Master 2 Economie
Parcours Econometrie, Big Data - Statistique

DEVELOPMENT OF A COMPOSITE INDICATOR TO MEASURE REAL ESTATE TENSION IN FRANCE AND IMPROVE CURRENT ZONING

Head of track:
Prof. Christian Schluter

Candidate:
Eleonora Fiorentino

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

Academic year 2023/2024
19 September 2024

NON-PLAGIARISM UNDERTAKING

I, (full name of the undersigned) ……Eleonora Fiorentino……

declare being fully aware that plagiarism by copying documents or a portion of a document published in all forms and media, including online publications, constitutes copyright infringement and related rights, as well as outright fraud.

Consequently, I hereby undertake to quote all sources and authors that I've used to write my internship report and its appendices.

Date:
10/09/2024

Signature:
Eleonora Fiorentino

CESAER
Dijon, France

Data Analytics for Business and Society
Dipartimento di Economia
Venice, Italy

Master 2 Economie : parcours Econometrie, big data - statistique
Aix-Marseille School of Economics
Marseille, France

Abstract

Since 2003, for all questions concerning housing policies, the French territory has been divided into zones: specifically the Abis, A, B1, B2 and C zones. Over the years, however, this division has been revised several times and nowadays it has been deemed necessary to develop a new zoning method that is more transparent and suited to the needs of the various municipalities.
This is where this stage comes into play: the objective of this research is to build for the first time a Composite Indicator for real estate tension with the aim of providing a basis to design the future zoning of the entire French territory.
This research was conducted within CESAER, a mixed research unit that brings together teacher-researchers from the Institut Agro Dijon and researchers from INRAE, whose research activities focus mainly on rural and peri-urban spaces and on agricultural transformations.

Summary

In this thesis, the method of developing a Composite Indicator for real estate tension in the case of France is shown, analyzed and applied to the data collected, following a series of very precise steps aimed at implementing an indicator that can be robust, reliable and reproducible.
During this study, several hypotheses were examined and numerous methods were discussed and applied.

In this document, divided into three main chapters, the following topics will be addressed: in the first chapter there will be a general introduction regarding the context in which this mission was conducted and what the assumptions and state of the art were. In particular, I have tried to make as clear as possible both the purpose of this research and the reason why it was so necessary to change the zoning. The workplace, the team and my role in the project are also discussed in this chapter. The chapter then concludes with an explanation of the composite indicators, which is essential to understand the methodological approach followed later.

In the second chapter there will be the actual description of the procedure followed until the indicator was created: it will be divided into the various stages of development and will include both a brief theoretical explications and the actual implementation of the method chosen.

Finally, the conclusion will present the possible applications of the Indicator, together with some final considerations on the knowledge acquired during the master's degree that were useful for the purposes of the project and the various challenges faced that allowed me to expand my knowledge and skills.
I will focus in particular on some important reflections regarding the work and school path, finding connections between the two both at a methodological and conceptual level.

Acknowledgements

First of all, I would like to thank my family, my mother Antonella, my father Stefano and my brother Lorenzo for all the love and support they have always given me, as well as the precious advice and the opportunity to have this exchange.
I also thanks the rest of the family, my grandparents, Luigina, Rosanna and Felice, my aunt Beatrice and Franco.
Thanks also to all the professors I have had in recent years, both in Venice and Marseille, but also those of my three-year course: all of them have given me the strong theoretical and practical basis that allowed me to carry on this project with all the necessary skills.
I would also like to thank the entire CESAER team, in particular Marie and Julie, for choosing me for this project and always following me with extreme kindness, availability and competence.
I thank Camille, for the fantastic work done with the data and his great expertise, and Martin, who provided us with an additional source of data and worked on it intensely.
I also thank my partner in this scholastic and professional adventure, Aldo, who brought a huge contribution to this project, and with whom I shared this beautiful experience, first in Marseille with the Double Joint degree programme, and then also in Dijon, working at CESAER.

CONFIDENTIALITY NOTE:

Ce projet étant jugé à caractère confidentiel par le CESAER, le(s) tuteur(s) enseignant(s) et le président du jury sont tenus de veiller, chacun pour ce qui le concerne, à la non divulgation des informations concernant ce stage

As this project is considered confidential by the CESAER, the teaching tutor(s) and the jury president are required to ensure, each for their part, that information concerning this internship is not disclosed.

Table of content

List of appendices

List of Figures

Context

1.1 The zoning system

The zoning conventionally called ABC, defined in article D304-1 of the Building and Housing Code, implements a "classification of municipalities in the national territory into geographical zones based on the imbalance between supply and demand for housing".
The geographical zones, arranged in descending order of intensity, are: A (which includes zones Abis and A), B (grouping zones B1 and B2), and C.

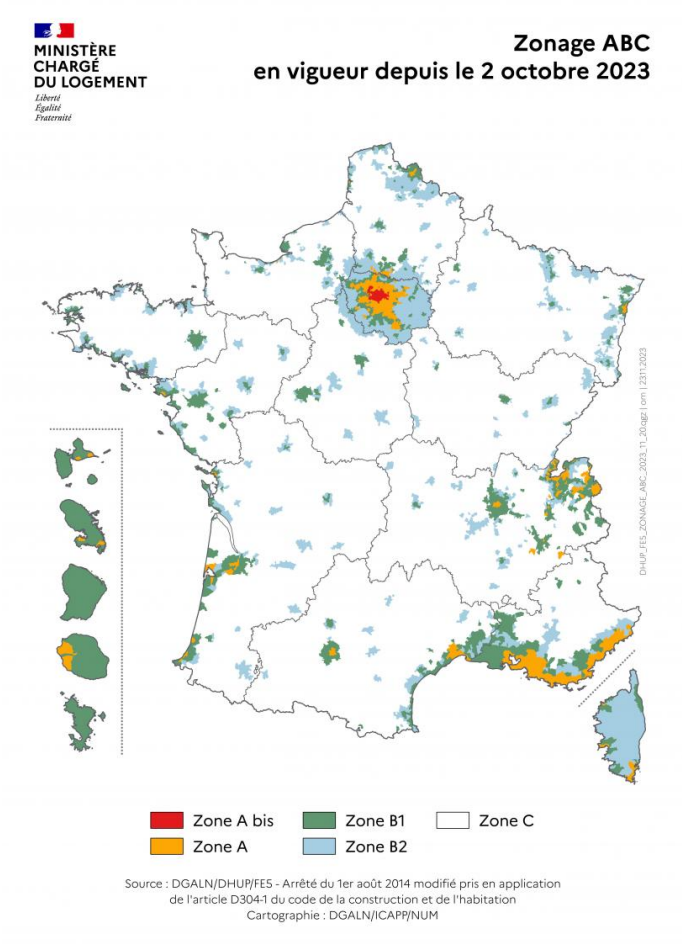
- Zone A bis: includes Paris and 75 municipalities of Yvelines, Hauts de-Seine, Seine-Saint-Denis, Val-de-Marne and Val-d'Oise;
- Zone A: the urban agglomeration of Paris (including the Abis zone), the Côte d'Azur, the French part of the Geneva agglomeration, some urban agglomerations or municipalities (for example Lille, Strasbourg, Lyon, Marseille, Montpellier, Toulouse, Bordeaux) and 8 municipalities in overseas departments where rents and house prices are very high;
- Zone B1: includes some large urban areas and some communes where rents and house prices are high, a part of the large Parisian suburbs not located in the Abis or A zone, the tense provincial cities and the municipalities of the overseas departments not classified in zone A;
- Zone B2: the central cities of some large urban agglomerations, the large Parisian suburbs not located in the Abis, A and B1 zones, some municipalities where rents and house prices are quite high, the municipalities of Corsica not located in the A or B1 zones;
- Zone C: rest of the territory.

This division has not always been like this: the zoning has in fact undergone numerous changes over the years. Initially established in 2003 under the "Robien" rental investment scheme, the ABC zoning was revised in 2006, 2009, and 2014. Subsequently, since 2014, it has been subject to three partial reviews in 2019, 2022, and the last in October 2023.
Despite these updates, the effectiveness of the zoning system remains questionable, especially now that France is experiencing a real estate crisis. This crisis disproportionately impacts certain regions, such as major cities and urban centers, which face substantial real estate pressures due to various factors. These include a dearth of new housing, particularly in high-demand areas, escalating construction costs, a significant proportion of vacant housing stock (8.5%), and many more.
The real problem with this division into zones, however, lies in the way it was created: few, rough indicators but above all a lack of transparency regarding the construction of this zoning [1] [1].

[1] In cited source this problem is explicitly underlined

Even the updates made were not better: instead of being only moved on a scientific basis and on actual data, they were also partly influenced by political lobbying.

Figure 1.1: Zonage in 2023



Zone	Abis	A	B1	B2	C
N. of municipalities	124	816	2216	3156	28623
Percentage of total	0,4%	2,3%	6,3%	9%	81,9%

[2]

It is important to underline how much transparency is lacking today and how much it is necessary for such a purpose. Dividing France into tension zones has in fact a great relevance not only from a political and legal point of view (for the ad hoc laws to be applied in each situation), but also from an urban point of view (construction of new houses), demographic (tension reflects migratory needs), social, economic (imbalance of supply/demand) and many others. Therefore, although perhaps not everyone is aware of

[2] Distribution of French municipalities with data updated to 5 July 2024

the zone of his municipality, this zoning enters the lives of every French person somehow through local housing policies implemented, and determines choices and balances that are sometimes very delicate.

A lack of transparency on the way in which such an instrument is calculated is a very serious problem. It's a problem also the fact that this zoning is calculated mostly with simple indicators that are unsuitable for capturing all the nuances of this concept (the 'tension') which, as seen before, is part of many aspects of society and therefore requires numerous diversified variables to be captured at its best.
As reported in 2012 by the Cour des Comptes itself (and the situation since then, as explained above, has not yet been properly addressed): "The policy of concentrating funding on the most strained areas, effectively implemented since 2010, is thus based on unsuitable instruments, sometimes even counterproductive, and its results are, to date, modest. These are the findings drawn by the Court from a recent investigation [...].".[1]

1.2 Tax and public policy implications of the zoning

Delimitation of the zoning has important implications for public and private actors. Each residential zone has different housing policies, tax incentives and real estate regulations associated with it.
To use the explanation provided by the Ministry of Ecological Transition and territorial cohesion: "ABC zoning is used in particular for the eligibility scope and for the applicable scales (rent and/or resource ceilings) for aid relating to rental investment (Denormandie, Pinel, Loc'avantages), for home ownership (social rental-purchase loan, zero-rate loan, reduced-rate VAT in ANRU zones and priority city districts, real solidarity lease), as well as for intermediate rental housing and for setting rent ceilings for social housing financed in PLS".
To get a clearer idea of what it means for a municipality to be categorized in one zone rather than another, I will briefly report some differences between the various zones:

1. Tax Incentives

    - **Tax devices for real estate investment:**
      Zone A: Real estate investments in new constructions or renovations benefit from the highest tax incentives. For example, the Pinel program offers tax breaks up to 21% of the investment over 12 years for purchasing rental properties in zone Abis and up to 18% in zone A. Additionally, programs like the Denormandie scheme provide substantial tax reductions for renovation projects, further boosting incentives.

      Zone B: Moderate tax incentives are available to stimulate the housing market. The Pinel program provides tax breaks up to 12% over 12 years in zone B1 and up to 10% in zone B2. The focus is on encouraging new investments while balancing the capacity of existing infrastructure.

      Zone C: Tax incentives are minimal or not applicable, reflecting the lower need to stimulate the real estate market. The Pinel program is generally not available

in many areas within Zone C. Any available incentives are targeted at specific projects or small-scale renovations, with potential savings significantly lower than in Zones Abis, A, B1, and B2. The focus is on maintaining economic stability rather than promoting rapid development.

2. State Housing Aid

   - **Prêt à Taux Zéro (PTZ):**
     Zone A: The PTZ is highly accessible with the most favorable conditions. First-time homebuyers in zone Abis can borrow up to €40,000 interest-free, with favorable repayment terms and eligibility extending to higher income thresholds, while in zone A they can borrow up to €30,000 interest-free, with eligibility extending to households earning up to €45,000.

     Zone B: The PTZ is available, though with less favorable conditions compared to zone A. In zone B1 the maximum loan amount is up to €25,000, with eligibility for households earning up to €40,000. In zone B2 the maximum loan amount is up to €20,000, and eligibility typically extends to households earning up to €35,000. Terms are adjusted to reflect moderate housing demand.

     Zone C: The PTZ is available to a much lesser extent or may not be available. In rural areas, the loan amount might be capped at €15,000 with more restrictive eligibility criteria, generally aimed at households earning up to €30,000.

   - **Renovation aid:**
     Zone A: Significant incentives are provided for building renovations aimed at increasing energy efficiency and improving housing quality. For example, in zone Abis tax credits and subsidies can cover up to 40% of renovation costs, with grants for energy-efficient upgrades up to €10,000. In zone A, tax credits can cover up to 30% of renovation costs, and grants for energy efficiency can reach up to €7,500.

     Zone B: Renovation aid is available but at reduced levels. In zone B1 incentives can cover up to 20% of renovation costs, with specific grants for energy efficiency up to €5,000. In zone B2, incentives can cover up to 15% of renovation costs, with smaller grants for energy improvements, typically up to €3,000.

     Zone C: Reduced renovation incentives are available, generally focused on targeted interventions rather than large-scale programs. Aid can cover up to 10% of renovation costs, with minimal grants for energy efficiency, often not exceeding €2,000

3. Urban Planning

   - **Urban planning and development:**
     Zone A: Urban planning policies are focused on maximizing housing density, improving urban quality of life, and ensuring comprehensive infrastructure.

For instance, in zone Abis, Paris has implemented policies that mandate minimum housing densities of 1,200 residents per square kilometer and prioritize green spaces with at least 20% of new developments dedicated to public parks. In zone A policies include requirements for a minimum of 800 residents per square kilometer and at least 15% of new developments allocated to communal spaces. Urban development plans often include extensive public transportation networks and mixed-use developments to support vibrant city life.

Zone B: Urban planning is focused on balanced development and infrastructure. In zone B1 the planning policies encourage medium-density housing, typically around 500 residents per square kilometer, and support infrastructure improvements such as local public transport and community facilities. In zone B2, policies include density targets of around 300 residents per square kilometer and investment in essential services.

Zone C: Urban planning prioritizes enhancing rural areas and small towns with less emphasis on high-density development. Policies often focus on improving local amenities and infrastructure in a way that supports sustainable growth. For instance, planning may involve projects to improve rural road networks and local facilities, with density targets generally below 200 residents per square kilometer.

4. Support for Social Housing Construction

   - **Social Housing**
     Zone A: Significant support and funding are provided for social housing to address high demand. In zone Abis the government allocates approximately €200 million annually for new social housing projects, with a focus on increasing affordable units by 5,000 each year. In zone A funding might be around €150 million annually, with efforts aimed at increasing affordable housing by around 3,500 units per year. Policies include incentives for developers and substantial subsidies for construction.

     Zone B: Support for social housing is available but less prioritized compared to zone A. In zone B1, annual funding is approximately €100 million, with a target of adding around 2,000 new social housing units per year, while in zone B2 annual funding is around €50 million annually, aiming to add around 1,000 new social housing units per year. Support includes moderate subsidies and incentives for construction.

     Zone C: Support for social housing is minimal, with specific attention to local needs. In rural areas, funding might be limited to around €20 million annually, with a focus on small-scale projects or refurbishments. Efforts are often concentrated on addressing critical shortages rather than extensive new construction.

These differences were designed to try to solve the problem of housing tension existing in some French municipalities and help to meet the different needs and dynamics of the

housing market in different regions, balancing the housing supply and improving access to housing throughout all France.

1.3 Motivations of the research

Due to the lack of transparency and all the other problems already listed, the 'Cour des comptes' strongly criticized this zoning and in fact recommended that the A/B/C zoning be "updated in order to better reflect current real estate tensions" (Court of Auditors, 2023). The National Council for Housing Reconstruction called for a flash revision of ABC zoning last June "to bring more municipalities into tense areas during a more structural reform on zoning" with the aim of relaunching production of housing, and in particular to "release constraints to produce more intermediate housing where it is relevant". But despite this being done, the main problems remained.
As Madec (2018) shows, in fact, with the current system, neither the share of tenants assisted nor the amount of aid received increases with the level of tension, which leads to the conclusion that A/B/C zoning does not sufficiently correct territorial heterogeneities due to insufficiently discriminating segmentation of the French territory into tension zones.
To deal with these issues, the Ministry contacted CESAER, and in particular Julie Le Gallo and Marie Breuillé, for the construction of the new zoning.
The main objective of this work was in fact to build an indicator of real estate pressure which can serve as an objective basis for updating the zoning of the French territory.
This team, in particular Marie, Julie and Camille, was chosen because they already have experience with housing public policies (evaluation of the impact of rent control, rental indicators and many more) and have already previously worked with the Ministry of Housing.

1.4 The workplace and the team

This research thesis was conducted, as already mentioned, within CESAER, a mixed research unit that brings together teacher-researchers from the Institut Agro Dijon and researchers from INRAE.
INRAE is a national research institute (originally INRA, created in 1946, which merged in 2020 with IRSTEA) organized in several centers[3].
The INRAE center (Institut national de recherche pour l'agriculture, l'alimentation et l'environnement) located in the Bourgogne-Franche-Comté region is a multidisciplinary center that focuses its research on 3 specific scientific areas:

- Agroecology: biodiversity, biotic interactions and cropping systems
- Food, taste, sensoriality
- Economics and sociology of the development of rural and peri-urban territories

This brief digression on INRAE was necessary to make clear the multidisciplinarity of CESAER, that was fundamental to have a team with different and complementary points of view and skills. In fact, various figures with very specific and distributed roles participated in the construction of this indicator, but at the same time with great cooperation and availability.

[3] https://www.inrae.fr/nous-connaitre/organigramme

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 project was conducted during 5 months, from April to August 2024 in Dijon.

The team members worked mostly individually but frequently exchanged opinions and ideas via Discord, papers and books through Zotero and organized weekly calls on Google Meet.

> **Note.** Zotero is a free and open source software for managing bibliographic references and related materials (e.g. PDF files). During this stage, Zotero was used as a 'library' of papers to read, which were divided into 7 main categories: 'Determinants of housing prices', 'Impact of zoning', 'Institutional settings', 'Land pressure indicators', 'Methodology', 'Rent pressure zones abroad', 'Reports'.

Within the team, I performed many different roles, and this gave me the opportunity to test my skills and the knowledge I learned during my master's degree in different situations. The first few months served to build a solid theoretical base, through reading numerous academic papers, books and articles regarding zoning in different parts of the world, French legislation, methodology, implementation techniques and environment. The subsequent phases were then those of discussion and formulation of strategies to best address the research and, finally, implementation. In this sense, I worked a lot with R both for data analysis and research, and also for the construction of the indicator itself.

The best thing about this internship was the fact that we had great freedom on the tools to use: in fact, we mainly used the R software, but we were free to also use other programs. I was thus able to use Python, Matlab, Qlik. Another thing that I really appreciated was the great trust that was given to us. At all times we were encouraged to give an opinion and to actively participate in decisions, both from a theoretical and methodological point of view.

On the following page I have included two graphs, the first (Fig. 1.2) represents the evolution of the work during these months of internship, the second (Fig. 1.3) better explains the division of tasks within the team.

Figure 1.2: Gantt Chart of the project

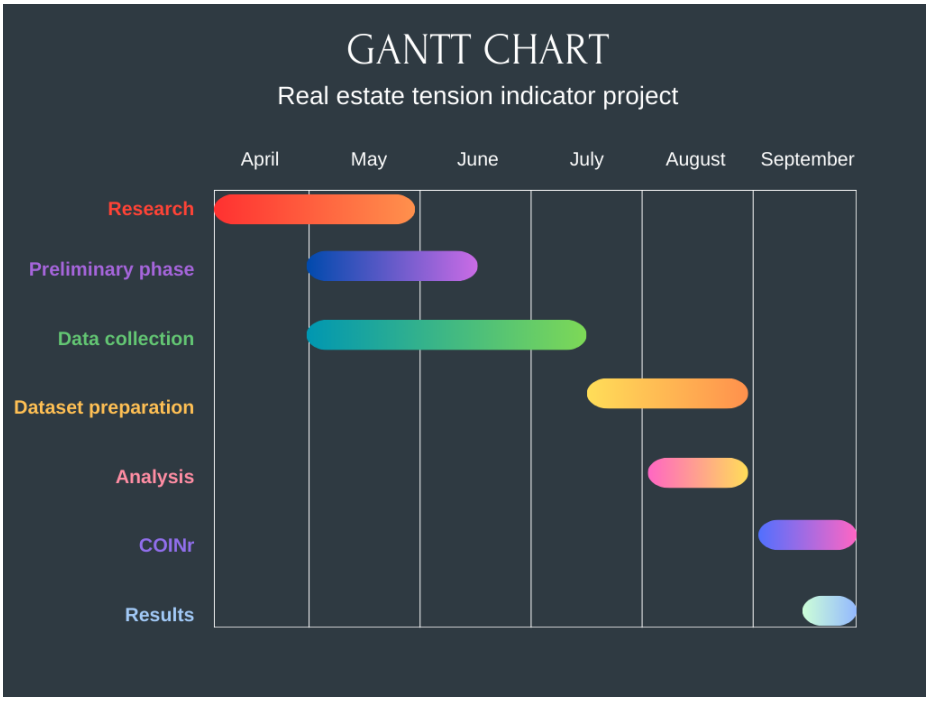
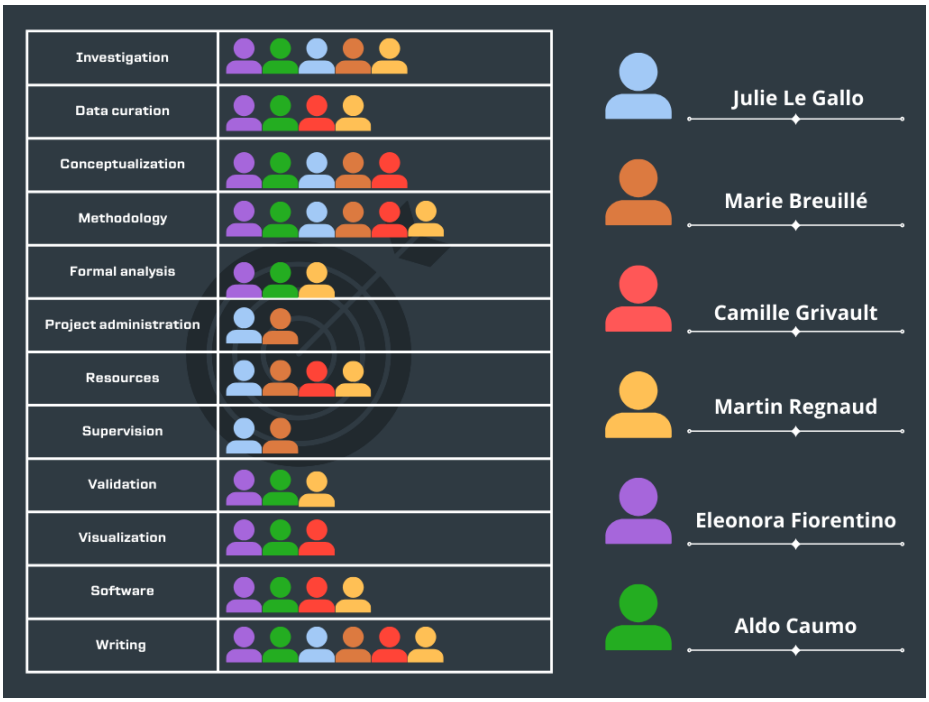


Figure 1.3: Division of tasks among project members



1.5 What is a Composite Indicator

> *"A CI is the result of a mathematical combination of individual indicators that together act as a proxy of the phenomena being measured".*
>
> **— Mazziotta and Pareto 2013**

Composite indicators (CI) are tools that can have multiple purposes.
In our case it will be 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).
Thanks to the CIs, even phenomena that would otherwise be very difficult to evaluate (such as the well-being of a country) can be reduced to a simple score and then be compared with the score of all other countries, in order to obtain a proper classification.
This tool, especially in the last two decades, has been used in many cases: an unfortunately recent example is Swiss Re's Covid-19 vulnerability index (Swiss Re 2020), but many more will be provided in the section 'Applications'.
Some might wonder why it is preferable in this case to use a Composite Indicator instead of a more common PCA or clustering. A Composite Indicator (CI) provides a single, interpretable index that aggregates multiple dimensions, making it easier to evaluate performance or classify entities (in our case, municipalities). 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, theory-driven weighting and non-linear aggregation. This flexibility makes CIs usable in a variety of contexts, while PCA and clustering are more exploratory, aimed at discovering patterns rather than providing an overall score.
Index development is not a trivial task as it involves numerous key steps that can have crucial implications on the results. The beauty of this tool actually lies precisely in its dual nature: on one hand, the output of the composite indicator has great interpretability, and can therefore be easily understood by everyone, even non-experts; on the other hand, it is a sophisticated synthesis of many different factors, which bring deep and complex nuances of the phenomenon we want to analyse.

1.5.1 The history of CIs

The evolution of composite indicators has been marked by significant advancements over time.
Initially, in the early 20th century, composite indicators were rudimentary, combining simple averages of basic measures to assess economic and social phenomena.
By the mid-20th century, indices like the Human Development Index (HDI, 1990) began to emerge, incorporating multiple dimensions such as health, education, and income to provide a more holistic view of development.
The late 20th and early 21st centuries saw an explosion of specialized indices, driven by increasing concern for the environment and society, resulting in tools like the Ecological Footprint and the Social Progress Index.
In the end of the OECD[2] Handbook, there is an interesting table showing the number of hits obtained by searching for "Composite Indicators" through Google and Google Scholar. The table shows these numbers for three different times: October 2005, June 2006

and September 2007. I will show this table here adding the same data for the present, specifically August 2024.

Year	Google	Google Scholar
October 2005	35.500	992
June 2006	80.800	1.440
September 2007	2.000.000	167.000
August 2024	62.000.000	4.080.000

As we can clearly see from the table, the interest of scholars from all fields in composite indicators has increased significantly over the last 20 years. The vast literature has therefore allowed us to get a general idea of the various methods of constructing these indicators and above all led us to reflect on the many problems and limitations that necessarily arise when several complex variables are combined to obtain a simple score to interpret and use.
Despite the vast literature, however, we have found very few, if any, composite indicators that measure real estate tension. It is therefore clear that this is a sector in which it is not yet established practice to use this kind of tools, and it is also evident that the scarcity of papers on the topic has been both a limitation and a challenge.

1.5.2 Construction

The construction of an indicator has been well described and analysed in numerous scientific papers, but it has been formally defined in the OECD Handbook (see [2]): there the various steps necessary to create a CI are explained and for each of them different techniques and approaches have been analysed.
The steps to follow are:

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

Of course, not all steps are always necessary: in some cases, for example, it is not necessary to normalize the data if you used weighting and aggregation methods that do not require having normalized data.
All these steps will be explored in the second chapter, where they will be explained in more detail and where I will expose the choices and steps taken during this research to obtain an indicator that was as close as possible to the reflection of the phenomenon to be analyzed (real estate tension).

1.5.3 Applications

As I mentioned before, CIs are widely used tools, and can turn out to be useful in many different cases. Here are presented CI for different fields (an example for each):

- **Economics**

  Human Development Index (HDI): Combines indicators of life expectancy, education, and per capita income to rank countries into different tiers of human development.

- **Environmental Science**

  Environmental Performance Index (EPI): Uses indicators such as air quality, water and sanitation, biodiversity and habitat, and climate and energy to rank countries on their environmental health and ecosystem vitality.

- **Social Sciences**

  Social Progress Index (SPI): Uses indicators like basic human needs, foundations of wellbeing, and opportunity to measure the social and environmental performance of different countries.

- **Health**

  Health Care Access and Quality Index (HAQ Index): Combines indicators related to the quality and accessibility of healthcare services to assess the performance of healthcare systems in different countries.

- **Education**

  Education Index: Part of the HDI, it combines indicators of mean years of schooling for adults and expected years of schooling for children.

- **Business and Finance**

  Dow Jones Industrial Average (DJIA): Combines the stock prices of 30 large, publicly-owned companies in the United States to provide a general overview of the market's performance.

- **Technology**

  Digital Economy and Society Index (DESI): Uses indicators such as connectivity, human capital, use of internet services, integration of digital technology, and digital public services to measure digital performance across Europe.

Finally, there are also examples of composite indicators that span multiple fields. The Sustainable Society Index (SSI), for instance, primarily fits within Environmental Science due to its strong focus on sustainability and environmental indicators. However, it also incorporates aspects of human wellbeing and economic factors, making it a multidisciplinary composite indicator that bridges several domains.

Mission

We now come to the central part of the work, namely the construction of the indicator itself.
The construction of a composite indicator is not trivial and requires several steps to reach a satisfactory and robust solution.
The next paragraphs will be dedicated to the various steps faced during our journey and will include a brief theoretical introduction on the topic before moving on to the actual implementation.

> **Note.** For the construction of the indicator we mainly followed the guidelines proposed by the well-known in this sector "Handbook on Constructing Composite Indicators" (OECD [2], 2006)

2.1 Conceptual framework

What is badly defined is likely to be badly measured

An indicator is a complex evaluation tool of a phenomenon which is attributed a close link with a series of other variables which are part of a model decided a priori or an interpretative scheme (a posteriori).
The phenomena are observable and measurable only after having defined a conceptual reference model which has the aim of making clear exactly what is intended to be measured and how it is intended to be done.
The definition of the conceptual framework represents a complex moment as it requires the identification and definition of constructs that are generally abstract and theoretical but which in this context must have concrete references.

One of the most important aspects of a CI is the fact that it is a single indicator that collects multiple different variables. In order to avoid confusion and make the CI an even more useful tool, it is good practice to divide the variables that compose it into macro categories, also called 'pillars', each of which represents a theme or aspect that contributes to the final indicator. Each of these categories then contains some variables, and the number of categories and variables per category can be chosen from time to time based on the available data and needs.
This part was a very complex part of the construction process. Initially we were unsure about the number of 'layers' to create. The idea was to initially enclose the variables in some categories, and then merge these into other macro-categories (e.g. supply/demand), but we soon realized that this classification would introduce several problems, such as the need to include the same variable in more than one pillar (which would have greatly distorted the correlation between two categories), or the impossibility of giving a clear definition in terms of supply/demand to some other categories (for example those related to natural constraints or laws).

Given all the problems that would arise from such an operation, we decided to proceed with a single intermediate layer between the variables and the CI, and we created these 9 categories: *Housing affordability, Supply/Demand mismatch, Distance jobs/equipment, Demand, Market dynamics, Economic attractiveness, Supply, Regulation, Tourist attraction*.[1]
The categories and the variables assigned to each of them can be seen in Fig. 2.1, in the Sunburst chart.

Figure 2.1: Sunburst chart of variables by dimensions

2

Note. Ideally, the question and dimensions are defined before selecting variables. However, discussions with stakeholders may begin with the variables already defined. In this case, it is still valuable to work backwards to define the dimensions and questions, as this process will help with analysis decisions, such as weighting of variables, which will be covered in subsequent steps.

[1]There is also the "Identifier" category, a short explanation of it can be found in the Appendix C, Section II

[2]For further information on each variable see the Appendix C

Why these categories in particular?

Why were these chosen as macro-categories? Some of these categories, such as 'Supply' and 'Demand' were chosen quite intuitively, as they are certainly very determining factors for the final purpose (measuring real estate tension), and many of the variables fall into one of these two categories.

Another very important category is the one that represents a mismatch between Supply and Demand and which contains all those variables that represent a ratio between D/S and that in our research play a crucial role, as they effectively represent tension.

The category 'Regulation' is the one concerning the legislative aspects that in some way influence the real estate market: it specifically contains two types of variables: those representing rates, and binary variables that indicate for each municipality whether or not a certain law/regulation has been applied.

To measure the adequacy of rental costs, a new category ('Housing affordability') has been added, which measures the rental capacity of a tenant with respect to different types of housing.

Clearly, however, there are many other factors that push people to prefer a certain municipality or residential area rather than another: these are mainly economic factors (categories 'Economic attractiveness' and 'Market dynamics'), motivations linked to tourism, both summer and winter (category 'Tourist attractiveness'), and also a certain propensity for municipalities that are well served in terms of infrastructure and transport ('Distance jobs/equipment').

2.2 Variable selection and the creation of the dataset

The creation of a composite indicator, as already mentioned, is divided into a series of pre-established steps that allow not only greater clarity of exposition but also a conceptual rigor that ensures that each fundamental step has been respected without having committed shortcomings.
When we talk about 'Variable Selection' we are referring to the process that leads to the creation of the initial database.
A more in-depth screening of the variables will then be performed after the imputation of missing values, when a clean data set is available and any redundancy between the variables can be analyzed using a variety of statistical and participatory methods. This step is called "Multivariate Analysis".
In our work, the creation of the initial dataset was accomplished by collecting data from numerous 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).
All information on these data sources can be found in the Appendix A I
The construction of the dataset is what underwent the greatest changes during the first months of work: based on readings of papers and research done, numerous variables that

seemed pertinent to the work to be carried out were added.
This work on the variables was mainly developed in two phases, the first of which was the one that required the most time, and the second that required the most effort.
The first phase was the collection of all the variables that could be useful: we started from a dataset used for another survey, aimed at creating a dashboard of real estate tensions [3], to which other variables specific to our survey were then added, from different sources and sometimes even from different years.
The second phase, however, was more complicated at a conceptual level than the first, as we had to move from a dataset with 200 or more variables to a smaller one. Warning: this is not yet a skimming based on statistical methods of multivariate analysis, but rather a simple initial selection that allows us to create a sensible dataset, with variables that cover the various aspects of the phenomenon but without producing duplicate or clearly useless information. This step is fundamental: although at first it might seem intuitive to think that the more variables you have the better, upon reflection it becomes clear that a large quantity is not necessarily synonymous with greater explainability of the phenomenon to be measured. An excessively large dataset in fact, not only makes the initial analysis more complicated, but can also be misleading and create confusion, making it difficult to divide the variables into categories (or 'pillars') and assign a 'direction' (also called 'sign', see section 2.3) to each variable with respect to the final indicator.[4]
To understand which variables were necessary and which were not, we had to analyze some 'macro-arguments' related to this project and which certainly have an influence on the real estate tension, and which were therefore part of this construction process:

- **Main and secondary houses**

  As regards the problem of second and vacant houses, it is discussed in great detail in the report: "Lutte contre l'attrition des résidences principals dans les zones touristiques en Corse et sur le territoire continental[3]" [5], where the main reasons for the strong housing pressure are described, especially in areas considered 'touristic'.

  To find a criterion for defining an area as 'touristic', an exhaustive statistical analysis was carried out on all metropolitan municipalities to objectively characterize their "tourism". After this analysis, they were able to identify 1,027 municipalities as touristic and belonging to the tense zone.

  The problem of real estate tension is closely linked to the crisis in the real estate sector, both in terms of renting and buying a house. This crisis is due in part to an insufficient supply compared to demand, and in part to an increase in prices for both renting and buying.

  This phenomenon is more marked in some metropolises and in some tourist areas that attract and concentrate the population, which is why it is useful not only to divide cities based on the number of people who reside there permanently, but also based on those who are attracted 'seasonally' by that municipality, for work, study or vacation reasons.

  In fact, some tourist areas concentrate the demand for housing both as primary and secondary residences due to favorable climatic and geographical conditions and

[3] `https://shiny.observatoire-des-territoires.gouv.fr/tdb.tensions.immobilieres/`

[4] The Excel file in which all the variables, source, description, polarity and other essential information for the project are collected can be consulted in Appendix C

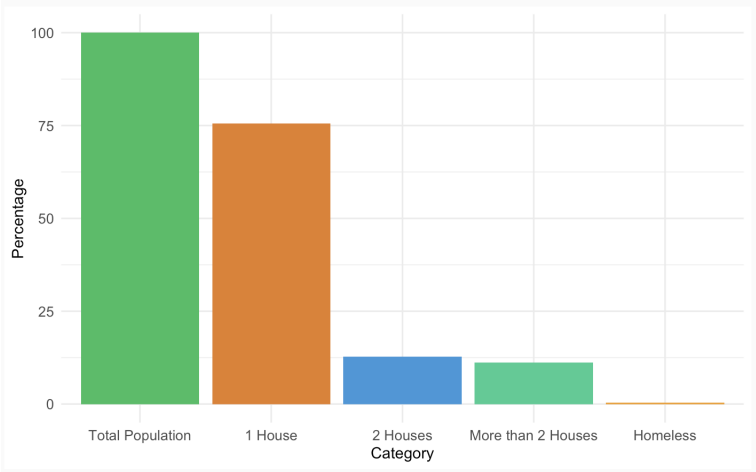
[5] The dashboard of real estate tensions that we talked about earlier was created exactly in the context of this report

economic development that attracts new workers. This strong demand inevitably leads to even higher prices in the real estate market, especially when the supply (and possibly the construction of new houses) cannot satisfy the great demand.

The distribution of houses is certainly not linear: approximately one eighth of families (12.8%) own 2 houses, while 11.2% own more than two, the rest of the families own only one or none at all (0.44% of the population, approximately 300,000 people).

Figure 2.2 shows well this situation from a visual point of view. More information can be found in A II.

Figure 2.2: Housing Statistics in France



> **Note.** In our dataset, there are variables related to second homes in more than one category: in fact, they are present in the pillars of the variables related to Tourism, Regulation and Supply/demand mismatch.

- **Municipalities**

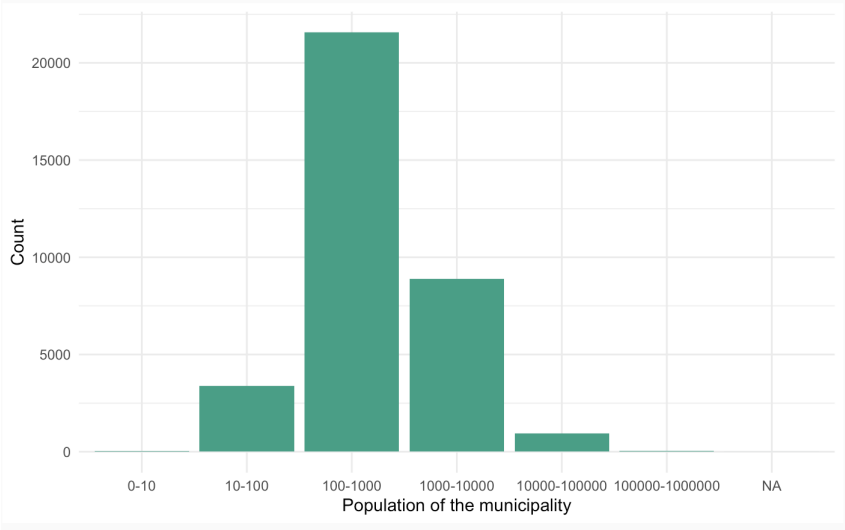
  Another aspect to pay close attention to is certainly the large number of French municipalities and the asymmetrical distribution of the population within them. France, simplifying a bit the complexity of the geo-political divisions of the nation, is in fact divided into regions, which in turn are divided into departments, which are finally divided into municipalities (there are also other divisions like EPCI or maille habitat but we will talk about it later).

  The French regions are divided into 18 administrative entities, of which 13 are located in Europe and 5 are overseas regions. Each region has its own regional council and a president of the regional council, and plays a significant role in the management of local affairs, including economic development, education and transport.

  Regarding the French overseas territories, an important clarification is to be made: in this work, in fact, we have examined exclusively the continental territories of France, thus excluding those overseas. It was in fact the main interest of the Court of Auditors to focus on continental France in this investigation on tension. This of course does not exclude that in the future they may also be added to this CI.

The distribution of the population among these regions is highly variable. Some regions, such as Île-de-France, which includes Paris, are densely populated and urbanized, while others, such as Corsica and many of the overseas regions, are more sparsely populated and often have a lower population density.

Figure 2.3: Population Distribution



The number of departments is 101, of which 96 are located in metropolitan France (Europe) and 5 are overseas departments. Departments are second-level administrative entities that manage functions such as social welfare, infrastructure maintenance and the management of secondary schools.

Finally, there are more than 34 thousand communes, which are the basic administrative unit. Each municipality has a mayor and a city council and manages local services such as urban planning, primary schools and the municipal police. 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.

Note. All this information was then used to create the 'Identifiant' category, which collects all those variables that have a mostly identifying purpose but which in fact represent geographical or administrative denominations.

It is very important to underline that all this geographical complexity has translated into a real challenge, the challenge of selecting the right scale to maintain an adequate trade-off between the availability of data (which are obviously more difficult to obtain and often completely missing the more you try to get into the specifics) and a fine scale. At the level of data availability, there are other non-institutional scales, such as the 'maille habitat', used in our research for some specific purposes, such as the treatment of missing values, the 'bassin d'emploi', and some others that we have not covered.

2.3 Impact of variables and COINr package

During the study of the scientific literature preliminary to the beginning of the work of building the CI we had the opportunity to read numerous methodological papers written by the researcher William Becker. In addition to having published numerous studies on CIs, he also created a very useful tool, namely a package, called 'COINR'[6], for the Rstudio software, through which to build composite indicators.

Having studied its potential, we therefore decided to rely on it to build the Tension Indicator.

In order to use this package, it is necessary to provide some input data: obviously a dataset is needed (even with NA, they can in fact be processed directly within the program), it is necessary to already have in mind the division into dimensions or pillars, and finally it is necessary to provide the SIGN of the variables on the dimension and of the dimensions on the CI. With 'sign', also called 'impact of variable' I refer to the impact that a variable has on the dimension to which it belongs. This impact can be of 2 types: positive (represented by a plus) or negative (represented by a minus).

This sign can be calculated empirically (with correlations or other techniques) or theoretically. In our study, we have decided to proceed with this second path. The reason for this decision is that in our case it was more effective to define the effect (positive or negative) of a variable on a category through a shared reflection of the team members, rather than looking for a variable that somehow represented the category and calculating a correlation between these elements.

The aspect in the process of defining the sign that needed to be paid more attention to was to maintain coherence between the combined effect of <variable + dimension> on the CI and the effect that the variable would have on the CI without the category.

Explanatory example

I will try to clarify this aspect with an example: **(example 1)** let's take the variable 'population growth', and suppose we put it in the category 'demand (for housing)'. Clearly the variable has a POSITIVE effect on the category (more people, more demand), and the category has a POSITIVE effect on the CI (more demand, more housing tension). If we take the direct effect of the variable on the CI, it is also POSITIVE. This is a sensible and logical scenario.
Now let's take another example: **(example 2)** suppose we have a variable that has a POSITIVE effect on the category it is in, but a NEGATIVE effect on tension, and that it is put in a category that also has a POSITIVE effect on the CI. In this case a problem arises, as there is no match between the expected effect of the variable on the CI and the effect it will actually have.

An important step was therefore to verify that each indicator had coherent signs and that they did not lead to internal contradictions.

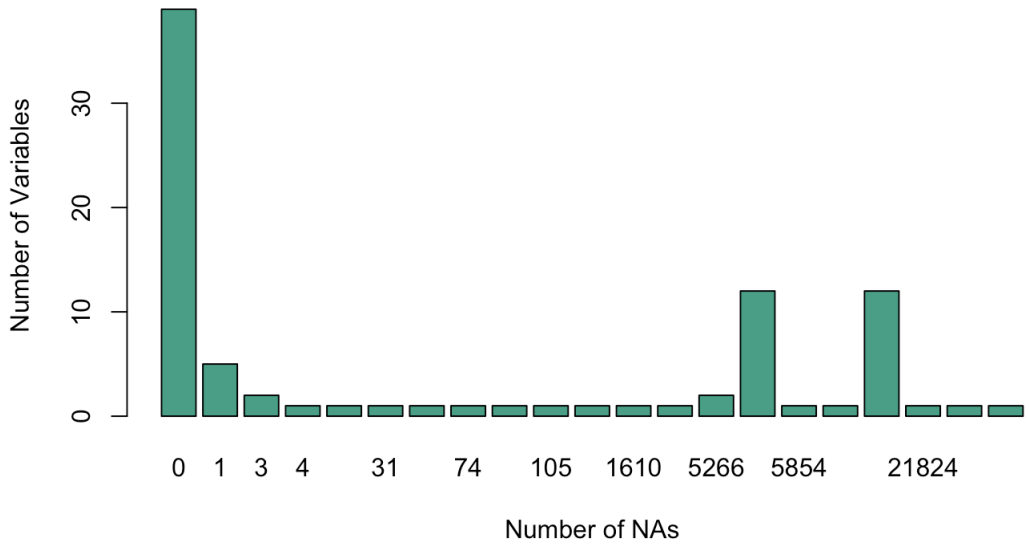
Finally, the positive or negative value of each sign was verified and validated by the correlation analysis performed on the variables within each category (see the 'Multivariate Analysis' section and Appendix B for the correlation tables).

[6] https://www.willbecker.me/projects/coinr/

2.4 Imputation of Missing Data

Imputation of missing data is an important and necessary step to even begin preliminary analyses on the collected data. Dealing with NAs is an art and sometimes requires applying different strategies based on the situation and the variable to be treated. It is important to know the reasons for missing data well in order to be able to replace it in the most correct way from a theoretical point of view.
Unfortunately, when you do analyses at the level of municipalities, which are very numerous here in France and often have a very small population, it is always difficult to obtain homogeneous, complete, uncensored and monitored statistical data over time for some of the variables of interest. In the case of this research, significant efforts were made to acquire data from different sources and that could be as complete as possible, but the resulting dataset, with more than 80 variables for over 34,000 observations, was inevitably full of missing data (specifically around 409000 NAs).
The main problem of this dataset was precisely the lack of data. This lack is found especially for small municipalities (with less than 100 inhabitants), for islands, and for the municipalities of 3 specific departments, Moselle (57), Bas-Rhin (67) and Haut-Rhin (68). Moselle, Bas-Rhin and Haut-Rhin are part of the Alsace-Lorraine region, which has a unique legal status due to its historical shifts between French and German control. This region was part of Germany from 1871 to 1918 (after the Franco-Prussian War) and again during World War II. As a result, some German laws and customs were retained even after it was reintegrated into France. For this reason, not all data for this region is as accessible as for other departments of France.
With regard to the municipalities with a small population, on the other hand, one of the main reasons for the NAs is statistical secrecy.

Figure 2.4: Number of variables with NAs



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

To treat the NAs we mainly used two methods, one developed by Aldo and one by me. Among the variables present in the dataset, 48 had NAs. Of these, 11 had less than 50 NAs, and we therefore decided to fill this missing data in the most precise way possible, that is, by looking at the neighbors (in a geographical sense) and averaging their values. However, this was not possible for all 11, since some of them were islands (and therefore had no neighbors).
The remaining variables, i.e. those with more than 50 NAs and the islands, were treated with an algorithm developed by me. The operation of this algorithm is as follows: for each variable we consider 3 levels of aggregation in increasing order: maille habitat, EPCI and department. For each level we calculate the average of the variable value of the other municipalities belonging to the same level and the percentage of missing data (for example, if an EPCI is made up of 10 municipalities and among these 4 are NA, we will have a percentage of NA of 40% for that level). After these operations we obtain a dataframe that contains: the variable of interest, with a row for each municipality and all the NAs still to be treated, and then 3 columns containing the average value of the municipalities of the same maille habitat, EPCI and department and another 3 containing the percentage of NA for each level (in Figure 2.6 an example of this table).

Figure 2.5: Example of table for NAs

CODGEO <chr>	pricem2_med_2020_2022 <dbl>	DEP <chr>	EPCI <chr>	code_maille_habitat <chr>	mean_epci <dbl>	perc_na_epci <dbl>	mean_maille <dbl>	perc_na_maille <dbl>	mean_dep <dbl>	perc_na_dep <dbl>
02624	5.534206	02	200071983	Reg_I1_18	5.534206	0	27.97330	23.8	45.78213	9
02625	5.534206	02	200071983	Reg_I1_18	5.534206	0	27.97330	23.8	45.78213	9
02626	93.177080	02	240200576	Reg_I1_18	93.207628	0	27.97330	23.8	45.78213	9
02627	93.177080	02	240200576	Reg_I1_42	93.207628	0	75.78019	0.0	45.78213	9
02628	38.756868	02	200071991	Reg_I1_14	43.059763	0	246.18493	0.0	45.78213	9
02629	34.813953	02	240200444	Reg_I1_18	34.813953	0	27.97330	23.8	45.78213	9
02631	25.381855	02	200071785	Reg_32_18	25.998900	0	38.50322	0.0	45.78213	9
02632	37.000000	02	200071769	Reg_32_80	37.000000	0	44.76894	0.0	45.78213	9
02633	39.944367	02	240200501	Reg_I1_5	39.944367	0	55.52945	0.0	45.78213	9
02634	NA	02	240200634	Reg_I1_18	NaN	100	27.97330	23.8	45.78213	9
02635	44.071354	02	240200493	Reg_32_71	44.071354	0	49.39626	0.0	45.78213	9

At this point the algorithm proceeds as follows:

- If the percentage of NA in the maille habitat is < 50%, use the maille mean.
- Otherwise, if the percentage of NA in the EPCI < 50%, use the EPCI mean.
- If neither of the above conditions are true, use the department mean. However, if the department mean is NA (it can happen), use the maille mean, and if this is also NA, use the EPCI mean.

Clearly the logic that we tried to apply here was to always take, when possible, the data using the closest municipalities, and therefore the most similar in terms of level of tourism, attractiveness, dynamics and other factors.
At the end of this process, we went from 409,000 to about 2,000 NAs. The remaining NAs, as mentioned, are from those 'particular' departments (57, 67, 68) for which we do not have any data, and will be treated separately.
Treating NAs means finding the most appropriate technique based on the context to obtain a clean dataset introducing the smallest possible error, which is exactly what we tried to do in this project, analyzing the structure of the data and all its geographical and administrative peculiarities.

2.5 Removing Outliers

Outliers are data points that deviate significantly from the majority of a data set, distinguishing them as anomalies or extreme values. They can occur for a variety of reasons, such as measurement errors, data entry errors, or natural variability in the data. While outliers can sometimes represent important findings or unique cases, more often they can distort the analysis, leading to misleading conclusions. The treatment of outliers is essential in data analysis to ensure the accuracy and reliability of the results. The goal of this part of the work is to identify and appropriately handle outliers, in order to improve the quality of the model as much as possible.

There are many techniques that allow to deal with outliers in the distribution, but I won't bother to list them all. Instead, I would like to focus on the technique that we decided to implement in this study, namely **Winsorization**.

Winsorization is a statistical technique used to reduce the effect of outliers by limiting the extreme values in a data set. Instead of removing these outliers, Winsorization fits them in the lower and upper tails to a specified percentile. This helps minimize the influence of extreme data points, making the data set more robust and suitable for analysis, while preserving the overall structure of the data.

Of course, it would be wrong to apply this technique to the entire data set indiscriminately.

The choice of variables to apply this technique to certainly has a subjective component, but an attempt was made to select census variables that concern small numbers. Since the population census is not exhaustive but is based on a sample, INSEE suggests handling small numbers with care. Here is what INSEE says about it: "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".

We therefore selected a few variables[7] and applied Winsorization to these variables only in municipalities with a population of less than 500 people.

This is because these municipalities are the ones that have the most outliers, since the small population influences the data of many other variables in our data set. For these variables, we applied Winsorization to the values within the lower and upper 5% tails of the normal distribution curve, fitting them to the closest value that falls outside these tails.

This process effectively compressed the distribution toward the center, thus reducing the influence of outliers and creating a more robust dataset for analysis.

2.6 Multivariate Analysis

At this point in the process, it is necessary, also following the OECD guide, to implement some multivariate analysis techniques. These methods can be used to study the relationships between variables and categories, better understand the weight and importance of each variable, and do a more in-depth study on which variables really have an influence on the model.

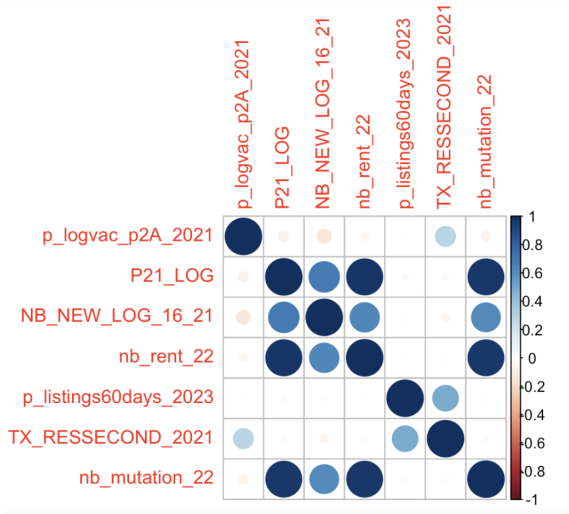
Below I will list the various analyses carried out, the motivations and the results.

[7] Suroccupation, Etudiants, Evol Seul, Evol Monoparent, Transfrontaliers, Travailleurs ext, Chômage

2.6.1 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 (corrplot, Figure 2.6) that can be observed in Appendix B.
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. The problem is therefore that we are dealing with a spurious and non-linear correlation, in which the real relationship between the variables and their influence on the category is not very clear.
Because of this problem, it is difficult to determine from a simple correlation analysis the real importance of each variable on the composite indicator. As noted by Becker et al.[4], 'The importance of xi (ed. the single variable) is also strongly dependent on its (possibly nonlinear) correlations with other variables, which are in turn correlated with each other. Therefore determining and isolating the effect of xi on y (ed., the CI) is by no means trivial'.

Figure 2.6: Correlation of the category 'Offre'



2.6.2 K-means

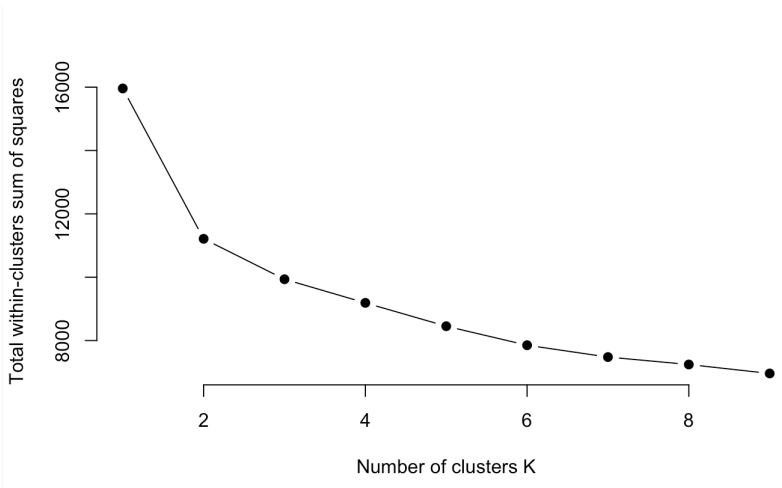
After studying the correlations, I started to get interested in clustering algorithms: I thought that by applying them to the dataset we had created, we could generate a partition of the municipalities that was at least interesting to analyze and perhaps to compare with the final scores of our CI.
The idea I had was in fact the following: if the variables in our dataset are the best to represent the real estate tension of each municipality, then it is possible that a combination of these variables on an n-dimensional plane can generate points close to each other for municipalities with a similar tension, thus identifying an approximate zoning.
I therefore decided to start with the K-means algorithm, one of the most popular clustering methods. Initially, it was necessary to select only the numerical variables and normal-

ize them: this is because, being an algorithm based on the calculation of the Euclidean distance between the points, if the variables have very different scales, those with larger values will influence the distance more and, consequently, the clustering result. Normalization instead makes all the variables comparable, ensuring that none of them dominates the calculation of the distances.
Then I tried to create 5 partitions (like the current zoning, Abis, A, B1, B2, C) and see for each group, how many municipalities belonged to each zone.
The results are shown in the table below:

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

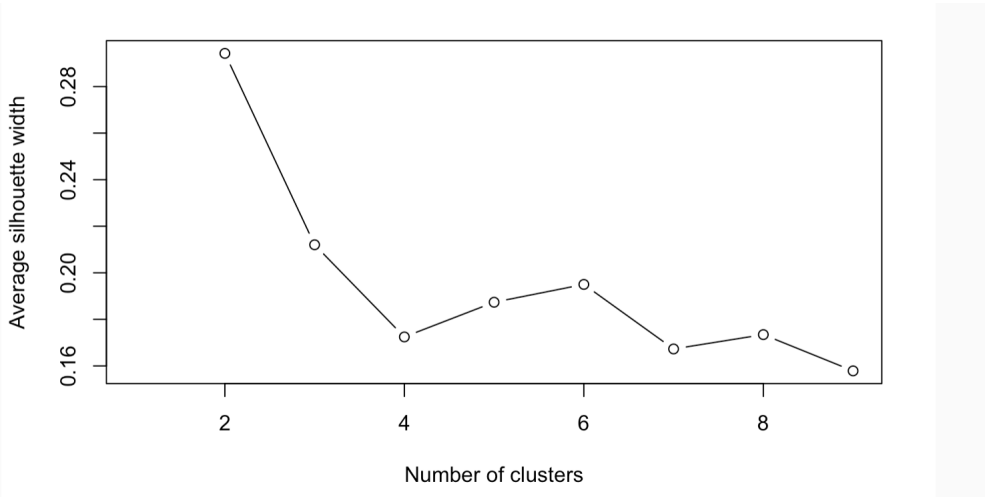
In the yellow boxes I wanted to highlight for each zone in which group the greatest number of observations fell: both the majority of the Abis and A zones fall into the first group, the B1 zone falls mainly into the third cluster and the B2 and C zones also fall predominantly into the same cluster, namely the fifth.
Although this subdivision effectively highlights a proximity between some areas (Abis and A, and B2 and C), the algorithm is not able to clearly separate them.
Of course, an arbitrary choice of the number of clusters would not be correct from a theoretical point of view. I therefore performed a check with the Elbow Method to identify the point at which adding additional clusters does not bring significant improvements to the data subdivision.
However, as you can see in Figure 2.7, this method does not give us a precise answer on the optimal k to use.

Figure 2.7: Elbow method



To finally identify the ideal k, I decided to use the Silhouette method. As we can see from Fig.2.8, this method suggests us to use 2 as the number of clusters.

Figure 2.8: Silhouette method



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

The reason why k-means doesn't perform well is because we are dealing with variables that do not have linear relationships between them. That means that our data cannot be separated into clusters by a straight line or a simple geometric shape. Applying k-means to this data results in clusters that cut off the natural structure of the data, resulting in poor performance and accuracy.
K-means cannot handle nonlinear relations in the data well because it relies on the Euclidean distance metric, which may not accurately capture the similarity between data points.

2.6.3 Spectral clustering

An algorithm that may yield better results is spectral clustering.
Spectral clustering is a clustering algorithm that attempts to find k clusters in a data set using the eigenvalues and eigenvectors of a similarity matrix. The algorithm works by first computing a similarity matrix that measures how similar each pair of data points is, based on a kernel function. It then applies a dimensionality reduction technique, such as Laplacian eigenmap or singular value decomposition, to the similarity matrix to obtain k eigenvectors corresponding to the k smallest eigenvalues. Next, each eigenvector is treated as a feature vector and stacked into a matrix. Finally, k-means or another clustering algorithm is applied to the rows of the matrix to obtain k clusters.
Unlike k-means, spectral clustering can handle nonlinear data well by using a kernel function, which can map the data into a higher-dimensional space where clusters are more separable.
I personally wrote a code to evaluate the results the spectral clustering, which is similar to the one used for k-means and allows to see for each group (cluster) to which zone the

municipalities that are part of it belong (I used the old zoning for this purpose, with data from July 2024). Also in this case I proceeded to normalize the data before applying the algorithm, and then I let the Spectral identify possible clusters.
In this way, if there are clearly visible patterns or correlations between a group and an area, they can be easily identified.
Since it is a fairly computationally expensive algorithm, I did it using only 15% [8] of the entire dataset. This certainly produced less precise results, but we can still notice the clear separation between Abis, A, and a third cluster in which, unfortunately, there is no real demarcation between zones B1, B2 and C.
The results are shown in the table below:

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

As can be seen from this table, which also highlight which cluster has the largest number of observations for each zone, it is not possible to observe well-defined patterns even using this method.
Given these outcomes, I decided to move on to something more complex, and in the next paragraph I will show the results of the Random Forest.

2.7 Random Forest

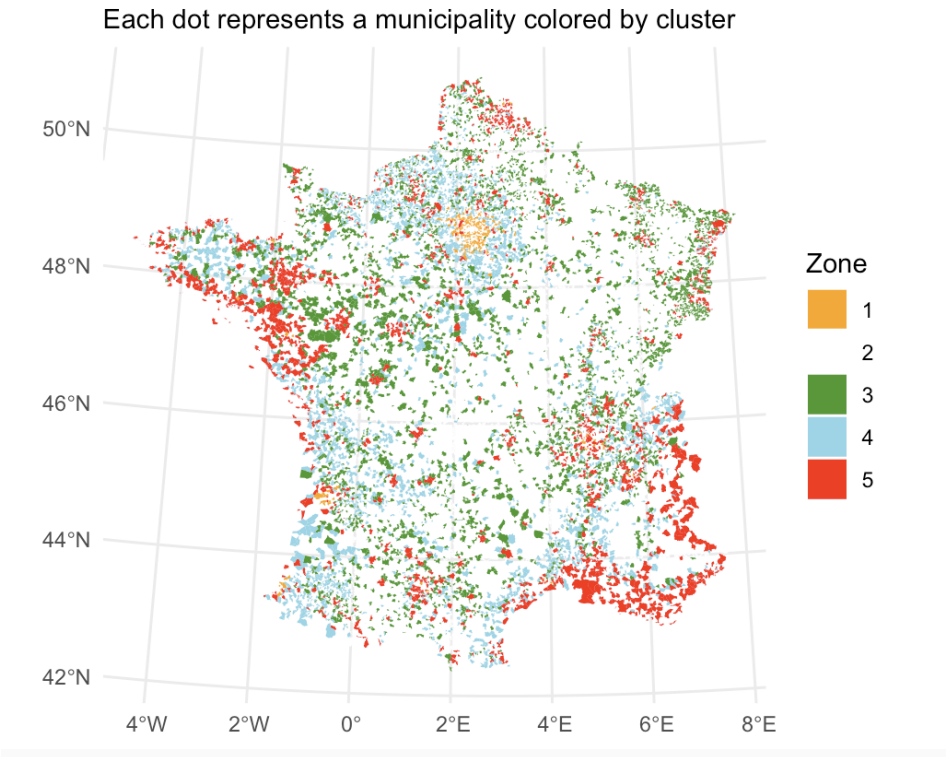
As a final clustering method I applied a Random Forest.
Random Forest is an ensemble learning method mainly used for classification and regression tasks.
However, it can also be adapted for unsupervised learning by measuring the proximity between data points. That happens thanks to the proximity matrix: in unsupervised learning with Random Forest, in fact, the algorithm generates a proximity matrix, which reflects the similarity between pairs of data points based on how often they end up in the same leaf node in the set of trees. This matrix can then be used as input for a clustering algorithm to identify groups of similar municipalities.
Furthermore, unlike K-means, Random Forest does not make assumptions about the linearity or shape of the data distribution and does not require prior normalization. It is suitable for capturing complex and non-linear relationships between variables, which is probably the case for our dataset.
As in the other cases I wanted to look for patterns in the clustering and, as for Spectral Clustering, also here the computational difficulty limited the study to a part of the original dataset.
What I found is shown in the table below:

[8] Please note that the total number of municipalities in the dataset is 34816

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

The result of the random forest is interesting because each group is predominantly assigned to one zone, with the exception of zone A and B1 which are mostly in the same cluster. It can also be noted in other cases that zone Abis and A are grouped together and in general municipalities 'nearby' in the zoning are more likely to end up in the same cluster.
This clearly reflects a pattern and tells us that, although not precisely, the algorithm is identifying the main characteristics of each zone.
To better visualize this result, I produced a geographical map of France (Fig 2.9), coloring each municipality according to the zone to which it was assigned, so that it can more easily be understood the result of the Random Forest.

Figure 2.9: Results of 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.

2.8 Inside COINr

From this point on, all the analysis will be done thanks to the R package 'COINr'.
To explain what COINr is, I would like to quote directly the words of the author of this package, the data scientist, policy analyst and researcher William Becker: "COINr is a high-level R package which is the first fully-flexible development and analysis environment for composite indicators and scoreboards. The main features can be summarised as features for building, features for analysis and features for visualisation and presentation.".[9]
In this section the data will first be normalized, then there will be a description of the weighting process and then the aggregation process. From the result of the latter, we will obtain partitions in the dataset through the Jenks algorithm and we will use them to simulate a new zoning.

2.8.1 Normalization

Normalization is a necessary step in this process, as it allows us to adjust the scale of the variables, ensuring that they contribute equally to the analysis, especially those that have different units or scales, and preventing those with larger ranges from disproportionately influencing the result.
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]. This is achieved by subtracting the minimum value of each feature and dividing by the interval (maximum - minimum).

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

This is a technique widely used in other similar works and ensures good normalization without losing important information.

2.8.2 Weighting

Another critical phase in the construction of the Composite Indicator is the choice of weights. For this reason it is important to clearly explain the methodology adopted and

[9] https://www.willbecker.me/projects/coinr/

the reason why it was chosen to adopt it, as well as what results would have been obtained by adopting other equally valid methodologies.
A first choice to make when assigning weights is whether to use equal weights (EW) or different weights (DW). It is however important to specify that "equal weights" does not necessarily mean "zero weights".
The choice of using equal weights is often used, as it proves to be the best choice in some specific situations. These are for example when:

- the theoretical framework underlying the construction of the composite indicator attributes the same importance to each variable in defining the composite concept to be measured,
- there is not sufficient statistical or empirical knowledge on which to base the definition of weights (for example: insufficient knowledge on the existence of causal relationships),
- the theoretical structure does not allow the formulation of hypotheses that give a specific and different weight to each variable,
- there is no agreement on the alternatives that can be adopted.

However, the choice to use EW can in other cases prove risky: one of the most common examples is when there is a strong correlation between two variables. This correlation could indicate that they represent the same underlying phenomenon, and therefore giving both an equal weight could lead to the issue of 'double counting', and consequently to have an unbalanced indicator.

In case we have enough information to be able to define different weights, they can be calculated by choosing between the following approaches: **statistical methods** (as the Correlation and Regression Analysis, the Principal Component Analysis (PCA), the Data Envelopment Analysis (DEA) or the Unobserved Components Models (UCM)), **multi-attribute models** (divided into 2 categories, the Multi-Attribute Decision Making, as the Analytic Hierarchy Processes (AHP), and the Multi-Attribute Compositional Model, as the Conjoint Analysis (CA)), and even **participatory methods** (Budget Allocation (BAL) or Public Opinion (PO), opinion polls).

In our specific case, we decided to opt for Equal weights for our first analysis.
The reason for this choice is that in this phase of the project any other approach could have proved to be reckless and risky from a conceptual point of view, because it could have assigned weights in a too subjective way and exposed to the risk of inconsistencies and distortions of the data.
However, one of the future projects could be to create an interface that allows the user to adjust the weights to their liking and observe how the zoning changes accordingly. Such a tool could be used by the government to run simulations, but also by the research community to test economic theories.

2.8.3 Aggregation

According to the OECD[2] Handbook on constructing composite indicators, aggregation methods can be divided into three main categories: linear, geometric, and multi-criteria.

Another crucial point in the aggregation phase is the choice between a “compensatory”, “partially compensatory” and “non-compensatory” approach.

The difference between these approaches consists in the degree of compensation: in the case of a **compensatory** method, a low value in one variable can be completely compensated by a high value in another. The variables are therefore considered interchangeable. In **partially compensatory** methods, on the other hand, there is limited compensation, i.e. there are minimum thresholds that cannot be exceeded through compensation.
Finally, in **non-compensatory** methods, no compensation is foreseen between the variables and each variable must reach a minimum level independently of the others.

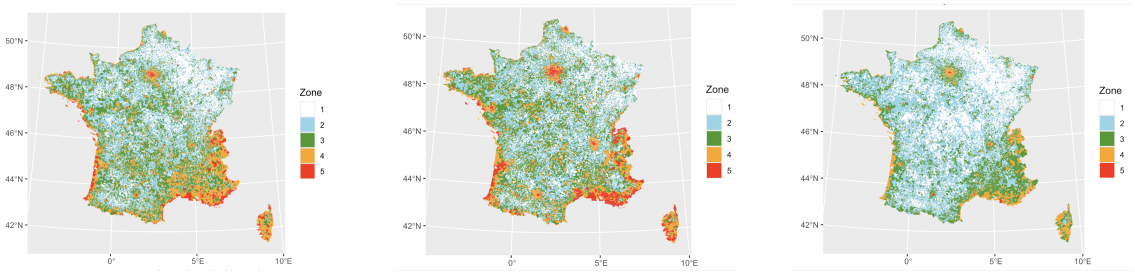
In our project, we decided early on to adopt a partially or non-compensatory approach to ensure a more precise and rigorous assessment. Adopting these approaches allows us to highlight problematic variables without the discrepancies between dimensions being masked or attenuated by compensation. This helps us to clearly identify areas of weakness and make informed decisions based on a complete view of performance.
We decided to start with the most common aggregation functions, present in the COINr package. These include aggregation via arithmetic mean, geometric mean, and harmonic mean. Below is a short table that highlights their main characteristics (Fig. 2.10):

Figure 2.10: Comparison of Aggregation Techniques

Mean Type	Description	Formula	Characteristics	Degree of Compensation
Arithmetic Mean	The sum of all values divided by the number of values.	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	Sensitive to extreme values; gives equal weight to all values.	Fully compensatory; low values can be offset by high values.
Geometric Mean	The nth root of the product of all values, where n is the number of values.	$\bar{x} = (\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$	Less sensitive to extreme values; useful for data with multiplicative effects.	Moderately compensatory; low values influence the mean significantly, but not fully.
Harmonic Mean	The reciprocal of the arithmetic mean of the reciprocals of the values.	$\bar{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$	Useful for rates and ratios; gives more weight to smaller values.	Less compensatory; very low values have a strong impact on the mean.

Although it may seem like a simple choice, it is important to always take into account several factors: first of all, these are normalized and Winsorized data, which therefore do not have values on different scales and do not present outliers. Furthermore, it is necessary to evaluate how the variables were constructed and finally choose the method with the best degree of compensation (for this specific case).
In this first result, I decided to choose the method that compensates the least among the three, namely the harmonic mean. It will be interesting later to try not only more complex

methods, but also to use two different methods in the two aggregation steps: our CI in fact provides for two aggregation stages, the first when the variables are aggregated in the 9 dimensions, the second when the dimensions are aggregated to create the CI. In this thesis I did not have time to properly investigate the various implications of combining two different aggregation methods, however I would like to show a graphical example to show how much this choice can influence the final output.
In Figure 2.11 we can in fact see the map of France with the various areas highlighted based on three different aggregation methods.
As methods we have: in figure a) the harmonic mean, used in both aggregation steps, in figure b) arithmetic and then geometric mean aggregation, finally in c) harmonic and then geometric mean aggregation.



(a) Harmonic Mean (b) Aritmetic and Geometric (c) Harmonic and Geometric

Figure 2.11: Three different aggregation methods confronted

It is important to underline that the chosen method also depends a lot on the objective that we want to achieve. In our case, for example, we would like to highlight the critical issues of real estate tension in many municipalities and ideally have more areas considered 'tense', so as to be adequately treated.
Also for this reason, I will present and study in the final analyses, precisely case a), as well as for the reasons explained before, also because it is the most balanced in terms of real estate tension and is able to effectively highlight the critical areas.

2.8.4 Jenks's Algorithm for partitions

In the previous section I did not explain an important part: when you perform the aggregation in COINr you get a dataset with the final value of the Indicator (Index) for each municipality. This Index is a continuous numerical value and allows you to have a 'ranking' of the various municipalities. However, it does not indicate how to partition them, in how many parts or based on which criteria. In fact, even in our project, the final word on how to divide the Indicator will be up to the Court of Auditors. However, proposing a subdivision method based on scientific criteria is very important and in this paragraph I would like to present the method I chose.
Jenks Algorithm, also known as 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. It is commonly used in data visualization, especially in mapping and geographic information systems (GIS), to create meaningful breaks in the data for classification.

In the case of this project, I 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 presented in the previous section and in the final results.

2.8.5 Sensitivity Analysis

The final step in building a Composite Indicator (CI) is actually a review step rather than a construction step: Sensitivity Analysis, in fact, is the process through which we evaluate how sensitive the CI is to the different methodological choices made during the construction process, such as the selection of variables, the normalization method, the weighting of individual indicators, or the aggregation method. This allows us to understand the impact of these choices on the final results and to ensure that the CI is robust and reliable.

A more rigorous description is provided to us once again by the OECD: " [...] sensitivity analysis is the study of how the variation in the output can be apportioned, qualitatively or quantitatively, to different sources of variation in the assumptions, and of how the given composite indicator depends upon the information fed into it. Sensitivity analysis is thus closely related to uncertainty analysis, which aims to quantify the overall uncertainty in country rankings as a result of the uncertainties in the model input. A combination of uncertainty and sensitivity analysis can help to gauge the robustness of the composite indicator ranking, to increase its transparency, to identify which countries are favoured or weakened under certain assumptions and to help frame a debate around the index.".[2]

As far as our research is concerned, we have not yet reached this point in the analysis and therefore the procedures that we will implement in this phase will not be present here for the moment, nor any concrete results.
While I cannot yet provide a critical analysis of the results in a data-driven manner, I can certainly make a brief summary of the possible **limitations** of this research.

First of all, we must certainly take into account a certain amount of subjectivity that inevitably was a component of almost all the phases: from the choice of the methodological approach, to the choices on which and how many categories to implement, which variables to select, which to winsorize and how to clean the dataset from NAs, up to the choice of weights and the aggregation method. In each stage there were 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.

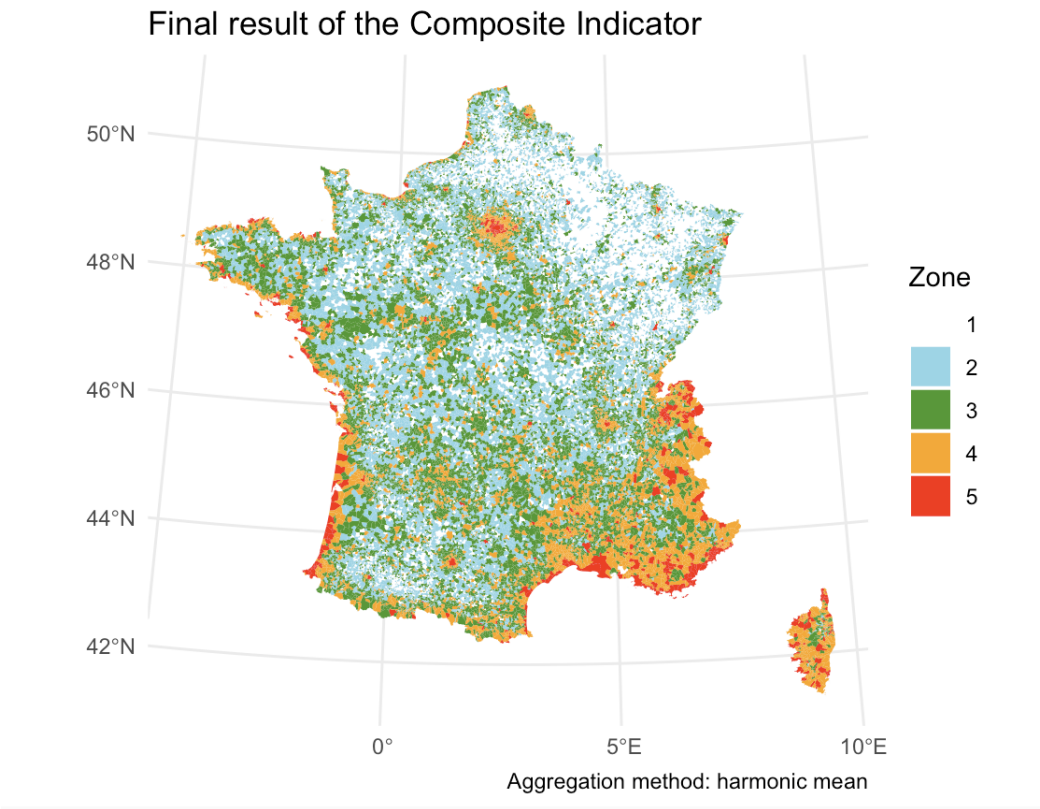
Another limitation is in the data. As I have already explained during this report and as I will also explore in the section dedicated to the challenges of this thesis, in the next chapter, the lack of much data has certainly constituted a limitation and is therefore another factor to be included among the limitations of this work, although the great attention given to this point has certainly helped a lot to mitigate this aspect.

A final limitation is time: 5 months is not a lot 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.

2.9 Final results

Finally in this conclusive part of the construction process I can present the results of the work done in these months, namely the finished Composite Indicator.
I would like to point out once again that this is only a first result and that there is still work to be done before it can be presented to the Court of Auditors, however this first output is certainly very important and summarizes well the work done during this internship.
The result obtained, clearly and directly visible in the geographical map in Fig. 2.12, shows a level of tension that is, in general, higher than the current one.

Figure 2.12: Harmonic Mean Aggregation



This first observation leads us to reflect on the discrepancy between the current Zoning and the actual data, which show us a more complex situation with many more facets. In fact, from the current map depicting the areas of real estate tension in France, it appears that almost 82% of the French territory is in zone C and therefore does not present tension. Comparing this data with the new zoning, we immediately notice how this percentage is incredibly modified:

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%

In fact, as we can see from the following table, many municipalities have been reclassified from lower tension zones to higher ones in the ranking:

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

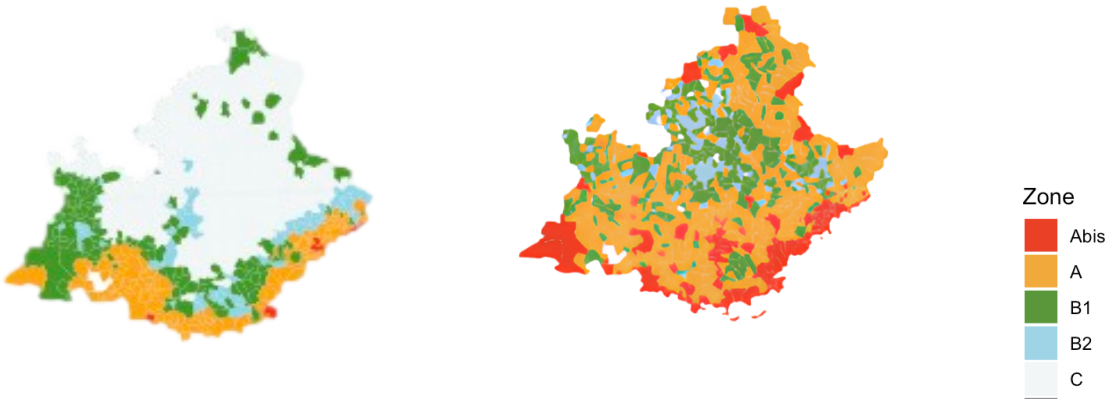
One of the questions I asked myself was whether this general increase was proportional between the various regions or whether there were some that had more drastic changes than others from this new zoning.
To verify this, I built a ranking of the various regions, from the most tense to the least tense, in the old and new zoning.
To do this, I assigned a score (from 1 to 5) to each zone (Abis=5, A=4, B1=3, B2=2, C=1) and calculated an average between the scores of each municipality in each region for the old and new zoning, adding a column representing the percentage change between these 2 values. The result is shown in Fig. 2.13

Figure 2.13: Percentage change in region's tension before/after

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% 🟢

As you can see, most regions have seen a high increase in tension but if we compare them, the increases have been quite proportional to each other (the range between 70 and 90% change), with the exception of some for which there has been a strong increase in tension (Brittany, Occitania, Nouvelle-Aquitaine) and others for which there has been a slight increase (Grand Est and Hauts-de-France). The only exception is the Ile de France region, which is the only one that has seen a slight **decrease** in overall tension.

This general increase can be seen even better by analyzing more closely one region in particular, the Provence-Alpes-Côte d'Azur region.
The difference between the current zoning and the one proposed in this thesis is depicted in Fig. 2.14



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

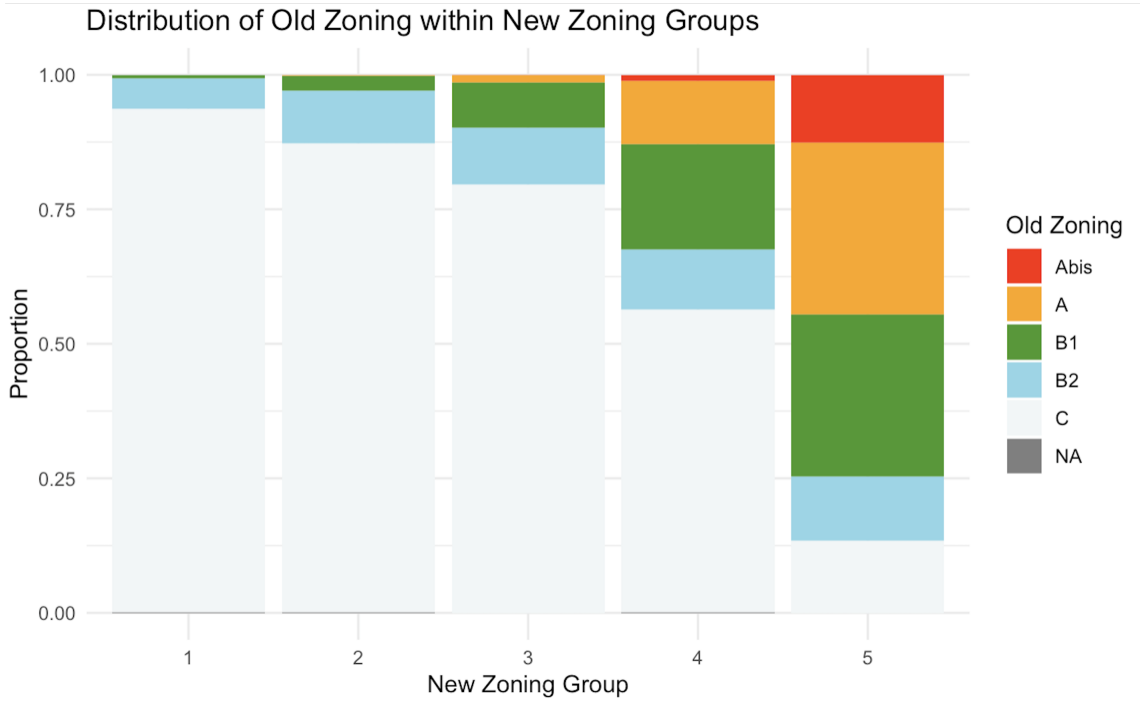
Figure 2.14: Before and after new zoning for region 93: Provence-Alpes-Côte d'Azur

In fact, this region, which is also one of the most tense in absolute terms, has much higher tension values in the new zoning compared to the old one.
In fact, it is a very interesting region from a tourist point of view: especially during the summer months, in this area there are numerous requests for housing from both French citizens and foreign tourists, many of whom come from Italy, given the proximity to the border.
However, even in winter this area is quite interesting and therefore in demand, especially in the Alps.
Furthermore, Marseille is the second most populous metropolis in France and has always represented (also thanks to the famous port) a place of great interest both from a cultural and work point of view.
It is important that the zoning knows how to adequately classify all the municipalities taking into account every possible factor of tension.

To conclude this analysis I would like to show, through the graph in Fig. 2.15, how the old zones have been redistributed in the new zoning.

The development of this Composite Indicator has thus allowed us not only to provide the State with a new tool to be able to make more informed and correct decisions, and

Figure 2.15: Distribution of old Zoning in new one



above all guided by data, but it has also been an opportunity for critical reflection and a very in-depth study on the critical issues but also the potential of each area and each municipality.

Conclusion

3.1 Applications of the CI

As mentioned in the introduction and explained in the first chapter, the importance of this research lies in the fact that it is a response to a concrete need, the problem of the division into residential zones and real estate tension, which required a rigorous statistical tool to be studied and corrected in the best possible way.
This composite indicator not only allows us to assign a score to each municipality based on its real estate tension, but also offers a look at the possible reasons for this tension, i.e. which sub-categories of the CI need to be fixed and which do not represent a danger. The investigation of the individual sub-categories offers a clearer picture of the situation and naturally, starting from this CI, it will be possible in the future to carry out studies and analyses on the correlation between the categories and the trend over time of tension in different areas of France.
In addition to these direct applications, there are others that are a little less straightforward. Among these we find:

- **'Zero artificialisation nette' (ZAN)**[5]

  A concept partially linked to this study is that of 'Zéro artificialisation net', established by the "Climate and Resilience" law of 22 August 2021.

  This law aims to better take environmental consequences into account during the construction and development of the territory, without neglecting the needs of the territory in terms of housing, infrastructure and activities, setting the objective of zero net artificialisation (ZAN) by 2050. This reduction in the rate of artificialisation directly affects real estate pressure, as it leads to a shortage of rents which benefits owners but penalizes tenants.

  A possible application for a Composite Indicator for real estate tension in France for ZAN could be that it serves as a comprehensive tool to monitor and evaluate the balance between urban development and environmental conservation. The composite indicator would provide a nuanced understanding of the pressures and trade-offs involved in achieving the ZAN goal.

- **Secondary houses**

  A clearer vision of the real estate tension situation would allow for an updated review of the regulation of secondary houses

- **Long-term Monitoring and Forecasting**

  The CI can provide a foundation for long-term monitoring and forecasting of real estate trends. By regularly updating the CI and analyzing changes over time, it will be possible to gain insights into the effectiveness of policies and the evolving dynamics of the housing market. This could be particularly useful to predict future

areas of tension before they become critical, allowing for proactive measures to be taken.

3.2 How I used the knowledge learned in the Master

This thesis, focused on measuring real estate tension in France through a composite indicator, served as a comprehensive application of the diverse skill set and knowledge base that I developed during my Master's degree.
This research was a direct test of the academic training and practical tools that I acquired during this last year, demonstrating my ability to synthesize, apply and extend what I learned in school in a more complex and real-world setting.
The knowledge provided to me by the master and put into practice during this internship can be divided into some macro areas:

1. **Econometric Modeling and Analysis**

   The core of this thesis involved the application of econometric analysis to evaluate at each step the correctness of the applied methods. The econometric courses provided me with a deep understanding of the various models implemented, and allowed me to have a more critical look during all the steps and the various development phases, all essential elements in the construction and refinement of the Composite Indicator.

   I was able to apply correlation analysis methods, clustering, Random Forest and various diagnostic tests to ensure the robustness and validity of the models used in the study.

2. **Statistical methods and theory**

   Throughout the thesis, statistical methods played a fundamental role in the construction and validation of the composite indicator.

   The knowledge acquired from the courses on statistical theory was applied in various phases of the research. For example, 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.

   Of course, this is just an example, but in reality our statistical knowledge was often put to the test during the design of the indicator, the selection of the relevant variables and the validation of the results.

3. **Data Visualization and Interpretation**

   Effective communication of complex results is crucial to any analytical work. My studies included significant training in data visualization, which I applied extensively in this thesis. Tools such as R and Python (with libraries such as Ggplot for the first and Seaborn for the second) were used to create clear and informative visualizations that aided in the interpretation of the composite indicator results.

   These visual aids not only supported the analysis, but also made the results more accessible to a wider audience (when the indicator was periodically presented to ministers for approval during various stages).

4. **Practical application of theoretical knowledge**

   During this last year, in M2 we had to produce numerous group projects. Although this modality is not always appreciated, it has proven to be very useful in the world of work: it 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

The successful completion of this thesis is a testament to the comprehensive training I have received during my Master 2 in Econometrics, Big Data and Statistics. Each stage of this research has been an opportunity to apply and improve the knowledge and skills developed during my studies, demonstrating not only my ability to handle complex analytical tasks, but also the practical relevance and applicability of the program's curriculum in addressing real-world econometric-statistical challenges.

3.3 Challenges and new learnings from this Internship

This internship certainly presented many challenges.
First of all, since we are not French citizens, it was necessary to do a preliminary study on some specific mechanisms of France: in particular, we studied its political geography (division into regions, departments, EPCI etc.), the functioning of the laws on housing, Social housing, the distribution of the population and the problem of the high number of municipalities, just to name a few.
Regarding this last point, another great difficulty of this work was related to the data, and in particular to their cleanliness: in fact, having so many municipalities implies that many of them have a very large population, while others have a very small one. This, in addition to bringing imbalance, also means that for many municipalities we did not have the exact data, either because they were classified as sensitive data (statistical secrecy) or because they did not exist at all. Historical-political reasons also influence the data: for the departments of Moselle (57), Bas-Rhin (67) and Haut-Rhin (68), for example, some data relating to rents were not available and therefore it was necessary to obtain them from other sources (Airdna, SeLoger etc).
Of course, there were also many daily challenges, for example to find the most effective method to do a certain analysis, or to improve some code and make it more performant.
However, all these challenges were overcome, also thanks to the help and discussion with our supervisors and the other team members. This is precisely what allowed us to grow and learn many new things, both specific to the objective of this work (we learned a lot about composite indicators, how to develop a research, the problems with these types of data and with these geographical specifications), but also other more general knowledge, such as how to work in a team, what are the roles and what is expected from each one and how important it is to always continue to review what you know while learning new things.

3.4 A step-back final analysis

This research started from a very specific objective: the creation of a Composite Indicator to measure real estate tension. Having to make a critical analysis of my contribution to the project and in general of the work carried out in these months, I would divide it into two parts: the initial part (more theoretical) and the central part (more practical but still with a theoretical component).

During the initial part, which corresponded approximately to the first two months, the main objective was to build a framework, that is, to have a general idea of all the aspects relevant to this research and define the subsequent steps and possible resolution approaches. In this phase my task was to critically study the existing literature and contribute ideas and reflections during the weekly meetings. I believe that my contribution helped in some situations to make important decisions for the project, having always actively participated and brought new information or insights to each new meeting. If I had to improve something, now that I have more experience, I would immediately apply more efficient methods of reading and summarizing the papers, so that I can find the information more easily later.
In the second phase of the project this component of study and in-depth analysis has always remained, but an important practical component has naturally been added. While on some parts there was no clear division (for example the study of the sign, the choice of categories and other initial steps were done collaboratively by the team, discussing each personal proposal of the members), other parts were divided, and both me and Aldo were assigned specific tasks.
Among these, I was responsible for doing an important part of the cleaning of the dataset from missing values (explained in Chapter 2), which was used to fix more than 90% of the dataset, aggregating at different levels based on the availability of data in the neighbors. I believe that this part was carried out very effectively and carefully to the specifications and, having allowed the analyses to continue, it was also a fundamental part of the process.
I then carried out the Multivariate Analysis, also choosing the algorithms to apply: initially a classic correlation made however by categories, then a cluster analysis implementing in order: K-means, Spectral Clustering and Random Forest. A great difficulty in this phase was represented by the high computational cost of these last two algorithms. However, the analysis produced very interesting results, which I visualized graphically through tables and geographical maps.
Finally, I used COINr to create a simulation of the results and I also carried out investigations and analyses on them, although the time left was not enough to do an even more in-depth study. On this part, there are certainly many things still to be reviewed, changed and fixed, however I am satisfied with the result, although still embryonic, that we obtained.

3.5 Conclusion

This experience has certainly been one of the most important and valuable of my life so far. It was not only my first work experience in this field, but it was also a bit of a discovery of different opportunities from those I had always imagined. The field of research was a pleasant discovery and, thanks also and above all to the commitment and great kindness and availability of Marie and Julie, to the fantastic team, and to a serene and satisfying work environment, I will certainly have a beautiful memory of this first work experience.
Having the opportunity to study in France, in a university recognized throughout the world, has made this last year of the Master much more intense and satisfying, and has allowed me to develop many new sides of myself, to practice a new and very beautiful language, to also learn new things and with different educational methods compared to Italy. In short, the last year has been full of growth in every way, and I am grateful to everyone for this great opportunity that I have had.

In addition, thanks again to Marie and Julie, we will also have the pleasure of concluding this project that certainly requires a lot of time, since we have been offered the opportunity to stay and work for CESAER for another six months.
This thesis will therefore remain partially unfinished but the research will of course continue and we are very happy about this.

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 interactive 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. Springer-Verlag GmbH Germany, part of Springer Nature 2018, 2018.

[8] E. Fusco, M. P. Libório, H. Rabiei-Dastjerdi, F. Vidoli, C. Brunsdon, and P. I. Ekel. *Harnessing Spatial Heterogeneity in Composite Indicators through the Ordered Geographically Weighted Averaging (OGWA) Operator*. Geographical Analysis. Geographical Analysis, 2023.

[9] S. Greco, A. Ishizaka, M. Tasiou, and G. Torrisi. *On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness*. The Author(s), 2018.

[10] L. D. Phillips, A. D. Pearman, M. Spackman, and J. S. Dodgson. *Multi-criteria analysis: a manual*. Department for Communities and Local Government, January 2009.

[11] M. Schlossarek, M. Syrovátka, and O. Vencálek. *The Importance of Variables in Composite Indices: A Contribution to the Methodology and Application to Development Indices*. Springer Nature, May 2019.

[12] F. Ferreira and J. Gyourko. *Heterogeneity in Neighborhood-Level Price Growth in the United States, 1993–2009*. American Economic Review, 2012.

[13] J-M.A. Martíneza, S. Suárez-Seoanea, and E. De Luis Calabuiga. *Modelling the risk of land cover change from environmental and socio-economic drivers in heterogeneous and changing landscapes: The role of uncertainty*. Elsevier B.V., January 2011.

[14] R. Attardi, M. Cerreta, V. Sannicandro, and C.M. Torre. *The Multidimensional Assessment of Land Take and Soil Sealing*. ResearchGate, June 2015.

[15] E. Burton. *Measuring urban compactness in UK towns and cities*. Oxford Centre for Sustainable Development, 2001.

[16] J. Chica-Olmo, R. Cano-Guervos, and M. Chica-Rivas. *Estimation of Housing Price Variations Using Spatio-Temporal Data*. Article, March 2019.

[17] S. Holly, M. H. Pesaran, and T. Yamagata. *A Spatio-Temporal Model of House Prices in the US*. IZA Discussion Paper, September 2006.

[18] B. Huang, L. Zhang, and B. Wu. *Spatiotemporal analysis of rural–urban land conversion*. International Journal of Geographical Information Science, October 2010.

[19] E.G. Irwin and J. Geoghegan. *Theory, data, methods: developing spatially explicit economic models of land use change*. Elsevier, October 2010.

[20] T. Lan, G. Shao, Z. Xu, L. Tanga, and Sun L. *Measuring urban compactness based on functional characterization and human activity intensity by integrating multiple geospatial data sources*. Ecological Indicators. Elsevier, 2021.

[21] K. Beaubrun-Diant and T.P. Maury. *Implications of homeownership policies on land prices: the case of a French experiment*. Economics Bulletin, Volume 41, Issue 3, 2021.

Appendix A

I Data Sources

Here is an explanation of the acronyms and terms you've mentioned, which are related to data and real estate in France:

1. COG (Code Officiel Géographique): A coding system managed by INSEE (Institut national de la statistique et des études économiques) that provides official geographic codes for French municipalities, departments, and regions. It is essential for structuring and analyzing territorial data.

2. INSEE (Institut National de la Statistique et des Études Économiques): The French National Institute of Statistics and Economic Studies, responsible for producing and analyzing statistical data about the French economy and population, including data used in real estate and housing studies.

3. RP (Recensement de la Population): Refers to the French population census, conducted by INSEE. It gathers demographic information crucial for urban planning, real estate analysis, and public services.

4. Fichier Détail Logement: A dataset from the population census, focusing on detailed information about housing characteristics, such as the size, type of housing, and occupancy status, which is used for real estate analysis.

5. DV3F (Demande de Valeurs Foncières): A dataset provided by the French government that contains information on real estate transactions, specifically sales values of land and buildings. It is used to track real estate trends and property values.

6. Cerema (Centre d'Études et d'Expertise sur les Risques, l'Environnement, la Mobilité et l'Aménagement): A public institution that provides expertise on environmental risks, urban planning, and mobility. It also provides data on housing and urban development, contributing to real estate projects and policy-making.

7. Filosofi (Fichier Localisé Social et Fiscal): A dataset produced by INSEE, detailing localized social and fiscal information, including household income and socioeconomic data. It helps in analyzing the social aspects of real estate markets.

8. Carte des Loyers: Refers to rental price maps in France, typically produced by various real estate data providers or public institutions, showing rental trends across different regions and cities.

9. Anil (Agence Nationale pour l'Information sur le Logement): The National Housing Information Agency, a public body that provides legal and practical information about housing in France. It offers guidance for tenants, landlords, and homeowners.

10. Sitadel2 (Système d'Information et de Traitement Automatisé des Données Élémentaires sur les Logements et les Locaux): A database that contains detailed information about building permits and construction in France, managed by the Ministry of Ecological Transition. It's useful for analyzing new housing supply.

11. SDES (Service de la Donnée et des Études Statistiques): The Statistical and Data Service, a branch of the French Ministry for the Ecological and Inclusive Transition, responsible for providing statistical data on housing, environment, and energy.

12. Fichier Détail Mobilité Professionnelle: A dataset that tracks professional mobility, showing where people are moving for work. It's important for understanding housing demand in regions with high job mobility.

13. Lovac (Localisation des Opérations et Valeurs): A dataset that provides geographical information about real estate transactions, including prices and types of properties sold.

14. La Poste (Réexpéditions de Courrier): La Poste is the French postal service, and the réexpéditions de courrier refers to data on mail forwarding requests. This information can provide insight into residential mobility and changes in residence.

15. Airdna: A private company that collects and analyzes data from short-term rental platforms like Airbnb and Vrbo. It provides insights into the short-term rental market, occupancy rates, and pricing trends.

16. REE (Référentiel Energies et Environnement): A dataset or framework focusing on energy and environmental data, often used to analyze the energy performance and environmental impact of buildings and urban areas.

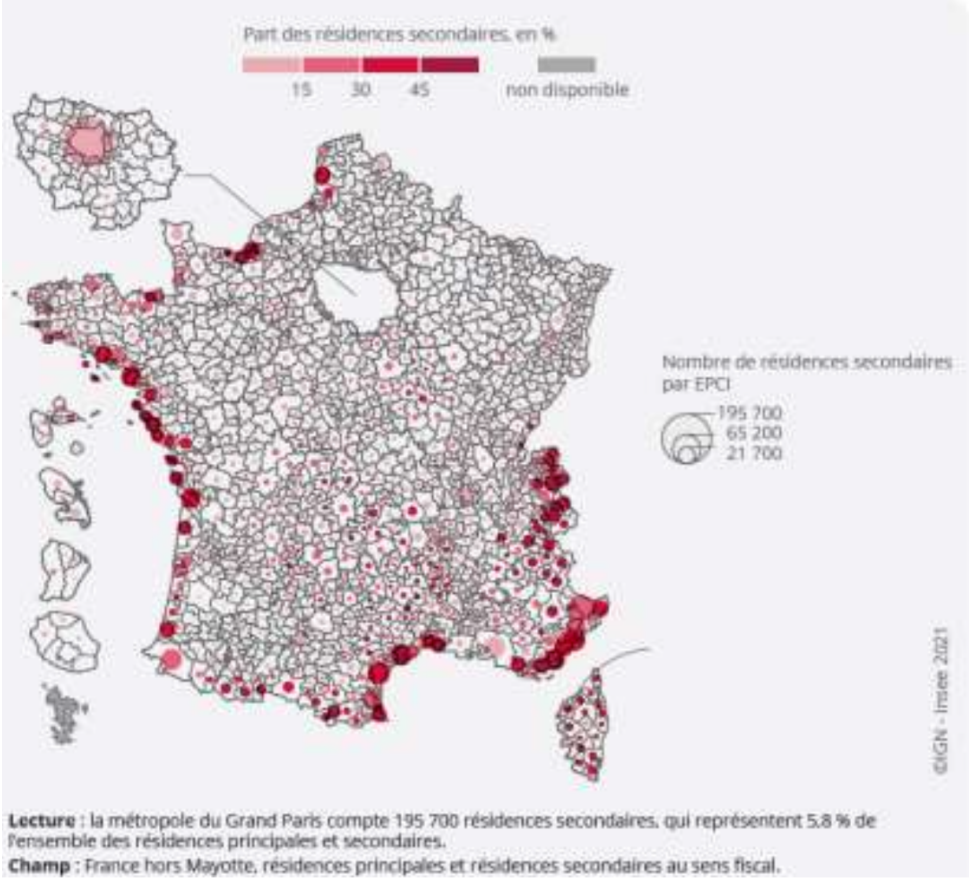
17. Observatoire DPE - AUDIT: DPE stands for Diagnostic de Performance Énergétique (Energy Performance Diagnosis), and the Observatoire DPE monitors and audits the energy efficiency of buildings in France, crucial for sustainability efforts in real estate.

18. ADEME (Agence de l'Environnement et de la Maîtrise de l'Énergie): The French Environment and Energy Management Agency, which promotes sustainable development and provides data and resources on energy efficiency in buildings and real estate projects.

19. AVIV Group: A digital real estate company that owns and manages some of the main real estate platforms in France, including SeLoger and MeilleursAgents. These platforms provide property listings, real estate evaluations, and services for buyers, sellers, and renters.

II Second houses

Figure A.1: Number and share of second homes by municipality



Data source : Insee, Fideli 2017.
Graph source: [3]

Appendix B

I Correlation plots

Figure B.1: Correlation of the category 'Abordabilité du logement'

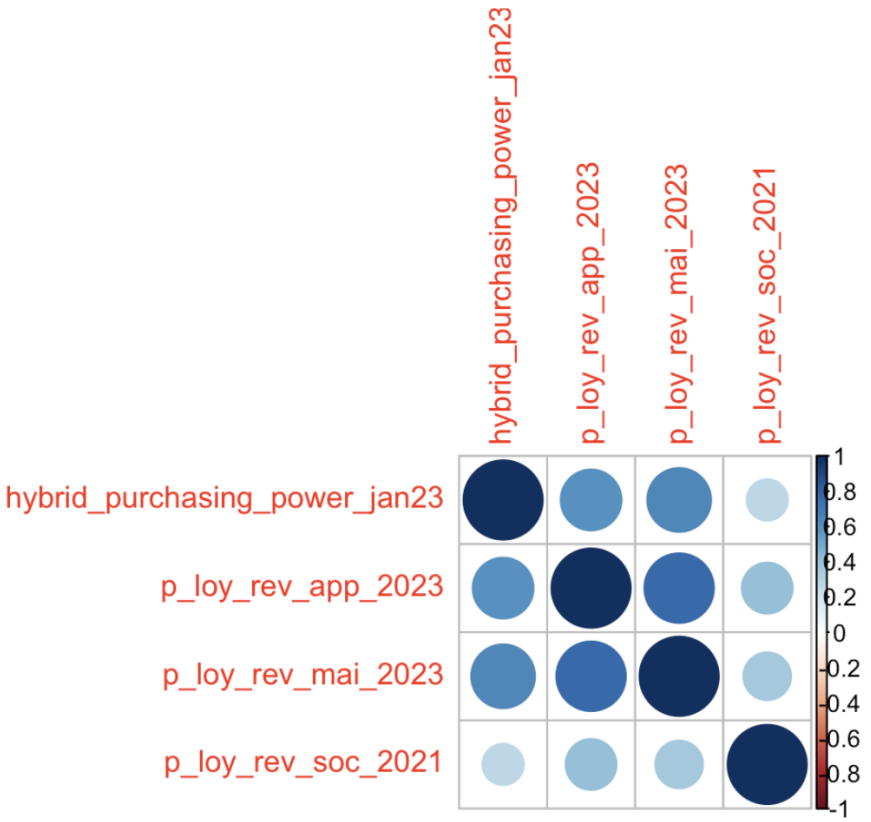


Figure B.2: Correlation of the category 'Attractivite economique'

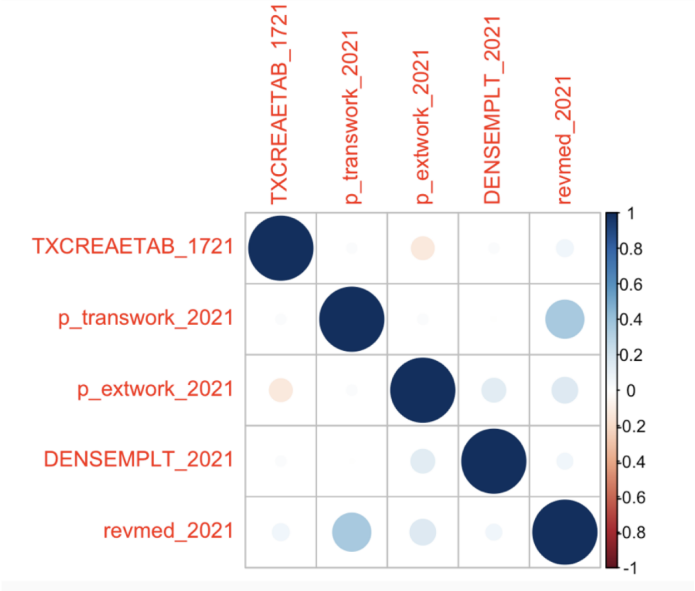


Figure B.3: Correlation of the category 'Attractivite touristique'

Figure B.4: Correlation of the category 'Demande'

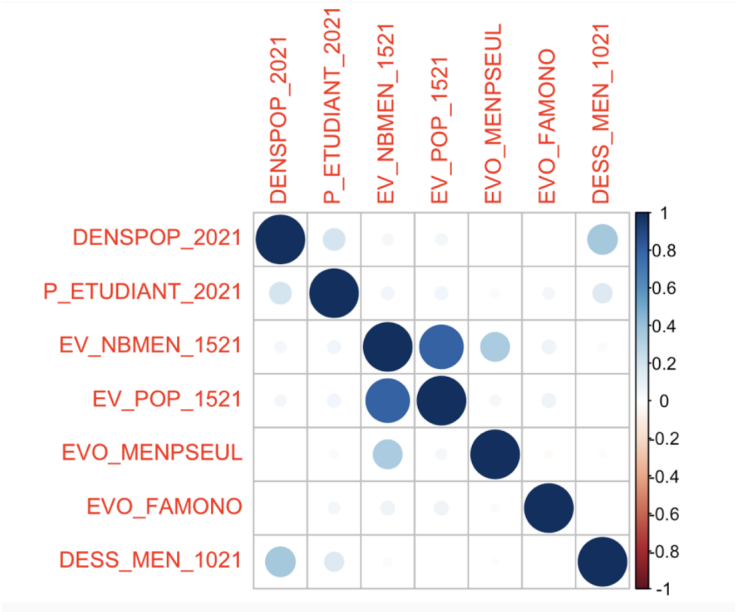


Figure B.5: Correlation of the category 'Distance emploi:equipements'

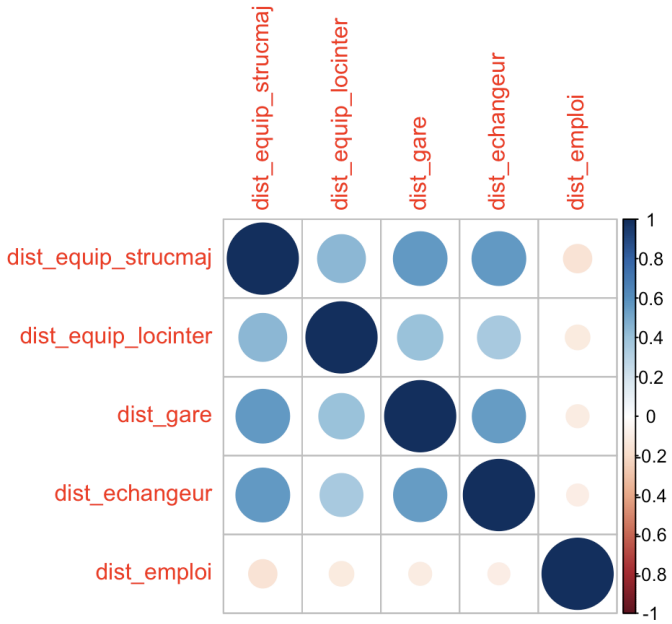


Figure B.6: Correlation of the category 'Dynamique du marche'

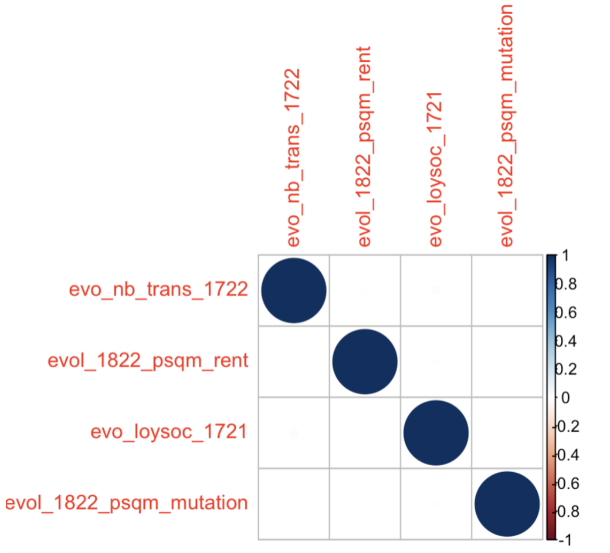


Figure B.7: Correlation of the category 'Inadequation offre:demande'

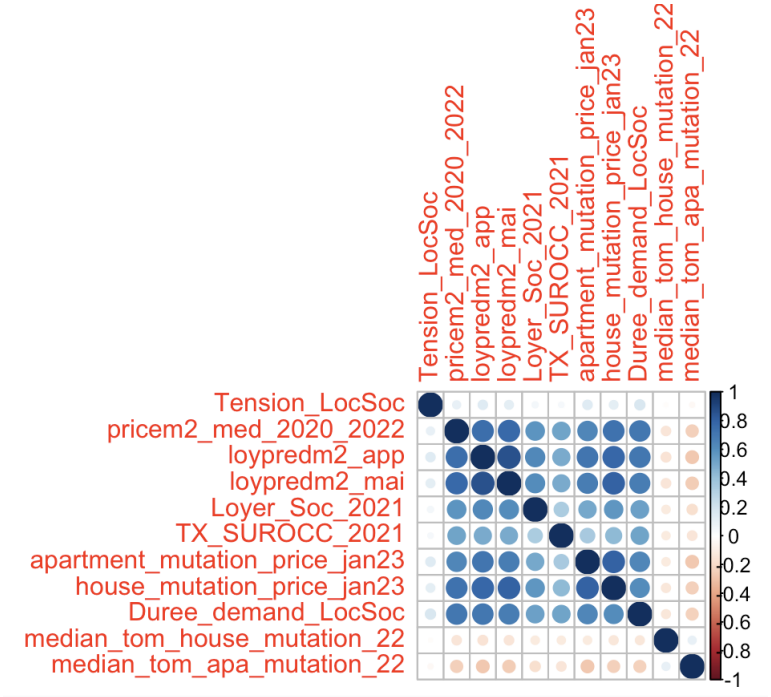


Figure B.8: Correlation of the category 'Offre'

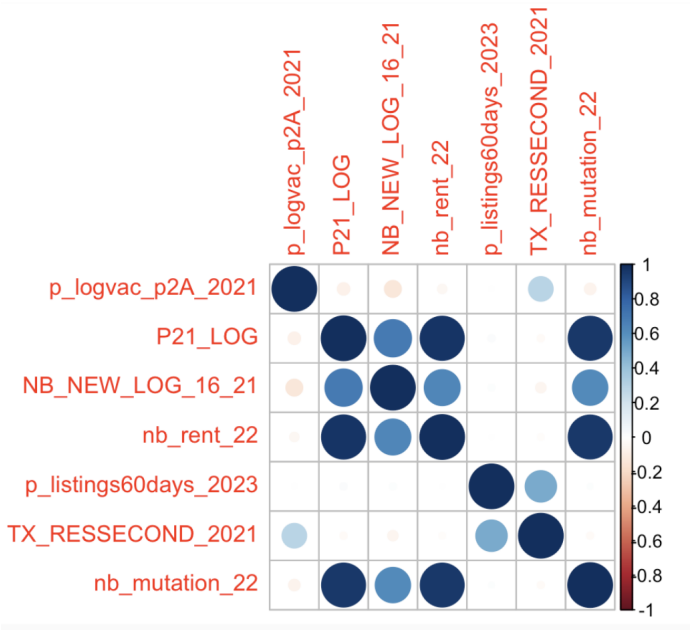
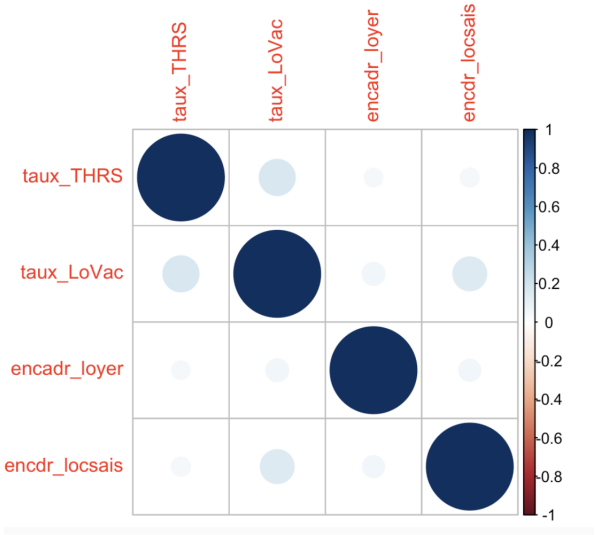


Figure B.9: Correlation of the category 'Regulation'



Appendix C

I Sheet of the dataset

Figure C.1: Excel with all the information about the dataset

Variable	Nom	Dimension	Polarité sur CI	Polarité sur la dimension	Polarité dimension sur CI	Description	Source	to_normalize	Outliers
CODGEO		identifiant				Code insee de la commune	COG 2023, Insee		
LIBGEO		identifiant				Libellé de la commune	COG 2023, Insee		
DEP		identifiant				Code insee du département	COG 2023, Insee		
REG		identifiant				Code insee de la région	COG 2023, Insee		
EPCI		identifiant				Code insee de l'EPCI	COG 2023, Insee		
ZE2020		identifiant				Code insee de la zone d'emploi	COG 2023, Insee		
aav2020		identifiant				Code insee de l'aire d'attraction des villes	COG 2023, Insee		
code_insee_habitat		identifiant				Code insee habitat	COG 2023, Insee		
hybrid_purchasing_power_jan23	Taux Effort Immo.	Abordabilité du logement	+	-	-	Prix au m2 médian en 2023 / (capacité d'emprunt maximale (sur la base du revenu médian par UC 25	AVIV		
p_loy_rev_app_2023	Taux Effort Loc Appart.	Abordabilité du logement	+	-	-	Loyer annuel d'un appartement de 52 m² en 2023 / revenu médian par UC des locataires du parc priv	Carte des Loyers 2023, Anil; Filosofi 2021, Insee		
p_loy_rev_mai_2023	Taux Effort Loc Maison	Abordabilité du logement	+	-	-	Loyer annuel d'une maison de 92 m² en 2023 / revenu médian par UC des locataires du parc privé en	Carte des Loyers 2023, Anil; Filosofi 2021, Insee		
p_loy_rev_soc_2021	Taux Effort Loc Social	Abordabilité du logement	+	-	-	Loyer annuel d'un logement social de 60 m² en 2021 / revenu médian par UC des locataires du parc	RPLS 2021, Anil; Filosofi 2021, Insee		
Tension_LocSoc	Ratio demande social	Inadéquation offre/demande	+	+	+	Nombre de demandes de logement social en cours / nombre de demandes satisfaites en 2022	SNE 2022		
prixm2_med_2020_2022	Prix terrain	Inadéquation offre/demande	+	+	+	Prix médian au m² des terrains à bâtir entre 2020 et 2022 (euros)	DVSF 2020-2022		
loyprdm2_app	Loyer appart	Inadéquation offre/demande	+	+	+	Loyer/m2 pour un appartement de 52 m² en 2023 (euros)	Carte des Loyers 2023, Anil		
loyprdm2_mai	Loyer maison	Inadéquation offre/demande	+	+	+	Loyer/m2 pour une maison de 92 m² en 2023 (euros)	Carte des Loyers 2023, Anil		
Loyer_Soc_2021	Loyer social	Inadéquation offre/demande	+	+	+	Loyer moyen/m² dans le parc social en 2021 (euros)	RPLS 2021		
TX_SUROCC_2021	Suroccupation	Inadéquation offre/demande	+	+	+	Taux de suroccupation des résidences principales (hors studio de 1 personne) en 2021 (%)	RP 2021, Insee	Yes	
appartement_mutation_price_jan23	Prix appart	Inadéquation offre/demande	+	+	+	Prix d'un appartement de 40 m² au 2ème étage en janvier 2023	AVIV		
house_mutation_price_jan23	Prix maison	Inadéquation offre/demande	+	+	+	Prix d'une maison de 100 m² en janvier 2023	AVIV		
Duree_demand_LocSoc	Attribution social	Inadéquation offre/demande	+	-	-	Temps d'attribution d'un logement social en 2022 (mois)	SNE 2022		
median_tomi_house_mutation_22	Temps vente appart.	Inadéquation offre/demande	+	-	-	Temps de vente médian d'un appartement en 2022	AVIV		
median_tomi_apa_mutation_22	Temps vente maison	Inadéquation offre/demande	-	+	+	Temps de vente médian d'une maison en 2022	AVIV		
dist_emploi	Dist emploi	Distance emploi/équipements	+	-	-	Accessibilité potentielle à l'emploi (accessible routière en min)	RP 2021, Insee, Metric OSRM,Insee		
dist_equip_structurag	Dist équip majeurs	Distance emploi/équipements	-	+	+	Temps d'accès au centre d'équipement structurant ou majeur le plus proche (accessible routière en	ANCT, INSEE, BPE, Metric OSRM,Insee		
dist_equip_bcenter	Dist équip bcenter	Distance emploi/équipements	-	+	+	Temps d'accès à la gare la plus proche (accessible routière en min)	Réseau ferré (Gare Voyageur SNCF, BD Carto IGN, Metric OSRM,Insee		
dist_gare_plus	Dist gare plus	Distance emploi/équipements	-	+	+	Temps d'accès à la gare la plus proche (accessible routière en min)	Réseau ferré (Gare Voyageur SNCF, BD Carto IGN, Metric OSRM,Insee		
dist_echangeur	Dist échangeur	Distance emploi/équipements	-	+	+	Temps d'accès à l'échangeur le plus proche (accessible routière en min)	BD Carto, IGN, Metric OSRM,Insee		
DENSPOP_2021	Densité pop	Demande	+	+	+	Densité de population à l'hectare en 2021	RP 2021, Insee		
P_ETUDIANT_2021	Etudiants	Demande	+	+	+	Part des étudiants dans la population dès 15-64 ans en 2021 (%)	RP 2021, Insee	Yes	
EV_NBMEN_1521	Evol Menages	Demande	+	+	+	Taux d'évolution du nombre de ménages entre 2015 et 2021 (%)	RP 2021, Insee		
EV_POP_1521	Evol Pop	Demande	+	+	+	Taux d'évolution de la population entre 2015 et 2021 (%)	RP 2021, Insee	Yes	
EVO_MENPSEUL	Evol Seul	Demande	+	+	+	Taux d'évolution du nombre de ménages composés d'une seule personne entre 2015 et 2021 (%)	RP 2021, Insee	Yes	
EVO_FAMMONO	Evol Monoparent	Demande	+	+	+	Taux d'évolution du nombre de familles monoparentales entre 2015 et 2021 (%)	RP 2021, Insee	Yes	
DESS_MEN_1021	Desserrement ménages	Demande	+	+	+	Desserrement des ménages : (Population en 2010/taille moyenne des ménages en 2021) - (Populati	RP 2010, 2021, Insee	? With logement	
leads_per_day_rent	Leads annonce loc	Demande	+	+	+	Nombre de leads (prise de contact par email ou téléphone) par annonce de location	AVIV		
leads_per_day_mutation	Leads annonce vente	Demande	+	+	+	Nombre de leads (prise de contact par email ou téléphone) par annonce de vente	AVIV		
evo_nb_trans_1722	Evol transactions	Dynamique du marché	+	+	+	(Nombre de transactions immobilières en 2022 - Nombre de transactions immobilières en 2017) / N	DVSF 2017, 2022		
evol_1822_psqm_rent	Evol loyer annonce	Dynamique du marché	+	+	+	Evolution des loyers d'annonces entre 2018 et 2022	AVIV		
evo_loysoc_1721	Evol loyer social	Dynamique du marché	+	+	+	Evolution du loyer moyen/m² dans le parc social entre 2017 et 2021	RPLS 2021		
evol_1822_psqm_mutation	Evol prix annonce	Dynamique du marché	+	+	+	Evolution des prix d'annonces 2018-2022	AVIV		
TXCREATER8_1721	Creation etab.	Attractivité économique	+	+	+	Taux de création moyen d'établissements entre 2017 et 2021 (moyenne des créations d'établisseme	REE 2021, Insee		
p_transwork_2021	Transfrontaliers	Attractivité économique	+	+	+	Part des travailleurs transfrontaliers en 2021 (%)	RP richer detail mobilité professionnelle 2021, Insee	Yes	
p_navetsortie_2021	Travailleurs ext.	Attractivité économique	+	+	+	Part des travailleurs en provenance de l'extérieur de la commune en 2021	RP richer detail mobilité professionnelle 2021, Insee	Yes	
p_navetentre_2021	Travailleurs int.	Attractivité économique	+	+	+	Part des travailleurs internes à la commune en 2021	RP richer detail mobilité professionnelle 2021, Insee	Yes	
TXC_CHOMT1564_2021	Chomage	Attractivité économique	+	+	+	Taux de chômage des 15-64 ans en 2021 (%)	RP 2021, Insee	Yes	
nbemploi_2021	Revenu	Attractivité économique	+	+	+	Revenu médian disponible par UC en 2021	Filosofi 2021, Insee		
p_logvac_p2A_2021	Part vacance +2 ans	Offre	-	+	+	Part des logements vacants depuis plus de 2 ans en 2021 (%)	LOVAC 2021, Cerema		
P21_LOG	Nb logements	Offre	-	+	+	Nombre de logements en 2021	RP 2021, Insee	? With population	
NB_NEW_LOG_16_21	Nb constructions	Offre	-	+	+	Nombre de logements construits entre 2017 et 2021	Sitadel, 2017-2021	? With logement	
nb_rent_22	Volume annonces loc	Offre	-	+	+	Volume d'annonces proposées à la location en 2022	AVIV		
p_listing60days_2023	Airbnb reserv 60 jours	Offre	-	+	+	Nb de listings réservés plus de 60 jours par an en 2023 / Nb de logements du parc privé en 2021 * 10	Airbnb, 2023; RP 2021, Insee		
TX_RESSECOND_2021	Part secondaire	Offre	-	+	+	Part des résidences secondaires dans le parc de logements en 2021 (%)	RP 2021, Insee		
nb_mutation_22	Volume annonces vente	Offre	-	+	+	Volume d'annonces proposées à la vente en 2022	AVIV		
potentiel_construct	Potentiel constructible	Offre	-	+	+	Part de la surface communale potentiellement constructible (%)	CEREMA, BDAtlt, INPN		
taux_THRS	Tax Hab Res Second	Régulation	-	+	+	Taux cumulé de la taxe d'habitation sur les résidences secondaires en 2022 (%)	REI 2022		
taux_LoVac	Tax Hab Log Vac	Régulation	-	+	+	Taux cumulé de la taxe sur les logements vacants en 2022 (TLV ou THLV) (%)	REI 2022		
encadr_loyer	Encadrement	Régulation	-	+	+	Commune appliquant actuellement l'encadrement des loyers (1: oui, 0: non)	https://www.service-public.fr/particuliers/vosdroits/F3314		
encadr_locsais	Reg. meublés	Régulation	-	+	+	Commune réglementant actuellement l'activité des plateformes de location saisonnière (1: oui, 0: no	LNPTU, 2024		
p_listing_2023	Part AirBnB	Attractivité touristique	+	+	+	Nombre de listings actifs en 2023 / Nombre de logements du parc privé en 2021 * 100 (%)	Airbnb, 2023; RP 2021, Insee		
TXFONC_TOURIS_2023	Ta fonction tourist.	Attractivité touristique	+	+	+	Taux de fonction touristique en 2023 (nb lits touristiques / 1000 hab)	Capacité d'accueil touristique 2023, Insee-Partenaires territoriaux; RP Insee 2021		
entree_temp_hiv	Part tourisme HIVER	Attractivité touristique	+	+	+	Nombre de réexpéditions temporaires de courrier entrantes pour 100 logements HIVER (1er décembr	Réexpéditions de courrier la Poste; RP 2021 Insee		
entree_temp_ete	Part tourisme ETE	Attractivité touristique	+	+	+	Nombre de réexpéditions temporaires de courrier entrantes pour 100 logements ETE (1er juin 2023	Réexpéditions de courrier la Poste; RP 2021 Insee		

II Identifier category

There is also another category, made up of character type variables and which are therefore not actually used in calculations but which are still used to provide the information needed for groupings, maps, analyses etc.
This is the category called 'Identifiant', and includes the following variables:

Variable	Description
CODGEO	INSEE code of the municipality
LIBGEO	Name of the municipality
DEP	INSEE department code
REG	INSEE region code
EPCI	EPCI INSEE code
ZE2020	INSEE code of the employment area
BV2022	INSEE code of the catchment area
code_maille_habitat	'Maille habitat' code