A methodology for local housing price index in France

Carmelo Micciche[1,2], Michel Baroni[1], Pierre Vidal[1,2]

[1] CY Cergy Paris Université, Laboratoire THEMA, Cergy, France
[2] Meilleurs Agents, Paris, France

Abstract

Real estate accounts for 61% of France's national net wealth. Housing is the largest item of expenditure of French households. Indices that track real estate prices evolution are thus crucial instruments for decision makers of all kinds: households, investors, the scientific community, local governments, etc. Yet, the available public statistics fail to cope with the heterogeneity of the housing prices dynamics across the country.

In France, Notaire-Insee indices are considered as the reference, especially because their methodology and indices are open source. Quarterly, the institute produces indices for apartments and houses in big agglomerates. With 9 indices for house prices in France, the division proposed by this methodology hides a lot of disparities. For instance, the “Province” house index includes more than 25000 cities as diverse as Toulouse (450k inhabitants) and Malroy (350 inhabitants), which represents 36% of the French housing stock. This indicator does not make it possible to highlight the differences in dynamics between cities geographically distinct and drived by different fundamentals due to different economic conditions.

This work aims at producing a library of open data real estate price indices that track price evolution at fine geographical scale. To do so we develop a methodology for real estate price index computation, and then apply it on geographical clusters close to local markets. We want to be part of an open-source approach. Indeed, the methodology will be published, and all the indices will be made available for free to all.

The proposed method is applied on the fiscal database of real estate transactions DV3F, containing all the transactions in France (except Alsace, Moselle and Mayotte) between 2010 and 2020.

Our approach is based on classic hedonic price index methods. Each aspect and hypothesis of the hedonic method have been justified to produce precise indices.

Producing indices close to local markets requires working in a low data environment, and increases the probability of encountering outliers. Hedonic methods being very sensitive to outliers, we tackle this issue by testing the impact of different dynamics filters methods. To reduce the heteroscedasticity and improve the precision of the model, different forms and combinations of the regression have been tested.

This method is applied to 2 divisions of France: one for apartments, another for houses. In order to produce indices close to local markets, a clusterization of cities of France is computed as finely as possible and based on socio economic and local housing stock criteria. To preserve the quality of indexes, all clusters respect constraints of minimum transaction volumes. This division is based on a clusterization of urban areas thanks to Ascending Hierarchical Classification and Kohonen algorithms.

This clusterization resulted in the computation of 350 apartments and 400 houses indices. The application of our approach on these geographical clusters reveals a great diversity of house price dynamics. For instance, the “Province” index produced by Notaire-Insee is divided into 220 clusters, with variations between 2015 and 2020 of 2% and 29% respectively for the first and ninth decile of these indices.

By highlighting the plurality of real estate price dynamics in France and urban centers, our approach emphasizes the need for indices to be computed on a local scale to be useful.

Introduction

Real estate accounts for 61% of France’s national wealth and represents the largest item of expenditure for households. Thereby, price evolutions can affect all stakeholders from households to investors or policymakers. Real estate price indexes are a statistical representation of these evolutions. And this paper proposes a methodology for price index computation at a fine scale covering all of France. This method allows us to highlight the diversity of housing price evolutions of submarkets while maintaining a statistical significance of indexes computed.

A real estate index is a statistic that should be computed locally. Indeed, these price evolutions are driven by local conditions like the unemployment rate, local infrastructure developments, or city attractiveness. These fundamentals imply a large diversity in price evolutions and require a local description of price changes. However, although price evolutions are well described and documented in big cities like Paris, on large scales like big aggregates, or in the whole of France, the local evolution of prices is often poorly described.

In France, many indexes are produced and displayed by private actors like real estate agencies or real estate websites. But only Notaire-INSEE indexes are considered as the reference. The French national institute, using the complete BIEN and Perval databases, displays 30 indexes for apartments and houses in the whole territory. Although a city like Paris is well described by a single index, other indexes cover a large part of the country. For instance, the institute produces only 9 indexes for house prices in France, which cannot represent the disparities in terms of price evolutions. The « Province » house index describes more than 25.000 cities as diverse as Toulouse (450K inhabitants) and Malroy (350 inhabitants), and represent more than 36% of total housing stock. Thereby, although these indexes can well describe all these big aggregates, it hides the differences in dynamics between cities geographically distinct and driven by different fundamentals due to different economic conditions.

This paper aims to build a complete database of price indexes in France at a local scale between 2010 and 2020. The main objective is to describe as precisely as possible each real estate market. Moreover, these indexes allow us to answer to the relevance of the interpretability of local price evolutions, and can enlighten us on the right scale from which real estate markets need to be analyzed by highlighting the diversity of dynamics while keeping a large-scale logic. A complete methodology is proposed, from the definition of every real estate market to the construction of indexes on each. To ensure the spatial proximity of entities of each market, they will result from the construction of a regionalization of France based on socio-demographic criteria computed on open data statistics. To ensure the precision of indexes, all clusters resulting from this regionalization need to be as small as possible, but, also, large enough to compute good-quality indexes. Price index computation is based on a hedonic method, with an outlier detection method in pre-treatment to ensure the robustness of indexes and to make it possible to compute indexes at a fine scale.

The work presented in this paper is based on the DV3F database. This database, published by the CEREMA, comes from the “DVF” database from the Direction Générale des Finances Publiques (DGFIP) and Land Values (fichiers fonciers). It contains almost all property transfers that occurred in France (except Alsace Moselle and Mayotte) between 2010 and 2020. This database is rich in details concerning each property, which makes it possible to correctly filter relevant mutations and control quality effects. Furthermore, the geolocalisation at the address level allows controlling for geographical effects and computing indexes at a fine scale.

The method proposed here is based on a classic hedonic regression method. Each aspect of the regression has been studied. A classic forward selection and an outlier detection methods are applied to improve the model performances against outliers and so against incoherent variations and excessive volatility.

This price index method is applied in the regionalization of France. This regionalization is built on the geographical mesh of the city and arrondissements of the 3 biggest cities. To compute coherent indexes close to local markets, this clusterization needs to respect 3 criteria: the spatial continuity within each cluster, the minimum size of clusters, and building clusters as small as possible. To ensure these criteria, the algorithm used comes from the Max-P region problem.

The main results are the construction of 876 indexes in all of France between 2010 and 2020. These indexes offer a wide variety of price variations during this period and make it possible to analyze in detail the real estate market price evolutions geographically and temporally. Indeed, each city of France is described by an index close to its own market, and global tendencies seem to be coherent. Price evolution maps confirm the relevancy of these indexes by showing several price trends inside subregions or regions. The comparison of price index aggregations shows that the global tendency seems to be very similar to French reference indexes.

The first section of this paper focuses on the description of the DV3F dataset and highlights its main necessary treatments. Then, in the second section, the price index methodology is developed. The hedonic approach and the pretreatment outlier detection method are presented. The third section develops the regionalization of the territory and all the results. Finally, the index production results are detailed in the fourth section. To justify the quality of these indexes, a comparison to French reference real estate price indexes is highlighted. Then, the contribution in terms of the diversity of price variations over the territory is underlined.

Bibliographie

Index bibliography

The specificities of the real estate market make it complicated to compute precise indexes. Indeed, mean or median indexes could be noisy due to the heterogeneity of dwellings and the illiquidity of the market. The two major real estate index methodologies usually used to overcome these difficulties are the hedonic price index by Bailey, M. R. Muth, and H. Nourse [2] which uses ordinary least squares to control any effects which can influence the price and compute an index thanks to time dummies but requires an exhaustive description of each dwelling and the repeat sales method by Case, K. E., & Shiller, R. J.[7] which consists of reducing the dataset to multiple sales and deducing the price evolution from that.

In France, the methodology produced by INSEE institute is based on a hedonic regression applied to a clusterization of France based on socio-demographic and housing stock data [18]. Then, all indexes are aggregated and only 30 indexes are displayed, which means that each index will describe a large part of the territory. Although these indexes offer a good description of global evolutions of prices, it makes it complicated to understand the price changes at a local scale. In Paris, a repeat sales index was tested by M. Baroni & F. Barthélémy & M. Mokrane [3] and compared to classic hedonic methods. This paper highlights that this index is robust but significantly different from the one produced by Notaire-INSEE during the periods of market reversal. In the same way, the paper written by Thion, B., Riva, F., & Chameeva, T. [23] consists of building an improved repeat sales methodology in the Bordeaux region which avoids the revision of previous variations.

At the international scale, the repeat sale method was initially applied in the US to compute the monthly real estate index [7]. These methods have been tested in several other countries like the repeat sales method in the Netherlands [16] or the hedonic method to build an Irish index [10]. For each case, the choice of the method depends on the type of data and the usage.

Choosing among these methods requires being aware of the potential bias or hypothesis depending on the dataset. Even though the repeat sales method doesn't require a lot of characteristics, the sample reduction to repeat sales can cause bias. Gatzlaff, D. H., & Haurin, D. R. [13] found a bias in a data sample of transactions in Dade County, and as well as M. Baroni & F. Barthélémy & M. Mokrane [3] in Paris, this paper underlines that this bias is highly correlated with changes in economic conditions. In 5 metropolitan areas of the US, Clapp, J. M., Giaccotto, C., & Tirtiroglu, D. (1991) [8] showed that even if this bias causes differences in short-term price trends, the long-term tendency stays very similar. To go further, potential bias can come by using only sold houses according to the paper written by Gatzlaff, D. H., & Haurin, D. R. (1998) [14] which can underestimate the price changes.

Apart from the traditional methods, other price indexes were built to overcome these potential biases or based on more complex models. For instance, the paper presented by Case, B., & Quigley, J. M. (1991) [6] developed a methodology using all transactions and both information about the property and potential resales. Other methods based on hybrid or improved repeat sales models to overcome the biases were proposed by Quigley, J. M. (1995) [20] or Goetzmann, W. N. (1992) [15]. An improved hedonic model was proposed by Meese, R., & Wallace, N. (1991) [19] which is more robust to the influence of unusual observations. More complex methods use neural networks [17] or gradient boosting techniques [4] which can offer relevant alternatives to classic methods.

Regionalization bibliography

Although there exist many clusterization techniques, relatively few are suitable for geographical regionalization and can integrate spatial correlations or impose spatial continuity. SKATER algorithm, from Assunção, R. M., Neves, M. C., Câmara, G., & da Costa Freitas, C. [1] paper is a common and performant regionalization algorithm based on a partition of a Minimum Spanning Tree. In France, this algorithm was tested in a paper written by Collin, Roussez [9] to build a regionalization of all of France to have a better understanding of local markets. Based on socio-demographic data, this clusterization made it possible to compute 800 homogeneous regions. But, even if this algorithm provides homogeneous and contiguous regions, it requires imposing the number of clusters in advance and makes it impossible to control their sizes.

The Max-P Regions Problem, a method which comes from the paper written by Duque, J. C., Anselin, L., & Rey, S. J. [12] responds perfectly to this problem of a minimum size of clusters. Indeed, this algorithm provides a heuristic solution to the problem of computing a maximum of homogeneous regions while respecting a threshold criterion. This method was chosen in the paper from Breuillé, Grivault, and Le Gallo [5] to compute a regionalization of all of France to evaluate the average rent of each market. They highlight the need to respect a minimum size for each cluster to gather enough data to compute all rent indicators.

DV3F Data

Overview & Definition

Property sales data is extracted from the “DV3F” database published by the CEREMA. This database is composed of the open data “DVF” database (“Demande de Valeur Foncière”, Request for Land Value) published by the Direction Générale des Finances Publiques (DGFIP) enriched structured by Land Values (fichiers fonciers).

This database contains 12,700,787 property transfers ("mutations") recorded by the notaries during the period 2010-2020. These mutations cover all departments of France except Alsace, Moselle, and Mayotte.

Classic sale filtering

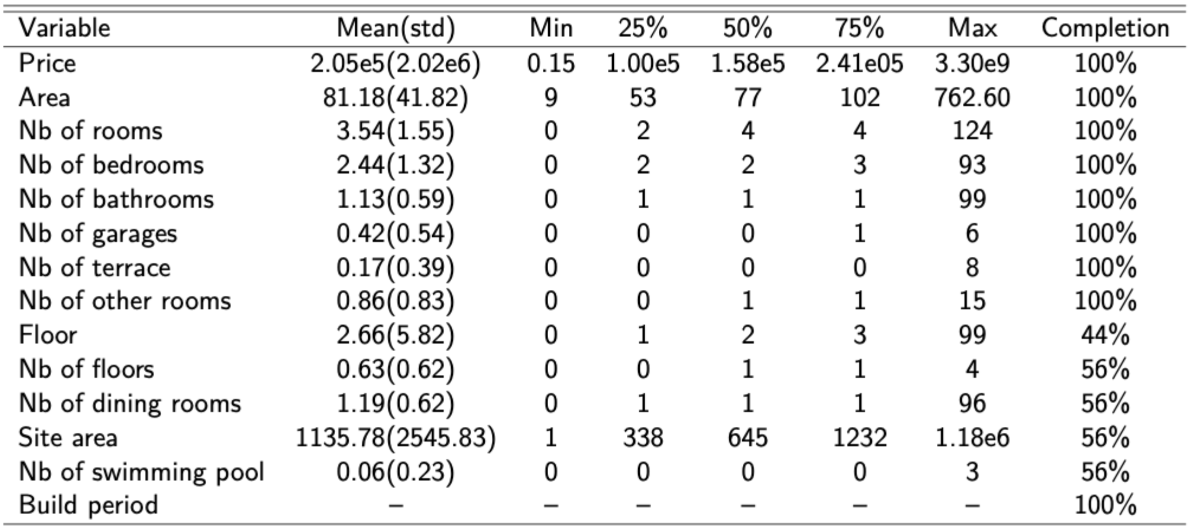
All kinds of property transfers are registered in this database, including housing unit transactions as well as land transactions, expropriations, or exchanges. For our purpose, we need to keep only transfers that come from a classic sale. A classic sale is considered as a sale where the price was fixed by the market.
DV3F database contains a lot of information about the transaction, the context, and the stakeholders. These characteristics allow us to remove transactions that are outside of the scope of our study.
Observations are removed based on these criteria :

- New constructions
- Multiple property transfers
- Parcel that has undergone a significant change (information extracted from the Fichiers Fonciers)
- Type of mutations: Exchange, Auction, Social Housing, Expropriation
- Rare or exceptional dwelling
- Abnormal conditions of sale (2 sales of the same property on the same day, …)
- Property located in several parcels
- Transfers between public entities

Our final dataset consists of more than 2M apartment sales in 8,465 cities and 2,5M house sales located in 33,869 cities.

Descriptive statistics

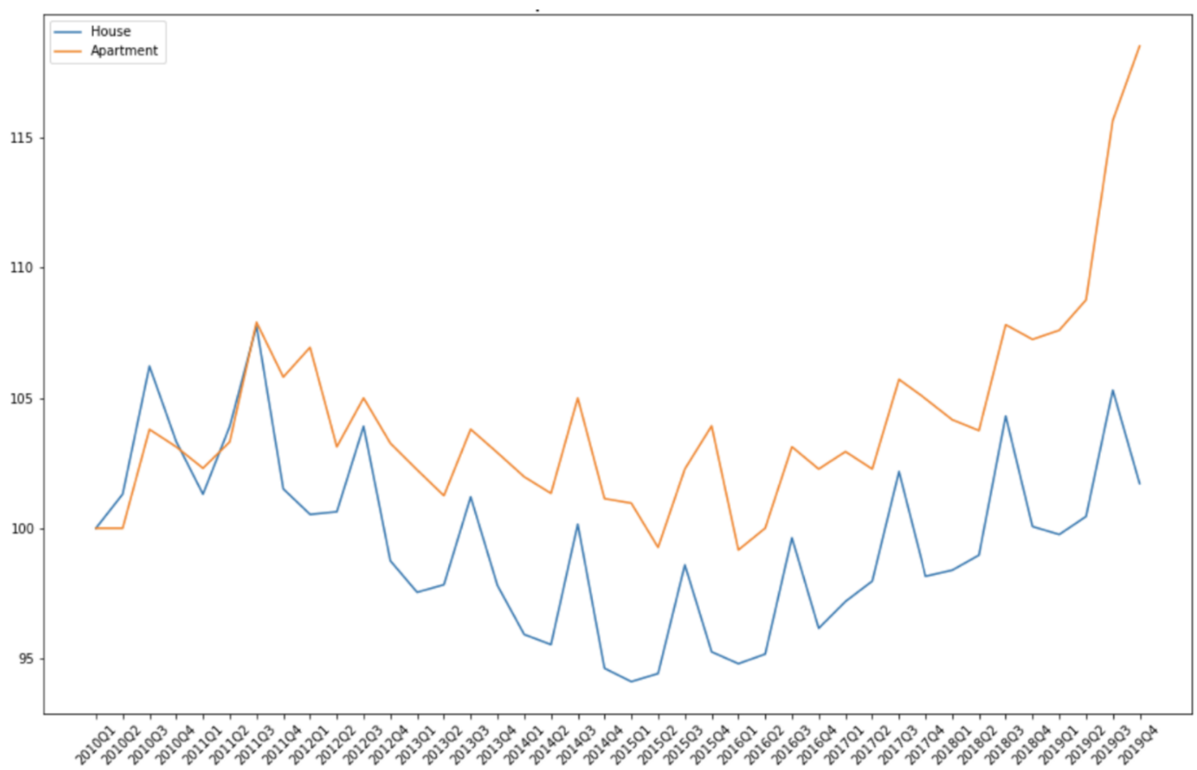
Table 1 - Descriptive table of DV3F dataset



Variable	Mean(std)	Min	25%	50%	75%	Max	Completion
Price	2.05e5(2.02e6)	0.15	1.00e5	1.58e5	2.41e05	3.30e9	100%
Area	81.18(41.82)	9	53	77	102	762.60	100%
Nb of rooms	3.54(1.55)	0	2	4	4	124	100%
Nb of bedrooms	2.44(1.32)	0	2	2	3	93	100%
Nb of bathrooms	1.13(0.59)	0	1	1	1	99	100%
Nb of garages	0.42(0.54)	0	0	0	1	6	100%
Nb of terrace	0.17(0.39)	0	0	0	0	8	100%
Nb of other rooms	0.86(0.83)	0	0	1	1	15	100%
Floor	2.66(5.82)	0	1	2	3	99	44%
Nb of floors	0.63(0.62)	0	0	1	1	4	56%
Nb of dining rooms	1.19(0.62)	0	1	1	1	96	56%
Site area	1135.78(2545.83)	1	338	645	1232	1.18e6	56%
Nb of swimming pool	0.06(0.23)	0	0	0	0	3	56%
Build period	–	–	–	–	–	–	100%

Each transaction is described by the features displayed in Table 1, and the geolocalisation is done at the parcel level. Every feature is completed at 100%. Floor characteristic is completed only for apartments and Nb floors, Nb of dining rooms, Site area, and Nb of swimming pool are completed only for houses.
Maximum and minimum prices seem not to be consistent. This table reveals the importance of the outlier detection part of the index methodology detailed in the *Outlier detection* part.

Table 2 - Median price index in France



Median index

Indexes displayed in Table 2 are simple median indexes built with all apartment or house transactions in France. The volume of transactions allows us to build good-quality indexes.
We note the presence of a strong seasonality, with a high point each third semester, which could correspond to the beginning of the school year.
Both indexes highlight three periods of evolution:

- 2010Q1 - 2011Q3: A period of price growth with a catch-up effect following the fall in real estate prices after the 2007 economic crisis
- 2011Q3 - 2016Q1: A period of price decrease following by the European debt crisis
- 2016Q1 - 2019Q4: A period of price growth following to the European quantitative easing policy held by the ECB

The results will be analyzed with regard to these three periods.

Price index methodology

Overview

The methodology is based on a hedonic price index method, which consists of a regression of the logarithm of price per square meter by the characteristics of the dwelling. This regression is applied in a dataset where outliers are removed thanks to a Cook outlier detection method. Then, the characteristics are selected with a classic forward method adapted to our special case.

Hedonic method

The main methods to build a real estate price index are the Hedonic price index and the repeat sales method.
Housing market is not liquid and the turnover is slow, around 3%, that's why it requires working with large periods. But ten years of data is not a sufficient time range and it could lead to potential bias problems by selecting a specific type of dwellings and building an index for a submarket, like first-time buyers' property type.
For this paper, indexes are computed using a hedonic method. This method is based on the valuation of each element of a dwelling. A price is fixed for all the elements, the geolocalisation (at the city or borough level), and the quarter dummies. The regression follows this equation:

$$log(pm^2) = Cste + \sum_i \alpha_i X_i + \sum_t \beta_t Q_t + \varepsilon$$

With : pm^2 the price per square meter
X_i the dummies for each element of the dwelling and α_i the corresponding coefficient
Q_t the dummies for quarter t and β_t the corresponding coefficient

The index is computed with the coefficients associated to the quarters:

$$i_t = exp(\beta_t) \times 100$$

With i_t the value of the index for quarter t

The final index is the time series $(i_t)_{t \in N}$ in base 100 with the first quarter as reference (i_0=100).

Outlier detection

As highlighted in the *Descriptive statistics* section, an outlier detection method is necessary to build a methodology robust to outliers in a context of low data volume. Indeed, linear regressions are very sensitive to this type of data. And, in order to compute indexes close to local markets, in restricted geographical areas, regressions are computed with low data volume and the presence of outliers strongly affects the quality of indexes.
Several outlier detection methods have been tested. Papers written by Rousseeuw, P. J., & Hubert, M. (2011) [21], and Wisnowski, J. W., Montgomery, D. C., & Simpson, J. R. (2001) [24] offer a comparative study of robust statistics for outlier detection. The advantages of these methods are the consideration of the features in the outlier detection procedure. Indeed, a large difference in prices is not always considered an outlier and can be explained by the dwelling characteristics. For instance, 2 apartments located in different neighborhoods may have large differences in prices which can be explained by neighborhood qualities.[1]
Cook's outlier detection method is based on a distance measure that reflects the influence of each observation on the regression. Observations with a large distance are considered outliers and removed from the dataset. The distance is computed as follows:

$$D_i = \frac{(b_i - b)^T (X^T X)(b_i - b)}{ps^2}$$

With : D_i distance associated with the observation i

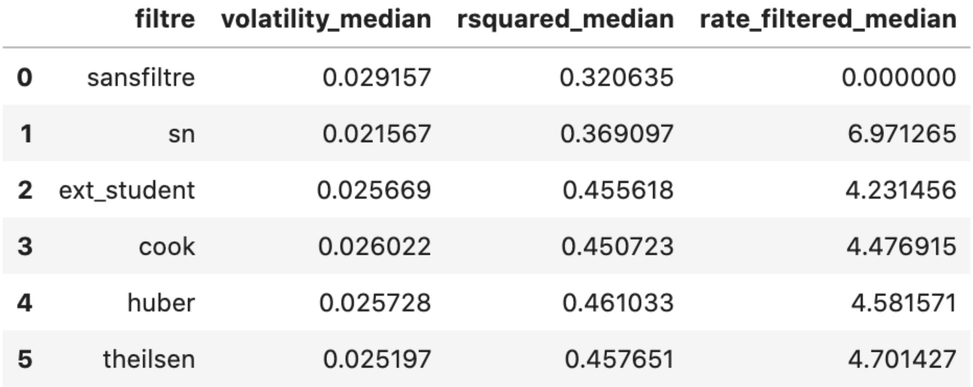
b vector of coefficients
b_i vector of coefficients computed without observation i
p the sum of elements of H, the Hat matrix
s the variance of the error

Outliers are observations that respect this condition:

$$D_i < \frac{4}{N_{obs}}$$

With N_{obs} the number of observations

Table 3 - Outlier detection methods results



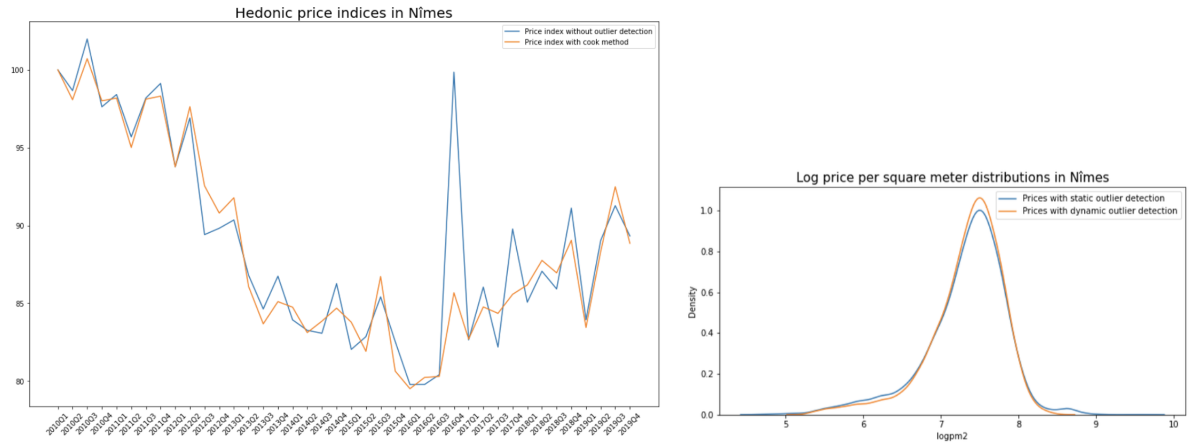
	filtre	volatility_median	rsquared_median	rate_filtered_median
0	sansfiltre	0.029157	0.320635	0.000000
1	sn	0.021567	0.369097	6.971265
2	ext_student	0.025669	0.455618	4.231456
3	cook	0.026022	0.450723	4.476915
4	huber	0.025728	0.461033	4.581571
5	theilsen	0.025197	0.457651	4.701427

All these outlier detection methods have been compared to regressions without any removal of outliers. They all improve the R^2 clearly and decrease the volatility.

[1] For our purpose, these methods, cited by these papers have been evaluated: Cook's outlier detection, a method based on an influence measure of an observation on coefficients ; Huber and Theilsen regressions, based on robust regressions ; Externally studentized residuals, which consist of a residuals analysis ; Robust scaling, a method based on the robust scaling of the target variable and the removal of extreme values.
Cook's outlier detection method has been chosen after an analysis of 35 linear regressions on apartment observations over 35 French cities. The overall quality of each index has been analyzed, and three statistics have been computed: the volatility, the r squared, and the rate of observations filtered. The volatility reflects the quality of the index, by shedding light on the indexes without strong inconsistent variations caused by outliers.

Indexes computed with Sn outlier detection method have the lowest volatility, but the worst R squared. This method is rejected because a better R^2 is preferred, indeed it reflects the quality of the regression.
All other methods obtain equivalent results, and the cook's method is chosen because of its performances on regressions and capacity of avoiding incoherent variations. Furthermore, this statistic is often used for outlier detection.

Table 4 - Outlier detection results in Nîmes



Indexes shown in Table 4 show that Cook's method improves the quality of index by smoothing and avoiding inconsistent increases like the one in the 4th quarter of 2016. A detailed analysis of this perturbation revealed a set of apartments sold during these periods at a very high price due to recent renovations. These observations are considered outliers since no features are available to control these quality effects. In the distribution, these apartments are located around 8.6 logarithm of price per square meter. The algorithm removed only 5% of the data and managed to remove the correct outliers. Furthermore, the R^2 went from 0.54 to 0.65, which means an improvement in the quality of the regression.

Variable selection method

DV3F dataset contains a lot of descriptive variables, and a variable selection method allows to keep only relevant variables. The geographical control is done at the city or borough level. If a city is alone in its cluster, the geographical control is done at the borough level.
This variable selection method is inspired by a classic forward selection, with R^2 adjusted as a target variable. The only difference is that the initialization of R^2 adjusted is computed with temporal dummies because these variables are necessary to build an index. Then, iteratively, the variable that most improves the R^2 adjusted is added until the target variable no longer progresses. [2]

Regionalization

Overview

The methodology described above is applied to regionalization of France. To compute indexes close to local markets, it is necessary to divide the country into small and consistent clusters. Indeed, it is impossible to compute an index for each individual city because of the lack of data, and the department scale is too large for many departments and does not describe well the diversity of price changes. The three main criteria of this clustering are: Geographical continuity; similarity in terms of socio-demographic criteria; And building clusters as small as possible, but large enough to compute a regression with significant coefficients.

Geographical proximity is important for this purpose. Indeed, it seems not coherent to compute indexes over cities geographically very distant, and impose geographical continuity to ensure spatial proximity. This type of

[2] A hypothesis of the hedonic method is the constancy of coefficients of variables during all the periods of interest. This hypothesis can be questioned within a 10 year period. An empirical study shows that even if this hypothesis is violated for a significant part of coefficients, the final index is very similar to an index computed over a shorter period.

clusterization has been adopted for the papers “Élaboration d’une maille géographique pour l’habitat” [22] and the one written by Breuillé, Grivault, Le Gallo (2020) [5] where, for both papers, this division is realized for a housing market analysis. And for both papers, clusterization criteria are based on socio-demographic variables. Likewise, Notaire-INSEE indexes are built on a clusterization based on the same type of characteristics.
Building clusters as close as possible to the local markets requires them to be as small as possible. But, regressions need enough volume to compute significant coefficients, especially enough volume by quarter dummies. Notaire-INSEE methodology [18] imposes a minimum of 110 observations per quarter. For this project, this threshold is fixed at 100.

Algorithm

These criteria are exactly those of the Max-P Regions Problem [12]. Thus, the algorithm of regionalization is the Max-P algorithm, with a threshold of 100 observations per quarter and the regionalization of all cities of France (except Alsace-Moselle).
This algorithm requires a distance between entities. The distance used for this paper is the Kohonen distance. Kohonen is a Self-Organizing-Map algorithm, which was already used for geographical clusterization problems in the paper written by Cottrell, M., Olteanu, M., Randon-Furling, J., & Hazan, A. (2017, June) [11]. The advantages of this method are that it maintains the topology between input and output spaces, and manages the collinearity between variables. This clusterization algorithm provides a distance that is used for Max-P regionalization.

There are three specific treatments for these geographical entities: Little islands; Corse; and the arrondissements of the three biggest cities Paris, Marseille, and Lyon.
Max-P is a regionalization algorithm, thus it cannot treat geographical entities not contiguous like islands. Furthermore, islands do not reach the threshold of the volume of observations. That is why Islands are attached to the nearest cluster.
Corse is an island, but it reaches the threshold of the amount of data. The regionalization algorithm is applied only for this region. The algorithm is thus applied two times: One for mainland France and the other for the region of Corse.
Arrondissements of the three biggest cities of France are large enough to be treated like cities. So the arrondissement of Paris, Marseille, and Lyon are integrated into the regionalization algorithm like every other city.

Data

Kohonen algorithm requires characteristics about each geographical area. In line with the papers computing clusterization of France for a housing purpose (*Les indicateurs de loyers dans le parc locatif privé, Une nouvelle grille de lecture des territoires pour le logement, la maille habitat, and Hedonic housing price indexes: the French experience*) the characteristics used for this paper are from 3 different types: Demographic, Economic, and about Housing stock. They are all in open data, produced and hosted by INSEE in 2014.
Cities with missing data are completed with the median value of neighboring cities.

Demographic statistics

- Number of persons per household
- Density of population
- Percentage of the population between 25 and 54

Economic statistics

- Median income
- Unemployment rate

Housing stock statistics

- Percentage of second home
- Percentage of vacant housing
- Percentage of owners
- Percentage of social housing
- Percentage of recent housing (>1990)

- Percentage of recent housing (<1945)
- Housing turnover rate
- Percentage of apartments

Results

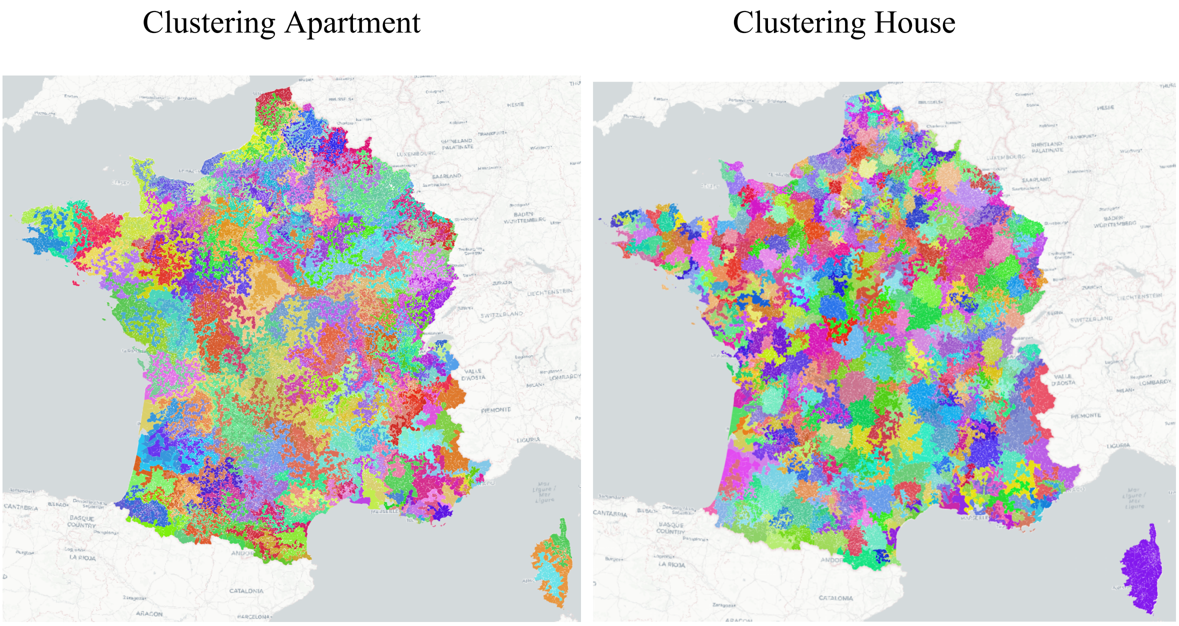
Volume of observations for each city is computed during the period 2010-2020, and the threshold is fixed at 2,400. A volume is computed for apartments and houses, and regionalization is done for each type.

Algorithm parameters:

- Max-P
  - Max iteration constr. : 99
  - Max iterations sa : 1
  - Policy: Single
- Kohonen
  - Grid size: 30

Results are summarized in Table 6. 876 clusters are computed for apartment and house, and our threshold criteria is respected with a minimum average observation volume of 112 for both types. The median number of cities per cluster is higher for apartments than for houses because in France many cities have no, or very few, apartments.
The city of Paris, for apartment regionalization, is divided into 15 clusters. More or less every arrondissement is alone in a cluster, thus they can be considered like other cities.

Table 5 - Clusterings maps



Clustering Apartment

Clustering House

Table 6 - Clusterings main statistics

		Apartment	House
Nb of clusters		298	578
Median nb of cities per cluster		81	49
Nb of sales per cluster per quarter	Min	112	112
	Median	144	114
	Max	1584	142

Results

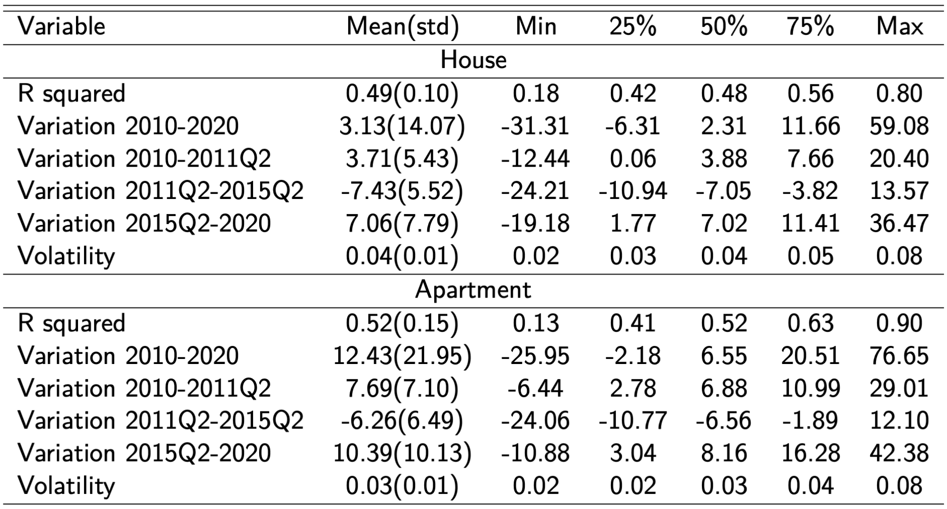
Method

Price index methodologies are then applied to all apartment and house clusters. 876 indexes are built, and statistics about their precision are stored: R squared, volatility, variations throughout the period, and variations over the periods of interest.
Indexes are computed with a single regression during the period 2010-2020. As it is shown in *Descriptive statistics* section, the analysis is done during the 3 periods of interest: 2010-2011Q2 ; 2011Q2-2015Q2 ; 2015Q2-2020.

The variation of each index is computed during the entire period and the three periods of interest for apartments and houses.
Then, a hybrid variation is computed for each city in France. The variation of the cluster for apartment and house is associated with the city, and the hybrid variation is the mean of these variations weighted by the housing stock.

In order to compare each period, this variation is then divided by the number of quarters of this period.

Table 7 - Distributions of house indices characteristics



Variable	Mean(std)	Min	25%	50%	75%	Max
			House			
R squared	0.49(0.10)	0.18	0.42	0.48	0.56	0.80
Variation 2010-2020	3.13(14.07)	-31.31	-6.31	2.31	11.66	59.08
Variation 2010-2011Q2	3.71(5.43)	-12.44	0.06	3.88	7.66	20.40
Variation 2011Q2-2015Q2	-7.43(5.52)	-24.21	-10.94	-7.05	-3.82	13.57
Variation 2015Q2-2020	7.06(7.79)	-19.18	1.77	7.02	11.41	36.47
Volatility	0.04(0.01)	0.02	0.03	0.04	0.05	0.08
			Apartment			
R squared	0.52(0.15)	0.13	0.41	0.52	0.63	0.90
Variation 2010-2020	12.43(21.95)	-25.95	-2.18	6.55	20.51	76.65
Variation 2010-2011Q2	7.69(7.10)	-6.44	2.78	6.88	10.99	29.01
Variation 2011Q2-2015Q2	-6.26(6.49)	-24.06	-10.77	-6.56	-1.89	12.10
Variation 2015Q2-2020	10.39(10.13)	-10.88	3.04	8.16	16.28	42.38
Volatility	0.03(0.01)	0.02	0.02	0.03	0.04	0.08

Results

Table 7 show the distributions of R squared, volatility, and variations of each period of interest. Results about R squared are equivalent between house and apartment, regressions explain about 50% of the variance on average. Variations over each period of interest are consistent with the results of Table 2, indeed, on average, for apartments and houses, the variation over 2010-2011Q2 is positive, then negative during 2011Q2-2015Q2, and again positive during the last period 2015Q2-2020. These results are also consistent with Notaire-INSEE trends. According to a A note on the situation from INSEE, the global increase after 2010 is explained by the decrease in interest rates and the fact that the French real estate market was not affected that much by the 2008 crisis in comparison with the United States or Spain.

Table 8 - Yearly hybrid variation between 2010 and 2020

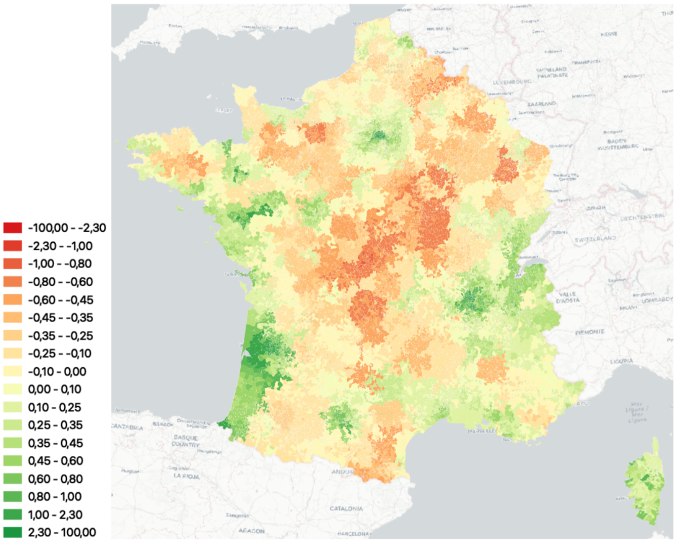
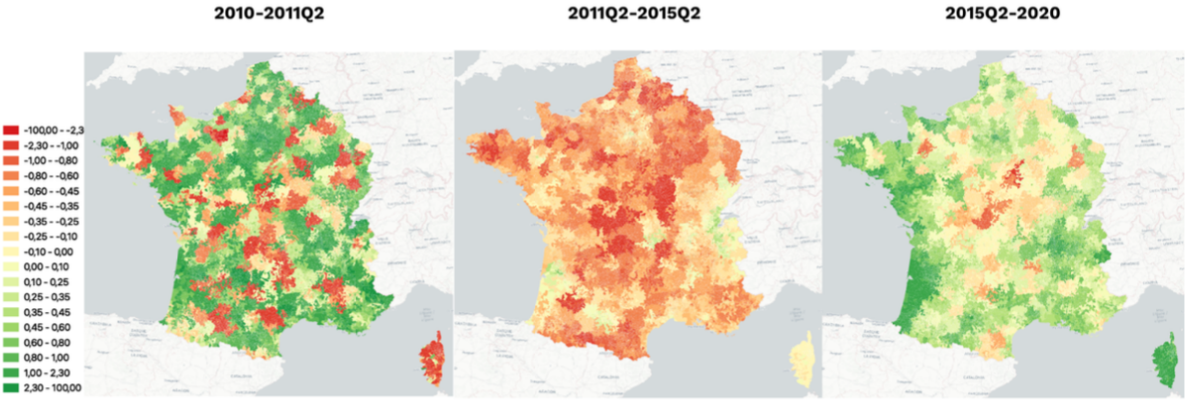


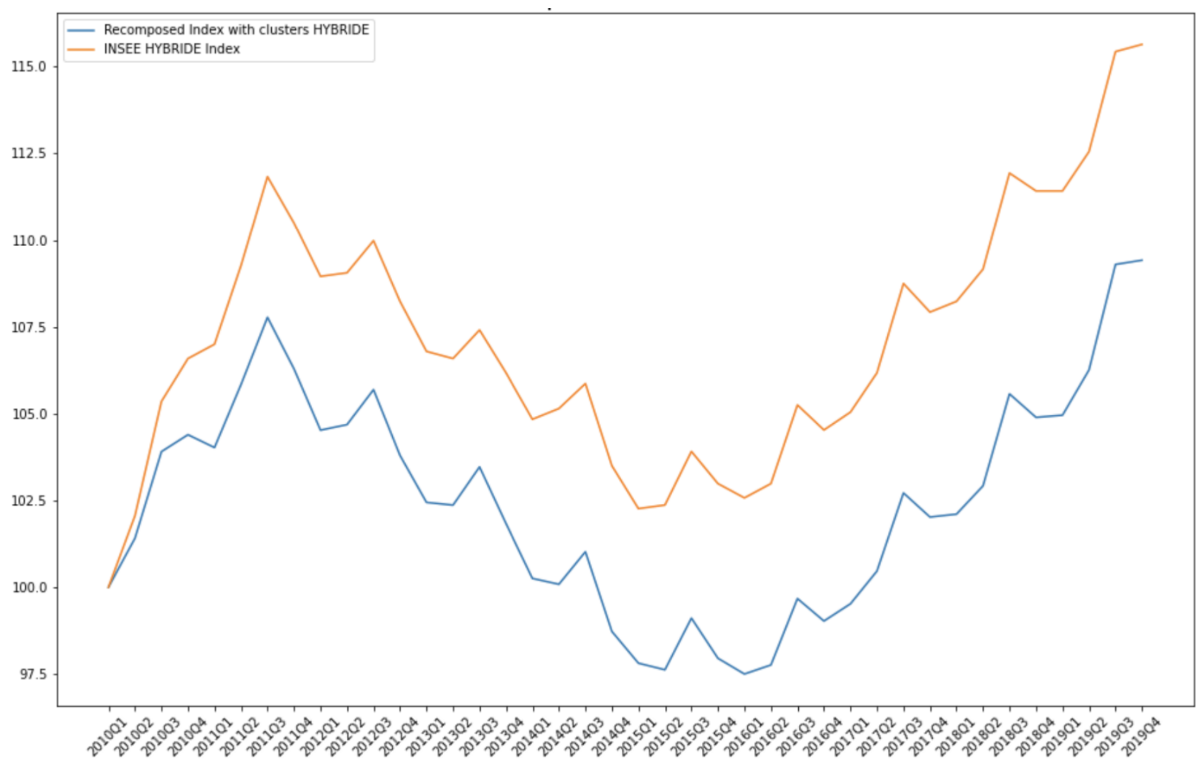
Table 9 - Hybrid variation for each period of variation



Hybrid variation map shown in Table 8 reveals the diversity of situations for the period 2010-2020. First of all, a long band of fairly strong price decreases is present in the center of the territory, along the well-known sparsely populated diagonal.

Urban areas of the biggest cities of France have a positive and strong variation. Almost every green spot on the map is centered on a city like Paris, Lyon, Bordeaux, Toulouse, or Rennes. However, the variation map over the second period shown in table 9 nuances these findings. Indeed, 2011Q2-2015Q2 was a period of global price decline in France, and almost all territory was affected except some urban areas like cities close to Switzerland and Bordeaux urban area.

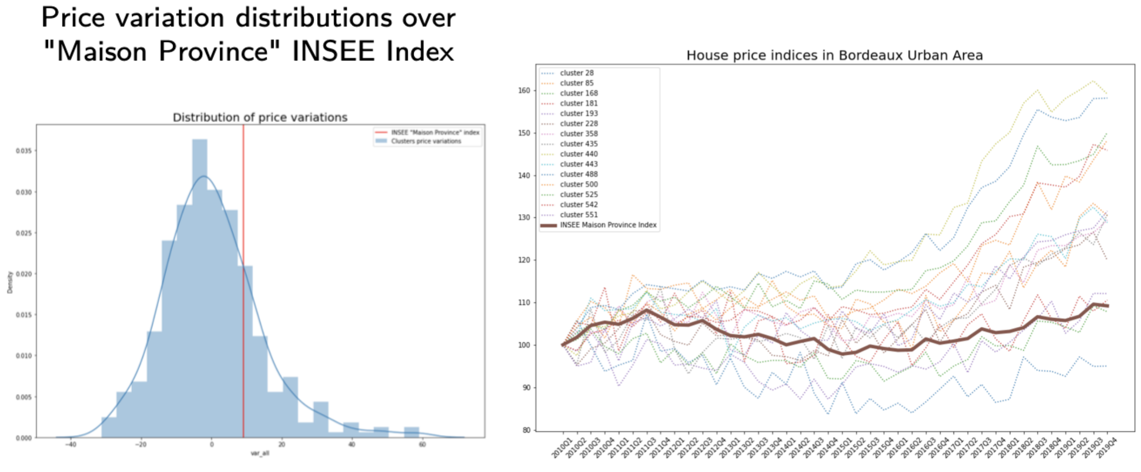
Table 10 - Comparison with Notaire-INSEE index



One way of validating cluster indexes is to compare global indexes to those produced by INSEE-Notaires. INSEE-Notaires produce a France apartment and house index. In this project, an equivalent index was built as follows: First, an apartment and house France index is built with the weighted average of all cluster indexes by the housing stock of each type. Then, the France hybrid index is the result of the weighted average of these two indexes by the housing stock of each type. The main difference between Notaire-INSEE and cluster France hybrid index is the absence of Alsace Moselle in cluster indexes, but it represents only 3 departments among the 96 of metropolitan France.

Indexes in Table 10 have the same trend over the period. The only difference is during the period 2010Q1-2011Q2, where the Notaire-INSEE index is much more dynamic, with a difference of 5 points at the end. But this difference remains constant until the end. This difference could be due to the methodologies, where Notaire-INSEE indexes follow a stock of dwellings, which could be biased during a period of strong increase.

Table 11 - Clusters indexes in comparison with “Maison Province” INSEE Index



With a low number of indexes, and some that describe a large number of cities in France, the price changes in France are not well described at a local level with Notaire-INSEE indexes. For instance, the “Maison Province” index describes the evolution of house prices for more than 25,000 cities. The clusterization allows to build a large number of indexes while respecting the volume of data criteria. Table 11 displays the distribution of variations throughout the period for all indexes built inside the “Maison Province” Notaire-INSEE cluster. The red vertical bar represents the variation of the single Notaire-INSEE index. Cluster indexes reveal a great diversity of price variations during this period, while these evolutions are only described by a single variation. This example is also illustrated for the Bordeaux urban area in Table 11, all house price evolutions are described with the single Notaire-INSEE index, while this area is divided into 15 clusters with this model. Furthermore, as it is illustrated in the variation map, this graph shows a decrease in price variation according to distance to major urban centers. Indeed, clusters 440 and 488 which are the clusters of Bordeaux and surrounding towns have the strongest variations.

Conclusion

The methodology developed here is the first one built based on the DV3F dataset in all of France. It requires selecting a subset of classic transactions, with as little bias as possible. Based on a hedonic method, this method is robust to small datasets thanks to an outlier detection method. Indeed, this can help to filter observations with specific attributes unobservable with the characteristics of this dataset.

This method is applied to a clustering of France based on socio-demographic data at the city or arrondissement level. The clusterization method ensures a minimum volume to compute a regression on each cluster and forces the spatial proximity by the spatial contiguity. Finally, 876 indexes are built for houses and apartments in France.

Hybrid variation maps show a lot of nuances in the overall territory. During the period 2010-2020, the urban areas of the biggest cities of France, a big part of the west coast, and the cities near Switzerland have seen their prices go up. Contrariwise, prices in rural areas decrease during this period. However, this finding is not constant during all the periods. Indeed, 2011Q2-2015Q2 was a period of price decline in all of France, except in the Bordeaux urban area and in cities near the Swiss border. Except in rural areas, 2015-2020 was a period of price increase, in all cities of France, but especially in the biggest cities.

This analysis could not be possible with Notaire-INSEE indexes, because of the size of clusters. Indeed, for the majority of Notaire-INSEE indexes, cluster indexes offer a great variety of price evolutions. This methodology helps to have a better understanding of the price evolutions of the territory while obtaining similar results on aggregate indexes.

References

1. Assunção, R. M., Neves, M. C., Câmara, G., & da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. International Journal of Geographical Information Science, 20(7), 797-811.
2. Bailey, M. R. Muth, and H. Nourse. (1963). “A Regression Method for Real Estate Price Index Construction,” Journal of the American Statistical Association 58, 933-942.
3. Baroni, M., Barthélémy, F., & Mokrane, M. (2011). A repeat sales index robust to small datasets. Journal of Property Investment & Finance.
4. Barr, J. R., Ellis, E. A., Kassab, A., Redfearn, C. L., Srinivasan, N. N., & Voris, K. B. (2017). Home price index: a machine learning methodology. International Journal of Semantic Computing, 11(01), 111-133.
5. Breuillé, Grivault, Le Gallo (2020), “Les indicateurs de loyers dans le parc locatif privé”
6. Case, B., & Quigley, J. M. (1991). The dynamics of real estate prices. The review of economics and statistics, 50-58.
7. Case, K. E., & Shiller, R. J. (1987). Prices of single family homes since 1970: New indexes for four cities.
8. Clapp, J. M., Giaccotto, C., & Tirtiroglu, D. (1991). Housing price indices based on all transactions compared to repeat subsamples. Real Estate Economics, 19(3), 270-285.
9. Collin, Roussez (2019), “Une nouvelle grille de lecture des territoires pour le logement, la maille habitat”
10. Conniffe, D., & Duffy, D. (1999). Irish house price indices? methodological issues. Vol. XX, No. XX, Issue, Year.
11. Cottrell, M., Olteanu, M., Randon-Furling, J., & Hazan, A. (2017, June). Multidimensional urban segregation: an exploratory case study. In 2017 12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM) (pp. 1-7). IEEE.
12. Duque, J. C., Anselin, L., & Rey, S. J. (2012). The max-p-regions problem. Journal of Regional Science, 52(3), 397-419.
13. Gatzlaff, D. H., & Haurin, D. R. (1997). Sample selection bias and repeat-sales index estimates. The Journal of Real Estate Finance and Economics, 14(1), 33-50.
14. Gatzlaff, D. H., & Haurin, D. R. (1998). Sample selection and biases in local house value indices. Journal of Urban Economics, 43(2), 199-222.
15. Goetzmann, W. N. (1992). The accuracy of real estate indices: Repeat sale estimators. The Journal of Real Estate Finance and Economics, 5(1), 5-53.
16. Jansen, S. J. T., de Vries, P. A. U. L., Coolen, H. C. C. H., Lamain, C. J. M., & Boelhouwer, P. J. (2008). Developing a house price index for the Netherlands: A practical application of weighted repeat sales. The Journal of Real Estate Finance and Economics, 37(2), 163-186.
17. Kershaw, P., & Rossini, P. (1999). Using neural networks to estimate constant quality house price indices (Doctoral dissertation, INTERNATIONAL REAL ESTATE SOCIETY).
18. Laferrère, A. (2005). Hedonic housing price indexes: the French experience. Press & Communications CH-4002 Basel, Switzerland E-mail: publications@ bis. org Fax:+ 41 61 280 9100 and+ 41 61 280 8100, 271.
19. Meese, R., & Wallace, N. (1991). Nonparametric estimation of dynamic hedonic price models and the construction of residential housing price indices. Real Estate Economics, 19(3), 308-332.
20. Quigley, J. M. (1995). A simple hybrid model for estimating real estate price indexes. Journal of Housing Economics, 4(1), 1-12.
21. Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. Wiley interdisciplinary reviews: Data mining and knowledge discovery, 1(1), 73-79.
22. Solène COLIN, Vivien ROUSSEZ (2018), Élaboration d’une maille géographique pour l’habitat
23. Thion, B., Riva, F., & Chameeva, T. (2006). Repeat Sales and Urban Price Indices: a New Approach. Cahier de recherche, 7.
24. Wisnowski, J. W., Montgomery, D. C., & Simpson, J. R. (2001). A comparative analysis of multiple outlier detection procedures in the linear regression model. Computational statistics & data analysis, 36(3), 351-382.