WORKING PAPER

DANMARKS NATIONALBANK

11 AUGUST 2022 — NO. 188

Segmentation of the Housing Market with Internet Data: Evidence from Denmark

Jakob Feveile Adolfsen
DANMARKS NATIONALBANK

Bjarke Mørch Mønsted
DANMARKS NATIONALBANK

Adrian Michael Bay Schmith
DANMARKS NATIONALBANK

Alessandro Tang-Andersen Martinello
alem@nationalbanken.dk
DANMARKS NATIONALBANK

Simon Gudiksen
BOLIGSIDEN A/S

Kasper Fredløv Sonberg
kfs@boligsiden.dk
BOLIGSIDEN A/S

Segmentation of the Housing Market with Internet Data: Evidence from Denmark

Abstract
In this paper, we introduce a novel tool for housing market analysis developed on the basis of online listings data from the largest real estate listing site in Denmark. The tool uses a combination of machine learning techniques to provide a data-driven segmentation of the housing market into meaningful submarkets that differ from administrative classifications. We demonstrate how the tool can support monitoring and research of underlying housing market developments in Denmark.

Resume
I dette papir introduceres et nyt værktøj til boligmarkedsanalyse, der er udviklet på baggrund af salgsopstillinger fra den største hjemmeside for boligannoncering i Danmark. Værktøjet bruger en kombination af forskellige machine learning-teknikker til at foretage en datadrevet inddeling af boligmarkedet i relevante undermarkeder frem for forskellige administrative inddelinger, som fx landsdele. I papiret vises, hvordan værktøjet kan understøtte overvågning og analyse af underliggende udviklinger på boligmarkedet i Danmark.

Key words
Housing market, internet data, neural network

Acknowledgements
The authors wish to thank Birgit Daetz, Thomas Harr, Peter Storgaard, Lars Mayland Nielsen, Birgitte Vølund Buchholst, Denis Gorea, Martin Nygaard Jørgensen, Niels Framroze Møller, Paul Lassenius Kramp, Simon Juul Hviid, Thais Lærkholm Jensen, and seminar participants at Danmarks Nationalbank for their valuable comments.

The authors alone are responsible for any remaining errors.

SEGMENTATION OF THE HOUSING MARKET WITH INTERNET DATA: EVIDENCE FROM DENMARK

WORKING PAPER

Jakob Feveile Adolfsen
Macroeconomic Analysis
Economics and Monetary Policy
Danmarks Nationalbank

Simon Gudiksen
Boligsiden A/S

Alessandro Tang-Andersen Martinello*
Data Analytics and Science
Financial Statistics
Danmarks Nationalbank

Bjarke Mørch Mønsted
Data Analytics and Science
Financial Statistics
Danmarks Nationalbank

Adrian Michael Bay Schmith
Macroeconomic Analysis
Economics and Monetary Policy
Danmarks Nationalbank

Kasper Fredløv Sonberg†
Boligsiden A/S

July 5, 2022

ABSTRACT

Examining submarkets in the housing market can be useful to identify underlying developments that might be masked in aggregate data. In this paper, we introduce a novel tool for housing market analysis developed on online listings data from the largest real estate listing site in Denmark. The tool uses a combination of advanced machine learning techniques to provide a data-driven segmentation of the aggregate housing market into submarkets that are based on characteristics important for price. This approach differs from traditional housing market research where segmentation is typically based on administrative or other pre-defined classifications, and offers new insights beyond that of publicly available information. We demonstrate how the tool can support monitoring and research of underlying housing market developments through two case studies of the urban housing market in Denmark.

Keywords Housing market · Internet data · Neural network

*Langelinie Allé 47, 2100 København Ø, E-mail: alem@nationalbanken.dk

†Kronprinsensgade 6, 2., 1114 København K, E-mail: kfs@boligsiden.dk

The viewpoints and conclusions stated are the responsibility of the individual contributors, and do not necessarily reflect the views of Danmarks Nationalbank nor Boligsiden A/S.

The authors wish to thank Birgit Daetz, Birgitte Vølund Buchholst, Denis Gorea, Martin Nygaard Jørgensen, Niels Framroze Møller, Paul Lassenius Kramp, Simon Juul Hviid, Thais Lærkholm Jensen, and seminar participants at Danmarks Nationalbank for their valuable comments.

1 Introduction

The housing market is crucial for most economies. In Denmark, the majority of household assets are linked to the housing market, either directly through home ownership or indirectly through pension funds where savings are invested in covered mortgage bonds.[1] Monitoring the housing market is therefore important for assessing the general macroeconomic state of the economy. Because of the complexity of this market, monitoring segments (or submarkets, i.e. groups of housing units that are exposed to similar supply and demand shocks) separately can be useful to identify developments that might be masked at the aggregate level. In housing market analysis, these submarkets are often defined by administrative classifications, such as postal codes and types of housing. However, these classifications risk on the one hand ignoring important dimensions for the development of house prices, and on the other being too granular to be informative.

In this paper, we document a novel tool to segment the housing market based on the physical characteristics of a housing unit that are important for prices. This framework can be used to construct, explore, and monitor any number of housing market segments, depending on the desired granularity. Illustratively, we use it to analyze micro-level developments in the urban housing market of Denmark. This explorative exercise shows the perspective in using internet data for detailed housing market analysis.

The segmentation tool is trained on a unique dataset covering housing sales between 2009 and 2019 provided by the largest online real estate listing service in Denmark, Boligsiden.dk. The sample period reflects how long Boligsiden has been operating, and hence covers the entire availability of the data up until 2019. The data contains housing characteristics and market outcomes (i.e. price, time on market, and number of clicks on the listing) of half a million housing units listed on Boligsiden.dk. Based on these characteristics, the tool assigns housing units to specific submarkets based on their similarity. However, the performance of segmentation algorithms in identifying meaningful subsets of the data decreases in high-dimensional settings such as this, where a single housing unit is characterized by a large number of observable characteristics.

We therefore construct the segmentation algorithm in two steps. In the first step, we reduce a high-dimensional vector of housing characteristics into a low-dimensional space. This mapping is crucial for ensuring a good and stable segmentation performance, and consists in a combination of two approaches. We begin by training a simple neural network which uses prices as its target. The representation layer of this network informs us on the combination of characteristics that best explains house prices out of sample. We then structure these combinations into a two-dimensional space using Uniform Manifold Approximation and Projection, a novel algorithm for dimensionality embedding that outperforms standard alternatives (McInnes et al., 2018).

In the second step, we group housing units, that are now represented in a two-dimensional space, by using a hierarchical density-based algorithm (McInnes et al., 2017). This grouping allows us to assign any given housing unit to a specific market segment, and thereby analyze the role of submarkets in aggregate price developments.

This approach provides two main advantages. First, it allows us to select the relevant combinations of housing characteristics that are most predictive of price developments. This provides a weighting of the many characteristics of a house when grouping them according to their similarity. As a consequence, we are able to potentially account for many more characteristics than geography and housing type alone. Second, it allows us to control the desired granularity of the submarkets. In Denmark, there are over 900 postal codes, which interacted by even a rough classification of housing types will return thousands of housing segments. Our approach allows us to group houses in 157 segments, 40 of which represent over half of the housing stock in the country.

Our analysis confirms that geographical location is very important for prices. Yet, our approach captures a higher share of the variance in housing prices with just a few geographic categories compared to a large number of municipality dummies. Moreover, our tool identifies submarkets in the housing market that exhibit different price trends across and within administrative boundaries. Focusing on urban areas, we provide two specific insights on the Danish housing market: We show that while apartments in different urban areas with different characteristics (e.g. size and year of construction) exhibit similar price developments over time, a similar pattern does not apply for single-family houses. In the urban market for single-family houses, price developments are more heterogeneous. Specifically, prices on high-end single-family housing (defined as distinctively larger and older) in the Greater Copenhagen area have been accelerating relative to prices on other single-family houses in Greater Copenhagen.

Being able to identify these submarkets enables users to easily follow heterogeneity in price developments which would not be possible on the back of traditional housing data. To illustrate the potential usefulness of such findings, we use an empirical framework inspired by Dam et al. (2011) to show that prices on high-end houses are more sensitive to the

[1] For an overview of life-cycle wealth statistics for Denmark, see e.g. Statistics Denmark, NYT december 2020.

costs of owning a house (e.g. interest rates and taxes) than to income relative to other prices on single-family homes in the area.

Our findings are based on data from 2009 to 2019 and illustrate possible applications of our tool. However, since internet data can be updated on a high frequency and without a time lag, our tool has an onward potential for supporting real-time surveillance of the housing market.

The paper contributes to the economic literature on housing market dynamics. In the hedonic housing market literature, the price of a housing unit is modeled on different housing characteristics, including size, number of bathrooms, heating system, geographic location etc., which forms an implicit hedonic price (Rosen, 1974). Geographical location is typically recognized as an important driver of differences in market dynamics, especially within urban areas where different neighborhoods tend to contain specific amenities and types of housing units with a relatively fixed supply of properties (Bayer et al., 2016; Bourassa et al., 2003; Goodman, 1978, 1981; Goodman and Thibodeau, 1998; Li and Brown, 1980; Straszheim, 1975; Schnare and Struyk, 1976). In our paper, we build on determinants of house prices previously found in the literature to inform our input to the segmentation algorithm.

In Danish housing market research, submarkets are typically based on geography. Denmark is a small country with only a few large business activity centers. A typical approach to housing market analysis in Denmark has therefore been to focus on the country divided into Copenhagen (i.e. the capital of Denmark) or Copenhagen and large cities versus the rest of the country (Ho, 2016; Dam et al., 2014; Hviid, 2017a; Danmarks Nationalbank, 2019a,b, 2018; Finansministeriet, 2019; Økonomi- og Indenrigsministeriet, 2018; De Økonomiske Råd, 2019). Another approach has been to divide the country into geographical groups based on administratively defined borders (Heebøll, 2014). Hviid (2017b) divides Denmark into 13 geographical groups based on country parts defined by Statistics Denmark and builds a regional housing market model for Denmark.

We also contribute to the literature by exploiting data from the internet to explore housing market dynamics. Several papers have used internet data and machine learning techniques to obtain new insights on housing market dynamics in different countries and cities (Piazzesi et al., 2020; van Dijk and Francke, 2015; Loberto et al., 2018; Rae and Sener, 2016). We are the first to use machine learning techniques on high-dimensional internet data to segment the Danish housing market.

The paper is structured as follows. Section 2 provides a brief description of the data from Boligsiden.dk. In section 3, we outline the segmentation algorithm. Section 4 describes the outcome of the segmentation algorithm and explores how the algorithm can be used for housing market surveillance.

2 Data

We exploit a micro-level data set on housing sales between 2009 and 2019 from the country's largest online real estate listing service, Boligsiden.dk. The country-wide listing service is owned by real estate agents in Denmark, who pass on information to Boligsiden.dk on all housing transactions and attempted sales. Boligsiden.dk has information on around 80,000 transactions and 115,000 new listings per year depending on the business cycle. We enrich the data from Boligsiden.dk with information from the Central Register of Buildings and Dwellings (BBR), which is a public register established in 1976 with information on all buildings in Denmark. In the following subsections, we will briefly discuss data coverage and quality as well as features of the data.

Data coverage

Because of the direct link between Boligsiden.dk and real estate agents, Boligsiden.dk has the most updated and frequent information on the Danish housing market, covering around 85 per cent of Danish home ownerships, according to their estimates.[2] Moreover, Denmark is a small country with low competition on the market for online real estate listings relative to other countries.[3] Denmark is a highly digital country with 93 per cent of Danish families having access to the internet compared to an EU average of 87 per cent (Danmarks Statistik, 2018) and the largest reported regular use of internet services in the EU (European Commission, 2019).

[2]The market for co-operative housing is not sufficiently covered in our dataset. Therefore, we leave out co-operative housing in this paper. Summerhouses, too, are left out of the analysis in this paper. Instead, focus is solely on owner-occupied villas, terraced houses and apartments.

[3]The only competitor to Boligsiden.dk of significant size is Boliga.dk.

Data quality

Another advantage of our dataset is data quality. In the data collection process, Boligsiden.dk performs thorough validation of the reported data to ensure that both information and statistics on Boligsiden.dk are correct. This process provides an advantage compared to other data sources from internet providers. An example can be found in a paper by Loberto et al. (2018) who investigate dynamics on the Italian housing market based on data from Immobiliare.it. The authors of this paper use machine learning techniques to deal with the problem of multiple postings appearing for the same listing - a technique used by agents to maximize exposure of their listing on the website. We do not have to deal with such a problem due to the data validation process at Boligsiden.dk. Therefore, we only use very simple measures to clean data for what seems to be reporting errors that might not have been caught in the initial validation process, or very extreme outlier listings which are not the focus of this paper.[4] Our data provides a good representation of the entire Danish market for owner-occupied housing.

Data features

Our data covers individual housing units for sale in Denmark and contains around 500,000 transactions of listings posted on Boligsiden.dk. This corresponds to around 400,000 unique single-family houses and apartments, as some of the units have been listed for sale more than once within the period. We can track each housing unit by a unique identifier.

A unique feature of this data is the combination of information on housing supply and information on housing demand. As a proxy for demand, we use the number of clicks re-directing users to the website of the real estate agent, and the supply side is represented by each listing. The aggregate number of clicks from Boligsiden.dk is approximately 150 millions per year. In this paper, we only use this variable to test whether our geographic representation of the sample can explain differences in observable features related to housing demand. Nonetheless, the ability of following clicks opens for new research possibilities, which we do not pursue in this paper.

For each listing, we also observe a list of physical features (such as size, number of rooms, energy label, etc.) and outcomes (i.e. price and time on market).[5] Each listing is accompanied by GIS coordinates. It gives us the exact location of the housing unit and the opportunity of calculating distances that could be of importance to demand and price, e.g. distance to coastline, distance to railroad, etc. Furthermore, we use the location to calculate noise pollution from larger roads based on sound data from the Danish Environmental Protection Agency. Overall, it gives us a good indication of how attractive the location of the housing unit is.

This wide range of information helps us track the main determinants of market outcomes such as sales price and time on the market, including the interactions of both determinants and market outcomes. Figure 1 shows some examples of the relationships between square meter sales price, standardized to have a mean of zero and a standard deviation of one within each calendar year, and four other variables in the dataset.

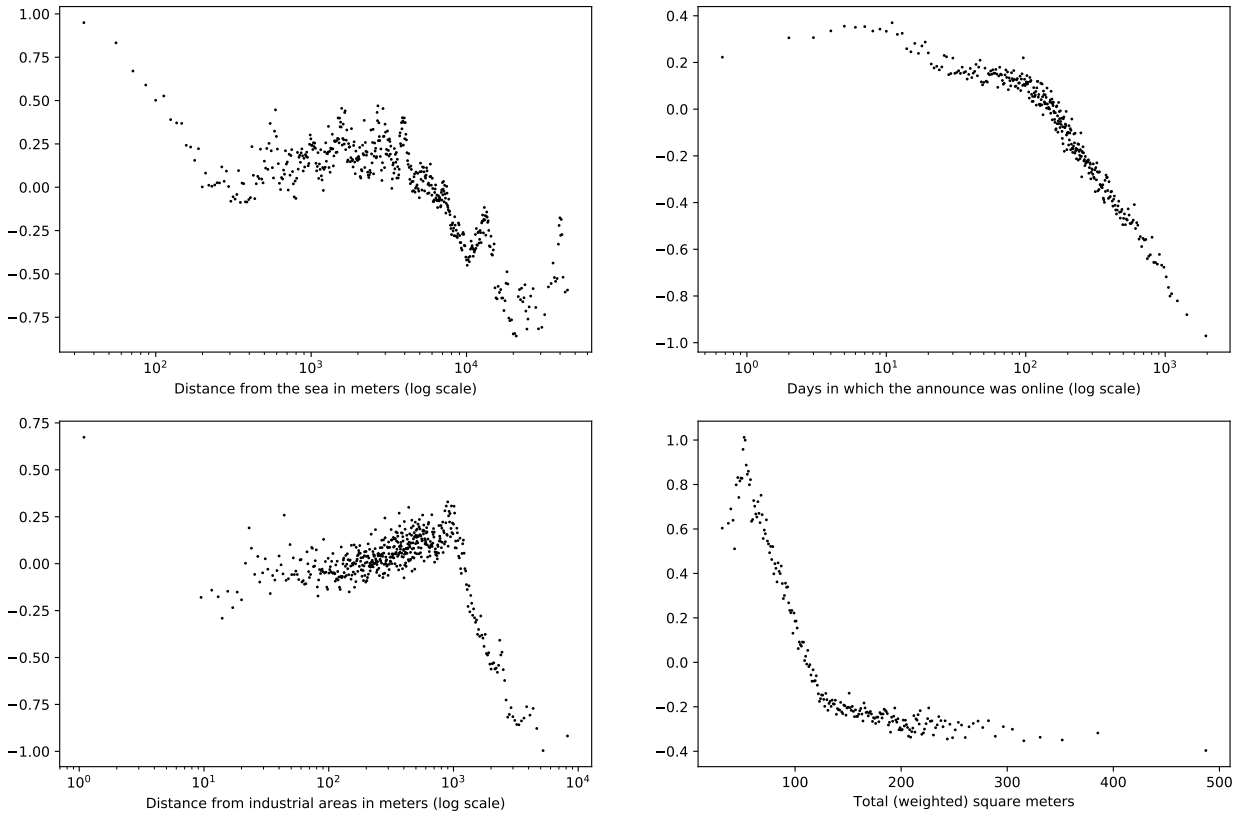
The panels on the left of the figure show the relationship between standardized square meter price and distance to the coast and industrial areas. While initial relationships are as expected, with prices initially decreasing as the distance from sea increases, and prices increasing as the distance from an industrial area increases, both relationships are highly non-linear. Also, the sign of the relationship inverts after approximately 400 meters from the sea and 1 kilometer from an industrial area. These non-linearities are likely caused by interactions with other observable characteristics such as urban density. Nonetheless, these figures stress how modeling of market outcomes in the real estate market requires models that can handle non-linear relationships and interactions between a wide range of features in a flexible way.

The top and bottom right panels show the relationship of standardized square meter price with time on market and total square meters, respectively. Both these correlations are well-known to market operators. Longer time on market typically indicates a less favorable sales price, and the square meter price typically peaks at approximately 60 square meters. Our data allows us to observe and learn these patterns directly and incorporate them in our segmentation approach.

[4] See appendix A for data-cleaning procedures.

[5] The list of variables can be found in appendix B.

Figure 1
Correlation between standardized square meter price and selected features in the dataset.



NOTE: Each panel plots the correlation of a selected variable in the dataset with a housing unit's square meter sale price, standardized to have a mean of zero and a standard deviation of one within each calendar year. Each dot represents a centile of the data, ordered according to the variable plotted in the horizontal axis. For each such centile, the figure plots the means of standardized square meter price and the selected feature in a scatterplot.

3 Segmentation approach: constructing clusters

Our goal is to segment the Danish housing market into groups that represent separate submarkets based on their observable characteristics. We do that by generating clusters of housing units, where characteristics of housing units are very similar within each cluster. We are not trying to directly predict the market outcomes of these clusters such as price or time on market. It is therefore an unsupervised machine learning problem that cannot be tackled by supervised learning approaches (e.g. a regression analysis). Unsupervised machine learning problems typically leads to the application of clustering algorithms.

While clustering algorithms are widespread, this specific application presents two main challenges. First, our data are high-dimensional. Housing units and submarkets are characterized by a large number of features, and most clustering algorithms perform poorly in high-dimensional settings without initial data processing. They converge on local maxima resulting in mixed clusters providing little additional information with respect to random groupings. Second, the problem of assigning housing units with a high degree of internal similarity into different clusters raises the problem of how to define and quantify such similarity in a multi-dimensional setting. Take the example of a villa with a 150 square meter liveable area. A guiding principle is required to determine whether such a housing unit is most similar to a villa with a 200 square meter livable area located a few hundred meters away, or to a villa of exactly the same size, but located at the other end of the country. This question is further complicated as the number of relevant dimensions increase.

To solve these issues, we combine several machine learning techniques to two ends. First, they reduce the dimensionality of the raw data, strongly improving the performance of the final clustering algorithms. Second, they help us determine a guiding principle for weighting 'raw' characteristics when determining similarity between housing units. We choose to

assign the highest weights to combinations of characteristics that are most predictive for final sales prices. Consistent with the literature, we allow these characteristics to enter the model non-linearly (Zietz et al., 2008; Malpezzi et al., 1980).

Figure 2 shows a representation of the full pipeline of machine learning steps, which we will call the segmentation algorithm, as it generates clusters from the initial, unprocessed data. The pipeline is roughly divided into the following three steps: feature engineering, dimensionality reduction, and density-based clustering. This section describes each of these three steps, with focus on the embedding and dimensionality reduction steps described in appendix B.

Feature engineering

We begin by preprocessing the information contained in the raw data and transforming it into a set of features, which are digestible for the rest of the pipeline. The data includes both information that describes the housing unit itself, such as housing type (apartment, villa, etc.), number of rooms, square meters, etc., and the location of the housing unit expressed in geographic coordinates.

We use geographic coordinates in two ways. First, on the basis of the coordinates, we calculate distance measures of a housing unit to prominent landscape features, such as distance to the coast or to freeways. We treat these features as additional physical characteristics of a housing unit. Second, we map the coordinates into five geographical areas that are characterized by markedly different price levels across the country. These price areas represent a compact mapping of coordinate data into five separate features.

We construct price areas by ranking housing units in five equal-sized groups based on their square meter price normalized within each sales year. The choice of five groups is arbitrary. Choosing more groups can improve precision at the cost of model complexity and risk of overfitting. While other choices are certainly possible, we find that five groups represents price areas flexibly enough for our purposes. We then train an Xgboost classifier (Chen and Guestrin, 2016) to predict the group to which a specific housing unit most likely belongs given its coordinates. The output of the classifier is a vector of five probabilities for each housing unit, which represents a flexible mapping of location independent of administrative boundaries.

A mapping of these price areas appears in figure 3. While each price area represents 20 per cent of the housing units in the sample, they cover very different areas on the country map. The most expensive price areas are concentrated in the urban areas of Denmark's two largest cities, while price areas 2 and 1 cover small towns and rural areas, respectively. The medium price area is located in large cities, excluding the two largest cities, or the remote suburbs of the two largest cities. Intuition on the location of the price areas can be drawn from the so-called ripple effect suggested in Meen (1999) and Oikarinen (2006), and examined in Denmark by Hviid (2017b). According to this effect, developments in submarkets close to a city center are likely to transmit to suburbs with longer distances to the city center or other cities of smaller size. This effect is likely to transmit in ripples until affecting the most rural parts of the housing market.

The resolution of these mappings is very high, with price areas running along waterways, major roads and coastlines. While this high resolution might give the impression of overlapping maps when plotting them over the whole country, it flexibly allows for multiple price areas to coexist within the same administrative boundaries when zooming in on particular areas. Panel (f) of the figure shows how several price areas are distributed within the municipality of Aarhus, with adjacent neighborhoods belonging to different price areas.

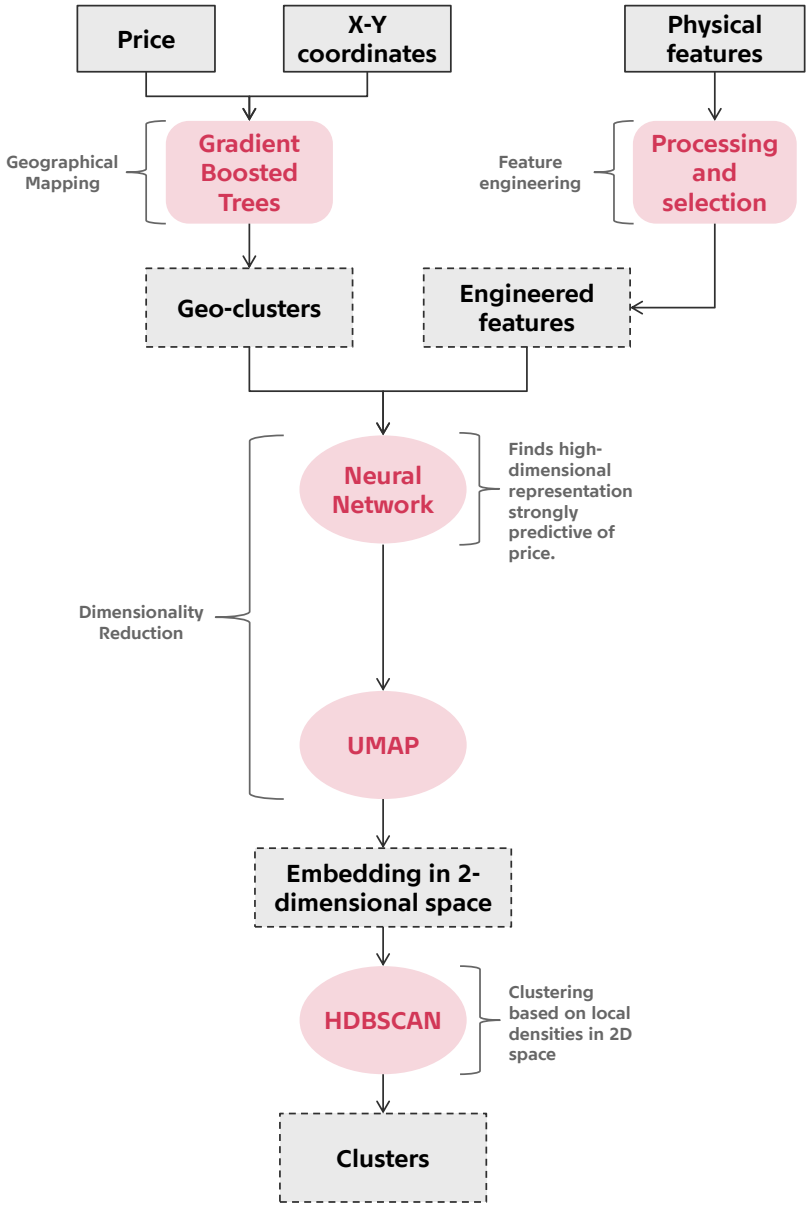
While in our algorithm we include price areas as vectors of probabilities, thereby allowing for a soft transitioning from one area to another, even these sharp boundaries capture an important dimension of the development in housing prices that improves our understanding of geographical price dispersion relative to arbitrary administrative boundaries alone. Table 1 shows the predictive power for house prices, search intensity and time on market of these five price areas, compared to municipalities and other features entering our clustering algorithm. We measure the predictive power of each set of features by computing the out-of-sample R^2 of an unregularized OLS regression taking these features as input. [6]

We conduct these regressions out-of-sample to alleviate reflection problems. The model producing price areas is itself a function of coordinates and house prices. By training and testing these models in different samples, we ensure that the house sales for which we'd like to predict prices using price areas are not the same sales informing the price areas model. Intuitively, it would be like using past sale prices in the neigbouring area to predict current sale prices. While we expect a strong correlation, the relationship is not mechanical.

A model based on our five price areas alone is able to explain 82 per cent of the out-of-sample variation in square meter prices in 2019. In comparison, an alternative model using 98 municipality dummies is only able to explain 76 per

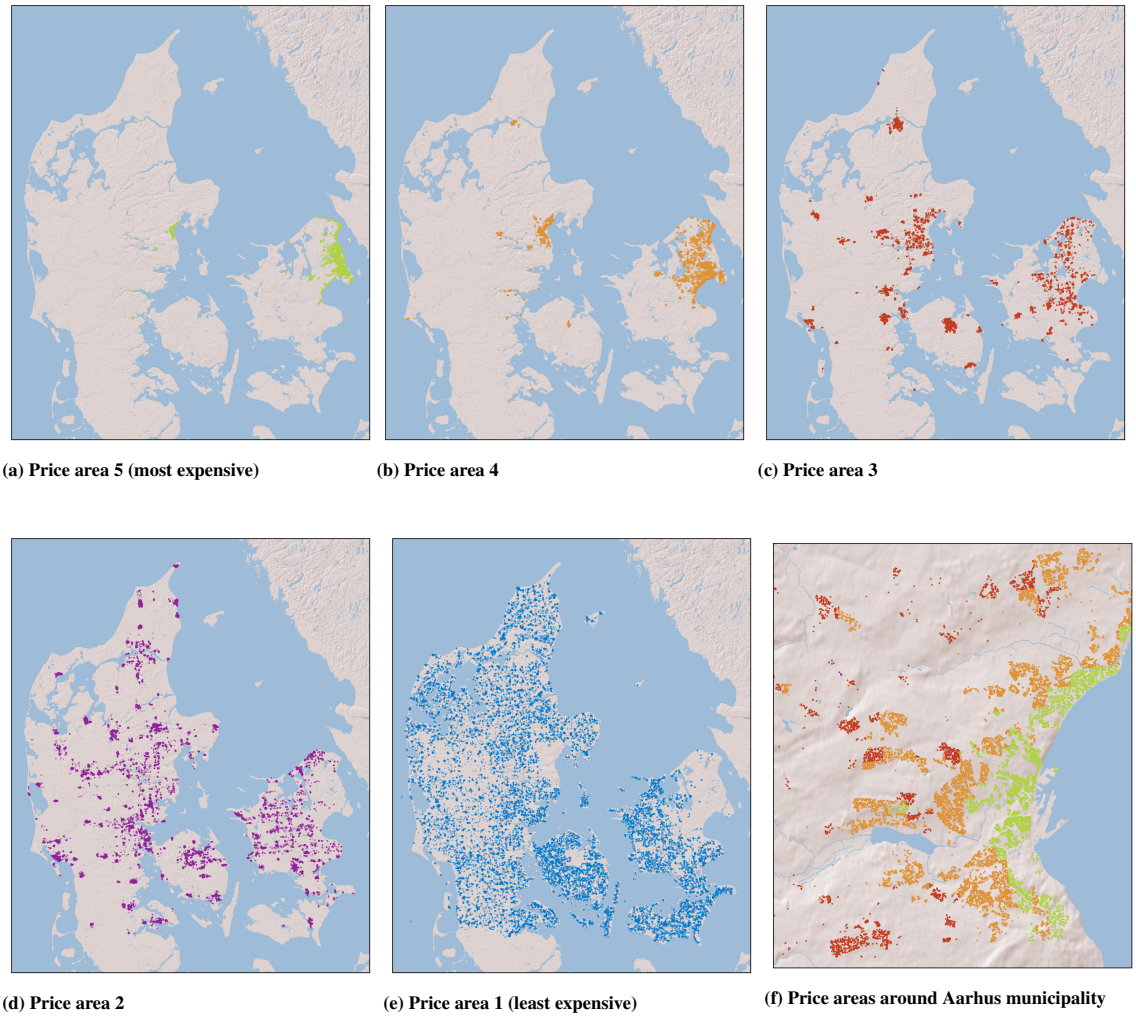
[6] Advanced machine learning models that allow for more complex relationships in the data show the same patterns.

Figure 2
Illustration of the full pipeline that processes housing data and identifies clusters.



NOTE: UMAP = Uniform Manifold Approximation and Projection. HDBSCAN = Hierarchical Density-Based Spatial Clustering of Applications with Noise.

Figure 3
Price areas in Denmark



(a) Price area 5 (most expensive) **(b) Price area 4** **(c) Price area 3**

(d) Price area 2 **(e) Price area 1 (least expensive)** **(f) Price areas around Aarhus municipality**

NOTE: Each dot in the plots represent a sale in our dataset. Dots are color-coded according to the price area they belong to. Panels 3a to 3e show the Danish mainland with housing units belonging to specific price areas plotted on top of the map. The last panel shows a map of Aarhus, the second largest city in Denmark, with housing units belonging to all price areas existing within the municipality and plotted according to their specific color code.

cent of the out-of-sample variation. While this result is expected and not surprising, as we explicitly construct price areas to describe house price developments, this result translates to other dimensions of the housing market, with fewer but flexibly defined price areas being able to explain more variation than arbitrary administrative boundaries. Higher granularity does not necessarily help to describe the housing market, but meaningful geographical aggregations can be constructed by allowing for enough flexibility.

A model with all physical features except geographic information is able to explain only 66 per cent of the out-of-sample variation. These results highlight the importance of location for the determination of house prices, in line with commonly-used research approaches in Denmark and internationally (Hviid, 2017b; Hviid et al., 2016a; Heebøll, 2014; Claeys et al., 2017; Hilber and Vermeulen, 2014; Holly et al., 2010; Huang, 2019; Oikarinen and Engblom, 2014). Nonetheless, other measurable characteristics of a housing unit add predictive power over and beyond that of geography alone, especially for other market outcomes such as search intensity and time on the market. These results support the construction of clusters based on a larger set of features.

Table 1
Out-of-sample variance explained by groups of input

	Normalized sqm price	Clicks per day	Time on market
All information	0.8437	0.1870	0.1105
Price areas and housing type	0.8253	0.1634	0.0749
Only price areas (5)	0.8248	0.1577	0.0699
Municipalities and housing type	0.7625	0.1526	0.0710
Only municipalities (98)	0.7552	0.1510	0.0642
Only regions (11)	0.4131	0.0692	0.0276
Only zipcodes (909)	0.3077	0.0606	0.0210
Only other features	0.6605	0.1272	0.0974

NOTE: We standardize each outcome within each calendar year. Then, for each standardized outcome, we train an unregularized OLS model on a specific set of features with all sales in our sample before 2019. Categorical features are transformed into dummies, with each dummy entering the regression as a separate regressor. We then use these models to predict the standardized outcome in 2019. The table reports the resulting out-of-sample R^2 for each of these calculations on a hold-out dataset.

For constructing clusters, we therefore combine price areas, geographic measure of distances to specific landmarks, features based on physical characteristics of the housing units, and information on their surrounding area.[7] Examples of housing characteristics include size, number of rooms, energy certification, and whether the property has a garage or not. Examples of information about the area include the municipal tax rate and the number of kindergartens relative to the size of the population.

In this set of characteristics, we also include zip codes, as on Boligsiden.dk customers typically search for units within a single zip code. Thus, two housing units located geographically near to each other may exhibit different dynamics simply because they are distributed over different zip codes that appear differently in the search results of the customers. Our approach is nonetheless independent of the features included in the clustering procedure, and alternative feature specifications can be constructed, leading to different cluster constructions. In this paper, we focus on the clustering resulting from including the most complete set of physical characteristics observable in our data.

Dimensionality reduction and embedding

The resulting dataset has a high number of dimensions (variables) with respect to the number of sales we observe. Clustering algorithms typically work best in low-dimensionality settings. Therefore, we reduce the dimensionality of our dataset and condense the information contained in it into a few representative features in three steps. In practice, this part of the pipeline takes in data on each housing unit – that is high-dimensional given all its attributes – and returns coordinates in a two-dimensional embedding space.

First, we represent the features in terms of their principal components to avoid correlations between them. Second, the principal components are fed to a neural network[8], which is trained to predict the logged square meter price. Finally, we embed the neurons of the representation layer into a two-dimensional space using UMAP (Uniform Manifold Approximation and Projection), a novel algorithm developed by McInnes et al. (2018). More details on these steps appear in appendix B.

Hierarchical clustering

We cluster our datapoints in the two-dimensional embedding space through a density-based clustering approach. Some approaches, such as DBSCAN (Ester et al., 1996), work by iteratively combining points that are located within a specified distance ϵ of each other into clusters. Decreasing ϵ thus increases the granularity of the clustering process, and vice versa.

We choose to employ a more recent approach called hierarchical DBSCAN (HDBSCAN) developed by Campello et al. (2013). Here, clusters are computed for a range of values of ϵ, allowing clusters to be selected based on their robustness

[7] We log-transform features with pronounced heavy tails, including e.g. quantities like distance to the nearest railroad, and the area of the housing unit. We also normalize municipality-level features, such as the number of schools in the municipality, with the municipality population. Continuous features are then scaled to the unit interval.

[8] For a concise introduction to neural networks, see e.g. chapter 5 in Bishop (2006).

to the choice of ϵ.[9] In addition, this approach allows us to run HDBSCAN only once and extract cluster outcomes of varying granularity on the fly (McInnes et al., 2017).

We choose a specification that balances the trade-off between cluster granularity and tractability.[10] This specification produces 157 clusters of Danish housing units, which allows for investigations of differences within traditionally defined submarkets (such as e.g. the Greater Copenhagen Area or a given zip code). With respect to classification based on administrative boundaries such as postal codes, this clustering allows us to be more parsimonious in terms of selected categories (there are over 900 postal codes in Denmark).

Algorithm output

Figure 4 provides a general characterization of the outcome of our preferred segmentation specification for the Danish housing market. The outcome contains both large and general clusters that include housing units with different characteristics across large geographical areas as well as small clusters with very specific characteristics.

Clusters typically span across multiple municipalities, although smaller clusters tend to be contained within a single municipality. The median municipality contains five clusters, while the median cluster spans across two municipalities.[11] Sorting clusters by their maximum within-municipality representation, 78 per cent of units of the median cluster is contained in a single municipality. A representation of the overlap between municipalities and clusters appears in appendix A.2.

Half of the housing stock[12] is covered by 40 clusters, corresponding to 25 per cent of all clusters. Most clusters are predominantly tied to a specific geographical area, with 34 per cent of all clusters being located in the Copenhagen area. The division of clusters is more granular in the large cities than in the countryside. Nonetheless, 43 per cent of clusters represent countryside locations. The number of clusters predominantly consisting of apartments make up 23 per cent of the total number of clusters. In comparison, 22.5 per cent of the Boligsiden.dk housing stock are apartments.

As the algorithm constructs clusters based on what is important for prices, prices within clusters are in general quite similar: 44 per cent of clusters consist of housing units with a very small spread in price levels, while 4 per cent of clusters have a substantial spread.[13]

Based on inspection of features within the cluster, they can be attributed a qualitative description. The largest cluster can e.g. be described as single-family houses in rural areas. It consists of 19 per cent of the entire housing stock, and it is geographically spread across the country, excluding larger cities and the northern part of Zealand. It consists of single-family homes in the size range of 100 square meters, with a median construction year around 1950. The second and third largest clusters are also spread across the country. However, houses in these clusters are newer. The differences between these clusters concern their size and energy rating: While the second largest cluster consists of large houses with high energy ratings (i.e. low energy consumption), the third largest cluster consists predominantly of smaller houses with below-average energy ratings. These types of housing are quite common across the country. The size of the clusters and their geographical spread might reflect that once out of the larger cities and suburban areas, housing preferences might be less location-specific and more related to the amenities of the housing unit.

The fourth-largest cluster is made up by terraced houses in cities, excluding Copenhagen and Aarhus. These terraced houses are constructed around year 1970, and their energy rating is higher than average. In comparison to one of the smaller clusters made up of terraced houses in Copenhagen and Aarhus, the plot sizes of terraced houses outside the big cities are larger (i.e. garden), but the size of the house and the time of construction is very similar.

[9] The HDBSCAN instance ran with a minimum cluster size of 5.

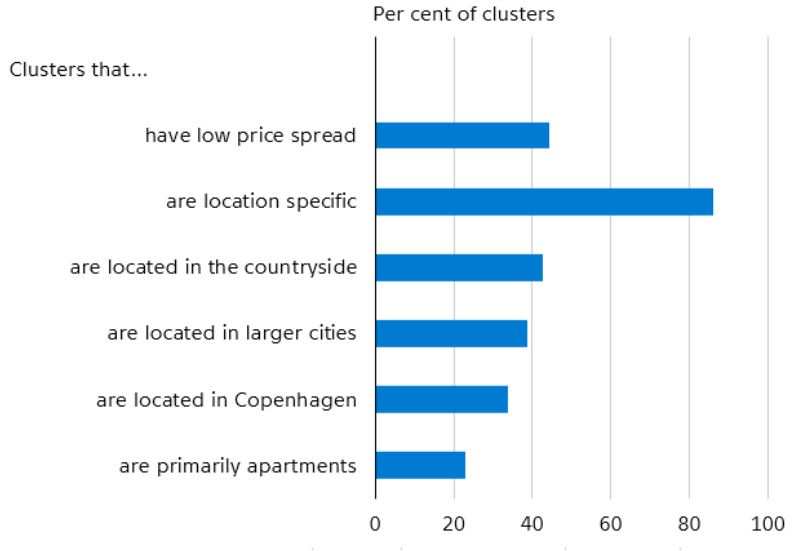
[10] A visualization of our preferred specification and an example of an alternative specification appear in appendix B.

[11] For these calculations, we consider a cluster to be represented in a municipality if at least 1 per cent of its housing units are located in that municipality.

[12] In this subsection, we refer to all houses in the data set as the housing stock. This covers houses that have been sold on Boligsiden over the past decade

[13] A low price spread is defined by a standard deviation below 30 per cent of the average square meter price, while a high price spread is defined by a standard deviation above 80 per cent over the average square metre price.

Figure 4
Characteristics of clusters in the preferred segmentation



NOTE: Large cities are Copenhagen, Aarhus, Odense and Aalborg. A low price spread is defined by a standard deviation below kr. 5,000 per square metre, while a high price spread is defined by a standard deviation above kr. 10,000 per square metre.

4 Economic application of the segmentation

Our segmentation can easily be used to investigate price trends in specific parts of the housing market. It is useful for zooming in on micro-level developments and for highlighting heterogeneous trends that might potentially stand out from the overall assessment of the housing market. As shown in the previous section, our approach identifies submarkets through a data-driven algorithm and on the basis of physical characteristics that impact the house price. It implies that assessments are not restricted by administrative or other pre-defined divisions of the housing market, although they might be influenced by them. To illustrate how the tool may be used, we provide two case studies on how to obtain micro-level insights from the algorithm. Going forward, the potential of the tool is to use it for real-time surveillance of heterogeneity in the housing market by feeding the algorithm with updated data on new housing sales that go beyond our estimation period.

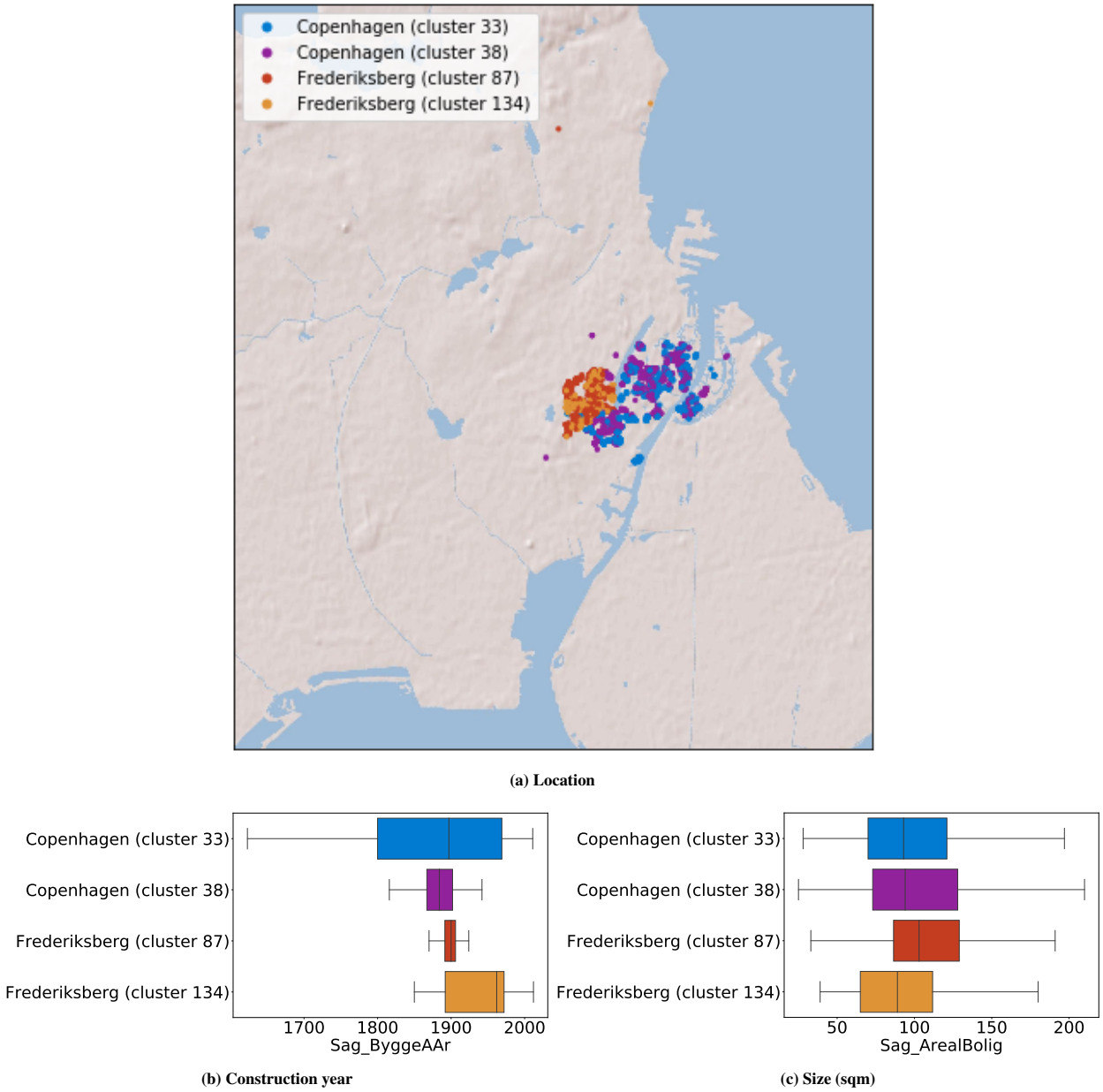
Case study 1: Price divergence on city apartments?

One area of particular interest in Denmark is the apartment market in the two largest cities, Copenhagen and Aarhus. In general, apartment prices in Copenhagen and Aarhus have increased markedly over the last decade. Our algorithm suggests that the apartment market in Copenhagen and Aarhus can be divided into 23 clusters (21 clusters in Copenhagen and 2 clusters in Aarhus).[14] Clusters in Copenhagen are typically linked to specific zip codes.[15] That is not surprising in a large city where different neighborhoods are likely to differ in amenities and popularity (Straszheim, 1975; Goodman, 1978, 1981; Goodman and Thibodeau, 1998; Rae and Sener, 2016; Palm, 1978; Piazzesi et al., 2020). Nonetheless, we also find distinct clusters within the same zip codes.

[14] In order to narrow down descriptive interpretation of clusters, we label clusters to be in Copenhagen if more than 80 per cent of the housing units in the cluster are located in the municipalities that constitute Copenhagen City or Copenhagen and suburbs according to the definitions of country parts by Statistics Denmark. The municipality codes are 101, 147, 155, 185, 165, 151, 153, 157, 159, 161, 163, 167, 169, 183, 173, 175 and 187. A cluster in Aarhus is a cluster where more than 80 per cent of the housing units are located in the municipality of Aarhus (i.e. municipality code 751). The amount of clusters in the larger cities are subject to vary with the size of the 'labelling' parameter. The lower the parameter, the more clusters in each large city. The current parameter of 80 per cent is based on qualitative judgement.

[15] In some areas of Copenhagen, a large amount of zip codes coexist within the same neighbourhoods, e.g. in the neighborhood of Vesterbro. We replace these zip codes with one code that represents the neighborhood.

Figure 5
Selected apartment clusters in inner Copenhagen and Frederiksberg



(a) Location

(b) Construction year

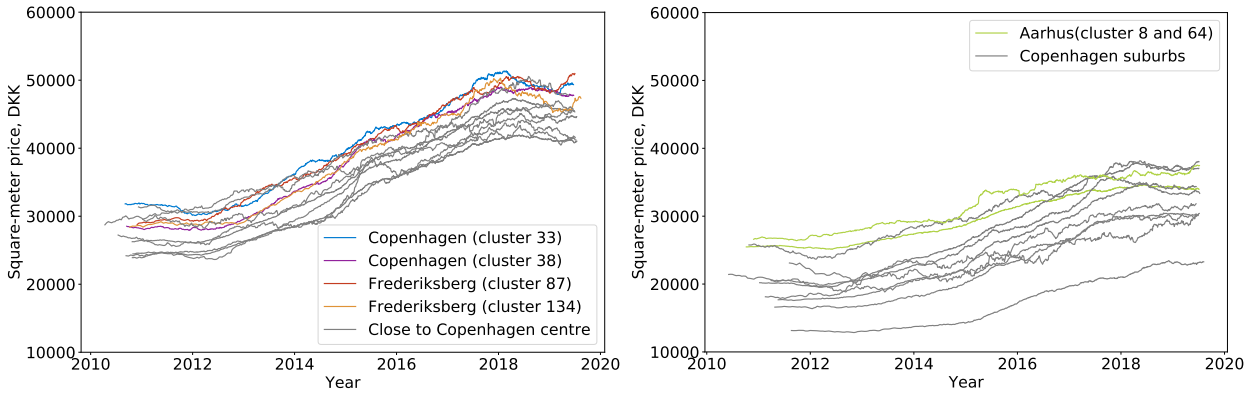
(c) Size (sqm)

NOTE: The upper figure maps a selection of four clusters with apartments located in Copenhagen. Each dot in the plot represents a sale in our dataset. Dots are color-coded according to the cluster they belong to. The lower figures show the distribution of construction year (left-hand figure) and size of the livable area in square meters (right-hand figure) for the same four clusters.

Figure 5 shows two examples from the city center of Copenhagen and Frederiksberg, two popular areas of Copenhagen with a lot of old apartments. Here, the clusters seem to be divided into old apartments of a particularly large size[16] and all other apartments in the area. Figure 6 shows some differences in the square meter price levels across urban apartment clusters, but developments since 2009 have been very similar. In particular, prices of apartments close to

[16]What in Danish would be called a 'herskabslejlighed', referring to apartments of a certain size constructed around 1900 with a specific type of decorative style.

Figure 6
Square meter prices on apartment clusters in Copenhagen and Aarhus



(a) Copenhagen city center

(b) Aarhus city center and Copenhagen suburbs

NOTE: The figure shows the development of average square meter prices of clusters where a minimum of 80 per cent of the housing units are apartments located in municipalities included in Statistics Denmark's definition of Copenhagen and suburbs or Aarhus. In the left-hand figure, apartments are located in zip codes close to the city center of Copenhagen. In the right-hand figure, apartments are located in the suburbs of Copenhagen or Aarhus city center.

the city center of Copenhagen (i.e. Østerbro, Vesterbro, Nørrebro[17], Islands Brygge, and Hellerup) have increased markedly. Prices in clusters located more distant to the Copenhagen city center as well as in Aarhus have followed a similar path in relative terms.

This simple case study illustrates that our algorithm supports a quick and easily accessible overview of developments within particular submarkets that are traditionally masked in aggregate measures. In this case, one might hypothesize that there are distinct apartment markets within Copenhagen that potentially have different price developments. Nevertheless, we conclude that while containing some differences in terms of price level, the development of apartment prices in Copenhagen and Aarhus since 2009 has been relatively uniform across different clusters within the cities. Hence, we do not find empirical evidence for the hypothesis above.

However, we do not have a period of significant price drops in our sample. If prices start to decline, the algorithm will be able to provide input to an analysis of whether different types of apartments exert uniform sensitivity to an overall negative shock to housing demand.

Case study 2: Price divergence in villa segments?

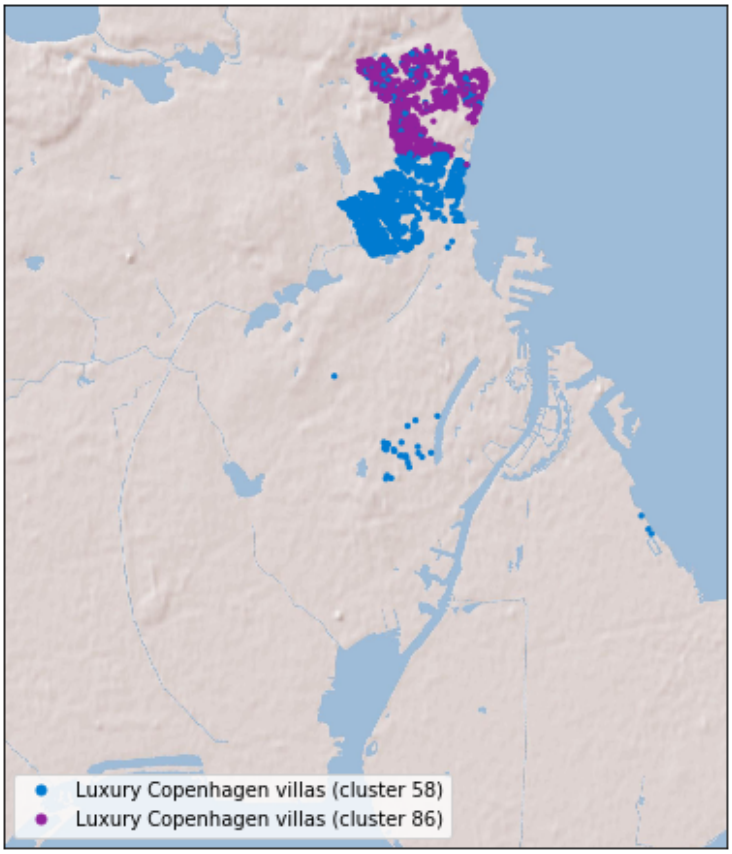
The market for single-family houses in the Greater Copenhagen area (i.e. Copenhagen and suburban municipalities) is also of great interest for at least three reasons. First, single-family houses in Greater Copenhagen alone constitute 13 per cent of the entire housing market in Denmark.[18] Second, the housing market is often segmented into Greater Copenhagen and the rest of Denmark in existing housing market research (Ho, 2016; Dam et al., 2014; Hviid, 2017a; Danmarks Nationalbank, 2019a,b, 2018; Finansministeriet, 2019; Økonomi- og Indenrigsministeriet, 2018; De Økonomiske Råd, 2019). Third, the market for single-family houses in Greater Copenhagen has in particular benefited from an unchanged housing tax base in Denmark since 2001 following a change in the Danish real estate tax system (Hviid and Kramp, 2017; De Økonomiske Råd, 2016). We demonstrate how the output of our algorithm can be used to unfold heterogeneity in the market for single-family houses in Greater Copenhagen. Moreover, we show similar results for Denmark's third largest city, Odense.

The algorithm singles out several clusters consisting of single-family houses in Greater Copenhagen. However, two clusters shown in figure 7 stand out. These clusters contain a higher share of houses that are distinctively larger and older relative to other houses in the Greater Copenhagen area. The clusters are not tied strictly to a certain location within the Greater Copenhagen area: The houses are predominantly in the towns of Hellerup and Gentofte, but are also scattered around inner Frederiksberg and Amager Strandpark.

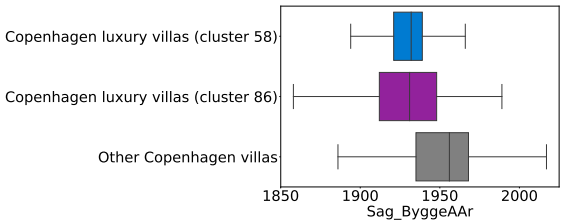
[17] The areas of Østerbro, Vesterbro and Nørrebro are all close to the city center of Copenhagen.

[18] Measured as the share of the total market sales value in 2019.

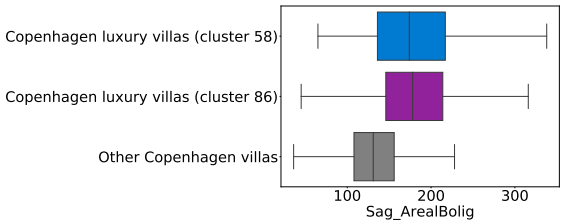
Figure 7
Luxury villas in Greater Copenhagen



(a) Location



(b) Construction year



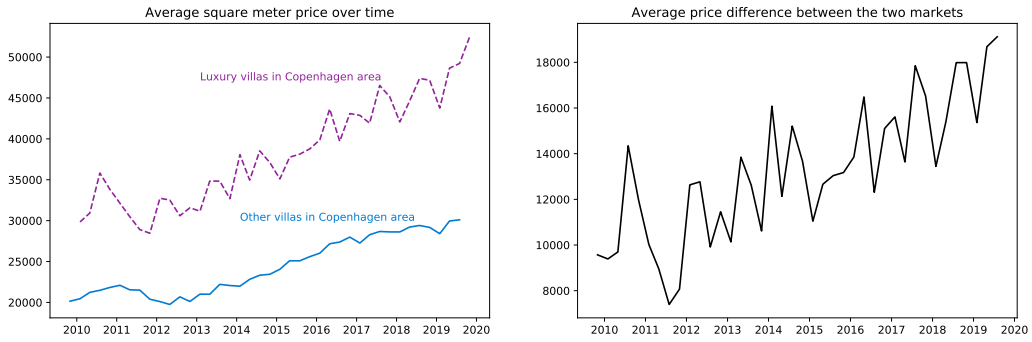
(c) Size (sqm)

NOTE: The upper figure maps a selection of two clusters with single-family houses located in Greater Copenhagen. Each dot in the plot represents a sale in our dataset. Dots are color-coded according to the cluster they belong to. The lower figures shows the distribution of construction year (left-hand figure) and size of the livable area in square meters (right-hand figure) for the same four clusters.

As shown in figure 8, the two identified clusters also appear to be distinctively more expensive than other clusters with single-family houses in Copenhagen. Hence, we dub them *luxury villas* and view them as a separate submarket.

In 2019, the price gap was kr. 26,000 per square meter between these luxury villas and other single-family houses in the Copenhagen area. The luxury villas are not only more expensive, their prices also seem to diverge from other houses. Since 2012, square meter prices in these two clusters have increased by 6 per cent per year on average compared to 4 per cent for other houses in Greater Copenhagen. Hence, there seems to be different price trends in the market for single-family houses in Greater Copenhagen. Such developments are not clearly visible in traditional housing market data.

Figure 8
Square meter prices on single-family houses in Copenhagen



NOTE: The figure shows the development of average square meter prices on two groups of clusters with single-family houses in Greater Copenhagen (left-hand figure) and the difference between the two groups (right-hand figure). Luxury villas are defined as clusters 58 and 86 in figure 7. Other villas in the Copenhagen area are the rest of the clusters where a minimum of 80 per cent of the housing units are single-family houses in Statistics Denmark's definition of Copenhagen and suburbs.

The data further allows us to investigate differences in drivers of price developments of luxury villas relative to other houses by using an econometric approach. To do so, we enrich the micro-level data set with public information on fundamental drivers of housing prices (e.g. income levels, taxes, and interest rates) to form a panel data set of the Danish housing market based on the identified segments.[19] This information is only publically available at municipality level. Therefore, we calculate the shares of housing units across municipalities in each cluster and use them as weights when calculating the cluster-specific variables. Thus, we make the simplifying assumption that a specific household can be represented by the average municipality level of fundamental variables within the municipality of the household. We use the panel data set to propose a demand expression for the Danish housing market in error correction form based on previous literature on the subject, e.g. Hviid (2017b). By using this approach, we can identify drivers of prices specific to a particular submarket.

The error correction model consists of housing prices, income, and costs of owning the house. The cost of owning a house is defined as an average of the user costs, including interest rate payments and taxes, and minimum first-year repayments. Income is defined as aggregate disposable income in order to account for both income levels and demographics in the municipality that the housing unit belongs to. In the literature, the housing supply is traditionally included in the model, but due to our short data sample we are unable to gain insights from this slow-moving variable. The error correction model can be formulated as

$$\Delta logP_t = c + \alpha_1 \Delta logY_{t-1} + \alpha_3 \Delta C_{t-1} + \beta_1 logP_{t-1} + \beta_2 logY_{t-1} + \beta_3 C_{t-1} + \epsilon_t \quad (1)$$

where P is the real sales price, Y is real aggregate disposable income, and C are the costs of owning a house in real terms. The error term is assumed to be independent and identically distributed (iid) across observations.

Table 2 provides the estimation results from our error correction model. The first column shows results from a regression based on the aggregate housing market without taking into account differences in drivers across specific clusters. These results serve as a baseline for assessing price sensitivity and allows for a comparison of our results to previous housing market research in Denmark, e.g. Hviid et al. (2016b) and Hviid (2017b). The baseline results are largely in line with the literature. It confirms that on the aggregate level, we are able to produce similar results to studies that are based on more traditional housing market data. The second column shows estimation results which single out effects for luxury villas and for other single-family houses in Greater Copenhagen. We note that the cluster-specific interactions are statistically significant.

To aid the interpretation of the results, we rewrite the estimation results of table 2 into a steady-state representation[20]

[19] See appendix D for the construction of the panel data set.

[20] The steady-state representation is found by removing all short-term expressions in the error correction model, that is changes to prices, income and costs.

Table 2
Estimation results for error correction model

		Baseline	Copenhagen villa market
Real house prices, change	$\Delta \log P_t$		
Real income, change	$\Delta \log Y_t$	1.72^{***}	1.70^{***}
		(0.31)	(0.29)
Costs, change	ΔC_t	-10.05^{***}	-9.82^{***}
		(1.68)	(1.69)
Real house prices, lagged	$\log P_{t-1}$	-0.88^{***}	-0.89^{***}
		(0.02)	(0.02)
Real income, lagged	$\log Y_{t-1}$	1.74^{***}	1.67^{***}
		(0.08)	(0.09)
- *for luxury villas*	$\log Y_{t-1} \times luxury$		-0.27^{**}
			(0.12)
- *for other Copenhagen villas*	$\log Y_{t-1} \times other$		0.75^{***}
			(0.22)
Costs, lagged	C_{t-1}	-11.46^{***}	-11.22^{***}
		(1.02)	(1.14)
- *for luxury villas*	$C_{t-1} \times luxury$		-2.83^{**}
			(1.31)
- *for other Copenhagen villas*	$C_{t-1} \times other$		1.87
			(2.40)
Constant		-8.39^{***}	-8.75^{***}
		(1.02)	(1.03)
Fixed effects		✓	✓
Observations		4,144	4,144

NOTE: Numbers in parentheses represent robust standard errors.
***,**, and * represent significance at a 1, 5, and 10 per cent confidence interval, respectively.

$$logP_{luxury} = c + 1.58 \times logY - 15.89 \times C + \epsilon \tag{2}$$

$$logP_{other} = c + 2.73 \times logY - 12.61 \times C + \epsilon, \tag{3}$$

where the coefficients on Y and C express the elasticity of housing prices to income and the semi-elasticity of housing prices to costs, respectively.

The semi-elasticity of prices on luxury villas to costs is significantly higher than for other Greater Copenhagen houses. In other words, prices on luxury villas seem more sensitive to changes in interest rates and tax costs relative to the rest of the single-family house stock in Greater Copenhagen as well as to the rest of the housing market in Denmark. In contrast, prices on luxury villas are also less sensitive to changes in income.

Although explorative, this exercise illustrates that the combination of data from Boligsiden and machine learning tools in our algorithm can be used to reveal differences in underlying trends and dynamics in distinct parts of the housing market. The explicit findings of this exercise may be relevant for further research on housing markets and inequality or a further deep dive into the heterogeneity of housing market dynamics.

5 Concluding remarks

In this paper, we develop a novel tool for segmentation of the Danish housing market based on a combination of housing characteristics that impact sales prices the most, as determined by our model. We make use of advanced machine learning techniques on unique internet-based data, which cover both housing characteristics and market outcomes on the housing unit level. Our method differs from traditional housing market analysis where administrative classifications such as postal codes or municipalities are often used to segment the housing market.

We show that this new segmentation algorithm supports the ability to monitor heterogeneity at the micro-level. We argue that such segmentation tools can be used as inputs to the ongoing monitoring of the housing markets across the country, either by following the price development of single clusters over time, or by identifying whether cluster compositions themselves change over time as markets heat up and cool down.

Through specific case studies of urban housing markets in Denmark, we show that underlying heterogeneity of price dynamics exists in the villa market, with e.g. prices for luxury villas in Copenhagen being significantly more sensitive to cost shocks than other villas in the same geographic location. Although our data sample does not cover boom-bust-cycles in the Danish housing market, the ability to monitor such heterogeneity could prove useful if the cycle turns.

As our data sample covers housing sales between 2009 and 2019, this paper does not address the developments after the outbreak of covid-19. However, house prices in Denmark have soared during the pandemic as in most western countries. Prices on single-family houses in Denmark have increased by 4.2 per cent in 2020 and 12.7 per cent again in 2021. Increases have been greater than what fundamentals such as interest rates and incomes would suggest. Instead, an increased preference for housing among households is expected to have contributed to large price increases and marked house trading activity. This preference is believed to be partly temporary. See Hviid et al. (2021) for an account of drivers behind developments on the Danish housing market during the pandemic. The algorithm introduced in this paper will have a potential for detecting heterogeneous trends during the pandemic when feeding it with updated data.

References

BAYER, P., R. MCMILLAN, A. MURPHY, AND C. TIMMINS (2016): "A Dynamic Model of Demand for Houses and Neighborhoods," *Econometrica*, 84, 893–942.

BISHOP, C. M. (2006): *Pattern recognition and machine learning*, Springer.

BOURASSA, S. C., M. HOESLI, AND V. S. PENGD (2003): "Do housing submarkets really matter?" *Journal of Housing Economics*, 12, 12–28.

CAMPELLO, R. J., D. MOULAVI, AND J. SANDER (2013): "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*, Springer, 160–172.

CHEN, T. AND C. GUESTRIN (2016): "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, ACM, 785–794.

CLAEYS, G., K. EFSTATHIOU, AND D. SCHOENMAKER (2017): "Spotting excessive regional house price growth and what to do about it," *Bruegel, Policy Constribution*, 26.

DAM, N. A., T. S. HVOLBØL, E. H. PEDERSEN, P. B. SØRENSEN, AND S. H. THAMSBORG (2011): "Developments in the Market for Owner-Occupied Housing in Recent Years – Can House Prices be Explained?" *Danmarks Nationalbank Monetary Review*, 1st Quarter - Part 2.

DAM, N. A., T. S. HVOLBØL, AND M. H. RASMUSSEN (2014): "A multispeed housing market," *Danmarks Nationalbank Monetary Review*, 3rd Quarter.

DANMARKS NATIONALBANK (2018): "Outlook for the Danish Economy: Boom with no signs of imbalances," *Danmarks Nationalbank Analysis*, 15.

——— (2019a): "Outlook for the Danish Economy: Slightly lower growth in the coming years," *Danmarks Nationalbank Analysis*, 29.

——— (2019b): "Outlook for the Danish Economy: The Danish economy is heading deeper into the boom," *Danmarks Nationalbank Analysis*, 7.

DANMARKS STATISTIK (2018): "It-anvendelse i befolkningen 2018," Report, Danmarks Statistik.

DE ØKONOMISKE RÅD (2016): "Ejerboligbeskatning: Principper og erfaringer," Tech. rep., De Økonomiske Råd.

——— (2019): "Dansk Økonomi," Tech. rep., De Økonomiske Råd.

ESTER, M., H.-P. KRIEGEL, J. SANDER, X. XU, ET AL. (1996): "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, 226–231.

EUROPEAN COMMISSION (2019): "Use of Internet Services," Report, European Commission.

FINANSMINISTERIET (2019): "Økonomisk Redegørelse," Tech. rep., Finansministeriet.

GOODMAN, A. C. (1978): "Hedonic Prices, Price Indices, and Housing Markets," *Journal of Urban Economics*, 5, 471–484.

——— (1981): "Housing Submarkets Within Urban Areas: Definitions and Evidence," *Journal of Regional Science*, 21, 175–185.

GOODMAN, A. C. AND T. G. THIBODEAU (1998): "Housing Market Segmentation," *Journal of Housing Economics*, 7, 121–143.

HEEBØLL, C. (2014): "Regional Danish housing booms and the effects of financial deregulation and expansionary economic policy," *Kraka - Finanskrisekommissionen*.

HILBER, C. A. L. AND W. VERMEULEN (2014): "The Impact of Supply Constraints on House Prices in England," *The Economic Journal*, 116, 358–405.

HO, G. (2016): "House Prices in Denmark's Cities: The Role of Supply," *IMF Country Report*, 13–23.

HOLLY, S., M. H. PESARAN, AND T. YAMAGATA (2010): "A spatio-temporal model of house prices in the USA," *Journal of Econometrics*, 158, 160–173.

HUANG, M. (2019): "A Nationwide or Localized Housing Crisis? Evidence from Structural Instability in US Housing Price and Volume Cycles," *Computational Economics*, 53.

HVIID, S. J. (2017a): "A leading indicator of house-price bubbles," *Danmarks Nationalbank Working Paper*, 114.

——— (2017b): "A regional model of the Danish housing market," *Danmarks Nationalbank Working Paper*, 121.

HVIID, S. J., T. S. HVOLBØL, A. KLEIN, P. L. KRAMP, AND E. H. PEDERSEN (2016a): "House price bubbles and the advantages of stabilising housing taxation," *Danmarks Nationalbank Monetary Review*, Quarter 3.

HVIID, S. J., T. S. HVOLBØL, AND E. H. PEDERSEN (2016b): "Regional aspects of the housing market," *Danmarks Nationalbank Monetary Review*, 4th Quarter, 47–59.

HVIID, S. J. AND P. L. KRAMP (2017): "Housing Taxation Agreement Stabilises House Prices," *Danmarks Nationalbank Analysis*, 14.

HVIID, S. J., A. M. B. SCHMITH, S. H. THINGGAARD, AND J. PEDERSEN (2021): "Housing market robustness should be strengthened," *Danmarks Nationalbank Analysis*, 16.

LI, M. M. AND H. J. BROWN (1980): "Micro-Neighborhood Externalities and Hedonic Housing Prices," *Land Economics*, 56, 125–141.

LOBERTO, M., A. LUCIANI, AND M. PANGALLO (2018): "The potential of big housing data: an application to the Italian real-estate market," *Bank of Italy Working Paper*.

MALPEZZI, S., L. OZANNE, AND T. THIBODEAU (1980): *Characteristic prices of housing in fifty-nine metropolitan areas*, Urban Institute.

MCINNES, L., J. HEALY, AND S. ASTELS (2017): "hdbscan: Hierarchical density based clustering." *J. Open Source Software*, 2, 205.

MCINNES, L., J. HEALY, AND J. MELVILLE (2018): "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*.

MEEN, G. (1999): "Regional House Prices and the Ripple Effect: A New Interpretation," *Housing Studies*, 14, 733–753.

OIKARINEN, E. (2006): "The diffusion of housing price movements from center to surrounding areas," *Journal of Hounsing Research*, 15, 3–28.

OIKARINEN, E. AND J. ENGBLOM (2014): "Regional differences in housing price dynamics: Panel data evidence," *Aboa Centre for Economics, Discussion Papers*, 94.

PALM, R. (1978): "Spatial Segmentation of the Urban Housing Market," *Economic Geography*, 54, 210–221.

PAN, S. J. AND Q. YANG (2009): "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, 22, 1345–1359.

PIAZZESI, M., M. SCHNEIDER, AND J. STROEBEL (2020): "Segmented Housing Search," *American Economic Review*, 110, 720–759.

RAE, A. AND E. SENER (2016): "How website users segment a city: The geography of housing search in London," *Cities*, 52.

ROSEN, S. (1974): "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 82, 34–55.

SCHNARE, A. AND R. STRUYK (1976): "Segmentation in Urban Housing Markets," *Journal of Urban Economics*, 3, 146–166.

STRASZHEIM, M. R. (1975): "An Econometric Analysis of the Urban Housing Market," *New York: National Bureau of Economic Research*.

VAN DER MAATEN, L. AND G. HINTON (2008): "Visualizing data using t-SNE," *Journal of machine learning research*, 9, 2579–2605.

VAN DIJK, D. AND M. FRANCKE (2015): "Internet search behavior, liquidity and prices in the housing market," *DNB Working Papers*.

ZIETZ, J., E. N. ZIETZ, AND G. S. SIRMANS (2008): "Determinants of house prices: a quantile regression approach," *The Journal of Real Estate Finance and Economics*, 37, 317–333.

ØKONOMI- OG INDENRIGSMINISTERIET (2018): "Økonomisk Redegørelse," Tech. rep., Økonomi- og Indenrigsministeriet.

Appendix

A Data cleaning

In order to remove extreme outliers from the dataset, we imposed the following conditions on the raw data:

- The housing unit must be available in BBR.
- The housing unit must have more than zero clicks.
- The sales price must be in the interval kr. $]100,000; 100,000,000]$.
- The housing unit must have a total living area of $]10; 1,000]$ square meters.
- The construction year of the housing unit must be after year 1200.
- The number of rooms must be below 20
- The number of toilets/bathrooms must be below 10

Subsequently, we run a series of preprocessing steps to perform the following operations on the data:

- The construction year for each house is set to the maximum of either the time of construction or the time of any major renovation, i.e. houses that are renovated are counted as 'new'.
- The number of school's in the municipality is normalized with the municipality population.
- The number of childcare facilities in the municipality is aggregated accross categories (daycare units ('vuggestue'), kindergarden ('børnehave'), and compound institutions ('integrerede institutioner')), and normalized in a similar fashion.
- For apartments, the floor is truncated to the $[-1, 4]$ range.
- The number of bathrooms is truncated to the $[0, 2]$ range.
- Indicators for the house having a garage or a carport are converted into the same boolean value, indicating if it has either.
- The construction year is set to unknown (NaN) if lower than 1800.
- Houses have an energy rating, most commonly a letter from A to G, with A being more energy efficient. These are converted into a numbered list with A mapped to 1, B to 2, and so forth. Some ratings have more granular 'subratings' called A2010, A2015. These were aggregated into one.
- Housing areas were truncated to the 0-$1,000$ square meter range.
- Distances from housing to railroads, industry, and highways, as well as estate areas, have been log-transformed.

B List of variables

Features	Description
Sag_SagtypeNr_100	Dummy: Villas
Sag_SagtypeNr_200	Dummy: Apartments
Sag_SagtypeNr_300	Dummy: Townhouses
Sag_SagtypeNr_900	Dummy: Villa apartments
Bygning_GOP_OpfoerselesAAr	Construction year
Afstand_Kyst	Distance to coast
Afstand_Vindmoelle	Distance to wind turbine
Afstand_Landingsbane	Distance to airport
Sag_Etage	Floor
Enhed_GOP_AntVaerelseBebo	Number of rooms
Bygning_IOP_VarmeinstalKode	Type of heating
Enhed_GOP_AntToilet	Number of bathrooms
Kommune_SkatteProcent	Tax rate on personal income in municipality
Kommune_Grundskyld	Tax rate on land ownership in municipality
Sag_ArealVaegtet	Size of unit in square meters (weighted)
Sag_EnergiMaerke	Energy rating
Sag_ArealBolig	Living unit's m^2
geocluster_*	Geocluster dummies
Kommune_FolkeSkoler/1k	Number of primary schools per capita in municipality
Kommune_Vuggestuer/1k	Number of daycare units per capita in municipality
Garage	Dummy for a garage
Afstand_JernbaneSynlig_Log	Distance to surface railroad (log)
Afstand_Industri_Log	Distance to industrial complex (log)
Afstand_Motorvej_Log	Distance to highway (log)
Sag_ArealGrund_Log	Plot size in square meters
Sag_GeoPostNr_*	Zip code dummies

C Machine learning pipeline

This section of the appendix contains details about our implementation of the machine learning steps introduced in section 3.

C.1 Neural network representation

To avoid overfitting, the number of neurons in the two hidden layers of the network are determined so as to keep the ratio of degrees of freedom in the principal components and the neural network above 5, resulting in 131 neurons in each hidden layer. After fitting, the network may be thought of as a large composite function which maps the input principal components to an estimated price via a series of steps, each one converting the output from one layer to the values at the next layer. The last hidden layer is often called the *representation layer*, because it contains a representation of the data in terms of complex patterns which are optimized to best predict the target value (square meter price). Using a neural network to convert data into another representation in this way has proven extremely useful in a number of machine learning applications (Pan and Yang, 2009).

C.2 UMAP and two-dimensional space embedding

UMAP assigns to each data point a set of low-dimensional coordinates such that the neighborhood of the point is similar to its neighborhood in the original, high dimensional space. Similarity in the sense used here is defined in the following way: At each data point in the high dimensional space, a probability distribution over the remaining points is defined such that probabilities decrease with distance.

A similar distribution is defined in the low-dimensional space, where the distances, and hence the probabilities, depend on the embedding coordinates. The embedding coordinates are then chosen to minimize a difference measure between the two probability distributions.

In the original space, UMAP defines for each data point i the probability distribution over other points j as

$$p_{ij} = \mathrm{e}^{-(d(x_i,x_j)-\rho_i)/\sigma_i}, \tag{A.1}$$

where $d(\cdot,\cdot)$ denotes the distance between two points, ρ_i denotes the distance to the point closest to x_i, and σ_i is computed such that the probability distribution has a desired number of effective neighbors

$$k = 2^{\sum_j p_{ij}}, \tag{A.2}$$

which is specified as a hyperparameter to the algorithm. In the lower-dimensional embedding space, the probability distributions are given by

$$q_{ij} = \frac{1}{1 + a\,|y_i - y_2|^{2b}}, \tag{A.3}$$

where a and b are specified through a hyperparameter. The difference measure which is then minimized using a gradient-descent approach is the cross-entropy, defined as

$$CE(X,Y) = \sum_i \sum_j \left(p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1-p_{ij}) \log \frac{1-p_{ij}}{1-q_{ij}} \right). \tag{A.4}$$

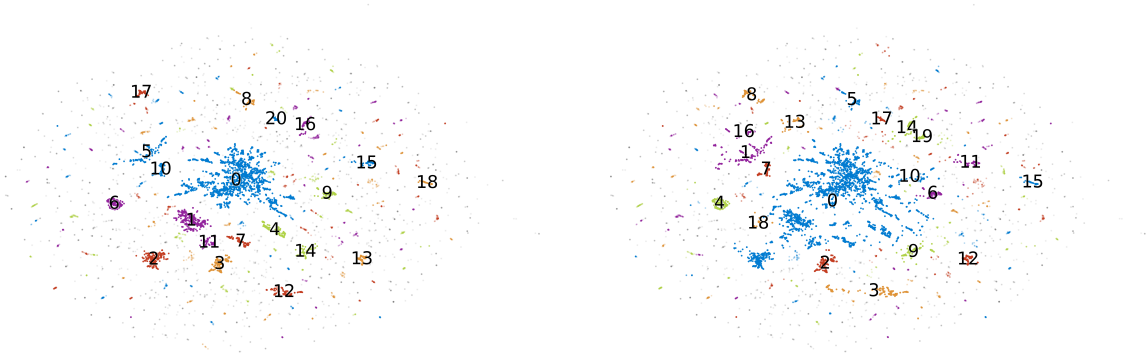
This approach shares some similarities with another technique, t-distributed stochastic neighbor embedding (tSNE) (van der Maaten and Hinton, 2008), which has been in use for a longer period of time but is less suited for clustering purposes, because it only retains the local, not global, structure, i.e. points that are close together in the high-dimensional space will end up close together in the embedding space, but the converse is not ensured. tSNE works in a similar way and defines a Gaussian probability distribution in the original space and a Cauchy distribution in the embedding space, and it then minimizes the Kullbeck-Leibler divergence between them

$$\sum_i D_{\mathrm{KL}}(q||p) = -\sum_{i,j} q_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right). \tag{A.5}$$

The relevant qualitative difference is that whereas both eq. (A.4) and eq. (A.5) penalize situations where points that are close in the original space end up far from each other in the embedding space, only eq. (A.4) penalizes points being far from each other in the original space but close together in the embedding space, thus preserving global structure in the data.[21] For this reason, UMAP is well suited for density-based clustering algorithms. The UMAP instance in the present application used the default settings in the python implementation (McInnes et al., 2018), and was set to map its input into a two-dimensional space.

[21] An excellent explanation of this difference can be found at https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668.

Figure A.1
Clustering housing units based on their 2-dimensional embedding via HDBSCAN



(a) Preferred clustering parametrization **(b) Alternative clustering parametrization (fewer, larger clusters)**

Note: The figure plots a random 10 per cent sample of our data in the two-dimensional artificial embedding space we use for final clustering. Each dot corresponds to a housing unit. Housing unit are color-coded according to the assigned cluster under a given specification of the HDBSCAN clustering algorithm, with the specification shown on the right producing fewer, larger clusters. Only the 20 largest clusters in each specification are labeled.

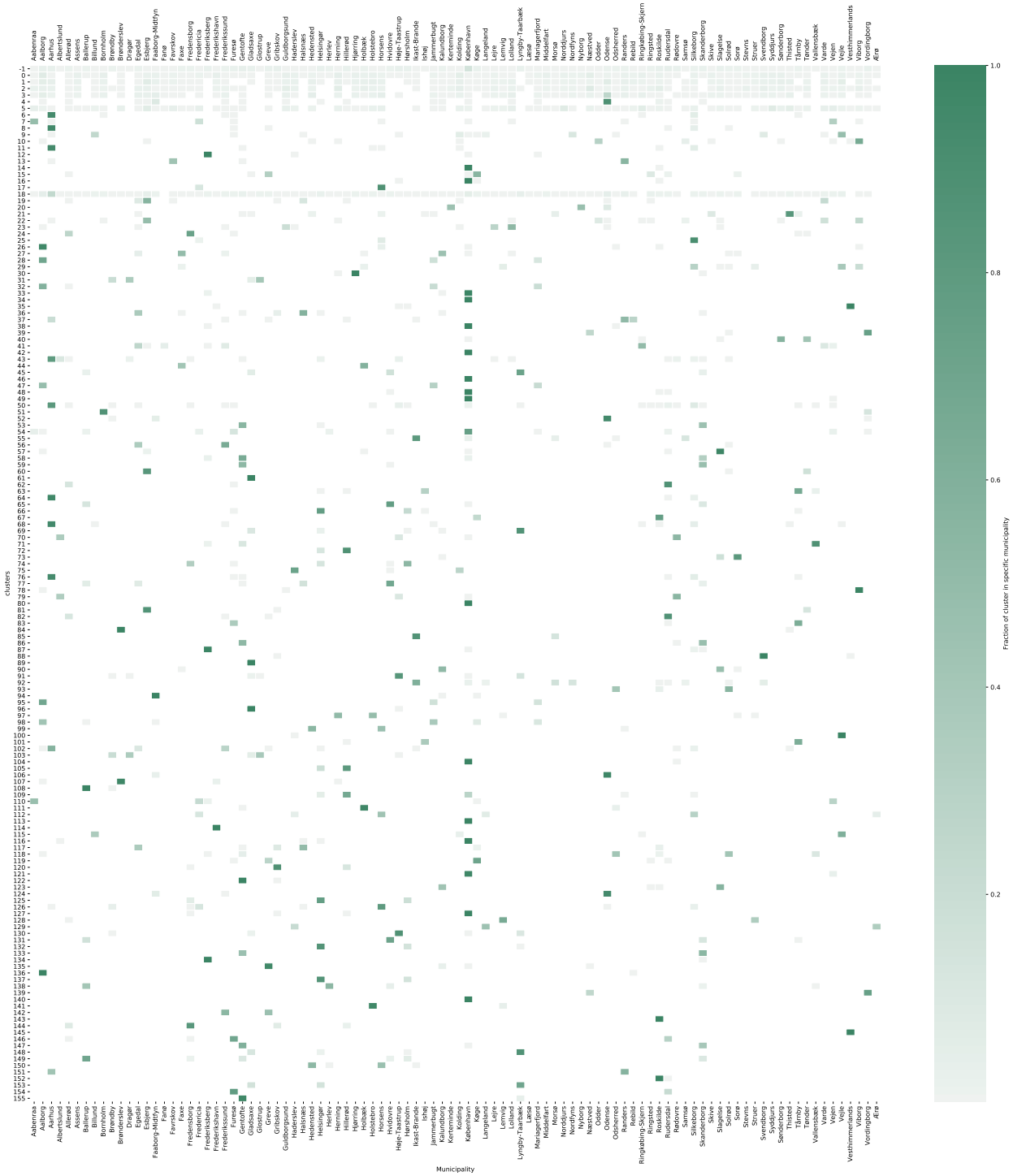
C.3 Hierarchical clustering with HDBSCAN

An example of extracting clusters with varying granularities appears in figure A.1. The figure shows a random sample of 10,000 housing units in our sample in the embedding space produced by UMAP. While these units are visibly clustered in lumps in this space, the tolerance with which the researcher considers lumps to be separated from each other can vary.

The left panel shows our preferred clustering configuration. This configuration produces 157 clusters, where all clusters contain at least 500 housing units, and we consider as outliers all the smallest clusters that together represent less than 1 per cent of the housing units in our sample. The central, largest cluster makes up 18.7 per cent of our sample and represents mostly single-family homes in rural areas.

In the example in the right panel, large gaps in densities between points in the space must exist in order to be separated into different clusters. This approach creates few large clusters, but which might include heterogeneous types of housing units. For example, the central large cluster expands to 32.7 per cent of the sample, absorbing other types of single-family homes that are slightly larger and are located in small towns. Some of these absorbed groups are also significantly more expensive, and their price has increased relatively more in the past ten years with respect to the houses in the original cluster.

Figure A.2
Overlap between clusters and municipalities



NOTE: Combinations of clusters and municipalities are colored according to the proportion of a cluster located in the specific municipality. Clusters are ordered by size, with the largest at the top of the figure.

D Constructing a panel data set of the housing market

In order to make a econometric analysis of house price drivers across certain clusters, we construct a panel data set for clusters at the quarterly frequency. To do so, we start by collecting and constructing measures of income, demographics, and costs, which reflect short-run house price fundamentals. We collect and construct the variables at the geographically most granular level possible. We go through these in the following.

We collect aggregate disposable income from Statistics Denmark *INDKP106* at the annual frequency at municipality level. Aggregate disposable income contains information on both income and demographic developments. For each housing unit, we assign the annual level of disposable income of the municipality of the housing unit in the year in which the unit was sold.

We construct costs by following the same methodological approach. Costs are an average of user costs and minimum first year repayments.

User costs reflect the benchmark interest and contribution rate on a 30-year fixed-rate covered mortgage bond after tax in real terms, as well as the effective real estate tax and inflation expectations. Inflation expectations are modeled through Muth-Pischke smoothing. Minimum first year repayments reflect the interest rate and contribution rate on a 1-year covered mortgage bond after tax as well as the effective real estate tax and minimum amortization requirements. The latter are subject to the fact that interest-only loans can be obtained for a large part of financing in Denmark.

The effective real estate tax is computed as the nominal proceeds from real estate taxes[22] relative to the market value of the housing stock.

The frequency of all the variables is quarterly. Interest rates, contribution rates, and inflation expectations are measured at the national level, while effective tax rates are computed at the housing unit level according to the municipality of the housing unit. In general, variables in real terms are deflated by the consumer price index.

Having attributed the observations of the variables to each housing unit, we collapse data from the housing unit level to the cluster level at a quarterly frequency. We collapse by taking cluster averages, implying that the observations represent the cluster level average of the variable in question for sold houses within the cluster in a specific quarter. This leaves us with a panel set with clusters along the id dimension and quarters along the time dimension.

[22] Proceeds are compiled from registry data at Statistics Denmark, where public valuations are observed. From these, the paid taxes - and thus public proceeds - can be imputed.

PUBLICATIONS

NEWS

News offers quick and accessible insight into an Analysis, an Economic Memo, a Working Paper or a Report from Danmarks Nationalbank. News is published continuously.

ANALYSIS

Analyses from Danmarks Nationalbank focus on economic and financial matters. Some Analyses are published at regular intervals, e.g. *Outlook for the Danish economy* and *Financial stability*. Other Analyses are published continuously.

REPORT

Reports comprise recurring reports and reviews of the functioning of Danmarks Nationalbank and include, for instance, the *Annual report* and the annual publication *Danish government borrowing and debt.*

ECONOMIC MEMO

An Economic Memo is a cross between an Analysis and a Working Paper and often shows the ongoing study of the authors. The publication series is primarily aimed at professionals. Economic Memos are published continuously.

WORKING PAPER

Working Papers present research projects by economists in Danmarks Nationalbank and their associates. The series is primarily targeted at professionals and people with an interest in academia. Working Papers are published continuously.

DANMARKS NATIONALBANK
LANGELINIE ALLÉ 47
DK-2100 COPENHAGEN Ø
WWW.NATIONALBANKEN.DK

DANMARKS NATIONALBANK

As a rule, Working Papers are not translated, but are available in the original language used by the contributor.

Danmarks Nationalbank's Working Paper are published in PDF format at www.nationalbanken.dk. A free electronic subscription is also available at the website. The subscriber receives an e-mail notification whenever a new Working Paper is published.

Please direct any enquiries directly to the contributors or to Danmarks Nationalbank, Communications, *Kommunikation@nationalbanken.dk.*