BANCA D'ITALIA

EUROSISTEMA

# Temi di discussione

(Working Papers)

## The potential of big housing data: an application to the Italian real-estate market

by Michele Loberto, Andrea Luciani and Marco Pangallo

April 2018

Number 1171

BANCA D'ITALIA
EUROSISTEMA

# Temi di discussione

(Working Papers)

## The potential of big housing data: an application to the Italian real-estate market

by Michele Loberto, Andrea Luciani and Marco Pangallo

Number 1171 - April 2018

ISSN 1594-7939 (print)
ISSN 2281-3950 (online)

*Printed by the Printing and Publishing Division of the Bank of Italy*

# THE POTENTIAL OF BIG HOUSING DATA: AN APPLICATION TO THE ITALIAN REAL-ESTATE MARKET

by Michele Loberto*, Andrea Luciani* and Marco Pangallo**

## Abstract

We present a new dataset of housing sales advertisements (ads) taken from Immobiliare.it, a popular online portal for real estate services in Italy. This dataset fills a big gap in Italian housing market statistics, namely the absence of detailed physical characteristics for houses sold. The granularity of online data also makes possible timely analyses at a very detailed geographical level. We first address the main problem of the dataset, i.e. the mismatch between ads and actual housing units - agencies have incentives for posting multiple ads for the same unit. We correct this distortion by using machine learning tools and provide evidence about its quantitative relevance. We then show that the information from this dataset is consistent with existing official statistical sources. Finally, we present some unique applications for these data. For example, we provide first evidence at the Italian level that online interest in a particular area is a leading indicator of prices. Our work is a concrete example of the potential of large user-generated online databases for institutional applications.

**JEL Classification**: C44, C81, R31.
**Keywords**: big data, machine learning, housing market.

## Contents



---

* Bank of Italy, Directorate General for Economics, Statistics and Research.

** Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford OX2 6ED, UK. Mathematical Institute, University of Oxford, Oxford OX1 3LP, UK.

# 1 Introduction

The attention of economists and policy makers to the functioning of housing markets has certainly increased in the recent years.[1] Due to the large exposure of the banking sector to the real estate, housing is extremely relevant for financial stability.[2] But the housing market also affects the real economy throughout multiple channels. The temporal real-estate price trends substantially influence the construction industry, which in turn affects GDP growth. Housing is "the democratic asset" (Glaeser and Nathanson, 2015), constituting in Italy about 85% of household wealth, and thus shapes the consumption patterns of the population through expectations and wealth effects (Mian et al., 2013). Finally, the spatial real-estate price trends may represent a serious obstacle for economic growth. Indeed, extremely high prices and residential income segregation may prevent a large fraction of workers to reside in the most productive areas (Hsieh and Moretti, 2017).

Yet, there are both theoretical and empirical limitations to our understanding of housing markets. On the theoretical side, one major problem is that dwellings are heterogeneous goods exchanged in decentralized and segmented markets. As a consequence spatial and informational frictions make the Walrasian equilibrium concept unsuitable for this market.[3] On the empirical side, lack of comprehensive data has always represented a shortcoming, both for research and for policy. Microdata on actual housing transactions are available only in a few countries and unfortunately many of these sources show limitations in the spatial or in the temporal dimension. Much more challenging is finding comprehensive information about the full history of housing transactions, from the moment the dwelling goes on the market up the actual transaction.[4]

Online data are starting to fill some gaps. For example, Piazzesi et al. (2017) use online search data from trulia.com – a popular online real-estate portal in the United States – to analyze housing demand, so far ignored, and the behavior of buyers across different segments. Anenberg and Laufer (2017) show that a house price index constructed from online listing data is more timely than standard price indexes based on administrative data, because transaction deeds are registered with a lag of several months. Our paper is the first exploiting online sales advertisements for the analysis of the Italian housing market. We highlight the strengths and the weaknesses of these data and discuss in particular their complementarities with existing sources.

In Italy, the main provider of spatially disaggregated data on housing markets is Osservatorio del Mercato Immobiliare (OMI), namely the real estate market observatory of the Italian Tax Office. Among the several administrative datasets maintained by OMI, two are particularly relevant for the analysis of the housing market. First, OMI disseminates twice per year estimates of minimum and maximal house values within so-called OMI micro-zones, which are uniform socio-economic areas roughly corresponding to neighborhoods. The OMI estimates are based on a limited sample of actual transactions and sales offers collected by real estate agencies. Second, OMI disseminates statistics on the volume of housing transactions at city level.

OMI maintains also the database of microdata on all the transaction deeds recorded in the notary registries, but unfortunately this database is not public. The main shortcoming of the OMI datasets is the limited information about the physical characteristics of the transacted housing units. This hinders a serious analysis of market segmentation and potentially creates

[1] We are extremely grateful to Immobiliare.it for providing the data and for their assistance. All mistakes are our own.

[2] See for example (Ciocchetta et al., 2016) for an analysis of the Italian case.

[3] Search theory addressed the problem of the lack of Walrasian equilibrium and showed that price adjustments also occur along the time dimension (Han and Strange, 2014). Moreover, the peculiarities of houses as investment goods make them really difficult to be studied in the standard asset pricing paradigm (Piazzesi and Schneider, 2016).

[4] In this respect a notable exception is the database exploited by Merlo and Ortalo-Magne (2004), which includes the full histories of a sample of housing transactions, coming from four British real-estate agencies. However, their sample size is limited to 780 transactions.

difficulties in the construction of quality-adjusted price indexes. Moreover, the OMI estimates of house values only become available with a lag of several months, and in some situations might not be representative of the universe of transactions (due to the small sample size). Additionally, the neighborhood level of geographical aggregation is sufficient for most applications, but prevents the study of more localized phenomena.[5] Finally, information about demand and time on market is lacking from administrative data.

We potentially address all these shortcomings by analyzing a new "big data" database, containing housing sales advertisements (ads) on Immobiliare.it, a popular online portal for real-estate services in Italy. The data cover the period since early 2015 up to the end of June 2017 and consist of both residential and commercial units (for sale or for rent/lease). In this paper we only focus on sales of residential units in the 110 provincial capitals, which include all major cities and comprise about 18 million inhabitants in total.[6] Our database consists of about one million unique ads, that we monitor from the time they were created up to the time they were removed from the database.

Every record comprises detailed information on the listed housing unit (asking price, floor area, energy class, maintenance status, number of rooms, etc.), on the building (elevator, garage, garden, etc.), on the ad (publication and removal date, number of visits – clicks –, etc.), and a short description. From the publication and removal dates, we can construct an estimate of the time on market, under the assumption that when the ad was definitively removed from the database the house was sold. We also know the latitude and longitude of the listed housing unit, so we can study the housing market at an arbitrarily fine geographical level. By aggregating the visits (clicks) over the neighborhoods, and dividing by the total supply, we can construct a proxy for the demand tightness.

Having already described the advantages of analyzing this dataset, we focus on the problems we encountered. The main issue with this dataset is that there is a significant fraction of duplicates, namely more than one ad referring to the same dwelling. This issue affects more or less all the "marketplace" websites, because, for example, the sellers post multiple ads for the same good or remove and post again the ad multiple times in order to make it appear always as recently published. Here this problem is trickier, since the same good can be sold by multiple sellers: the owner of a house can entrust more than one real estate agency for the sale of her dwelling, leading to a duplication of this dwelling in the dataset.

We correct this distortion using machine learning tools. These algorithms autonomously learn the criteria that identify the duplicates after they are given pairs of ads that certainly refer to the same housing unit. Machine learning algorithms are mostly effective thanks to the large amount of data they can learn from – which is why they were not widespread before the recent explosion of large granular datasets. After running the "deduplication" procedure, we end up with a dataset of about 650 thousand housing units. We show that the distortion implied by duplicates has a significant magnitude mainly when we focus on the short-term housing market dynamics in small geographical aggregates. If a researcher is interested only in the heterogeneity across cities or in the dynamics at a relatively broad geographical level, the overall distortion seems less relevant.

We then validate the dataset, by comparing its summary statistics to those coming from official sources, such as OMI, or the quarterly Italian Housing Market Survey (conducted by Bank of Italy, OMI and Tecnoborsa). We find that after the deduplication procedure online ads provide a picture of the housing market broadly consistent with official sources.

Finally, we analyse a number of issues that could not have been addressed using currently

[5]For example, the edification of a prestigious building could trigger a gentrification process that first diffuses within the neighborhood and then to the bordering neighborhoods.

[6]In 2016 the number of actual housing transactions in the provincial capitals was 183,000 units (about one-third of all housing transactions in Italy). In cities the majority of transactions is brokered by real estate agents – who are more likely to upload an ad on Immobiliare.it than private citizens –, whereas in small towns and in the countryside sales are less likely to need brokerage and so representativeness is potentially a problem.

available public data sources on the Italian real-estate market. First of all, we are able to provide quantitative evidence about the evolution of the stock of dwellings for sale and about the composition of supply in terms of physical characteristics. Then, we now-cast the aggregated price level and show that it is possible to anticipate by several months the evolution of actual average housing prices, constructed from OMI data with the methodology proposed in Cannari and Faiella (2007). After performing hedonic regressions, we calculate the quality-adjusted price indexes in Rome and Milan and show the price evolution in multiple segments. We also provide first evidence at the Italian level that online interest for a particular neighborhood – the demand tightness described above – is a leading indicator of prices.

This paper is organized as follows. Section 2 describes the Immobiliare.it ads dataset, while in Section 3 we show how we create the final dataset of housing units. In Section 4 we compare the information coming from our dataset with the official statistical sources available. In Section 5 we show the potential applications of the dataset, and Section 6 concludes.

## 2 Description of the dataset

We analyze a novel dataset provided by Immobiliare.it (`www.immobiliare.it`), the largest online portal for real-estate services in Italy. The primary purpose of Immobiliare.it is to ease the match between buyers and sellers in the housing market. Indeed, the core of the website is the search engine that allows to browse thousands of advertisements (ads) of dwellings. While Immobiliare.it deals with both sales and rents, in this study we only focus on sales.

The sellers are private citizens or real estate agencies. They upload an ad for the property they are selling and, if the ad is set as visible, every internet user can visualize the ad without the need to sign up on the website. On the contrary, the sellers have to first register for an account. Private citizens can hold an account for free, whereas agencies have to pay a fee that depends on the number of ads they post.

Potential buyers can search by geographical criteria (map search) and by the physical characteristics of the dwellings. They can also specify a price range and look at the pictures or at the textual description of the ad. Once the potential buyers identify properties they are interested into, they can contact the seller who posted the ad. Immobiliare.it provides phone and email contacts of the users.

### 2.1 Construction of the ads dataset

Our data consist of weekly snapshots of ads located in the Italian provincial capitals. By weekly snapshots we mean the ads that are visible on the website every Friday, since 2016, January 5 until 2017, July 6. For 2015 only quarterly snapshots are available, therefore in our analysis we rely mostly on ads visible from early 2016 on.

In practice, most ads remain unchanged between two weekly snapshots. The average turnover is about 5%, meaning that 5% of the ads are removed from the dataset between two snapshots and every weekly snapshot contains on average 5% new ads. Some of the physical characteristics of the dwellings reported in the ads change between two snapshots. We always rely on the latest available features, because we assume that the sellers correct the mistakes they might have made when posting the ad.

There are three variables in the snapshots whose trends are mostly meaningful for our analysis. These are price, number of visits (clicks) on the ad and number of times potential buyers contacted the seller through the website.[7] For example, it is important to know the sequence of price revisions if one is interested in bargaining dynamics, and an upward trend of clicks and contacts in a certain neighborhood could unveil a gentrification process.

[7]In the webpage of each ad there is a form that can be used to send a message to the seller asking information about that particular dwelling.

Therefore, we keep all information about price, clicks and contacts by saving all values together with the modification date. We finally construct the main dataset by keeping unique ads, as selected by their unique identifier.

### 2.2 Content of the ads dataset

The full set of available information is summarized in Table 1. A more exhaustive description can be found in Appendix A. Here we first discuss some of the variables, we then explain how we fill in some missing values using the textual description of the ads, and we finally describe the preliminary cleaning of the dataset.

| **Type of data** | **Variables** |
|---|---|
| Numerical | Price, floor area, *rooms*, *bathrooms* |
| Categorical | Property type, furniture, kitchen type, heating type, *maintenance status*, *balcony*, *terrace*, *floor*, air conditioning, energy class, *basement*, *utility room* |
| Related to the building | *Elevator*, *type of garden*, *garage*, *porter*, building category |
| Contractual | *Foreclosure auction*, contract type |
| Related to the seller | Publisher type (private citizen or real estate agency), agency name and address |
| Visual | Hash codes of the pictures, pictures count |
| Geographical | Longitude, latitude, address |
| Related to the ad | Visits, contacts |
| Temporal | Ad posted, ad removed, ad modified |
| Textual | Description |

Table 1: Information contained in the dataset provided by Immobiliare.it. Variables in italic are complemented using semantic analysis on the textual description of the ad.

For each ad, together with the asking price, we are given detailed information about the physical characteristics of the housing unit. These include floor area, number of rooms and bathrooms, whether the property is furnished, the condition of the kitchen (eat-in or kitchenette) and the heating system (autonomous or centralized). We also know if the dwelling has an air conditioning system and a balcony or terrace, and if the building where the housing unit is in has an elevator, a garden and a garage.

Some other characteristics are not objective and left to the best judgment of the seller. For example, the ad reports the maintenance status, the property type (e.g. apartment or attic, detached house or villa) and the building category (luxury, average, cheap). The energy class is instead certified officially.

From a contractual point of view, we also know if the dwelling is sold through a foreclosure auction and the type of contract, i.e. entire ownership, bare ownership, usufruct, etc. As already mentioned, we know if the user posting the ad is a private citizen or a real estate agency, and in the latter case we are given the name and address of the agency.

Ads can be uploaded with some pictures, a video, a virtual tour and the floor plan of the property. We are just given the hash codes of the pictures, which serve as unique identifiers of the images, and the number of pictures.

Among the most important variables are the geographical coordinates of the dwelling. When uploading the ad, sellers can either select the position of the dwelling on the map, or write the address of the property, in which case the coordinates are automatically selected. We match the location of the ad with the perimeters of the OMI microzones[8] and of the census areas, so we obtain very detailed information on the socio-economic characteristics of the neighborhood and on the stock of buildings in the surrounding area.

Some information is related to the ad itself. For example, we are given the number of visits (clicks) on the ad and the number of times potential buyers contacted the seller for that particular dwellings. We also know the date the ad was created on the website, and the day when it was removed. We construct a new variable that keeps track of the last time the ad was modified. Moreover, by looking at the weekly variation in the number of visits we are able to identify the periods when the ad was not visible (the seller may turn off the visibility of the ad, e.g. in case she is negotiating with a potential buyer). We take the time difference between the removal and the upload of the ad as a proxy of the time on market, implicitly assuming that the removal corresponds to the sale of the property.

Finally, almost all sellers write a brief description of the dwelling. This is usually a short paragraph that contains the same information that is stored in the other variables, but also provides more details about the neighborhood and the agency that sells the property, or mentions some characteristics that are not explicitly considered by Immobiliare.it (e.g. basement).

We use the textual description to fill missing data for specific and relevant variables. In particular, we extract information from the description only when the variable is missing in the dataset and only for a subset of the variables: terrace, balcony, elevator, garage, garden type, floor level, number of bathrooms, number of rooms and maintenance status.

For the first five variables if nothing is said about them in the description we assume they are absent; since these characteristics are almost all dichotomous and have an impact on house valuation, we are implicitly assuming that if they were present they would have been surely mentioned in the description or among the characteristics. For the remaining variables, instead, we fill the missing data only if exact information can be extracted.

Finally, we use the textual description to extract information about the existence of a proprietary basement, a utility-room and a janitor (*porter*). The textual description is also useful to identify the foreclosure auctions and new construction homes, i.e. sales in buildings still in progress.

In our analysis we focus on ads with entire ownership and in which the type of property is one of the following: apartments, attics, detached and semi-detached houses, loft and open spaces. Moreover, we consider only the ads for which both geographical coordinates and asking price are present. We eliminate also ads that are removed in less than a week and those related to dwellings sold through foreclosure auctions or in buildings still in progress. The set of ads we will work on counts 1,037,095 ads. About 92% of those ads are posted by real estate agents, the remaining by private users.

# 3 Duplicated ads and construction of the housing units dataset

The main issue with the dataset is that several ads referring to the same dwelling can be simultaneously or at different points in time posted by the users, meaning that the number of ads is by far bigger than the actual number of dwellings on the market. In this section we use machine learning methods to cluster all ads that refer to the same housing unit, so to create a "housing units dataset" in place of an "ads dataset".

There are many reasons why multiple ads are posted. First of all, in Italy there is no legal

[8] Osservatorio del Mercato Immobiliare (OMI) is the real estate market observatory of the Italian Tax Office. OMI manages so-called OMI microzones, which correspond to homogeneous areas for what concerns socioeconomic and geographic characteristics and cover almost the whole Italian territory (they are about 30 thousand). For each microzone OMI provides biannual estimates of the minimal and maximal house price per m2 and expected minimum and maximum rent per m2 for each type of housing unit (provided that in the microzone there is a sufficiently high number of dwellings for a particular typology).

obligation for owners to entrust at most one real estate agent for the sale of their property. This means that two or more real estate agents may be selling the same dwelling.[9] Then, in many cases only one real estate agency is entrusted to sell a dwelling, but this agency posts more than one ad for the same house.[10] Finally, we should also consider the case in which the mandate of the agent ceases and the owner of the house entrusts a new agent. In this case the two ads are not simultaneously present in the dataset, but we still need to know that these ads refer to the same dwelling.[11]

Keeping duplicated ads in the sample leads to a misrepresentation of the supply and can produce a bias in the subsequent analysis, as the presence of more ads for the same dwelling is far to be random and possibly associated with a need to sell soon or difficulties to find a buyer. Moreover, when different ads referring to the same dwelling are not clustered we incur in the issue of underestimating the time a dwelling has been on the market.

We show that duplicated ads are particularly harmful to measure growth rates in small samples – because multiple duplicated ads may over-represent the specific housing unit they are associated to –, whereas the problem is less serious when working with levels or large samples. If the effect of duplicated ads averages out, it is sufficient to correct for the biases on supply and time on market. We conclude that, as long as the duplication process is stationary, working with ads is fine at the aggregate level.

### 3.1 Evidence of the problem

A first assessment of the quality of the dataset of online ads can be made comparing the information coming from those data with similar statistics coming from existing and reliable sources. Here we focus on volume of transactions and time on market (which are the most critical variables), and then we show that the presence of duplicated ads is not random.

In our dataset there is no information regarding the actual sale of the dwellings, so we assume that ads removed from the website potentially represent house sales.[12] We compute for each quarter and each city the number of ads removed and we compare those with the actual number of housing transactions provided by OMI. On average we find a high correlation (the adjusted $R^2$ is equal to 0.96), and this result holds also if we look at several sub-samples of cities.[13]

However, focusing on the biggest eight Italian cities, Table 2 shows that for 2016 as a whole the number of ads removed from the website is by far bigger than the number of actual transactions. Moreover, the ratio between those quantities shows huge volatility: it goes from 268% of Florence to 116.7% of Palermo. Intuitively, this becomes a big issue when a researcher is interested in analyzing the evolution of housing market at a very fine geographical level, as dwellings with duplicated ads are over-represented. We will provide more evidence on this point once we have explained how we create the dataset of housing units.

A further important issue with duplicates emerges from Table 3. Here, we compare the average time on market of ads removed in each quarter (measured as number of months between

[9]One possible explanation for the owners' behavior is that they want to reduce the time to sell the house and by increasing the number of real estate agents they increase the probability to find a buyer.

[10]For example, real estate agents know that some potential buyers search on the website starting from the most recently published ads; this implies that after some period they need to post a new ad for the same dwelling, in order to get the attention of more buyers on a particular dwelling.

[11]A final case that is worth mentioning is the one where both the agency and the owner post an ad.

[12]This is one of the main issues of the dataset. Sometimes agencies keep the ads posted online also if the housing unit has been already sold. In other cases after the dwelling is sold the ad is no more visible, but the agencies do not remove it definitively from the website. Finally, sellers can decide to withdraw their housing units from the market.

[13]This result has been quite surprising, because OMI identifies for each sale the reference period based on the date of the property deed, that, according to the Italian Housing Market Survey, is on average 3 month later than the date of the agreement between buyer and seller (and this should be in principle what we observe in our data).

| City | IMM | OMI | IMM/OMI*100 |
|---|---|---|---|
| Turin | 20263 | 12322 | 164.4 |
| Genoa | 10358 | 6601 | 156.9 |
| Milan | 40342 | 21909 | 184.1 |
| Bologna | 7655 | 5507 | 139.0 |
| Florence | 12833 | 4786 | 268.1 |
| Rome | 73070 | 30173 | 242.2 |
| Naples | 9764 | 6650 | 146.8 |
| Palermo | 5504 | 4718 | 116.7 |

Table 2: Transactions. Comparison between ads and OMI data.

the day the ad was posted and removed from the website) with the equivalent statistics coming from the Italian Housing Market Survey.[14] As expected, the statistics computed on the dataset of ads underestimate the actual time of market, as the sale of a single dwelling can be associated to several ads posted in different periods of time.

| Year | Quarter | IMM | Survey BI |
|---|---|---|---|
| 2016 | 1 | 4.6 | 7.5 |
| 2016 | 2 | 4.0 | 7.7 |
| 2016 | 3 | 4.4 | 7.9 |
| 2016 | 4 | 4.4 | 6.8 |
| 2017 | 1 | 4.8 | 6.4 |
| 2017 | 2 | 4.2 | 6.4 |

Table 3: Time on market. Comparison between ads and the Italian housing market survey, conducted by the Bank of Italy.

An additional source of concern when using the original data is that the existence of duplicates is not random. To prove this point we build a binary variable that takes value 1 if there is more than one ad associated to a housing unit (as determined using the algorithm described in Section 3.2). We then run a logit regression of this variable on several characteristics of the dwelling and variables measuring the relative demand for the dwelling and its relative price compared to other dwellings in the neighborhood.

Overall, we do not find any meaningful correlation between the presence of duplicates and the physical characteristics of the housing unit. However, Table 4 shows that the presence of duplicates is correlated with the relative demand for that particular housing unit and with its relative price.[15] While the sign of the correlation with the price changes among different specifications, the correlation with demand variables is more robust and clearly shows that dwellings with many duplicates are also those with relatively lower demand.

[14]The Italian Housing Market Survey is a quarterly survey conducted by Banca d'Italia, OMI and Tecnoborsa that covers a sample of real estate agents and describes their opinions regarding the evolution of the Italian residential real estate market. More information about the survey is available at `http://www.bancaditalia.it/pubblicazioni/sondaggio-abitazioni/index.html`. As the survey does not provide evidence regarding all the provincial capitals, but only about cities with a population greater than 250 thousands persons, we restricted the comparison to these cities.

[15]The variable *demand* is defined starting from the ratio between the total number of visits to all ads associated with the housing units and the number of ads. Then, we take the ratio between this measure and its average in the OMI micro-zone of the housing unit. This is a measure of relative demand for a particular housing unit compared to the other dwellings in the OMI micro-zone. *Daily demand* is constructed in a similar manner, but now the number of visits is divided also by the number of days the housing unit has been on the market. The last two variables are respectively the ratio between the asking price per square meter of the housing unit and the average in the OMI micro-zone (*overvaluation*) and the inverse ratio between the predicted (according to the hedonic regression) and the actual asking price (*hedonic overvaluation*).

These correlations are not a by-product of our deduplication procedure, because we do not use these variables to identify duplicates (see Section C). Therefore, there exists a predictable over-sampling of particular dwellings and this calls for particular attention in the analysis of the original ads data.

Table 4: Determinants of duplicated ads

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Ad is duplicated | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| Demand | $-0.597^{***}$ | | | | $-0.295^{***}$ |
| | (0.006) | | | | (0.007) |
| Daily demand | | $-1.824^{***}$ | | | $-1.786^{***}$ |
| | | (0.011) | | | (0.011) |
| Overvaluation | | | $0.092^{***}$ | | $-0.529^{***}$ |
| | | | (0.024) | | (0.044) |
| Hedonic overvaluation | | | | $0.132^{***}$ | $-0.944^{***}$ |
| | | | | (0.023) | (0.045) |
| Observations | 197,860 | 197,860 | 197,860 | 197,860 | 197,860 |
| *Note:* | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | | |

## 3.2 Construction of the housing units dataset

Given the considerations in the previous section, our goal is to depart from the original dataset of ads and to build a new dataset of housing units on sale. This means that we should identify the duplicates among the ads and collapse them as if they were a single ad.

Here we just sketch the working of the algorithm. The whole procedure is carefully explained is Appendix C, and we also provide the pseudo-codes in Appendix D. Our approach is based on the standard methodologies adopted for the deduplication of datasets and, in particular, on Naumann and Herschel (2010) and Christen (2012). Loosely speaking, the operation occurs in three steps (Figure 1).

First, we pre-process the ads. We associate to each textual description a numeric vector, in order to meaningfully measure the distance between two descriptions. Some algorithms that accomplish this task just consider the multiplicity of the words. We use instead the Paragraph Vector (or *doc2vec*) algorithm (Le and Mikolov, 2014), in which a neural network learns about the order and semantics of the words. The numeric vectors computed by *doc2vec* accurately evaluate the similarity between descriptions. In some cases we are even able to tell that different ads for different housing units are posted by the same agency, because the style of writing the description is similar.

We also convert the class of some variables from categorical to numerical to meaningfully calculate the difference between some characteristics. For example, if two ads report that the maintenance status is either good or excellent, it is possible that they refer to the same dwelling. If instead one ad reports that the dwelling should be renovated, it is unlikely that the two ads are duplicates.

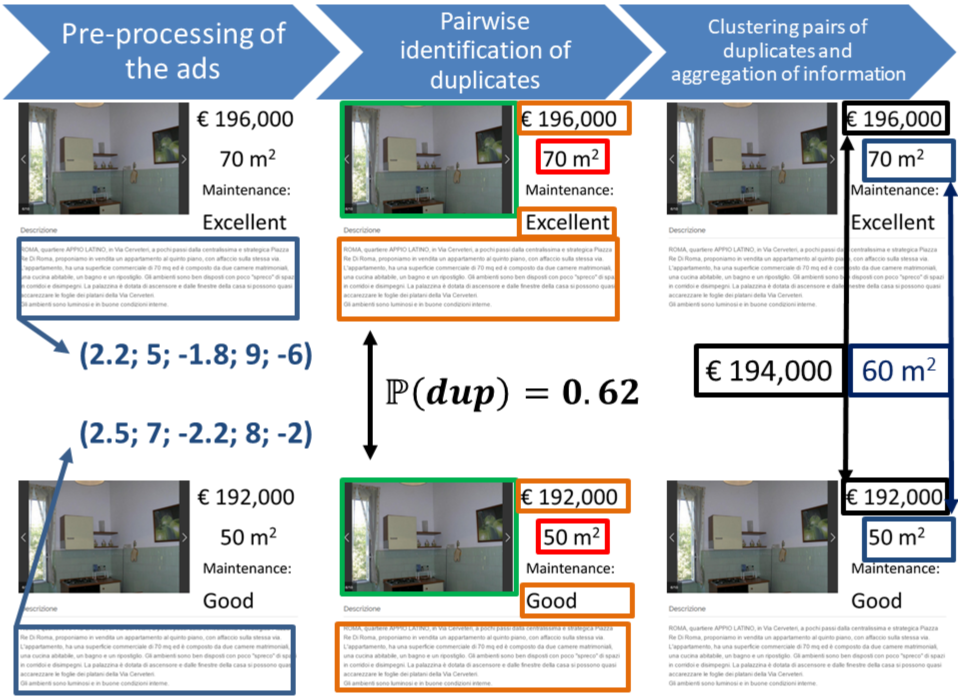



Figure 1: Simplified example illustrating the deduplication procedure. Each row shows a different ad. Both ads refer to the same housing unit, although some characteristics and the description are different. The first column shows an example of the pre-processing of the ads: Using the algorithm *doc2vec*, we transform the textual descriptions in numeric vectors. The second column shows how the pairwise identification of duplicates occurs. The frames are colored according to the similarity of the characteristics between the two ads (red $\rightarrow$ dissimilar; orange $\rightarrow$ similar; green $\rightarrow$ identical). The descriptions are compared by calculating the distance between the numeric vectors. A machine learning algorithm uses all similarity information to output a probability that the two ads refer to the same housing unit. If this probability is larger than 0.5, we consider the two ads as duplicates. In this case, the algorithm correctly identifies the duplicates. Finally, the third column gives an intuition of how we aggregate the information coming from the two ads.

In the second step, we perform a pairwise comparison of all pairs of ads that can potentially be duplicates – e.g. because they are geographically close or their price is relatively similar – and identify which pairs are likely to be duplicates (i.e. refer to the same housing unit).[16]

One option to determine this is manually coding a fixed set of rules and then applying a threshold. For example, we could compare the price, the geographical location and some physical characteristics of the housing units, aggregate this information with arbitrarily determined weights (e.g. 0.5 for the price difference, 0.2 for the geographical distance and 0.05 for some physical characteristics) and finally check if the so-defined similarity measure is above an arbitrarily defined threshold.

We use instead a machine learning algorithm, the C5.0 classification tree proposed by Quinlan (1993). The advantage of machine learning (James et al., 2013) is that it is not necessary to hard-code all the above rules. It is the algorithm that autonomously learns which variables are most relevant to identify duplicates, once it is supplied with a sufficiently large *training sample*, i.e. a dataset of pairs of ads of which we know with certainty if they are duplicates or not. We manually construct the training sample by looking at pairs of ads on the website and using the pictures and our best judgment to decide whether the two ads refer to the same dwelling. We

[16] To keep the pairwise comparison computationally feasible, we proceed iteratively for every weekly snapshot and only compare the newly created ads to the previously identified housing units. We name this procedure "time machine approach" (see Section C.4 in Appendix C).

then run the classification tree which outputs a probability that the two ads are duplicates. If this probability is larger than 0.5, we consider the two ads as referring to the same housing unit.

In the last step we start with a list of pairs of duplicated ads and we create clusters of ads that refer to unique housing units. Indeed, in the simplified example in Figure 1 we consider only two dwellings, but we can easily incur in groups of ads that refer to the same housing unit and some ad is not estimated to be a duplicate of another. Suppose for instance that ads A, B and C refer to the same dwelling. It is possible that the pairs (A,B) and (B,C) are classified as duplicates, but the pair (A,C) is not. In this case we use methods from graph theory and consider a cluster of ads as referring to the same housing unit if an internal similarity condition is satisfied. Finally, we aggregate information coming from the different ads by considering the average of the values (as in Figure 1) or the most frequent characteristics.

We apply the deduplication algorithm to the dataset of ads. According to our procedure, the total number of dwellings is about 654,000 units. The number of related ads is instead equal to 1,037,095 units, meaning that the number of effective dwellings is only 63% of the total number of posted ads. Looking to Table 5, it should be noted that the large majority of dwellings have only one associated ad, while the duplicates are concentrated over a smaller number of houses.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 or more |
|---|---|---|---|---|---|---|---|
| Number of dwellings | 465,041 | 113,365 | 37,566 | 15,981 | 7,723 | 4,264 | 9,559 |

Table 5: Distribution of dwellings by number of associated ads

According to our procedure, the main trouble with duplicates does not arise when we look at a single day, in which they account for about 20% of total ads. The real issue is that duplicates accumulate across several weeks, possibly for the reasons we discussed at the beginning of this section.

We find that the share of duplicates over total ads increases with city size and there is significant variability across cities.[17]

After the deduplication process we make additional controls on the dataset to address for potential errors in the data. First of all we keep only the dwellings that have been on the market for almost two weeks. Then, we drop from the dataset those dwellings for which the price is not sufficiently consistent with the characteristics of the housing units. In this way we are also able to identify foreclosure auctions that were not previously identified, because for example in the textual description the auction was not reported.

Our approach consists of running a hedonic regression, estimating for each dwelling the ratio between actual and predicted price and eliminating the housing units with a ratio between asking and predicted price lower than 0.5 or higher than 1.5.[18]

The cleaned sample that we will consider in most applications consists of those dwellings that have been on the market at least after January, 1, 2016 and it amounts to about 465,000 housing units.

### 3.3 Comparison between ads and housing units datasets

In this section we compare the original datasets of ads and the one we derived on housing units, in order to find out under what circumstances omitting the deduplication procedure would entail a bias in the results

[17] For example, the ratio between the number of ads and housing units is equal to 1.75 for Naples and 2.15 for Milan.

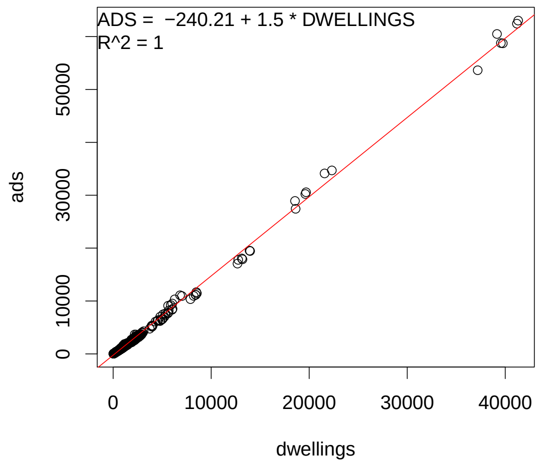
[18] We keep a relatively large range because the hedonic regression is limited to a small set of housing unit characteristics, those less affected by missing data issues. In this step we impute missing characteristics for each housing unit using the approach proposed by Honaker, King and Blackwell (`https://gking.harvard.edu/amelia`).

We will compare the information coming from the datasets along two different dimensions. Firstly, we look at the levels and growth rates for stocks of dwellings for sale, potential transactions and asking prices. Secondly, we compute these statistics at different levels of geographical aggregation: city level and OMI micro-zones.

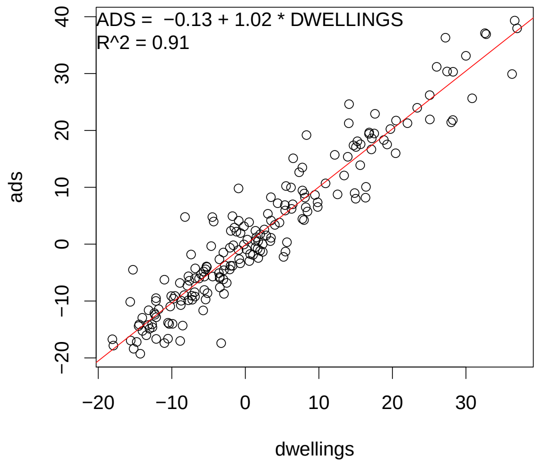
As first comparison, we compute the stock of dwellings for sale for each quarter in each city between 2016Q1 and 2017Q2. As can be seen from Figure 2(a), at city level there is a perfect correlation between the stocks computed over the two datasets, although the number of ads is on average 1.5 times the effective number of dwellings. We find a good correlation between the two datasets also comparing the year-on-year growth rates of the stock of houses for sale in each city in 2017Q1 and 2017Q2 (Figure 2(b)). Similar insights derive from the comparison of the number of potential transactions and asking prices. Looking to the levels of transactions and prices, the two datasets provide similar information about the heterogeneity between cities and quarters (Figures 2(c) and 2(e)). The correlation is weaker when we look at year-on-year growth rates, but the coefficient of determination is still above 0.6 (Figures 2(d) and 2(f)).

The overall picture is somewhat different when we compute the same statistics at a finer geographical level, namely OMI micro-zones. As can be seen from Figures 3(a), 3(c) and 3(e), when we look at stocks, potential transactions and asking prices in levels the two datasets are equally informative. However, the correlation between year-on-year growth rates proves to be relatively low for all the computed variables: in many cases the two datasets provide opposite indications about the evolution of the variable of interest compared to one year before (Figures 3(b), 3(d) and 3(f)).

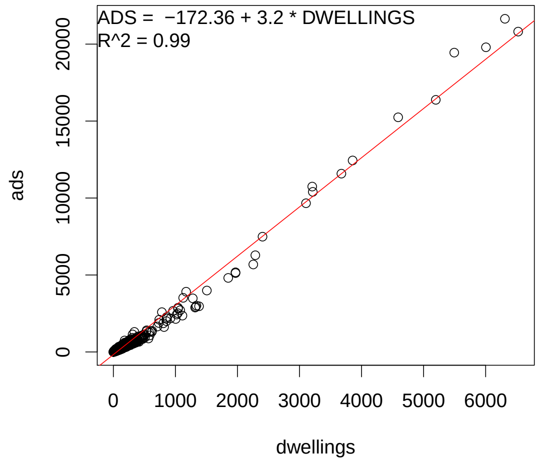
All in all, we conclude that the distortion implied by keeping duplicated ads in the sample is relevant mainly when we look at the short-term dynamics of the housing market and, obviously, when we look at small geographical aggregates. If a researcher is interested only in the heterogeneity across cities there seems no need to run a deduplication process, as original ads provide broadly the same information. The same applies if the analysis should be made at a sufficiently aggregate level. This is good news, since in some cases the analysis of the evolution of the housing market does not require to perform in advance the time-consuming deduplication process.



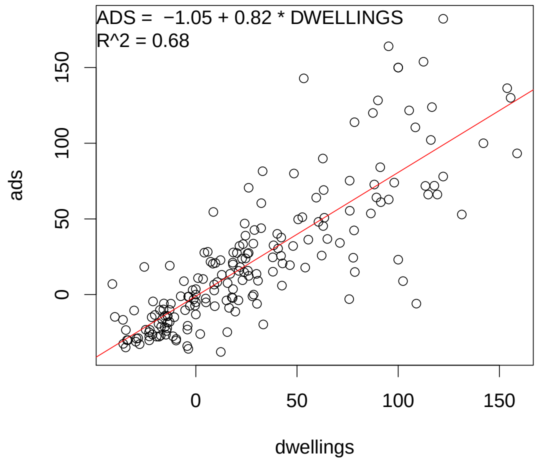

(a) Stocks - Levels



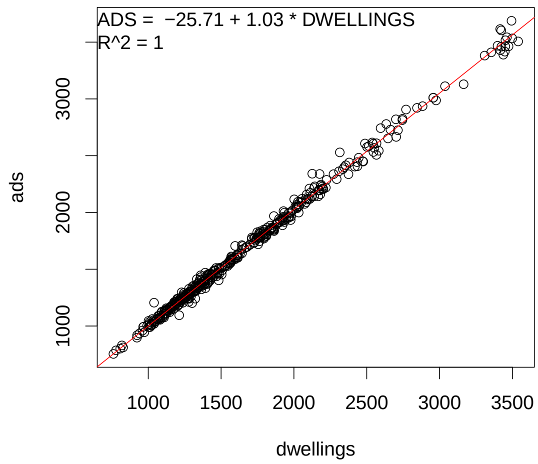

(b) Stocks - Growth rates



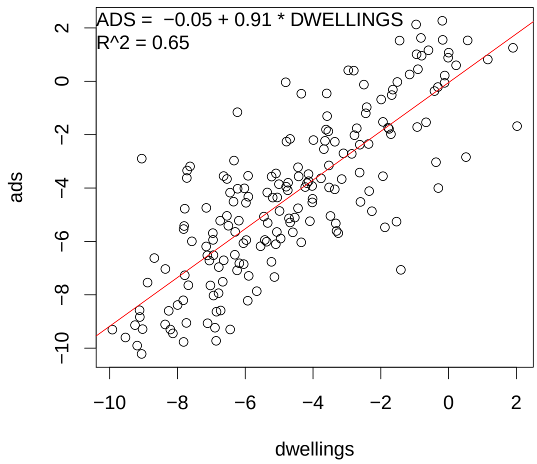

(c) Transactions - Levels




(d) Transactions - Growth rates

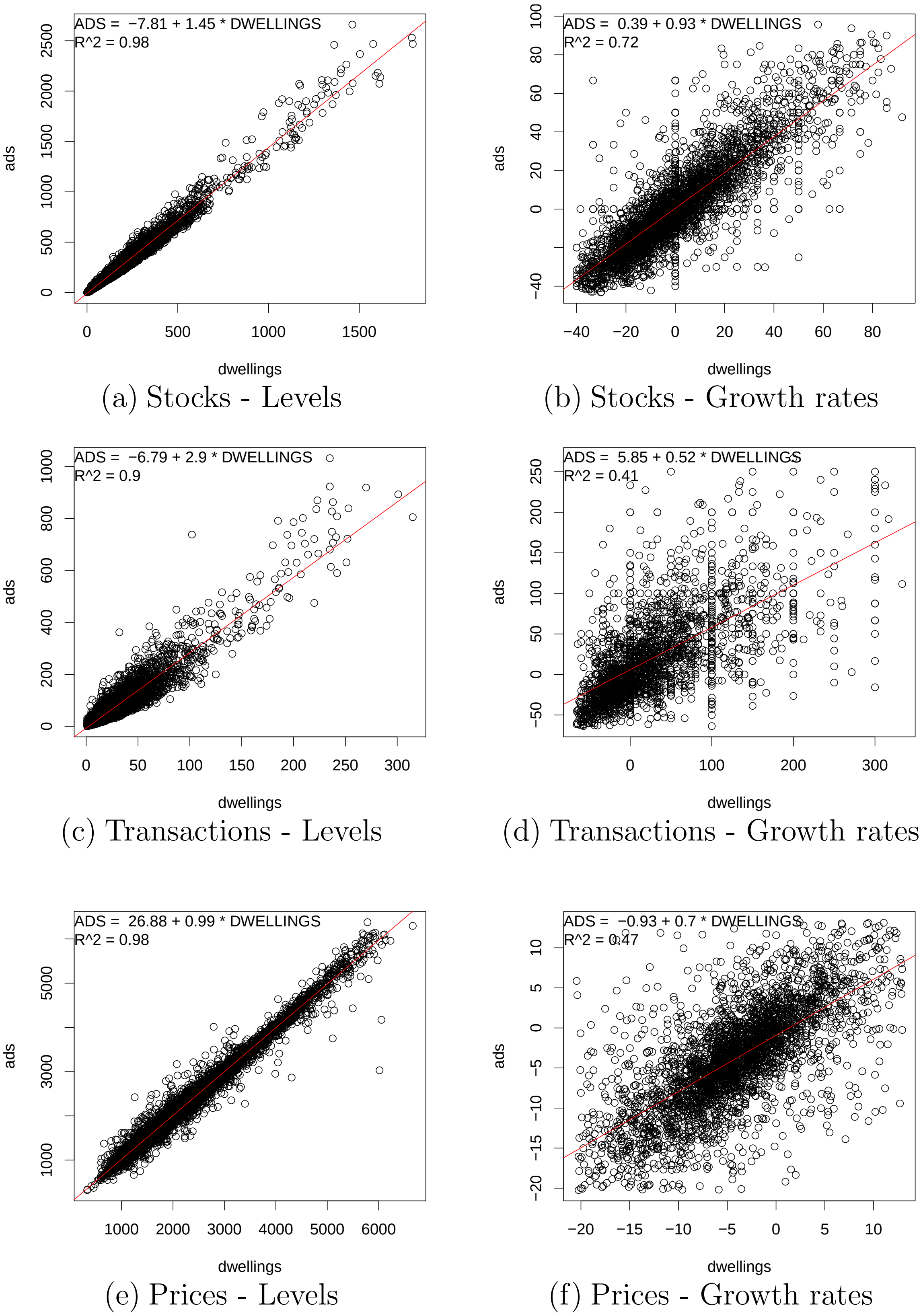



(e) Prices - Levels




(f) Prices - Growth rates

Figure 2: Comparison between original ads and final housing units datasets. Each data point is obtained aggregating information over a city in a specific quarter. Growth rates refer to y-o-y changes in 2016-2017-Q1 and 2016-2017-Q2.




(a) Stocks - Levels

(b) Stocks - Growth rates

(c) Transactions - Levels

(d) Transactions - Growth rates

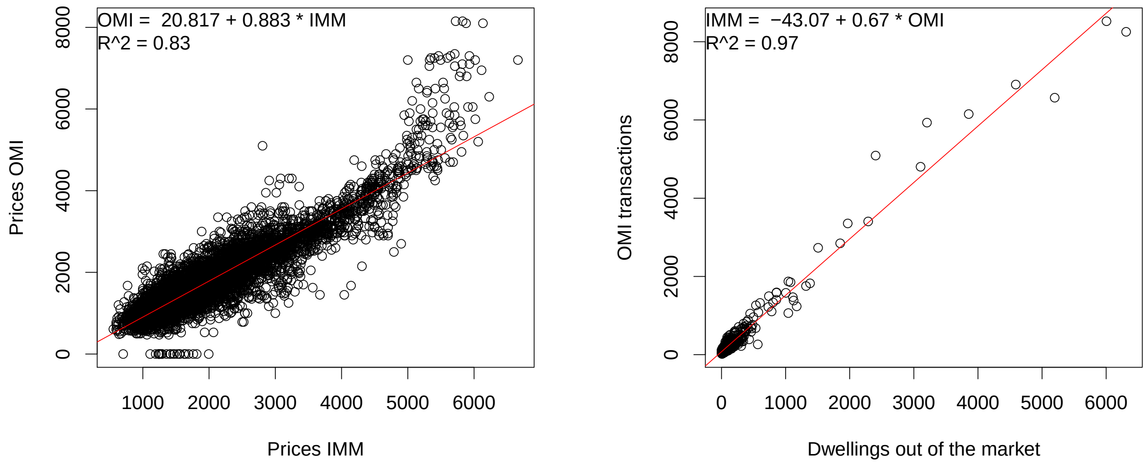
(e) Prices - Levels

(f) Prices - Growth rates

Figure 3: Comparison between original ads and final housing units datasets. Each data point is obtained aggregating information over an OMI micro-zone in a specific quarter. Growth rates refer to y-o-y changes in 2016-2017-Q1 and 2016-2017-Q2.

# 4 Validation

The descriptive statistics of the dataset of housing units are presented in Appendix B. Here we assess the quality of the deduplication process by checking if the information coming from the dataset is coherent with other well established sources of statistics for the Italian real estate market.




(a) Prices (b) Transactions

Figure 4: Prices and transactions. Comparison of final housing units dataset and OMI data. Each data point is obtained aggregating information over an OMI micro-zone in a specific semester.

Our first check is between the quarterly data on housing transactions for each city disseminated by OMI and the number of dwellings going out of the market in the same quarter in each city of our sample. Looking to Figure 4(b) it can be seen that the two statistics show a quite good correlation, as for the original dataset of ads. However, as can be seen from the slope of the regression line and from Table 6, the number of dwellings going out of the market in our dataset is lower than the official sales.

This is more reasonable than sales from our sample being larger than the official sales, as was the case in Section 3.1. First of all, only a fraction of housing transactions is brokered by real estate agents,[19] who are the main users of Immobiliare.it. Secondly, in our processing of the dataset we drop several ads that can refer to actual transactions included in the OMI data (ads with no price, foreclosure auctions, ads without geographical coordinates, etc.).

| City | IMM | OMI | Coverage |
|---|---|---|---|
| Turin | 7615 | 12322 | 61.8 |
| Genoa | 4816 | 6601 | 73.0 |
| Milan | 12570 | 21909 | 57.4 |
| Bologna | 2866 | 5507 | 52.0 |
| Florence | 4122 | 4786 | 86.1 |
| Rome | 22103 | 30173 | 73.3 |
| Naples | 3832 | 6650 | 57.6 |
| Palermo | 2427 | 4718 | 51.4 |

Table 6: Transactions. Comparison between the housing units dataset and OMI data.

[19] According to the Italian Housing Market Survey, in Italy as a whole the share of housing transaction brokered by real estate agencies is about 50%. This share is most likely higher in the provincial capitals, as in small and rural areas there is less need of the brokerage services provided by real estate agents, making our estimates quite plausible.

Looking to prices, in Figure 4(a) we show that the correlation with OMI prices in each micro-zone is also quite good. We compare the average asking price on Immobiliare.it in each OMI micro-zone and for each semester with the relative mean of the estimates of minimal and maximal house values produced by OMI. The two sources of information are coherent: the adjusted $R^2$ of the regression of OMI prices on the average asking prices coming from our dataset is equal to 0.83.

The slope is instead 0.88. This coefficient indicates an average discount on asking prices of about 12%, that is coherent with the evidence provided by the quarterly Italian Housing Market Survey.

The third comparison we make is related to the time on market, as computed from our dataset and as taken from the quarterly Italian Housing Market Survey. The results are shown in Table 7. Up to 2016Q3 it seems that it is possible to build a reliable measure of the time on market, as we almost replicate the results of the Italian Housing Market Survey. However, we fail to catch the downward trend started in the last quarter of 2016. Unfortunately, this result is quite robust to different estimation strategies and does not seem related to the deduplication process, as the same dynamics can be retrieved looking to the original ads (see Table 3).

Moreover, this result is at odds with our finding on market liquidity (see below): in the same period liquidity has been increasing.[20] We believe that the time series are still too short to draw conclusions about time on market statistics and so we leave this as an open issue. We should also remember that absence of information about the actual sale of a dwelling is plausibly more harmful for the estimation of time on market as compared to other statistics.

| Year | Quarter | IMM | Survey BI |
|---|---|---|---|
| 2016 | 1 | 7.3 | 7.5 |
| 2016 | 2 | 7.1 | 7.7 |
| 2016 | 3 | 7.7 | 7.9 |
| 2016 | 4 | 7.8 | 6.8 |
| 2017 | 1 | 8.2 | 6.4 |
| 2017 | 2 | 7.9 | 6.4 |

Table 7: Time on market. Comparison between the housing units dataset and the Italian Housing Market Survey, conducted by the Bank of Italy.

Our fourth comparison regards the price revisions. If unsuccessful in selling their property, sellers may decide to lower the asking price. Occasionally, sellers may choose instead to increase the asking price, either if the ad triggers an auction or if the sale conditions change (e.g. if they decide to sell the garage together with the apartment).

Table 8 shows the relative price difference after first, second and third price revisions. Only 12% of the first price revisions are positive, and the mean first price revision is -6.3%. Mean second and third price revisions are smaller, respectively -4.1% and -3.3%. Comparable evidence is reported in Merlo and Ortalo-Magne (2004), who analyze a hand-collected dataset of housing transactions in England. The authors show that the mean first and second price revisions are -5.3% and -4.4% respectively, in line with our findings.

| Variable | N | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|---|
| First price revision | 164385 | -0.304 | -0.098 | -0.062 | -0.063 | -0.035 | 0.314 |
| Second price revision | 59697 | -0.253 | -0.079 | -0.048 | -0.041 | -0.020 | 0.322 |
| Third price revision | 22604 | -0.241 | -0.072 | -0.041 | -0.033 | -0.012 | 0.296 |

Table 8: Relative price difference between subsequent price revisions.

[20] In principle there can be economic explanations that justify the increase of the time on market in a context of improving housing market conditions. For example, this can be transitory and due to the fact that also dwellings that have been for a long time on the market have been finally sold. However, we believe that before to investigate the several hypotheses that can justify this fact more data are needed.

Finally, we discuss some stylized facts about housing supply in Italy, focusing not on the stock of existing housing units, but rather on the amount of dwellings for sale.[21] In Table 9 we show the evolution of the number of dwellings for sale starting from 2016Q1 up to 2017Q2. This is reasonably only a fraction of the total number of dwellings for sale, as for transactions. However, if we rule out the possibility of structural breaks in the coverage of our sample, these figures should provide trustful indications about the evolution of supply. Indeed, in the first semester of 2017 the number of dwellings for sale (third column) were by 4.4% lower compared to the same period of the previous year. This evolution is consistent with evidence coming from other sources: also according to the Italian Housing Market Survey the dynamics of housing supply have been subdued (see Banca d'Italia (2017)).

| Period | Sales | For sale | Liquidity | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | All cities | To | Ge | Mi | Bo | Fi | Rm | Na | Pa |
| 2016Q1 | 24655 | 205771 | 12.0 | 13.2 | 13.2 | 13.9 | 13.8 | 14.9 | 12.6 | 14.1 | 11.5 |
| 2016Q2 | 31618 | 208413 | 15.2 | 16.4 | 16.2 | 17.9 | 16.9 | 16.5 | 15.3 | 17.9 | 15.0 |
| 2016Q3 | 22647 | 187864 | 12.1 | 11.9 | 12.7 | 12.9 | 11.9 | 13.3 | 12.3 | 16.4 | 9.3 |
| 2016Q4 | 29689 | 199600 | 14.9 | 15.0 | 16.0 | 16.3 | 15.4 | 18.7 | 15.2 | 18.4 | 11.9 |
| 2017Q1 | 33260 | 199898 | 16.6 | 17.1 | 15.9 | 18.7 | 20.0 | 21.2 | 16.4 | 18.6 | 16.5 |
| 2017Q2 | 28272 | 195771 | 14.4 | 15.4 | 15.7 | 17.3 | 16.5 | 17.2 | 14.0 | 18.1 | 12.9 |

Table 9: Transactions, supply and liquidity.

In Table 9 (second column) we also report the number of potential sales (i.e., housing units removed from the dataset) aggregated over all the provincial capitals. This quantity increased by 9.3% in the first semester of 2017, as compared to the same period of 2016 (according to OMI the variation of the actual transactions over the same period was 5.3%). We use these estimates to assess the liquidity of the housing market in the main Italian cities.[22] Liquidity is computed as the percentage ratio of sales over the stock of housing units for sale in a given period. In the full sample we observe that in the first half of 2017 the liquidity of the market has been on average higher than in the same period of the previous year. Moreover, the liquidity has been quite heterogeneous across cities. For example, since early 2016 it has been higher in Milan as compared to Rome or Turin.

# 5 Applications

In this section we present a number of concrete applications that highlight the potential of this dataset.

First, we explore the spatial heterogeneity of the housing market. Second, we perform accurate hedonic regressions by taking into account all physical characteristics of the housing units. Third, we analyze the evolution of the housing market. Among other things, we show how to now-cast the aggregated house price level and anticipate by several months a house price index constructed from administrative data. Fourth, we provide first evidence at the Italian

[21]In the economic literature it is standard to define housing supply as the total number of dwellings, independently if they are on sale or are currently inhabited (see for example Glaeser and Gyourko (2017) and for the Italian case Loberto and Zollino (2016)). As a consequence, variation in housing supply is represented by new constructions and is generally non-negative because of the durable nature of dwellings. Depending on the issue at stake, this definition is not necessarily the most suitable, especially if we are interested in the short-run effects of housing supply on the housing market. At the opposite, in this paper we define as housing supply only houses on sale. We believe it is fair to say that this distinction is the same that arises in labour market economics, in which only people that are already working or searching actively for a job are considered inside the labour supply.

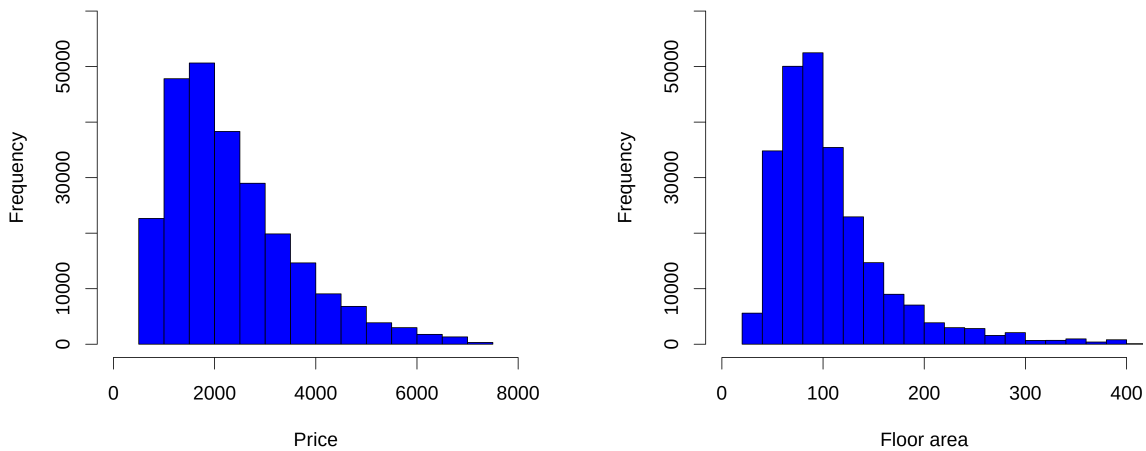
[22]In principle we could use the number of actual sales provided by OMI, but this would not allow a comparison of the liquidity of the market in different cities, as the coverage of the housing market by our data is not homogeneous across cities.

level that online interest for a particular area is a leading indicator of prices. Finally, we test a prediction of search theory, finding no significant support.

The common denominator of these exercises is that they would not be possible with any other currently available public data source on the Italian real-estate market.

### 5.1 Heterogeneity

Heterogeneity is a key property of the housing market. For example, certain segments of the market may be disproportionally affected by evolving credit conditions (Landvoigt et al., 2015). Heterogeneity occurs between and within cities, but also between and within neighborhoods.




(a) Price per m2 (b) Floor area

Figure 5: Heterogeneity in the distributions of price per m2 and floor area.

Figures 5(a) and 5(b) show the distribution of asking prices per m2 and floor area respectively. Both distributions are skewed, with heavy right tails, indicating the existence of housing units with extremely high values. We represent the price per m2 in spatial form in Figure 6, where we focus on the cities on Rome and Milan. In order to smooth the spatial distribution and mitigate the problem of outliers, we plot a kernel approximation of the prices.

An important difference between the distributions in the two cities is that in the case of Milan the prices decline radially from the center, whereas in the case of Rome we observe hotspots of high prices in peripheral neighborhoods (Appia Antica and Eur). Moreover, in Rome the prices do not decline radially from the center, because prices north of the center are larger than prices south of the center. This difference can be traced back to historical, infrastructural and geographical reasons.

The levels of the prices are similar in Rome and Milan, and the prices are among the highest within Italian cities. In Appendix E we show similar maps for eight other major cities, namely Turin, Naples, Genoa, Palermo, Venice, Florence, Bari and Bologna. The trends are similar, with high heterogeneity within and between cities. The cheapest city is Palermo, with prices ranging from 611 to 3242 euros per m2, while the most expensive is Milan, whose price range is 1600-9200 euros per m2.

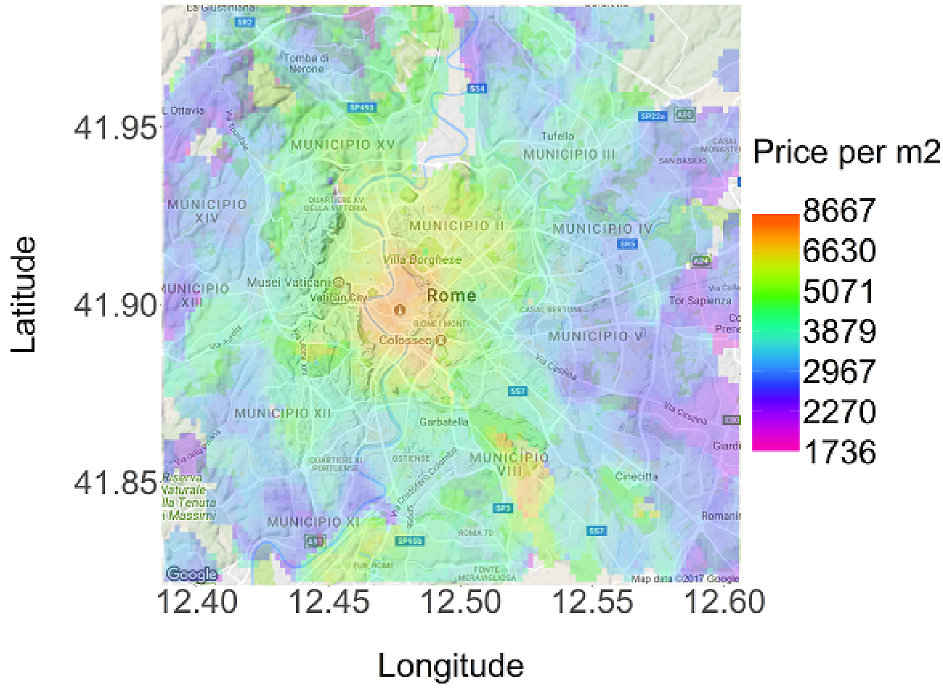
In Figure 7 we plot other variables. Instead of plotting a kernel approximation of their values, we aggregate these quantities over OMI micro-zones and color the OMI polygons according to the quartiles of the distribution. Figure 7(a) represents the median number of clicks on housing units, which are a proxy of demand. Comparing to Figure 6(a), we see that demand is highly correlated to price per m2, probably because both are correlated to an intrinsic attractiveness of the neighborhoods. There are some exceptions though. Consider the OMI micro-zone in the center of Rome in Figure 7(a), as indicated by the black arrow. This area includes some of the most famous touristic attractions in Rome, yet the number of clicks is below median.

In Figure 7(b) we look at the relative supply, namely the ratio between the number of ads in the OMI micro-zones and the stock of dwellings, as obtained from 2011 Census data.[23] The relative supply looks larger in the north of Rome, but the correlation with other variables is less clear. Interestingly, the same central OMI micro-zone in which demand was low has a comparatively large relative supply.

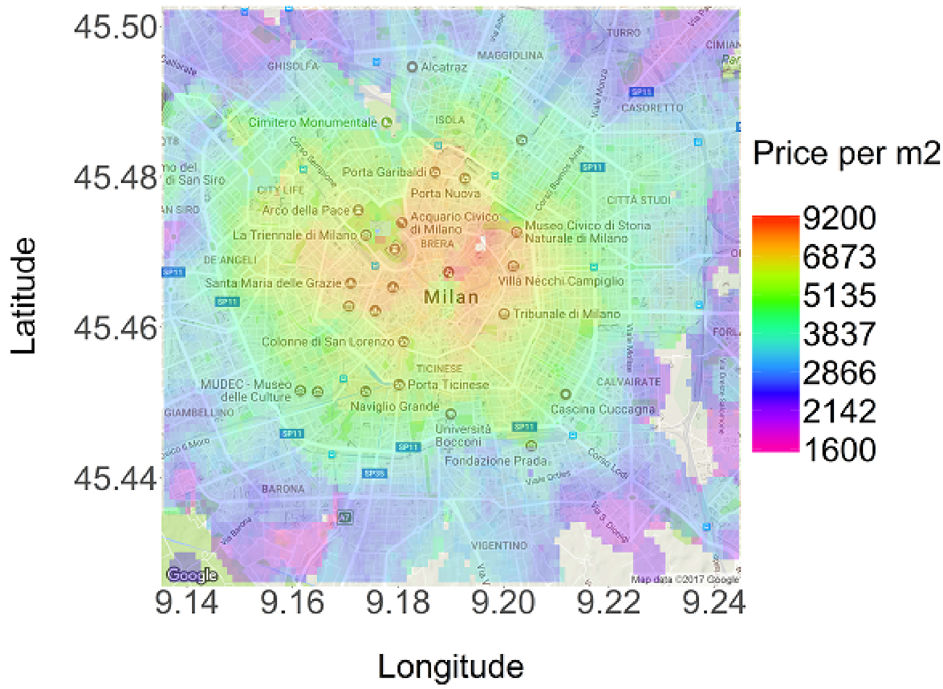
Figure 7(c) shows instead the median floor area. In this case there is again a strong correlation with the price per m2, indicating that the total price of the most expensive dwellings is much larger than the prices of the other houses. This suggests a possible explanation for the low interest towards the central OMI micro-zone which we mentioned previously. It could just be that dwellings located there are too expensive and few buyers can afford them, hence the high relative supply too. Finally, in Figure 7(d) we report the average maintenance status.[24] While the maintenance status is quite good in the center, the highest values can be found in the peripheries, because most dwellings on sale in those areas are new.

[23]Istat census tracts are much smaller than OMI micro-zones (indeed, there are approximately 400,000 Istat census tracts over the Italian territory, as compared to 27,000 OMI micro-zones) and do not necessarily coincide with them. We perform spatial matching of the polygons representing the tracts and the micro-zones and impute the Istat variables to the OMI micro-zones according to the overlap percentage of the polygons. For example, if an Istat census tract comprises 2,000 housing units and it straddles two OMI micro-zones, such that there is a 50% overlap for both, we impute 1,000 housing units to each of the two OMI micro-zones.

[24]Maintenance status is a categorical ordinal variable, so we transform it into a numerical variable with the conversion reported in Table 17 (Appendix C).

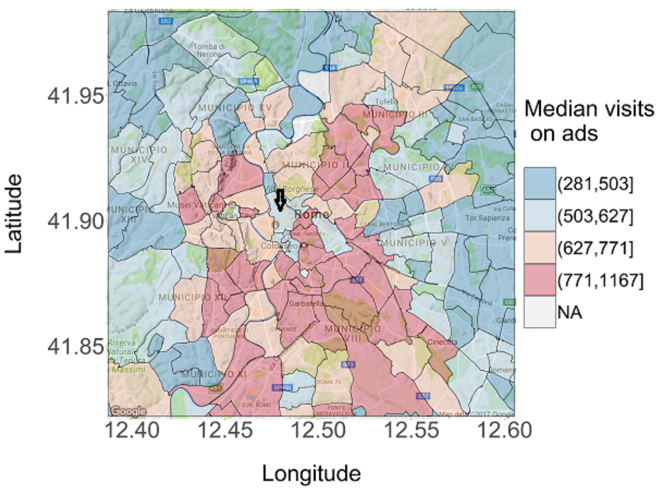



(a) Rome

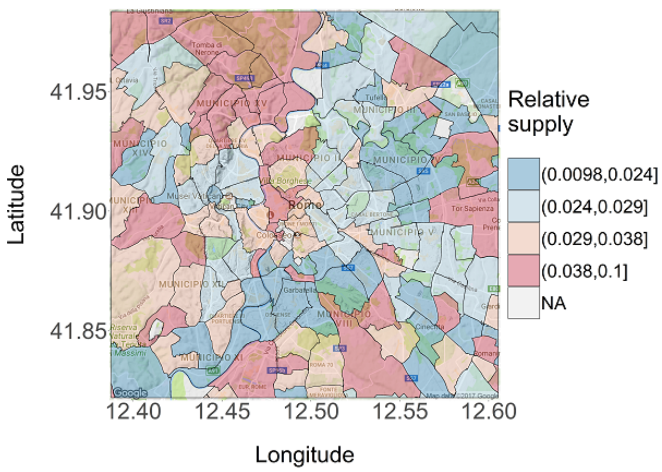



(b) Milan

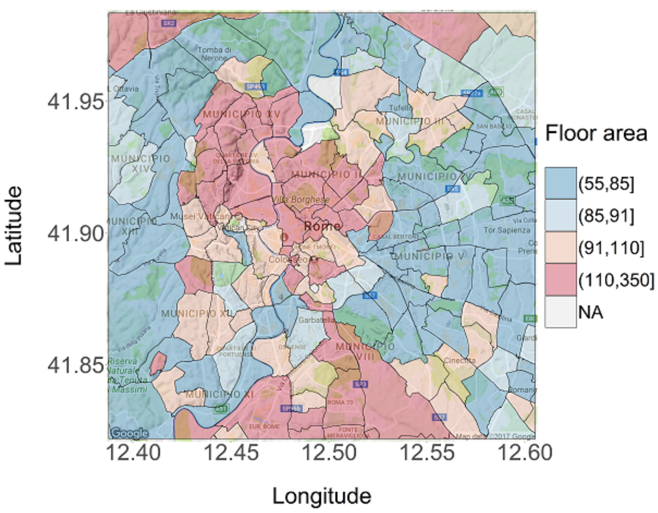
Figure 6: Kernel approximation of the (asking) price per m2 during 2017Q1.



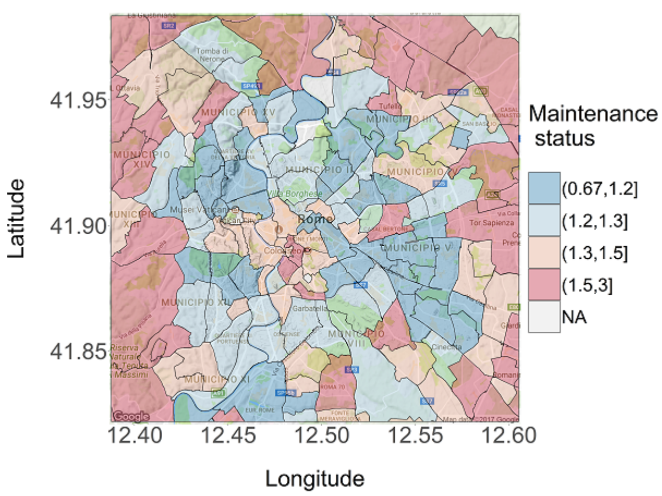

(a) Median demand




(b) Relative supply




(c) Floor area




(d) Maintenance status

Figure 7: Spatial distribution of selected variables. Data are aggregated over OMI micro-zones, represented by the polygons in the maps. We plot the center of Rome during 2017Q1.

### 5.2 Hedonic regressions

In Table 10 we show the results of the regression of the price per m2 on the physical characteristics of the housing units. The sample size shrinks to 174,143 housing units, because here we make no imputation for missing variables and want to have the largest set of characteristics in order to understand their contribution to the price of the dwelling. Moreover, most missing values are due to variables introduced from a relatively shorter period in the dataset, such as "energy class" (since 2016 it shows rate of missingness comparable to other variables), or to variables that can be discarded with minor impact on the fit of the regression, such as air conditioning or heating type.[25] The regressions are run for the whole dataset (comprising all 110 provincial capitals) and for the cities of Rome and Milan, with OMI micro-zone and quarter dummies (in order to control for geolocation and common trends, respectively).[26]

First of all, the coefficients are similar in the three cases. The coefficients always have the same sign (except for one case) and are of the same order of magnitude. For some variables, such as air conditioning, the coefficients are also quantitatively similar, while this is not true for other variables such as number of bathrooms. The value of most coefficients is expected, including the negative coefficient on the floor area (larger apartments are on average cheaper per m2). The unexpected coefficients are those on number of rooms, type of kitchen and utility room.

Indeed, we find that housing units with few rooms, a kitchenette and without utility room are more expensive per m2 than dwellings with many rooms, an eat-in kitchen and with one or more utility rooms. Since we are already controlling for floor area, a possible explanation we can propose is that buyers prefer less fragmented housing units.[27]

Finally, we believe is interesting to highlight that the energy performance of the dwelling has a significant impact on prices, also controlling for all the other characteristics (such as the maintenance status).

In addition to the identification of the contribution of the physical characteristics of the dwelling to house prices, the hedonic regression is a tool to control for composition effects when assessing the evolution of house prices. Indeed, dwellings are heterogeneous goods and the composition of housing supply or transactions can change qualitatively from period to period. This volatility is even greater at finer geographical areas. However, when we compute a house price index we would like to look in each period to the price of exactly the same pool of dwellings. Since this is of course not feasible in the real world, the linear hedonic model is a tool that allows to control for the different characteristics of the dwellings taken in consideration.

[25] In this section we consider all variables listed in Table 10. In the following sections, when we need to control for the physical characteristics of the housing units we drop the variables with most missing values, so to increase the sample size.

[26] Quarter dummies take value 1 if the housing unit was visible on Immobiliare.it during the quarter.

[27] There are also other possible explanations. For example, this can reflect preferences for houses where the living area includes the kitchen. It can be also be correlated with lower fertility: as families are smaller in size there is no need to have a plenty of rooms.

Table 10: Hedonic regression

| | *Dependent variable:* | | |
|---|---|---|---|
| | Price per m2 | | |
| | Italy | Rome | Milan |
| | (1) | (2) | (3) |
| Number of bathrooms | 154.577*** (3.743) | 72.178*** (8.034) | 208.084*** (12.031) |
| Floor of the apartment | 32.363*** (0.926) | 51.505*** (2.077) | 48.876*** (2.338) |
| Floor area [m2] | −1.281*** (0.048) | −2.811*** (0.102) | −0.141 (0.174) |
| Number of rooms | −22.093*** (2.103) | −48.384*** (4.904) | −21.860*** (8.086) |
| Air conditioning | 106.661*** (3.499) | 126.300*** (7.545) | 122.863*** (10.211) |
| Balcony | −30.398*** (3.693) | 11.483 (7.912) | 0.764 (10.935) |
| Elevator | 168.574*** (3.969) | 144.780*** (9.708) | 223.752*** (12.614) |
| Energy class; Ref: EG; Level: AB | 360.102*** (7.560) | 267.614*** (18.907) | 393.328*** (21.635) |
| Energy class; Ref: EG; Level: CD | 114.052*** (5.742) | 92.140*** (20.624) | 135.590*** (16.620) |
| Garage; Ref: No; Level: Single | 120.817*** (5.838) | 127.548*** (11.158) | 108.439*** (27.483) |
| Garage; Ref: No; Level: Double | 205.992*** (4.330) | 319.284*** (10.700) | 149.507*** (12.491) |
| Heating type; Ref: No; Level: Centralized | 47.737*** (10.439) | 156.469*** (29.439) | 225.680** (88.304) |
| Heating type; Ref: No; Level: Autonomous | 135.250*** (9.878) | 222.994*** (29.050) | 281.075*** (88.417) |
| Kitchen type; Ref: Kitchenette; Level: Small eat-in kitchen | −49.458*** (5.350) | −23.742** (11.736) | −89.975*** (14.268) |
| Kitchen type; Ref: Kitchenette; Level: Large eat-in kitchen | −77.403*** (4.477) | −86.146*** (10.131) | −57.177*** (12.522) |
| Status; Ref: To renovate; Level: Good | 192.238*** (5.661) | 164.210*** (11.707) | 186.031*** (16.280) |
| Status; Ref: To renovate; Level: Excellent | 480.591*** (5.874) | 462.987*** (12.252) | 535.513*** (16.606) |
| Status; Ref: To renovate; Level: New | 610.984*** (8.808) | 479.285*** (21.080) | 584.823*** (25.053) |
| Terrace | 143.539*** (3.761) | 229.414*** (8.255) | 217.474*** (12.834) |
| Utility room | −56.913*** (3.516) | −84.680*** (8.234) | −77.461*** (10.396) |
| Basement | 11.039*** (3.660) | 57.360*** (8.388) | 24.695** (9.814) |
| Porter | 99.019*** (5.557) | 73.887*** (10.685) | 62.285*** (10.476) |
| Constant | 718.810*** (241.357) | 2,026.932*** (724.520) | 457.625 (749.341) |
| Observations | 174,143 | 43,945 | 27,212 |
| $R^2$ | 0.783 | 0.675 | 0.731 |
| Adjusted $R^2$ | 0.781 | 0.673 | 0.730 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
OMI micro-zone and quarter dummies

### 5.3 Evolution of prices

In this section we analyze the dynamics of the Italian housing market from the second semester of 2015 to the first semester of 2017.

We first construct an aggregate house price index based on our dataset. To this end, for each semester we aggregate the average asking price in each city, using as weights the stock of dwellings in each city as taken from the 2011 Census. We also consider a house price index constructed according to the methodology proposed in Cannari and Faiella (2007) (CF hereafter), based on OMI and *Il Consulente Immobiliare* data, using the same weights of the asking prices index.[28] In order to compare the two indexes more meaningfully, we obtain the average discount on asking prices from the Italian Housing Market Survey.

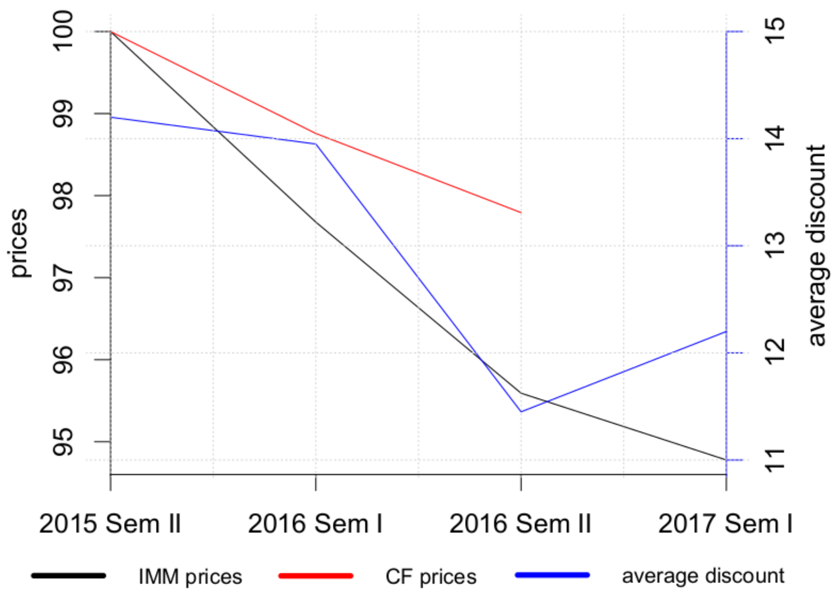



Figure 8: Evolution of prices and average discount from 2015S2 to 2017S1. Prices are compared to the reference level in 2015S2 (=100 on the left axis). We compare the transaction prices estimated using the Cannari-Faiella methodology (Cannari and Faiella, 2007) and the asking prices obtained from our dataset. The discrepancy between asking and transaction prices is consistent with the variation of the average discount, whose scale is shown on the right axis.

In Figure 8 we show the evolution of the two indexes between 2015S2 and 2017S1. CF-prices have declined by about 2% up to 2016S2, while according to our dataset asking prices have declined by 4.5% in the same period. This dynamics is coherent with the reduction of the average discount on asking prices, cumulatively equal to 3 percentage points. Indeed, a lower discount implies that transaction prices decreased less than asking prices.

Our estimates extend to 2017S1, whereas at the time of writing this manuscript the CF-prices are not available for the same semester. Therefore, the use of online data coming from Immobiliare.it can be regarded as a simple now-casting exercise. Based on our estimates for asking prices and the evolution of average discount according to the Italian Housing Market Survey, in 2017S1 house prices should have continued to decline, to a greater extent than asking prices.

Thanks to the rich set of characteristics that are available, we can build quality adjusted house price indexes or we can look at the evolution of average prices in specific market segments.

[28] *Il Consulente Immobiliare* (CI) is an industry-related review published by Il Sole 24 Ore media group that collects information on actual sales from real-estate agents in more than 1,000 Italian municipalities. CI estimates are combined at city level with those of OMI by a weighted average. The details of the weighting scheme are explained in (Cannari and Faiella, 2007).

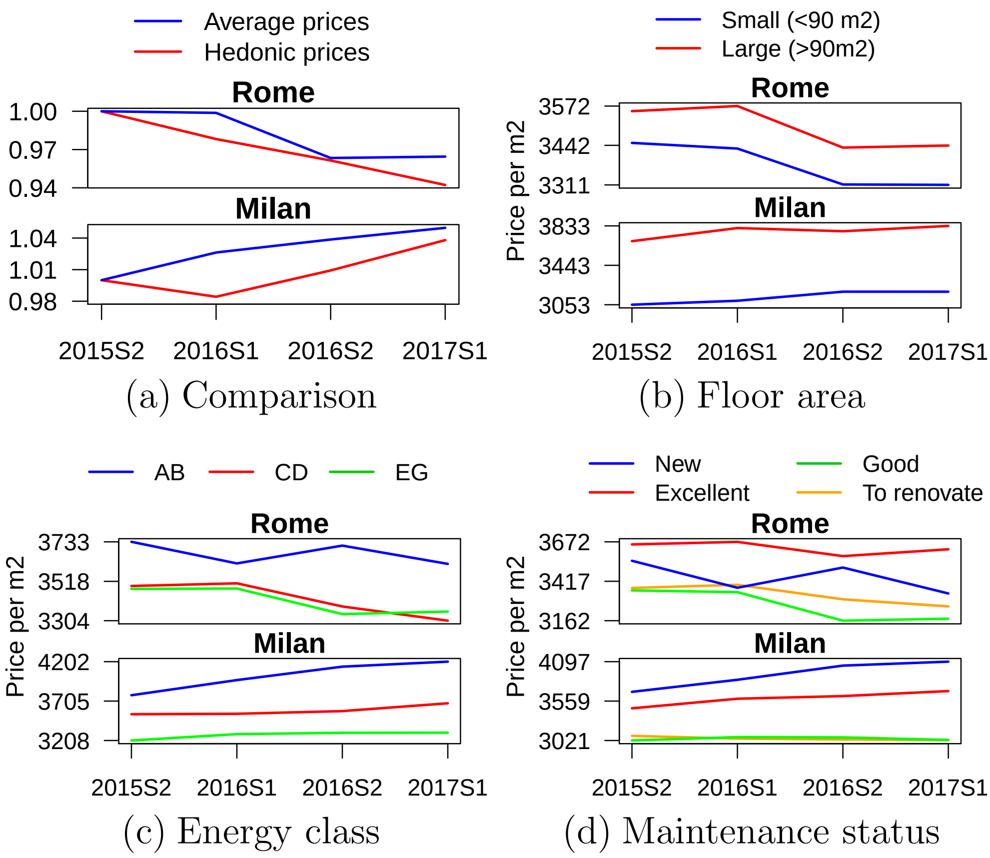



Figure 9: Comparison of average and hedonic prices and evolution of the average prices across several market segments. In the top-left panel the vertical axis shows values of the prices with respect to the reference level in 2015S1 (= 1.00), in the other panels we plot the price levels.

Focusing on the cities of Rome and Milan, in Figure 9(a) we compare the trends of average prices and hedonic prices. The hedonic prices are calculated from the regressions described in Section 5.2, implementing a time dummy approach. In particular, we consider the intercept of the regression as the reference value in 2015S2 and calculate the percentage variation using the coefficients on the semester dummies in 2016S1-S2 and 2017S1.[29] Average and hedonic prices decreased in Rome but increased in Milan during this time period. Hedonic prices were on average lower. This suggests that the quality of some physical characteristics of the housing units improved.

In Figures 9(b), 9(c) and 9(d) we plot the average price per m2, as disaggregated by floor area, energy class and maintenance status respectively. We observe that the price dynamics is generally very similar in Rome and Milan across these market segments. Small apartments are more expensive (per m2) than large apartments, but this result is not in contradiction with the negative coefficient on floor area in the hedonic regression in Table 10. When performing hedonic regressions we control for all variables simultaneously, whereas in Figure 9 we do not. Most likely, large apartments are located in the center or in the most expensive areas (Figure 7(c)), and this explains why prices (per m2) are higher. Interestingly, new dwellings are more expensive than apartments with excellent maintenance status in Milan, but the reverse is true

[29] Because here we want to associate each housing unit with one and only one semester dummy, we only consider the dwellings that went out of the market and associate them to a semester depending on the removal date. The sample size is 18245 housing units in Rome and 12896 in Milan.

in Rome. This is also consistent with Figure 7(d), as most new housing units in Rome are located in the peripheries.

| Year | 5th | 15th | 25th | 50th | 75th | 85th | 95th | Mean |
|---|---|---|---|---|---|---|---|---|
| | | | | Full sample | | | | |
| 2016 | 915 | 1235 | 1500 | 2167 | 3125 | 3778 | 5143 | 2487 |
| 2017 | 857 | 1161 | 1417 | 2067 | 3000 | 3672 | 5053 | 2398 |
| y-o-y variation | -6.3 | -6.0 | -5.6 | -4.6 | -4.0 | -2.8 | -1.8 | -3.6 |
| | | | | Milan | | | | |
| 2016 | 1667 | 2083 | 2409 | 3200 | 4400 | 5182 | 6861 | 3588 |
| 2017 | 1635 | 2067 | 2388 | 3192 | 4474 | 5333 | 6957 | 3617 |
| y-o-y variation | -1.9 | -0.8 | -0.9 | -0.3 | 1.7 | 2.9 | 1.4 | 0.8 |
| | | | | Turin | | | | |
| 2016 | 863 | 1107 | 1311 | 1750 | 2316 | 2684 | 3500 | 1900 |
| 2017 | 824 | 1060 | 1253 | 1700 | 2256 | 2659 | 3500 | 1860 |
| y-o-y variation | -4.6 | -4.2 | -4.4 | -2.9 | -2.6 | -0.9 | 0.0 | -2.1 |
| | | | | Rome | | | | |
| 2016 | 1731 | 2222 | 2548 | 3278 | 4231 | 4900 | 6316 | 3557 |
| 2017 | 1627 | 2106 | 2438 | 3173 | 4133 | 4780 | 6212 | 3448 |
| y-o-y variation | -6.0 | -5.2 | -4.4 | -3.2 | -2.3 | -2.4 | -1.6 | -3.1 |
| | | | | Naples | | | | |
| 2016 | 1154 | 1532 | 1833 | 2667 | 3875 | 4600 | 6047 | 3007 |
| 2017 | 1078 | 1435 | 1722 | 2544 | 3750 | 4465 | 5900 | 2894 |
| y-o-y variation | -6.6 | -6.3 | -6.1 | -4.6 | -3.2 | -2.9 | -2.4 | -3.8 |

Table 11: Evolution of the asking prices per square meter between 2016S1 and 2017S1 across several quantiles of the price distribution. In the top lines we show the aggregate dynamics, below we consider the four largest Italian cities. The dynamics are similar.

We finally analyze the evolution of the prices across several quantiles of the price distribution. Table 11 shows that between the first half of 2017 and the corresponding period of 2016 the decline of asking prices was stronger in the left tail of the distribution. Indeed, the year-on-year variation is monotonically increasing with the position in the distribution, in the sense that the evolution of prices has been less negative in the upper tail of the distribution (except for Milan, where prices increased in absolute value). Note that the aggregate dynamics is not due to composition effects, as it has been similar across the biggest four Italian cities.

### 5.4 Market tightness as leading indicator of prices

An advantage of using online data for the analysis of the housing market is that they also convey information about the evolution of housing demand and, therefore, can possibly improve the forecasting of housing prices and sales (Carrillo et al., 2015). Wu and Brynjolfsson (2015) show how this goal can be attained using Google search data. Here we follow van Dijk and Francke (2017) and we use the information coming from each ad to build a measure of demand conditions.

We construct a proxy of market tightness by simply considering the number of clicks on housing units within a specific OMI micro-zone, and dividing that number by the total number of housing units for sale in the same micro-zone (in practice, we are considering the average number of clicks per housing unit). We test whether market tightness is a leading indicator of prices by running the regression in Eq. (1).

$$\log(P_{i,t}) = \alpha + \beta_1 \log(D_{i,t-1}) + \beta_2 \log(D_{i,t-2}) + \gamma T + \delta_0 \log(P_{i,t-1}) + \delta X_i + \epsilon_{i,t}, \qquad (1)$$

where $i$ indicates OMI micro-zones, $t$ is a quarter, $P_{i,t}$ is the average price per m2 in zone $i$ at time $t$, $D$ is the market tightness as just defined, $T$ represents quarter dummies, $X_i$ is a vector of OMI micro-zone $i$ characteristics.[30]

The results of this regression are shown in Table 12. The most significant control is the first lag of the asking price (we do not report the coefficients on the other control variables), but we also see that the first and second lags on the tightness are significant, with an elasticity around 4-5%.

Table 12: Tightness predictive power

| | *Dependent variable:* |
|---|---|
| | Price per m2 (t) [log] |
| Price per m2 (t-1) [log] ($\delta_0$) | 0.506*** (0.010) |
| Tightness (t-1) [log] ($\beta_1$) | 0.044*** (0.013) |
| Tightness (t-2) [log] ($\beta_2$) | 0.048*** (0.013) |
| Constant ($\alpha$) | 3.211*** (0.120) |
| Observations | 6,288 |
| $R^2$ | 0.811 |
| Adjusted $R^2$ | 0.807 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
City and quarter dummies. Other controls include OMI area characteristics.
Data are aggregated over OMI areas and quarters, from 2016Q1 to 2017Q2.

## 5.5 Atypicality

As a final application, we test a result from search theory, namely that ceteris paribus atypical housing units sell at a higher price, and take longer to sell (Haurin, 1988).

Haurin et al. (2010) test this prediction in a small dataset purchased from real estate agencies. Their identification strategy is as follows. First, they run hedonic regressions in order to assess the importance of the physical characteristics of the housing units. Second, they construct an atypicality measure for each characteristic by considering the difference between the housing unit characteristic and the average characteristics in the neighbourhood. For instance, if the floor area is 90m2 and the average floor area in the neighborhood is 70m2, the floor area atypicality is 20. Third, the authors aggregate the various measures of atypicality using the coefficients of the hedonic regression as weights. Finally, they regress the price against the atypicality measure (without controlling for any other characteristic), showing that there is a positive significant coefficient on atypicality.

We dispute the validity of this identification strategy. The main problem is the lack of controls. If atypicality was correlated with, e.g., the number of bathrooms, which have an important influence on the price, the results would not be valid. Moreover, it would not be possible to control for the physical characteristics of the housing units because these are correlated with the weights used to construct the atypicality measure, which would then be endogenous.

Therefore, we take a different econometric approach. We construct measures of heterogeneity at the neighborhood (OMI micro-zone) level, and then regress the price of each housing unit on the neighborhood heterogeneity (with controls). The underlying assumption in Haurin (1988) is that buyers have more difficulty assessing the value of an atypical dwelling, and so the variance

[30] Again, we obtain information on these characteristics mainly from the 2011 Census. Included in $X$ are city dummies, fraction of population with a university degree, stock of dwellings, total population, unemployment rate, fraction of owned dwellings (as opposed to rented dwellings) and share of foreigner population.

of the distribution of offers is greater. This should also be true if the housing units in a given neighborhood are highly heterogeneous.

As measures of heterogeneity, we consider the coefficient of variation for price and floor area, and the information entropy for number of rooms and floor. We do not aggregate these measures, because using the hedonic prices as weights would make the measures endogenous. We instead consider the four measures separately as covariates. We do not calculate the heterogeneity from the stock of housing units in the neighborhood, but from the sample of dwellings on sale. As this sample changes over time, we cannot use static values.

To this end, we calculate the four measures of heterogeneity for each weekly snapshot based on an average over the past ten weekly snapshots (so we obtain a smooth evolution of the variables). For each housing unit, we consider the upload date of the first ad and impute the measures of heterogeneity corresponding to the closest weekly snapshot. In the same way, we impute to each housing unit the tightness (defined as in Section 5.4) and average price in the neighborhood.

We then regress the posted asking price and the time on market on the heterogeneity measures. We control for the physical characteristics of the dwelling and for the tightness and average price in the neighborhood, and use OMI micro-zone and quarter dummies.

The results of this regression are shown in Table 13. No measure of heterogeneity is significant, despite the large number of observations.[31]

Table 13: Atypicality

| | *Dependent variable:* | |
|---|---|---|
| | Price per m2 | Time on market |
| Price heterogeneity | 1.133 (18.692) | 4.914 (4.914) |
| Floor heterogeneity | −52.128 (31.700) | −3.206 (8.546) |
| Floor area heterogeneity | −3.998 (23.342) | 8.695 (6.118) |
| Rooms heterogeneity | 4.335 (37.742) | −0.662 (10.305) |
| Constant | 1,662.831*** (462.390) | 147.532* (89.070) |
| Observations | 168,073 | 53,208 |
| $R^2$ | 0.780 | 0.235 |
| Adjusted $R^2$ | 0.778 | 0.211 |

*Note:* *p<0.1; **p<0.05; ***p<0.01
OMI micro-zone and quarter dummies. Other controls include housing unit characteristics, and the tightness and mean price per m2 in the neighborhood (lagged).

# 6 Conclusion

Big data are becoming ubiquitous in business and academia, and increasingly in institutions. There are many reasons for their success: big data aim to cover the universe of entities under consideration (without the need for sampling), provide a lot of information which can be integrated by textual analysis and image processing, if coming from online sources are frequently available (on a much shorter timescale than administrative data) and rely on observations rather than surveys. There are disadvantages too: big data may well fail to provide universal coverage (and so lead to non-representative results), are less structured and controlled (there might be

[31] There are only 53,208 observations for the time on market, because we only consider dwellings that have been removed from the dataset.

hidden factors influencing the generation of the data) and could have other sorts of measurement errors.

This study provides a concrete example of the strengths and weaknesses of big data for institutional applications. We analyze a dataset consisting of more than one million online sales advertisements for residential units posted on the website Immobiliare.it between the beginning of 2015 up to June 2017 in all Italian provincial capitals. This dataset allows to overcome the limitations of existing administrative data. Most importantly, our dataset has information about the physical characteristics of the housing units, previously lacking. We also construct new variables that were previously unavailable. For instance, we construct a proxy of the time on market by counting the days the housing unit on sale has been listed on the website, and a proxy for the demand tightness in a given neighborhood by considering the average number of visits (clicks) on ads located in that neighborhood.

We validate this dataset against official statistical sources and show that it matches core indicators about the Italian housing market. However, some indicators are only matched once we resolve the main problem with the dataset, namely that two or more ads listed at the same time could refer to the same housing unit. We use machine learning techniques – which are effective thanks to the amount of data – to identify the duplicates and construct a final dataset of housing units.

We finally provide a number of potential applications of this dataset. These include the now-casting of aggregate and local temporal price trends in Italy and the detailed study of heterogeneity and segmentation. By correcting for the physical characteristics of the housing units, we construct a quality-adjusted price index for the cities or Rome and Milan, which can potentially be extended at the national level. We also provide first evidence at the Italian level that the number of visits (clicks) on ads located in a given neighborhood is a leading indicator of prices in the same area. The so-constructed demand tightness can potentially be used to predict price trends at the national and local level, and thereby inform policies dealing with the construction and financial industries.

# References

Anenberg, E. and Laufer, S. (2017) "A More Timely House Price Index," Forthcoming in *The Review of Economics and Statistics.*

Banca d'Italia (2017) "Annual Report,"Technical report, Banca d'Italia.

Cannari, L. and Faiella, I. (2007) "House prices and housing wealth in Italy."

Carrillo, P. E., de Wit, E. R., and Larson, W. (2015) "Can Tightness in the Housing Market Help Predict Subsequent Home Price Appreciation? Evidence from the United States and the Netherlands," *Real Estate Economics*, Vol. 43, pp. 609–651, URL: http://dx.doi.org/10.1111/1540-6229.12082, DOI: http://dx.doi.org/10.1111/1540-6229.12082.

Christen, P. (2012) *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*: Springer Publishing Company, Incorporated.

Ciocchetta, F., Cornacchia, W., Felici, R., and Loberto, M. (2016) "Assessing financial stability risks from the real estate market in Italy," Questioni di Economia e Finanza (Occasional Papers) 323, Bank of Italy, Economic Research and International Relations Area.

van Dijk, D. W. and Francke, M. K. (2017) "Internet Search Behavior, Liquidity and Prices in the Housing Market," *Real Estate Economics*, pp. n/a–n/a, URL: http://dx.doi.org/10.1111/1540-6229.12187, DOI: http://dx.doi.org/10.1111/1540-6229.12187.

Glaeser, E. and Gyourko, J. (2017) "The Economic Implications of Housing Supply."

Glaeser, E. L. and Nathanson, C. G. (2015) “Housing Bubbles,” Vol. 5: Elsevier, Chap. Chapter 11, pp. 701–751, URL: http://EconPapers.repec.org/RePEc:eee:regchp:5-701.

Han, L. and Strange, W. (2014) “The microstructure of housing markets: Search, bargaining, and brokerage,” *Handbook of Regional and Urban Economics*, Vol. 5.

Harris, Z. S. (1954) “Distributional structure,” *Word*, Vol. 10, pp. 146–162.

Haurin, D. (1988) “The duration of marketing time of residential housing,” *Real Estate Economics*, Vol. 16, pp. 396–410.

Haurin, D. R., Haurin, J. L., Nadauld, T., and Sanders, A. (2010) “List prices, sale prices and marketing time: an application to us housing markets,” *Real Estate Economics*, Vol. 38, pp. 659–685.

Hsieh, C.-T. and Moretti, E. (2017) “Housing Constraints and Spatial Misallocation.”

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An introduction to statistical learning*, Vol. 112: Springer.

Landvoigt, T., Piazzesi, M., and Schneider, M. (2015) “The Housing Market(s) of San Diego,” *American Economic Review*, Vol. 105, pp. 1371–1407, URL: http://www.aeaweb.org/articles?id=10.1257/aer.20111662, DOI: http://dx.doi.org/10.1257/aer.20111662.

Le, Q. and Mikolov, T. (2014) “Distributed representations of sentences and documents,” in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1188–1196.

Loberto, M. and Zollino, F. (2016) “Housing and credit markets in Italy in times of crisis,” Temi di discussione (Economic working papers) 1087, Bank of Italy, Economic Research and International Relations Area.

Merlo, A. and Ortalo-Magne, F. (2004) “Bargaining over residential real estate: evidence from England,” *Journal of Urban Economics*, Vol. 56, pp. 192–216.

Mian, A., Rao, K., and Sufi, A. (2013) “Household Balance Sheets, Consumption, and the Economic Slump,” *The Quarterly Journal of Economics*, Vol. 128, pp. 1687–1726, URL: https://ideas.repec.org/a/oup/qjecon/v128y2013i4p1687-1726.html.

Naumann, F. and Herschel, M. (2010) *An Introduction to Duplicate Detection*: Morgan and Claypool Publishers.

Piazzesi, M. and Schneider, M. (2016) “Chapter 19 - Housing and Macroeconomics,” Vol. 2 of Handbook of Macroeconomics: Elsevier, pp. 1547 – 1640, URL: http://www.sciencedirect.com/science/article/pii/S1574004816300167, DOI: http://dx.doi.org/https://doi.org/10.1016/bs.hesmac.2016.06.003.

Piazzesi, M., Schneider, M., and Stroebel, J. (2017) “Segmented Housing Search,” NBER Working Papers 20823, National Bureau of Economic Research, Inc.

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Wu, L. and Brynjolfsson, E. (2015) “The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales,” in *Economic Analysis of the Digital Economy*: National Bureau of Economic Research, Inc, pp. 89–118, URL: https://ideas.repec.org/h/nbr/nberch/12994.html.

# A Description of the dataset

The source data which we obtained from Immobiliare.it are contained in weekly files. Starting from these snapshots, we construct six datasets. The main dataset is the one with the unique ads. Then there are three datasets that track the change of price, visits and favorites. Every record is a distinct value of these three variables, the unique identifier of the corresponding ad and the modification date. The last two datasets contain information about real estate agents and the the list of hash codes of the pictures associated to each ad.

For each ad the database comprises the following information:

- `ad_id` *integer* Unique ad identifier.
- `ad_db_added` *date* Date in which the ad was created in the database.
- `ad_db_removed` *date* Date in which the ad was removed from the database.
- `ad_update` *date* Date in which one of the characteristics of the ad was modified for the last time.
- `publisher_type` *categorical* Publisher of the ad. Levels: "agency" or "private citizen".
- `agency_id` *integer* Unique identifier for the agency. Note that the identifier corresponds to an account on the website, so the same agency could create multiple accounts.
- `city_istat_code` *integer* Istat code for the municipality the housing unit is in.
- `crc_codes` *integer* Hash codes of the pictures associated with the ad.
- `contract_type` *categorical* Type of sale contract. Levels: "Full ownership", "Partial ownership", "Leasehold estate", "Usufruct".
- `address` *character* Address of the housing unit .
- `agency_name` *character* Name of the agency posting the ad.
- `agency_address` *character* Address of the agency posting the ad.
- `air_conditioning` *boolean* True if the housing unit has an air conditioning system.
- `auction` *boolean* True if the housing unit is sold through a foreclosure auction.
- `balcony` *boolean* True if the housing unit has a balcony.
- `bathrooms` *integer* Number of bathrooms in the housing unit.
- `building_category` *categorical* Category of the building the housing unit is in. Levels: "Luxury", "Cheap", "Average".
- `city` *categorical* Municipality the ad is in.
- `content` *character* Description of the housing unit . It contains both a repetition of the features in the other fields (e.g. air conditioning, balcony, etc.) and some additional information. There is usually a promotional message for the agency which uploaded the ad.
- `elevator` *boolean* True if there is an elevator in the building the housing unit is in.
- `energy_class` *categorical* Energy class of the housing unit . Energy classes are assigned according to APE values. Levels: "A+", "A1-4", "A", "B", "C", "D", "E", "F", "G", "Not classifiable".

- `floor` *categorical* Floor of the housing unit. Levels: “1-10”, ‘Ground floor‘”, “Basement”, “On multiple floors”, “Highest”
- `floor_area` *integer* Floor area of the housing unit.
- `garage` *categorical* Type of garage for the housing unit. Levels: “Single”, “Double”, “Parking space”. It is “Missing” if the housing unit does not have a garage, it is NULL if this piece of information is not provided.
- `garden_type` *categorical* Type of garden for the building the housing unit is. Levels: “Shared”, “Private”. It is “Missing” if the building does not have a garden, it is NULL if this piece of information is not provided.
- `heating_type` *categorical* Type of heating system for the housing unit. Levels: “Central”, “Autonomous”. It is “Missing” if the housing unit does not have a heating system, it is NULL if this piece of information is not provided.
- `kitchen_type` *categorical* Type of kitchen for the housing unit . Levels: “Large eat-in kitchen”, “Small eat-in kitchen”, “Kitchenette”.
- `latitude` *float* Latitude of the housing unit.
- `leads` *integer* Times the creator of the ad has been contacted through the website by a potential buyer. For each ad we observe this variable weekly and we store it only when it changes from compared to the week before, together with the relative date.
- `longitude` *float* Longitude of the housing unit.
- `price` *float* Asking price of the housing unit. For each ad we observe this variable weekly and we store it only when it changes from compared to the week before, together with the relative date.
- `property_type` *categorical* Type of the housing unit. Levels: “Apartment”, “Villa”, “Attic”, “Semi-detached house”, “Detached house”, “Loft/open space”.
- `rooms` *integer* Number of rooms of the housing unit. It is upper limited by 5.
- `status` *categorical* Maintenance status of the housing unit. Levels: “New”, “Excellent”, “Good”, “To renovate”.
- `terrace` *boolean* True if the housing unit has a terrace.
- `visits` *integer* Number of visits on the ad. For each ad we observe this variable weekly and we store it only when it changes from compared to the week before, together with the relative date.
- `basement` *boolean* True if the housing unit has a basement (information recovered from the textual description).
- `janitor` *integer* True if the housing unit has a janitor (information recovered from the textual description).
- `utilityroom` *integer* True if the housing unit has a utility room (information recovered from the textual description).

# B Summary statistics

| Variable | N | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Missing |
|---|---|---|---|---|---|---|---|---|
| Start date | 653499 | 2006-02-25 | 2014-12-30 | 2015-11-24 | 2015-10-12 | 2016-09-11 | 2017-07-02 | 0 |
| End date | 301532 | 2015-01-05 | 2015-09-16 | 2016-05-06 | 2016-04-22 | 2016-12-12 | 2017-07-02 | 351967 |

Table 14: Variables of type `Date`. If end date is missing, it means that the housing unit has not disappeared from the website by the latest available weekly snapshot.

| Variable | N | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA |
|---|---|---|---|---|---|---|---|---|
| Floor area | 647,223 | 27.0 | 70.0 | 93.0 | 108.9 | 130 | 550.0 | 6,276 |
| Number of rooms | 630,220 | 1.0 | 3.0 | 3.0 | 3.3 | 4 | 5.0 | 23,279 |
| Number of bathrooms | 640,468 | 1.0 | 1.0 | 1.0 | 1.5 | 2 | 3.0 | 13,031 |
| Ads that refer to the same housing unit | 653,499 | 1.0 | 1.0 | 1.0 | 1.6 | 2 | 105.0 | 0 |
| Time on market | 298,843 | 8.0 | 92.0 | 189.0 | 268.4 | 360 | 1,512.0 | 354,656 |
| Price | 647,330 | 25,000 | 128,000 | 196,000 | 269,900 | 320,000 | 2,120,000 | 6,169 |
| Price per m2 | 640,752 | 393.3 | 1,467.2 | 2,142.9 | 2,466.5 | 3,125 | 9,166.7 | 12,747 |
| Visits on ads that refer to the same housing unit | 646,989 | 43.0 | 487.0 | 1,102.0 | 1,834.8 | 2,340 | 15,435.0 | 6,510 |
| Favorites on ads that refer to the same housing unit | 649,970 | 0.0 | 0.0 | 1.0 | 2.5 | 3 | 39.0 | 3,529 |

Table 15: Numeric variables. In the cases of floor area, time on market, price, price per m2, visits and favorites we remove the upper and lower 0.5% of the distribution. Extreme values are often outliers due to misreporting by the real estate agents. The time on market is calculated only for the housing units that have disappeared from the dataset. Note also that the number of rooms is limited to "5 or more".

| **Variable** | **Levels** | **N** | % |
|---|---|---|---|
| Geographical area | Center | 219,616 | 33.6 |
| | North-West | 207,319 | 31.7 |
| | North-East | 108,563 | 16.6 |
| | South | 66,904 | 10.2 |
| | Islands | 41,663 | 6.4 |
| | Missing values | 9,434 | 1.4 |
| | all | 653,499 | 100.0 |
| Region | Lazio | 140,534 | 21.5 |
| | Lombardia | 112,193 | 17.2 |
| | Toscana | 60,191 | 9.2 |
| | Piemonte | 60,130 | 9.2 |
| | Emilia-Romagna | 59,299 | 9.1 |
| | Veneto | 38,166 | 5.8 |
| | Sicilia | 36,227 | 5.5 |
| | Liguria | 34,028 | 5.2 |
| | Campania | 26,498 | 4.0 |
| | Puglia | 22,906 | 3.5 |
| | (Others) | 60,462 | 9.2 |
| | Missing value | 2,865 | 0.4 |
| | all | 653,499 | 100.0 |
| City | Rome | 131,967 | 20.2 |
| | Milan | 70,222 | 10.8 |
| | Turin | 43,424 | 6.6 |
| | Genoa | 26,502 | 4.1 |
| | Florence | 21,842 | 3.3 |
| | Naples | 20,600 | 3.1 |
| | Bologna | 17,625 | 2.7 |
| | Palermo | 16,066 | 2.5 |
| | Padua | 11,110 | 1.7 |
| | Venice | 9,058 | 1.4 |

| | | | |
|---|---|---|---|
| | (Others) | 282,218 | 43.2 |
| | Missing values | 2,865 | 0.4 |
| | all | 653,499 | 100.0 |
| Energy class | G | 256,031 | 39.2 |
| | F | 53,544 | 8.2 |
| | E | 33,309 | 5.1 |
| | D | 23,521 | 3.6 |
| | A | 19,479 | 3.0 |
| | C | 16,078 | 2.5 |
| | Not available | 15,622 | 2.4 |
| | B | 15,080 | 2.3 |
| | Missing values | 220,835 | 33.8 |
| | all | 653,499 | 100.0 |
| Maintenance status | Excellent | 242,410 | 37.1 |
| | Good | 238,711 | 36.5 |
| | To renovate | 81,434 | 12.5 |
| | New | 63,994 | 9.8 |
| | Missing values | 26,950 | 4.1 |
| | all | 653,499 | 100.0 |
| Elevator | True | 354,420 | 54.2 |
| | False | 299,079 | 45.8 |
| | all | 653,499 | 100.0 |
| Kitchen type | Large eat-in kitchen | 389,902 | 59.7 |
| | Kitchenette | 127,798 | 19.6 |
| | Small eat-in kitchen | 87,022 | 13.3 |
| | Missing values | 48,777 | 7.5 |
| | all | 653,499 | 100.0 |
| Heating type | Autonomous | 387,215 | 59.2 |
| | Centralized | 193,337 | 29.6 |
| | Missing | 30,103 | 4.6 |
| | Missing values | 42,844 | 6.6 |
| | all | 653,499 | 100.0 |
| Floor | 1 | 165,954 | 25.4 |
| | 2 | 109,625 | 16.8 |
| | 3 | 81,225 | 12.4 |
| | 0 | 79,492 | 12.2 |
| | On multiple floors | 58,522 | 9.0 |
| | 4 | 51,981 | 8.0 |
| | 5 | 30,628 | 4.7 |
| | 6 | 16,055 | 2.5 |
| | 7 | 8,919 | 1.4 |
| | -1 | 5,144 | 0.8 |
| | (Others) | 8,624 | 1.3 |
| | Missing values | 37,330 | 5.7 |
| | all | 653,499 | 100.0 |
| Air conditioning | True | 183,937 | 28.1 |
| | False | 153,196 | 23.4 |
| | Missing values | 316,366 | 48.4 |
| | all | 653,499 | 100.0 |
| Property type | Apartment | 586,805 | 89.8 |
| | Villa | 66,694 | 10.2 |
| | all | 653,499 | 100.0 |
| Balcony | True | 405,513 | 62.0 |
| | False | 247,986 | 38.0 |
| | all | 653,499 | 100.0 |
| Terrace | False | 432,474 | 66.2 |
| | True | 221,025 | 33.8 |
| | all | 653,499 | 100.0 |
| Garden type | Missing | 401,905 | 61.5 |
| | Shared | 136,340 | 20.9 |

| | | | |
|---|---|---|---|
| | Private | 115,254 | 17.6 |
| | all | 653,499 | 100.0 |
| Garage | Missing | 413,928 | 63.3 |
| | Double | 194,895 | 29.8 |
| | Single | 44,676 | 6.8 |
| | all | 653,499 | 100.0 |
| Porter | False | 603,057 | 92.3 |
| | True | 50,442 | 7.7 |
| | all | 653,499 | 100.0 |
| Basement | False | 414,792 | 63.5 |
| | True | 238,707 | 36.5 |
| | all | 653,499 | 100.0 |
| Utility room | False | 468,868 | 71.8 |
| | True | 184,631 | 28.2 |
| | all | 653,499 | 100.0 |

Table 16: Categorical variables

# C Construction of the housing units dataset

In this section we fully describe the algorithm we implemented to remove the duplicated ads. In the next section we also show the pseudo-codes of the procedure.

## C.1 Pre-processing of the ads dataset

We want to use the description of the housing unit to identify potential duplicates, but we first need to transform the text into a numeric vector using semantic analysis. There exist standard algorithms in natural language processing that accomplish this task by considering the multiplicity of the words, such as bag-of-words (Harris, 1954), but we cannot use these algorithms here. Indeed, two different real estate agents can describe the same dwelling using different words or sentences and this makes standard measures of distance among texts useless. For this reason we resort to the recent *Paragraph Vector* (or *doc2vec*) algorithm proposed by Le and Mikolov (2014), that allows to represent a document by a $N$-dimensional vector taking into account both the order and the semantic of the words.

We also convert the class of some variables to alleviate the issue of misreporting of dwellings characteristics. Indeed, two different agents can report information partially different but not completely at the opposite regarding the characteristics of the same housing unit. Consider for example the case of maintenance status: one real estate agent can report that the dwelling must be completely renovated, while the other agent writes that only a partial renovation is necessary. However, it is not plausible that the second agent says that the housing unit is new. As maintenance status takes only 4 possible ordered categories, we convert the categorical variable to an integer variable that takes value from 1 to 4 and a greater value means a better maintenance status. In this way when we compare two dwellings we take the absolute difference between the two variables and we can easily allow for partial matching. We do this operation for several other ordered categorical variables other than maintenance status: energy class, garage, type of garden, type of kitchen. We report the details in Table 17.

| Variable | Original levels | Transformation |
|---|---|---|
| *Garage* | Missing, Single, Double | Integer: Missing = 0, Single = 1, Double = 2 |
| *Garden* | Missing, Shared, Private | Integer: Missing = 0, Shared = 1, Private = 2 |
| *Maintenance status* | To renovate, Good, Excellent, New | Integer: To renovate = 0, Good = 1, Excellent = 2, New = 3 |
| *Kitchen Type* | Kitchenette, Small eat-in kitchen, Large eat-in kitchen | Integer: Kitchenette = 0, Small eat-in kitchen = 1, Large eat-in kitchen = 2 |
| *Energy Class* | A+, A, B, C, D, E, F, G | Integer: A+ = 0, A = 0, B = 1, C = 2, D = 3, E = 4, F = 5, G = 6 |
| *address* | Text of the address | Vector of words in the address (removing prepositions and articles) |

Table 17: Variable transformations for the classification trees

## C.2 Identification of duplicates

We identify the duplicated ads based on a pairwise comparison, meaning that we compare each ad with all other ads that are potentially duplicates.

First of all, in order to reduce the computational complexity of the pairwise approach we identify for each ad its potential duplicates. We define as potential duplicates those ads that refer to dwellings distant less than 400 meters and with a difference in asking price lower than

25% in absolute value.[32] In this way we end up with a long list of pairs of ads and for each of them we have to decide if they are duplicates.

We classify each pair of ads as duplicates (TRUE) or distinct housing units (FALSE) based on a supervised classification tree. The algorithm adopted here is the C5.0 classification tree proposed by Quinlan (1993) (`http://www.rulequest.com/see5-info.html`). This algorithm handles autonomously missing data, is faster than similar algorithms and allows for boosting.

For each pair of ads we provide to the algorithm a vector of predictors (covariates in the jargon of machine learning) and based on this information the classification tree returns the probability that the two ads are duplicates. We consider a pair of ads to be duplicates if the estimated probability is greater than 0.5.

Among the predictors we consider: floor area, price, floor, energy class, garage, garden type, air conditioning, heating type, maintenance status, kitchen type, number of bathrooms, number of rooms, janitor, utility room, location, elevator, balcony and terrace. For continuous variables, such as price and floor area, we use both the percentage and the absolute difference; for geolocation, we take the distance in meters between the geographical coordinates of the two dwellings. For binary variables, such as elevator or basement, the predictor is a dummy variable, that takes value equal to 1 if both ads share the same characteristic. For discrete ordered multinomial variables (such as maintenance status) we consider instead different degrees of similarity, by taking the absolute difference between the two variables.

We use as predictor also the distance between the textual description of the two ads. For this variable we consider two different measures, depending on whether the ads are posted by the same agency or not. In the first case we use the Levenshtein distance, otherwise we compute the cosine similarities between the vectors produced using the *Paragraph Vector* algorithm.

We implement two different C5.0 models, depending on whether the ads are posted by the same agency or not. This choice is motivated by the observation that when an agency posts two ads for the same dwelling the characteristics in the ads are almost equal. On the contrary, when the ads are posted by different agencies (or by a private user) sometimes you can tell they refer to the same dwelling only thanks to the pictures on the website. This means that duplicated ads are less similar if posted by different agencies than if created by the same agency. As a consequence, a unique model for both cases could lead to an excess of ads considered as duplicates among those published by the same agency.

C5.0 is a supervised method that requires an initial training sample of pairs of ads of which we know with certainty whether they are duplicates or not. We construct two different training samples, one for each model, by manually checking the ads on the website, in particular comparing the pictures. The training sample for the ads of different agencies is made up of 9997 pairs of ads; among them 3483 are duplicates (true positive, TP). The training sample for the ads of the same agency is made up of 8688 observations and 1473 are duplicates. These samples are constructed by iterating the following steps: (i) estimation of the model based on the initial training sample; (ii) out-of-sample validation of the models; (iii) using the results of the out-of-sample exercise to increase the training sample. This three step approach is repeated several times, until we reach a sufficiently low misclassification error.

In order to assess the performance of the two models we randomly split each training sample in two different sub-samples: the first one (90% of the observations) is used to estimate the models, the second one (10% of the observations) is used for the out-of-sample assessment of the classification performance. We repeat the operation 1,000 times and we evaluate the performance based on average results. Since the number of true negatives (ads that are not duplicates) is much larger than the number of true positives, using the classic accuracy rate can be misleading about the actual performance of the models. For this reason we consider measures

[32] The difference in asking price is computed dividing the absolute difference between the two asking prices with the lowest of the two. Since this condition can be quite restrictive when considering dwellings with low asking prices, we consider as potential duplicates also those ads with absolute difference lower than 50,000 euro.

| | Observations | Duplicates | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| Different agency | 9997 | 3483 | 0.923 | 0.892 | 0.907 |
| Same agency | 8688 | 1473 | 0.952 | 0.963 | 0.957 |

Precision = TP/(TP+FP). Recall = TP/(TP+FN). F-measure = 2*(Precision*Recall)/(Precision+Recall). TP = true positive; FP = false positive; FN = false negative.

Table 18: Assessment of C5.0 models

of classification performance that do not rely on the number of true negatives, namely: precision, recall and F-measure.[33]

We show the results in Table 18. As expected, the model for ads of the same agency is significantly more precise than the one for ads of different agencies. As we said before, ads posted from the same agency and related to the same dwelling have almost all the characteristics in common, therefore it is easier to identify them. However, as the F-measure is equal to .907, also the C5.0 model for ads of different agencies has a quite good classification performance. We should remark that the variables used in the two models are not the same and have been selected in order to maximize the F-measure.[34] We report the set of variables for each model in Table 19.

### C.3 Creation of clusters of duplicates and information aggregation

Once we have identified the pairs of ads that are duplicates, we need a procedure to cluster all the ads that are considered related to the same housing unit and to aggregate the information in the ads.

Let us suppose for example that we have only three ads: A, B and C. It is possible that the pairs (A,B) and (B,C) are considered as duplicates, but (A,C) is not. How should we manage this case? A simple solution is to assume transitivity: this means that since A is a duplicate of B and B is a duplicate of C, we assume that C is a duplicate of A and all these ads are considered related to the same dwelling. However, this approach can bring several issues: let us suppose for example that the probability of being duplicates for the pair (A,B) is 0.95 and the probability for the pair (B,C) is 0.51. How reliable is in this case the assumption of transitivity?

Here we abstract from the assumption of transitivity and we decide whether a cluster of ads refers to the same housing unit based on a measure of internal similarity of the cluster. In order to illustrate our approach we consider a simple example. Assume we have ten ads, we compute for each of the 45 possible pairs the probability that they are duplicates and we remove all pairs with probability smaller than 0.5. The remaining pairs are shown in Table 20.

Starting from the results of the pairwise classification step in Table 20, we represent the information as a graph, in order to form clusters. The output of this step is represented in Figure 10(a). The identifiers of the ads (here assumed to be integers between 1 and 10) are the nodes of the graph. Two nodes are connected if the probability that they are duplicates is greater than 0.5.

The tuples of ads (2,3) and (1,7,8) are considered to refer to two distinct dwellings, as each

[33]The precision rate is defined as the ratio between the number of true positives and the sum of true and false positives; it thus measures how precise a classifier is in classifying true matches. The recall rate is defined as the ratio of true positives over the sum of true positives and false negatives; it measures the proportion of true matches that have been classified correctly. As there is a trade-off between precision and recall, we consider also a third additional measure, the F-measure, that calculates the harmonic mean between precision and recall.

[34]We started for both models with only five predictors: percentage difference between prices, absolute difference between prices, percentage difference between floor areas, absolute difference between floor areas, difference between floors. Then we added each candidate predictor one-by-one updating the initial model only if the variable provided an improvement of the F-measure (computed on the out-of-sample observations in a Monte Carlo experiment with 1,000 draws). We repeated the operation iteratively as long as there was no performance improvement from adding an additional predictor.

| Variable | Model 1 | Model 2 | Description of the variable |
|---|---|---|---|
| *price_abs* | Yes | Yes | Absolute difference between asking prices |
| *price_per* | Yes | Yes | Percentage difference between asking prices |
| *floorarea_abs* | Yes | Yes | Absolute difference between floor area |
| *floorarea_per* | Yes | Yes | Percentage difference between floor area |
| *floor* | Yes | Yes | Absolute difference between floor level (integer) |
| *distance* | Yes | Yes | Absolute distance in meters between households |
| *address* | Yes | Yes | Indicator function: 1 if the two addresses have at least one common word |
| *isnew* | Yes | Yes | Indicator function: 1 if at least one of the ads refers to a new house |
| *balcony* | Yes | Yes | Indicator function: 1 if the feature balcony is the same |
| *terrace* | Yes | Yes | Indicator function: 1 if the feature terrace is the same |
| *distdays1* | Yes | Yes | Number of days between the dates the ads have been added |
| *distdays2* | Yes | Yes | Number of days between the dates the characteristics have been updated |
| *status* | Yes | Yes | Absolute difference (integer) between categories |
| *elevator* | Yes | No | Indicator function: 1 if the feature elevator is the same |
| *energy_class* | Yes | No | Absolute difference (integer) between categories |
| *isdetached* | Yes | No | Indicator function: 1 if at least one of the ads refers to a detached or semi-detached house |
| *bathrooms* | Yes | No | Absolute difference between number of bathrooms (integer) |
| *kitchen_type* | Yes | No | Absolute difference (integer) between categories |
| *heating_type* | Yes | No | Indicator function: 1 if the feature heating type is the same |
| *distcontent1* | Yes | No | Cosine distance of vectors (*Paragraph vectors*) representing textual descriptions |
| *distcontent2* | No | Yes | Levenshtein distance between textual descriptions |
| *air_conditioning* | No | Yes | Indicator function: 1 if the feature air conditioning is the same |
| *rooms* | No | Yes | Absolute difference between number of rooms (integer) |
| *garage* | No | Yes | Absolute difference (integer) between categories |
| *garden* | No | Yes | Absolute difference (integer) between categories |
| *distdays3* | No | Yes | Number of days between the dates the prices have been updated |
| *utility_room* | No | Yes | Indicator function: 1 if the feature utility room is the same |
| *janitor* | No | Yes | Indicator function: 1 if the feature janitor is the same |

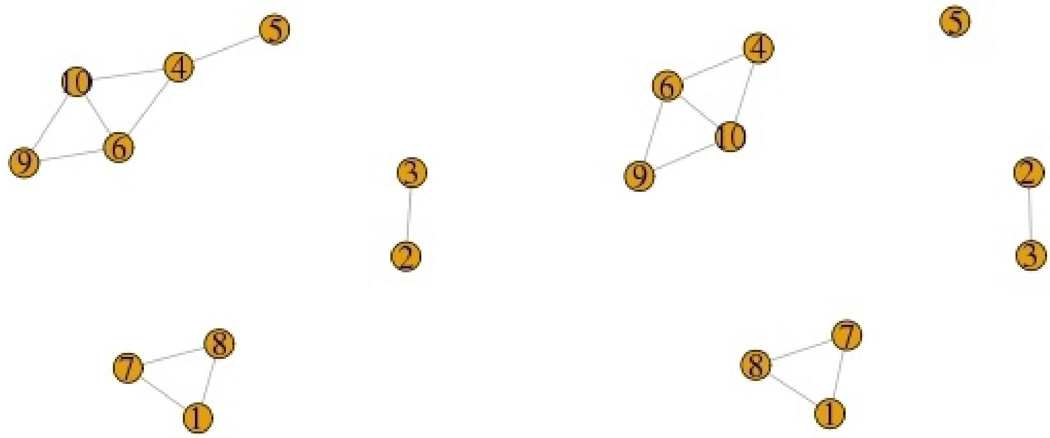
Table 19: Variables for the classification trees

of the ads in the tuple is a duplicate of all the other ads. The troubles come with the tuple (4,5,6,9,10). Here, differently than before, it is not true that each ad is a duplicate of all the others. In particular this sub-graph only has 6 edges, while in order to be defined as a fully connected graph we would need 10 edges. More generally, an indirect graph is said to be fully connected if the number of edges is equal to $\frac{N(N-1)}{2}$, where $N$ is the number of the nodes of

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| id.x | 1 | 1 | 2 | 4 | 4 | 4 | 6 | 6 | 7 | 9 |
| id.y | 7 | 8 | 3 | 6 | 10 | 5 | 9 | 10 | 8 | 10 |
| Prob. | 0.92 | 0.81 | 0.73 | 0.98 | 1.00 | 0.52 | 0.87 | 0.70 | 0.93 | 0.86 |

Table 20: Example of clusters

the graph (in our case the number of ads).




(a) Step 1 (b) Step 2

Figure 10: Clustering of the ads

The tuples (2,3) and (1,7,8) are clearly fully connected, while the tuple (4,5,6,9,10) is not. We consider a cluster as representing a single housing unit if it is a group of ads with a sufficiently high internal similarity, i.e. the number of edges is at least a fraction 5/6 of the maximum number of edges that we can have in the cluster. At each step we verify for each cluster if this condition is verified or not; if it is not satisfied we remove the weakest edge, that we define as the one with the lowest duplicate probability among those in the cluster.

Since for the tuple (4,5,6,9,10) the condition is not satisfied, we delete the weakest link, that in this case is represented by the edge between nodes (4) and (5), whose associated probability is 0.52. The new set of clusters after this operation is represented in Figure 10(b), in which the node (5) is now considered as referring to a distinct housing unit. If we look at the new tuple (4,6,9,10), we see that it has 5 edges out of 6 possible edges. Since our internal similarity condition is satisfied, we consider also this last tuple as a distinct dwelling.

Summing up our example, we started with 10 ads and we ended up with only 4 housing units. Based on this approach, we estimate for our training week (ads visible in 21 December 2016) that real dwellings were only 78% of the total ads (130 thousands housing units out of 168 thousands ads).

Once we have created the clusters of ads identifying different dwellings, an additional issue that must be considered is to collapse the information contained in multiple ads related to the same dwelling. Here, we adopt as a general rule that for each characteristic we take the one with highest absolute frequency. We deviate from this rule in the case of latitude and longitude (we compute the mean across the coordinates of all ads) and when we compute the dates of entry and exit of the dwelling into the housing market (for the entry we take the date of creation of the first ad associated to the dwelling, for the exit we consider the date of removal from the database of the last ad).[35]

[35] An additional exception to the general rule is done for asking prices. In this case we take the most frequent

## C.4 Time machine approach

The approach delineated above has the limit to be computationally unfeasible once the number of ads rises, because the number of pairwise comparisons increases exponentially. For this reason the procedure described in the previous section will be applied using an iterative approach ("time machine approach"), illustrated in detail in Appendix D.

We process the ads progressively as soon as they are published on the website. At the first iteration of the process we run the deduplication procedure on all the ads that have been added before the end of the first week we are considering. Once we apply the deduplication procedure, we end up with a new dataset where each row corresponds in principle to a unique dwelling and the characteristics of these housing units are derived from those of the associated ads.

At the second iteration we take as an input the datasets of ads and housing units of the first week. We check for duplicates only among the new ads added during the second week or the ads posted before but for which the price or other characteristics have been updated during the second week. For all these ads we look for duplicates both among new or updated ads and the dataset of housing units from the first week. The ads that are updated are preliminarily removed from the dataset of dwellings (that must be updated accordingly).

The decision on whether the ads are duplicates is still based on a pairwise comparison, but now we can have pairs with two ads or pairs with one ad and one housing unit. Once we compute for each pair the probability that they are duplicates we cluster the results as explained in Section C.3. Differently than before, we impose the additional condition that in each cluster there can be at most one housing unit that was already identified in the previous week. This additional condition is necessary to avoid that clusters of ads that have been considered as referring to different dwellings in the past processing can be considered now as duplicates, because there are new ads that are potential duplicates of both of them.

observation only among ads that have not been removed.

## D Pseudo code for deduplication

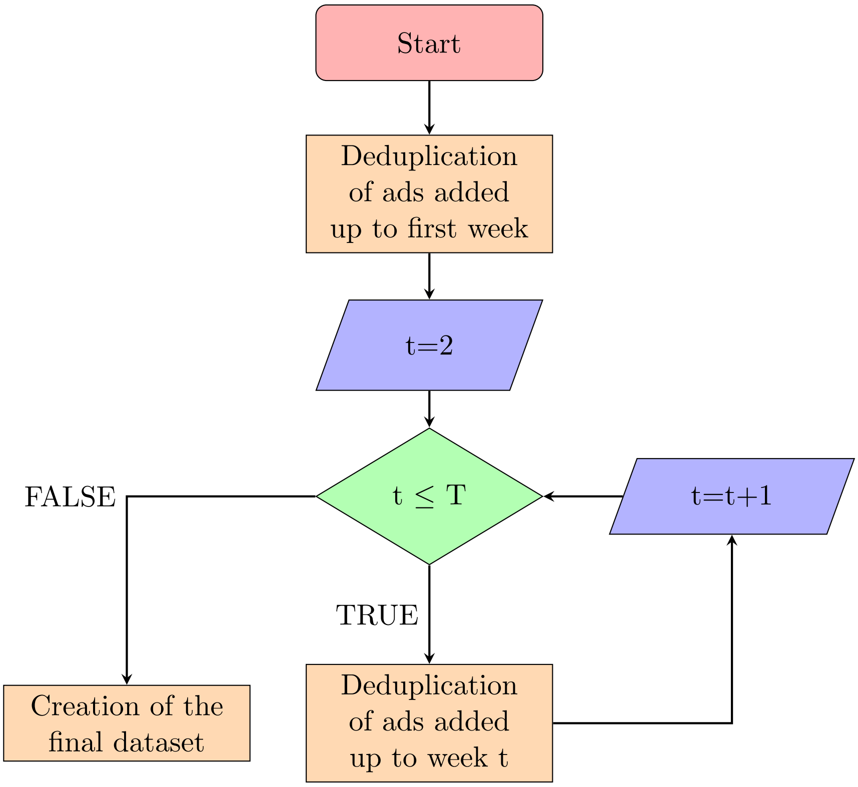



Figure 11: General approach

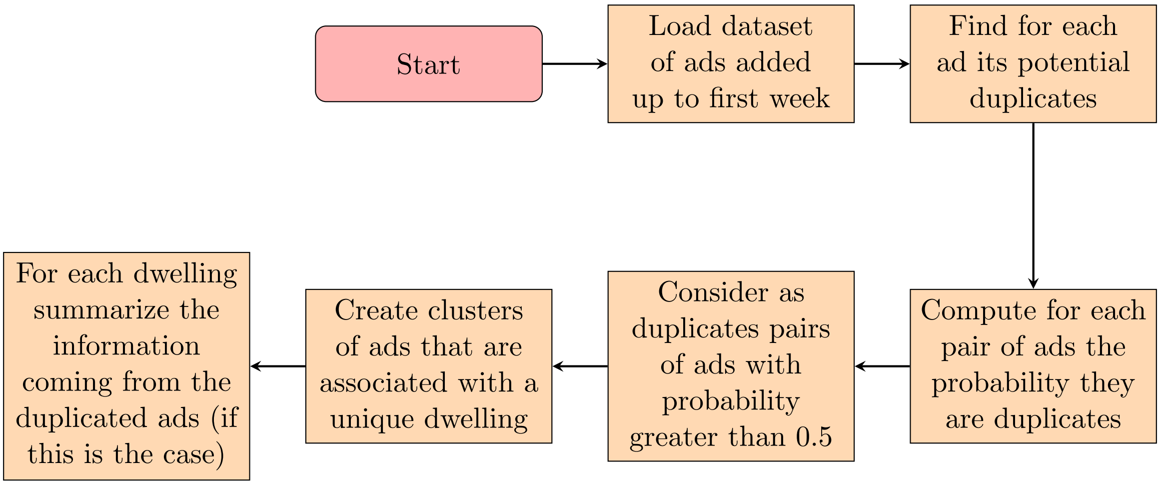



Figure 12: Deduplication - week 1

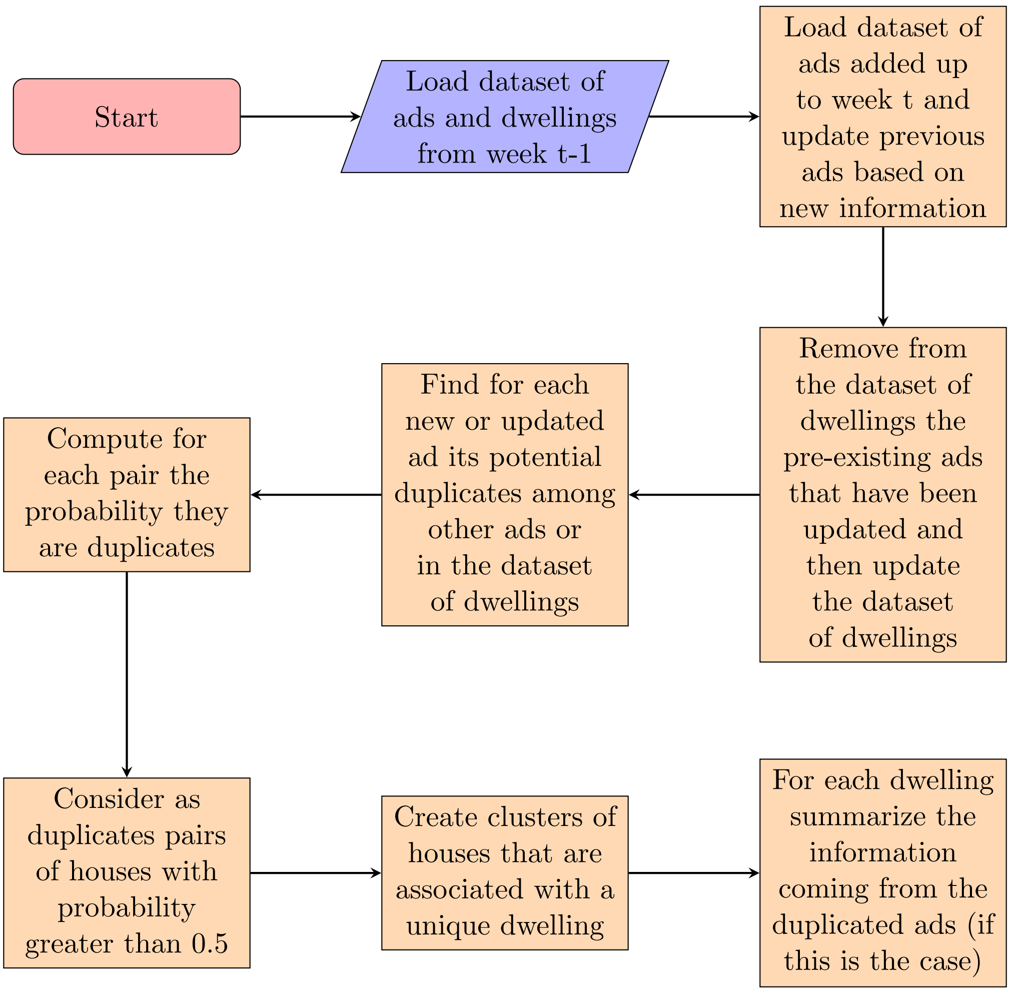



Figure 13: Deduplication - week >1

**Algorithm 1** First processing of the data

**procedure** CLEANING(DT) ▷ *DT is the dataset of original data on sales offers*
  DT ← DT[latitude ≠ NULL] ▷ *Keep only geo-referentiated ads*
  DT ← DT[longitude ≠ NULL]
  DT ← DT[floorarea ≠ NULL] ▷ *Keep only ads where the surface is not NULL*
  DT ← DT[entireprop = TRUE] ▷ *Keep only sales of the full property*
  DT ← DT[proptype in (Apartments, Detached and semi-detached dwellings, Penthouses, Loft-Open spaces )] ▷ *Keep only the most common types of properties*
  DT ← DT[(dbremoved-dbadded)>7] ▷ *Keep only the ads that last at least one week*
  DT ← DT[price ≠ NULL] ▷ *Keep only the ads where the price is not missing*

  *Extract information from the description of the ad to determine if the house is under foreclosure, it has still to be built or is not a residential unit. These ads will be drop. Information extraction is always performed looking for keywords in the description*
  **for** $i$ in 1:nrow(DT) **do**
    DT[$i$, auction] ← ISAUCTION(DT[$i$, descr])
    DT[$i$, tobebuilt] ← INPROGRESS(DT[$i$, descr])
    DT[$i$, nonresid] ← ISCOMMERCIAL(DT[$i$, descr])
  **end for**
  DT ← DT[auction = FALSE & tobebuilt = FALSE & nonresid = FALSE]

  *Extract additional information from the description of the ad*
  **for** $i$ in 1:nrow(DT) **do**
    DT[$i$, janitor] ← ANYJANITOR(DT[$i$, descr]) ▷ *TRUE if the building has a janitor*
    DT[$i$, basement] ← ANYBASEMENT(DT[$i$, descr]) ▷ *TRUE if the house has a basement*
    DT[$i$, utilityroom] ← ANYUTIL(DT[$i$, descr]) ▷ *TRUE if the house has an utility room*
  **end for**

  *For some variables recover information from the description of the ad if missing*
  **for** $i$ in 1:nrow(DT) **do**
    **if** DT[i, baths] = NULL **then**
      DT[$i$, baths] ← FINDBATH(DT[$i$, descr]) ▷ *Number of bathrooms*
    **else if** DT[$i$, rooms]= NULL **then**
      DT[$i$, rooms] ← FINDROOMS(DT[$i$, descr]) ▷ *Number of rooms*
    **else if** DT[$i$, floor]= NULL **then**
      DT[$i$, floor] ← FINDFLOOR(DT[$i$, descr]) ▷ *Floor level*
    **else if** DT[$i$, status]= NULL **then**
      DT[$i$, status] ← FINDSTATUS(DT[$i$, descr]) ▷ *Maintenance status*
    **end if**
    *Same operation is done for garage, garden, balcony, terrace and elevator. Only for these variables if no information is provided in the description we assume they do not exist, otherwise we let the missing data*
  **end for**
  DT[, ZonaOMI] ← FINDOMI(DT[, latitude],DT[, longitude]) ▷ *Find the OMI micro-zone of each dwelling using coordinates of the ads, maps available at* `http://wwwt.agenziaentrate.gov.it/geopoi_omi/` *and the* **sp** *package*

  *As last step we convert some ordinal variables from factor to integer; those variables are: floor level, garage, garden, energy class and maintenance status.*
  **return** DT
**end procedure**

**Algorithm 2** Deduplication for t=1

**procedure** DEDUPLICATION1(DT)
DT ← DT[dbadded ≤ t=1] ▷ *Keep only the ads added before the end of the first week. Prices, visits and leads are those at t=1*

*Identify potential duplicates*
POTDUPL ← NULL ▷ *Initialize an empty matrix POTDUPL with 2 cols*
**for** $i$ in 1: nrow(DT) **do**
**for** $j$ in $i$: nrow(DT) **do**
**check** if house $j$ is distant less than 400 meters from house $i$ and if $\frac{\max(P(i),P(j))}{\min(P(i),P(j))} - 1 < 0.25$ or $|P(i) - P(j)| < 50,000$
**return** FINDT ▷ *Pairs of ads that satisfy the conditions. FINDT is a dataset with 2 cols (adidx,adidy): adidx and adidy are the id of the ads*
**end for**
POTDUPL ← append(POTDUPL,FINDT) ▷ *Matrix with all pairs of potential duplicates*
POTDUPL ← POTDUPL[adidx ≠ adidy] ▷ *Drop the rows where adidx is equal to adidy*
**end for**

*For each pair of ads compute the probability they are duplicates and keep if prob > 0.5*
**for** $i$ in 1: nrow(POTDUPL) **do**
POTDUPL[i,prob] = FINDDUPL(POTDUPL[i,prob],DT) ▷ *Probability computed using C5.0 algorithm. Details in algorithm 4*
**end for**
POTDUPL ← POTDUPL[prob>0.5]

*Create clusters of ads that refer to the same dwelling. Details in algorithm 5*
DWELL ← CREATECLUSTERS(POTDUPL) ▷ *DWELL is a list containing the vectors of ads that are duplicates*
DWELL2 ← ads in DT that are not in DWELL ▷ *Find the ads that are not duplicated*
DTDWELL ← append(DWELL,DWELL2) ▷ *DTDWELL is now the list of all dwellings*
DTDWELL[, idunique = 1:nrow(DTDWELL)] ▷ *Assign to each dwelling a unique identifier. Now we have a data table with two columns: idunique, listads (listads contains the vector of ads associated with the idunique)*

*Compute for each dwelling its characteristics based on the information provided by the associated ads. For dwellings associated with a single ads the characteristics are those of the ad. When the dwelling is associated with more than one ads we assign for each characteristics the most frequent observation among the ads (excluding NaN). For latitude and longitude we take the mean*
**for** $i$ in 1: nrow(DTDWELL) **do**
**for** $j$ in 1: Nchar **do** ▷ *Nchar is the number of characteristics*
obsnew ← $\max_N$ DT[id in DTDWELL[$i$, listads],characteristic $j$]
DTDWELL[$i$, characteristic $j$] ← obsnew
**end for**
**end for**
**return** DTDWELL
**end procedure**

**Algorithm 3** Deduplication for t>1

**procedure** DEDUPLICATION1(DT,DTDWELL)

*Keep only the dwellings still on the market or those that have been retired by less than 10 weeks. Removed dwellings are saved in a separate dataset*

DTDWELL $\leftarrow$ DTDWELL[enddate=NaN or (t-enddate)<10]

DT $\leftarrow$ DT[dbadded $\leq$ t] $\triangleright$ *Keep only the ads added before the end of the week t. Prices, visits and leads are those most updated up to t*

DTnew $\leftarrow$ ADUPDATE(DT) $\triangleright$ *Create a list of id of the ads that have been added or updated (change of price or characteristics) in the current week*

*Remove from DTDWELL the ads in DTnew: if a dwelling was associated only to a single ad the entire dwelling is removed from DTDWELL. Then update DTDWELL based on the new information*

**for** $i$ in 1: nrow(DTnew) **do**

  **find** idunique $j$ s.t. DTnew[$i$, id] in DTDWELL[idunique= $j$, listads]

  **remove** DTnew[$i$, id] from DTDWELL[idunique= $j$, listads]

  **if** length(DTDWELL[idunique= $j$,listads])=0 **then**

    **remove** DTDWELL[idunique= $j$] from DTDWELL

  **end if**

**end for**

*Compute for each dwelling its characteristics based on the information provided by the associated ads (see algorithm 2)*

**update** DTDWELL

*Identify potential duplicates of each ad in DTnew among other ads in DTnew and dwellings in DTDWELL (see algorithm 2)*

**create** POTDUPL $\triangleright$ *Data table with 2 cols (adidx,adidy). Now we have both id of ads and idunique of the dwellings in DTDWELL*

*For each pair of dwellings compute the probability they are duplicates and keep if prob > 0.5 (see algorithm 4)*

**create** POTDUPL[,prob]

POTDUPL $\leftarrow$ POTDUPL[prob>0.5]

*Create clusters of ads that refer to the same dwelling. Details in algorithm 5*

DWELL $\leftarrow$ CREATECLUSTERS(POTDUPL)

DWELL2 $\leftarrow$ ads in DTnew and dwellings in DTDWELL that are not in DWELL

DTDWELL $\leftarrow$ append(DWELL,DWELL2)

**assign** idunique to all elements in DTDWELL $\triangleright$ *For dwellings already in DTDWELL mantain the same idunique. For new dwellings assign a new idunique*

*Compute now for each dwelling its characteristics based on the information provided by the associated ads (see algorithm 2)*

**update** DTDWELL

**return** DTDWELL

**end procedure**

**Algorithm 4** Identify if two houses are the same

```
procedure FINDDUPLICATES(POTDUPL,DT,DTDWELL)
    if t=1 then
        for i in 1:nrow(POTDUPL) do
            FEATURES ← differences between the characteristics of the dwellings in POTDUPL[i]    ▷ For the list of variables see Table 19
            if agencyid(POTDUPL[i, adidx])=agencyid(POTDUPL[i, adidy]) then    ▷ Case when the ad is published by the same agency
                POTDUPL[i, prob] = PREDC50SAMEAGENCY(FEATURES)    ▷ from C5.0 algorithm
            else
                POTDUPL[i, prob] = PREDC50DIFFAGENCY(FEATURES)
            end if
        end for
    else
        for i in 1:nrow(POTDUPL) do
            if POTDUPL[i, adidx] & POTDUPL[i, adidy] are both new or updated ads then
                apply the same procedure when t=1
            else
                Suppose that POTDUPL[i, adidx] is the idunique of a dwelling in DTDWELL. If the agency that have published the ad with id POTDUPL[i, adidy] is the same of one of the ads already associated with dwelling POTDUPL[i, adidx], then compare POTDUPL[i, adidy] with that ad. Otherwise, compare with the dwelling
                isdwell ← ISIDUNIQUE(POTDUPL[i, adidx])
                if isdwell=TRUE then
                    listagencies ← RECOVERAGENCY(DTDWELL[idunique = adidx, listads])    ▷ Recover the list of agencies that published the ads associated with dwelling adidx
                    if DTnew[id = adidy, agency] in listagencies then
                        FEATURES ← differences between the characteristics of ad adidy and the ad associated with dwelling i published by the same agency
                        POTDUPL[i, prob] = PREDC50SAMEAGENCY(FEATURES)
                    else
                        FEATURES ← differences between the characteristics of ad adidy and dwelling adidx
                        POTDUPL[i, prob] = PREDC50DIFFAGENCY(FEATURES)
                    end if
                else
                    FEATURES ← differences between the characteristics of ad adidy and dwelling adidx
                    POTDUPL[i, prob] = PREDC50DIFFAGENCY(FEATURES)
                end if
            end if
        end for
    end if
    return POTDUPL
end procedure
```

**Algorithm 5** Create clusters of duplicates

**procedure** CREATECLUSTERS(POTDUPL)

*Using POTDUPL as input create an undirected graph. The unique elements in POTDUPL[,adidx] and POTDUPL[,adidy] are the vertex of the graph. Each row of POTDUPL are the edges of the graph and the probability in that row is an attribute of that edge*

net ← GRAPH(POTDUPL) ▷ *Create the graph. All the procedure is done using the igraph library*

net ← DECOMPOSE(net) ▷ *Creates a separate subgraph for each component of a graph and return a list of graphs*

*Consider a subgraph as a unique dwelling if: 1) the number of edges is at least 5/6 of those necessary for the graph to be connected; 2) there is at most one idunique*

dwellist ← NULL ▷ *Initialize an empty list*

k=1

**while** length(net)>0 **do**

  **for** $i$ in 1:length(net) **do**

    Nedges ← COMPUTEDGES(net[$i$]) ▷ *Compute the number of edges of the graph*

    N ← COMPUTVERTEX(net[$i$]) ▷ *Compute the number of vertex of the graph*

    NN ← number of idunique among vertex

    **if** Nedges >= $\frac{5}{6}\frac{N(N-1)}{2}$ & NN <2 **then** ▷ *An undirected graph has* $\frac{N(N-1)}{2}$ *edges*

      dwellist[k] ← VERTEX(net[$i$]) ▷ *Add to dwellist the vector of id of the duplicates*

      net[$i$] ← NULL ▷ *Delete the graph from the list of graphs*

      k=k+1

    **else**

      xx ← edge with lower associated probability

      net[$i$] ← REMOVEDGE(net[$i$], xx) ▷ *Remove the weakest link from the graph*

    **end if**

  **end for**

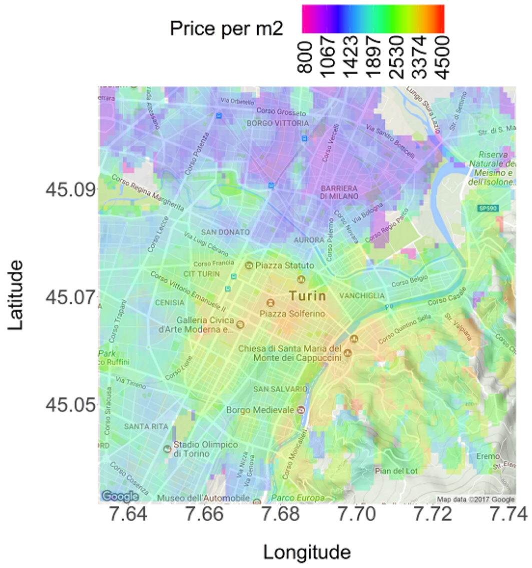
  net ← DECOMPOSE(net)

**end while**

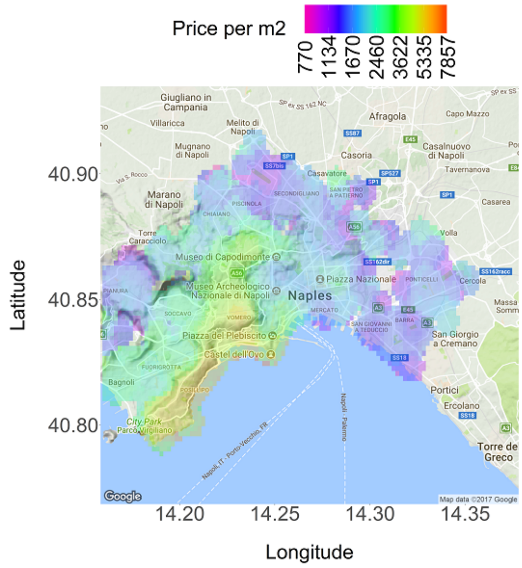
**return** dwellist

**end procedure**

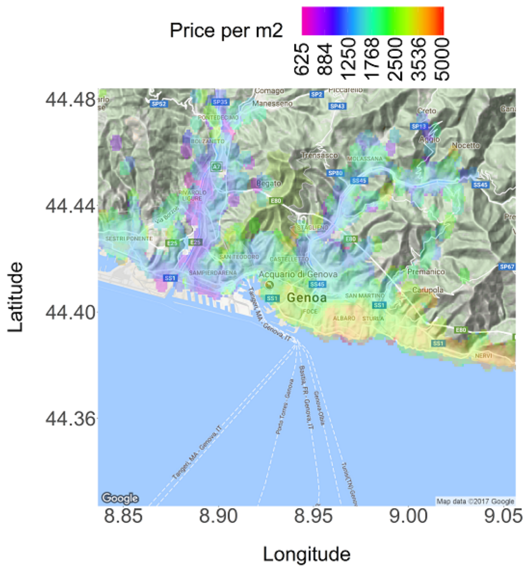
# E Additional maps



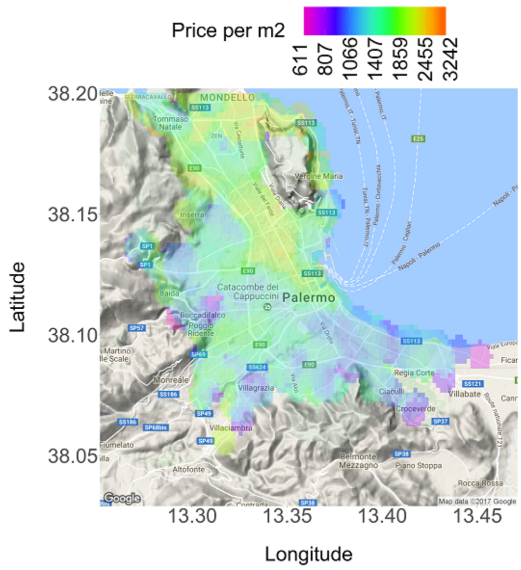

(a) Turin




(b) Naples

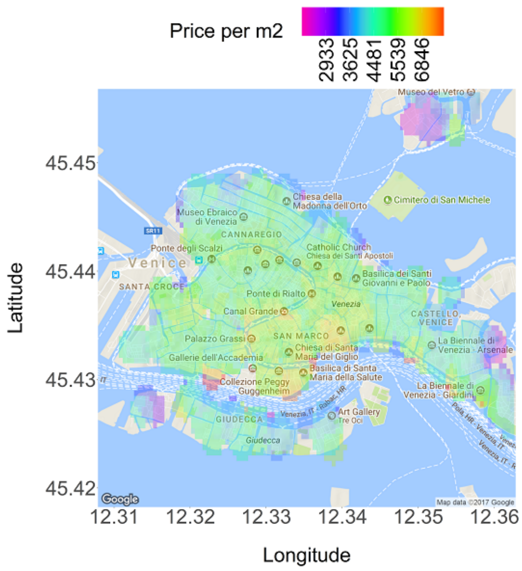



(c) Genoa

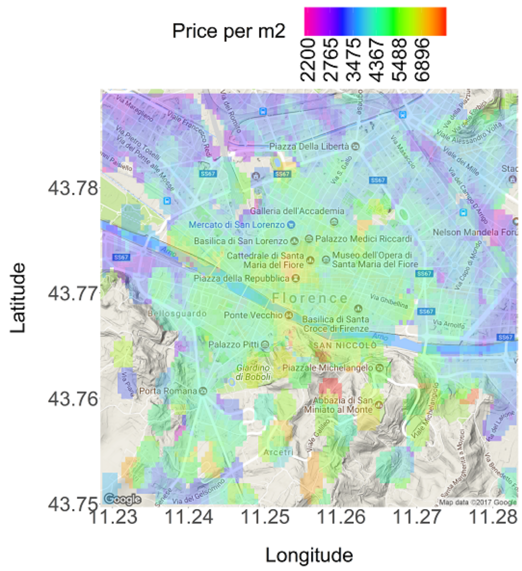



(d) Palermo

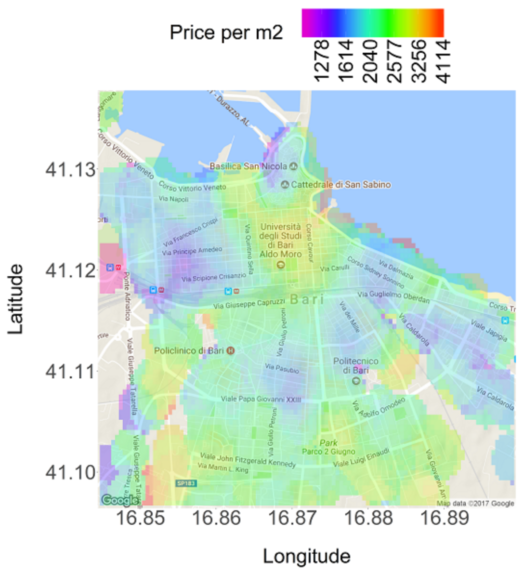
Figure 14: Kernel approximation of the (asking) price per m2 during 2017Q1.



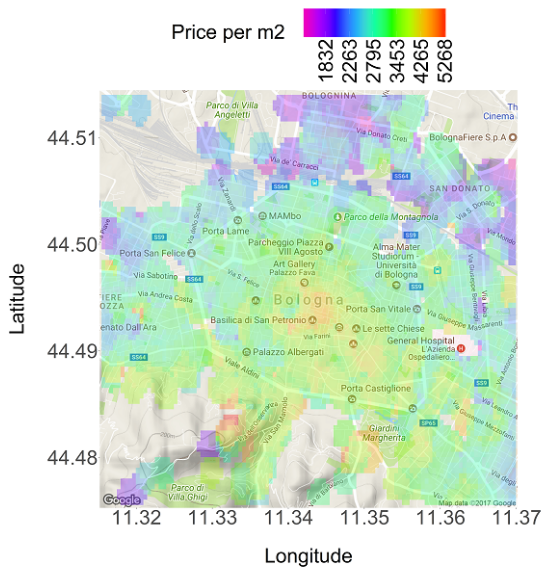

(a) Venice




(b) Florence




(c) Bari




(d) Bologna

Figure 15: Kernel approximation of the (asking) price per m2 during 2017Q1.

RECENTLY PUBLISHED "TEMI" (*)

N. 1147 – *Consistent inference in fixed-effects stochastic frontier models*, by Federico Belotti and Giuseppe Ilardi (October 2017).

N. 1148 – *Investment decisions by European firms and financing constraints*, by Andrea Mercatanti, Taneli Mäkinen and Andrea Silvestrini (October 2017).

N. 1149 – *Looking behind the financial cycle: the neglected role of demographics*, by Alessandro Ferrari (December 2017).

N. 1150 – *Public investment and monetary policy stance in the euro area*, by Lorenzo Burlon, Alberto Locarno, Alessandro Notarpietro and Massimiliano Pisani (December 2017).

N. 1151 – *Fiscal policy uncertainty and the business cycle: time series evidence from Italy*, by Alessio Anzuini, Luca Rossi and Pietro Tommasino (December 2017).

N. 1152 – *International financial flows and the risk-taking channel*, by Pietro Cova and Filippo Natoli (December 2017).

N. 1153 – *Systemic risk and systemic importance measures during the crisis*, by Sergio Masciantonio and Andrea Zaghini (December 2017).

N. 1154 – *Capital controls, macroprudential measures and monetary policy interactions in an emerging economy*, by Valerio Nispi Landi (December 2017).

N. 1155 – *Optimal monetary policy and fiscal policy interaction in a non-Ricardian economy*, by Massimiliano Rigon and Francesco Zanetti (December 2017).

N. 1156 – *Secular stagnation, R&D, public investment and monetary policy: a global-model perspective*, by Pietro Cova, Patrizio Pagano, Alessandro Notarpietro and Massimiliano Pisani (December 2017).

N. 1157 – *The CSPP at work: yield heterogeneity and the portfolio rebalancing channel*, by Andrea Zaghini (December 2017).

N. 1158 – *Targeting policy-compliers with machine learning: an application to a tax rebate programme in Italy*, by Monica Andini, Emanuele Ciani, Guido de Blasio, Alessio D'Ignazio and Viola Salvestrini (December 2017).

N. 1159 – *Banks' maturity transformation: risk, reward, and policy*, by Pierluigi Bologna (December 2017).

N. 1160 – *Pairwise trading in the money market during the European sovereign debt crisis*, by Edoardo Rainone (December 2017).

N. 1161 – *Please in my back yard: the private and public benefitsof a new tram line in Florence*, by Valeriia Budiakivska and Luca Casolaro (January 2018).

N. 1162 – *Real exchange rate misalignments in the euro area*, by by Michael Fidora, Claire Giordano and Martin Schmitz (January 2018).

N. 1163 – *What will Brexit mean for the British and euro-area economies? A model-based assessment of trade regimes*, by Massimiliano Pisani and Filippo Vergara Caffarelli (January 2018).

N. 1164 – *Are lenders using risk-based pricing in the consumer loan market? The effects of the 2008 crisis*, by Silvia Magri (January 2018).

N. 1165 – *Listening to the buzz: social media sentiment and retail depositors' trust* by Matteo Accornero and Mirko Moscatelli (February 2018)

N. 1166 – *Banks' holdings of and trading in government bonds*, by Michele Manna and Stefano Nobili (March 2018).

N. 1167 – *Firms' and households' investment in Italy: the role of credit constraints and other macro factors*, by Claire Giordano, Marco Marinucci and Andrea Silvestrini (March 2018).

N. 1168 – *Credit supply and productivity growth*, by Francesco Manaresi and Nicola Pierri (March 2018).

N. 1169 – *Consumption volatility risk and the inversion of the yield curve*, by Adriana Grasso and Filippo Natoli (March 2018).

---

(*) Requests for copies should be sent to:
Banca d'Italia – Servizio Studi di struttura economica e finanziaria – Divisione Biblioteca e Archivio storico – Via Nazionale, 91 – 00184 Rome – (fax 0039 06 47922059). They are available on the Internet www.bancaditalia.it.

*2016*

ALBANESE G., G. DE BLASIO and P. SESTITO, *My parents taught me. evidence on the family transmission of values,* Journal of Population Economics, v. 29, 2, pp. 571-592, **TD No. 955 (March 2014).**

ANDINI M. and G. DE BLASIO, *Local development that money cannot buy: Italy's Contratti di Programma,* Journal of Economic Geography, v. 16, 2, pp. 365-393, **TD No. 915 (June 2013).**

BARONE G. and S. MOCETTI, *Inequality and trust: new evidence from panel data,* Economic Inquiry, v. 54, pp. 794-809, **TD No. 973 (October 2014).**

BELTRATTI A., B. BORTOLOTTI and M. CACCAVAIO, *Stock market efficiency in China: evidence from the split-share reform,* Quarterly Review of Economics and Finance, v. 60, pp. 125-137, **TD No. 969 (October 2014).**

BOLATTO S. and M. SBRACIA, *Deconstructing the gains from trade: selection of industries vs reallocation of workers,* Review of International Economics, v. 24, 2, pp. 344-363, **TD No. 1037 (November 2015).**

BOLTON P., X. FREIXAS, L. GAMBACORTA and P. E. MISTRULLI, *Relationship and transaction lending in a crisis*, Review of Financial Studies, v. 29, 10, pp. 2643-2676, **TD No. 917 (July 2013).**

BONACCORSI DI PATTI E. and E. SETTE, *Did the securitization market freeze affect bank lending during the financial crisis? Evidence from a credit register,* Journal of Financial Intermediation , v. 25, 1, pp. 54-76, **TD No. 848 (February 2012).**

BORIN A. and M. MANCINI, *Foreign direct investment and firm performance: an empirical analysis of Italian firms,* Review of World Economics, v. 152, 4, pp. 705-732, **TD No. 1011 (June 2015).**

BRAGOLI D., M. RIGON and F. ZANETTI, *Optimal inflation weights in the euro area,* International Journal of Central Banking, v. 12, 2, pp. 357-383, **TD No. 1045 (January 2016).**

BRANDOLINI A. and E. VIVIANO, *Behind and beyond the (headcount) employment rate,* Journal of the Royal Statistical Society: Series A, v. 179, 3, pp. 657-681, **TD No. 965 (July 2015).**

BRIPI F., *The role of regulation on entry: evidence from the Italian provinces,* World Bank Economic Review, v. 30, 2, pp. 383-411, **TD No. 932 (September 2013).**

BRONZINI R. and P. PISELLI, *The impact of R&D subsidies on firm innovation,* Research Policy, v. 45, 2, pp. 442-457, **TD No. 960 (April 2014).**

BURLON L. and M. VILALTA-BUFI, *A new look at technical progress and early retirement,* IZA Journal of Labor Policy, v. 5, **TD No. 963 (June 2014).**

BUSETTI F. and M. CAIVANO, *The trend–cycle decomposition of output and the Phillips Curve: bayesian estimates for Italy and the Euro Area,* Empirical Economics, V. 50, 4, pp. 1565-1587, **TD No. 941 (November 2013).**

CAIVANO M. and A. HARVEY, *Time-series models with an EGB2 conditional distribution,* Journal of Time Series Analysis, v. 35, 6, pp. 558-571, **TD No. 947 (January 2014).**

CALZA A. and A. ZAGHINI, *Shoe-leather costs in the euro area and the foreign demand for euro banknotes,* International Journal of Central Banking, v. 12, 1, pp. 231-246, **TD No. 1039 (December 2015).**

CESARONI T. and R. DE SANTIS, *Current account "core-periphery dualism" in the EMU,* The World Economy, v. 39, 10, pp. 1514-1538, **TD No. 996 (December 2014).**

CIANI E., *Retirement, Pension eligibility and home production,* Labour Economics, v. 38, pp. 106-120, **TD No. 1056 (March 2016).**

CIARLONE A. and V. MICELI, *Escaping financial crises? Macro evidence from sovereign wealth funds' investment behaviour,* Emerging Markets Review, v. 27, 2, pp. 169-196, **TD No. 972 (October 2014).**

CORNELI F. and E. TARANTINO, *Sovereign debt and reserves with liquidity and productivity crises,* Journal of International Money and Finance, v. 65, pp. 166-194, **TD No. 1012 (June 2015).**

D'AURIZIO L. and D. DEPALO, *An evaluation of the policies on repayment of government's trade debt in Italy,* Italian Economic Journal, v. 2, 2, pp. 167-196, **TD No. 1061 (April 2016).**

DE BLASIO G., G. MAGIO and C. MENON, *Down and out in Italian towns: measuring the impact of economic downturns on crime,* Economics Letters, 146, pp. 99-102, **TD No. 925 (July 2013).**

DOTTORI D. and M. MANNA, *Strategy and tactics in public debt management,* Journal of Policy Modeling, v. 38, 1, pp. 1-25, **TD No. 1005 (March 2015).**

LIBERATI D., M. MARINUCCI and G. M. TANZI, *Science and technology parks in Italy: main features and analysis of their effects on hosted firms*, Journal of Technology Transfer, v. 41, 4, pp. 694-729, **TD No. 983 (November 2014).**

MARCELLINO M., M. PORQUEDDU and F. VENDITTI, *Short-Term GDP forecasting with a mixed frequency dynamic factor model with stochastic volatility,* Journal of Business & Economic Statistics , v. 34, 1, pp. 118-127, **TD No. 896 (January 2013).**

RODANO G., N. SERRANO-VELARDE and E. TARANTINO, *Bankruptcy law and bank financing,* Journal of Financial Economics, v. 120, 2, pp. 363-382, **TD No. 1013 (June 2015).**

ZINNA G., *Price pressures on UK real rates: an empirical investigation,* Review of Finance,v. 20, 4, pp. 1587-1630, **TD No. 968 (July 2014).**

*2017*

ADAMOPOULOU A. and G.M. TANZI, *Academic dropout and the great recession,* Journal of Human Capital, V. 11, 1, pp. 35–71, **TD No. 970 (October 2014).**

ALBERTAZZI U., M. BOTTERO and G. SENE, *Information externalities in the credit market and the spell of credit rationing,* Journal of Financial Intermediation, v. 30, pp. 61–70, **TD No. 980 (November 2014).**

ALESSANDRI P. and H. MUMTAZ, *Financial indicators and density forecasts for US output and inflation,* Review of Economic Dynamics, v. 24, pp. 66-78, **TD No. 977 (November 2014).**

BARBIERI G., C. ROSSETTI and P. SESTITO, *Teacher motivation and student learning,* Politica economica/Journal of Economic Policy, v. 33, 1, pp.59-72, **TD No. 761 (June 2010).**

BENTIVOGLI C. and M. LITTERIO, *Foreign ownership and performance: evidence from a panel of Italian firms,* International Journal of the Economics of Business, v. 24, 3, pp. 251-273, **TD No. 1085 (October 2016).**

BRONZINI R. and A. D'IGNAZIO, *Bank internationalisation and firm exports: evidence from matched firm-bank data,* Review of International Economics, v. 25, 3, pp. 476-499 **TD No. 1055 (March 2016).**

BRUCHE M. and A. SEGURA, *Debt maturity and the liquidity of secondary debt markets,* Journal of Financial Economics, v. 124, 3, pp. 599-613, **TD No. 1049 (January 2016).**

BURLON L., *Public expenditure distribution, voting, and growth,* Journal of Public Economic Theory,, v. 19, 4, pp. 789–810, **TD No. 961 (April 2014).**

BURLON L., A. GERALI, A. NOTARPIETRO and M. PISANI, *Macroeconomic effectiveness of non-standard monetary policy and early exit. a model-based evaluation,* International Finance, v. 20, 2, pp.155-173, **TD No. 1074 (July 2016).**

BUSETTI F., *Quantile aggregation of density forecasts,* Oxford Bulletin of Economics and Statistics, v. 79, 4, pp. 495-512, **TD No. 979 (November 2014).**

CESARONI T. and S. IEZZI, *The predictive content of business survey indicators: evidence from SIGE,* Journal of Business Cycle Research, v.13, 1, pp 75–104, **TD No. 1031 (October 2015).**

CONTI P., D. MARELLA and A. NERI, *Statistical matching and uncertainty analysis in combining household income and expenditure data,* Statistical Methods & Applications, v. 26, 3, pp 485–505, **TD No. 1018 (July 2015).**

D'AMURI F., *Monitoring and disincentives in containing paid sick leave,* Labour Economics, v. 49, pp. 74-83, **TD No. 787 (January 2011).**

D'AMURI F. and J. MARCUCCI, *The predictive power of google searches in forecasting unemployment,* International Journal of Forecasting, v. 33, 4, pp. 801-816, **TD No. 891 (November 2012).**

DE BLASIO G. and S. POY, *The impact of local minimum wages on employment: evidence from Italy in the 1950s,* Journal of Regional Science, v. 57, 1, pp. 48-74, **TD No. 953 (March 2014).**

DEL GIOVANE P., A. NOBILI and F. M. SIGNORETTI, *Assessing the sources of credit supply tightening: was the sovereign debt crisis different from Lehman?,* International Journal of Central Banking, v. 13, 2, pp. 197-234, **TD No. 942 (November 2013).**

DEL PRETE S., M. PAGNINI, P. ROSSI and V. VACCA, *Lending organization and credit supply during the 2008–2009 crisis,* Economic Notes, v. 46, 2, pp. 207–236, **TD No. 1108 (April 2017).**

DELLE MONACHE D. and I. PETRELLA, *Adaptive models and heavy tails with an application to inflation forecasting,* International Journal of Forecasting, v. 33, 2, pp. 482-501, **TD No. 1052 (March 2016).**

FEDERICO S. and E. TOSTI, *Exporters and importers of services: firm-level evidence on Italy,* The World Economy, v. 40, 10, pp. 2078-2096, **TD No. 877 (September 2012).**

GIACOMELLI S. and C. MENON, *Does weak contract enforcement affect firm size? Evidence from the neighbour's court,* Journal of Economic Geography, v. 17, 6, pp. 1251-1282, **TD No. 898 (January 2013).**

LOBERTO M. and C. PERRICONE, *Does trend inflation make a difference?,* Economic Modelling, v. 61, pp. 351–375, **TD No. 1033 (October 2015).**

MANCINI A.L., C. MONFARDINI and S. PASQUA, *Is a good example the best sermon? Children's imitation of parental reading,* Review of Economics of the Household, v. 15, 3, pp 965–993, **D No. 958 (April 2014).**

MEEKS R., B. NELSON and P. ALESSANDRI, *Shadow banks and macroeconomic instability,* Journal of Money, Credit and Banking, v. 49, 7, pp. 1483–1516, **TD No. 939 (November 2013).**

MICUCCI G. and P. ROSSI, *Debt restructuring and the role of banks' organizational structure and lending technologies,* Journal of Financial Services Research, v. 51, 3, pp 339–361, **TD No. 763 (June 2010).**

MOCETTI S., M. PAGNINI and E. SETTE, *Information technology and banking organization,* Journal of Journal of Financial Services Research, v. 51, pp. 313-338, **TD No. 752 (March 2010).**

MOCETTI S. and E. VIVIANO, *Looking behind mortgage delinquencies,* Journal of Banking & Finance, v. 75, pp. 53-63, **TD No. 999 (January 2015).**

NOBILI A. and F. ZOLLINO, *A structural model for the housing and credit market in Italy,* Journal of Housing Economics, v. 36, pp. 73-87, **TD No. 887 (October 2012).**

PALAZZO F., *Search costs and the severity of adverse selection,* Research in Economics, v. 71, 1, pp. 171-197, **TD No. 1073 (July 2016).**

PATACCHINI E. and E. RAINONE, *Social ties and the demand for financial services,* Journal of Financial Services Research, v. 52, 1–2, pp 35–88, **TD No. 1115 (June 2017).**

PATACCHINI E., E. RAINONE and Y. ZENOU, *Heterogeneous peer effects in education,* Journal of Economic Behavior & Organization, v. 134, pp. 190–227, **TD No. 1048 (January 2016).**

SBRANA G., A. SILVESTRINI and F. VENDITTI, *Short-term inflation forecasting: the M.E.T.A. approach,* International Journal of Forecasting, v. 33, 4, pp. 1065-1081, **TD No. 1016 (June 2015).**

SEGURA A. and J. SUAREZ, *How excessive is banks' maturity transformation?,* Review of Financial Studies, v. 30, 10, pp. 3538–3580, **TD No. 1065 (April 2016).**

VACCA V., *An unexpected crisis? Looking at pricing effectiveness of heterogeneous banks,* Economic Notes, v. 46, 2, pp. 171–206, **TD No. 814 (July 2011).**

VERGARA CAFFARELI F., *One-way flow networks with decreasing returns to linking,* Dynamic Games and Applications, v. 7, 2, pp. 323-345, **TD No. 734 (November 2009).**

ZAGHINI A., *A Tale of fragmentation: corporate funding in the euro-area bond market,* International Review of Financial Analysis, v. 49, pp. 59-68, **TD No. 1104 (February 2017).**

*2018*

BELOTTI F. and G. ILARDI *Consistent inference in fixed-effects stochastic frontier models,* Journal of Econometrics, v. 202, 2, pp. 161-177, **TD No. 1147 (October 2017).**

CARTA F. and M. DE PHLIPPIS, *You've Come a long way, baby. husbands' commuting time and family labour supply,* Regional Science and Urban Economics, v. 69, pp. 25-37, **TD No. 1003 (March 2015).**

CARTA F. and L. RIZZICA, *Early kindergarten, maternal labor supply and children's outcomes: evidence from Italy,* Journal of Public Economics, v. 158, pp. 79-102, **TD No. 1030 (October 2015).**

CECCHETTI S., F. NATOLI and L. SIGALOTTI, *Tail co-movement in inflation expectations as an indicator of anchoring,* International Journal of Central Banking, v. 14, 1, pp. 35-71, **TD No. 1025 (July 2015).**

NUCCI F. and M. RIGGI, *Labor force participation, wage rigidities, and inflation,* Journal of Macroeconomics, v. 55, 3 pp. 274-292, **TD No. 1054 (March 2016).**

SEGURA A., *Why did sponsor banks rescue their SIVs?,* Review of Finance, v. 22, 2, pp. 661-697, **TD No. 1100 (February 2017).**

*FORTHCOMING*


ADAMOPOULOU A. and E. KAYA, *Young Adults living with their parents and the influence of peers,* Oxford Bulletin of Economics and Statistics, **TD No. 1038 (November 2015).**

ALBANESE G., G. DE BLASIO and P. SESTITO, *Trust, risk and time preferences: evidence from survey data,* International Review of Economics, **TD No. 911 (April 2013).**

BARONE G., G. DE BLASIO and S. MOCETTI, *The real effects of credit crunch in the great recession: evidence from Italian provinces,* Regional Science and Urban Economics, **TD No. 1057 (March 2016).**

BELOTTI F. and G. ILARDI, *Consistent inference in fixed-effects stochastic frontier models,* Journal of Econometrics, **TD No. 1147 (October 2017).**

BERTON F., S. MOCETTI, A. PRESBITERO and M. RICHIARDI, *Banks, firms, and jobs,* Review of Financial Studies, **TD No. 1097 (February 2017).**

BOFONDI M., L. CARPINELLI and E. SETTE, *Credit supply during a sovereign debt crisis,* Journal of the European Economic Association, **TD No. 909 (April 2013).**

BRILLI Y. and M. TONELLO, *Does increasing compulsory education reduce or displace adolescent crime? New evidence from administrative and victimization data,* CESifo Economic Studies, **TD No. 1008 (April 2015).**

CASIRAGHI M., E. GAIOTTI, L. RODANO and A. SECCHI, *A "Reverse Robin Hood"? The distributional implications of non-standard monetary policy for Italian households,* Journal of International Money and Finance, **TD No. 1077 (July 2016).**

CIPRIANI M., A. GUARINO, G. GUAZZAROTTI, F. TAGLIATI and S. FISHER, *Informational contagion in the laboratory,* Review of Finance, **TD No. 1063 (April 2016).**

D'AMURI F., *Monitoring and disincentives in containing paid sick leave,* Labour Economics, **TD No. 787 (January 2011).**

FEDERICO S. and E. TOSTI, *Exporters and importers of services: firm-level evidence on Italy,* The World Economy, **TD No. 877 (September 2012).**

GIACOMELLI S. and C. MENON, *Does weak contract enforcement affect firm size? Evidence from the neighbour's court,* Journal of Economic Geography, **TD No. 898 (January 2013).**

NATOLI F. and L. SIGALOTTI, *Tail co-movement in inflation expectations as an indicator of anchoring,* International Journal of Central Banking, **TD No. 1025 (July 2015).**

RIGGI M., *Capital destruction, jobless recoveries, and the discipline device role of unemployment,* Macroeconomic Dynamics, **TD No. 871 (July 2012).**

SEGURA A., *Why did sponsor banks rescue their SIVs?,* Review of Finance, **TD No. 1100 (February 2017).**