What Do Online Listings Tell Us about the Housing Market?*

Michele Loberto,[a] Andrea Luciani,[a] and Marco Pangallo[b]
[a]Banca d'Italia
[b]Sant'Anna School of Advanced Studies

Since the Great Recession, central banks and macroprudential authorities have been devoting much more attention to the housing market. To properly assess trends and risks, policymakers need detailed, timely, and granular information on demand, supply, and transactions. This information is hardly provided by traditional survey or administrative data. In this paper, we argue that data coming from housing sales advertisements (ads) websites can be used to overcome some existing deficiencies. Using a large data set of ads in Italy, we provide the first comprehensive analysis of the problems and potential of these data. We show how machine learning tools can correct a specific bias of online listings, namely the proliferation of duplicate ads that refer to the same housing unit, increasing the representativeness and reliability of these data. We then show how the timeliness, granularity, and online nature of these data make it possible to monitor in real-time housing demand, supply, and prices.

JEL Codes: C44, C81, C31, R21, R31.

*The views expressed in this paper are those of the authors and do not reflect the views of Banca d'Italia. We are extremely grateful to Immobiliare.it for providing the data and for their assistance. We thank all the seminar and workshop participants at Banca d'Italia, Bank of England, Danmarks Nationalbank, Eurostat, and JRC Centre for Advanced Studies (JRC-CAS), for useful comments and discussions. Special thanks go to Attilio Mattiocco, Juri Marcucci, and Diana Nicoletti for their assistance and feedback in all the phases of this project. An earlier version of this paper circulated under the title "The Potential of Big Housing Data: An Application to the Italian Real-Estate Market." Marco Pangallo performed most of his work on this project while he was affiliated with the Mathematical Institute and Institute for New Economic Thinking at the University of Oxford. Marco Pangallo acknowledges support from the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Dynamic and Multiscale Systems — Postdoctoral Fellowship Award. All mistakes are our own. Author e-mails: michele.loberto@bancaditalia.it, andrea.luciani@bancaditalia.it, marco.pangallo@santannapisa.it.

1. Introduction

Following the Great Recession of 2007–09, policymakers have been following housing market trends more closely. Housing is the main component of household wealth and is one of the main drivers of private consumption (Mian, Rao, and Sufi 2013). Moreover, housing markets are highly relevant for financial stability, as housing price bubbles have historically been dramatically damaging events (Jordà, Schularick, and Taylor 2015). Therefore, assessing the trends and risks in the housing markets is crucial for central banks.

Choosing the right policy mix hinges on the availability of detailed, granular, and timely information about housing demand and supply. For instance, policymakers may want to choose different policies depending on whether excessive housing price growth is due to exuberance on the demand side or to low supply. Additionally, this information should be available at the country level but also at a more granular level. Indeed, housing bubbles may occur in local markets or even in specific market segments (Landvoigt, Piazzesi, and Schneider 2015), even if no signs of imbalances are detected at the country level. Furthermore, to take timely action, information on demand and supply should ideally be available in real time.

Collecting detailed, granular, and timely data on the housing market has been traditionally challenging. Microdata on home sales are available to researchers only in a few countries and with a significant temporal lag. Moreover, they may show limitations in the spatial and temporal dimension, or in their informational content.[1] Most importantly, extracting information on demand and supply from home sales may require strong identifying assumptions, as transactions represent "equilibrium" points.

In this paper, we show how online data from marketplace websites (such as Zillow) can fill some gaps, providing valuable

[1] For example, the United Kingdom is one of the few countries where microdata are available. However, those data contain inadequate information about the physical characteristics of homes. Many papers on the U.S. housing market have used data from Multiple Listing Services (MLS), which are pools of real estate brokers sharing information about properties to make the matching between buyers and sellers more efficient. These data report details on the price and physical characteristics of homes. Yet, because there exist many different MLSs, studies usually focused on a limited geographical area (for example, see Han and Strange 2015). In Italy, administrative microdata on housing transactions are not available for research because of privacy concerns.

information to support policy choices. Our aim is twofold. First, we investigate the measurement issues and show how to improve the representativeness and reliability of online listings data. Second, we show how these data can provide detailed, timely, and granular information on housing demand, supply, and prices, which would be harder to get from traditional sources.

Our analysis is based on a large database containing all housing sales advertisements (ads) published on Immobiliare.it, the most popular online portal for real estate services in Italy. Similar analyses could be performed using data from similar websites, such as Zillow or Trulia in the United States or Zoopla in the United Kingdom. From these sources, we can retrieve real-time and detailed information about listed dwellings, including physical characteristics, location, time on market, and asking prices. Compared with traditional listing data collected by real estate professional associations, online listing data also allow for real-time monitoring of buyers' search behavior (Piazzesi, Schneider, and Stroebel 2020), as we discuss below.

Despite the wealth of information on the housing market that these data provide, data generation could be biased in several ways. As with all non-survey or non-universal administrative data, online listings data may lead to non-representative results or feature measurement error. Additionally, online listings posted on marketplace websites may have a peculiar issue: There could be two or more *duplicate* ads referring to the same housing unit. This is a common problem in our data set, but we think that it is not just a website-specific or Italian-specific issue. For instance, Kolbe et al. (2021) report the same issue for ads on ImmobilienScout24, the largest real estate platform in Germany. To identify duplicates, we propose a procedure using machine learning algorithms, as duplicate identification based on geographical coordinates or heuristic rules is not sufficiently precise. We show that the duplicate bias is not a serious issue for a few applications, such as monitoring housing market trends at the country level. However, we also demonstrate that it is a serious problem when granularity and high frequency matter for identification. As a consequence, the results of regressions that use duplicate ads instead of unique listings may be greatly biased.[2]

[2] For instance, we show that the odds of reducing the asking price if the property on sale does not attract enough interest is reduced tenfold when considering

In terms of applications of online listings data, we first show that the number of views to the ads' webpages is a good proxy of housing demand. Indeed, when individuals visit the webpage of an ad, they convey information about the characteristics, location, and price of the home they are searching for. By aggregating all this information, we can understand which area households are searching more intensely and what they are looking for.[3] At the micro (dwelling) level, high online interest predicts lower time on market and lower probability that a downward revision of the asking price occurs. By aggregating the number of page views, we can build a measure of market tightness. We show that this indicator is a good predictor of housing prices, as suggested by the recent literature (Carillo, de Wit, and Larson 2015; Wu and Brynjolfsson 2015; van Dijk and Francke 2018).[4]

We also show that online listings are an effective tool for monitoring the number of homes for sale (so-called market inventory). Although housing supply is usually defined as the total stock of homes (Glaeser and Gyourko 2018), policymakers should focus on market inventory as a measure of short-medium run housing supply. Home sales variation is mainly driven by changes in the number of listings, and households take market conditions into account before deciding whether to put their home up for sale (Ngai and Sheedy 2020). Moreover, also the composition of market inventory changes over time with market conditions. We show that the quality of listed existing homes improves with better market conditions, as measured by housing price growth.

Finally, we discuss under which conditions listing prices can be used to nowcast and forecast sale prices (Anenberg and Laufer 2017). We stress that a good estimate of the average discount to the initial asking price is needed. When this discount is constant, asking

duplicate ads instead of unique listings. This is because brokers are likely to post a new ad when revising the price, and if this is not taken into account price revisions look excessively rare.

[3] This proxy of housing demand is complementary to web searches, which have been used recently by Piazzesi, Schneider, and Stroebel (2020).

[4] We already investigated the possibility of using webpages' views as a proxy of housing demand in a previous publication directed at a different audience (Pangallo and Loberto 2018). Here, we adopt a different econometric approach, and the sample is twice as long, highlighting the robustness of our findings.

prices are a good proxy of transaction prices. However, since discount changes with market conditions, asking prices may be a poor predictor of sale prices. In this case, auxiliary information is needed to improve the forecast (Anenberg and Laufer 2017; Lyons 2019).

This paper is organized as follows. Section 2 illustrates the main institutional details and trends of the Italian housing market. Section 3 describe the Immobiliare.it ads data set and discuss the main issues with online ads. In Section 4 we show how online listing data can be used to measure demand, supply, and housing prices. Section 5 concludes.

2. The Italian Housing Market

In this section, we describe the main trends and institutions of the Italian housing market.

The 2011 sovereign debt crisis had a strong impact on the housing market. From 2011 to 2013, home purchases and sales fell by about 30 percent and only resumed growth in 2014 (Figure B.1 in Appendix B). Housing prices experienced a more moderate but more persistent decline (Figure B.1, panel B): Between 2011 and 2018, they fell cumulatively by about 20 percent. The average time on market surged from seven to nine months between 2010 and 2015, but returned to pre-crisis levels since mid-2016 (Figure B.1, panel C). The average discount obtained by buyers relative to sellers' asking prices has followed a similar pattern, varying between 10 and 15 percent. Trends in home sales and prices diverged across geographic areas. In 2016–18, which is the period we primarily focus on in this paper, home prices were still declining in most cities. However, they had returned to growth in many large cities (Figure B.1, panel D).

In Italy, about half of all households' home purchases are financed through a mortgage loan. The relative amount of the mortgage is generally not very large: The average loan-to-value is about 60 percent. Transaction costs associated with purchasing a home depend on several factors. Costs include transaction taxes, notary fees, brokerage fees, and mortgage-related costs. Estimating the impact of transaction costs on the value of a purchased home is difficult.[5]

[5] Some costs are not proportional to the value of the home. Other costs are partially tax deductible. Moreover, many of these costs are lower if the new owner

Considering a home to be occupied by the owner and worth 100,000 euros, transaction costs can be up to 13 percent. In other cases, transaction costs can be up to 20 percent (e.g., dwellings purchased for investment purposes).

Most importantly for the focus of this paper, about half of total home sales are intermediated by real estate brokers. Real estate brokers are essential in cities and metropolitan areas. By contrast, in suburbs and rural areas most transactions do not involve an intermediary. Moreover, in Italy open listings agreements are possible, in the sense that two or more real estate agents are entitled to sell the same dwelling.

List prices are not legally binding, and the seller can always refuse to sell to a potential buyer. In general, the buyer and the seller negotiate the final price and other contractual arrangements. When a broker is involved in a sale, the seller cannot simultaneously negotiate with multiple buyers, which rules out bidding wars. Usually, the final price is below the listing price. Indeed, during 2016–18, the average discount compared with the initial asking price was about 12 percent, and the final price was equal to or higher than the initial asking price only in about 5 percent of transactions (Italian Housing Market Survey).[6]

3. Data

We analyze a data set of home listings published on Immobiliare.it, the largest online portal for real estate services in Italy. This data set covers the whole country. However, since small towns and villages may have representativeness issues, we only consider listings in the 109 main cities that are capitals of the NUTS-3 Italian regions. About 18 million people live in these cities, and the number of home sales is about one-third of all transactions in Italy.

bought the home as a primary residence. The total cost depends on home value, buyer income, and the reason for the purchase.

[6] The sale price may be higher than the asking price for various reasons other than bidding wars. For instance, the buyer may have particular requirements for finalizing the sale or taking possession of the home. Alternatively, the transaction includes additional amenities compared with the initial offer (e.g., a garage).

Immobiliare.it provides us with weekly snapshots of all ads visible on their website every Monday, from January 4, 2016 until December 31, 2018. For 2015 only quarterly snapshots are available.[7]

For each ad, we have detailed information about the physical characteristics and exact location of the dwelling (see Appendix A for the complete list of variables). We keep track of all variations concerning asking prices and number of times that the webpage of the ad has been visited (*clicks*). We also know the date when the ad was created and the date when it was removed. Unfortunately, we do not know if a property was sold or withdrawn from the market.

The data set counts 1,402,798 ads. Since we observe ads at a weekly frequency, the total number of records is almost 28 million. Most ads remain unchanged between two weekly snapshots, as the average turnover is about 5 percent. About 92 percent of the ads are posted by real estate agents; the remaining ads are posted by households or construction firms.

We divide the territory of each city into local housing markets using the partition developed by OMI, a branch of the Italian Tax Office. The elements of this partition are contiguous areas of the city that satisfy strict requirements in terms of homogeneity of housing prices, urban characteristics, socioeconomic characteristics, and the endowment of services and urban infrastructures. This partition is periodically revised to satisfy these criteria and better approximate local housing markets. The latest revision dates back to 2014. Thus, unlike census tracts, these zones can be considered as "local housing markets." For each of these zones, OMI estimates the minimum and maximum housing price per square meter on a six-month basis. Table B.3 in Appendix B reports some descriptive statistics about these local housing markets.

Finally, we use information coming from the Italian Housing Market Survey, a quarterly survey covering a large sample of real estate agents. A detailed discussion about all data sources can be found in Appendix A.

[7] Data are available for the following days in 2015: January 5, April 25, September 7, December 28.

3.1 Duplicate Listings

The use of new, unconventional data sources is becoming increasingly common. However, using these data requires identifying potential biases that could make the data unrepresentative of the phenomenon under analysis.

The main concern with several housing marketplace websites—such as Immobiliare.it, Craigslist, Zoopla, ImmobilienScout24, Idealista, and many others—is the difficulty to strictly control the content of the ads published by the users. These websites are market platforms that allow home sellers and brokers to advertise the sale of a home in exchange for a fee. Rigorous checks on ads published by users are costly or even unfeasible. Consequently, before using online listings for economic analysis, it is necessary to assess their reliability.

A key issue is that multiple ads can be associated with the same dwelling. That may be due to various reasons. First of all, under open listing agreements, each broker could publish a different ad. Additionally, a broker could post multiple ads for the same home. In particular, the broker may delete the old ad and create a new one to refresh the time on market of the listing.[8] Furthermore, when a mandate to sell expires, the home seller may sign a listing agreement with a new agent that publishes a new ad.[9]

We are concerned with duplicates for several reasons. First, duplicate ads may provide a biased representation of housing supply, especially at granular levels. Second, the presence of duplicates may not be random but associated with the physical characteristics of the home, the urgency of the owner to sell soon, or difficulties in finding a buyer. Third, the disappearance of a duplicate ad does not necessarily correspond to a sale or a withdrawal. Likewise, new ads do not necessarily correspond to new properties entering the market.

[8] Indeed, many potential buyers search on the website from the most recently published ad to the oldest. Moreover, posting a new ad provides greater visibility to the listing because potential buyers receive notifications about new listings through the email-alert service.

[9] If the old agent does not immediately delete the ad, and the new agent posts a new one, two ads for the same dwelling exist simultaneously. Even if this does not happen, and the two ads are not simultaneously visible on the website, we still need to know that these ads refer to the same dwelling.

We identify duplicate ads using machine learning tools.[10] We depart from the original data set of ads and build a new data set of listings. In the latter, the unit of observation is a home instead of an ad. We use machine learning tools because there is no exact matching between characteristics of the homes reported in two duplicate ads. In our experimentation, using pre-specified heuristic rules (such as, consider apartments whose price difference is smaller than 5 percent) to identify duplicates was not particularly successful. Instead, machine learning algorithms autonomously learn the best criteria that identify duplicates provided the training sample is sufficiently large. Moreover, these algorithms can effectively exploit the partial similarity between dwellings' characteristics, which is crucial because different brokers can provide partially different information about the same feature. The primary input for our algorithm is location. However, other variables play a significant role (e.g., asking price, size, amenities).

After identifying duplicates, we combine them as if they were a single ad. Our final data set includes about 940,000 homes, which we will also refer to as "listings."[11] Tables B.1 and B.2 in Appendix B report descriptive statistics about the sample. In Appendix C, we provide all details about the cleaning procedure.

Once we get rid of duplicates, listing data are much more consistent with official statistics than the original ads (Table 1). The average time-on-market measure on listings data is consistent with the results of the Italian Housing Market Survey and is about two

[10] In general, it is not possible to identify duplicate ads by the address. Both in urban and in rural areas, addresses—as generally reported in the ads—may not uniquely identify homes. For example, for condo apartments in cities, sellers usually report the address of the building, and multiple apartments from the same building could be simultaneously on sale. In rural areas, non-unique addresses are also common. Georeferencing the ads may help in rural settings, where houses are more spread out. However, it is less useful in urban settings with a high concentration of homes.

[11] Duplicates are associated with a small share of listings (about 20 percent). Open listing agreements with many agents seem to be a primary source of duplicate ads. We also observe that the duplicate ads of a property appear over time: new ads are created while old ads are deleted, giving rise to a considerable number of delistings and new listings. Finally, the share of duplicates over total ads increases with city size, and there is significant variability across cities. More details about the distribution of duplicates can be found in Appendix C.2.

Table 1. Sales and Time on Market (months), for Ads, Listings (homes), and Official Data

	Sales			Time on Market		
Year	Ads	Listings	Italian Tax Office	Ads	Listings	Survey
2016	335,181	207,120	178,690	5.1	6.7	7.5
2017	312,584	187,443	186,657	4.9	6.7	6.3
2018	321,840	189,505	197,506	4.4	6.3	6.6

months longer than the average duration of ads.[12] Also, the number of delistings is much lower than the number of removed ads and broadly in line with the number of home sales, once considering that (i) delistings include withdrawals; (ii) not all the homes sold have been listed online. The correlation between the number of delistings and home sales at the city level is 0.96 (Figure 1, panel A). The fit is also excellent for housing prices. The correlation between average asking and sale prices of apartments at the local housing market level is 0.93 (Figure 1, panel B).[13]

3.2 Assessing the Distortions Generated by Duplicates

The presence of duplicates does not introduce significant distortions when estimating the trend of prices and delistings at the country level (Figure 2). However, the measurement error could be more significant at more granular levels.

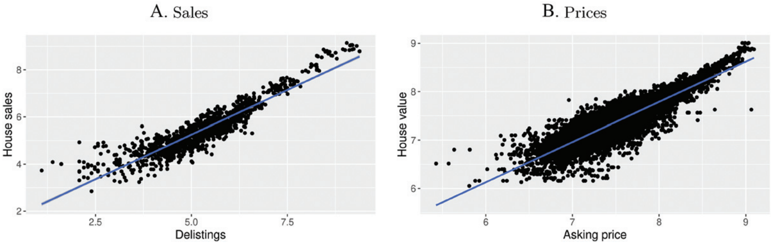
To quantify the measurement error for average asking prices and delistings, we estimate the following ordinary least squares (OLS) regressions:

$$Y_{it} = \alpha + \beta X_{it} + \varepsilon_{it}, \tag{1}$$

[12] We find a significant deviation only for 2016, when listings underestimate time on market. That is plausible because some of the homes listed in 2016 may have been initially listed in 2015. Since we only observe quarterly snapshots for 2015, we may not reconstruct the complete history of dwellings delisted in 2016.

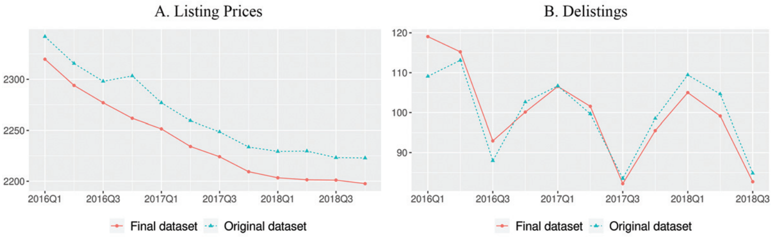
[13] Furthermore, we compute the ratio between the listing prices and actual home values per square meter for each local housing market. On average, during 2016–18, the discount on asking prices was about 12 percent, a value consistent with evidence provided by the Italian Housing Market Survey.

Figure 1. Home Sales, Delistings, and Prices



Note: Official data on home sales and prices are provided by the Italian Tax Office. Home sales and delistings are at city level and data are quarterly. Prices are at the local housing market level and data are half-yearly. All variables are in logs.

Figure 2. Housing Prices and Sales in Italy



Note: Asking prices (panel A) are measured in euros per s.m. Delistings (panel B) were converted in index numbers, where 100 is the average number of estimated delistings (or removed ads) between 2016 and 2018.

where Y_{it} is the value of a statistic computed on the final data set of listings and referring to geographical area i during quarter t. Y can be the average asking price or the number of delistings. For both variables, we consider levels and year-on-year growth rates at a quarterly frequency. X_{it} is the same statistic as calculated on the original data set of ads. The geographical area can be a city or a local housing market because the measurement error can have a different magnitude depending on spatial granularity.

Table 2 reports the results. When considering the levels of asking prices and delistings, the distortion due to duplicate ads seems negligible. There is much heterogeneity between cities and local markets, both in terms of prices and number of delistings. Both data sets can

Table 2. Measurement Error Due to Duplicate Ads

	Levels			**Y-o-y Growth Rates**		
	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{R^2}$	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{R^2}$
City Level: Asking Prices	25.374	0.981	0.999	–0.295	0.916	0.917
City Level: Delistings	92.878	0.469	0.984	–3.786	0.863	0.816
Local Market Level: Prices	30.329	0.990	0.994	–0.332	0.817	0.742
Local Market Level: Delistings	2.981	0.513	0.932	–1.660	0.785	0.639

account very well for this spatial heterogeneity, and this explains the almost perfect correlation observed in the left panel of Table 2.

However, the regressions with quarterly year-on-year growth rates show that the measurement error is higher for delistings and is always significantly larger for local housing markets. Figures B.2 and B.3 reveal that the measurement error for asking prices occurs in markets with low number of listings. Delistings are harder to measure because their number is generally low, even in the largest local markets.[14] This prevents the use of the original data for most analyses where granularity and high frequency matter for identification.[15]

Moreover, the presence of multiple ads related to the same dwelling is not random. Indeed, home sellers or brokers post multiple ads to attract more attention. In Appendix D we show that using the original data set of ads implies an oversampling of homes that are relatively expensive and less attractive given their location and characteristics. This implies that, by using ads, we overestimate average listing prices. Moreover, lower attractiveness is associated with higher time on market and propensity to revise the asking price downward. Therefore, using the original data would imply severe distortions when analyzing the microstructure of the housing market (see footnote 23 for a concrete example).

[14] Table B.3 in Appendix B shows that the annual median number of delistings across local markets is 28. As local markets must be homogenous areas, their size is necessarily small.

[15] For example, Anenberg and Kung (2014) assess the impact of foreclosures in small neighborhoods by exploiting the timing of listings' entry and exit into the market. The presence of duplicate listings would seriously impair the representativeness of listing data for similar analyses.

Summing up, the measurement error implied by keeping duplicate listings in the sample is sizable at the granular level, particularly when we look at dynamics. However, it is possible to use the original data set without incurring in significant measurement error in several cases. For example, it is possible to use ads to monitor housing market trends at the country level or for sufficiently large areas. Unfortunately, the presence of duplicates is a substantial disadvantage that prevents the full exploitation of these data.

4. Measuring Demand, Supply, and Prices

This section discusses the potential of online listing data and illustrates their complementarity with traditional statistical sources.

Based on online listings, we can build timely indicators on market inventories (homes on sale), liquidity, and asking prices. By exploiting the richness of details about home characteristics and location, we can detect any diverging pattern across market segments or geographical areas. Yet, similar high-frequency data can be retrieved from some traditional providers, such as MLS or real estate broker associations. We argue that the most significant potential of these data is in the information generated by users as they browse the site, which provides insight into the search activity of potential home buyers, i.e., housing demand. Therefore, compared with traditional sources, online listings allow monitoring both sides of the housing market.

4.1 Demand

Online activity leaves digital traces of human behavior. When individuals visit an ad's webpage, they convey information about the characteristics, location, and price interval of the home they are searching for. By aggregating all this information, we can understand which area households are searching more intensely and what they are looking for. We can observe housing demand.

In our data set, we know how often website users visited the webpage of an ad during each week (clicks). Clicks are complementary to information about online housing demand that has been used in other studies (see, e.g., Piazzesi, Schneider, and Stroebel 2020), namely web searches, i.e., queries where the user specifies the location, characteristics, and price range of the home she is looking for. In principle, web searches and clicks do not convey the same information. People may search for homes with a bundle of characteristics that cannot be found in the market. In this case, we cannot observe clicks that map to those preferences. So, we do not observe the actual preferences of potential buyers. However, it is plausible that potential buyers would somewhat adapt their preferences to the composition of supply. Thus, we think that there is no loss of generality in using clicks instead of web searches for a large class of applications. Moreover, clicks are easier to be used than web searches. Home listing websites usually allow "map search," letting users specify a polygon on the map to look for a home. Extracting and aggregating this type of information about buyers' preferences is hard (Rae 2015; Piazzesi, Schneider, and Stroebel 2020). However, this problem does not arise when considering visits to webpages. Finally, clicks are available for each listing. They can be used to proxy the interest of potential buyers for each home.

To show that online interest is a proxy of housing demand, we proceed as follows. We test whether online interest for a dwelling is correlated with the time it has been on the market and with price revisions. If the webpage of a listed dwelling gets many views, it is plausible that many households are searching for that type of home. Therefore, our first hypothesis is that high online interest is associated with a shorter time on market. Moreover, it is plausible that the price interval is a key searching criterion set by all potential buyers. Suppose many households search in a given price interval for a dwelling with a particular bundle of characteristics, ceteris paribus. Then, it is less plausible to observe downward revisions of the asking price for these dwellings. Therefore, our second hypothesis is that higher online interest implies a lower propensity to revise the asking price.

We build the variable $ONLINT$ to quantify the relative interest in a particular dwelling compared with the other dwellings in the same local housing market.[16] $ONLINT$ is the average daily number of clicks on the home in the first three weeks since its initial listing, divided by the average daily number of clicks in its local

[16] We cannot use the variable $CLICKS$ because homes are listed at different times and for different periods.

housing market during the same period. Thus, when $ONLINT > 1$ it means that the home received more online interest than the average home in the same local housing market, and when $ONLINT < 1$ the reverse is true.

We consider the number of clicks in the first three weeks, as it strikes a compromise between two different problems. If we look at number of clicks over a period that is very long, say two months, online interest may be endogenous. For instance, downward price revisions that occur after a month could likely trigger a change in online interest. By contrast, a period that is too short, say a week, leads to more noise, as we observe ads only once per week.[17] Three weeks is a period that is sufficiently long to mitigate measurement error, while short enough to make it unlikely that price revisions occurred.

We restrict our sample to dwellings that have been initially listed between January 2016 and June 2018 because the observation period for any price revisions or delisting ends in December 2018. We also drop listings with duplicate ads to avoid the bias identified in Appendix D.

To test the relation between time on market and online interest, we estimate the following Weibull regression model:

$$\log(TOM_i) = \beta ONLINT_i + \delta \mathbf{X}_i + \sigma \eta_i, \tag{2}$$

where η_i are i.i.d. random variables following an extreme value distribution. TOM is the time on market—measured as the number of days between the delisting and the first listing.[18] The vector $\mathbf{X}$ includes the physical characteristics of the dwelling. We control for the relative asking price per square meter because relatively more expensive homes are less viewed.[19] We add year-quarter dummies

[17] We observe ads every Monday, but an ad could have been posted on any day of the previous week. Thus, the number of days on which online interest is measured may differ between ads.

[18] Unfortunately, we do not observe if a home has been withdrawn from the market or sold. Then, our variable TOM may be a poor proxy of the time on market. We believe that this is not the case because our measure of the time on market is consistent with survey estimates on average.

[19] The relative asking price is defined similarly to $ONLINT$ and is the ratio of the initial listing price per square meter to the average price in the local housing market during the first three weeks since initial listing.

Table 3. Online Interest

	Dependent Variable		
	TOM **(AFT)** **(1)**	***PRICEREV*** **(LOGIT)** **(2)**	***PRICE*** **(OLS)** **(3)**
ONLINT	–0.069*** (0.002)	–0.080*** (0.005)	
$DEMAND_{t-1}$			0.038*** (0.009)
$AVPRICE_{t-1}$			0.193*** (0.069)
Log(*Scale*)	–0.098*** (0.001)		
Fixed Effects		Local Mkt.	Local Mkt.
Temporal Dummies	Year-Quarter	Year-Quarter	Year-Quarter
Observations	324,906	313,777	427,165
R^2	—	—	0.78

referring to the period of first listing of a home to control for common time-varying unobservables.

In column 1 of Table 3, we report the results. The coefficient associated with *ONLINT* is statistically significant, and its sign confirms our hypothesis. A one-standard-deviation increase in online interest in the early stage of the listing period implies a $e^{-0.069} = 0.93$ times shorter time on market.[20] Notice that the same factor would shrink to 0.70 if online interest were measured over the whole lifetime of the listings. However, in this case, the claim of exogeneity would be hard to support. As we show below, lower online interest implies a greater propensity to revise the asking price downward. Price revision affects time on market (de Wit and van der Klaauw 2013), and likely the online interest of potential buyers.

[20] The results of the Weibull regression can be alternatively interpreted in terms of a proportional hazard model. The hazard ratio associated with a one-standard-deviation increase in online interest is computed as $e^{-\left(\frac{-0.069}{0.907}\right)} = 1.08$, where 0.907 is the scale parameter.

To test whether online interest predicts the occurrence of price revisions, we introduce a binary variable $PRICEREF$. This variable takes value one if the asking price of the dwelling is revised downward and zero if it is not revised or revised upward.[21] Then, we run the following logistic regression:[22]

$$\log\left(\frac{p_{ijt}}{1-p_{ijt}}\right) = \beta ONLINT_{ijt} + \delta\mathbf{X}_{ijt} + \varepsilon_{ijt}, \tag{3}$$

where $p \equiv Prob(PRICEREF = 1)$ and, as in the previous regression, we control for the relative asking price per square meter and the physical characteristics of the dwelling. We also add local housing market and year-quarter fixed effects. We estimate that a one-standard-deviation increase in the relative number of clicks is associated with a 7 percent reduction in the odds of a downward price revision (Table 3, column 2).[23]

Finally, we test if online interest predicts aggregate housing market dynamics. We build an indicator of housing demand in each local housing market. We expect that aggregate online interest is correlated with housing prices. In particular, we hypothesize that stronger demand is associated with higher growth in housing prices, as suggested in the recent literature (Carrillo, de Wit, and Larson 2015; Wu and Brynjolfsson 2015; van Dijk and Francke 2018).

We construct the quarterly variable $DEMAND$, defined as the average daily number of clicks per listing in a local housing market. To deal with the potential endogeneity of this measure of demand to prices, we choose the following econometric strategy. We investigate whether the entry price of a new listing is positively affected by

[21] We consider only the case of downward price revisions for two reasons. First, the number of upward revisions is relatively small. Second, a price increase can be motivated by changing terms of trade or some unobserved change in dwelling quality.

[22] Pangallo and Loberto (2018) show that the relation between prices and online interest also works the other way around. We find that a 1 percent higher price is associated with a 0.66 percent lower number of clicks. We also show that this elasticity has a causal interpretation.

[23] If we did not run the deduplication procedure, running the same logistic regression on the ads data set would yield a 0.7 percent reduction in the odds of a downward price revision instead of a 7 percent reduction. This difference is explained by the fact that brokers are likely to post a new ad when revising the price, as it would attract more attention.

the intensity of search activity in the local housing market in previous months. Suppose online searches are a proxy for actual visits to homes for sale. In that case, real estate agents observe an increase in the market's tightness. Consequently, they may likely suggest higher listing prices to new sellers. We consider the entry prices of new listings; otherwise, average search activity in period $t-1$ would be correlated with prices in period t because of dwellings listed in both periods. This would be problematic, especially in smaller local markets.

We run the following OLS regression:

$$\begin{aligned}\log(P_{ijt}) = \alpha_j + \zeta_t + \beta_1 \log(DEMAND_{j,t-1}) \\ + \beta_2 \log(\bar{P}_{j,t-1}) + \delta \mathbf{X}_i + \varepsilon_{ijt},\end{aligned} \tag{4}$$

where P_{ijt} is entry price of new listing i, located in local market j in quarter t. We control for past average asking prices, $\bar{P}$, and dwellings' characteristics. α_j control for local housing market unobservables; ζ_t is a set of year-quarter dummies. The results reported in column 3 of Table 3 confirm that online interest is a good leading indicator of prices. The elasticity of the entry prices of new listings with respect to past average search activity is about 4 percent.

In sum, webpage clicks are a valuable tool for measuring housing demand in real time and understanding buyers' preferences. Moreover, differently from buyers' web searches, clicks are easy to handle. They allow building a measure of demand for a specific home, not only for a neighborhood or a typology of dwellings.[24]

4.2 Supply

Housing supply is usually defined as the total number of dwellings, without considering whether they are on sale or not (Glaeser and Gyourko 2018). Consequently, housing supply increases because new homes are built, and it is downwardly rigid because of the durable nature of dwellings.

In the short or medium run, this definition is not necessarily the most suitable. Indeed, the number of homes potentially available for

[24] In an earlier version of this paper, we showed that the variable $DEMAND$ is a good predictor to forecast the trends of average asking price and liquidity of a local housing market. The results are available upon request.

sale changes over time, at least for two reasons. First, homeowners' decision to move into a new home can depend on macroeconomic developments and housing market conditions (Anenberg and Bayer 2020; Ngai and Sheedy 2020). Second, new homes may enter the housing market because of worsening conditions in the rental market. Owners of vacant homes always have the option to search for either a buyer or a tenant (Krainer 2001; Head, Lloyd-Ellis, and Sun 2014; Liberati and Loberto 2019).

Since the number and type of homes that are for sale may not correlate with the total number of homes, in some cases it is more reasonable to look at listings as a measure of housing supply (see Mian, Sufi, and Trebbi 2015; Piazzesi, Schneider, and Stroebel 2020).[25] For example, Ngai and Sheedy (2020) show that home sales variation is mainly driven by listings instead of a change in matching efficiency in the housing market. Here, we show that the housing supply composition is not time invariant and may change over the housing market cycle.

To show that the average quality of the homes offered for sale changes with the real estate cycle, we consider four variables that measure the average quality of listings in each city at a half-yearly frequency. We define $FLOORAREA$ as the logarithm of the average floor area of listings (measured in square meters); $BATH$ is the share of listings with at least two bathrooms; $GARDEN$ is the share of listings having a private garden; $TERRACE$ is the share of listings having a terrace. To measure the timing of the housing market cycle in each city, we use the logarithm of a hedonic asking price index ($HEDON$). We consider this variable because hedonic price indices are by construction not affected by the physical characteristics of dwellings. Therefore, they are uncorrelated with changes in average home size and quality.[26] Finally, we consider only existing dwellings. In this way, we can show that the home supply composition changes with housing prices and does not depend on the characteristics of newly built houses.

[25] It is fair to say that this distinction is the same that arises in labor market statistics, in which only people that are already working or searching actively for a job are considered inside the labor supply.

[26] Otherwise, an increase in the home average size is associated with a decrease in average asking prices. Indeed, larger homes are ceteris paribus priced at a lower price per square meter.

Table 4. Quality of Listed Dwellings and House Prices (half-yearly data)

	Dependent Variable			
	FLOORAREA **(1)**	***BATH*** **(2)**	***GARDEN*** **(3)**	***TERRACE*** **(4)**
HEDON	0.114** (0.050)	0.143*** (0.031)	10.491*** (3.001)	8.946** (4.096)
Fixed Effects	City	City	City	City
Temporal Dummies	Year-Semester	Year-Semester	Year-Semester	Year-Semester
Observations	546	546	546	546
R^2	0.153	0.057	0.036	0.087

Note: Results of a panel fixed-effect estimation, using the *within* transformation. *HEDON* is the logarithm of a hedonic city-level house asking price index.

We estimate the following model for city i and half-year t:

$$Y_{i,t} = \alpha_i + \zeta_t + \beta HEDON_{i,t} + \varepsilon_{i,t}. \tag{5}$$

The dependent variable Y is one among $FLOORAREA$, $BATH$, $GARDEN$, and $TERRACE$. We add city fixed effects and time dummies. Table 4 reports the results of a panel fixed-effect estimation, using the *within* transformation. Housing supply in cities with stronger housing price dynamics is characterized by a larger average floor area and a higher number of bathrooms of listed homes. We also find an increase in the share of listings with a private garden or a terrace. Results would be qualitatively similar when using the housing prices series estimated by the Italian Tax Office (see Table B.4 in Appendix B).[27] Therefore, housing price increases are associated with a better quality of housing supply.

[27] The limited temporal dimension of our data set prevents a comprehensive analysis of potential non-stationarity in the data. However, we believe that introducing city-level fixed effects eases those concerns. Table B.5 in Appendix B reports consisting evidence.

4.3 Listing Prices

Another potential strength of listing data is the observation of sellers' asking prices. Anenberg and Laufer (2017) show that listing prices can be used to predict a standard housing price index over a short-term horizon. Indeed, listing prices are observed in real time, while sale prices are usually available with a significant lag. However, the determination of the listing price is ultimately a seller's decision possibly made in conjunction with a listing broker. Therefore, it is reasonable to question the ability of listing prices to track sale prices.

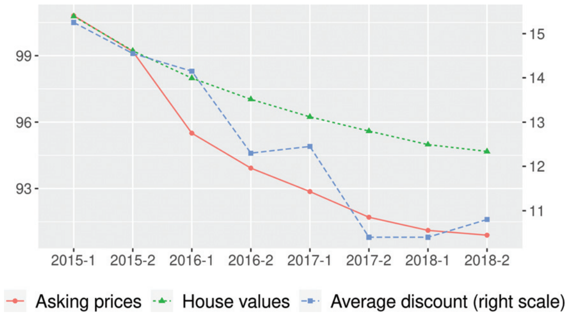
In Italy, the sale price is almost always below the list price. In this case, the asking price dynamics is a good proxy for sale price variations only if the average discount to asking prices obtained by buyers in the bargaining process were stable.[28] Since the outside option of both buyers and sellers in the bargaining process is affected by the general market conditions, the average discount on asking prices changes over time. In Italy, the discount increases during market downturns and decreases during market recoveries. Consequently, a decrease in the discount implies that sale prices decrease less or increase more than asking prices. Potentially, in some periods, sale prices may increase while asking prices decline.

We illustrate this issue in Figure 3. Between the first semester of 2015 and the second semester of 2018, average home values declined by about 6.0 percent, while asking prices diminished by 9.8 percent. That is consistent with the observation that the average discount on asking prices decreased cumulatively by 4.4 percentage points over the same period.

In sum, using asking prices to predict sale prices may require auxiliary variables to improve the fit. For example, Anenberg and Laufer (2017) show that including variables correlated with the discount—such as time on market—improves the forecasting performance of listing prices. Similarly, Lyons (2019) shows that an index based on

[28] We can express the relation between asking prices, P_t^a, and sale prices, P_t, as $P_t^a = (1 - d_t)\, P_t$, where d_t is the average discount. The dynamic relation between asking and sale prices is therefore given by $P_t^a - P_{t-1}^a = P_t - P_{t-1} - (P_t d_t - P_{t-1} d_{t-1})$.

Figure 3. Prices (index 2015S1=100) and Average Discount (percentage points)



list prices that accounts for time on market can track sale prices very well.[29]

5. 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). They provide a lot of information that can be integrated by textual analysis and image processing. If coming from online sources, they are frequently available (on a much shorter timescale than administrative data). They 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). They are less structured and controlled (there might be hidden factors influencing the data-generation process). They 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 large data set of housing ads published on the leading online portal

[29] Lyons (2019) shows that the spread between the asking and the transaction prices can be decomposed in four components corresponding to distinct market processes that take place between the time of listing and when the transaction takes place.

for real estate services in Italy. We provide a comprehensive analysis of the strengths and weaknesses of these data to study housing markets. The main issue is the existence of a substantial share of duplicate ads, leading to a misrepresentation of the volume and composition of the housing supply. However, once this issue is fixed, the potential of these data is enormous, particularly in analyzing housing demand. For example, using these data Guglielminetti et al. (2021) show how the COVID-19 pandemic has influenced the demand for housing heterogeneously across different market segments in Italy.

Although our analysis is specific to the data set we use, we think our insights could be employed more generally as economists increasingly rely on online listings websites. For example, duplicates are likely to affect all listings websites that have no incentive to control the proliferation of duplicates, e.g., because they profit from the number of ads rather than from data quality. For home listings websites, this problem is exacerbated by open mandate agreements. In all countries where these agreements are possible, duplicate ads could arise from different agencies. We find it unlikely that website administrators could correct this bias. Yet, this paper shows that machine learning techniques can correct this distortion and make online listings a powerful tool for the real-time analysis of housing markets.

Appendix A. Data Sources

Listings. The source data which we obtained from Immobiliare.it are contained in weekly files. Starting from these snapshots, we construct six data sets. The main data set is the one with unique ads. Three data sets track the weekly change of asking prices, visits, and uses of the form to contact the agency that is shown on each ad (we do not use information on contacts in this paper; in Pangallo and Loberto 2018 we show that it provides equivalent information to the number of visits). The last two data sets contain information about real estate agents and the list of hash codes of the pictures associated to each ad (we will not use these data in this paper). The information available for each ad is reported in Table A.1.

Housing Prices. Twice per year, OMI (a branch of the Italian Tax Office) disseminates estimates of minimum and maximum

Table A.1. Information Contained in the Database Provided by Immobiliare.it

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

Note: For a complete description of the meaning of the variables, see Loberto, Luciani, and Pangallo (2018). Italics indicates that if variables are missing, we perform semantic analysis on the textual description of the ads to recover missing information.

home values in euros per square meter, P_l and P_h, at a very granular level. Home values are available for all OMI microzones—which are uniform socioeconomic areas roughly corresponding to neighborhoods—in Italian cities. P_l and P_h are estimated based on a limited sample of home sales and valuations by real estate experts. Further information is available at https://www.agenziaentrate.gov.it/wps/content/Nsilib/Nsi/Schede/FabbricatiTerreni/omi.

We define the average home value in neighborhood (OMI microzone) j as $\bar{P}_j = \frac{P_{lj}+P_{hj}}{2}$. The average home values at city level are estimated as a simple average of the $\bar{P}_j$. For further aggregation above the city level, we compute weighted averages of the cities' average home values. As weights, we use the stock of homes measured in the 2011 census. OMI estimates are not designed for statistical purposes, and the index we construct must not be considered as equivalent to a quality-adjusted price index.

In Italy, quality-adjusted (hedonic) housing price indices are disseminated by Istat, but their reference area is not consistent with our listing data, apart from three city-level indices that refer to the main Italian cities: Rome, Milan, and Turin.

Home Sales. Quarterly data about the volume of home sales in each city are disseminated by OMI.

Italian Housing Market Survey. The Italian Housing Market Survey is a quarterly survey that has been conducted by Banca d'Italia, OMI, and Tecnoborsa since 2009. It covers a sample of real estate agents and reports their opinions regarding the current and expected course of home sales, price trends, time on market, and terms of trade. See https://www.bancaditalia.it/pubblicazioni/sondaggio-abitazioni/ for further information.

Census Data. We retrieve detailed information on socioeconomic characteristics and stock of buildings in OMI microzones from the 2011 census. Istat census tracts are much smaller than OMI microzones (quantitatively, there are approximately 400,000 Istat census tracts over the Italian territory, as compared with 27,000 OMI microzones) and do not necessarily coincide with them. We perform spatial matching of the polygons representing the tracts and the microzones and impute the Istat variables to the OMI microzones 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 microzones, such that there is a 50 percent overlap for both, we impute 1,000 housing units to each of the two OMI microzones.

Appendix B. Additional Tables and Figures

Table B.1. Descriptive Statistics: Physical Characteristics and Location

Number of Observations	**936,126**
Surface (sm)	
Minimum	30
25th	70.00
Median	93.00
75th	126.00
Maximum	600
Mean	108.68
Std. Dev.	64.15
Type of Property	
Multi-family Residential Dwelling	847,008
Single-Family Home	89,118
Floor Level	
Ground Floor	122,670
Floor Level: 1–3	521,223
Floor Level: 4–	168,812
Multi-level	70,058
NA	53,363
Rooms	
Number of Rooms: 1	29,417
Number of Rooms: 2	194,115
Number of Rooms: 3	295,953
Number of Rooms: 4	240,358
Number of Rooms: 5 or More	147,063
NA	29,220
Bathrooms	
Number of Bathrooms: 1	548,843
Number of Bathrooms: 2	307,287
Number of Bathrooms: 3 or More	63,116
NA	16,880
Terrace	
Terrace: No	631,324
Terrace: Yes	304,802
Balcony	
Balcony: No	346,474
Balcony: Yes	589,652

(*continued*)

Table B.1. (Continued)

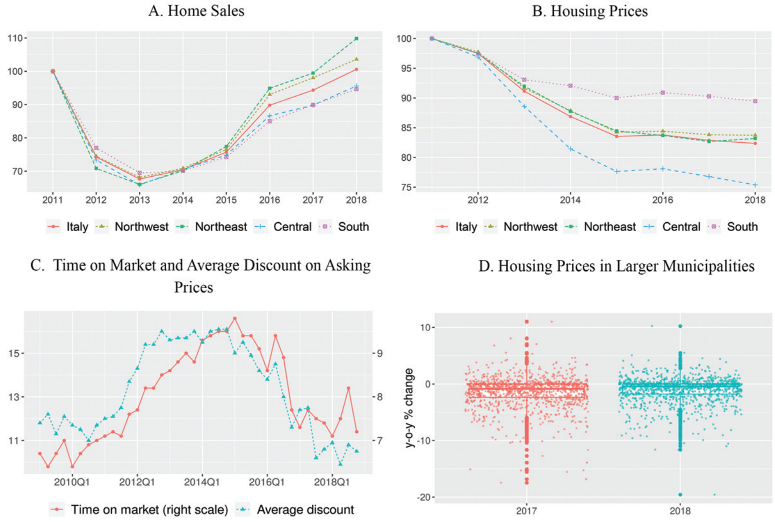
Number of Observations	**936,126**
Maintenance Status	
To Be Renovated	119,236
Good Conditions	349,691
Very Good Conditions	338,544
New-Built	85,188
NA	43,467
Kitchen Type	
Cooking Corner	165,086
Small Kitchen	121,955
Large Kitchen	558,580
NA	90,505
Utility Room	
Utility Room: No	664,806
Utility Room: Yes	271,320
Basement	
Basement: No	585,508
Basement: Yes	350,618
Garage	
No Parking Slot/Private Garage	598,023
Parking Slot	66,348
Private Garage	271,755
Garden	
Without Garden	582,787
Shared Garden	195,773
Private Garden	157,566
Janitor	
Janitor: No	861,184
Janitor: Yes	74,942
Elevator	
Elevator: No	423,983
Elevator: Yes	512,143
Air Conditioning	
Air Conditioning: No	204,779
Air Conditioning: Yes	221,551
NA	509,796
Heating	
Centralized Heating System	282,466
Autonomous Heating System	545,506
NA	108,154

(*continued*)

Table B.1. (Continued)

Number of Observations	936,126
Energy Efficiency	
Energy Efficiency: High	50,984
Energy Efficiency: Intermediate	108,744
Energy Efficiency: Low	476,659
NA	299,739
NUTS-1	
Northwest (ITC)	301,455
Northeast (ITH)	163,613
Central (ITI)	306,086
South and Insular (ITF-G)	164,972

Figure B.1. Main Trends in the Italian Housing Market



A. Home Sales

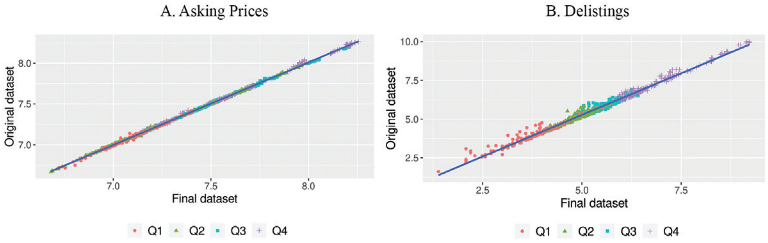
B. Housing Prices

C. Time on Market and Average Discount on Asking Prices

D. Housing Prices in Larger Municipalities

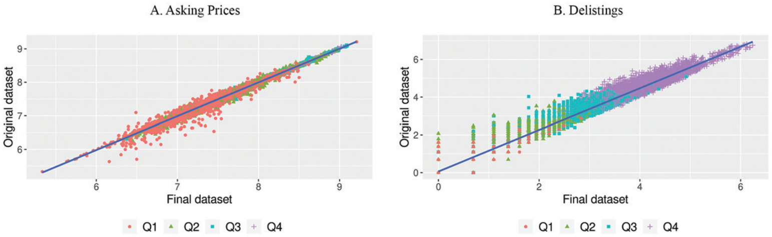
Note: Panel A: home sales, annual data from OMI (a branch of the Italian Tax Office), index 2011 = 100. Panel B: housing prices, annual data from Istat (Institute of Statistics), index 2011 = 100. Panel C: time on market (months) and average discount on the asking price obtained by the buyer (percentage points), quarterly data from the Italian Housing Market Survey. Panel D: housing prices (year-on-year percentage changes), annual data from OMI for 1,174 municipalities with a population of at least 10,000 individuals. This representation shows both a boxplot and raw data (points)—the horizontal position of a point within a year does not carry any meaning and is just needed for graphical representation. In panels A and B we report home sales and prices at country and NUTS-1 level. The other panels report quantities at country level.

Figure B.2. Comparison between the Original and the Final Data Set at City Level



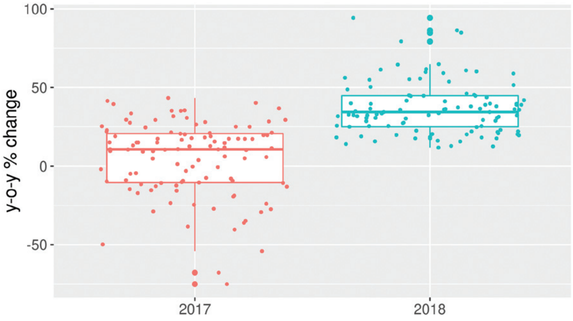
Note: Quarterly data between 2017:Q1 and 2018:Q4. Cities are ranked according to the number of listings in the final data set. The different colors of the dots represent the quartile to which the city belongs (for figures in color, see the online version of the paper at http://www.ijcb.org).

Figure B.3. Comparison between the Original and the Final Data Set at Local Housing Market Level



Note: Quarterly data between 2017:Q1 and 2018:Q4. Local markets are ranked according to the number of listings in the final data set. The different colors of the dots represent the quartile to which the local market belongs (for figures in color, see the online version of the paper at http://www.ijcb.org).

Figure B.4. Online Attention



Note: Average daily number of clicks per ad at city level.

Table B.2. Descriptive Statistics: Asking Prices and Time on Market

	Percentiles						
	5	**25**	**50**	**75**	**95**	**Mean**	**Std. Dev.**
Full Sample							
Price	63,000	122,500	190,000	305,000	690,000	201,531	257,228
Price per s.m.	834	1,389	2,049	3,004	5,000	2,378	1,377
Time on Market	20	68	147	274	680	217	235
Years							
2016							
Price	67,000	128,000	195,762	320,000	728,571	273,148	269,200
Price per s.m.	891	1,471	2,131	3,100	5,122	2,461	1,392
Time on Market	27	90	192	376	872	286	286
2017							
Price	63,549	123,850	190,000	312,077	712,593	266,542	264,460
Price per s.m.	835	1,389	2,045	3,000	5,015	2,379	1,384
Time on Market	25	89	182	336	768	259	262
2018							
Price	60,000	120,000	187,500	305,009	692,000	260,542	260,779
Price per s.m.	797	1,333	1,989	2,951	4,985	2,321	1,378
Time on Market	19	62	145	265	656	212	239
NUTS-1 Regions							
Northwest							
Price	55,000	105,000	169,000	290,540	741,000	254,214	283,708
Price per s.m.	811	1,365	1,988	2,915	5,227	2,354	1,443
Time on Market	20	69	155	293	711	228	242
Northeast							
Price	68,000	115,434	170,000	270,000	550,000	223,849	183,426
Price per s.m.	831	1,271	1,743	2,394	3,800	1,951	959
Time on Market	19	62	140	281	759	226	260
Central							
Price	91,000	166,400	246,000	373,835	800,000	322,894	287,372
Price per s.m.	1,114	1,986	2,773	3,716	5,604	2,989	1,427
Time on Market	20	68	146	272	664	215	232
South and Insular							
Price	53,000	101,732	155,000	240,000	490,000	200,444	175,545
Price per s.m.	708	1,083	1,490	2,067	3,710	1,737	1,006
Time on Market	22	69	140	242	559	193	198

Table B.3. Descriptive Statistics for Local Housing Markets

	Percentiles						
	5	**25**	**50**	**75**	**95**	**Mean**	**Std. Dev.**
Population	56.0	1,666.0	4,652.5	10.801.5	49,503.9	8,182.6	10,440.2
Households	25.0	724.0	1,954.5	4,672.5	23,467.4	3,590.4	4,809.3
Housing Units	39.0	933.5	2,499.5	5,577.0	26,856.2	4,235.9	5,392.8
Share of Owner-Occupied (perc.)	30.4	63.0	70.0	75.6	86.7	68.1	11.1
Average Asking Price	697.4	1,282.6	1,717.0	2,380.2	5,817.2	1,987.9	1,041.5
Delistings	0.2	6.8	28.0	83.8	494.9	66.5	109.2
Delistings/Housing Units (perc.)	0.0	0.8	1.4	1.9	4.8	1.5	2.3

Note: Data on the number of residents (populations), households, housing units, and owner-occupied homes are from the 2011 Census. Average asking prices are computed over the period 2016–18. For delistings, we show the average annual number during the period 2016–18.

Table B.4. Quality of Listed Dwellings and House Prices (half-yearly data)

	Dependent Variable			
	FLOORAREA **(1)**	***BATH*** **(2)**	***GARDEN*** **(3)**	***TERRACE*** **(4)**
PRICE	0.073* (0.041)	0.068*** (0.025)	10.986*** (2.525)	5.460 (3.388)
Temporal Dummies	Year-Semester	Year-Semester	Year-Semester	Year-Semester
Observations	534	534	534	534
R^2	0.168	0.043	0.052	0.083

Note: Results of a panel fixed-effect estimation, using the *within* transformation. *PRICE* is the logarithm of the housing prices as estimated by the Italian Tax Office.

Table B.5. Stationarity of Variables on Housing Supply (half-yearly data)

	$\Delta_1 \boldsymbol{HEDON}_t$ **(1)**	$\Delta_1 \boldsymbol{PRICE}_t$ **(2)**	$\Delta_1 \boldsymbol{FLOORAREA}_t$ **(3)**	$\Delta_1 \boldsymbol{BATH}_t$ **(4)**	$\Delta_1 \boldsymbol{GARDEN}_t$ **(5)**	$\Delta_1 \boldsymbol{TERRACE}_t$ **(6)**
$HEDON_{t-1}$	–0.222*** (0.025)					
$PRICE_{t-1}$		–0.543*** (0.045)				
$FLOORAREA_{t-1}$			–0.614*** (0.049)			
$BATH_{t-1}$				–0.775*** (0.049)		
$GARDEN_{t-1}$					–0.688*** (0.053)	
$TERRACE_{t-1}$						–0.592*** (0.052)
Fixed Effects	City	City	City	City	City	City
Observations	455	443	455	455	455	455
R^2	0.180	0.291	0.303	0.411	0.316	0.266

Note: Results of a panel fixed-effects estimation, using the *within* transformation.

Appendix C. Construction of the Housing Units Data Set

Considering the initial data set of ads, during 2016–18 the number of home sales was about 60 percent of the number of delistings (Table C.1), with significant volatility across different cities.[30] Although this statistic is broadly consistent with studies on the U.S. housing market, alternative evidence from the United Kingdom suggests that this estimate is too low.[31] Since our data set is mostly representative of home sales brokered by real estate agents—the largest share of all transactions—the assumption that each ad is associated with a different dwelling would imply that the share of sales over delistings could be well below 60 percent. Moreover, the average time on market computed on listings—as the number of months between the initial listing and the delisting—is about two months lower than the estimates provided by real estate agents in the Italian Housing Market Survey (Table C.1).

Table C.1. Number of Delistings, House Sales, and Time on Market (months)

			Time on Market	
Year	**Delistings**	**Sales**	**Listings**	**Survey**
2016	335,181	178,690	5.1	7.5
2017	312,584	186,657	4.9	6.3
2018	321,840	197,506	4.4	6.6

Note: Data on sales and time on market come from the Immobiliare.it data set and from the OMI and Italian Housing Market Survey (see Appendix A).

Given these issues, we follow the procedure to clean the original data set described in the next section.

[30] This statistic ranges between 40 percent in Florence and 70 percent in Naples.

[31] According to Anenberg and Laufer (2017) and Carrillo and Williams (2019), about half of the delistings in the United States result in withdraws. In a sample of listings from the United Kingdom, Merlo and Ortalo-Magne (2004) find that withdraws are about 25 percent of the delistings.

C.1 Deduplication at a Glance

We adopt standard methodologies for data deduplication (see Naumann and Herschel 2010; Christen 2012), which we adapt to tackle the specifics of our data set better. The deduplication process consists of three steps.

Data Preparation. To identify if two ads refer to the same dwelling, we have to compare the locations and characteristics of the homes described in the ads. This operation is complicated because the geographical coordinates or the address may not be precise enough. Moreover, some information is not accurate, but based on the best judgment of the home seller/broker.[32]

Thus, we cannot look for perfect matching between home characteristics and have to build partial similarity measures. Moreover, we use the textual description of the home provided in the ad to impute missing data and to extract information useful to identify the duplicates.

Classification. For each pair of ads, we have to decide if the ads refer to the same housing unit. To do so, we compare the characteristics of the dwellings described in the ads and based on some rules, we classify them as *duplicates* or *not duplicates*. To identify these rules, we use a machine learning algorithm, the C5.0 classification tree proposed by Quinlan (1993). The algorithm 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.

Clusterization. The output of the previous step is a list of pairs of duplicate ads. Since multiple pairs can refer to the same dwelling, we have to create clusters of all ads referring to the same home. To do so, we use methods from graph theory and consider a cluster of ads as referring to the same housing unit if an internal similarity condition for the cluster is satisfied. Finally, for each variable, we aggregate information coming from different ads by computing the average or the most common feature observed across ads.

[32] The seller/broker of the home can identify the location on a map or provide the address as an input. The fact that two different tools are available—and the user's lack of precision—gives rise to the possibility that the same dwelling has slightly different geolocation in different ads. That is not an issue in rural areas, but in urban areas with a high concentration of housing units.

Below, we fully describe the algorithm we implemented to remove the duplicate ads. In Loberto, Luciani, and Pangallo (2018) we also show the pseudo-codes of the procedure.

C.1.1 Data Preparation

The textual description of the home provided in the ad performs a dual role. First, by using semantic analysis, information extracted from the textual description allows imputing missing data. That is important because the best way to identify duplicates is to retrieve as much information as possible from the ads. Second, we use the textual description as a further variable to identify if two ads refer to the same dwelling.

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). However, 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 paragraph vector (or *doc2vec*) algorithm proposed by Le and Mikolov (2014), an algorithm based on neural networks that allow representing a document by an N-dimensional vector taking into account both the order and the semantic of the words. In this way, we can measure the "distance" between two descriptions by computing the associated vectors' cosine distance.

We also convert the class of some variables to alleviate the issue of misreporting dwellings' characteristics. Indeed, two different agents can report information partially different but not completely at odds regarding the characteristics of the same housing unit. For example, consider 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 four possible ordered categories, we convert the categorical variable to an integer variable that takes value from one to four (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

Table C.2. Variable Transformations for the Deduplication Algorithm (classification tree)

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)

other ordered categorical variables other than maintenance status: energy class, garage, type of garden, and kitchen type. We report the details in Table C.2.

C.1.2 Classification

We identify duplicate ads based on pairwise comparisons, meaning that we compare each ad with all other ads that are potential duplicates.

First of all, for each ad, we identify its potential duplicates to reduce the computational complexity of the pairwise approach. We define as potential duplicates those ads that refer to dwellings closer than 400 meters to each other and with a difference in asking price lower than 25 percent in absolute value.[33] In this way, we end up

[33] We compute the difference in asking price by dividing the absolute difference between the two asking prices with the lowest of the two. This condition can be quite restrictive when considering dwellings with low asking prices. Then, we

with a long list of pairs of ads, and we have to decide which pairs 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 as an input to the algorithm a vector of predictors (covariates in the jargon of machine learning). Based on this information, the classification tree returns the probability that the two ads are duplicates. We consider a pair of ads as duplicates if the estimated probability is higher than 0.5.

Among the predictors, we consider the following variables: 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 two dwellings' geographical coordinates. For binary variables, such as elevator or basement, the predictor is a dummy variable that takes value equal to one if both ads share the same characteristic. For discrete ordered multinomial variables (such as maintenance status), we consider different degrees of similarity instead, taking the absolute difference between the two variables.

We also use the distance between the textual description of the two ads as a predictor. For this variable, we consider two different measures, depending on whether the same agency posted the ads. 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 same agency posted the ads. This choice is motivated by the observation that when an agency posts two ads for the same dwelling,

consider as potential duplicates also those ads with absolute difference lower than 50,000 euros.

its characteristics 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 by the pictures on the website. Then, duplicate ads are less similar if posted by different agencies than if created by the same agency. Consequently, 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, comparing the pictures. The training sample for the ads of different agencies is made of 8,296 pairs of ads; among them, 3,711 are duplicates (true positive, TP). The training sample for the ads of the same agency includes 9,844 observations, and 1,850 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. We repeat this three-step approach several times until we reach a sufficiently low misclassification error.

To assess the performance of the two models, we randomly split each training sample into two different subsamples. We use the first sample (90 percent of the observations) to estimate the models. The second one (10 percent 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 standard accuracy rate can be misleading about the models' actual performance. For this reason, we consider measures of classification performance that do not rely on the number of true negatives, namely, precision, recall, and F-measure.[34]

[34] The precision rate is the ratio between the number of true positives and the sum of true and false positives. Thus, it measures how accurate a classifier is in classifying true matches. The recall rate is 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

Table C.3. Assessment of C5.0 Models

	Observations	Duplicates	Precision	Recall	F-measure
Different Agency	8,296	3,711	0.930	0.924	0.927
Same Agency	9,844	1,850	0.952	0.946	0.949

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

We show the results in Table C.3. As expected, the model for ads of the same agency is 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 are almost the same. Therefore, it is easier to identify them. However, as the F-measure is equal to .927, 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 to maximize the F-measure.[35] We report the set of variables for each model in Table C.4.

C.1.3 Clusterization

Once we have identified the pairs of ads that are duplicates, we need a procedure to cluster all ads that are considered related to the same housing unit and to aggregate the information in the ads. Here, we follow a standard procedure in the computer science literature (Naumann and Herschel 2010; Christen 2012).

recall, we also consider a third additional measure, the F-measure, that calculates the harmonic mean between precision and recall.

[35] We started for both models with only five predictors: the percentage difference between prices, the absolute difference between prices, the percentage difference between floor areas, the absolute difference between floor areas, and the 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 a new predictor.

Table C.4. Variables for the Classification Trees

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	No	Indicator function: 1 if the feature balcony is the same
distdays1	Yes	Yes	Number of days between the dates the ads have been added
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)
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
rooms	Yes	No	Absolute difference between number of rooms (integer)
garage	Yes	Yes	Absolute difference (integer) between categories
garden	Yes	No	Absolute difference (integer) between categories
utility_room	Yes	No	Indicator function: 1 if the feature utility room is the same
janitor	Yes	No	Indicator function: 1 if the feature janitor is the same.

(*continued*)

Table C.4. (Continued)

Variable	Model 1	Model 2	Description of the Variable
basement	Yes	No	Indicator function: 1 if the feature basement is the same
pricemq_abs	No	Yes	Absolute difference in the asking price per square meter
pricemq_min	Yes	Yes	Minimum of the two asking prices per square meter
pricemq_max	No	Yes	Maximum of the two asking prices per square meter

Table C.5. Example of Clusters

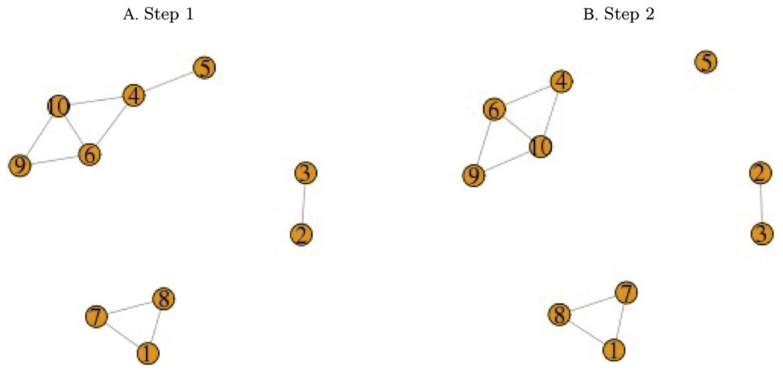
Id.x	1	1	2	4	4	4	6	6	7	9
Id.y	7	8	3	6	10	5	9	10	8	10
Prob.	0.92	0.81	0.73	0.98	1.00	0.52	0.87	0.70	0.93	0.86

Let us suppose 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. 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 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. The assumption of transitivity in this case may not be reliable.

Here, we abstract from the assumption of transitivity. 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 10 ads. We compute for each of the 45 possible pairs the probability that they are duplicates, and we remove all pairs with a probability smaller than 0.5. The remaining pairs are in Table C.5.

Starting from the results of the pairwise classification step in Table C.5, we represent the information as a graph, in order to form

Figure C.1. Clustering of the Ads



clusters. The output of this step is in Figure C.1. 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 in each tuple ads are all pairwise duplicates. 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 for it to be 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 the graph (in our case the number of ads).

The tuples (2,3) and (1,7,8) are 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 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, which we define as the one with the lowest duplicate probability among those in the cluster.

For the tuple (4,5,6,9,10), the condition is not satisfied. In this case, we delete the weakest link, represented by the edge between nodes 4 and 5 because the associated probability is 0.52. The new set of clusters is in Figure 1B, where node 5 now refers to a distinct housing unit. If we look at the new tuple (4,6,9,10), we see five edges out of six possible edges. Since our internal similarity condition is satisfied, we consider this last tuple as a distinct dwelling.

Summing up our example, we started with 10 ads, and we ended up with only four housing units.

Once we have created the clusters of ads identifying different dwellings, we must combine the information contained in multiple ads related to the same dwelling. As a general rule, for each characteristic, we take the one with the 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 with the dwelling; for the exit we consider the date of removal from the database of the last ad).[36]

C.1.4 Implementation

This approach becomes computationally unfeasible once the number of ads rises. Indeed, the number of pairwise comparisons increases exponentially. Thus, the procedure described in the previous section will be applied using an iterative approach ("time machine approach").

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 ads published before the first week. Once we apply the deduplication procedure, we end up with a new data set. Each row corresponds to a unique dwelling.

At the second iteration, we take as an input the data sets 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. We look for duplicates for all these ads among new or updated ads and the data set of housing

[36] We make a further exception to the general rule for asking prices. In this case, we take the most frequent observation among ads that have not been removed.

units from the first week. The ads that are updated are preliminarily removed from the data set of dwellings (that must be updated accordingly).

Whether the ads are duplicates is still based on the pairwise comparison, but now we can have pairs with two ads or pairs with one ad and one housing unit. Once we compute the probability that they are duplicates for each pair, we cluster the results, as explained in Section C.1.3. We impose the additional condition that in each cluster there can be at most one housing unit 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.

C.1.5 Additional Controls

After the deduplication procedure, we make additional controls on the data set to address potential errors. First of all, we keep only the dwellings that have been on the market for at least two weeks. Then, we drop from the data set those dwellings for which the price is not sufficiently consistent with the characteristics of the housing units. In this way, we can also identify foreclosure listings that were not previously identified because the ads did not report the foreclosure status.

Our approach consists of running a hedonic regression, estimating the ratio between actual and predicted price for each dwelling and eliminating the housing units with a ratio between asking and predicted price lower than 0.5 or higher than 1.5.[37]

C.2 Final Data Set and Representativeness

The number of homes—or "true" listings—is only 67 percent of the number of ads (about 940,000 housing units). Looking at the

[37] We keep a relatively broad 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).

distribution of homes per number of associated ads, we find that duplicates are concentrated over a small share of homes: about 77 percent of dwellings have one associated ad, 13 percent have two duplicate ads, and 10 percent have more than two duplicates.

Open listing agreements with many agents seem to be the main source of duplicate ads. To see that, consider that only 15 percent of homes were listed with more than one agency, but these homes account for 35 percent of ads.

Considering a single daily snapshot, the number of listed homes is 87 percent of the number of ads on average. Thus, by taking a snapshot of the data on any specific day, we expect that only 13 percent of the ads are duplicates. These figures are consistent with those concerning the full sample because duplicate ads for the same listed home grow over time: new ads are created while old ads are deleted, and that gives rise to a huge number of delistings and new listings. We find confirming evidence when we consider only homes with multiple corresponding ads. Every week, for 90 percent of them, at most two duplicate ads are on average visible. This figure can be compared with the share of homes with two ads among those with multiple ads in the full sample, which is $10/(10 + 13) = 57\%$.

Finally, the share of duplicates over total ads increases with city size, and there is significant variability across cities. For example, the ratio between the number of ads and housing units is equal to 1.4 for Naples and 1.8 for Rome. Therefore, an additional implication of duplicates is that they can make the comparability across cities difficult.

To validate the quality of our deduplication procedure, we compare information coming from the final data set with other well-established statistical sources.

First, we compute the number of delistings and home sales in each city (obtained from OMI) at a quarterly frequency, and we find that these two variables are strongly correlated (Figure 1): their correlation coefficient is 0.94. Now, a delisting is an effective exit of a home from the market. Table 1 compares the absolute number of delistings and home sales. Compared with Table C.1, the numbers seem more plausible once we take into account that not all homes sold during these years have been listed on Immobiliare.it.

We find a strong correlation with official data when we consider prices (the correlation is 0.82; Figure 1). Our results are even stronger, because we have official estimates from OMI for each local housing market, so we can compare listing prices and average home values per square meter at a finer granularity. The non-linearity observed for very high home values is probably because OMI estimates refer to the average value of all homes in the local market. In contrast, the most expensive and prestigious homes are likely to be less liquid and, therefore, less represented among listed homes.

Moreover, we compute the ratio between listing prices and actual home values per square meter for each local housing market. On average, we find that the discount on asking prices was about 12 percent during 2016–18, a value consistent with the evidence provided by the Italian Housing Market Survey.

Finally, we look at time on the market. After our deduplication procedure, listings provide an estimate of the time on market overall consistent with the Italian Housing Market Survey (see Table 1). We find a significant deviation only for 2016 when listings underestimate time on market. That is plausible because that is the first year for which we have weekly data. Some of the homes listed in 2016 may have been initially listed in 2015. However, for 2015 we only observe quarterly snapshots, and we may not be able to reconstruct the full history of these listings due to difficulties in identifying duplicates.

Overall, information coming from our final data set of listings is consistent with official statistical sources. We consider this as evidence of the efficacy of our deduplication procedure.

Appendix D. Duplicate Ads and Systematic Bias

The presence of multiple ads related to the same dwelling is not random. In particular, we focus on two hypotheses. First, duplicate ads are more likely among those homes for which potential buyers show little interest, i.e., demand for these homes is relatively small. Intuitively, home sellers would choose to increase search intensity—through open listing agreements with multiple agencies or more generally by posting numerous ads—to compensate for the scarcity of buyers potentially interested in their homes. Second, the presence of duplicates is correlated with the listing price. It is reasonable that homes whose listing price is too high compared with similar nearby homes may have multiple associated ads because the seller increases their odds of finding a buyer.

To test for these hypotheses, we estimate the following linear probability model:[38]

$$DUPL_{ijt} = \alpha_{jt} + \beta CLICKS_{ijt} + \gamma PRICE_{ijt} + \delta \mathbf{X}_i + \varepsilon_{ijt}, \quad (6)$$

where $DUPL_{ijt}$ is an indicator variable equal to one if more than one ad is associated with home i in week t; the index j refers to the home's local housing market. $CLICKS_{ijt}$ is average daily number of visits to the webpages (*clicks*) related to dwelling i during week t.[39] Intuitively, the most-searched homes are likely to be those for which the owner or the broker receives more calls or emails from potential buyers. $PRICE_{ijt}$ is the listing price per square meter of dwelling i during the week t. We control for spatial and temporal heterogeneity at the local housing market level through the set of dummies α_{jt}. Finally, $\mathbf{X}$ is a vector of dwellings' physical characteristics: floor area (square meters), type of property (apartment, detached dwelling), floor level, number of bathrooms, maintenance status, presence of a balcony or a terrace, garage, and elevator.[40]

Since duplicates are identified through machine learning tools, any inefficiency in this first step could invalidate our analysis. However, we believe that this is not an issue for the following two reasons. First, we estimate the classification trees over a large sample of couples of ads for which we know for sure whether they are duplicates. Standard measures of performance for classification tasks used in the machine learning literature suggest that our approach is very effective (see Section C.1.2 and Table C.3). Second, duplicates' identification relies on the similarity between physical characteristics or listing prices and geographical proximity. Since the visits to the webpages and the relative (to the neighborhood) listing prices do not affect the identification of duplicates, any results of our analysis are not a consequence of the deduplication procedure.

[38] We use a linear probability model instead of a logit model because of computational convenience.

[39] When multiple ads refer to dwelling i, $CLICKS$ is computed in two steps. First, we compute the average daily number of clicks for each ad. Second, we compute the mean of the daily number of clicks across all ads.

[40] Given the inclusion of time-varying fixed effects and physical characteristics, there is no need to control for the housing price level in the local market to identify overpriced listings. In our context, we only need to estimate if a listing has an asking price higher than those of properties with similar characteristics.

Table D.1. Determinants of Duplicates

	Multiple Ads (1)	**New Duplicate (2)**	**New Duplicate (3)**	**New Duplicate (4)**
Listing Price t	0.0198*** (0.0011)			
Clicks t	–0.1221*** (0.0010)			0.3003*** (0.0006)
Clicks t–1		–0.0015*** (0.0003)	–0.0027*** (0.0003)	–0.1744*** (0.0007)
Clicks t–2				–0.0463*** (0.0007)
Clicks t–3				–0.0262*** (0.0007)
Clicks t–4				–0.0227*** (0.0006)
Listing Price t–1		0.0024*** (0.0004)		0.0112*** (0.0005)
Listing Price t–4			0.0013*** (0.0002)	
Observations	16,042,720	15,450,398	14,374,903	13,452,978
Adjusted R^2	0.0036	0.0004	0.0004	0.0178

Note: Coefficients and standard errors reported in the table have been multiplied by 100.

Column 1 in Table D.1 reports the results. The coefficients associated with $CLICKS$ and $PRICE$ are statistically significant, and their sign confirms our initial hypotheses. The estimated coefficient for $CLICKS$ is negative (–0.12), and the coefficient for $PRICE$ is positive (0.02). The presence of multiple ads is associated with lower interest by potential buyers and a relatively higher listing price. Although we cannot claim any causal relation based on model (6), the evidence is consistent with the hypothesis that the home seller increases his effort to find a buyer to compensate for a high asking price or unattractive characteristics of the home.

To identify a causal effect of demand and listing prices on the propensity to post multiple ads, we create a new indicator variable called *NEWDUPL*. This variable is equal to one if the number of ads associated with a home already on the market increases during week t. Then, we estimate the following linear probability model:

$$\begin{aligned} NEWDUPL_{ijt} = \alpha_{jt} + \beta CLICKS_{ijt-1} + \gamma PRICE_{ijt-1} \\ + \delta \mathbf{X}_i + \zeta z_{it} + \varepsilon_{ijt}. \end{aligned} \tag{7}$$

Compared with (6), we take as regressors the one-week lag for both demand and listing price. This model allows us to test if the home seller's or the broker's propensity to increase advertising during the week t by posting a new ad is affected by asking price and buyers' demand during the previous week.[41] We also control for the number of days dwelling i has been listed up to week t (z_{it}).

Column 2 in Table D.1 shows that our previous results are qualitatively confirmed. The propensity to post a new ad for a previously listed home decreases when online interest for that home goes up; this propensity is also increasing in the listing price. Notice that these coefficients are statistically significant, although we include many controls, and the phenomenon we are considering is not very frequent at a weekly frequency. In particular, the unconditional probability that during week t a new ad is posted for a previously listed home is 0.9 percent.

In this regression, clicks can be considered as exogenous because potential buyers cannot know the sellers' strategies a week before. Moreover, since we control for the listing price and dwellings characteristics, we deduce that the lower online attention is determined not only by an excessively high price asked by the seller but also by a genuine mismatch with potential buyers' preferences. Unfortunately, we cannot resort to this argument to claim that the lagged value of the listing price is exogenous.[42] However, in column 3, we show that replacing the one-week lagged listing price with the four-weeks lag, we still find a positive and significant effect on the propensity to post a new ad.

Finally, after showing that the listings that receive little online attention are those with the highest probability of having multiple

[41] Controlling for higher-order lags (up to $t-4$) would not affect our results.

[42] Indeed, home sellers/brokers set both the listing price and the advertising strategy, and when changing the listing price, they may have already decided to post a new ad.

ads, we want to evaluate the effectiveness of this advertising strategy. To do that, we estimate the following extension of model (7):

$$\begin{aligned} NEWDUPL_{ijt} = \alpha_{jt} + \sum_{i=0}^{4} \beta_i CLICKS_{ijt-i} + \gamma PRICE_{ijt-1} \\ + \delta \mathbf{X}_i + \zeta z_{it} + \varepsilon_{ijt}, \end{aligned} \tag{8}$$

where we add as regressors the contemporaneous value of the variable $CLICKS$ and all its lags up to four weeks. The results are reported in column 4 in Table D.1. We find that during the four weeks before the seller posts a new ad, his home gets a relatively poor online interest ($\beta_i < 0$ for $i = -1, -2, -3, -4$). Clicks are low especially in the previous week (β_{-1}). Then, following the publication of the new ad, a spike in clicks occurs ($\beta_0 > 0$). These results, which must be interpreted as correlations, are consistent with the hypothesis that potential buyers may believe that this is a new listing. Homebuyers may not easily recognize that the new ad refers to a previously listed home, and this is especially true when a new broker posts the ad.

Finally, homes with multiple ads show further systematic differences compared with other dwellings. We estimate the OLS regression of time on market over a dummy taking value one if a home had multiple ads, and we find that those with many ads stay longer on the market (see Table D.2). We also estimated a linear probability model where the dependent variable is an indicator variable taking value one if the home seller revised downward the initial asking price, and zero otherwise. As expected, it is more plausible to observe a price change for homes with multiple ads (Table D.2). These results are consistent with previous evidence: by using the ads, we underestimate the time on the market, and dwellings with multiple ads are overpriced (therefore more subject to price reductions).

The main conclusion is that using the original data set of ads implies an oversampling of relatively expensive homes—given their location and characteristics—and less attractive homes. Moreover, lower attractiveness is associated with higher time on market and propensity to revise downward the asking price. Therefore, using the original data would imply severe distortions when analyzing the microstructure of the housing market.

Table D.2. Duplicates, Time on Market, and Price Changes

	Time on Market **(1)**	**Price Change** **(2)**
Multiple Ads	125.30580*** (0.74404)	0.17805*** (0.00150)
Fixed Effects	OMI Microzone	OMI Microzone
Temporal Effects	Quarter	Quarter
Observations	512,246	512,246
Adjusted R^2	0.06827	0.09316

References

Anenberg, E., and P. Bayer. 2020. "Endogenous Sources of Volatility in Housing Markets: The Joint Buyer–Seller Problem." *International Economic Review* 61 (3): 1195–1228.

Anenberg, E., and E. Kung. 2014. "Estimates of the Size and Source of Price Declines Due to Nearby Foreclosures." *American Economic Review* 104 (8): 2527–51.

Anenberg, E., and S. Laufer. 2017. "A More Timely House Price Index." *Review of Economics and Statistics* 99 (4): 722–34.

Carrillo, P. E., E. R. de Wit, and W. Larson. 2015. "Can Tightness in the Housing Market Help Predict Subsequent Home Price Appreciation? Evidence from the United States and the Netherlands." *Real Estate Economics* 43 (3): 609–51.

Carrillo, P. E., and B. Williams. 2019. "The Repeat Time-on-the-Market Index." *Journal of Urban Economics* 112 (July): 33–49.

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

de Wit, E. R, and B. van der Klaauw. 2013. "Asymmetric Information and List-Price Reductions in the Housing Market." *Regional Science and Urban Economics* 43 (3): 507–20.

Glaeser, E., and J. Gyourko. 2018. "The Economic Implications of Housing Supply." *Journal of Economic Perspectives* 32 (1): 3–30.

Guglielminetti, E., M. Loberto, G. Zevi, and R. Zizza. 2021. "Living on My Own: The Impact of Covid-19 Pandemic on Housing Preferences." Questioni di Economia e Finanza (Occasional Papers) No. 627, Bank of Italy.

Han, L., and W. C. Strange. 2015. "The Microstructure of Housing Markets: Search, Bargaining, and Brokerage." In *Handbook of Regional and Urban Economics,* Vol. 5, ed. G. Duranton, J. V. Henderson, and W. C. Strange, 813–86 (chapter 13). North-Holland.

Harris, Z. S. 1954. "Distributional Structure." *Word* 10 (2–3): 146–62.

Head, A., H. Lloyd-Ellis, and H. Sun. 2014. "Search, Liquidity, and the Dynamics of House Prices and Construction." *American Economic Review* 104 (4): 1172–1210.

Jordà, O., M. Schularick, and A. M. Taylor. 2015. "Leveraged Bubbles." *Journal of Monetary Economics* 76 (Supplement): S1–S20.

Kolbe, J., R. Schulz, M. Wersing, and A. Werwatz. 2021. "Real Estate Listings and Their Usefulness for Hedonic Regressions." *Empirical Economics* 61 (6): 3239–69.

Krainer, J. 2001. "A Theory of Liquidity in Residential Real Estate Markets." *Journal of Urban Economics* 49 (1): 32–53.

Landvoigt, T., M. Piazzesi, and M. Schneider. 2015. "The Housing Market(s) of San Diego." *American Economic Review* 105 (4): 1371–1407.

Le, Q., and T. Mikolov. 2014. "Distributed Representations of Sentences and Documents." In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, Vol. 32, ed. E. P. Xing and T. Jebara, 1188–96.

Liberati, D., and M. Loberto. 2019. "Taxation and Housing Markets with Search Frictions." *Journal of Housing Economics* 46 (December): Article 101632.

Loberto, M., A. Luciani, and M. Pangallo. 2018. "The Potential of Big Housing Data: An Application to the Italian Real-Estate Market." Temi di discussione (Economic Working Papers), No. 1171, Bank of Italy.

Lyons, R. C. 2019. "Can List Prices Accurately Capture Housing Price Trends? Insights from Extreme Markets Conditions." *Finance Research Letters* 30 (September): 228–32.

Merlo, A., and F. Ortalo-Magne. 2004. “Bargaining over Residential Real Estate: Evidence from England.” *Journal of Urban Economic*s 56 (2): 192–216.

Mian, A., K. Rao, and A. Sufi. 2013. “Household Balance Sheets, Consumption, and the Economic Slump.” *Quarterly Journal of Economics* 128 (4): 1687–1726.

Mian, A., A. Sufi, and F. Trebbi. 2015. “Foreclosures, House Prices, and the Real Economy.” *Journal of Finance* 70 (6): 2587–2634.

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

Ngai, L. R., and K. D. Sheedy. 2020. “The Decision to Move House and Aggregate Housing-Market Dynamics.” *Journal of the European Economic Association* 18 (5): 2487–2531.

Pangallo, M., and M. Loberto. 2018. “Home Is Where the Ad Is: Online Interest Proxies Housing Demand.” *EPJ Data Science* 7 (1): Article 47.

Piazzesi, M., M. Schneider, and J. Stroebel. 2020. “Segmented Housing Search.” *American Economic Review* 110 (3): 720–59.

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

Rae, A. 2015. “Online Housing Search and the Geography of Submarkets.” *Housing Studies* 30 (3): 453–72.

van Dijk, D. W., and M. K. Francke. 2018. “Internet Search Behavior, Liquidity and Prices in the Housing Market.” *Real Estate Economics* 46 (2): 368–403.

Wu, L., and E. Brynjolfsson. 2015. “The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales.” In *Economic Analysis of the Digital Economy*, ed. A. Goldfarb, S. E. Greenstein, and C. E. Tucker, 89–118 (chapter 3). Chicago: University of Chicago Press.