Norwegian University of Life Sciences (NMBU)

Price and Hedonic Heterogeneity Measures in Local Housing Markets

Dag Einar Sommervoll

Norwegian University of Life Sciences
Centre for Land Tenure Studies

Centre for Land Tenure Studies Working Paper 02/23
ISBN: 978-82-7490-310-4

Price and Hedonic Heterogeneity Measures in Local Housing Markets

Dag Einar Sommervoll[a]

[a]School of Economics and Business, Norwegian University of Life Sciences and NTNU Trondheim Business School

Abstract

In this paper, we develop a local housing stock heterogeneity measure. This measure may be used to monitor housing stock heterogeneity over time and in combination with other measures of policy interest. We illustrate the latter by looking at local housing stock heterogeneity, house price variation (measured by local Gini coefficients), and affordability measures.

Keywords:
Housing market, Heterogeneity, segregation
JEL: R31, R21, C43

1. Introduction

The world sees rapid urbanization, and ever more of us live in cities. Edvard Glaeser argues in his critically acclaimed book "Triumph of the city" ([Gla11]) that cities are diverse, environmentally friendly, and creates a vibrant place to live and prosper. At the same time, cities may become diverse in a sinister way. Urban segregation is today a concern in many cities ([FCM+07],[TO14]). Deprived neighborhoods may be focal points for crime and radicalization. Moreover, such neighborhoods may be downward spiraling, as those who can settle elsewhere do.

From a housing market policy perspective, it is arguably of principal importance to be proactive regarding segregation and downward spiraling neighborhoods. One way to mitigate segregation is to create socio-demographic heterogeneous neighborhoods. In a society where its citizens decide and finance themselves where to live, anti-segregation policy measures boil down to facilitate socio-demographic mixing. A mix can be achieved by targeted policies. A natural first bliss is to create a heterogeneous housing stock.

In order to ensure housing stock heterogeneity, we need to rely on some measure of heterogeneity. This is our point of departure. We seek to construct a local housing marked heterogeneity measure. Interestingly, dwelling homogeneity is easy to define

Email address: dag.einar.sommervoll@nmbu.no (Dag Einar Sommervoll)

and understand; two flats with the same hedonic characteristics are homogeneous. On the other hand, dwelling heterogeneity is subtle; is a flat of 120 square meters more different from a house of 120 square meters than a dwelling of 40 square meters? If we want to address spatial housing market heterogeneity and develop a spatial measure for heterogeneity, we must compare such differences by a numerical measure.

This paper will develop a local housing heterogeneity measure for every dwelling in a given geographical area. From a policy perspective, such a measure can be of importance as a way to monitor heterogeneity. Furthermore, this monitoring provides a way to evaluate if policy measures increase heterogeneity. Our contribution is connected to research regarding the spatial dimension of segregation ([Vau07],[VA11]).

As a measure must be judged by its value in use, we include a brief discussion of the measure applied to the metropolitan area of Oslo. From a policy point of view, the litmus test of such a measure is whether it can be integrated with other analytic tools and unravel this spatial variation that would be otherwise hard to assess. We illustrate the value added by the local housing market heterogeneity measure by focusing on housing affordability. From an affordability perspective, a locally diverse housing stock only helps a little if all dwellings are too expensive. Therefore, we pair the local heterogeneity measure with a local measure of price heterogeneity. We use this to analyze which local housing markets are diverse along both dimensions and which are not. It may be of importance from a policy perspective to know if lack of heterogeneity is primarily a housing stock issue or whether limited price heterogeneity is the driver of segregation.

Another important use of these measures, which we will not illustrate this paper, is the monitoring local housing markets over time. To ensure diverse neighborhoods requires both identification, monitoring and counter-measures for potential trouble spots. It is well known that ex ante measures in general tend to be easier and less costly, than ex post interventions. Given this insight, may monitoring be the most important application of the measures developed and discussed in this paper.

The remainder of the paper is organized as follows. In Section 2, we develop the local housing market heterogeneity measure. Section 3 gives an overview of the technical details of the measure. Comparing two dwelling characteristics, like size and floor, is non-trivial and requires converting two very different features to one common scale. The technical term for this is feature engineering, and care must be taken to do this coherently. The latter part of the paper illustrates the potential uses of the heterogeneity measure in the case of Oslo. We present heterogeneity maps in Section 4 and proceed in Section 5 with Gini coefficients as a measure of house price heterogeneity. The local housing stock measure is paired with the Gini measure in Section 6. In the latter part of Section 6, develop an ad hoc affordable housing measure and pair this med the local housing stock heterogeneity measure. Section 7 concludes.

2. A hedonic heterogeneity measure

In this section, we will construct the heterogeneity measure for a local housing market around a given dwelling. Quantitative measures have a slight domain-specific variation. An index tends to refer to a quotient in statistics. The quotient's denominator refers to a quantity in a base period, whereas the nominator is the corresponding quantity at another (usually later) period. A typical example is a house price index, where an index of 1.15 gives a price increase of 15 percent. The finer details of index construction have been an area of intensive research; see, for example, [Die78].

In economics, using an index does not necessarily refer to a quotient. For example, the Herfindahl index ([Mar50])is a measure of market concentration that is just the squared sum of the firms' market shares. Our heterogeneity measure resembles the Herfindahl index as it is a weighted squared sum (of feature distances) and thus could be called an index. However, our measure is as the Herfindahl index dimensionless and thus independent of the actual measuring units related to housing features, like the size in square meters. We abstain from using the term index as we seek to distance ourselves from the field of economic index research.

A central idea dates back to Rosen ([Ros74]). We assume a house can be decomposed into a "sum" of characteristics. It must be stressed that as we for the heterogeneity index, we are not concerned with the respective pricing of hedonic characteristics. Our only concern is to measure and compare characteristics that will have a varying appeal to prospective home seekers, particularly different household types. It is essential to make these building blocks dimensionless, and we rely on feature engineering tools ([Mar11].

In the following, we will construct a heterogeneity measure for a local housing market around a dwelling. The construction relies on defining a local housing market. This local market consists of all dwellings within a given radius. Furthermore, a measure of the hedonic difference between this house and all houses in this local housing market is calculated. Then the measure is a (weighted) sum of all these hedonic differences.

Before we discuss these steps in detail, let us first fix some ideas. Let us assume that we have a set of houses $\{h_i, i \in I\}$. Furthermore, let X_{ij} be a set of feature-engineered characteristics/features[1] for the house i. The feature distance between

[1] We will give the details regarding feature engineering in the next section. Here we only rely on the fact that some numerical measure allow us to compare different hedonic features.

two houses i_1 and i_2 is the (normalized) weighted euclidean distance:[2]

$$d_f(h_{i_1}, h_{i_2})^2 = \sum_{i,j} w_j (x_{i_1 j} - x_{i_2 j})^2$$

This feature distance mirrors our customary physical distance measure, and as such similar houses have a smaller feature distance than less similar houses for every feature dimension. It must be stressed that this is true irrespective of feature engineering. Feature engineering only concerns comparisons across hedonic variables.

2.1. The local housing market for a given house

Any notion of local heterogeneity relies on considering which houses are spatially close enough to contribute to heterogeneity. Moreover, the houses closer to the house in question should contribute more as they are "more" local. A natural way to accommodate both requirements is to have a proximity function that weights closer houses more than more distant houses less. We choose the Epanechikov function $d_E(x) = \frac{3}{4}(1 - \frac{x}{r}^2)$, where x is the distance to the house in question and r is the "cut off"-radius. The cut-off radius is 500 meters. We refer to this function as the Epanechikov geo-weighing function.

The precise value of the cut off-radius is not critical; the important thing is that it serves as an average proxy for a distance so far away from the house in question that it belongs to another local housing market.

The definition of the heterogeneity index at house h_i is the following:

$$het(h_i) = \sum d_E(h_{i_1}, h_{i_2}) \cdot d_f(h_{i_1}, h_{i_2})$$

where $d_E()$ is the Epanechikov geoweighing function and $d_f()$ is the hedonic feature distance.

A few comments are in order. Houses that are hedonically different from the given house contributes more to to $het()$. Moreover, a hedonically different house close to house in question contributes even more. If we had pitched this measure construction as axiomatically, these properties would have been the axiom list. Another natural requirement is that the introduction of an extra hedonically different house within the cut-off radius increase the heterogeneity. The measure $het()$ has this property by construction as it is sums positive terms one for each house. Moreover, this measure can be calculated for every dwelling in a house market, hence we can assess the pointwise heterogeneity for a given area, say a metropolitan area. Oftentimes for analytic purposes we may choose to aggregate this heterogeneity measure some

[2] In this paper, all weights $w_j = 1$.

spatial units/administrative units. We will illustrate such aggregations in the latter part of this paper.

3. Feature Engineering

More often than not, feature engineering is required to assess the variables one by one to ensure that differences are measured consistently and in a way that makes comparisons across different variables meaningful. The case of hedonic variables for dwellings is no exception. In particular, finesse is needed as some variables vary considerably between dwelling types. (We have detached houses, duplexes, row houses, and flats.) In short, most variables must be feature engineered within each dwelling type. All details of feature engineering by dwelling type are given in the appendix. Here we will provide an overview and briefly discuss the feature engineering used.

This paper's heterogeneity measure constructed and discussed relies on four hedonic characteristics. These are size (in square meters), lot size (in square meters) for detached houses, Floor (for flats), and dwelling type.

A much-used feature engineering is to use unit interval normalization, that is, normalize a variable so that the max value is 1 and the min value is 0. This is achieved by making the variable transformation

$$x\text{unorm} = \frac{x - min(x)}{max(x) - min(x)}.$$

We rely on this normalization for Floor (for the dwelling type flats).

We rely on quartile normalization for each dwelling type for the continuous variable size. For lot size, we do this for detached houses and duplexes (lot size is zero for other dwelling types).

The formula for quartile normalization is:

$$x\text{qnorm} = \frac{x - Q_1(x)}{Q_3(x) - Q_1(x)},$$

where Q_1 and Q_3 is the 1st and 3rd quantile respectively.

The reason for relying on quartile normalization rather than normalization using the max and the min is due to extreme outliers that would truncate variation, say from 60 sq.m. to 70 sq. m. to a minor variation. In order to illustrate this challenge, assume that there is a 1000 sqm dwelling in the housing stock. It may be a mansion or a data entry mistake, either way; it is to normalize the characteristic sqm so that the presence of this observation does not dwarf significant variation from a household perspective.

The quartile normalization ensures that the middle half of the observations is between 0 and 1. We cap outliers at -1 and 2. It must be noted that the outlier capping only affects the few neighborhoods with outliers and that quartile normalization is vital in this feature engineering. This ensures that variation in the most common range is comparable across features.

The dwelling type is also a dimension that should count in the heterogeneity measure. A row house and a dwelling with equal hedonics apart from dwelling type are different. We order dwelling type the following way: Detached house (1), duplex (2), row house (3), and flat (4), and use this ordering in a unit interval normalization.

The feature engineering outlined above and given in full detail in the appendix (in the case of the Oslo Metropolitan area housing stock) ensure that the lion's share of dwelling characteristics are in the same range. However, it is not evident that all characteristics are equally important for heterogeneity. Therefore, we have introduced a weight in the feature difference d_f. In the analysis presented here, the weights are kept equal to one.

3.1. Properties of the heterogeneity measure

The review of the heterogeneity measure thus far has been technical. Furthermore, though the construction follows the well-trodden path of feature engineering, it may not be obvious that this measure has desirable properties. The measure has the following properties:

1. The measure is dimensionless. (It does not rely on the units we use for size and lot size.)

2. Hedonically similar houses have $d_f = 0$

3. Monotonicity: A new house in a local housing market can only increase the heterogeneity.

4. Distance property: Houses close to the local housing market center contribute more to the heterogeneity measure than more distant houses.

4. Illustration of the heterogeneity measure: The municipality of Oslo

In this section, we will consider maps that display the heterogeneity in Oslo by taking average heterogeneity at the smallest administrative unit (Grunnkrets). We will refer to these as administrative units or, for short, AUs.

Before we go into the specifics maps, some background regarding the Oslo housing market may be helpful. Historically there has been an east-west division. The western suburbs are generally more affluent than the eastern part. Along the north-south axis, houses tend to be more expensive in the north compared to the south, but this price differential is less clear and with more exceptions than in the east-west division. The most notable exceptions to low prices in the south and east stem from proximity to or a panoramic view of the fiord. This almost fractal dimension to the housing market is creating possibilities for a socio-demographic mix that could prevent segregation; at the same time, it also may mask local housing markets that are on a downward spiral as they may be close but not close enough to benefit from a less deprived neighborhood.

In the following, it is important to have this insight in mind. We are not interested in general trends in the east-west and north-south dimensions. We are not interested in the trends at the city district level (bydel level); we are interested in local housing markets. As such, we will, for display purposes, focus on the administrative units.[3]

Figure 1 is a heterogeneity map of Oslo. The heterogeneity is colored from red (low heterogeneity) to blue (high heterogeneity). We notice that dominant color on the map is red. This is because large AUs have few houses. Most of the northern part of Oslo is forest and has few houses and little housing stock heterogeneity. More interesting are the finer details in areas closer to the city center. Figure 2 zooms in on a smaller part of the Oslo municipality. This map shows the city center and western and eastern AUs close to the city center. Again, we see a striking local variation. AUs with low heterogeneity are immersed in AUs with considerably more heterogeneity. From an urban planner/urban policy perspective, the flagging of the AUs in red could be a good idea. The flagged AUs could be studied in more detail and targeted in future urban development.

[3] In Norwegian: Grunnkretser.

Figure 1: Oslo Overview

The dots in the plot are the heterogeneity measure for the dwelling at that location. The color of the area is the average heterogeneity at basic area ("Grunnkrets" the smallest administrative unit).

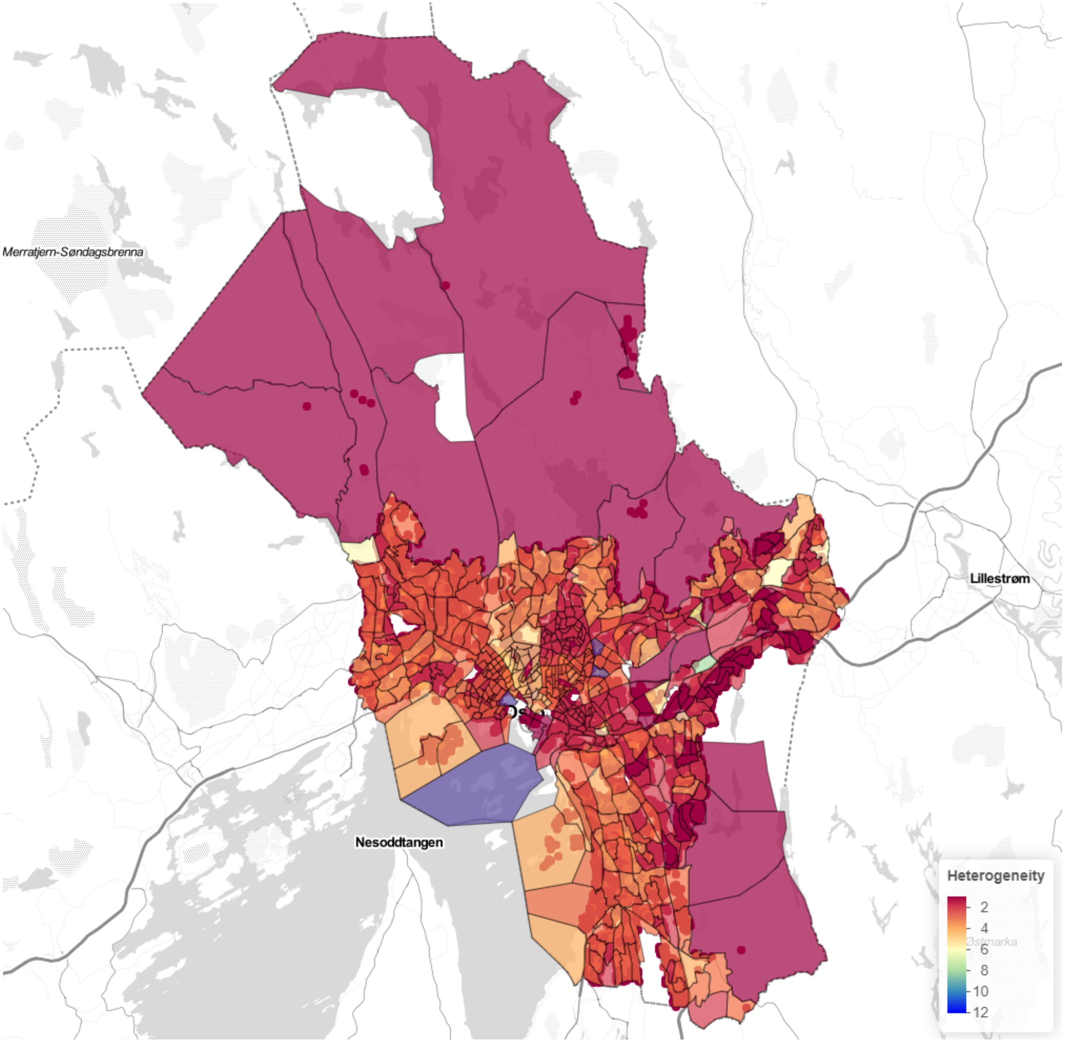
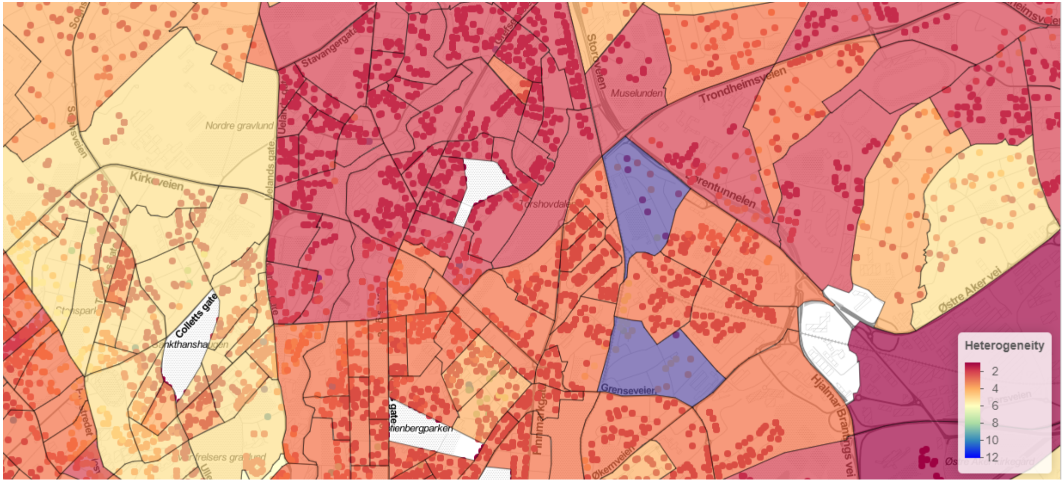


Figure 2: Oslo Detail

The dots in the plot are the heterogeneity measure for the dwelling at that location. The area's color is the average heterogeneity at AU level ("Grunnkrets" the smallest administrative unit).



5. Price heterogeneity

Housing stock heterogeneity is the first step in providing housing for households with differing housing needs and preferences. On the other hand, housing stock heterogeneity may amount to little if the housing stock heterogeneity does not translate to price heterogeneity. In this section, we will consider price heterogeneity at the AU level. A natural candidate for an (economic) variable like dwelling price is the Gini coefficient.[4] Note that we can only calculate the Gini coefficient for prices of transacted dwellings. Moreover, only "recent" transactions may be deemed relevant.[5]. We calculate the Gini coefficient at the AU level for transactions in 2019-2020 for AUs with more than 25 sales in these two years. Figure 3 gives the birds eye view of the municipality as a whole. The most notable difference compared with the corresponding map with housing stock heterogeneity is that the large AUs with few dwellings are not present. We also note that there is no clear center periphery dimension in the Gini coefficient map. It is interesting to note that both the housing stock heterogeneity map and the Gini map display considerable variation both close to and far from the city center.

Figure 4 zooms in on the city center and high-end western suburbs. We see that the Gini coefficient varies considerably even in these attractive neighborhoods. It must be stressed that this does not translate to a spread of transactions from the affordable to the too-expensive range for a "representative home buyer". The Gini could, and is, for some high-end local markets, driven by expensive dwellings. We will explore this in the next section and present a measure of affordable houses that emphasizes the frequency of (relatively) low-priced dwellings in the local housing market in question.

[4] The definition of the Gini coefficient is $G = \frac{\sum_i \sum_j |x_i - x_j|}{2n^2 \bar{x}}$.In other words, this is half of the relative mean difference. This lies between 0 and 1 when all x_i's are positive.

[5] One way to include more dated transactions may be to use a price index to adjust all transactions to a given (recent) time. This is unattractive for several reasons. One is that it is not obvious what index to use. A broad index, like a city index, may mask idiosyncratic price movements that may be important. Another reason is that if the Gini coefficient aims to tell what can be purchased, now or in the near future, dated sales may be irrelevant.

Figure 3: Oslo Overview

The area's color gives the Gini coefficient of arms-length sales in 2019-2020 at AU level (AU is "Grunnkrets" in Norwegian).

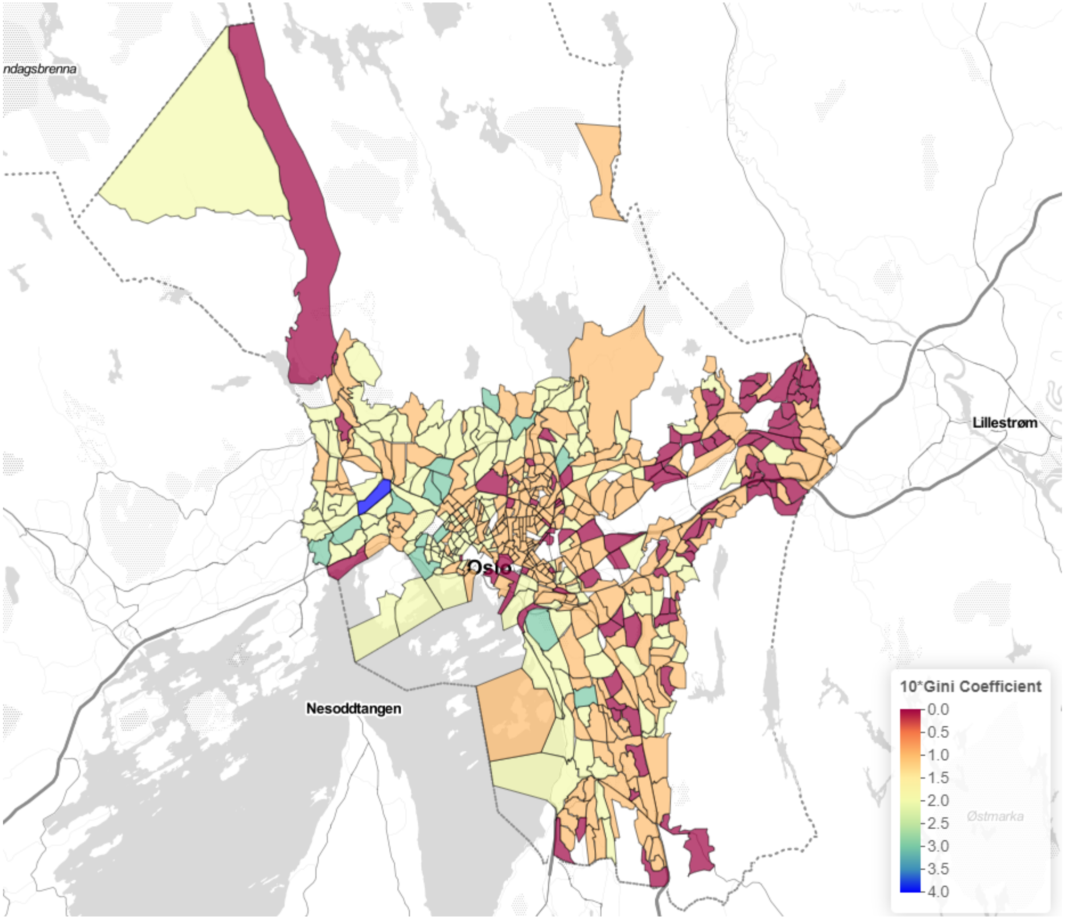
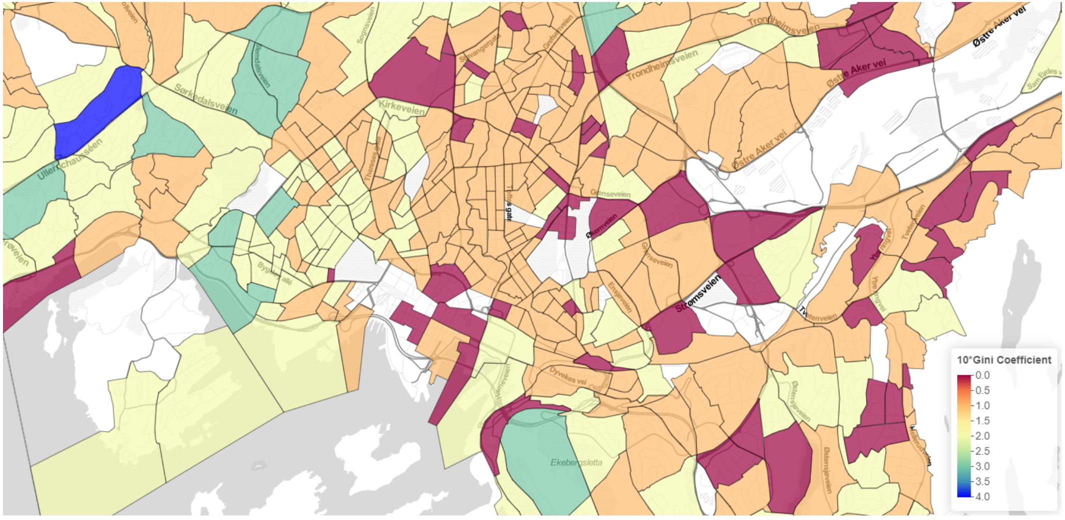


Figure 4: Oslo Detail

The area's color gives the Gini coefficient of arms-length sales in 2019-2020 at AU level (AU is "Grunnkrets" in Norwegian).



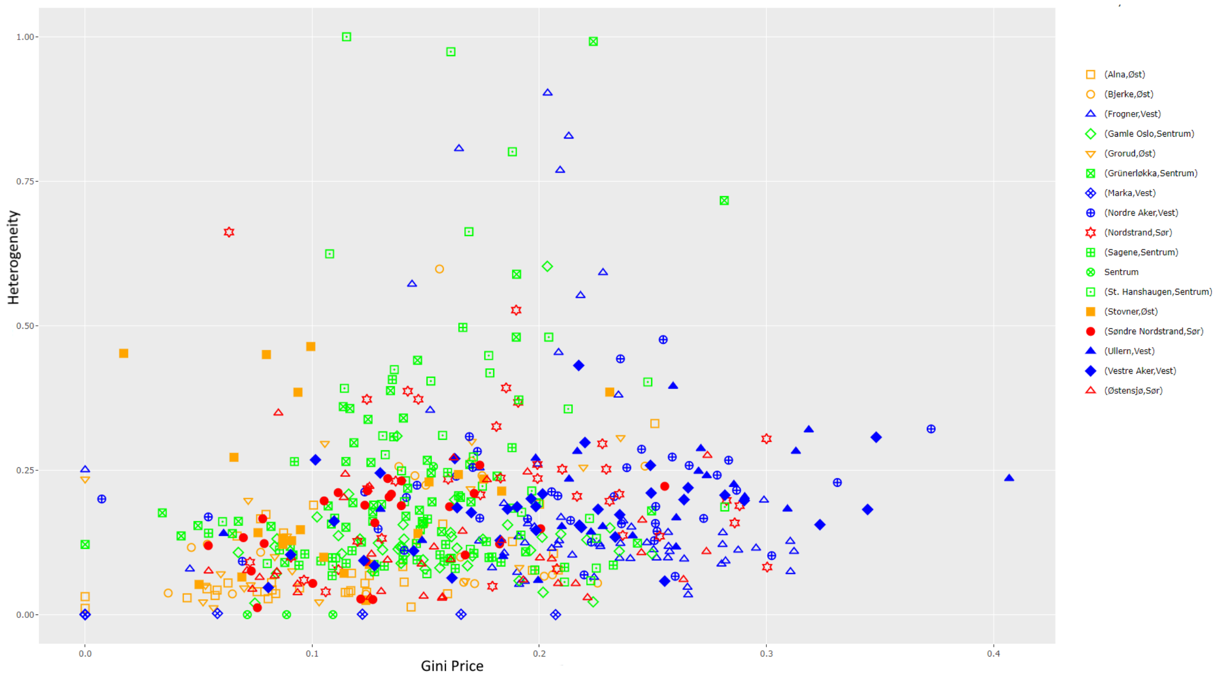
6. Heterogeneity in housing stock and price

In the introduction, we argued that in order to avoid segregation, socio-demographic heterogeneity in local housing markets may be a goal. Moreover, that housing stock heterogeneity with respect to hedonic characteristics and price is a reasonable first step to achieving such a goal. Moreover, we may have heterogeneity in the local housing stock but slight price variation, and vice versa. From a policy perspective, the countermeasures will depend on whether the heterogeneity challenge lies in the stock, the prices, or both.

In this section, we will shed light on the interplay between the two heterogeneity measures in the case of Oslo. Figure 5 displays all AUs in Oslo (that have dwellings and transaction enough to calculate the measures). The AUs far from the origo along both dimensions have a heterogeneous housing market and a high Gini coefficient for price. Furthermore, the figure has colored the different parts of the city by color and gives each city district (bydel in Norwegian) a unique symbol.

Here we see that the western part of the city (blue) generally has a considerable price variation and less hedonic variation. Note that even though this is a trend, quite a few areas in the western part are heterogeneous along both dimensions. Moreover, the city center has a substantial housing stock heterogeneity but less Gini variation. The strength of this visualization is that it allows us to assess AUs in the west with little diversity along both dimensions. However, we see that for western suburbs in general, heterogeneity along both dimensions dominates.

Figure 5: Oslo Hedonic and Price heterogeneity



6.1. Affordable housing

Price heterogeneity does not imply that the least pricey dwellings are affordable. Mansions in an affluent neighborhood can drive heterogeneity. Some of the AUs with high Gini coefficients are attractive AUs by the Oslo fiord.

A good definition of affordable housing requires some assessment of the income and wealth required to get mortgage financing. This is beyond the scope of this paper. We will instead rely on a relative measure. That is, we will consider the number of below-the-median-priced dwellings. Moreover, we will weigh the number of houses in the lower quartile double. This is just an (ad hoc) measure of available, affordable housing. We will use this to shed light on AUs that have affordable housing.[6]

Figure 6: Oslo Hedonic and Price heterogeneity

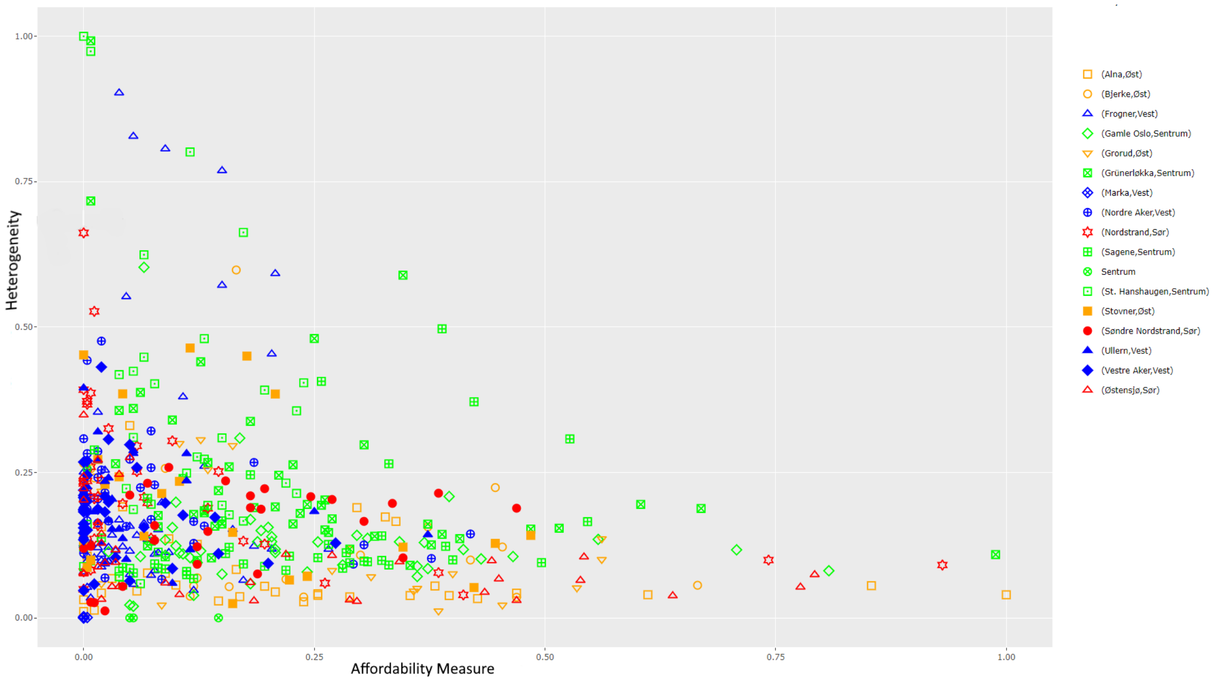


Figure 6 gives that relatively affordable housing plotted against local housing stock heterogeneity. We notice that most of the price heterogeneity observed by the Gini measure for western AUs is lost when focusing on affordable housing. This is not surprising, but the local heterogeneity within city districts is the main takeaway. We observe that AUs in the city center score high both on this relative affordability measure and housing stock heterogeneity. A potential use of such plots is to inspect AUs that score high on both measures and see if this translates to socio-demographic heterogeneity. Another interesting follow-up is to look closely at AUs that score low

[6] The following equation gives the relative affordability measure: $Afford(AU) =$ No. dwellings in 1. quartile + No. dwellings below median.

on one or both of these measures. How are these local housing markets doing? Are there signs of segregation?

From an urban planning perspective, these are the key questions. The measures discussed in this section are just a systematic and quantitative way to single out which neighborhoods may benefit from extra attention.

7. Conclusion

In this paper, we have constructed heterogeneity measures for housing markets. The main goal was to develop and illustrate a local housing stock heterogeneity measure. We relied on domain-specific feature engineering to measure the hedonic difference between any two dwellings. The hedonic measure was then used with an Epanechnikov distance to give the local housing stock heterogeneity measure. This measure satisfies several natural monotonicity properties. The most important is that additional dwellings in a local housing market can only increase heterogeneity. Moreover, the contribution is higher if it is closer to the dwelling.

The devil is in the details for a heterogeneity measure of this kind. In this respect, is the feature engineering that is diverted to the appendix the most important and domain-specific (housing market domain) academic contribution.

The latter part of this paper is devoted to price heterogeneity and its relation to local housing market stock heterogeneity. A price heterogeneity measure is at first bliss an easy task. As it is just one variable, no feature engineering is necessary. Moreover, the Gini coefficient is a well-known and much-used measure that can be applied directly.

A fundamental insight is that a measure can only be judged by considering its intended use. If the Gini measure is intended to shed light on affordable housing, we show that it may miss the mark. The reason is that the price variation tends to be high in the high-end housing market, as prices range from expensive to ultra-expensive. In statistical terms, the distribution of house prices has a fat right tail. As the left tail, the least pricey houses are curbed by 0; this asymmetry creates a "bias" towards more heterogeneity in affluent neighborhoods. We developed an ad hoc "affordability" measure, a weighted count of below-median house price transactions. This measure, combined with the hedonic heterogeneity measure, showed non-trivial variation across AUs.

The takeaway from this is that tailored approaches relying on the heterogeneity measure can help to understand and visualize spatial patterns that otherwise may be hard to unravel. On a higher level, if urban planners aim to create heterogeneous local housing markets, the first and necessary step is to define and measure local housing market heterogeneity.

This paper gives just one example of a housing stock heterogeneity measure and discusses this with house price and (house affordability) measures. We must judge a measure by its use. At best, the measures provided here would give urban planners a tool for singling out neighborhoods that may require special attention. At best, these measures will also provide a navigation tool and countermeasures in case little or loss of heterogeneity may be undertaken. One thing is sure, any policy action to prevent neighborhood degradation driven by a loss of housing stock heterogeneity requires some measure to monitor if the policy intervention is working according to plan. A famous quote by Seneca states: "If one does not know which port one is sailing, no wind is favorable." A proper housing stock heterogeneity measure may provide a navigational tool for urban planners.

References

[Die78] W Erwin Diewert. Superlative index numbers and consistency in aggregation. *Econometrica: Journal of the Econometric Society*, pages 883–900, 1978.

[FCM$^+$07] Flávia F Feitosa, Gilberto Camara, Antônio Miguel Vieira Monteiro, Thomas Koschitzki, and Marcelino PS Silva. Global and local spatial indices of urban segregation. *International Journal of Geographical Information Science*, 21(3):299–323, 2007.

[Gla11] Edward Glaeser. *Triumph of the city: How urban spaces make us human.* Pan Macmillan, 2011.

[Mar50] Stephen Marsland. *Concentration in the US Steel Industry.* Doctoral Dissertation Unpublished, Columbia University, 1950.

[Mar11] Stephen Marsland. *Machine learning: an algorithmic perspective.* Chapman and Hall/CRC, 2011.

[Ros74] Sherwin Rosen. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1):34–55, 1974.

[TO14] Joel Thibert and Giselle Andrea Osorio. Urban segregation and metropolitics in l atin a merica: The case of b ogotá, c olombia. *International journal of urban and regional research*, 38(4):1319–1343, 2014.

[VA11] Laura Vaughan and Sonia Arbaci. The challenges of understanding urban segregation. *Built Environment*, 37(2):128–138, 2011.

[Vau07] Laura Vaughan. The spatial syntax of urban segregation. *Progress in Planning*, 67(3):199–294, 2007.

8. Appendix

8.1. Feature engineering

In the data preparation, we rely on two normalizations:

Unit interval normalization

A much-used normalization is to use

$$norm(x) = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

This ensures that the normalized variable is between 0 and 1.

This measure may be adequate for some variables, for example, floors in our data set. However, due to extreme observations, much of the relevant variation may be truncated in an undesirable way. An example is dwelling size; a dwelling of 500 sqm will then dwarf the significant variation of 60 to 70 sqm. To address this concern, we adopt quartile normalization. **Quartile normalization**

The formula for quartile normalization is:

$$norm(x) = \frac{x - Q_1(x)}{Q_3(x) - Q_1(x)},$$

where Q_1 and Q_3 is the 1st and 3rd quantile respectively.

This ensures that the 1st quantile is at 0, and the 3rd is at 1. This does not address the potential problem of outliers with excessively high values that will bias our heterogeneity measure. Therefore, we cap values less than -1 at -1 and above 1 at 1. In concrete terms, in the dataset, the largest house was at 2127, it could be a mansion, and it could be a false registration, that is, it really is 212.7. Assume that the 3rd quartile is 258 sqm. This is normalized to 1, and thus the house/mansion is set to 2 (in contrast to 8.24). This corresponds to 516 sqm. In other words, we believe this house to be huge, but when it comes to the heterogeneity measure, there is no difference between a large and a huge house. This ensures that extreme hedonics, accurate or an error, do not bias the measure too much.

8.2. Preparation of the data set

Preparing the data set is challenging as we are not at liberty to exclude dwellings due to extreme values (that may be mistakes/typos) nor exclude observations due to missing variables. This calls for careful imputation and capping of extreme values to keep all observations and, at the same time, avoid those extremes bias our heterogeneity measure. Moreover, some variables only apply to some dwelling types.

Default variables for variables that only apply to a subset of dwelling types

The variable Floor is an essential dimension for flats but has no informational value for rowhouses and detached houses. (Duplexes may be horizontally divided.) This is because the default value for detached houses and rowhouses is zero.

If present, lot size refers to the lot associated with the building where the dwelling is located. These variables are important and of good data quality for houses and detached houses. The lot size in the data set is the total area for rowhouses and flats. This also includes common grounds that can not be attributed to a given dwelling. Therefore, we assign the default value zero.

Imputation of missing values for dwelling size

We will rely on the BRA number[7] for the size. The second measure PRom[8]is available for 22635 dwellings where the BRA number is missing. We regress the BRA on PRom for the entire sample (where both are non-missing). Table 1 gives the result of the regression. As expected, the PRom-coefficient is higher than 1. In concrete terms, the BRA measure is estimated to be, on average, 13 percent higher than the PRom measure. We use this model to impute BRA for (22 635) observations with PRom. For the remaining 25 895 dwellings that lack both BRA and PRom, we impute with the median BRA of the corresponding EstateType (detached, Row, duplex, and flat)

Table 1

	Dependent variable:
	BRA
PRom	1.129^{***}
	(0.001)
Constant	-6.192^{***}
	(0.065)
Observations	171,103
R^2	0.941
Adjusted R^2	0.941
Residual Std. Error	13.345 (df = 171101)
F Statistic	$2{,}748{,}482.000^{***}$ (df = 1; 171101)
Note:	*p<0.1; **p<0.05; ***p<0.01

[7] The technical definition is the total interior area. For example, for a detached one-story house without a cellar, it would be the dwelling's footprint minus exterior walls.

[8] The technical definition is the total interior area of rooms designated for human presence over time. This implies that closets, areas with low ceilings, etc., are excluded.

Table 2

dwelling type	Enebolig	Leilighet	Rekkehus	Tomannsbolig
No Observations	2, 364	203	1, 169	2, 159
Median BRA	211	66	122	164

Duplex

Floor

The floor in a duplex is usually 1 or 2 and a few times 3. Moreover, cases where the floor is not recorded, are most likely due to a similar reason as detached houses that the entry-level is the ground floor. We impute missing values different from 1,2,3 to floor=1.

Table 3

-1	0	1	2	3	4	5	NA's
8	44	1, 436	1, 521	166	11	3	12, 914

Lot size

2384 duplexes do not have lot sizes. We impute those by median lot size for duplexes in the City district in question. Moreover, lot sizes above 30 000 sqm. are most likely incorrect; we cap these at 30000. This applies to 415 duplexes.

Build year

There are 372 duplexes that do not contain build years; these are imputed by the median build year in the city district in question. (In the heterogeneity measure, the reported build year is not part of our hedonic characteristics.)

The duplex subset of observations is then feature-engineered in the following way: The variables size, build year, and lot size are quartile normalized. The floor is min-max normalized.

Detached houses

Lot size

1585 houses do not have lot size. We impute those by median lot size for houses in the City district in question. Moreover, lot sizes above 30 000 sqm. are most likely incorrect; we cap these at 30000. This applies to 37 houses.

Build year

610 houses do not contain build years; these are imputed by the median build year in the city district in question. (In the heterogeneity measure, the reported build year is not part of our hedonic characteristics.)

The variables size, build year, and lot size are quartile normalized.

Row house

Build year

418 row houses do not contain build years; these are imputed by the median build year in the city district in question. (In the heterogeneity measure, the reported build year is not part of our hedonic characteristics.)

The variables size and build year are quartile normalized.

Flats

Floor

16 517 flats do not have floor numbers. We impute this with the median floor. There are no apartment buildings higher than 20 floors in Oslo. That means that floors above this threshold need to be corrected. We impute with the median floor in these cases. This applies to 33 flats.

Build year

2000 flats do not have a building year; these are imputed by the median build year in the city district in question. (In the heterogeneity measure, the reported build year is not part of the hedonic characteristics we use.)

The variables size and build year are quartile normalized. The floor is min-max normalized.

Dwelling type

We order estate type the following way: Detached house (1), duplex (2), row house (3), and flat (1), and use this ordering to give the Estate type distance in the same way as for floor variable. That is, we use unit interval normalization.