ESPON Big Data for Territorial Analysis and Housing Dynamics

Wellbeing of European citizens regarding the affordability of housing.

Monitoring and tools

Draft Technical Guidance Document
Interim Report 2

This monitoring and tools activity is conducted within the framework of the ESPON 2020 Cooperation Programme.

The ESPON EGTC is the Single Beneficiary of the ESPON 2020 Cooperation Programme. The Single Operation within the programme is implemented by the ESPON EGTC and co-financed by the European Regional Development Fund, the EU Member States and the Partner States, Iceland, Liechtenstein, Norway and Switzerland.

This delivery does not necessarily reflect the opinion of the members of the ESPON 2020 Monitoring Committee.

Authors
Renaud Le Goix, UMS RIA – University Paris Diderot (France) – Project coordinator
Ronan Ysebaert, UMS RIA – University Paris Diderot (France)
Timothée Giraud, UMS RIA – CNRS (France)
Lambert Nicolas, UMS RIA – CNRS (France)
Marianne Guérois, UMS RIA – University Paris Diderot (France)
Malika Madelin, University Paris Diderot (France)
Thomas Louail, Géographie-cités (France)
José Ravier Ramasco, FISC-CSIC (Spain)
Pere Colet, FISC-CSIC (Spain)
Thierry Teurillat, Haute-Ecole Arc (Switzerland)
Alain Segessemann, Haute-Ecole Arc (Switzerland)
Szymon Marcinczak, University Lodz (Poland)
Bartosz Bartosiewicz, University Lodz (Poland)
Elisabete Silva, Cambridge University (UK)
Sølve Baerug, Norwegian University of Life Sciences (Norway)
Terje Holsen, Norwegian University of Life Sciences (Norway)
Guilhem Boulay, University of Avignon (France)
Marco Kellenberg, Office Fédéral du Développement Territorial (Switzerland)

Advisory Group
ESPON EGTC: Marjan Van Herwijnen (project expert), Caroline Clause (financial expert)

Information on ESPON and its projects can be found on www.espon.eu.

The web site provides the possibility to download and examine the most recent documents produced by finalised and ongoing ESPON projects.

© ESPON, 2019

Printing, reproduction or quotation is authorised provided the source is acknowledged and a copy is forwarded to the ESPON EGTC in Luxembourg.

Contact: info@espon.eu
Interim Report 2

ESPON Big Data for Territorial Analysis and Housing Dynamics

Wellbeing of European citizens regarding the affordability of housing.

Draft Technical Guidance Document
26/04/2018

Disclaimer:
This document is a interim report.

The information contained herein is subject to change and does not commit the ESPON EGTC and the countries participating in the ESPON 2020 Cooperation Programme.

The final version of the report will be published as soon as approved.
# Table of contents

List of Maps .................................................................................................................. ii
List of Figures .................................................................................................................. ii
List of Tables .................................................................................................................. ii
Abbreviations ................................................................................................................... ii

1 Introduction to the guidance document .................................................................... 3

2 Using institutional data to analyze the dynamics on property markets : data, methods, sample results ................................................................................................................. 6

2.1 Unconventional institutional data : data from Paris Chamber of the Notaries ........ 6

2.2 A spatial analysis of price dynamics : why a grid ? .............................................. 6

2.3 A narrative of the methodology used ..................................................................... 7

3 Harmonized indicators and methodology to analyse real-estate markets – first results ... 9

3.1 Comparative approach of households harmonized indicators .............................. 9

3.2 Average price for transactions .............................................................................. 10

3.3 Income data ........................................................................................................... 10

3.4 Price to income ratio ............................................................................................. 11

3.5 Debt-to-value ratio ................................................................................................ 13

4 Unconventional (“Big data”) vs institutional data sources ...................................... 16

4.1 From harvested data to harmonized indicators – overall process ....................... 16

4.2 Narrative of a working example of data collection ................................................ 19

5 Output : some examples of indicators to compare and analyze unequal access to housing markets and living conditions ................................................................. 20

5.2 Harmonized variables and geographical levels at which data can be prepared and delivered .................................................................................................................. 22

References ...................................................................................................................... 23
List of Maps

Map 4-1 – Paris Average price per square meters ........................................... 10
Map 4-2 – Paris Average Income ................................................................. 11
Map 4-3 – Price to income ratio, 1km grid ...................................................... 12
Map 4-4 – Price to income ratio, LAU2 ......................................................... 13
Map 4-5 – Debt to value ratio, 1km grid ......................................................... 14
Map 4-5 – Debt to value ratio, LAU2 ............................................................. 15

List of Figures

Figure 1.1 - Overview of the workflow ......................................................... 4
Figure 4.1 - Step 1 - Get links to all ads of a given real estate Website .............. 17
Figure 2.2 - Step 2 – Collect and sparse geospatial information ...................... 17
Figure 4.3 - Step 3 – Data cleaning to obtain an accurate dataset with geocoded transaction location ................................................................. 18
Figure 4.4. Summary of harvested data ......................................................... 20
Figure 4.4. Comparison of aggregated harvested and institutional data at the LAU2 level ................................................................. 21

List of Tables

No table of figures entries found.

Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESPON</td>
<td>European Territorial Observatory Network</td>
</tr>
<tr>
<td>ESPON EGTC</td>
<td>ESPON European Grouping of Territorial Cooperation</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>FUA</td>
<td>Functional Urban Area</td>
</tr>
<tr>
<td>IDS</td>
<td>Internet Data Sources</td>
</tr>
<tr>
<td>OECD</td>
<td>Organisation for Economic Co-operation and Development</td>
</tr>
<tr>
<td>TOR</td>
<td>Terms of reference</td>
</tr>
</tbody>
</table>
1 Introduction to the guidance document

This document demonstrates the methodological framework to map and analyze the spatial dynamics of unequal local affordability, to analyze the impact on neighborhoods. It prefigurates the final guidance document (delivery #3), and elaborates on a workable example applied to a case study in Paris regions, to demonstrate and discusses implementation methods on local disaggregated urban data. The methodology will be used throughout the ESPON Big Data for Territorial Analysis of Housing Dynamics“ 2018-19 project.

This document elaborates on the prefigurative case of Paris, as an ideal case study because of the availability of institutional data. We consider a variety of datasets: property-level data from the Paris Chamber of Notaries (1996-2012, a sample of 1 million rows), public data, compared to possible harvested big data sources (real estate websites; opendata resources and warehouses). Harmonized and standardized variables are proposed, down to the local level (1 km grid). Methods are applied to a subset of data available to test and demonstrate the reproducibility of the methodology.

The guidance document is composed as follows:

- The main body of the guidance document literaly describes the data selection, harvesting and analysis process, with examples from the Paris case study.

- The technical part of the guidance document is provided as an appendix, written a RMarkdown document, a standard programing language in spatial analysis, big data analysis, and statistic. It uses a workable subset of data in the western suburbs of Paris (Yvelines, 78). We deliberately chose this case study in order to demonstrate (1) the mix of harversted data already collected in this area, so as to to describe (2) the workflow on how to bridge heterogenous data sets from various sources (institutional, census and harversted data) to analyze unequalities and wellbeing at local scales. The current draft version of this document is available online (version 1.2: http://bit.ly/2UHIEIF)

The goals of this guidance document outline are :

- To describe the data collection process, both conventional and unconventional.

- To describe and document the methodology employed to harvest datasets, using APIs, and R packages s.a. ‘Cartography’, ‘SpatialPostion’, ‘rvest’, and ‘httr’. The R code is documented as an Appendix, so as to ensure reproducibility and transferability of the protocols. The dicussion includes an overview of data harvesting methodologies.

- To describe a set of harmonized variables, and also which data sources are used, what methods have been applied. Harmonized variables should be made comparable between European cities, and over time. Ratios and standardized indices, such as affordability ratio will
be considered as valuable alternatives to rough stock variables (s.a. price, surface), that are structurally contingent to each country, city and local market contexts.

- To analyze for each datasets and variables how it can be used to document the dynamics of markets, over time. Spatio-temporal information is sensitive to two types of sampling issues: in space, and in time, therefore requiring the use of various interpolation and estimations procedures to ensure the quality and representativeness of the spatial information produced. Spatial interpolation procedures are described, using for instance the ‘Spatialposition’ R Package.

- The guidance document is structured according to the workflow of the analysis, and describes methods and R code used to implement the case study, in order to provide ESPON with the conditions of reproducibility of the methodology used.

**Figure 1.1 - Overview of the workflow.**

To illustrate the project workflow, the guidance document covers the following aspects of the implementation:

1. The case of institutional data to describe residential property markets, with some examples of the variables useful to describe unequal access to housing. The section covers aspects from **WP1** (sources and transaction data), the harmonization of spatial definitions into 1 km cells, by the means of interpolation (**WP2**).
2. Section 2 covers the construction of **harmonized variables**, that will be produced at the local fine-grain geographical city level (grid and LAU2), as in **WP3**, using a mix of institutional and census data.

3. Finally, we use a sample of harvested data from *leboncoin.fr* website (target websites to be analyzed in **WP1**), to discuss the methodology used to harvest data (**WP2**), cleanup and produce usable datasets for the project, that can be compared and benchmarked with conventional data (**WP3**).

A conclusion will highlights how such data can be aggregated together to produce harmonized variables at the three geographical levels of interest for this study, *depending of data availability*: local grids, LAU2, central cities, FUAs.

For the purpose of the *Wellbeing of European citizens regarding the affordability of housing*, we go beyond the aggregated territorial levels to understand intra-urban inequalities between the cities. So as to deliver harmonized statistics for all the selected cities, 2 geographical levels will be used to aggregate collected data on housing dynamics: the LAU2 level, and the 1km European reference grid. The other challenge of the project will be to cover the entire Functional Urban Areas, despite missing data and incomplete datasets, as displayed for a very basic statistic s.a. population density. Where data is found missing, case studies will be conducted at list within the delineation of the core cities (surrounded in black on the figure below).
2 Using institutional data to analyze the dynamics on property markets: data, methods, sample results

2.1 Unconventional institutional data: data from Paris Chamber of the Notaries

We use property-level data from the Paris Chamber of Notaries (1996-2012), provided to the lead researcher by the Paris Notaires Services, a subsidiary of the Chamber of the Notaries. This sample contains transactions for the region and its suburbs, within the administrative limits of Ile-de-France (1 million rows). All records contain information on the property amenities and pricing, and series of understudied interesting variables on sellers and buyers, such as age, sex, socio-economic status, national origin, place of residence, and some credit history related to the transaction. This preliminary analysis is conducted on single family homes in the Yvelines department (west of Paris), that consists in a transect of the many residential type strata found in Ile-de-France, from the denser mixed fabric of the inner suburbs. The subsample, once cleaned (single family homes - detached tract houses), without null geographical coordinates, represents 35000 observations.

2.2 A spatial analysis of price dynamics: why a grid?

The focus being to analyze the geography of affordability through home-ownership inequalities with transactions, several issues have to be dealt with regarding the spatial level of aggregation. This section details the reasons why the 1 km grid as been chosen as a relevant aggregation level for the delivery of the data and their computation. The information displayed below is aggregated at a very low level of spatial granularity (1km grid, LAU2). On the one hand, it provides information in a “official” territorial division, as such (LAU2). But on the other hand, these datasets are subject to important outliers affecting the quality of the results, especially for territorial units described by a small number of real estate transactions, requiring the use of a grid to perform interpolation and estimation spatial statistics (1km grid).

- For spatial analysis purpose, a grid allows us to integrate datasets with various spatial definitons;
- Data secrecy, privacy control and legal and/or ethic requirements regarding the confidentiality of individual transactions ;
- The MAUP (Modifiable Areal Unit Problem), related to the spatial distribution of transactions and aggregation ;

The weakness of the sample and missing data issues. Grid interpolation allows us to estimate a potential price in adjacent cells, with assumptions regarding the spatial interactions between transactions. To offset these limitations, we use a combination of a
1km grid and techniques of interpolation, following the assumptions of Stewart’s potential, using the `SpatialPosition R package (Commenges et al., 2015). For examples and detailed discussion of methodology regarding data processing, gridding, interpolation, and mapping, see (Le Goix et al., 2019).

This is the reason the report details results using interpolation and estimation techniques, described in the Appendice 1 of the Guidance document, part 1.

2.3 A narrative of the methodology used

To better understand inequalities on housing markets, we start with nominal price (Figure 1a), and then produce harmonized variables, based on ratios, s.a. price-to-income, to analyze affordability; and debt-to-value, a proxy for inequalities stemming from equity capital availability of households. The ultimate goal of harmonized variables is to be able to compare between cities, and between countries, ceteris paribus. This section provides examples on how to match data contained in transaction datafiles (debt), and external sources s.a. income from 2011 census data.

An example of a harmonized variable produced with the transaction dataset only is debt-to-value ratio (Figure 1b). It is computed, for each grid cell, as the total amount of debt contracted and the total value of transactions. Two regimes can be highlighted on this kind of map, that show a very unequal spatial distribution of wealth and households’ vulnerability:

Asset accumulation, especially in the more affluent south-east side of the district, and

A detrimental effect for households purchasing a property in lower brackets submarkets, especially along the Seine river and the North-West part of the map. Because of their lack of assets, they are more likely to rely on a higher debt-to-value ratio when contracting the mortgage¹.

Another example consists in matching transaction data and income census data. Figure 1c shows the integration of such heterogeneous datasets, so as to produce a harmonized indicator describing unequal affordability of home-ownership: price to income ratio. Data show very high price-to-income ratio in the eastern part of the district, where prices are higher; and an apparent better affordability on the western part. Two contrasted regimes coexist:

Where prices are higher, transfers of assets (for instance from a previously owned property) allow households to better offset the effects of higher prices on income. Mortgages represent a lower share of the paid price.

¹ Information is incomplete and dataset contains more information with mortgages. Data may be missing for other conventional forms of loans, securitized by insurance mutual companies.
On the western side of outer suburbs, lower price-to-income spatially correlates with higher debt-to-value ratio. Where price are lower, households tend to rely mostly on mortgage to access ownership.

Such maps clearly depict the effects of property price, income, debt, and assets on unequal access to home-ownership.

Figure 1. Price and other Harmonized Variables

![Map of potential price (EUR) of single family homes, within 10 min. neighborhoods.](image)

![Map of debt to value ratio, within 10 min. neighborhoods.](image)

![Map of home price to household's income ratio.](image)
3 Harmonized indicators and methodology to analyse real-estate markets – first results

3.1 Comparative approach of households harmonized indicators

The first series of variables that we are to harmonize and match with price, to compare between and within case studies are variables of interest to standardize the understanding of inequalities: income, debt, for instance. This section provides an example of what can be done, adapted to the case of Paris FUA, with the data originating from unconventional institutional data (Paris Chamber of Notaries). To better understand inequalities on housing markets, we start with nominal price, and then produce harmonized variables, based on ratios, s.a. price-to-income, to analyze affordability; and debt-to-value, a proxy for inequalities stemming from equity capital availability of households. The ultimate goal of harmonized variables is to be able to compare between cities, and between countries, ceteris paribus. This section provides a series of examples on how to match and harmonize unconventional and conventional data to produce relevant indicators.

This sample contains transactions for the region and its suburbs, within the administrative limits of Ile-de-France (1 million records). All records contain information on the property amenities and pricing, and series of understudied interesting variable on sellers and buyers, such as age, sex, socio-economic status, national origin, place of residence, and some credit history related to the transaction. To be comparable between cities Europe, the transactions data are weighted by income levels (INSEE provides income data). The spatial analysis proposed offer many ways to compare within case studies, between neighborhoods where strong inequalities of affordability have to be monitored: Paris, Geneva, Avignon, Madrid, Lodz are expected to provide interesting insights. This example provides a global framework which aims at being extended for the rest of the case-studies of the project in term of data collection.

By the means of bringing together conventional and unconventional data, the policy relevance of the case-study is threefold:

- To monitor the spatial effects of pro-ownership policies on socio-economic inequalities, and the attendant risks of market-based exclusion.
- To analyze the spatial patterns of inequalities stemming from unequal capitalization of housing wealth some areas, vs. vulnerability of households in others.
- To better inform and to map the increased affordability gap, a critical issue for social cohesion sustainability in metropolitan areas in Europe.
The first results delivered in D2 are not yet available in their interpolated / harmonized version at the 1km grid cell level. The subsequent section discusses temporary results at the LAU2 level.

### 3.2 average price for transactions

Using the Chamber of Notaries database, it is firstly possible to propose the average price per square meters for the period 2011-2012. After cleaning the database (outliers, wrong values), it covers 88480 records for the Paris FUA (ordinary transactions, residential only). This data has been aggregated into the 1km European reference grid and into the LAU2 units.

*Map 3-1 – Paris Average price per square meters*

Obviously, the dataset includes a lot of missing values. It is mainly due to the fact that the number of transactions in most of the grid cells and the LAU2 in periphery of the Paris FUA are described by a few number of real estate transactions (are displayed on the map only geographical objects with a transaction number above 5). The corrections for missing values will be dealt with by the means of interpolation (see Appendix 1, *commented code for the Guidance Document, section 1.3*).

Nevertheless, such maps display a high heterogeneity of situations: average prices are comprised from 1507 euros per square meters to more than 6000 euros. Not surprisingly, the higher prices are located around the Paris core city.

### 3.3 Income data

However, this difference of housing prices is also closely linked to the level of wealth of the Parisian population. Income maps reveal contrasting situations between the North-East of the Paris Core city and the periphery with low level of income (below 18900 euros per inhabitant), as regards to the Western Part of the metropolitan area with income levels generally above 25000 euros per inhabitant.
It raises consequently the question of the relationship between housing prices and the level of wealth of the population.

3.4 Price to income ratio

Price-to-income ratio is used as a first harmonized indicator which can be produced starting from Notarial databases and census ones. The values displayed on the map display very high heterogeneity inside the Paris FUA: values come from 3 to 30 at LAU2 level, and from 3 to 50 at 1 km grid level.

The interest of going into a 1km grid, set aside the question of spatial harmonization at European level, consists in being able to observe inequalities at the municipal level, so as to highlight local peaks (high values surrounded by low ones for instance).
Data mapped show a striking figure of spatial inequality in the structure of affordability. For the owner occupied market, data show very high price-to-income ratio center and western part of the region, where prices are higher; and an apparent better affordability on the peripheries and to the east: the overall picture contrasts with previous maps displayed in 4.1.1 and 4.2.2.

The price to income ratio is characterised by highest values in the center of Paris and in its immediate suburbs (around 5 km). The highest values are concentrated in the Western Part of Paris (index 29.9 for Neuilly-sur-Seine and 21.6 for Paris), but Seine-Saint-Denis department, globally characterised by low income levels is also characterised by high price to income ratio (indexes 13.3 for Pré-Saint-Gervais, Montreuil and Bagnolet). It is more in the periphery of Paris FUA lowest values are found (20 km around of Paris), but with some interesting exceptions: Versailles (15.6), Saint-Germain-en-Laye (15.3), Saint-Maur-des-Fossés and Cachan (11.9), Chessy (11.4) or Massy (11.1).
3.5 Debt-to-value ratio

An example of a harmonized variable produced with the transaction dataset only is debt-to-value ratio. It is computed, for each grid cell / LAU2, as the total amount of debt contracted and the total value of transactions. Two regimes can be highlighted on this kind of map, that show a very unequal spatial distribution of wealth and households’ vulnerability:

- Asset accumulation, especially in the more affluent center and western part of the study area.
- A detrimental effect for households purchasing a property in lower brackets submarkets, especially along the Seine river and the North-West part of the map, and in North-East suburbs. Because of their lack of assets, they are more likely to rely on a higher debt-to-value ratio when contracting the mortgage.

A more general way to understand the interplay of both indices is to consider that where prices are higher, transfers of assets (for instance from a previously owned property) allow
households to better offset the effects of higher prices on income. Mortgages represent a lower share of the paid price. Lower price-to-income spatially often correlates with higher debt-to-value ratio. Where price are lower, households tend to rely mostly on mortgage to access ownership.

Map 3-5 – Debt to value ratio, 1km grid

Debt to value, 2011–2012

Source: ESPON Big Data for Territorial Analysis and Housing Dynamics, 2019
© EuroGeographics for the administrative boundaries
4 Unconventional (“Big data”) vs institutional data sources

This section discusses how to compare and benchmark institutional and unconventional data sources. Institutional data can be compared to possible harvested big data sources (e.g. real estate websites; open-data resources and warehouses; Airbnb…). To do so, we are currently testing how to harvest property listings from a real estate sale web service. Data collection has started in Poland, Spain, and will start after a testing stage in April 2019 in France and Switzerland. Our targets are leading platform of real-estate advertisements, by individuals as well as real-estate agents and companies to advertise residential properties to sell and to let.

4.1 From harvested data to harmonized indicators – overall process

Unconventional data are often viewed as interesting proxies to measure, and better understand spatial behaviors and territorial dynamics (Kitchin 2013; Gallotti, et al. 2016), and also as a means of providing higher spatio-temporal resolution data when compared to institutional data sources (FP7 EUNOIA final report, 2015).

Prior to relying upon the unconventional data sources, it is important to assess their reliability, and if they provide accurate information when compared to the long established, statistically robust information collection data. Studies address the representativeness of IDS (Internet Data Sources) compared to conventional data sources.

Follows a general technical narrative of the procedure we follow to harvest data, as detailed in the Appendix 1 Commented code for the guidance document.

- The first step consists in getting the total number of ads and deduce the number of pages to scrape after having identified the relevant query: in Figure 4.1 it corresponds to a geographical locator (name of the NUTS3) and the type of offer (real estate transactions). Then, the method consists in exploring automatically all the adds included in all pages of the query result. At the end, the aim is to obtain a list of URLs to be harvested (one URL by offer).

- A second step consists in identifying all the relevant and required information included in the ad’s. Scripts need to be created to gather the data automatically. This is a tedious, costly and time-consuming process. The cost and duration of the project allows for test drives and a few months of collection, and some platforms. We deliver a general methodology (cf. Appendix 1 Commented code for a Guidance document, section 3) that will be reproducible and adaptable. It is obvious that a script is valid for one real estate Website, considering the fact that they are coded differently. Moreover, if the real estate Website change the organisation of the Web page, the tags used in the script must be re-written. Such an iterative procedure is hand-made, highly artisanal, and highly consuming in qualified worked-force, therefore costly.
• **The thirst step consists in cleaning up the data gathered.** The most common mistakes errors are duplicated ad’s (sometimes a real estate ad can be published several times), absence of location coordinates or mistakes when entering the real estate ad (area, price, etc.). Consequently, results obtained through the Web scraping must be filtered. As an example for a case-study located in Yvelines in France, the 9934 observations collected resulted in 7460 unique and accurate records, with correct location down to the municipality. The geocoding resulting from this procedure is in many regards of poor quality compared to the locations provided by institutional data (transactions). In the preliminary study, location data is provided by the website either as the city or municipality (78%), and only a few ads are geocoded down to the address. In some other cases location appear to be the location of the agency.

![Figure 4.1 - Step 1 - Get links to all ads of a given real estate Website](image)

**Search homepage of a real estate Website**

https://www.leboncoin.fr/recherche/categorie=98&regions=12&departments=76&real_estate_type=12

![Number of Web pages of results](image)

**List of urls to be harvested**

https://www.leboncoin.fr/ventes_immobilieres/1548490497.htm
https://www.leboncoin.fr/ventes_immobilieres/1560521232.htm
https://www.leboncoin.fr/ventes_immobilieres/1500812641.htm

![Figure 4.2 - Step 2 – Collect and sparse geospatial information](image)
Figure 4.3 - Step 3 – Data cleaning to obtain an accurate dataset with geocoded transaction location
• The fourth step consists in interpolating point sample data into the EU reference layers of analysis (LAU2 and 1km grid). Considering the structure of the data collected, we estimate the best generalization of the spatial information gathered will be to apply the same spatial interpolation to the harvested data points, and map them down to the 1km grid and LAU2. The interpolation methods used, that produces estimates e.g. potential price or potential density of transactions in 1km neighborhoods, allows to compare, visualize, and benchmark the harvested dataset vs. the institutional transaction datasets. (cf. Appendix 1 Commented code for a Guidance document, section 3).

4.2 Narrative of a working example of data collection

This section provides a narrative of the example of data harvesting provided in Appendix 1 Commented code for a Guidance document, section 3.

A total 9,934 observations for houses to sell have been collected online by the end of October 2018. We used standard methodologies with rvest R language libraries, and applied it to the Yvelines to demonstrate the method, for one department only (Yvelines)2. Data collected included advertised price, the main characteristics of the property, location. Location generally describes that of the property, but not systematically. As in Figure 4.4a, location data is provided by the website either as the city or municipality (78%), and only a few ads are geocoded down to the address. In some other cases location appear to be the location of the agency of seller, and does not provide any usable location information for the scope of our study.

The map of extracted results (Figure 4.4b) shows a good coverage of the study area, with a wide range of municipalities (LAU2) covered. But data also show that alternative methods for geocoding and cleanup will be necessary if we intend to remove all outliers. For the case study, the 9934 observations collected resulted in 7460 unique and accurate records, as showed on Figure 4.4b. For the purpose of comparison, two subsets are produced, houses and apartments. We create a spatial dataframe from latitude and longitude. CRS for harvested data being WGS84, harmonization with institutional datasets requires to convert coordinates into a Lambert-93 RFG83 CRS.

To benchmark the results of harvested data, we also applied an interpolation to the harvested data points, and map it with the 1km grid (Figure 4.4c). The interpolation methods used, that produces estimates e.g. potential price or potential density of transactions in 1km

---

2 Portion of code for harvesting “leboncoin.fr” website has been initially prepared by L. Vaudor, S. Rey-Coyrehourcq, F. Pfaender for the PIA Labex Dynamite SummerSchool 2018, and was modified, with the authorization of the authors, which has to be acknowledged.
neighborhoods, allows to compare, visualize, and benchmark the harvested dataset vs. the institutional transaction datasets. The harvested data vs. institutional transaction data plot (Figure 4.4c) compares values collected with both methods.

### Figure 4.4. Summary of harvested data

5 Aggregating data at LAU2 levels, to produce relevant indicators for policy making

One interest of the analysis a micro-data level relies in the induced flexibility in aggregating data to better inform policy. This section provides some examples for aggregation and further analysis documented in Append. 1 Commented code for a Guidance document, section 4.

5.1 Output: some examples of indicators to compare and analyze unequal access to housing markets and living conditions

Among interesting policy analysis to characterize housing inequalities, well-being and segregation in European cities, income Gini index denotes the level of eveness or
concentration of income, while debt to value denotes the vulnerability of households in localities according to the value of real estate assets, compared to the debt contracted. The plot on Figure 4.5 shows that in the case study, the lower income inequalities are, the more households living in these areas are on average likely to accumulate assets, with lower debt-to-value ratios.

Other type of data collection will be used in the project, s.a. rental websites, or Airbnb data for central cities only. The interpolation methods used, that produces estimates s.a. potential price or potential density of transactions in 1km neighborhoods, allows to compare, visualize, and benchmark the harvested dataset vs. the institutional transaction datasets. Some examples of LAU2 aggregates are displayed on Figure 4.5. Plots show for instance a comparison between values, and densities of transactions (i.e. market activity), collected with both methods. Although the quality of price estimation is good, the unexplained variance is significantly higher with regards to density, i.e. the spatial distribution of transactions vs. advertised properties.

Data also show that the price to income ratio, i.e. pressure on income linked to the cost of housing is an interesting variable to analyze inequalities of access to housing markets. Regressed by median price (here, advertised price), it clearly shows the effect of price on sustainability of living conditions for households. However, comparison of Figures 1, 2 & 3 also show that residents in wealthier neighborhoods are more likely to also consider the purchase of more expensive houses, with higher price-to-income ratio, while assets are mobilized, leading to lower debt-to-income ratio.

Aggregates can therefore be used for other various analysis related to policy issues for the Wellbeing report. For instance, harmonized variables such as Median transaction price, Median advertised price, Price interquartile range, Mean price to income ratio, Mean debt to value ratio, Income gini index will be relevant to analyze inequalities of access to housing markets, and unequal vulnerability of neighborhoods in terms of price dynamics vs. endebtness of households. All these issues have been introduced according to the policy relevance in the background section.

Figure 4.5. Comparison of aggregated harvested and institutional data at the LAU2 level.
5.2 harmonized variables and geographical levels at which data can be prepared and delivered

The proposed indicators collected for the case study in the course of the project will be as follow. This proposal is subject to the actual collection of data, according to the data sources surveyed and subject to limits of data availability. Such limits are explicited in the appended documents Table of data sources available and D1_case_study_presentation.

Considering the caveat aforementioned, and elaborating on the code and methodology developed in this document, we propose to target the delivery of the following variables for the case study:

At the finer aggregate level possible (target : 1km grid):

- price paid : transactions price paid (institutional data). Aggregated in median price Q25, Q75, IQR at LAU2 and FUAs
- price asked : advertised price (harvested data). Aggregated in median price Q25, Q75, IQR at LAU2 and FUAs
- surface of property : advertised or transactions (institutional data and/or harvested data). Aggregated in median price Q25, Q75, IQR at LAU2 and FUAs
- number of rooms : advertised or transactions (institutional data and/or harvested data). Aggregated in median price Q25, Q75, IQR at LAU2 and FUAs
- density of transactions : advertised or transactions (institutional data and/or harvested data). Aggregated as average density
- debt to value ratio : transactions where debt info available (institutional data only). Ratio. Limited availability of data for several case studies
- price to income ratio : advertised or transactions for price. Institutional data for income. Ratio. Limited availability of data for several case studies.

At the LAU2 level:

- Aggregates of variables listed at lower geographical levels.
- Gini index of HH income in case studies where lower levels geographies of income are available
- Relevant census data (s.a. median income, households income)

At the core-cities and FUAs levels:

- Aggregates of variables listed at lower geographical levels only.
- Gini index of HH income in case studies where only LAU2 income data are available
References

