Aix-Marseille Université
CESAER

Master 2 Economie
Parcours Econometrie, Big Data - Statistique

19 SEPTEMBER 2024

Development of a Composite Indicator to measure real estate tension in France and improve current zoning

ELEONORA FIORENTINO

Head of track:
Prof. Christian Schluter

Internship supervisors:
Prof. Julie Le Gallo
Research director Marie Breuillé

Overview

CONTEXT

MISSION

RESULTS

CONCLUSION

The context

THE ZONING SYSTEM

What is zoning, why it needs to be redesigned

ASSIGNMENT OF THE TASK

Who asked for this project, how was the team chosen, CESAER

THE COMPOSITE INDICATOR

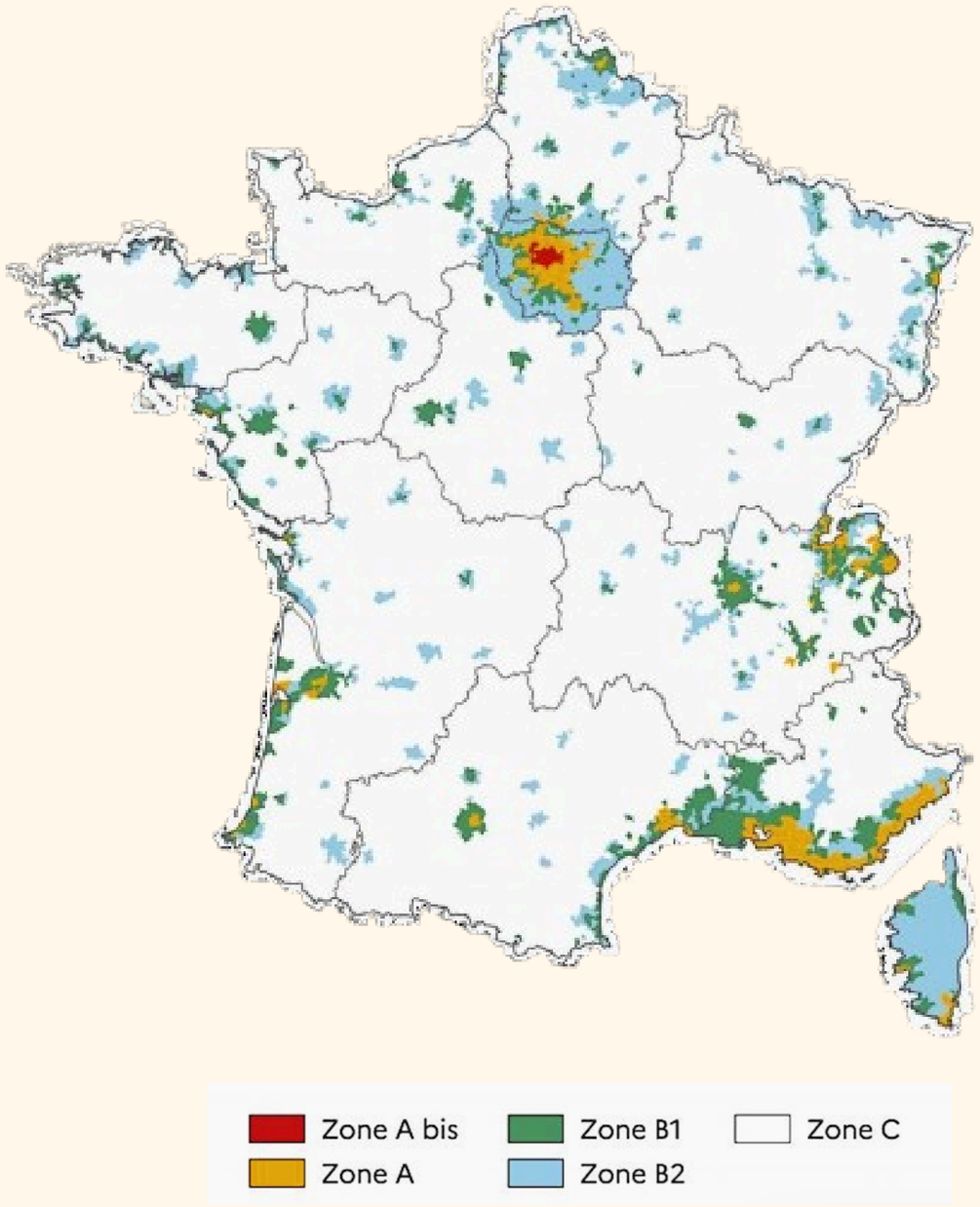
What is a CI, construction phases, applications

The zoning system

WHAT IS ZONING Since 2003 for all questions regarding housing policies, France has been divided into tension zones. The zoning is a "classification of municipalities in the national territory into geographical zones based on the imbalance between supply and demand for housing".

WHEN WAS IT MADE Initially established in 2003, the ABC zoning was revised in 2006, 2009, and 2014. Then it has been subject to three partial reviews in 2019, 2022, and the last in October 2023.

WHY IT NEEDS TO BE CHANGED The real problem with this zoning lies in the way it was created: few, rough indicators but above all a lack of transparency.



Assignment of the task

In 2023 the Court of Auditors strongly criticized this zoning and recommended that the A/B/C zoning be "updated in order to better reflect current real estate tensions".

The Ministry contacted CESAER, and in particular Julie Le Gallo and Marie Breuillé, for the construction of the new zoning, because of their experience with housing public policies (evaluation of the impact of rent control, rental indicators and many more) and also because they had already previously worked with the Ministry of Housing.

CESAER is a mixed research unit that brings together teacher-researchers from the Institut Agro Dijon and researchers from INRAE.

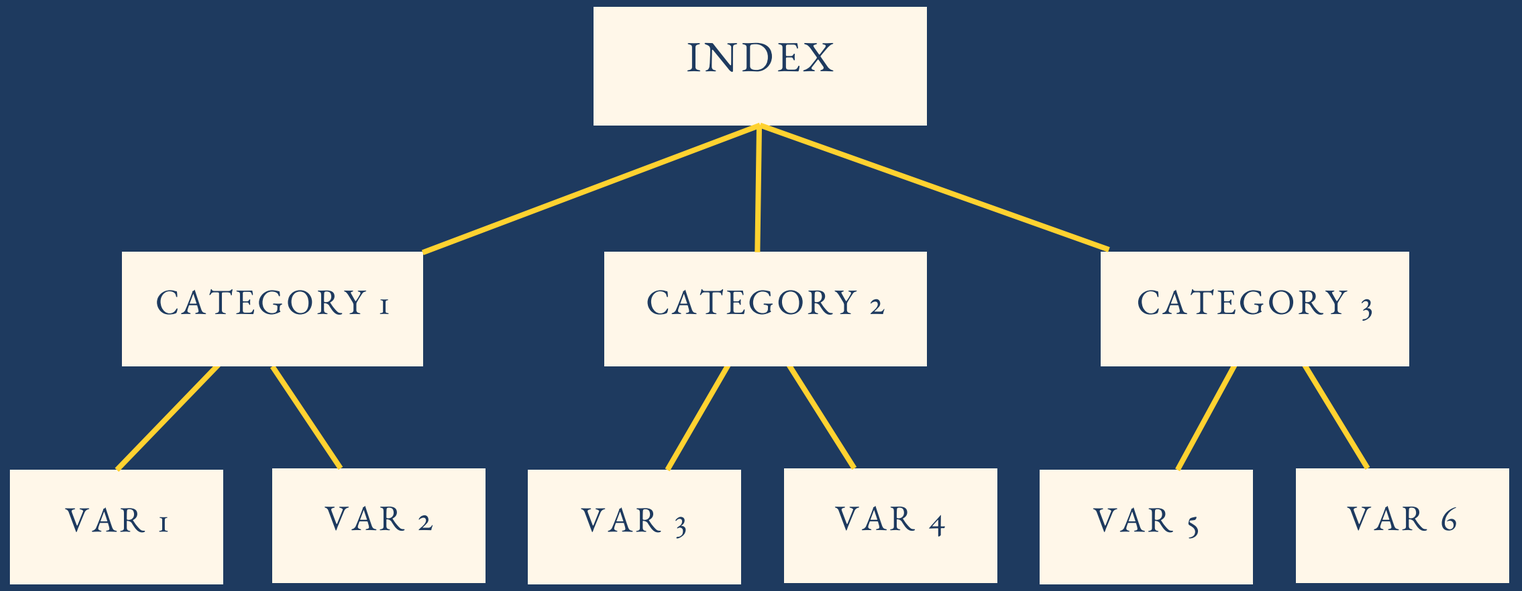
The team

The team was made up of 6 people, specifically:

- Julie Le Gallo, Economics Professor at the Institut Agro Dijon, CESAER
- Marie Breuillé, Research director in economics, INRAE, CESAER
- Camille Grivault, Geographer (self-employed)
- Martin Regnaud, a CIFRE PhD student at CESAER who works for Se Loger/MeilleursAgents
- Aldo Caumo, student of the Aix-Marseille University doing the M2 Economie: Parcours Econometrie, big data - statistique.
- Me, Eleonora Fiorentino, student of the Aix-Marseille University doing the Master 2 Economie: Parcours Econometrie, big data - statistique.

The Composite Indicator

Composite indicators (CI) are tools that can have multiple purposes. In our case it was used to assign to each municipality a continuous value, weighing and aggregating all the characteristics that define the concept of real estate tension, and using these values to then create subdivisions of these municipalities (the new zones).



CONSTRUCTION STEPS

1. A quality framework
2. Variable selection
3. Imputation of Missing Data
4. Multivariate Analysis
5. Normalization
6. Weighting
7. Aggregation
8. Sensitivity Analysis

COMPOSITE INDICATOR

Applications of the CI

Unlike a PCA, which focuses on maximizing the variance of the data through linear combinations, or a clustering algorithm, which groups the data without creating a unified index, CIs allow for flexible, easier to interpret and clearer results. This flexibility makes CIs usable in a variety of contexts

Field	Composite Indicator	Purpose
Economics	Human Development Index (HDI)	Measures key dimensions of human development, such as health, education, and income, to assess the overall well-being and development of countries.
Environmental Science	Environmental Performance Index (EPI)	Assesses countries' environmental health and ecosystem vitality by measuring factors like air quality, water resources, and biodiversity.
Social Sciences	Social Progress Index (SPI)	Evaluates the social and environmental conditions of nations, focusing on human well-being beyond economic metrics.
Health	Health Care Access and Quality Index (HAQ)	Assesses the quality and accessibility of healthcare services worldwide, reflecting how well healthcare systems meet population needs.
Education	Education Index	Tracks progress in education by measuring expected years of schooling for children and the mean years of schooling for adults, reflecting educational attainment.
Business and Finance	Dow Jones Industrial Average (DJIA)	Provides a snapshot of the U.S. stock market's performance by tracking 30 large publicly traded companies.
Technology	Digital Economy and Society Index (DESI)	Measures digital performance in European countries, focusing on connectivity, human capital, use of internet services, and integration of digital technology.

The mission

VARIABLES AND CATEGORIES

Conceptual framework, variable selection and the creation of the dataset

PRELIMINARY ANALYSIS

Impact of variables, COINr package, imputation of Missing Data, removing outliers

MULTIVARIATE ANALYSIS

Correlation, K-means, Spectral clustering, Random Forest

INSIDE COINR

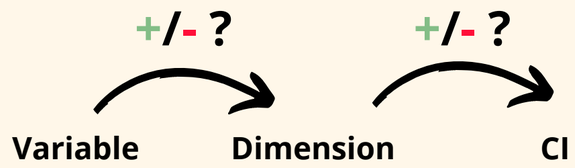
Normalization, weighting, aggregation

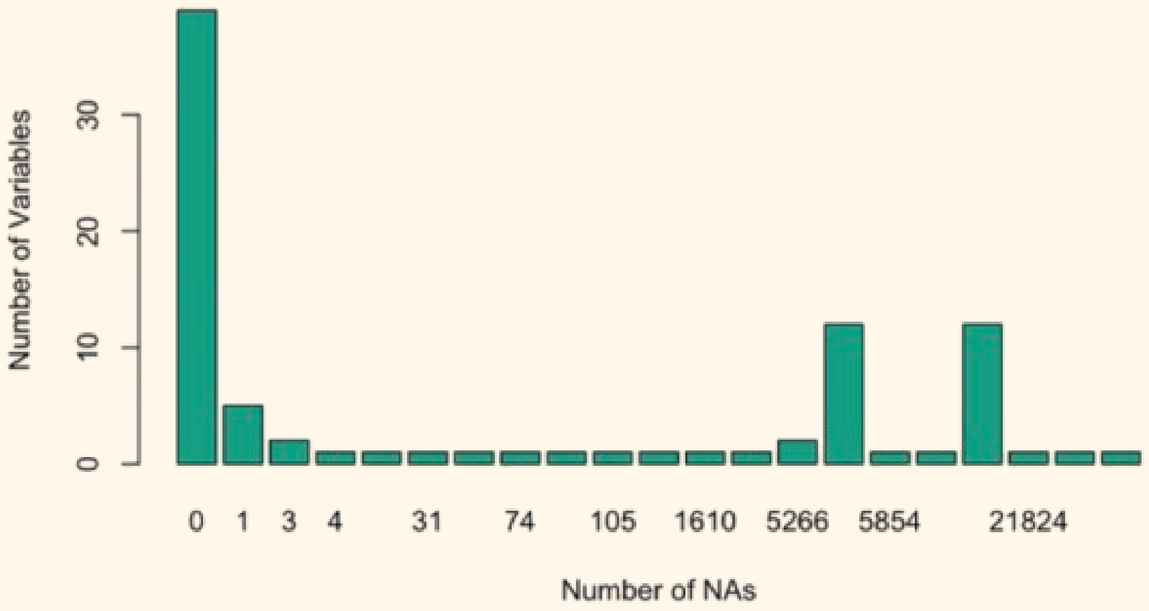
Variables and Categories

DATA SOURCES: COG, INSEE, RP, Fichier détail logement, DV3F, Cerema, Filosofi, Carte des Loyers, Anil, Sitadel2, Sdes, fichier détail mobilité professionnelle, Lovac, La Poste, rééxpéditions de courrier, Airdna, REE, Observatoire DPE - AUDIT, Ademe and AVIV (group that owns and manages some of the main real estate portals in France, including SeLoger and MeilleursAgents).

Preliminary analysis

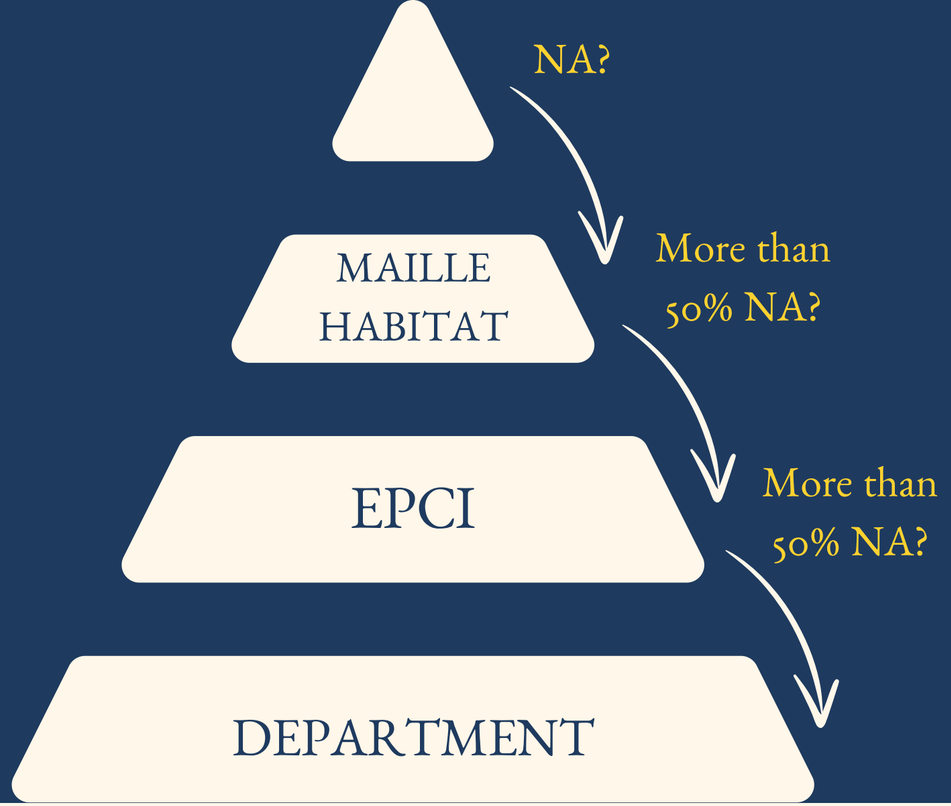
- Impact of variables
- Imputation of NAs
- Outliers





Total number of variables: 87
Number of variables with 0 NAs: 39
Number of variables with more than 100 NAs: 36

Imputation of missing data



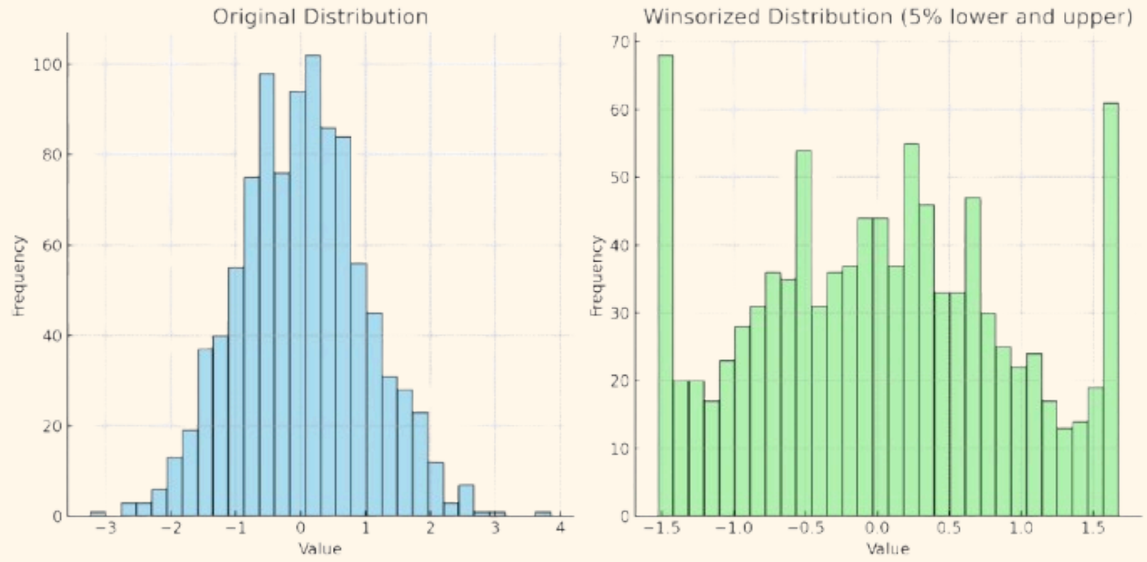
Reasons for NAs:

- Statistical secrecy for small municipalities and islands
- Legal status: departments of Moselle (57), Bas-Rhin (67) and Haut-Rhin (68).
- Lack of data

Removing outliers

“Numbers over 500 can normally be used with confidence.
Numbers below 200 should be treated with caution, as they may not be significant due to the imprecision of the survey. Comparisons between small territories should be avoided”. (INSEE)

Winsorization was applied only to 7 variables, and only to municipalities with less than 500 inhabitants

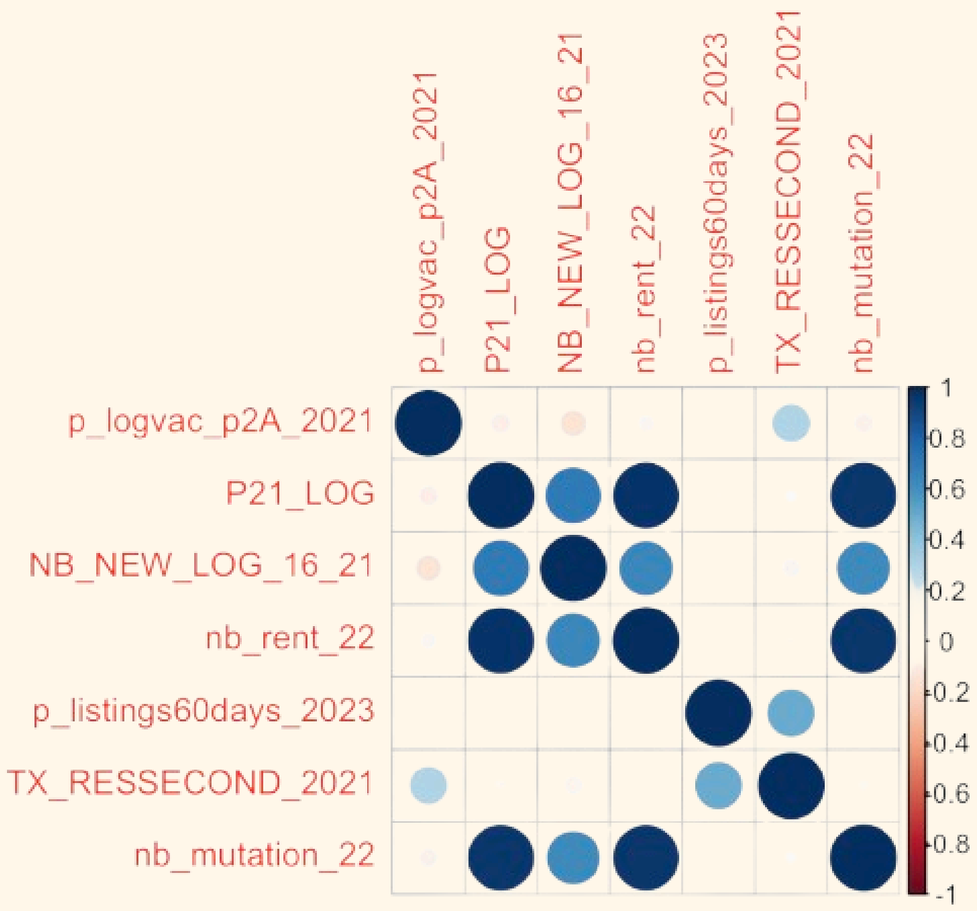


The left plot displays the original data, which includes the natural outliers from the normal distribution.

The right plot shows the winsorized distribution, where the extreme values have been capped at the 5th and 95th percentiles, reducing the impact of outliers on the data.

This method preserves the overall shape while limiting the influence of extreme values

Multivariate analysis



Correlation

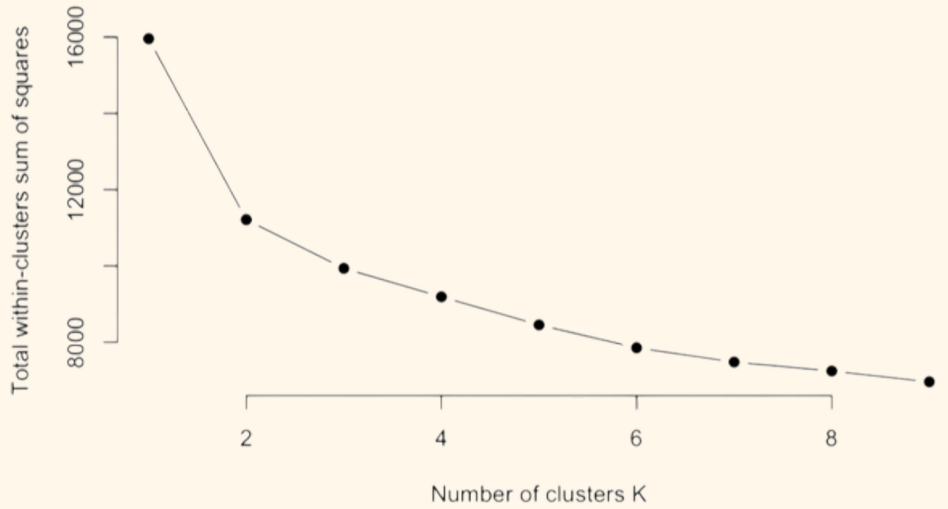
A first correlation analysis was carried out by studying the level of correlation between variables belonging to the same category. This produced 9 correlation graphs.
However, we are clearly dealing with a biased correlation: many variables have in fact been constructed as a ratio between two other variables present in the dataset, or there are distinct variables but divided by the same variable used as a normalizer. Because of this spurious and non-linear correlation, the real relationship between the variables and their influence on the category cannot be clearly understood from this correlation graph.

K-means

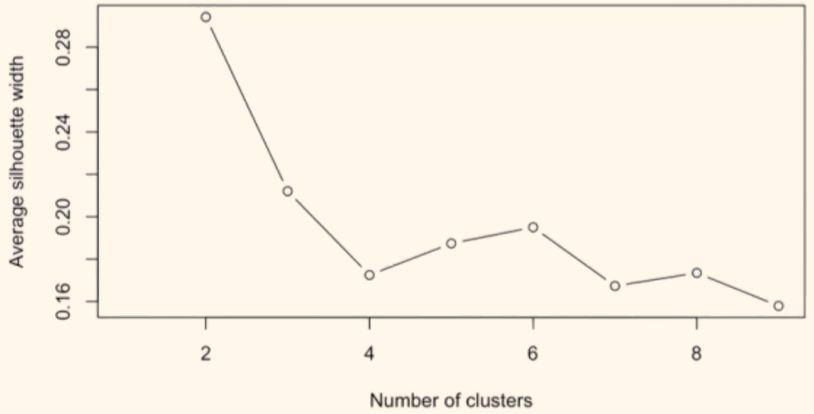
Zone	Abis	A	B1	B2	C
Cluster1	99	410	462	221	52
Cluster2	0	21	152	337	5710
Cluster3	25	359	843	574	5742
Cluster4	0	2	41	437	7278
Cluster5	0	14	599	1586	9839

K-means with 5 clusters

Elbow method



Siholouette method



K-means 2

Zone	Abis	A	B1	B2	C
Cluster1	25	384	1005	903	11451
Cluster2	99	422	1092	2252	17170

Spectral clustering

Zone	Abis	A	B1	B2	C
Cluster1	5	109	37	275	273
Cluster2	0	4	51	135	2133
Cluster3	0	0	26	39	991
Cluster4	0	0	8	48	1539
Cluster5	27	15	9	0	2

15% of the dataset

- Effective in identifying clusters that are not linearly separable
- High Computational Cost

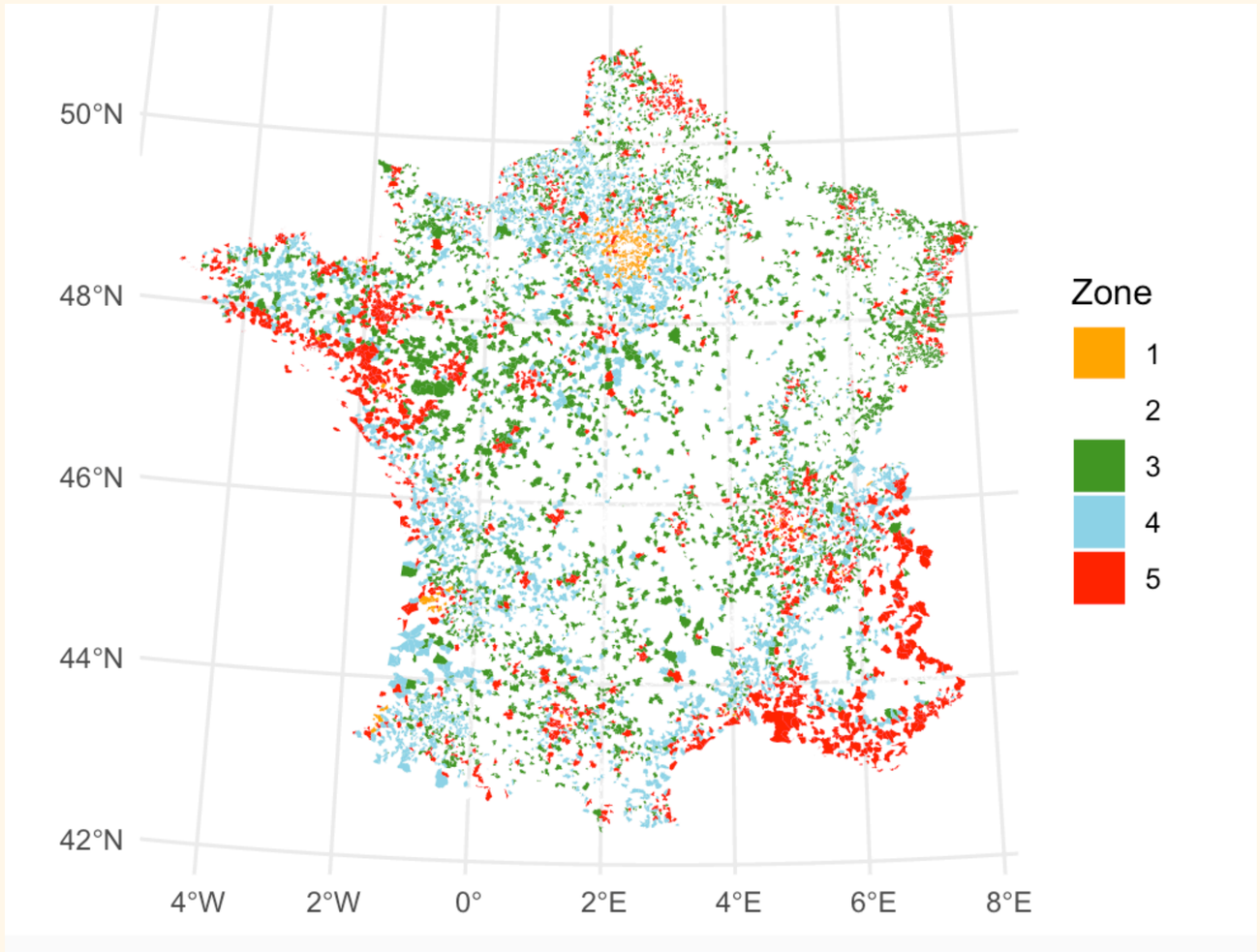
Random Forest

Zone	Abis	A	B1	B2	C
Cluster1	16	223	574	275	273
Cluster2	0	2	6	44	6409
Cluster3	0	2	104	365	2069
Cluster4	4	73	344	778	5521
Cluster5	41	95	4	0	0

50% of the dataset

- Work well with high-dimensional data
- Can capture complex, non-linear relationships between variables
- Robust to noise and outliers
- For large dataset can be memory and time-intensive

MAP FROM THE RANDOM FOREST



From this spatial representation, it is possible to notice 2 things: first, that the areas on the south coast, in particular the Provence-Alpes-Côte d'Azur region and the Alps on the border with Italy, as well as the coasts of Bretagne, are in the same cluster and represent rather tense territories.

Second, it is also clearly visible the diagonal that goes from Spain to Belgium, which notoriously has a much lower population density than the other areas of France.

This result, although obtained using only half the dataset (this is why Paris does not appear among the red zones), still presents very interesting information and certainly with greater computing power and a few more analyses it could produce an even more precise classification.

INSIDE COINr

Normalization

The algorithm we decided to apply, and which I had already used in the previous section for k-means and spectral clustering, is the "min-max" algorithm. Min-max normalization scales the data to a fixed interval, typically [0, 1], in our case [1,100] to be suitable with the chosen aggregation method. This is achieved by subtracting the minimum value of each feature and dividing by the interval (maximum - minimum).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Weighting

A first choice to make when assigning weights is whether to use equal weights (EW) or different weights (DW).
In our specific case, we decided to opt for Equal weights for our first analysis.
Reasons:

- Simplicity and Transparency
- Useful in cases of lack of clear relationship between variables
- Avoid Subjectivity and Bias
- Robustness and Comparability

Aggregation

Aggregation methods can be divided into three main categories: linear, geometric, and multi-criteria.
Another crucial point in the aggregation phase is the choice betweena “compensatory”, “partially compensatory” and “non-compensatory” approach.
In our project, we decided early on to adopt a partially or non-compensatory approach and to start with the most common aggregation functions, present in the COINr package: arithmetic mean, geometric mean, and harmonic mean.
In this first result, I decided to choose the method that compensates the least among the three, namely the harmonic mean.

Results

JENKS'S ALGORITHM

Jenks's algorithm for partitions

SENSITIVITY ANALYSIS

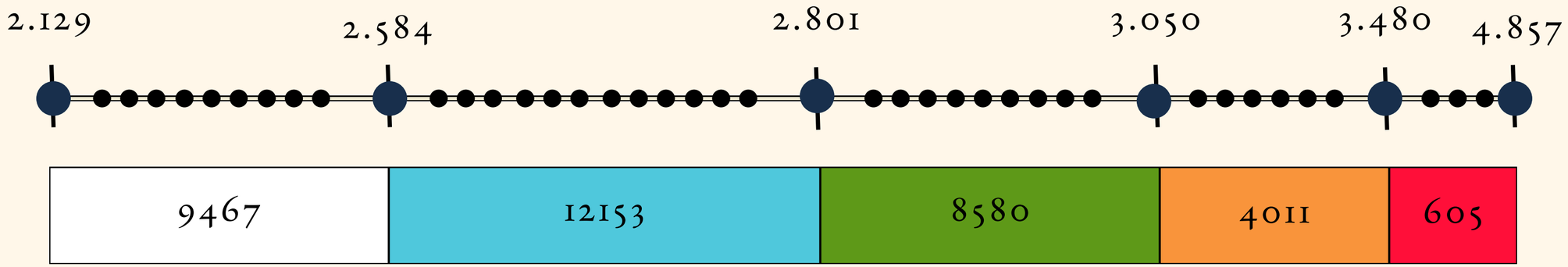
Limitations of this project

PRESENTATION OF THE CI

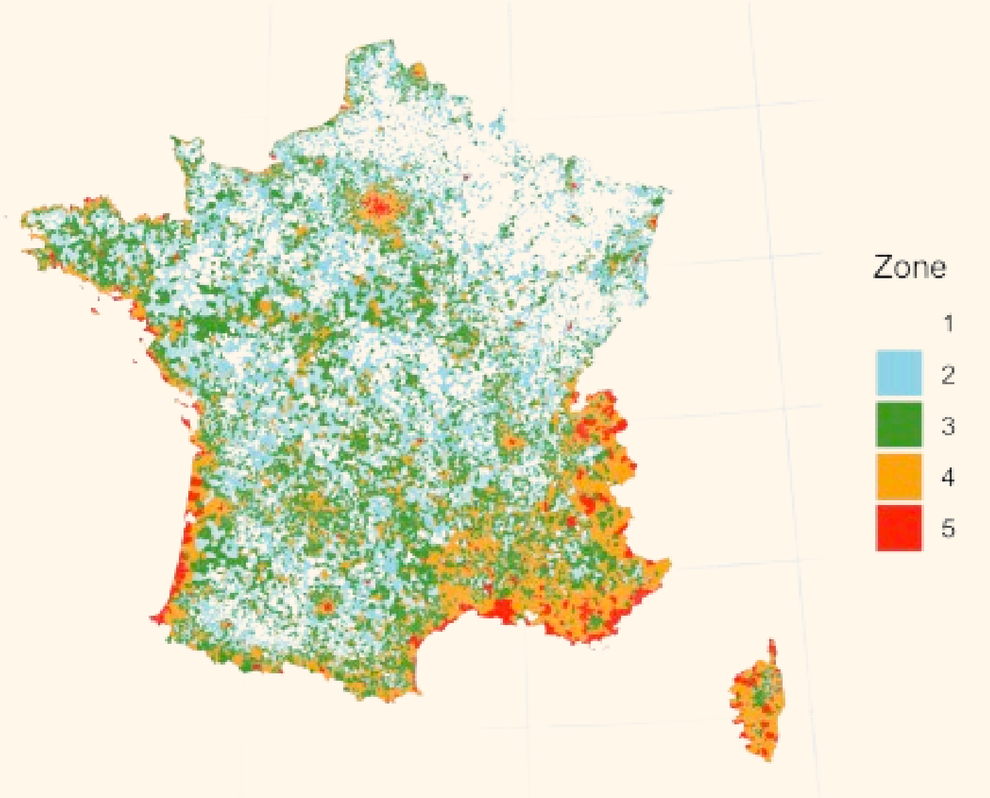
Proposed Zoning Map, Analysis at Municipal and Regional Levels

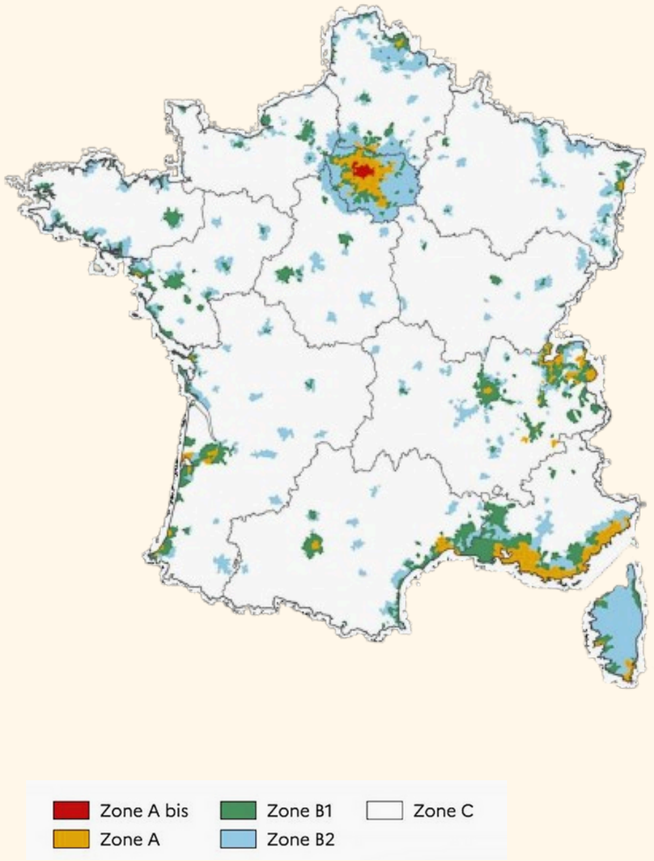
Jenks's Algorithm

The Jenks Natural Breaks Classification is a data clustering method designed to partition continuous data into groups (or classes) such that the variance within each group is minimized, while the variance between groups is maximized.
II used this algorithm to identify the break points needed to partition the data into 5 clusters. I then used this partition to create the final output.



Presentation of the CI





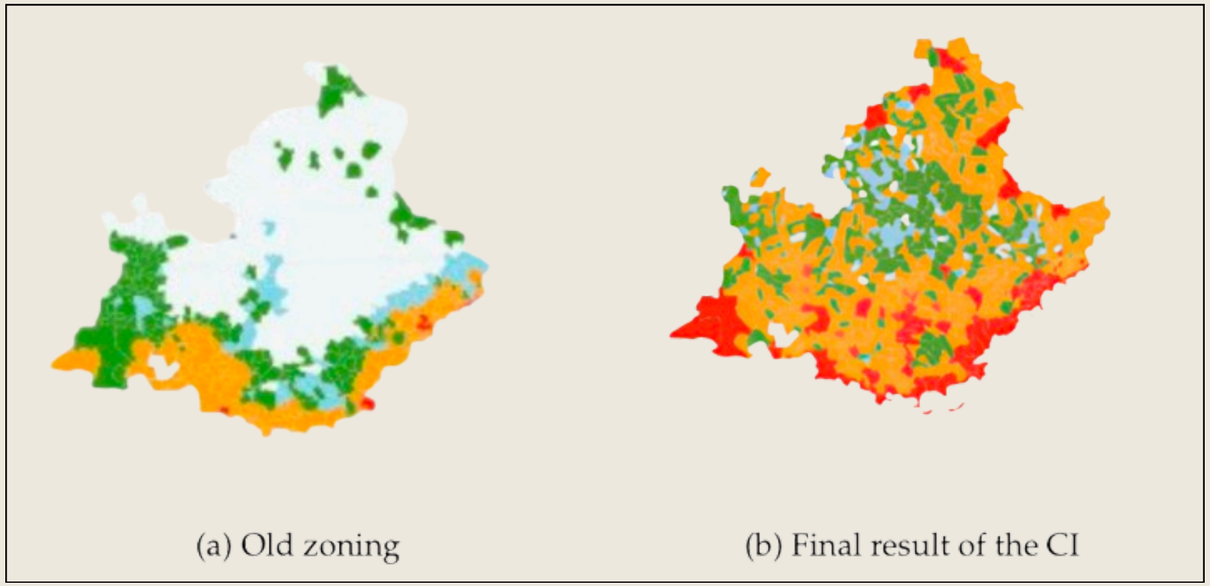
The fact that in our zoning, the levels of tension are much more evenly distributed across the entire territory is related to the fact that the current zoning is based on a morphological distinction between urban and rural areas, which restricts the classification in high tension zones to dense urban areas.

Our typology does not inherently include such a distinction and allow less densely populated municipalities to have hight level of tension.

Region: Provence-Alpes-Côte d'Azur

Close up

In the current zoning, high-pressure areas are confined to a densely populated coastal strip, and to a lesser extent, to the lower Rhône Valley and a few Alpine municipalities.
In our zoning, these areas extend significantly into the hinterland, which experiences intense tourist pressure and high real estate prices

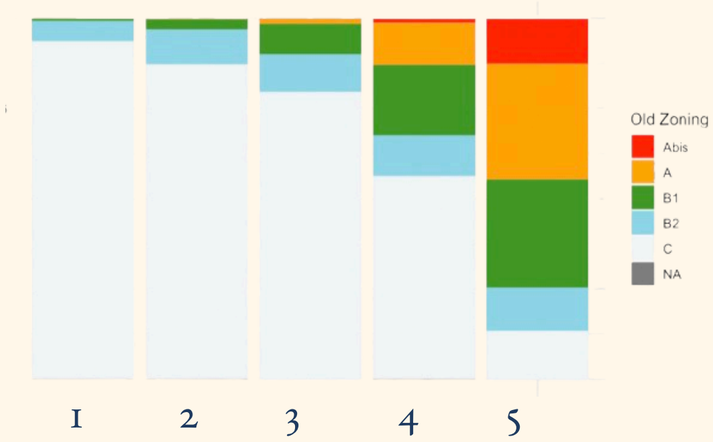


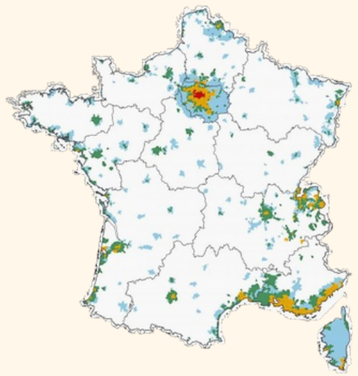
(a) Old zoning (b) Final result of the CI

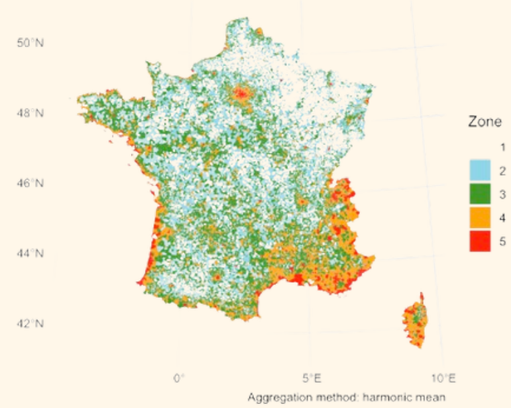
It is important that the zoning knows how to adequately classify all the municipalities taking into account every possible factor of tension.

Redistribution of municipalities

Zone	Abis	A	B1	B2	C
Cluster1	0	0	71	534	8855
Cluster2	1	21	335	1202	10590
Cluster3	3	119	721	903	6834
Cluster4	44	472	788	444	2261
Cluster5	76	194	182	72	81

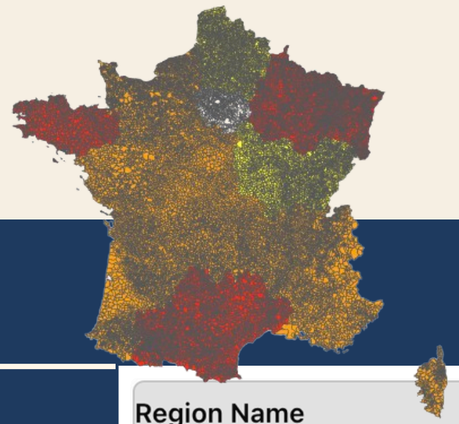






Percentage change

Zone	Abis	A	B1	B2	C
Percentage in old zoning	0,4%	2,3%	6,3%	9%	81,9%
Percentage in new zoning	1,7%	11,5%	24,6%	34,9%	27,2%



Tension by region

Region Name	Old Tension	New Tension	Percentage Change (%)
Île-de-France	3.108129	3.067877	-1.29%
Corse	2.216667	3.866667	74.53%
Provence-Alpes-Côte d'Azur	1.982030	3.589852	81.18%
Auvergne-Rhône-Alpes	1.384157	2.673703	93.19%
Hauts-de-France	1.308952	1.708476	30.52%
Bretagne	1.277778	2.714760	112.48%
Pays de la Loire	1.231707	2.312195	87.74%
Normandie	1.172388	2.116937	80.59%
Centre-Val de Loire	1.175399	2.214123	88.41%
Nouvelle-Aquitaine	1.142625	2.463879	115.55%
Occitanie	1.133169	2.593757	128.88%
Grand Est	1.131692	1.538882	35.96%
Bourgogne-Franche-Comté	1.065458	1.892345	77.57%

- The traditionally less attractive regions in the northeast of France (Hauts-de-France and Grand Est) are experiencing a modest increase.
- Regions that attract significant tourist and residential flows, on the other hand, are experiencing the most noticeable increase (Bretagne, Nouvelle Aquitaine, Occitanie)

Conclusions

SUM UP

Objective and result

APPLICATIONS

Possible applications of this work

LIMITATIONS

Limitations of this work

CONTRIBUTIONS AND CHALLENGES

How I contributed and which were the challenges

HOW THIS MASTER WAS USEFUL

Recap on all the learnings of the M2 that were useful for this Internship

Sum up

Objective of this Internship

Create a new proposal of zoning that redesign the actual ABC zoning by keeping into consideration the real estate tension of each french municipality.

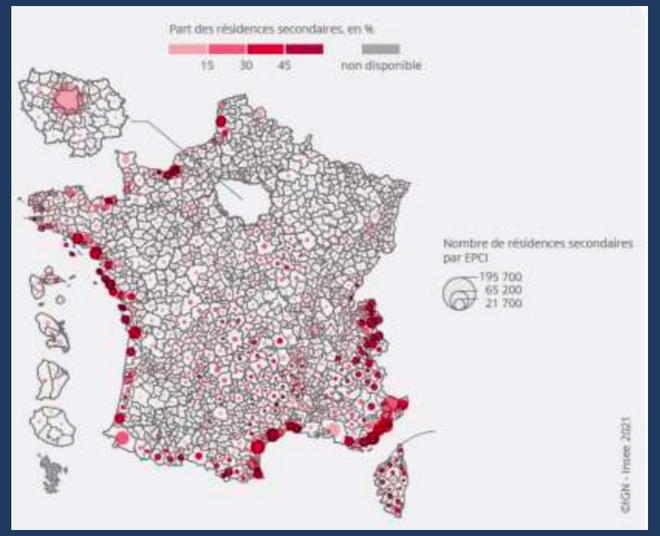
How my results answered to the research objective

The Composite Indicator built in this project is able to capture the actual tension and to provide the Ministry with a data driven tool to make more informed decisions and keep the real estate tension situation under control

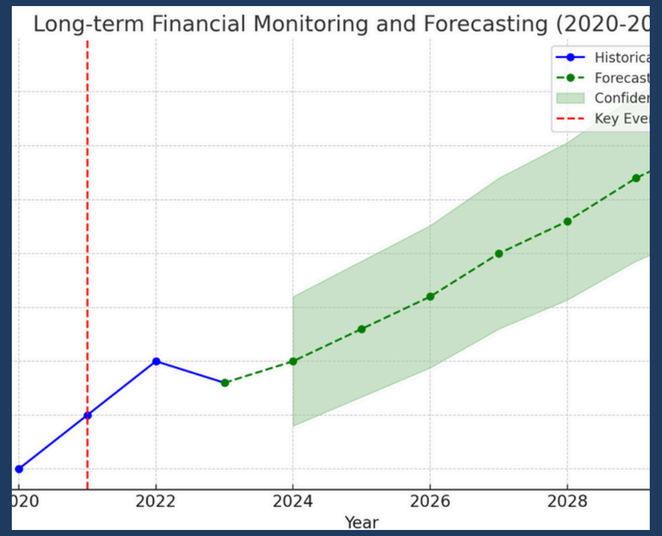
APPLICATIONS

Zero artificialisation nette (ZAN)

Secondary houses



Long-term Monitoring and Forecasting



Limitations

- A certain amount of subjectivity. In each stage we had to make choices that, although based on theoretical models, scientific readings, and many reflections, necessarily brought a high degree of subjectivity and left a certain margin of error.

- The data. The lack of data has certainly constituted a limitation and is therefore another factor to be included among the limitations of this work, although the great attention given to this point has certainly helped a lot to mitigate this aspect.

- The time. 5 months is not a large amount of time for such a large and important project and certainly in the coming months many advances will be made that however naturally go beyond the content of this thesis.

Contributions

- Critically studied the existing literature and contributed ideas and reflections to make important decisions for the project during the weekly meetings.
- Removed 90% of the missing values in the dataset, aggregating at different levels based on the availability of data in the neighbors.
- Carried out the Multivariate Analysis: Correlation, K-means, Spectral Clustering and Random Forest. Visualized the results graphically through tables and geographical maps.
- Used COINr to create a simulation of the results and carried out investigations and analyses on them.

Challenges

- Political geography of France (division into regions, departments, EPCI etc)
- Functioning of the laws on housing, Social housing, distribution of the population and the problem of the high number of municipalities
- Data cleaning preserving the geographical patterns
- Find most effective methods to use

HOW I USED THE KNOWLEDGE
learned in this M2

1.Econometric Modeling and Analysis

I was able to apply econometric models during this project and also to understand them when presented in the papers and in the literature studied.

2.Statistical methods and theory

The course 'Models for truncated and censored variables' has been very useful in dealing with data that was truncated or censored due to statistical secrecy or standardization reasons.

3.Data Visualization and Interpretation

Visual tools learned during this M2 were used to create clear and informative visualizations that aided in the interpretation of the composite indicator results.

4.Practical application of theoretical knowledge

The numerous group projects done in this M2 taught us how to work in a team, manage a project, deal with instructions and deliver on time for a deadline, presenting a work done to the best of our ability

Thanks for the attention

Aix*Marseille université

Appendices

GANTT CHART

DASHBOARD

GEOPOLITICS AND USEFUL GRAPHS

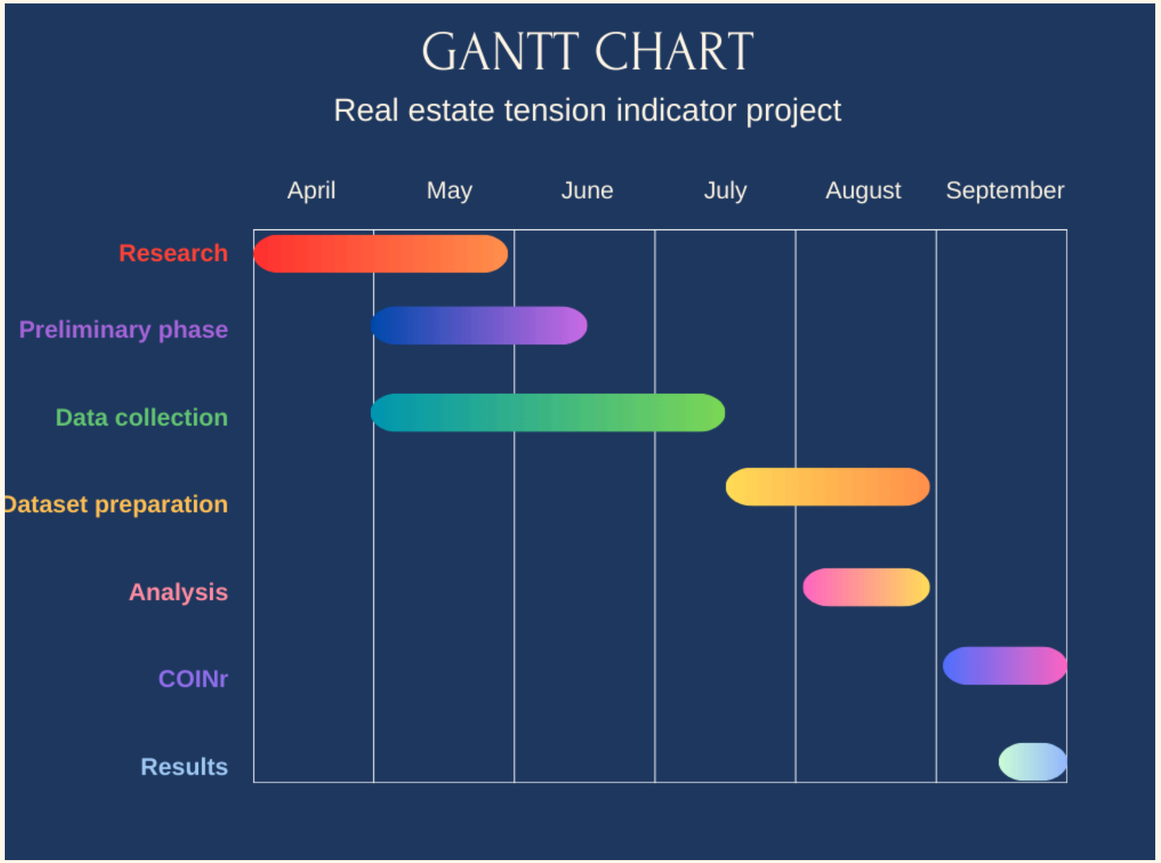
DISTRIBUTION AFTER AGGREGATION

RESULTS OF THE AGGREGATION PROCESS

BIBLIOGRAPHY

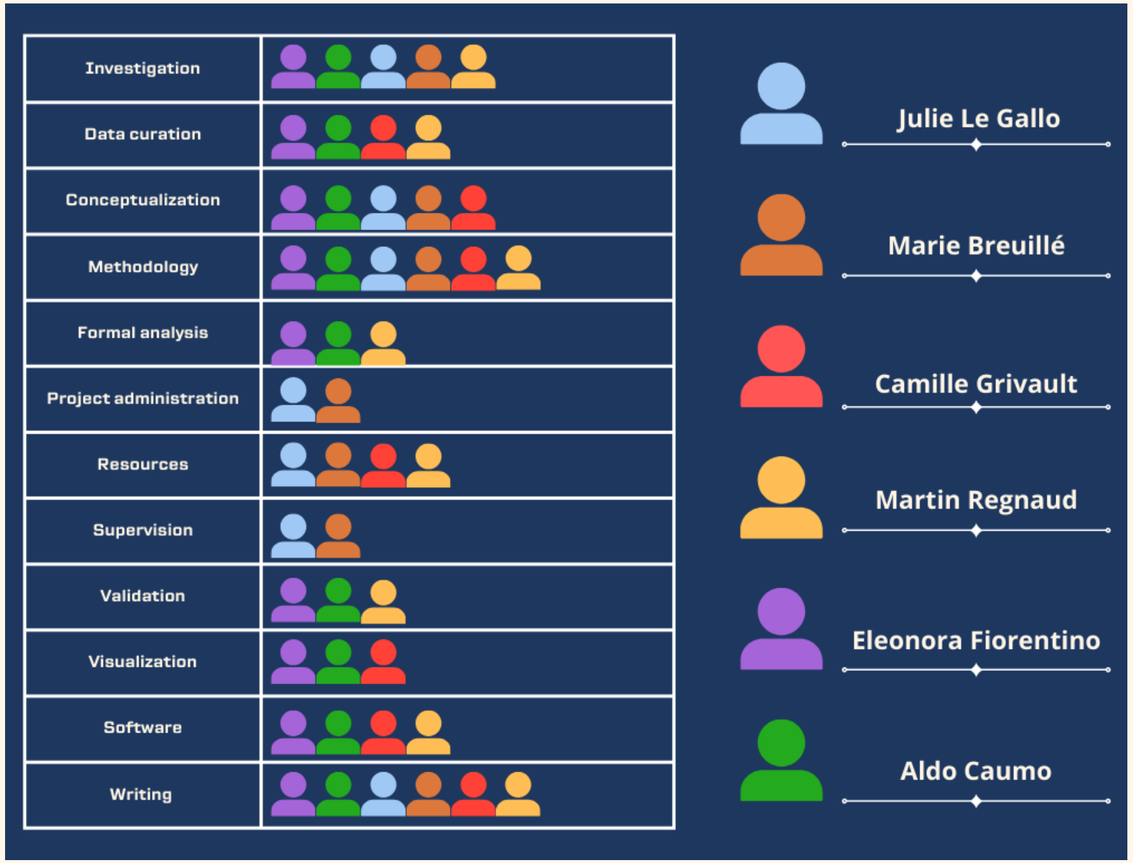
Project timeline

April/September

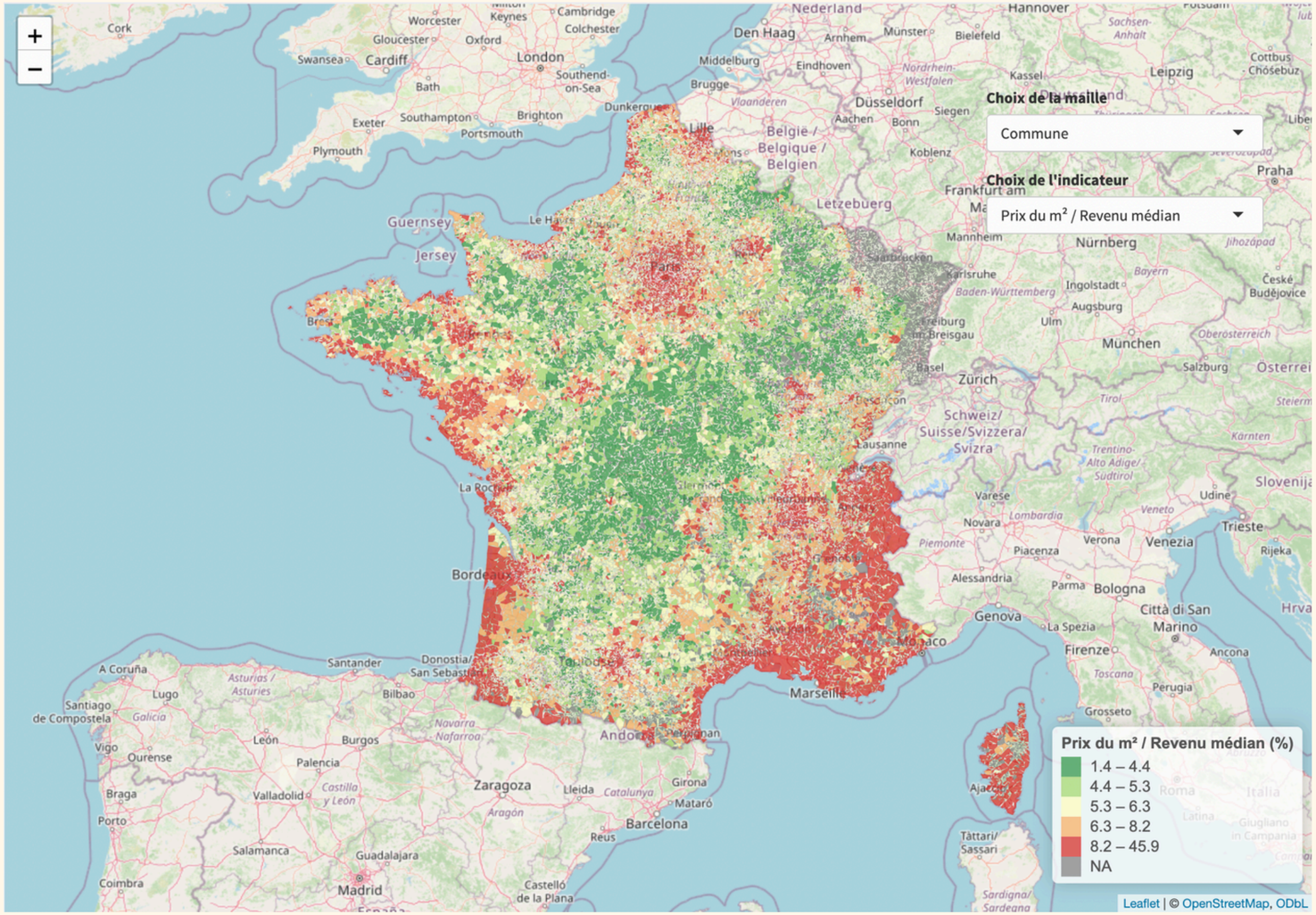


Division of roles

among team members



Dashboard of local authorities - "Attrition of main residences" mission



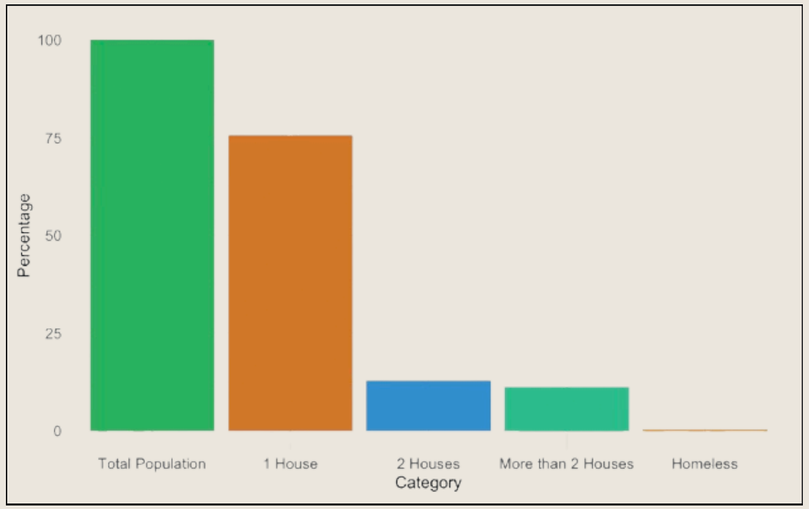
Purpose of the application

This application is a dashboard of real estate tensions, which provides its users with information on the housing stock and its use. It is primarily intended for local authorities, to enable them to monitor the situation of their housing market and facilitate the control of the rules applicable to different rentals.

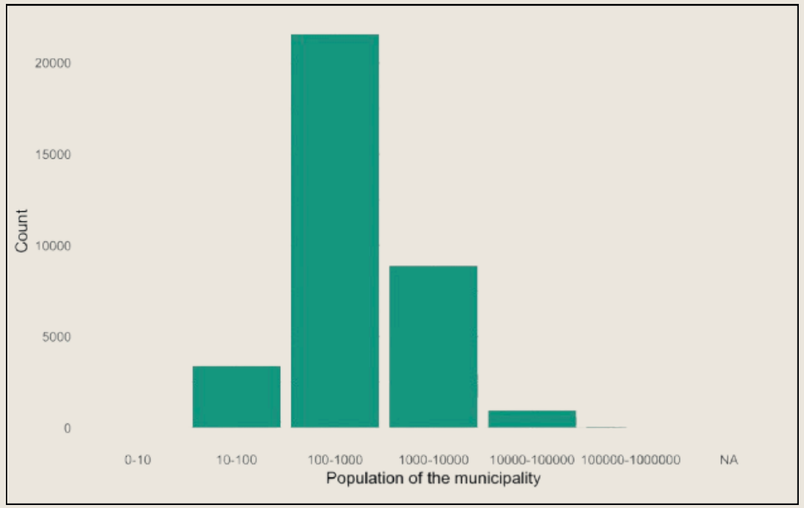
This dashboard was produced as part of the "Attrition of main residences in tourist areas" mission of the General Inspectorate of Finance. The mission initiated a "hackathon" to identify relevant variables to include in the dashboard to best characterize real estate tensions in the municipalities.

This hackathon brought together several housing experts from different administrations (CEREMA, CGDD, DHUP, Insee and DGFiP), whom we would like to thank. This is a knowledge tool and not a regulatory tool, produced for demonstrative and informative purposes.

Housing Statistics in France



Population Distribution

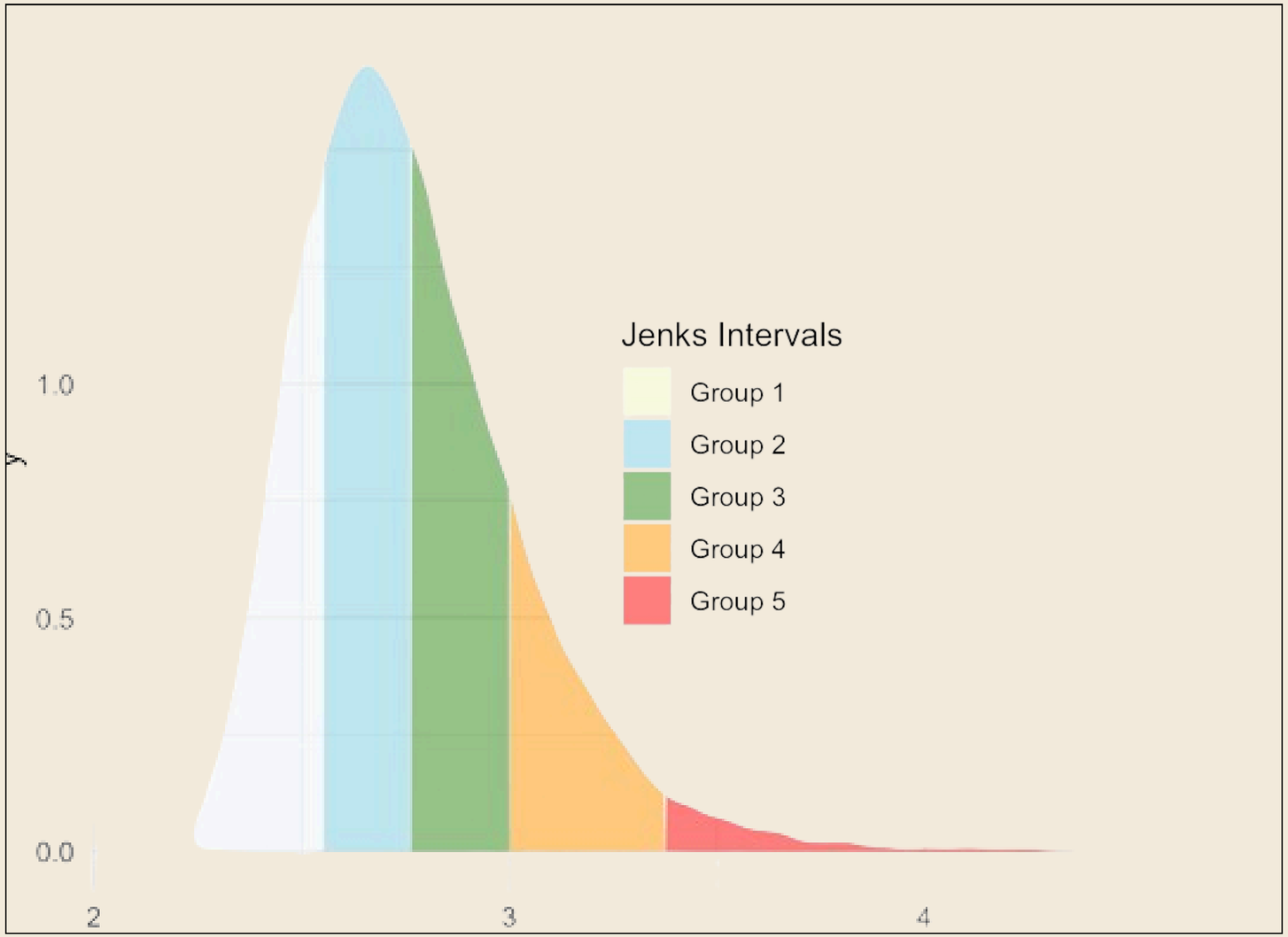


The French regions are divided into 18 administrative entities, of which 13 are located in Europe and 5 are overseas regions.

The number of departments is 101, of which 96 are located in metropolitan France (Europe) and 5 are overseas departments.

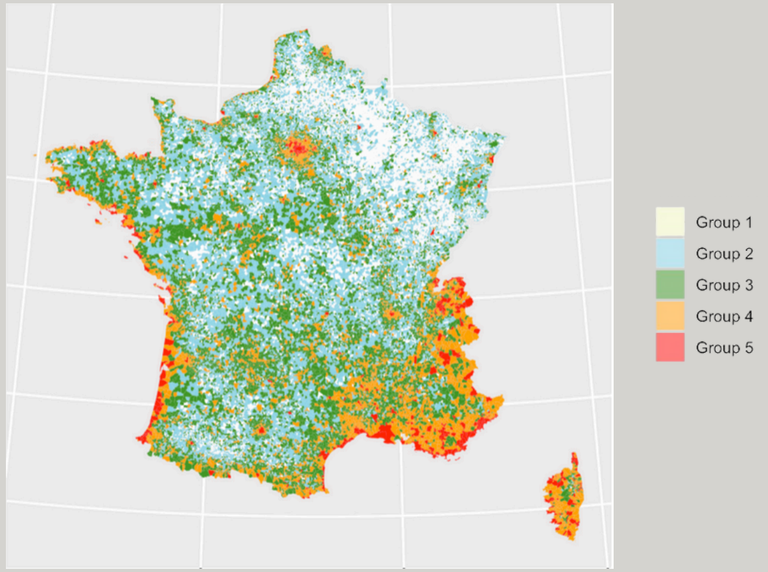
Finally, there are more than 34 thousand communes, which are the basic administrative unit. The distribution of the population among these communes is very uneven, with some large cities such as Paris, Marseille and Lyon having millions of inhabitants, while many rural communes have less than 1000.

Distribution after aggregation and results of the Jenks's Algorithm

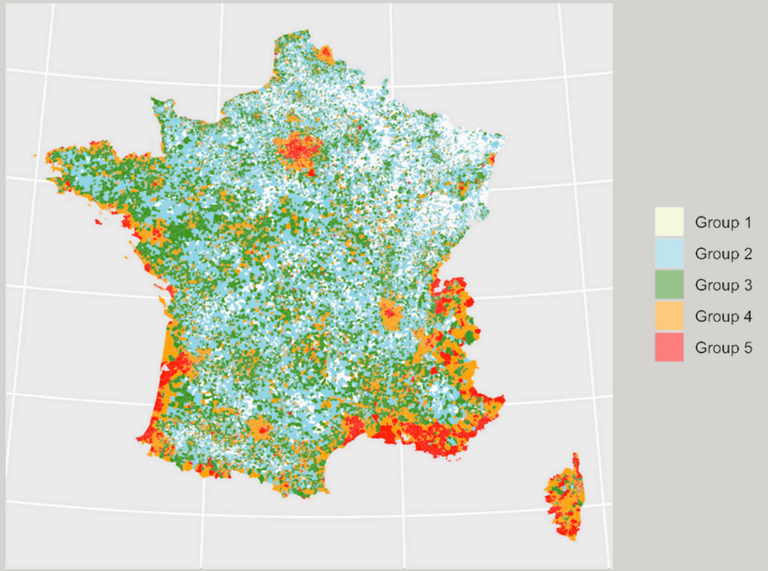


RESULTS OF THE AGGREGATION PROCESS

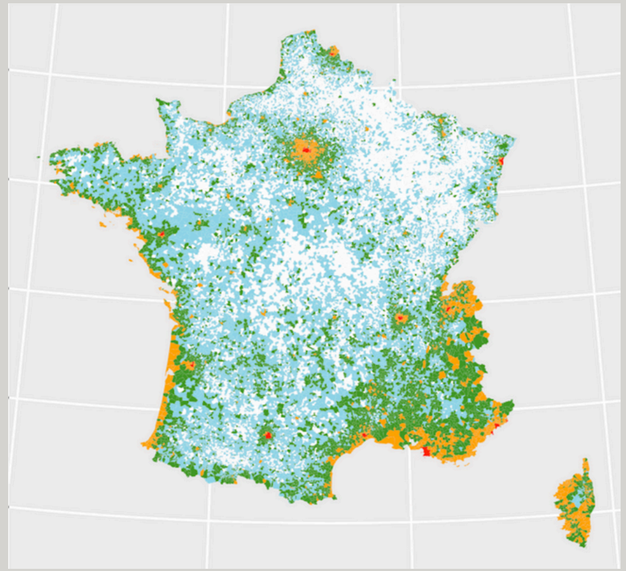
done with different aggregation methods



a) Harmonic mean



b) Aritmetic and Geometric



c) Harmonic and Geometric

Bibliography

[1] Cour des Comptes (2012). Rapport Public Annuel 2012, Tome I. Tech. rep. Cour des Comptes. Cour des Comptes, 2012.

[2] OECD. Handbook on constructing composite indicators: Methodology and user guide. OECD publishing, Paris, 2008.

[3] République Française. Lutte contre l'attrition des résidences principales dans les zones touristiques en Corse et sur le territoire continental. Elsevier, 2022.

[4] W. Becker, M. Saisana, P. Paruolo, and I. Vandecasteele. Weights and importance in composite indicators: Closing the gap. Ecological Indicators. European Commission, Joint Research Centre, 2017.

[5] C. Bizau. Développement d'un indicateur de pression foncière à l'échelle communale. Rapport de stage, September 2023.

[6] M. Cinelli, M. Spada, W. Kim, Y. Zhang, and P. Burgherr. MCDA Index Tool: an inter- active software to develop indices and rankings. The Author(s), 2020.

[7] S. El Gibari, T. Gómez, and F. Ruiz. Building composite indicators using multicriteria methods: a review. Journal of Business Economics. SpringerVerlag GmbH Germany, part of Springer Nature 2018, 2018.