(x^*)$ is a convex combination of all $\mathbf{Y}(x_i)$, because some weights can still be negative.

Under this constraint of weights summing to 1, the following proposition gives the optimal weights. One retrieves similar formulae as in ordinary Kriging, but the involved elements in matrices \mathbf{K} and \mathbf{H} are different: they are computed taking into account all p mutually dependent output variables over all observations.

Proposition 4 (Ordinary Joint Kriging weights). *Under the constraint of Equation (III.6), the optimal weights $\boldsymbol{\alpha}(x^*)$ minimising the loss of Equation (III.2) are given by:*

$$\boldsymbol{\alpha}(x^*) = \mathbf{K}^{-1} (\mathbf{h}(x^*) + \lambda(x^*) \mathbf{1}_n) \in \mathbb{R}^n.$$

Equivalently, using matrix expressions, one gets

$$\mathbf{A} = \mathbf{K}^{-1} (\mathbf{H} + \mathbf{1}_n \boldsymbol{\lambda}^\top) \in \mathcal{M}_{n \times q}(\mathbb{R})$$

where

$$\begin{aligned} \delta &:= \mathbf{1}_n^\top \mathbf{K}^{-1} \mathbf{1}_n \in \mathbb{R}, \\ \lambda(x^*) &:= \frac{1}{\delta} \left(\mathbf{1} - \mathbf{1}_n^\top \mathbf{K}^{-1} \mathbf{h}(x^*) \right) \in \mathbb{R}, & \mathbf{K} &:= \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y} \right] \in \mathcal{S}_n^+(\mathbb{R}), \\ \boldsymbol{\lambda} &:= \left(\lambda(x_1^*), \dots, \lambda(x_q^*) \right)^\top \in \mathbb{R}^q, & \mathbf{h}(x^*) &:= \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y}(x^*) \right] \in \mathbb{R}^n, \\ \boldsymbol{\lambda}^\top &= \frac{1}{\delta} \left(\mathbf{1}_q^\top - \mathbf{1}_n^\top \mathbf{K}^{-1} \mathbf{H} \right), & \mathbf{H} &:= \left[\mathbf{h}(x_1^*), \dots, \mathbf{h}(x_q^*) \right] \in \mathcal{M}_{n \times q}(\mathbb{R}). \end{aligned}$$

If furthermore, for all output variables $i = 1, \dots, p$, for all locations $x \in \chi$, $\mathbb{E}[Y_i(x)] = \mu_i$, then $\mathbf{M}(x^*)$ is unbiased.

Proof. The proof is given in Supplementary Material Subsection 10.2, page 246. \square

The originality of the presentation of the result is that matrices can be expressed indifferently with compact expressions, using \mathbf{K} and $\mathbf{h}(x^*)$, or with more classical covariances, using $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{h}}(x^*)$, as stated in the following remark.

Remark 2 (Covariance matrices). *Let us define the true unknown values of \mathbf{Y} at all prediction points by $\mathbf{Y}^* := [\mathbf{Y}(x_1^*), \dots, \mathbf{Y}(x_q^*)]$. Assume $\mathbb{E}[\mathbf{Y}(x)] = \boldsymbol{\mu}$ for all $x \in \chi$. Furthermore, assume that either weights sum to one, that is, $\boldsymbol{\alpha}(x^*)^\top \mathbf{1}_n = 1$, or $\boldsymbol{\mu} = \mathbf{0}_p$.*

Then the matrices \mathbf{K} , \mathbf{H} , and the vector $\mathbf{h}(x^)$ can be replaced by*

$$\begin{aligned} \tilde{\mathbf{K}} &= \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y} \right] - \mathbb{E} \left[\mathbf{Y}^\top \right] \mathbf{W} \mathbb{E} \left[\mathbf{Y} \right] \\ \tilde{\mathbf{H}} &= \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y}^* \right] - \mathbb{E} \left[\mathbf{Y}^\top \right] \mathbf{W} \mathbb{E} \left[\mathbf{Y}^* \right] \\ \tilde{\mathbf{h}}(x^*) &= \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y}(x^*) \right] - \mathbb{E} \left[\mathbf{Y}^\top \right] \mathbf{W} \mathbb{E} \left[\mathbf{Y}(x^*) \right] \end{aligned}$$

everywhere in Proposition 4, without changing the optimal weights $\boldsymbol{\alpha}(x^*)$.

Proof. The proof is given in Supplementary Material Subsection 10.3, page 247. \square

2.3 Optimal Weights With Constraint on Predictions

The constraint we consider here is more original than the previous one: we would like that, given observations \mathbf{Y} , the average of the predicted values has some prescribed value. Formally, we introduce X^* , a random variable taking values in prediction locations

$\{x_1^*, \dots, x_q^*\}$, and we introduce the constraint:

$$\mathbb{E}[\mathbf{M}(X^*) | \mathbb{Y}] = \mathbf{m} \text{ for some } \mathbf{m} \in \mathbb{R}^p. \quad (\text{III.7})$$

This constraint relies on predicted values for a given set of observations. The idea is to force the optimal weights to take into account this *a posteriori* constraint. The interpretation of this constraint is that there is a secondary source of information that gives knowledge of the output expectation over the points to predict. A typical case would be one where the observations and the points to predict both form a representative sample of the territory.

Notice the importance of conditioning by \mathbb{Y} . Otherwise, if all $Y_i(x)$ are centred, then the constraint would not be possible to satisfy in general since all $\mathbf{M}(x^*)$ would be centred. We will see that this kind of constraint is particularly useful for fuzzy classification when one wishes to force the proportions of classes, whatever the observed values.

Gathering all predictors in a single matrix \mathbf{M} , we have:

$$\begin{aligned} \mathbf{M} &:= [\mathbf{M}(x_1^*), \dots, \mathbf{M}(x_q^*)] \\ \text{i.e. } \mathbf{M} &= \mathbb{Y}\mathbf{A}, \\ \text{and denoting } \pi_{x^*} &:= \mathbb{P}[X^* = x^*], \\ \boldsymbol{\pi} &:= (\pi_{x_1^*}, \dots, \pi_{x_q^*})^\top, \\ \text{we have } \mathbb{Y}\mathbf{A}\boldsymbol{\pi} &= \mathbf{m}. \end{aligned} \quad (\text{III.8})$$

One specificity is that the resulting weights in the matrix \mathbf{A} will have to be solved all at once for all q prediction locations. This is different from usual Kriging settings, where prediction locations can be treated separately if desired.

The constraint (III.7) is cumulated with the above constraint (III.6), so that the new system of constraints is:

$$\begin{cases} \mathbf{A}^\top \mathbf{1}_n = \mathbf{1}_q \\ \mathbb{Y}\mathbf{A}\boldsymbol{\pi} = \mathbf{m} \end{cases} \quad (\text{III.9})$$

In general, those constraints are linearly independent. The necessary and sufficient condition for those constraints to be linearly dependent is:

$$\exists \boldsymbol{\omega} \in \mathbb{R}^p, \text{ such that } \mathbb{Y}^\top \boldsymbol{\omega} = \omega_0 \mathbf{1}_n$$

It is the case, in particular, if \mathbb{Y} is a matrix of membership degrees in a fuzzy classification context, where $\mathbb{Y}^\top \mathbf{1}_p = \mathbf{1}_n$.

In the following proposition, we give the matrix of optimal weights \mathbf{A} when both constraints are considered at the same time: the constraint (III.7) on predicted values and the constraint (III.6) on weights summing to one. We treat both cases: when the system of Equations (III.9) is of full rank $q + p$ and when its rank is $q + p - 1$.

Proposition 5 (Joint Kriging weights under a predicted values constraint). *The Joint Kriging weights minimising the loss of Equation (III.3) under the constraint of weights summing to one of Equation (III.6), and prescribed average predicted values of Equation (III.7) write:*

$$\mathbb{A} = \mathbf{K}^{-1} \left(\mathbb{H} + \mathbf{1}_n \boldsymbol{\lambda}^\top + \mathbb{Y}^\top \boldsymbol{\lambda}' \boldsymbol{\pi}^\top \right) \quad (\text{III.10})$$

— If the system of Equations (III.9) is of full rank $q + p$, Lagrange multipliers are:

$$\begin{aligned} \boldsymbol{\lambda}' &= \frac{1}{\gamma} \left(\frac{1}{\delta} \mathbf{u} \mathbf{u}^\top - \mathbb{Y} \mathbf{K}^{-1} \mathbb{Y}^\top \right)^{-1} \left(\mathbb{Y} \mathbf{K}^{-1} \mathbb{H} \boldsymbol{\pi} + \frac{1}{\delta} \mathbf{u} \left(1 - \mathbf{1}_n^\top \mathbf{K}^{-1} \mathbb{H} \boldsymbol{\pi} \right) - \mathbf{m} \right) \in \mathbb{R}^p \\ \boldsymbol{\lambda} &= \frac{1}{\delta} \left(\mathbf{1}_q - \mathbb{H}^\top \mathbf{K}^{-1} \mathbf{1}_n - \boldsymbol{\pi} \boldsymbol{\lambda}'^\top \mathbf{u} \right) \in \mathbb{R}^q \end{aligned}$$

— If the system of Equations (III.9) is of rank $n + p - 1$, we remove arbitrarily the first constraint of the first equation and Lagrange multipliers become:

$$\begin{aligned} \boldsymbol{\lambda}' &= \left(\frac{\gamma_1}{\delta} \mathbf{u} \mathbf{u}^\top - \gamma \mathbb{Y} \mathbf{K}^{-1} \mathbb{Y}^\top \right)^{-1} \\ &\quad \left(\mathbb{Y} \mathbf{K}^{-1} \mathbb{H} \boldsymbol{\pi} + \frac{1 - \pi_1}{\delta} \mathbf{u} - \frac{1}{\delta} \mathbf{u} \mathbf{1}_n^\top \mathbf{K}^{-1} \mathbb{H}_1 \boldsymbol{\pi}_1 - \mathbf{m} \right) \in \mathbb{R}^p \\ \boldsymbol{\lambda}_1 &= \frac{1}{\delta} \left(\mathbf{1}_{q-1} - \mathbb{H}_1^\top \mathbf{K}^{-1} \mathbf{1}_n - \boldsymbol{\pi}_1 \boldsymbol{\lambda}'^\top \mathbf{u} \right) \in \mathbb{R}^{q-1} \\ \boldsymbol{\lambda} &= \begin{pmatrix} 0 \\ \boldsymbol{\lambda}_1 \end{pmatrix} \in \mathbb{R}^q \end{aligned}$$

$$\begin{aligned} \text{where } \pi_1 &:= \pi_{x_1^*} && \in \mathbb{R} , \\ \boldsymbol{\pi}_1 &:= \left(\pi_{x_2^*}, \dots, \pi_{x_q^*} \right)^\top && \in \mathbb{R}_+^{q-1} , \\ \mathbf{u} &:= \mathbb{Y} \mathbf{K}^{-1} \mathbf{1}_n && \in \mathbb{R}^p , \\ \gamma &:= \boldsymbol{\pi}^\top \boldsymbol{\pi} && \in \mathbb{R} , \\ \gamma_1 &:= \boldsymbol{\pi}_1^\top \boldsymbol{\pi}_1 && \in \mathbb{R} , \\ \delta &:= \mathbf{1}_n^\top \mathbf{K}^{-1} \mathbf{1}_n && \in \mathbb{R} , \\ \mathbb{H} &:= \left[\mathbf{h}(x_1^*), \dots, \mathbf{h}(x_q^*) \right] && \in \mathcal{M}_{n \times q}(\mathbb{R}) , \\ \mathbb{H}_1 &:= \left[\mathbf{h}(x_2^*), \dots, \mathbf{h}(x_q^*) \right] && \in \mathcal{M}_{n \times (q-1)}(\mathbb{R}) . \end{aligned}$$

Proof. The proof is given in Supplementary Material 10.4, page 247. \square

Note that it is also possible to compute a model with the only constraint $\mathbb{Y} \mathbb{A} \boldsymbol{\pi} = \mathbf{m}$, but without requiring weights summing to one. In view of further classification applications, we do not develop it here and keep both constraints.

Again, an originality is that the previous result can be expressed using compact expressions for \mathbf{K} and \mathbb{H} or more classical covariances, as stated in the following remark. The covariance functions that can be used to fill those matrices are detailed in Section 4.

Remark 3 (Covariance matrices with two constraints). *Under the assumptions of Remark 2 and using the same notations, the matrices \mathbf{K} and \mathbf{H} can be replaced by $\widetilde{\mathbf{K}}$ and $\widetilde{\mathbf{H}}$ everywhere in Proposition 5, without changing the optimal weights \mathbb{A} .*

Proof. The proof is given in Supplementary Material 10.5, page 252. □

Notice that the constraint on predicted values depends on prediction locations, which is the innovative aspect of this work. And the constraints become obviously too strong with a single predicted location; the prediction would be entirely prescribed. In practice, such a constraint is typically applied either on a given static grid of locations or on problems where the prediction locations are known (e.g., when the observed locations constitute a subset of some given finite set).

2.4 Optimal Weights With Affine Extension

A well-known characteristic of Simple Kriging is that the Kriging weights and the Kriging mean both tend to zero far from observed locations. In our setting, predicted values should be \mathbf{m} on average. Hence, one may desire that predictions return to \mathbf{m} far from the observed locations. This behaviour is similar to what one may expect from a Simple Kriging model applied on $\mathbf{Y} - \mathbf{m}$, where predictions' weights far from the observations tend to 0. However, it is important to keep in mind that since we want the sum of a prediction's weights to be equal to 1, it is incompatible with Simple Kriging with a null limit. We present in this section an affine extension of the Joint Kriging model, which is useful when one needs, at the same time, weights summing to 1 and a tunable behaviour far from the observations.

Up to this point, one has only considered linear predictors, where a predictor is a linear combination of observed responses $\mathbf{Y}(x_1), \dots, \mathbf{Y}(x_n)$, under various constraints. We now consider the case where the prediction involves one additional term.

The constraint on predicted values in the Joint Kriging model suggests that there is an external source of information giving a hint on the prediction. In addition to the observations, one knows that predicted values should be \mathbf{m} on average. This information may come, for instance, from some known overall statistics on the territory, some expert knowledge, or from an expectancy estimator. Let us denote \mathbf{Z} the $p \times 1$ random vector containing this external source of information.

With this in mind, we define an affine prediction:

$$\mathbf{M}^+(x^*) := \alpha_0(x^*)\mathbf{Z} + \sum_{i=1}^n \alpha_i(x^*)\mathbf{Y}(x_i), \quad (\text{III.11})$$

given $\mathbf{Z} = \mathbf{m}$, a constant term is included in the sum, hence the name ‘‘affine prediction’’.

The sum of weights constraint on the new vector $\boldsymbol{\alpha}^+ = (\alpha_0(x^*), \dots, \alpha_n(x^*))$ can be written:

$$\mathbf{1}_{n+1}^\top \boldsymbol{\alpha}^+(x^*) = 1.$$

This way, if the p components of \mathbf{m} and $\mathbf{Y}(x_i)$, $i = 1, \dots, n$ are probabilities summing to one, then the p components of the predictor $\mathbf{M}(x^*)$ will also sum to one.

For the second constraint on average predicted values, previously detailed in Equation (III.7), there is an implicit conditioning by $\mathbf{Z} = \mathbf{m}$. This constraint may write, with X^* a r.v. defined on $\{x_1^*, \dots, x_q^*\}$:

$$\mathbb{E} [\mathbf{M}^+(X^*) \mid \mathbf{Z} = \mathbf{m}, \mathbf{Y}] = \mathbf{m}. \quad (\text{III.12})$$

Finally, provided covariances between \mathbf{Z} and $\mathbf{Y}(x)$ are given for all $x \in \chi$, then the setting is the same as in previous Propositions 3, 4, and 5, except that one observation $\mathbf{Z} = \mathbf{m}$ is added in the vectors of observations $\mathbf{Y}(x_1), \dots, \mathbf{Y}(x_n)$. The covariance matrices are also updated. This is detailed in the following Proposition.

Proposition 6 (Affine version of predictors). *Assume that the following covariance vectors are given:*

$$\begin{aligned} \mathbf{P}^\top &:= \mathbb{E} [\mathbf{Z}^\top \mathbf{W} \mathbf{Y}] - \mathbb{E} [\mathbf{Z}^\top] \mathbf{W} \mathbb{E} [\mathbf{Y}], \\ \mathbf{Q}^\top &:= \mathbb{E} [\mathbf{Z}^\top \mathbf{W} \mathbf{Y}^*] - \mathbb{E} [\mathbf{Z}^\top] \mathbf{W} \mathbb{E} [\mathbf{Y}^*], \\ \sigma_Z^2 &:= \mathbb{E} [\mathbf{Z}^\top \mathbf{W} \mathbf{Z}] - \mathbb{E} [\mathbf{Z}^\top] \mathbf{W} \mathbb{E} [\mathbf{Z}]. \end{aligned}$$

Then, affine predictors corresponding to the simple unconstrained case, to the ordinary case with one constraint, and to the case with two constraints can be obtained by replacing \mathbf{Y} , \mathbf{K} , and \mathbf{H} by

$$\mathbf{Y}^+ = \begin{pmatrix} \mathbf{m} & \mathbf{Y} \end{pmatrix}, \quad \mathbf{K}^+ = \begin{pmatrix} \sigma_Z^2 & \mathbf{P}^\top \\ \mathbf{P} & \mathbf{K} \end{pmatrix}, \quad \mathbf{H}^+ = \begin{pmatrix} \mathbf{Q}^\top \\ \mathbf{H} \end{pmatrix},$$

in Propositions 3, 4, and 5 respectively.

Proof. The proof is straightforward, hence not appearing in the Supplementary Material. \square

Notice that the previous Proposition 6 can be easily extended to several sources of information: $\mathbf{Z}_1, \mathbf{Z}_2, \dots$. For the sake of simplicity, this is not developed here.

As detailed in Section 4, the matrices \mathbf{K}, \mathbf{H} can be derived from simple correlation functions. Now it remains to derive one expression for \mathbf{P} and \mathbf{Q} .

Remark 4 (Extra covariances for affine prediction). *Let $\mathbf{P} = (P_1, \dots, P_n)$, $\mathbf{Q} = (Q_1, \dots, Q_q)$, and $\mathbf{Z} = (Z_1, \dots, Z_p)$.*

Let us assume that \mathbf{P} and \mathbf{Q} do not depend on x_i nor on x_j^ , which means that the general source of information informs about the whole process, not about a particular location. then one can propose*

$$\begin{aligned} P_i &= \rho \sigma \sigma_Z, & i &= 1, \dots, n \\ Q_j &= \rho \sigma \sigma_Z, & j &= 1, \dots, q \end{aligned}$$

This happens, for instance, when $Y_k(x) = \rho \frac{\sigma}{\sigma_Z} Z_k + G_k(x)$, $k = 1, \dots, p$, where all $G_k(x)$ are independent of all Z_k .

The parameter $\rho \in [-1, 1]$ measures how redundant the information provided by \mathbf{Z} is, and can even be set to 0 if one considers that the external information source is completely independent of observations. The parameter σ_Z measures how certain the external information is: when σ_Z is high, the added information cannot be trusted, and one retrieves the linear predictor; when σ_Z is low, the added information is trustable, so that far from observed locations, $\mathbf{M}(x)$ gets nearer to \mathbf{m} . In practice, one can set $0 < \sigma_Z \ll \sigma$ to see the maximal difference with the linear predictor. One can even optimise this parameter σ_Z to smoothly switch from a linear to an affine model.

Other assumptions can be made, leading to different vectors \mathbf{P} and \mathbf{Q} .

Far from observations, all the weights in Equation (III.11) tend to predict the external source of information \mathbf{Z} . By choosing specific values of \mathbf{Z} , the default behaviour of the output variables is tunable. The affine predictor hence satisfies both weights summing to one and the default limit of the output variables, which is chosen here to be $\mathbf{Z} = \mathbf{m}$.

2.5 Joint Kriging Mean and Variance

In this subsection, we derive the mean predictor and the prediction error, assuming the optimal weights have been calculated with chosen constraints, as detailed in previous subsections.

Consider $\mathbf{M}(x^*)$ and $\boldsymbol{\alpha}(x^*)$ a Joint Kriging predictor and the associated weights with or without constraints. In the following, we call **Joint Kriging mean** the value of the predictor $\mathbf{M}(x^*)$ and **Joint Kriging variance** the value of the quadratic error $\Delta(x^*)$. Let us recall that:

$$\begin{aligned} \mathbf{M}(x^*) &:= \mathbf{Y}\boldsymbol{\alpha}(x^*), \\ \Delta(x^*) &:= \mathbb{E} \left[\|\mathbf{M}(x^*) - \mathbf{Y}(x^*)\|_{\mathbf{W}}^2 \right]. \end{aligned}$$

where $\mathbf{Y} = [\mathbf{Y}(x_1), \dots, \mathbf{Y}(x_p)]$ is the $p \times n$ matrix of observations. If $p = 1$ and if \mathbf{W} is the identity matrix, Joint Kriging mean and Joint Kriging variance are exactly the Kriging mean and the Kriging variance usually known in Kriging.

The following Proposition gives a closed formula to compute the Joint Kriging variance.

Proposition 7 (Joint Kriging variance with arbitrary weights). *Let $\boldsymbol{\alpha}(x^*)$ be any vector of weights, possibly satisfying supplementary constraints. The associated Joint Kriging variance writes:*

$$\Delta(x^*) = \boldsymbol{\alpha}(x^*)^\top \mathbf{K}\boldsymbol{\alpha}(x^*) - 2\boldsymbol{\alpha}(x^*)^\top \mathbf{h}(x^*) + v(x^*), \quad (\text{III.13})$$

or using a matrix expression, denoting $\boldsymbol{\Delta} := (\Delta(x_1^*), \dots, \Delta(x_1^*))^\top$, we get

$$\boldsymbol{\Delta} = \text{diag} [\mathbf{A}^\top \mathbf{K}\mathbf{A}] - 2 \text{diag} [\mathbf{A}^\top \mathbf{H}] + \text{diag} [\mathbf{K}^*],$$

$$\begin{aligned}
\text{where} \quad \mathbf{K} &:= \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y} \right] && \in \mathcal{S}_n^+(\mathbb{R}), \\
\mathbf{K}^* &:= \mathbb{E} \left[\mathbf{Y}^{*\top} \mathbf{W} \mathbf{Y}^* \right] && \in \mathcal{S}_q^+(\mathbb{R}), \\
\mathbf{h}(x^*) &:= \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y}(x^*) \right] && \in \mathbb{R}^n, \\
v(x^*) &:= \mathbb{E} \left[\mathbf{Y}(x^*)^\top \mathbf{W} \mathbf{Y}(x^*) \right] && \in \mathbb{R}
\end{aligned}$$

are assumed to be known.

$\text{diag}[\cdot]$ is the vector whose entries are the diagonal of the considered matrix.

Proof. The proof is given in Supplementary Material 10.6, page 252. \square

Note that the above Proposition 7 can be directly adapted to the affine case of Proposition 6 by replacing $\mathbf{Y}, \mathbf{K}, \mathbb{H}$ by $\mathbf{Y}^+, \mathbf{K}^+, \mathbb{H}^+$, v being unchanged: one can interpret the predictor to be a linear predictor with one more observation, with correct covariances.

As previously stated in Remarks 2 and 3, and using the same notation, one can replace $\mathbf{K}, \mathbb{H}, \mathbf{h}$ with $\widetilde{\mathbf{K}}, \widetilde{\mathbb{H}}, \widetilde{\mathbf{h}}$, provided that the following new quantities are defined:

$$\begin{aligned}
\widetilde{\mathbf{K}}^* &= \mathbb{E} \left[\mathbf{Y}^{*\top} \mathbf{W} \mathbf{Y}^* \right] - \mathbb{E} \left[\mathbf{Y}^{*\top} \right] \mathbf{W} \mathbb{E} \left[\mathbf{Y}^* \right] \\
\widetilde{v}(x^*) &= \mathbb{E} \left[\mathbf{Y}(x^*)^\top \mathbf{W} \mathbf{Y}(x^*) \right] - \mathbb{E} \left[\mathbf{Y}(x^*)^\top \right] \mathbf{W} \mathbb{E} \left[\mathbf{Y}(x^*) \right].
\end{aligned}$$

The result is stated in Remark 5 below. Hence, in practice, all these covariances can be filled using a given covariance function $k(x, x')$, under suitable assumptions, as detailed in Section 4.

Remark 5 (Covariance matrices in Joint Kriging mean and variance). *Under the assumptions of Remark 2 and using the same notation, the matrices $\mathbf{K}, \mathbb{H}, \mathbf{K}^*$, the vector $\mathbf{h}(x^*)$ and the scalar $v(x^*)$ can be replaced by $\widetilde{\mathbf{K}}, \widetilde{\mathbb{H}}, \widetilde{\mathbf{K}}, \widetilde{\mathbf{h}}(x^*)$ and $\widetilde{v}(x^*)$ everywhere in Proposition 7, without changing the Joint Kriging mean and variance.*

Proof. The proof is given in Supplementary Material 10.7, page 253. \square

Now, remark that Proposition 7 gives only an overall error:

$$\Delta(x^*) := \mathbb{E} \left[\|\mathbf{M}(x^*) - \mathbf{Y}(x^*)\|_{\mathbf{W}}^2 \right]$$

which is a weighted sum of errors over all components of $\mathbf{M}(x^*)$.

This is a strength of the method since the quantity to optimise is real-valued, which allows using standard covariance functions as detailed in Example 7. This is also an important limitation, because in practice, one surely needs prediction errors for each component of $\mathbf{M}(x)$:

$$\delta_i(x^*) := \mathbb{E} \left[\|M_i(x^*) - Y_i(x^*)\|^2 \right], \quad i = 1, \dots, p.$$

The following Proposition 8 shows that one can get this error $\delta_i(x^*)$ for each component $i = 1, \dots, p$. It relies on a supplementary assumption on the matrix \mathbf{W} , but this

assumption is only useful for determining the confidence bands for each component of the predictor $\mathbf{M}(x)$, not for computing $\mathbf{M}(x)$ itself.

Proposition 8 (Variance sharing). *Assume that transformed observations $\widetilde{\mathbf{Y}}(x) := \mathbf{W}^{1/2}\mathbf{Y}(x)$ are such that components of $\widetilde{\mathbf{Y}}$ are uncorrelated and bear the same share of the covariance function k , that is to say:*

$$\text{Cov} [\widetilde{Y}_i(x), \widetilde{Y}_j(x')] = \frac{1}{p}k(x, x')\mathbf{1}_{\{i=j\}}, \quad i, j \in \{1, \dots, p\}, \quad x, x' \in \chi,$$

Assume also that $\mathbb{E}[\mathbf{Y}(x)] = \boldsymbol{\mu}$ for all $x \in \chi$. Furthermore, assume that either the weights sum to one or $\boldsymbol{\mu} = \mathbf{0}_p$. Then, the local errors write:

$$\delta_i(x^*) = \frac{\sigma_i^2}{\sigma^2}\Delta(x^*), \quad i = 1, \dots, p. \quad (\text{III.14})$$

where $\sigma_i^2 := \text{Var}[Y_i(x)]$ is the variance of the component $Y_i(x)$, assumed to be constant over x .

Proof. The proof is given in Supplementary Material 10.8, page 253. □

The result of Proposition 8 states that for a well-chosen matrix \mathbf{W} , the error $\delta_i(x^*)$ is proportional to the unit global error $\sigma^{-2}\Delta(x^*)$: one has to apply the variance σ_i^2 of the component instead of the variance σ^2 of the aggregated weighted components.

3 Constrained Classification

In this section, we now apply multi-output prediction to membership degrees for fuzzy classification. We show that the Joint Kriging predictor, together with constraints on weights and predicted values, is especially suited to this task, and the above constraints make sense in a classification setting.

3.1 Prescribed Constraints

We aim here at proposing a fuzzy classification with a prescribed average of predicted membership degrees. Either because one requires that predicted values are overall distributed like the observed ones or because an external source of information gives the expected label percentages on a higher scale. It can be the case for a regional study, knowing some statistics at a national level. It can also be used for modelling adverse scenarios, as discussed in the [Introduction](#) of this chapter.

Consider a classification problem with p possible labels. Labels depend on some explanatory variables $x \in \chi$, so that one may observe labels $\ell(x_1), \dots, \ell(x_n)$ taking values in $\{1, \dots, p\}$. Assume that, at a prediction point x^* , a fuzzy classification method provides membership degrees of the p classes, gathered in a vector $\mathbf{M}(x^*)$. Consider q prediction points x_1^*, \dots, x_q^* , and bind all predicted membership degrees in a $p \times q$ matrix:

$$\mathbf{M} := [\mathbf{M}(x_1^*), \dots, \mathbf{M}(x_q^*)].$$

Components of $\mathbf{M}(x^*)$ should be positive and sum to one at each prediction point x^* , and the weighted average of predictions $\mathbf{M}(x_1^*), \dots, \mathbf{M}(x_q^*)$ is prescribed. As a result, we must satisfy positivity and the following system of constraints:

$$\begin{cases} \mathbf{1}_p^\top \mathbf{M} = \mathbf{1}_q^\top & \text{(probabilistic constraint)} \\ \mathbf{M}\boldsymbol{\pi} = \mathbf{m} & \text{(higher scale constraint)} \end{cases} \quad (\text{III.15})$$

where $\boldsymbol{\pi}$ and \mathbf{m} are two vectors of positive weights summing to one (i.e., two distributions). Hence, the set of predictions is subject to both constraints on the prescribed sum of rows and the prescribed sum of columns.

In Table III.1, we show an example of a confusion matrix, deriving from the previous constraints: the distribution of predicted classes is chosen to be identical to the one of actual classes, assumed to be given (or estimated). This is especially useful in situations where a class is dominant: all models tend to predict this dominant class, ensuring good accuracy, but the study of other classes thus becomes very difficult. The constraint forces the model to predict the right class probabilities, providing a way to study rarer classes.

		Predicted Classes				Sum
		A	B	C	D	
Actual Classes	A	52.30	37.00	20.50	10.20	120.00
	B	23.60	65.40	44.90	16.10	150.00
	C	37.00	38.40	72.90	21.70	170.00
	D	7.10	9.20	31.70	42.00	90.00
	Sum	120.00	150.00	170.00	90.00	530.00

Table III.1 – A Constrained Confusion Matrix: the sum of predicted classes, i.e., the sum of predicted membership degrees, is, here, equal to the sum of actual classes. For example, knowing that one must predict 120 labels A (higher scale constraint), the sum of predicted membership degrees for the actual class A is forced to be exactly 120.

3.2 Application of the Joint Kriging Model

We have considered specific constraints, such as higher-scale constraints. We show here that other predictors are usually not suited to satisfy such constraints, but that the Joint Kriging model naturally satisfies them.

Predictors of the literature may be unsuited to deal with considered constraints: at an unobserved location x^* , for a predictor $L(x^*) \in \{1, \dots, p\}$ of a label $\ell(x^*)$, the reader may convince himself that, for given probabilities p_j , $j \in \{1, \dots, p\}$, constraints on predicted classification such as

$$\mathbb{P}[L(X^*) = j \mid \text{observed labels}] = p_j,$$

are not so easy to handle, even if X^* is a uniformly distributed random variable over prediction points x_1^*, \dots, x_q^* . This is because such constraints usually act in a non-linear way on the predictor $L(X^*)$, and the predictor $L(\cdot)$ itself can be some complicated function of observed labels. Existing predictors, such as Indicator Kriging, may be unable to deal with such constraints. Furthermore, they may not be appropriate in cases where the considered labels do not correspond to ordinal classes.

Now, let us adapt the classification problem to the Joint Kriging model. In a classification problem, each label $\ell \in \{1, \dots, p\}$ can be converted into a $p \times 1$ vector of indicator functions, namely

$$\mathbf{Y} := \left(\mathbb{1}_{\{j=\ell\}} \right)_{j=1, \dots, p}.$$

This transformation is well known in the machine learning community as **label binarisation** (see also dummy variables or one-hot encoding), and is implemented in many languages. It also appears in some contexts of multiple outputs [68, Section 3.1]. With this representation, the equality $\mathbf{1}_p^\top \mathbf{Y} = 1$ is verified.

In practice, it is common to observe true label values depending on some explanatory variables $x \in \chi$. But it may also happen that one observes uncertain labels: multiple and distinct observed labels for the same $x \in \chi$, uncertainty in the value of x , etc. To handle this problem, one generalises slightly the previous label binarisation: one assumes here that observations consist in a distribution of possible labels, so that one observes n vectors $\mathbf{Y}(x_1), \dots, \mathbf{Y}(x_n)$, such that the components of each vector are summing to one: $\mathbf{1}_p^\top \mathbf{Y}(x_i) = 1$, $i = 1, \dots, n$. In other words, the p components of $\mathbf{Y}(x_i)$ represent the membership degrees of the p possible classes at an observed location x_i , $i = 1, \dots, n$. Using the previous notation, recall that $\mathbf{Y} := [\mathbf{Y}(x_1), \dots, \mathbf{Y}(x_n)]$, so that observed membership degrees satisfy

$$\mathbf{1}_p^\top \mathbf{Y} = \mathbf{1}_n^\top. \quad (\text{III.16})$$

Finally, using a Joint Kriging model, one can infer the membership degree of an unobserved location x^* using the predictor of Equation (III.1):

$$\mathbf{M}(x^*) := \sum_{i=1}^n \alpha_i(x^*) \mathbf{Y}(x_i) \quad (\text{III.17})$$

The next remark details the impact of both constraints, weights summing to one (probabilistic constraint) and prescribed average predicted values (higher scale constraint), in the particular setting of membership degrees that are summing to one. It shows that the Joint Kriging model naturally satisfies the considered system of constraints in the fuzzy classification setting.

Remark 6 (Constraints' impact). *Consider the membership degree assumption given in Equation (III.16), $\mathbf{1}_p^\top \mathbf{Y} = \mathbf{1}_n^\top$. Consider also the two previous constraints on weights and predicted values, namely the constraints of Equation (III.6) and Equation (III.7). Then the Joint Kriging model implies that:*

— *Predicted membership degrees are summing to one:*

$$\mathbf{1}_p^\top \mathbf{M}(x^*) = 1,$$

for all prediction points $x^ \in \chi$. In particular, $\mathbf{1}_p^\top \mathbf{M} = \mathbf{1}_q^\top$.*

— *The average membership degree over prediction points can be chosen:*

$$\mathbb{E}[\mathbf{M}(X^*) \mid \mathbb{Y}] = \mathbf{m},$$

where \mathbf{m} is a prescribed average of predicted membership degrees of each class, with $\mathbf{1}_p^\top \mathbf{m} = 1$, and X^ a random variable over all prediction points.*

Proof. The proof is given in Supplementary Material 10.9, page 255. □

3.3 Positivity Requirement

As noticed before, although predicted membership degrees are summing to one, there is no guarantee of positivity for the predicted values yet.

For hard clustering, it should be noted that, even without a positive weights requirement, predictions can still be used, as they can be interpreted as more general membership scores. It is quite natural to predict a class by selecting the one with the highest membership score. As an example, in a two-class problem, this leads to choosing one class or another if the associated prediction is greater than 0.5 or not, and even more so when the prediction is below 0 or greater than 1.

For fuzzy clustering, it is, of course, highly desirable to force the positivity of the Kriging weights. This way, when summing to one, the weights belong to a simplex, ensuring that predictions can be seen as probabilistic membership degrees. In practice, as is the case with numerous machine learning methods, a post-treatment of results may be required in specific cases involving negative membership degrees (see, e.g., the use of the softmax function with neural networks). With such a positivity requirement, the prediction falls within the framework of compositional data analysis, often treated by transformations of the observations and the predictions [67, Section 5.7.4]. Recent papers on this topic are [90], [91]. Some usual transformations require the strict positivity of values and are thus unsuited to the one-hot encoding that is used here. Hence, the set of usable transformations is restricted to the very recent literature on the topic.

A more prominent difficulty in our setting is that applying a transformation on weights in order to keep them in the $[0, 1]$ interval would alter the prescription of the predictions' expectation, so that it is not so straightforward to prescribe \mathbf{m} and to ensure the weights to be positive and summing to one, as higher scale constraints of Section 3.1. Furthermore, the distance to be minimised would also be altered, and the predictor would lose some minimal variance properties. Another approach would be to do an optimisation with such a positivity constraint, leading to quadratic programming [67, Section 3.9.1], but losing the closed-form expressions that are presented here.

In the numerical illustrations presented in Section 5, we have chosen another approach: we found empirically that adding a small nugget effect (i.e., adding a constant to the diagonal of the covariance matrix \mathbf{K}) was sufficient to ensure the positivity of weights when needed. Indeed, this empirical finding is supported by the following Proposition 9. Furthermore, in our investigations of the numerical Section 5, the added nugget effects were small enough, so that prediction accuracies were not significantly altered.

Proposition 9 (Nugget ensuring positive weights). *Assume that \mathbf{K} is replaced by $\mathbf{K}_{\text{nug}} := \mathbf{K} + \eta \mathbb{I}_n$ in Proposition 4 and Proposition 5, where $\eta > 0$ is a nugget parameter and \mathbb{I}_n is the identity matrix of size n . Assume furthermore that the prescribed vector $\mathbf{m} = \frac{1}{n} \mathbb{Y} \mathbf{1}_n$ in the latter Proposition 5, so that \mathbf{m} contains the proportion of each label in the observations. Then, for the predictors given by both Propositions 4 and 5, there exists a nugget η large enough ensuring that all weights in \mathbf{A} are positive and summing to one.*

Proof. The proof is given in Supplementary Material 10.10, page 255. □

In practice for the classification problem, it means that adding a sufficient nugget effect ensures the positivity of weights so that the predictions can be considered as membership degrees, summing to one and positive.

This approach was sufficient for the tested numerical illustrations. Otherwise, optimisation under the positivity constraint would still be possible, but would require either the use of quadratic programming or the extension of recent methods to higher scale constraints.

4 Filling Cross-Covariances

In the previous sections, we have seen the Joint Kriging model and its applications to constrained classification. The assumptions on the underlying random fields were very general: dependent components of $\mathbf{Y}(\cdot)$, with the existence of the first two cross-moments, without any Gaussian assumption. The obtained results were derived from specific covariances, in particular in matrices \mathbf{K} and \mathbf{H} . In the present section, we discuss practical strategies that can be used to fill the needed covariance matrices.

The main result of previous sections is the expression of optimal weights, with weights summing to one and a constraint on average predicted values, with an affine extension. However, despite this rather general model, the predictor is simplified, with weights applying jointly to all components in Equation (III.1). The application to constrained classification is justifying this simplifying assumption: indeed, applying different weights to components of $\mathbf{Y}(\cdot)$ would make it far more difficult to preserve classification higher scale constraints as presented in Section 3.1 and Table III.1. To the authors knowledge, existing more general multi-output models are not conceived to handle such higher scale constraints.

It is worth mentioning that, once the Joint Kriging simplifying assumption is accepted, there is no obstacle to using the cross-covariance function of any multi-output process $\mathbf{Y}(\cdot)$, even a non-stationary one. In particular, the reader may refer to the books [66, Chapter 20] and [67, Chapter 5] for general considerations about cross-covariances for multivariate models. Important articles on the topic are [68] and [69], where the estimation is also discussed.

In order to fill the matrices $\widetilde{\mathbf{K}}$ and $\widetilde{\mathbf{H}}$, for a given positive definite matrix \mathbf{W} , we need a function $k(\cdot, \cdot)$ that gives:

$$k(x, x') := \mathbb{E} \left[\mathbf{Y}(x)^\top \mathbf{W} \mathbf{Y}(x') \right] - \mathbb{E} \left[\mathbf{Y}(x)^\top \right] \mathbf{W} \mathbb{E} \left[\mathbf{Y}(x') \right]. \quad (\text{III.18})$$

Then the elements of the covariance matrices $\widetilde{\mathbf{K}}$ and $\widetilde{\mathbf{H}}$ can be derived from the covariance function $k : \chi \times \chi \rightarrow \mathbb{R}$ by setting:

$$\widetilde{\mathbf{K}}_{ij} = k(x_i, x_j) \quad (\text{III.19})$$

$$\widetilde{\mathbf{H}}_{ik} = k(x_i, x_k^*) \quad (\text{III.20})$$

Denoting

$$c_{i,j}(x, x') := \text{Cov} [Y_i(x), Y_j(x')],$$

one can derive:

$$k(x, x') = \sum_{i=1}^p \sum_{j=1}^p c_{i,j}(x, x') \mathbf{W}_{i,j}.$$

By gathering covariances in a matrix, we denote:

$$\mathbf{C}(x, x') = [c_{i,j}(x, x')]_{\substack{i \in \{1, \dots, p\} \\ j \in \{1, \dots, p\}}} \in \mathcal{S}_p^+(\mathbb{R}).$$

One can also check that the covariance function $k(x, x')$ is equal to the trace of the transformed process $\mathbf{W}^{1/2} \mathbf{Y}(\cdot)$ cross-covariances:

$$k(x, x') = \text{Tr} \left[\mathbf{W}^{1/2} \mathbf{C}(x, x') \mathbf{W}^{1/2 \top} \right].$$

If all $c_{i,j}(\cdot, \cdot)$ are known, then the latter quantity can be used to fill all needed covariance matrices. It can however rely on many parameters, as each $c_{i,j}$ can come with its own hyperparameters, namely $\mathcal{O}(p^2)$ hyperparameters.

Through the following examples, we investigate the link with several classical cross-covariance models.

Example 5 (Separable cross-covariances). *Let us consider the simplifying assumption of multi-output separable kernel functions [68, Section 4], for which there exists a covariance function $c(x, x')$ and a $p \times p$ real matrix $\mathbf{S} = [S_{i,j}]_{\substack{i \in \{1, \dots, p\} \\ j \in \{1, \dots, p\}}}$, such that:*

$$c_{i,j}(x, x') = S_{i,j} c(x, x'), \quad (i, j) \in \{1, \dots, p\} \times \{1, \dots, p\}.$$

Then

$$k(x, x') = c(x, x') \sum_{i=1}^p \sum_{j=1}^p S_{i,j} \mathbf{W}_{i,j}.$$

One sees that the role played by the matrix \mathbf{W} is quite similar to the one played by the separability matrix \mathbf{S} , and that, in this simplified setting, $k(x, x')$ is also proportional to the driving covariance function $c(x, x')$.

Example 6 (Linear model of coregionalisation). *The Linear Model of Coregionalisation (LMC) is a classical approach for combining several univariate covariances, [67, Section 5.6.4] and [69, Sections 2.1 and 4]. Let $\mathbf{R}(x, x') = \text{diag}[\rho_1(x, x'), \dots, \rho_r(x, x')]$ be a diagonal matrix of r univariate correlation functions. Assuming the output variables are generated by a linear combination of r independent univariate random fields with correlation functions ρ_i , $i \in \{1, \dots, r\}$, the LMC combines the univariate covariances in \mathbf{R} by setting:*

$$\mathbf{C}(x, x') = \mathbf{B}\mathbf{R}(x, x')\mathbf{B}^\top, \quad (\text{III.21})$$

where \mathbf{B} is a $p \times r$ full rank matrix. As a consequence, we get:

$$k(x, x') = \text{Tr} \left[\mathbf{W}^{1/2} \mathbf{B}\mathbf{R}(x, x')\mathbf{B}^\top \mathbf{W}^{1/2} \right].$$

One sees that the role played by the matrix $\mathbf{W}^{1/2}$ is quite similar to the one played by the LMC matrix \mathbf{B} , especially in the case where $r = p$.

Whatever the underlying model of cross-covariances $c_{i,j}(x, x')$, the model uses a single covariance function $k(x, x')$, which can be seen as the covariance function of a weighted sum of all output variables. This is due to the simplifying assumption and the optimisation of a single scalar error. The next example shows that in some cases, the latter covariance function depends on fewer parameters than the whole set of cross-covariance functions $\{c_{i,j}(x, x') : 1 \leq i, j \leq p\}$.

Example 7 (Isotropic k). *Let us recall Equation (III.18):*

$$k(x, x') := \mathbb{E} \left[\mathbf{Y}(x)^\top \mathbf{W} \mathbf{Y}(x') \right] - \mathbb{E} \left[\mathbf{Y}(x)^\top \right] \mathbf{W} \mathbb{E} \left[\mathbf{Y}(x') \right].$$

Let us assume that there exists a positive definite matrix \mathbf{W} such that the covariance function $k(\cdot, \cdot)$ is isotropic. That is to say, $k(x, x')$ depends only on some distance between x and x' . Then k can be written in a simplified form, so one does not need to estimate \mathbf{W} .

In this case, with one variance hyperparameter $\sigma^2 > 0$ and d positive hyperparameters $\theta_1, \dots, \theta_d$, usually referred to as characteristic length scales [54, page 14], one can set, for example:

$$k(x, x') = \sigma^2 r_0(\|x - x'\|_\theta),$$

where r_0 is a correlation function and $\|x - x'\|_\theta^2 = \sum_{i=1}^d \left(\frac{x_i - x'_i}{\theta_i} \right)^2$ is a rescaled Euclidean norm. Notice that this expression does not depend on \mathbf{W} any more, so that when using the above assumption, we do not have to estimate \mathbf{W} .

As a consequence of the previous Example 7, for a given matrix W , a noticeable advantage of the Joint Kriging method is the possibility to use a limited number of hyperparameters that need to be optimised. In that case, despite the multivariate output of the $p \times 1$ response vector $\mathbf{Y}(x)$, $x \in \chi$, there are only a few hyperparameters required for defining the covariances: for instance, σ^2 , $\boldsymbol{\theta}$, and the covariance family. This is quite different from co-Kriging techniques where all cross-covariances between components $Y_i(x)$ and $Y_j(x')$ should be defined for all $i, j \in \{1, \dots, p\}$, and $x, x' \in \chi$, which ends up in an order of $O(p^2)$ covariance functions and many associated hyperparameters. Furthermore, the method satisfies all prescribed constraints.

The relaxation of the simplifying assumption of Equation (III.1) would surely exploit more precisely the cross-covariance structure of output variables and increase the generality and accuracy of the model. However, the preservation of higher scale constraints makes the use of other existing models in the literature difficult.

5 Numerical Illustrations

In this section, one considers different numerical illustrations for both prediction and classification. The first illustration focuses on the impact of constraints with one output, and the second one on the behaviour of the predictor with multiple outputs. The third illustration gives an application to classification and a benchmark with numerous competitors. All the illustrations are created in `R markdown` notebooks, one per subsection, available as online supplementary material at <https://gitlab.emse.fr/marc.grossouvre/jointkrigingsupplementary/> [88]. Notebooks are given in both an executable format and an executed `html` format. The presented figures are directly extracted from the notebooks, and the results are fully reproducible.

5.1 A Simplified Toy Example

One considers here the very simple case where there is a single output variable: the output $\mathbf{Y}(x)$ is belonging to \mathbb{R}^p , with $p = 1$. The interest of testing the Joint Kriging with one single output variable is to discuss the impact of the constraint on predicted values and the impact of the affine prediction. For $p = 1$, Simple Joint Kriging and Ordinary Joint Kriging are identical to common Simple Kriging and Ordinary Kriging, but the constraint on predicted values leads to a new original predictor. We keep here the vector bold font for vectors $\mathbf{Y}(x) \in \mathbb{R}^p$ and $\mathbf{m} \in \mathbb{R}^p$, even though $p = 1$, in order to keep the very same notation as in the rest of the chapter.

Let us consider that the process $\mathbf{Y}(x)$ aims at approximating a hidden function, with say $a = 1$ and $b = 4$,

$$f(x) := a + \sin(x/b).$$

The observed locations x_1, \dots, x_n are randomly chosen with a uniform distribution over the interval $[-10, 5]$, and q prediction locations x_1^*, \dots, x_q^* are chosen regularly spaced over the interval $[-3, 10]$. Both intervals are purposely shifted so that some prediction points are far from observations, and vice versa. It also seeks to illustrate

how the constraint on average predictions is affected by the prediction sites. Observed responses in \mathbb{R}^p , with $p = 1$, are $\mathbf{Y}(x_i) = f(x_i)$, $i = 1, \dots, n$, with $n = 10$. Prediction is made over a set of $q = 100$ points. One defines X^* as a discrete and uniform random variable over all prediction points. The purpose here is not to interpolate as precisely as possible the hidden function f given a few observations, but only to illustrate the differences between various possible interpolators and the impact of requiring a prescribed average values for predicted values.

The prescribed value for $\mathbf{m} \in \mathbb{R}^p$, with $p = 1$, is the scalar $\mathbf{m} = 1.5$. The covariances between $\mathbf{Y}(\cdot)$ are modelled as prescribed in Example 7, from a single covariance function, using a squared exponential kernel. One could also pick a kernel that reflects f periodicity. However, the purpose is not to make the best possible prediction but, rather, to understand the impact of various constraints.

$$\text{Cov}[\mathbf{Y}(x), \mathbf{Y}(x')] = k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\theta^2}\right).$$

We set $\sigma^2 = 0.6$ mainly for the visibility of the confidence band in the presented figures, and $\theta = 1.2$.

In Figure III.1, one exclusively considers the constraint of sum of weights, which is assumed to be one: $\mathbf{1}_n^\top \boldsymbol{\alpha} = 1$. The predictor $\mathbf{M}(x)$ appears in a thick red line, together with confidence intervals built from the variance $\Delta(x)$.

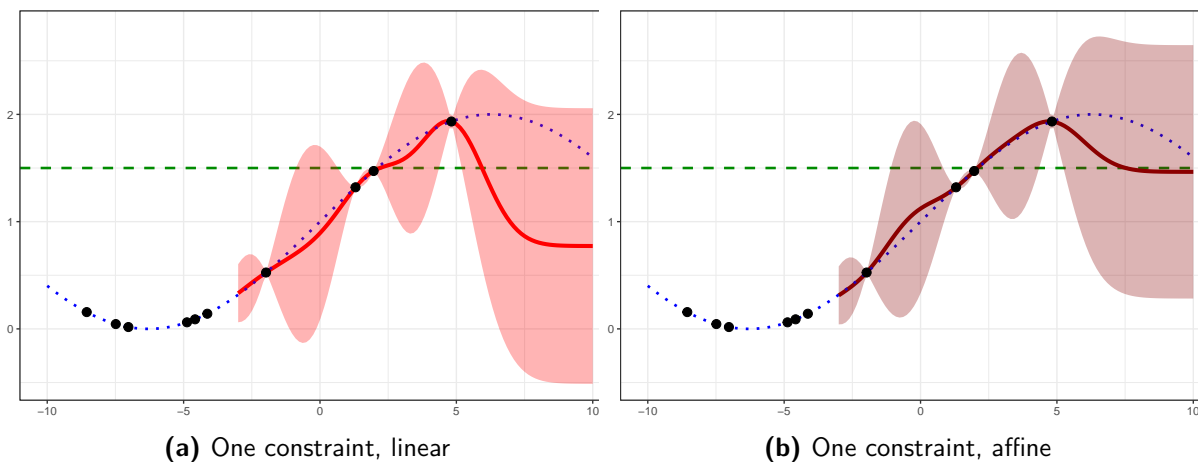


Figure III.1 – Prediction with one constraint: weights summing to one. Left: linear predictor, and right: affine predictor. In both cases, the average of the predicted values is distinct from the prescribed value $\mathbf{m} = 1.5$ (horizontal dashed line). The observations are the black dots. The thin, dotted, blue line is the underlying function. In the right panel, one applies the assumption in Remark 4 with $\rho = 0$ and $\sigma_Z = \sigma/10$.

Figure III.1a presents the result of ordinary Kriging exposed in Proposition 4. As is well known, when the location x is large (and far from observed locations), the ordinary Kriging mean tends to return to the estimated mean of the observations. The average value of the Kriging mean $\mathbb{E}[\mathbf{M}(X^*) | \mathbf{Y}] \simeq 1.12$ is different from the value $\mathbf{m} = 1.5$ (horizontal dashed line), which is natural as this constraint has not been taken into account yet.

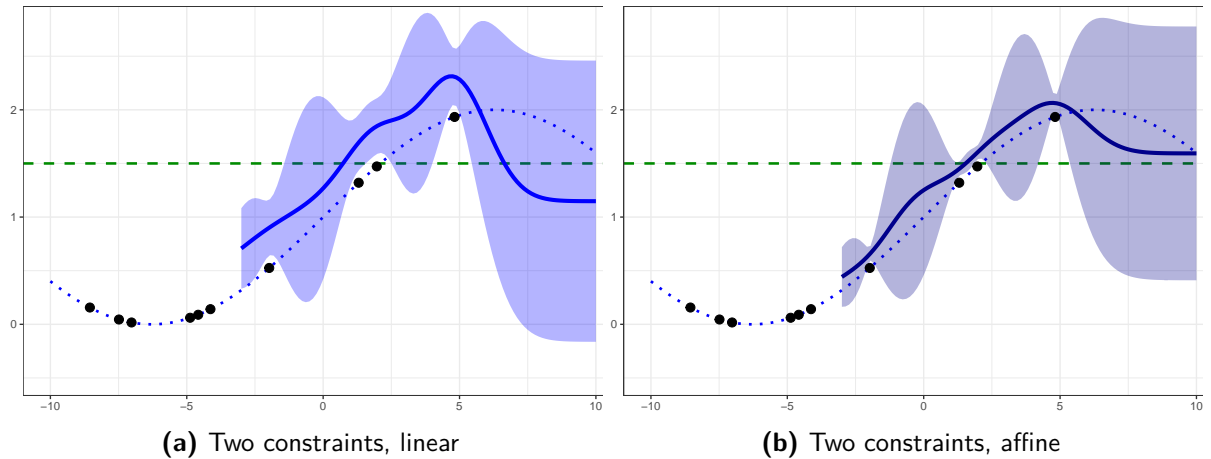


Figure III.2 – Prediction with two constraints: weights summing to one and the average of predicted values set to $m = 1.5$ (horizontal dashed line). The average of predictions is equal to this value $m = 1.5$ in both cases. The observations are the black dots. The thin, dotted blue line is the underlying function. In the right panel, one applies the assumption in Remark 4 with $\rho = 0$ and $\sigma_Z = \sigma/10$.

Figure III.1b uses the Proposition 6 to add a supplementary affine term to the linear combination, while preserving the sum of weights equal to one. The affine term is derived from a random variable \mathbf{Z} , and we choose $\sigma_Z = \sigma/10$ so that this external information is assumed to be trustworthy (small variance). Given $\mathbf{Z} = \mathbf{m}$, the consequence is that, far from observed locations, the prediction tends to put all weight on this external source of information, so that the prediction gets closer to \mathbf{m} , as one can see at the extreme right of Figure III.1b. This also makes the average $\mathbb{E}[\mathbf{M}(X^*) | \mathbf{Y}] \simeq 1.37$ closer to $\mathbf{m} = 1.5$, but the values of those two quantities remain distinct. Another consequence of the affine term is the reduction of the confidence band width, as a new source of information has been added.

In Figure III.2, one considers both the constraint of sum of weights, which is assumed to be one: $\mathbf{1}_n^\top \boldsymbol{\alpha} = 1$, together with the prescribed average of predicted values $\mathbb{E}[\mathbf{M}(X^*) | \mathbf{Y}] = \mathbf{m}$. The predictor $\mathbf{M}(x)$ appears in a thick blue line, together with confidence intervals built from the variance $\Delta(x)$.

Figure III.2a presents the result of ordinary Kriging exposed in Proposition 5. The average value of the Kriging mean $\mathbb{E}[\mathbf{M}(X^*) | \mathbf{Y}] = 1.5$ is exactly the prescribed one $\mathbf{m} = 1.5$ (horizontal dashed line), which is natural as this constraint has been taken into account during the joint optimisation of all $\boldsymbol{\alpha}(x_j^*)$, $j = 1, \dots, q$. However, the predictor is no longer interpolating. This is logical: if $q = 1$, one has one only prediction point x_1^* , and the constraint $\mathbb{E}[\mathbf{M}(X^*) | \mathbf{Y}] = \mathbf{m}$ becomes $\mathbf{M}(x_1^*) = \mathbf{m}$, which is distinct from an observation $\mathbf{Y}(x_n)$, even if x_1^* gets closer to x_n . Another example: if on the one hand observation points and prediction points are the same, if on the other hand \mathbf{m} is not the average value of observations, then at least one prediction must be different from the associated observation to satisfy the constraint.

Figure III.2b uses Proposition 6 to add a supplementary affine term to the previous

linear predictor of Figure III.2a while preserving the sum of weights being equal to one. The affine term is derived from a random variable \mathbf{Z} , and we choose, as previously, $\sigma_Z = \sigma/10$, so that this external information is assumed to be trustworthy. As above, the average of predicted values is exactly the prescribed one, by construction. Again, given $\mathbf{Z} = \mathbf{m}$, the consequence is that, far from observed locations, the prediction tends to put all weights on this external source of information, so that the prediction gets closer to \mathbf{m} , as one can see at the extreme right of Figure III.2b. Another consequence of the affine term is the reduction of the confidence band width, as a new source of information has been added. With the prescribed average of the predicted value, the predictor is not interpolating, but adding the affine term helps the prediction get closer to observations.

Notice that the constraint \mathbf{m} is purposely set to an arbitrary value, so that adding this constraint does not necessarily improve the prediction in this toy example: it is not the aim of such a constraint. The reader may imagine the case of an adverse scenario, which can worsen the prediction, or the case of external useful information, which can improve it.

In this simple toy example, one can check numerically that each prediction satisfies the constraints that it should. One can also clearly visualise the impact of the specific constraint on average predicted values and the behaviour of the predictor when adding an affine term.

The illustrations that have been presented in this subsection are available in the notebook [Application1D](#) of the online supplementary material.

5.2 A Multi-Output Time Series Example

In the previous example, we illustrated the impact of constraints on the prediction of a one-dimensional output. Hence, the *joint* aspect of the estimation was not discussed. In the present example, one considers multi-output data so as to illustrate the specificity of the single hyperparameter estimation with multiple outputs. We choose one-dimensional inputs in \mathbb{R} to facilitate the interpolation representation, but considering more general inputs in \mathbb{R}^d , $d > 1$, would be easy. It would only change the number of hyperparameters to estimate, d instead of 1.

Imagine the following situation: a city wants to infer the history of the concentration of some pollutants at a particular crossroad based on a small series of measurements. This simple problem requires a model that takes time as input and multiple concentrations as output. Obviously, the end purpose would be to have a model with space and time as input, but this is outside this illustration's framework.

Using the data *air quality* [92], one tries to infer the concentration of several pollutants from only a few values. The studied pollutants in the data were chosen arbitrarily: CO, C₆H₆, NO_x and NO₂. The time range of learning data has been selected so that visually there is not too much missing data in the period (sensor stuck to an inferior

bound or missing). It corresponds to hourly measurements from 23/04/2004 18.00.00 to 28/04/2004 17.00.00. Missing values are tagged with the value -200 in this data. They were all filtered before the study, as if they were not informative at all. The challenge is to predict all hourly measurements in the selected period from only $n = 10$ values.

The purpose here is not to give specific conclusions about the measured pollution but only to illustrate the capacity of the Joint Kriging model to handle complex multivalued data with very few hyperparameters to optimise. The idea is to create a *joint* model that is as simple as possible. Many refinements of the model could be suggested, but this is not the purpose of this example.

Let us model the covariances between components of $\mathbf{Y}(\cdot)$ using Example 7. The proposed method does not require the definition of each cross-correlation between a pollutant concentration at one location and a different pollutant concentration at a different location. It just takes one covariance function $k(x, x')$ between an implicitly weighted sum of all output variables. We use the multiplication of two covariance kernels (hence it is positive semi-definite): a periodic kernel with a period of one day and a kernel of the Matérn 3/2 family [54, Chapter 4 and Equation (4.31)]:

$$k(x, x') = \sigma^2 \exp\left(-\sin^2(\pi|x - x'|)\right) \left(1 + \frac{|x - x'|}{\theta}\right) \exp\left(-\frac{|x - x'|}{\theta}\right). \quad (\text{III.22})$$

The parametrisation has been simplified, e.g. factors $\sqrt{3}$ in Matérn covariance expressions are not used here because they have the same effect as a rescaling of the characteristic length-scale θ . Notice that despite the p dimensional output where $p = 4$ is the number of studied pollutants, the kernel $k(x, x')$ in Equation (III.22) depends only on two hyperparameters θ and σ^2 . Since σ^2 impacts the uncertainty in the prediction but not the prediction itself, it is set to $\sigma^2 = 1$.

Let us first consider one single constraint: the sum of weights should be one. It corresponds to the Joint Ordinary Kriging predictor.

Figure III.3 shows the optimisation of the *single* length-scale hyperparameter θ . As this study does not aim at comparing the prediction accuracy with other methods, we did not use a separate test sample but only a validation sample, keeping in mind that it may lead to overfitting. The validation data used for this single hyperparameter estimation is set to all hourly measurements in the selected period. For the hyperparameter optimisation, a specific error has been chosen, where one optimises the worst standardised mean absolute error over all $p = 4$ series: The errors have been standardised in order to make them unitless and scale-invariant. The best estimation is $\hat{\theta} \simeq 1.4$. It is kept for all other illustrations in the subsection.

The optimisation here depends quite heavily on the chosen observation locations, so that in practice, an averaged error on several training and validation datasets would probably be more stable. In many situations on real data, the error function is monotonic, either increasing and leading to extremely small optimised hyperparameter θ (the

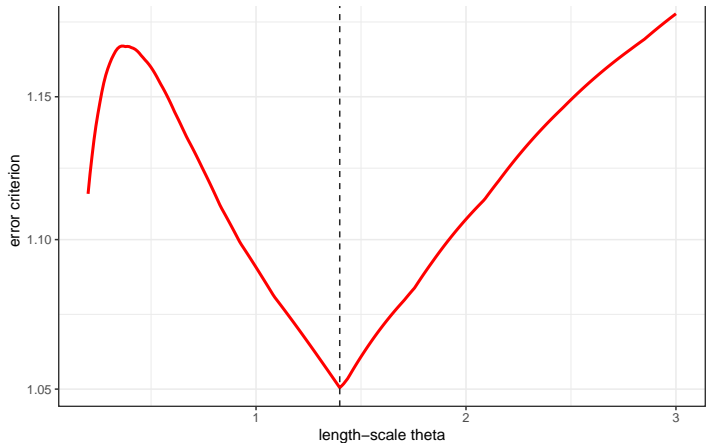


Figure III.3 – optimisation of the single correlation hyperparameter θ for the four selected pollutants, data extracted from Air quality data set.

prediction then tends to return quickly to an average value), or either decreasing, leading to a very large value of θ (the prediction then tends to smooth data a lot). Classical co-Kriging strategies that define a large number of cross-covariance hyperparameters would probably worsen the situation, highlighting the utility of a small number of hyperparameters.

Figure III.4 presents the simultaneous predictions of the four pollutant concentrations with Joint Kriging, the only constraint being that weights sum to 1. The confidence band associated with a given pollutant is proportional to the standard deviation of this pollutant’s concentration, as detailed in Proposition 8. Pollutant concentrations have very different orders of magnitude, but when applying Proposition 8, the obtained confidence bands look quite comparable between series, as desired.

With very few hyperparameters and a rough covariance model, the result has a lot of room for improvement. Nevertheless, despite the single model hyperparameter θ and, considering the limited number of observations $n = 10$, the predictions of the $p = 4$ concentrations seem quite reasonable. By construction, each prediction is a combination of observed values of the considered pollutant, with weights summing to one. In Figure III.4, no other constraint is added, so that the average of predictions does not correspond at all to a specific prescribed value.

Figure III.5 presents the simultaneous predictions of the four pollutant concentrations with Joint Kriging, on which both constraints on the weights and on the predicted values are imposed using the affine model of Remark 6. The left panels show an adverse scenario where the average of predictions is set to 130% of the true average. The right panels show a normal scenario where the average of predictions is set to 100% of the true average. Using this setting, the interpolation property is lost, as seen in the previous example of Section 5.1, but the $n = 10$ observations still have a large influence, and the global shape of the prediction is preserved. By construction, the average of predicted

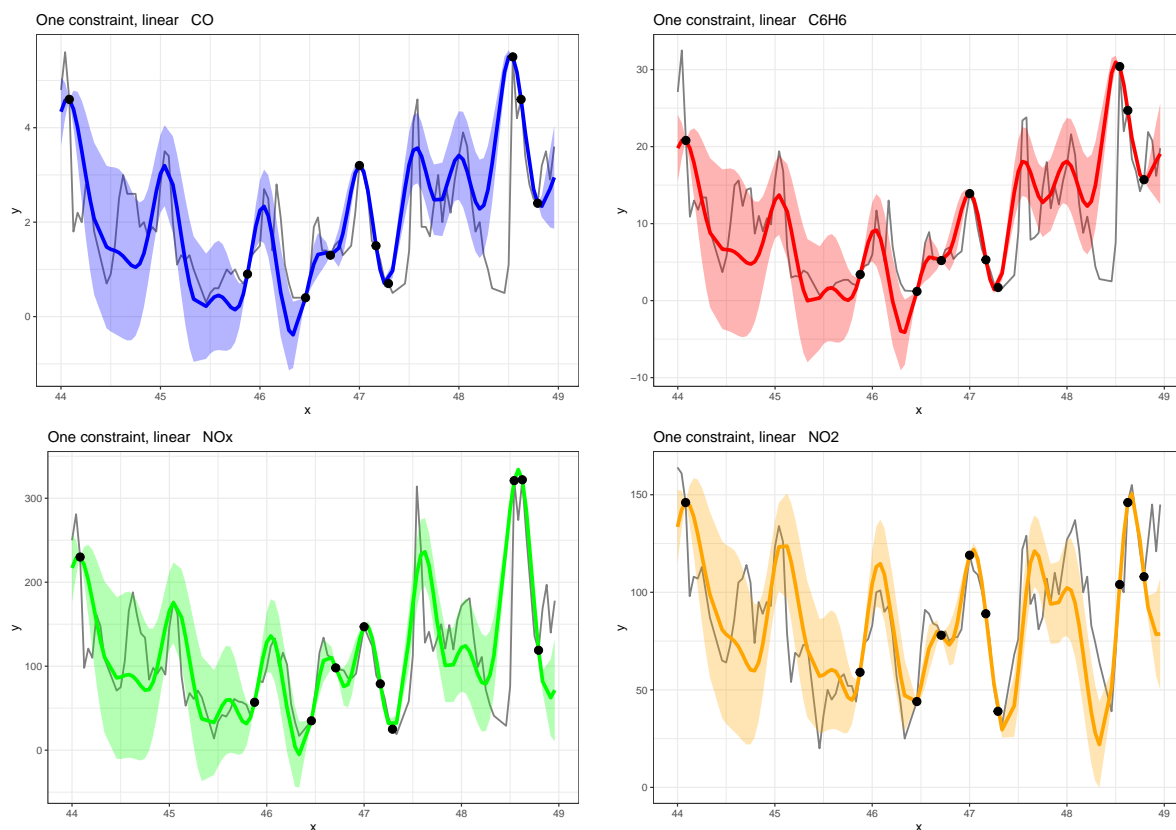


Figure III.4 – Joint Kriging interpolation: using the joint ordinary model with weights summing to one, with very few data points (black dots) and a single optimised length-scale hyperparameter obtained in Figure III.3. Upper left: CO, upper right: C₆H₆, lower left: NO_x, lower right: NO₂. Predictions are in thick solid lines, and true values are in thin black solid lines.

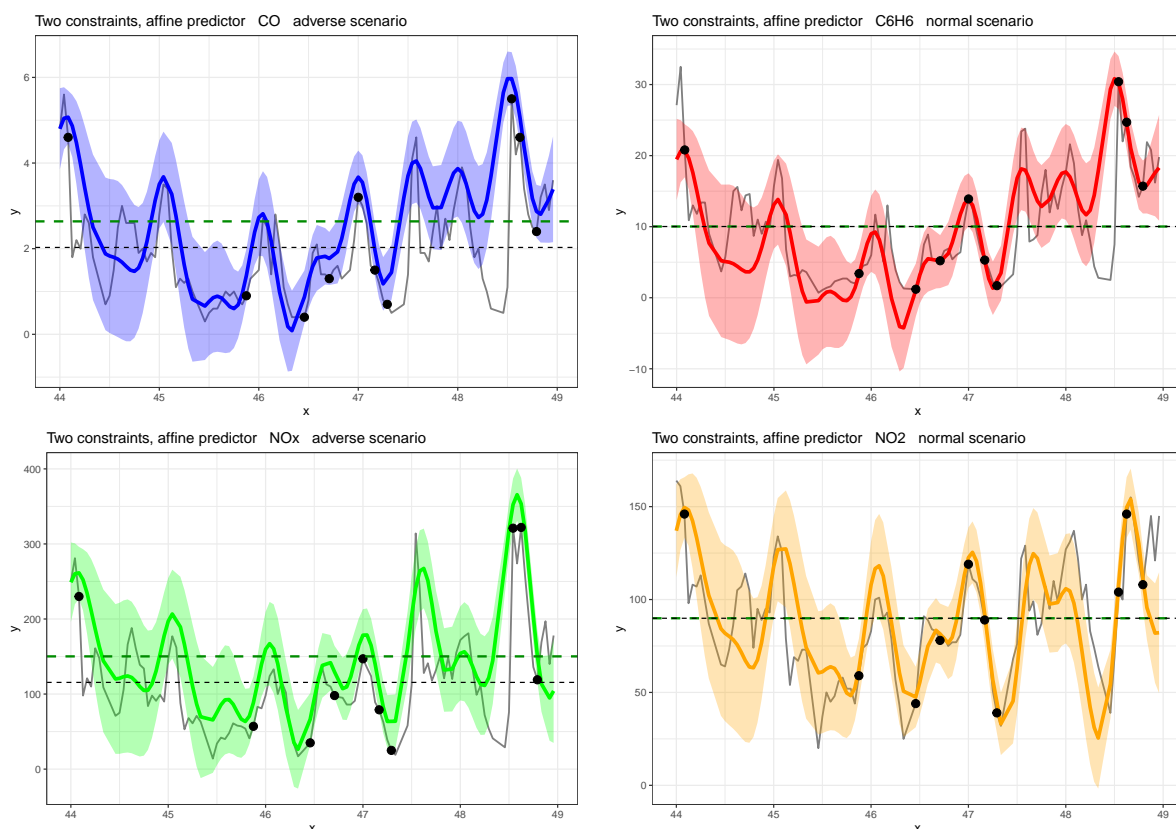


Figure III.5 – Adverse scenarios: interpolation using the joint affine model with two constraints (weights summing to one, prescribed average predictions), with very few data points (black dots), and a single optimised length-scale hyperparameter obtained in Figure III.3. Upper left: CO, upper right: C₆H₆, lower left: NO_x, lower right: NO₂. Predictions are in thick, solid lines, and true values are in thin, black, solid lines. Left panels are adverse scenarios where the average of predictions (thick dark green horizontal dashed line) is set to 130 % of the true average (thin black horizontal dashed line). Right panels are scenarios where the average of predictions is set to 100 % of the true average.

values (thick solid line) is exactly the prescribed one (horizontal thick dashed dark green line).

Considering this single hyperparameter model with a basic covariance model, the results also seem reasonable when using two constraints. In the left panels, the average of predicted values is exactly set to 30% more than the observed average of pollutant concentration, which is a lot. However, the visual differences between true and predicted sequences look surprisingly moderate, even in this adverse scenario. Despite satisfying all constraints, the model still offers a good fit with observations.

The goal of this numerical experiment is to demonstrate the Joint Kriging model's ability to handle complex multivalued data. It also illustrates the advantage of having a limited number of hyperparameters. One sees here that with a quite simple model, in a difficult problem (predicting four quite erratic time series from 10 observations), the model performs reasonably well. Furthermore, it allows for introducing some constraints, like setting an adverse scenario of a 30% increase in the pollutant concentration.

The illustrations that have been presented in this subsection are available in the notebook [ApplicationAirQuality](#) of the online supplementary material.

5.3 A Constrained Classification Example

We present in this subsection the specific case of multi-dimensional outputs derived from a classification problem. As presented in Section 3, Joint Kriging can be implemented for fuzzy classification. Different modalities of a classification variable are regarded as multiple output variables with values in $[0, 1]$.

Imagine the case of an event with measurable intensity that may occur at a given location in a territory. We are interested in classifying the intensity of this event, if it occurs, into multiple classes, depending on some thresholds. In the following, this event is an earthquake, and its intensity is its Richter magnitude.

The Quake data set given in [93], visualised in Figure III.6, describes 2178 earthquakes with their latitude, longitude, focal depth, and magnitude. A given location x has three coordinates: latitude, longitude, and focal depth. For a single observation at location x , the target $\mathbf{Y}(x) = (Y_1(x), Y_2(x))^T$ is equal to $(1, 0)^T$ if an earthquake is occurring here with a magnitude above the data set average magnitude, or $(0, 1)^T$ otherwise. If a location x is observed repeatedly, the membership degrees at x are averaged out over observations. It makes sense to impose that membership degrees are summing to one, so that $\mathbf{1}_p^T \mathbf{Y}(x) = 1$. Extensions with more thresholds is easy to conduct, as in Figure III.11. We keep here $p = 2$ for comparison to existing benchmarks. The binarised data is available at www.openml.org/search?type=data&id=772, on the openML website [94].

The purpose here is to compare the performance of Joint Kriging with a set of 69 other models' performances. The study available at www.openml.org/search?type=task&id=4516 (data retrieved on the 28th of June

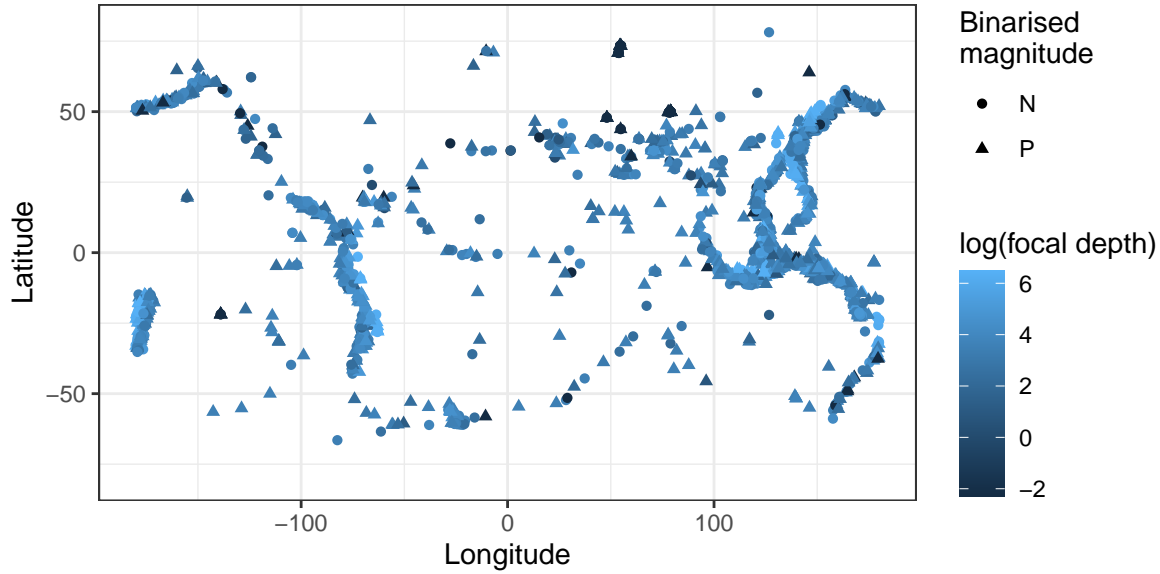


Figure III.6 – Earthquakes observations. An earthquake is a point with latitude, longitude, and focal depth (given by the colour) as its coordinates. Triangles represent earthquakes whose magnitude is above average. Circles represent earthquakes whose magnitude is below average.

2024) compares models, called flows in openML, performing 10 times a 10-fold cross-validation and computing the predictive accuracy as a performance indicator (see tab Analysis, measure `predictive_accuracy`).

Remember from Example 7 that although we are constructing a bivariate model, we need a single covariance kernel. The latter should be periodic with respect to latitude and longitude, not with respect to focal depth. A simple way to define an admissible kernel is to multiply the three kernels associated with the three dimensions [54]:

$$k(x, x') = \sigma^2 \exp \left(-2 \frac{\sin^2((x_1 - x'_1)/2)}{\theta_1^2} - 2 \frac{\sin^2((x_2 - x'_2)/2)}{\theta_2^2} - 2 \frac{(x_3 - x'_3)^2}{\theta_3^2} \right)$$

The hyperparameters estimation has been treated separately on other train/test splits to avoid overfitting the data. The resulting values for θ are 2.3 for latitude, 0.9 for longitude, and 196.8 for focal depth.

In order to visualise the algorithm’s behaviour, we predict on a grid of latitude, longitude, and focal depth values. In addition to imposing the sum of membership degrees to be 1, we set the output mean expectation to be the same as in the data set. Predictions on a grid of latitude, longitude, and focal depth are presented in Figure III.7. One can observe that maps representing membership degrees (first two rows) can be deduced from each other by $y = 1 - x$. The third row shows a segmentation of the plane into areas where the membership degree for “P: magnitude is greater than average” is greater than 0.5, and areas where the converse is true. This segmentation depends on the focal depth: a small focal depth on the top row (21 km, first quartile) and a greater

one on the bottom row (68 km, third quartile). For instance, looking at the bottom left corner of the map, which is around the Fiji archipelago, one can predict that earthquakes with small focal depths are more likely to be of large magnitude than deep earthquakes. However, the converse is true in the South Atlantic area (bottom-centre part of the map). Moreover, the predictor achieves circular coherence along longitude due to the periodicity of covariance. Periodicity along latitude is more difficult to observe because it covers only 180° .

Performances are evaluated using Predictive Accuracy which is the percentage of instances that are classified correctly. It is measured on binarised predicted membership degrees, on a ten times 10-fold cross-validation, as in the OpenML benchmark, in order to get comparable results. Prior to that, the hyperparameters optimisation has been treated separately on other train/test splits in order not to overfit the data. Figure III.9 presents, from top to bottom, two results found in openML, i.e., the best recorded model, which is the Kernel Logistic Regression with Radial Basis Function Kernel and Random Forest for reference. Below are the results of the Joint Kriging models: the simple model without constraint, the model with weights summing to 1, the model with constraint on the prediction and weights summing to 1, the affine model with weights summing to 1, and the affine model with constrained output.

For the ten runs, each diagram shows the Predictive Accuracy of each run (coloured points), the minimum, first quartile, median, 3rd quartile, and maximum, as well as the mean value materialised by a cross. Although the runs' performances stay in the range of those observed for Random Forest and Kernel Logistic Regression, the average values obtained with Joint Kriging are greater: the average is 0.558 ± 0.002 for the best model in the OpenML benchmark and 0.5660 ± 0.0038 for the best Joint Kriging model. The latter was even slightly greater, 0.5669, during hyperparameter optimisation, due to a slight overfit that has been reduced when using different train/test splits. Benchmark being based on this average value, it means that Joint Kriging has a better performance than the 69 models tested in the OpenML benchmark.

One can expect the multiplication of constraints to have an adverse effect on performance, as a constrained optimisation has less degree of freedom than an unconstrained one. On the other hand, injecting useful information through constraints may improve the performance. Figure III.9 shows that overall, the performance is improved, especially when adding the constraint on the output. Figure III.10 shows the distribution of the mean predictive accuracies for the 69 models tested in the OpenML website. None of them is above 0.56, while all Joint Kriging models are.

In Figure III.8, one uses the affine version of Joint Kriging with two constraints: weights summing to one and prescribed average prediction. In the left panels, an adverse scenario forces the average predicted membership degrees of the first class (large magnitude events) to be equal to 65%. In the right panels, this percentage is set to the observed percentage of large-magnitude events, 55%. This illustrates the usefulness of the constraint for adverse modelling.

In order to compare the results with existing benchmarks, we studied above the $p = 2$ binary classification problem. But the method can handle more classes as well. As an example, in Figure III.11, we give a prediction for $p = 4$ classes. Observations have been converted into four classes using three Richter magnitude thresholds: 5.85, 5.95 and 6.15. Specific thresholds have been chosen for this illustration in order to get enough observations in each class (at least 17% observations), but a seismology study might focus on other thresholds. Once again, the predictor achieves a circular coherence along longitude, and one can observe complex patterns that would be difficult to catch with classification trees. The presented classification task was constructed from indicators deriving from an underlying real value, the Richter magnitude, and from thresholds, thus creating ordinal classes. But the prediction can also be derived from observations of non ordinal class labels without any underlying process or thresholds.

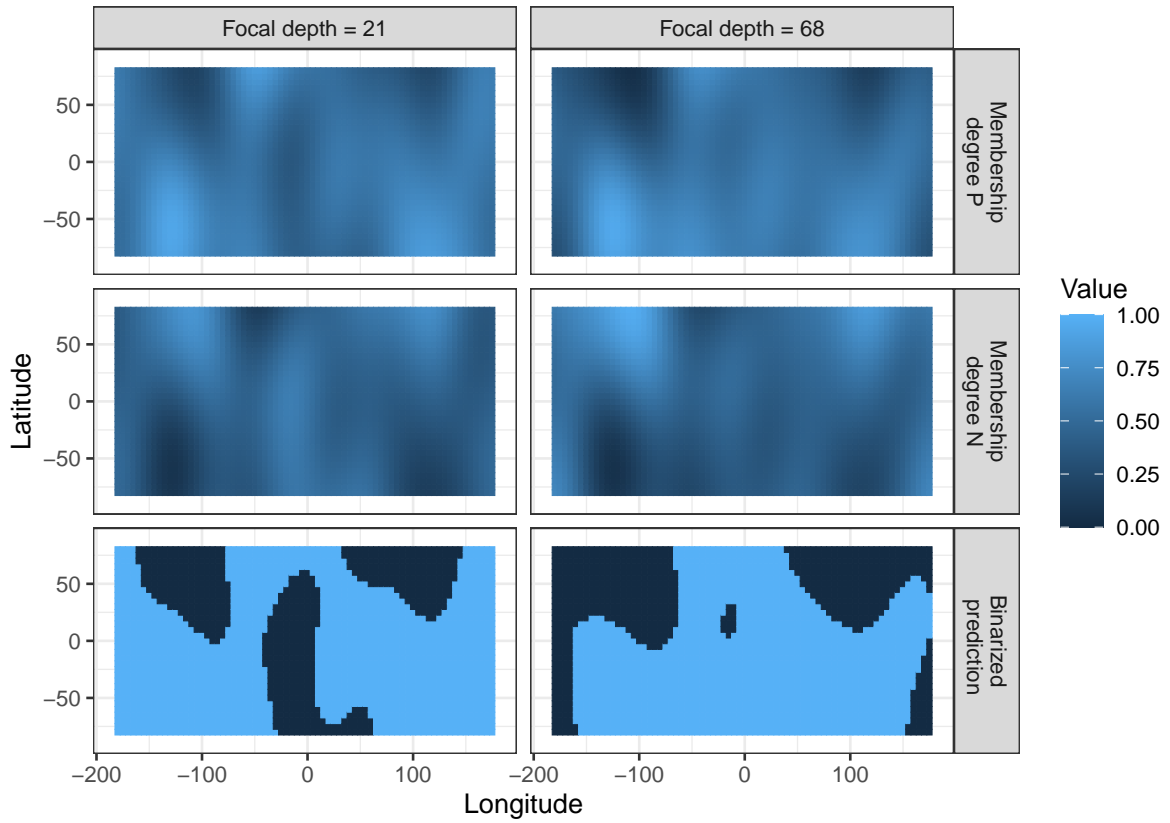


Figure III.7 – Joint Kriging with 2 constraints, earthquakes' magnitude prediction into 2 classes. From top to bottom: membership degree of “P: magnitude is above average”, membership degree of “N: magnitude is below average”, binarised prediction (1 if membership degree of P is greater than 0.5). Left: focal depth of 21 km. Right: focal depth of 68 km.

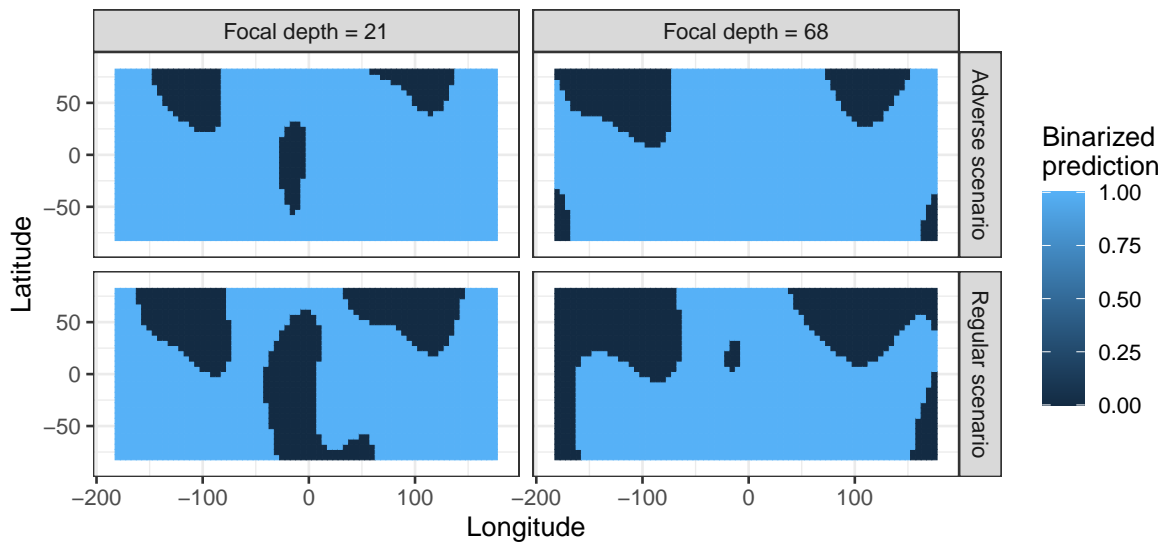


Figure III.8 – Adverse scenario: predicted membership degrees of earthquakes' magnitude using Joint Kriging with two constraints. Top panels: adverse scenario, first-class output average constrained to be 65%. Bottom panels: regular scenario, output average constrained to 55.5%. Left: focal depth of 21 km. Right: focal depth of 68 km.

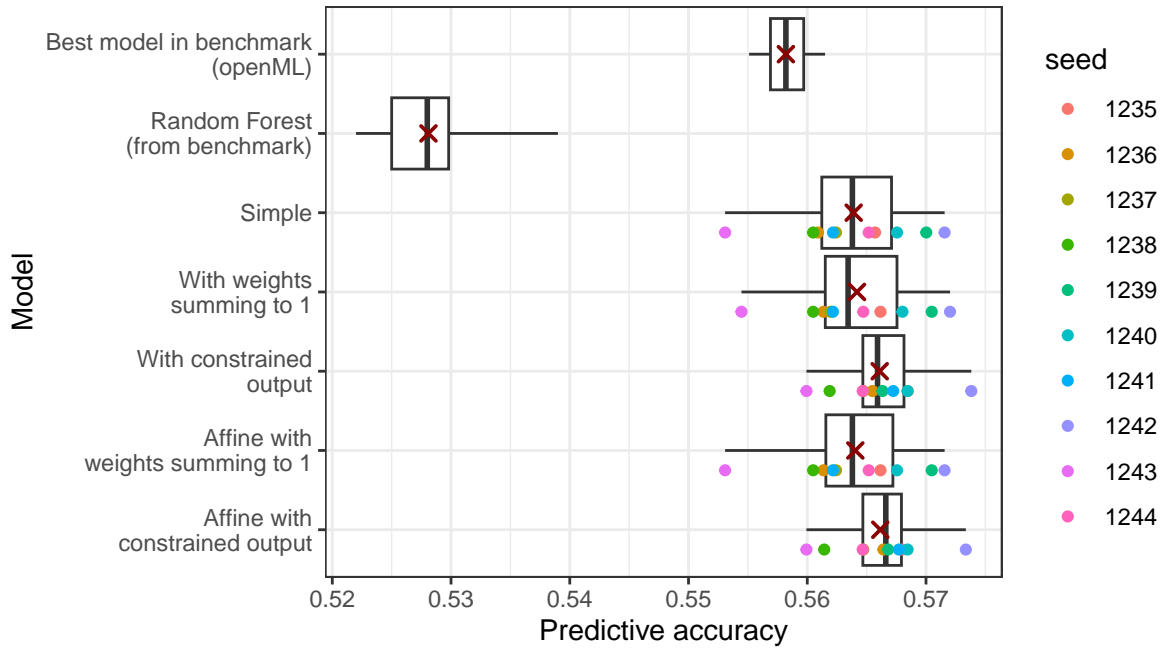


Figure III.9 – Distribution of performances for 10 runs of two OpenML models (the best model in the benchmark among 69 models, and Random Forest), and the 5 different types of Joint Kriging model. The whisker plots give the minimum, first quartile, median, third quartile, and maximum. The dark red cross indicates the average predictive accuracy; the higher, the better. The average is 0.5660 ± 0.0038 for the best Joint Kriging model and 0.558 ± 0.002 for the best model in the OpenML benchmark. Other 67 models of the Benchmark are omitted here. The script used to retrieve data from the OpenML database is provided in Supplementary Material 10.11, page 257. Data were extracted on June 28th, 2024.

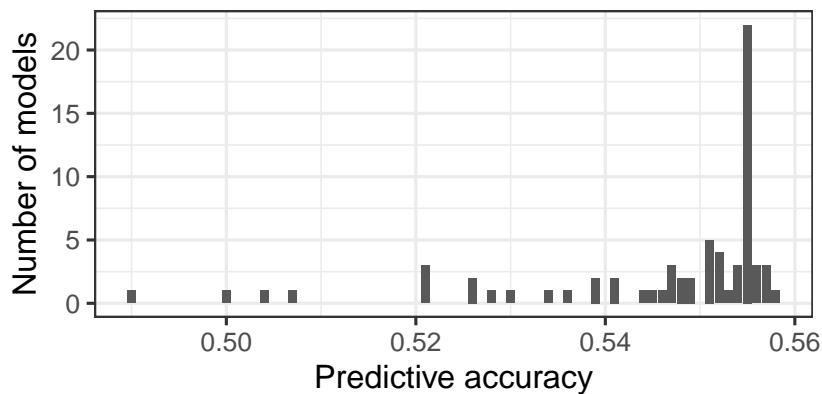


Figure III.10 – Distribution of the mean predictive accuracies for the 69 models tested on the Quake dataset, in the OpenML framework. Note that some models have been run multiple times, in which case we select only the best run. The graphic is a bar plot of the predictive accuracies rounded to the nearest third digit. Data were extracted on June 28th, 2024.

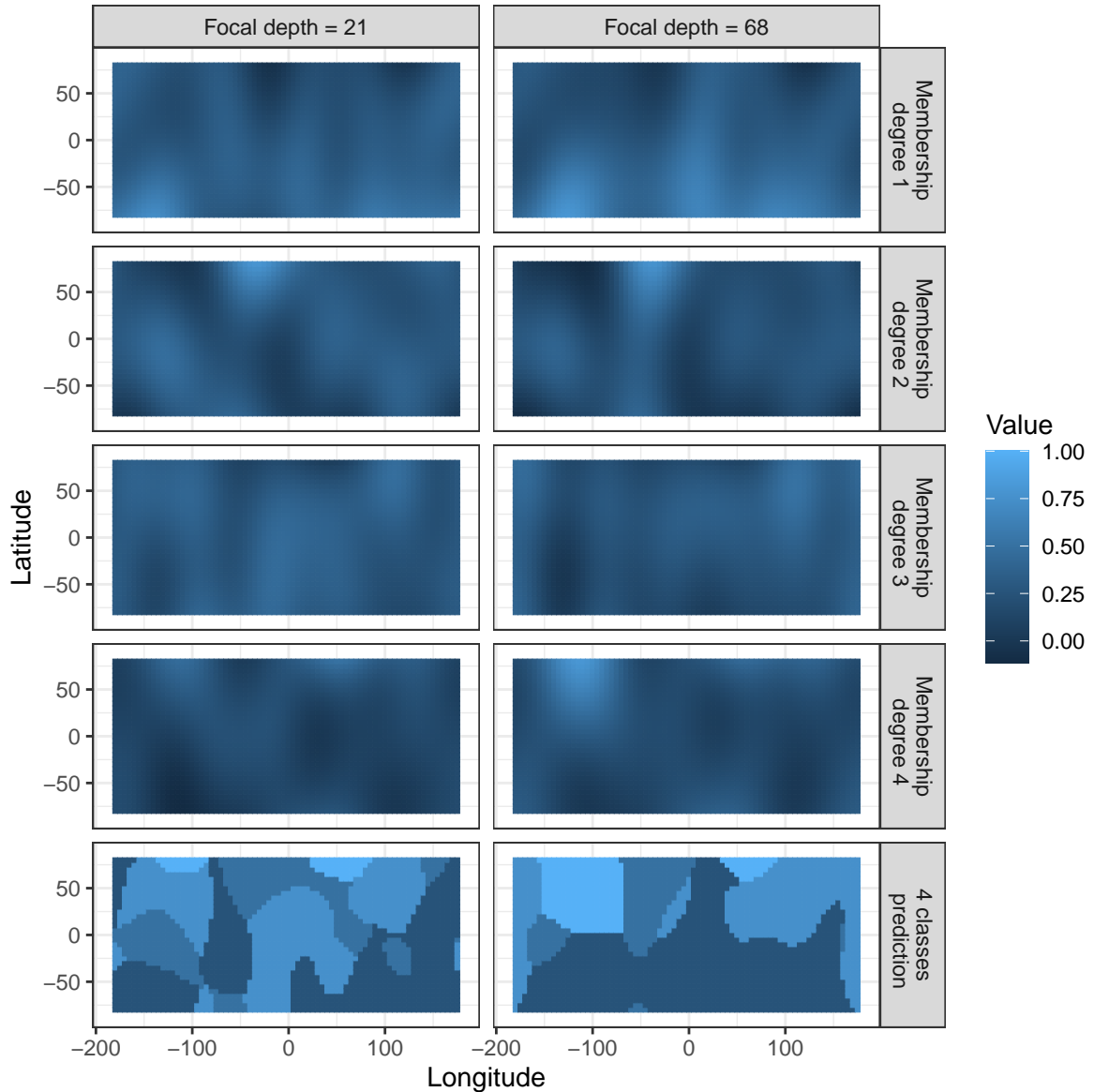


Figure III.11 – Affine Joint Kriging with 2 constraints. Earthquakes’ magnitude is divided into 4 classes. From top to bottom: membership degrees of “1: magnitude is smaller than 5.85”, “2: magnitude is between 5.85 and 5.95”, “3: magnitude is between 5.95 and 6.15”, “4: magnitude is greater than 6.15” and class of greatest membership degree in the 5th row coloured by increasing magnitude from dark to light blue. From left to right: focal depth of 21 km, focal depth of 68 km.

In this classification example, we used a direct implementation of the model with a single covariance family. Nevertheless, the average performance is the best one among the whole OpenML benchmark. Fine tuning of the model would surely lead to performance improvements. The illustration above aims at demonstrating that, with very basic assumptions, the method is competitive with an open benchmark that has numerous competitors, as shown in Figure III.9. It also aims at showing that it can model adverse scenarios, as in Figure III.8, or multiple classes, as in Figure III.11.

The illustrations that have been presented in this subsection are available in the notebook [ApplicationQKmain](#) of the online supplementary material.

6 Conclusion

A Joint Kriging model on multiple outputs has been presented, where at each prediction location, the same weights apply to all outputs. This simplification was necessary to handle all the considered constraints. It also allows for easy covariance modelling with very few hyperparameters even though the number of outputs p is large. Still, the model benefits from Kriging advantages: interpretability, ability to interpolate data, prediction of the uncertainty in each prediction, and specific covariance modelling. As with any simplification, the model can surely be improved and may have some limitations compared to heavily parametrised models. For instance, co-Kriging with many cross-covariance functions might be more flexible for dealing with time series with different regularities, or models with parametrised distortions of locations might be more convenient for dealing with non-stationarities. However, the limited number of hyperparameters and the simplicity of their estimation are an asset of the model, while allowing specific model characteristics such as periodicity. Furthermore, the model is not limited to Gaussian Processes, as it only relies on the existence of moments of order one and two.

An original constraint on predicted values was also introduced. It appears to be useful for using external information, for adverse modelling, for homogenising results, or for considering fairness constraints. To handle this constraint, all weights of predicted points need to be computed at the same time, unlike usual Kriging techniques. But the resulting predictor itself is quite simple to derive since it is given by a closed formula. Some extensions using an affine term were also proposed to account for external information and provide more control over the behaviour of the predictor far from observations.

Ultimately, an application to classification was developed. Applying a multi-output Kriging model on classification is feasible through the prediction of membership degrees. Even without constraints, it is in itself interesting: it allows for interpretability, modelling uncertainty's estimation, and interpolating data. Using Joint Kriging with the proposed constraints easily ensures that membership degrees sum to one and allows for prescribed percentages of each predicted class. The simplified covariance model greatly

eases the hyperparameters' estimation. At the same time, with Joint Kriging, classification tasks benefit from the diversity of covariance kernels, including periodicity. The resulting classification performs especially well in the investigated practical case: in the earthquake numerical example, the model competes with the best-provided approaches on an open data set with numerous competitors.

Multiple extensions to the model can be imagined. For instance, the model with constrained predicted values does not guarantee continuous interpolation, so that further work may fix this problem. A specific estimation procedure for the underlying joint covariance structure could also be of interest. Moreover, once applied to classification, membership degrees summing to one do not imply the combinations to be convex. Some weights can still be negative or greater than 1 so that an adjustment of the nugget effect may be required. Ensuring the combinations to be convex without any nugget effect adjustment could also be an improvement. Eventually, one may be interested in searching for a way to relax the simplifying assumption while keeping the constraints.

Enhancing Prediction's Performances

The content of this chapter is published, under the title *Enhancing buildings' energy efficiency prediction through advanced data fusion and fuzzy classification*, in the *Energy and Buildings* journal, volume 313, 15 June 2024 [95]. It is published in open access under the license Creative Commons Attribution 4.0 International.

	Résumé en français136
1	Introduction137
2	Data Presentation139
2.1	Information About Dwellings139
2.2	Energy Efficiency Observations141
3	Methodology144
3.1	Performance Indicators144
3.2	Variable Selection146
3.3	Fuzzy Classification with KNN149
3.4	Fuzzy Classification with Kriging149
4	Results and Discussion150
5	Conclusion153

Résumé en français

Ce chapitre traite de la prédiction des étiquettes **DPE** (*Diagnostic de Performance Énergétique – Energy Performance Certificate*) des bâtiments en France en utilisant des informations descriptives disponibles, sans nécessiter de visite physique. Cette approche répond aux politiques européennes sur la durabilité, qui encouragent l'amélioration de la performance énergétique pour réduire la consommation énergétique et les émissions de gaz à effet de serre des bâtiments résidentiels et tertiaires. La France, avec ses 40 millions de bâtiments résidentiels, dont 22 millions construits avant 1975, présente un défi significatif¹. Ces bâtiments anciens sont souvent mal isolés et énergivores, représentant 55 % du secteur du logement et plus de 75 % de sa consommation d'énergie. Leur rénovation est donc prioritaire pour atteindre les objectifs climatiques tout en améliorant la qualité de vie.

La prédiction du **DPE** sans visite sur place implique plusieurs défis, notamment la collecte et la fusion de données dispersées à travers diverses administrations, ainsi que la mise au point de modèles prédictifs. Le chapitre détaille l'utilisation de modèles d'apprentissage machine, en particulier les modèles de classification floue et de régression, pour estimer les étiquettes **DPE** à partir de données socio-économiques et de caractéristiques des bâtiments.

Les approches de classification floue, comme **FKNN** (*Fuzzy k-Nearest Neighbours*) et Joint Kriging, montrent une capacité particulière à traiter la complexité et l'hétérogénéité des données des bâtiments, permettant des prédictions à la fois précises et informatives sur le plan énergétique. Ces modèles permettent de catégoriser efficacement les bâtiments selon leur efficacité énergétique prédite, sans les coûts associés aux évaluations sur place. Il est aussi montré que la classification floue permet d'apporter un supplément d'information utile à la détection des passoires énergétiques.

En conclusion, le chapitre souligne l'importance de recherches pour orienter les efforts de rénovation énergétique, en alignant les actions de lutte contre le changement climatique avec les améliorations de la qualité de vie, tout en respectant les cadres réglementaires européens. La prédiction précise du **DPE**, à partir d'informations facilement accessibles, peut soutenir la gestion de l'efficacité énergétique dans le secteur du bâtiment, favorisant une transition énergétique plus efficace et ciblée.

1. Statistiques calculées à partir de la base **IMOPE**, en comptant les identifiants bâtiments issus du **RNB** (*Référentiel National des Bâtiments – National Buildings' Inventory*).

Abstract

This study proposes a method to predict buildings' energy efficiency based on available descriptive information, without a physical visit, by merging diverse datasets and employing advanced classification techniques. By integrating geographical, structural, legal, and socio-economic data with **EPC (Energy Performance Certificate)** observations, our approach yields a rich learning set. Through variable selection methods like forward selection with **KNN** and **Simultaneous Perturbation Stochastic Approximation** for **FKNN**, we refine the model's variables. Comparing fuzzy and hard classification using **KNN**, Kriging (see Chapter III) or Random Forest approaches, we find fuzzy classification more adept at capturing nuanced energy inefficiency indicators. Our study highlights the importance of mass energy efficiency prediction for sustainable renovation efforts.

Keywords – fuzzy classification, Kriging, constrained classification, spatial Prediction, energy efficiency, sustainability.

1 Introduction

As per the data provided by the French Ministry of Finance in 2023, France has 40 million residential buildings, encompassing a total land area of 12 000 km². Of these, 22 million dwellings were constructed before the first thermal regulations were introduced in 1975². These older buildings are highly demanding in energy, possess poor thermal properties, and lack insulation [96]. They represent 55 % of the housing sector and contribute to over 75 % of its energy consumption. Their renovation has therefore been a priority for the last fifteen years [97] for several reasons: The building stock presents significant potential for energy saving [98]; refurbishing buildings is the most profitable sector in terms of CO₂ decrease per invested Euro [99]; the long lifespan of buildings amplifies the consequences of a wrong design. Studies also suggest that refurbishment, rather than demolition, is more effective based on time, cost, community impact, prevention of urban sprawl, reuse of existing infrastructure, and protection of established communities. By renovating buildings to high efficiency standards, ambitious climate change mitigation actions align with improvements in living quality.

However, undertaking energy efficiency refurbishment is a complex process involving many stakeholders. This aspect challenges city planners and municipal decision-makers, who have a crucial role to play [100]. Effective policy-making necessitates a comprehensive understanding of the building stock [101]. Yet, the technical literature [102] and our empirical experience suggest that data collection is a barrier among the local administrations for strategic and political decisions. This is due to the dispersion of data among several municipal offices and other administrative entities, the lack of interoperability among the data collection systems, and notably the absence of energy efficiency

2. Statistics computed for the **IMOPE** database, counting buildings identifiers as per the **RNB (Référentiel National des Bâtiments – National Buildings' Inventory)** database.

assessments for each and every building and dwelling in the town [103], [104].

To overcome these difficulties, models for predicting the buildings' energy consumption have been developed over the last 15 years [103], [105]. Simplified and data-driven approach models, also known as “bottom-up” models, provide valuable methods for assessing the consumption of a city's building stock [106], [107]. Bottom-up energy models are classified into two main subcategories: engineering models and statistical models [108], [109]. On the one hand, engineering, or numerical, models address the energy consumption questions with a dynamic approach based on equations describing the physical and thermal behaviour of the building [110]. However, these approaches are often driven by archetypes and sampling methods that rarely consider the local variability of building characteristics [107], nor do they incorporate the incremental energy measures implemented in older buildings due to renovation strategies applied by municipalities. On the other hand, statistical models based on machine learning algorithms rely on numerous ground observations and can yield high prediction capabilities provided that physical indicators describing a building are available. A comprehensive review can be found in [111]. While some efforts have been made in feature reduction [18], the remaining features are still challenging to infer without a physical visit to the building [19].

Defining categories of buildings based on their energy efficiency is now widespread. In the **USA**, the Energy Star programme evaluates the energy efficiency of homes based on criteria such as insulation, windows, heating, and cooling systems. China has adopted a holistic indicator known as the “Three Star” building rating system, which assesses the overall sustainability of a building in terms of land efficiency, energy efficiency, water efficiency, resource efficiency, and environmental quality, among others. In Canada, dwellings are classified using the EnerGuide rating system, which provides the energy consumption per square metre and per year. A similar building classification according to their energy performances, whether real or theoretical, has been defined in all E.U. countries. These labels are used to identify the energy sieves and target the renovation efforts; they may also be used to assess present and future energy needs. Some countries have opted for more qualitative indicators, while France has chosen an indicator based on quantitative measures. An **EPC (Energy Performance Certificate)** is defined in France as the building's energy consumption for standard use, associated with a qualitative labelling letter ranging from A to G. Similarly, a greenhouse gas emission label is defined. The final **EPC** label is the worst of both. For instance, if a building is labelled C for energy consumption and D for gas emissions, the **EPC** label is D.

The main goal of the present work is to quantitatively predict the EPC label of each building in France based on available descriptive information without requiring a physical visit. Unlike the publications cited previously, our interest extends beyond assessing the distribution of labels or energy consumption at the area level, such as a city; we aim for the most accurate prediction at the individual building level. It is important to note that **EPC** is determined by simulating a standard

use for the building. Therefore, our work differs from those studies focusing on actual energy consumptions, as, for instance, [112], which provides a comprehensive bibliography of recent works in this area. Furthermore, while real past energy consumption data is typically readily available and valuable for predictions in commercial buildings [113], our focus lies on residential buildings, where access to real consumption data is constrained by privacy regulations. **In terms of methodology, this chapter aims to demonstrate the feasibility of estimating a building’s EPC label without requiring a technician visit.** We illustrate that socio-economic features can compensate, to a certain extent, for the lack of technical information about a building. From a technical standpoint, we introduce fuzzy classification as a valuable tool in this context.

The next section introduces the dataset used in our study, consolidating information from various sources to comprehensively characterise residential buildings. We will then detail our methodology for variable selection and the process of learning and predicting EPC labels. Finally, the fourth section presents and discusses the results obtained.

2 Data Presentation

This section describes available data sources and the way to merge them to obtain a table that can be used as input for a learning algorithm.

2.1 Information About Dwellings

Among the various French institutions collecting information about dwellings, the MoF (Ministry of Finances) is a major player as it requires data to compute property taxes. MoF manages a database of all dwellings in France, geolocated by address and land plot identifier. It provides structural, historical, and qualitative information about each dwelling, including the surface area, number of rooms of each type (bedroom, kitchen, bathroom, etc.), construction materials for the roof and main wall, and year of construction. Some qualitative variables, such as the comfort level and the maintenance quality, are also provided. Another set of variables informs about ownership and occupancy, including date of acquisition, type of owner (private/public, individual/company), occupancy status (owner-occupied, rented, vacant), and rental value. This database has very limited information about energy consumption except for an indicator that identifies the dwellings that are connected to the city gas supply.

The main advantage of the MoF’s database is its comprehensiveness, as it inventories all dwellings. Moreover, it provides up-to-date documentation that includes the reliability level of each variable. However, it has limitations, primarily stemming from missing data and a lack of data updating. These issues affect the performance of the algorithms that learn from this database.

In addition to this restricted-access database, open data is also available. The National Institute of Geography provides a 2D model of the territory, outlining the ground print of all buildings. And it is in the process of acquiring LIDAR (Laser Imaging Detection And Ranging) data for a 3D model. Other databases contain information about altitude,

climate zones, and areas subject to specific regulations, such as heritage protection. Although historically, the government has collected a lot of information about dwellings, there has been minimal focus on energy consumption. To address the need for improved national energy consumption management, all **EPCs** are now collected and available for research (see Subsection 2.2).

The above databases are merged to gather all information about buildings themselves, not only dwellings. They are presented in Figure IV.1. This fusion process results in a table where each row is an address. This table comprises 22 million rows and 275 columns. It is the main table of the **IMOPE** database produced by **U.R.B.S.**. Many of these columns have characteristics that prevent us from using them in a machine learning algorithm: they may have a specific type, like strings or geometries, or they may contain too much missing data. Among the columns that are valuable to us, we also have to select those that are truly relevant for predicting the **EPC** label as explained in Subsection 3.2.

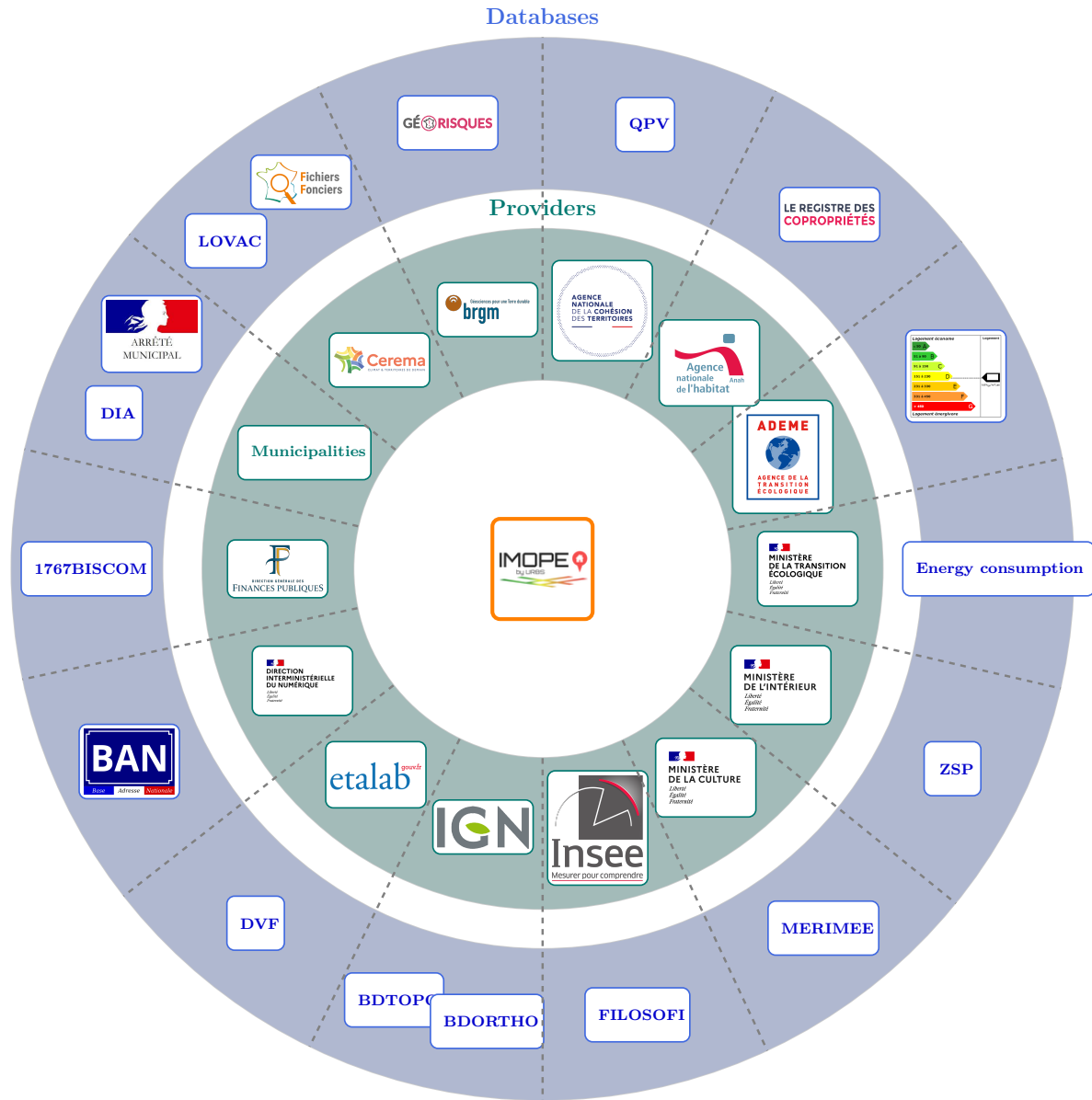


Figure IV.1 – Outer circle: Main databases that are merged to form the **IMOPE** (*Inventaire Multi-Objets du Parc Existant – Multi-Object Inventory of the Existing Building Stock*) database. Inner circle: data provider. The databases’ names as well as the providers’ names are clickable. See also Section 1 of Chapter I, page 52, for technical details about the data fusion process.

2.2 Energy Efficiency Observations

When a dwelling or a building is sold or put up for rent, an **EPC** must be available for the buyer or tenant. To establish an **EPC**, a technician visits the dwelling or building, creates a floor plan, gathers information about construction materials, insulation type, windows’ specifications and orientation, heating system, air conditioning (if any), hot sanitary water system, and other relevant indicators. This information is entered as input parameters into software that models energy consumption. It calculates a standardised energy consumption, making assumptions about the occupants and their

behaviour, neighbours' behaviour (in the case of flats), and climate conditions. An **EPC** presents two figures: energy consumption expressed in kWh/m²/year and greenhouse gas emissions given in kg_{CO₂}/m²/year. Two labels are derived from these quantities, ultimately resulting in an **EPC** label as described in the Introduction 1 and in Figure IV.2.

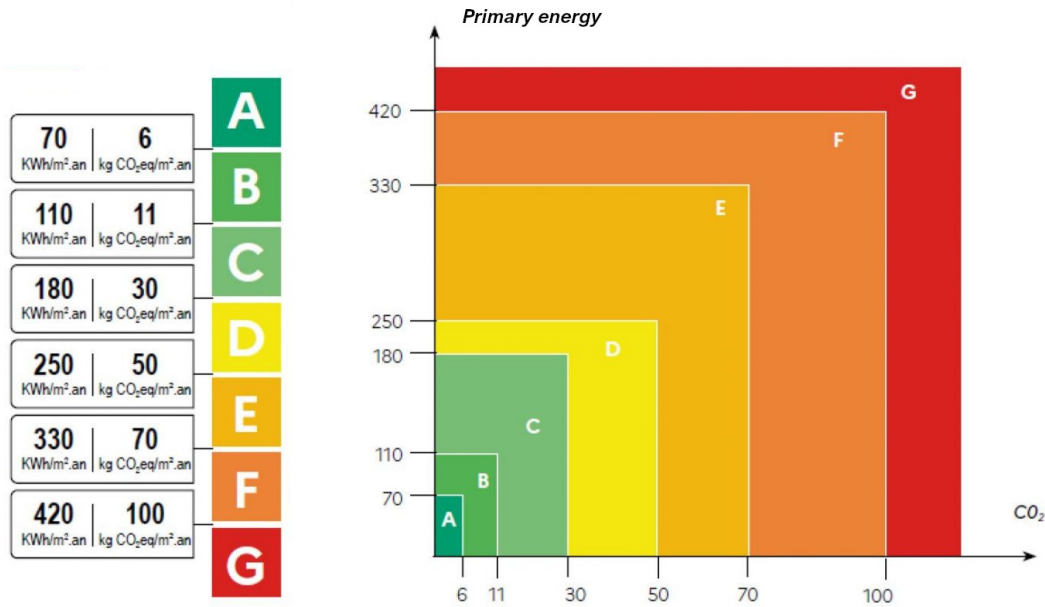


Figure IV.2 – Double threshold process used to determine the French **EPC** label of a building.

The database containing all diagnostics, encompassing every structural, quantitative, and qualitative detail about dwellings, is publicly available as open data and continuously updated by the **ADEME** (*Agence Française pour la Transition Écologique – French Agency For Ecological Transition*). These observations are important in several aspects. They result from direct visual inspections of the buildings, which enhances their reliability. They are used to generate a legal document for which the technician is held accountable. Additionally, they include a set of recommendations to improve energy efficiency. The main limitation of this database is the difficulty in precisely geolocating the visited buildings. This is because only an address is provided, without any land plot specification. Technicians input this address manually without connecting to the national address database. This is a source of ambiguities, as can be seen in Figure IV.3. Consequently, addresses may lack standardisation, contain ambiguities, and have missing information.

These observations of dwellings' energy efficiency form the learning set for attempting to predict the energy efficiency of all French buildings. In the following, we focus on the French region called Pays de Loire, in the west of France. This region presents a homogeneous climatic environment and comprises a mix of mid-size cities and rural areas. The distribution of observations is presented in Figure IV.4 for a representative sample of 50 000 buildings.



EPC observed at
"La Montagne 22350 Plumaudan"

Ministry of Finances says that there are
6 houses numbered 1, 2, 3, 4, 6, 8.

Figure IV.3 – Problematic case: Matching EPC observations with Ministry of Finances database. A technician lists a visit to a house in "La Montagne" hamlet, which, in fact, comprises 6 separate houses. It is difficult to find out which one of the 6 houses was visited.

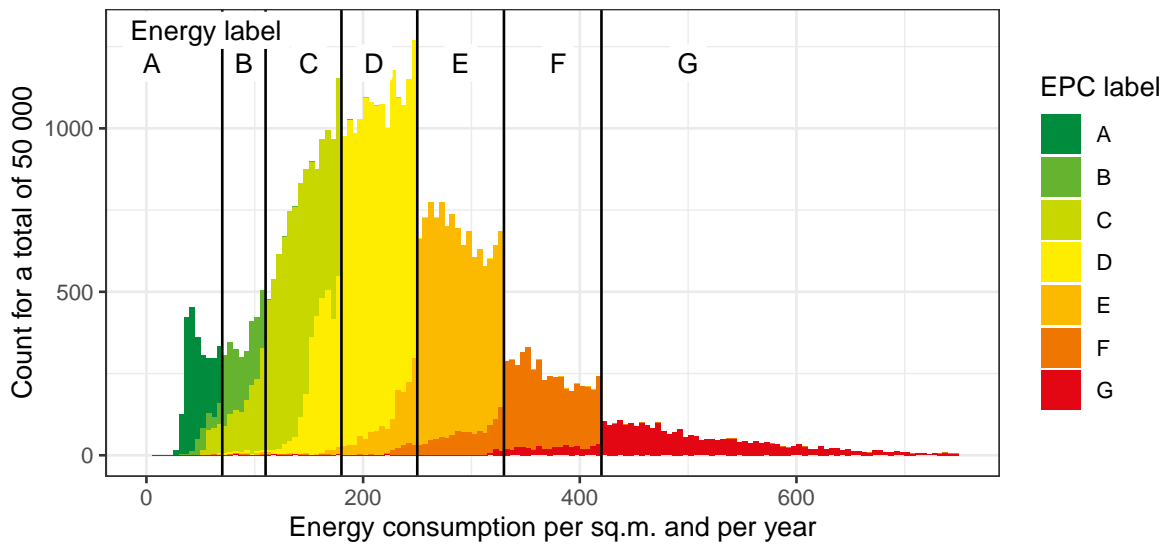


Figure IV.4 – Histogram of observed energy consumption for 50 000 buildings in the French region of Pays de la Loire. Vertical lines indicate energy label thresholds. Colours represent the worst label between energy and GHG labels. Threshold effects are evident between energy labels D to G. See also Section 1 of Chapter I, page 52, for details about threshold effects.

3 Methodology

Two approaches are possible when trying to predict the **EPC**. The first is to treat it as a regression problem, in which case the target variable is the standardised energy consumption or the **GHG** emissions quantity. The second approach is to consider it as a classification problem. The main goal of this work is to be able to detect energy-inefficient dwellings, also known as energy sieves (**EPC** labels F and G), and energy-saving dwellings (**EPC** labels A and B). We do not intend to compute energy consumption or a confidence interval. In fact, considering the known variability of the **EPC** depending on the technician, we can assume that predicting energy consumption would come with such a large confidence interval that it would cover more than a label span. In this section, we therefore treat the **EPC** prediction problem as a classification problem. We intend to predict the **EPC** label at the address level as accurately as possible, respecting, as much as possible, the overall distribution of **EPCs** over a territory. To measure the model’s performance, quantitative indicators are introduced in Subsection 3.1.

The data fusion process summarised in Section 1 produces a table of addresses that contains more than 250 variables. A lot of them are either irrelevant for the **EPC**, such as the distance of the address to the nearest school, or impossible to value, such as the identifier of the census tract where the address is located. Among the numerical or categorical variables that can be used in a predictive model, there is still some variable selection to perform to reduce as much as possible the noise in the input data. This process is described in Subsection 3.2. Eventually, we propose two supervised fuzzy classification models in Subsections 3.3 and 3.4. One is based on **KNN**, the other on Kriging.

3.1 Performance Indicators

Following early works on fuzzy sets, Ruspini introduced in 1969 a fuzzy classification approach where an individual, in our case a vector in the feature (input) space, is assigned a “degree of membership” for each of the possible classes (fuzzy sets), which are in our case labels A to G. Ruspini introduces the condition for membership degrees to be of sum equal to 1, so that they represent the probability of each class knowing the feature vector. However, his approach has turned out to be very computationally intensive [115]. The algorithms presented in this work, using the **KNN** or Kriging approaches, reach suboptimal results as compared to Ruspini’s, but in a very reasonable time.

For supervised hard classification, meaning classification that predicts a single class, the base indicator is the confusion matrix, after which other indicators are computed. However, it is seldom used because of its complexity, and, depending on the problem to solve, more synthetic performance indicators can be derived. As far as hard classification is concerned, the confusion matrix is defined to be the matrix of which element (i, j) counts the number of **EPCs** observed as label i and predicted as label j . We propose here

a new definition of the confusion matrix in order to generalise it to fuzzy classification, which involves predicting membership degrees between 0 and 1 for each class instead of predicting classes. For a given EPC, the seven membership degrees associated with the seven classes sum to 1.

Definition 3 (Confusion matrix, accuracy, balanced accuracy). *Let \mathbf{M} and $\hat{\mathbf{M}}$ be two matrices associated, respectively, with true membership degrees and predicted membership degrees of a given set of buildings, with one building per row and $c = 7$ columns each. The associated confusion matrix is:*

$$\mathcal{C}_{\mathbf{M},\hat{\mathbf{M}}} = \mathbf{M}^\top \hat{\mathbf{M}}$$

The accuracy of the prediction is the sum of the diagonal elements of $\mathcal{C}_{\mathbf{M},\hat{\mathbf{M}}}$ divided by the sum of all its elements.

$$\text{Acc}_{\mathbf{M},\hat{\mathbf{M}}} = \frac{\text{diag}[\mathcal{C}_{\mathbf{M},\hat{\mathbf{M}}}] \mathbf{1}_c}{\mathbf{1}_c^\top \mathcal{C}_{\mathbf{M},\hat{\mathbf{M}}} \mathbf{1}_c} \text{ where } \mathbf{1}_c \text{ is a vector of } c \text{ ones.}$$

The balanced accuracy of the prediction is the mean value of each label's accuracy, which is an element of $\mathcal{C}_{\mathbf{M},\hat{\mathbf{M}}}$'s diagonal divided by the sum of the elements in its row.

$$\text{BA}_{\mathbf{M},\hat{\mathbf{M}}} = \frac{1}{c} \mathbf{1}_c^\top \left(\frac{\text{diag}[\mathcal{C}_{\mathbf{M},\hat{\mathbf{M}}}] \mathbf{1}_c}{\mathcal{C}_{\mathbf{M},\hat{\mathbf{M}}} \mathbf{1}_c} \right) \text{ where the second fraction denotes a term-wise division.}$$

For instance, let us assume that we have $c = 3$ classes and 5 observations:

$$\text{If } \mathbf{M} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}; \hat{\mathbf{M}} = \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.7 & 0.1 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.8 & 0 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}, \text{ then we have } \mathcal{C}_{\mathbf{M},\hat{\mathbf{M}}} = \begin{pmatrix} 1.2 & 0.4 & 0.4 \\ 0.4 & 1.4 & 0.2 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$$

$$\text{Acc}_{\mathbf{M},\hat{\mathbf{M}}} = \frac{1.2 + 1.4 + 0.3}{5} = 0.58 \text{ and } \text{BA}_{\mathbf{M},\hat{\mathbf{M}}} = \frac{1}{3} \left(\frac{1.2}{2} + \frac{1.4}{2} + \frac{0.3}{1} \right) = 0.53.$$

In the case of hard classification, true classes and predicted classes are specific instances of fuzzy classification, wherein one membership degree is 1 and the others are null. One can verify that Definition 3 aligns with the usual definitions of accuracy and balanced accuracy for hard classification.

It is worth noting that, as depicted in Figure IV.4, labels C, D, and E are much more frequent than labels A, B, F, and G. However, decision-makers have a particular interest in identifying buildings labelled F or G (energy sieves) and A or B (energy-efficient buildings). A model that exclusively predicts labels C, D, and E may exhibit good accuracy but could still be irrelevant for decision-makers. Therefore, the balanced accuracy indicator aids in identifying models with both good accuracy and relevance.

Since **EPC** labels A to G are ordered, it is also pertinent to assess the proximity of predictions to the true values. We define the accuracy ± 1 label ratio as follows:

$$\text{Acc}_{\pm 1} = \frac{\sum_{-1 \leq i-j \leq 1} \mathcal{C}_{\mathbb{M}, \hat{\mathbb{M}}}[i, j]}{\mathbf{1}_c^\top \mathcal{C}_{\mathbb{M}, \hat{\mathbb{M}}} \mathbf{1}_c}.$$

For example, if an observation is classified as C, we are interested in knowing if the prediction falls within the range of B, C, or D, and not A, E, F, or G. Similarly, we can extend this concept to define accuracy for ranges beyond ± 1 label, such as ± 2 , 3, 4, 5, or 6 labels.

In the above example, it yields:

$$\text{Acc}_{\pm 1} = \frac{1.2 + 1.4 + 0.3 + 0.4 + 0.4 + 0.4 + 0.2}{5} = 0.86$$

Eventually, since we are interested in producing predictions that reflect the population according to the overall distribution of predicted labels, we compare the label distribution for a representative sample with the distribution of predicted labels for the same sample. This distribution is estimated by computing the membership degree's mean value for each label.

3.2 Variable Selection

After completing the data fusion process, each building exhibits a large number of features, not all of which are usable or relevant for predicting the **EPC**. We have identified the potentially useful features as presented in Table **IV.1**.

Variable selection is first implemented using forward selection with **KNN** for hard classification. Each variable is tested separately; the best one is selected, say v_1 ; each of the remaining variables is tested with v_1 ; and the best pair, v_1, v_2 , is selected. And so on, as long as the performance indicator, in our case, balanced accuracy, improves. The process stops when balanced accuracy does not improve anymore. This variable selection process has been performed for each one of the 12 French regions separately using fifty thousand **EPC** observations, forming a representative sample of the building stock with regard to the construction period and status (house/block of flats. For each feature, we have identified whether it was selected and its rank. Those who have been selected only once, never, or only in the last steps of the selection process have been ignored.

Based on this first subset of variables, a second variable selection was implemented using the algorithm presented as “A stochastic approximation approach to simultaneous feature weighting and selection for nearest neighbour learners” in [116], maximising the balanced accuracy. This algorithm optimises the variables' weights in the distance measure that is used to compute the distance between two individuals, in our case, two buildings. It computes an approximated gradient based on the averaging of multiple directional derivatives. The algorithm performs simultaneously weights' optimisation

and variable selection, therefore providing a powerful tool, especially calibrated for **KNN**. In this process, the weight of some variables tends towards 0, making it handy for variable selection. Out of the 47 variables, 18 coefficients end up nearly null, while 29 coefficients have non-null values. Figure IV.5 presents the final weights. In the following, we work with those 29 variables.

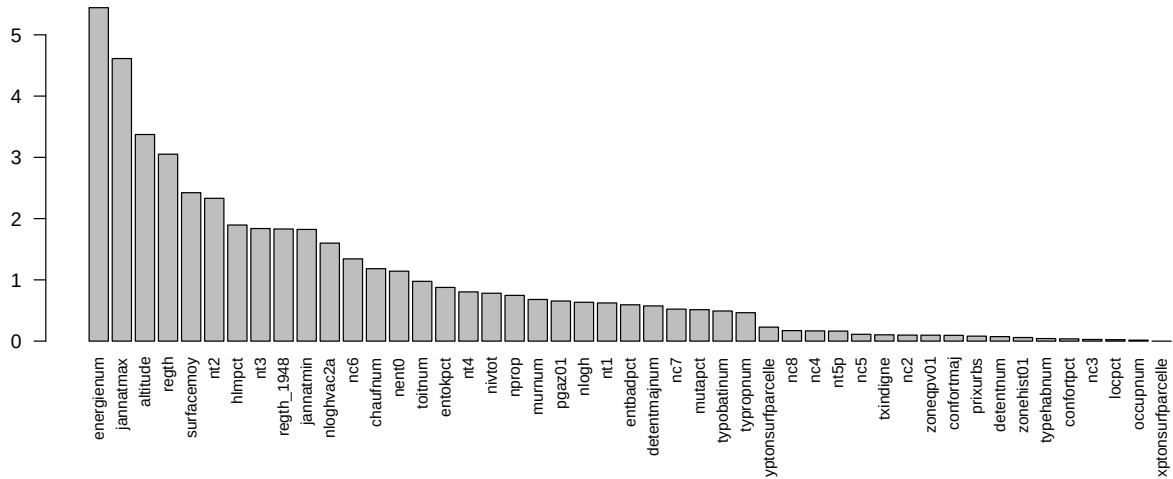


Figure IV.5 – Optimal features' weights for the **KNN** algorithm. See the dictionary of variables in Supplementary Material 11, page 264.

Type of feature	Features
Geographic location	Latitude, longitude, and altitude.
Physical features	Roof and main walls material, total living space, number of storeys, house/block of flats
Descriptive features	Number of dwellings, average living space of dwellings, number of flats of each type (with 1, 2, 3, 4+ rooms), year of construction for the oldest part of the building, type of energy saving regulation at construction time (identified by an integer increasing for each new regulation), year of construction for the newest part of the building.
Heating system	Individual or collective system, heating source of energy (electricity, city gas, wood, oil, other), availability of a city gas connection in the building.
Dwellings' quality	Comfort level, maintenance level.
Surroundings	Indication of a nearby national heritage building, indication that the building is located in a priority area (meaning a qualified underprivileged area).
Inhabitants & owners	Type of owner (individual, private company, state), type of occupant (owner/tenant), number of vacant dwellings, tax status regarding the occupation (occupied for free, occupied by a farming worker, rented free of furniture, rented as a fully furnished dwelling), number of dwellings that are unfit for renting, indication that a dwelling has been sold in the last year, price per square metre, number of social housing units, number of different owners owning dwellings in the building.

Table IV.1 – Features identified as potentially useful to predict the **EPC**.

3.3 Fuzzy Classification with KNN

The fuzzy k -nearest neighbour classifier known as **FKNN** and presented in [59] assigns a membership degree to each class of a given categorical feature for any unobserved individual. In the case of **EPC** prediction, this means that for each unobserved building, membership degrees can be predicted for all 7 **EPC** labels from A to G. These membership degrees are positive and sum to 1.

The algorithm proceeds as follows:

1. Begin with a labelled set consisting of buildings with known **EPC** labels, forming the observed buildings.
2. Select an unobserved building y .
3. For each **EPC** label L ranging from A to G, find the k -nearest neighbours of y that have label L in the labelled set. They are denoted x_1^L, \dots, x_k^L .
4. For each **EPC** label L , compute the membership degree $u_L(y)$ of y in class L :

$$u_L(y) = \frac{\sum_{j=1}^k 1/\|y - x_j^L\|^2}{\sum_{L=A}^G \sum_{j=1}^k 1/\|y - x_j^L\|^2} \quad (\text{IV.1})$$

In Equation (IV.1), $\|\cdot\|^2$ represents the squared Euclidean distance. However, refining this model by applying rescaling factors to feature variables is of interest. In this case, each feature is divided by a positive number. These factors are optimised with the same stochastic method employed for variable selection, maximising balanced accuracy. Additionally, if one is interested in hard classification, the label with the largest membership degree is attributed to y .

Although membership degrees in **FKNN** are not strictly defined as probabilities, they possess properties akin to probabilities, allowing for interpretation as such. Notably, each class has a strictly positive membership degree. However, given the ordered nature of classes in our scenario, if an individual is very likely to be of class A, it might be likely to be of class B, but it should be very unlikely to observe it in class G. The model presented in Subsection 3.4 introduces the possibility of having negative membership degrees.

3.4 Fuzzy Classification with Kriging

KNN is a classification algorithm that predicts the class of a given individual by considering a finite number of its neighbours. It is reasonable to assume that results could improve if we consider all neighbours, assigning them decreasing importance as they are further from the individual being predicted. Thus, instead of considering the number of nearest neighbours, Kriging replaces this with a characteristic distance, often referred to as range. While the statistical interpretation of **KNN** is challenging, Kriging minimises the predicted mean error, ensuring the best predictor in a well-defined sense. Moreover, Kriging can be expressed with a close formula that allows for the inclusion of constraints. However, Kriging is not inherently a classification algorithm, and additional conditions are needed to use it as a fuzzy classifier, which is the objective of Joint Kriging.

The Joint Kriging algorithm [117], [57] assigns a classification score for each EPC label to an unobserved building. These classification scores sum to 1 and, when positive, can be interpreted as probabilistic membership degrees. However, these classification scores may also take negative values, indicating both positive and negative confidence levels for each class. When assessing the model, one can encounter a confusion matrix with negative elements. For instance, considering the example presented to illustrate Definition 3, one could observe a confusion matrix such as presented in Equation (IV.2). In this context, predicting a negative classification score for label G for a given individual is interpreted as the individual being a counter-example of label G. In terms of ordered classes, the individual is “far” from being in class G.

$$\mathcal{C}_{\mathbf{M}, \hat{\mathbf{M}}} = \begin{pmatrix} 1.4 & 0.8 & -0.2 \\ 0.4 & 1.4 & 0.2 \\ -0.1 & 0.4 & 0.7 \end{pmatrix} \quad (\text{IV.2})$$

When predicting a set of unobserved buildings, the model can be further constrained so that the average classification scores for each label are defined by the user. This is particularly valuable for predicting EPC labels because the labelled set is large enough to extract a subset that is representative of the complete building stock. Consequently, an estimated average distribution of labels can be derived. Moreover, as mentioned in Subsection 3.1, it is desirable to minimise the risk of exclusively predicting labels C, D, and E. By constraining the average membership degree for each label, the model ensures sufficient weight is assigned to each label.

This model offers flexibility, as it may be constrained to simulate multiple scenarios of the predicted output, which FKNN can't do. It takes into account the ordered nature of the classes. However, the probabilistic aspect is lost due to the negative classification scores. Positive scores can be retrieved with an appropriate setting, as detailed in Proposition 9, page 114.

4 Results and Discussion

Here is the learning/test process for each algorithm:

KNN For the region of Pays de Loire in France, a **FKNN (Fuzzy k-Nearest Neighbours)** model coupled with **SPSA** pseudo-gradient descent was run based on the 29 selected variables. The model was trained on a sample of 15 000 randomly selected buildings from observations. Predictions were made using 10-fold cross-validation, selecting the 3 nearest neighbours for each label. The resulting weights were tested using the learning set to predict a test set of 50 000 observations, representative of the complete building stock. Membership degrees (positive and summing to 1) were predicted for fuzzy classification and binarised for hard classification.

Joint Kriging Based on the 29 selected variables, a joint Kriging model was run on a balanced sample (the same number of observations for each label) of

5000 observations. A preliminary step of variable selection reduced the number of variables to 9. The learning sample was then used to predict a test set of 6000 observations, representative of the building stock. Joint Kriging predicts classification scores, which can also be binarised.

Random Forest The same learning and test sets as for Joint Kriging were used. Variable selection was performed based on the same 29 selected variables using the **VSURF (Variable Selection Using Random Forest)** algorithm [118], resulting in 9 selected variables. Random Forest is a hard classifier and does not predict membership degrees.

The list of selected variables can be found in Supplementary Material 12, page 265, and the dictionary of variables is available in Supplementary Material 11, page 11. The selection of variables informing about the building's age or about the heating system and source of energy is expected. However, neither Joint Kriging nor Random Forest select variables informing about the building's material (walls, roof). Instead, these models favoured socio-economic variables such as the number of owners occupying their dwellings and the percentage of dwellings under the social housing system. Moreover, both Joint Kriging and Random Forest select latitude and longitude as meaningful variables, indicating that **EPCs** are geographic information in the sense that a building located near an observed building is likely to have the same **EPC** label as the observed building.

In addition to the results given in Tables IV.2 and IV.3, the complete confusion matrices are available in Supplementary Material 13, page 266. These results demonstrate a diversity of behaviours among models. When considering balanced accuracy, our key indicator for this study, Joint Kriging performs best (0.434) for fuzzy classification, while Random Forest performs best (0.451) for hard classification. However, while Joint Kriging maintains its superior performance in terms of accuracy for fuzzy classification, Random Forest is surpassed by both **KNN** and Joint Kriging in the case of hard classification. This suggests that Random Forest struggles to predict the more common classes accurately but performs well in rare classes. Consequently, Random Forest predicts labels A, B, F, and G more frequently than their actual occurrence, with labels F and G being predicted 76% more frequently than their actual frequency.

Joint Kriging demonstrates strong scores across all indicators for fuzzy classification and predicts a label distribution identical to the actual population distribution, making it the most effective model overall. However, when classification scores are binarised for hard classification, there is a decrease in performance for Joint Kriging, although its performance remains consistent between **KNN** and Random Forest. The frequency of predicted energy sieves (labels F and G) is only 18% higher than their actual frequency.

While **FKNN** is outperformed by Joint Kriging for fuzzy classification, its performance significantly improves when membership degrees are binarised for hard classification. Particularly, it achieves the best accuracy within one label. Although **FKNN** underestimates the frequencies of labels A, B, F, and G, the discrepancy with observed

Fuzzy classification									
Indicator	FKNN + SPSA			Joint Kriging			Random Forest		
Model optimisation criterion	Balanced accuracy			Balanced accuracy			Gini impurity		
Balanced accuracy	0.269			0.434			N.A.		
Accuracy	0.284			0.387			N.A.		
Accuracy ± 1 label	62.5 %			82.6 %			N.A.		
Accuracy of A or B	45.0 %			94.3 %			N.A.		
Accuracy of C, D or E	66.7 %			85.1 %			N.A.		
Accuracy of F or G	35.0 %			46.7 %			N.A.		
Adequacy between learnt and predicted distributions		True	Predicted		True	Predicted			
	A	0.040	0.080	A	0.034	0.034			
	B	0.035	0.079	B	0.027	0.027			
	C	0.191	0.189	C	0.191	0.191			N.A.
	D	0.334	0.245	D	0.347	0.347			
	E	0.229	0.190	E	0.234	0.234			
	F	0.109	0.127	F	0.104	0.104			
	G	0.060	0.090	G	0.063	0.063			
Hellinger distance	0.119			0.000			N.A.		

Table IV.2 – Performances of the 3 compared models for fuzzy classification.

Hard classification									
Indicator	FKNN + SPSA			Joint Kriging			Random Forest		
Balanced accuracy	0.358			0.383			0.451		
Accuracy	0.409			0.387			0.371		
Accuracy ± 1 label	79.5 %			73.6 %			72.4 %		
Accuracy of A or B	49.7 %			63.3 %			78.3 %		
Accuracy of C, D or E	86.6 %			76.1 %			64.2 %		
Accuracy of F or G	36.5 %			45.2 %			66.0 %		
Adequacy between learnt and predicted distributions		True	Predicted		True	Predicted		True	Predicted
	A	0.040	0.036	A	0.034	0.066	A	0.037	0.075
	B	0.035	0.028	B	0.027	0.044	B	0.032	0.083
	C	0.191	0.181	C	0.191	0.130	C	0.198	0.161
	D	0.334	0.370	D	0.347	0.411	D	0.337	0.187
	E	0.229	0.242	E	0.234	0.150	E	0.229	0.201
	F	0.109	0.093	F	0.104	0.082	F	0.105	0.148
	G	0.060	0.050	G	0.063	0.119	G	0.062	0.146
Hellinger distance	0.038			0.133			0.180		

Table IV.3 – Performances of the 3 compared models for hard classification.

frequencies is much smaller than that of Random Forest.

Overall, it is noteworthy that these models can reasonably predict the **EPC** label with a minimal number of variables compared to the parameters required to compute a building's energy efficiency.

We are also interested in extracting information from fuzzy classification predictions, membership degrees for **FKNN**, and classification scores for Joint Kriging. Table IV.4 illustrates that fuzzy classification effectively captures class orders. For a set of buildings with a given label, we compute the mean values of the fuzzy indicators, membership degree of **KNN**, and classification score for Joint Kriging. The mean fuzzy indicator is consistently highest for the true label, with the true label's neighbours being given the two next largest values. For example, according to Joint Kriging, buildings with true label F have a mean classification score of 0.24 for F, 0.23 and 0.16 for G and E, with the four other scores considerably smaller. This observation raises questions about the probability of finding the true label among the top two or three fuzzy indicators. In the case of **FKNN**, the true label is among the top three membership degrees in 83 % of the sample studied. Similarly, for Joint Kriging, 76 % of energy sieves (true labels F or G) have F or G among their top two classification scores. These results highlight the added value of fuzzy classification for detecting potential energy sieves.

5 Conclusion

After presenting the scientific context and the main goals of this work, a data fusion approach has been implemented to construct a data table that gathers all available information about dwellings, including geographical, structural, legal, or socio-economic aspects. This data table has been matched with **EPC** observations, creating a learning set comprising millions of observations and hundreds of features. To learn from this dataset, variable selection was necessary. Forward selection with **KNN** reduced the number of variables to 47. **SPSA** for **FKNN** reduced this subset to 29. Forward selection with Joint Kriging further reduced the number of variables down to 9, with the same number of variables selected by the **VSURF** algorithm. The results of fuzzy classification and hard classification were compared using the same parameters, thanks to a generalisation of confusion matrices.

Results indicate that for hard classification, if an **EPC** label is predicted, there is a 70 % to 80 % probability that the true label matches the predicted label or one of the two adjacent labels. While this may not be sufficient for legal purposes, it is adequate for identifying buildings likely to be energy-inefficient or energy-efficient, which is the primary objective of this article. Although Random Forest appears to have promising results for hard classification, it significantly distorts the distribution of labels, resulting in a considerable overestimation of energy-inefficient buildings, which is a notable drawback. Joint Kriging and **FKNN** fairly reproduce the overall distribution of labels, but energy-inefficient buildings remain challenging to predict, as half of them are not detected. These challenges justify the decision to employ fuzzy classification, which proves

true EPC	mean classification score in Fuzzy Joint Kriging						
	A	B	C	D	E	F	G
A	0.87	0.28	0.05	0.00	-0.03	-0.02	-0.02
B	0.25	0.44	0.10	0.01	-0.04	-0.05	-0.05
C	0.02	0.16	0.40	0.20	0.11	0.10	0.06
D	0.11	0.19	0.33	0.43	0.36	0.29	0.22
E	-0.03	0.05	0.15	0.25	0.32	0.27	0.23
F	-0.11	-0.05	0.01	0.08	0.16	0.24	0.23
G	-0.10	-0.07	-0.03	0.03	0.11	0.16	0.35

true EPC	mean membership degree in Fuzzy KNN						
	A	B	C	D	E	F	G
A	0.32	0.14	0.07	0.07	0.07	0.06	0.08
B	0.16	0.28	0.09	0.07	0.06	0.06	0.06
C	0.14	0.20	0.34	0.18	0.13	0.12	0.12
D	0.15	0.16	0.23	0.31	0.22	0.20	0.18
E	0.11	0.10	0.13	0.19	0.26	0.22	0.20
F	0.07	0.07	0.08	0.11	0.15	0.20	0.19
G	0.05	0.05	0.05	0.08	0.11	0.15	0.17

Table IV.4 – Fuzzy classification in relation with true labels.

efficient in capturing secondary information indicative of energy-inefficient buildings.

Despite our efforts, we were unable to find any quantitative results to compare this work with. However, the mass prediction of buildings energy efficiency for enhancing renovation efforts is undeniably a major challenge, and we hope that this work will encourage other teams to publish their methodologies and results. Only then can we truly advance sustainability.

Acknowledgments This work has been jointly funded by Mines Saint-étienne School of Engineering, **U.R.B.S.** company, and a Ph.D. grant from **ANRT (French National Agency for Research and Technology)**. The data fusion process presented in Section 2 is the result of teamwork in which **U.R.B.S.** engineers played a crucial role, and their contribution is acknowledged. The authors also thank Nathan Seychal for implementing the **SPSA** algorithm for **FKNN**. Additionally, they would like to express their gratitude to Maximilien Brossard for his advice and support.

Conclusion and Future Directions

1	An Improved Knowledge...	.156
2	Significance of These Results for Decision Makers.	.157
3	Limitations	.158
4	Future directions.	.159
	Afterword	.161

1 An Improved Knowledge of the Buildings' Energy Efficiency

The initial material for this work is a set of 8 million **EPC (Energy Performance Certificate)** representing less than 20% of the French building stock and the need for providing some knowledge about the energy efficiency of each and every building. The purpose is to support the massification of buildings' renovations to improve the energy sustainability of the country. To serve our purpose, we convene most of the data that is available at the national scale. We try to identify to what extent this data informs us about the energy efficiency of the buildings, even though no physical visit has been made. The data fusion process, its challenges, and the resulting uncertainties are described in Chapter I. We also show how to homogenise buildings' features by using normalised pseudo-coordinates.

In Chapter II, we propose a model that takes into account the uncertainty of the buildings' position, and we show that this very position contains information about energy efficiency. Based on a building's position and age, it is possible to predict the **EPC** label with a balanced accuracy of 23% and to predict the standardised energy consumption with a **RMSE** of 106 kWh/m²/year. This first step in our journey turns out to be very informative about the potential of our data. But, unfortunately, the upscaling of the model is too costly, and it is not possible to implement Mixture Kriging at the scale of a department or, even less, a region.

At the same time, a model to be implemented in production at **U.R.B.S.** was designed. It is based on a **KNN** model, takes a lot of input variables on which it performs variable selection, and selects an optimal data transformation as well as a distance model. In December 2022, this model showed a balanced accuracy ranging from 23% to 29% depending on the region in France. This result proves the benefit of variable selection in improving predictions. These results have not been published, but they are provided as Supplementary Material 14, page 270.

We found that a limiting factor in improving these models was the regression approach. Since the thresholds defining the **EPC** labels are not linearly distributed, the same error in terms of energy consumption may result in a major label error if we are dealing with small energy consumption and a minor label error if we are dealing with large energy consumption. This is introducing a bias in the optimisation process that is difficult to remove. And a consequence is that the distribution of predicted labels may diverge too much from their real distribution. We tried unsuccessfully to work with a transformed version of energy consumption. Then we decided to turn towards classification. Classification is also more relevant after the 2021 legal changes because it allows the direct prediction of the **EPC** without going through the double thresholding process.

After testing fuzzy classification with a modified **KNN** process, we found that, on the one hand, the performance is improved, but on the other hand, the predicted distribution is still diverging from the real distribution of labels. **KNN** does not offer the flexibility

for imposing such constraints as a known predicted distribution. But since Kriging is based on a closed formula, theoretical tools are available to impose constraints on it. This is the basis for Joint Kriging presented in Chapter III. This model works with a double system of constraints: it predicts membership degrees that should sum to 1 for each prediction, and it predicts multiple points at the same time with a constraint on the predicted distribution. Joint Kriging yields a balanced accuracy equal to 38%, which is a significant improvement as compared with previous models.

This work improves the prediction of EPCs from 23% to 38% regarding the balanced accuracy. The accuracy is also improved, as is the detection of energy sieves. These quantitative results show that knowledge of the buildings' energy efficiency has improved. The comparative study of the models we tested is presented in Chapter IV.

From a qualitative point of view, our work proves that EPCs can be regarded as geolocated data and that socio-economic factors can improve the prediction of the EPCs.

These research findings have been peer-reviewed, published, and presented at conferences as follows:

- Chapter II: The model has been presented as a poster in the *MASCOT-NUM 2022* conference in Clermont-Ferrand, where comments brought up by the poster showed that some mathematicians were already wondering why geospatial interpolation had not been implemented for buildings yet. It was also presented in an oral presentation on the occasion of the *Spatial Statistics 2023* conference in Denver, Colorado. And it was published in the *Energy and AI* journal in 2024 [52]. This journal has been in the first quarter for engineering, energy, and artificial intelligence since 2021. This model was also presented from the programming point of view in the *Rencontres R 2023* in Avignon. It was an opportunity to discuss computational complexity and efficiency with peers.
- Chapter III: The model has been presented in an oral presentation on the occasion of the *Statistics Seminar* of the Mathematics Laboratory of Avignon (LMA – Laboratoire de Mathématiques d'Avignon). The chapter itself is submitted as an article.
- Chapter IV: This study is published in the *Energy and Buildings* journal [95], which has been in the first quartile for building and construction journals since 2000.

2 Significance of These Results for Decision Makers

The decision to renovate a building is based on a set of qualitative and quantitative factors. In spite of financial incentives, it is known that the owners do not play a proactive enough role. Decision-makers can help in this matter by providing human services to facilitate the renovation process. But the raw knowledge of the building stock does not allow for a good identification of energy sieves and their owners. Our research provides a way to do it. Thanks to the U.R.B.S. software, the ONB (*Observatoire*

National des Bâtiments – National Buildings Observatory), our predictions are directly accessible to the institutions for queries.

To our knowledge, there is only one other laboratory in France that focuses on the same topic: *CSTB (Centre Scientifique et Technique du Bâtiment – Scientific and Technical Centre for Building)*. They have a very different approach; they focus on predicting the buildings’ structural indicators to compute an *EPC* based on a simplified thermal engineering approach to avoid a physical visit to the building. The advantage of this method is its ability to test renovation scenarios and their effects. But the propagation of errors generated by the aggregation of multiple predicted indicators increases prediction uncertainty. Unfortunately, to the extent of our knowledge, the *CSTB* did not publish any quantitative results that would allow a comparison of performances.

Therefore, at this point, *U.R.B.S.* is the only actor able to predict *EPCs* with a known measurable performance. As a result, many institutions join the *U.R.B.S.* ecosystem, and we benefit from their feedback. At this point, the quantity of data *U.R.B.S.* is learning from and the models’ performance make it difficult to improve our prediction models. But thanks to the users’ feedback, the algorithms’ optimisation, and maybe the use of black box models, it is possible to continuously improve these prediction models.

3 Limitations

The main limitation of our work is that some information potentially useful for better predictions is missing. In particular, there are multiple renovation actions that are not submitted to any official authorization and are difficult to detect from the data, although they may have an impact on a building’s energy efficiency. Before starting the research, we expected our results to allow for the definition of a renovation indicator. But it turns out that the uncertainty of our prediction does not allow for predicting with a good level of confidence whether a building has been renovated or not, and to what extent. An unpublished study by Jonathan Villot tends to suggest that such information would significantly improve the knowledge of buildings’ energy efficiency.

Another limitation of this work is the initial constraint that was set to work with explainable models. The knowledge we have acquired about the data and the familiarisation of the users with this kind of model suggest that we could relax this constraint. The performance of Random Forest, for instance, is described in Chapter IV. In particular, we closely followed the fast advances in neural networks for tabular data. For the record, we have tested the neural network *tabnet* for the region Pays de la Loire, and the results show a good balanced accuracy (40 %) and an average accuracy (36 %). But, similarly to Random Forest, it appears that *tabnet* strongly distorts the predicted distribution as it predicts three times more labels A and G than in the original distribution. This is a preliminary study that would probably deserve more investigation.

4 Future directions

The first task that results from this work is the implementation of Joint Kriging in production at **U.R.B.S.** We already know that it is possible from the tests presented in Chapter **IV**, but there remain unknowns about the variability of the model depending on the French regions.

As mentioned above, we believe that the explainability constraint can be relaxed considering the amount of knowledge we gained during these three years of research. Our tests with Random Forest and **tabnet** show that these models have good accuracy but a tendency to distort output distributions. But, in particular for **tabnet**, there are many ways to organise both the training (pre-training and sequential trainings) and the output (fuzzy classification or hard classification), and to avoid overfitting. So that there are many directions to explore and maybe improve our current results.

We have shown that the level of renovation of a building cannot be predicted as a secondary output of an energy efficiency prediction. However, there remains much unexploited data in the **EPCs** database provided by **ADEME**. In particular, the technicians include natural language data in their diagnostics to make recommendations for potential improvements to the buildings' energy efficiency. We believe that exploiting this data, together with other secondary indicators in the database, could help us derive a learnable renovation level for buildings.

Afterword

I remember our preliminary discussions about the relevance of geostatistics to describe the **EPCs**, about the descriptions of French cities with concentric circles, and about the extent of thermal engineering influence on the **EPC** as compared to politics or climate. We also started imagining ways to give flexibility to spatial interpolation to take into account position uncertainty. Land plots on one side, mixture distributions on the other side, expert knowledge, and its model. And those two fields of research, energy efficiency and geostatistics, gradually came to understand each other so much that we could imagine, test, criticise, and improve new models, i.e., Mixture Kriging and Joint Kriging models. The new knowledge brought up by this research therefore emerges from a strong belief in interdisciplinarity.

Interdisciplinary research is often encouraged but is difficult to publish. At first, we wrote an article describing Mixture Kriging that was sincerely trying to show how a model could emerge from the data. But we encountered difficulties. In particular, a reviewer gave this very explicit demand: “Either the paper is (i) about the presentation of a general **BLUP** approach or (ii) about the **EPC** prediction problem.” I was really puzzled by this comment because I had personally put some effort into doing precisely what we were blamed for. The thing becomes really interesting when, at the same time, emerges a new journal, Energy and **AI**, describing itself as follows: “Energy and **AI** provides a fast and authoritative open access platform to disseminate the latest research progress in the cross-disciplinary area of energy and artificial intelligence.” This review was an immediate success. This shows both the need for and the benefits of publishing interdisciplinary research.

Our research brings new knowledge in the field of geospatial interpolation by showing that position uncertainty can be taken into account as such in Kriging, and by showing that Kriging can be efficiently constrained for fuzzy classification. But it also brings new knowledge in the field of energy efficiency research by showing that **EPC** can be regarded as geolocated data and by showing that socio-economic predictors can also be used to predict an **EPC**. Eventually, **EPC** is not only a thermal engineering object, but it is also a legal concept that both informs about and impacts the building stock and its geography.

Conclusion et perspectives (en français)

1	Une meilleure connaissance...	.164
2	Importance de ces résultats pour les décideurs	.166
3	Limites	.166
4	Perspectives	.167

1 Une meilleure connaissance de l'efficacité énergétique des bâtiments

La matière initiale de ce travail est un ensemble de 8 millions **DPE** (*Diagnostic de Performance Énergétique – Energy Performance Certificate*) représentant moins de 20 % du parc immobilier français et la nécessité de fournir des connaissances sur l'efficacité énergétique de chaque bâtiment. L'objectif est de soutenir la massification des rénovations des bâtiments afin d'améliorer la durabilité énergétique du pays. Pour servir notre objectif, nous rassemblons la plupart des données disponibles à l'échelle nationale. Nous essayons d'identifier dans quelle mesure ces données nous informent sur l'efficacité énergétique des bâtiments, même si aucune visite physique n'a été effectuée. Le processus de fusion des données, ses défis et les incertitudes qui en résultent sont décrits au Chapitre I. Nous montrons également comment homogénéiser les caractéristiques des bâtiments en utilisant des pseudo-coordonnées normalisées.

Dans le Chapitre II, nous proposons un modèle qui prend en compte l'incertitude de la position des bâtiments, et nous montrons que cette position contient des informations sur l'efficacité énergétique. Sur la base de la position et de l'âge d'un bâtiment, il est possible de prédire l'étiquette **DPE** avec une précision équilibrée de 23 % et de prédire la consommation d'énergie standardisée avec une **RMSE** de 106 kWh/m²/an. Cette première étape de notre parcours s'avère très instructive quant au potentiel de nos données. Mais, malheureusement, le passage à l'échelle du modèle est trop coûteux, et il n'est pas possible d'implémenter le modèle Mixture Kriging à l'échelle d'un département ou, encore moins, d'une région.

Dans le même temps, un modèle à mettre en œuvre en production à **U.R.B.S.** a été conçu. Il est basé sur un modèle de type **KNN**, prend un grand nombre de variables d'entrée sur lesquelles il effectue une sélection de variables, et sélectionne une transformation de données optimale ainsi qu'un modèle de distance. En décembre 2022, ce modèle a montré une précision équilibrée allant de 23 % à 29 % en fonction de la région en France. Ce résultat prouve l'intérêt de la sélection des variables pour améliorer les prévisions. Ces résultats n'ont pas été publiés, mais ils sont fournis en tant que Supplementary Material 14, page 270.

Nous avons constaté que l'approche par régression constituait un facteur limitant pour l'amélioration de ces modèles. Étant donné que les seuils définissant les étiquettes **DPE** ne sont pas distribués linéairement, la même erreur en termes de consommation d'énergie peut entraîner une erreur d'étiquette majeure si nous avons affaire à une petite consommation d'énergie et une erreur d'étiquette mineure si nous avons affaire à une grande consommation d'énergie. Cela introduit un biais dans le processus d'optimisation qu'il est difficile d'éliminer. En conséquence, la distribution des étiquettes prédites peut fortement diverger de leur distribution réelle. Nous avons essayé sans succès de travailler avec une version transformée de la consommation d'énergie. Nous avons alors décidé de nous tourner vers la classification. La classification devient aussi plus pertinente après

les changements juridiques de 2021, car elle permet de prédire directement le **DPE** sans passer par le processus de double seuillage.

Après avoir testé la classification floue avec un algorithme **KNN** modifié, nous avons constaté que, d'une part, les performances sont améliorées, mais que, d'autre part, la distribution prédite diverge toujours de la distribution réelle des étiquettes. **KNN** n'offre pas la flexibilité nécessaire pour imposer de telles contraintes comme une distribution prédite connue. Mais comme le Krigeage est basé sur une formule fermée, des outils théoriques sont disponibles pour lui imposer des contraintes. C'est la base du modèle Joint Kriging présenté au Chapitre **III**. Ce modèle fonctionne avec un double système de contraintes : il prédit des degrés d'appartenance dont la somme doit être égale à 1 pour chaque prédiction, et il prédit plusieurs points d'un territoire en même temps avec une contrainte sur la distribution prédite. Le modèle Joint Kriging permet d'obtenir une précision équilibrée égale à 38 %, ce qui représente une amélioration significative par rapport aux modèles précédents.

Ce travail améliore la prédiction des **DPE** de 23 % à 38 % pour ce qui est de la précision équilibrée. La précision est également améliorée, de même que la détection des passoires énergétiques. Ces résultats quantitatifs montrent que la connaissance de l'efficacité énergétique des bâtiments s'est améliorée. L'étude comparative des modèles testés est présentée au Chapitre **IV**.

D'un point de vue qualitatif, notre travail prouve que le **DPE** peut être considéré comme une donnée microlocalisée et que des facteurs socio-économiques peuvent améliorer la prédiction des **DPE**.

Les résultats de ces recherches ont été évalués par des pairs, publiés et présentés lors de conférences :

- Chapitre **II** : Le modèle a été présenté sous forme de poster lors de la conférence *MASCOT-NUM 2022* à Clermont-Ferrand, où les commentaires suscités par le poster ont montré que certains mathématiciens se demandaient déjà pourquoi l'interpolation géospatiale n'avait pas encore été mise en œuvre pour les bâtiments. Il a fait l'objet d'une présentation orale à l'occasion de la conférence *Spatial Statistics 2023* à Denver, Colorado. Il a également été publié dans la revue *Energy and AI* en 2024 [52]. Cette revue est dans le premier quartile pour l'ingénierie, l'énergie et l'intelligence artificielle depuis 2021. Ce modèle a également été présenté du point de vue de la programmation aux *Rencontres R 2023* d'Avignon. Ce fut l'occasion de discuter de la complexité et de l'efficacité des calculs avec des pairs.
- Chapitre **III** : Le modèle a fait l'objet d'une présentation orale à l'occasion du *séminaire de statistique* du Laboratoire de Mathématiques d'Avignon (LMA). Le chapitre lui-même est soumis en tant qu'article.
- Chapitre **IV** : Cette étude est publiée dans la revue *Energy and Buildings* [95], qui se situe dans le premier quartile des revues sur le bâtiment et la construction depuis 2000.

2 Importance de ces résultats pour les décideurs

La décision de rénover un bâtiment repose sur un ensemble de facteurs qualitatifs et quantitatifs. Malgré les incitations financières, on sait que les propriétaires ne jouent pas un rôle assez proactif. Les décideurs peuvent les aider dans ce domaine en mettant en place des services de conseil personnalisés pour faciliter le processus de rénovation. Mais la connaissance brute du parc immobilier ne permet pas une bonne identification des passoires énergétiques et de leurs propriétaires. Notre recherche fournit un moyen de le faire. Grâce au logiciel **U.R.B.S.**, l'ONB (*Observatoire National des Bâtiments – National Buildings Observatory*), nos prédictions sont directement accessibles aux collectivités clientes pour des requêtes.

A notre connaissance, il n'existe qu'un seul autre laboratoire en France qui se concentre sur le même sujet : le **CSTB** (*Centre Scientifique et Technique du Bâtiment – Scientific and Technical Centre for Building*). Son approche est très différente ; il se concentre sur la prévision des indicateurs structurels des bâtiments pour calculer un **DPE** basé sur une approche d'ingénierie thermique simplifiée afin d'éviter une visite physique du bâtiment. L'avantage de cette méthode est qu'elle permet de tester des scénarios de rénovation et leurs effets. Mais la propagation des erreurs générées par l'agrégation de multiples indicateurs prédits augmente l'incertitude de la prédiction. Malheureusement, à notre connaissance, le **CSTB** n'a pas publié de résultats quantitatifs qui permettraient de comparer les performances.

Par conséquent, à ce stade, **U.R.B.S.** est le seul acteur capable de prédire les **DPE** avec une performance mesurable connue. Ainsi, de nombreuses institutions rejoignent l'écosystème **U.R.B.S.**, et nous bénéficions de leur retour d'information. À ce stade, la quantité de données dont **U.R.B.S.** dispose et les performances des modèles rendent difficile l'amélioration de nos modèles de prédiction. Mais grâce au retour d'information des utilisateurs, à l'optimisation des algorithmes et peut-être à l'utilisation de modèles de boîte noire, il sera possible d'améliorer continuellement ces modèles de prédiction.

3 Limites

La principale limite à notre travail est que certaines informations potentiellement utiles pour de meilleures prédictions sont manquantes. En particulier, de nombreuses actions de rénovation ne sont soumises à aucune autorisation officielle et sont difficiles à détecter à partir des données, bien qu'elles puissent avoir un impact sur l'efficacité énergétique d'un bâtiment. Avant de commencer la recherche, nous nous attendions à ce que nos résultats permettent de définir un indicateur de rénovation. Mais il s'avère que l'incertitude de notre prédiction ne permet pas de prédire avec un bon niveau de confiance si un bâtiment a été rénové ou non, et dans quelle mesure. Une étude non publiée de Jonathan Villot amène à suggérer qu'une telle information améliorerait significativement la connaissance de l'efficacité énergétique des bâtiments.

Une autre limite de ce travail est la contrainte initiale de travailler avec des modèles

explicables. Les connaissances que nous avons acquises sur les données et la familiarisation des utilisateurs avec ce type de modèle suggèrent que nous pourrions relâcher cette contrainte. Les performances de Random Forest, par exemple, sont décrites au Chapitre IV. En particulier, nous avons suivi de près les progrès rapides des réseaux neuronaux pour les données tabulaires. Pour mémoire, nous avons testé le réseau neuronal `tabnet` pour la région Pays de la Loire, et les résultats montrent une bonne précision équilibrée (40 %) et une précision moyenne (36 %). Mais, comme pour Random Forest, il apparaît que `tabnet` déforme fortement la distribution prédite puisqu'il prédit trois fois plus d'étiquettes A et G que dans la distribution originale. Il s'agit d'une étude préliminaire qui mériterait d'être approfondie.

4 Perspectives

La première mission qui résulte de ce travail est la mise en œuvre du modèle Joint Kriging en production à U.R.B.S. Nous savons déjà que cela est possible grâce aux tests présentés au Chapitre IV, mais il reste des inconnues sur la variabilité du modèle en fonction des régions françaises.

Comme indiqué ci-dessus, nous pensons que la contrainte d'explicabilité peut être assouplie compte tenu de la quantité de connaissances que nous avons acquises au cours de ces trois années de recherche. Nos tests avec Random Forest et `tabnet` montrent que ces modèles ont une bonne précision mais une tendance à déformer les distributions de sortie. Mais, en particulier pour `tabnet`, il existe de nombreuses façons d'organiser à la fois l'entraînement (pré-entraînement et entraînements séquentiels) et la sortie (classification floue ou classification dure), et d'éviter le surapprentissage. Il y a donc de nombreuses directions à explorer et pour peut-être améliorer nos résultats actuels.

Nous avons constaté que le niveau de rénovation d'un bâtiment ne peut pas être prédit en tant que résultat secondaire d'une prédiction d'efficacité énergétique. Cependant, la base de données DPE fournie par ADEME contient encore de nombreuses données inexploitées. En particulier, les techniciens intègrent des données en langage naturel dans leurs diagnostics afin de formuler des recommandations sur les améliorations potentielles à apporter à l'efficacité énergétique des bâtiments. Nous pensons que l'exploitation de ces données, ainsi que d'autres indicateurs secondaires dans la base de données, pourrait nous aider à trouver un indicateur de niveau de rénovation des bâtiments susceptible d'être appris.

Bibliography

- [1] *Directive (EU) 2018/844 of the european parliament and of the council of 30 may 2018 amending directive 2010/31/EU on the energy performance of buildings and directive 2012/27/EU on energy efficiency (text with EEA relevance)*, Accessed: Jan. 8, 2024. [Online]. Available: <http://data.europa.eu/eli/dir/2018/844/oj/eng>.
- [2] *Directive (EU) 2023/1791 of the european parliament and of the council of 13 september 2023 on energy efficiency and amending regulation (EU) 2023/955 (recast) (text with EEA relevance)*, Accessed: Jan. 8, 2024. [Online]. Available: <http://data.europa.eu/eli/dir/2023/1791/oj/eng>.
- [3] « Europe's buildings under the microscope, BPIE - buildings performance institute europe », BPIE - Buildings Performance Institute Europe, Accessed: Jan. 8, 2024. [Online]. Available: <https://www.bpie.eu/publication/europes-buildings-under-the-microscope/>.
- [4] « In focus: energy efficiency in buildings - european commission », Accessed: Jan. 8, 2024. [Online]. Available: https://commission.europa.eu/news/focus-energy-efficiency-buildings-2020-02-17_en.
- [5] « ADEME - observatoire DPE - audit (diagnostic de performance énergétique audit énergétique) », Accessed: Jun. 20, 2024. [Online]. Available: <https://observatoire-dpe-audit.ademe.fr/donnees-dpe-publiques>.
- [6] *Sous-section 2 : diagnostic de performance énergétique (articles l126-26 à l126-33) - légifrance*, Accessed: Apr. 8, 2024. [Online]. Available: https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006074096/LEGISCTA000043967326/#LEGISCTA000043967326.
- [7] *Arrêté du 15 septembre 2006 relatif au diagnostic de performance énergétique pour les bâtiments ou parties de bâtiment autres que d'habitation existants proposés à la vente en france métropolitaine*, Accessed: Nov. 12, 2021. [Online]. Available: <https://www.legifrance.gouv.fr/loda/id/LEGIARTI000025624077/2012-03-16/>.
- [8] S. Kahouli and S. Okushima, « Regional energy poverty reevaluated: a direct measurement approach applied to france and japan », *Energy Economics*, vol. 102, p. 105491, Oct. 1, 2021, ISSN: 0140-9883. DOI: [10.1016/j.eneco.2021.105491](https://doi.org/10.1016/j.eneco.2021.105491). Accessed: Apr. 8, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0140988321003777>.
- [9] H. Watt. « Comparer DPE et consommation a-t-il un sens ? », Hello Watt, Accessed: Apr. 8, 2024. [Online]. Available: <https://www.hellowatt.fr/blog/etude-comparaison-dpe-et-consommation/>.
- [10] *LOI n° 2009-967 du 3 août 2009 de programmation relative à la mise en œuvre du grenelle de l'environnement (1)*, Accessed: Apr. 11, 2024.
- [11] *LOI n° 2010-788 du 12 juillet 2010 portant engagement national pour l'environnement (1)*, Accessed: Apr. 11, 2024.

- [12] Commissariat général au développement durable. « Les émissions des gaz à effet de serre du secteur tertiaire », *notre-environnement*, Accessed: May 30, 2024. [Online]. Available: <https://www.notre-environnement.gouv.fr/>.
- [13] *LOI n° 2021-1104 du 22 août 2021 portant lutte contre le dérèglement climatique et renforcement de la résilience face à ses effets (1)*, Accessed: Apr. 10, 2024.
- [14] S. Hempel, « John snow », *The Lancet*, vol. 381, no. 9874, pp. 1269–1270, Apr. 13, 2013, Publisher: Elsevier, ISSN: 0140-6736, 1474-547X. DOI: [10.1016/S0140-6736\(13\)60830-2](https://doi.org/10.1016/S0140-6736(13)60830-2). Accessed: Apr. 29, 2024. [Online]. Available: [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(13\)60830-2/fulltext?amp=&code=lancet-site](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(13)60830-2/fulltext?amp=&code=lancet-site).
- [15] ADEME, *Méthode de calcul 3cl-DPE 2021*, 2021.
- [16] I. Ballarini, V. Corrado, F. Madonna, S. Paduos, and F. Ravasio, « Energy refurbishment of the italian residential building stock: energy and cost analysis through the application of the building typology », *Energy Policy*, vol. 105, pp. 148–160, Jun. 1, 2017, ISSN: 0301-4215. DOI: [10/gs2gpg](https://doi.org/10/gs2gpg). [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0301421517301015>.
- [17] J.-P. Rathle, « Les réductions des émissions de gaz à effet de serre liées aux rénovations. résultats de l'enquête TREMI 2020. », Observatoire national de la rénovation énergétique, Sep. 2022.
- [18] U. Ali *et al.*, « A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings », *Applied Energy*, vol. 267, Jun. 1, 2020, ISSN: 0306-2619. DOI: [10/gtn38c](https://doi.org/10/gtn38c). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306261920303731>.
- [19] P. Schetelat, L. Lefort, and N. Delgado, « Urban data imputation using multi-output multi-class classification », *Building to Buildings: Urban and Community Energy Modelling*, Nov. 12, 2020.
- [20] B. J. Smith, J. Yan, and M. K. Cowles, « Unified geostatistical modeling for data fusion and spatial heteroskedasticity with r package ramps », *Journal of Statistical Software*, vol. 25, pp. 1–21, Apr. 29, 2008, ISSN: 1548-7660. DOI: [10/gtn364](https://doi.org/10/gtn364). [Online]. Available: <https://doi.org/10.18637/jss.v025.i10>.
- [21] L. A. Zadeh, « Toward a generalized theory of uncertainty (GTU)—an outline », *Information Sciences*, vol. 172, no. 1, pp. 1–40, Jun. 9, 2005, ISSN: 0020-0255. DOI: [10/btmcw2](https://doi.org/10/btmcw2). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S002002550500054X>.
- [22] W. Pedrycz, *Granular computing : analysis and design of intelligent systems* (Industrial electronics series). Boca Raton: Taylor & Francis, 2013, ISBN: 978-1-4398-8681-6. [Online]. Available: <https://doi.org/10.1201/9781315216737>.
- [23] N. A. Cressie, *Statistics for spatial data revised edition*, John Wiley & Sons Inc. New York: Wiley series in probability, mathematical statistics. Applied probability, and statistics., 1993, ISBN: 0-471-00255-0. [Online]. Available: <https://onlinelibrary.wiley.com/doi/book/10.1002/9781119115151>.
- [24] A. Comber and W. Zeng, « Spatial interpolation using areal features: a review of methods and opportunities using new forms of data with coded illustrations », *Geography Compass*, vol. 13, no. 10, 2019, ISSN: 1749-8198. DOI: [10/gg4gbh](https://doi.org/10/gg4gbh). [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/gec3.12465>.
- [25] « La performance énergétique du parc résidentiel privé dans la Communauté urbaine Le Havre Seine Métropole », AURH. Section: Observatoires et études, Accessed: Apr. 30, 2024. [Online]. Available: <https://www.aurh.fr/observatoires-et-etudes/performance-energetique-parc-prive>.

- [26] W. Tobler, « Smooth pycnophylactic interpolation for geographic regions », *Journal of the American Statistical Association*, vol. 74, pp. 519–30, Feb. 1, 1979. DOI: [10/ghz78f](https://doi.org/10/ghz78f).
- [27] C. Williams and C. Rasmussen, « Gaussian processes for regression », in *Proceedings of the 1995 Conference*, Cambridge, Mass: The MIT Press, Jun. 1996, ISBN: 978-0-262-20107-0. [Online]. Available: <https://mitpress.mit.edu/9780262201070/advances-in-neural-information-processing-systems-8/>.
- [28] G. Matheron, « Principles of geostatistics », *Economic Geology*, vol. 58, no. 8, pp. 1246–1266, Dec. 1, 1963, ISSN: 0361-0128. DOI: [10/fdsdjx](https://doi.org/10/fdsdjx). Accessed: Feb. 6, 2023. [Online]. Available: <https://doi.org/10.2113/gsecongeo.58.8.1246>.
- [29] N. S.-N. Lam, « Spatial interpolation methods: a review », *The American Cartographer*, vol. 10, no. 2, pp. 129–150, Jan. 1, 1983, ISSN: 0094-1689. DOI: [10/cjx96x](https://doi.org/10/cjx96x). [Online]. Available: <https://doi.org/10.1559/152304083783914958>.
- [30] C. Gotway and L. Young, « Combining incompatible spatial data », *Journal of the American Statistical Association*, vol. 97, pp. 632–648, Feb. 1, 2002. DOI: [10/ctnp97](https://doi.org/10/ctnp97).
- [31] E. H. Isaaks and R. M. Srivastava, *An Introduction to Applied Geostatistics*. New York: Oxford University Press, 1989, ISBN: 978-1-61583-082-4. Accessed: Feb. 6, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0098300491900551>.
- [32] P. Kyriakidis, « A geostatistical framework for area-to-point spatial interpolation », *Geographical Analysis*, vol. 36, Aug. 1, 2004. DOI: [10/bf43m2](https://doi.org/10/bf43m2).
- [33] P. Goovaerts, « Kriging and semivariogram deconvolution in the presence of irregular geographical units », *Mathematical Geology*, vol. 40, no. 1, pp. 101–128, 2008, ISSN: 0882-8121.
- [34] A. Briz-Redon, « A bayesian shared-effects modeling framework to quantify the modifiable areal unit problem », *Spatial Statistics*, vol. 51, Oct. 1, 2022, ISSN: 2211-6753. DOI: [10/gtn379](https://doi.org/10/gtn379). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2211675322000537>.
- [35] C. Li, Z. Lu, T. Ma, and X. Zhu, « A simple kriging method incorporating multiscale measurements in geochemical survey », *Journal of Geochemical Exploration*, vol. 101, no. 2, pp. 147–154, May 1, 2009, ISSN: 0375-6742. DOI: [10/bpb9ps](https://doi.org/10/bpb9ps). [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0375674208000666>.
- [36] L. Poggio and A. Gimona, « Downscaling and correction of regional climate models outputs with a hybrid geostatistical approach », *Spatial Statistics, Spatial and Spatio-Temporal Models for Interpolating Climatic and Meteorological Data*, vol. 14, pp. 4–21, Nov. 1, 2015, ISSN: 2211-6753. DOI: [10/gdm76n](https://doi.org/10/gdm76n). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2211675315000305>.
- [37] S. N. Wood, *Generalized Additive Models: An Introduction with R, Second Edition*. Boca Raton: CRC Press, May 18, 2017, 497 pp., ISBN: 978-1-4987-2834-8.
- [38] R. Kerry, P. Goovaerts, I. P. Smit, and B. 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)*, vol. 27, no. 1, pp. 47–67, 2013, ISSN: 1365-8816. DOI: [10/gtn37x](https://doi.org/10/gtn37x). [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341904/>.
- [39] P. Truong and G. Heuvelink, « Bayesian area-to-point kriging using expert knowledge as informative priors », *International Journal of Applied Earth Observation and Geoinformation*, vol. 30, p. 2291, Apr. 1, 2013. DOI: [10/gc3f55](https://doi.org/10/gc3f55).

- [40] E.-H. Yoo and P. C. Kyriakidis, « Area-to-point kriging with inequality-type data », *Journal of Geographical Systems*, vol. 8, no. 4, pp. 357–390, Nov. 2, 2006, ISSN: 1435-5930, 1435-5949. DOI: [10/fh7ccg](https://doi.org/10.1007/s10109-006-0036-7). [Online]. Available: <http://link.springer.com/10.1007/s10109-006-0036-7>.
- [41] X. Zhang, W. Zuo, S. Zhao, L. Jiang, L. Chen, and Y. 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*, vol. 7, no. 1, Jan. 2018. DOI: [10/gcztrd](https://doi.org/10/gcztrd). [Online]. Available: <https://www.mdpi.com/2220-9964/7/1/33>.
- [42] Y. Jin, Y. Ge, J. Wang, G. Heuvelink, and L. Wang, « Geographically weighted area-to-point regression kriging for spatial downscaling in remote sensing », *Remote Sensing*, vol. 10, Apr. 9, 2018. DOI: [10/gdw9x7](https://doi.org/10/gdw9x7).
- [43] O. J. R. Pereira, A. J. Melfi, C. R. Montes, and Y. Lucas, « Downscaling of ASTER thermal images based on geographically weighted regression kriging », *Remote Sensing*, vol. 10, no. 4, p. 633, Apr. 2018. DOI: [10/gtn36z](https://doi.org/10/gtn36z). [Online]. Available: <https://www.mdpi.com/2072-4292/10/4/633>.
- [44] Q. Wang, W. Shi, and P. M. Atkinson, « Area-to-point regression kriging for pan-sharpening », *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 151–165, Apr. 1, 2016, ISSN: 0924-2716. DOI: [10/f8nwt](https://doi.org/10/f8nwt). [Online]. Available: <http://adsabs.harvard.edu/abs/2016JPRS..114..151W>.
- [45] P. Moraga, S. M. Cramb, K. L. Mengersen, and M. Pagano, « A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE », *Spatial Statistics*, vol. 21, pp. 27–41, Aug. 1, 2017, ISSN: 2211-6753. DOI: [10/gg3rqv](https://doi.org/10/gg3rqv). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2211675317301318>.
- [46] « Base de données nationale IMOPE », U.R.B.S, Accessed: Jun. 6, 2024. [Online]. Available: <https://www.urbs.fr/data/>.
- [47] Y. Abdelouadoud. « Sur le nombre de passoires énergétiques en france », Alternatives Énergétiques, Accessed: Nov. 22, 2021. [Online]. Available: <https://www.energy-alternatives.eu/2021/11/10/DPE-passoires.html>.
- [48] R. Quiblier and M. Grossouvre, *Rapport de stage de spécialité, imputation et prédiction des valeurs manquantes*. [Online]. Available: <https://seafire.urbs.fr/f/48b6664720fd4d5686d2/?d1=1>.
- [49] S. v. Buuren, *Flexible Imputation of Missing Data, Second Edition*, 2nd ed. New York: Chapman and Hall/CRC, Jul. 12, 2018, 444 pp., ISBN: 978-0-429-49225-9. DOI: [10.1201/9780429492259](https://doi.org/10.1201/9780429492259).
- [50] D. J. Stekhoven, *Using the missForest package*, Apr. 2022.
- [51] A. B. Hassanat, « Dimensionality invariant similarity measure », *arXiv:1409.0923 [cs]*, Sep. 2, 2014. arXiv: [1409.0923](https://arxiv.org/abs/1409.0923). Accessed: Oct. 9, 2020. [Online]. Available: <http://arxiv.org/abs/1409.0923>.
- [52] M. Grossouvre, D. Rullière, and J. Villot, « Predicting missing energy performance certificates: spatial interpolation of mixture distributions », *Energy and AI*, vol. 16, p. 100 339, May 1, 2024, ISSN: 2666-5468. DOI: [10.1016/j.egyai.2024.100339](https://doi.org/10.1016/j.egyai.2024.100339). Accessed: Jun. 21, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666546824000053>.
- [53] L. d. C. Godoy, M. O. Prates, and J. Yan, *An unified framework for point-level, areal, and mixed spatial data: the hausdorff-gaussian process*, Aug. 16, 2022. DOI: [10.48550/arXiv.2208.07900](https://doi.org/10.48550/arXiv.2208.07900). [Online]. Available: <http://arxiv.org/abs/2208.07900>.

- [54] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning* (Adaptive computation and machine learning). Cambridge, Mass: MIT Press, 2006, 248 pp., ISBN: 978-0-262-18253-9. [Online]. Available: <https://gaussianprocess.org/gpml/chapters/RW.pdf>.
- [55] M. Rocas, A. García-González, S. Zlotnik, X. Larráyoiz, and P. Díez, « Nonintrusive uncertainty quantification for automotive crash problems with VPS/pamcrash », *Finite Elements in Analysis and Design*, vol. 193, p. 103 556, Oct. 1, 2021, ISSN: 0168-874X. DOI: [10/gjkkgc](https://doi.org/10.1016/j.finel.2021.102472). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0168874X21000408>.
- [56] M. Gösgens, A. Zhiyanov, A. Tikhonov, and L. Prokhorenkova, « Good classification measures and how to find them », in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., vol. 34, New York: Curran Associates, Inc., 2021, pp. 17 136–17 147, ISBN: 978-1-71384-539-3. [Online]. Available: <https://proceedings.neurips.cc/paper/2021/file/8e489b4966fe8f703b5be647f1cbae63-Paper.pdf>.
- [57] D. Rullière and M. Grossouvre, *A joint kriging model with application to constrained classification*, Sep. 27, 2023. Accessed: Jan. 24, 2024. [Online]. Available: <https://hal.science/hal-04208454>.
- [58] A. de Castro Mazarro, S. K. Sikder, and A. A. Pedro, « Spatializing inequality across residential built-up types: a relational geography of urban density in são paulo, brazil. », *Habitat International*, vol. 119, p. 102 472, Jan. 1, 2022, ISSN: 0197-3975. DOI: [10.1016/j.habitatint.2021.102472](https://doi.org/10.1016/j.habitatint.2021.102472). Accessed: May 7, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0197397521001612>.
- [59] M. Mailagaha Kumbure, P. Luukka, and M. Collan, « A new fuzzy k-nearest neighbor classifier based on the bonferroni mean », *Pattern Recognition Letters*, vol. 140, pp. 172–178, Dec. 1, 2020, ISSN: 0167-8655. DOI: [10/gtn36j](https://doi.org/10.1016/j.patrec.2020.12.011). Accessed: Apr. 24, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167865520303792>.
- [60] N. Cressie and G. Johannesson, « Fixed rank kriging for very large spatial data sets », *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 70, no. 1, pp. 209–226, Jan. 2008, eprint: https://academic.oup.com/jrsssb/article-pdf/70/1/209/49796487/jrsssb_70_1_209.pdf, ISSN: 1369-7412. DOI: [10/fq9f68](https://doi.org/10.1093/bjstat/70.1.209).
- [61] A. Banerjee, D. B. Dunson, and S. T. Tokdar, « Efficient gaussian process regression for large datasets », *Biometrika*, vol. 100, no. 1, pp. 75–89, 2013, Publisher: Oxford University Press. DOI: [10/f4q3jt](https://doi.org/10.1093/bjstat/70.1.209).
- [62] D. Rullière, N. Durrande, F. Bachoc, and C. Chevalier, « Nested kriging predictions for datasets with a large number of observations », *Statistics and Computing*, vol. 28, pp. 849–867, 2018, Publisher: Springer. DOI: [10/gmt4z6](https://doi.org/10.1007/s11222-018-0968-1).
- [63] F. Bachoc, « Cross validation and maximum likelihood estimations of hyper-parameters of gaussian processes with model misspecification », *Computational Statistics & Data Analysis*, vol. 66, pp. 55–69, 2013, Publisher: Elsevier. DOI: [10/gmt4z5](https://doi.org/10.1016/j.csda.2013.06.001).
- [64] R. Furrer and M. G. Genton, « Aggregation-cokriging for highly multivariate spatial data », *Biometrika*, vol. 98, no. 3, pp. 615–631, 2011, Publisher: Oxford University Press. DOI: [10/djxb6d](https://doi.org/10.1093/bjstat/70.1.209).
- [65] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, « Fairness constraints: a flexible approach for fair classification », *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 2737–2778, 2019, Publisher: JMLR. org.
- [66] H. Wackernagel, *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media, 2003.

- [67] J.-P. Chiles and P. Delfiner, *Geostatistics: modeling spatial uncertainty*. John Wiley & Sons, 2012, vol. 713.
- [68] M. A. Alvarez, L. Rosasco, N. D. Lawrence, *et al.*, « Kernels for vector-valued functions: a review », *Foundations and Trends® in Machine Learning*, vol. 4, no. 3, pp. 195–266, 2012, Publisher: Now Publishers, Inc. DOI: [10/gmmw42](https://doi.org/10/gmmw42).
- [69] M. G. Genton and W. Kleiber, « Cross-covariance functions for multivariate geostatistics », *Statistical Science*, vol. 30, no. 2, May 1, 2015, ISSN: 0883-4237. DOI: [10/gh4j8k](https://doi.org/10/gh4j8k). arXiv: [1507.08017\[stat\]](https://arxiv.org/abs/1507.08017). Accessed: Dec. 1, 2022. [Online]. Available: <http://arxiv.org/abs/1507.08017>.
- [70] A. Leroy, P. Latouche, B. Guedj, and S. Gey, « MAGMA: inference and prediction using multi-task gaussian processes with common mean », *Machine Learning*, vol. 111, no. 5, pp. 1821–1849, 2022, Publisher: Springer. DOI: [10/gtn36k](https://doi.org/10/gtn36k).
- [71] A. Leroy, P. Latouche, B. Guedj, and S. Gey, « Cluster-specific predictions with multi-task gaussian processes », *Journal of Machine Learning Research*, vol. 24, no. 5, pp. 1–49, 2023.
- [72] P. Goovaerts, « Ordinary cokriging revisited », *Mathematical Geology*, vol. 30, pp. 21–42, 1998, Publisher: Springer.
- [73] J. M. Ver Hoef and N. Cressie, « Multivariable spatial prediction », *Mathematical Geology*, vol. 25, pp. 219–240, 1993, Publisher: Springer. DOI: [10/bxj7qj](https://doi.org/10/bxj7qj).
- [74] C. Rasmussen and Z. Ghahramani, « Occam’s razor », *Advances in neural information processing systems*, vol. 13, 2000.
- [75] A. Gordon, « A survey of constrained classification », *Computational Statistics & Data Analysis*, vol. 21, no. 1, pp. 17–29, 1996, Publisher: Elsevier. DOI: [10/d4zv5x](https://doi.org/10/d4zv5x).
- [76] P. S. Bradley, K. P. Bennett, and A. Demiriz, « Constrained k-means clustering », *Microsoft Research, Redmond*, vol. 20, 2000.
- [77] F. Höppner and F. Klawonn, « Clustering with size constraints », in *Computational Intelligence Paradigms: Innovative Applications*, L. C. Jain, M. Sato-Ilic, M. Virvou, G. A. Tsihrintzis, V. E. Balas, and C. Abeynayake, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 167–180, ISBN: 978-3-540-79474-5. DOI: [10.1007/978-3-540-79474-5_8](https://doi.org/10.1007/978-3-540-79474-5_8).
- [78] N. Ganganath, C.-T. Cheng, and C. K. Tse, « Data clustering with cluster size constraints using a modified k-means algorithm », in *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 2014, pp. 158–161. DOI: [10/gtn36d](https://doi.org/10/gtn36d).
- [79] K. A. Benatti, L. G. Pedroso, and A. A. Ribeiro, « Theoretical analysis of classic and capacity constrained fuzzy clustering », *Information Sciences*, vol. 616, pp. 127–140, 2022, ISSN: 0020-0255. DOI: [10/gtn36f](https://doi.org/10/gtn36f).
- [80] C. K. Williams and D. Barber, « Bayesian classification with gaussian processes », *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 12, pp. 1342–1351, 1998, Publisher: IEEE. DOI: [10/dt7m5q](https://doi.org/10/dt7m5q).
- [81] A. Dahl and E. V. Bonilla, « Grouped gaussian processes for solar power prediction », *Machine Learning*, vol. 108, no. 8, pp. 1287–1306, 2019, Publisher: Springer. DOI: [10/gjhjht](https://doi.org/10/gjhjht).
- [82] A. Panos, P. Dellaportas, and M. K. Titsias, « Large scale multi-label learning using gaussian processes », *Machine Learning*, vol. 110, pp. 965–987, 2021, Publisher: Springer. DOI: [10/gtn36h](https://doi.org/10/gtn36h).
- [83] A. G. Journel, « Nonparametric estimation of spatial distributions », *Journal of the International Association for Mathematical Geology*, vol. 15, pp. 445–468, 1983, Publisher: Springer. DOI: [10/dn3nf3](https://doi.org/10/dn3nf3).

- [84] F. V. D. Meer, « Classification of remotely-sensed imagery using an indicator kriging approach: application to the problem of calcite-dolomite mineral mapping », *International Journal of Remote Sensing*, vol. 17, no. 6, pp. 1233–1249, 1996, Publisher: Taylor & Francis. DOI: [10/ckz384](https://doi.org/10.1080/01448799608839384).
- [85] P. Goovaerts, « AUTO-IK: a 2d indicator kriging program for the automated non-parametric modeling of local uncertainty in earth sciences », *Computers & Geosciences*, vol. 35, no. 6, pp. 1255–1270, 2009, ISSN: 0098-3004. DOI: [10/cktgxr](https://doi.org/10.1016/j.cageo.2009.05.011).
- [86] J.-L. Chiang, J.-J. Liou, C. Wei, and K.-S. Cheng, « A feature-space indicator kriging approach for remote sensing image classification », *IEEE transactions on geoscience and remote sensing*, vol. 52, no. 7, pp. 4046–4055, 2013, Publisher: IEEE. DOI: [10/gdp8zr](https://doi.org/10.1109/TGRS.2013.2268878).
- [87] G. Agarwal, Y. Sun, and H. J. Wang, « Copula-based multiple indicator kriging for non-gaussian random fields », *Spatial Statistics*, vol. 44, p. 100 524, 2021, Publisher: Elsevier. DOI: [10/gtn36g](https://doi.org/10.1016/j.spatstat.2021.100524).
- [88] M. Grossouvre and D. Rullière, *Supplementary material to: a joint kriging model with application to constrained classification*, Publication Title: GitHub repository, 2023. [Online]. Available: <https://gitlab.emse.fr/marc.grossouvre/jointkrigingsupplementary/>.
- [89] N. Cressie, « Spatial prediction and ordinary kriging », *Mathematical geology*, vol. 20, pp. 405–421, 1988, Publisher: Springer. DOI: [10/d8hfzg](https://doi.org/10.1007/BF01531221).
- [90] L. Clarotto, D. Allard, and A. Menafoglio, « A new class of alpha-transformations for the spatial analysis of compositional data », *Spatial Statistics*, vol. 47, p. 100 570, 2022, Publisher: Elsevier. DOI: [10/gtn36m](https://doi.org/10.1016/j.spatstat.2022.100570).
- [91] J. Martínez-Minaya and H. Rue, « A flexible bayesian tool for CoDa mixed models: logistic-normal distribution with dirichlet covariance », *arXiv preprint arXiv:2308.13928*, 2023.
- [92] S. Vito, *Air quality*, Published: UCI Machine Learning Repository, 2016.
- [93] J. S. Simonoff, *Smoothing Methods in Statistics*. Springer-Verlag, 1996.
- [94] B. Bischl *et al.*, « OpenML benchmarking suites », in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: <https://openreview.net/forum?id=0CrD8ycKjG>.
- [95] M. Grossouvre, D. Rullière, and J. Villot, « Enhancing buildings’ energy efficiency prediction through advanced data fusion and fuzzy classification », *Energy and Buildings*, vol. 313, 2024, ISSN: 0378-7788. DOI: [10.1016/j.enbuild.2024.114243](https://doi.org/10.1016/j.enbuild.2024.114243).
- [96] C. Petersdorff, T. Boermans, J. Harnisch, O. Stobbe, S. Ullrich, and S. Wartmann, « Cost-effective climate protection in the EU building stock », ECOFYS, Cologne, Germany, Feb. 2005. [Online]. Available: https://www.eurima.org/uploads/files/modules/articles/1577099802_ecofysIII_report_EN.pdf.
- [97] P. Van de Maele, « French know-how in the field of energy efficiency in buildings; le savoir-faire français dans le domaine de l’efficacité énergétique des bâtiments », Sep. 15, 2010. Accessed: Jan. 22, 2024. [Online]. Available: <https://www.osti.gov/etdweb/biblio/22777397>.
- [98] I. Ballarini, S. P. Corgnati, and V. Corrado, « Use of reference buildings to assess the energy saving potentials of the residential building stock: the experience of TABULA project », *Energy Policy*, vol. 68, pp. 273–284, May 1, 2014, ISSN: 0301-4215. DOI: [10/gq64fv](https://doi.org/10.1016/j.enpol.2014.03.039). Accessed: Jan. 8, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0301421514000329>.

- [99] M. Storck, S. Slabik, A. Hafner, and R. Herz, « Towards assessing embodied emissions in existing buildings LCA—comparison of continuing use, energetic refurbishment versus demolition and new construction », *Sustainability*, vol. 15, no. 18, p. 13981, Jan. 2023, Number: 18 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2071-1050. DOI: [10/gtn37r](https://doi.org/10/gtn37r). Accessed: Jan. 8, 2024. [Online]. Available: <https://www.mdpi.com/2071-1050/15/18/13981>.
- [100] S. Hrabovszky-Horváth, T. Pálvölgyi, T. Csoknyai, and A. Talamon, « Generalized residential building typology for urban climate change mitigation and adaptation strategies: the case of hungary », *Energy and Buildings*, vol. 62, pp. 475–485, Jul. 1, 2013, ISSN: 0378-7788. DOI: [10/f4392t](https://doi.org/10/f4392t). Accessed: Jan. 8, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0378778813001795>.
- [101] F. Cappelletti, T. D. Mora, F. Peron, P. Romagnoni, and P. Ruggeri, « Building renovation: which kind of guidelines could be proposed for policy makers and professional owners? », *Energy Procedia*, 6th International Building Physics Conference, IBPC 2015, vol. 78, pp. 2366–2371, Nov. 1, 2015, ISSN: 1876-6102. DOI: [10/gtn376](https://doi.org/10/gtn376). Accessed: Jan. 8, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1876610215019219>.
- [102] T. Johansson, T. Olofsson, and M. Mangold, « Development of an energy atlas for renovation of the multifamily building stock in sweden », *Applied Energy*, vol. 203, pp. 723–736, Oct. 1, 2017, ISSN: 0306-2619. DOI: [10/gbn5s6](https://doi.org/10/gbn5s6). Accessed: Oct. 1, 2020. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0306261917307857>.
- [103] L. G. Swan and V. I. Ugursal, « Modeling of end-use energy consumption in the residential sector: a review of modeling techniques », *Renewable and Sustainable Energy Reviews*, vol. 13, no. 8, pp. 1819–1835, Oct. 1, 2009, ISSN: 1364-0321. DOI: [10/bkg3hx](https://doi.org/10/bkg3hx). Accessed: Oct. 1, 2020. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1364032108001949>.
- [104] P. Caputo and G. Pasetti, « Overcoming the inertia of building energy retrofit at municipal level: the italian challenge », *Sustainable Cities and Society*, vol. 15, pp. 120–134, Jul. 1, 2015, ISSN: 2210-6707. DOI: [10/ggxkmr](https://doi.org/10/ggxkmr). Accessed: Oct. 1, 2020. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S2210670715000025>.
- [105] C. F. Reinhart and C. Cerezo Davila, « Urban building energy modeling – a review of a nascent field », *Building and Environment*, vol. 97, pp. 196–202, Feb. 15, 2016, ISSN: 0360-1323. DOI: [10/f792m6](https://doi.org/10/f792m6). Accessed: Jan. 22, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0360132315003248>.
- [106] M. W. Ahmad, M. Mourshed, and Y. Rezgui, « Trees vs neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption », *Energy and Buildings*, vol. 147, pp. 77–89, Jul. 15, 2017, ISSN: 0378-7788. DOI: [10/gg2csg](https://doi.org/10/gg2csg). Accessed: Oct. 1, 2020. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378778816313937>.
- [107] A. Mastrucci, P. Pérez-López, E. Benetto, U. Leopold, and I. Blanc, « Global sensitivity analysis as a support for the generation of simplified building stock energy models », *Energy and Buildings*, vol. 149, pp. 368–383, Aug. 15, 2017, ISSN: 0378-7788. DOI: [10/gbsw2c](https://doi.org/10/gbsw2c). Accessed: Sep. 30, 2020. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378778816320023>.
- [108] N. Fumo, « A review on the basics of building energy estimation », *Renewable and Sustainable Energy Reviews*, vol. 31, pp. 53–60, Mar. 1, 2014, ISSN: 1364-0321. DOI: [10/ggq48x](https://doi.org/10/ggq48x). Accessed: Jan. 22, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1364032113007892>.

- [109] A. Fouquier, S. Robert, F. Suard, L. Stéphan, and A. Jay, « State of the art in building modelling and energy performances prediction: a review », *Renewable and Sustainable Energy Reviews*, vol. 23, pp. 272–288, Jul. 1, 2013, ISSN: 1364-0321. DOI: [10/f4z4jq](https://doi.org/10/f4z4jq). Accessed: Jan. 22, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1364032113001536>.
- [110] H. Wang and Z. (Zhai, « Advances in building simulation and computational techniques: a review between 1987 and 2014 », *Energy and Buildings*, vol. 128, pp. 319–335, Sep. 15, 2016, ISSN: 0378-7788. DOI: [10/f83c2q](https://doi.org/10/f83c2q). Accessed: Jan. 22, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0378778816305692>.
- [111] A. A. Al-Shargabi, A. Almhafdy, D. M. Ibrahim, M. Alghieth, and F. Chiclana, « Buildings’ energy consumption prediction models based on buildings’ characteristics: research trends, taxonomy, and performance measures », *Journal of Building Engineering*, vol. 54, p. 104577, Aug. 15, 2022, ISSN: 2352-7102. DOI: [10/gqqbjd](https://doi.org/10/gqqbjd). Accessed: Mar. 19, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352710222005903>.
- [112] D. Khafaga, E.-S. El-kenawy, A. Alhussan, and M. Eid, « Forecasting energy consumption using a novel hybrid dipper throated optimization and stochastic fractal search algorithm », *Intelligent Automation & Soft Computing*, vol. 37, no. 2, pp. 2117–2132, 2023, Publisher: Tech Science Press, ISSN: 1079-8587, 2326-005X. DOI: [10/gtc3tf](https://doi.org/10/gtc3tf). Accessed: Jan. 22, 2024. [Online]. Available: <https://www.techscience.com/iasc/v37n2/53228>.
- [113] Y. Zhang, B. K. Teoh, M. Wu, J. Chen, and L. Zhang, « Data-driven estimation of building energy consumption and GHG emissions using explainable artificial intelligence », *Energy*, vol. 262, p. 125468, Jan. 1, 2023, ISSN: 0360-5442. DOI: [10/grdfnv](https://doi.org/10/grdfnv). Accessed: Jan. 22, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0360544222023507>.
- [114] E. H. Ruspini, « A new approach to clustering », *Information and Control*, vol. 15, no. 1, pp. 22–32, Jul. 1, 1969, ISSN: 0019-9958. DOI: [10/bx88kc](https://doi.org/10/bx88kc). Accessed: Feb. 5, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0019995869905919>.
- [115] A. Amo, J. Montero, G. Biging, and V. Cutello, « Fuzzy classification systems », *European Journal of Operational Research*, vol. 156, no. 2, pp. 495–507, Jul. 16, 2004, ISSN: 0377-2217. DOI: [10/d27gkd](https://doi.org/10/d27gkd). Accessed: Feb. 5, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S037722170300002X>.
- [116] G. F. A. Yeo and V. Aksakalli, « A stochastic approximation approach to simultaneous feature weighting and selection for nearest neighbour learners », *Expert Systems with Applications*, vol. 185, p. 115671, Dec. 15, 2021, ISSN: 0957-4174. DOI: [10/gtn37q](https://doi.org/10/gtn37q). Accessed: May 13, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417421010605>.
- [117] D. Rullière and M. Grossouvre, « On multi-output kriging and constrained classification », in *Séminaire Statistique LMA*, Avignon (FR), France, Oct. 2023. Accessed: Mar. 18, 2024. [Online]. Available: <https://hal.science/hal-04227155>.
- [118] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, « VSURF: an r package for variable selection using random forests », *The R Journal*, vol. 7, no. 2, pp. 19–33, Dec. 2015. DOI: [10/gj3hq7](https://doi.org/10/gj3hq7). Accessed: Feb. 19, 2021. [Online]. Available: <https://hal.archives-ouvertes.fr/hal-01251924>.
- [119] W. J. K. Aldworth, « Spatial prediction, spatial sampling, and measurement error », Doctor of Philosophy, Iowa State University, 1998. [Online]. Available: <https://lib.dr.iastate.edu/rtd/11842/>.

Supplementary Material

Supplementary Material for Chapter I

1	About EPCs Uncertainties	181
2	Mathematical Presentation of Variables Normalisation	195
2.1	Cumulative Distribution, Quantile, Pseudo-Observations, Pseudo-Quantile, Definitions and Estimators.	195
2.2	Quantile and Pseudo-Quantile Estimators..	198
3	Normalised Pseudo-Observations	199
4	Distribution of Distances According to the Data Distribution	201
4.1	1-Dimensional problem.	201
4.2	2-Dimensional problem.	201
4.3	More About the Triangular Distribution.	204
5	Supplementary Material About the Hassanat Distance	209

1 About EPCs Uncertainties

This document contains the slides used for an oral presentation given in Institut Henri Fayol in march 2023.




vendredi 24 mars 2023,
Institut Henri Fayol

L'incertitude sur le DPE et sa prise en compte pour une prédiction à grande échelle

Marc Grossouvre¹ sous la direction de Didier Rullière²
et la supervision de Jonathan Villot³


¹Doctorant CIFRE, U.R.B.S. SAS, marcgrossouvre@urbs.fr
²Mines Saint-Etienne - LIMOS - Univ Clermont Auvergne
³Mines Saint-Etienne - U.R.B.S. SAS



PLAN

1. Que représentent les DPE fournis par l'ADEME ?
2. Une étiquette incertaine :
mesure humaine, effets de seuil, valeurs manquantes
3. Une localisation incertaine :
adresse, bâtiment et parcelle
4. Intégrer ces aléas à un modèle géostatistique

20/03/23 **L'incertitude sur les DPE** 2

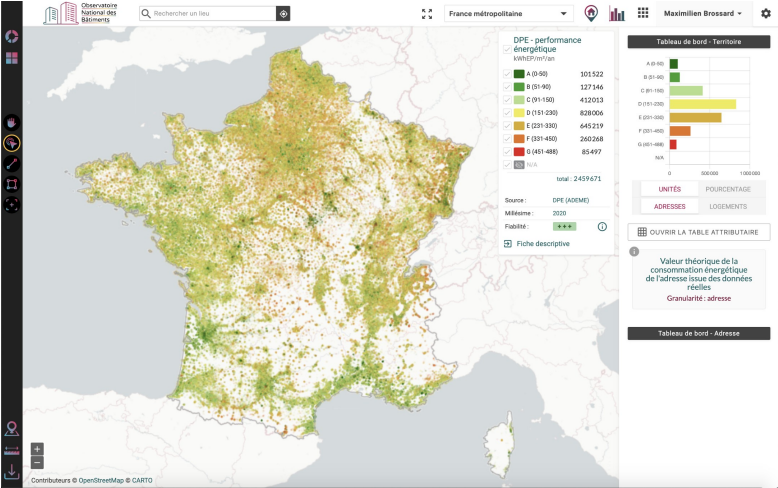


PLAN

1. Que représentent les DPE fournis par l'ADEME ?
2. Une étiquette incertaine :
mesure humaine, effets de seuil, valeurs manquantes
3. Une localisation incertaine :
adresse, bâtiment et parcelle
4. Intégrer ces aléas à un modèle géostatistique

20/03/23
L'incertitude sur les DPE
3

Environ 2 millions de DPE collectés par an



Classe	Nombre
A (0-50)	101 522
B (51-90)	127 346
C (91-150)	412 013
D (151-230)	828 006
E (231-330)	645 219
F (331-450)	262 268
G (451-488)	85 497
Tota	2 459 671

Logement économe

≤ 50 A

51 à 90 B

91 à 150 C

151 à 230 D

231 à 330 E

331 à 450 F

> 450 G

Logement énergivore

Logement

XXX

kWh/m².an

20/03/23
L'incertitude sur les DPE
4

1. Que représentent les DPE fournis par l'ADEME ?

Code de la construction et de l'habitation, article L126-226



RÉPUBLIQUE FRANÇAISE
*Liberté
Égalité
Fraternité*



Légifrance
Le service public de la diffusion du droit

Loi du 22 août 2021

Le diagnostic de performance énergétique **d'un bâtiment ou d'une partie de bâtiment** est un document qui comporte la quantité d'énergie effectivement consommée ou estimée, exprimée en énergie primaire et finale, ainsi que les émissions de gaz à effet de serre induites, pour une utilisation standardisée du bâtiment ou d'une partie de bâtiment et une classification en fonction de valeurs de référence permettant de comparer et évaluer sa performance énergétique et sa performance en matière d'émissions de gaz à effet de serre. Il comporte une information sur les conditions d'aération ou de ventilation. [...]
Sa durée de validité est fixée par voie réglementaire.

20/03/23
L'incertitude sur les DPE

5

...un bâtiment ou une partie de bâtiment...

- La méthode n'est pas la même pour un logement ou pour un immeuble.
- Les étiquettes dans un même immeuble peuvent varier.



1 rue d'Etrembières, Annemasse

65 apparts en étiquette C,
7 en étiquette D,
5 en étiquette E,
4 en étiquette F =
passoires énergétiques.



Appart 705 :
diagnostiqué C en 2018,
diagnostiqué E en 2020...

20/03/23
L'incertitude sur les DPE

6

1. Que représentent les DPE fournis par l'ADEME ?

Code de la construction et de l'habitation, article L126-226

Loi du 22 août 2021

Le diagnostic de performance énergétique d'un bâtiment ou d'une partie de bâtiment est un document qui comporte la quantité d'énergie effectivement consommée ou estimée, exprimée en énergie primaire et finale, ainsi que les émissions de gaz à effet de serre induites, **pour une utilisation standardisée** du bâtiment ou d'une partie de bâtiment et une classification en fonction de valeurs de référence permettant de comparer et évaluer sa performance énergétique et sa performance en matière d'émissions de gaz à effet de serre. Il comporte une information sur les conditions d'aération ou de ventilation. [...]

Sa durée de validité est fixée par voie réglementaire.

20/03/23

L'incertitude sur les DPE

7

...pour une utilisation standardisée...

- **Quels habitants respectent les standards ?**
 - Il y a ceux qui "surchauffent".
 - Il y a aussi ceux qui renoncent au chauffage.
 - Et ceux qui sont chauffés par les voisins...



1 rue d'Etrembières
à Annemasse

Consommation e.f. ENEDIS :
360 MWh/an

Surface habitable :
2 221m²

Consommation e.p. :
418kWh/m²/an => F

20/03/23

L'incertitude sur les DPE

8

1. Que représentent les DPE fournis par l'ADEME ?

Code de la construction et de l'habitation, article L126-226




Loi du 22 août 2021

Le diagnostic de performance énergétique d'un bâtiment ou d'une partie de bâtiment est un document qui comporte la quantité d'énergie effectivement consommée ou estimée, exprimée en énergie primaire et finale, ainsi que les émissions de gaz à effet de serre induites, pour une utilisation standardisée du bâtiment ou d'une partie de bâtiment et une classification en fonction de valeurs de référence permettant de comparer et évaluer sa performance énergétique et sa performance en matière d'émissions de gaz à effet de serre. Il comporte une information sur les conditions d'aération ou de ventilation. [...]

Sa durée de validité est fixée par voie réglementaire.

20/03/23
L'incertitude sur les DPE
9

... Sa durée de validité est fixée par voie réglementaire.

Et pendant ce temps

- La laine de verre se tasse,
- La maison est agrandie,
- L'appartement est rénové,
- La chaudière est changée,
- Le chauffe-eau a lâché.

Le DPE est toujours **"valable"**.

Un DPE (sauf exceptions ci-dessous) est valable 10 ans.


Exceptions :

DPE réalisés entre le 1er janvier 2013 et le 31 décembre 2017 inclus : valables jusqu'au 31 décembre 2022 ;

DPE réalisés entre le 1er janvier 2018 et le 30 juin 2021 inclus : valables jusqu'au 31 décembre 2024.

20/03/23
L'incertitude sur les DPE
10

1. Que représentent les DPE fournis par l'ADEME ?
Une valeur réglementaire qui informe sur une partie du parc bâti (~15% du parc)




La cohérence entre les valeurs au logement et les valeurs au bâtiment n'est pas assurée.

Le DPE ne donne pas de consommation réelle.

Le DPE représente l'état d'un logement ou d'un bâtiment à un moment du passé.

20/03/23 **L'incertitude sur les DPE** 11

PLAN



1. Que représentent les DPE fournis par l'ADEME ?

2. Une étiquette incertaine : mesure humaine, effets de seuil, valeurs manquantes

3. Une localisation incertaine : adresse, bâtiment et parcelle

4. Intégrer ces aléas à un modèle géostatistique

20/03/23 **L'incertitude sur les DPE** 12

2. Une étiquette incertaine : mesure humaine, effets de seuil, valeurs manquantes

Un seuil est une limite réglementaire du domaine associé à une étiquette DPE.
L'effet de seuil regroupe les phénomènes qui apparaissent à proximité de cette valeur.

Distribution réelle des DPE réalisés sur des logements

20/03/23
L'incertitude sur les DPE
13

Peut-on “lisser” les effets seuil ? Légimité et validation ?

Source : Alternatives énergétiques, Yassine Abdelouadoud

Que mesure-t-on ?

Année de construction : 1946 à 1974
Méthode : 3CL

Type logement : Maison
Combustible chauffage : Fioul

Correction DPE A et B : Oui

Distribution originale
Part de passoires : 38.1 %

Taille de l'échantillon : 161658 dpe

Année de construction : 1946 à 1974
Méthode : 3CL

Type logement : Maison
Combustible chauffage : Fioul

Correction DPE A et B : Oui

Après correction étape 0
Part de passoires : 47.2 %

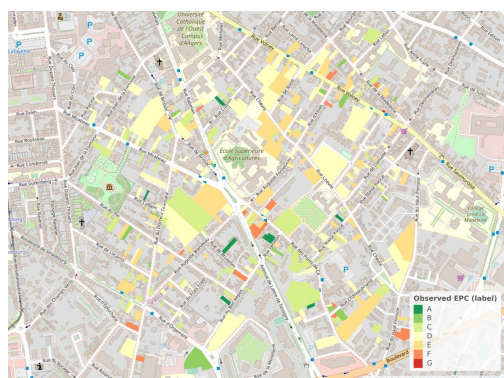
Erreur maximale : 1.1 %
Erreur maximale attendue : 7.9 %
Erreur moyenne : 0.6 %
Erreur moyenne attendue : 4.2 %

Taille de l'échantillon : 161658 dpe

20/03/23
L'incertitude sur les DPE
14

Beaucoup de valeurs manquantes.

Un quartier d'Angers



- Environ 15% des adresses ont eu au moins un logement diagnostiqué.
- Que peut-on dire des adresses qui n'ont pas été observées ?
- Peut-on détecter les passoires thermiques ?

20/03/23

L'incertitude sur les DPE

15

PLAN



1. Que représentent les DPE fournis par l'ADEME ?

2. Une étiquette incertaine :
mesure humaine, effets de seuil, valeurs manquantes

3. Une localisation incertaine :
adresse, bâtiment et parcelle

4. Intégrer ces aléas à un modèle géostatistique

20/03/23

L'incertitude sur les DPE

16

3. Une localisation incertaine : adresse, bâtiment et parcelle

nom_rue
28 & 30 Rue Gabriel Vicaire \nRue Prosper Convert

Comprendre : 28 et 30 rue Gabriel Vicaire,
à l'angle de la rue Prosper Convert

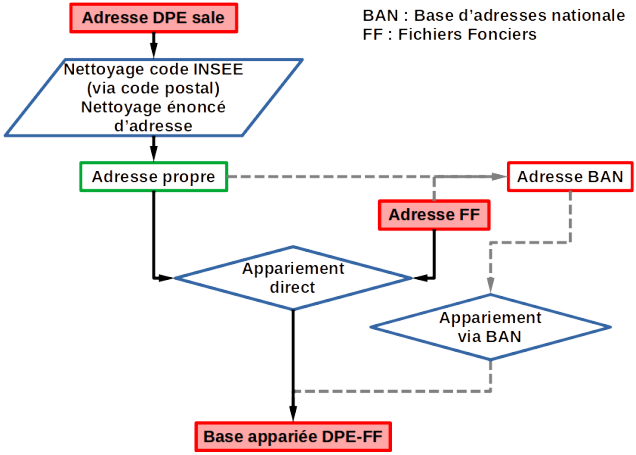


17

Il faut appairer les DPE aux fichiers fonciers pour les localiser

- Les fichiers fonciers permettent de relier le DPE observé à une parcelle.

Le DPE renseigne sur l'un des logements de l'un des bâtiments de la parcelle.

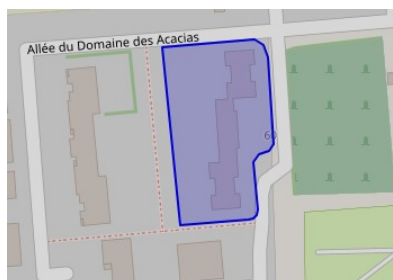


18

Une fois la parcelle associée à l'adresse identifiée, difficulté de géolocalisation sur la parcelle

14, 22, 30, 40, 52 impasse des Acacias
à Ars-sur-Formans (Ain).

- 5 maisons en bande,
- 1 seule parcelle,
- Observations : 3 étiquettes E, 2 étiquettes D.
- On ne sais pas où sont les logements observés.

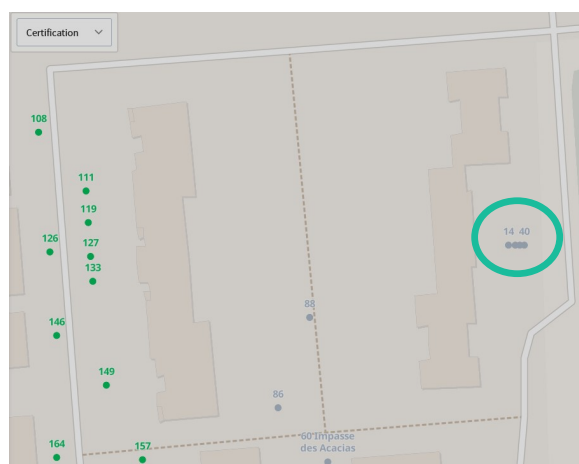


20/03/23

L'incertitude sur les DPE

19

Qu'en dit la Base d'Adresses Nationale (BAN) ?



20/03/23

L'incertitude sur les DPE

20

PLAN

1. Que représentent les DPE fournis par l'ADEME ?

**2. Une étiquette incertaine :
mesure humaine, effets de seuil, valeurs manquantes**

**3. Une localisation incertaine :
adresse, bâtiment et parcelle**

4. Intégrer ces aléas à un modèle géostatistique

20/03/23 **L'incertitude sur les DPE** 21

4. Intégrer ces aléas à un modèle géostatistique

Supposons que l'on veuille déterminer le DPE en fonction de la position géographique et de l'année de construction.

Nous devons gérer une incertitude à la fois sur les données d'entrée (position incertaine) et sur les données de sortie (incertitude sur l'observation).

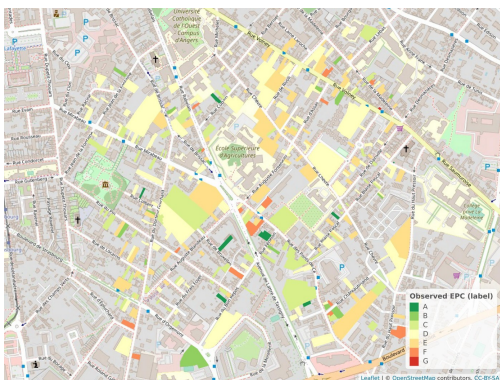
20/03/23 **L'incertitude sur les DPE** 22

L'ADEME observe une distribution de mélange des DPE

- Chaque logement a un DPE aléatoire : probablement C mais peut-être B ou D selon les travaux réalisés, le diagnostiqueur, le seuil...
 - Chaque DPE de l'ADEME est associé à un logement aléatoire parmi les logements de la parcelle.
 - Chaque DPE de l'ADEME est donc une valeur aléatoire associée à une position aléatoire.
- **Que peut-on faire de ça ?**

Spatial interpolation using mixture distributions: A Best Linear Unbiased Predictor

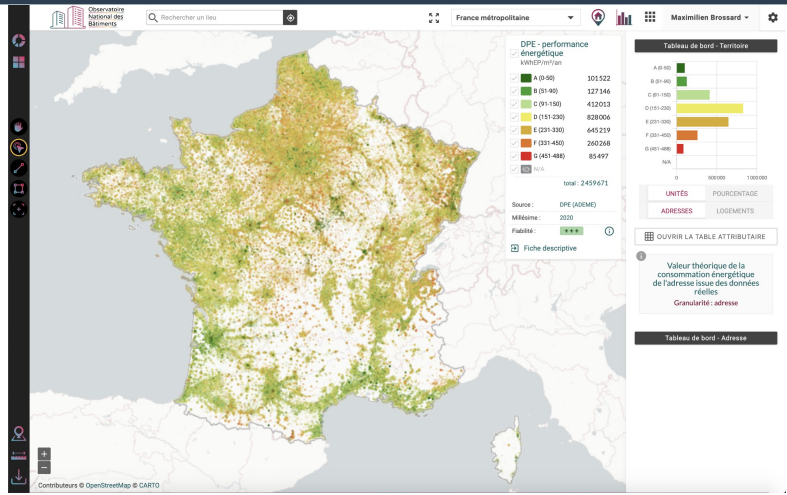
Un quartier de la ville d'Angers



Résultat des prédictions

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

Merci de votre attention. Vous avez des questions ?



20/03/23 **L'incertitude sur les DPE** 25

2 Mathematical Presentation of Variables Normalisation

In order to address the issue of choosing a distance for a dataset containing heterogeneous variables in order to feed a **KNN** model, we study the impact of pseudo-observations and pseudo-quantiles on predictions.

2.1 Cumulative Distribution, Quantile, Pseudo-Observations, Pseudo-Quantile, Definitions and Estimators

Definition 4. Given a real number a , we call **indicator function** of $]-\infty, a]$ the function:

$$\mathbb{1}_{]-\infty, a]} : \mathbb{R} \rightarrow \{0, 1\}$$

$$x \mapsto \mathbb{1}_{]-\infty, a]}x = \mathbb{1}_{x \leq a} = \begin{cases} 1 & \text{if } x \leq a \\ 0 & \text{otherwise} \end{cases}$$

Similarly, for a given interval $[a, b]$, we define $\mathbb{1}_{[a, b]}(x) = \mathbb{1}_{x \in [a, b]}$ and so on.

2.1.a Samples and Weighted Samples

Definition 5. Let X be a real random variable.

The **cumulative distribution function** of X is :

$$F_X : \mathbb{R} \rightarrow [0, 1]$$

$$t \mapsto F_X(t) = P(X \leq t)$$

Proposition 10. F_X is non-decreasing and right continuous on \mathbb{R} . Moreover, if X is continuous, $F_X(X)$ is a uniform random variable on $[0, 1]$.

Proof. See proof https://people.math.ethz.ch/~embrecht/ftp/generalized_inverse.pdf. □

Definition 6. Let X be a real random variable.

The **quantile function** of X is:

$$F_X^{-1} : [0, 1] \rightarrow \mathbb{R}$$

$$u \mapsto F_X^{-1}(u) = \inf\{t \in \mathbb{R} : F_X(t) \geq u\}$$

Proposition 11. The quantile function is an almost sure inverse of the cumulative distribution function:

$$F_X^{-1}(F_X(t)) = t \text{ almost everywhere.}$$

Definition 7. A set $\underline{X}_n = (X_1, \dots, X_n)$ of n pairwise independent, identically distributed random variables with same cumulative distribution as X is called a **sample** of X of size n , or an n -sample of X .

Definition 8. Given an n -sample $\underline{X}_n = (X_1, \dots, X_n)$ of X , we call **i -th order statistic** and denote $X_{(i)}$ the variable yielding the i -th value of the sample when reordered in increasing order: $X_{(1)} \leq \dots \leq X_{(n)}$.

Definition 9. Given an n -sample $\underline{X}_n = (X_1, \dots, X_n)$ of X , we associate with \underline{X}_n a **rank function**:

$$\begin{aligned} \text{rk}_n &: \underline{X}_n \rightarrow \{1, \dots, n\} \\ X_i &\mapsto \text{rk}_n(X_i) = \sum_{k=1}^n \mathbb{1}_{X_k \leq X_i} \end{aligned}$$

And we denote $\text{rk}(\underline{X}_n) = (\text{rk}(X_1), \dots, \text{rk}(X_n))$.

Let us now assume that X is continuous but instead of observing values of X , we observe classes given by intervals.

Definition 10. Let X be a real continuous random variable.

Let $A_n = (a_i)_{i \in \{1, \dots, m\}}$ be a set of real numbers such that $a_1 < a_2 < \dots < a_m$.

We define a sequence of intervals $I = (I_i)_{i \in \{0, \dots, m\}}$ such that:

$$\begin{aligned} I_0 &=]-\infty, a_1] \\ \forall i \in \{1, \dots, m-1\}, I_i &=]a_i, a_{i+1}] \\ I_m &=]a_m, +\infty[\end{aligned}$$

For a given set of values (x_0, \dots, x_m) such that $\forall i \in \{0, \dots, m\}, x_i \in I_i$, we call **discretised bounded form of X** a random variable X_I :

$$X_I = \sum_{i=0}^m x_i \mathbb{1}_{X \in I_i}$$

We can summarise observations in a table by associating with each unique value in (x_0, \dots, x_m) the number of times it occurs in the sample, say $\underline{X}'_{I,n} = ((x_0, n_0), \dots, (x_m, n_m))$ where $\sum_{i=0}^m n_i = n$. We call it a **weighted sample** of X .

Remark 7. By construction, $\underline{X}_{I,n}$ is a sample of X_I .

Remark 8. The main interest of introducing weighted samples is to be able to generalise to counts that are not integer. For instance, census data is published with weighted individuals where weight is a decimal number. Considering that the population is the sum of weights, all indicators that are defined below (weighted pseudo-observations in particular) can be computed although it is impossible to rebuild a theoretical original sample.

Remark 9.

$$\forall i \in \{0, \dots, m\}, \lim_{n \rightarrow +\infty} \frac{n_i}{n} = P(X \in I_i)$$

2.1.b Estimators for Cumulative Distribution and Pseudo-Observation

Proposition 12. *Given an n -sample $\underline{X}_n = (X_1, \dots, X_n)$ of X , an unbiased estimator of F_X is given by the **empirical distribution function**:*

$$\begin{aligned} \hat{F}_n &: \mathbb{R} \rightarrow [0, 1] \\ t \mapsto \hat{F}_n(t) &= \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{X_k \leq t} \end{aligned}$$

In particular $\forall i \in \{1, \dots, n\}$, $\hat{F}_n(X_i) = \frac{\text{rk}_n(X_i)}{n}$.

However this estimator's tail behaviour is not symmetric in the sense that:

$$\begin{aligned} \min_{i \in \{1, \dots, n\}} (\hat{F}_n(X_i)) &= \hat{F}_n \left(\min_{i \in \{1, \dots, n\}} (X_i) \right) = \frac{1}{n} \\ \text{while } \max_{i \in \{1, \dots, n\}} (\hat{F}_n(X_i)) &= \hat{F}_n \left(\max_{i \in \{1, \dots, n\}} (X_i) \right) = 1 \end{aligned}$$

This is also awkward because $(F_X(X_1), \dots, F_X(X_n))$ is a sample of a uniformly distributed random variable on $[0, 1]$ and we have the following result:

Proposition 13. *Let U be a uniformly distributed random variable on $[0, 1]$. Let $\underline{U}_n = (U_1, \dots, U_n)$ be a n -sample of U .*

$$\begin{aligned} \mathbb{E} \left[\max_{i \in \{1, \dots, n\}} (U_i) \right] &= \frac{n}{n+1} \\ \mathbb{E} \left[\min_{i \in \{1, \dots, n\}} (U_i) \right] &= \frac{1}{n+1} \end{aligned}$$

Therefore, it makes sense to rescale the estimator and define:

Definition 11. *Given an n -sample $\underline{X}_n = (X_1, \dots, X_n)$ of X , an asymptotically unbiased estimator of F_X is given by the **pseudo-observation function**:*

$$\begin{aligned} \hat{G}_n &: \mathbb{R} \rightarrow [0, 1] \\ t \mapsto \hat{G}_n(t) &= \frac{1}{n+1} \sum_{k=1}^n \mathbf{1}_{X_k \leq t} \end{aligned}$$

In particular $\forall i \in \{1, \dots, n\}$, $\hat{G}_n(X_i) = \frac{\text{rk}_n(X_i)}{n+1}$.

The images of the sample $(\hat{G}_n(X_1), \dots, \hat{G}_n(X_n))$ are called **pseudo-observations** of the sample \underline{X}_n .

Remark 10. *Spearman's covariance can be computed with ranks of pseudo-observations equivalently.*

Definition 12. Given a weighted sample $\underline{X}'_{I,n}$ of a continuous random variable X , we call **weighted empirical distribution function**:

$$\begin{aligned} \hat{F}'_n &: \mathbb{R} \rightarrow [0, 1] \\ t \mapsto \hat{F}'_n(t) &= \frac{1}{n} \sum_{i=1}^m n_i \mathbf{1}_{x_i \leq t} \end{aligned}$$

Remark 11. Note that in the above definition, n_i is a random number and x_i is a deterministic value.

We did not have enough time to prove it but it seems that the weighted empirical distribution function asymptotically converges to the distribution function if the supports of X is finite, m, n goes to ∞ , $a_1 < \min(X)$, $a_m > \max(X)$, $a_{i+1} - a_i$ goes to 0.

Definition 13. Given a weighted sample $\underline{X}'_{I,n}$ of a continuous random variable X , we call **weighted pseudo-observation function**:

$$\begin{aligned} \hat{G}'_n &: \mathbb{R} \rightarrow [0, 1] \\ t \mapsto \hat{G}'_n(t) &= \frac{1}{n+1} \sum_{i=0}^m n_i \mathbf{1}_{x_i \leq t} \end{aligned}$$

In particular, $\hat{G}'_n(x_{(k)}) = \frac{1}{n+1} \sum_{i=0}^k n_i$ are the **weighted pseudo-observations** of the weighted sample $\underline{X}'_{I,n}$.

2.2 Quantile and Pseudo-Quantile Estimators.

There are multiple quantile estimators among which is the generalised inverse of empirical distribution function:

Definition 14. Given a positive real number x , we call **ceiling function**:

$$\begin{aligned} \lceil \cdot \rceil &: \mathbb{R}^+ \rightarrow \mathbb{N} \\ x \mapsto \lceil x \rceil &= \min \{k \in \mathbb{N} : k \geq x\} \end{aligned}$$

2.2.a Estimator on Samples

Definition 15. Given an n -sample $\underline{X}_n = (X_1, \dots, X_n)$ of X we call **empirical quantile function**:

$$\begin{aligned} \hat{F}_n^{-1} &: [0, 1] \rightarrow \mathbb{R} \\ u \mapsto \hat{F}_n^{-1}(u) &= X_{(\lceil nu \rceil)} \end{aligned}$$

It means that the estimator is a step function left continuous

Proposition 14. Given an n -sample $\underline{X}_n = (X_1, \dots, X_n)$ of X , we have:

$$\forall i \in \{1, \dots, n\}, \hat{F}_n^{-1}(\hat{F}_n(X_i)) = X_i$$

Definition 16. We can define a **pseudo-quantile function**:

$$\hat{G}_n^{-1} : [0, 1] \rightarrow \mathbb{R}$$

$$u \mapsto \hat{G}_n^{-1}(u) = \begin{cases} X_{(\lceil (n+1)u \rceil)} & \text{if } u \leq \frac{n}{n+1} \\ X_{(n)} & \text{if } u > \frac{n}{n+1} \end{cases}$$

Proposition 15. Given an n -sample $\underline{X}_n = (X_1, \dots, X_n)$ of X , we have:

$$\forall i \in \{1, \dots, n\}, \hat{G}_n^{-1}(\hat{G}_n(X_i)) = X_i$$

2.2.b Estimator on Weighted Samples

Definition 17. Given a weighted sample $\underline{X}'_{I,n}$ of a continuous random variable X , we call **weighted quantile function**:

$$\hat{F}'_n^{-1} : [0, 1] \rightarrow \{x_0, \dots, x_m\}$$

$$u \mapsto \hat{F}'_n^{-1}(u) = x_{\min \left\{ k: \sum_{i=1}^k n_i \geq nu \right\}}$$

Proposition 16. $\forall i \in \{1, \dots, m\}, \hat{F}'_n^{-1}(\hat{F}'_n(x_i)) = x_i$

Definition 18. Given a weighted sample $\underline{X}'_{I,n}$ of a continuous random variable X , we call **weighted pseudo-quantile function**:

$$\hat{G}'_n^{-1} : [0, 1] \rightarrow \{x_0, \dots, x_m\}$$

$$u \mapsto \hat{G}'_n^{-1}(u) = \begin{cases} x_{\min \left\{ k: \sum_{i=1}^k n_i \geq (n+1)u \right\}} & \text{if } u \leq \frac{n}{n+1} \\ x_m & \text{if } u > \frac{n}{n+1} \end{cases}$$

Proposition 17. $\forall i \in \{1, \dots, m\}, \hat{G}'_n^{-1}(\hat{G}'_n(x_i)) = x_i$

3 Normalised Pseudo-Observations

Proposition 18. Such as defined above, the pseudo-observations, whether they be from samples or weighted samples are in $]0, 1[$.

We denote Φ^{-1} the quantile function of the standard normal distribution $\mathcal{N}(0, 1)$.

Definition 19. Given a n -sample $\underline{X}_n = (X_1, \dots, X_n)$ of a random variable X , we call **normalised pseudo-observation function** of \underline{X}_n the function $\Phi^{-1} \circ \hat{G}_n$. Given a weighted n -sample $\underline{X}'_{I,n} = ((x_0, n_0), \dots, (x_m, n_m))$ of a random variable X , we call **normalised weighted pseudo-observations** of \underline{X}_n the function $\Phi^{-1} \circ \hat{G}'_n$.

In particular, $(\Phi^{-1}(\hat{G}_n(X_1)), \dots, \Phi^{-1}(\hat{G}_n(X_n)))$ are the **normalised pseudo-observations**.

Proposition 19. *For a given sample $\underline{X}_n = (X_1, \dots, X_n)$ of a random variable X , we denote $\hat{F}_{X,n}$ its pseudo-observation function. Given a strictly increasing function f on \mathbb{R} , we denote $\underline{Y}_n = (f(X_1), \dots, f(X_n))$. Then \underline{Y}_n is a n -sample of $Y = f(X)$. We denote $\hat{F}_{Y,n}$ its pseudo-observation function. Then we have:*

$$\forall t \in \mathbb{R}, \hat{F}_{X,n}(t) = \hat{F}_{Y,n}(f(t))$$

We say that pseudo-observations are invariant by increasing transformation. The same holds for normalised pseudo-observations.

We have the same result for weighted pseudo-observations and normalised weighted pseudo-observations.

4 Distribution of Distances According to the Data Distribution

In Supplementary Material 2, page 195, we define a way to transform a real random variable into either a uniform or a normal distribution. We are now interested in understanding the effect of these transformations on the distances between individuals.

4.1 1-Dimensional problem

We assume that 2 random variables X_1 and X_2 , independent and identically distributed, follow a standard normal distribution $\mathcal{N}(\mu = 0, \sigma^2 = 1)$ then one can prove that:

Variable	Distribution	Expectation	Median	Mode	Variance
X_1, X_2	$\mathcal{N}(0, 1)$	0	0	0	1
$X_1 - X_2$	$\mathcal{N}(0, 2)$	0	0	0	2
$\frac{(X_1 - X_2)^2}{2}$	$\chi_1^2 (1)$	1	$\approx (\frac{7}{9})^3 \approx 0.471$	0	2
$ X_1 - X_2 $	$\mathcal{H}(\sigma^2 = 2) (2)$	$\frac{2}{\sqrt{\pi}} \approx 1.128$	$2\text{erf}^{-1}(1/2) \approx 1.041$	0	$2(1 - \frac{2}{\pi}) \approx 0.727$

(1) Chi-square distribution with 1 degree of freedom.

(2) Half normal distribution of variance 2.

Now, if we assume that Y_1 and Y_2 , independent and identically distributed, follow a uniform distribution $\mathcal{U}(0, 1)$

Variable	Distribution	Expectation	Median	Mode	Variance
Y_1, Y_2	$\mathcal{U}(0, 1)$	$\frac{1}{2}$	0.5	[0, 1]	$\frac{1}{12} \approx 0.083$
$Y_1 - Y_2$	$\mathcal{T}(a = -1, b = 1, c = 0) (3)$	0	0	0	$\frac{1}{6} \approx 0.167$
$(Y_1 - Y_2)^2$	$f(t) = \frac{1}{\sqrt{t}} - 1$	$\frac{1}{6}$	$\frac{3}{2} - \sqrt{2} \approx 0.086$	0	$\frac{7}{180} \approx 0.039$
$ Y_1 - Y_2 $	$\mathcal{T}(a = 0, b = 1, c = 0)$	$\frac{1}{3} \approx 0.333$	$\frac{\sqrt{2}}{2} \approx 0.707$	0	$\frac{1}{18} \approx 0.056$

(3) Triangular distribution.

For X , the distance between values located in an area of high density is much smaller than between values located in area of small density. For Y the order of magnitude of distances are the same in both cases. We expect the introduction of pseudo-observations to be more inclusive of large values probably leading to a higher variability of predictions.

4.2 2-Dimensional problem

So as to have a qualitative approach of the problem, we visualise the clouds of nearest neighbours computed with raw observations values vs pseudo-observations values, for observations of a 2-dimensional Gaussian vector.

- First, we plot the clouds using raw values as coordinates (see Figure 1): When computing Euclidean distances with raw observations, neighbours fill circles, and squares when computing Manhattan distances, as expected. But when we use pseudo observations, we observe a distortion of the clouds in a radial direction from the plot origin, that is to say parallel with the density gradient. The shapes vary from distorted circles/squares (areas of high density) to spray shape as soon as we reach an area of higher density gradient. In the latter case, the cloud is

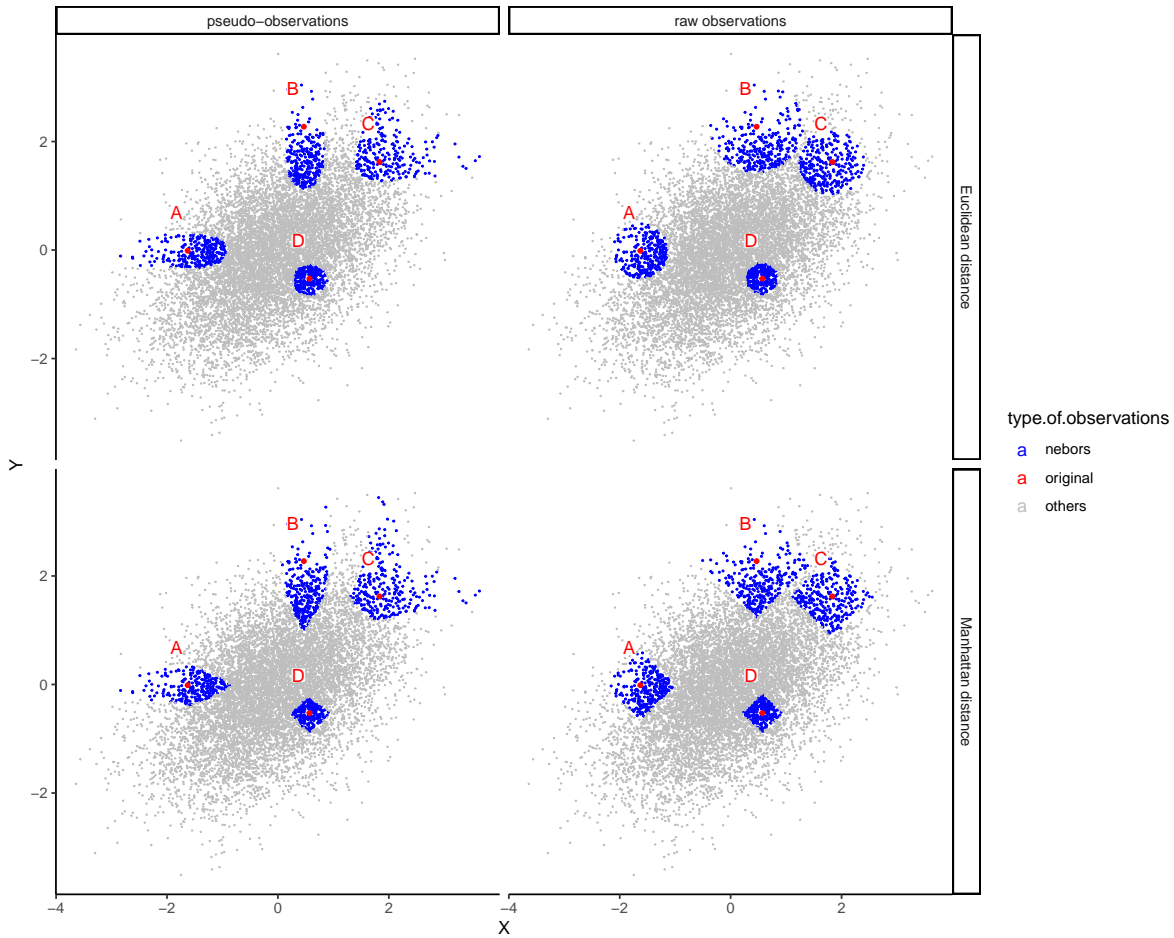


Figure 1 – Effect of pseudo observations as compared to raw observations on the shape of the nearest neighbours cloud according to the distance (scale is raw values). We generate a sample of 10,000 points following a centred Gaussian vector of dimension 2 with covariance matrix $\begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}$. We pick 4 observations in this set (red colour) and visualise their 300 nearest neighbours (blue colour) with Manhattan or Euclidean distance and computing distances on raw observations or on pseudo-observations.

going a little deeper towards areas of high density and further towards areas of high density.

- Now, we plot the clouds using pseudo-observations values as coordinates (see Figure 2): As expected, we have circular and square shapes when plotting the clouds with distances computed on pseudo-observations. When computing distances with raw observations, we observe a directional distortion effect. This is especially true when we look at neighbours of observations that have coordinates of different order of magnitude and in general for off-centred observations. The clouds are spread along the global observations cloud border.

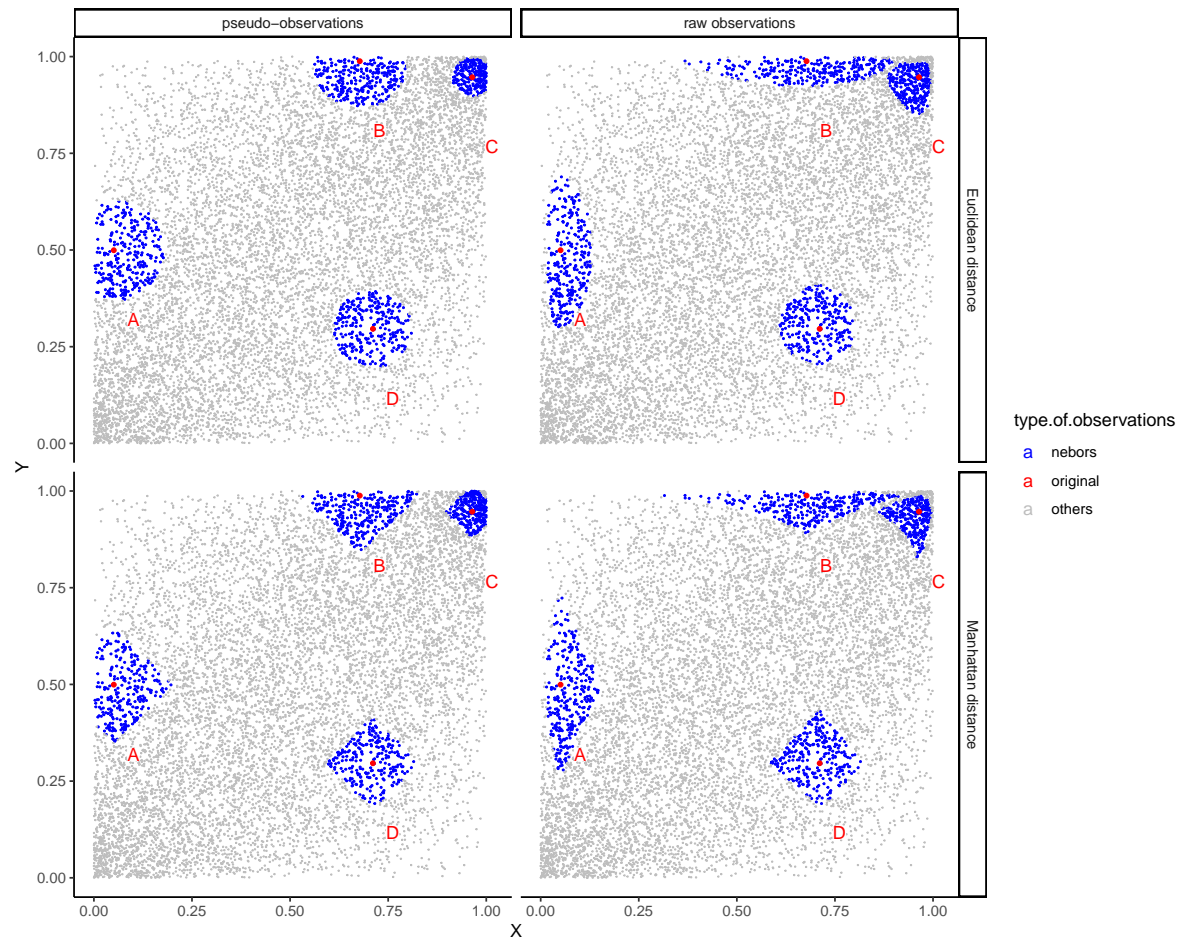


Figure 2 – Effect of pseudo observations as compared to raw observations on the shape of the nearest neighbours cloud according to the distance (scale is pseudo-observations values). The blue clouds include the same individuals as in Figure 1, but the referential is different..

4.3 More About the Triangular Distribution

For 3 real numbers a, b, c such that $a \leq b \leq c$, let X be a random variable of density:

$$\begin{cases} f_X(x) = 0 & \text{for } x < a \\ f_X(x) = \frac{2(x-a)}{(b-a)(c-a)} & \text{for } a \leq x \leq c \\ f_X(x) = \frac{2(b-x)}{(b-a)(b-c)} & \text{for } c < x \leq b \\ f_X(x) = 0 & \text{for } b < x \end{cases}$$

$$\begin{cases} F_X(x) = 0 & \text{for } x < a \\ F_X(x) = \frac{(x-a)^2}{(b-a)(c-a)} & \text{for } a \leq x \leq c \\ F_X(x) = 1 - \frac{(b-x)^2}{(b-a)(b-c)} & \text{for } c < x \leq b \\ F_X(x) = 0 & \text{for } b < x \end{cases}$$

We say that X follows a triangular distribution $\mathcal{T}(a, b, c)$.

Its density is continuous piecewise linear and its mode is obviously c .

We now consider $Y = X - c$ and denote: $\alpha = a - c$, $\beta = b - c$, we have:

$$\begin{cases} f_Y(y) = 0 & \text{for } y < \alpha \\ f_Y(y) = \frac{2(y-\alpha)}{(\beta-\alpha)(-\alpha)} & \text{for } \alpha \leq y \leq 0 \\ f_Y(y) = \frac{2(\beta-y)}{(\beta-\alpha)(\beta)} & \text{for } 0 < y \leq \beta \\ f_Y(y) = 0 & \text{for } \beta < y \end{cases}$$

$$\begin{cases} F_Y(y) = 0 & \text{for } y < \alpha \\ F_Y(y) = \frac{(y-\alpha)^2}{(\beta-\alpha)(-\alpha)} & \text{for } \alpha \leq y \leq 0 \\ F_Y(y) = 1 - \frac{(\beta-y)^2}{(\beta-\alpha)(\beta)} & \text{for } 0 < y \leq \beta \\ F_Y(y) = 1 & \text{for } \beta < y \end{cases}$$

We can compute the moment of Y of order k :

$$\begin{aligned} \mathbb{E}[Y^k] &= \int_{\alpha}^0 y^k \frac{(y-\alpha)}{(\beta-\alpha)(-\alpha)} dy + \int_0^{\beta} y^k \frac{2(\beta-y)}{(\beta-\alpha)(\beta)} dy \\ &= \frac{2}{(\beta-\alpha)(-\alpha)} \left[\frac{y^{k+2}}{k+2} - \alpha \frac{y^{k+1}}{k+1} \right]_{\alpha}^0 + \frac{2}{(\beta-\alpha)(\beta)} \left[\beta \frac{y^{k+1}}{k+1} - \frac{y^{k+2}}{k+2} \right]_0^{\beta} \\ &= \frac{2}{(\beta-\alpha)(-\alpha)} \left(-\frac{\alpha^{k+2}}{k+2} + \frac{\alpha^{k+2}}{k+1} \right) + \frac{2}{(\beta-\alpha)(\beta)} \left(\frac{\beta^{k+2}}{k+1} - \frac{\beta^{k+2}}{k+2} \right) \\ \mathbb{E}[Y^k] &= \frac{2(\beta^{k+1} - \alpha^{k+1})}{(\beta-\alpha)(k+1)(k+2)} \end{aligned}$$

It follows that:

$$\begin{aligned}
 k = 1 : \quad \mathbb{E}[Y] &= \frac{\beta + \alpha}{3} \\
 \mathbb{E}[X] &= \frac{a + b + c}{3} \\
 k = 2 : \quad \mathbb{E}[Y^2] &= \frac{\beta^2 + \beta\alpha + \alpha^2}{6} \\
 \text{Var}[Y] &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{\beta^2 + \beta\alpha + \alpha^2}{6} - \left(\frac{\beta + \alpha}{3}\right)^2 \\
 \text{Var}[Y] &= \frac{3(\beta^2 + \beta\alpha + \alpha^2) - 2(\beta^2 + 2\beta\alpha + \alpha^2)}{18} \\
 \text{Var}[Y] &= \frac{\beta^2 - \beta\alpha + \alpha^2}{18} \\
 \text{Var}[X] = \text{Var}[Y] &= \frac{a^2 + b^2 + c^2 - (ab - bc - ca)}{18}
 \end{aligned}$$

4.3.a Symmetric Centred Case $a = -b$, $c = 0$

$$\begin{cases}
 f_X(x) = 0 & \text{for } x < -b \\
 f_X(x) = \frac{x+b}{b^2} & \text{for } -b \leq x \leq 0 \\
 f_X(x) = \frac{b-x}{b^2} & \text{for } 0 < x \leq b \\
 f_X(x) = 0 & \text{for } b < x
 \end{cases}$$

$$\begin{cases}
 F_X(x) = 0 & \text{for } x < -b \\
 F_X(x) = \frac{(x+b)^2}{2b^2} & \text{for } -b \leq x \leq 0 \\
 F_X(x) = 1 - \frac{(b-x)^2}{2b^2} & \text{for } 0 < x \leq b \\
 F_X(x) = 1 & \text{for } b < x
 \end{cases}$$

$$\mathbb{E}[X] = 0$$

$$\mathbb{E}[X^2] = \text{Var}[X] = \frac{b^2}{6}$$

4.3.b Squared Value in the Symmetric Centred Case $a = -b$, $c = 0$

Let $X \sim \mathcal{T}(a = -b, b = b, c = 0)$

Let $Y = X^2$, note that $0 \leq Y \leq b^2$

$$\begin{aligned}
 P(Y \leq y) &= 2 \int_0^{\sqrt{y}} \frac{b-x}{b^2} dx \\
 &= 2 \left[-\frac{(b-x)^2}{2b^2} \right]_0^{\sqrt{y}} \\
 &= 2 \left(\frac{1}{2} - \frac{(b-\sqrt{y})^2}{2b^2} \right) \\
 F_Y(y) &= 1 - \frac{(b-\sqrt{y})^2}{b^2}
 \end{aligned}$$

Check: $F_Y(0) = 0, F_Y(b^2) = 1$

$$\begin{aligned}
 \text{Med}[Y] &= F_Y^{-1}\left(\frac{1}{2}\right) \\
 \frac{1}{2} &= 1 - \frac{(b-\sqrt{y})^2}{b^2} \\
 (b-\sqrt{y})^2 &= \frac{b^2}{2} \text{ (squared terms or positive)} \\
 b-\sqrt{y} &= \frac{b}{\sqrt{2}} \\
 y &= \left(1 - \frac{\sqrt{2}}{2}\right)^2 \\
 y &= \frac{3}{2} - \sqrt{2}
 \end{aligned}$$

Therefore:

$$\begin{aligned}
 f_Y(y) &= -\frac{2}{b^2} (b-\sqrt{y}) \left(-\frac{1}{2\sqrt{y}}\right) \\
 f_Y(y) &= \frac{1}{b^2} \left(\frac{b}{\sqrt{y}} - 1\right)
 \end{aligned}$$

$$\begin{aligned}
 \mathbb{E}[Y^k]_\epsilon &= \int_\epsilon^{b^2} \frac{y^k}{b^2} \left(\frac{b}{\sqrt{y}} - 1 \right) dy \\
 \mathbb{E}[Y^k]_\epsilon &= \frac{1}{b^2} \int_\epsilon^{b^2} by^{k-\frac{1}{2}} - y^k dy \\
 \mathbb{E}[Y^k] &= \frac{1}{b^2} \int_0^{b^2} by^{k-\frac{1}{2}} - y^k dy \\
 &= \frac{1}{b^2} \left[b \frac{2}{2k+1} y^{k+\frac{1}{2}} - \frac{y^{k+1}}{k+1} \right]_0^{b^2} \\
 &= \frac{1}{b^2} \left(\frac{2}{2k+1} b^{2k+2} - \frac{1}{k+1} b^{2k+2} \right) \\
 \mathbb{E}[Y^k] &= \frac{b^{2k}}{(2k+1)(k+1)}
 \end{aligned}$$

And in particular:

$$\begin{aligned}
 \mathbb{E}[Y] &= \frac{b^2}{6} \\
 \mathbb{E}[Y^2] &= \frac{b^4}{15} \\
 \text{Var}[Y] &= \frac{b^4}{15} - \left(\frac{b^2}{6} \right)^2 = b^4 \left(\frac{1}{15} - \frac{1}{36} \right) = \frac{7}{180} b^4
 \end{aligned}$$

4.3.c Absolute Value in the Symmetric Centred Case $a = -b$, $c = 0$

Let $X \sim \mathcal{T}(a = -b, b = b, c = 0)$.

Let $Z = |X|$, note that $0 \leq Z \leq b$.

$$\begin{aligned}
 P(Z \leq z) &= 2 \int_0^z \frac{b-x}{b^2} dx \\
 &= 2 \left[-\frac{(b-x)^2}{2b^2} \right]_0^z \\
 &= 2 \left(\frac{1}{2} - \frac{(b-z)^2}{2b^2} \right) \\
 F_Z(z) &= 1 - \frac{(b-z)^2}{b^2}
 \end{aligned}$$

Therefore

$$f_Z(z) = \frac{2(b-z)}{b^2}$$

We recognise a triangular distribution: $a = 0, b = b, c = 0$.

The mode of Z is 0.

The median is $\frac{\sqrt{2}}{2}b$.

And we can compute (note that we can also make a direct application of the above general formula):

$$\begin{aligned} \mathbb{E}[Z^k] &= \int_0^b z^k \frac{2(b-z)}{b^2} dz \\ &= \frac{2}{b^2} \int_0^b bz^k - z^{k+1} dz \\ &= \frac{2}{b^2} \left[b \frac{z^{k+1}}{k+1} - \frac{z^{k+2}}{k+2} \right]_0^b \\ \mathbb{E}[Z^k] &= \frac{2b^k}{(k+1)(k+2)} \end{aligned}$$

In particular:

$$\begin{aligned} \mathbb{E}[Z] &= \frac{b}{3} \\ \mathbb{E}[Z^2] &= \frac{b^2}{6} \\ \text{Var}[Z] &= \frac{b^2}{6} - \frac{b^2}{9} \\ \text{Var}[Z] &= \frac{b^2}{18} \end{aligned}$$

5 Supplementary Material About the Hassanat Distance

Démonstration que la distance de Hassanat est effectivement une distance

Marc Grossouvre

29 mai 2024

1 Prérequis

1.1 Distance sur un ensemble quelconque

Définition 1.1. Une application d d'un ensemble E^2 sur \mathbb{R} , symétrique positive et vérifiant l'inégalité triangulaire est appelée une distance sur E . Lorsque d est aussi définie, on dit que d est une distance séparante sur E .

Symétrique :

$$\forall (x, y) \in E^2, d(x, y) = d(y, x)$$

Positive :

$$\forall (x, y) \in E^2, d(x, y) \geq 0$$

Inégalité triangulaire :

$$\forall (x, y, z) \in E^3, d(x, z) \leq d(x, y) + d(y, z)$$

Définie :

$$\forall (x, y) \in E^2, d(x, y) = 0 \Leftrightarrow x = y$$

Propriété 1.1.1. Une somme finie de distances définies sur un même ensemble E est elle-même une distance.

Démonstration. Soit $(d_i)_{i \in \llbracket 1, n \rrbracket}$ une famille finie de distances, définies sur un ensemble E .

Soit D l'application :

$$D : E \times E \longrightarrow \mathbb{R}^+$$

$$(x, y) \longmapsto D(x, y) = \sum_{i=1}^n d_i(x, y)$$

Cette application est symétrique par symétrie de chacune des distances.

Elle est aussi positive car la somme de fonctions positives reste positive.

En sommant les inégalités triangulaires écrites pour chacune des distance, on montre que D satisfait aussi l'inégalité triangulaire. ■

Propriété 1.1.2. Si au moins une des distances d'une famille finie de distances est séparante, alors la somme des distances de cette famille est une distance séparante.

Démonstration. Sachant qu'une somme de nombres positifs est nulle si et seulement si chacun de ses termes est nul, le résultat en découle directement. ■

1.2 Norme sur un espace vectoriel

Définition 1.2. Une application \mathcal{N} d'un \mathbb{R} -espace vectoriel E dans \mathbb{R} définie positive, absolument homogène et sous-additive, est appelée une norme sur E .

La norme est notée $\|\cdot\|$ s'il n'y a pas d'ambiguïté.

Définie :

$$\forall x \in E, \|x\| = 0 \Leftrightarrow x = 0$$

Positive :

$$\forall x \in E, \|x\| \geq 0$$

Absolument homogène :

$$\forall x \in E, \forall \lambda \in \mathbb{R}, \|\lambda x\| = |\lambda| \|x\|$$

Sous-additive :

$$\|x + y\| \leq \|x\| + \|y\|$$

Propriété 1.2.1. Si $\|\cdot\|$ est une norme alors $(x, y) \mapsto \|x - y\|$ est une distance séparante.

1.3 Applications concaves

Définition 1.3. Une application f d'un \mathbb{R} -espace vectoriel E dans \mathbb{R} est dite concave lorsqu'elle vérifie la propriété :

$$\forall (c, y) \in \mathbb{R}, \forall \lambda \in [0, 1], f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y)$$

Si $E = \mathbb{R}$, on dit que le graphe de la fonction est au dessus de ses cordes.

Propriété 1.3.1. Une fonction deux fois dérivables sur un intervalle ouvert de \mathbb{R} est concave si et seulement si sa dérivée seconde est négative ou nulle sur cet intervalle.

Démonstration. Ce résultat est une application classique du théorème dit des accroissements finis. Nous renvoyons le lecteur au cours en ligne de Bernard Ycart, professeur à l'université de Grenoble-Alpes [1]. ■

Propriété 1.3.2. En remplaçant y par 0 dans la formule de concavité, il apparaît que pour une fonction f concave positive à l'origine, on observe aussi une forme d'homogénéité :

$$f \text{ concave, } f(0) \geq 0 \text{ et } \lambda \in [0, 1] \Rightarrow f(\lambda x) \geq \lambda f(x)$$

Propriété 1.3.3. Comme conséquence de la propriété précédente, dans le cas particulier $E = \mathbb{R}$, il apparaît qu'une fonction concave, positive à l'origine est sous-additive :

$$f \text{ concave et } f(0) \geq 0 \Rightarrow \forall (x, y) \in \mathbb{R}^2, f(x + y) \leq f(x) + f(y)$$

Démonstration. Soit f une fonction concave sur E et positive à l'origine. Soit $(x, y) \in \mathbb{R}^2$.

$$\begin{aligned} f(x) + f(y) &= f\left(\frac{x}{x+y}(x+y)\right) + f\left(\frac{y}{x+y}(x+y)\right) \\ f(x) + f(y) &\geq \frac{x}{x+y}f(x+y) + \frac{y}{x+y}f(x+y) \quad (\text{par la propriété 1.3.2}) \\ f(x) + f(y) &\geq f(x+y) \end{aligned}$$

■

On en déduit le résultat suivant :

Lemme 1.1. Soit d une distance sur un ensemble E quelconque, soit f une fonction concave et croissante sur \mathbb{R}^+ nulle à l'origine, alors la composée $f \circ d$ définit une distance sur E . De plus, lorsque d est séparante et f strictement croissante, la distance ainsi définie est séparante.

Démonstration. Notons d'abord que cette composition ne pose pas de problème de définition :

$$\begin{array}{ccc} E \times E & \xrightarrow{d} & \mathbb{R}^+ & \xrightarrow{f} & \mathbb{R}^+ \\ (x, y) & \mapsto & d(x, y) & \mapsto & f \circ d(x, y) = f(d(x, y)) \end{array}$$

La symétrie de d entraîne celle de $f \circ d$.

f est nulle à l'origine et strictement croissante sur R^+ donc f est positive sur R^+ et par suite, $f \circ d$ est aussi positive sur R^+ .

Concernant l'inégalité triangulaire, soient $(x, y, z) \in E^3$. D'après la propriété 1.3.3 f est sous-additive donc on a :

$$f(d(x, y)) + f(d(y, z)) \geq f(d(x, y) + d(y, z)) \quad (1)$$

Par ailleurs, d satisfait l'inégalité triangulaire donc on a :

$$d(x, y) + d(y, z) \geq d(x, z)$$

Et f est croissante donc on peut appliquer f à cette inégalité :

$$f(d(x, y) + d(y, z)) \geq f(d(x, z)) \quad (2)$$

Et finalement par transitivité des inégalités 1 et 2 :

$$f(d(x, y)) + f(d(y, z)) \geq f(d(x, z))$$

Ce qui est l'inégalité triangulaire pour $f \circ d$.

$f \circ d$ est symétrique, positive et vérifie l'inégalité triangulaire, c'est donc une distance.

De plus, dans le cas où f est strictement croissante, l'équation $f(x) = 0$ a au plus une solution. Or $f(0) = 0$ donc :

$$f(x) = 0 \Leftrightarrow x = 0$$

On a donc :

$$f(d(x, y)) = 0 \Leftrightarrow d(x, y) = 0$$

Et lorsque d est séparante, cela conduit à :

$$f(d(x, y)) = 0 \Leftrightarrow x = y$$

Ce qui montre que $f \circ d$ est séparante. ■

Ce résultat combiné à la propriété 1.2.1 donne immédiatement un résultat similaire pour les normes :

Lemme 1.2. *Soit $\|\cdot\|$ une norme sur un \mathbb{R} -espace vectoriel E , soit f une fonction concave et croissante sur \mathbb{R}^+ nulle à l'origine, alors $(x, y) \mapsto f(\|x - y\|)$ est une distance sur E . De plus, lorsque f est strictement croissante, la distance ainsi définie est séparante.*

2 La fonction de Hassanat est-elle une distance ?

Dans un article publié en 2014 [2], Ahmad Basheer Hassanat cherche à étendre la distance dite de "Wave-Hedges" définie sur \mathbb{R}^+ à l'ensemble des réels. Nous allons reprendre ici son résultat principal, à savoir la définition d'une nouvelle distance mais nous proposons une démonstration complète du fait que nous avons effectivement affaire à une distance. La démonstration proposée par Hassanat comporte une erreur de logique dans la partie sur l'inégalité triangulaire et passe sous silence des cas incontournables pour une preuve complète.

Définition 2.1. Nous définissons la fonction de Hassanat d_H comme la fonction définie de \mathbb{R}^2 dans \mathbb{R} par les deux formules :

$$\begin{aligned} \forall (x, y) \in \mathbb{R}^2 \text{ tel que } \min(x, y) \geq 0, \\ d_H(x, y) = 1 - \frac{1 + \min(x, y)}{1 + \max(x, y)} \\ \forall (x, y) \in \mathbb{R}^2 \text{ tel que } \min(x, y) < 0, \\ d_H(x, y) = 1 - \frac{1 + \min(x, y) + |\min(x, y)|}{1 + \max(x, y) + |\min(x, y)|} \end{aligned}$$

Il convient d'abord de confirmer que cette écriture définit correctement une fonction sur \mathbb{R}^2 .

On note :

$$\begin{aligned} P_1 &= \{(x, y) \mid \min(x, y) \geq 0\} \\ P_2 &= \{(x, y) \mid \max(x, y) < 0\} \end{aligned}$$

P_1 et P_2 forment une partition de \mathbb{R}^2 . P_1 est le quadrant supérieur droit du plan, fermé et convexe. P_2 regroupe les 3 autres quadrants et forme un ensemble ouvert connexe mais non-convexe. d_H est donc définie de façon univoque.

De plus, sur P_1 , $\min(x, y) \geq 0$ donc a fortiori $\max(x, y) \geq 0$, ce qui entraîne

$$1 + \max(x, y) \geq 1 > 0$$

Donc le dénominateur de la fraction utilisée dans la formule n'est jamais nul et la fraction est correctement définie.

Sur P_2 , $\min(x, y) < 0$ donc $|\min(x, y)| = -\min(x, y)$. On peut alors récrire la formule servant à définir d_H :

$$d_H(x, y) = 1 - \frac{1 + \min(x, y) - \min(x, y)}{1 + \max(x, y) - \min(x, y)}$$

qui peut se simplifier en :

$$d_H(x, y) = 1 - \frac{1}{1 + \max(x, y) - \min(x, y)} \quad (1)$$

$$d_H(x, y) = 1 - \frac{1}{1 + |x - y|} \quad (2)$$

Le nombre $|x - y|$ est toujours positif donc :

$$1 + \max(x, y) - \min(x, y) \geq 1 > 0$$

Donc le dénominateur de la fraction utilisée dans la formule n'est lui non-plus jamais nul et la fraction est correctement définie.

Proposition 2.1. *La fonction de Hassanat définit une distance séparante sur \mathbb{R} .*

Démonstration. Les fonctions \min et \max sont symétriques sur \mathbb{R}^2 donc par composition, d_H est aussi symétrique. Ce qui nous garantit aussi que $d_H(x, y)$ et $d_H(y, x)$ sont toujours définis par la même formule.

La fonction \max est supérieure à la fonction \min sur \mathbb{R}^2 donc dans les deux formules de la définition 2.1, les fractions sont inférieures à 1 donc leurs compléments à 1 sont positifs donc d_H est positive.

Par ailleurs, quelle que soit la formule utilisée, si la fonction de Hassanat est nulle, alors le quotient de la formule utilisée doit être égal à 1 et nécessairement $\min(x, y) = \max(x, y)$, ce qui est équivalent à $x = y$. d_H est donc définie.

L'inégalité triangulaire est beaucoup plus délicate à prouver. Nous devons disjoindre plusieurs situations. Soient 3 éléments de \mathbb{R}^3 que l'on nomme dans un ordre croissant de sorte que $x \leq y \leq z$. Alors 4 situations sont possibles :

$$0 \leq x \leq y \leq z$$

$$x < 0 \leq y \leq z$$

$$x \leq y < 0 \leq z$$

$$x \leq y \leq z < 0$$

Dans la suite, on note :

$$d_1 = d_H(x, y)$$

$$d_2 = d_H(y, z)$$

$$d_3 = d_H(x, z)$$

2.0.1 Cas $0 \leq x \leq y \leq z$

Les formules de la définition 2.1 se simplifient après réduction au même dénominateur :

$$d_1 = \frac{y-x}{1+y} ; d_2 = \frac{z-y}{1+z} ; d_3 = \frac{z-x}{1+z}$$

Du fait que x, y et z sont ordonnés, ils ne sont pas interchangeables dans les formules et nous devons prouver l'inégalité triangulaire dans les 3 configurations possibles.

$$\begin{aligned} d_1 + d_2 - d_3 &= \frac{y-x}{1+y} + \frac{z-y}{1+z} - \frac{z-x}{1+z} \\ d_1 + d_2 - d_3 &= \frac{y-x}{1+y} + \frac{x-y}{1+z} \\ d_1 + d_2 - d_3 &= (y-x) \left(\frac{1}{1+y} - \frac{1}{1+z} \right) \\ d_1 + d_2 - d_3 &= \frac{(y-x)(z-y)}{(1+y)(1+z)} \\ \mathbf{d_1 + d_2 - d_3} &\geq \mathbf{0} \end{aligned}$$

De la même façon, on a :

$$\begin{aligned} d_1 + d_3 - d_2 &= \frac{y-x}{1+y} + \frac{z-x}{1+z} - \frac{z-y}{1+z} \\ d_1 + d_3 - d_2 &= (y-x) \left(\frac{1}{1+y} + \frac{1}{1+z} \right) \\ d_1 + d_3 - d_2 &= \frac{(y-x)(2+y+z)}{(1+y)(1+z)} \\ \mathbf{d_1 + d_3 - d_2} &\geq \mathbf{0} \end{aligned}$$

Enfin :

$$\begin{aligned} d_2 + d_3 - d_1 &= \frac{z-y}{1+z} + \frac{z-x}{1+z} - \frac{y-x}{1+y} \\ d_2 + d_3 - d_1 &= \frac{(z-y)(1+y) + (z-x)(1+y) - (y-x)(1+z)}{(1+y)(1+z)} \\ d_2 + d_3 - d_1 &= \frac{(z-y)(1+y) + z + y - x - y - y + x + z}{(1+y)(1+z)} \\ d_2 + d_3 - d_1 &= \frac{(z-y)(2+x+y)}{(1+y)(1+z)} \\ \mathbf{d_2 + d_3 - d_1} &\geq \mathbf{0} \end{aligned}$$

L'inégalité triangulaire est donc vérifiée dans ce cas.

2.0.2 Cas $x < 0 \leq y \leq z$

$$d_1 = \frac{y-x}{1+y-x} ; d_2 = \frac{z-y}{1+z} ; d_3 = \frac{z-x}{1+z-x}$$

$$d_1 + d_2 - d_3 = \frac{y-x}{1+y-x} + \frac{z-y}{1+z} - \frac{z-x}{1+z-x}$$

On applique la fonction `factor(simplify())` à l'expression et on obtient :

$$d_1 + d_2 - d_3 = \frac{(-y+z)(x^2 - xy - xz - 2x + yz + y)}{(-z-1)(x-z-1)(y-x+1)}$$

qui peut se récrire :

$$d_1 + d_2 - d_3 = \frac{(z-y)(x^2 + yz + y - x(2+y+z))}{(z+1)(1+z-x)(1+y-x)}$$

En tenant compte de l'ordre de x , y et z et de leurs signes respectifs, on obtient :

$$\mathbf{d_1 + d_2 - d_3 \geq 0}$$

$$d_1 + d_3 - d_2 = \frac{y-x}{1+y-x} + \frac{z-x}{1+z-x} - \frac{z-y}{1+z}$$

On applique la fonction `factor(simplify())` à l'expression et on obtient :

$$d_1 + d_3 - d_2 = \frac{x^2y + x^2z + 2x^2 - xy^2 - 2xyz - 4xy - xz^2 - 2xz - 2x + y^2z + y^2 + yz^2 + 3yz + 2y}{(-z-1)(x-z-1)(y-x+1)}$$

qui peut se récrire :

$$d_1 + d_3 - d_2 = \frac{x^2y + x^2z + 2x^2 + y^2z + y^2 + yz^2 + 3yz + 2y - x(y^2 + 2yz + 4y + z^2 + 2z + 2)}{(1+z)(1+z-x)(1+y-x)}$$

En tenant compte de l'ordre de x , y et z et de leurs signes respectifs, on obtient :

$$\mathbf{d_1 + d_3 - d_2 \geq 0}$$

$$d_2 + d_3 - d_1 = \frac{z - y}{1 + z} + \frac{z - x}{1 + z - x} - \frac{y - x}{1 + y - x}$$

On applique la fonction `factor(simplify())` à l'expression et on obtient :

$$d_2 + d_3 - d_1 = \frac{(-y + z)(x^2 - xy - xz - 2x + yz + y + 2z + 2)}{(-z - 1)(x - z - 1)(y - x + 1)}$$

qui peut se récrire :

$$d_2 + d_3 - d_1 = \frac{(z - y)(x^2 + yz + y + 2z + 2 - x(2 + y + z))}{(1 + z)(1 + z - x)(1 + y - x)}$$

En tenant compte de l'ordre de x , y et z et de leurs signes respectifs, on obtient :

$$\mathbf{d_2 + d_3 - d_1 \geq 0}$$

2.0.3 Cas $x \leq y < 0 \leq z$ ou $x \leq y \leq z < 0$

Dans ce cas, seule la deuxième formule est utilisée pour déterminer les 3 distances d_1 , d_2 et d_3 .

D'après l'égalité 2 page 6 :

$$d_H(x, y) = 1 - \frac{1}{1 + |x - y|} = f(|x - y|)$$

où f est la fonction définie sur \mathbb{R}^+ par la formule :

$$f(x) = 1 - \frac{1}{1 + x}$$

Cette fonction est nulle à l'origine, infiniment dérivable sur son domaine de définition, en particulier :

$$f'(x) = \frac{1}{(1 + x)^2} > 0$$

et

$$f''(x) = -\frac{2}{(1 + x)^3} < 0$$

donc f est concave et strictement croissante sur \mathbb{R}^+ .

De plus nous avons vu que $|x - y|$ définit une distance sur \mathbb{R} .

Donc d'après le lemme 1.1, $d_H(x, y) = f(|x - y|)$ définit une distance sur \mathbb{R} et en particulier, vérifie l'inégalité triangulaire.

d_H est donc une fonction symétrique, définie, positive qui vérifie l'inégalité triangulaire : c'est une distance séparante sur \mathbb{R}^2 . ■

À partir de ce résultat, il est possible d'étendre en 2 temps le résultat à un espace multi-dimensionnel pour retrouver la distance telle que définie par Hassanat en 2014.

Corollaire 2.1.1. *Pour un entier $n \geq 1$ et pour $i \in \llbracket 1, n \rrbracket$, l'application :*

$$d_{H,i} : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$$

$$(X, Y) = \left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right) \longmapsto d_H(x_i, y_i)$$

définit une distance (non-séparante) sur \mathbb{R}^n

Corollaire 2.1.2. *Pour un entier $n \geq 1$, l'application :*

$$d_H^n : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$$

$$(X, Y) = \left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right) \longmapsto d_H^n(X, Y) = \sum_{i=1}^n d_H(x_i, y_i)$$

définit une distance séparante sur \mathbb{R}^n .

Démonstration. Nous avons déjà vu dans la propriété 1.1.1 page 1 que le passage à la somme ne pose de difficulté.

De plus, une somme de nombre positifs est nulle si et seulement si chacun de ses termes est nul. Or dans la somme que nous avons définie, un terme est nul si et seulement si les deux coordonnées associées sont égales. La distance entre X et Y est donc nulle si et seulement si toutes leurs coordonnées sont 2 à 2 égales, c'est à dire si et seulement si $X = Y$. Ce qui prouve que la distance ainsi définie est séparante. ■

Dans la suite, nous appelons distance de Hassanat la distance ainsi définie.

3 Cercle de Hassanat

En dimension 1 Supposons $0 \leq x \leq y$, $c > 0$, $c \neq 1$.

$$1 - \frac{1+x}{1-y} = c$$

$$y = -\frac{1}{1-c}x + \frac{c}{1-c}$$

Références

- [1] B. Ycart, “Fonctions convexes,” 2011. [Online]. Available : <https://ljk.imag.fr/membres/Bernard.Ycart/mel/dc/node7.html>
- [2] A. B. Hassanat, “Dimensionality Invariant Similarity Measure,” *arXiv :1409.0923 [cs]*, Sep. 2014, arXiv : 1409.0923. [Online]. Available : <http://arxiv.org/abs/1409.0923>

Table des matières

1	Prérequis	1
1.1	Distance sur un ensemble quelconque	1
1.2	Norme sur un espace vectoriel	2
1.3	Applications concaves	3
2	La fonction de Hassanat est-elle une distance ?	5
2.0.1	Cas $0 \leq x \leq y \leq z$	7
2.0.2	Cas $x < 0 \leq y \leq z$	8
2.0.3	Cas $x \leq y < 0 \leq z$ ou $x \leq y \leq z < 0$	9
3	Cercle de Hassanat	11

Supplementary Material for Chapter II

6	Proof of Proposition 2222
7	Cross-Errors225
8	Operations on Granularities227
9	Detailed Implementation of Mixture Kriging for EPCs231
9.1	Mixture Variables: Covariance, Likelihood and Conditional Likelihood	231
9.2	Functions of Mixtures235
9.3	Observations and Repeated Observations239
9.4	Modeling241
9.5	Data Processing242
9.6	Computing the Model243

6 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:

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

And similarly, we denote:

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

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 3). But it also means that the sets of unbiased linear predictors for each output variable are parallel:

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

Now, given that we are minimising 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 optimisation process is geometrically a projection of $Y_i(g)$ on $F_i(g)$. This approach is illustrated in Figure 3.

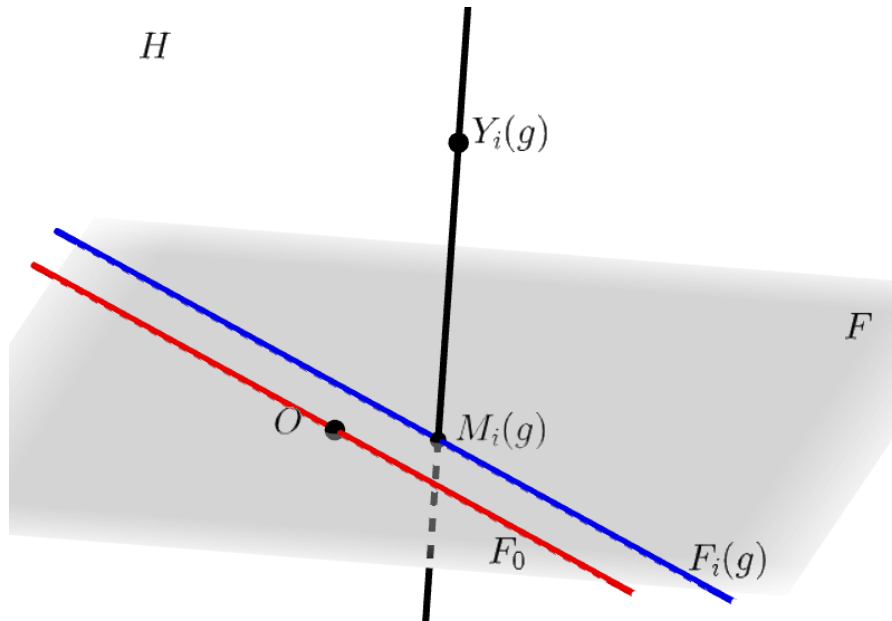


Figure 3 – Geometrical interpretation of the prediction process.

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

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

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

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

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

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

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

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

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

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

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

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

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 minimising 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:

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

Due to unbiasedness condition, it means that:

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

Which rewrites:

$$c_{i,j}(g, g') = k_{i,j}(g, g') - \boldsymbol{\alpha}_j(g')^\top \mathbf{h}_i(g) - \boldsymbol{\alpha}_i(g)^\top \mathbf{h}_j(g') + \boldsymbol{\alpha}_i(g)^\top \mathbf{K} \boldsymbol{\alpha}_j(g'). \quad (\text{IV.3})$$

Note that equation (IV.3) is true for all linear unbiased predictors.

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. \square

7 Cross-Errors and Conditional Covariances

Proposition 20 (Cross-errors and conditional covariances). *Consider the assumption*

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

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

$$c_{i,j}(g, g') = \mathbb{E}[\text{Cov}[Y_i(g), Y_j(g') \mid \underline{\mathbf{Y}}]]. \quad (\text{IV.4})$$

If $\text{Cov}[Y_i(g), Y_j(g') \mid \underline{\mathbf{Y}}]$ does not depend on $\underline{\mathbf{Y}}$, as it is the case for conditional Gaussian vectors, Equation simplifies: $\mathbb{E}[\text{Cov}[Y_i(g), Y_j(g') \mid \underline{\mathbf{Y}}]] = \text{Cov}[Y_i(g), Y_j(g') \mid \underline{\mathbf{Y}}]$.

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 Supplementary Material 6, page 222, 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 $v_i^*(g) := \mathbb{E}[(\epsilon_i^*(g))^2]$.

We have:

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

The minimum value of this polynomial expression is reached for:

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

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

$$\forall \delta \in F_0, \quad \forall i \in \{1, \dots, p\}, \quad \forall g \in \chi, \quad \mathbb{E}[\epsilon_i^*(g)\delta] = \text{Cov}[\epsilon_i^*(g), \delta] = 0. \quad (\text{IV.5})$$

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 [119], 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 (IV.5), we have:

$$\text{Cov}[\epsilon_i^*(g) + \delta, \epsilon_j^*(g') + \delta'] = c_{i,j}^*(g, g') + 0 + 0 + \text{Cov}[\delta, \delta'] \quad (\text{IV.6})$$

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

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

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

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

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

$$\text{Cov} [\epsilon_i^*(g) + \delta, \epsilon_j^*(g') + \delta'] = \mathbb{E} [\text{Cov} [Y_i(g), Y_j(g') \mid \underline{\mathbf{Y}}]] + \text{Cov} [\delta, \delta'] \quad (\text{IV.7})$$

Identifying the equations (IV.6) and (IV.7), we get the expected result.

□

8 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 20 (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 21 (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:*

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

\mathcal{G} is said to be *thinner than* \mathcal{H} , or equivalently \mathcal{H} *coarser than* \mathcal{G} , see Figure 4. 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.

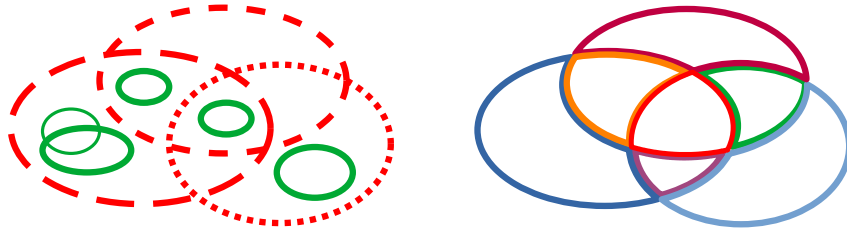


Figure 4 – *Thinner granularity and maximal thinner non-overlapping granularity. Left:* The granularity comprising the 5 green grains (solid lines) is thinner than the granularity comprising the 3 red grains (dashed lines). **Right:** The granularity comprising 7 non overlapping grains is the maximal non-overlapping granularity that is thinner than the red granularity on the left.

Proposition 21 (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 $h \cup \bigcup_{g \in \mathcal{G}} g$.

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 $\mathcal{G} = \{g_1, \dots, g_n\}$, there is a unique **maximal non-overlapping granularity** \mathcal{G}^\oplus and we can construct it iteratively with the insertion operator.

$$\mathcal{G}^\oplus := \{g_1\} \oplus \dots \oplus \{g_n\}. \quad (\text{IV.8})$$

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 $g_+ \in \bigcup_{g \in \mathcal{G} \cup \{h\}} g$. Moreover:

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

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:

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

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

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

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

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:

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

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 21 and build \mathcal{G}^\oplus directly from any finite granularity \mathcal{G} , possibly overlapping. However, we will see in the rest of the chapter 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 21 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 8 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 21, see Supplementary Material 8, page 227.

Example 8 (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.*

- *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:

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

- *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:

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

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

$$\Delta := \text{Cov} [Y_i(X_g), Y_j(X_{g'}^\perp)] - \text{Cov} [Y_i(X_g), Y_j(X_{g'})]. \quad (\text{IV.9})$$

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

$$\forall x \in g \cup g', \mu_i(x) = \mu_i(g) = \mu_i(g') \text{ and } \mu_j(x) = \mu_j(g) = \mu_j(g') \quad (\text{IV.10})$$

one can show that:

$$|\Delta| \leq \text{P} [B = B' = 1] \text{P} [X_{g_0} \neq X_{g_0}^\perp] \rho_{\max}. \quad (\text{IV.11})$$

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 8. Under given assumptions on the means μ_i and μ_j , Applying the total covariance formula on $\text{Cov} [Y_i(X_g), Y_j(X_{g'}^\perp)]$ and $\text{Cov} [Y_i(X_g), Y_j(X_{g'})]$, we get

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

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

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

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 ρ_{\max} , hence the result. \square

9 Detailed Implementation of Mixture Kriging for EPCs

This section was not included in the published article. It presents an in-depth understanding of Mixture Kriging in the Gaussian case and the details of the model that was implemented. The main points are:

- Subsection 9.1: In the Gaussian case, likelihood of the observations and of the conditional observations can be expressed with closed formulae.
- Subsection 9.2: It is also possible to express functions of mixtures, in particular linear combinations of mixtures, so that it is possible to express in a closed formula the density of the leave-one-out cross-validation Mixture Kriging predictor. And it is possible to transform input and output variables to fall in this Gaussian case.
- Subsection 9.3: Repeated observations can be used in the learning set with no contradiction.
- Subsections 9.4, 9.5 and 9.6 define granularities, data processing and model optimisation for a model that takes into account geographical position, year of construction and median inhabitants' income.

9.1 Mixture Variables: Covariance, Likelihood and Conditional Likelihood

We work on a **territory** $\chi \subset \mathbb{R}^d$. The considered output

$$\mathbf{Y} = \left\{ \mathbf{Y}(x) = (Y_1(x), \dots, Y_p(x))^\top : x \in \chi \right\}$$

is a p -dimensional multivariate random field over χ . We consider a non necessarily finite **granularity** \mathcal{G} . For each **grain** g of \mathcal{G} , X_g is a random variable that gives a random location in g and we define a p -dimensional random vector that is the value of \mathbf{Y} at a random location $X_g \in g$:

$$\begin{aligned} \mathbf{Y}(g) &= \mathbf{Y}(X_g) \\ &= (Y_1(g), \dots, Y_p(g))^\top \\ &= (Y_1(X_g), \dots, Y_p(X_g))^\top \end{aligned}$$

By definition, \mathbf{Y} is a p -dimensional **mixture** distribution.

We consider a given finite set of n grains $\{g_1, \dots, g_n\}$ in \mathcal{G} , not necessarily pairwise distinct, for which we use the shorthand $X_{g_j} = X_j$ and we denote $\mathbf{X} = (X_1, \dots, X_n)^\top$. We also consider a set of n indices $\{i_1, \dots, i_n\} \in \{1, \dots, p\}^n$. An **observation** of $Y_{i_j}(g_j)$ is denoted Y^j . Eventually, we consider the vector of n observations $\mathbf{Y} = (Y^1, \dots, Y^n)^\top$, it is itself an n -dimensional **mixture** variable. Let $\Sigma_{\mathbf{Y}}$ be a covariance matrix gathering all the elements of the set:

$$\left\{ \text{Cov} \left[Y_{i_k}(x), Y_{i_{k'}}(x') \right] : (k, k') \in \{1, \dots, n\}^2, (x, x') \in g_k \times g_{k'} \right\}.$$

For a given vector of points $\mathbf{x} = (x_1, \dots, x_n)^\top \in g_1 \times \dots \times g_n$, we denote:

$$\begin{aligned} \underline{\mathbf{Y}}(\mathbf{x}) &= (Y_{i_1}(x_1), \dots, Y_{i_n}(x_n))^\top \\ \underline{\mu}_{\mathbf{x}} &= (\mu_{i_1}(x_1), \dots, \mu_{i_n}(x_n))^\top \\ \Sigma_{\mathbf{x}} &= \left(\text{Cov} \left[Y_{i_j}(x_j), Y_{i_k}(x_k) \right] \right)_{\substack{1 \leq j \leq n \\ 1 \leq k \leq n}} \in \mathcal{M}_{n,n}(\mathbb{R}) \end{aligned}$$

$\Sigma_{\mathbf{x}}$ is symmetric and we assume that it is always definite positive. It is a matrix extracted from $\Sigma_{\mathbf{Y}}$. For the sake of simplicity and applicability, we assume that \mathbf{X} is discrete and takes a finite number of values with probabilities $p_{\mathbf{x}} = P(\mathbf{X} = \mathbf{x})$. Note that there is no loss of generality in restricting a grain g to the support of the associated X_g random variable, in which case, the above assumption means that $g_{i_1} \times \dots \times g_{i_n}$ is finite.

Proposition 22 (General case: Covariance matrix of observations). *With the above framework and notations, we have:*

$$\begin{aligned} \text{Cov} \left[Y^j, Y^k \right] &= \sum_{(x_j, x_k) \in g_j \times g_k} p_{(x_j, x_k)} \left(\text{Cov} \left[Y_{i_j}(x_j), Y_{i_k}(x_k) \right] + \mu_{i_j}(x_j) \mu_{i_k}(x_k) \right) \\ &\quad - \sum_{x \in g_j} p_x \mu_{i_j}(x) \sum_{x \in g_k} p_x \mu_{i_k}(x) \end{aligned}$$

and

$$\text{Var} \left[Y^j \right] = \sum_{x_j \in g_j} p_{x_j} \left(\text{Var} \left[Y_{i_j}(x_j) \right] + \mu_{i_j}(x_j)^2 \right) - \left(\sum_{x \in g_j} p_x \mu_{i_j}(x) \right)^2$$

Proof.

$$\begin{aligned} \text{Cov} \left[Y^j, Y^k \right] &= \text{Cov} \left[Y_{i_j}(g_j), Y_{i_k}(g_k) \right] \\ &= \mathbb{E} \left[\text{Cov} \left[Y_{i_j}(X_j), Y_{i_k}(X_k) \right] \mid X_j, X_k \right] + \text{Cov} \left[\mathbb{E} \left[Y_{i_j}(X_j) \mid X_j \right], \mathbb{E} \left[Y_{i_k}(X_k) \mid X_k \right] \right] \\ &= \sum_{(x_j, x_k) \in g_j \times g_k} p_{(x_j, x_k)} \left(\Sigma_{(x_j, x_k)} + \left(\mu_{i_j}(x_j) - \sum_{x \in g_j} p_x \mu_{i_j}(x) \right) \left(\mu_{i_k}(x_k) - \sum_{x \in g_k} p_x \mu_{i_k}(x) \right) \right) \\ &= \sum_{(x_j, x_k) \in g_j \times g_k} p_{(x_j, x_k)} \left(\text{Cov} \left[Y_{i_j}(x_j), Y_{i_k}(x_k) \right] + \mu_{i_j}(x_j) \mu_{i_k}(x_k) \right) - \sum_{x \in g_j} p_x \mu_{i_j}(x) \sum_{x \in g_k} p_x \mu_{i_k}(x) \end{aligned}$$

$$\begin{aligned}
 \text{Var} [Y^j] &= \text{Cov} [Y_{i_j}(g_j), Y_{i_j}(g_j)] \\
 &= \mathbb{E} [\text{Cov} [Y_{i_j}(X_j), Y_{i_j}(X_j)] | X_j] + \text{Cov} [\mathbb{E} [Y_{i_j}(X_j) | X_j], \mathbb{E} [Y_{i_j}(X_j) | X_j]] \\
 &= \sum_{x_j \in g_j} p_{x_j} \left(\Sigma_{x_j} + \left(\mu_{i_j}(x_j) - \sum_{x \in g_j} p_x \mu_{i_j}(x) \right)^2 \right) \\
 \text{Var} [Y^j] &= \sum_{x_j \in g_j} p_{x_j} \left(\text{Var} [Y_{i_j}(x_j)] + \mu_{i_j}(x_j)^2 \right) - \left(\sum_{x \in g_j} p_x \mu_{i_j}(x) \right)^2
 \end{aligned}$$

□

Proposition 23 (Gaussian case: Likelihood of observations). *With the above framework and notations, assuming that \mathbf{Y} is a p -variate Gaussian field over the points of χ , the density of $\underline{\mathbf{Y}}$ at a given point $\underline{\mathbf{y}} = (y^1, \dots, y^n)^\top$ is:*

$$f_{\underline{\mathbf{Y}}}(\underline{\mathbf{y}}) = \sum_{\underline{\mathbf{x}} \in g_1 \times \dots \times g_n} p_{\underline{\mathbf{x}}} \frac{1}{\sqrt{(2\pi)^n |\Sigma_{\underline{\mathbf{x}}}|}} e^{-\frac{1}{2}(\underline{\mathbf{y}} - \underline{\mu}_{\underline{\mathbf{x}}})^\top \Sigma_{\underline{\mathbf{x}}}^{-1} (\underline{\mathbf{y}} - \underline{\mu}_{\underline{\mathbf{x}}})} \quad (\text{IV.12})$$

Sketch of the proof. For all $\underline{\mathbf{x}}$, the vector $\underline{\mathbf{Y}}(\underline{\mathbf{x}})$ is a Gaussian vector as a subvector of \mathbf{Y} . Therefore, its density can be explicitated. Moreover, for a given event A , we have:

$$P(\underline{\mathbf{Y}} \in A) = \mathbb{E} [P(\underline{\mathbf{Y}} \in A | \underline{\mathbf{X}})] .$$

If we take A as a neighbourhood in \mathbb{R}^n of a point $\underline{\mathbf{y}} = (y^1, \dots, y^n)^\top$ we derive the density of $\underline{\mathbf{Y}}$:

$$\begin{aligned}
 f_{\underline{\mathbf{Y}}}(\underline{\mathbf{y}}) &= \sum_{\underline{\mathbf{x}} \in g_1 \times \dots \times g_n} p_{\underline{\mathbf{x}}} f_{\underline{\mathbf{Y}}(\underline{\mathbf{x}})}(\underline{\mathbf{y}}) \\
 \text{and therefore } f_{\underline{\mathbf{Y}}}(\underline{\mathbf{y}}) &= \sum_{\underline{\mathbf{x}} \in g_1 \times \dots \times g_n} p_{\underline{\mathbf{x}}} \frac{1}{\sqrt{(2\pi)^n |\Sigma_{\underline{\mathbf{x}}}|}} e^{-\frac{1}{2}(\underline{\mathbf{y}} - \underline{\mu}_{\underline{\mathbf{x}}})^\top \Sigma_{\underline{\mathbf{x}}}^{-1} (\underline{\mathbf{y}} - \underline{\mu}_{\underline{\mathbf{x}}})} . \\
 \text{In particular } f_{Y_i(g)}(y) &= \sum_{x \in g} p_x \frac{1}{\sqrt{2\pi \text{Var} [Y_i(x)]}} e^{-\frac{(y - \mu_x)^2}{2 \text{Var} [Y_i(x)]}} .
 \end{aligned}$$

□

Remark 12. $\underline{\mathbf{Y}}$ is therefore a mixture of Gaussian vectors. It does not follow a Gaussian distribution in general. It follows a Gaussian distribution in specific cases such as: each grain is restricted to a single point; \mathbf{Y} is a white noise; any situation where all exponential terms are equal.

On Figure 5, we compare likelihood of a 2 dimensional observation of mixtures and the theoretical density of a bivariate Gaussian vector having same covariance matrix.

We observe that the mixture’s likelihood is not elliptic, showing a cross shape along the eigen vectors of the covariance matrix. We compare it with the more compact elliptic levels of a Gaussian vector. In Figure 5, granularity has 2 grains $g_1 = \{x_{11}, x_{12}\}$ and $g_2 = x_2$. The law of X_1 is $P(X_1 = x_{11}) = 0.475, P(X_1 = x_{12}) = 0.525$. The output is 1-dimensional and we observe $(Y_1(g_1), Y_2(g_2))$.

The complete covariance matrix is:

$$\begin{pmatrix} & Y_1(x_{11}) & Y_1(x_{12}) & Y_1(x_2) \\ Y_1(x_{11}) & 0.011 & 0.011 & -0.068 \\ Y_1(x_{12}) & 0.011 & 0.473 & -0.043 \\ Y_1(x_2) & -0.068 & -0.043 & 0.817 \end{pmatrix}.$$

Hence the covariance matrix of the observation, used to compute the likelihood of the bivariate Gaussian vector is $\begin{pmatrix} 0.253 & -0.055 \\ -0.055 & 0.817 \end{pmatrix}$.

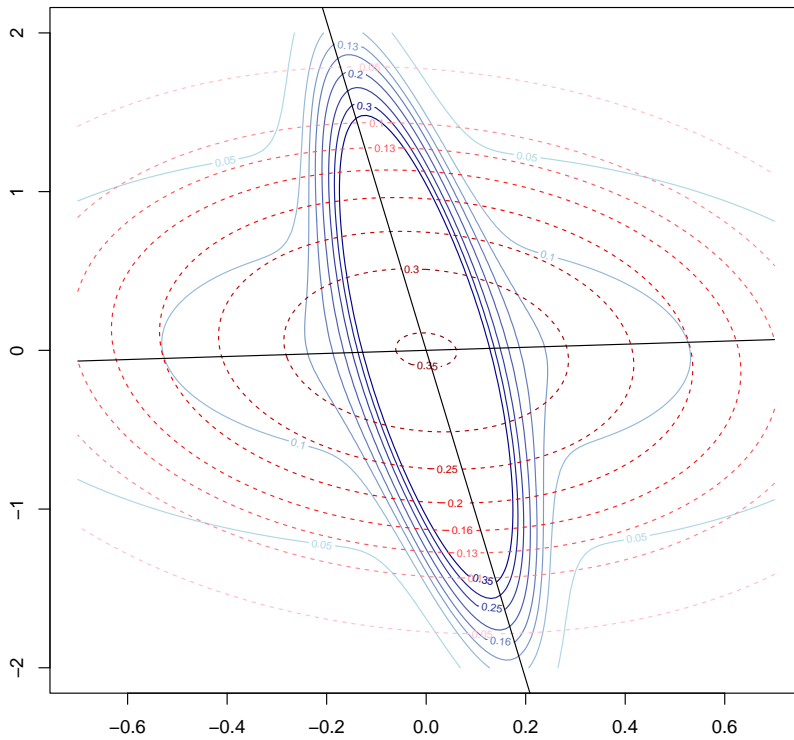


Figure 5 – Comparing levelplots of likelihood between a bivariate mixture and a bivariate Gaussian vector with same covariance matrix. Solid lines represent the mixture’s likelihood, dashed lines the Gaussian likelihood. Straight line represent the eigen directions of the covariance matrix. Levels are 0.05, 0.10, 0.13, 0.16, 0.20, 0.25, 0.30, 0.35.

Given two sets of observations denoted $\underline{\mathbf{Y}}$ and $\underline{\mathbf{Y}}^*$, we introduce the following no-

tations: We denote: n^* the length of $\underline{\mathbf{Y}}^*$, $g_{i_1}^*, \dots, g_{i_{n^*}}^*$ the grains associated to the observation $\underline{\mathbf{Y}}^*$. $\underline{\mathbf{xx}}^*$ is a vector composed of a vector of points drawn from the grains of $\underline{\mathbf{Y}}$ to which is appended a vector of points drawn from the grains of $\underline{\mathbf{Y}}^*$ and $p_{\underline{\mathbf{xx}}^*} = P(\underline{\mathbf{X}} = \underline{\mathbf{x}} \cap \underline{\mathbf{X}}^* = \underline{\mathbf{x}}^*)$. We define similarly $\underline{\mathbf{yy}}^*$ built appending a vector of values of $\underline{\mathbf{Y}}$ and $\underline{\mathbf{Y}}^*$. Their associated covariance matrices and expected values are denoted $\Sigma_{\underline{\mathbf{xx}}^*}$ and $\underline{\boldsymbol{\mu}}_{\underline{\mathbf{xx}}^*}$.

Proposition 24 (Gaussian case: Likelihood of conditional observations). *With the above framework and notations, assuming that \mathbf{Y} is a p -variate Gaussian field over the points of χ , The likelihood of $\underline{\mathbf{Y}}$ conditionally to $\underline{\mathbf{Y}}^*$ is:*

$$f_{\underline{\mathbf{Y}}|\underline{\mathbf{Y}}^*}(\underline{\mathbf{y}}, \underline{\mathbf{y}}^*) = \frac{\sum_{\substack{\underline{\mathbf{x}} \in g_1 \times \dots \times g_n \\ \underline{\mathbf{x}}^* \in g_1^* \times \dots \times g_{n^*}^*}} p_{\underline{\mathbf{xx}}^*} \frac{\exp\left(-\frac{1}{2}(\underline{\mathbf{yy}}^* - \underline{\boldsymbol{\mu}}_{\underline{\mathbf{xx}}^*})^\top \Sigma_{\underline{\mathbf{xx}}^*}^{-1}(\underline{\mathbf{yy}}^* - \underline{\boldsymbol{\mu}}_{\underline{\mathbf{xx}}^*})\right)}{\sqrt{(2\pi)^{n+n^*} |\Sigma_{\underline{\mathbf{xx}}^*}|}}}{\sum_{\underline{\mathbf{x}}^* \in g_1^* \times \dots \times g_{n^*}^*} p_{\underline{\mathbf{x}}^*} \frac{\exp\left(-\frac{1}{2}(\underline{\mathbf{y}}^* - \underline{\boldsymbol{\mu}}_{\underline{\mathbf{x}}^*})^\top \Sigma_{\underline{\mathbf{x}}^*}^{-1}(\underline{\mathbf{y}}^* - \underline{\boldsymbol{\mu}}_{\underline{\mathbf{x}}^*})\right)}{\sqrt{(2\pi)^{n^*} |\Sigma_{\underline{\mathbf{x}}^*}|}}} \quad (\text{IV.13})$$

Sketch of the proof. For given events A and B , we want to compute $P(\underline{\mathbf{Y}} \in A | \underline{\mathbf{Y}}^* \in B)$. We use the conditional probability formula:

$$P(\underline{\mathbf{Y}} \in A | \underline{\mathbf{Y}}^* \in B) = \frac{P(\underline{\mathbf{Y}} \in A \cap \underline{\mathbf{Y}}^* \in B)}{P(\underline{\mathbf{Y}}^* \in B)}$$

Keeping in mind that subvectors of Gaussian vectors are also Gaussian, we have:

$$f_{\underline{\mathbf{Y}}|\underline{\mathbf{Y}}^*}(\underline{\mathbf{y}}, \underline{\mathbf{y}}^*) = \frac{\sum_{\substack{\underline{\mathbf{x}} \in g_1 \times \dots \times g_n \\ \underline{\mathbf{x}}^* \in g_1^* \times \dots \times g_{n^*}^*}} p_{\underline{\mathbf{xx}}^*} \frac{\exp\left(-\frac{1}{2}(\underline{\mathbf{yy}}^* - \underline{\boldsymbol{\mu}}_{\underline{\mathbf{xx}}^*})^\top \Sigma_{\underline{\mathbf{xx}}^*}^{-1}(\underline{\mathbf{yy}}^* - \underline{\boldsymbol{\mu}}_{\underline{\mathbf{xx}}^*})\right)}{\sqrt{(2\pi)^{n+n^*} |\Sigma_{\underline{\mathbf{xx}}^*}|}}}{\sum_{\underline{\mathbf{x}}^* \in g_1^* \times \dots \times g_{n^*}^*} p_{\underline{\mathbf{x}}^*} \frac{\exp\left(-\frac{1}{2}(\underline{\mathbf{y}}^* - \underline{\boldsymbol{\mu}}_{\underline{\mathbf{x}}^*})^\top \Sigma_{\underline{\mathbf{x}}^*}^{-1}(\underline{\mathbf{y}}^* - \underline{\boldsymbol{\mu}}_{\underline{\mathbf{x}}^*})\right)}{\sqrt{(2\pi)^{n^*} |\Sigma_{\underline{\mathbf{x}}^*}|}}}$$

In particular:

$$f_{Y_i(g)|\underline{\mathbf{Y}}^*}(y, \underline{\mathbf{y}}^*) = \frac{\sum_{\substack{x \in g \\ \underline{\mathbf{x}}^* \in g_1^* \times \dots \times g_{n^*}^*}} p_x p_{\underline{\mathbf{x}}^*} \frac{\exp\left(-\frac{1}{2}(y \underline{\mathbf{y}}^* - \underline{\boldsymbol{\mu}}_{x \underline{\mathbf{x}}^*})^\top \Sigma_{x \underline{\mathbf{x}}^*}^{-1}(y \underline{\mathbf{y}}^* - \underline{\boldsymbol{\mu}}_{x \underline{\mathbf{x}}^*})\right)}{\sqrt{(2\pi)^{1+n^*} |\Sigma_{x \underline{\mathbf{x}}^*}|}}}{\sum_{\underline{\mathbf{x}}^* \in g_1^* \times \dots \times g_{n^*}^*} p_{\underline{\mathbf{x}}^*} \frac{\exp\left(-\frac{1}{2}(\underline{\mathbf{y}}^* - \underline{\boldsymbol{\mu}}_{\underline{\mathbf{x}}^*})^\top \Sigma_{\underline{\mathbf{x}}^*}^{-1}(\underline{\mathbf{y}}^* - \underline{\boldsymbol{\mu}}_{\underline{\mathbf{x}}^*})\right)}{\sqrt{(2\pi)^{n^*} |\Sigma_{\underline{\mathbf{x}}^*}|}}}$$

□

9.2 Functions of Mixtures

Proposition 25 (Functions of mixtures). *A function of a multidimensional mixture variable is a mixture of the images of point multidimensional variables by this same function.*

Proof. Let us consider an observation $\underline{\mathbf{Y}}$, it is a mixture of multiple $\underline{\mathbf{Y}}(\underline{\mathbf{x}})$ variables. We also consider a function f from \mathbb{R}^q to \mathbb{R}^n . For any event B :

$$P(f(\underline{\mathbf{Y}}) \in B) = \mathbb{E}[P(f(\underline{\mathbf{Y}}) \in B \mid \underline{\mathbf{X}})]$$

Therefore, $f(\underline{\mathbf{Y}})$ is a mixture distribution of the different $f(\underline{\mathbf{Y}}(\underline{\mathbf{x}}))$ for $\underline{\mathbf{x}} \in g_{i_1} \times \dots \times g_{i_n}$. \square

Proposition 26 (Linear combination of mixture variables). *A linear combination of mixture variables is a mixture of the same linear combination of point variables.*

Proof. It is a particular case of Proposition 25 :

We consider a real matrix $\mathbf{A} = (a_{i,j})_{\substack{1 \leq i \leq q \\ 1 \leq j \leq n}} \in \mathcal{M}_{q,n}(\mathbb{R})$.

For any event B :

$$P(\mathbf{A}\underline{\mathbf{Y}} \in B) = \mathbb{E}[P(\mathbf{A}\underline{\mathbf{Y}} \in B \mid \underline{\mathbf{X}})]$$

Therefore, $\mathbf{A}\underline{\mathbf{Y}}$ is a mixture of the different $\mathbf{A}\underline{\mathbf{Y}}(\underline{\mathbf{x}})$ for $\underline{\mathbf{x}} \in g_{i_1} \times \dots \times g_{i_n}$. \square

Proposition 27 (Linear combination of Gaussian mixture variables). *In the case where $\underline{\mathbf{Y}}$ is a Gaussian vector, we have:*

- (i) $\mathbf{A}\underline{\mathbf{Y}}$ is a mixture of Gaussian vectors and its density can be explicitly computed.
- (ii) The simple Kriging predictor (best linear unbiased predictor) of mixtures of a Gaussian field yields a mixture of Gaussian vectors.

Proof. (i) $\mathbf{A}\underline{\mathbf{Y}}(\underline{\mathbf{x}})$ is a vector of linear combinations of components of a Gaussian field. It is therefore a Gaussian vector. Therefore, $\mathbf{A}\underline{\mathbf{Y}}$ is a mixture of Gaussian vectors. Because of linearity of expectation, the expectation of $\mathbf{A}\underline{\mathbf{Y}}_{\underline{\mathbf{x}}}$ is $\mathbf{A}\mu_{\underline{\mathbf{x}}}$. Because of bilinearity of covariance, the covariance matrix of $\mathbf{A}\underline{\mathbf{Y}}_{\underline{\mathbf{x}}}$ is $\mathbf{A}\Sigma_{\underline{\mathbf{x}}}\mathbf{A}^\top$. Replacing the corresponding variables in Equation (IV.12), we get:

$$f_{\mathbf{A}\underline{\mathbf{Y}}}(\mathbf{z}) = \sum_{\underline{\mathbf{x}} \in g_1 \times \dots \times g_n} p_{\underline{\mathbf{x}}} \frac{1}{\sqrt{(2\pi)^q |\mathbf{A}\Sigma_{\underline{\mathbf{x}}}\mathbf{A}^\top|}} e^{-\frac{1}{2}(\mathbf{z} - \mathbf{A}\mu_{\underline{\mathbf{x}}})^\top (\mathbf{A}\Sigma_{\underline{\mathbf{x}}}\mathbf{A}^\top)^{-1} (\mathbf{z} - \mathbf{A}\mu_{\underline{\mathbf{x}}})} \quad (\text{IV.14})$$

This is a mixture of Gaussian distributions. Additionnally, in case $q = n$ and \mathbf{A} is invertible, this formula can be simplified. Denoting $\mathbf{y} = \mathbf{A}^{-1}\mathbf{z}$, we can write:

$$\begin{aligned} f_{\mathbf{A}\underline{\mathbf{Y}}}(\mathbf{z}) &= f_{\mathbf{A}\underline{\mathbf{Y}}}(\mathbf{A}\underline{\mathbf{y}}) = \sum_{\underline{\mathbf{x}} \in g_1 \times \dots \times g_n} p_{\underline{\mathbf{x}}} \frac{1}{\sqrt{(2\pi)^n |\mathbf{A}|^2 |\Sigma_{\underline{\mathbf{x}}|}} e^{-\frac{1}{2}(\mathbf{A}\underline{\mathbf{y}} - \mathbf{A}\mu_{\underline{\mathbf{x}}})^\top (\mathbf{A}^\top)^{-1} \Sigma_{\underline{\mathbf{x}}}^{-1} \mathbf{A}^{-1} (\mathbf{A}\underline{\mathbf{y}} - \mathbf{A}\mu_{\underline{\mathbf{x}}})} \\ &= \sum_{\underline{\mathbf{x}} \in g_1 \times \dots \times g_n} p_{\underline{\mathbf{x}}} \frac{1}{\sqrt{(2\pi)^n |\mathbf{A}|^2 |\Sigma_{\underline{\mathbf{x}}|}} e^{-\frac{1}{2}(\underline{\mathbf{y}} - \mu_{\underline{\mathbf{x}}})^\top \Sigma_{\underline{\mathbf{x}}}^{-1} (\underline{\mathbf{y}} - \mu_{\underline{\mathbf{x}}})} \\ f_{\mathbf{A}\underline{\mathbf{Y}}}(\mathbf{z}) &= \frac{1}{\sqrt{|\mathbf{A}|^2}} f_{\underline{\mathbf{Y}}}(\mathbf{A}^{-1}\mathbf{z}) \end{aligned}$$

- (ii) The simple kriging predictor is linear, therefore item (ii) is a particular case of item (i). □

Proposition 28 (Leave one-out cross validation for simple Kriging). *For simple Kriging, given a set of observations $\underline{\mathbf{Y}}$, the leave one out cross-validation predictor and its associated cross-validation error are both linear transformations of $\underline{\mathbf{Y}}$.*

Proof. Remember:

$$\mathbf{K} = \left(\text{Cov} \left[Y^j, Y^{j'} \right] \right)_{\substack{1 \leq j \leq n \\ 1 \leq j' \leq n}}$$

$$\mathbf{h}_i(g) = \left(\text{Cov} \left[Y_i(g), Y^j \right] \right)_{1 \leq j \leq n}$$

For simple mixture Kriging, the general formula for predictor of $Y_i(g)$ is:

$$M_i(g) = \boldsymbol{\alpha}_i(g)^\top \underline{\mathbf{Y}}$$

with

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

If we want to predict $Y^j = Y_{i_j}(g_j)$, we build a new observation vector:

$${}_j \underline{\mathbf{Y}} = \left(Y^1, \dots, Y^{j-1}, Y^{j+1}, \dots, Y^n \right)^\top$$

and its associated matrices:

$${}_j \mathbf{K} = \left(\text{Cov} \left[Y^k, Y^{j'} \right] \right)_{\substack{1 \leq k \leq n \text{ \& } k \neq j \\ 1 \leq j' \leq n \text{ \& } j' \neq j}}$$

$${}_j \mathbf{h} = \left(\text{Cov} \left[Y^j, Y^{j'} \right] \right)_{1 \leq j' \leq n \text{ \& } j' \neq j}$$

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

$${}_j M = {}_j \boldsymbol{\alpha}^\top {}_j \mathbf{K}$$

${}_j\mathbf{K}$ and ${}_j\mathbf{h}$ are extracted from \mathbf{K} with the following linear transformation:

Denote $\mathbf{I}_k =$ identity matrix of rank k ,

denote $\mathbf{J}_{k,l} =$ matrix of zeros of dimension $k \times l$,

denote $\mathbf{L}_j = \begin{pmatrix} \mathbf{I}_{j-1} & \mathbf{J}_{j-1,1} & \mathbf{J}_{j-1,n-j} \\ \mathbf{J}_{n-j,j-1} & \mathbf{J}_{n-j,1} & \mathbf{I}_{n-j} \end{pmatrix} \in \mathcal{M}_{n-1,n}(\mathbb{R})$,

then ${}_j\mathbf{Y} = \mathbf{L}_j\mathbf{Y}$,

and ${}_j\mathbf{K} = \mathbf{L}_j\mathbf{K}\mathbf{L}_j^\top$ (invertible).

Denote $\mathbf{l}_j = \begin{pmatrix} \mathbf{J}_{j-1,1} \\ 1 \\ \mathbf{J}_{n-j,1} \end{pmatrix} \in \mathbb{R}^n$ ($\mathbf{L}_j\mathbf{l}_j = \mathbf{0}_{n-1}$, $\ker \mathbf{L}_j = \mathbb{R}\mathbf{l}_j$),

then ${}_j\mathbf{h} = \mathbf{L}_j\mathbf{K}\mathbf{l}_j$,

and therefore ${}_j\boldsymbol{\alpha} = (\mathbf{L}_j^\top\mathbf{K}\mathbf{L}_j)^{-1}\mathbf{L}_j\mathbf{K}\mathbf{l}_j$.

and ${}_jM = \left((\mathbf{L}_j^\top\mathbf{K}\mathbf{L}_j)^{-1}\mathbf{L}_j\mathbf{K}\mathbf{l}_j \right)^\top \mathbf{L}_j\mathbf{Y} = \mathbf{N}_j\mathbf{Y}$.

Therefore, the predictor is linear and we can apply the Corollary 27, therefore it is a mixture of Gaussian variables and we can compute the density of ${}_jM$. \mathbf{N}_j is a row vector of length n . We build a matrix \mathbf{N} which rows are $(\mathbf{N}_j)_{1 \leq j \leq n}$. That way we can build the complete cross validated simple Kriging predictor $\mathbf{M}_{cv} = \mathbf{N}\mathbf{Y}$ and we can compute its density with Equation (IV.14). Similarly, we can also compute the density of the cross validation error $\mathbf{M}_{cv} - \mathbf{Y} = (\mathbf{N} - \mathbf{I}_n)\mathbf{Y} = \mathbf{C}\mathbf{Y}$. \square

Proposition 29 (Gaussian case: Leave one-out cross validation error). *In the Gaussian case, for simple Kriging, given a set of observations \mathbf{Y} , the leave one out cross-validation predictor and its associated cross-validation error both are mixtures of Gaussian variables and their density can be explicitly computed:*

$$f_{\mathbf{C}\mathbf{Y}}(\mathbf{z}) = \sum_{\mathbf{x} \in g_1 \times \dots \times g_n} p_{\mathbf{x}} \frac{1}{\sqrt{(2\pi)^q |\mathbf{C}\Sigma_{\mathbf{x}}\mathbf{C}^\top|}} e^{-\frac{1}{2}(\mathbf{z} - \mathbf{C}\boldsymbol{\mu}_{\mathbf{x}})^\top (\mathbf{C}\Sigma_{\mathbf{x}}\mathbf{C}^\top)^{-1} (\mathbf{z} - \mathbf{C}\boldsymbol{\mu}_{\mathbf{x}})} \quad (\text{IV.15})$$

And if we assume that \mathbf{C} is invertible:

$$f_{\mathbf{C}\mathbf{Y}}(\mathbf{z}) = \frac{1}{\sqrt{|\mathbf{C}|^2}} f_{\mathbf{Y}}(\mathbf{C}^{-1}\mathbf{z})$$

Definition 22 (Pseudo-variable, normalised variable). *Given a real random variable X , given F_X its cumulative distribution function, the random variable $F_X(X)$ takes values in $[0, 1]$. We call it the pseudo-variable associated with X .*

Let F_0^{-1} be the quantile function of the standard normal distribution $\mathcal{N}(0, 1)$. The random variable $F_0^{-1}(F_X(X))$ takes values in \mathbb{R} . We call it the normalised variable associated with X .

Remark 13. *If f is a strictly increasing function on \mathbb{R} then X and $f(X)$ are associated with the same pseudo-variable and the same normalised variable.*

Definition 23 (Granularity output). *For a given granularity \mathcal{G} , we define $X_{\mathcal{G}}$ to be the value of x at a random point on a random grain G of \mathcal{G} . We assume that G is known for any granularity \mathcal{G} . As a consequence, we can define $Y_{\mathcal{G}} = Y(X_{\mathcal{G}})$.*

Definition 24 (Pseudo-input, pseudo-output). *For a given granularity \mathcal{G} , for a given coordinate $X_{\mathcal{G},i}$ of $X_{\mathcal{G}}$, we define U_i to be the pseudo-variable associated with $X_{\mathcal{G},i}$. Similarly for a coordinate $Y_{\mathcal{G},i}$, we define V_i to be the normalised variable associated with $Y_{\mathcal{G},i}$. We call them pseudo-input, pseudo-output, normalised input, and normalised output variables.*

Remark 14. *Given a granularity \mathcal{G} , for any point $x \in \chi$ it is now possible to associate not only its coordinates $(x_i)_{i \in \{1, \dots, d\}}$ but also its pseudo-coordinates $(u_i = F_{X_{\mathcal{G},i}}(x_i))_{i \in \{1, \dots, d\}}$ and its normalised-coordinates $(n_i = F_0^{-1}(u_i))_{i \in \{1, \dots, p\}}$. And for an output y , we can similarly define $(y_i)_{i \in \{1, \dots, p\}}$ but also its pseudo-coordinates $(v_i = F_{Y_{\mathcal{G},i}}(y_i))_{i \in \{1, \dots, p\}}$ and its normalised-coordinates $(n_i = F_0^{-1}(v_i))_{i \in \{1, \dots, p\}}$.*

Remark 15. *$X_{\mathcal{G}}$ is a p -dimensional random variable. $Y_{\mathcal{G}}$ is a d -dimensional random variable.*

9.3 Observations and Repeated Observations

Assume that a territory $\chi \subset \mathbb{R}$ comprises 3 points $\chi = \{x_1, x_2, x_3\}$.

We define on this territory a bivariate output variable $\mathbf{Y} = (Y_1, Y_2)^\top$ such that:

$$\forall x \in \chi, \boldsymbol{\mu}_x = (0, 0)^\top$$

$$\exists (\sigma_{11}, \sigma_{22}, \sigma_{12} = \sigma_{21}) \in \mathbb{R}_+^* \times \mathbb{R}_+^* \times \mathbb{R} : \forall x \in \chi, \forall (i, j) \in \{1, 2\}^2, \text{Cov}[Y_i(x), Y_j(x)] = \sigma_{i,j}$$

$$\exists (\theta_{11}, \theta_{22}, \theta_{12} = \theta_{21}) \in \mathbb{R}_+^* \times \mathbb{R}_+^* \times \mathbb{R}_+^* : \forall (x, y) \in \chi^2, \forall (i, j) \in \{1, 2\}^2, \text{Cov}[Y_i(x), Y_j(y)] = \sigma_{i,j} e^{-\frac{|x-y|}{2}}$$

We also define a granularity \mathcal{G} comprising 2 grains g_1 and g_2 such that $g_1 = \{x_1, x_2\}$ and $g_2 = \{x_3\}$. On those grains, we define independant random positions X_{g_1} and X_{g_2} . Note that X_{g_2} is constant equal to x_3 . We denote $p_{x_1} = P(X_{g_1} = x_1)$ and $p_{x_2} = P(X_{g_1} = x_2)$.

We can eventually define $\mathbf{Y}(g_1) = (Y_1(X_{g_1}), Y_2(X_{g_1}))^\top$ and $\mathbf{Y}(g_2) = (Y_1(x_3), Y_2(x_3))^\top$.

Let us now define $X_{g_1}^\perp$ an i.i.d. copy of X_{g_1} .

We observe

$$\underline{\mathbf{Y}} = \left(Y_1(X_{g_1}), Y_2(X_{g_1}), Y_1(X_{g_1}^\perp), Y_2(X_{g_1}^\perp), Y_1(X_{g_2}), Y_2(X_{g_2}) \right)^\top$$

The covariance matrix of $\underline{\mathbf{Y}}$ is:

$$\mathbf{K} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & c_{11} & c_{12} & d_{11} & d_{12} \\ \sigma_{12} & \sigma_{22} & c_{12} & c_{22} & d_{12} & d_{22} \\ c_{11} & c_{12} & \sigma_{11} & \sigma_{12} & d_{11} & d_{12} \\ c_{12} & c_{22} & \sigma_{12} & \sigma_{22} & d_{12} & d_{22} \\ d_{11} & d_{12} & d_{11} & d_{12} & \sigma_{11} & \sigma_{12} \\ d_{12} & d_{22} & d_{12} & d_{22} & \sigma_{12} & \sigma_{22} \end{pmatrix}$$

where:

$$\begin{aligned} c_{11} &= \sigma_{11} \left(p_{x_1}^2 + p_{x_2}^2 + 2p_{x_1}p_{x_2} \exp \left(-\frac{(x_1 - x_2)^2}{2\theta_{11}^2} \right) \right) \\ c_{22} &= \sigma_{22} \left(p_{x_1}^2 + p_{x_2}^2 + 2p_{x_1}p_{x_2} \exp \left(-\frac{(x_1 - x_2)^2}{2\theta_{22}^2} \right) \right) \\ c_{12} &= \sigma_{12} \left(p_{x_1}^2 + p_{x_2}^2 + 2p_{x_1}p_{x_2} \exp \left(-\frac{(x_1 - x_2)^2}{2\theta_{12}^2} \right) \right) \\ d_{11} &= \sigma_{11} \left(p_{x_1} \exp \left(-\frac{(x_1 - x_3)^2}{2\theta_{11}^2} \right) + p_{x_2} \exp \left(-\frac{(x_2 - x_3)^2}{2\theta_{11}^2} \right) \right) \\ d_{22} &= \sigma_{22} \left(p_{x_1} \exp \left(-\frac{(x_1 - x_3)^2}{2\theta_{22}^2} \right) + p_{x_2} \exp \left(-\frac{(x_2 - x_3)^2}{2\theta_{22}^2} \right) \right) \\ d_{12} &= \sigma_{12} \left(p_{x_1} \exp \left(-\frac{(x_1 - x_3)^2}{2\theta_{12}^2} \right) + p_{x_2} \exp \left(-\frac{(x_2 - x_3)^2}{2\theta_{12}^2} \right) \right) \end{aligned}$$

9.4 Modeling

We consider a territory $\chi = \mathbb{R}^3$. A point $x = (x_1, x_2, x_3)^\top \in \chi$ represents a square metre of a dwelling with its coordinate x_1 representing the date when the part of the dwelling containing x was built, coordinates x_2, x_3 represent the 2D geographical coordinates where the dwelling's square metre is located.

On this territory, we define a real bivariate random field $Y(x) = (Y_1(x), Y_2(x))$.

$Y_1(x)$ represents the energy consumption per year associated with x .

$Y_2(x)$ represents the median of the yearly income of the household living in the dwelling where x lies.

We assume that Y_1 and Y_2 are centred and have constant expectancy over the territory.

An **address** is a grain of χ containing all points located in dwellings associated with a unique postal address. The definition of an address depends only on x_2 and x_3 , therefore an address is a cylinder in χ . The set of all addresses form the ‘‘addresses granularity’’ \mathcal{G}_a . See Figure 6.

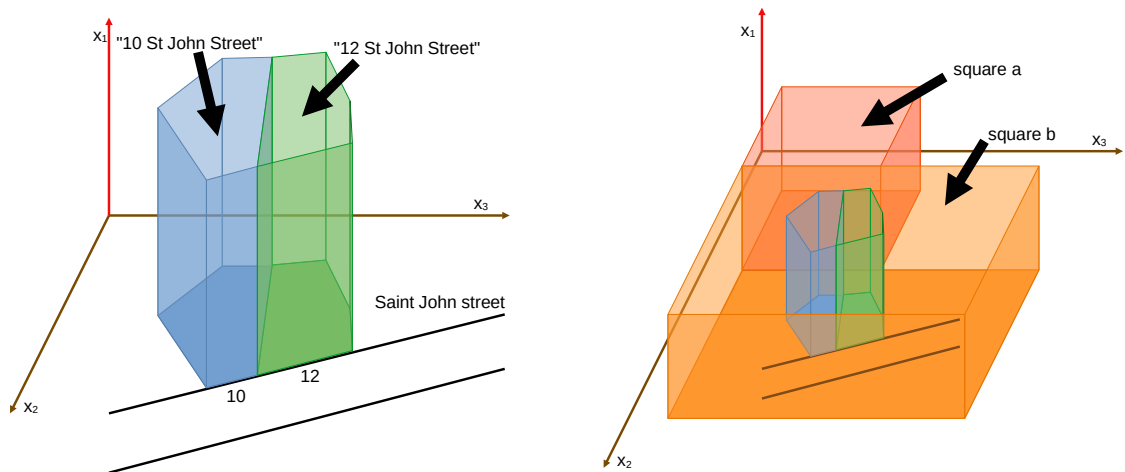


Figure 6 – Addresses granularity (left figure) Each grain is a cylinder in χ containing the geographical coordinates of dwellings located at a given postal address and all construction dates. **squared granularity** (right figure) Each grain is a cylinder in χ containing the geographical coordinates of dwellings located in a geographical square and all construction dates.

For anonymisation reasons, we also define grains that form cylinders with a square basis in χ . We simply call it a **square**. Squares are of different sizes. Together, those grains form the ‘‘squared granularity’’ \mathcal{G}_s . See Figure 6.

For $g \in \mathcal{G}_a$, we define a random variable X_g picking a random point in g with a uniform probability. Note that by construction, it is the same as a random variable that would pick a random dwelling with a probability proportionate to its surface area and then pick a random point in this dwelling with uniform probability.

For $g \in \mathcal{G}_s$, we define similarly a random variable X_g picking a random point in g with a uniform probability.

For any pair of grains $(g, g') \in (\mathcal{G}_a \cup \mathcal{G}_s)^2$, we set $X_g \perp X_{g'}$.

We define the output variables at grain levels to be: $\forall g \in \mathcal{G}_a \cup \mathcal{G}_s, Y_i(g) = Y_i(X_g)$.

The energy consumption Y_1 is observed on some grains of the addresses granularity.

The median yearly income Y_2 is observed on all grains of the squared granularity.

We assume that output variables Y_i and Y_j at points $x = (x_1, x_2, x_3)$ and $x' = (x'_1, x'_2, x'_3)$ are correlated according to a kernel that depends only on the norm $\|x - x'\|_{\theta_{i,j}} = \sum_{k=1}^3 \frac{(x_k - x'_k)^2}{(\theta_{i,j}^k)^2}$ where $\theta_{i,j} = (\theta_{i,j}^k)_{k \in \{1, \dots, 3\}}$ is a triplet of parameters to be found where $\theta_{i,j} = \theta_{j,i}$.

The uncertainty in the output variable is modelled with a centred noise of constant variance $\sigma_{\epsilon_i}^2$ called the nugget effect and denoted ϵ_x .

We assume that variances of Y_1 and Y_2 (without the nugget effect) are constant over the territory and denoted σ_1^2, σ_2^2 .

For instance, if we assume that the kernel is Gaussian (squared exponential), we have:

$$\forall (i, j) \in \{1, \dots, 2\}^2, \text{Cov}[Y_i(x), Y_j(x')] = \sigma_i \sigma_j e^{-\|x - x'\|_{\theta_{i,j}}} + \delta(x, x') \delta(i, j) \sigma_{\epsilon_i}^2$$

where $\delta(a, b) = 1$ if a, b' and 0 otherwise.

Therefore, the model has 13 parameters to be found:

$$(\sigma_1, \sigma_2, \theta_{1,1} = (\theta_{1,1}^k)_{k \in \{1, \dots, 3\}}, \theta_{1,2} = \theta_{2,1}, \theta_{2,2}, \sigma_{\epsilon_1}, \sigma_{\epsilon_2})$$

And the covariance between 2 grains output variables is:

$$k_{i,j}(g, g') = \text{Cov}[Y_i(g), Y_j(g')] = \begin{cases} \frac{1}{|g||g'|} \sum_{x \in g} \sum_{x' \in g'} \sigma_i \sigma_j e^{-\|x - x'\|_{\theta_{i,j}}} & \text{if } g \neq g' \text{ (i.e. } X_g \perp X_{g'}) \\ \sigma_i \sigma_j + \delta(i, j) \sigma_{\epsilon_i}^2 & \text{if } g = g' \text{ (i.e. } X_g = X_{g'}) \end{cases}$$

Note that as a result, whenever the nugget effect is not null, the observations covariance matrix is strictly dominated by its diagonal, therefore it is invertible.

9.5 Data Processing

9.5.a Points/Grains/Random Position

For each dwelling, we know:

- its construction year;
- its postal address;
- the geographical perimeter of the land plot where the dwellings having the same postal address are located.

We assume that the square metres forming the dwelling are distributed uniformly on the land plot of the address. For the sake of computation simplification we only generate 1 point for 10 square metres in the dwelling.

For computation, an address is the union of all points generated by all dwellings that are located at a same address.

Theoretically, a square should be the union of all points generated by all dwellings that are located on it perimeter. But that would make it impossible to compute correlations between squares in a reasonable time. Therefore, we remove randomly 90% of the points which leaves us still with thousands of points in each square that are distributed according to the dwellings surface density.

Note that with this definition, some addresses may intersect 1 to 4 squares. For simplification, we associate points in an address with the square that has the largest geographical intersection with the address.

9.5.b Observations

The energy efficiency Y_1 is observed for some grains of \mathcal{G}_a . If a same grain is observed several times (multiple diagnostics performed at the same address), we keep only the last observation which is assumed to be more accurate.

The median yearly income Y_2 is observed for some grains of \mathcal{G}_s .

9.6 Computing the Model

Our goal is to be able to predict $Y_1(g)$ for any $g \in \mathcal{G}_a$.

We define:

- a learning set with 75% of Y_1 observations and all Y_2 observations;
- a validation set with 25% of Y_1 observations.

Since the value of Y_1 on a grain is observed only once in our dataset, the set of grains observed in the learning set is disjoint from the set of grains observed in the validation set.

The optimisation is implemented stepwise:

Estimation of $(\sigma_1, \theta_{1,1}, \sigma_{\epsilon_1})$:

- We build a model with observations made on \mathcal{G}_a that is with only Y_1 as output variable and using the projection of the territory on the first axis. We denote $g_{|1}$ the projection of g on the first axis:

$$k_{1,1}(g_{|1}, g'_{|1}) = \text{Cov}[Y_1(g), Y_1(g')] = \begin{cases} \frac{1}{|g||g'|} \sum_{\substack{x \in g \\ x' \in g'}} \sigma_1^2 e^{-\|x_1 - x'_1\|_{\theta_{1,1}^1}} & \text{if } g \neq g' \text{ (i.e. } X_g \perp X_{g'}) \\ \sigma_1^2 + \delta(i, j) \sigma_{\epsilon_1}^2 & \text{if } g = g' \text{ (i.e. } X_g = X_{g'}) \end{cases}$$

This model has 3 parameters $(\sigma_1, \theta_{1,1}^1, \sigma_{\epsilon_1})$. We compute the cross validated error of this model on a 3 dimensional grid of parameters and keep the best triplet.

- Once this is done, we keep the estimated values of σ_1 and σ_{ϵ_1} and use them to estimate $\theta_{1,1}^2$ on a model predicting Y_1 with $g_{|2}$. Similarly we estimate $\theta_{1,1}^3$

We repeat the same process with 3 models predicting Y_2 to estimate $(\sigma_2, \theta_{2,2}, \sigma_{\epsilon_2})$.

Eventually, we keep these previous estimates and build 3 models to estimate $\theta_{1,2}$.

With all these estimated parameters, we predict on the validation set to assess the performance of the model.

Supplementary Material for Chapter III

10	Proofs	.246
10.1	Proposition 3 Simple Joint Kriging weights	.246
10.2	Proposition 4 Ordinary Joint Kriging weights	.246
10.3	Remark 2 Covariance matrices	.247
10.4	Proposition 5 Joint Kriging weights under a predicted values constraint	.247
10.5	Remark 3 Covariance matrices with two constraints	.252
10.6	Proposition 7 Joint Kriging variance with arbitrary weights	.252
10.7	Remark 5 Covariance matrices in Joint Kriging mean and variance	.253
10.8	Proposition 8 Variance sharing	.253
10.9	Remark 6 Constraints' impact	.255
10.10	Proof of Proposition 9 Nugget ensuring positive weights	.255
10.11	Procedure to Retrieve OpenML Results for the Quake Dataset	.257

10 Proofs

10.1 Proposition 3 Simple Joint Kriging weights

Proof of Proposition 3. The proof is very similar to the geo-statistical proof of Simple Kriging model. It does not rely on any Gaussian assumption, but just on existing moments of order two. Recall that \mathbf{W} is a symmetrical positive definite matrix, so that $\mathbf{W} = \mathbf{W}^\top$. Let us calculate the gradient of $\Delta(x^*)$ with respect to $\boldsymbol{\alpha}(x^*)$:

$$\begin{aligned} & \nabla_{\boldsymbol{\alpha}(x^*)} \mathbb{E} \left[\|\mathbf{M}(x^*) - \mathbf{Y}(x^*)\|_{\mathbf{W}}^2 \right] \\ &= \nabla_{\boldsymbol{\alpha}(x^*)} \mathbb{E} \left[(\mathbf{M}(x^*) - \mathbf{Y}(x^*))^\top \mathbf{W} (\mathbf{M}(x^*) - \mathbf{Y}(x^*)) \right] \\ &= \nabla_{\boldsymbol{\alpha}(x^*)} \mathbb{E} \left[(\mathbf{Y}\boldsymbol{\alpha}(x^*) - \mathbf{Y}(x^*))^\top \mathbf{W} (\mathbf{Y}\boldsymbol{\alpha}(x^*) - \mathbf{Y}(x^*)) \right] \\ &= \nabla_{\boldsymbol{\alpha}(x^*)} \mathbb{E} \left[\boldsymbol{\alpha}(x^*)^\top \mathbf{Y}^\top \mathbf{W} \mathbf{Y} \boldsymbol{\alpha}(x^*) - 2\boldsymbol{\alpha}(x^*)^\top \mathbf{Y}^\top \mathbf{W} \mathbf{Y}(x^*) + \mathbf{Y}(x^*)^\top \mathbf{W} \mathbf{Y}(x^*) \right] \\ &= 2 \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y} \right] \boldsymbol{\alpha}(x^*) - 2 \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y}(x^*) \right]. \end{aligned}$$

Thus,

$$\nabla_{\boldsymbol{\alpha}(x^*)} \Delta(x^*) = 2\mathbf{K}\boldsymbol{\alpha}(x^*) - 2\mathbf{h}(x^*). \quad (\text{IV.16})$$

Where $\mathbf{K} := \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y} \right]$ is a $n \times n$ matrix and $\mathbf{h}(x^*) := \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y}(x^*) \right]$ is a $n \times 1$ vector, thus leading to a $n \times 1$ gradient. Hence $\boldsymbol{\alpha}(x^*) = \mathbf{K}^{-1}\mathbf{h}(x^*)$ in Equation (III.4) when the gradient is zero. The matrix expression $\mathbf{A} = \mathbf{K}^{-1}\mathbf{H}$ of Equation (III.5) is obtained by binding column vectors of Equation (III.4), for all prediction locations. Remark that, under assumption that $\mathbb{E}[\mathbf{Y}(x)] = \mathbf{0}_p$ for all $x \in \chi$, it is clear that $\mathbb{E}[\mathbf{M}(x^*)] = \mathbb{E}[\mathbf{Y}(x^*)] = \mathbf{0}$, so that the predictor is unbiased.

In that case, the (i, j) component of the matrix $\mathbf{Y}^\top \mathbf{Y}$ is

$$\left(\mathbb{E} \left[\mathbf{Y}^\top \mathbf{Y} \right] \right)_{ij} = \sum_{k=1}^n \mathbb{E} [Y_i(x_k) Y_j(x_k)] = \sum_{k=1}^n \text{Cov} [Y_i(x_k), Y_j(x_k)].$$

Hence $\mathbf{Y}^\top \mathbf{Y}$ is a symmetric positive semi-definite matrix. The same holds for \mathbf{K} : writing $\mathbf{K} = \left(\mathbf{W}^{1/2} \mathbf{Y} \right)^\top \left(\mathbf{W}^{1/2} \mathbf{Y} \right)$, it is clear that for any vector \mathbf{v} , $\mathbf{v}^\top \mathbf{K} \mathbf{v} = \tilde{\mathbf{v}}^\top \tilde{\mathbf{v}} \geq 0$, where the vector $\tilde{\mathbf{v}} := \mathbf{W}^{1/2} \mathbf{Y} \mathbf{v}$. Thus \mathbf{K} is a symmetric semi-definite positive matrix, i.e. a covariance matrix. \square

10.2 Proposition 4 Ordinary Joint Kriging weights

Proof of Proposition 4. Under the constraint (III.6), and using a Lagrange multiplier $\lambda \in \mathbb{R}$, the loss to minimise is

$$\Delta_1(x^*) := \Delta(x^*) - 2\lambda(x^*) \left(\boldsymbol{\alpha}(x^*)^\top \mathbf{1}_n - 1 \right)$$

Using Equation (IV.16), the gradient of $\Delta_1(x^*)$ with respect to $\boldsymbol{\alpha}(x^*)$ is

$$\nabla_{\boldsymbol{\alpha}(x^*)} \Delta_1(x^*) = 2 \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y} \right] \boldsymbol{\alpha}(x^*) - 2 \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y}(x^*) \right] - 2\lambda(x^*) \mathbf{1}_n \quad (\text{IV.17})$$

Setting this $\nabla_{\alpha(x^*)}\Delta_1(x^*)$ to be zero for all of its p components, we get

$$\mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y} \right] \boldsymbol{\alpha}(x^*) = \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y}(x^*) \right] + \lambda(x^*) \mathbf{1}_n.$$

and finally,

$$\mathbf{K} \boldsymbol{\alpha}(x^*) = \mathbf{h}(x^*) + \lambda(x^*) \mathbf{1}_n.$$

Once $\boldsymbol{\alpha}(x^*)$ is written as a function of $\lambda(x^*)$, one easily gets the value of $\lambda(x^*)$ by setting $\mathbf{1}_n^\top \boldsymbol{\alpha}(x^*) = 1$. Hence the result. Matrix expressions are obtained by binding column vectors for all x^* in $\{x_1^*, \dots, x_q^*\}$ \square

10.3 Remark 2 Covariance matrices

Proof of Remark 2. Recall that $\mathbf{K} = \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y} \right]$ and $\mathbf{h}(x^*) = \mathbb{E} \left[\mathbf{Y}^\top \mathbf{W} \mathbf{Y}(x^*) \right]$. Under the chosen mean assumption, both $\mathbb{E} \left[\mathbf{Y}(x^*) \right] = \boldsymbol{\mu}$ and $\mathbb{E} \left[\mathbf{Y} \right] = \boldsymbol{\mu} \mathbf{1}_n^\top$. Thus, under the given constraint $\boldsymbol{\alpha}(x^*)^\top \mathbf{1}_n = 1$, or when $\boldsymbol{\mu} = \mathbf{0}_p$,

$$\mathbb{E} \left[\mathbf{Y}^\top \right] \mathbf{W} \mathbb{E} \left[\mathbf{Y} \right] \boldsymbol{\alpha}(x^*) = \mathbb{E} \left[\mathbf{Y}^\top \right] \mathbf{W} \mathbb{E} \left[\mathbf{Y}(x^*) \right] = \mathbf{1}_n \boldsymbol{\mu}^\top \mathbf{W} \boldsymbol{\mu}.$$

Hence the gradient of $\Delta(x^*)$ in Equation (IV.16) also writes

$$\nabla_{\alpha(x^*)} \Delta(x^*) = 2\mathbf{K} \boldsymbol{\alpha}(x^*) - 2\mathbf{h}(x^*) = 2\widetilde{\mathbf{K}} \boldsymbol{\alpha}(x^*) - 2\widetilde{\mathbf{h}}(x^*).$$

As a consequence, the gradient of $\Delta_1(x^*)$ in Equation (IV.17) is unchanged when replacing both (\mathbf{K}, \mathbf{h}) by $(\widetilde{\mathbf{K}}, \widetilde{\mathbf{h}})$. Thus, one can freely replace both (\mathbf{K}, \mathbf{h}) by $(\widetilde{\mathbf{K}}, \widetilde{\mathbf{h}})$ in the rest of the proof of Proposition 4, without changing the result. \square

10.4 Proposition 5 Joint Kriging weights under a predicted values constraint

Let us first study the rank of the system of constraints (III.9):

$$\begin{cases} \mathbf{A}^\top \mathbf{1}_n = \mathbf{1}_q \\ \mathbf{Y} \mathbf{A} \boldsymbol{\pi} = \mathbf{m} \end{cases} \quad (\text{IV.18})$$

We denote:

$$a_{ij} := \alpha_i x_j^*$$

$$y_{ij} := Y_j(x_i)$$

$$\pi_i := \pi_{x_i^*}$$

m_i the i -th component of \mathbf{m}

The system of constraints rewrites:

$$\begin{cases} \forall j \in \{1, \dots, q\}, \sum_{i=1}^n a_{ij} = 1 \\ \forall k \in \{1, \dots, p\}, \sum_{j=1}^q \sum_{i=1}^n \pi_j y_{ki} a_{ij} \end{cases}$$

Which can be expressed in matrix form as:

$$\begin{pmatrix} 1 & \dots & 1 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & \dots & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & 1 & \dots & 1 \\ \pi_1 y_{11} & \dots & \pi_1 y_{1n} & \pi_2 y_{11} & \dots & \pi_2 y_{1n} & \dots & \pi_q y_{11} & \dots & \pi_q y_{1n} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \\ \pi_1 y_{p1} & \dots & \pi_1 y_{pn} & \pi_2 y_{p1} & \dots & \pi_2 y_{pn} & \dots & \pi_q y_{p1} & \dots & \pi_q y_{pn} \end{pmatrix} \begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \\ a_{12} \\ \vdots \\ a_{n2} \\ \vdots \\ a_{1q} \\ \vdots \\ a_{nq} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \\ m_1 \\ \vdots \\ m_p \end{pmatrix}$$

Now we have to study the matrix of constraints. The reader can recognise that the last p rows are an aggregation of q times the matrix \mathbb{Y} , each time multiplied by a different factor.

- It is clear that the first q rows are linearly independent.
- If the p last rows are linearly dependent then, taking into account that $\boldsymbol{\pi}$ is strictly positive, it means that there exists at least one vector $\boldsymbol{\omega} \in \mathbb{R}^p$ such that:

$$\mathbb{Y}^\top \boldsymbol{\omega} = \mathbf{0}_n = \mathbf{0}\mathbf{1}_n$$

- If the p last rows are linearly dependent with the q first one, it means that there exists at least one vector $\boldsymbol{\omega} \in \mathbb{R}^p$ and one scalar ω_0 such that:

$$\mathbb{Y}^\top \boldsymbol{\omega} = \omega_0 \mathbf{1}_n$$

Therefore, the system is not of full rank if and only if there exists $\boldsymbol{\omega}$ and ω_0 such that:

$$\mathbb{Y}^\top \boldsymbol{\omega} = \omega_0 \mathbf{1}_n$$

Depending on the situation, the matrix of constraints can be of rank ranging from $q + 1$ up to $q + p$. In the following, we are interested in two important cases, when the matrix of constraints is of full rank $q + p$ and when it is of rank $q + p - 1$. The second case is useful for fuzzy classification. Other cases are of no interest for our study although they could also be treated removing a sufficient number of constraints in the system. Note that a system of rank lower than $q + p - 1$ corresponds to the cases where \mathbb{Y} carries little information and a variable selection should be implemented.

Proof of Proposition 5 when the system of Equations (III.9) is of full rank $q + p$.

$$\Delta_2(x^*) := \Delta(x^*) - 2\lambda(x^*) \left(\boldsymbol{\alpha}(x^*)^\top \mathbf{1}_n - 1 \right) - 2\boldsymbol{\lambda}'^\top (\mathbb{Y}\mathbb{A}\boldsymbol{\pi} - \mathbf{m}),$$

where $\boldsymbol{\lambda}'$ is a $p \times 1$ vector of Lagrange multipliers. The gradient of the last term, with respect to $\boldsymbol{\alpha}(x^*)$ is

$$\begin{aligned}
 & \nabla_{\boldsymbol{\alpha}(x^*)} 2\boldsymbol{\lambda}'^\top (\mathbb{E}[\mathbf{M}(X^*) | \mathbb{Y}] - \mathbf{m}) \\
 &= \nabla_{\boldsymbol{\alpha}(x^*)} 2\boldsymbol{\lambda}'^\top (\mathbb{P}[X^* = x^*] \mathbb{E}[\mathbf{M}(x^*) | \mathbb{Y}] + \mathbb{P}[X^* \neq x^*] \mathbb{E}[\mathbf{M}(X^*) | X^* \neq x^*, \mathbb{Y}] - \mathbf{m}) \\
 &= \nabla_{\boldsymbol{\alpha}(x^*)} 2\boldsymbol{\lambda}'^\top (\mathbb{P}[X^* = x^*] \mathbb{E}[\mathbf{M}(x^*) | \mathbb{Y}] - \mathbf{m}) + 0 \\
 &= \nabla_{\boldsymbol{\alpha}(x^*)} 2\boldsymbol{\lambda}'^\top \mathbb{P}[X^* = x^*] \mathbb{E}[\mathbb{Y}\boldsymbol{\alpha}(x^*) - \mathbf{m} | \mathbb{Y}] \\
 &= \nabla_{\boldsymbol{\alpha}(x^*)} 2\boldsymbol{\lambda}'^\top (\pi_{x^*} \mathbb{Y}\boldsymbol{\alpha}(x^*) - \mathbf{m}) \\
 &= 2\pi_{x^*} \mathbb{Y}^\top \boldsymbol{\lambda}'
 \end{aligned}$$

Hence using the gradient of $\Delta(x^*)$ in Equation (IV.16), one gets

$$\nabla_{\boldsymbol{\alpha}(x^*)} \Delta_2(x^*) = 2\mathbf{K}\boldsymbol{\alpha}(x^*) - 2\mathbf{h}(x^*) - 2\lambda(x^*)\mathbf{1}_n - 2\pi_{x^*} \mathbb{Y}^\top \boldsymbol{\lambda}' \quad (\text{IV.19})$$

Setting the gradient to be equal to a $n \times 1$ vector of zeros, we get for all prediction locations $x^* \in \{x_1^*, \dots, x_q^*\}$

$$\begin{cases} \mathbf{K}\boldsymbol{\alpha}(x^*) = \mathbf{h}(x^*) + \lambda(x^*)\mathbf{1}_n + \pi_{x^*} \mathbb{Y}^\top \boldsymbol{\lambda}' \\ \mathbf{1}_n^\top \boldsymbol{\alpha}(x^*) = 1 \\ \mathbb{Y}\mathbb{A}\boldsymbol{\pi} = \mathbf{m} \end{cases}$$

As optimal weights are gathered in the $n \times q$ matrix $\mathbb{A} := [\boldsymbol{\alpha}(x_1^*), \dots, \boldsymbol{\alpha}(x_q^*)]$, if one defines the $n \times q$ matrix $\mathbb{H} := [\mathbf{h}(x_1^*), \dots, \mathbf{h}(x_q^*)]$, then the previous system can be written, by binding columns for all $x^* \in \{x_1^*, \dots, x_q^*\}$:

$$\begin{cases} \mathbf{K}\mathbb{A} = \mathbb{H} + \mathbf{1}_n \boldsymbol{\lambda}^\top + \mathbb{Y}^\top \boldsymbol{\lambda}' \boldsymbol{\pi}^\top \\ \mathbf{1}_n^\top \mathbb{A} = \mathbf{1}_q^\top \\ \mathbb{Y}\mathbb{A}\boldsymbol{\pi} = \mathbf{m} \end{cases} \quad (\text{IV.20})$$

with $q \times 1$ Lagrange multiplier $\boldsymbol{\lambda}$, and $p \times 1$ Lagrange multiplier $\boldsymbol{\lambda}'$.

If \mathbf{K} is invertible, then the first equation writes

$$\mathbb{A} = \mathbf{K}^{-1}\mathbb{H} + \mathbf{K}^{-1}\mathbf{1}_n \boldsymbol{\lambda}^\top + \mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}' \boldsymbol{\pi}^\top$$

Injecting this value of \mathbb{A} into the first constraint $\mathbf{1}_n^\top \mathbb{A} = \mathbf{1}_q^\top$, denoting $\gamma := \boldsymbol{\pi}^\top \boldsymbol{\pi} \in \mathbb{R}$ and $\delta := \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbf{1}_n \in \mathbb{R}$ one gets:

$$\begin{aligned}
 \mathbf{1}_q^\top &= \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H} + \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbf{1}_n \boldsymbol{\lambda}^\top + \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}' \boldsymbol{\pi}^\top \\
 \mathbf{1}_q^\top \boldsymbol{\pi} &= \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbf{1}_n \boldsymbol{\lambda}^\top \boldsymbol{\pi} + \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}' \boldsymbol{\pi}^\top \boldsymbol{\pi} \\
 1 &= \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \delta \boldsymbol{\lambda}^\top \boldsymbol{\pi} + \gamma \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}' \\
 \delta \boldsymbol{\lambda}^\top \boldsymbol{\pi} &= 1 - \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} - \gamma \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}'
 \end{aligned}$$

Now injecting the value of \mathbb{A} into the second constraint $\mathbb{Y}\mathbb{A}\boldsymbol{\pi} = \mathbf{m}$, and using the last equation, denoting the $p \times 1$ vector $\mathbf{u} := \mathbb{Y}\mathbf{K}^{-1}\mathbf{1}_n$, one gets

$$\begin{aligned} \mathbf{m} &= \mathbb{Y}\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \mathbb{Y}\mathbf{K}^{-1}\mathbf{1}_n\boldsymbol{\lambda}^\top\boldsymbol{\pi} + \mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top\boldsymbol{\lambda}'\boldsymbol{\pi}^\top\boldsymbol{\pi} \\ &= \mathbb{Y}\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \mathbb{Y}\mathbf{K}^{-1}\mathbf{1}_n\frac{1}{\delta}\left(1 - \mathbf{1}_n^\top\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} - \gamma\mathbf{1}_n^\top\mathbf{K}^{-1}\mathbb{Y}^\top\boldsymbol{\lambda}'\right) + \gamma\mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top\boldsymbol{\lambda}' \\ &= \mathbb{Y}\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \frac{1}{\delta}\mathbf{u}\left(1 - \mathbf{1}_n^\top\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi}\right) - \gamma\frac{1}{\delta}\mathbf{u}\mathbf{u}^\top\boldsymbol{\lambda}' + \gamma\mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top\boldsymbol{\lambda}' \end{aligned}$$

and finally, the vector $\boldsymbol{\lambda}'$ must satisfies

$$\gamma\left(\frac{1}{\delta}\mathbf{u}\mathbf{u}^\top - \mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top\right)\boldsymbol{\lambda}' = \mathbb{Y}\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \frac{1}{\delta}\mathbf{u}\left(1 - \mathbf{1}_n^\top\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi}\right) - \mathbf{m}.$$

Hence, provided the matrix factor is invertible,

$$\boldsymbol{\lambda}' = \gamma^{-1}\left(\frac{1}{\delta}\mathbf{u}\mathbf{u}^\top - \mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top\right)^{-1}\left(\mathbb{Y}\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \frac{1}{\delta}\mathbf{u}\left(1 - \mathbf{1}_n^\top\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi}\right) - \mathbf{m}\right)$$

Once $\boldsymbol{\lambda}'$ computed, one gets for $\boldsymbol{\lambda}$

$$\begin{aligned} \mathbf{1}_q^\top &= \mathbf{1}_n^\top\mathbf{K}^{-1}\mathbb{H} + \delta\boldsymbol{\lambda}^\top + \mathbf{1}_n^\top\mathbf{K}^{-1}\mathbb{Y}^\top\boldsymbol{\lambda}'\boldsymbol{\pi}^\top \\ \delta\boldsymbol{\lambda}^\top &= -\mathbf{1}_n^\top\mathbf{K}^{-1}\mathbb{H} - \mathbf{1}_n^\top\mathbf{K}^{-1}\mathbb{Y}^\top\boldsymbol{\lambda}'\boldsymbol{\pi}^\top + \mathbf{1}_q^\top \end{aligned}$$

And finally, using $\mathbf{u} = \mathbb{Y}\mathbf{K}^{-1}\mathbf{1}_n$,

$$\boldsymbol{\lambda} = \delta^{-1}\left(\mathbf{1}_q - \mathbb{H}^\top\mathbf{K}^{-1}\mathbf{1}_n - \boldsymbol{\pi}\boldsymbol{\lambda}'^\top\mathbf{u}\right)$$

□

The above proof of Proposition 5 in the case where all constraints are independent relies on the invertibility of the $p \times p$ matrix $\mathbb{S} = \frac{1}{\delta}\mathbf{u}\mathbf{u}^\top - \mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top$. Let us now discuss this condition.

Let us assume take a vector $\boldsymbol{\omega}$, not null, such that $\mathbb{S}\boldsymbol{\omega} = \mathbf{0}_p$. It implies that $\boldsymbol{\omega}^\top\mathbb{S}\boldsymbol{\omega} = 0$. And rewriting \mathbb{S} , we get:

$$0 = \boldsymbol{\omega}^\top\left(\frac{1}{\mathbf{1}_n^\top\mathbf{K}^{-1}\mathbf{1}_n}\mathbb{Y}\mathbf{K}^{-1}\mathbf{1}_n\mathbf{1}_n^\top\mathbf{K}^{-1}\mathbb{Y}^\top - \mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top\right)\boldsymbol{\omega}$$

We denote $\mathbf{w}_n = \mathbb{Y}^\top\boldsymbol{\omega}$ and get:

$$\frac{\left(\mathbf{w}_n^\top\mathbf{K}^{-1}\mathbf{1}_n\right)\left(\mathbf{1}_n^\top\mathbf{K}^{-1}\mathbf{w}_n\right)}{\mathbf{1}_n^\top\mathbf{K}^{-1}\mathbf{1}_n} = \mathbf{w}_n^\top\mathbf{K}^{-1}\mathbf{w}_n$$

And since \mathbf{K} is a definite positive symmetric matrix, its inverse too and this inverse can be seen as a scalar product denoted $\langle \cdot, \cdot \rangle$. We get:

$$\langle \mathbf{w}_n, \mathbf{1}_n \rangle^2 = \langle \mathbf{w}_n, \mathbf{w}_n \rangle \langle \mathbf{1}_n, \mathbf{1}_n \rangle$$

Due to Cauchy-Schwartz inequality, this is possible if and only if we, $\mathbf{w}_n = \omega_0\mathbf{1}_n$ for some scalar ω_0 . Which means that $\mathbb{Y}^\top\boldsymbol{\omega} = \omega_0\mathbf{1}_n$. But this case has been excluded because it implies that the system of constraints is not of full rank (see Equation (10.4)). Therefore, the matrix \mathbb{S} is always invertible.

Proof of Proposition 5 when the system of Equations (III.9) is of rank $q + p - 1$.

Theory of Lagrangian factors holds only in the case of regular constraints, meaning that constraints' gradients should be independent. In particular, constraints are not regular if the constraints themselves are not regular. Let us show how to solve the optimisation problem, removing one of the conditions on the optimal weights. For the sake of simplicity, we remove the first one. Keeping the above notations, we denote:

$$\begin{aligned}\boldsymbol{\lambda}_0 &= (0, \lambda_2, \dots, \lambda_q)^\top \in \mathbb{R}^{q-1} \\ \boldsymbol{\lambda}_1 &= (\lambda_2, \dots, \lambda_q)^\top \in \mathbb{R}^{q-1} \\ \boldsymbol{\pi}_1 &= (\pi_{x_2^*}, \dots, \pi_{x_q^*})^\top \in \mathbb{R}_+^{q-1} \\ \pi_1 &= \pi_{x_1^*} \\ \mathbb{A}_1 &= [\boldsymbol{\alpha}(x_2^*, \dots, \boldsymbol{\alpha}_{x_q^*})] \in \mathbb{R}^{p \times (q-1)} \\ \gamma_1 &= \boldsymbol{\pi}_1^\top \boldsymbol{\pi}_1 = \gamma - \pi_1^2 \in \mathbb{R}_+ \\ \mathbb{H}_1 &= [\mathbf{h}(x_2^*, \dots, x_q^*)]\end{aligned}$$

The constraints (III.6) rewrite:

$$\mathbf{1}_n^\top \mathbb{A}_1 = \mathbf{1}_q^\top$$

And the system of Equations (IV.20) rewrites:

$$\begin{cases} \mathbf{K}\mathbb{A} = \mathbb{H} + \mathbf{1}_n \boldsymbol{\lambda}_0^\top + \mathbb{Y}^\top \boldsymbol{\lambda}' \boldsymbol{\pi}^\top \\ \mathbf{1}_n^\top \mathbb{A}_1 = \mathbf{1}_{q-1}^\top \\ \mathbb{Y}\mathbb{A}\boldsymbol{\pi} = \mathbf{m} \end{cases} \quad (\text{IV.21})$$

The first equation implies:

$$\begin{aligned}\mathbb{A} &= \mathbf{K}^{-1}\mathbb{H} + \mathbf{K}^{-1}\mathbf{1}_n \boldsymbol{\lambda}_0^\top + \mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}' \boldsymbol{\pi}^\top \\ \text{therefore } \mathbb{A}_1 &= \mathbf{K}^{-1}\mathbb{H}_1 + \mathbf{K}^{-1}\mathbf{1}_n \boldsymbol{\lambda}_1^\top + \mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}' \boldsymbol{\pi}_1^\top\end{aligned}$$

We replace \mathbb{A}_1 in the second equation:

$$\begin{aligned}\mathbf{1}_{q-1}^\top &= \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H}_1 + \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbf{1}_n \boldsymbol{\lambda}_1^\top + \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}' \boldsymbol{\pi}_1^\top \\ \mathbf{1}_{q-1}^\top \boldsymbol{\pi}_1 &= \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H}_1 \boldsymbol{\pi}_1 + \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbf{1}_n \boldsymbol{\lambda}_1^\top \boldsymbol{\pi}_1 + \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}' \boldsymbol{\pi}_1^\top \boldsymbol{\pi}_1 \\ 1 - \pi_1 &= \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H}_1 \boldsymbol{\pi}_1 + \delta \boldsymbol{\lambda}_1^\top \boldsymbol{\pi}_1 + \gamma_1 \mathbf{u}^\top \boldsymbol{\lambda}' \\ \delta \boldsymbol{\lambda}_1^\top \boldsymbol{\pi}_1 &= 1 - \pi_1 - \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H}_1 \boldsymbol{\pi}_1 - \gamma_1 \mathbf{u}^\top \boldsymbol{\lambda}' \\ \boldsymbol{\lambda}_0^\top \boldsymbol{\pi} &= \frac{1}{\delta} \left(1 - \pi_1 - \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H}_1 \boldsymbol{\pi}_1 - \gamma_1 \mathbf{u}^\top \boldsymbol{\lambda}' \right)\end{aligned}$$

We can also replace \mathbb{A} in the third equation and replace $\boldsymbol{\lambda}_0^\top \boldsymbol{\pi}$ with the last result:

$$\begin{aligned} m &= \mathbb{Y}\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \mathbb{Y}\mathbf{K}^{-1}\mathbf{1}_n\boldsymbol{\lambda}_0^\top \boldsymbol{\pi} + \mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}'\boldsymbol{\pi}^\top \boldsymbol{\pi} \\ m &= \mathbb{Y}\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \mathbf{u}\boldsymbol{\lambda}_0^\top \boldsymbol{\pi} + \gamma\mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}' \\ m &= \mathbb{Y}\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \frac{1}{\delta}\mathbf{u}\left(1 - \pi_1 - \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H}_1\boldsymbol{\pi}_1 - \gamma_1\mathbf{u}^\top \boldsymbol{\lambda}'\right) + \gamma\mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}' \\ m &= \mathbb{Y}\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \frac{1 - \pi_1}{\delta}\mathbf{u} - \frac{1}{\delta}\mathbf{u}\mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H}_1\boldsymbol{\pi}_1 - \frac{\gamma_1}{\delta}\mathbf{u}\mathbf{u}^\top \boldsymbol{\lambda}' + \gamma\mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}' \end{aligned}$$

Which yields:

$$\left(\frac{\gamma_1}{\delta}\mathbf{u}\mathbf{u}^\top - \gamma\mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top\right)\boldsymbol{\lambda}' = \mathbb{Y}\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \frac{1 - \pi_1}{\delta}\mathbf{u} - \frac{1}{\delta}\mathbf{u}\mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H}_1\boldsymbol{\pi}_1 - m$$

Assuming that $\frac{\gamma_1}{\delta}\mathbf{u}\mathbf{u}^\top - \gamma\mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top$ is invertible:

$$\boldsymbol{\lambda}' = \left(\frac{\gamma_1}{\delta}\mathbf{u}\mathbf{u}^\top - \gamma\mathbb{Y}\mathbf{K}^{-1}\mathbb{Y}^\top\right)^{-1} \left(\mathbb{Y}\mathbf{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \frac{1 - \pi_1}{\delta}\mathbf{u} - \frac{1}{\delta}\mathbf{u}\mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H}_1\boldsymbol{\pi}_1 - m\right)$$

But we know that:

$$\begin{aligned} \mathbf{1}_{q-1}^\top &= \mathbf{1}_n^\top \mathbf{K}^{-1}\mathbb{H}_1 + \delta\boldsymbol{\lambda}_1^\top + \mathbf{u}^\top \boldsymbol{\lambda}'\boldsymbol{\pi}_1^\top \\ \text{therefore } \boldsymbol{\lambda}_1 &= \frac{1}{\delta} \left(\mathbf{1}_{q-1} - \mathbb{H}_1^\top \mathbf{K}^{-1}\mathbf{1}_n - \boldsymbol{\pi}_1\boldsymbol{\lambda}'^\top \mathbf{u}\right) \end{aligned}$$

And \mathbb{A} can finally be computed with the equation:

$$\mathbb{A} = \mathbf{K}^{-1}\mathbb{H} + \mathbf{K}^{-1}\mathbf{1}_n\boldsymbol{\lambda}_0^\top + \mathbf{K}^{-1}\mathbb{Y}^\top \boldsymbol{\lambda}'\boldsymbol{\pi}^\top$$

□

10.5 Remark 3 Covariance matrices with two constraints

Proof of Remark 3. - The proof is similar to the one of Remark 2 and uses the fact that, under chosen assumptions and for any prediction point x^* ,

$$\mathbf{K}\boldsymbol{\alpha}(x^*) - \mathbf{h}(x^*) = \widetilde{\mathbf{K}}\boldsymbol{\alpha}(x^*) - \widetilde{\mathbf{h}}(x^*).$$

Hence the gradient of $\Delta_2(x^*)$ in Equation (IV.19) is unchanged when replacing \mathbf{K} and $\mathbf{h}(x^*)$ by $\widetilde{\mathbf{K}}$ and $\widetilde{\mathbf{h}}(x^*)$, and all further expressions follows the same way in the proof of Proposition 5. □

10.6 Proposition 7 Joint Kriging variance with arbitrary weights

Proof of Proposition 7. The first equation is a simple vector rewriting of Equation (III.1). For the prediction error, one simply write, whatever the weights $\boldsymbol{\alpha}(x^*)$,

$$\begin{aligned} \Delta(x^*) &= \mathbb{E} \left[\|\mathbf{M}(x^*) - \mathbf{Y}(x^*)\|_{\mathbb{W}}^2 \right] \\ &= \mathbb{E} \left[(\mathbf{M}(x^*) - \mathbf{Y}(x^*))^\top \mathbb{W} (\mathbf{M}(x^*) - \mathbf{Y}(x^*)) \right] \\ &= \mathbb{E} \left[(\mathbb{Y}\boldsymbol{\alpha}(x^*) - \mathbf{Y}(x^*))^\top \mathbb{W} (\mathbb{Y}\boldsymbol{\alpha}(x^*) - \mathbf{Y}(x^*)) \right] \\ &= \mathbb{E} \left[\boldsymbol{\alpha}(x^*)^\top \mathbb{Y}^\top \mathbb{W} \mathbb{Y} \boldsymbol{\alpha}(x^*) - 2\boldsymbol{\alpha}(x^*)^\top \mathbb{Y}^\top \mathbb{W} \mathbf{Y}(x^*) + \mathbf{Y}(x^*)^\top \mathbb{W} \mathbf{Y}(x^*) \right]. \end{aligned}$$

Hence the result. □

10.7 Remark 5 Covariance matrices in Joint Kriging mean and variance

Proof of Remark 5. The case where $\boldsymbol{\mu} = \mathbf{0}_p$ is straightforward, as in that case $\widetilde{\mathbf{K}} = \mathbf{K}$, $\widetilde{\mathbf{h}}(x^*) = \mathbf{h}(x^*)$ and $\widetilde{v}(x^*) = v(x^*)$, whatever the weights $\boldsymbol{\alpha}(x^*)$. It remains the case where weights are summing to one. As in previous remarks, under chosen assumptions one gets

$$\mathbf{K}\boldsymbol{\alpha}(x^*) - \mathbf{h}(x^*) = \widetilde{\mathbf{K}}\boldsymbol{\alpha}(x^*) - \widetilde{\mathbf{h}}(x^*),$$

and moreover one can show that

$$-\boldsymbol{\alpha}(x^*)^\top \mathbf{h}(x^*) + v(x^*) = -\boldsymbol{\alpha}(x^*)^\top \widetilde{\mathbf{h}}(x^*) + \widetilde{v}(x^*).$$

Hence the result. □

10.8 Proposition 8 Variance sharing

Proof of Proposition 8. The difficulty here is to derive the cross-covariance $k_{ij}(x, x') = \text{Cov}[Y_i(x), Y_j(x')]$ from the expression of $k(x, x')$ that is detailed in Remark 7

$$k(x, x') := \mathbb{E}[\mathbf{Y}(x)^\top \mathbf{W} \mathbf{Y}(x')] - \mathbb{E}[\mathbf{Y}(x)^\top] \mathbf{W} \mathbb{E}[\mathbf{Y}(x')]$$

Denoting $\widetilde{\mathbf{Y}}(x) := \mathbf{W}^{1/2} \mathbf{Y}(x)$, $x \in \chi$, this scalar covariance writes

$$k(x, x') = \mathbb{E}[\widetilde{\mathbf{Y}}(x)^\top \widetilde{\mathbf{Y}}(x')] - \mathbb{E}[\widetilde{\mathbf{Y}}(x)^\top] \mathbb{E}[\widetilde{\mathbf{Y}}(x')] \quad (\text{IV.22})$$

One would like to compute the $p \times p$ cross-covariance matrix between $\mathbf{Y}(x)$ and $\mathbf{Y}(x')$, using $\mathbf{Y}(x) = \mathbf{W}^{-1/2} \widetilde{\mathbf{Y}}(x)$, $x \in \chi$:

$$\begin{aligned} \mathbf{K}_Y(x, x') &:= \mathbb{E}[\mathbf{Y}(x) \mathbf{Y}(x')^\top] - \mathbb{E}[\mathbf{Y}(x)] \mathbb{E}[\mathbf{Y}(x')^\top] \\ &= \mathbf{W}^{-1/2} \left(\mathbb{E}[\widetilde{\mathbf{Y}}(x) \widetilde{\mathbf{Y}}(x')^\top] - \mathbb{E}[\widetilde{\mathbf{Y}}(x)] \mathbb{E}[\widetilde{\mathbf{Y}}(x')^\top] \right) \mathbf{W}^{-1/2\top} \\ &= \mathbf{W}^{-1/2} \mathbf{K}_{\widetilde{\mathbf{Y}}}(x, x') \mathbf{W}^{-1/2\top} \end{aligned} \quad (\text{IV.23})$$

where one defines $\mathbf{K}_{\widetilde{\mathbf{Y}}}(x, x') := \mathbb{E}[\widetilde{\mathbf{Y}}(x) \widetilde{\mathbf{Y}}(x')^\top] - \mathbb{E}[\widetilde{\mathbf{Y}}(x)] \mathbb{E}[\widetilde{\mathbf{Y}}(x')^\top]$.

Now assume that:

$$\text{Cov}[\widetilde{Y}_i(x), \widetilde{Y}_j(x')] = 0 \quad \text{whenever } i \neq j, x, x' \in \chi.$$

This implies that $\mathbf{W}^{1/2}$ is proportional to a whitening transformation, so that all components of $\widetilde{Y}_1(x), \dots, \widetilde{Y}_p(x)$ are uncorrelated.

Assume furthermore that:

$$\text{Cov}[\widetilde{Y}_1(x), \widetilde{Y}_1(x')] = \dots = \text{Cov}[\widetilde{Y}_p(x), \widetilde{Y}_p(x')], \quad x, x' \in \chi.$$

Then one easily sees from Equation (IV.22) that the scalar $k(x, x')$ satisfies

$$k(x, x') = \sum_{i=1}^p \text{Cov} [\tilde{Y}_i(x), \tilde{Y}_i(x')] = p \text{Cov} [\tilde{Y}_j(x), \tilde{Y}_j(x')], \quad j = 1, \dots, p$$

Hence under these assumptions, denoting \mathbb{I}_p the $p \times p$ identity matrix,

$$\mathbf{K}_{\tilde{\mathbf{Y}}}(x, x') = \frac{1}{p} k(x, x') \mathbb{I}_p.$$

As a consequence, from Equation (IV.23),

$$\mathbf{K}_{\mathbf{Y}}(x, x') := \mathbb{E} [\mathbf{Y}(x) \mathbf{Y}(x')^\top] - \mathbb{E} [\mathbf{Y}(x)] \mathbb{E} [\mathbf{Y}(x')^\top] = \frac{1}{p} k(x, x') \mathbf{W}^{-1} \quad (\text{IV.24})$$

$$\text{Cov} [Y_i(x), Y_j(x')] = \frac{1}{p} k(x, x') (\mathbf{W}^{-1})_{ij} \quad (\text{IV.25})$$

Now from this, one can derive the local cross errors

$$\delta_{ij}(x, x') := \mathbb{E} [(M_i(x) - Y_i(x)) (M_j(x') - Y_j(x'))]$$

Let us denote by \mathbf{Y}_i the i th row vector of the matrix \mathbf{Y} . We get

$$\begin{aligned} \delta_{ij}(x, x') &= \mathbb{E} [(\mathbf{Y}_i \boldsymbol{\alpha}(x) - Y_i(x)) (\mathbf{Y}_j \boldsymbol{\alpha}(x') - Y_j(x'))] \\ &= \boldsymbol{\alpha}(x)^\top \mathbb{E} [\mathbf{Y}_i^\top \mathbf{Y}_j] \boldsymbol{\alpha}(x') - \boldsymbol{\alpha}(x)^\top \mathbb{E} [\mathbf{Y}_i^\top Y_j(x')] \\ &\quad - \mathbb{E} [Y_i(x)^\top \mathbf{Y}_j] \boldsymbol{\alpha}(x') + \mathbb{E} [Y_i(x) Y_j(x')] \end{aligned}$$

Now assume $\mathbb{E} [\mathbf{Y}(x)] = \boldsymbol{\mu}$ for all $x \in \chi$. Then from Equation (IV.25),

$$\mathbb{E} [Y_i(x)^\top Y_j(x')] - \mu_i \mu_j = \frac{1}{p} k(x, x') (\mathbf{W}^{-1})_{ij}$$

which implies, using the matrix $\tilde{\mathbf{K}}$ defined in Equations (III.19), page 115:

$$\mathbb{E} [\mathbf{Y}_i^\top \mathbf{Y}_j] - \mu_i \mu_j \mathbf{1}_n \mathbf{1}_n^\top = \frac{1}{p} (\mathbf{W}^{-1})_{ij} \tilde{\mathbf{K}} \quad \text{and} \quad \mathbb{E} [Y_i^\top Y_j(x')] - \mu_i \mu_j \mathbf{1}_n = \frac{1}{p} (\mathbf{W}^{-1})_{ij} \tilde{\mathbf{h}}(x').$$

Furthermore, assume that either weights sum to one or $\boldsymbol{\mu} = \mathbf{0}_p$, then terms in $\mu_i \mu_j$ vanish and one gets:

$$\delta_{ij}(x, x') = \frac{1}{p} (\mathbf{W}^{-1})_{ij} \left(\boldsymbol{\alpha}(x)^\top \tilde{\mathbf{K}} \boldsymbol{\alpha}(x') - \boldsymbol{\alpha}(x)^\top \tilde{\mathbf{h}}(x') - \tilde{\mathbf{h}}^\top(x) \boldsymbol{\alpha}(x') + k(x, x') \right).$$

In particular from Proposition 7, using Remark 7 and Remark 5,

$$\delta_i(x^*) = \frac{1}{p} (\mathbf{W}^{-1})_{ii} \Delta(x^*). \quad (\text{IV.26})$$

From Equation (IV.24), when $k(x, x) = \sigma^2$ for all x , one can write

$$\mathbf{K}_Y(x, x) = \frac{1}{p} \sigma^2 \mathbf{W}^{-1}.$$

Using $\sigma_i^2 := \text{Var}[Y_i(x)]$, assumed to be constant over x ,

$$\frac{1}{p} (\mathbf{W}^{-1})_{ii} = \frac{(\mathbf{K}_Y(x, x))_{ii}}{\sigma^2} = \frac{\sigma_i^2}{\sigma^2}.$$

Hence from Equation (IV.26),

$$\delta_i(x^*) = \frac{\sigma_i^2}{\sigma^2} \Delta(x^*).$$

□

10.9 Remark 6 Constraints' impact

Proof: Nugget for positive weights. [Proof of Remark 6] The result is a very straightforward rewriting and interpretation of constraints (III.6) and (III.7). From $\mathbf{Y} \mathbf{A} \boldsymbol{\pi} = \mathbf{m}$ one derives $\mathbf{1}_p^\top \mathbf{m} = \mathbf{1}_p^\top \mathbf{Y} \mathbf{A} \boldsymbol{\pi} = \mathbf{1}_n^\top \mathbf{A} \boldsymbol{\pi} = \mathbf{1}_q^\top \boldsymbol{\pi} = 1$, hence the natural constraint on prescribed average membership degrees in \mathbf{m} , that must sum to one. □

10.10 Proof of Proposition 9 Nugget ensuring positive weights

Proof of Proposition 9.

We have	$\mathbf{K}_{\text{nug}} := \mathbf{K} + \eta \mathbb{I}_n .$
We denote	$\varepsilon := \frac{1}{\eta} .$
therefore	$\mathbf{K}_{\text{nug}} = \mathbf{K} + \frac{1}{\varepsilon} \mathbb{I}_n ,$ $= \frac{1}{\varepsilon} (\mathbb{I}_n + \varepsilon \mathbf{K})$
and	$\mathbf{K}_{\text{nug}}^{-1} = \varepsilon (\mathbb{I}_n - \varepsilon \mathbf{K} + o(\varepsilon))$ $= \varepsilon \mathbb{I}_n + o(\varepsilon) .$

Following notations of Proposition 4, we have:

$$\begin{aligned} \delta &= \mathbf{1}_n^\top \mathbf{K}_{\text{nug}}^{-1} \mathbf{1}_n \\ \delta &= n\varepsilon + o(\varepsilon) \\ \delta &= n\varepsilon(1 + o(1)) \\ \delta^{-1} &= \frac{1}{n\varepsilon} (1 + o(1)) \\ \boldsymbol{\lambda}^\top &= \frac{1}{n\varepsilon} (1 + o(1)) \left(\mathbf{1}_q^\top - \mathbf{1}_n^\top (\varepsilon \mathbb{I}_n + o(\varepsilon)) \mathbb{H} \right) \\ \boldsymbol{\lambda}^\top &= \frac{1}{n\varepsilon} \left(\mathbf{1}_q^\top + o(1) \right) \end{aligned}$$

And eventually:

$$\begin{aligned}\mathbb{A} &= (\varepsilon \mathbb{I}_n + o(\varepsilon)) \left(\mathbb{H} + \frac{1}{n\varepsilon} \mathbf{1}_n \mathbf{1}_q^\top + \frac{1}{n\varepsilon} o(1) \right) \\ \mathbb{A} &= \frac{1}{n} \mathbf{1}_n \mathbf{1}_q^\top + \varepsilon \mathbb{H} - o(1) \\ \lim_{\varepsilon \rightarrow 0} \mathbb{A} &= \frac{1}{n} \mathbf{1}_n \mathbf{1}_q^\top\end{aligned}$$

Which is the expected result. □

10.11 Procedure to Retrieve OpenML Results for the Quake Dataset

OpenML statistics for the Quake dataset

Marc Grossouvre

2024-06-28

```

library(OpenML)

## Warning: le package 'OpenML' a été compilé avec la version R 4.3.3

library(dplyr)

## Warning: le package 'dplyr' a été compilé avec la version R 4.3.2

##
## Attachement du package : 'dplyr'

## Les objets suivants sont masqués depuis 'package:stats':
##
##   filter, lag

## Les objets suivants sont masqués depuis 'package:base':
##
##   intersect, setdiff, setequal, union

setOMLConfig(apikey = "e3fe50169b55509840741ed93bd954dc")

## OpenML configuration:
##   server      : http://www.openml.org/api/v1
##   cachedir    : C:\Users\marc\AppData\Local\Temp\RtmpcZwn0k/cache
##   verbosity   : 1
##   arff.reader  : farff
##   confirm.upload : TRUE
##   apikey      : *****954dc

task <- getOMLTask(task.id = 4516)

## Downloading from 'http://www.openml.org/api/v1/task/4516' to 'C:\Users\marc\AppData\Local\Temp\RtmpcZ
## Downloading from 'https://api.openml.org/api_splits/get/4516/Task_4516_splits.arff' to 'C:\Users\mar
## Downloading from 'http://www.openml.org/api/v1/data/948' to 'C:\Users\marc\AppData\Local\Temp\RtmpcZl
## Downloading from 'https://api.openml.org/data/v1/download/53482/quake.arff' to 'C:\Users\marc\AppData

```

SUPPLEMENTARY MATERIAL FOR JOINT KRIGING

```
## Warning in getOMLDataSetById(data.id = data.id, cache.only = cache.only, : Data
## set has been deactivated.

## Le chargement a nécessité le package : readr

## Warning: le package 'readr' a été compilé avec la version R 4.3.3

run <- listOMLRuns(task.id = 4516)

## Downloading from 'http://www.openml.org/api/v1/json/run/list/task/4516/limit/5000' to '<mem>'.

flows <- listOMLFlows()

## Downloading from 'http://www.openml.org/api/v1/json/flow/list/' to '<mem>'.

# For Kernel Logistic Regression
flowid <- c(logistic = 1174, randomforest = 1079)

stats_best <- list()

for(i in 1:2) {
  fid <- flowid[i]

  res <- listOMLRunEvaluations(task.id = 4516L, flow.id = fid)

  all_runs <-
    lapply(X = res$run.id,
           FUN = getOMLRun)

  all_runs_evaluations <-
    lapply(X = all_runs, FUN = \(x) {
      x$output.data$evaluations
    })

  all_runs_acc <-
    lapply(X = all_runs_evaluations, FUN = \(x) {
      x %>%
        filter(name == "predictive_accuracy") %>%
        select('repeat', fold, value)
    })

  all_runs_acc2 <-
    all_runs_acc %>%
    do.call(what = rbind)

  colnames(x = all_runs_acc2)[1] <- "rep"

  all_runs_acc2 %>%
    filter(is.na(x = fold))

  res <- all_runs_acc2 %>%
```

```

group_by(rep) %>%
  summarise(
    rep = min(rep),
    mean = mean(value),
    min = min(value),
    med = median(value),
    max = max(value)
  ) %>%
  arrange(min)

stats_best[[names(x = flowid)[i]]] <-
  list(summary = summary(res$mean),
        sd = sd(res$mean))
}

## Suggestion: Use the 'evaluation.measure' argument to restrict the results to only one measure.

## Downloading from 'http://www.openml.org/api/v1/json/evaluation/list/task/4516/flow/1174' to '<mem>'.

## Downloading from 'http://www.openml.org/api/v1/run/191079' to 'C:\Users\marc\AppData\Local\Temp\Rtmp

## Downloading from 'https://api.openml.org/data/download/631543/weka_generated_predictions705498022027:

## Downloading from 'http://www.openml.org/api/v1/run/350782' to 'C:\Users\marc\AppData\Local\Temp\Rtmp

## Downloading from 'https://api.openml.org/data/download/1208403/weka_generated_predictions14087414881:

## Downloading from 'http://www.openml.org/api/v1/run/390395' to 'C:\Users\marc\AppData\Local\Temp\Rtmp

## Downloading from 'https://api.openml.org/data/download/1362685/weka_generated_predictions55298222345:

## Downloading from 'http://www.openml.org/api/v1/run/444283' to 'C:\Users\marc\AppData\Local\Temp\Rtmp

## Downloading from 'https://api.openml.org/data/download/1555842/weka_generated_predictions37275019620:

## Downloading from 'http://www.openml.org/api/v1/run/466818' to 'C:\Users\marc\AppData\Local\Temp\Rtmp

## Downloading from 'https://api.openml.org/data/download/1641694/weka_generated_predictions84159826790:

## Suggestion: Use the 'evaluation.measure' argument to restrict the results to only one measure.

## Downloading from 'http://www.openml.org/api/v1/json/evaluation/list/task/4516/flow/1079' to '<mem>'.

## Downloading from 'http://www.openml.org/api/v1/run/186561' to 'C:\Users\marc\AppData\Local\Temp\Rtmp

## Downloading from 'https://api.openml.org/data/download/618019/weka_generated_predictions632420274369:

## Downloading from 'http://www.openml.org/api/v1/run/340756' to 'C:\Users\marc\AppData\Local\Temp\Rtmp

```

```

## Downloading from 'https://api.openml.org/data/download/1170087/weka_generated_predictions24185739005!
## Downloading from 'http://www.openml.org/api/v1/run/356272' to 'C:\Users\marc\AppData\Local\Temp\Rtmp
## Downloading from 'https://api.openml.org/data/download/1229867/weka_generated_predictions82325074099:
## Downloading from 'http://www.openml.org/api/v1/run/386757' to 'C:\Users\marc\AppData\Local\Temp\Rtmp
## Downloading from 'https://api.openml.org/data/download/1349383/weka_generated_predictions57142944906'
## Downloading from 'http://www.openml.org/api/v1/run/440931' to 'C:\Users\marc\AppData\Local\Temp\Rtmp
## Downloading from 'https://api.openml.org/data/download/1543404/weka_generated_predictions16553982629!
## Downloading from 'http://www.openml.org/api/v1/run/458077' to 'C:\Users\marc\AppData\Local\Temp\Rtmp
## Downloading from 'https://api.openml.org/data/download/1608421/weka_generated_predictions60540145002!

stats_best

## $logistic
## $logistic$summary
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.5551 0.5569 0.5582 0.5582 0.5597 0.5615
##
## $logistic$sd
## [1] 0.002042705
##
##
## $randomforest
## $randomforest$summary
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.5220 0.5250 0.5280 0.5281 0.5298 0.5390
##
## $randomforest$sd
## [1] 0.004902082

rr <- run[1,]

all_predictiveaccuracies <-
  apply(X =run,
        MARGIN = 1,
        FUN = function(rr){
          # print(rr)

          rid <- as.integer(x = rr[["run.id"]])
          fid <- as.integer(x = rr[["flow.id"]])

          res <- listOMLRunEvaluations(task.id = 4516L, flow.id = fid, run.id = rid, evaluation.measure = "pr
          res$predictive.accuracy
        })

```

```

library(ggplot2)

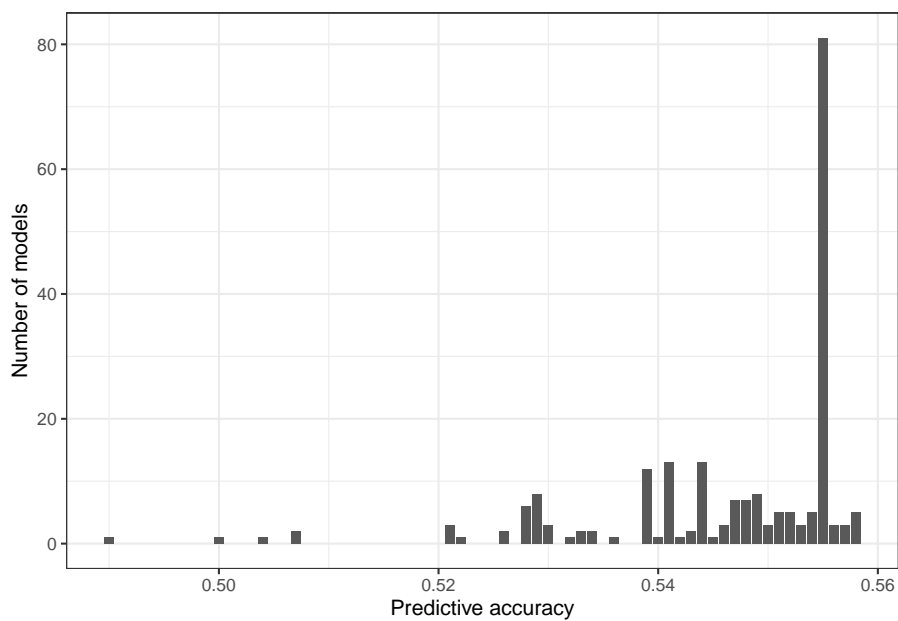
## Warning: le package 'ggplot2' a été compilé avec la version R 4.3.3

r_all_predictiveaccuracies <-
  data.frame(accuracy = round(x = all_predictiveaccuracies, digits = 3))

allopenmplot <-
  ggplot(data = r_all_predictiveaccuracies, mapping = aes(x = accuracy)) +
  geom_bar() +
  xlab(label = "Predictive accuracy") +
  ylab(label = "Number of models") +
  theme_bw()

allopenmplot

```



```

ggsave(filename = "D:/Seafire/Seafire/RenovationEnergetique/60_THESE/V08/JointKriging/allopenmplot.pdf")

```

Supplementary Material for Chapter IV

11	Dictionary of Main Variables264
12	Details of Selected Variables per Model265
13	Confusion Matrices.266

11 Dictionary of Main Variables

Variables are presented in the order in which they were selected by the **KNN** algorithm.

- altitude Altitude of the building.
- chaufnum Type of heating system for the dwelling/building.
- detentmajnum Most frequent level of maintenance in the building, ranging from 1 to 5.
- energienum Main source of energy for heating.
- entbadpct Percentage of dwellings with bad level of maintenance (4 or 5 out of 5).
- entokpct Percentage of dwellings with good level of maintenance (1, 2 or 3 out of 5).
- hlmptct Number of dwellings in the building that are under the social housing system.
- jannatmax Year of construction of the most recent part of the building/dwelling.
- jannatmin Year of construction of the oldest part of the building/dwelling.
- lat Latitude of the building.
- lon Longitude of the building.
- murnum Main walls material.
- mutapct Percentage of dwellings in the building that have been sold in the last year.
- nc6 Number of dwellings with comfort level 6 on a scale ranging from 1 to 8, from good to bad.
- nc7 Number of dwellings with comfort level 7 on a scale ranging from 1 to 8, from good to bad.
- nent0 Number of dwellings with unknown level of maintenance. Maintenance levels range from 1 to 5, from good to bad.
- nivtot Number of floors in the building.
- nlogh Number of dwellings in the building.
- nloghvac2a Number of dwellings in the building that have been vacant for at least 2 years.
- nprop Number of owners that occupy their dwelling in the building.
- nt1 Number of dwellings in the building comprising 1 rooms and a bathroom.
- nt2 Number of dwellings in the building comprising 2 rooms and a bathroom.
- nt3 Number of dwellings in the building comprising 3 rooms and a bathroom.
- nt4 Number of dwellings in the building comprising 4 rooms and a bathroom.
- pgaz01 Availability of city gas connection in the building/dwelling.
- regth Thermal regulation at the time of construction.
- regth_1948 Boolean indicating that the building construction started before 1948.
- surfacemoy Average surface area of the dwellings in the building.
- toitnum Roof material.
- typobatinum Type of occupation: only housing or also with some professional activities.
- typopronum Typology of dwellings' owners: only one private owner, multiple private owners, mix of private and public owners, only public owners.

12 Details of Selected Variables per Model

Variables selected by FKNN: energienum, jannatmax, altitude, regth, surfacemoy, nt2, hlmptct, nt3, regth_1948, jannatmin, nloghvac2a, nc6, chauffnum, nent0, toitnum, entokpct, nt4, nivtot, nprop, murnum, pgaz01, nlogh, nt1, entbadpct, detentmajnum, nc7, mutapct, typobatinum, typropnum.

Variables selected by Joint Kriging: lat, lon, jannatmax, jannatmin, energienum, chauffagenum, nlogh, entbadpct, nprop.

Variables selected by Random Forest: jannatmin, energienum, surfacemoy, lon, lat, nprop, nivtot, hlmptct, detentmajnum.

13 Confusion Matrices

True values	Predicted values							
	A	B	C	D	E	F	G	
A	648.7	320.2	288.2	297.2	217.4	150.6	104.8	2 027
B	249.2	491.0	361.2	283.7	183.1	121.4	82.4	1 772
C	667.6	844.5	3 262.9	2 236.1	1 228.9	785.9	512.0	9 538
D	1 097.8	1 100.5	3 018.4	5 255.0	3 098.8	1 873.3	1 266.2	16 710
E	750.7	702.3	1 527.0	2 560.4	2 947.7	1 764.3	1 211.5	11 464
F	338.7	304.9	637.9	1 071.7	1 224.5	1 092.6	801.8	5 472
G	231.4	185.0	348.8	554.7	616.3	569.0	511.9	3 017
	3 984.1	3 948.4	9 444.3	12 258.8	9 516.7	6 357.1	4 490.6	50 000

Table IV.5 – Confusion matrix of the Fuzzy KNN model.

True values	Predicted values							
	A	B	C	D	E	F	G	
A	933	189	231	343	237	64	30	2 027
B	225	542	447	368	144	24	22	1 772
C	205	348	4 353	3 227	1 028	255	122	9 538
D	219	184	2 671	8 558	3 623	982	473	16 710
E	134	79	890	3 978	4 312	1 415	656	11 464
F	53	27	306	1 350	1 849	1 230	657	5 472
G	49	15	148	676	918	689	522	3 017
	1 818	1 384	9 046	18 500	12 111	4 659	2 482	50 000

Table IV.6 – Confusion matrix of the binarised KNN model, used for hard classification.

SUPPLEMENTARY MATERIAL FOR FUZZY CLASSIFICATION

True values	Predicted values							
	A	B	C	D	E	F	G	
A	179.8	50.6	3.4	22.1	-5.5	-23.1	-21.2	206
B	45.5	71.1	25.5	30.7	8.5	-8.4	-10.9	162
C	53.5	114.0	454.7	375.3	173.8	14.2	-39.5	1 146
D	-5.0	24.2	417.6	885.9	523.0	175.2	60.1	2 081
E	-44.0	-49.9	159.8	500.1	447.6	228.9	159.4	1 402
F	-14.4	-28.2	63.5	182.9	169.1	149.6	100.5	623
G	-9.3	-19.8	21.4	83.9	85.6	86.6	131.6	380
	206	162	1 146	2 081	1 402	623	380	6 000

Table IV.7 – Confusion matrix of the Fuzzy Joint Kriging model.

True values	Predicted values							
	A	B	C	D	E	F	G	
A	120	18	4	27	10	11	16	206
B	39	56	17	32	7	3	8	162
C	89	86	423	360	103	35	50	1 146
D	96	74	262	1 092	284	100	173	2 081
E	33	17	51	607	340	147	207	1 402
F	16	5	15	233	109	137	108	623
G	3	7	5	113	44	56	152	380
	396	263	777	2 464	897	489	714	6 000

Table IV.8 – Confusion matrix of the binarised Joint Kriging model, used for hard classification.

True values	Predicted values							
	A	B	C	D	E	F	G	
A	147	36	4	6	10	7	14	224
B	27	115	20	8	15	2	4	191
C	56	173	513	192	131	64	58	1 187
D	109	111	340	610	430	240	180	2 020
E	67	48	69	221	436	284	249	1 374
F	17	9	13	59	140	211	181	630
G	25	4	6	22	46	80	191	374
	448	496	965	1 118	1 208	888	877	6 000

Table IV.9 – Confusion matrix of Random Forest hard classification.

Supplementary Material for the Conclusion

14	U.R.B.S. model 2022.	270
----	------------------------------	-----

14 Performances of the U.R.B.S. Prediction Model for EPCs in Production in December 2022



URBS

Urban Retrofit Business Services

Bilan des prédictions DPE V2.2.0 au niveau national

12 décembre 2022

urbs.fr



Table des matières

1	Performance générale	3
1.1	Scores	3
1.2	Étalement des scores	3
1.3	Choix de la distance et du type de transformation de données	4
1.4	Variables sélectionnées	5
2	Statistiques détaillées	8
2.1	Indicateurs de tendance centrale de l'erreur	8
2.2	Bienvenue dans la matrice	9
2.3	Un bon coup de fourchette	10
3	Conclusion	10
4	ANNEXE	11
4.1	Auvergne Rhône-Alpes	12
4.2	Bourgogne Franche-Comté	13
4.3	Bretagne	14
4.4	Corse	15
4.5	Centre Val de Loire	16
4.6	Grand-Est	17
4.7	Hauts de France	18
4.8	Ile de France	19
4.9	Nouvelle Aquitaine	20
4.10	Normandie	21
4.11	Occitanie	22
4.12	Provence Alpes Côte d Azur	23
4.13	Pays de Loire	24



1 Performance générale

1.1 Scores

Table 1 – Score de validation par région

Territoire	Pourcentage moyen de bonnes prédictions par étiquette
<i>Corse</i>	75.8
Centre Val de Loire	35.0
Pays de Loire	29.8
Ile de France	29.2
Occitanie	28.7
Grand-Est	28.4
Bourgogne Franche-Comté	26.1
Hauts de France	26.0
Bretagne	25.7
Normandie	25.6
Auvergne Rhône-Alpes	25.5
Provence Alpes Côte d Azur	24.1
Nouvelle Aquitaine	23.3

On s'intéresse avant tout au score qui a été choisi pour l'optimisation. Pour calculer ce score, on calcule le pourcentage de bonnes prédictions pour chaque étiquette puis on fait la moyenne de ces pourcentages. Pour mémoire, on a choisi ce score pour favoriser la bonne prédiction des étiquettes rares ou difficiles à prédire (A, B, F, G) et éviter que l'algorithme se focalise sur les étiquettes C, D, E.

Sur la Table 1 on voit que les scores sont globalement étalés entre 23% et 35%. La dispersion est donc importante. Pour la France entières, les données agrégées et pondérées par la population des régions amènent à un score global de **28.8**.

Il semble y avoir un problème sur la Corse pour laquelle on avait essayé de faire un algorithme spécial du fait d'un trop petit nombre de données d'apprentissage. Dans la suite, on ignore cette région, il faudra prévoir de refaire tourner l'algorithme ou étudier en détail les données appariées.

1.2 Étalement des scores

Pour un échantillon de 1000 adresses, quel est le score observé ? Cette information nous permet de connaître la variabilité des scores, selon les régions et globalement.

Sur la Figure 1, on voit que

- la grande majorité des échantillons a un score entre 25% et 30% ;
- on a un histogramme en cloche avec une queue plus étalée à droite qu'à gauche ;
- la variabilité des scores pour une région donnée est raisonnable ;
- toutes les régions ont sensiblement la même variabilité.

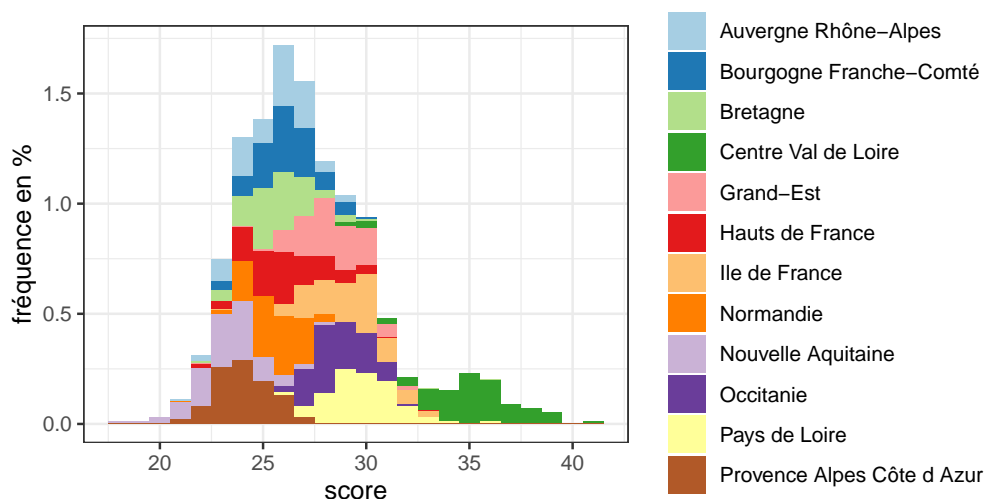


Figure 1 – Étalement des scores de prédiction du DPE pour des échantillons d'environ 1000 adresses.

1.3 Choix de la distance et du type de transformation de données

On compare les modèles sélectionnés pour la prédiction de chaque région.

Pour mémoire, les distances proposées sont : la distance euclidienne classique (norme 2), la distance de Manhattan (norme 1) et la distance de Hassanat.

On teste aussi les transformation de variables :

- raw : les variables sont laissées inchangées
- scaled : les variables sont centrées normées (transformation linéaire)
- positive : les variables sont ramenées entre 0 et 1 (transformation linéaire)
- pseudo_observations : les variables sont ramenées à leurs pseudo-observations. Par exemple, une variable avec les valeurs $(-3.5, -2.1, 0.3, 0.4, 2)$ deviendra $(1/6, 2/6, 3/6, 4/6, 5/6)$, seul le rang compte.
- normal_pseudo_observations : il s'agit d'une version normalisée des pseudo-observations (les queues de distribution sont plus étalées).
- pca : composantes principales.

De plus, on teste le nombre de voisins pris en compte pour des valeurs entre 1 et 30.

Sur la Table 2, on constate que :

- dans le peloton de tête, c'est la distance euclidienne qui a été sélectionnée ;
- dans ce même peloton de tête, ce sont en majorité les pseudo-observations qui sont favorisées ;
- les composantes principales ne sont jamais sélectionnées ;
- la situation du Centre Val de Loire fait figure d'exception avec la transformation "positive" sélectionnée.



Table 2 – Meilleurs modèles sélectionnés

Territoire	Meilleure transformation des données	Meilleure distance	Score de validation	Nombre de voisins
Centre Val de Loire	positive	Euclid	35.0	15
Pays de Loire	pseudo_observations	Euclid	29.8	5
Ile de France	pseudo_observations	Euclid	29.2	15
Occitanie	normal_pseudo_observations	Euclid	28.7	24
Grand-Est	pseudo_observations	Hassanat	28.4	29
Bourgogne Franche-Comté	scaled	Manhattan	26.1	23
Hauts de France	positive	Hassanat	26.0	1
Bretagne	raw	Hassanat	25.7	11
Normandie	pseudo_observations	Euclid	25.6	22
Auvergne Rhône-Alpes	normal_pseudo_observations	Hassanat	25.5	8
Provence Alpes Côte d Azur	scaled	Hassanat	24.1	14
Nouvelle Aquitaine	scaled	Hassanat	23.3	5

1.4 Variables sélectionnées

Intéressons-nous aux variables sélectionnées. Dans la Table 3, on a classé les régions de la plus performante, à gauche, à la moins performante à droite. La sélection de variable se fait par étapes, on indique par le nombre 1, la première variable sélectionnée (la plus significative), par le nombre 2, la 2ème variable sélectionnée, etc.

Table 3 – Meilleure variable sélectionnée à chaque étape

varName	Centre Val de Loire	Pays de Loire	Ile de France	Occitanie	Grand-Est	Bourgogne Franche-Comté	Hauts de France	Bretagne	Normandie	Auvergne Rhône-Alpes	Provence Alpes Côte d Azur	Nouvelle Aquitaine
chauf_individuel	9
confortpct	6
detent_mauvais	13
detent_mediocre	6	.	.	.
detent_passable	9
energie_autre	5	.	9	.	3	.	.	3
energie_electricite	2	2	2	5	2	2	5	.	2	2	3	2
energie_fioul	15	.	4	3	.	.	.
energie_gaz	10	3	.	3	.	.	2	2	.	.	2	.
energie_mixte	11
entbadpct	14
entokpct	8
himpct	18
jannatmax	1	.	.	.	1
jannatmin	4	.	1	1	.



Table 3 – Meilleure variable sélectionnée à chaque étape (*continued*)

varName	Centre Val de Loire	Pays de Loire	Ile de France	Occitanie	Grand-Est	Bourgogne Franche-Comté	Hauts de France	Bretagne	Normandie	Auvergne Rhône-Alpes	Provence Alpes Côte d Azur	Nouvelle Aquitaine
locpct	4	4	.
maison	7	.	.
mur_agglomere-autres	.	.	.	6	.	14	13
mur_agglomere-beton	17
mur_agglomere-bois	16	6
mur_agglomere-briques	.	10
mur_agglomere-meuliere	24
mur_agglomere-pierre	.	.	10
mur_beton-autres	7
mur_beton-meuliere	17
mur_bois	7	5	.	.	10	.	.
mur_bois-autres	6	7	.	.	.
mur_bois-pierre	12
mur_briques-meuliere	11	.	.
mur_indetermine	6	.
mur_meuliere	.	.	9	.	6	13
mur_meuliere-autres	16
mur_meuliere-pierre	12	25	15	.	5	.	.	.
mur_pierre-autres	23
mutapct	.	.	.	4
nc3	19
nc4	5	3
nc6	8
nc7	4
nent0	.	12
nivtot	.	3
nlogh	3	.	.	.
nloghvac2a	9	.	.
nprop	17
nt1	.	7	.	7	8
nt2	4	.	.	.
nt3	.	5
nt4	4
nt5p	4
occup_bairural	9
occup_gratuit	15
occup_Locmeuble	.	14	8	8	7	.



Table 3 – Meilleure variable sélectionnée à chaque étape (*continued*)

varName	Centre Val de Loire	Pays de Loire	Ile de France	Occitanie	Grand-Est	Bourgogne Franche-Comté	Hauts de France	Bretagne	Normandie	Auvergne Rhône-Alpes	Provence Alpes Côte d Azur	Nouvelle Aquitaine
occup_taxepro	10
pgaz01	3
regth	.	1	.	1	.	1	1	1	1	1	.	1
regth_1974	.	.	.	8
regth_1982	14	18	16
regth_1988	20
regth_2000	.	.	.	11	.	.	.	7
regth_2007	.	8	6
regth_2012	6	.	8	6	.	7	9	.	.	.	6	.
surfacemoy	.	.	.	2
surfacetot	7	5	.	.
toit_ardoises	.	.	.	7	3
toit_ardoises-autres	10
toit_beton	11	11
toit_beton-autres	.	.	.	12
toit_beton-tuiles	.	.	11
toit_tuiles-autres	12
toit_tuiles-zincaluminium	8
toit_zincaluminium	5
txindigne	.	6
typobati_habitation	18	.	.	5
typrop_coproprietemixte	21
typrop_etablissementpublic	.	9	13	8	.	22	7	5
typrop_monopropriete	5
typrop_officehlm	10	4
xcoord	.	.	.	4
zonehist01	3
zoneqpv01	4

On remarque que tous modèles confondus, un très grand nombre de variables sont susceptibles d’apparaître, ce qui indique la pertinence des variables proposées. Cependant, il semble difficile d’envisager un modèle à 80 variables, il pourrait donc être utile de réduire la dimension en projetant les données dans un espace de plus petite dimension mais qui conserve la même quantité d’information.

On a aussi une confirmation indirecte de la qualité de l’appariement par le fait que les variables attendues sont bien sélectionnées : âge du bâtiment et variables associées (période de réglementation thermique), matériaux des murs et du toit, niveau de confort et niveau d’entretien.

On note aussi que le nombre de variables sélectionnées varie beaucoup d’un modèle à l’autre.



2 Statistiques détaillées

2.1 Indicateurs de tendance centrale de l'erreur

Dans la Table 4, on observe les mesures d'erreur moyenne sur le DPE (en nombre d'étiquettes d'erreur) et sur le ndpe (en nombre de kWh/m²/an).

Table 4 – Mesures d'erreurs par région

Territoire	Score d'optimisation	Erreur absolue moyenne DPE	RMSE DPE	Erreur absolue moyenne ndpe	RMSE ndpe
Centre Val de Loire	35.00	0.73	1.08	70.00	97.80
Pays de Loire	29.86	0.83	1.19	70.47	96.24
Ile de France	29.14	0.77	1.12	74.82	103.52
Occitanie	28.71	0.76	1.14	62.44	87.33
Grand-Est	28.43	0.77	1.09	75.16	102.18
Hauts de France	26.14	1.00	1.37	92.28	121.90
Bourgogne Franche-Comté	26.00	0.84	1.18	80.86	108.23
Bretagne	25.71	0.80	1.16	68.82	92.79
Normandie	25.57	0.76	1.11	72.16	98.48
Auvergne Rhône-Alpes	25.43	0.89	1.26	82.77	110.92
Provence Alpes Côte d Azur	24.29	0.80	1.15	66.26	91.14
Nouvelle Aquitaine	23.43	0.98	1.36	82.81	110.69
FRANCE	28.80	0.83	1.20	75.73	103.38

Ce tableau appelle plusieurs remarques. D'abord, certaines régions font des contre-performances en terme de RMSE sur le DPE : Hauts de France, Nouvelle Aquitaine, Auvergne-Rhône-Alpes et dans une moindre mesure Pays de Loire. Or ce sont justement les régions qui ont sélectionné les plus petits nombres de voisins (respectivement 1, 5, 8, 5). Au contraire, les régions qui ont de plus petites erreurs moyennes ont sélectionné un plus grand nombre de voisins. Pour mémoire, on teste toutes les possibilités de nombres de voisins entre 1 et 30. Il semble apparaître un optimum autour de 15 voisins.

On remarque aussi que quelques régions ont une très petite erreur moyenne sur le ndpe : Occitanie, Provence Alpes Côté d'Azur et Bretagne sans qu'on trouve de point commun particulier entre les modèles.



2.2 Bienvenue dans la matrice

Voici la matrice de confusion moyenne pour 10000 prédictions sur toute la France.

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	88	40	32	84	42	12	1
B	38	109	114	161	60	15	1
C	25	44	503	863	222	48	1
D	24	35	377	1930	869	201	7
E	9	20	111	1084	1113	310	14
F	2	8	27	318	497	214	12
G	1	3	5	75	144	76	8

La matrice est correctement équilibrée autour de sa diagonale. Il persiste une grosse difficulté à prédire les étiquettes F et G. La prédiction des étiquettes A et B étant plutôt correcte au regard des difficultés passées.

Voici la même matrice de confusion en pourcentages par lignes.

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	29.5	13.5	10.7	27.9	14	4	0.4
B	7.7	21.9	22.8	32.3	12	3.1	0.2
C	1.5	2.6	29.5	50.6	13	2.8	0.1
D	0.7	1	10.9	56	25.2	5.8	0.2
E	0.3	0.8	4.2	40.7	41.8	11.6	0.5
F	0.2	0.7	2.5	29.5	46.1	19.9	1.1
G	0.3	0.9	1.7	24.1	46.1	24.3	2.7

On voit bien qu'une partie des bâtiments réellement en A ou B est prédite en C ou D, ce qui pourrait correspondre aux bâtiments "bien" rénovés. On sait qu'il nous manque un indicateur de rénovation.

Voici la même matrice de confusion en pourcentages par colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	46.9	15.5	2.7	1.9	1.4	1.4	2.5
B	20.3	42.1	9.7	3.6	2	1.8	2.6
C	13.5	16.9	43	19.1	7.5	5.5	3.2
D	12.8	13.6	32.2	42.7	29.5	23	16.4
E	4.9	7.8	9.5	24	37.8	35.3	30.3
F	1.2	3.1	2.3	7	16.9	24.4	26.3
G	0.5	1	0.4	1.7	4.9	8.6	18.7

On voit que du point de vue de l'utilisateur, les prédictions sont plutôt bonnes. Par exemple, si un utilisateur sélectionne les bâtiments A ou B, il trouvera à plus de 60% des bâtiments réellement en A ou B.



2.3 Un bon coup de fourchette

Dans la Table 5, on observe qu'en moyenne pour tout le pays, on a une prédiction à + ou - une étiquette qui est correcte à 83%.

Table 5 – Part de bonnes prédictions dans une fourchette d'erreur DPE

Prédictions correctes	Part de toutes les prédictions
+/- 0 étiquettes	39.7
+/- 1 étiquettes	82.9
+/- 2 étiquettes	95.5
+/- 3 étiquettes	99.0
+/- 4 étiquettes	99.8
+/- 5 étiquettes	100.0
+/- 6 étiquettes	100.0

3 Conclusion

En conclusion, les statistiques nationales sont conformes à nos attentes, très similaire aux tests réalisés sur AURA. C'est plutôt rassurant sur la stabilité du modèle. Cependant on constate une variabilité assez importante entre les régions. Evidemment, il y a des variations dues au fait que les régions sont plus ou moins homogènes, avec des données de qualité variables. Mais on peut aussi penser que le modèle a trop de libertés et ne devrait pas aboutir à des situation extrêmes du type "1 seul voisin sélectionné". On peut donc imaginer d'aller vers plus de contraintes (moins de modèles testés). On sait qu'il nous reste aussi une marge significative d'amélioration puisque les variables ne sont pas pondérées à ce stade bien que l'on sache qu'elles n'ont pas toutes la même importance pour la prédiction.



4 ANNEXE



4.1 Auvergne Rhône-Alpes

Matrice de confusion brute

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	607	74	417	393	643	79	1
B	796	285	1670	1665	1076	261	2
C	411	184	3553	10637	2248	841	3
D	490	114	3134	18317	7466	3032	36
E	325	57	1179	9440	10899	4577	58
F	47	13	277	2768	5068	3207	36
G	20	7	63	762	1614	1128	20

Pourcentages en lignes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	27	3	19	18	29	4	0
B	14	5	29	29	19	5	0
C	2	1	20	60	13	5	0
D	2	0	10	56	23	9	0
E	1	0	4	36	41	17	0
F	0	0	2	24	44	28	0
G	1	0	2	21	45	31	1

Pourcentages en colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	23	10	4	1	2	1	1
B	30	39	16	4	4	2	1
C	15	25	35	24	8	6	2
D	18	16	30	42	26	23	23
E	12	8	11	21	38	35	37
F	2	2	3	6	17	24	23
G	1	1	1	2	6	9	13

##

##	Prédictions correctes	Effectif	Part du total
##	+/- 0 étiquettes	36888	0.36888
##	+/- 1 étiquettes	81098	0.81098
##	+/- 2 étiquettes	94604	0.94604
##	+/- 3 étiquettes	98536	0.98536
##	+/- 4 étiquettes	99844	0.99844
##	+/- 5 étiquettes	99979	0.99979
##	+/- 6 étiquettes	100000	1.00000



4.2 Bourgogne Franche-Comté

Matrice de confusion brute

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	4	656	23	203	264	86	1
B	2	1915	122	1259	726	229	7
C	0	763	777	8713	2043	654	3
D	0	691	730	18815	8590	3456	7
E	0	328	219	11052	11494	6246	28
F	0	86	68	3777	5975	5000	23
G	0	26	23	1010	1970	1909	27

Pourcentages en lignes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	0	53	2	16	21	7	0
B	0	45	3	30	17	5	0
C	0	6	6	67	16	5	0
D	0	2	2	58	27	11	0
E	0	1	1	38	39	21	0
F	0	1	0	25	40	33	0
G	0	1	0	20	40	38	1

Pourcentages en colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	67	15	1	0	1	0	1
B	33	43	6	3	2	1	7
C	0	17	40	19	7	4	3
D	0	15	37	42	28	20	7
E	0	7	11	25	37	36	29
F	0	2	3	8	19	28	24
G	0	1	1	2	6	11	28

##

##	Prédictions correctes	Effectif	Part du total
##	+/- 0 étiquettes	38032	0.38032
##	+/- 1 étiquettes	82813	0.82813
##	+/- 2 étiquettes	96279	0.96279
##	+/- 3 étiquettes	99275	0.99275
##	+/- 4 étiquettes	99880	0.99880
##	+/- 5 étiquettes	99999	0.99999
##	+/- 6 étiquettes	100000	1.00000



4.3 Bretagne

Matrice de confusion brute

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	165	769	216	2653	666	12	1
B	159	1793	1376	2796	623	23	0
C	37	336	7477	7926	2568	62	0
D	40	180	3407	18748	11079	180	2
E	7	42	590	11685	13591	195	2
F	4	17	67	3061	5247	84	3
G	0	5	11	615	1449	30	1

Pourcentages en lignes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	4	17	5	59	15	0	0
B	2	26	20	41	9	0	0
C	0	2	41	43	14	0	0
D	0	1	10	56	33	1	0
E	0	0	2	45	52	1	0
F	0	0	1	36	62	1	0
G	0	0	1	29	69	1	0

Pourcentages en colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	40	24	2	6	2	2	11
B	39	57	10	6	2	4	0
C	9	11	57	17	7	11	0
D	10	6	26	39	31	31	22
E	2	1	4	25	39	33	22
F	1	1	1	6	15	14	33
G	0	0	0	1	4	5	11

```
##
## Prédictions correctes Effectif Part du total
## +/- 0 étiquettes 41859 0.41859
## +/- 1 étiquettes 84071 0.84071
## +/- 2 étiquettes 95150 0.95150
## +/- 3 étiquettes 99254 0.99254
## +/- 4 étiquettes 99978 0.99978
## +/- 5 étiquettes 99999 0.99999
## +/- 6 étiquettes 100000 1.00000
```



4.4 Corse

Matrice de confusion brute

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	59	5	11	13	3	2	2
B	4	118	13	18	5	3	1
C	12	9	407	55	19	4	0
D	7	20	57	496	47	13	1
E	4	4	22	31	275	11	1
F	1	1	9	11	9	134	3
G	0	0	1	2	0	4	31

Pourcentages en lignes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	62	5	12	14	3	2	2
B	2	73	8	11	3	2	1
C	2	2	80	11	4	1	0
D	1	3	9	77	7	2	0
E	1	1	6	9	79	3	0
F	1	1	5	7	5	80	2
G	0	0	3	5	0	11	82

Pourcentages en colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	68	3	2	2	1	1	5
B	5	75	3	3	1	2	3
C	14	6	78	9	5	2	0
D	8	13	11	79	13	8	3
E	5	3	4	5	77	6	3
F	1	1	2	2	3	78	8
G	0	0	0	0	0	2	79

##

##	Prédictions correctes	Effectif	Part du total
##	+/- 0 étiquettes	1520	0.7763023
##	+/- 1 étiquettes	1768	0.9029622
##	+/- 2 étiquettes	1895	0.9678243
##	+/- 3 étiquettes	1940	0.9908069
##	+/- 4 étiquettes	1952	0.9969356
##	+/- 5 étiquettes	1956	0.9989785
##	+/- 6 étiquettes	1958	1.0000000



4.5 Centre Val de Loire

Matrice de confusion brute

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	904	394	115	379	403	97	22
B	398	1257	508	1020	584	103	28
C	103	329	4412	5975	1949	249	16
D	102	152	2409	17888	9846	1351	99
E	65	65	586	9524	16897	3407	301
F	20	16	140	2612	7775	3084	344
G	6	8	24	570	2045	1149	270

Pourcentages en lignes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	39	17	5	16	17	4	1
B	10	32	13	26	15	3	1
C	1	3	34	46	15	2	0
D	0	0	8	56	31	4	0
E	0	0	2	31	55	11	1
F	0	0	1	19	56	22	2
G	0	0	1	14	50	28	7

Pourcentages en colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	57	18	1	1	1	1	2
B	25	57	6	3	1	1	3
C	6	15	54	16	5	3	1
D	6	7	29	47	25	14	9
E	4	3	7	25	43	36	28
F	1	1	2	7	20	33	32
G	0	0	0	2	5	12	25

##

##	Prédictions correctes	Effectif	Part du total
##	+/- 0 étiquettes	44712	0.44712
##	+/- 1 étiquettes	86770	0.86770
##	+/- 2 étiquettes	97004	0.97004
##	+/- 3 étiquettes	99192	0.99192
##	+/- 4 étiquettes	99819	0.99819
##	+/- 5 étiquettes	99972	0.99972
##	+/- 6 étiquettes	100000	1.00000



4.6 Grand-Est

Matrice de confusion brute

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	0	361	43	119	76	14	4
B	1	1936	508	1086	554	62	11
C	0	552	3585	7884	2739	303	18
D	0	267	2354	18438	9696	1460	75
E	0	179	491	11238	13765	3194	132
F	0	69	113	3692	7324	2988	89
G	0	29	31	932	2318	1130	140

Pourcentages en lignes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	0	59	7	19	12	2	1
B	0	47	12	26	13	1	0
C	0	4	24	52	18	2	0
D	0	1	7	57	30	5	0
E	0	1	2	39	47	11	0
F	0	0	1	26	51	21	1
G	0	1	1	20	51	25	3

Pourcentages en colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	0	11	1	0	0	0	1
B	100	57	7	3	2	1	2
C	0	16	50	18	8	3	4
D	0	8	33	42	27	16	16
E	0	5	7	26	38	35	28
F	0	2	2	9	20	33	19
G	0	1	0	2	6	12	30

```
##
## Prédictions correctes Effectif Part du total
## +/- 0 étiquettes 40852 0.40852
## +/- 1 étiquettes 85183 0.85183
## +/- 2 étiquettes 97411 0.97411
## +/- 3 étiquettes 99686 0.99686
## +/- 4 étiquettes 99942 0.99942
## +/- 5 étiquettes 99996 0.99996
## +/- 6 étiquettes 100000 1.00000
```




4.7 Hauts de France

Matrice de confusion brute

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	83	478	260	230	114	168	9
B	145	1401	1164	1057	422	362	13
C	58	937	6378	5516	1849	1262	36
D	148	1266	7880	12610	6286	6041	170
E	56	1103	3630	7695	6781	8811	271
F	9	525	997	2338	2915	5084	209
G	3	170	198	596	847	1337	82

Pourcentages en lignes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	6	36	19	17	8	13	1
B	3	31	26	23	9	8	0
C	0	6	40	34	12	8	0
D	0	4	23	37	18	18	0
E	0	4	13	27	24	31	1
F	0	4	8	19	24	42	2
G	0	5	6	18	26	41	3

Pourcentages en colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	17	8	1	1	1	1	1
B	29	24	6	4	2	2	2
C	12	16	31	18	10	5	5
D	29	22	38	42	33	26	22
E	11	19	18	26	35	38	34
F	2	9	5	8	15	22	26
G	1	3	1	2	4	6	10

```
##
## Prédictions correctes Effectif Part du total
## +/- 0 étiquettes 32419 0.32419
## +/- 1 étiquettes 75792 0.75792
## +/- 2 étiquettes 93409 0.93409
## +/- 3 étiquettes 98337 0.98337
## +/- 4 étiquettes 99628 0.99628
## +/- 5 étiquettes 99988 0.99988
## +/- 6 étiquettes 100000 1.00000
```



4.8 Ile de France

Matrice de confusion brute

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	201	628	37	295	440	126	35
B	204	1555	313	962	662	157	37
C	23	337	2398	5913	1993	226	18
D	35	220	1625	20056	11051	1359	156
E	19	111	331	11980	15489	2848	332
F	11	54	98	3642	7126	2349	325
G	5	18	12	803	2100	1049	236

Pourcentages en lignes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	11	36	2	17	25	7	2
B	5	40	8	25	17	4	1
C	0	3	22	54	18	2	0
D	0	1	5	58	32	4	0
E	0	0	1	39	50	9	1
F	0	0	1	27	52	17	2
G	0	0	0	19	50	25	6

Pourcentages en colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	40	21	1	1	1	2	3
B	41	53	7	2	2	2	3
C	5	12	50	14	5	3	2
D	7	8	34	46	28	17	14
E	4	4	7	27	40	35	29
F	2	2	2	8	18	29	29
G	1	1	0	2	5	13	21

##

##	Prédictions correctes	Effectif	Part du total
##	+/- 0 étiquettes	42284	0.42284
##	+/- 1 étiquettes	85683	0.85683
##	+/- 2 étiquettes	96682	0.96682
##	+/- 3 étiquettes	99068	0.99068
##	+/- 4 étiquettes	99768	0.99768
##	+/- 5 étiquettes	99960	0.99960
##	+/- 6 étiquettes	100000	1.00000



4.9 Nouvelle Aquitaine

Matrice de confusion brute

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	2096	125	535	1909	297	568	14
B	622	152	1174	2955	386	309	6
C	435	213	4319	11834	2709	1330	41
D	469	136	4837	20291	7149	3986	77
E	94	36	1630	11620	5578	3328	75
F	27	7	354	3222	1894	1457	46
G	10	3	64	799	343	423	16

Pourcentages en lignes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	38	2	10	34	5	10	0
B	11	3	21	53	7	6	0
C	2	1	21	57	13	6	0
D	1	0	13	55	19	11	0
E	0	0	7	52	25	15	0
F	0	0	5	46	27	21	1
G	1	0	4	48	21	26	1

Pourcentages en colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	56	19	4	4	2	5	5
B	17	23	9	6	2	3	2
C	12	32	33	22	15	12	15
D	12	20	37	39	39	35	28
E	3	5	13	22	30	29	27
F	1	1	3	6	10	13	17
G	0	0	0	2	2	4	6

```
##
## Prédictions correctes Effectif Part du total
## +/- 0 étiquettes 33909 0.33909
## +/- 1 étiquettes 77174 0.77174
## +/- 2 étiquettes 93200 0.93200
## +/- 3 étiquettes 98560 0.98560
## +/- 4 étiquettes 99372 0.99372
## +/- 5 étiquettes 99976 0.99976
## +/- 6 étiquettes 100000 1.00000
```



4.10 Normandie

Matrice de confusion brute

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	0	1063	19	242	801	14	0
B	0	2245	324	1495	994	23	0
C	0	708	1229	8487	2449	30	1
D	0	989	910	17826	12543	139	1
E	0	402	161	8605	21014	254	8
F	0	88	34	2383	10531	170	3
G	0	33	6	475	3237	63	1

Pourcentages en lignes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	0	50	1	11	37	1	0
B	0	44	6	29	20	0	0
C	0	5	10	66	19	0	0
D	0	3	3	55	39	0	0
E	0	1	1	28	69	1	0
F	0	1	0	18	80	1	0
G	0	1	0	12	85	2	0

Pourcentages en colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	NaN	19	1	1	2	2	0
B	NaN	41	12	4	2	3	0
C	NaN	13	46	21	5	4	7
D	NaN	18	34	45	24	20	7
E	NaN	7	6	22	41	37	57
F	NaN	2	1	6	20	25	21
G	NaN	1	0	1	6	9	7

```
##
## Prédictions correctes Effectif Part du total
## +/- 0 étiquettes 42485 0.42485
## +/- 1 étiquettes 85976 0.85976
## +/- 2 étiquettes 96856 0.96856
## +/- 3 étiquettes 99034 0.99034
## +/- 4 étiquettes 99953 0.99953
## +/- 5 étiquettes 100000 1.00000
```



4.11 Occitanie

Matrice de confusion brute

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	2537	145	812	1803	539	29	4
B	682	410	2369	1693	277	16	0
C	923	329	10881	10604	1579	56	4
D	551	124	5994	23599	5908	266	19
E	152	42	1273	11372	6781	416	41
F	47	12	244	2825	2831	270	37
G	27	5	49	613	653	124	33

Pourcentages en lignes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	43	2	14	31	9	0	0
B	13	8	43	31	5	0	0
C	4	1	45	44	6	0	0
D	2	0	16	65	16	1	0
E	1	0	6	57	34	2	0
F	1	0	4	45	45	4	1
G	2	0	3	41	43	8	2

Pourcentages en colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	52	14	4	3	3	2	3
B	14	38	11	3	1	1	0
C	19	31	50	20	9	5	3
D	11	12	28	45	32	23	14
E	3	4	6	22	37	35	30
F	1	1	1	5	15	23	27
G	1	0	0	1	4	11	24

```
##
## Prédictions correctes Effectif Part du total
## +/- 0 étiquettes 44511 0.44511
## +/- 1 étiquettes 85322 0.85322
## +/- 2 étiquettes 95511 0.95511
## +/- 3 étiquettes 99116 0.99116
## +/- 4 étiquettes 99888 0.99888
## +/- 5 étiquettes 99969 0.99969
## +/- 6 étiquettes 100000 1.00000
```



4.12 Provence Alpes Côte d Azur

Matrice de confusion brute

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	1233	206	953	1441	300	20	3
B	331	583	2575	2066	302	15	0
C	414	546	9517	12814	1902	80	3
D	240	232	7107	24664	5471	321	8
E	97	80	1981	12199	4553	322	10
F	37	22	474	3545	1709	158	5
G	11	6	86	843	474	41	0

Pourcentages en lignes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	30	5	23	35	7	0	0
B	6	10	44	35	5	0	0
C	2	2	38	51	8	0	0
D	1	1	19	65	14	1	0
E	1	0	10	63	24	2	0
F	1	0	8	60	29	3	0
G	1	0	6	58	32	3	0

Pourcentages en colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	52	12	4	3	2	2	10
B	14	35	11	4	2	2	0
C	18	33	42	22	13	8	10
D	10	14	31	43	37	34	28
E	4	5	9	21	31	34	34
F	2	1	2	6	12	17	17
G	0	0	0	1	3	4	0

##

##	Prédictions correctes	Effectif	Part du total
##	+/- 0 étiquettes	40708	0.40708
##	+/- 1 étiquettes	84034	0.84034
##	+/- 2 étiquettes	95932	0.95932
##	+/- 3 étiquettes	99400	0.99400
##	+/- 4 étiquettes	99923	0.99923
##	+/- 5 étiquettes	99986	0.99986
##	+/- 6 étiquettes	100000	1.00000



4.13 Pays de Loire

Matrice de confusion brute

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	3217	361	283	1134	582	59	6
B	1001	565	1238	2054	485	85	1
C	472	483	7082	7779	3172	253	11
D	664	351	3867	18180	11012	1037	54
E	152	103	541	12517	10088	1839	66
F	40	12	92	3848	2561	910	34
G	20	4	10	871	552	240	12

Pourcentages en lignes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	57	6	5	20	10	1	0
B	18	10	23	38	9	2	0
C	2	3	37	40	16	1	0
D	2	1	11	52	31	3	0
E	1	0	2	49	40	7	0
F	1	0	1	51	34	12	0
G	1	0	1	51	32	14	1

Pourcentages en colonnes

Vraies valeurs	Valeurs prédites						
	A	B	C	D	E	F	G
A	58	19	2	2	2	1	3
B	18	30	9	4	2	2	1
C	8	26	54	17	11	6	6
D	12	19	29	39	39	23	29
E	3	5	4	27	35	42	36
F	1	1	1	8	9	21	18
G	0	0	0	2	2	5	7

```
##
## Prédictions correctes Effectif Part du total
## +/- 0 étiquettes 40054 0.40054
## +/- 1 étiquettes 82986 0.82986
## +/- 2 étiquettes 95362 0.95362
## +/- 3 étiquettes 99018 0.99018
## +/- 4 étiquettes 99870 0.99870
## +/- 5 étiquettes 99974 0.99974
## +/- 6 étiquettes 100000 1.00000
```

Table of contents

	Acknowledgements	iii
	Foreword	v
	Short table of contents	vii
	Acronyms	ix
	Mathematical notations	xi
	Specific variables	xii
	Introduction (en français)	1
1	Contexte institutionnel et environnemental	2
2	Vue d'ensemble du projet de recherche	3
3	Méthodologie et structure générale du document	6
4	État des lieux	9
5	État de l'art	17
	Introduction	27
1	Institutional and Environmental Context	28
2	Overview of the Research Project	29
3	Methodology and Overall Structure	32
4	State of Things	34
5	State of the Art	42
I	Pre-processing Data	51
1	Data Fusion and Uncertainties	52
2	Imputation of Missing Values	55
3	From Categories to Numbers	59
4	Ranks and Quantiles	60
5	Nearest Neighbours and Distance	61
	Conclusion	61
II	Handling Multi-Scale Uncertainties	65
	Résumé en français	66
	Abstract	67
1	EPC Prediction Problem	67
2	Optimal Interpolation	68
3	Illustration	76
	Discussion and Conclusion	88

III	Constrained Classification	91
	Résumé en français	94
1	Introduction	95
2	Joint Kriging Model	100
3	Constrained Classification	110
4	Filling Cross-Covariances	114
5	Numerical Illustrations	117
6	Conclusion	132
IV	Enhancing Prediction's Performances	135
	Résumé en français	136
1	Introduction	137
2	Data Presentation	139
3	Methodology	144
4	Results and Discussion	150
5	Conclusion	153
	Conclusion and Future Directions	155
1	An Improved Knowledge...	156
2	Significance of These Results for Decision Makers.	157
3	Limitations	158
4	Future directions	159
	Afterword	161
	Conclusion et perspectives (en français)	163
1	Une meilleure connaissance...	164
2	Importance de ces résultats pour les décideurs	166
3	Limites	166
4	Perspectives	167
	Bibliography	169
	Supplementary Material	181
	Supplementary Material for Pre-processing data	181
1	About EPCs Uncertainties	181
2	Mathematical Presentation of Variables Normalisation	195
3	Normalised Pseudo-Observations	199
4	Distribution of Distances According to the Data Distribution	201
5	Supplementary Material About the Hassanat Distance	209
	Supplementary Material for Mixture Kriging	221
6	Proof of Proposition 2	222
7	Cross-Errors	225

8	Operations on Granularities227
9	Detailed Implementation of Mixture Kriging for EPCs231
	Supplementary Material for Joint Kriging245
10	Proofs246
	Supplementary Material for Fuzzy Classification263
11	Dictionary of Main Variables264
12	Details of Selected Variables per Model265
13	Confusion Matrices.266
	Supplementary Material for the Conclusion269
14	U.R.B.S. model 2022.270
	Table of Contents297
	List of Tables.298
	List of Figures300

List of Tables

Handling Multi-Scale Uncertainties	65
II.1 Parameters and performances of Kriging and Mixture Kriging with rounded inputs	76
II.2 Observations of the simulated Gaussian random field.	77
II.3 Comparison of models quality for different types of granularities.	79
II.4 Optimal parameters for $M1$ and $M2$	84
II.5 Optimal performances achieved by $M1$ and $M2$ with 3 input variables and no output covariate.	84
II.6 Confusion matrix of $M1$ predictions.	85
II.7 Confusion matrix of $M2$ predictions.	85
II.8 Confusion matrix of $M1'$ predictions	85
Constrained Classification	91
III.1 A Constrained Confusion Matrix	111
Enhancing Prediction's Performances	135
IV.1 Features identified as potentially useful to predict the EPC.	148
IV.2 Performances of the 3 compared models for fuzzy classification.	152
IV.3 Performances of the 3 compared models for hard classification.	152
IV.4 Fuzzy classification in relation with true labels.	154
IV.5 Confusion matrix of the Fuzzy KNN model.	266
IV.6 Confusion matrix of the binarised KNN model, used for hard classification.	266
IV.7 Confusion matrix of the Fuzzy Joint Kriging model.	267
IV.8 Confusion matrix of the binarised Joint Kriging model, used for hard classification.	267
IV.9 Confusion matrix of Random Forest hard classification.	268

List of Figures

	Introduction (en français)	1
1	Comparaison entre le parc immobilier des DPE observés et le parc immobilier total en France	12
2	L'étiquette qui donne le résultat d'un diagnostic de performance énergétique.	13
3	Le processus de double seuil utilisant la consommation d'énergie primaire et les émissions de GES pour le calcul de l'étiquette DPE en France depuis 2021.	13
4	Un résumé visuel de la loi portant lutte contre le dérèglement climatique.	16
5	Carte de France des DPE répertoriés.	18
6	Première carte choroplèthe, par Dupin.	24
7	Carte choroplèthe des passoires thermiques.	25
	Introduction	27
8	Observed EPCs vs. total building stock in France: proportions of the populations built before or after 2000, that are houses or flats.	37
9	The French official vignette, which gives the result of the energy efficiency diagnostic.	38
10	The double thresholding process involving primary energy consumption and GHG emissions for computing the EPC label in France since 2021.	38
11	A visual summary of the Law on combating climate change.	41
12	Map of French inventoried EPCs	43
13	First choropleth map by Dupin.	48
14	Chloropleth map of energy sieves.	49
	Pre-processing Data	51
I.1	Uncertainty of the addresses in the EPCs database	53
I.2	Representation of the algorithm used to match EPCs database addresses with MoF addresses.	54
I.3	Map of Saint-Etienne showing the effects of variables normalisation	63
I.4	Effect of variables normalisation on an integer variable.	64
	Handling Multi-Scale Uncertainties	65
II.1	Distribution of EPCs collected from 2014 to 2021	71
II.2	Kriging and Mixture Kriging predictions for rounded inputs	78
II.3	Mixture Kriging and varying grain sizes	80
II.4	Map of buildings' construction years in an urban area in Angers	82
II.5	Map of EPC observations in an urban area in Angers	87
II.6	Map of all predicted values	87
	Constrained Classification	91
III.1	Joint Kriging prediction with one constraint	118
III.2	Joint Kriging prediction with two constraints	119

III.3	optimisation of the correlation hyperparameter for Air quality data set	122
III.4	Joint Kriging interpolation for the Air quality dataset	123
III.5	Prediction of air quality with Joint Kriging in an adverse scenario	124
III.6	Earthquakes observations	126
III.7	Joint Kriging interpolation for earthquakes	129
III.8	Prediction of earthquakes in an adverse scenario	129
III.9	Distribution of earthquakes prediction performances	130
III.10	Distribution of OpenML predictive accuracies for the Quake dataset	130
III.11	Earthquakes prediction with Joint Kriging using four classes	131
	Enhancing Prediction's Performances	135
IV.1	Schema of all data sources that are merged in IMOPE	141
IV.2	Double threshold process used to determine the French EPC label of a building.	142
IV.3	A problematic case for matching EPC observations with MoF database	143
IV.4	Histogram of observed energy consumption	143
IV.5	Optimal features' weights for the KNN algorithm	147
	Supplementary Material	181
1	Effect of pseudo observations as compared to raw observations (scale is raw values).	202
2	Effect of pseudo observations as compared to raw observations(scale is pseudo-observations values).	203
3	Geometrical interpretation of Mixture Kriging prediction process	222
4	Thinner granularity and maximal thinner non-overlapping granularity	227
5	Comparing levelplots of likelihood between a bivariate mixture and a bivariate Gaussian vector with same covariance matrix.	234
6	Addresses granularity description	241

Résumé

Cette thèse s'intéresse à la question de l'amélioration de l'efficacité énergétique des bâtiments en France, essentielle pour atteindre les objectifs nationaux de durabilité et réduire les émissions de gaz à effet de serre. Le principal objectif est de développer un modèle prédictif explicable pour estimer l'efficacité énergétique de chaque bâtiment en France en utilisant les **DPE** (*Diagnostic de Performance Énergétique – Energy Performance Certificate*) observés. L'étude s'attache à dépasser les difficultés posées par des données disponibles incertaines et incomplètes. Le premier chapitre décrit le processus de fusion des données mis en œuvre pour rassembler toutes les informations disponibles sur les logements dans une base de données unique. Le deuxième chapitre propose un modèle de prédiction qui prend en compte l'incertitude des emplacements des bâtiments. Ce modèle de régression, appelé Mixture Kriging, met en œuvre une approche de Krigeage avec des distributions de mélange. Il donne de bons résultats à l'échelle d'une ville mais est difficile à étendre à l'échelle du pays en raison de la quantité de calculs nécessaires. Dans le troisième chapitre, l'approche est orientée vers une approche de classification contrainte. Le modèle Joint Kriging réalise une classification floue sous deux contraintes simultanées : la somme des poids utilisés pour la régression est égale à 1, et la moyenne de toutes les prédictions peut être prescrite. Bien qu'il ne prenne pas en compte l'incertitude de position, ce modèle fournit efficacement des prédictions précises à grande échelle. Enfin, le quatrième chapitre se concentre sur l'amélioration des performances de prédiction en comparant Joint Kriging, **FKNN** (*Fuzzy k-Nearest Neighbours*) et Random Forest. Et les résultats sont prometteurs.

Ce travail de recherche démontre que prédire l'efficacité énergétique des bâtiments est faisable en utilisant des approches géostatistiques. Il souligne aussi l'importance de prendre en compte les données de géolocalisation. L'étude suggère que les **DPE** peuvent être considérés non seulement comme des résultats de calculs d'ingénierie thermique, mais aussi comme des données géolocalisées. Ce travail propose également des modèles explicables et adaptés à une mise en œuvre dans le logiciel **ONB** (*Observatoire National des Bâtiments – National Buildings Observatory*) d'**U.R.B.S.**. Les modèles prédictifs aident à identifier les bâtiments énergétiquement inefficaces (« passoires énergétiques ») sans inspections physiques, réalisant des améliorations significatives dans les taux de détection. Cette thèse contribue à la fois aux sciences environnementales et aux mathématiques appliquées en fournissant des méthodes innovantes pour évaluer et améliorer la performance énergétique du parc immobilier en France. Les modèles développés sont destinés à aider les décideurs, en particulier les collectivités locales, à prendre des décisions éclairées pour atteindre les objectifs d'efficacité énergétique et de durabilité.

Mots clés – Efficacité énergétique, bâtiments, data fusion, krigeage, classification floue, durabilité.

Abstract

This thesis is concerned with the issue of improving the energy efficiency of buildings in France, which is essential for meeting national sustainability goals and reducing greenhouse gas emissions. The primary objective is to develop an explainable predictive model for estimating the energy efficiency of every building in France using observed **EPC (Energy Performance Certificate)**. The study focuses on overcoming the challenges posed by the uncertain and incomplete nature of available data. The first chapter describes the data fusion process implemented to gather all available data about dwellings in a single data base. The second chapter proposes a prediction model that takes into account the uncertainty of buildings' locations. This regression model, called Mixture Kriging implements a Kriging approach with mixture distributions. It performs well at the city scale but is challenging to upscale nationally due to computational complexity. In the third chapter, the approach is shifted towards a constrained classification methodology. The Joint Kriging model performs a fuzzy classification under two simultaneous constraints: the weights used for the regression sum to 1 and the average over all predictions can be prescribed. Although it does not account for positional uncertainty, this model offers efficient and accurate predictions at large scales. Eventually, the fourth chapter focuses on enhancing prediction's performance by comparing Joint Kriging, **FKNN (Fuzzy k-Nearest Neighbours)** and Random Forest with promising results.

This research demonstrates the feasibility of predicting building energy efficiency using geostatistical approaches and stresses the importance of considering geolocation data. The study suggests that **EPCs** can be viewed not only as engineering outputs but also as geolocated data. This work also provides models that are explainable, and suitable for implementation in the **U.R.B.S.** software **ONB (Observatoire National des Bâtiments – National Buildings Observatory)**. The predictive models help identify energy-inefficient buildings ("energy sieves") without physical inspections, achieving significant improvements in detection rates. This thesis contributes to both fields of environmental sciences and applied mathematics by providing innovative methods to evaluate and improve the energy performance of the building stock in France. The models developed are expected to assist policymakers and urban planners in making informed decisions for achieving energy efficiency and sustainability targets.

Keywords – Energy efficiency, buildings, data fusion, Kriging, fuzzy classification, sustainability.