NTNU

Norwegian University of
Science and Technology

Combining Property Price Predictions from Repeat Sales and Spatially Enhanced Hedonic Regressions

Simen Nygaard Hansen

Tobias Rosenvinge Pettrem

Industrial Economics and Technology Management
Submission date: June 2018
Supervisor: Are Oust, IØT

Norwegian University of Science and Technology
Department of Industrial Economics and Technology Management

Problem description

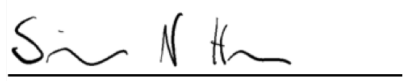
The purpose of this thesis is to study potential benefits of combining two acknowledged methods in housing analysis – hedonic regression and repeat sales – for property valuation. We enhance the classic hedonic pricing model applying three spatial models from literature as well as an ad hoc outlier-robust spatial model. Using a simple weighting approach, these estimates are subsequently combined with previous sales prices adjusted to account for the expected market growth. The investigation is conducted on the housing market of Oslo, Norway.

Preface

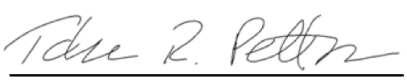
This thesis is written as part of achieving our Master of Science degrees in the field of Industrial Economics and Technology Management at the Norwegian University of Science and Technology (NTNU). The thesis is motivated by the complexity of automatic and precise valuation of residential property. It embodies original and independent work performed by Simen Nygaard Hansen and Tobias Rosenvinge Pettrém during the spring of 2018.

We would like to extend our sincere gratitude to our supervisor over the last year, Associate Professor Are Oust at NTNU Business School. His constructive guidance has been vital in the process of conducting this study, and we wish him the best of luck with his continued research and any other endeavours he may pursue. Further, we thank Henrik Langeland and the rest of Alva Technologies for providing key data and industry knowledge, and hope this article works as a contribution to their vision of creating a more functioning housing market. Finally, we would like to thank Elling Aronsen Oftedal for offering his expertise in geographic information systems.

Trondheim, June 2018



Simen Nygaard Hansen



Tobias Rosenvinge Pettrém

Abstract

Two common methods in real estate analysis are hedonic regression and repeat sales. While research has pointed out the merits of combining these models constructing house price indices, no study to our knowledge has examined this invention for property valuation. This paper investigates potential benefits of the described combination in a price prediction context, constructing house value estimates by combining predictions from repeat sales and various hedonic regression specifications enhanced to account for spatial effects. Three of these enhancements – *regression kriging*; *mixed regressive, spatial autoregressive model* and *geographically weighted regression* – are acknowledged spatial econometric models. Further, the article proposes a fourth augmentation which addresses systematic residual patterns in regressions with district indicator variables and the presence of outliers in housing data. Running the models on a data set containing 16,417 transactions in Oslo, Norway, we find that the repeat sales combination reduces the median absolute percentage error of the hedonic models with 6.8 % to 9.5 %, where larger gains are observed for less accurate spatial enhancements. We attribute the improvements to both spatial and non-spatial information inherent in previous sales prices. While the former has limited utility for well specified spatial models, we believe the non-spatial information in previous sales prices is able to capture otherwise hardly observable phenomena, making its contribution potentially highly valuable in automated valuation models.

Sammendrag

To utbredte metoder for boligprisanalyse er hedonisk regresjon og repetert salg. Flere studier har påvist nytteverdien av å kombinere disse to modellene ved boligprisindekskonstruksjon, men ingen har etter vår kjennskap undersøkt kombinasjonen for verdsetting av boliger. Denne artikkelen undersøker potensielle fordeler ved å kombinere boligprisprediksjoner fra repetert salg og ulike hedoniske regresjonsspesifikasjoner utvidet for å ta høyde for lokasjonseffekter. Tre av disse utvidelsene – *regresjonskriging*; *blandet regressiv, romlig-autoregressiv modell* og *geografisk vektet regresjon* – er anerkjente romlig-økonometriske metoder. Artikkelen introduserer dessuten en fjerde utvidelse, ment for å håndtere romlige residualmønstre forårsaket av bydelsindikatorer samt begrense påvirkning fra anomale observasjoner. Metodene ble testet på et datasett med 16 417 boligtransaksjoner i Oslo, hvor vi finner at kombinasjonen reduserer medianen av den absolutte prosentfeilen hos alle regresjonsmodellene med mellom 6,8 % og 9,5 %, hvor større forbedringer observeres for mindre nøyaktige regresjonsutvidelser. Vi argumenterer for at disse forbedringene kommer som et resultat av ulike typer iboende informasjon i repetert salgsprediksjoner, både relatert og urelatert til lokasjon. For velspesifiserte romlige utvidelser av den hedoniske regresjonen har lokasjonsinformasjonen i repetert salgsprediksjoner mindre nytte, men vi mener at den ikke-romlige informasjonen potensielt fanger opp egenskaper ved boliger som ellers er svært vanskelig å observere. Dette gjør bidraget fra repetert salgsmetoden potensielt svært verdifull i en automatisk verdsettingsmodell.

Contents

List of Figures

List of Tables

1

Introduction

A central aspect of uncertainty in housing transactions is accurate property valuation. As the acquisition of housing represents the largest investment in most people's lives, uncertainty tolerance is low, arguably underpinning the real estate agent industry and its business of human appraisal of property market value (Levin (2001)). If accurately computed, automatically generated housing value estimates represent an efficient and cost-effective alternative, potentially contributing to a more transparent housing market (Corcoran and Liu (2014)).

There is consensus in the literature that a hedonic price regression is a suitable approach when conducting house price analysis. First described by Rosen (1974) to value composite goods, this model assumes housing value equals the summarised market value of its individual characteristics. Generating precise hedonic house price predictions requires an apt representation of the main determinants of housing value: Structural characteristics, time and location. Although the two first-mentioned classes require considerate specification, proper modelling of location has proven particularly demanding in a classic hedonic regression framework. This is largely due to spatial interaction in cross-sectional housing data, introducing simultaneity and feedback effects that explicitly require spatial econometric modelling (Anselin (2010)). This has been a long-neglected fact, arguably because spatial analysis is deep-rooted in disciplines like geography and geology, far from the studies of economics (Dubin (1998)).

Another central methodology in real estate analysis is the repeat sales model, where previous sales prices of the same property are used to estimate the market development over a period of time (Bailey et al. (1963)). Similar logic suggests that up-to-date price estimates can be obtained by multiplying previous sale prices with the expected market growth. Case and Quigley (1991) demonstrated the merits of combining repeat sales with hedonic regression in the construction of house price indices, findings later confirmed and discussed by Case et al. (1991). Interestingly, no studies to our knowledge have investigated broader applicability of this combination.[1]

This paper applies the invention of Case and Quigley (1991) in a house price prediction context, hypothesising increased accuracy of a combined price estimate from the two

[1] We found no public available research on this topic, however some companies advertise automated valuation models with both models implemented, e.g. Home Value Explorer® by Freddie Mac (2017).

methodologies compared to non-combined value estimates. For our result to be robust against spatial misspesification of the hedonic model, repeat sales price predictions are combined with predictions from regression models enhanced by both traditional and state of the art spatial models from the literature, as well as an ad hoc spatial model proposed by this paper. As a result, in addition to our primary contribution on house price valuation methods, we provide further empirical evidence on spatial econometric modelling of the housing market, tying this paper to one of the most prevalent research trends in real estate valuation (Krause and Bitter (2012)).

The methods are tested on a proprietary data set of 16,417 transactions of residential property in Oslo, Norway, between August 2016 and December 2017. Repeat sales price predictions are constructed for 80 % of the dwellings in the data set, as these have data on previous sales prices. For this purpose, we apply a house price index published by Statistics Norway.

Combining regression predictions with estimates from the repeat sales model resulted in increased accuracy for all hedonic models on all metrics, even with simple combination techniques, which we mainly attribute to diversification effects (Bates and Granger (1969)). The geographically weighted regression outperformed the other spatial specifications, indicating that spatial non-stationarity is more prominent than spatial dependence in the Oslo housing market. Further, the combination resulted in higher improvements for regression models with low pre-combination accuracy. As the models differed in terms of location modelling only, we infer that repeat sales estimates contribute with, at least, some spatial information. While the value of this contribution shrinks for well specified spatial models, we argue that previous sales prices also contain a certain amount of hardly observable non-spatial information. If this assumption holds, previous sales prices could be particularly valuable in automated property valuation, as few alternatives to detect such information exist besides human inspection. The data requirements to construct repeat sales estimates are substantial, potentially reducing the practicability of our findings. Finally, the scale of the combination improvement is somewhat limited with respect to the effort required, leading us to suggest some concepts for further studies.

The rest of this article is organised as follows: Chapter 2 presents background theory on forecast combination, spatial effects and hedonic valuation. Chapter 3 introduces the housing market of Norway and Oslo. An overview of the source data is given in Chapter 4. Chapter 5 outlines the different models. The results are presented and discussed in Chapter 6, and Chapter 7 concludes the article and presents ideas for further research.

2

Theory

This chapter initially provides a theoretical backdrop for point forecast combination, emphasising notable considerations when combining estimates from repeat sales and hedonic regression. Further, we describe different spatial effects in the housing market, before outlining econometric methods for modelling location in a regression framework. Finally, two measures of spatial autocorrelation are introduced.

2.1 Combination of point forecasts

Clemen (1989) provides a summary of research on forecast combination conducted in the 20 years following the seminal paper by Bates and Granger (1969). He acknowledged their initial finding indicating that in general, the error of a linear combination of two competing point forecasts is smaller than those of the two individual predictions. He also drew a second, more counterintuitive conclusion: The simple average between forecasts tends to perform just as well as more sophisticated combination methods, further discussed by Armstrong (1989).

Several studies have investigated the reason behind the empirically proven gains of forecast combination: Granger (1989) argues that combinations providing substantial improvements likely consist of forecasts partly based on non-overlapping information sets, stressing the fundamental value of incorporating new information. Another view is given by Chan et al. (1999) who focus on the impact of outliers on a model, arguing that forecast combination could be used to identify and handle these observations.

In the context of combining hedonic regression estimates and estimates based on former sales prices in the housing market, two elements should be considered particularly important. First, house transaction data tend to include a non-negligible share of outliers, which should be addressed in the combination algorithm. Second, the high heterogeneity of dwellings implies that some price-influencing factors are unobservable and therefore omitted in the hedonic model. Such information is potentially captured in former sales prices, which would fulfil the criteria specified by Granger (1989) and imply the value of such a combination (Wallis (2011)).

2.2 Spatial effects and hedonic house price models

Two broad classes of spatial effects may be distinguished, referred to as *spatial dependence* and *spatial heterogeneity* (Anselin, 1988, p. 8). Spatial dependence, also denoted *spatial autocorrelation*, refers to the lack of independence between observations in cross-sectional data, formally defined in Cliff and Ord (1970). This effect might be caused by a variety of measurement problems as well as the presence of spatial externalities and spill-over effects ((Anselin, 1988, p 11), Griffith (1992)). Spatial heterogeneity, also denoted *spatial non-stationarity*, refers to systematic variation in the behaviour of and relationships between variables across space, implying that functional form and parameter values vary across locations (Fotheringham, 2009, p. 243).

Housing market analysis is often conducted using a hedonic house price model, regressing price on structural characteristics (e.g., size and age) and location attributes[1] (Can (1990)). Location variables are often classified as either *accessibility* variables or *neighbourhood* variables (Chica Olmo (1995)). Accessibility variables – usually a distance gradient to a central location – are meant to depict a spatial pattern, while neighbourhood variables are specified to indicate discrete differences between areas. In practice, the latter tend to be dummy variables denoting a dwelling's location by predefined submarkets, or variables such as educational level, where all dwellings in the same census area receive the same value (Fik et al. (2003)). Vital data on location attributes are in practice inaccessible, and if obtained, it is unlikely that a correct or *true* functional relationship will be specified. The consequence is spatial dependence between the error terms of closely located observations, resulting in a display of spatial patterns and autocorrelated regression residuals. Independently, a spatial non-stationary price process modelled assuming fixed regression parameters also results in autocorrelated residuals displaying a spatial pattern. As a consequence, distinguishing between spatial dependence and spatial non-stationarity based on residual characteristics is futile, a problem commonly known as the *inverse problem* (Anselin (2010)). Both types of spatial effects frequently coexist in spatial processes, also in the housing market (Anselin (2007)), and different enhancements of the regression model are recommended reliant on the root cause of the problem. As a result, identifying the more prominent spatial effect in the given housing market is of interest.

Below, we briefly outline how location attributes traditionally are incorporated in the hedonic regression, with emphasis on using location unit dummy variables, since this approach is taken in our ordinary hedonic regression specification. Further, we succinctly outline the different procedures proposed for augmenting this regression, depending on whether one assumes spatial dependence or spatial heterogeneity is the more prominent.

[1] As time also influences housing prices, auxiliary time indicators are often specified or real prices are utilised. Although impractical, modelling of time-dependent parameters is more in line with the theory of Rosen (1974).

2.2.1 Embedding spatial variation in an ordinary hedonic regression

Traditional house price models regress price on, among other variables, a set of location attributes. However, there is no clear consensus regarding which to include (Can (1990), Dubin (1992)), where different attribute types have distinct properties. Accessibility variables are hard to specify, posing questions such as which and how many central locations to include, and whether the effect of living close to a given place varies with direction. Fik et al. (2003) and Richardson (1988) argue that a predetermined, limited number of central locations inadequately represents a housing market, suggesting that the use of such variables in practice is ineffective. Neighbourhood variables represent an alternative method, corresponding to the idea that dwellings in a given area share a set of local conditions. Embedding data on externalities like crime rate is theoretically attractive, but in practice hard due to data limitations.[2] Including area indicators is more feasible, where the argument is that combined location effects are revealed as a "premium" for living in a given neighbourhood (Maser et al. (1977)). In this paper, the latter approach is taken in specifying our basic hedonic regression.

Intercept neighbourhood dummy variables are easily implementable and empirically increase prediction accuracy (Goodman and Thibodeau (2003), Helbich et al. (2013)). However, Fik et al. (2003) recognise several pitfalls associated with this approach. First, the processes determining local conditions vary over short distances, and such differences get averaged out by district dummies. Second, most externalities vary continuously over space, not displaying discrete jumps. Hence, dwellings on the edge of an area might have more in common with the adjoining neighbourhood dwellings than dwellings in their own area, resulting in large residuals close to district edges. Third, defining true submarkets is difficult (if they even exist), and in practice, administrative units are often used. If the area boundaries are misspecified, we get biased estimates and spatial autocorrelation in the residuals (Fotheringham et al. (2002)).

Several methods to identify housing submarket boundaries have been developed, e.g., hierarchical models (Goodman and Thibodeau (1998)), factor reduction through principal component analysis combined with cluster analysis (Bourassa et al. (1999)) or expanding the submarket to be defined in a multi-dimension space (Dale-Johnson (1982)). In this paper, we follow Case et al. (2004) in that *k-means clustering* is used to construct new city districts. The algorithm is a stochastic, unsupervised clustering method designed to divide a set of data points into k partitions based on multi-dimensional proximity (MacQueen et al., 1967, p 281). As a part of the algorithm, the haversine formula (Sinnott (1984)) is used to calculate distance between two longitude and latitude points. Subsequently, the *k-nearest neighbour*[3] algorithm is used to classify out-of-sample entries. First suggested by Fix and Hodges Jr (1952), it assigns unclassified points to the cluster most heavily represented among its k nearest neighbours. Together, these techniques enable dwellings to be grouped into artificial districts based on proximity in geography as well as in other variable spaces.

[2] Data tend to be non-existent or published aggregated due to privacy concerns (Kain and Quigley (1970)).
[3] The two algorithms work independently; their k values are not related and need not be equal.

2.2.2 Enhancing the hedonic regression to model spatial dependence

In the presence of spatial autocorrelation, the Gauss-Markov assumption of i.i.d. error terms in the classical regression model is violated (Brooks, 2014, p 91). To correct this behaviour, three general methods can be applied:

i **Spatial error models:** The covariance structure of the random error term is modelled. We give an example of this approach, called *regression kriging*, outlined in Section 5.2, following Dubin (1998) and Hengl et al. (2007) among others. An alternative approach to kriging is using what is called a lattice model; for more information we refer to Colwell et al. (1983) and Pace and Gilley (1997).

ii **Spatial lag models:** Rather than modelling the error covariance structure, the correlation of the dependent variable is modelled by including a function of the dependent variable observed at other locations on the right-hand side of the regression. When this autoregressive term is included, the error is assumed to be i.i.d.. We give an example of this approach, called the *mixed regressive, spatial autoregressive model*, outlined in Section 5.3, following Can and Megbolugbe (1997), Haider and Miller (2000) and Farber and Yeates (2006). We also refer to this as the *autoregressive* model.

iii **Incorporate omitted spatial variables:** The i.i.d. error assumption will begin to hold if all omitted location attributes are specified. As discussed in Subsection 2.2.1, this approach is highly challenging due to data constraints. Moreover, it requires no modification of the model specification or the estimation technique, and as a result, will not be discussed further. More details are provided in Florax and Folmer (1992).

2.2.3 Enhancing the hedonic regression to model spatial heterogeneity

A different conceptualisation of the housing price determination process is assuming the original regression is correctly specified, but a spatial non-stationary process implies space-varying parameter values. Localised models describe such relationships, with examples including the spatial expansion model (Casetti (1972)), adaptive filtering (Foster and Gorr (1986)), multilevel modelling (Congdon (1995)) and moving window regression (Farber and Yeates (2006)), all used for house price valuation. This paper implements a *geographically weighted regression* (GWR), described in Section 5.4, following (Fotheringham et al., 2002, p. 15) who argue for the superiority of this model when analysing non-stationary spatial relationships.

2.3 Indicators of spatial association

2.3.1 Moran's I

Moran's I applied on regression residuals is arguably the best known test statistic used to detect spatial autocorrelation (Fotheringham, 2009, p. 261). Defined by Moran (1950) and extended to the spatial domain by Cliff and Ord (1970), the test statistic is defined as:

$$I = \frac{N}{W}\frac{\Sigma_i\Sigma_j w_{ij}(\epsilon_i-\bar{\epsilon})(\epsilon_j-\bar{\epsilon})}{\Sigma_i(\epsilon_i-\bar{\epsilon})^2} = \frac{N}{W}\frac{\Sigma_i\Sigma_j w_{ij}\epsilon_i\epsilon_j}{\Sigma_i\epsilon_i^2}, \quad i,j=1,2,...,N; \quad W=\Sigma_i\Sigma_j w_{ij} \tag{2.1}$$

where N is the number of observations; $\epsilon_{i\ (j)}$ is the residual for observation *i (j)*; the mean of ϵ_i, $\bar{\epsilon}$, is by definition zero; w_{ij} are spatial weights, where $w_{ii}=0$, and W is the total distance used to normalise the statistic. Similar to Pearson's correlation coefficient, possible values are in the range of -1 to 1. In the spatial domain, 0 implies perfect randomness, -1 implies perfect clustering of dissimilar values (perfect dispersion), and 1 implies perfect clustering of similar values. Inference is usually conducted following a normality assumption ($\frac{I-E[I]}{\sqrt{Var[I]}} \sim N(0,1)$[4]), with the null hypothesis usually being perfect randomness.

2.3.2 Geary's C

Geary's C, also denoted *Geary's contiguity ratio*, first defined by Geary (1954) and generalised by Cliff and Ord (1970), is a statistic used for detecting spatial autocorrelation, structurally akin to the Durbin-Watson test in the time domain (Sokal and Oden (1978)). It is defined as:

$$c = \frac{(N-1)}{2W}\frac{\Sigma_i\Sigma_j w_{ij}(\epsilon_i-\epsilon_j)^2}{\Sigma_i(\epsilon_i-\hat{\epsilon})^2} = \frac{(N-1)}{2W}\frac{\Sigma_i\Sigma_j w_{ij}(\epsilon_i-\epsilon_j)^2}{\Sigma_i\epsilon_i^2}, i,j=1,2,...,N; W=\Sigma_i\Sigma_j w_{ij} \tag{2.2}$$

where the notation corresponds to the definitions specified for Equation (2.1). Possible values range from 0, implying perfect clustering of similar values, to an unspecified positive number above 1, where 1 implies perfect randomness. Similarly to Moran's I, inference is usually conducted following a normality assumption ($\frac{c-E[c]}{\sqrt{Var[c]}} \sim N(0,1)$[4]), with the null hypothesis usually being perfect randomness.

Evident from the equations above, Moran's I and Geary's C are sensitive to the choice of weighting function. In this paper, we use the distance decay function outlined in Equation (5.7). Although the two statistics are highly similar, Moran's I can be understood as a measure of global spatial autocorrelation, while Geary's C is more sensitive towards local spatial autocorrelation (Sokal and Oden (1978)).

[4] E[I] and Var[I] are defined in (Fotheringham, 2009, p.262), while E[c] and Var[c] are defined in Chen (2016).

3

Background

3.1 The property market of Norway

The Norwegian housing market has some noteworthy characteristics making it highly suitable for studies on property pricing in general. First, the sales process can be characterised as an *English auction* where the price is determined in a near perfect bidding context (Olaussen et al. (2017)). Second, most properties for sale in Norway are announced with standardised adds on the classifieds site FINN.no.[1] This facilitates comparison between dwellings and provides high-quality data for market participants. Third, Norwegians have a strong preference for home ownership as opposed to renting, with a 2016 ownership rate of 82.7 %[2] (Eurostat (2016)). As a result, the Norwegian housing market is dominated by non-professional buyers and sellers with low purchasing power, contributing to a fairly efficient housing market.

3.2 The property market of Oslo

Oslo is the capital of Norway with a 2018 population of approximately 670,000. Historically, the city has been demographically divided between east and west; industry workers were based around the river Akerselva in the central and eastern areas, while wealthier families mainly resided in western parts (Amundsen (2015)). Today, some former working class districts like Grünerløkka and Gamle Oslo are becoming increasingly popular (Faksvåg (2015)), but the historical pattern with higher prices in western areas is still evident, as shown in Figure 3.1. Figure 3.2 shows the square metre price development for dwellings in Oslo and Bærum[3] from 1993 to 2018. It indicates that prices have been highly volatile in recent years – a rationale behind our decision to draw the test sample randomly with respect to time, further discussed in Subsection 5.1.1.

[1] FINN covers approximately 70 % of the Norwegian housing market (Eiendom Norge, Eiendomsverdi and FINN.no (2017)). All properties in our data set were announced on the site.

[2] Higher than comparable figures like those of Sweden (65.2 %), Denmark (62.0 %) or EU average (69.2 %).

[3] Bærum is a suburb of Oslo, located west of the city.

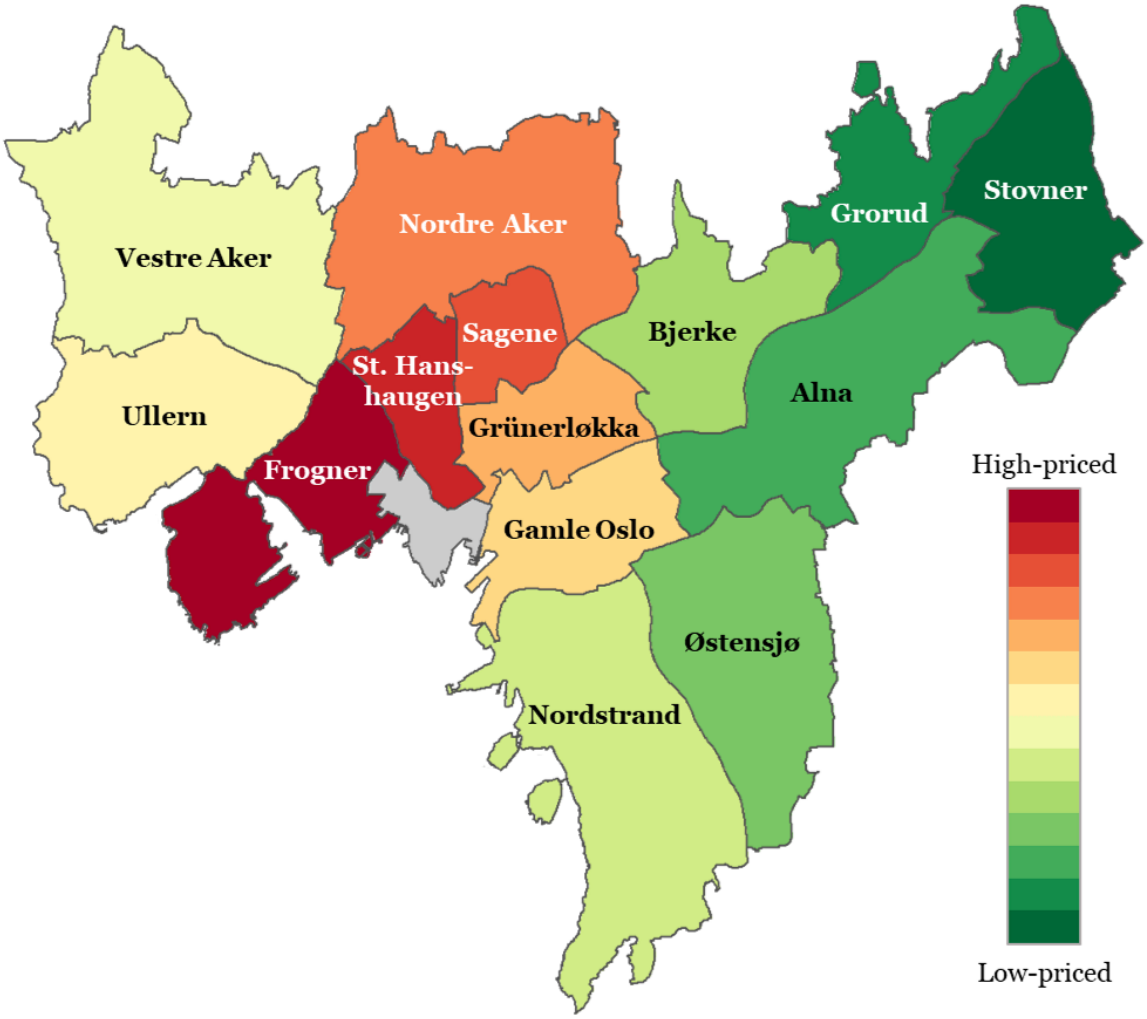


Figure 3.1: Administrative districts of Oslo with a 2017 square metre price ranking (Humberset (2018)). Data for district Sentrum (grey area) are not available; district Søndre Nordstrand is not represented in our data set and as a result excluded from this figure.

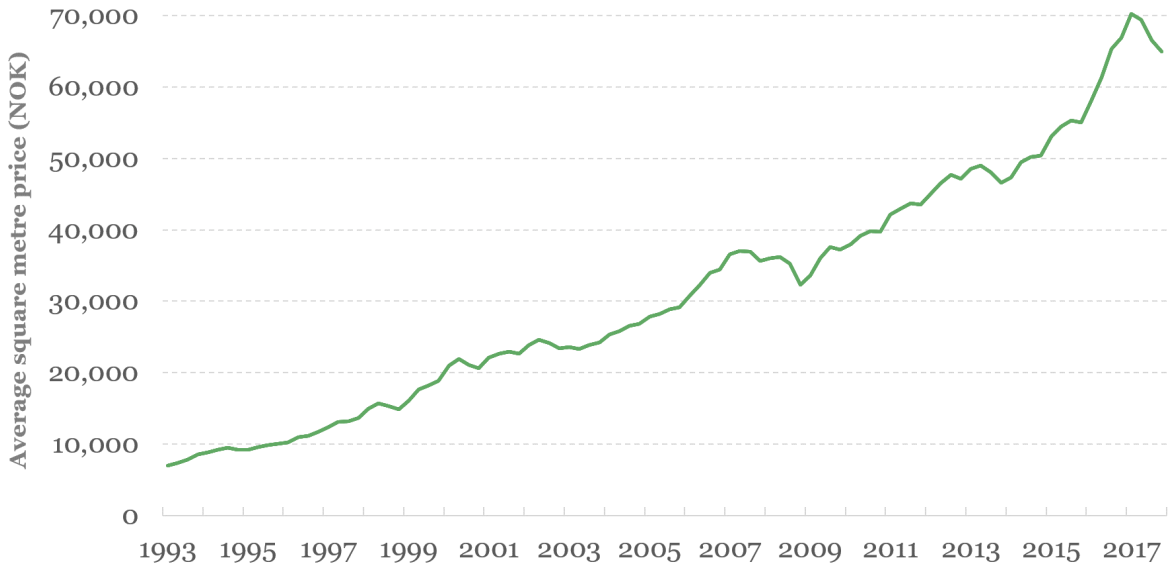


Figure 3.2: Statistic Norway's Price index for existing dwellings in Oslo and Bærum expressed in square metre price (Monsrud and Takle (2018)). 1 NOK ≈ 0.11 EUR (June 2018).

4

Data

Our models were run on a proprietary data set, which we refer to as *Virdi*,[1] provided by the firm Alva Technologies (Alva), with data on housing transactions in Oslo occurring between August 2016 and December 2017. The data set had accurate and comprehensive information on building characteristics for each transaction, including longitude and latitude data. In the data preparation process, entries in district *Marka* were discarded, dwellings from district *Sentrum* were reassigned to *St. Hanshaugen* and dwellings marked with *Other unit type* were reassigned as *Apartment*.[2] This left *Virdi* with 16,417 entries. We augmented the data set further by mapping administrative district information from Oslo Kommune (2018), adding up to 3 previous sales of the *Virdi* dwellings using a second proprietary data set received from Alva,[3] and finally retrieving additional data through corresponding FINN advertisements. An overview of the variables included in the regression models is provided in Tables 4.1 - 4.6, where all attributes are specified as indicator variables. Variables stemming from FINN are described in Table 4.6.

Parts of the information sourced from FINN was obtained through word recognition applied on the advertisement title. This implies that only characteristics highlighted by the seller/agent were obtainable, potentially causing some dwellings to lack data on attributes they in fact possess. However, these titles are fairly comprehensive; the *Virdi* dwellings have titles with almost 17 words on average, allowing the promotion of multiple property characteristics. Further, since data on price-increasing variables like *Has a garden* were retrieved, we consider this problem limited as promoting such attributes is in the seller's interest.[4]

Statistics Norway's *Price index for existing dwellings for Oslo and Bærum* (Monsrud and Takle (2018)), with data going back to 1993, was used as a proxy for the expected price appreciation of dwellings between sales in the repeat sales method. We refer to Hansen and Pettrém (2017) for an overview of house price indices available in Norway. A distribution of the number of previous sales for the *Virdi* dwellings used in the repeat sales method is provided in Table 4.7.

[1] Named after Alva's value estimate application. For more information see: `https://virdi.no`.
[2] The discarded dwellings were too remotely located to be relevant; reassignments due to sample size.
[3] An unprocessed data set with more than 300,000 historical transactions of dwellings in Oslo.
[4] Data on refurbishment need were obtained. In our experience, the title always promotes this if relevant.

Table 4.1: Construction year for dwellings in *Virdi*

Construction year	Dwellings
1820 - 1989	12,894
1990 - 2004	1,120
2005 - 2014	2,063
2015 -	340
Total	16,417

Table 4.2: House type distribution for dwellings in *Virdi*

House type	Dwellings
Apartment	14,592
Semi-detached house	367
Detached house	385
Serial house	1,073
Total	16,417

Table 4.3: District distribution for dwellings in *Virdi*[5]

District	Dwellings
Alna	1,317
Bjerke	719
Frogner	1,616
Gamle Oslo	1,647
Grorud	771
Grünerløkka	2,079
Nordre Aker	816
Nordstrand	1,038
Sagene	1,949
St. Hanshaugen	1,193
Stovner	556
Ullern	641
Vestre Aker	725
Østensjø	1,350
Total	16,417

Table 4.4: Sales month for dwellings in *Virdi*

Time of sale	Dwellings
August 2016	92
September 2016	1,036
October 2016	1,234
November 2016	1,094
December 2016	545
January 2017	729
February 2017	1,112
March 2017	1,189
April 2017	724
May 2017	1,283
June 2017	1,478
July 2017	863
August 2017	1,112
September 2017	1,098
October 2017	1,184
November 2017	1,140
December 2017	504
Total	16,417

Table 4.5: Size distribution for dwellings in *Virdi*

Dwelling Size	Dwellings
10 - 29 m^2	578
30 - 39 m^2	1,513
40 - 49 m^2	1,933
50 - 59 m^2	2,886
60 - 69 m^2	3,277
70 - 79 m^2	1,901
80 - 89 m^2	1,303
90 - 99 m^2	757
100 - 109 m^2	578
110 - 119 m^2	348
120 - 129 m^2	292
130 - 139 m^2	203
140 - 149 m^2	165
150 - 179 m^2	328
Above 180 m^2	355
Total	16,417

[5]The district *Søndre Nordstrand* was excluded from the data set by Alva.

Table 4.6: Variables retrieved from FINN advertisements for dwellings in *Virdi*

	Dwellings	% of total
High monthly shared cost	1,642	10.0 %
Two bedrooms & size < 60 m^2	932	5.7 %
Three bedrooms & size < 85 m^2	835	5.1 %
Housing cooperative	8,615	52.5 %
Needs refurbishment	1,158	7.1 %
Is a penthouse	2,299	18.2 %
Has a garden	1,659	10.1 %
Has a terrace	1,139	6.9 %

High monthly shared cost is defined as having the top 10 % shared cost of dwellings in *Virdi*, the threshold being NOK 4,713 per month; *Two bedrooms & size* < 60 m^2 is the number of dwellings being smaller than 60 m^2 and having exactly two bedrooms; *Three bedrooms & size* < 85 m^2 is the number of dwellings being smaller than 85 m^2 and having exactly three bedrooms; *Housing cooperative* is whether the dwelling is part of a housing cooperative. The self-explanatory variables *Needs refurbishment*, *Is a penthouse*, *Has a garden* and *Has a terrace* were retrieved by word recognition.

Table 4.7: Previous sales count for dwellings in *Virdi*

Number of sales	Dwellings	% of total
No previous sales	3,279	20.0 %
One previous sale	5,631	34.3 %
Two previous sales	4,109	25.0 %
Three previous sales	3,398	20.7 %
Total	16,417	100.0 %

5

Methodology

We first outline an ordinary hedonic regression model with emphasis on intercept area dummies constructed using the k-means and k-nearest neighbour algorithms. We then move on to describe the following four extensions to the basic regression model:

1. **Regression kriging:** A spatial error model assuming *spatial dependence* in the error terms.
2. **Mixed regressive, spatial autoregressive model:** A spatial lag model assuming *spatial dependence* in the dependent variable.
3. **Geographically weighted regression:** A local regression model assuming *spatial non-stationarity*.
4. **Vicinity-based residual tuning:** A simple, outlier-robust procedure adjusting hedonic regression predictions based on residuals of nearby dwellings.

Finally, we outline how estimates from a repeat sales model are combined with the hedonic regression estimates. Our model scheme is plotted in Figure 5.1 below, where the geographically weighted regression has a dashed line since district variables must be omitted in this model.

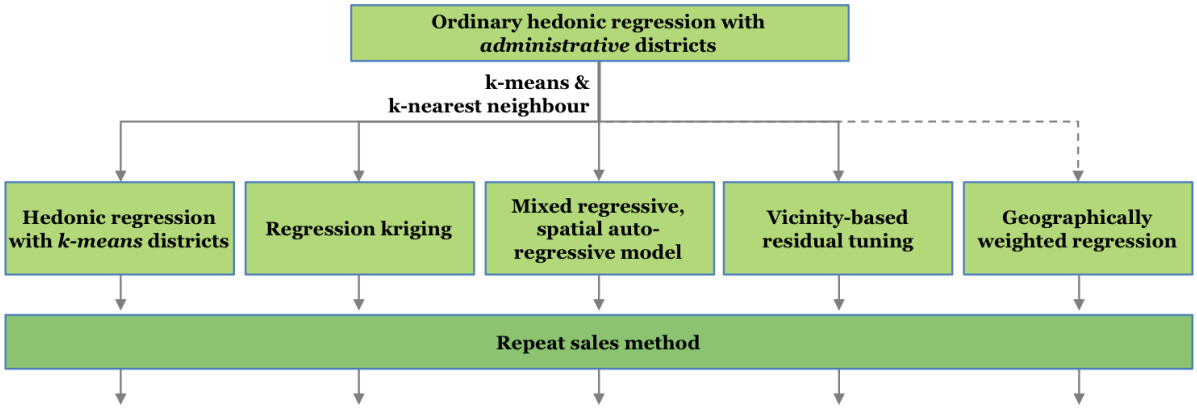


Figure 5.1: Overview of spatial models and extensions used in this paper. The dashed line indicates that district indicator variables cannot be specified in the GWR model.

5.1 Basic hedonic regression model

The hedonic regression model was first introduced by Rosen (1974). It is a commonly used model in property valuation, based on the assumption that the value of a property equals the summarised market value of its parts. In our model, the value of a given dwelling is represented by its common debt at sales date plus sales price divided by house area in m^2, formally:

$$P_i = \frac{sales\ price_i + common\ debt_i}{house\ area_i}. \tag{5.1}$$

The natural logarithm of P_i from Equation (5.1) is estimated by evaluating contributions to the price by each utility-bearing attribute using multiple linear regression. The general equation to be estimated is written

$$ln(P_i) = \beta_0 + \sum_k \beta_k X_{ki} + \sum_n \delta_n D_{ni} + \epsilon_i, \quad \epsilon \sim i.i.d. \tag{5.2}$$

where P is the price variable as defined in Equation (5.1); X_k is a set of explanatory variables[1] describing a presence of utility-bearing characteristic k; D_n is a set of n area indicator variables; ϵ is the error term, and β_0, β_k and δ_n are the parameters to be estimated with estimates denoted $\hat{\beta}_0$, $\hat{\beta}_k$ and $\hat{\delta}_n$. Our data only span 17 months and cover a single city, so we follow the common assumption of parameter vectors invariant across space and time (de Haan and Diewert (2013)). Equation (5.2) is estimated using *least absolute deviation* (LAD) following Koenker and Bassett Jr (1978), as LAD is more robust towards outliers than ordinary least squares (OLS) as well as other estimators based on distributional assumptions (Yoo (2001)). An overview of the structural explanatory variables used in Equation (5.2) is found in Tables 4.1 - 4.6. To incorporate spatial and temporal variability in Equation (5.2), we use intercept indicator variables, further discussed in the following subsections.

5.1.1 Isolating time – indicator variables and nonchronological sampling

House prices are volatile, subject to seasonality effects, and generally substantially influenced by time (Reichert (1990)). However, the focus of this article is modelling spatial effects, not the temporal dimension. To get results unbiased from the market price development, we isolate this effect by including monthly time dummies as explanatory variables in all regression models. Furthermore, our test sample is constructed by randomly drawing 20 % of the observations from the whole sample. This implies that the two samples we use for estimating and testing the models span the same time horizon, as opposed to a sequential sampling.

[1] Includes both building characteristics and dummy variables for time.

5.1.2 Constructing district boundaries with k-means

The k-means algorithm is applied to our training data sample to construct artificial market districts with more homogeneous property pricing processes while retaining cohesiveness. Distance between dwellings is measured as a function of longitude, latitude and price, the latter as defined in Equation (5.1). After clustering the training set, k-nearest neighbour is used to classify dwellings in the test sample based on the newly constructed districts, measuring distance in classical, geographic sense using the haversine formula. An algorithmic outline of the methods is provided in Appendix A.2. Empirical trial gave best results for values of k in k-means between 14 and 20, and was set to 18 in the final model as this provided the most stable results. An illustrative plot comparing a k-means clustering with k = 14 and administrative borders is shown in Figure 6.1. The k in k-nearest neighbour was set to 3, based on empirical trial as well as visual inspection of district shapes produced by k-means.

5.2 Regression kriging

As argued by Dubin (1988) and Basu and Thibodeau (1998) among others, spatial dependence in the housing price process can be modelled by assuming the original functional relationship (Equation (5.2)) holds, while abandoning the assumption of the error term being i.i.d. This requires the modelling of the error covariance structure. Adopting this approach for prediction builds on the statistical interpolation technique kriging.[2] Following the previously outlined notation, Equation (5.2) is reformulated into

$$ln(P_i) = \beta_0 + \sum_k \beta_k X_{ki} + \sum_n \delta_n D_{ni} + \epsilon_i, \qquad \epsilon \sim N(0, \sigma^2 \mathbf{C}) \tag{5.3}$$

where $\mathbf{C}$ is the error correlation matrix. To estimate Equation (5.3), a functional form for the error term's covariance structure must be assumed. The parameters of this function, along with the normal regression coefficients, are simultaneously estimated using maximum likelihood.[3] The estimation of (5.3) can become very complex in the presence of nonlinear explanatory variables (Hengl et al. (2007)), and in practice, parameter instability is a major concern (Goovaerts (1999)).

[2] When auxiliary variables (here X and D) are included, the terms *kriging with external drift* (KED) or *universal kriging* (UK) are used. Often, the latter term is limited to the case where the auxiliary variables are coordinates only.

[3] We refer to Dubin (1988) for a definition of the maximum likelihood function.

To mitigate this, the estimation can be divided into two parts. First, the linear regression parameters β_0, β_k and δ_n are estimated using a less complex estimator, in our case, LAD. Then, the error covariance function parameters are estimated by *simple kriging*[4] with zero mean on the residuals from the first regression. The prediction is finally calculated by adding the fitted residual from the simple kriging model to the fitted value from the linear regression. In mathematical terms, the predicted value for dwelling i in the test sample having structural characteristics X' and D' is given by

$$ln(\hat{P}_i) = \hat{\beta}_{0,LAD} + \sum_k \hat{\beta}_{k,LAD} X'_{ki} + \sum_n \hat{\delta}_{n,LAD} D'_{ni} \quad + \quad \sum_j w_{ij} \hat{\epsilon}_{j,LAD}, \qquad j \neq i \tag{5.4}$$

where $\hat{\epsilon}$ are the LAD residuals from the training sample, and w_{ij} are elements in the weight matrix $\mathbf{W}$, determined by an assumed covariance function. This two-step procedure is called *regression kriging*, first named by Odeh et al. (1995). Predictions from (5.4) and predictions made from directly estimating (5.3) are mathematically equivalent; as long as the assumed covariance function is identical, the difference lies in the computational steps only (Hengl et al. (2003)).

Several structural covariance functions are applicable in kriging, with the common feature that correlation between observations decreases with increased physical distance. We assume that the error covariance follows the negative exponential form, following Case et al. (2004)

$$c_{ij} = b_1 + e^{-\frac{d_{ij}}{b_2}}, \qquad j = 1,2,3,...,100; \;\; j \neq i$$

where the parameters b_1 and b_2 are estimated in the second step of the regression kriging procedure outlined above; d_{ij} are euclidean distances between dwelling i and dwelling j, and c_{ij} are entries in the $\mathbf{C}$-matrix from Equation (5.3). To calculate the weights based on the covariance matrix, we use the relationship $\mathbf{W} = \mathbf{C}^{-1}\mathbf{c}$, where $\mathbf{c}$ is a vector of covariances between the training data points and the estimation point (Bohling (2005)). For computational reasons, we limit the maximum number of neighbours taken into account to 100 for each dwelling.

We give a final remark regarding our use of LAD. Generalised least squares (GLS) is recommended as the proper estimator in the first step of regression kriging, to account for spatial autocorrelation in the error term (Cressie (1990)). Despite this, Kitanidis (1993) shows that the difference between several iterations of GLS and a single iteration (OLS) is too small to have any notable effect on the final result. We tried using both GLS and LAD and observed just marginal differences in line with Kitanidis (1993), and chose to use LAD to get a consistent choice of estimator across all models evaluated in this paper.

[4]The term simple kriging is used when the mean of the dependent variable is assumed known (Cressie (1990)).

5.3 Mixed regressive, spatial autoregressive model

As argued by Can (1992), spatial dependence in the housing price determination process can be modelled by including a function of the dependent variable as an autoregressive term in the standard hedonic regression (Equation (5.2)). Using the specification by (Fotheringham, 2009, p. 257) the model can be expressed as

$$ln(P_i) = \beta_0 + \rho \sum_j w_{ij} ln(P_j) + \sum_k \beta_k X_{ki} + \sum_n \delta_n D_{ni} + \epsilon_i, \qquad j \neq i \tag{5.5}$$

where ρ is a measure of the overall level of spatial dependence among $(ln(P_i), ln(P_j))$ pairs for which $w_{ij} > 0$, and w_{ij} are spatial weights we give to the sales price of dwelling j. Other variables are as described in Section 5.1. Including the dependent variable on the right-hand side induces *simultaneity*,[5] so estimating Equation (5.5) with OLS or LAD returns biased estimates. However, this is commonly done, as appropriate estimation using maximum likelihood is a challenge (Farber and Yeates (2006)). A different solution following Can and Megbolugbe (1997) is to include an additional constraint, rephrasing Equation (5.5) into

$$ln(P_{it}) = \beta_0 + \rho \sum_j w_{ij} ln(P_{j,t-m}) + \sum_k \beta_k X_{ki} + \sum_n \delta_n D_{ni} + \epsilon_{it}, \quad m = 1,2,...; \quad j \neq i. \tag{5.6}$$

The distinction between Equation (5.5) and (5.6) is that the dependent variables in the latter are determined at time t and are hence exogenous, resulting in OLS and LAD being unbiased estimators. We define the weighting function, again following Can and Megbolugbe (1997):

$$w_{ij} = \begin{cases} \frac{1/d_{ij}}{\Sigma_j (1/d_{ij})} & : d_{ij} < 1.5\ km \\ 0 & : d_{ij} \geq 1.5\ km \end{cases} \qquad j = 1,2,3,...,15; \quad j \neq i \tag{5.7}$$

where d_{ij} are euclidean distances between dwelling i and dwelling j, j representing the 15 dwellings located closest to dwelling i, and with earlier sales dates than dwelling i. In the special case where two dwellings share the same location, we set $d_{ij} = 10$ metres so Equation (5.7) is defined for all observations.[6]

A final remark is given about remotely located dwellings and dwellings with the oldest transactions of the sample, having no relevant neighbours available to define their autoregressive term. To retain the same number of observations for all models considered in this paper, we set the autoregressive term of these dwellings equal to the average log price for the relevant district, with price defined as in Equation (5.1).[7]

[5] For definitions of simultaneity and simultaneous equations bias we refer to (Brooks, 2014, p. 308).
[6] Mainly for academic interest, as it accounts for less than 0.5 % of the sample.
[7] Mainly for academic interest, as it accounts for less than 1.0 % of the sample.

5.4 Geographically weighted regression

As argued by Wheeler and Calder (2007), the housing price process is non-stationary over space and the coefficients in the traditional hedonic regression represent the global "average" only. As a result, accurate predictions must stem from an enhanced regression model where the parameters are allowed to vary over space (Yao and Fotheringham (2016)). The geographically weighted regression method enables such a local parameter estimation. We follow the notation of (Fotheringham et al., 2002, p. 52) and rephrase the traditional regression framework into

$$ln(P_i) = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) X_{ki} + \epsilon_i \tag{5.8}$$

where u_i and v_i denote the coordinates of the ith point in space, and $\beta_k(u_i, v_i)$ is a realisation of the continuous function $\beta_k(u, v)$ at point i. Note that the location area indicator variable D from Equation (5.2) is omitted in Equation (5.8). The equation above contains more unknown than observed variables, so at point i, Equation (5.8) is approximated by:

$$ln(P_i) = \beta_0 + \sum_k \beta_k X_{ki} + \epsilon_i. \tag{5.9}$$

The parameters β_0 and β_k are independently estimated for all i locations with dwellings in the test sample. Estimation is conducted by weighting the observations in accordance with their proximity to location i, and the parameters are chosen to minimise the weighted sum of squared residuals. We follow Fotheringham et al. (2002) and estimate Equation (5.9) with weights calculated using a Gaussian kernel function:

$$w_{ij} = e^{-0.5\frac{d_{ij}}{b}^2}, \quad i \neq j \tag{5.10}$$

where d_{ij} are the euclidean distances between point i and j; b is referred to as bandwidth and chosen by a cross-validation optimisation approach following Cleveland (1979). Practically speaking, this implies that just a small subset of the observations in the training sample is used to estimate Equation (5.9) at the different points i, making the estimate for a given dwelling vulnerable to anomalies in the data of nearby located dwellings.

5.5 Vicinity-based residual tuning

We introduce an automated variant of a valuation method commonly used by real estate agents where a limited number of recently sold properties in the immediate neighbourhood (usually 3 - 6) are used to provide a house value estimate (Can and Megbolugbe (1997), Pace et al. (2000)). The procedure outlined here utilises the fact that differences between properties are already controlled for in the residuals of a hedonic regression. Pitfalls from including district intercept dummies in a regression, as outlined in Fik et al. (2003), are also addressed. The method is referred to as *Vicinity-based residual tuning* or *VRT*.

5.5.1 Procedure

Obtain fitted values for dwellings in the test set by using regression coefficients estimated on the training set. Then, for each dwelling in the test set, with sales date denoted τ:

i) Identify up to the κ closest neighbours from the training set sold before time τ, located within the same district[8] and within a radius of maximum μ metres.

ii) Extract the residuals of the neighbours and calculate their median. This median residual is multiplied by a deflation factor α, and another deflation factor β if the number of neighbours is below λ. Finally, add this residual to the fitted value to obtain the VRT estimate.

Table 5.1: Parameter values, VRT method

κ	μ	α	β	λ
6	150	0.7	0.5	3

5.5.2 Reasoning behind parameter values and neighbour constraints

Specifying area intercept dummies in a hedonic regression often results in low prediction accuracy close to district borders, where residuals with different magnitudes and signs are clustered on each side of the borders. Figure A.2 provides an example of such effects. To address this, we include the district constraint in i). Further, an outlier with extreme residual value included as a neighbour can have a severe impact on the model accuracy. We limit this effect by using the median[9] and including the (λ, β) clause in ii), where $\lambda = 3$ corresponds to the lowest number of neighbours where the smallest and largest neighbour residual value are discarded in the calculation of the median. The rest of the parameters are set by the following reasoning: μ is chosen intuitively, α and β are set based on empirical trial and the selection of κ follows Can and Megbolugbe (1997).

[8] Either administrative or generated by k-means, depending on what variable the regression uses.
[9] As opposed to, e.g., regular or distance weighted average usually proposed in the literature (Dubin (1998)).

5.6 Constructing and combining repeat sales predictions

Hedonic house price models have a shortcoming stemming from the high heterogeneity of dwellings, making the inclusion of all price-influencing attributes infeasible (Case et al. (1991)). By using the concept of repeat sales analysis, we try to capture such hardly observable effects through the former sales prices of a given dwelling. The model assumes that prices have developed in line with the overall market, as described by a house price index, implying that the quality of each dwelling is assumed comparable at the transaction times. As outlined in Chapter 4, Statistic Norway's *Price index for existing dwellings for Oslo and Bærum* with quarterly data going back to 1993 is used in this paper. We consider a maximum of three previous transactions for each dwelling, where the most recent transactions are retained, excluding all sales from before 1993 as a result of the index span.

The premise that a dwelling's quality is highly similar at different transaction times is a vulnerable assumption at best. If previous sales conditions are unrepresentative for the dwelling's condition at resale, the repeat sales estimate is likely to mispredict gravely. To remove such outliers, all repeat sales estimates deviating more than 25 % from the regression estimate with which they are to be combined, are discarded, following Chan et al. (1999) and OECD et al. (2013). To obtain one final prediction, the remaining estimates are combined following the procedure outlined in detail in Appendix A.1; in short, the weight given to the hedonic regression estimate will amount to at least 60 %, and heavier weighting is given to predictions based on more recent previous sales than earlier transactions. Following Clemen (1989), only simple linear combination techniques are used.

6

Results and discussion

The performance[1] of an ordinary hedonic regression without any location attributes is displayed in the top row of Table 6.1. This model includes no spatial information whatsoever, and consequently represents a benchmark for all our enhancements addressing the spatial aspect. A comparison of the results confirms the strong influence of location on housing value: The most basic effort to model location – adding administrative district indicator variables (row 2) – reduces the median error from 12.1 % to 8.05 %, an improvement of 33.5 %. Interestingly, augmenting the benchmark model with either regression kriging (row 5) or mixed regressive, spatial autoregressive model (row 9) yields similar improvements – from 12.1 % to 8.18 % and 7.70 %, respectively. We deduce that district intercept dummies relatively accurately incorporate the effect of location, although several methods can tackle this matter. The extensive use of indicator variables is likely driven by the intuitive interpretation of the parameters, as well as ease of implementation. However, the use of such variables, particularly when based on administrative districts, disregards intradistrict variation and tends to result in irregular residual patterns close to borders, as discussed in Subsection 2.2.1.[2]

Statistically generated districts can reduce these issues. A comparison of the administrative and a k-means based division of Oslo is visualised in Figure 6.1.[3] An interesting case is the administrative district Alna, where k-means classifies the dwellings into four different districts, indicating large internal price differences. We limit ourselves to highlighting this example and encourage examining Figure 6.1 in light of Figures 3.1, A.1 and A.2. Improved performance from using k-means districts is evident comparing row 2 and 3 in Table 6.1; the median absolute percentage error improves from 8.05 % to 7.67 %, and the values of Moran's I and Geary's C indicate reduced spatial autocorrelation in the residuals. As stated in Subsection 5.1.2, k-means is set to divide the city into a higher number of districts (18) than the administrative division (14), due to more stable performance. Results based on different values of k are shown in Appendix A.3, supporting the algorithm's conceptual advantages, as the improvement from implementing k-means with 14 districts is relatively high compared to the improvement stemming from finer district fragmentation.

[1] Generally, we measure model performance by *median absolute percentage error* ($Q_{0.5}$).
[2] The resulting residual pattern from using administrative borders is plotted in Figure A.2.
[3] As k-means works independently of administrative districts, any area similarities are coincidental.

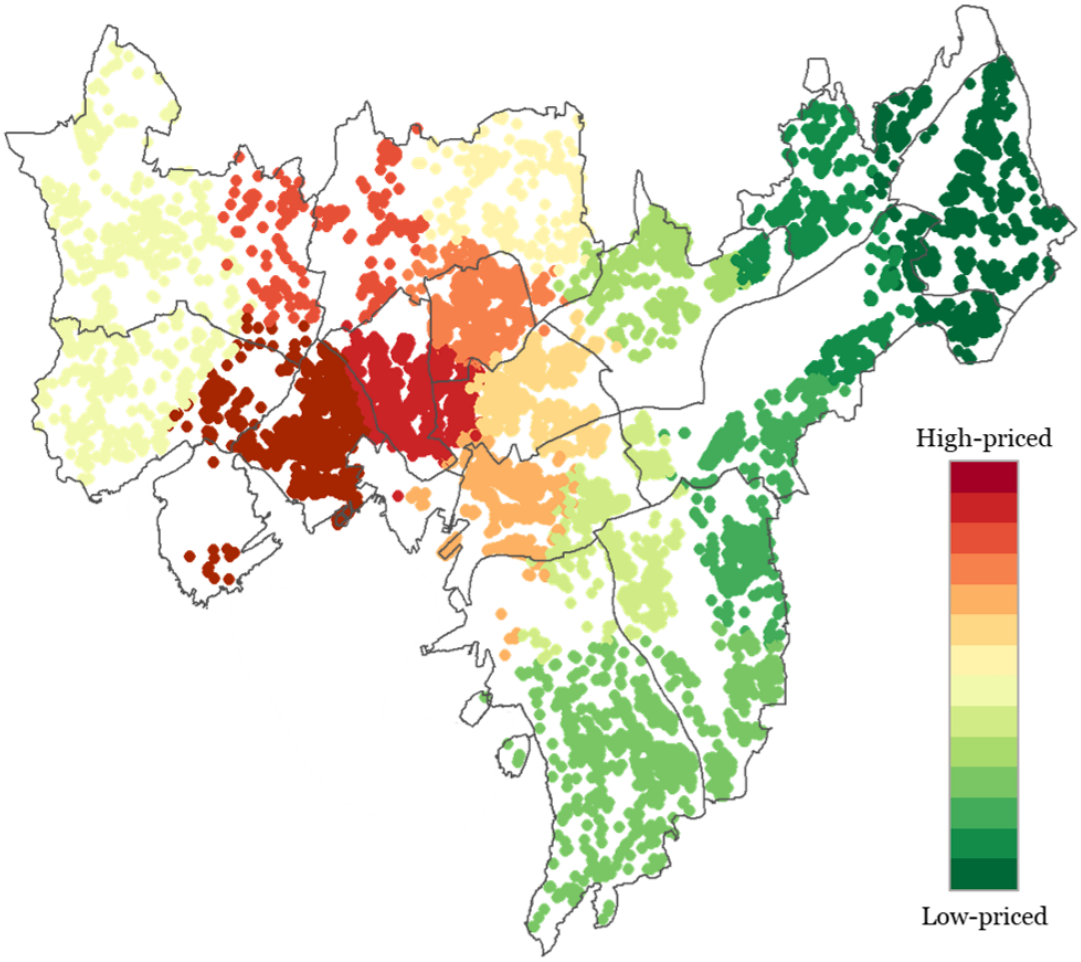


Figure 6.1: Comparison of administrative districts (lines) and statistically made districts constructed using k-means (coloured markers) for the city of Oslo. Final k-means models use k = 18, but this figure shows a partition using k = 14 for easier visual inspection of algorithm functioning. Colour gradient indicates the average square metre price for each k-means district, directly comparable with Figure 3.1. Number of observations is 13,133.

Table 6.1: Model performances

Model	Admin district	K-means	Repeat sales	$Q_{0.25}$	$\mathbf{Q_{0.5}}$	$Q_{0.75}$	Within 10 %	Moran's I	Geary's C	Row no.
Ordinary regression	☐	☐	☐	5.69%	**12.1**%	21.3%	42.4%	0.524	0.449	1
	☑	☐	☐	3.74%	**8.05**%	14.1%	59.6%	0.143	0.826	2
	☐	☑	☐	3.55%	**7.67**%	13.5%	61.6%	0.107	0.860	3
	☐	☑	☑	3.25%	**6.94**%	12.3%	66.0%	0.099	0.867	4
Regression kriging	☐	☐	☐	3.72%	**8.18**%	14.6%	58.5%	0.045	0.940	5
	☑	☐	☐	3.55%	**7.72**%	13.9%	61.1%	0.038	0.949	6
	☐	☑	☐	3.53%	**7.72**%	14.0%	61.2%	0.037	0.950	7
	☐	☑	☑	3.23%	**7.00**%	12.5%	65.3%	0.036	0.950	8
Auto-regressive	☐	☐	☐	3.52%	**7.70**%	13.7%	61.0%	0.086	0.889	9
	☑	☐	☐	3.40%	**7.31**%	13.0%	63.7%	0.068	0.902	10
	☐	☑	☐	3.39%	**7.28**%	12.9%	63.6%	0.065	0.906	11
	☐	☑	☑	3.18%	**6.77**%	11.8%	67.4%	0.063	0.906	12
VRT	☐	☐	☐	4.03%	**8.76**%	15.7%	55.6%	0.204	0.761	13
	☑	☐	☐	3.26%	**7.12**%	12.7%	64.6%	0.049	0.918	14
	☐	☑	☐	3.25%	**7.06**%	12.6%	64.8%	0.044	0.926	15
	☐	☑	☑	3.01%	**6.54**%	11.6%	68.2%	0.051	0.919	16
GWR			☐	3.14%	**6.65**%	11.7%	68.1%	0.072	0.896	17
			☑	2.93%	**6.20**%	11.0%	71.1%	0.059	0.909	18

Model refers to the methods outlined in Section 5.1 - 5.5; *Admin district* and *K-means* indicate if the boundaries for the area dummy variables are administrative districts or generated by k-means, respectively (irrelevant for the GWR model); *Repeat sales* indicates whether the results are obtained after combining with repeat sales predictions; $Q_{0.25}$, $Q_{0.5}$ and $Q_{0.75}$ denote the first, second and third quartile of the errors, respectively, where $Q_{0.5}$ is boldfaced for emphasis; *Within 10 %* specifies the fraction of errors below 10 %; *Moran's I* and *Geary's C* are calculated as detailed in Section 2.3, and *Row no.* is row number provided for convenience when referring to this table.
The results shown are average values from 10 runs for each implementation. The number of observations used for model training is 13,133, while the number of observations out-of-sample is 3,284.

District indicators are insufficient to appropriately model refined spatial patterns. We first consider the performance of the global augmentations regression kriging and mixed regressive, spatial autoregressive model. Without district variables, both models display improved prediction accuracy compared to the benchmark model, as already pointed out. With district variables, accuracy increases further, although not substantially. A less intuitive result is that the two spatial models seem indifferent to the choice of district representation,[4] in contrast to the clear advantage of applying k-means to the ordinary regression. We suggest two possible explanations. First, the influence of district dummy variables shrinks when location is concurrently modelled by several methods.[5] Second, the two enhancements correct some spatial abnormalities caused by the administrative district, reducing the need for k-means. A final observation is that the autoregressive model outperforms regression kriging. No clear trend in similar research concurs with this finding, nor did applying different weighting functions in the regression kriging model affect the result (in line with LeSage and Pace (2014)). However, while this might reduce the credibility of our kriging implementation, the effect of combining these predictions with repeat sales estimates (discussed later in this section) coincides with the remaining spatial models.

The VRT model performs second best among the spatial enhancements (rows 14 and 15 in Table 6.1). To explain these results, we refer to the arguments by Chan et al. (1999) about the severe impact from outliers on most models, given that VRT is constructed to be more outlier-robust. We also note that the model only performs well for specifications including district variables[6] and attribute this behaviour to VRT being unable to distinguish more district-wide trends when only considering a very limited number of neighbours. As opposed to adjusting the model to capture such trends, we stress that it is intrinsically tailored to address spatial residual patterns emerging from the use of intercept dummy variables in a regression. As a result, the method probably has limited use in general forecasting, but proves to be highly effective in this specific context. VRT also seems quite indifferent to the choice of district representation, most likely by the same reasons suggested for regression kriging and the autoregressive model.

The geographically weighted regression emerges as the most precise spatial enhancement (row 17 in Table 6.1). Since this model assumes and addresses spatial non-stationarity, the large improvement strongly suggests that this is the more prominent spatial effect in the Oslo housing market. The fact that GWR seems to outperform other spatial models for out-of-sample predictions corresponds with the findings of Farber and Yeates (2006) and Páez et al. (2008), but contrasts Harris et al. (2010) and Harris et al. (2011), recommending universal kriging. Although GWR tends to provide precise predictions, it has received criticism based on its limited value for making inferences. Furthermore, the method is sensitive to outliers on a local level, particularly problematic in housing valuation where outliers pose a permanent challenge.

[4] Comparing row 6 with row 7, and row 10 with row 11 in Table 6.1.
[5] Observable by comparing the absolute values of the location dummy parameters in Tables A.4 and A.5.
[6] Row 13 in Table 6.1 shows unsatisfactory performance by VRT where district variables are omitted.

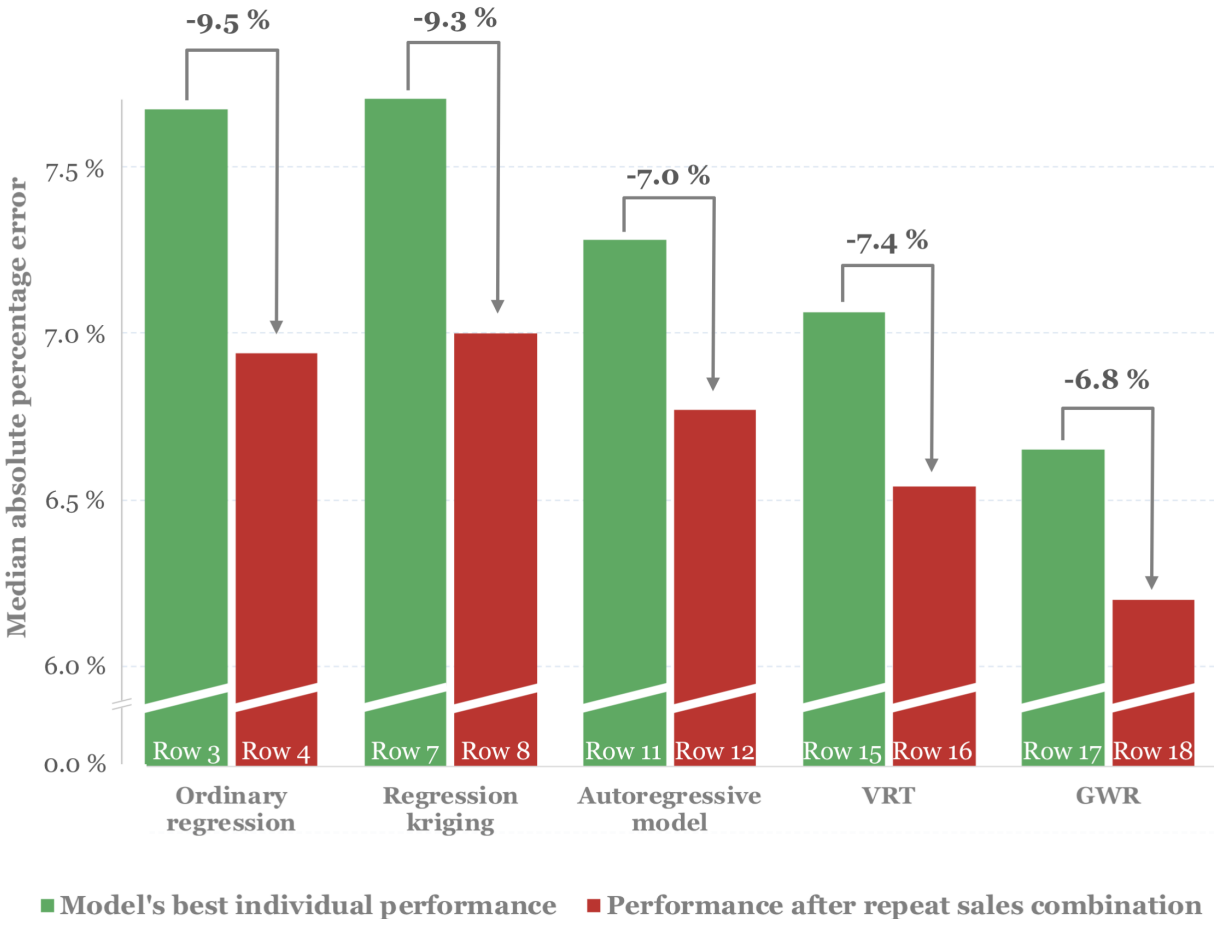


Figure 6.2: Visualisation of improved performance by combining repeat sales predictions with hedonic regression predictions. The bold number above the arrows indicates the reduction in median absolute percentage error in percentage terms. The row number at the bottom of each column indicates the corresponding row in Table 6.1.

The gain from combining repeat sales predictions with hedonic regression forecasts is evident in Table 6.1 and further emphasised in Figure 6.2, where the median absolute error achieved by the different regression models are plotted pre- and post-combination.[7] In fact, the table reveals that the combination improves accuracy by every metric and for every variation of the hedonic regression.[8] To support diversification as the main driver of the improvement (as argued by Bates and Granger (1969)), as opposed to a deterioration of highly accurate repeat sales predictions, independent repeat sales results are provided in Appendix A.4. This table makes it evident that uncombined repeat sales predictions perform worse than all regression models used for combination, supporting the diversification argument. Also note the considerable effect of outlier removal on the repeat sales estimates, evident by comparing the two rows in Appendix A.4 – arguably a necessity in order to replicate the level of improvement from the repeat sales/hedonic regression combination.

[7] Equivalent to comparing rows 4, 8, 12, 16 and 18 with their prior in Table 6.1.
[8] Moran's I and Geary's C do not describe model accuracy.

Apart from overall increased accuracy, Figure 6.2 shows that improvements from the repeat sales combination vary between the regression models, where a more substantial effect is observed when the initial regression error is large. Considering the fact that the regression models only differ in terms of location modelling, we infer that repeat sales contribute with at least some spatial information, with diminished value for more sophisticated spatial models. Arguably, location is fairly well modelled in the autoregressive, VRT and GWR models, where the repeat sales combination resulted in similar improvements of 0.51, 0.52 and 0.45 percentage points, respectively. Consequently, we strongly suppose that the predominant part of these improvements stems from the incorporation of non-spatial information omitted from the hedonic regression. Although not verifiable, this argument is supported by the inherent heterogeneity of dwellings, making inclusion of all price-influencing attributes infeasible in a regression framework (de Haan and Diewert (2013)).

Based on this, we argue that previous sales prices can provide specific value in two ways. Most importantly, they have the ability to incorporate information on hardly observable attributes. This could have a pivotal value in automated property valuation, as there are few alternatives to detect such information besides human inspection. Second, they enable the implementation of a scalable, parsimonious forecasting model, only specifying easily available attributes, and relying on previous sales prices to incorporate information on the omitted, more market-specific attributes.[9]

While we argue for conceptual advantages of the combination, some practical limitations should also be addressed. First, collecting previous sales price data reflecting current housing quality is hard or even impossible in particular cases. Newly built dwellings obviously lack such data, but very old sales prices are hardly better, as they rarely represent the current state of the property (Case and Shiller (1987)). As a result, the combination might be less useful for markets where houses are traded less frequently, such as rural or suburban areas containing more family homes (Clapp et al. (1991)). In addition, there will always be some houses lacking data, preventing the method's applicability to all dwellings. We make a final remark regarding the scale of the model improvement. As an example, the median error of GWR is reduced from 6.65 % to 6.20 % – a 6.8 % improvement – when combining the regression predictions with estimates from repeat sales. This visible, but rather marginal improvement might imply that the combination has little practical implication. Arguably, both models are good enough to get an approximate value estimate but, at the same time, not good enough to make end users confident in the result.

[9] No universal hedonic specification exists (Bowen et al. (2001)), resulting, to a certain degree, in the need for local expertise to identify relevant price-influencing attributes in a given market (Gelfand et al. (1998)).

7

Conclusion

A central aspect of uncertainty in housing transactions is accurate property valuation. This article has investigated benefits of combining property price predictions from two valuation methods: Repeat sales and hedonic regression. The models were tested on 16,417 historical transactions in Oslo, Norway. Due to spatial effects inherent in housing markets, the hedonic regression was enhanced with three acknowledged spatial econometric models and a fourth, outlier-robust model proposed by this paper. This was done to ensure a change in performance was caused by methodological effects from the combination, as opposed to the correction of a spatially misspecified regression.

The studied combination resulted in improved accuracy for all hedonic regressions on all metrics, assumed due to diversification, following Bates and Granger (1969). Models with lower pre-combination accuracy displayed higher improvements, where reduction in median absolute percentage error ranged from 9.5 % for the ordinary regression to 6.8 % for the most accurate enhancement, geographically weighted regression. We infer that this varying gain indicates that repeat sales predictions contribute with at least some spatial information. While this might have limited value for refined spatial models, we suggest that the existence of some non-locational information in previous sales prices could have pivotal value for automated property valuation, as there are few alternatives to detect such information besides human inspection.

We identify two main limitations of the combination. Non-existent or inapplicable previous sales price data in certain markets is inevitable. A more manageable problem is our limited improvement, which we propose for further studies. Optimising the simple combination scheme presented in this paper is a clear-cut path, e.g., through more considerate implementation of the temporal dimension of previous sales. A second approach is improving repeat sales accuracy, for example, by applying local price indices. Considering broader trends in automatic housing valuation, machine learning appears to be usurping the position as focal point of research at the expense of hedonic regression (Park and Bae (2015)). However, these tools are equally dependent on observable, quantifiable data (Trawiński et al. (2017)), and following our argument that previous sales prices incorporate some otherwise hardly observable information, a repeat sales/machine learning combination is an interesting direction for further research.

References

Amundsen, B. (2015). Rike og fattige flytter fra hverandre i Oslo. *Forskning.no*. Available at: https://forskning.no/samfunnsgeografi/2015/06/rike-og-fattige-flytter-fra-hverandre-i-oslo (accessed: 11 December 2017).

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, volume 4. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Anselin, L. (2007). Spatial econometrics. In Baltagi, B. H., editor, *A Companion to Theoretical Econometrics*, chapter 14. Blackwell Publishing Ltd, Oxford. Doi: 10.1002/9780470996249.ch15.

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in regional science*, 89(1): 3–25. Available at: https://doi.org/10.1111/j.1435-5957.2010.00279.x (accessed: 22 May 2018).

Armstrong, J. S. (1989). Combining forecasts: The end of the beginning or the beginning of the end? *International Journal of Forecasting*, 5(4): 585–588. Available at: https://doi.org/10.1016/0169-2070(89)90013-7 (accessed: 13 December 2017).

Bailey, M. J., Muth, R. F., and Nourse, H. O. (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association*, 58(304): 933–942. Available at: http://www.jstor.org/stable/2283324 (accessed: 16 November 2017).

Basu, S. and Thibodeau, T. G. (1998). Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics*, 17(1): 61–85. Available at: https://link.springer.com/article/10.1023%2FA%3A1007703229507 (accessed: 1 May 2018).

Bates, J. M. and Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, 20(4): 451–468. Available at: https://doi.org/10.1057/jors.1969.103 (accessed: 24 May 2018).

Bohling, G. (2005). Kriging. In *C&PE940 Data Analysis in Engineering and Natural Science*. Kansas Geological Survey. Available at: http://people.ku.edu/~gbohling/cpe940/Kriging.pdf (accessed: 8 June 2018).

Bourassa, S. C., Hamelink, F., Hoesli, M., and MacGregor, B. D. (1999). Defining housing submarkets. *Journal of Housing Economics*, 8(2): 160–183. Available at: https://doi.org/10.1006/jhec.1999.0246 (accessed: 13 May 2018).

Bowen, W. M., Mikelbank, B. A., and Prestegaard, D. M. (2001). Theoretical and empirical considerations regarding space in hedonic housing price model applications. *Growth and change*, 32(4): 466–490. Available at: https://doi.org/10.1111/0017-4815.00171 (accessed: 6 June 2018).

Brooks, C. (2014). *Introductory Econometrics for Finance*. Cambridge University Press, Cambridge, 3rd edition.

Can, A. (1990). The measurement of neighborhood dynamics in urban house prices. *Economic geography*, 66(3): 254–272. Available at: http://www.jstor.org/stable/143400 (accessed: 30 April 2018).

Can, A. (1992). Specification and estimation of hedonic housing price models. *Regional science and urban economics*, 22(3): 453–474. Available at: https://doi.org/10.1016/0166-0462(92)90039-4 (accessed: 5 December 2017).

Can, A. and Megbolugbe, I. (1997). Spatial dependence and house price index construction. *The Journal of Real Estate Finance and Economics*, 14(1-2): 203–222. Avaialble at: https://link.springer.com/content/pdf/10.1023/A:1007744706720.pdf (accessed: 18 April 2018).

Case, B., Clapp, J., Dubin, R., and Rodriguez, M. (2004). Modeling spatial and temporal house price patterns: A comparison of four models. *The Journal of Real Estate Finance and Economics*, 29(2): 167–191. Available at: https://doi.org/10.1023/B:REAL.0000035309.60607.53 (accessed: 8 May 2018).

Case, B., Pollakowski, H. O., and Wachter, S. M. (1991). On choosing among house price index methodologies. *Real estate economics*, 19(3): 286–307. Available at: https://search.proquest.com/docview/211137020?accountid=12870 (accessed: 12 April 2018).

Case, B. and Quigley, J. M. (1991). The dynamics of real estate prices. *The Review of Economics and Statistics*, 73(1): 50–58. Available at: http://www.jstor.org/stable/2109686 (accessed: 23 May 2018).

Case, K. E. and Shiller, R. J. (1987). Prices of single family homes since 1970: New indexes for four cities. NBER Working Paper Series 2393, National Bureau of Economic Research. doi: 10.3386/w2393.

Casetti, E. (1972). Generating models by the expansion method: applications to geographical research. *Geographical analysis*, 4(1): 81–91. Available at: https://doi.org/10.1111/j.1538-4632.1972.tb00458.x (accessed: 30 April 2018).

Chan, Y. L., Stock, J. H., and Watson, M. W. (1999). A dynamic factor model framework for forecast combination. *Spanish Economic Review*, 1(2): 91–121. Available at: https://doi.org/10.1007/s101080050005 (accessed: 13 December 2017).

Chen, Y. (2016). Spatial autocorrelation approaches to testing residuals from least squares regression. *PloS one*, 11(1): e0146865. Available at: https://doi.org/10.1371/journal.pone.0146865 (accessed: 5 May 2018).

Chica Olmo, J. (1995). Spatial estimation of housing prices and locational rents. *Urban studies*, 32(8): 1331–1344. Available at: https://doi.org/10.1080/00420989550012492 (accessed: 1 May 2018).

Clapp, J. M., Giaccotto, C., and Tirtiroglu, D. (1991). Housing price indices based on all transactions compared to repeat subsamples. *Real Estate Economics*, 19(3): 270–285. Available at: https://doi.org/10.1111/1540-6229.00553 (accessed: 6 June 2018).

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4): 559–583. Available at: https://doi.org/10.1016/0169-2070(89)90012-5 (accessed: 24 May 2018).

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368): 829–836. Available at: http://www.jstor.org/stable/2286407 (accessed: 30 April 2018).

Cliff, A. D. and Ord, K. (1970). Spatial autocorrelation: a review of existing and new measures with applications. *Economic Geography*, 46: 269–292. Available at: http://www.jstor.org/stable/143144?seq=1#page_scan_tab_contents (accessed: 18 April 2018).

Colwell, P. F., Cannaday, R. E., and Wu, C. (1983). The analytical foundations of adjustment grid methods. *Real Estate Economics*, 11(1): 11–29. Available at: https://doi.org/10.1111/1540-6229.00277 (accessed: 11 May 2018).

Congdon, P. (1995). The impact of area context on long term illness and premature mortality: an illustration of multi-level analysis. *Regional Studies*, 29(4): 327–344. Available at: https://doi.org/10.1080/00343409512331349003 (accessed: 2 May 2018).

Corcoran, C. and Liu, F. (2014). Accuracy of Zillow's home value estimates. *REAL ESTATE ISSUES®*, 39(1): 45–49. Available at: http://www.cre.org/wp-content/uploads/201605/39_1.pdf#page=47 (accessed: 14 December 2017).

Cressie, N. (1990). The origins of kriging. *Mathematical geology*, 22(3): 239–252. Available at: https://doi.org/10.1007/BF00889887 (accessed: 8 May 2018).

Dale-Johnson, D. (1982). An alternative approach to housing market segmentation using hedonic price data. *Journal of Urban Economics*, 11(3): 311–332. Available at: https://doi.org/10.1016/0094-1190(82)90078-X (accessed: 13 May 2018).

de Haan, J. and Diewert, E. (2013). Hedonic regression methods. In *Handbook on Residential Property Prices Indices (RPPIs)*. Pages: 49–64. OECD publishing, Paris. Available at: http://dx.doi.org/10.1787/9789264197183-7-en (accessed: 21 May 2018).

Dubin, R. A. (1988). Estimation of regression coefficients in the presence of spatially autocorrelated error terms. *The Review of Economics and Statistics*, 70(3): 466–474. Available at: http://www.jstor.org/stable/1926785 (accessed: 8 May 2018).

Dubin, R. A. (1992). Spatial autocorrelation and neighborhood quality. *Regional science and urban economics*, 22(3): 433–452. Available at: https://doi.org/10.1016/0166-0462(92)90038-3 (acccessed: 8 May 2018).

Dubin, R. A. (1998). Spatial autocorrelation: a primer. *Journal of housing economics*, 7(4): 304–327. Available at: https://doi.org/10.1006/jhec.1998.0236 (accessed: 8 May 2018).

Eiendom Norge, Eiendomsverdi and FINN.no (2017). Eiendom Norges boligprisstatistikk. Available at: http://eiendomnorge.no/wp-content/uploads/2017/11/Boligstatistikk_oktober_01.pdf (accessed: 4 December 2017).

Eurostat (2016). Distribution of population by tenure status, type of household and income group - EU-SILC survey. Available at: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_lvho02&lang=en (accessed: 20 May 2018).

Faksvåg, K. V. (2015). Derfor øker boligprisene på østkanten. *Boligmani.no*. Available at: http://www.boligmani.no/aktuelt/boligmarkedet/boligprisene-ostkanten/14171 (accessed: 11 Decembner 2017).

Farber, S. and Yeates, M. (2006). A comparison of localized regression models in a hedonic house price context. *Canadian Journal of Regional Science*, 29(3): 405–420. Available at: http://www.cjrs-rcsr.org/archives/29-3/6-Farber-Yeates.pdf (accessed: 18 April 2018).

Fik, T. J., Ling, D. C., and Mulligan, G. F. (2003). Modeling spatial variation in housing prices: a variable interaction approach. *Real Estate Economics*, 31(4): 623–646. Available at: https://search.proquest.com/docview/211133757?accountid=12870 (accessed: 1 May 2018).

Fix, E. and Hodges Jr, J. L. (1952). Discriminatory analysis-nonparametric discrimination: Small sample performance. Technical report, CALIFORNIA UNIV BERKELEY. Available at: http://www.dtic.mil/dtic/tr/fulltext/u2/a800391.pdf (accessed: 7 May 2018).

Florax, R. and Folmer, H. (1992). Specification and estimation of spatial linear regression models: Monte Carlo evaluation of pre-test estimators. *Regional science and urban economics*, 22(3): 405–432. Available at: https://doi.org/10.1016/0166-0462(92)90037-2 (accessed: 1 May 2018).

Foster, S. A. and Gorr, W. L. (1986). An adaptive filter for estimating spatially-varying parameters: Application to modeling police hours spent in response to calls for service. *Management Science*, 32(7): 878–889. Available at: https://doi.org/10.1287/mnsc.32.7.878 (accessed: 2 May 2018).

Fotheringham, A. S. (2009). Geographically weighted regression. In *The SAGE Handbook of Spatial Analysis*, pages 243–254. SAGE Publications Ltd, London. doi: 10.4135/9780857020130.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley.

Freddie Mac (2017). Home Value Explorer. Available at: http://www.freddiemac.com/hve/hve.html (accessed: 5 December 2017).

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3): 115–146. Available at: http://www.jstor.org/stable/2986645 (accessed: 5 May 2018).

Gelfand, A. E., Ghosh, S. K., Knight, J. R., and Sirmans, C. F. (1998). Spatio-temporal modeling of residential sales data. *Journal of Business & Economic Statistics*, 16(3): 312–321. Available at: https://doi.org/10.1080/07350015.1998.10524770 (accessed: 14 May 2018).

Goodman, A. C. and Thibodeau, T. G. (1998). Housing market segmentation. *Journal of housing economics*, 7(2): 121–143. Available at: https://doi.org/10.1006/jhec.1998.0229 (accessed: 1 May 2018).

Goodman, A. C. and Thibodeau, T. G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, 12(3): 181–201. Available at: https://doi.org/10.1016/S1051-1377(03)00031-7 (accessed: 12 May 2018).

Goovaerts, P. (1999). Using elevation to aid the geostatistical mapping of rainfall erosivity. *Catena*, 34(3-4): 227–242. Available at: https://doi.org/10.1016/S0341-8162(98)00116-7 (accessed: 11 May 2018).

Granger, C. W. (1989). Invited review combining forecasts—twenty years later. *Journal of Forecasting*, 8(3): 167–173. Available at: https://doi.org/10.1002/for.3980080303 (accessed: 24 May 2018).

Griffith, D. A. (1992). What is spatial autocorrelation? reflections on the past 25 years of spatial statistics. *L'Espace géographique*, 21(3): 265–280. Available at: http://www.jstor.org/stable/44381737 (accessed: 11 May 2018).

Haider, M. and Miller, E. (2000). Effects of transportation infrastructure and location on residential real estate values: application of spatial autoregressive techniques. *Transportation*

Research Record: Journal of the Transportation Research Board, 1722: 1–8. Available at: https://doi.org/10.3141/1722-01 (accessed: 1 May 2018).

Hansen, S. N. and Pettrém, T. R. (2017). Housing value estimation by combining hedonic regression and the repeat sales method. Unpublished project thesis, Department of Industrial Economics and Technology Management, NTNU, Trondheim, Norway.

Harris, P., Brunsdon, C., and Fotheringham, A. S. (2011). Links, comparisons and extensions of the geographically weighted regression model when used as a spatial predictor. *Stochastic environmental Research and Risk assessment*, 25(2): 123–138. Available at: https://doi.org/10.1007/s00477-010-0444-6 (accessed: 1 June 2018).

Harris, P., Fotheringham, A., Crespo, R., and Charlton, M. (2010). The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets. *Mathematical Geosciences*, 42(6): 657–680. Available at: https://doi.org/10.1007/s11004-010-9284-7 (accessed: 1 June 2018).

Helbich, M., Brunauer, W., Hagenauer, J., and Leitner, M. (2013). Data-driven regionalization of housing markets. *Annals of the Association of American Geographers*, 103(4): 871–889. Available at: https://doi.org/10.1080/00045608.2012.707587 (accessed: 30 April 2018).

Hengl, T., Heuvelink, G. B. M., and Rossiter, D. G. (2007). About regression-kriging: from equations to case studies. *Computers & geosciences*, 33(10): 1301–1315. Available at: https://doi.org/10.1016/j.cageo.2007.05.001 (accessed: 8 May 2018).

Hengl, T., Heuvelink, G. B. M., and Stein, A. (2003). Comparison of kriging with external drift and regression kriging. Technical note, ITC. Available at: https://webapps.itc.utwente.nl/librarywww/papers_2003/misca/hengl_comparison.pdf (accessed: 11 May 2018).

Humberset, K. (2018). Her må du punge ut 85.000 kroner for én kvadratmeter. *aftenposten.no*. Available at: https://www.aftenposten.no/bolig/Her-ma-du-punge-ut-85000-kroner-for-n-kvadratmeter-10982b.html (accessed: 4 June 2018).

Kain, J. F. and Quigley, J. M. (1970). Measuring the value of housing quality. *Journal of the American statistical association*, 65(330): 532–548. Available at: http://www.jstor.org/stable/2284565 (accessed: 18 May 2018).

Kitanidis, P. K. (1993). Generalized covariance functions in estimation. *Mathematical Geology*, 25(5): 525–540. Available at: https://doi.org/10.1007/BF00890244 (accessed: 11 May 2018).

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 46(1): 33–50. Available at: http://www.jstor.org/stable/1913643 (accessed: 14 May 2018).

Krause, A. L. and Bitter, C. (2012). Spatial econometrics, land values and sustainability: Trends in real estate valuation research. *Cities*, 29: S19–S25. Available at: https://doi.org/10.1016/j.cities.2012.06.006 (accessed: 26 May 2018).

LeSage, J. P. and Pace, R. K. (2014). The biggest myth in spatial econometrics. *Econometrics*, 2(4): 217–249. Available at: https://doi.org/10.3390/econometrics2040217 (accessed: 15 May 2018).

Levin, J. (2001). Information and the market for lemons. *RAND Journal of Economics*, pages 657–666. Available at: http://www.jstor.org/stable/2696386 (accessed: 27 May 2018).

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. University of California Press.

Maser, S. M., Riker, W. H., and Rosett, R. N. (1977). The effects of zoning and externalities on the price of land: An empirical analysis of Monroe County, New York. *The Journal of Law and Economics*, 20(1): 111–132. Available at: http://www.jstor.org/stable/725089 (accessed: 8 May 2018).

Monsrud, I. J. and Takle, M. (2018). Price index for existing dwellings. *Statistics Norway*. Availbale at: https://www.ssb.no/en/priser-og-prisindekser/statistikker/bpi (accessed: 7 June 2018).

Moran, P. A. (1950). A test for the serial independence of residuals. *Biometrika*, 37(1/2): 178–181. Available at: http://www.jstor.org/stable/2332162 (accessed: 5 May 2018).

Odeh, I. O. A., McBratney, A. B., and Chittleborough, D. J. (1995). Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*, 67(3-4): 215–226. Available at: https://doi.org/10.1016/0016-7061(95)00007-B (accessed: 8 May 2018).

OECD, Eurostat, International Labour Organization, International Monetary Fund, The World Bank, and United Nations Economic Commission for Europe (2013). Repeat sales methods. In *Handbook on Residential Property Prices Indices (RPPIs)*. OECD publishing, Paris. Available at: http://dx.doi.org/10.1787/9789264197183-8-en (accessed: 21 May 2018).

Olaussen, J. O., Oust, A., and Solstad, J. T. (2017). Energy performance certificates–Informing the informed or the indifferent? *Energy Policy*, 111: 246–254. Available at: https://doi.org/10.1016/j.enpol.2017.09.029 (accessed: 19 May 2018).

Oslo Kommune, . (2018). Finn bydelen din her. Available at: https://www.oslo.kommune.no/politikk-og-administrasjon/bydeler/bydelsvelger/ (accessed: 13 May 2018).

Pace, R. K., Barry, R., Gilley, O. W., and Sirmans, C. F. (2000). A method for spatial–temporal forecasting with an application to real estate prices. *International Journal of Forecasting*, 16(2): 229–246. Available at: https://doi.org/10.1016/S0169-2070(99)00047-3 (accessed: 15 May 2018).

Pace, R. K. and Gilley, O. W. (1997). Using the spatial configuration of the data to improve estimation. *The Journal of Real Estate Finance and Economics*, 14(3): 333–340. Available at: https://doi.org/10.1023/A:1007762613901 (accessed: 11 May 2018).

Páez, A., Long, F., and Farber, S. (2008). Moving window approaches for hedonic price estimation: an empirical comparison of modelling techniques. *Urban Studies*, 45(8): 1565–1581. Available at: https://doi.org/10.1177/0042098008091491 (accessed: 1 June 2018).

Park, B. and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, 42(6): 2928–2934. Available at: https://doi.org/10.1016/j.eswa.2014.11.040 (accessed: 7 June 2018).

Reichert, A. K. (1990). The impact of interest rates, income, and employment upon regional housing prices. *The Journal of Real Estate Finance and Economics*, 3(4): 373–391. Available at: https://doi.org/10.1007/BF00178859 (accessed: 8 June 2018).

Richardson, H. W. (1988). Monocentric vs. policentric models: The future of urban economics in regional science. *The Annals of Regional Science*, 22(2): 1–12. Available at: https://doi.org/10.1007/BF01287319 (accessed: 14 May 2018).

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1): 34–55. Available at: http://www.journals.uchicago.edu/doi/pdfplus/10.1086/260169 (accessed: 5 December 20179).

Sinnott, R. W. (1984). Virtues of the haversine. *Sky and Telescope*, 68(2): 159.

Sokal, R. R. and Oden, N. L. (1978). Spatial autocorrelation in biology: 1. Methodology. *Biological journal of the Linnean Society*, 10(2): 199–228. Available at: https://doi.org/10.1111/j.1095-8312.1978.tb00013.x (accessed: 5 May 2018).

Trawiński, B., Lasota, T., Kempa, O., Telec, Z., and Kutrzyński, M. (2017). Comparison of ensemble learning models with expert algorithms designed for a property valuation system. In *Conference on Computational Collective Intelligence Technologies and Applications*, pages 317–327. Springer. Available at: https://doi.org/10.1007/978-3-319-67074-4_31 (accessed: 6 June 2018).

Wallis, K. F. (2011). Combining forecasts–forty years later. *Applied Financial Economics*, 21(1-2): 33–41. Available at: https://doi.org/10.1080/09603107.2011.523179 (accessed: 29 May 2018).

Wheeler, D. C. and Calder, C. A. (2007). An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. *Journal of Geographical Systems*, 9(2): 145–166. Available at: https://doi.org/10.1007/s10109-006-0040-y (accessed: 11 May 2018).

Yao, J. and Fotheringham, A. S. (2016). Local spatiotemporal modeling of house prices: a mixed model approach. *The Professional Geographer*, 68(2): 189–201. Available at: https://doi.org/10.1080/00330124.2015.1033671 (accessed: 30 April 2018).

Yoo, S.-H. (2001). A robust estimation of hedonic price models: least absolute deviations estimation. *Applied Economics Letters*, 8(1): 55–58. Available at: https://doi.org/10.1080/135048501750041303 (accessed: 14 May 2018).

A

Appendix

A.1 Algorithm combining repeat sales and hedonic regression predictions

1. **No previous sales available:**
   (Applies to 20.0 % of the dwellings in Virdi (Table 4.7))

   (a) In this situation, the regression estimate is set as final estimate.

2. **One previous sale available:**
   (Applies to 34.3 % of the dwellings in Virdi (Table 4.7))

   (a) We discard the repeat sales estimate when this value deviates more than 25 % from the regression estimate.[1] In this case, the regression estimate is set as final estimate.

   (b) When the repeat sales estimate remains, the final estimate is a weighted sum of the two estimates, giving the regression estimate a weight of 60 %.

3. **Two previous sales available:**
   (Applies to 25.0 % of the dwellings in Virdi (Table 4.7))

   (a) For both repeat sales estimates, we remove outliers following procedure 2a.

   (b) When discarding both repeat sales estimates, the regression estimate is used.

   (c) When discarding only one repeat sales estimate, the procedure in 2b is used for the estimate remaining.

   (d) When both estimates remain, we compose the final price estimate weighting the regression 60 %. The remaining 40 % is based on the two repeat sales estimates, where we let the estimate from the most recent sale be weighted 90 %.

[1] This threshold is chosen so the resulting removal is in accordance with Chan et al. (1999)

4. **Three previous sales available**:
   (Applies to 20.7 % of the dwellings in Virdi (Table 4.7))

   (a) For all repeat sales estimates, we remove outliers following the procedure 2a.

   (b) When discarding all repeat sales estimates, the regression estimate is used.

   (c) When discarding two repeat sales estimates, the procedure in 2b is used for the estimate remaining.

   (d) When one estimate is discarded, the procedure in 3d is used for the estimates remaining.

   (e) When all estimates remain, the regression estimate is given a weight of 60 %. The remaining 40 % is based on the three repeat sales estimates, where the most recent is given 85 %, the second 12 % and the least recent 3 %.

A.2 K-means and k-nearest neighbour algorithms

input : Data set with longitude, latitude and price; the number of desired clusters, k
output: Clustered data set
begin;
Select longitude, latitude and price values from k random entries in the data set, and initialise district centroids with these values;
while *Centroids still updating* **do**
    Assign each observation to its closest centroid using distance function Δ_m defined below;
    Reposition all centroids to the average longitude, latitude and price values of their belonging observations;
end

Algorithm 1: K-means. $\Delta_m(s,t) = \sqrt{hav((u_s, v_s),(u_t, v_t))^2 + \mu(p_s - p_t)^2}$, where s and t represent two dwellings; $u_{s\ (t)}$ and $v_{s\ (t)}$ represent the longitude and latitude of s (t), respectively; hav is the haversine function returning the kilometre distance between two (longitude, latitude) points; $p_{s\ (t)}$ is the price of s (t), on the form specified in Equation (5.1), and μ is a scaling factor set to 9.[3]

input : Clustered training set with longitude and latitude; unclustered test set with longitude and latitude; the number of desired neighbours, k
output: Predicted clustering of the test set
begin;
for $i \leftarrow 1$ **to** *test set size* **do**
    for $j \leftarrow 1$ **to** k **do**
        Identify the j^{th} nearest neighbour in the training set, n, to dwelling i in the test set, using distance function Δ_n defined below;
        Store the clustering of n;
    end
    Assign dwelling i to the cluster shared by the largest number of neighbours;
end

Algorithm 2: K-nearest neighbour. $\Delta_n(s,t) = hav((u_s, v_s),(u_t, v_t))$, where s and t represent two dwellings; $u_{s\ (t)}$ and $v_{s\ (t)}$ represent the longitude and latitude of s (t), respectively, and hav is the haversine function returning the kilometre distance between two (longitude, latitude) points.

[3] Based on trial; higher values of μ repeatedly resulted in geographically non-cohesive clusters.

A.3 K-means results for different values of k

Table A.1: K-means results with varying k, compared to administrative districts

District type	$Q_{0.25}$	$\boldsymbol{Q_{0.5}}$	$Q_{0.75}$	Within 10 %	Moran's I	Geary's C
Administrative	3.74%	**8.05%**	14.1%	59.6%	0.143	0.826
K-means (k = 14)	3.61%	**7.78%**	13.9%	60.8%	0.125	0.839
K-means (k = 16)	3.66%	**7.78%**	13.9%	60.8%	0.115	0.853
K-means (k = 18)	3.55%	**7.67%**	13.5%	61.6%	0.107	0.860
K-means (k = 20)	3.54%	**7.62%**	13.6%	61.5%	0.098	0.870
K-means (k = 22)	3.57%	**7.64%**	13.6%	61.5%	0.103	0.866

District type refers to the type of district indicator used; $Q_{0.25}$, $Q_{0.5}$ and $Q_{0.75}$ denote the first, second and third quartile of the errors, respectively, where $Q_{0.5}$ is boldfaced for emphasis; *Within 10 %* specifies the fraction of errors below 10 %, and *Moran's I* and *Geary's C* are calculated as detailed in Section 2.3.
The results shown are average values from 10 runs for each implementation. The number of observations used for model training is 13,133, while the number of observations out-of-sample is 3,284.

A.4 Independent repeat sales results

Table A.2: Independent repeat sales results

Model	$Q_{0.25}$	$\boldsymbol{Q_{0.5}}$	$Q_{0.75}$	Within 10 %	Number of observations	Percentage of total
Repeat sales	4.11 %	**8.88 %**	16.1 %	54.8 %	13,138	80.0 %
Repeat sales after outlier removal	3.68 %	**7.77 %**	13.4 %	61.5 %	11,690	71.2 %

Model refers to whether the repeat sales estimates have undergone outlier removal; $Q_{0.25}$, $Q_{0.5}$ and $Q_{0.75}$ denote the first, second and third quartile of the errors, respectively, where $Q_{0.5}$ is boldfaced for emphasis; *Within 10 %* specifies the fraction of errors below 10 %; *Number of observations* denotes the number of entries with previous sales prices, and *Percentage of total* indicates the latter number as a percentage of the entire data set.
We remark the coincidental similarity between the number of entries with previous sales before outlier removal (13,138) and the number of observations used for model training of regressions in this paper (13,133).

A.5 Hedonic regression coefficients tables

Table A.3: Hedonic regression coefficients with admin districts

	Coefficients	Robust Std. Err	P > |t|
Intercept	11.5659***	(0.014)	0.0000
Size group: 10 - 29 m^2	0.2944***	(0.008)	0.0000
Size group: 30 - 39 m^2	0.1307***	(0.005)	0.0000
Size group: 50 - 59 m^2	−0.1090***	(0.005)	0.0000
Size group: 60 - 69 m^2	−0.1221***	(0.005)	0.0000
Size group: 70 - 79 m^2	−0.1560***	(0.005)	0.0000
Size group: 80 - 89 m^2	−0.1786***	(0.006)	0.0000
Size group: 90 - 99 m^2	−0.1543***	(0.007)	0.0000
Size group: 100 - 109 m^2	−0.1739***	(0.008)	0.0000
Size group: 110 - 119 m^2	−0.1759***	(0.010)	0.0000
Size group: 120 - 129 m^2	−0.2038***	(0.010)	0.0000
Size group: 130 - 139 m^2	−0.2230***	(0.012)	0.0000
Size group: 140 - 149 m^2	−0.2367***	(0.013)	0.0000
Size group: 150 - 179 m^2	−0.2864***	(0.011)	0.0000
Size group: Above 180 m^2	−0.3454***	(0.012)	0.0000
Month sold: August 2016	−0.0590***	(0.018)	0.0010
Month sold: September 2016	−0.0441***	(0.008)	0.0000
Month sold: October 2016	−0.0343***	(0.007)	0.0000
Month sold: November 2016	−0.0211***	(0.008)	0.0050
Month sold: December 2016	0.0056	(0.009)	0.5350
Month sold: January 2017	0.0111	(0.008)	0.1830
Month sold: February 2017	0.0196**	(0.008)	0.0100
Month sold: March 2017	−0.0023	(0.008)	0.7630
Month sold: May 2017	−0.0126*	(0.007)	0.0870
Month sold: June 2017	−0.0545***	(0.007)	0.0000
Month sold: July 2017	−0.0754***	(0.008)	0.0000
Month sold: August 2017	−0.0805***	(0.008)	0.0000
Month sold: September 2017	−0.0771***	(0.008)	0.0000
Month sold: October 2017	−0.0976***	(0.007)	0.0000
Month sold: November 2017	−0.0968***	(0.008)	0.0000
Month sold: December 2017	−0.1326***	(0.009)	0.0000
District: Alna	−0.4747***	(0.006)	0.0000
District: Bjerke	−0.3463***	(0.007)	0.0000
District: Grünerløkka	−0.1706***	(0.006)	0.0000
District: Gamle Oslo	−0.2086***	(0.006)	0.0000
District: Grorud	−0.5306***	(0.007)	0.0000
District: Nordre Aker	−0.1582***	(0.007)	0.0000
District: Nordstrand	−0.3214***	(0.007)	0.0000
District: Østensjø	−0.3788***	(0.006)	0.0000
District: Sagene	−0.1298***	(0.006)	0.0000
District: St. Hanshaugen	−0.0633***	(0.006)	0.0000
District: Stovner	−0.6528***	(0.008)	0.0000
District: Ullern	−0.1544***	(0.008)	0.0000
District: Vestre Aker	−0.2238***	(0.007)	0.0000
House type: Apartment	−0.0992***	(0.011)	0.0000
House type: Semi-detached house	−0.0184	(0.012)	0.1380
House type: Serial house	−0.0617***	(0.011)	0.0000
Construction year: 1990 - 2004	0.0595***	(0.005)	0.0000
Construction year: 2005 - 2014	0.1062***	(0.004)	0.0000
Construction year: 2015 -	0.1681***	(0.009)	0.0000
Needs refurbishment	−0.1028***	(0.005)	0.0000
Has a garden	0.0397***	(0.005)	0.0000
Housing cooperative	−0.0098***	(0.003)	0.0020
Is a penthouse	0.0449***	(0.005)	0.0000
Has a terrace	0.0182***	(0.003)	0.0000
High monthly shared cost	−0.0546***	(0.004)	0.0000
Two bedrooms & size < 60 m^2	0.0477***	(0.006)	0.0000
Three bedrooms & size < 85 m^2	0.0262***	(0.006)	0.0000

*** Significant at the 1 % level. ** Significant at 5 % level. * Significant at 10 % level.
Coefficients estimated using LAD, Pseudo R-squared = 0.5290, number of observations is 13,133. Dependent variable is price as defined in Equation (5.1). All explanatory variables are indicator variables as described in Table (4.1 - 4.6). The baseline variables are: *Size group 40 - 49 m^2*; *Month sold: April 2017*; *District: Frogner*; *House type: Detached house* and *Construction year: 1820 - 1889*.

Table A.4: Hedonic regression coefficients with k-means districts

	Coefficients	Robust Std. Err	P>|t|
Intercept	10.8891***	(0.014)	0.0000
Size group: 10 - 29 m^2	0.2913***	(0.007)	0.0000
Size group: 30 - 39 m^2	0.1322***	(0.005)	0.0000
Size group: 50 - 59 m^2	−0.0934***	(0.005)	0.0000
Size group: 60 - 69 m^2	−0.1200***	(0.005)	0.0000
Size group: 70 - 79 m^2	−0.1403***	(0.005)	0.0000
Size group: 80 - 89 m^2	−0.1549***	(0.006)	0.0000
Size group: 90 - 99 m^2	−0.1458***	(0.007)	0.0000
Size group: 100 - 109 m^2	−0.1530***	(0.008)	0.0000
Size group: 110 - 119 m^2	−0.1744***	(0.009)	0.0000
Size group: 120 - 129 m^2	−0.1925***	(0.010)	0.0000
Size group: 130 - 139 m^2	−0.1801***	(0.012)	0.0000
Size group: 140 - 149 m^2	−0.2266***	(0.013)	0.0000
Size group: 150 - 179 m^2	−0.2668***	(0.011)	0.0000
Size group: Above 180 m^2	−0.3437***	(0.012)	0.0000
Month sold: August 2016	−0.0647***	(0.016)	0.0000
Month sold: September 2016	−0.0371***	(0.007)	0.0000
Month sold: October 2016	−0.0292***	(0.007)	0.0000
Month sold: November 2016	−0.0111	(0.007)	0.1300
Month sold: December 2016	0.0097	(0.009)	0.2690
Month sold: January 2017	0.0204**	(0.008)	0.0110
Month sold: February 2017	0.0269***	(0.007)	0.0000
Month sold: March 2017	0.0092	(0.007)	0.2020
Month sold: May 2017	−0.0071	(0.007)	0.3160
Month sold: June 2017	−0.0454***	(0.007)	0.0000
Month sold: July 2017	−0.0717***	(0.008)	0.0000
Month sold: August 2017	−0.0781***	(0.007)	0.0000
Month sold: September 2017	−0.0683***	(0.007)	0.0000
Month sold: October 2017	−0.0854***	(0.007)	0.0000
Month sold: November 2017	−0.0914***	(0.007)	0.0000
Month sold: December 2017	−0.1242***	(0.009)	0.0000
K-means district 1	0.2991***	(0.007)	0.0000
K-means district 2	0.5303***	(0.006)	0.0000
K-means district 3	0.3158***	(0.007)	0.0000
K-means district 4	0.4531***	(0.006)	0.0000
K-means district 5	0.6171***	(0.007)	0.0000
K-means district 6	0.2566***	(0.007)	0.0000
K-means district 7	0.3250***	(0.008)	0.0000
K-means district 8	0.6652***	(0.007)	0.0000
K-means district 9	0.5428***	(0.008)	0.0000
K-means district 10	0.4454***	(0.007)	0.0000
K-means district 11	0.4140***	(0.007)	0.0000
K-means district 12	0.1361***	(0.007)	0.0000
K-means district 13	0.5137***	(0.006)	0.0000
House type: Apartment	−0.0919***	(0.011)	0.0000
House type: Semi-detached house	−0.0181	(0.012)	0.1330
House type: Serial house	−0.0474***	(0.011)	0.0000
Construction year: 1990 - 2004	0.0519***	(0.005)	0.0000
Construction year: 2005 - 2014	0.1035***	(0.004)	0.0000
Construction year: 2015 -	0.1411***	(0.009)	0.0000
Needs refurbishment	−0.0952***	(0.005)	0.0000
Has a garden	0.0298***	(0.004)	0.0000
Housing cooperative	−0.0031	(0.003)	0.3180
Is a penthouse	0.0359***	(0.005)	0.0000
Has a terrace	0.0204***	(0.003)	0.0000
High monthly shared cost	−0.0498***	(0.004)	0.0000
Two bedrooms & size < 60 m^2	0.0465***	(0.006)	0.0000
Three bedrooms & size < 85 m^2	0.0294***	(0.006)	0.0000

*** Significant at the 1 % level. ** Significant at 5 % level. * Significant at 10 % level.
Coefficients estimated using LAD, Pseudo R-squared = 0.5559, number of observations is 13,133. Dependent variable is price as defined in Equation (5.1). All explanatory variables are indicator variables as described in Table (4.1 - 4.6). The baseline variables are: *Size group 40 - 49m^2*; *Month sold: April 2017*; *K-means district 14*; *House type: Detached house* and *Construction year: 1820 - 1889*.

Table A.5: Hedonic regression coefficients with k-means districts and autoregressive term

	Coefficients	Robust Std. Err	P>|t|
Intercept	7.3865***	(0.144)	0.0000
Size group: 10 - 29 m^2	0.2656***	(0.007)	0.0000
Size group: 30 - 39 m^2	0.1218***	(0.005)	0.0000
Size group: 50 - 59 m^2	−0.0926***	(0.004)	0.0000
Size group: 60 - 69 m^2	−0.1149***	(0.004)	0.0000
Size group: 70 - 79 m^2	−0.1331***	(0.005)	0.0000
Size group: 80 - 89 m^2	−0.1443***	(0.006)	0.0000
Size group: 90 - 99 m^2	−0.1348***	(0.006)	0.0000
Size group: 100 - 109 m^2	−0.1480***	(0.007)	0.0000
Size group: 110 - 119 m^2	−0.1710***	(0.009)	0.0000
Size group: 120 - 129 m^2	−0.2011***	(0.010)	0.0000
Size group: 130 - 139 m^2	−0.1863***	(0.011)	0.0000
Size group: 140 - 149 m^2	−0.2200***	(0.012)	0.0000
Size group: 150 - 179 m^2	−0.2569***	(0.01)	0.0000
Size group: Above 180 m^2	−0.3596***	(0.011)	0.0000
Month sold: August 2016	−0.0450***	(0.016)	0.0060
Month sold: September 2016	−0.0264***	(0.007)	0.0000
Month sold: October 2016	−0.0227***	(0.007)	0.0010
Month sold: November 2016	−0.0087	(0.007)	0.2060
Month sold: December 2016	0.0090	(0.008)	0.2670
Month sold: January 2017	0.0146*	(0.007)	0.0510
Month sold: February 2017	0.0287***	(0.007)	0.0000
Month sold: March 2017	0.0043	(0.007)	0.5230
Month sold: May 2017	−0.0075	(0.007)	0.2640
Month sold: June 2017	−0.0518***	(0.007)	0.0000
Month sold: July 2017	−0.0758***	(0.007)	0.0000
Month sold: August 2017	−0.0783***	(0.007)	0.0000
Month sold: September 2017	−0.0728***	(0.007)	0.0000
Month sold: October 2017	−0.0845***	(0.007)	0.0000
Month sold: November 2017	−0.0887***	(0.007)	0.0000
Month sold: December 2017	−0.1113***	(0.008)	0.0000
K-means district 1	−0.0009	(0.007)	0.8940
K-means district 2	0.1470***	(0.008)	0.0000
K-means district 3	−0.0629***	(0.008)	0.0000
K-means district 4	−0.0763***	(0.007)	0.0000
K-means district 5	0.0345***	(0.007)	0.0000
K-means district 6	−0.1110***	(0.007)	0.0000
K-means district 7	−0.2666***	(0.009)	0.0000
K-means district 8	−0.0782***	(0.007)	0.0000
K-means district 9	0.1042***	(0.007)	0.0000
K-means district 10	0.0483***	(0.007)	0.0000
K-means district 11	0.0771***	(0.008)	0.0000
K-means district 12	−0.1926***	(0.008)	0.0000
K-means district 13	−0.0132*	(0.007)	0.0540
House type: Apartment	−0.1190***	(0.01)	0.0000
House type: Semi-detached house	−0.0266**	(0.011)	0.0180
House type: Serial house	−0.0666***	(0.01)	0.0000
Construction year: 1990 - 2004	0.0555***	(0.005)	0.0000
Construction year: 2005 - 2014	0.0884***	(0.004)	0.0000
Construction year: 2015 -	0.1347***	(0.008)	0.0000
Needs refurbishment	−0.0959***	(0.005)	0.0000
Has a garden	0.0237***	(0.004)	0.0000
Housing cooperative	0.0020	(0.003)	0.4870
Is a penthouse	0.0442***	(0.004)	0.0000
Has a terrace	0.0234***	(0.003)	0.0000
High monthly shared cost	−0.0457***	(0.004)	0.0000
Two bedrooms & size < 60 m^2	0.0478***	(0.005)	0.0000
Three bedrooms & size < 85 m^2	0.0283***	(0.006)	0.0000
Autoregressive term	0.3553***	(0.013)	0.0000

*** Significant at the 1 % level. ** Significant at 5 % level. * Significant at 10 % level.
Coefficients estimated using LAD, Pseudo R-squared = 0.5710, number of observations is 13,133. Dependent variable is price as defined in Equation (5.1). All explanatory variables are indicator variables as described in Table (4.1 - 4.6). The baseline variables are: *Size group 40 - 49m^2*; *Month sold: April 2017*; *K-means district 14*; *House type: Detached house* and *Construction year: 1820 - 1889*.

A.6 Residual maps

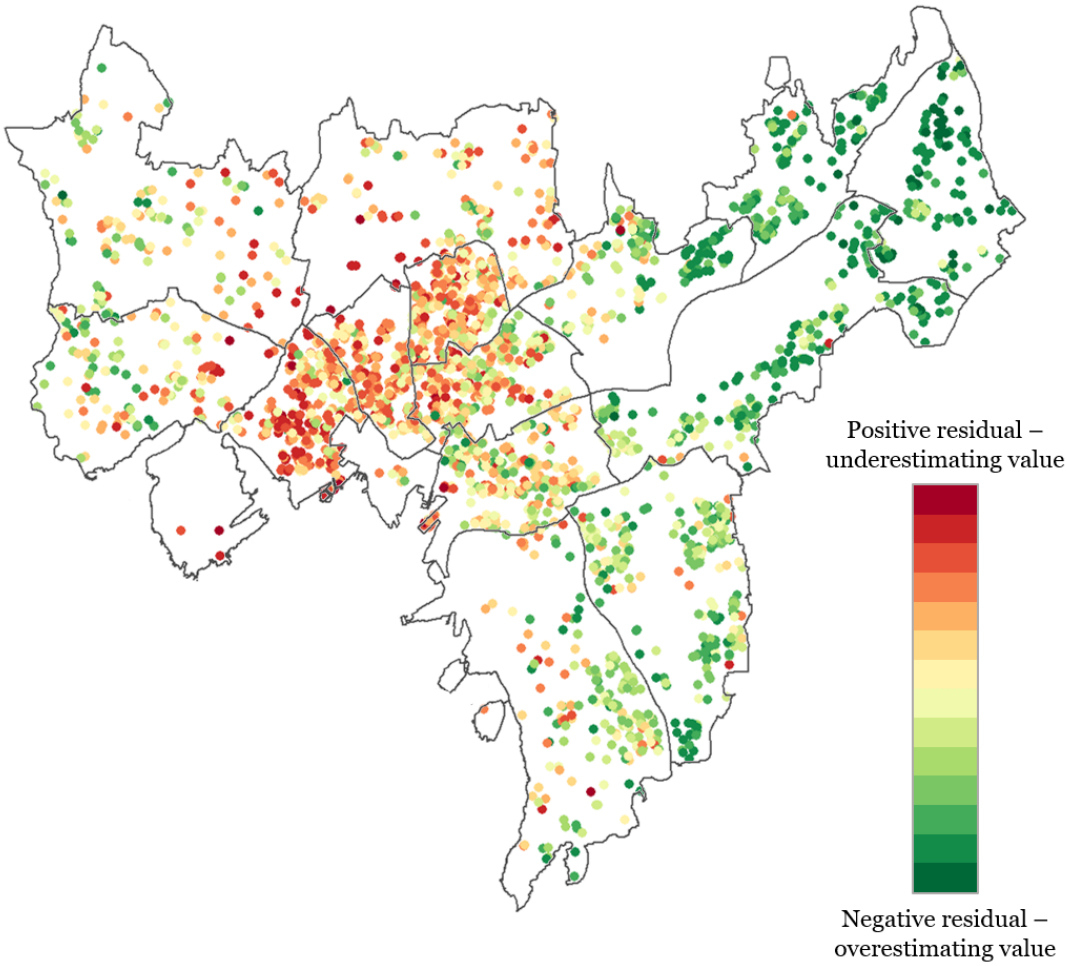


Figure A.1: Out-of-sample residuals from the hedonic regression without spatial enhancements or district indicators (referred to in Chapter 6 as the benchmark model) on a map of Oslo. Lines represent administrative district boundaries; marker colour represents residual value for each dwelling. Observations: 3,284.

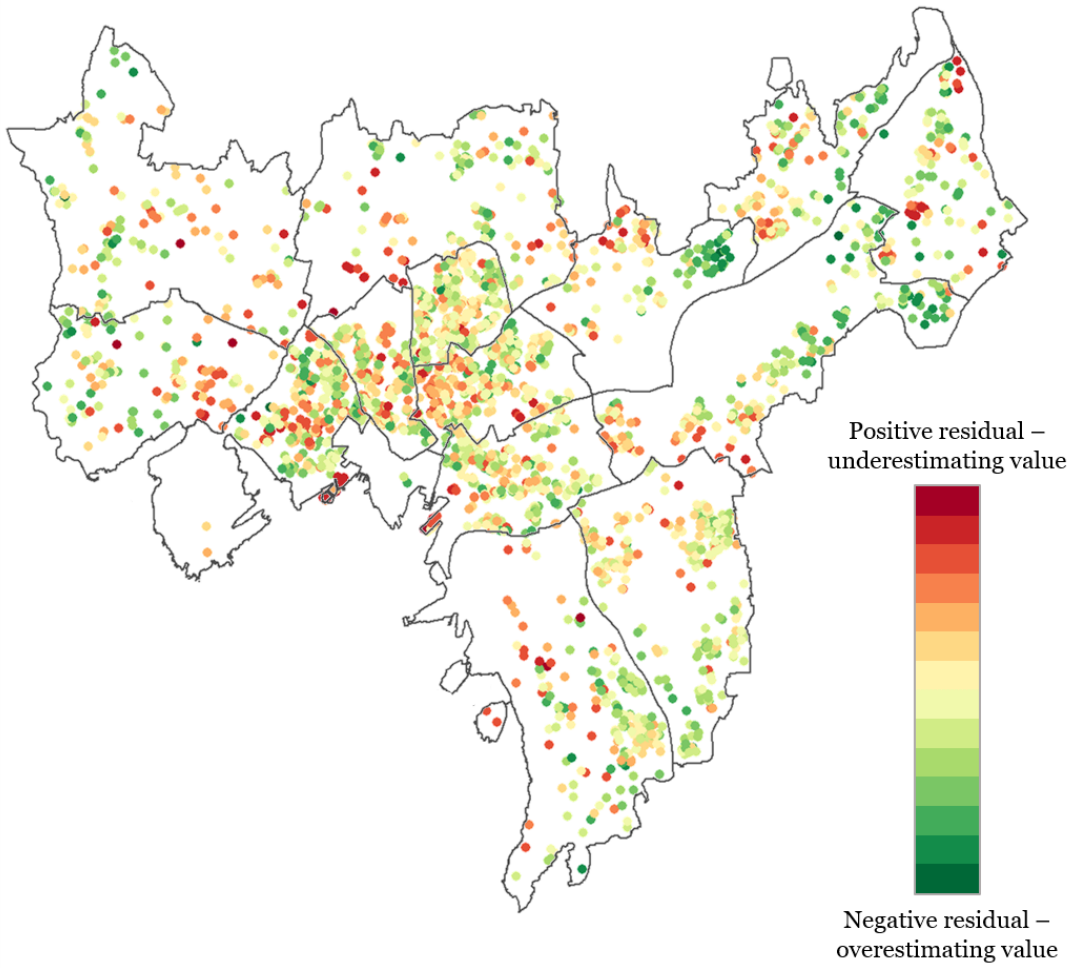


Figure A.2: Out-of-sample residuals from the hedonic regression with administrative district indicators on a map of Oslo. Lines represent administrative district boundaries; marker colour represents residual value for each dwelling. Observations: 3,284.

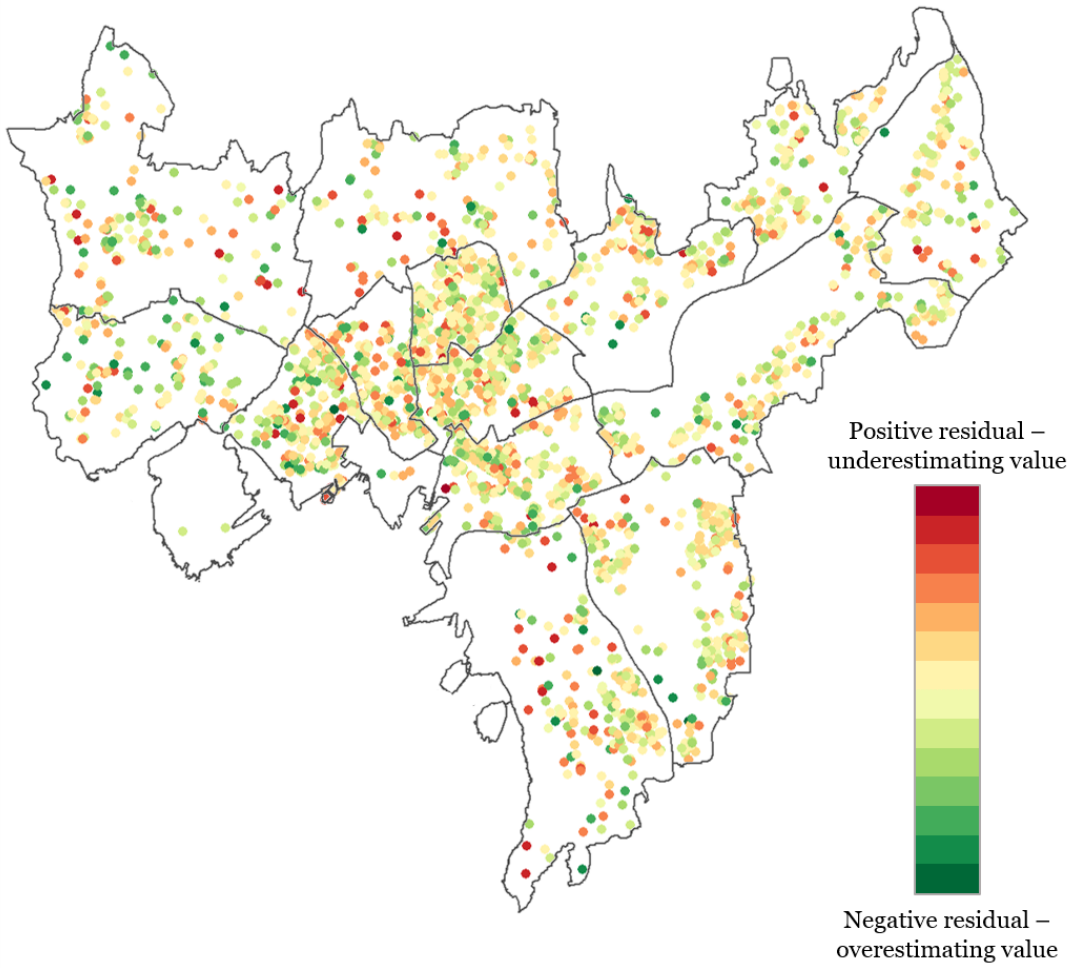


Figure A.3: Out-of-sample residuals from our most accurate model, combining geographically weighted regression and repeat sales, on a map of Oslo. Lines represent administrative district boundaries; marker colour represents residual value for each dwelling. Observations: 3,284.