Web-scraping housing prices in real-time: The Covid-19 crisis in the UK

Jean-Charles Bricongne, Baptiste Meunier, Sylvain Pouget

Web-scraping housing prices in real-time: the Covid-19 crisis in the UK *

Jean-Charles Bricongne[†], Baptiste Meunier[‡], Sylvain Pouget[§]

June 1, 2022

Abstract:

While official statistics provide lagged and aggregate information on the housing market, extensive information is available publicly on real-estate websites. By web-scraping them for the UK on a daily basis, this paper extracts a large database from which we build timely and highly granular indicators. One originality of the dataset is to focus on the supply side of the housing market, allowing to compute innovative indicators reflecting the sellers' perspective such as the number of new listings posted or how prices fluctuate over time for existing listings. Matching listing prices in our dataset with transacted prices from the notarial database, using machine learning, also measures the negotiation margin of buyers. During the Covid-19 crisis, these indicators demonstrate the freezing of the market and the "wait-and-see" behaviour of sellers. They also show that listing prices after the lockdown experienced a continued decline in London but increased in other regions.

Keywords: housing, real time, big data, web-scraping, high frequency, United Kingdom
JEL classification: E01, R30

*We are very grateful to an anonymous referee, Kieran McQuinn (discussant), Jean Clark, Nigel Henretty, Karim Abdalla, Francoise Drumetz, Olivier de Bandt, and Nicolas Chatelais for useful comments. We thank Jerome Coffinet and Etienne Kintzler for excellent assistance regarding the automatizing of the data acquisition process. The views expressed here are those of the authors and do not necessarily represent those of the European Central Bank, Banque de France, LEO, LIEPP, AMSE or ENSIMAG. This work was supported by the French National Research Agency Grant ANR-17-EURE-0020, and by the Excellence Initiative of Aix-Marseille University - A*MIDEX.

[†]Banque de France, LEO, LIEPP. jean-charles.bricongne@banque-france.fr

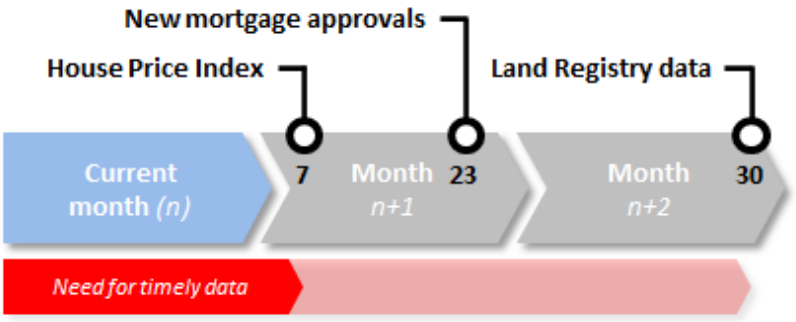
[‡]European Central Bank and Aix Marseille Univ, CNRS, AMSE, Marseille, France.
baptiste.meunier@ecb.europa.eu

[§]ENSIMAG. pougetsylvain12@gmail.com

1 Introduction

Official statistics on the residential housing market are relatively scarce and available with a significant delay. In the UK – the reference market in this article – official transacted prices are published in the Land Registry two months after the end of the corresponding month, or even later for transactions that may be registered with some delay, as shown in **Figure 1**. The earliest indicators (House Price index and new mortgage approvals) are available in the following month but are proxy statistics based on mortgage data. Hence, timelier information might complement these official statistics. In addition, most of the existing early indicators are provided at aggregate level while discrepancies between urban – most notably London – and rural areas have been documented (e.g. Poon and Garratt, 2012). In the meantime however, a lot of information is available publicly and in real-time on real-estate websites, particularly for the residential segment where 92% of real-estate firms post ads on the Internet.[1] Getting timely information might be even more critical during *crisis* episodes such as the Covid-19 pandemics as the publication delays of official statistics would not allow to grasp swiftly dramatic and sudden turning points in the cycle.

Figure 1: Timeline of official statistics
Source: Thomson Reuters, e.g. for March 2020



Against this background, the purpose of this paper is to exploit such publicly available information online to build real-time and granular indicators for the housing market. Our approach builds on web-scraping the five main real-estate websites in the UK: Zoopla, Rightmove, and OnTheMarket – plus PropertyPal and S1Homes which focus on Northern Ireland and Scotland respectively. On average, we scrape around 1.5 million listings per day with extensive information on price, type of ads (for sale or to rent), location, area, number of rooms, type of good, and general description. Taking

[1]Source: Realtor, available at https://www.nar.realtor/sites/default/files/documents/2018-real-estate-in-a-digital-world-12-12-2018.pdf

advantage of this innovative dataset, we build real-time indicators for the supply side of the housing market that can complement the official statistics.

The originality of the web-scraped data lies in getting the supply side of the housing market through the listings that sellers (or the real-estate agencies they mandate) post on the Internet. This peculiar point of view – in the literature – brings a twofold value-added. It first allows to monitoring the housing market in real-time, notably price trends in listings that can be tracked daily and at a highly granular level. While not a substitute for transacted data – which reflect the outcome of the housing market equilibrium – such information centred on the supply side of the market can provide complementary, as well as timely and granular, information. In addition, this original dataset also allows for innovative indicators as regards sellers' perspective. A first example is the number of new listings – indicating sellers' willingness to put their properties on the market. A second one relates to price changes for an existing listing as our daily web-scraping makes it possible to track one listing over time and observe how its price adjusts. In a second endeavour, our dataset makes it possible to match listings with transactions. Discrepancies appear – showing frictions between buyers and sellers in reaching market equilibrium, reflected by transactions – and the difference between the two provides an indication on the negotiation margin of buyers.

Using these indicators in a conjectural fashion, we track the UK housing market during the Covid-19 crisis. Our supply-side indicators shows a clear 80% decline in the number of new listings during the first lockdown and also shows that during the lockdown, sellers refrained from changing their prices, suggesting that most adopted a "wait-and-see" approach. In the aftermath of the lockdown, average listing prices started to increase at country-level. However, this hides large discrepancies across regions: while listing prices increased steadily in rural areas, they declined in London – the region most affected by the virus and where evidence suggests that the housing market is the most tense. Such cross-region heterogeneities are further documented in our computation of the buyers' negotiation margin.

This paper contributes to the literature, first by providing new evidence for the on-going move to higher frequency statistics relying to some extent on alternative data (see Veronese et al., 2020). Rather than competing with official statistics, our web-scraped data however complement the latter by detecting trends with enhanced timeliness (daily indices in real-time *vs.* monthly / quarterly with lag) and high granularity (ZIP code level *vs.* at best region-level). In a broader perspective, this paper mirrors the recent endeavour across economists to design high-frequency indicators (e.g. Lewis et al., 2021) – which the Covid-19 has made more pressing. More specifically, we contribute to the literature using alternative data to monitor the real-estate market, such as Kulkarni et al., 2009 with Google Trends. Closer to us, we expand a recent strand of the literature

making use of web-scraping to analyze the housing market (e.g. Hanson and Santas, 2014). In this strand of the literature, this paper is close to Galesi et al., 2020 by relying on web-scraping to derive indicators of the supply side of the housing market. Finally, this paper stands out by focusing on the UK market in this Covid-19 / Brexit period – though the approach can be seamlessly expanded in other geographies.

The remainder of the paper is organised as follows. **Section 2** reviews the related literature, **section 3** describes how we retrieve and clean the data, **section 4** presents the real-time monitoring of the housing market during the Covid-19 crisis. **Section 5** explores alternative indicators, namely a measure of buyers' negotiation power as well as the rent-to-price ratio. Finally, **section 6** concludes.

2 Literature review

This paper first contributes to the literature on monitoring the housing market, a key topic for economists not only because of its importance in the transmission between credit and business cycles (Kiyotaki, 1998) but also as it may act as a propagation mechanism for shocks (Kiyotaki and Moore, 1997) or even be the source of larger crises (Cheng et al., 2014). House prices can become disconnected from fundamentals up to the point that agents sharply revise down their assessment (Case and Schiller, 1990; Case and Schiller, 2003), potentially leading to a deterioration of banks' balance sheets while inducing negative wealth effect for households (e.g. Slacalek, 2009) with both factors ultimately weighing on aggregate demand. In that vein, papers such as Rünstler and Vlekke, 2018 establish a strong correlation between the housing cycle and GDP components. Therefore, designing indicators tracking price misalignments (e.g. rent-to-price ratio, vacancy rate, pending sales, price-to-income ratio) has been a long-lasting endeavour in the literature (e.g. Miller and Sklarz, 1986, Flood, 1997, Quigley, 2001, Case and Wachter, 2005, Lind, 2009, Dujardin et al., 2015, Engsted et al., 2016, and Blot et al., 2018) with an emphasis in the interest for policy-makers (e.g. Cesa-Bianchi, 2013). Closer to us, a strand of the literature has also focused on measuring the bargaining power of buyers, i.e. the difference between asking and selling prices (e.g. Ngai and Tenreyro, 2014, Han and Strange, 2016). This has been in particular possible for the U.S. by relying on the U.S. Multiple Listing Services which provides, similarly to our web-scraping, information on the different listings in a region (e.g. Carrillo, 2013, Ozhegov and Sidorovykh, 2017). Interestingly, several studies have linked this indicator of frictions between supply and demand in the housing market with the housing cycle, demonstrating the informative power of such a metric (e.g. Carrillo et al., 2015, Bo, 2018).

We contribute to this literature by using new techniques to retrieve information at the fastest pace – in real-time – and in a very granular fashion – at ZIP code level.[2] In addition, our dataset provides us information on the sellers' perspective, making it possible to design innovative indicators reflecting the supply side of the housing market (number of new listings, price changes for existing listings). Getting such information also allows to compare listing prices with transacted prices and measure the negotiation margin of buyers, in line with the literature mentioned above. On a more conjectural standpoint, we contribute by monitoring and documenting the impact of the Covid-19 on the supply side of the UK housing market, mirroring similar efforts such as Giudice et al., 2020 for Italy.

This paper also contributes to the growing field of the literature focusing on tracking the economy in real-time. In the wake of the Covid-19 crisis, a number of innovative high-frequency datasets have emerged such as weekly labour statistics (Coibion et al., 2020), daily satellite pollution data (Bricongne et al., 2021), hourly electricity consumption (Chen et al., 2020), or marine traffic by the minute (Cerdeiro et al., 2020). Particularly illustrative of this search for original data, Chetty et al., 2020 have developed multiple partnerships with private entities to provide a vast amount of data -– untapped until now in the economic literature — on US employment, household spending and mobility. In the same vein, Bricongne et al., 2020 have proposed a number of indicators for the French economy available from public sources. These high-frequency data allow for a swift detection of turning point in economic activity and their signalling power has been used to develop activity trackers as in Lewis et al., 2021 or to nowcast macroeconomic variables as in Jardet and Meunier, 2022. We contribute by covering the housing market while most of this literature has rather focused on GDP, industrial production or households' consumption.

This paper more closely relates to the literature using alternative data, including through web-scraping, to analyse the housing market. Web-scraping is increasingly used in housing, notably to obtain the levels of house prices (Bricongne et al., 2019) or rents (Chapelle and Eymeoud, 2018). This innovative technique has also been used to address specific issues such as potential discrimination in rentals across the US (Hanson and Santas, 2014) or the impact of rent control in Germany (Mense et al., 2017). Web-scraping is also extensively used to take into account new digital players in housing not covered by official statistics, for example Airbnb (Horn and Merante, 2017 or Garcia-López et al., 2020). More broadly, a number of papers have relied on user-generated data available online to analyse the housing market such as Askitas, 2015 and McLaren and Shanbhogue, 2011. In particular, Wu and Brynjolfsson, 2015, Veldhuizen et al., 2016, Oust and Martin, 2018, and Pavlicek and Kristoufek, 2019 have demonstrated

[2]Keeping however in mind the need to conciliate granularity with the representativeness constraint, as further developed in **section 3**.

the capacity of Google searches to monitor the dynamics of the housing market. Closer to us would be both Boeing and Waddell, 2017, who use web-scraping to track the housing market but focus only on rentals in the US, and Loberto et al., 2018, who use also web-scraped ads for monitoring the Italian real-estate market. We contribute to this literature through a comprehensive effort to monitor in real-time and for all types of goods the housing market in the UK. In the effort to compute buyers' negotiation margin, this paper is close to Galesi et al., 2020 who however do not rely on web-scraping and do not compute real-time indicators.

Policy-wise and particularly relevant to our focus, we finally contribute to the on-going movement calling economists and statisticians to rely more on alternative data. While the Covid-19 crisis has amplified this burgeoning movement (Veronese et al., 2020), a number of initiatives had been taken beforehand. Interestingly, some relied on web-scraping such as Polidoro et al., 2015 who get consumer prices for the Italian National Institute of Statistics (ISTAT) or Dumbacher and Capps, 2016 who retrieve government tax revenue for the US Census Bureau. Owing to heterogeneities in the source data (posted listings reflecting the supply side *vs.* transactions reflecting the supply-demand market equilibrium) as well as in the statistical adjustments applied to raw data, the indicators developed in this paper however aim at complementing official statistics rather than supplanting them.

3 Data

3.1 Web-scraping

Data are retrieved through a daily web-scraping of the three most important real-estate websites in the UK: Zoopla, Rightmove, and OnTheMarket. To enhance our coverage on specific areas, we also scrape PropertyPal – a real-estate website specialized on Northern Ireland – and S1Homes – its *alter ego* for Scotland. Through these websites, we scrape on average more than 1,500,000 real-estate listings every day. Out of these, around two thirds are listings for sale. The average number of listings scraped per day can be found in **Table 1**. Data for Zoopla are scraped since the beginning of March 2020 while data for other websites are available since July 2020.[3] In the residential segment, newly built dwellings represent on average around 11% of the total listings – a broadly stable share over time as shown in **Figure 11** in **appendix B** and in line with the share of new buildings in official UK House Price Index, around 10%.

[3]For this reason, the analysis undertaken in **section 4** regarding the effect of the Covid-19 on the UK housing market relies only on Zoopla data.

Table 1: Average number of scraped listings per day (unique URL) in August 2020

Website name	Residential For sale	Residential To rent	Commercial For sale	Commercial To rent
Zoopla	324,958	178,461	14,469	36,919
Rightmove	447,183	162,028	19,534	26,549
OnTheMarket	244,519	94,899	4,872	8,836

For each listing, we scrape information describing in detail the dwelling, the type of transaction, the price, and its location at very granular level. For our main provider Zoopla, we get the following information for residential real-estate listings: city, ZIP code, address, surface area, price, type of transaction (auction *vs.* sale), type of building (new housing projects *vs.* existing dwellings), number of bedrooms, number of bathrooms, number of living rooms, general description (e.g. "lovely flat with a terrace in a safe neighbourhood"), and type of good (e.g. flat, duplex, property). For commercial real-estate listings, we also retrieve the type of use such as hotel, offices, pub, restaurant, retail, warehouse, parking. Having extensive information on a listing might allow to better isolate the market effect from other elements that can affect the pricing of a listing (location, type of good, area, etc.). In particular, retrieving the general description might allow to check whether it contains keywords referring to additional facilities (e.g. balcony, terrace, garage) that can add a premium on the price.

3.2 Data cleaning

Once data are scraped, we ensure that the data are consistent across listings. We first check that the units are the same. We test if the price is displayed in British pounds and not in other currencies and we give special attention to the area which can be expressed in squared feet, squared meters or even acres. Where appropriate, we transform it to have all observations in British pounds and squared meters. In addition, instead of a single value for the area, listings can have a range: in this case, we take the mean value.[4] We also address the fact that rents can be expressed as weekly amounts: in this case, we express it per calendar month. For string objects, we enhance the comparability of keywords across listings by lowering characters and eliminating extraneous space

[4]However, area data are missing for a large number of observations. While it might be possible to interpolate such data, the issue is that these data are not missing at random: area is often quoted for premium dwellings geared towards international buyers such as large apartments in London. Owing to this issue and the fact that area is provided for a very limited sample (around 7%), no interpolation is conducted. For the other categories, the share of missing data is very low – for example around 3% for the city or 0.6% for the price. The average share of missing observations per category can be found in **Table 3** in **appendix B**.

character, line break and tabulation.

We also filter our data in order to gain a more consistent picture of the UK real-estate market. We first exclude non-residential listings like garages, parkings, mobile homes, lands, and bungalows.[5] Focusing on listings for sale posted on Zoopla, these account for around 30,000 listings per day – roughly 10% of the total listings for sale in the residential segment.[6] Second, we handle duplicates that may arise even on the same website. For dwellings that share the same address,[7] title of the listing, number of rooms, elements of description, and price, we keep only one of them. However we do it only for existing dwellings: in the case of new dwellings, a seemingly similar listing can match different properties – as multiple dwellings with similar characteristics can be constructed simultaneously in the same building. Also for listings for sale on Zoopla, around 7,500 daily listings are duplicates – around 2.5% of the total listings for sale in the residential segment.[8] Finally, it appears that a number of real-estate deals in the UK are transacted by auctions: in this case, the price displayed is in fact the reserve price. Therefore, we drop this type of listings that amounts to around 3,000 daily listings (1% of total). In total, these steps eliminate around 15% of the listings. **Figure 12** in **appendix B** shows that the number of listings removed in those filtering steps remains broadly stable over time – an indication of the regularity of our dataset. In addition, **Figure 2** presents evidence that price trends remain similar after those filtering steps.[9] We finally smooth the remaining outliers by performing a winsorization at the 1% level.[10]

[5]Further specific analysis on land prices can be found in **appendix D**

[6]Zoopla is the main provider in the analysis below, since the scraping has started in March 2020 for this website *vs.* July 2020 for other websites

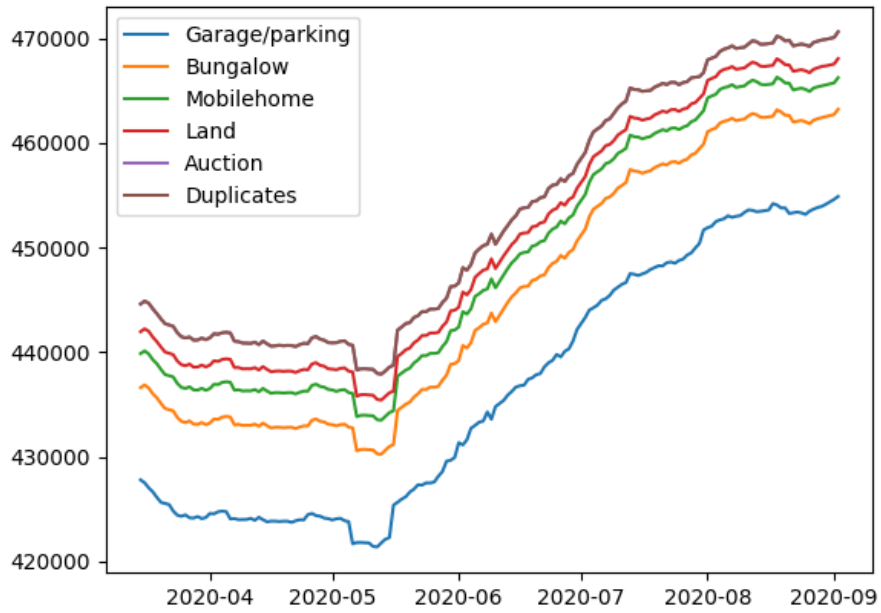
[7]The address can be given only at street level: in this case the comparison is done by relying on longitude and latitude coordinates. For more information on this procedure, please refer to the description of the KNN algorithm in **Section 5**

[8]In addition, it should be noted that a large share of listings are repeated from one day to the other: as regards our main provider Zoopla, on average 98.8% of listings in a given day were already online the day before.

[9]The figure presents the mean listing prices after removing one type of listings deemed irrelevant for our study. Note that the different removals are additive. For instance, the blue curve stands for the mean price after removing garage and parking, then the orange one after further removing the bungalows (so after removing garage, parking, and bungalows), and so on. The two last curves ("auction" and "duplicates") are superposed on the chart since the removal of duplicates does not significantly affect the mean listings price. All prices are expressed in GBP, as in the rest of the paper.

[10]Winsorizing a sample (symmetrically) at the n% level means that we replace each of the n% lowest and n% highest observations by the values of their nearest neighbors – i.e. respectively the values for the n% percentile and for the 100-n% percentiles. The purpose of "winsorization" – named in honour of the work of C. Winsor and proposed by Tukey and McLaughlin, 1963 and Dixon and Tukey, 1968 – is to make the statistical analysis more robust, by eliminating the extreme values to which statistics such as the mean are very sensitive. Alternative methods may be envisaged but winsorization is generally found in the literature as an adequate way to automatically process thousands of observations when it is impossible for a human to inspect each and every one (Jose and Winkler, 2008, Ruppert, 2014).

Figure 2: Mean price across listings after filtering steps
Source: Zoopla and authors' calculation



Finally, using text mining techniques on the description of the listing, we create dummy variables associated with the presence of keywords referring to additional facilities (e.g. "parking", "balcony", "garden") or to qualitative aspects (e.g. "viewing", "ground floor") that can affect the price.

3.3 Comparing with official statistics

A conceptual issue relates to whether alternative statistics should be substitutes or complements of official ones, i.e. whether or not the former should aim at providing the exact same information as the latter. In the particular case of this paper, it should be noted that web-scraped data represent a different aspect of the housing market from official data. Official statistics are indeed based on transacted prices while web-scraped data account for the price asked by the seller, which can markedly differ from the transacted price or might not even translate into a transaction at all. On top of this, there is a temporal discrepancy since listings data are taken at the start of the process, when the seller puts its dwelling on the market, while official statistics are taken at the very end of it. In addition, statistical discrepancies might arise since the House Price Index – the main index for housing prices in the UK – is computed as a geometrical mean with values corrected using hedonic regressions and mix-adjustment. This is in

contrast with our alternative indices which do not feature similar corrections due to the limited time span in our dataset and the very particular nature of the economic conjecture that might distort the results.[11] We however leave it as an avenue for future research when more data become available. As a consequence, in the particular context of this paper, discrepancies are inevitable between alternative and official statistics, and the former should rather be viewed as a complement of the latter.

While discrepancies with official statistics are inevitable, we however tried to limit biases in our web-scraped data. A key question relates to its representativeness: while a vast majority of real-estate firms post ads on real-estate websites (92% according to Realtor – *op. cit.*[12]), this might entail composition effects due notably to uneven geographical coverage. As shown in **appendix A**, the geographical coverage of our main provider Zoopla is indeed uneven with an over-representation of England. This is the main reason why we complement it with other websites, and in particular PropertyPal and S1Homes which specialize respectively in Northern Ireland and Scotland. It should however be noted that results in **section 4** are based only on Zoopla data since the web-scraping of other websites started only as of July 2020, a caveat to keep in mind. Consequently, the cross-sectional analysis in this section is limited to England for which the geographical coverage of Zoopla is more balanced. In addition, most of the statistics presented in this paper are at granular level, limiting such an issue of composition effects.[13]

[11] In addition, our alternative indices use the arithmetic mean *vs.* the geometric mean in official statistics. That being said, a figure with median prices, less influenced by high-value properties than arithmetic mean, can be found in **Figure 13** in **appendix B**: it conveys the same message as **Figure 6** based on the arithmetic mean. The median house price obtained from web-scraped data is also broadly similar to the statistically-adjusted geometric mean price in Land Registry as shown in **Figure 14**.

[12] This number however likely constitutes an upper bound to the share of "real-life" ads that are captured by web-scraping. Another upper bound is given by comparing the number of transactions from Land Registry during a month, with the number of web-scraped ads that disappear during the same month. Results are provided in **Table 4** in **appendix B**: both figures are on average broadly similar. A lower bound can be computed through the share of transactions in Land Registry matched in our dataset (see **section 5**): using only data from Zoopla and assuming conservative assumptions that limit the matching, such a lower bound is estimated at a relatively low level, e.g 23% in London though there are regional disparities. This relatively limited matching however also reflects the fact that some transactions were not yet registered in official statistics with data retrieved at the time of the study (Nov. 2020).

[13] For the mean price at national level, we compute an alternative index in a bottom-up fashion where regional indices are aggregated and weighted by their average share of dwellings sold from Land Registry data. This weighted mean is shown in **Figure 15** in **appendix B**, which compares to the simple mean of **Figure 5** in **section 4**. Both show the same trend.

4 Real-time monitoring of the supply side of the housing market: an application to the Covid-19 crisis

Using our web-scraped data, we produce daily indicators for the housing market – in advance of official statistics. Interestingly, we can also use location data to construct indicators at a very granular level. This allows to track several metrics (e.g. mean listing price, number of new construction projects) in a timely and granular fashion.

Getting data from listings gives us the perspective from the supply side – in contrast with official statistics based on transaction data. It offers an innovative standpoint: we track the number of new listings which indicates whether sellers are eager to put their properties on the market. A deviation from its "normal" value would signal that sellers are anticipating (or facing) a shock in the housing market. Our dataset also allows to track how the price of a listing fluctuates over time. Again, a deviation from the "normal" trend of corrections (even under stable economic conditions, sellers might adjust their prices over time to correct for initial under-/over-valuations) would indicate potential crisis/bubble episodes. Compared to official statistics based on transactions, an advantage of our approach based on listings is their timeliness – standing at an earlier stage of the housing sales process, possibly allowing for a swifter detection of turning points in the housing cycle.

While very timely, listings however offer only a picture of the supply side of the housing market. The housing market equilibrium relies also on the dynamics on the demand side – where incentives and behaviours might differ from those of the supply side (e.g. Ball et al., 2010, Caldera and Johansson, 2013, Han and Strange, 2015, Yoshino et al., 2015). By contrast, transactions inform on the outcome of the market equilibrium, providing a more balanced view. This is a reason why listings data can't substitute completely for transaction data (Kolbe et al., 2021), but literature still points to the informative power of listings.[14] For example, Bauer et al., 2013 argue that price trends can be measured, and at a higher frequency, with listings. As such, multiple studies have used listings to evaluate the effect of policies on housing prices (e.g. a nuclear plant closing in Bauer et al., 2017 or energy-efficiency regulations in Kholodilin et al., 2017). In the same vein, Perez-Rave et al., 2019 show how listings data can fit hedonic regression when transaction data are not available. More specifically for the UK, Korhonen, 2015 argues that changes in asking prices can be used to assess housing market conditions. On a more theoretical standpoint, search models for the housing market indicate a key role for asking price in directing the buyer search, notably in bust markets (Han and

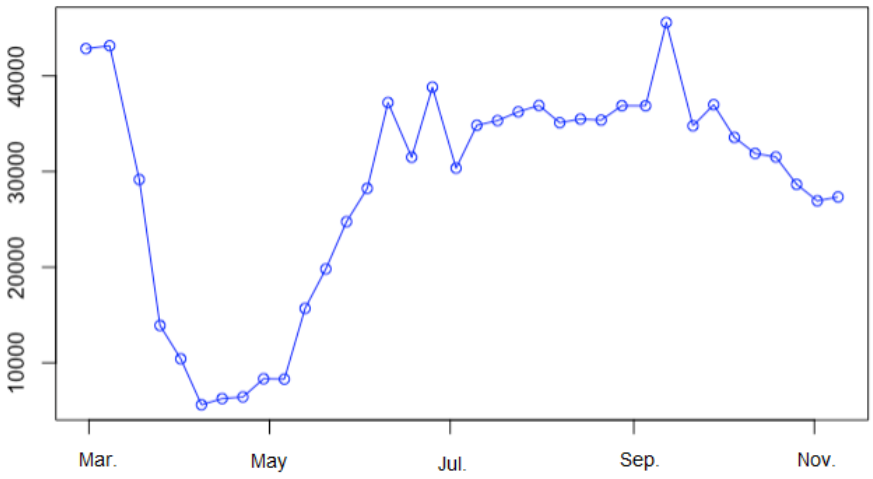
[14]It should be noted that the question of substitutability existed before real-estate websites, with for example Pollakowski, 1995 comparing transaction data with the U.S. Multiple Listing Service, and finding that the latter are well suited for local studies.

Strange, 2016). Literature more generally to the role of the supply side in shaping the overall housing cycle (Anundsen and Jansen, 2013, Accetturo et al., 2021). All-in-all, rather than a substitute to transaction data, listings provide a timely index somewhat capturing price trends and predicting turning points of the housing cycle – as shown empirically in Kolbe et al., 2021.[15]

4.1 An application: UK real-estate amidst the Covid-19 crisis

A first indicator is the number of new listings – which indicates whether sellers find it is the appropriate time to sell their property. In **Figure 3**, we display the weekly number of new listings.[16] This has clearly been 5 to 6 times less important during the lockdown than after or before this period.

Figure 3: Weekly new listings
Source: Zoopla and authors' calculation



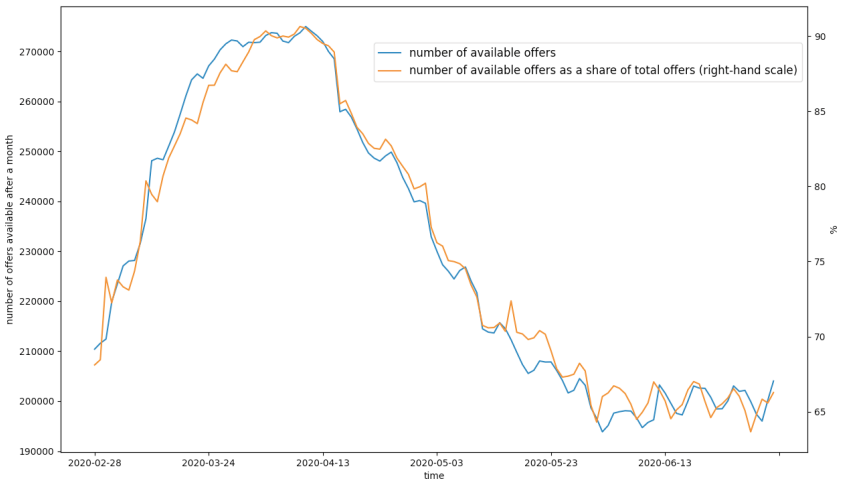
Another indicator is the number of listings that are still available after one month on the website. This is displayed in **Figure 4** on which a strong lockdown effect appears as

[15] **Section 5** also analyses the differences between transaction and listing prices, highlighting the frictions prevailing to the market equilibrium measured by transactions. In complement, **appendix C** uses real-time Internet-based information on the demand side, documenting a Covid-19-related drop in Internet searches for housing on Google Trends and SimilarWeb. Finally, **Section 3** also provides empirical evidence of the usefulness of listings data, since comparing price indices based on the latter with official statistics – based on transaction data – shows similar trends.

[16] While data are available at daily frequency, looking at a weekly sum has the advantage of taking care of intra-weekly seasonality in listing posting.

around 90% of the listings posted at the beginning of April were still available one month later. After the lockdown, the proportion drops progressively and stabilizes around its pre-lockdown level at about two thirds of the total.

Figure 4: Number of listings available after one month on the website
Source: Zoopla and authors' calculation



We now turn to daily prices in **Figure 5** showing the mean price at national level. This figure and the monthly growth rate recapitulated in **Table 2** show a slight decrease from March to the beginning of May, followed by a steady increase after the lockdown. This figure also highlights the role of filtering which removes non-residential listings (e.g. garage, parking, lands) whose average price tends to be lower.

The granularity of our dataset also allows us to explore listing price trends by categories: available in **appendix B**, **Figures 16**, **17**, and **18** analyse price level and evolution by quantiles while **Figure 19** shows the prices by dwelling categories (studio; 2-rooms, etc.). Those decompositions are particularly key to alleviate concerns over composition effects that might distort our indices since the lack of surface data impairs computing a price index per squared meter.[17] In the particular case of the Covid-19 crisis, breaking down the sample by price quantiles or dwelling categories suggests that price trends have been shared across all types of dwellings.

[17]It should however be noted that the same issue arises in official statistics which do not provide such price index per surface.

Figure 5: Evolution of the mean price of listings
Source: Zoopla and authors' calculation

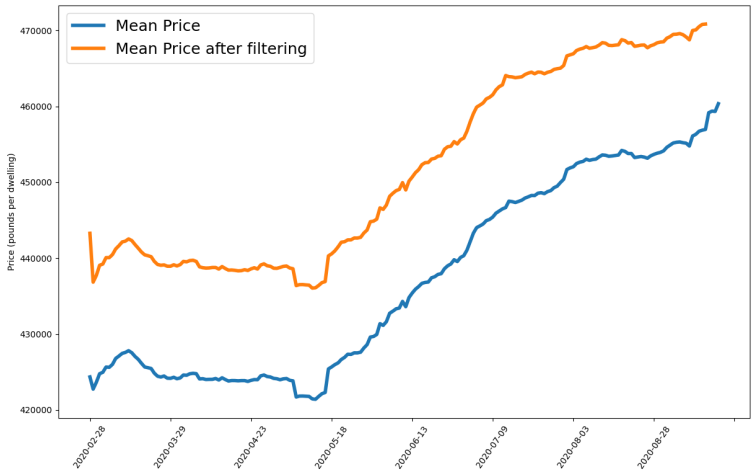
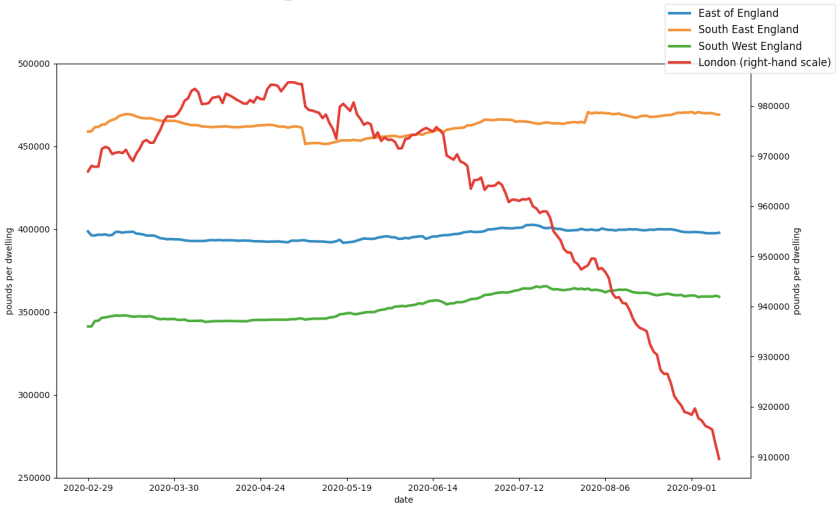


Table 2: Evolution of monthly mean listing price (filtered data)

Metrics	March	April	May	June	July	August
Mean price (pounds)	439,998	438,847	439,816	451,566	463,071	467,885
Monthly change (%)	-	-0.26	0.22	2.67	2.55	1.04

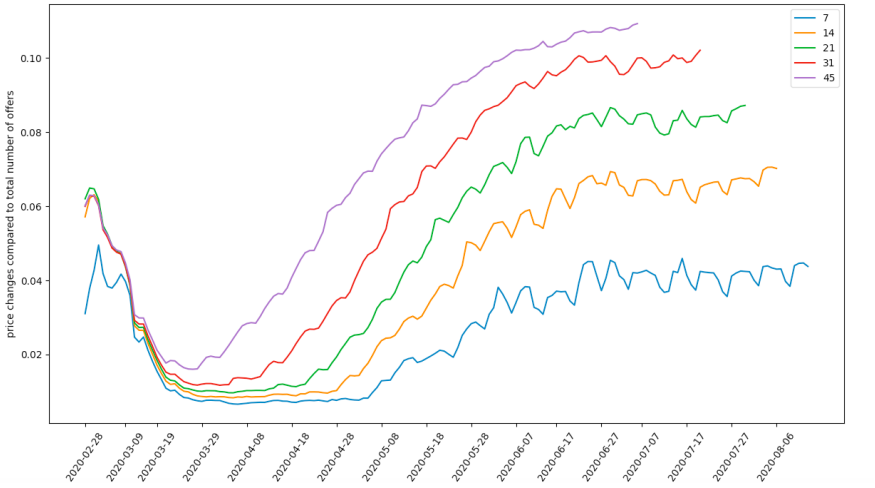
Taking advantage of the granularity of our dataset, prices can also be observed at local level. In **Figure 6** we represent price trends for four sub-regions: East England, South-East England, South-West England, and London. While trends are relatively similar for the first three, the case of London stands out with a continued decline from May to the end of August. This may be due to a fall in price per squared meter, a composition effect with more transactions taking place for smaller surfaces or a mix of the two.

Figure 6: Mean listing price per region
Source: Zoopla and authors' calculation



Using our web-scraped data, we can also focus on how the price fluctuates for each listing. Using a listing posted at a given day, we check if it is still available in the next day and whether its price has changed; we then repeat the same process day after day. Comparing listings for which the price changes over the total number of listings, it appears that only a minority of listings (between 8 and 10%) experience price changes with 90% of these changes being downward revisions (see **Figure 20** in **appendix B**). In **Figure 7**, we take a look at the share of listings with price changes after 7, 14, 21, 31 and 45 days on the market. It shows that at the beginning of the lockdown, sellers were not yet ready to change their prices and adopted a "wait-and-see" approach: a significantly lower share of listings experienced price changes compared to pre-lockdown period – even for listings that were in the market for an extended period of time (45 days). However, after the lockdown, the share of listings whose price changes has grown rapidly and stabilized around levels that appear even higher than pre-lockdown levels for listings already staying for a longer period in the market.

Figure 7: Share of listings with prices revisions, by number of days already on the market
Source: Zoopla and authors' calculation



Further analysis on the web traffic related to UK housing market, based on alternative data from Google Trends and SimilarWeb can be found in **appendix C**. In particular, it corroborates our finding that price trends have been quite dissimilar in London compared to rural areas since the lockdown.

5 Alternative indicators

On top of monitoring in real-time the supply side of the housing market through the prices of listings and sellers' behaviour, our dataset allows us to derive alternative indicators to measure complementary dimensions in the housing market. In particular, focusing on the supply side allows us to measure the difference between listings prices (asked by sellers) and transacted prices (eventually reached between a seller and a buyer), therefore providing an estimate for the buyers' negotiation margin. Such a metrics has been proven useful in the literature to assess conditions of the housing market (e.g. Han and Strange, 2016, Bo, 2018, and Ngai and Tenreyro, 2014). It notably provides, from a policy perspective, an indication of tensions in the housing market. Frictions on the housing market can also be proxied through other early indicators such as the rent-to-price or price-to-income ratios which have been extensively analyzed in the literature (e.g. Lind, 2009 or Bunda and Ca'Zorzi, 2010) – and put into use by regulators to anticipate housing bubbles (e.g. Philiponnet and Turrini, 2017 or Kelly et al., 2019).

5.1 Measuring buyers' negotiation margin

This paper sheds light on the supply side of housing by retrieving the price that sellers are expecting for their dwellings – in contrast with official statistics based on the prices of transactions. We have shown how these data can be exploited to monitor the supply side of the housing market in real-time in **section 4**. We now turn to comparing our dataset on listing prices with notarial data on transacted prices: while the latter reflects prices after negotiation, the former reflect the price initially expected by the seller. Comparing the two therefore provides information about the negotiation margin of the buyers $M_{r,t}$ for a region r at a time t as the median of all the individual negotiation margins for dwellings j transacted and defined – in line with Galesi et al., 2020[18] – as:

$$\frac{ListingPrice_{j,t} - TransactionPrice_{j,t}}{ListingPrice_{j,t}} \tag{1}$$

Land Registry data provide prices for all real-estate transaction in England and Wales. In this dataset, dwellings are identified by their exact address, the type of property (e.g. detached, semi-detached, flat), the type of dwelling (new-build or not), the transaction price, and the date of the transaction. Since there is no predefined key to match these notarial data with our web-scraped data, we design a matching algorithm based on the K-Nearest-Neighbours (KNN) algorithm first developed by Fix and Hodges, 1951 and widely used in the machine learning literature (see for example Hastie et al., 2009). A first element of comparison would naturally be the address, however listings in our web-scraped dataset are generally not identified by their exact location: in most cases the most granular locatioanl information is the name of the street. The first step is then to transform addresses in both datasets into longitude and latitude coordinates using *openstreetmap* – a collaborative project creating a free editable world map – in order to compare the distance between dwellings in Land Registry (exact location) and the web-scraped dataset (approximate location). The algorithm then takes the following sequential steps:

1. Starting from a transaction in the Land Registry, it searches whether observations in the web-scraped share the same longitude and latitude (i.e. at the highest locational precision possible provided by *openstreetmap*). If no ad is found, the algorithm stops and no matching is done for this transaction,

2. On the selected ads, the algorithm retains only those which share the same type of property (detached, semi-detached, etc.) and the same type of dwelling (new

[18]However, Galesi et al., 2020 compute the regional index as the average of the individuals margins.

or old). If no ad remains, the algorithm stops and no matching is done for this transaction,

3. On the remaining ads, the algorithm retains only those which share the same time range – more precisely those ads that are present at least one time in a range of three months before and two months after the date of the transaction. If no ad remains, the algorithm stops and no matching is done for this transaction,

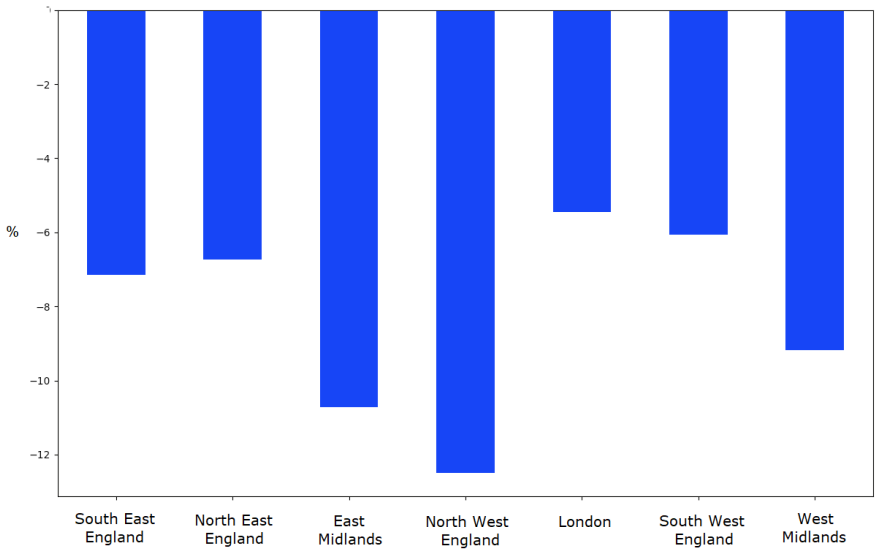
4. On the remaining ads, the algorithm retains only those which fall within a similar price range. Specifically, we assume that the transacted price should have experienced a maximum decrease of 35% and a maximum increase of 35% compared to the listing price. If no ad remains, the algorithm stops and no matching is done for this transaction,

5. At this stage, generally at most a single corresponding ad remains. In the few cases where several ads remain, the algorithm takes the average of their prices as they might be considered as indistinguishable with respects to the transaction.

As a consequence of these constraints, not all transactions are matched (around 30% when using only Zoopla data) but this ensures that only very similar dwellings are matched – a pattern that we manually verify. Step 1 in particular eliminates a number of potential matches given the imprecise location of the web-scraped data, notably in rural regions where a street might cover a wider area.

Based on this methodology, **Figure 8** shows an estimation of the buyers' negotiation margin at local level computed as the average percent change of the transacted price over the lastly-posted listing price – meaning that if the price on a dwelling changes over time, only the last listing price is kept in the matching with Land Registry. A first observation relates to the fact that transacted prices are lower than listing prices across all regions, as indicated by a negative negotiation margin. Second, buyers' negotiation margin tends to be largely lower in London – as would be expected in a tense housing market. On the opposite, the negotiation margin is the double of London's in North-West England and the East Midlands. These large discrepancies across regions validate *ex post* our granular approach. Finally, while no noticeable change in the negotiation margin can be detected over our time sample, the temporal dimension should come more relevant as more data become available – the web-scraping started only as of March 2020 – in complement of the cross-section analysis presented here.

Figure 8: Negotiation margin per region
Source: Real-estate websites, notarial data, and authors' calculation



5.2 Rent-to-price ratio

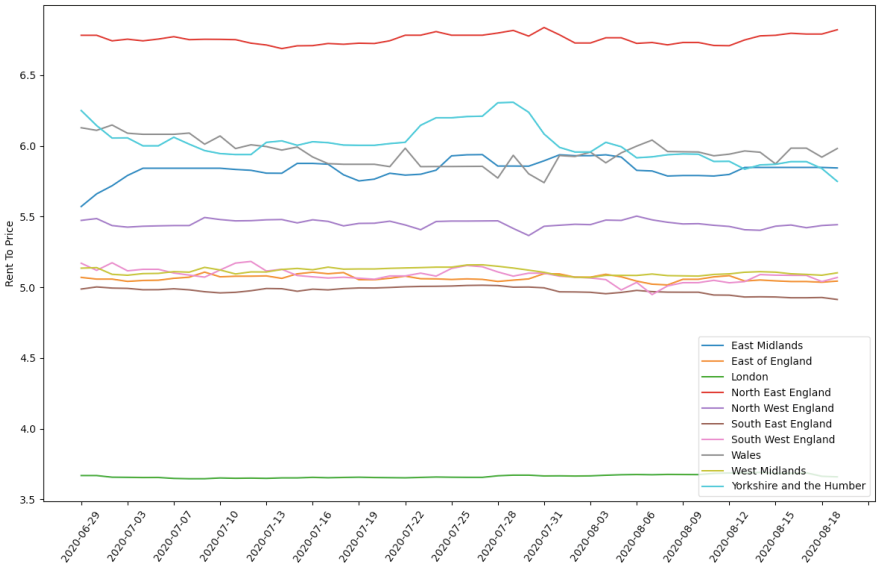
A well-established early indicator is the rent-to-price ratio (see for example Campbell et al., 2009 and Engsted and Pedersen, 2015) which measures whether it is cheaper to rent or own a property. It is used as an indicator of whether housing prices are overvalued and might also be viewed as a proxy for yields in the housing market – not taking into account taxation. The innovation brought by the web-scraped data is the ability to compute such indicators in real-time and more importantly in a very granular fashion. On top of the rent-to-price ratio, other indicators – the price-to-income ratio and a novel purchasing-power-capacity – are also shown in **appendix E**.

We produce a rent-to-price estimation by matching a listing to rent in our dataset with listings for sale sharing similar characteristics (ZIP code level, area, number of living rooms, bedrooms, and bathrooms). To construct such an indicator, we also rely on machine learning techniques and use again a KNN algorithm. We compute the rent-to-price for an individual listing i as:

$$RTP_i = \frac{AnnualRent_i}{\frac{1}{K}\sum_{j=1}^{K} HomePrice_j} \tag{2}$$

In our approach, we match every listing with its $K = 10$ closest ones: for the whole dataset, the rent-to-price ratio stands around 4.1%.[19] Our web-scraped data allows us to produce this indicator on a daily basis. Another contribution of our approach lies in the possibility to provide such an indicator at a very granular level. Regional rent-to-price ratios are represented in **Figure 9**. It validates our granular approach since large discrepancies are shown across regions. In particular, the London area stands out with a very low rent-to-price ratio which might signal to some extent an over-valuation of the housing market – corroborating the findings of Marsden, 2015 or more recently Petris et al., 2022 for some London's boroughs – or the fact that this location is considered less risky and that corresponding yield integrates a smaller risk premium.

Figure 9: Rent-to-price ratio per region
Source: Real-estate websites and authors' calculation



Due to heterogeneous coverage of UK regions in our web-scraped data,[20] the rent-to-price ratio for the whole dataset cannot however be considered as a rent-to-price at the national level due to composition effects. Using the regional rent-to-price ratios, we

[19]Note that this depends on K. Rent-to-price ratio varies between 3.7 and 4.1 when taking K between 10 and 50. However if K is larger, a rental listing will be paired with listings for sale that are less and less related. In addition, if K is larger, the matching would require more data at local level – When this is not the case, the matching is impossible. Therefore when K is larger, matching can only be performed in large cities, distorting the computation of the rent-to-price ratio. For those reasons, we choose $K = 10$ which appears sufficiently low to match a listing only with comparable ones but sufficiently large to have a statistically meaningful number of listings to compare with.

[20]See **appendix A** for the coverage of Zoopla. As explained above, the analysis afferent to the Covid-19 period is only computed based on Zoopla data since the web-scraping of other websites has only started later in July 2020.

compute a national index using the share of each region in the total number of dwellings at national level. We find a ratio around 5.1% – to be compared with the 4.1% obtained for the unweighted index. This reflects notably the fact that large urban areas with lower ratios are over-represented in our dataset.

6 Conclusion

By web-scraping main real-estate websites in the UK, this paper builds daily indicators that monitor the supply side of the housing market in real-time. This approach allows for timely and highly granular (ZIP code level) indicators. By monitoring the supply side of the housing market, we are able to define a number of innovative indicators pertaining to supply such as the number of new listings – indicating buyers' willingness to put their properties on the market – and price fluctuations for existing listing – showing whether sellers are eager to lower their prices to adapt to economic conditions. A key interest of our data is to provide a very early signal of the housing market dynamics. Finally, matching our web-scraped data on listings with notarial data on transactions allows us to assess the buyers' negotiation margin.

We use these innovative data to monitor the UK housing market during the Covid-19 crisis. The lockdown has been characterized by a freezing of activity in terms of new listings posted and of price adjustments for existing listings. It appears therefore that sellers adopted a "wait-and-see" approach. However, evidence shows that they have been more ready to lower their prices afterwards – in particular for listings that were already on the market for a long time. Mean price slightly decreases during the lockdown until the beginning of May 2020, after which it started to increase at national level. This trend however hides regional disparities as the London area has been experiencing a continued decline in mean listing price since the lockdown. We finally compute rent-to-price ratios at very granular level, an early indicator for imbalances on the housing market. While there is no evidence for changes in this indicator during the Covid-19 period, it displays large heterogeneities across regions – validating somehow our granular endeavour.

A potential limitation is however the lack of timespan since data have been collected only since March 2020. While it captures the Covid-19 crisis, the lack of timespan makes it challenging to properly correct for seasonality issues that might bias the indicators. It also makes it difficult to test for the predictive capacities of our innovative indicators. Another limitation is the fact that we cover only the UK, though the approach can be extended seamlessly to other countries – as long as their main real-estate websites allow for web-scraping.

Finally, this analysis can be usefully extended to land prices which, combined with other costs (especially construction costs), and possibly compared with the prices of new dwellings, can give insights about under-/over-investment in the construction sector – see for example Bricongne and Pontuch, 2017.[21] Another avenue for future work relates to computing prices indices for commercial real-estate – a category mostly uncovered by official statistics. Finally, web-scraped data can serve in econometric models – for example in nowcasting housing prices or modelling the housing price convergence at longer-term horizons.

[21] Preliminary analysis for land prices are performed in **appendix D**.

References

Accetturo, A., Lamorgese, A., Mocetti, S., & Pellegrino, D. (2021). Housing supply elasticity and growth: Evidence from Italian cities. *Journal of Economic Geography*, *21*, 367–396.

André, C., Gil-Alana, L., & Gupta, R. (2014). Testing for persistence in housing price-to-income and price-to-rent ratios in 16 OECD countries. *Applied Economics*, *46*(18), 2127–2138.

Anundsen, A., & Jansen, E. (2013). Self-reinforcing effects between housing prices and credit. *Journal of Housing Economics*, *22*, 192–212.

Askitas, N. (2015). Trend-spotting in the housing market. *Institute of Labor Economics (IZA) Discussion Papers*, *No 9427*.

Ball, M., Meen, G., & Nygaard, C. (2010). Housing supply price elasticities revisited: Evidence from international, national, local and company data. *Journal of Housing Economics*, *19*(4), 255–268.

Bauer, T., Braun, S., & Kvasnicka, M. (2017). Nuclear power plant closures and local housing values: Evidence from Fukushima and the German housing market. *Journal of Urban Economics*, *99*, 94–106.

Bauer, T., Feuerschutte, S., Kiefer, M., an de Meulen, P., Micheli, M., Schmidt, T., & Wilke, L. (2013). Ein hedonischer immobilienindex auf basis von internetdaten: 2007–2011. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *7*, 5–30.

Blot, C., Hubert, P., & Odry, R. (2018). Are there asset price bubbles in the euro zone? *Revue de l'OFCE*, *158*, 183–205.

Bo, E. (2018). Housing match quality and demand: What can we learn from comparing buyer characteristics? *Journal of Housing Economics*, *41*, 184–199.

Boeing, G., & Waddell, P. (2017). New insights into rental housing markets across the United States: Web scraping and analyzing Craigslist rental listings. *Journal of Planning Education and Research*, *37*, 457–476.

Bricongne, J.-C., Coffinet, J., Delbos, J.-B., Kaiser, V., Kien, J.-N., Kintzler, E., Lestrade, A., Meunier, B., Mouliom, M., & Nicolas, T. (2020). Tracking the economy during the Covid-19 pandemic: The contribution of high-frequency indicators. *Bulletin de la Banque de France*, *231*.

Bricongne, J.-C., Meunier, B., & Pical, T. (2021). Can satellite data on air pollution predict industrial production? *Banque de France Working papers*, *No 847*.

Bricongne, J.-C., & Pontuch, P. (2017). La crise s'est-elle accompagnée d'un sous-investissement dans l'immobilier résidentiel dans les pays de la zone euro ? *Revue Economique*, *68*, 129–157.

Bricongne, J.-C., Turrini, A., & Pontuch, P. (2019). Assessing house prices : Insights from « Houselev », a dataset of price level estimates. *European Commission Discussion Paper*, *No 101*.

Bunda, I., & Ca'Zorzi, M. (2010). Signals from housing and lending booms. *Emerging Markets Review*, *11*(1), 1–20.

Caldera, A., & Johansson, A. (2013). The price responsiveness of housing supply in OECD countries. *Journal of Housing Economics*, *22*(3), 231–249.

Campbell, S., Davis, M., Gallin, J., & Martin, R. (2009). What moves housing markets: A variance decomposition of the rent–price ratio. *Journal of Urban Economics*, *66*(2), 90–102.

Carrillo, P. (2013). To sell or not to sell: Measuring the heat of the housing market. *Real estate economics*, *41*(2), 310–346.

Carrillo, P., de Wit, E., & Larson, W. (2015). Can tightness in the housing market help predict subsequent home price appreciation? Evidence from the United States and the Netherlands. *Real estate economics*, *43*(3), 609–651.

Case, K., & Schiller, R. (1990). The efficiency of the market for single-family homes. *American Economic Review*, *79*(1), 125–137.

Case, K., & Schiller, R. (2003). Is there a bubble in the housing market? *Brookings Papers on Economic Activity*, *2*(2003), 299–362.

Case, K., & Wachter, S. (2005). Residential real estate price indices as financial soundness indicators: Methodological issues (BIS, Ed.). In BIS (Ed.), *Real estate indicators and financial stability*.

Cerdeiro, D., Komaromi, A., Liu, Y., & Saeed, M. (2020). World seaborne trade in real time: A proof of concept for building AIS-based nowcasts from scratch. *International Monetary Fund Working Papers*, *No 2020/57*.

Cesa-Bianchi, A. (2013). Housing cycles and macroeconomic fluctuations: A global perspective. *Journal of International Money and Finance*, *37*, 215–238.

Chapelle, G., & Eymeoud, J.-B. (2018). Can big data increase our knowledge of local rental markets? Estimating the cost of density with rents. *Sciences Po Paris*, *mimeo*.

Chen, S., Igan, D., Pierri, N., & Presbitero, A. (2020). Tracking the economic impact of Covid-19 and mitigation policies in Europe and the United States. *Covid Economics: Vetted and Real-Time Papers*, *36*, 1–24.

Cheng, I.-H., Raina, S., & Xiong, W. (2014). Wall Street and the housing bubble. *American Economic Review*, *104*(9), 2797–2829.

Chetty, R., Friedman, J., Hendren, N., Stepner, M., & Team, T. O. I. (2020). How did Covid-19 and stabilization policies affect spending and employment? A new real-time economic tracker based on private sector data. *National Bureau of Economic Research Working Papers*, *No 27431*.

Coibion, O., Gorodnichenko, Y., & Weber, M. (2020). Labor markets during the Covid-19 crisis: A preliminary view. *National Bureau of Economic Research Working Papers*, *No 27017*.

Dixon, W., & Tukey, J. (1968). Approximate behavior of the distribution of winsorized t (trimming / winsorization 2). *Technometrics*, *10*(1), 83–98.

Dujardin, M., Kelber, A., & Lalliard, A. (2015). Overvaluation in the housing market and returns on residential real estate in the euro area: Insights from data in euro per square metre. *Banque de France Bulletin Quarterly Selection*, *37*, 49–63.

Dumbacher, B., & Capps, C. (2016). Big data methods for scraping government tax revenue from the web, In *Proceedings of the joint statistical meetings, section on statistical learning and data science*.

Engsted, T., Hviid, S., & Pedersen, T. (2016). Explosive bubbles in house prices? Evidence from the OECD countries. *Journal of International Financial Markets, Institutions and Money*, *40*, 14–25.

Engsted, T., & Pedersen, T. (2015). Predicting returns and rent growth in the housing market using the rent-price ratio: Evidence from the OECD countries. *Journal of International Money and Finance*, *53*(100), 257–275.

Fix, E., & Hodges, J. (1951). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *USAF School of Aviation Medicine report.*

Flood, J. (1997). Urban and housing indicators. *Urban Studies*, *34*, 1635–1665.

Galesi, A., Mata, N., Rey, D., Schmitz, S., & Schuffels, J. (2020). Regional housing market conditions in Spain. *Maastricht University Graduate School of Business and Economics Research Memoranda*, *No 29*.

Gan, Q., & Hill, J. (2009). Measuring housing affordability: Looking beyond the median. *Journal of Housing Economics*, *18*(2), 115–125.

Garcia-López, M.-A., Jofre-Monseny, J., Martínez-Mazza, R., & Segú, M. (2020). Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona. *Journal of Urban Economics*, *119*.

Giudice, V. D., Paola, P. D., & Giudice, F. D. (2020). Covid-19 infects real estate markets: Short and mid-run effects on housing prices in Campania region (Italy). *Social Sciences*, *9*.

Goodman, A. (1988). An econometric model of housing price, permanent income, tenure choice, and housing demand. *Journal of Urban Economics*, *23*(3), 327–353.

Han, L., & Strange, W. (2015). Chapter 13 - the microstructure of housing markets: Search, bargaining, and brokerage (G. Duranton, J. Henderson, & W. Strange, Eds.). In G. Duranton, J. Henderson, & W. Strange (Eds.), *Handbook of regional and urban economics*.

Han, L., & Strange, W. (2016). What is the role of the asking price for a house? *Journal of Urban Economics*, *93*, 115–130.

Hanson, A., & Santas, M. (2014). Field experiment tests for discrimination against hispanics in the US rental housing market. *Southern Economic Journal*, *81*, 135–167.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning, second edition*. Springer.

Horn, K., & Merante, M. (2017). Is home sharing driving up rents? Evidence from Airbnb in Boston. *Journal of Housing Economics*, *38*, 14–24.

Jardet, C., & Meunier, B. (2022). Nowcasting world GDP growth with high-frequency data. *Journal of Forecasting*.

Jose, V., & Winkler, R. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, *24*, 163–169.

Kelly, J., Blanc, J. L., & Lydon, R. (2019). Pockets of risk in European housing markets: Then and now. *European Systemic Risk Board*, *No 87*.

Kholodilin, K., Mense, A., & Michelsen, C. (2017). The market value of energy efficiency in buildings and the mode of tenure. *Urban Studies*, *54*, 3218–3238.

Kiyotaki, N. (1998). Credit and business cycles. *The Japanese Economic Review*, *49*(1), 18–35.

Kiyotaki, N., & Moore, J. (1997). Credit cycles. *Journal of Political Economy*, *105*(2), 211–248.

Kolbe, J., Schulz, R., Wersing, M., & Werwatz, A. (2021). Real estate listings and their usefulness for hedonic regressions. *Empirical Economics*, *61*, 3239–3269.

Korhonen, P. (2015). Chance favours the prepared mind: What linked micro data can tell us about the housing market. *Bank Underground*, *20 July*. https://bankunderground.co.uk/2015/07/20/chance-favours-the-prepared-mind-what-linked-micro-data-can-tell-us-about-the-housing-market/

Kulkarni, R., Haynes, K., Stough, R., & Paelinck, J. (2009). Forecasting housing prices with Google Econometrics. *GMU School of Public Policy Research Paper*, *No 2009-10*.

Lewis, D., Mertens, K., Trivedi, M., & Stock, J. (2021). Measuring real activity using a weekly economic index. *Journal of Applied Econometrics*.

Lind, H. (2009). Price bubbles in housing markets: Concept, theory and indicators. *International Journal of Housing Markets and Analysis*, *1*(1), 78–90.

Loberto, M., Luciani, A., & Pangallo, M. (2018). The potential of big housing data: An application to the Italian real-estate market. *Banca d'Italia Working Papers*, *No 1171*.

Marsden, J. (2015). House prices in London – an economic analysis of London's housing market. *Greater London Authority Economics Working Papers*, *No 72*.

McLaren, N., & Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, *51*, 134–140.

Mense, A., Michelsen, C., & Kholodilin, K. (2017). Empirics on the causal effects of rent control in Germany. *FAU Discussion Papers in Economics*, *No 24*.

Miller, N., & Sklarz, M. (1986). A note on leaning indicators of housing market price trends. *Journal of Real Estate Research*, *1*(1), 99–109.

Ngai, L., & Tenreyro, S. (2014). Hot and cold seasons in the housing market. *American Economic Review*, *104*(12), 3991–4026.

Oust, A., & Martin, E. O. (2018). Can Google search data be used as a housing bubble indicator? *European Real Estate Society*, *No 2018-151*.

Ozhegov, E., & Sidorovykh, A. (2017). Heterogeneity of sellers in housing market: Difference in pricing strategies. *Journal of Housing Economics*, *37*, 42–51.

Pavlicek, J., & Kristoufek, L. (2019). Modeling UK mortgage demand using online searches. *Charles University Prague Working Papers IES*, *No 2019/18*.

Perez-Rave, J., Correa-Morales, J.-C., & Gonzalez-Echavarria, F. (2019). A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *Journal of Property Research*, *36*, 59–96.

Petris, P., Dotsis, G., & Alexakis, P. (2022). Bubble tests in the London housing market: A borough level analysis. *International Journal of Finance and Economics*, *27*(1), 1044–1063.

Philiponnet, N., & Turrini, A. (2017). Assessing house price developments in the eu. *European Commission Discussion Paper*, *No 048*.

Polidoro, F., Giannini, R., Conte, R., Mosca, S., & Rossetti, F. (2015). Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. *Statistical Journal of the IAOS*, *31*, 165–176.

Pollakowski, H. (1995). Data sources for measuring house price changes. *Journal of Housing Research*, *6*, 377–387.

Poon, J., & Garratt, D. (2012). Evaluating UK housing policies to tackle housing affordability. *International Journal of Housing Markets and Analysis*, *5*(3), 253–271.

Quigley, J. (2001). Real estate and the Asian crises. *Journal of Housing Economics*, *10*, 129–161.

Rünstler, G., & Vlekke, M. (2018). Business, housing, and credit cycles. *Journal of Applied Econometrics*, *33*, 212–226.

Ruppert, D. (2014). Trimming and winsorization (I. John Wiley Sons, Ed.). In I. John Wiley Sons (Ed.), *Wiley statsref: Statistics reference online*.

Slacalek, J. (2009). What drives personal consumption? The role of housing and financial wealth. *The B.E. Journal of Macroeconomics*, *9*, 1–37.

Tukey, J., & McLaughlin, D. (1963). Less vulnerable confidence and significance procedures for location based on a single sample (trimming / winsorization 1). *Sankhyā: The Indian Journal of Statistics, Series A*, *25*(3), 331–352.

Veldhuizen, S. V., Vogt, B., & Voogt, B. (2016). Internet searches and transactions on the Dutch housing market. *Applied Economics Letters*, *23*, 1321–1324.

Veronese, G., Biancotti, C., Rosolia, A., Kirchner, R., & Mouriaux, F. (2020). Covid-19 and official statistics: A wakeup call?, In *8th IMF statistical forum: Measuring the economics of a pandemic*.

Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales (NBER, Ed.). In NBER (Ed.), *Economic analysis of the digital economy*.

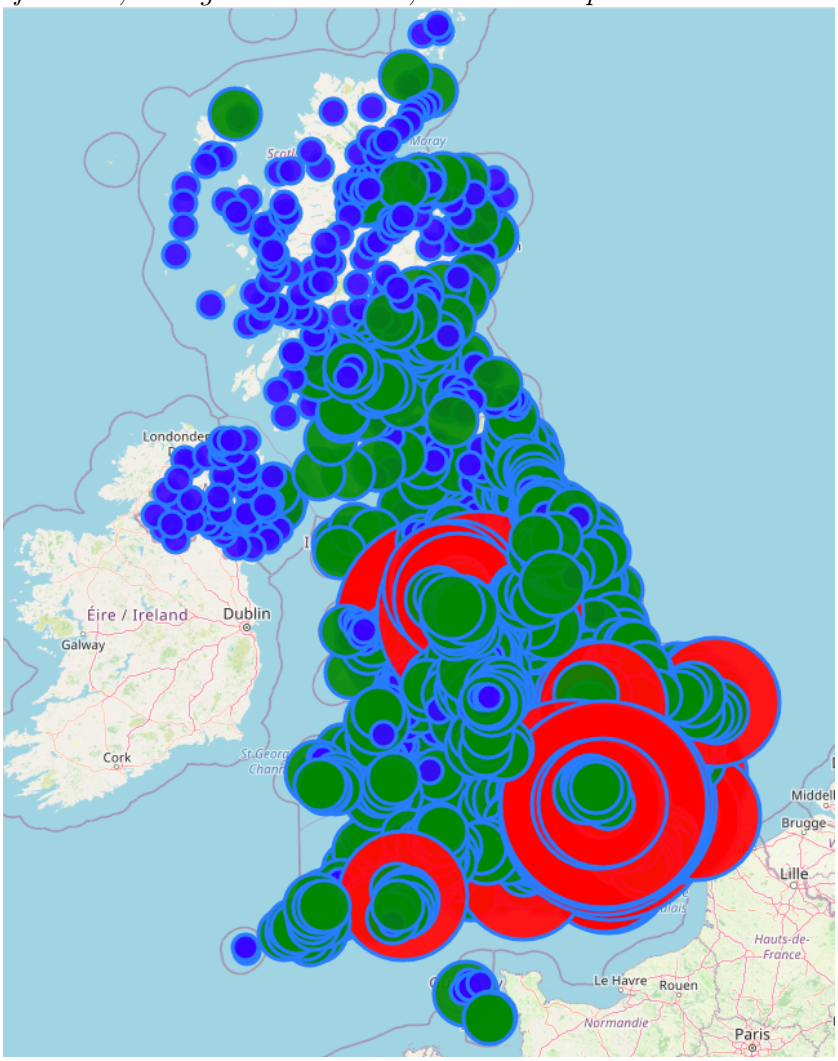
Yoshino, N., Helble, M., & Aizawa, T. (2015). Housing policies for Asia: A theoretical analysis by use of a demand and supply model. *Asian Development Bank Institute Working Papers*, *No 526*.

A Zoopla geographical coverage

In this paper, our main provider is Zoopla whose listings are web-scraped since March 2020 while it started only in July 2020 for other websites. In our analysis of the UK market, it should be kept in mind that the geographical coverage of this provider is however uneven across the UK territory: in the figure below, it appears in particular that data for Scotland and Northern Ireland are limited. This is why we also scrape Propertypal and S1homes, specialized in those two regions.

Figure 10: Zoopla's geographical coverage

The size of a circle represent the average number of daily ads per ZIP code, in red if > 500, blue if < 100, and green otherwise; source: Zoopla and authors' calculation



B Additional graphs

Figure 11: Share of newly built dwellings in the residential real-estate
Source: Zoopla and authors' calculation

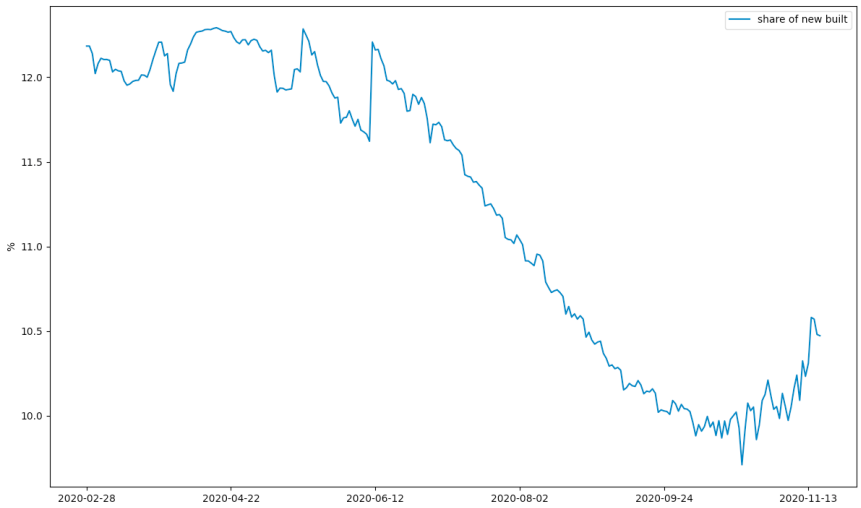


Figure 12: Share of listings filtered from the Zoopla dataset
Source: Zoopla and authors' calculation

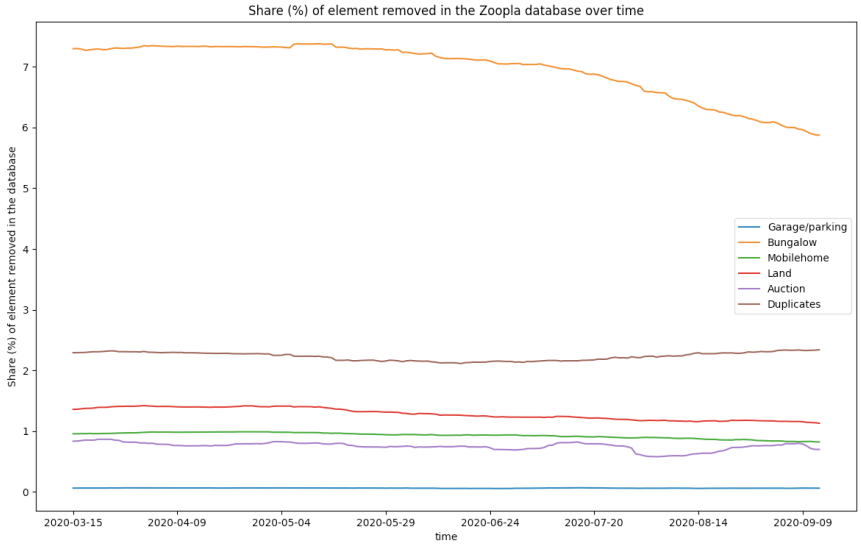


Table 3: Average proportion of missing data per category (%)

Category	Share of missing
City	2.8
ZIP code	0.0
Address	0.0
Price	0.6
Surface area	92.9
Type of transaction (auction *vs.* sale)	0.0
Type of building (new *vs.* existing)	0.0
Number of bedrooms	2.5
Number of bathrooms	16.9
Number of living rooms	19.4
General description	0.0
Type of good (e.g. flat, house)	0.6

Figure 13: Median listing prices by region
Source: Zoopla and authors' calculation

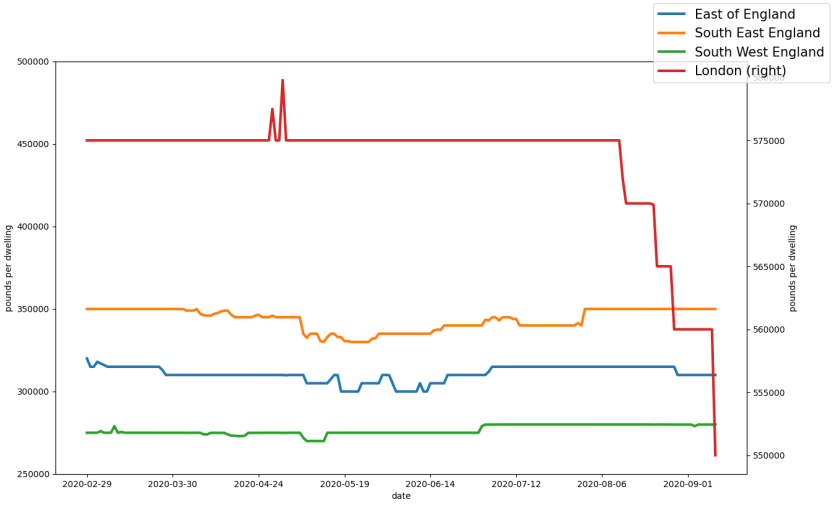


Figure 14: Statistically-adjusted geometric average listing prices by region
In pounds per dwelling, source: Land Registry

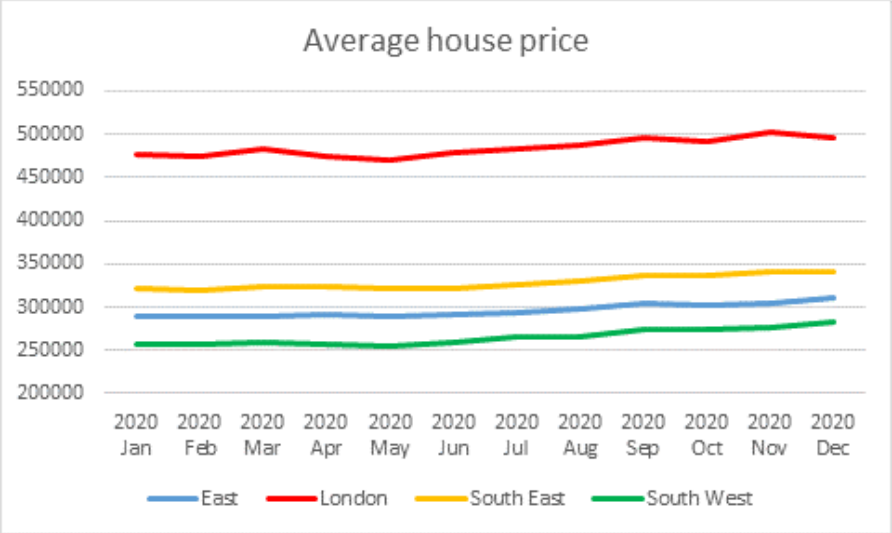


Figure 15: Mean listing prices as a weighted average of regional indices
Source: Zoopla and authors' calculation

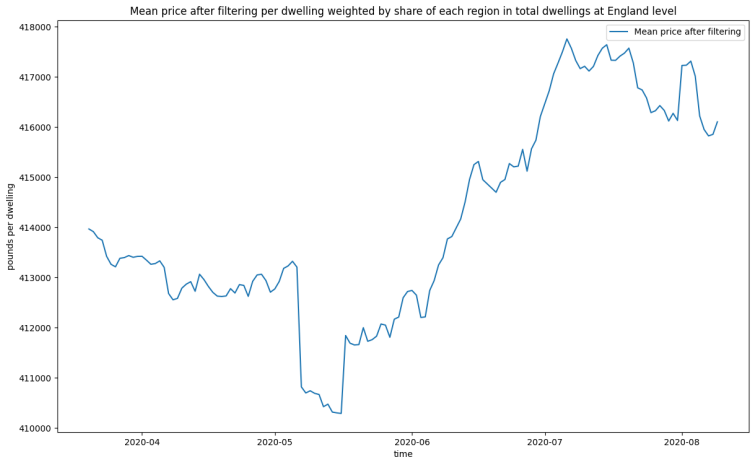


Table 4: Listings removed during a month compared to transactions

Average monthly data					
	Zoopla	Rightmove	OnThe Market	Transactions (Land Registry)	Total dwellings (ONS)
East Midlands	8,832	16,758	9,075	7,378	1,961,000
East of England	14,649	26,590	17,969	8,945	2,520,000
London	23,693	32,274	23,923	8,862	3,318,000
North East England	4,531	8,472	7,097	3,959	1,164,000
North West England	14,691	22,935	11,949	11,172	3,111,000
South East England	25,838	41,095	28,253	12,974	3,683,000
South West England	13,053	23,346	16,428	8,909	2,403,000
West Midlands	10,160	17,378	11,645	7,995	2,358,000
Yorkshire and the Humber	9,664	15,505	11,116	8,175	2,294,000

Figure 16: Weekly percentage change per quantile
Source: Zoopla and authors' calculation

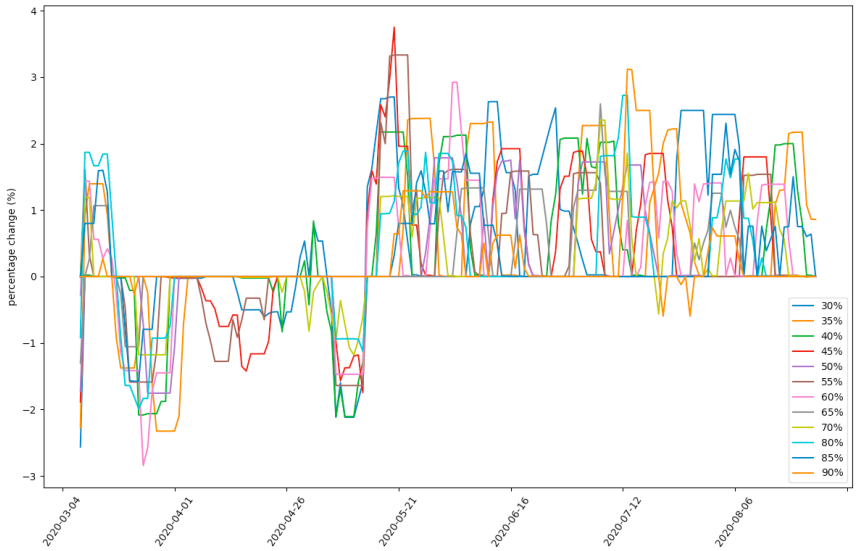


Figure 17: Monthly percentage change per quantile
Source: Zoopla and authors' calculation

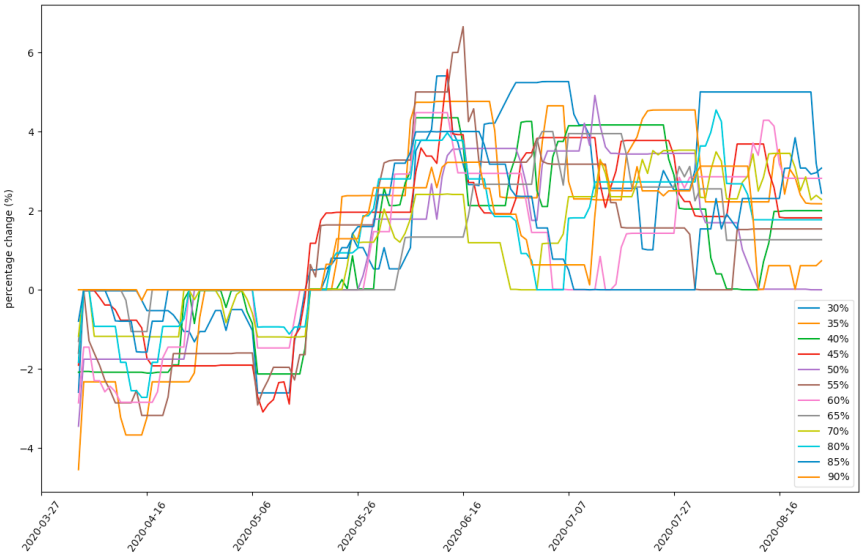


Figure 18: Price by quantile
Source: Zoopla and authors' calculation

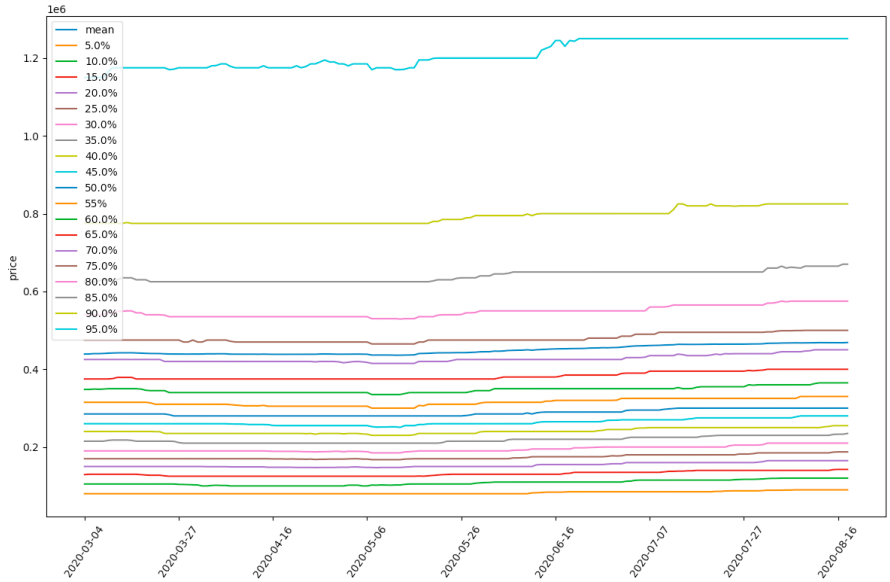


Figure 19: Mean price by dwelling type
Source: Zoopla and authors' calculation

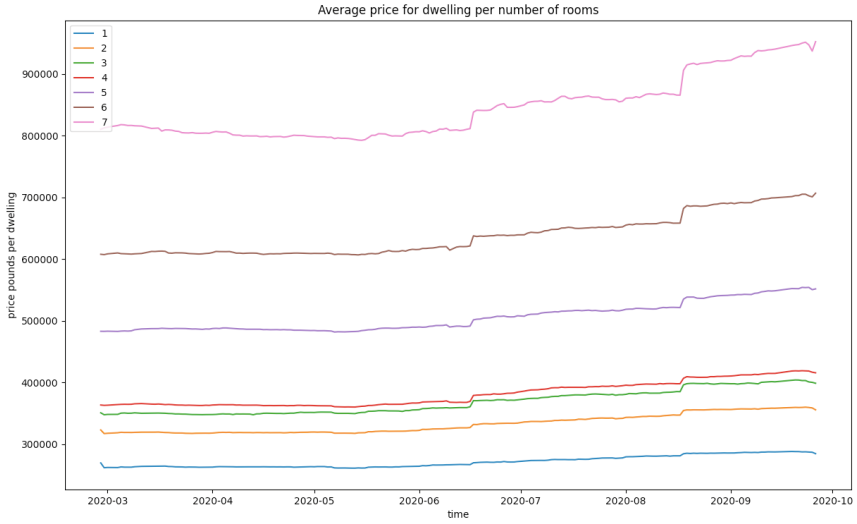
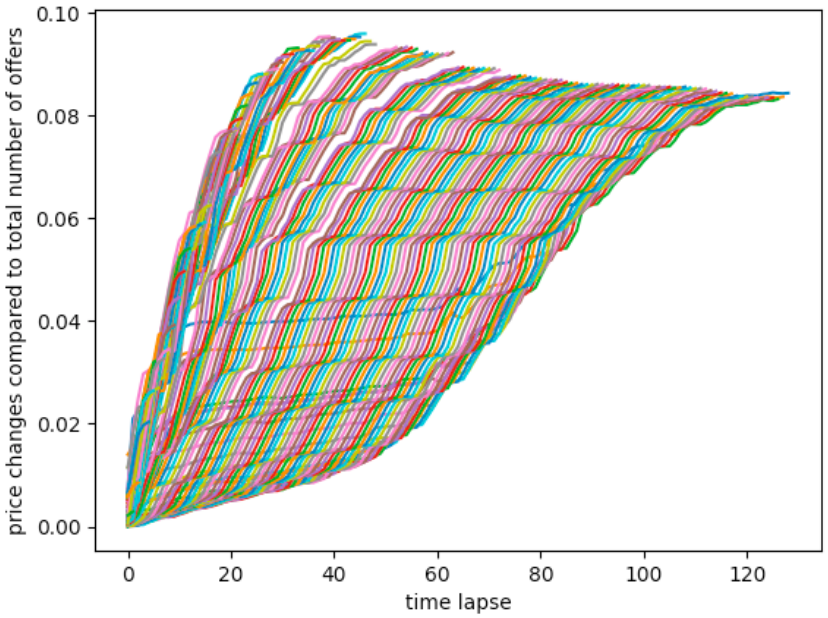


Figure 20: Price changes as the share of total listings (%)
Source: Zoopla and authors' calculation

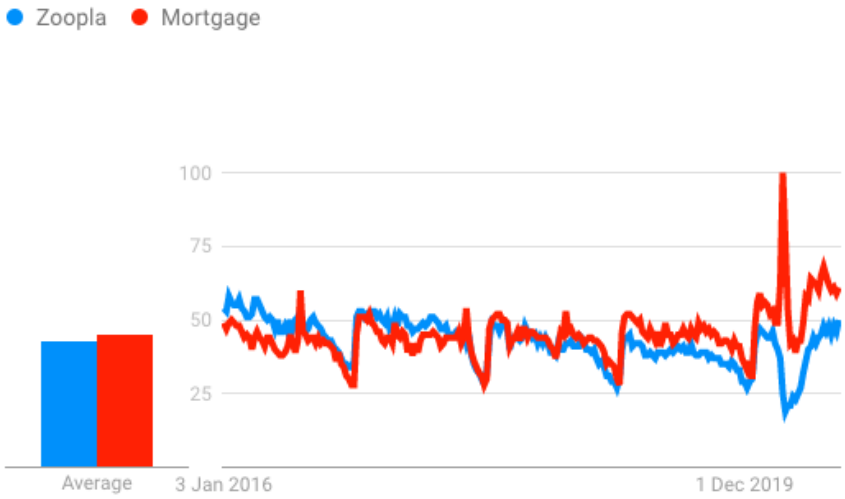


C Web traffic statistics during the Covid-19 crisis

C.1 Google Trends

Google Trends data show a strong effect of Covid-19 on real-estate-related Google searches. Searches for "Zoopla" was plummeting during the first lockdown while those related to "mortgage" peaked – potentially suggesting that some households had envisaged renegotiation. This peculiar Covid-19 period stands out with a clear decorrelation between these two terms, a unique feature in the last 5 years – see **Figure 21** .

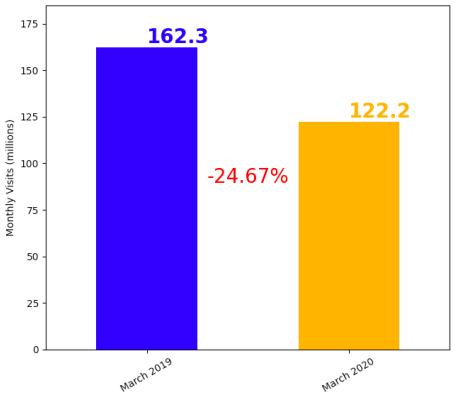
Figure 21: Google Trends for "mortgage" and "Zoopla" over the last 5 years
Source: Google Trends and authors' calculation



C.2 Similar Web

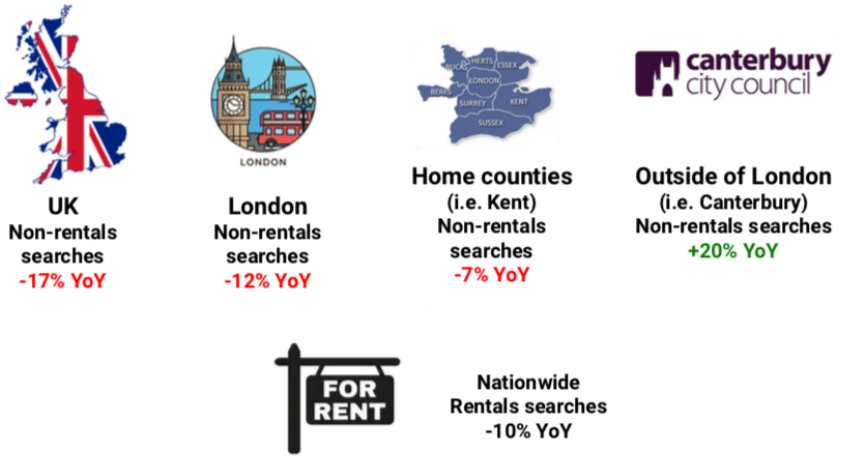
SimilarWeb, which provides audience data for websites, gives a complementary and congruent view. The loss of traffic during the Covid-19 period is visible in **Figure 22** as the real-estate industry in the UK lost a quarter of its traffic in March.

Figure 22: Real-estate Industry - Total monthly web traffic



SimilarWeb also shows geographical heterogeneities in this pattern, confirming to some extent the disparities observed in **section 4** regarding the evolution of the mean listing price. As shown in **Figure 23**, web traffic related to property sales in London (the area most affected by the virus) dropped by 12% in March 2020 compared to 2019, while the decline was only 7% for home counties (region immediately outside London). Further away from London, in the city of Canterbury, the traffic related to property sales instead grew by 20%. Also, the traffic for rentals proved more resilient than for non-rentals as the former declined only by 10% year-on-year *vs.* a 17% drop for the latter.

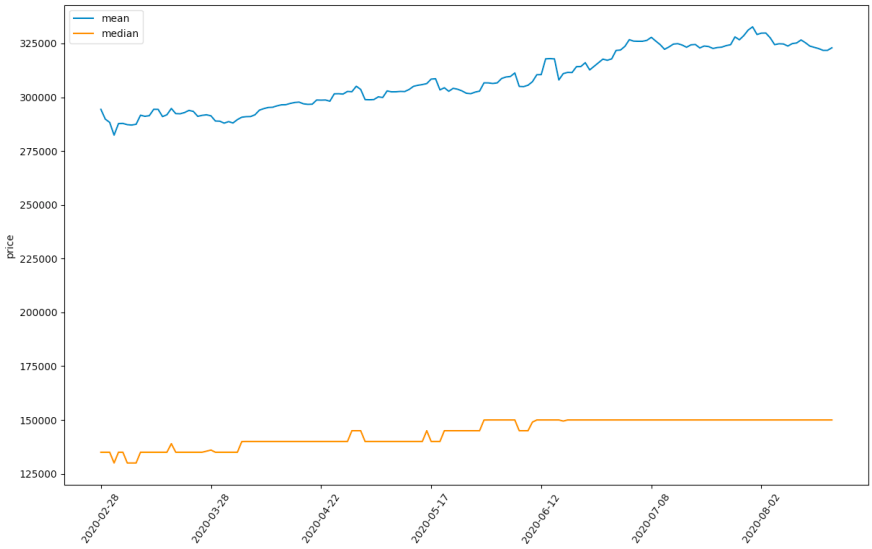
Figure 23: SimilarWeb summary



D Land prices

Part of the scraped data is also labelled as lands and farms. The evolution of land prices in levels is shown in **Figure 24**. It should however be noted that web-scraped listings for lands include agricultural lands and leisure lands. In future research, identifying building plots could inform on the evolution of this factor in the construction sector.

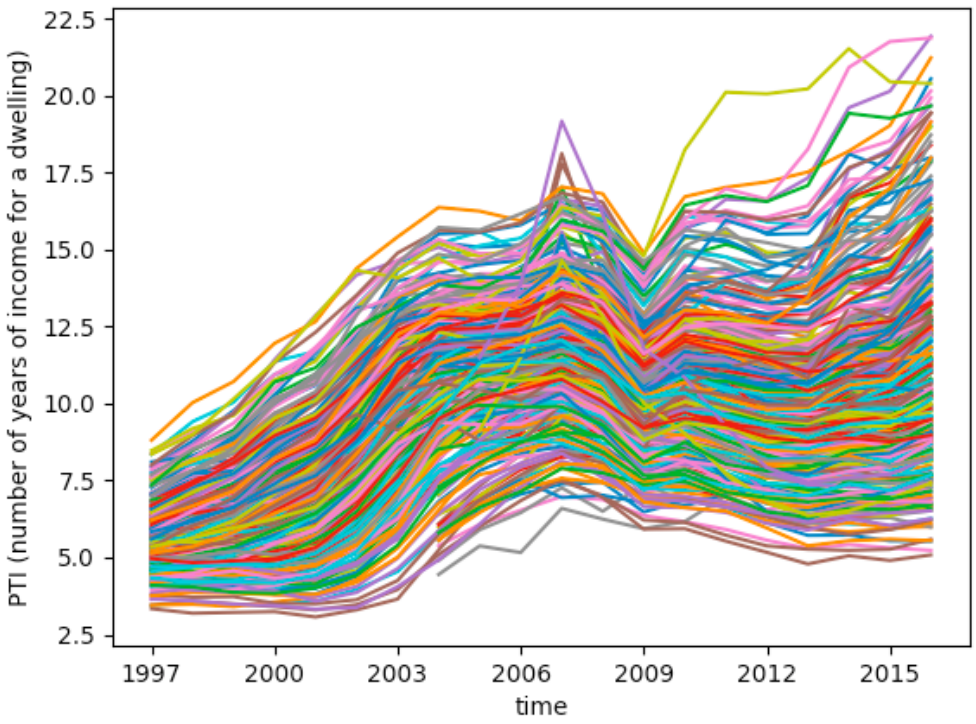
Figure 24: Land prices
Source: Zoopla and authors' calculation



E Price-to-income ratio and purchasing-power-capacity

Another indicator is the price-to-income ratio (André et al., 2014) which measures the number of years of median income necessary to buy a dwelling. It measures the affordability of the housing (the lower, the more affordable) and can be used to analyse the long-term equilibrium of the housing market (Goodman, 1988). Our main contribution to the existing literature is the capacity to compute those ratios at very granular level – building on findings such as those of Gan and Hill, 2009 that a finer picture is more relevant than the national median price-to-income ratio. This heterogeneity can be verified with **Figure 25** built on historical prices and incomes from the ONS.

Figure 25: Price-to-income ratio per locality
Source: ONS and authors' calculation



However the aforementioned price-to-income ratio does not take into account interest rates – as well as other factors that can impact housing prices (see Case and Schiller, 1990). To take into account mortgage rates, we propose the purchasing-power-capacity (PPC) based on the following formula:

D: median duration

ER: maximum effort rate

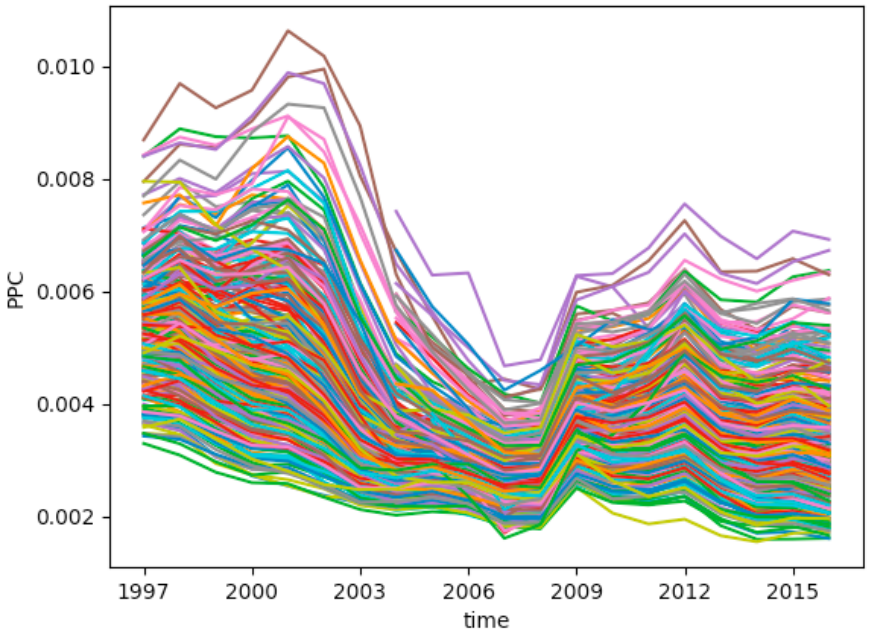
MR: mortgage rate

ITP: income-to-price ratio

$$PPC = \frac{ER}{MR} * ITP * \left[1 - \frac{1}{(1+MR)^D}\right] \tag{3}$$

Using historical data from the ONS about income and housing price, and mortgage rates from the Bank of England (or the Building Society Association if not available), it is possible to derive a PPC index for each of the 400 localities in the UK. To compute the index we assume a maximum effort (ER) rate of 33% – which is widely taken as the upper limit for the effort rate in France – and a median duration (D) of 15 years. Results are plotted in **Figure 26**.

Figure 26: PPC per locality
Source: ONS and authors' calculation



We then explore the signalling power of this new indicator. Considering data on PPC and on real-estate crisis per locality (defined as a strong and rapid correction of prices), we test whether there exists a threshold maximizing the signalling power of the PPC (i.e. a threshold value above which the PPC signals a potential risk of housing crisis

entailing an adjustment of housing prices). More formally, the signalling power is defined as follows:

TP: true positive, i.e. PPC $>$ threshold and a housing crisis actually occurs

FP: false positive, i.e. PPC $>$ threshold **but** no housing crisis actually occurs (incorrect prediction of a crisis)

TN: true negative, i.e. PPC $<$ threshold and no housing crisis actually occurs

TP: false negative, i.e. PPC $<$ threshold **but** a housing crisis actually occurs (incorrect prediction of no crisis)

$$SignallingPower = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1 \tag{4}$$

The signalling power (also named "informedness" or "bookmaker informedness") can reach 0.48 for an optimal threshold near 0,003.[22] It should be in particular noted that the signalling power of the PPC is higher than for the price-to-income ratio for which it stands only around 0.34. Comparing this threshold with actual values in **Figure 27**, it appears that many localities in the UK are still above this threshold.

[22] 15,000 potential thresholds have been tested.

Figure 27: PPC per locality compared with optimal threshold
Source: ONS and authors' calculation