The Pricing Implications of Housing Location

Samuel Abächerli
University of St. Gallen

Dr. Prof. Roland Füss
Swiss Institute of Banking and Finance s/bf-HSG
University of St. Gallen

17th of February, 2020

Abstract

Analysing the pricing implications of housing location, this thesis leverages a data set containing both the prices and structural characteristics of properties sold in New York City throughout 2018, which is augmented with locational data obtained from OpenStreetMaps as well as data available from New York City's open data portal. Predictions of locational value for all 28'826 observations were obtained by estimating a generalised additive model with cubic regression splines using ten knots, accounting for spatial dependence and nonlinearity. A residence at the best location in New York is estimated to be 3.5 times more expensive than a home at an average location with equal neighbourhood and structural characteristics. Moreover, it was found that 58 percent of house price variation can be explained by the locational value alone. As such, this thesis capitalises on an extensive set of locational characteristics obtained from openly available sources to further the understanding of location and its importance for house prices, and applies several econometric advances to address the shortcomings of the traditional hedonic approach.

JEL classification: R30, R32.

Keywords: Spatial autocorrelation; generalised additive models; locational amenities; points of interest; OpenStreetMap.

Contents

1 Introduction **1**

2 Literature review **2**

3 Theoretical background **10**

3.1 The housing market 10

3.2 The hedonic model 11

4 Econometric issues **14**

4.1 Spatial autocorrelation 14

4.2 Functional misspecification 19

4.3 Variable misspecification 25

4.4 Further issues 38

5 Empirical analysis **39**

5.1 Data description 39

5.2 Model specification 49

5.3 Descriptive statistics 51

5.4 Model comparison 54

5.5 Results 56

6 Discussion **60**

6.1 Limitations 61

7 Conclusion **62**

8 Acknowledgments **62**

A Variable discussion **64**

A.1 Distance to CBD . . . 64

A.2 School distance and quality . . . 64

A.3 Parks . . . 65

A.4 Waterfront . . . 66

A.5 Age . . . 67

A.6 Floor level . . . 67

A.7 Number of rooms . . . 68

A.8 Views . . . 69

A.9 Noise pollution . . . 69

A.10 Quality . . . 70

B Miscellaneous **71**

B.1 The importance of projections . . . 71

B.2 Inference for normality tests . . . 71

C Econometric expositions **72**

C.1 Effects of scaling and centring . . . 72

C.2 Bias of OLS with endogeneity . . . 73

C.3 Bias of OLS in spatial models . . . 74

C.4 The concentrated log-likelihood . . . 75

C.5 Geometric decay . . . 76

C.6 Spatial weights matrices . . . 77

C.7 MESS specification . . . 78

D Graphical representations **80**

D.1 Nonlinearity plots 89

E Tables **99**

F Code snippets **103**

F.1 Playground polygon and point data 103

List of Tables

1 Descriptive statistics of the numerical variables included in the processed data set used to analyse the pricing implications of housing location for New York City. In addition, there is one categorical variable, *Building class*, with 4 levels. 52

2 Global spatial statistics 53

3 As a means of identifying how well the different models capture the spatial effects in the data, Moran's I was calculated on the residuals of each model using four different spatial weights specifications. 55

4 A comparison of different model specifications measured by the median absolute percentage accuracy, along with the corresponding Breusch-Pagen and Anderson-Darling test statistics. 56

5 An overview of all the variables initially defined for the model. 57

6 Summary output of the estimated locational, structural, and neighbourhood values of an average one family house in Queens, along with the residual values. 58

7 List of the most expensive properties transacted in 2018. 99

8 List of detailed buildings types with their corresponding number of observations. 100

9 Two properties with near equivalent locations, located in the same neighbourhood, and with seemingly comparable structural characteristics trading at very different prices, showcasing the presence of anomalies remaining in the data set. 101

10 The 20 most highly correlated variables in the data set. 102

11 The predicted values for each explanatory variable for the first five observations of the data set using a GAM with cubic regression splines using ten knots. 103

List of Figures

1 Smoothing functions depicting the nonlinearity of the distance to fitness centres. 51

2 Local Moran's I depicted as a choropleth map for New York indicating strong spatial autocorrelation in Manhattan, and moderate spatial autocorrelation in the Bronx, the neighbourhood of Brooklyn Heights, and some parts of Queens and Brooklyn. 54

3 Two maps portraying the observed values and predicted values. 58

4 A map with the predicted locational value for each observation in the data set. 59

5 Graphical definition of rook contiguity, bishop contiguity, and queen contiguity spatial weights matrices. 77

6 A histogram of house prices per square metre depicting the distribution of the untransformed data. 80

7 Two maps used to determine the the maximum distance of all nearest neighbour distances in the data set (i.e. distance threshold). 80

8 A specific example of an area categorised by OpenStreetMap as a park, but likely not to be considered to be a park by prospective home buyers. 81

9 Upon further inspection of the OpenStreetMap polygon and point data for playgrounds, it was revealed that points data was incomplete (depicted by the green polygons and points dots), while point data was inaccurate and inconsistent (depicted by the blue points). 81

10 Locations of primary schools located in New York City showing a very good coverage. 82

11 A map of the main road network in New York City, where blue refers to highways, green to highway links, yellow to trunk roads, and red to main roads. 82

12 The individual significance levels of Local Moran's I depicted as a choropleth map for New York, once using the Bonferroni adjustment and once without. 83

13 A map showing the spatial distribution of the outliers measured by Cook's distance. 83

14 Diagnostics plot of the GAM model on the entire data set. 84

15 Diagnostics plot of the GAM model on the data set without outliers. 84

16 A scatterplot with distance to airport on the x-axis and distance to subway on the y-axis with observations in Staten Island marked red while all other observations are marked blue. 85

17 Residuals plots for the graphical detection of heteroskedasticity. 85

18 Normal Q-Q plot for the graphical detection of non-normality. 86

19 Histograms of prices for the graphical detection of non-normality. 86

20 Histograms of residuals for the graphical detection of non-normality. 87

21 A map portraying the predicted locational value of each observation. 87

22 A map portraying the predicted neighbourhood value of each observation. . . 87

23 A map portraying the predic- ted structural value of each observation. 88

24 A map portraying the residual value of each observation. 88

25 A map portraying all OpenStreetMaps available restaurants. 88

26 A map portraying all OpenStreetMaps available cafes. 88

27 A map portraying all OpenStreetMaps available bars. 89

28 A map portraying all OpenStreetMaps available parks. 89

29 Smoothing functions depicting the nonlinearity of parks. 89

30 Smoothing functions depicting the nonlinearity of building age. 90

31 Smoothing functions depicting the nonlinearity of the distance to airports. . 90

32 Smoothing functions depicting the nonlinearity of the distance to cinemas. . 91

33 Smoothing functions depicting the nonlinearity of the distance to convenience stores. 91

34 Smoothing functions depicting the nonlinearity of the distance to the stock exchange. 92

35 Smoothing functions depicting the nonlinearity of the distance to fitness centres. 92

36 Smoothing functions depicting the nonlinearity of the distance to highways. . 93

37 Smoothing functions depicting the nonlinearity of the distance to main roads. 93

38 Smoothing functions depicting the nonlinearity of the distance to playgrounds. 94

39 Smoothing functions depicting the nonlinearity of the distance to railways. . 94

40 Smoothing functions depicting the nonlinearity of the distance to shopping centres and department stores. 95

41 Smoothing functions depicting the nonlinearity of the distance to subway stations. 95

42 Smoothing functions depicting the nonlinearity of the distance to supermarkets. 96

43 Smoothing functions depicting the nonlinearity of the distance to waterfronts. 96

44 Smoothing functions depicting the nonlinearity of schooling quality. 97

45 Smoothing functions depicting the nonlinearity of neighbourhoods. 97

46 Smoothing functions depicting the nonlinearity of the number of restaurants, bars and cafes. 98

Listings

1 Normality tests for inference 72

2 Appropriate combination of polygon and point data. 103

1 Introduction

In light of the adage "location, location, location", this thesis takes a deeper look at the pricing implications of housing location, which are of considerable importance to several stakeholder groups, including homeowners, credit institutes, investors, tax authorities, and policy makers. For homeowners, the property often constitutes the single biggest asset in a private portfolio (Schulz and Werwatz, 2004, p. 37), while the the average Swiss household allocated 28 percent of total expenditure to housing in 2017 (Federal Statistical Office, 2019). Furthermore, Case, Quigley, and Shiller (2005) find a significant and large effect of housing wealth on consumption, as housing is widely accepted as collateral for consumer loans. Consequently, credit institutes have an inherent interest in being able to accurately measure the value of housing - an interest shared in particular by real estate investors seeking *alpha* and tax authorities levying property taxes. Furthermore, policy makers may have a vested interest in an improved understanding of locational quality to subsequently enhance the quality of life in their respective jurisdiction.

For the analysis of pricing implications of housing location, a data set containing the transaction prices for properties sold in New York City throughout 2018, and the respective basic structural characteristics, is augmented with locational data obtained from OpenStreetMaps as well as data available from New York City's open data portal, resulting in a model specification including 15 locational variables, two structural variables and two neighbourhood variables. To account for spatial autocorrelation and nonlinearity, a generalised additive model was employed and compared against a battery of alternative estimation techniques, including Moran eigenvector spatial filtering, geoadditive models, spatial two stage least squares and a matrix exponential spatial specification.

The results support the common belief that location is the most important factor in house prices. A residence at the best location in New York is estimated to be 3.5 times more expensive than a home at an average location with equal neighbourhood and structural characteristics. Moreover, a home with a locational value in the 75th percentile is estimated to demand a 41 percent locational value premium. The worst location is expected to be compensated by a 63 percent discount, while the 25th percentile is expected to receive a 23 percent discount. The spreads with regard to neighbourhood and structural values are considerably smaller, with the 75th percentile neighbourhood and structural values demanding an estimated 4 and 12 percent premium respectively, while the 25th percentile discounts are estimated at 6 and 13 percent, respectively. Furthermore, 58 percent of house price variation

can be explained by the locational value alone.

This thesis contributes two-fold to the second strand of housing demand literature, as categorised by Rothenberg (1991). Firstly, it capitalises on an extensive set of locational characteristics obtained from openly available sources to further the understanding of location and its importance for house prices. The lack of locational characteristics for housing data has been lamented throughout the literature and been mentioned as a critical issue impeding the progression of the literature (e.g. Calhoun (2001, p. 12)). Secondly, it simultaneously applies several econometric advances to address the shortcomings of the traditional hedonic approach proposed by Rosen (1974).

Section 2 categories the housing demand literature into four strands and points out major econometric developments that have contributed to the furthering of the second strand. After prodiving theoretical background on the housing market and the hedonic model in section 3, section 4 proceeds with discussing econometric challenges associated with the spatial analyses of housing data. Section 5 outlines the empirical analysis and presents the results, which are then discussed in section 6, along with the limitations and further research opportunities.

2 Literature review

Rothenberg (1991) categorises the literature on housing demand into the following four strands: (1) the demand for housing services, (2) the demand for individual housing attributes, (3) the demand for owner occupancy versus renting (tenure choice) and (4) the spatial allocation of households and other economic agents (Zabel, 2004, p. 17). The first strand of literature is concerned with the estimation of price and income elasticities of housing demand, which is modelled as a continous quantity that reflects the "flow of housing services" (Zabel, 2004, p. 17). It takes a macroeconomic perspective on the housing market and typically models housing demand as a function of housing, non-housing expenditures, and socioeconomic variables (Zabel, 2004, p. 17). The second strand, concerning the individual housing attributes, can further be subdivided into two distinct sub-strands. Rosen (1974) gave birth to the first sub-strand, which debates the validity of procedures for estimating the demand for specific structural characteristic (e.g. number of bedrooms), neighbourhood characteristics (e.g. crime rates), and locational attributes (e.g. proximity to public transport). The second sub-strand is centred around the median voter model and researches the demand for public goods. This sub-strand came into existence in the 1940s and 1950s with Bowen

(1943), Black et al. (1958) and Downs (1957) laying the early groundwork (Holcombe, 1989, p. 115). The third of the four overarching strands focuses on the binary choice of tenure, to rent or to buy, which plays a substantive role for policy makers who intend on increasing home ownership rates. Important contributors to the literature include Rosen (1979), Linneman (1985) and Doling (1973), who set forth an economic theory for tenure choice (Mills, 1990, p. 323). The last housing demand strand according to Rothenberg (1991) is centred around the application of the random utility model developed by McFadden (1978), who first applied it to neighbourhood choice four decades ago.

With the focus of this paper nested within the second strand of literature on housing demand (the demand for individual housing attributes), and specifically within the first sub-strand (the capitalisation of structural, neighbourhood and locational amenities), an overview of the seminal papers and major developments in this field will be given subsequently. The birth of this strand followed the seminal paper of Rosen (1974), in which he outlines a two-step hedonic model for differentiated products, such as housing, based on the hedonic hypothesis that goods are valued as the sum of the utility bearing attributes and characteristics (Rosen, 1974, p. 34). Each utility bearing characteristic is implicitly priced, referred to as the hedonic price (Rosen, 1974, p. 34). This is revealed by the observed prices of differentiated products (bundles of characteristics), such as housing, and their corresponding quantities of measurable characteristics (e.g. number of bedrooms) (Rosen, 1974, p. 34). Consequently, the revelation of hedonic prices requires sufficient variation among the composition of the bundled characteristics. Econometrically, the implicit prices can be estimated using multiple regression techniques where the observed price of housing is regressed on the measured characteristics. As such, the price of housing, $p_k(z)$, is deemed to be a function of a set of n objectively measurable characteristics $z = (z_1, z_2, ..., z_n)$, with z_i measuring the amount of the ith characteristic exhibited in the kth property (Rosen, 1974, p. 35).

The immense importance of Rosen (1974) for the housing demand literature is reflected in the countless citations it has garnered, with Google Scholar reporting over 12'500 thousand citations. Moreover, given that the approach is aimed at product differentiated markets in general, the hedonic approach has not only contributed to the housing market, but has also been applied to various different problems and markets, such as the market of washing machines, measuring the value of a good reputation in online auctions, or determining the price for clean air (Bajari and Benkard, 2005, p. 1240). The last application is of particular interest to this thesis, given its contribution to the second strand of literature. Within this second strand, the hedonic approach has also found popularity in studies of household valuation of shoreline, various measures of accessibility, crime rate, educational quality, air

quality, noise pollution, and many more characteristics (Follain and Jimenez, 1985, pp. 81-82). A more detailed review of amenities researched in the literature, that lack an explicit market, but are capitalised into house prices, can be found in section 4.3.

Particularly impactful early contributions to the second strand of literature, beyond Rosen's (1974) seminal work, include those of Freeman (1981), Follain and Jimenez (1985) and Epple (1987), who also provide useful reviews, examples or critiques of the application of the hedonic regression technique in the context of urban economics and housing valuation (Sheppard, 1999, p. 1596). Freeman (1981) reviews the theoretical contributions and examines a number of criticisms with regard to the validity of the hedonic price technique to measure demand for environmental variables in detail. While he observes that there are many aspects in which actual data diverges from the theoretical ideal and assumptions about the nature of the housing market and preferences are oversimplifications of reality, he acknowledges the substantial explanatory power the hedonic approach offers with respect to housing prices (Freeman, 1981, p. 171). Despite the theoretical validity and potential of the hedonic approach for estimating environmental demands, he believes that few studies published at that time were fully satisfactory in terms of their use of data, empirical technique and interpretation (Freeman, 1981, p. 171). After having concluded that the theoretical basis for analysis is sound, Follain and Jimenez (1985) shift the focus of the discussion towards issues in econometric applications of the hedonic approach with the intent of estimating demand rather than price. In addition to Rosen's (1974) two-step approach, they review a sample of popular empirical estimation techniques for assessing the household demand for structural and locational characteristics, consisting of the simple hedonic approach, the bid-rent approach, the Lancasterian index approach and discrete choice models. Summarising their findings, they state that the most appropriate approach depends very much on the economic issue under investigation and the nature of the available data. For instance, discrete choice models naturally lend themselves to the analysis of characteristics that are not continuous, but inherently discrete, such as the availability of type of water supply (Follain and Jimenez, 1985, p. 93). Furthermore, analysing the various empirical results used in their study, the authors infer from the substantial range of estimates reported that very little about the demand for accessibility and neighbourhood quality had been learned up until then. Lastly, albeit applauding the significant developments made to Rosen's (1974) two-step approach, Follain and Jimenez (1985, p. 101) conclude that in particular the second step suffers from issues of identification and simultaneity. Epple (1987) furthers this critique, stating that the second step, which intended to recover a demand function for each characteristic, suffers from a simultaneity problem because consumers with a high preference for a certain charac-

teristic would naturally purchase bundles that contain large amounts of this characteristic, causing inconsistent estimates in the second stage (Bajari and Benkard, 2005, p. 1240). As a consequence, the second stage of Rosen's famous hedonic framework has failed to gain popularity in the housing demand literature and is rarely used, standing in stark contrast to the ubiquitous first step regression (Bajari and Benkard, 2005, p. 1240).

A second major development in the second strand of literature was the introduction of spatial econometric models that account for spatial dependence and spatial heterogeneity. Over the past 40 years, spatial econometrics has develop from a marginal topic to mainstream econometrics, enjoying near exponential growth (Anselin, 2010). The early groundwork was layed during the mid 1970s and late 1980s by, for example, Cliff and Ord (1972), Anselin (1988) or Griffith (1989), in what Anselin (2010) describes as the period of *preconditions for growth*. It was also in this period where Cliff and Ord (1972) demonstrated the application of Moran's I test statistic for spatial autocorrelation to ordinary least squares residuals, while in the late 1980s the attention shifted towards maximum likelihood-based test statistics, such as the likelihood ratio and Lagrange multiplier statistics (Anselin, 2010, p. 11). In the 1990s, further test statistics were developed, such as Anselin's (1995) local Moran's I as a local indicator of spatial association and Getis and Ord's (1992) Gi^* as a means for hotspot analysis. The second stage of the development of spatial econometrics saw mainstream econometricians such as Bera, LeSage, and Kelejian begin to consider spatial problems (Anselin, 2010, p. 11). Major contributions were made by Kelejian and Prucha (1998) and Kelejian and Prucha (1999) with the development of the generalised moments and general method of moments estimators. Moving beyond standard linear regression models, spatial probit models, geographically weighted regression models, and spatial filtering began to see development (Anselin, 2010, p. 14), with Griffith being the major force behind the evolution of spatial filtering approaches (Griffith and Chun, 2014, 2016; Griffith, 2017, 2019). In the last stage, maximum likelihood and the generalised method of moments crystalise themselves as the two major estimators used in spatial econometrics along with their heteroskedasticity and autocorrelation consistent extensions and asymptotic properties, which were derived by Kelejian and Prucha (2007), Lee (2004), and Lee (2007) respectively (Anselin, 2010). Spatial panel regression has started to see a large gain in popularity and is actively developed by Baltagi (2008). Similarly, corresponding Lagrange multiplier tests (Baltagi, Song, and Koh, 2003) and their respective implementation in statistical software (Millo and Piras, 2012) have gained popularity. Advances in statistical software for spatial data analysis are driven by Bivand and Wong (2018) and the development of the Bivand, Pebesma, and Gomez-Rubio (2013) package. From a computational perspective, LeSage and Pace (2009) made significant

contributions both with their matrix exponential spatial specification (MESS) (LeSage and Pace, 2007) as a means of dealing with the Jacobian in the log-likelihood function, as well as their Chebyshev approximation of log-determiants of spatial weights matrices (Pace and LeSage, 2004).

These econometric developments were adopted by other authors within the second strand of housing demand literature, with, for example, Anselin, Lozano-Gracia, Deichmann, and Lall (2010) analysing the value of access to water on house prices in Bangalore, India, using a spatial hedonic approach. Other studies employing spatial hedonic approaches include Andersson, Jonsson, and Ögren (2010), who analyse the impact of noise on house prices, Kim, Phipps, and Anselin (2003), who analyse the valuation of improved air quality, Brasington and Hite (2005), who analyse the relationship between house prices and environmental disamenities, or Kim et al. (2003).

The third and most recent development within the second strand of the housing demand literature refers to the introduction of nonlinear estimation techniques. These include Generalised Additive Models (GAMs) and an extension thereof, geoadditive models. These techniques originally developed out of the fields of statistics and spatial statistics, but are starting to find their application in the applied economic literature, such as that of housing markets. GAMs were pioneered by Hastie and Tibshirani (1986, 1987) and later advanced by Wood, in particular Wood and Augustin (2002), Wood (2003), Wood (2004), Wood (2006), Wood (2008), Wood, Goude, and Shaw (2015) and Wood (2017). Geoadditive models were pioneered by Kammann and Wand (2003) within the geostatistical literature, resulting from the seamless merger of the geostatistical concept of kriging and generalized additive models. The major advantage of additive models is their nonparametric or semi-parametric nature, lending itself to the estimation of nonlinear covariate effects. Additionally, they offer a means for detecting an appropriate functional form in a data-driven manner (Hastie and Tibshirani, 1987, p. 371), offering an approach to the well-known issue of functional misspecification in the housing demand literature. Despite the advantages of nonparametric and semi-parametric approaches, they have yet to reach wide spread adoption (McMillen, 2010, p. 712).

Early uses of semi-parametric approaches in the 1990s are scattered, but can be found in, for example, Pace (1998), who apply the approach to the appraisal of real estate. A more recent study that has adopted the flexible approach within the second strand of housing demand literature is, for example, Bontemps, Simioni, and Surry (2008), who utlise a semiparametric approach to value the nonlinear influence of environmental indicators on

house prices. Similarly, Bin (2005) made use of a semi-parametric hedonic model to analyse the effects of proximity to wetlands on property values. Comparing parametric and semiparametric hedonic pricing models, Kwong, Cyr, Kushner, and Ogwang (2011) find a significantly superior fit for the latter, highlighting the improvement to Rosen's (1974) traditional approach. In similar vein, Del Giudice, Manganelli, and De Paola (2015) employed spline smoothing for an improved estimate of market prices of properties. Basile, Durbán, Mínguez, Montero, and Mur (2014) combine additive models with spatial autoregressive models, using the Lucas County house price data set, as a means of dealing with spatial dependence, spatial unobserved heterogeneity and nonlinearities, illustrating the performance of their semiparametric spatial hedonic approach.

Despite the extensive utilisation of the hedonic approach, it has garnered substantial criticism due to potential issues arising from the identification of supply and demand, market disequilibrium, the selection of variables, the choice of functional form of hedonic equation and market segmentation (Sheppard, 1999; Fan, Ong, and Koh, 2006; Selim, 2009). While the literature is attempting to address these issues with the extension of the traditional hedonic approach (e.g. semiparametric, spatial, or semiparametric spatial hedonic models), researchers have also ventured out to find alternative methodologies to accurately value individual housing attributes.

Arguea and Hsiao (2000, p. 104) argue that, in addition to the hedonic model, the discrete choice model has established itself in the housing demand literature, in particular in the fourth strand of the literature. The seminal paper of McFadden (1978) on modelling the choice of residential location gave birth to this strand, which stipulated that "the economically rational consumer will choose a residential location by weighting the attributes of each available alternative and by selecting the alternative that maximizes utility" (McFadden, 1978, p. 72). For this purpose McFadden (1978) introduces a family of probabilistic choice models, including the commonly employed joint multinomial logit model of choice, which permits a structure of similarities across alternatives. Despite discrete choice models being better suited to analyse the tenure choice (strand 3) or the impact of location and neighbourhood characteristics on the household location choice (strand 4) (Uyar and Brown, 2005, p. 428), they have also been applied to determine the willingness to pay for housing attributes (strand 2) (e.g. (Palmquist and Israngkura, 1999), (Earnhart, 2001), (Uyar and Brown, 2005), (Masiero, Heo, and Pan, 2015)). For further literature, see Lerman (1979), Friedman (1981), Ellickson (1981), Follain and Jimenez (1985),Quigley (1985), Nechyba and Strauss (1998) and Chattopadhyay (2000).

More recently, the literature has also started reviewing artificial neural networks as an alternative methodology to hedonic regression models (see Wong, So, and Hung (2002); Selim (2009); Limsombunchai (2004)). Selim (2009) acknowledges the great popularity of hedonic methods using multiple regression techniques for the analysis of locational value of housing. However, due to the potential nonlinearity in the hedonic functions, spatial dependence between observations or discontinuities in the data, Selim (2009) suggests an artificial neural network as an alternative and demonstrates that said methodology may provide better prediction results of house prices in Turkey. Furthermore, the neural network approach avoids having to make assumptions about the functional form, as well as the distributions of variables and error terms, which is one major pitfall of the hedonic approach (Limsombunchai, 2004, p. 7). Despite Limsombunchai (2004), Selim (2009), and Wong et al. (2002) all employing feed-forward networks with supervised learning supported by back-propagation, Wong et al. (2002) were only capable of reducing the standard deviation and failed to emulate Selim (2009) and Limsombunchai's (2004) improvements in accuracy (measured as the mean squared error) compared to the hedonic method. Furthermore, Wong et al. (2002, p. xxi) mention long computation time and absence of indicators of implicit prices of property characteristics as major drawbacks of the neural network approach.

Fan et al. (2006) utilise a non-parametric decision tree approach in their study on Singapore's resale public housing market. Fan et al. (2006) list the following advantages of the decision tree approach over the traditional hedonic approach. First, tree approaches are capable of dealing with classification as well as regression problems without the need of many assumptions regarding the distribution of the data or the built model. Secondly, the established decision tree allows for a straightforward interpretation allowing to identify major attributes in the observations. This is a major difference to the artificial neural network approach utilised by Selim (2009) Limsombunchai (2004) and Wong et al. (2002), where the model is treated as a black box. Thirdly, a major strength of the tree approach lies in its ability to analyse linear and nonlinear relationships between dependent and independent variables, as well as identifying the most significant explanatory variables. Lastly, tree approaches are able to deal with continous and categorical variables in classification and regression. (Fan et al., 2006, p. 2302)

Despite these formidable alternatives, the hedonic approach remains the most popular, albeit its flaws. Arguea and Hsiao (2000, p. 105) as well as Dubin (1988) mention the lack of universally accepted measures of neighbourhood quality within the literature as a significant issue with hedonic regressions that yet remains unsolved. Dubin and Sung (1990) attempt to address this issue by determining a set of neighbourhood characteristics that best describe

the variation in house prices. Their results indicate that neighbourhood characteristics, such as race and socio-economic variables, are more important than the quality of public services. However, they fail to include environmental variables, such as air pollution or number of restaurants in their analysis. As such, this thesis hopes to contribute to the determination of a universally accepted measure of neighbourhood and locational quality by accounting for and empirically analysing an array structural, neighbourhood, and locational variables simultaneously.

In a critique towards the literature of neighbourhood characteristics, Arguea and Hsiao (2000, p. 105) stipulate that variables measuring neighbourhood quality are more likely to be chosen due to the availability of data rather than an optimal reflection of consumers' perception of the amenity. This seems to be the case with the above mentioned study by Dubin and Sung (1990). Furthermore, one can argue that consumers are only vaguely capable of categorising neighbourhoods into different classes of, for example, safety. Much more obvious to potential homeowners is the accessibility of public transport, quantity and quality of nearby restaurants, availability of sport facilities or quality and proximity to recreational spaces, such as lakes or parks. Therefore, this thesis puts a focus on indicators of locational quality that are accessible by the public through mapping services, such as Google Maps or OpenStreetMaps.

Moreover, a common theme throughout the literature remains the need for better data (Calhoun, 2001, p. 12). In vein of these circumstances, this thesis adds to the literature by including novel data on locational amenities, such as quantity of restaurants, cafes and bars within a five minute walking distance, amount of park area in close proximity, or individual distances to fitness centres, shopping centres, playgrounds, supermarkets, and convenience stores. This vast array of detailed data, outlined in detail in section 4.3, is obtained from OpenStreetMaps and is to my knowledge the most exhaustive list of locational variables analysed in the literature.

Lastly, this thesis contributes to the growing literature of studies that apply relatively recent advances in econometric techniques to address the shortcomings of the traditional hedonic approach. Specifically, the techniques used in this thesis include eigenvector spatial filtering, GAMs, and geoadditive models to address the main issues of spatial autocorrelation, functional misspecification and nonlinearity.

3 Theoretical background

3.1 The housing market

From an economic perspective, the housing market is regarded as a unique market and is often described as completely product-differentiated, given that each product sold in the market (every house or piece of land) is unique. While neighbouring properties may act as close substitutes, the scarcity of land means that no two products will have exactly the same location. This is different to the majority of the economic markets, such as those for cars or washing machines, where a number of models are considered close, albeit not perfect, substitutes. Consequently, these markets are considered to be partially differentiated. The contrast is even stronger when comparing the housing market to traditional commodity markets, such as the markets for oil, corn, or metal, in which homogenous goods are traded in bulk quantities. In the housing market, no two plots are equivalent. (DiPasquale and Wheaton, 1996, p. 35)

Given the complete product-differentiation of the housing market, it becomes difficult to speak about supply and demand at any particular location. By definition, the supply of land is fixed at each location and thus considered price inelastic. The demand, on the other hand, is considered to be sensitive with respect to its price and thus considered elastic. This concludes from the observation that adjacent plots can be considered close substitutes and thus offer comparable value. Recognising these distinctive features of the housing market, economists developed a simple approach for valuing land and housing prices, where the cost of ownership of a specific plot of land is equivalent to its locational advantages. Consequently, understanding the advantages and disadvantages of locations and how consumers evaluate them is of paramount importance to understanding the spatial pattern of land and housing prices. Moreover, this theory of compensating differentials assumes that demand considerations determine the relative value of land and consequently housing at differing locations, while the supply of land determines the overall level of prices. (DiPasquale and Wheaton, 1996, pp. 35-36)

According to the standard urban real estate models of Alonso et al. (1964), Rosen (1979) and Roback (1982), house prices reflect a spatial equilibrium, where prices are determined by local wages and amenities, such that local heterogeneity is natural (Glaeser, Gyourko, Morales, and Nathanson, 2014, p. 45). Glaeser et al. (2014, p. 45) go as far as stating that "less than 8% of the variation in price levels and barely more than a quarter of the variation

in price changes across cities can be accounted for by national, year-specific fixed effects". Consequently, an abundance of local variation remains unexplained by macroeconomic variables such as interest rates and national income (Glaeser et al., 2014, p. 45).

Furthermore, Glaeser et al. (2014, p. 45) stipulate that the urban approach predicts that housing markets are local in nature, rather than national. This argumentation is rationalised by the principle of arbitrage, where consumers of housing would take advantage of favourable price differentials for a set of similar housing characteristics within a local market. On the contrary, price differentials for a set of similar housing characteristics in the national market may not be exploited due to the additional costs and uncertainty of finding new employment as well as the intangible costs revolving around the loss of existing social networks or emotional connections to the place of birth. Given these two reasons, it is reasonable to analyse the value of housing location in (1) a cross-sectional setting and (2) for a local area, such as a city.

3.2 The hedonic model

The work of Griliches (1961), Lancaster (1966), Griliches (1971), and Rosen (1974) laid the foundation for hedonic analysis and techniques for dealing with "commodity heterogeneity" (Sheppard, 1999, p. 1595). According to Rosen (1974, p. 34), the hedonic hypothesis states that "goods are valued for their utility-bearing attributes or characteristics". Moreover, a class of differentiated products, such as housing, is completely described by a vector of objectively measured characteristics. Each attribute or characteristic is governed by a hedonic price, which is revealed to economic agents from observed prices of differentiated products and the specific amounts of characteristics associated with them. (Rosen, 1974, p. 34)

Within a city, hedonic models are particularly valuable for mass appraisal of house prices as well as the valuation of non-market amenities that contribute to the price of housing (Anselin and Lozano-Gracia, 2009, p. 1214). Anselin and Lozano-Gracia (2009, p. 1214) explains the rationale for the second benefit as, in an efficient market, more desirable amenities, such as clean air, access to parks, or proximity to subway stations, should be capitalised into the value of the house and hence, ceteris paribus, houses with more desirable amenities should be more expensive. On a more theoretical level, Rosen (1974) describes the marginal prices of housing attributes as a result from an equilibrium between the demand and supply of those housing attributes (Linneman, 1980, p. 49). More specifically, the marginal

prices represent the intersections of the demand and supply curves of the specific attribute at various different quantities (Linneman, 1980, p. 49). At each location on the resulting hyperplane the amount of housing offered by providers matches the amount demanded by the population of consumers choosing the corresponding set of characteristics (Rosen, 1974, p. 35). Each individual's choice of a set of characteristics is deemed to be in accordance with utility maximising behaviour (Rosen, 1974, p. 35), such that no individual, buyer or seller, is better off choosing a different set of characteristics. Accordingly, the hyperplane of prices contains the market equilibrium prices for each and every set of characteristics.

Sheppard (1999, p. 1595) describe the hedonic analysis of the housing market by making the following analogy. Assume you are tasked to analyse the food market, but are forbidden to enter the supermarket to read the prices customers are willing to pay for individual items. Rather, you are only able to get a glimpse at the customers' shopping baskets at the cashier and the total amount they paid. You may not be able to accurately determine the amount of each food item that was purchased and some food items may have been hidden beneath other items. A similar situation is present when analysing the housing market. We are faced with imprecise observations of which attributes customers purchase in the housing market. As such, the hedonic approach deconstructs explicit markets of housing into implicit markets of housing attributes and characteristics, and argues that, while the bundle of characteristics sold on the explicit market may vary significantly in prices, their individual components share a common price (or at least a common price structure). (Sheppard, 1999, p. 1596)

Advantages of the hedonic model. One major advantage of the hedonic approach is its ability to estimate the capitalisation of such non-market goods (Bayer, Keohane, and Timmins, 2009; Anselin and Lozano-Gracia, 2009), given sufficient variation in the composition of the bundled products (Luttik, 2000; McMillen, 2010, pp. 165, 133). This property of hedonic regressions is critical to the objective of this thesis, namely to value the contribution of location to the price, or the rather estimated price of a house. Given a market value for each locational variable, the locational value can then be determined as the aggregate value of the individual locational variable predictions. Such broken down prices estimations are not possible with other popular valuation techniques, such as repeat sales techniques, discounted cash flow methods, or direct capitalisation approaches. Essentially, the hedonic approach, estimated by multiple regression or extensions thereof, is a data-driven sales comparison approach, in which the model estimates the variable coefficients and as such is able to provide artificial comparables (Mark and Goldberg, 1988, p. 93).

A second advantage of the hedonic approach is that it merely requires property prices,

composition of housing attributes and a proper specification for the functional relationships to discern the individual impact of a single housing characteristic on the overall price of the dwelling (Calhoun, 2001; Chau and Chin, 2003). While substantial difficulty pertaining to data availability may arise in practice, the theoretical data requirements are sparse and for the majority of relevant characteristics data can be obtained from public sources for certain housing markets. Another major advantage of the hedonic approach is its ubiquitous presence in the housing demand literature and the countless contributions to its theoretical development. With the first appearance of the hedonic concept as early as Waugh (1928), the approach is nearing a century of rigorous discussions and developments, making it one of the most proven econometric techniques in the literature. Furthermore, the econometric estimation techniques for hedonic regressions are ubiquitously implemented throughout most, if not all, statistical software, along with various extensions, such as spatial hedonic models, with estimation techniques implemented in the *spdep* package (Bivand et al., 2013) in R (R Core Team, 2013), or the *splm* package (Millo and Piras, 2012) for spatial panel linear models.

Disadvantages of the hedonic model. However, the approach is not without its imperfections. In the absence of theoretical guidance and a lack of a unanimously accepted measure of housing quality, in particular with regard to locational quality, the empirical literature has resorted to a plethora of different specifications and explanatory variables (Arguea and Hsiao, 2000, p. 105). In addition, it is argued that the use of the respective explanatory variables may be chosen due to their availability rather than their reflection of consumers' perception of valued amenities. A consequence of the use of such imperfect measures of consumer's perceptions of amenities may be a misspecification of hedonic models, leading to biased and inconsistent estimates of housing prices (Arguea and Hsiao, 2000, p. 105). On the other hand, viewing such variables as proxy variables for latent amenities raises the familiar measurement error, causing the same issues of bias and inconsistency (Arguea and Hsiao, 2000, p. 105).

In addition to these theoretical and practical issues, there are a multitude of statistical issues associated with estimating the impact of location on house prices. These issues include multicollinearity, endogeneity, heteroskedasticity, spatial autocorrelation, functional misspecification and nonlinearity, and variable misspecification. The following section will discuss these issues and elaborate on how they have been addressed in this thesis.

4 Econometric issues

4.1 Spatial autocorrelation

Tobler's first law of geography states that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Consequently, geographically near houses will tend to exhibit similar prices or characteristics compared to more distant ones, giving rise to the phenomenon termed spatial autocorrelation in the spatial econometric and spatial statistics literature (Anselin, 1988; Griffith and Chun, 2014).

Spatial dependence or spatial autocorrelation is deemed a "special case of cross-sectional dependence in which the covariance structure of observations at different locations is subject to spatial ordering" (Anselin and Lozano-Gracia, 2009, p. 1213). Importantly, this gives rise to a situation wherein observations lack independence throughout space (Anselin, 1988; Dubin, 1988). Spatial dependence can be caused by a variety of measurement problems, including the arbitrary delineation of spatial units (e.g. census tracts, postal codes, or district boundaries), issues stemming from spatial aggregation, and the existence of spatial externalities and spill-over effects (Anselin, 1988, p. 8). In addition to measurement error, "inherent spatial organisation and spatial structure of phenomena" (Anselin, 1988, p. 8) can give rise to cross-sectional data subject to spatial dependence. Examples for the latter cause of spatial dependence in respect to housing prices include, nearby properties having an increased likelihood of exhibiting similar structural features due to being developed during the same time (i.e. same architectural style and materials used), and sharing of locational amenities (Bourassa, Cantoni, and Hoesli, 2007, p. 144). As such, spatial dependence can occur due to theoretical considerations as well as statistical issues.

The Moran coefficient, or *Moran's I*, is the most commonly employed quantitative measure of spatial autocorrelation, and considered to be the spatial equivalent of Pearson's correlation coefficient (Griffith and Chun, 2014, p. 1478). It is calculated as the spatially weighted correlation coefficient, with the mathematical formula given by

$$I = \frac{N \sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{(\sum_i \sum_j w_{ij}) \sum_i (x_i - \bar{x})^2} \tag{1}$$

and its graphical representation portrayed by the Moran scatterplot. The exact range of the spatial correlation coefficient is determined by the eigenvalues of the spatial weights

matrix W, but is often between negative one and positive one, with negative values indicating negative spatial autocorrelation and positive values respectively positive autocorrelation (Griffith and Chun, 2014, p. 76). The expected value of *Moran's I* is given by $\mathbb{E}(I) = \frac{-1}{N-1}$ under the null hypothesis of zero spatial autocorrelation (Griffith and Chun, 2014, p. 76). Despite the rather complicated derivation of the standard deviation of *Moran's I*, it can be approximated well by $\sqrt{\frac{2}{\sum_{i=1}^{n}\sum_{i=1}^{n} w_{ij}}}$, such that significance tests using z-statistics $\frac{I-\mathbb{E}(I)}{\sqrt{Var(I)}}$ can be calculated under the assumption of normality (Griffith and Chun, 2014, p. 1480).

Geary (1954) formulated an alternative measure of spatial autocorrelation and is considered to be the second most popular measure, behind *Moran's I*. While *Moran's I* requires the differences to the overall mean, *Geary's C* is subject to the pairwise squared differences and consequently more sensitive to local autocorrelation patterns. The formula is given by

$$C = \frac{(n-1)\sum_i \sum_j w_{ij}(x_i - x_j)^2}{2(\sum_i \sum_j w_{ij}) \sum_i (x_i - \bar{x})^2} \tag{2}$$

with an expected value $\mathbb{E}(C) = 1$, while a value smaller than one indicates positive spatial autocorrelation and values larger than one indicate negative spatial autocorrelation (De Jong, Sprenger, and Van Veen, 1984, p. 18). Therefore, the interpretation of *Geary's C* is inverse of that of *Moran's I*. Similarly, a significance score can be calculated by $\frac{C-1}{\sigma_C}$, given that the statistic is near normally distributed (Geary, 1954, p. 115).

While both *Moran's I* and *Geary's C* are indicators of global spatial autocorrelation, Anselin (1995) set forth the concept of local indicators of spatial association (LISA), allowing for the decomposition of global indicators. Its development was motivated by the highly unrealistic assumption of stationary or structural stability over space in the face of large number of spatial observations (Anselin, 1995, p. 94). Consequently, a LISA is calculated for each individual observation, indicating the extent of significant spatial clustering of similar values around each observation (Anselin, 1995, p. 94). Part of the LISA family of statistics, *local Moran's I* is the local analog to the most popular global statistic, defined for an observation i as

$$I_i = z_i \sum w_{ij} z_j \tag{3}$$

where z_i, z_j are mean centred observations (Anselin, 1995, p. 98). The expected value is given by $\mathbb{E}(I_i) = -w_i/(n-1)$, where w_i is the sum of the row elements, $\sum_j w_{ij}$ (Anselin, 1995, p. 98). Similar to the global statistics, a test for significant local spatial association may be conducted using the expected value $\mathbb{E}(I_i)$ and its variance $Var(I_i)$. Alternatively

to the rather complex derivation of the variance (see Anselin (1995, Appendix A.)), a conditional randomisation approach may be taken in order to empirically obtain an estimate for the second moment, which can then be used in significance testing. However, due to the correlation between I_i and I_k, and the associated problem of multiple comparison, the conventional interpretation of significance is inapplicable (Anselin, 1995, p. 98). In order to obtain adequate significance levels, Anselin (1995) suggests using either a *Bonferroni* adjustment[1] of α/n, with n being the number of observations, or Šidák's (1967) proposed adjustment of $1-(1-\alpha)^{\frac{1}{n}}$ (see `?p.adjust` in `R`).

The calculation of all above statistics requires the definition of a spatial weights matrix W, where w_{ij} corresponds to the measure of spatial proximity between observation i and j. There are a plethora of different specifications used which can be broadly grouped into contiguity based matrices and distance based matrices. The former is governed by the existence of common borders between regions which are indicated in binary form (1 for the existence of a shared border and 0 otherwise)[2]. Naturally, this approach lends itself to aggregate data, such as districts, regions, or countries. For point data, distance based approaches (often in combination with threshold levels) are more obvious given the lack of borders, with a wide variety of definitions available. Examples include k-nearest neighbours, inverse distance, negative exponential, or kernel-based specifications, such as triangular, Epanechnikov or Gaussian (the mathematical specifications are accessible in the appendix in section C.6).[3] Due to the numerous options of specification and lack of formal guidance on the "correct" spatial weights matrix (Anselin and Lozano-Gracia, 2009, p. 1220), this thesis resorts to conducting robustness tests with an Epanechnikov and Gaussian kernel spatial weights matrix, as well as a negative exponential spatial weights matrix in addition to the standard inverse distance matrix when testing for spatial autocorrelation in the residuals of the models.

Measures of spatial autocorrelation can be utilised as diagnostic tools to detect functional misspecification in linear regression models, where spatial autocorrelation present in the residuals is suggestive of nonlinearities (Griffith and Chun, 2014, p. 1479). Cli and Ord

[1] Anselin (1995, p. 96) argues that the *Bonferroni* adjustment may be too conservative for LISA, and mentions that the optimal adjustment still requires further research.

[2] Further distinctions within contiguity based matrices are made with respect to the definition of a neighbour, with rook contiguity, bishop contiguity and queen contiguity as opportunities. However, due to the vast majority of regions being depicted by polygons, as opposed to rasters, queen contiguity assumed to be by far the most applied in practice.

[3] For a more detailed explanation on connectivity in space, please refer to Anselin (1988, p. 16).

(1981, p. 209) demonstrate the reduction of spatial autocorrelation in the residuals caused by a more accurate modelling of nonlinear relationships between dependent and independent variables. Moreover, spatial autocorrelation in the residuals may also indicate the omittance of variables with considerable explanatory power (Griffith and Chun, 2014, p. 1479). That is because the effects of omitted variables will be represented in the error term, and, given that house prices tend to exhibit spatially similar values, the error terms will likewise be spatially autocorrelated (Dubin, 1988; Griffith and Chun, 2014, pp. 468, 1479).

Traditional hedonic models, as set out by Rosen (1974), have commonly been estimated using ordinary least squares (OLS). However, in the presence of spatial autocorrelation as measured by the above methods and modelled by a spatially lagged dependent variable, the OLS estimator is biased and inconsistent due to the correlation between the spatial variable and the error term (Anselin, 1988, pp. 58, 82). For an econometric exposition see section C.3 in the appendix. In the face of inapplicability of OLS, spatial hedonic models are commonly estimated using maximum likelihood, while alternative approaches include instrumental variables estimation (Anselin, 1988, p. 81), eigenvector spatial filtering with subsequent OLS (Griffith and Chun, 2014), or generalised two stage least squares with a generalised moments estimator for the autoregressive parameter (Kelejian and Prucha, 1998, 1999). Subsequent sections will give a short explanation of the underlying concepts of the approaches, given their use for the selection of the best performing model in this thesis.

Maximum likelihood. Based on the assumption that the distribution of the residuals is known, maximum likelihood approaches can be used to estimate the model parameters, including the spatial autocorrelation parameter commonly denoted as ρ. However, the inclusion of a spatially lagged dependent variable as an explanatory variable causes a system of highly nonlinear first order conditions, which does not have an analytical solution, and hence must be solved by numerical methods (Anselin, 1988, p. 64). For spatial autoregressive models $y = \rho W y + X\beta + \epsilon$, among others, the first order conditions can be constructed as a concentrated likelihood function, which only remains nonlinear in one parameter, allowing the estimation of the parameters of interest in a step-wise procedure (see section C.4 in the appendix for a detailed explanation of the concentrated log-likelihood approach).

Instrumental variables. Endogeneity in traditional hedonic regression models are often dealt with using instrumental variables, where an appropriate instrument is correlated with the explanatory variable, but not with the error term (Mills and Patterson, 2009; Anselin, 1988, pp. xxviii, 82, 85). The main difficulty of applying two stage least squares in a spatial setting often lies with finding a set of instruments that fulfil the above criteria.

In regard to this issue, a major contribution was made by Kelejian and Prucha (1998), who suggested that the set of instruments $X, WX, W^2X, \ldots$ would satisfy the necessary requirements. Their suggestion is motivated by the fact that $\mathbb{E}(WY|X)$ can be seen to be formed as a linear combination of the columns of the matrices $X, WX, W^2X, \ldots$ using the Neumann series expansion:

$$\begin{aligned} \mathbb{E}(y_n) &= (I - \rho W)^{-1} X\beta \\ &= [\sum_{i=0}^{\infty} \rho^i W^i] X\beta, \end{aligned} \tag{4}$$

where $W^0 = I$ (see section C.5 in the appendix for an econometric exposition of the Neumann series expansion) (Kelejian and Prucha, 1998, p. 104).

Matrix Exponential Spatial Specification The matrix exponential spatial specification (MESS) was introduced by LeSage and Pace (2007) as a way of modelling spatially dependent data, providing an alternative to the spatial autoregressive process. At its core, the MESS specification replaces the geometric decay[4] throughout space, represented by the spatial autoregressive process, with exponential decay throughout space (LeSage and Pace, 2009, p. 237). However, the objective of the MESS specification is not the more accurate representation of the spatial decay process, but rather the technical simplification of the estimation, improving numerical performance and leading to a computational advantage[5], while providing similar estimates to the geometric decay specification. Specifically, it eliminates the need of calculating the inverse of the log-determinant in the maximum likelihood estimation and Bayesian model estimates, greatly reducing the computational resources required in estimations with large data sets (LeSage and Pace, 2009, p. 237).

Eigenvector Spatial Filtering. Eigenvector spatial filtering (ESF) offers a simpler approach for dealing with omitted variable bias in the spatial setting by constructing a spatial filter with eigenvectors emulating the underlying spatial autocorrelation pattern, serving as a proxy for missing spatially dependent variables (Griffith and Chun, 2016, p. 3). The synthetic proxy variables represented by the eigenvectors are derived from an adjusted spatial weights matrix and then added as control variables to an OLS regression specification (Griffith and Chun, 2016, p. 3). The adjustment of the spatial weights matrix W occurs according to $(I_n - \iota \cdot \iota^\intercal/n)W(I_n - \iota \cdot \iota^\intercal/n)$, with I being an $n \times n$ identity matrix and

[4] See section C.5 in the appendix for an exposition of the geometric decay represented by the spatial autoregressive process.

[5] See section C.7 in the appendix for an exposition of the computational advantage.

ι an $n \times 1$ vector of ones. The corresponding eigenvectors have the amicable properties of being orthogonal, due to the symmetry of W, and uncorrelated, due to the projection matrix $(I_n - \iota \cdot \iota^{\intercal}/n)$ (Griffith and Chun, 2016, p. 3). Furthermore, each eigenvector can be interpreted to represent distinct aspects of the spatial autocorrelation.

Griffith and Chun (2016) demonstrate how eigenvector spatial filtering (ESF) alleviates omitted variable bias, as measured by the Ramsey regression equation specification error test (RESET), a test introduced by Ramsey (1969) and validated by Shukur and Mantalos (2004). McCord, McCord, Davis, Haran, and Bidanset (2019) also find that the ESF approach increased model stability and reduced misspecification error. Moreover, Pace, LeSage, and Zhu (2013, p. 355) were also able to corroborate the findings that ESF does a good job of reducing the bias caused by spatial dependence. Moreover, a consequence of filtering out spatial dependence from the sample data is that standard linear model specifications may be applied, proving especially valuable where spatial alternatives are yet to be developed or estimation thereof renders computationally expensive or theoretically complex (Pace et al., 2013, p. 354). Due to these benefits, the popularity of spatial filtering approaches to address spatial dependence as an alternative to the conventional spatial regression models is increasing in the housing demand literature according to Pace et al. (2013, p. 353). For example, McCord et al. (2019) analyse the local variation and spatial heterogeneity of house prices in the Belfast housing market using the eigenvector spatial filtering approach. Xiao, Chen, Li, Yu, Chen, and Guo (2017) have employed ESF in their analysis of the effects of amenities on house prices in Beijing. Lastly, Chun (2014) extend the ESF method to a generalised linear mixed model (GLMM) to model space-time crime incidents in Texas.

4.2 Functional misspecification

Spatial hedonic models are no silver bullet. In addition to spatial dependence, locational data may suffer from a variety of other issues, such as heteroskedasticity, non-normality, multicollinearity, or misspecification for extreme values of the dependent variable (Pace, Barry, Slawson, and Sirmans, 2004, p. 213). For example, Su and Jin (2010) demonstrated that a lack of consideration of nonlinearity in spatial hedonic models can result in poorly estimated spatial parameters. The following section deals with issues arising from a functional misspecification of the hedonic form. Heteroskedasticity is tackled using the variance stabilising and normalising Box-Cox transformation, while the violation of linearity is addressed by semi-parametric GAMs, which in addition to addressing nonlinearity also reduces residual

spatial autocorrelation, due to reasons discussed above.

One important source of estimation bias, that doesn't receive the attention it deserves, is the functional misspecification of the hedonic regression function (Linneman, 1980, p. 48). There are theoretical reasons as to why determining the correct functional form is difficult. In general, Rosen (1974) stipulated that the marginal prices of housing attributes result from an equilibrium between the demand and supply of those housing attributes (Linneman, 1980, p. 49). More specifically, the marginal prices represent the intersections of the demand and supply curves of the specific attribute at various different quantities (Linneman, 1980, p. 49). Consequently, in a situation where an attribute is produced with constant returns to scale and under the assumption of a competitive housing market, the market clearing price is determined by the supply. Furthermore, if each attribute were to be produced by an independent production process, one would be able to a priori specify the marginal price functions as linear equations (Linneman, 1980, p. 49). However, these assumptions do not hold in practice. Many of the relevant housing attributes are not producible (e.g. distance to the nearest subway station) while others tend to be subjected to a joint production process (e.g. number of bedrooms and toilets) (Linneman, 1980, p. 49). Consequently, it is not possible to determine a precise functional form of hedonic price functions by theoretical reasoning (Halvorsen and Pollakowski, 1981, p. 37), resulting in a variety of functional forms employed throughout literature (Linneman, 1980, p. 52). The most common functional specification of hedonic house price regressions include linear, semi-logarithmic and log-log specifications (Anselin and Lozano-Gracia, 2009, p. 1218), while the literature has largely disregarded nonparametric transformations.

In general, the motivations of applying transformations to the variables are multitude. One major motive is that inference on OLS coefficients benefits from errors ϵ being normally distributed with a constant variance. In situations where this assumption cannot realistically be made[6], researchers have focused on finding transformations of the dependent variable

[6] To determine whether a variable violates the assumption of normality in the first place, one can apply the Kolmogorov-Smirnov or the Shapiro-Wilk test, where a p-value above .05 indicates that the distribution does not differ significantly from the standard normal distribution (Osborne, 2010, p. 2). However, the Shapiro-Wilk test is limited to a sample size of 5000 observations in *R* (see `?shapiro.test()`). Other tests available to the researcher include the standard Jarque-Bera test (`jb.norm.test` in the `normtest` package) or the Anderson-Darling test for normality (`ad.test` in the `nortest` package). Furthermore, with large data sets, the slightest deviation from normality will be reported as significant (see section B.2 in the appendix for empirical evidence), and hence the explanatory power of such tests is only limited, given that real world data is never perfectly normally distributed. A much more sensible approach is to plot a histogram of the variable in order to determine a resemblance with the normal distribution, such that p-values can still be relied on.

such that the assumption may hold nevertheless and consequently OLS be appropriately applied (Box and Tidwell, 1962, p. 531). Furthermore, much emphasis has been put on transformations of $\mathbb{E}(y)$ which may stablise the variance or improve the linearity of the explanatory variables (Box and Tidwell, 1962, p. 531).

Box and Cox (1964) developed a statistical approach to determine a functional specification that provides an optimal fit in terms of log-likelihood (Linneman, 1980, p. 52), while Osborne (2010, p. 1) described it as a potential best practice when normalising data or stabilising variance is desired. Their Box-Cox transformation represents a family of power transformations that is given by

$$y_i^\lambda = \frac{y_i^\lambda - 1}{\lambda} \text{where} \lambda \neq 0 y_i^\lambda = \ln(y_i) \text{where} \lambda = 0 \tag{5}$$

and thus includes a plethora of traditional transformations, such as:

- $\lambda = 1.00$: no transformation,
- $\lambda = 0.50$: square root transformation,
- $\lambda = 0.33$: cube root transformation,
- $\lambda = 0.00$: natural log transformation,
- $\lambda =$ -1.00: inverse transformation, etc.

Many authors in the literature employ the Box-Cox transformation in their analyses as a means to obtain a better model specification. A list of authors that all constructed a superior model using the Box-Cox transformation includes Can (1990), Teugels and Vanroelen (2004), Griffith (2019), Fischer (2014), Davidson and MacKinnon (1985), Cropper, Deck, and Mcconnell (1988), Wong et al. (2002), Rasmussen and Zuehlke (1990), Osborne (2010) and Mok, Chan, and Cho (1995).

The objective of transforming independent variables is to reduce the function in these transformed variables to as simple a form as possible (Box and Tidwell, 1962, p. 532). In most cases where economic theory suggests the existence of network dependence and spatial spillovers, it fails to provide a well structured model, leading to significant uncertainty of the functional form (Basile et al., 2014, p. 230). Early attempts at finding statistically optimal transformations include the parametric Box-Tidwell power transformations of the

independent variables as outlined in Box and Tidwell (1962).[7] In fact, the premise in the applied literature is that a linear structure coupled with some transformation of the data is sufficient to model the underlying data generating process (Basile et al., 2014, p. 230). However, Basile et al. (2014, p. 230) list 18 studies in the fields of regional growth, urban agglomeration economies, urban environment, land prices, urban sprawl, and house prices that have detected strong nonlinearities, underpinning the optimistic views that parametric transformations are sufficient. More modern approaches makes use of locally weighted regression, scatterplot smoothers, or various different splines.

Generalised additive models. In 1986, Hastie and Tibshirani (1986) published their seminal paper on GAMs, in the wake of the academic literatures' tendency to move away from linear models and towards semi-parametric and nonparametric models. GAMs have experienced continous development throughout the years, with Simon Wood being one of the most important contributors to the econometric success of the method (see Wood and Augustin (2002); Wood (2003, 2004, 2006, 2008); Wood et al. (2015); Wood (2017)). In particular, Wood (2017) is establishing itself as the theoretical reference for GAMs in general. As opposed to the linear combination of a set of covariates X_1, X_2, ..., X_p, GAM proposes to estimate the response variable by a set of smoothing functions of the respective covariates, written as $Y = s(X_1) + s(X_2) + \ldots + s(X_p) + \epsilon$. In its full form, a GAM, considered to be an extension of generalised linear models (GLMs), can be written as $g(\mu_i) = \mathbf{X_i^*}\beta^* + s_1(x_{1i}) + s_3(x_{3i}) + s_3(x_{3i}) + \ldots$ where $\mathbb{E}(Y_i) \equiv \mu_i$, $g(\cdot)$ is a known monotonic link function, $\mathbf{X_i^*}$ is the ith row of the model matrix for the strictly parametric covariates, including the categorical variables and the intercept, with the corresponding parameter vector β^*, and $s_j(\cdot)$ is a respective smoothing function (which need not be restricted to one covariate). The smoothing function may take on various forms, such as running mean, running median, running least squares line, kernel estimate, or splines (Hastie and Tibshirani, 1986, p. 297). The resulting smoothing function can then be used as a means of descriptive statistics to explore nonlinear relationships between explanatory variables and dependent variables, as might occur with distance based measures for locational quality (Geniaux and Napoléone, 2008, p. 101), as diagnostic tools to identify appropriate parametric covariate transformations, or directly for predictions (Hastie and Tibshirani, 1986, 1987). Especially in light of the lack of theoretical guidance with respect to the functional form, GAM offers the significant advantage of letting the data dictate the functional form (Hastie and Tibshirani, 1987, p. 371), addressing functional misspecification in a novel way compared to traditional

[7] The corresponding function in *R* is *box.tidwell*, located in the *car* package.

a-spatial and spatial parametric hedonic models.

A popular choice of splines are cubic splines. Constructing the smooth functions $s_j(\cdot)$ using cubic splines requires the choice of knots, where piecewise cubic polynomials are joined continuously and smoothly, determined by the second derivative (Wood and Augustin, 2002, p. 160). In practice, the choice of the number of knots as well as their locations can have substantial impact on modelling results, and hence can pose a critical issue with cubic splines (Wood and Augustin, 2002, p. 160). Penalised regression splines offer an alternative approach that avoid the researcher from specifying the number of knots, increasing the level of automation and the degree of data-drivenness. While the approach utilises a relatively large number of knots, a penalty term[8] is introduced in order to avoid excessive wiggle and consequently overfitting of the models (Wood and Augustin, 2002, p. 161). The amount of wiggle can be quantified by second order derivatives of the basis functions, and consequently a popular penalty term is given by $J(f) = \int [f''(x)]^2 dx$, such that the smoothing function for penalised splines may be represented by a basis expansion and the quadratic smoothing penalty (Wood and Augustin, 2002, p. 161).

Few studies to date have simultaneously considered spatial effects and functional form transformations, arguably due to its challenging nature (Pace et al., 2004, p. 197). While GAMs are prominently applied in other literature (e.g. biology and medicine) (Marra and Radice, 2010), Pace (1998) provided an early application thereof in the real estate literature, specifically for the appraisal of properties, outperforming naive parametric and polynomial models. Several years later, Pace et al. (2004) applied GAMs to housing data for Baton Rouge, where they were able to reduce the interquartile range of residuals by 38 percent in the untransformed variable space compared to the simple model. Moreover, the pattern of residuals in the transformed space improved significantly and effectiveness of the spatial transformation was increased (the spatial autoregressive parameter rose from 0.51 to 0.58) by reducing the functional form misspecification. Lastly, Basile et al. (2014) address the issues of spatial dependence, spatial unobserved heterogeneity and nonlinearities simultaneously using combinations of geoadditive and spatial hedonic models, resulting in, for example, a spatial autoregressive semi-parametric geoadditive model. Applying their innovative models to the Lucas County housing data, they were able to report significant model performance measured by a decrease in BIC value between 3.4 and 4.6 percent (Basile et al., 2014, p. 238).

[8] For a detailed exposition on the mathematics of the penalty term, directly refer to (Wood and Augustin, 2002).

Geoadditive models. Geoadditive models (Kammann and Wand, 2003) augment the hedonic function by additionally smoothing the coordinates of the observations, such that the full geoadditive model can be written as $g(\mu_i) = \mathbf{X}_\mathbf{i}^* \beta^* + s_1(x_{1i}) + s_3(x_{3i}) + s_3(x_{3i} + s_4(lat_i, lon_i) + \ldots$, where lat_i and lon_i are the latitude and longitude of the ith observation. McMillen (2012) show that the simple semi-parametric model with a smooth interaction between latitude and longitude, may be able to provide a solution to the issue of unobserved spatial heterogeneity. Therefore, geoadditive models offer a potential solution for the simultaneous consideration of nonlinearity and spatial heterogeneity, a fundamentally critical, but largely disregarded issue in spatial data analysis. Additionally, compared to more traditional spatial econometric techniques such as the SAR model, the geoadditive model has a major benefit of not requiring an N x N spatial weights matrix, which is often considered to be a significant bottleneck for traditional spatial models from a computational perspective, as well as a major point of critique due to the lack of guidance on the appropriate specification. Especially given the first point, it is argued that the geoadditive model is more suitable for working with large sample sizes (Geniaux and Napoléone, 2008, p. 106), although elegant solutions have also been proposed to circumvent the aforementioned computational issue, such as the matrix exponential spatial specification by LeSage and Pace (2007).

Despite the extensive documentation on issues concerning both nonlinearity and spatial effects individually, the simultaneous consideration is subject to a meagre academic literature base, with the main propositions provided by Lambert, Xu, and Florax (2014), who proposed a parametric model, and Griffith and Paelinck (2011), who proposed the Lotka-Volterra prey-predator model. In particular, the intense utilisation of maximum likelihood methods in spatial econometrics, which heavily depend on the assumption of a correct specification, has caused several prominent scholars to call for a reconsideration of the methodological basis of traditional spatial econometrics (Basile et al., 2014, p. 230). Gibbons and Overman (2012), for example, critique the dominant approach in spatial econometrics as unconvincing due to the many, sometimes unjustified, hypotheses made about the functional form, the presence of omitted factors, and the flexibility in choice of spatial weights, amongst others (Basile et al., 2014, p. 230). Similarly, due to the issue of functional form, McMillen (2010, p. 135) assert that fixed effects and semi-parametric models should always be preferred over spatial autoregressive and spatial error models, unless the objective is to explicitly model the effects of spatial lags of the dependent variable on itself.

4.3 Variable misspecification

An issue frequently associated with hedonic price models is the misspecification of the set of explanatory variables (Linneman, 1980; Mark and Goldberg, 1988), where either irrelevant variables are mistakenly included, known as over-specification, or relevant variables are mistakenly omitted, known as under-specification (Chau and Chin, 2003, p. 148). Over-specification results in unbiased and consistent coefficient estimates, but due to the inclusion of irrelevant variables, estimations are inefficient. On the other hand, hedonic models that are under-specified may produce biased and inconsistent coefficient estimates (Chau and Chin, 2003, p. 148), known more commonly as omitted variable bias, and results from the estimator attempting to compensate for the omitted variables (Griffith and Chun, 2016, p. 1). Omitting important independent variables that are uncorrelated with other predictor variables will negatively affect prediction accuracy of the model (Pace, Sirmans, and Slawson, 2002, p. 141). Hence, the issue of over-specification is regarded as less severe than the issue of under-specification for the objective of this thesis.

One major obstacle in overcoming the omitted variable bias has been the unavailability of locational characteristics (Linneman, 1980, p. 50). As recently as 2010, Pace and LeSage (2010, p. 17) have argued that locational variables, such as different types of amenities, highway accessibility, or school quality may plausibly influence the observed house prices, but are unlikely to be included in hedonic regressions, due to their improbable availability to researchers. For example, the American Housing Survey, a popular housing data set used in the literature and also employed by Linneman (1980), fails to include important measures of accessibility to various work and recreational sites, such as distance to parks, playgrounds or fitness centres, but also other location specific amenities, such as the number of restaurants, cafes, or grocery stores within walking distance of a property. The lack of locational variables has long been lamented in the literature, with examples including Mark and Goldberg (1988, p. 95) or Redfearn (2009, p. 298).

This thesis addresses this shortcoming by making use of the wealth of information provided by mapping services such as Google Maps and OpenStreetMap, as well as recent developments of statistical software packages, such as the packages *sf* (Pebesma, 2018), *tmap* (Tennekes, 2018), *nngeo* (Dorman, 2019), and *spdep* (Bivand et al., 2013) available in *R* (R Core Team, 2013), that ease the analytical workload of the researcher. With the inclusion of these commonly omitted locational variables, the omitted variable bias is addressed in a novel way, contributing to the housing demand literature. Furthermore, since over-specification of the hedonic model does not result in bias or inconsistent estimates, the general tendency to

over-specify rather than under-specify was deemed more appropriate for the objective of this thesis, especially with regard to the locational variables. This is reasoned by the fact that collinearity, or rather concurvity among individual locational variables is less of a concern due to the predictive objective of an overall locational value.

Furthermore, a distinct advantage of utilising publicly available map data is that consumers have equal access to the data and are likely to make use of it when evaluating the locational advantages of prospective future homes. Hence, the locational variables utilised in this analysis are most likely to reflect consumer perception of the locational value of the respective properties analysed, leading to a more accurate specification of the hedonic equation (Arguea and Hsiao, 2000, p. 105). In addition, the housing demand literature was scoured to determine variables that have commonly been employed and found significant throughout a plethora of studies, in an attempt to address the issue of omitted variable bias beyond locational variables. Furthermore, the literature's attempts at capturing the effects of locational amenities sets the foundation for the selection of locational variables. Given that the issue of variable selection is one of the major critiques of the hedonic approach (see e.g. Arguea and Hsiao (2000, p. 105) or Bowen, Mikelbank, and Prestegaard (2001, p. 470)), a substantial amount of time is devoted to summarising empirical findings and peculiarities of individual explanatory variables in the following section.

According to Dubin (1988, p. 468), housing characteristics can be categorised into three distinct, broad groups: (1) characteristics delineating the location of a property (location variables), (2) those describing the material construction of the house (structural variables), and (3) those depicting the neighbourhood in which the property is situated (neighbourhood variables). Location variables include, for example, the distance to the central business district (CBD), proximity to green areas or water bodies, or accessibility of playgrounds (Dubin, 1988, p. 469). Structural variables typically include interior floor area, exterior floor area, age of the building, number of bedrooms and bathrooms, as well as binary variables indicating the presence or absence of features, such as a garage, fireplace, or floor heating (Dubin, 1988, p. 468). These variables are often readily accessible and straightforward. Neighbourhood variables on the other hand are more difficult to measure and obtain. Furthermore, literature has not reached a consensus on which variables are to be considered appropriate and relevant measures of neighbourhood quality (Dubin, 1988; Arguea and Hsiao, 2000). Neighbourhood values found in the literature be be labelled as measurements of (1) the quality of public services, such as crime rates, local government expenditures, or elementary school aptitude test scores, (2) the characteristics of the individual neighbours, such as race, median income, or average education, and (3) the environmental variables, such as air pol-

lution or noise levels (Dubin, 1988; Goodman, 1986). In the hope of further clarifying the delineation between neighbourhood and locational variables, in this thesis locational characteristics are defined to refer to variables describing fixed attributes with respect to a specific location, while neighbourhood variables refer to attributes that are considered constant for an aggregate region, such as tax rates or racial composition, or that occur randomly within a given measure of regionality, such as shooting incidences. Albeit being merely a matter of definition, the aforementioned classification categorises school quality indicators, such as average SAT scores, as a locational variable due to the locationally fixed nature of a school, as opposed to being a neighbourhood variable.

4.3.1 Locational variables

The importance of locational variables is outlined by the first fundamental characteristic of housing markets, stipulating that advantageous locations are priced more expensively according to the compensation principle. This holds with regard to natural locational amenities, such as lakes or beaches, and man-made locational advantages, such as distance to employment or cultural centres. (DiPasquale and Wheaton, 1996, p. 36)

Distance to CBD. DiPasquale and Wheaton (1996, p. 36) illustrate the importance of the distance to employment as a locational variable by assuming a monocentric city[9], where the only locational advantage considered is the access or commuting distance (along a straight line) to that centre of employment. Furthermore, residents incur commuting costs k annually per kilometre and housing rent $R(d)$ dependent on the commuting distance d, while all residents receive the same income y which can be spent on either housing, commuting, or other goods x_0. With housing being allocated to those residents who offer the highest rents, the rents will be determined by

$$\begin{aligned} R(d) &= y - k \cdot d - x_0 \\ &= (r^a \cdot q + c) + k(b - d) \end{aligned} \tag{6}$$

in the equilibrium, with r^a reflecting the opportunity value per acre, q the acres of land, c the annualised cost of construction per unit, and b the distance of the border of the monocentric city from the city centre (see DiPasquale and Wheaton (1996) for a more extensive derivation

[9] The assumption of a monocentric city follows a long list of literature (see Alonso et al. (1964), Muth (1969), or Mills (1972)).

of the model). This model, referred to as the Ricardian rent model after the British political economist David Ricardo, implies that the decrease in rents as one moves away from the city centre directly offset the increases in commuting costs. (DiPasquale and Wheaton, 1996, pp. 37-38)

Using a multivariate regression analysis to control for the heterogeneity assumed to be present in the Ricardian rent model, DiPasquale and Wheaton (1996, p. 57) were able to confirm the importance of distance to CBD suggested by the Ricardian rent model. Other studies supporting these findings, to name a few, include Söderberg and Janssen (2001), Atack and Margo (1998), Yinger (2015, p. 16), McMillen, Jarmin, and Thorsnes (1992), Palmquist (1992), and Ridker and Henning (1967). A point worth mentioning is the fact that rent gradients may not be constant in all directions from the city centre. For instance, Söderberg and Janssen (2001, p. 77) found that the rent gradient was less step in the East of Stockholm compared to all other directions.

However, there have also been studies that have failed to support the theory of the Ricardian rent model. These studies include Kain and Quigley (1970, p. 540), Schmenner (1981), Anselin and Lozano-Gracia (2008, p. 11), Wilkinson (1973), Hamilton and Röell (1982), So, Tse, and Ganesan (1997, p. 40), and Ng (2008). For further details see section A.1 in the appendix.

Public transport. Mayer and Trevien (2017) found that upgrades made in and around the Parisian public transport network raised employment by 8.8 percent. The importance of public transport thus can be said to be evident and pronounced. So et al. (1997) found that access to minibuses and the MTR network has a significant positive effect on house prices. More specifically, they found that access to public transport options that allow for frequent transportation are valued. Similarly, Lehner (2011) researched the effect of access and distance to public transportation in Singapore as a locational variable, where it was found that longer distances to public transport results in significant negative impact on property prices (Lehner, 2011, p. 75). Armstrong and Rodríguez (2006, p. 38) were able to validate a positive significant effect of public transportation accessibility on house prices in Eastern Massachusetts, where properties located in municipalities with rail stations were priced 9.6 to 10.1 percent higher. In fact, properties that were located within a half mile radius of a station are priced 10.1 percent higher, and for each additional minute drive from a station, the price of properties decrease by 1.6 percent Armstrong and Rodríguez (2006, p.38).

However, the relationship with distance to public transportation and house prices need not be linearly increasing. In the theory it is often contested that the wealthy do not utilise public transportation and therefore would not capitalise access thereto into house prices. For example, Liao and Wang (2012, p. 26) argues that "wealthier households increasingly switch from public transportation to private vehicles". Consequently, the variable measuring the access to public transport may be expected to have a spatially varying coefficient, with a stronger coefficient expected in middle and low income neighbourhoods and weakened or possibly even negative coefficient in high income areas. The latter may also be argued to be due to the fact that proximity to public transport often coincides with noise pollution. However, the noise pollution is expected to strongly depend on the type of public transport (for example, subway station are less of a concern compared to bus stations) and is expected to drastically diminish with increasing distances, such that only very closely located properties would be severely negatively effected thereby.

Restaurants, cafes, and bars. As expected, studies employing variables indicating the availability of close-by restaurants, cafes, or bars are scarce to nonexistent. Even Xiao et al. (2017), who made use of open access point of interest data, omitted indicators of proximity to restaurants and the like. Nevertheless, it seems obvious that the proximity to a diverse offering of restaurants, cafes, or bars is attractive to prospective home owners. Teulings, Ossokina, and de Groot (2018) is a recent article mentioning the value of restaurants, wherein he argues that higher educated homebuyers are more willing to pay a higher price for the amenities of a city, including the availability of restaurants. However, they too, fail to include a measurement of availability of restaurants and the like in their empirical analysis. Arguably, the vast lack of consideration of said variables is due to the difficulty in obtaining accurate measurements. However, in the advent of Google Maps and OpenStreetMaps, such data has become accessible to the public. Consequently, this thesis answers the call of Ng (2008, p. 118) for future research to include endogenous amenities, such as restaurants, when analysing how households choose residential locations. Accordingly, this thesis aims at contributing to the literature with the inclusion of number of restaurants, cafes, and bars within walking distance of each property. Although, the importance of amenity quality is recognised in addition to the quantitative aspect, a measure thereof remained unavailable at the scale required (despite Google Maps collecting sufficient amounts of data). Further issues pertaining to the quality of data and potential solutions are discussed in the appendix (see section A.10).

Cinemas. Even less surprising is the inexistent literature assessing the capitalisation effects of proximity to cinemas, arguably due to the fact that such data has often been

considered unattainable. Additionally, it may be reasoned that consumers do not perceive proximity thereto as valuable given the lack of frequency with which such places are visited. An important aspect of this argumentation is that the importance of proximity to amenities is conditional on the frequent use thereof, an argument in line with the findings of Ng (2008). Furthermore, there may be strong preferential differences among citizens, given that not all people enjoy going to the movies. However, it may also be argued that cinemas often cluster together with other cultural offerings, and may hence serve as a proxy of cultural amenities in general.

Grocery stores. Picking up on the issues of frequency and preference, supermarkets and convenience stores are surely amenities frequented by the vast majority of home owners and hence can be expected to play a substantial role in housing choice. And indeed, Lehner (2011, p. 46) found that longer distances to food centres and supermarkets result in higher prices of properties. With respect to convenience stores, Xiao et al. (2017, p. 8) were able to find a significant positive effect on housing prices.

With regards to convenience stores specifically, Chiang, Peng, and Chang (2015) found that residents in neighbourhoods with lower prices property rather prefer access to convenience stores, as compared to neighbourhoods with higher prices properties. They suggest that this may be due to the fact that "those in the neighbourhoods with higher-priced property may be more mobile to access convenience stores [or supermarkets] in other suburbs en route from one place to another." (Chiang et al., 2015, p. 82). Consequently, homebuyers may not value proximity to supermarkets or convenience stores if they have grocery shopping opportunities along their daily commutes or in proximity to their work place. Furthermore, the size and quality of offering may significantly effect the valuation of grocery stores, in addition to distance.

School distance and quality. It has been widely accepted that proximity and quality of schooling affects house prices (Livy, 2017), with multiple studies reporting positive relationships, such as Oates (1969), Kain and Quigley (1970), Harrison Jr and Rubinfeld (1978), Li and Brown (1980), Jud and Watts (1981), Jud (1985), Haurin, Brasington, et al. (1996), or Bond and Seiler (1996). For more details, refer to section A.2 in the appendix.

With regard to school quality, most studies make use of standardised student examination scores or expenditures per pupil (Clauretie and Neill, 2000). For example, Downes and Zabel (2002) found that both per-pupil expenditures and test scores (as an indication of school quality) similarly impact house prices. However, Rosen and Fullerton (1977, p. 439)

found that test scores are better measures of school quality than per-pupil expenditures, because"inputs should not be used to measure outputs" (for more details regarding the influence of school quality on hous prices, refer to section A.2 in the appendix). Furthermore, when considering the role of schooling and education a distinction can be made between the different education levels i.e. primary, middle and high schools. For example, Yinger (2015, p. 15) mentions that quality perceptions of schools may differ between elementary and district schools. Namely, he proposes that "parents may not be as concerned about a low-scoring elementary school if their children will eventually attend a high-scoring middle school and high school" (Yinger, 2015, p. 15). In similar vein, Lehner (2011, p. 46) found that a larger distance from primary and secondary schools results in higher property prices, whereas a further distance from higher education schools result in lower property prices. Lastly, Clauretie and Neill (2000) points out that families without children may value school quality differently or not at all, in contrast to families with children. As such, the effect of schools may not have a uniform effect on housing prices. One bedroom or studio apartments, for example, may not be influenced or influenced significantly less by school availability and quality.

Sports facilities. Ng (2008, p. 116) observes that various authors find commuting distances to be higher than those predicted by standard urban models, while in near to all underlying theoretical models, household preferences for amenities are ignored. Thus, he stipulates, that it is likely that households are willing to accept longer commuting journeys given proximity to an amenity of great personal importance, such as sports facilities. Of the few studies that have analysed the proximity to sports facilities, Mok et al. (1995), So et al. (1997) Xiao et al. (2017), Ahlfeldt and Maennig (2010) and Feng and Humphreys (2018, p. 189) were all able to report positive and significant effects on house prices. For instance, Feng and Humphreys (2018, p. 189) found that "housing values increase by 1.75 percent for each 10 percent decrease in the distance from the house to the facility".

Parks. The most notable and earliest documentation of the relationship between public parks and housing values was by Frederick Law Olmsted (Fox, 1990), who used the proximate principle to explain how Central Park in New York would be self-financing. [10] The notable literature review by Crompton (2001) outlines that 20 out of 25 researched empirical studies reported a significant positive relationship between park proximity and real estate values.

[10] The proximate principle is based on the "process of capitalisation of park land into the value of nearby properties" (Crompton, 2001, p. 1), whereby the annual debt charges related to the development of the park would be sufficiently covered by the increased property tax revenue generated by the increased property values.

Moreover, four out of the five studies that did not find a significant relationship between the two are argued to suffer from methodological issues. Other studies who likewise report significant and positive effects of proximity to parks include So et al. (1997), Sirpal (1994), Cameron (2006), Wu, Ye, Du, and Luo (2017), Anselin and Lozano-Gracia (2008), Espey and Owusu-Edusei (2001, p. 487) and Tyrväinen (1997).

However, there have also been studies that have failed to support these findings. These studies include Smith, Poulos, and Kim (2002), Yinger (2015), Mahan, Polasky, and Adams (2000), Wong et al. (2002) and Molin (2011). Possible reasons for these conflicting findings include: (1) the definition of a park is vague and hence data in the different studies may not be comparable (Wu et al., 2017, p. 54); (2) the effect of parks may vary throughout the region, as suggested by Cho, Bowker, and Park (2006, p. 491); and (3) parks may vary significantly in quality, whereby unsafe parks may even be viewed as a disamenity. For further details on the conflicting effects found with regard to parks see section A.3 in the appendix.

Waterfront. Researches have long had an interest in measuring the benefits of access to the shoreline, the waterfront, or urban water parks. Darling (1973) set out to measure the benefits of urban water parks, which are mainly aesthetic quality and recreational opportunity, as early as 1973 in California, U.S.. He was able to show that the sales price regression had the expected negative sign for the explanatory variable of distance to urban water parks (Darling, 1973, p. 30). Several other authors found a similar effect, such as Anselin and Lozano-Gracia (2008), Brown and Pollakowski (1977), Walsh (2009), Mahan et al. (2000) and Walsh (2009). For further details see section A.4 in the appendix.

While a distinction between proximity to waterfronts and views thereof may be worthwhile, obtaining data on views and especially the quality thereof is rather cumbersome and subjective. Nevertheless, studies exist that measure both views and proximity, such as Stadelmann (2010, p. 184), who defines the proximity to a lake "with the number of hectare of the lake which can be over-looked from the single-family house". In this study in Zürich, a significant positive effect was found for proximity to and view of the lake Stadelmann (2010, p. 191). Fischer, Füss, and Stehle (2020) were also able to obtain data on views as well as proximity to the water surrounding Manhattan Island in New York in their analysis of local house price comovements.

Distance to playgrounds. Similar to the bars, cafes, and restaurants, literature on the effect of distance to playgrounds on housing prices is currently sparse, although some studies

do exist. Most notable is the study done by Breunig, Hasan, and Whiteoak (2018, p. 17), who find that the presence of a playground within 300 metres of a property adds just under 5 percent to property prices in urban Australia. This effect is larger for houses Breunig et al. (2018, p. 18), where authors argue this is due to demographic characteristics such as presence of children in houses. In a study done in Berlin, Ahlfeldt and Maennig (2010) found that proximity to playgrounds, as a control variable, did not have a significant effect on general land value. They argue that this is the case since playgrounds are an amenity "provided relatively evenly across locations [which results in the fact that] residents then no longer recognise it as a local amenity" (Ahlfeldt and Maennig, 2010, p. 214).

Distance to shopping centres. Shopping malls and department stores concentrate retail stores in one location, offering convenience to potential buyers. Furthermore, shopping is considered a leisure activity by many and hence is likely to influence the valuation of nearby housing. As one of the main locational factors in Stadelmann's (2010) study in Zürich, shopping centres were found to have a significant positive effect on house prices Stadelmann (2010, p. 191). Likewise, So et al. (1997), Xiao et al. (2017) and Yinger (2015) were able to reveal that the proximity to shopping centres is positive and statistically significant in determining house prices.

It is also important to consider the size of shopping centres, as a larger shopping centre may indicate more variety and higher chances of finding what is needed, thus being more beneficial and positively impacting house prices. To this end, Sirpal (1994) analysed the price differences of identical residential properties located around shopping centres of different sizes. Overall, the results show that "the value of a residential property at a radial distance from the outer periphery of a larger shopping centre is higher as compared to that of an otherwise identical residential property located at the same radial distance from the outer periphery of a smaller shopping centre" (Sirpal, 1994, p. 487). Similarly, (Des Rosiers, Lagana, Thériault, and Beaudoin, 1996, p. 51) found that each additional shop adds about $27 to the market value of the properties in the vicinity of the shopping centre.

4.3.2 Structural variables

As pointed out by Chau and Chin (2003, p. 10), "prices of properties are frequently related to their structural attributes". There is a large array of housing attributes that have been considered and found to be significant in the literature, raning from garages, heating, floor type to architectural attributes (see e.g. Haurin et al. (1996), McCord et al. (2019),

Xiao et al. (2017), Kain and Quigley (1970), Bateman, Day, Lake, and Lovett (2001), Schulz and Werwatz (2004), (Tyrväinen, 1997), Forrest, Glen, and Ward (1996)). Subsequently, the three major structural variables of floor space, age and building type will be discussed in more detail. Albeit carrying significant explanatory power, the variables of floor level, number of rooms, and views remained unobtainable for the studied area in New York. For completeness, those variables are still discussed in detail in the appendix in sections A.6, A.7 and A.8 respectively.

Floor space. Chau and Chin (2003) argues that the single most important structural variable is considered to be floor area. While finding a positive and significant effect of floor space on property prices in Beijing, China, Xiao et al. (2017, p. 8) argue that "the property value is dictated mainly by the floor space". Significant positive effects were also found by many other authors, including, but not limited to, Stadelmann (2010) for properties in Zürich, Tandon (2012) for properties in Nevada, Schulz and Werwatz (2004) for properties in Berlin, Choy, Mak, and Ho (2007), So et al. (1997) and Mok et al. (1995) for properties in Hong Kong,Tyrväinen (1997) in Helsinki, Kim and Lee (2018) for properties in Seoul, McCord et al. (2019) in Belfast, Liao and Wang (2012) in Changsha, China, Lehner (2011) in Singapore, Kain and Quigley (1970) in St. Louis, Helbich, Brunauer, Vaz, and Nijkamp (2014) in Austria, Forrest et al. (1996) in Greater Manchester, Chiang et al. (2015) in Taipei, Haurin et al. (1996) for properties in Ohio and finally Fan et al. (2006) in Singapore.

Age. The age of a building was cited by Lehner (2011) as one of the three most important determinants of property prices. In general, newer homes are preferred to older homes, due to the effect of depreciation. Chau and Chin (2003, p. 10) outlines some of the drawbacks of older buildings, which are argued to be worth less, due to the fact that they "incur more costs in maintenance and repair, and also have decreased usefulness due to changes in design, electrical and mechanical systems". In line with the aforementioned reasoning, a significant negative effect was found by So et al. (1997), Tyrväinen (1997), Kain and Quigley (1970), Tandon (2012), Schulz and Werwatz (2004), Choy et al. (2007), Mok et al. (1995), Helbich et al. (2014), Xiao et al. (2017), Kim and Lee (2018) and Haurin et al. (1996). For further details see section A.5 in the appendix.

However, older buildings can also be associated with positive aspects, such as charm or specific architectural styles. As such, age may exhibit a nonlinear relationship with house prices given that very old buildings often have a high value, due to historical significance, charm or vintage effects of the buildings (Li and Brown, 1980, p. 138). Another factor worth considering is the historical development of cities, where historical city centres often remain

highly attractive locations, and hence prospective home owners are willing to pay a premium for houses located in the city centre despite their old age.

Building type. There is considerable variation in the type of buildings and dwellings, ranging from studios to family townhouses, and from one story detached buildings to skyscrapers. Taking note of this significant variation, a number of studies that have accounted for building types include McCord et al. (2019), Kain and Quigley (1970), Bateman et al. (2001), Schulz and Werwatz (2004), (Tyrväinen, 1997), Forrest et al. (1996). For example, Bateman et al. (2001) include the property type as a structural explanatory variable in their quest of measuring the effect of road traffic on residential property value. Bateman et al. (2001, p. 163) argues that detatched houses will have the highest asking price compared to all other property types, and that the more flats in a single building, the lower the price for the flat would be. Moreover, "detached, semi-detached houses and terrace houses command relatively higher prices than sub-divided houses or four-block properties which in turn command higher prices than flats or tenements" (Bateman et al., 2001, p. 163). Similarly, Schulz and Werwatz (2004, p. 48) reports that "the price drops by 3.5 percent for row houses and rises by 7.6 percent for detached houses compared with a semi-detached house. As such, people are willing to pay a premium for privacy." In Helskini, houses made from brick are priced 7.0 percent higher than houses made form wood or concrete (or other materials) (Tyrväinen, 1997, p. 215). Finally, Forrest et al. (1996) found that, compared to terraced homes in Greater Manchester, a semi-detached house has a 16 percent higher price, a detached bungalow has a 50 percent higher price, a maisonette a 26 percent and a detached house 46 percent higher price.

4.3.3 Neighbourhood variables

Delineating neighbourhood variables from locational variables, I define the former to describe characteristics that apply constantly throughout a region, e.g. tax rates, or that occur randomly throughout a region, e.g. crime incidences. However, the definition of a neighbourhood is left to discretion of the researcher and is subject to debate within literature (see e.g. Mehaffy, Porta, and Romice (2015)).

Crime. Neighbourhood quality is expected to be strongly affected by crime rates and correspondingly many studies have set out to empirically underpin this expectation Arguea and Hsiao (2000, p. 105). For instance, Li and Brown (1980) found that buyers do not favour areas which are associated with crime rates, indicating a preference for safety. Arguea

and Hsiao (2000, p. 122) employ a latent variable approach to measuring the impact of neighbourhood value on house prices, finding a significant negative effect of crime rate on the quality of a neighbourhood, which in turn has a significant and positive effect on house prices. Clark and Herrin (2000) use the murder rate as an indicator of crime, and found that with each additional murder per 10'000 people in an area, properties in California are priced 7.28 percent lower. Besley and Mueller (2012) also found a significant negative relationship between house prices and killings in Northern Ireland. Interestingly, they also found evidence that there is a spillover effect of crime across different areas, which they state may indicate a "general equilibrium response to violence including migration across regions" (Besley and Mueller, 2012, p. 22).

However, the optimal measure of crime, or rather lack of safety, is debatable with measures varying throughout the literature, including robbery, rape, assault and theft Haurin et al. (1996, p. 13). Lynch and Rasmussen (2001, p. 1988) point out that the inclusion of all crimes in a neighbourhood distorts the picture of how public safety varies over space, advocating for a measure indicating their relative seriousness. For instance, theft is regarded by most as less severe of an issue than homocides. Furthermore, important for the willingness to pay of potential home buyers is the perceived safety of the neighbourhood. Hence, incidents such as burglary that are less obvious (unless police reports are analysed or word of mouth gets around) are less likely to influence house prices to the same extent that homocides are. However, given sufficient repeat offences of burglary in a specific neighbourhood, residence may be more likely to move out, thus lowering the price through supply and demand mechanisms. Furthermore, certain crimes, such as the aforementioned burglary, are expected to be correlated with expensive areas, given that more valuable objects are likely to be found in more expensive houses. Summarising, the choice of crime measures is not straightforward and needs to be given considerable attention.

Noise pollution. The sources of noise pollution are multitude. Commonly studied sources include aircrafts, railways, highways, main roads, and traffic noise (see e.g. Andersson et al. (2010), Bateman et al. (2001), Cohen and Coughlin (2008), Tomkins, Topham, Twomey, and Ward (1998), Feitelson, Hurd, and Mudge (1996), Forrest et al. (1996), Theebe (2004), Galilea and de Dios Ortuzar (2005) or Salvi (2008)). For instance, Theebe (2004), who conducted a study of noise in general, such as roads, highways, railways and airports, observe maximum discounts of approximately 12 percent in noisy areas, ignoring extremes, while particularly quiet areas may sell at premiums of up to 6.5 percent (for more details regarding this study, see the section A.9). Conducting a more specific study, Cohen and Coughlin (2008), believed to be the first study on airport noise in the literature that utilises

a spatial hedonic model, found a 0.15 percent decrease in house prices following a one percent increase in distance from the airport, all else held constant.

Simultaneously to the negative effect of exposure to noise, the distance measure may capture positive amenity values of proximity to the airport after the negative effects of noise pollution have levelled off after a certain distance. And indeed, Tomkins et al. (1998) and McMillen and McDonald (2004) have found evidence that proximity to airports in Manchester, England and Chicago, U.S., had positive effects on house prices due to the value of accessibility. Consequently, it is hypothesised that the relationship between distance to the nearest airport and house prices is nonlinear, with low distances being positively sloped before reaching a turning point at which the benefits of accessibility outweigh the drawbacks of additional noise. Furthermore, Bateman et al. (2001) point out that the impact of noise will not only depend on the magnitude of the noise pollution, but also on its intensity, frequency, duration, variability and time of occurrence during the day. For instance, heavy noise pollution during the night is hypothesised to cause greater loss in values than daytime noise.

Air pollution. Similarly to noise pollution, the implicit price of air quality is a well researched area of interest. Chattopadhyay (1999) found that air pollution is indeed a significant factor in determining house prices, such that residents in Chicago were willing to pay for a reduction in the pollution level of particulate matter (PM-10) and sulphur dioxide. Bayer et al. (2009) estimated that the median house would be willing to pay $336 - $418 (using an inflation rate of 2.26x for the period of 1982 until 2020) for a one unit reduction in average ambient concentration of particulate matter. Other studies that corroborate the significance of air quality in valuing house prices include Polinsky and Shavell (1976), Harrison Jr and Rubinfeld (1978), Palmquist and Israngkura (1999), Zabel and Kiel (2000), Kim et al. (2003) Chay and Greenstone (2005), Cameron (2006), Anselin and Lozano-Gracia (2008), Bayer et al. (2009), or Neill, Hassenzahl, and Assane (2007).

The estimation of the marginal price of air quality isn't straightforward, as it is plagued by endogeneity issues due to home buyers sorting according to their respective preference regarding environmental quality (Anselin and Lozano-Gracia, 2008, p. 7). One common approach in dealing with this type of endogeneity is to employ an instrumental variables approach (see e.g., Chay and Greenstone (2005) or Bayer et al. (2009)). Another type of endogeneity may arise from the use of spatially interpolated values for air quality, argued to result in a prediction error which may be correlated with the overall model disturbance term (Anselin and Lozano-Gracia, 2008, p. 7). Anselin and Lozano-Gracia (2008) propose the

inclusion of a spatially lagged dependent variable in addition to the instrumental variable for air quality, and estimate their hedonic specification using spatial two stage least squares. As the instrumental variable for air quality, they chose to use the latitude, longitude and their product, while arguing their ability to proxy the overall spatial pattern of the pollution as a global spatial trend (Anselin and Lozano-Gracia, 2008, p. 18). Furthermore, they argue that these instruments may also aid in correcting for endogeneity induced by other factors, such as sorting.

4.4 Further issues

Multicollinearity. Another considerable point of concern with regard to regression models is the issue of multicollinearity, in particular with models with a sizeable amount of explanatory variables. Multicollinearity occurs when two or more independent variables are linearly related and can result in considerably inflated standard errors of estimates, rendering hypothesis testing inconclusive (Cho et al., 2006, p. 489). Popular approaches for identifying the issue of multicollinearity include variance inflation factors and pairwise correlation coefficients between variables. Variance inflation factors are scaled multiple correlation coefficients between the variable j and the remaining explanatory variables and given by $VIF = \frac{1}{1-R_k^2}$, where R_k is the multiple correlation coefficient (Cho et al., 2006, p. 489). Both VIF and the pairwise correlation coefficient lack specific definitions for when multicollinearity becomes a serious issue. However, Judge (1982, p. 620) state that correlation coefficients greater than 0.8 or 0.9 are likely to cause serious issues, while VIFs greater than 5 or 10 are commonly used as thresholds for indicating the need for remediation (e.g. Cho et al. (2006) or Xiao et al. (2017)).

Principal component analysis (PCA) (Pearson, 1901; Hotelling, 1936) is a dimensionality reduction technique that can be employed in order to alleviate the issues of multicollinearity. Specifically, PCA is a statistical procedure that transforms a set of possibly correlated covariates into a set of linearly uncorrelated variables, called principle components. The transformation is defined such that the first principal component explains the largest possible variance, with each subsequent principal component accounting for the largest possible variance, conditional on being orthogonal to the preceding principal components. Consequently, a set of highly correlated explanatory variables can be replaced by first k orthogonal principal components, removing the issue of multicollinearity. However, the transformation comes at the cost of loss of interpretation. As an alternative to PCA, simpler approaches can be used,

while retaining the property of interpretability, such as dropping all but one variable in a set of highly correlated explanatory variables or relying on appropriate aggregation techniques.

With the introduction of nonparametric models, such as GAMs, the issue of collinearity of covariates is augmented to the concept of concurvity, which would require nonparametric analogues of linear principal components analysis for adequate remediation of the issue (Hastie and Tibshirani, 1986, p. 306). While some progress towards such statistical procedures have reportedly been made (Hastie, 1984; Young, Takane, and de Leeuw, 1978), the issue of collinearity and concurvity is deemed less severe for the purpose of prediction (as compared to the purpose of inference) (Hastie and Tibshirani, 1986, p. 306).

5 Empirical analysis

5.1 Data description

House prices. This thesis analyses the impact of location on house prices in the New York City housing market. The data contains transaction prices of properties sold in New York City in 2018, which is used for tax purposes by New York's Department of Finance. Importantly, all transactions are georeferenced, enabling the application of spatial econometric models and the derivation of distance based measures of locational benefits. In addition to the transaction price, latitude and longitude, the data set contains 26 variables[11]. The most important thereof are the corresponding building class category, the number of residential and commercial units, the gross square feet as a measure of floor area, the year in which the building was constructed, and the tax class at the time of the sale. The four building class categories made use of in this thesis can be roughly grouped into *one family homes*, *two family homes*, *three family homes*, and *elevator apartments*[12]. Every property in New York is assigned to one of four tax classes based on the use of the property. Given that properties allocated to different tax classes are subject to varying amount of taxes, the tax class should make a material impact on a property's valuation, and hence must be controlled for. This data set is provided by the New York City Department of Finance and is publicly available

[11] A detailed list and explanation of all variables in the data set can be found under https://www1.nyc.gov/assets/finance/downloads/pdf/07pdf/glossary_rsf071607.pdf

[12] A more detailed description of the building categories can be found under https://www1.nyc.gov/assets/finance/jump/hlpbldgcode.html#A

on New York City's open data platform. [13] It is said to be updated on an annual basis, with the data last being updated on the 29th of August, 2019.

Data cleansing. The unfiltered data set contains 262'248 observations for properties sold between 2016 and 2018. However, a vast majority thereof lack critical variables, such as the latitude or longitude, the floor area, the year of construction, the number of residential apartments, or have a price of zero. Filtering out such observations results in a remaining 99'735 observations. Given that the year of interest is 2018 and the focus is exclusively on residential properties, only properties sold in 2018 with zero commercial units and allocated to tax class 1 or 2 are retained in the data set, further shrinking the data set to a total of 36'162 observations. Despite the restriction to zero commercial units, 128 observations where the number of residential units did not match the total number of units remained, which were also dropped to ensure the quality of the data. After taking a look at the number of properties sold per building class category, categories with a sample size smaller than 120 observations were dropped, leaving four overarching building types, namely those mentioned in the above paragraph. A list of the detailed categories along with the corresponding number of observations can be found in table 8. Subsequently, square feet were converted into square metres and the price per square metre and the price per residential unit were calculated. However, despite this extensive data cleansing process over a thousand observations with transaction prices per square metre below $100 remained, which are highly unlikely to represent true market values.

The reasons for the very low values are believed to be twofold. Firstly, the possibility of human error exists, where unreliably low values occur due to misplaced decimal points or left out zeros. Secondly, residents of New York City are subject to a substantial property tax, given that property tax represented 45 percent of all the city tax dollars collected in the 2019 fiscal year (NYC Department of Finance, 2020). Moreover, sellers are obliged to pay a so-called Real Property Transfer Tax (RPTT) on sales, grants, assignments, transfers or surrenders of real property in New York City, including the sales or transfer of at least 50 percent of ownership in a corporation, partnership, trust, or other entity that owns/leases property and transfers of cooperative housing stock shares (NYC Department of Finance, 2020). Consequently, there is an incentive to keep sales prices below $25'000, especially for transfers to kin, as well as possibly other situations. Additionally, there may be properties sold under special circumstances, such as under financial distress or during foreclosures, and

[13] The data set containing the transaction prices can be found under `https://data.cityofnewyork.us/City-Government/NYC-Citywide-Annualized-Calendar-Sales-Update/w2pb-icbu`

therefore not reflect market prices. Next, attention is turned towards the very expensive properties and a visual inspection of the 25 most expensive objects is conducted (for a list thereof, see table 7 in the appendix). With a vast amount of buildings located at prime locations in Manhattan, and in renowned buildings, such as 432 Park Avenue, the observations are deemed legitimate. However, the question remains whether these observations are not subject to a distinct underlying data generating process, given that the preferences structure may considerably change for ultra-high net worth individuals. Furthermore, these building prices are only justifiable with exceptional structural qualities, such as views and materials used, for which data are not available, and hence the model will fail to pick up on these differences. The same holds true for very poorly constructed or deteriorated houses. While outlier detection (z-score, Cook's distance) and clustering (k-means, nearest neighbours using an average linkage algorithm) techniques were employed in an attempt to detect unrealistic transaction prices or sub housing markets subject to separate preference structures, the results were unsatisfactory and therefore only transaction prices under $25'000 were dropped, given the theoretical reasoning mentioned above. The distribution of the remaining house prices indicated by a histogram using small bins can be seen in 6.

The difference in preference structures between commercial and residential unit owners serves as the reasoning for the exclusion of commercial properties from the analysis. For example, noise due to the proximity of main roads is much less of a concern for commercial unit owners, given the fact that the property is mainly used during working hours and sleep disturbance isn't a concern. On the contrary, proximity to main roads are considered a positive amenity, due to the added visibility and accessibility for customers. Not only is the preference structure different for noise levels and distance to main roads, but in fact for most of the locational variables that are made use of. For instance, the area of parks within a two kilometre buffer zone around the commercial estate is for most businesses frankly irrelevant.

Before proceeding to analyse the data set, average prices are compared with the popular real estate platform *Zillow*, who report a median home value in New York City of $649'026, a median price of homes listed of $769'000 and a median price of homes sold of $563'800 as of November 30, 2019 (Zillow, 2020). Comparing the reported values of Zillow to the median price of the data set used in this thesis, namely $700'000, the data basis can be considered to be reasonable. Nevertheless, anomalies exist. For instance, the two properties depicted in figure 9 in the appendix exhibit near equivalent locations, are located in the same neighbourhood, and are more or less comparable with regard to their structural characteristics (at least judging from the street view and the data available), yet one was transacted for $2'605 per square meter while the other went for $13'248. There may be a vast difference

in the quality of materials used inside, additional desirable amenities, such as a sauna or a whirlpool (although such amenities should have little influence on the sales price due to the low cost thereof in the home improvement market, as argued by (Pace et al., 2002, p. 142)), special circumstances under which the properties were sold (e.g. foreclosure or sold to kin), or simply a lack of information by the seller or buyer. Whatever the case may be, the model will fail to pick up on these differences, given that the structural characteristic available, namely the type of building, the age and the floor area, are very comparable.

Education. The data set provided by the New York City Department of Finance is augmented by educational data, crime statistics, subway stations, noise complaints, measures of air quality, information of waterfront access, as well as a plethora of locational variables from OpenStreetMaps. The educational data are comprised of a quantitative and a qualitative measure. Regarding the quantitative measure, the location of all schools for the educational year of 2017 - 2018 are obtained. In addition to the locational information, the data set also provides a classification specifying the grade levels offered by the school, such that elementary schools (and others[14]) can be differentiated from high schools, as only the locations of high schools are of interest, given the great coverage of primary and secondary schools throughout entire New York City (see figure 10 in the appendix). The data set is published by the NYC Department of Education on an annual basis and is also publicly available through the open data platform.[15] Regarding the qualitative aspect, recent SAT results reported by schools are employed as means of measuring quality of education. The data set contains the school name, the number of students who took the SAT test and the average scores for critical reading, writing and maths. The data set is also published by the Department of Education.[16]

Crime. For crime data, a multitude of options were available. Specifically, the New York City government provides data on the crimes of burglary, felony assault, grand larceny, grand larceny of motor vehicles, murder, rape, and robbery. Not all types of crimes are equivalently suitable for measuring the negative effects of crime on house prices. For instance, burglary can be expected to occur in neighbourhoods with expensive properties, where it is more likely to find items worth to steal. Similarly, rape is a dissatisfactory proxy variable for the general safety of a neighbourhood, which is arguably the characteristic valued most by home

[14] For a detailed list of variables and categories, please refer directly to data set found under https://data.cityofnewyork.us/Education/2017-2018-School-Locations/p6h4-mpyy

[15] https://data.cityofnewyork.us/Education/2017-2018-School-Locations/p6h4-mpyy

[16] https://data.cityofnewyork.us/Education/2012-SAT-Results/f9bf-2cp4

owners. Moreover, potential homebuyers will have a hard time accessing the number of rapes in the neighbourhood without consulting the statistics, and as several researchers point out (see for example Arguea and Hsiao (2000, p. 105)), the correct measure of an environmental variable is that which reflects the perception of the households most accurately. Furthermore, there are issues with the rate of reportings, as not all rapes are reported to the authorities. This results in potentially inaccurate measurements on top of the lack of perception by potential home buyers (Bateman et al., 2001, p. 22). Finally, there doesn't seem to be a correlation between house prices and reported rapes, given the lack of spatial correlation with the latter, suggesting that it fails to capture the disamenity of crime.[17] A much more appropriate means of measuring neighbourhood safety seems to be the reported number of shootings. Firstly, one can assume that the vast majority of shootings are reported, either by witnesses, by neighbours hearing gun shots, by the victims themselves, or at the latest upon their visit to the hospital. Secondly, shooting incidences are reported with the exact location of occurrence in latitude and longitude, enabling the precise measurement of shootings within a given radius of a property. The data are provided by the New York Police Department, updated quarterly, and also available through the city's open data portal.[18]

Noise. One major issue in connection with measuring the value of quietness is that the sources of noise are manifold. Possible sources that have been researched in the literature include aircraft noise, railway noise, highway, road, and traffic noise, music venues, loud neighbours, amongst many others. Another issue in measuring the impact of noise on house prices is the availability of noise measurements for the multitude of emitting sources. For instance, Andersson et al. (2010), who analysed the impact of road and railway noise on property prices, obtained information on the noise levels from an independent study on the health effects of traffic noise. Another example that shows the limited availability of noise information is Salvi (2008), who made use of a proprietary aircraft noise model provided by the Swiss Federal Laboratories for Materials Testing and Research (EMPA) in their valuation of the impact of aircraft noise on house prices in Zürich, Switzerland. In contrast to such clearly identifiable sources of noise, others are less well documented. For instance, while the disamenity of living above a nightclub, in a street with a number of densely aligned bars, or in general a well frequented area, may seem obvious to individual prospective homeowners who are familiar with the area, it is much more challenging to identify such locations on a large scale with quantifiable data, especially considering the effect the building quality has

[17] A map of reported rapes can be found under https://maps.nyc.gov/crime/

[18] https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Year-To-Date-/5ucz-vwe8

on the level of the noise.

Considering these challenges, distance measures to airports, railways, highways, and main roads were used to capture the disamenities of noise pollution from the respective sources, while simultaneously capturing the amenities of accessibilities to such key transportation infrastructure. Respectively, the network of main roads, trunk roads, highways and highway links were obtained from OpenStreetMaps' API. Similarly, the light rail and rail network was obtained from OpenStreetMaps' API. Likewise, polygon data was obtained from OpenStreetMaps' API depicting the boundaries of New York's airports. With the availability of these data sets, the distance to each noise emitting transportation infrastructure could be measured for all properties.[19]

However, it is well documented that the perceived noise is not only a factor of distance, but also strongly dependent on for example the sound diffraction of the terrain, the availability of instrument approach procedures, or the weather (Salvi, 2008, p. 579). Hence, it has to be kept in mind that distance based measures are not the most accurate indicators of noise pollution for residents, nor for accessibility of the respective infrastructure, albeit being strongly correlated with both features of interest, and thus serving as satisfactory proxies.

With the intent of capturing the disamenities caused by music venues, loud neighbours and the like, the list of 311 service requests for the City of New York was employed. The data set[20], which is provided by the Department of Information Technology and Telecommunications and available on New York's open data platform, contains a list of all non-emergency complaints reported to the 311[21] service request hotline. The data set importantly contains the complaint type, of which one category is residential noise, along with the latitude and longitude of the incident location, allowing the calculation of number of complaints within a given radius of each property.

Air pollution. Similarly to noise pollution, air pollution poses considerable challenges for obtaining reliable measurements. Naturally, polluting particles diffuse throughout the air, whereas measurements are only able to occur at a finite set of individual locations. Consequently, spatial interpolation is required in hedonic house price models, as outlined

[19] All distances are measured in 100 metre units.

[20] `https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9`

[21] NYC 311 gives access to non-emergency City services and info about City government programs. More information can be found under `https://council.nyc.gov/data/311-services/`

by Anselin and Gallo (2006). For New York, measurements of total NO_x, $PM_{2.5}$, and SO_2 emissions were obtained for the five boroughs of New York City from the city's open data portal. The data[22] is provided by the Department of Health and Mental Hygiene with update frequencies being reported as needed. Albeit more granular measurements being available, the geographic entity chosen for reporting were in accordance to UHF 42 (United Hospital Fund) neighbourhoods, for which, according to my knowledge, no spatial data format exists, but is merely reported in a *pdf* format.[23] As a result of having to resort to the granularity level of boroughs, the issue of measurement errors is further augmented in the case of New York.

Additionally, air pollution can be hypothesised to be correlated with densely populated areas due to the increased amount of polluting cars. For instance, Kim et al. (2003, p. 33) report that NO_x emissions, which mainly arise from mobile sources, are greatest along the commuter corridors. However, densely populated areas often coincide with attractive neighbourhoods due to the number of close by restaurant, cafes, bars, and other leisure opportunities, but also access to supermarkets, convenience stores and cultural offerings. This concern was also voiced by Kim et al. (2003, p. 33), stating that negative effects of NO_x pollution may be offset by positive factors such as good job access and social and cultural amenities. In the case of New York, these observations hold particularly true, with Manhattan boasting by far the highest population density (27'826 people per km^2, as opposed to Brooklyn's 14'649 in second place (United States census Bureau, 2020)) and correspondingly the highest concentration of restaurants, bars, and cafes (see figures 25, 26, and 27 in the appendix).

Waterfront access. The New York City Department of City Planning offers a variety of different maps pertaining to topics such as street lines, land use, political and administrative boundaries, city facilities, and water related issues. One such issue is the provision of access to the shoreline for the city's citizens. The map, provided as an ESRI Shapefile[24], depicts the various parks and areas that offer access to the waterfront for public use and enjoyment.

Subway stations and OpenStreetMap data. The georeferenced subway stations of New York is the last data set that was obtained from the government, specifically its open

[22] https://data.cityofnewyork.us/Environment/Air-Quality/c3uy-2p5r

[23] The UHF42 neighbourhood map in pdf format can be found under https://www1.nyc.gov/site/doh/data/data-sets/maps-gis-data-files-for-download.page

[24] https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-waterfront.page

data portal, in the form of a Shapefile.[25] However, it may as well have been obtained from OpenStreetMaps, as was done with restaurants, bars, cafes, parks, highways, highway links, trunk roads, main roads, light rail and normal railway lines, shopping malls and department stores, supermarkets and convenience stores, playgrounds, cinemas, fitness centres and gyms, and airports. A full list of amenities made available though OpenStreetMaps can be found on the respective wiki page[26]. The data was obtained through their API with the help of the *osmdata* package in *R*. Important to point out is the fact that the API call returns data, depending on the feature requested, as point, line, polygon, multiline, or multipolygon data or a combination thereof. Taking the example of the playgrounds in New York, both point data and polygon data is returned. Relying on one or the other data set is troublesome, as becomes evident when taking a closer look at both data sets. Taking a look at figure 9 in the appendix, for example, where green refers to polygon data and blue to point data, one can see that simply taking the point data set would result in the count of 11 playgrounds within the walking distance from the objects in this area. On the other hand, simply taking the polygon data would result in measuring 2 instead of 3 playgrounds within walking distance. Consequently, points which overlap with polygon data were filtered out, centroids were calculated for each polygon, and subsequently both data sets were combined to obtain a data set as complete as possible (see section F.1 in the appendix for the corresponding *R* code).

For variables where variety is less of interest, which are most variables in the data set, distance to the closest amenity was calculated in metres and rescaled to 100 metres units. The projection used throughout the entire data analysis is *EPSG:32118* due to its unit being metres and its area of use being New York (epsg, 2020) (see section B.1 in the appendix for an explanation of the importance of projections). For variables where variety is of interest, as it is, for example, with restaurants, cafes and bars, the number of features were counted within a walking distance of each property. Walking distance in transportation literature and urban planning is commonly considered to be approximately 400 metres, which corresponds to a walking time of about five minutes (Mehaffy et al., 2015, p. 208). In the public transit industry, for example, buffers of 400 metres around bus stops and 800 metres around railway stations are commonly considered to be appropriate catchment areas (El-Geneidy, Grimsrud, Wasfi, Tétreault, and Surprenant-Legault, 2014, p. 193). Taking a specific example, the 85th percentile walking distance to bus transit services in Montreal, Canada was found to

[25] https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49

[26] https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-waterfront.page

be 524 metres (El-Geneidy et al., 2014, p. 193). Accordingly, 400 and 800 metres buffers were constructed around each property, and depending on the variable of interest, the more appropriate radius was chosen. For example, a 10 minute walking radius was chosen to count the number of crime incidences around each property, whereas 400 metres was chosen for the amount of restaurants located in close proximity.

To measure the access to parks, neither a distance based measure nor counts were employed, given the vast differences in sizes of parks and the corresponding different levels of utility they are expected provide. First of all, all parks with a size of less than 2'500 square metres were dropped from the data set, given that upon inspection with Google Street View it was revealed that elaborate sidewalks with one or two benches may already be labelled as parks by OpenStreetMaps (see figure 8 in the appendix for a specific example). Secondly, given that small parks offer considerably less utility to most people than larger parks, the most appropriate measure seemed to be the amount of square metres located within a given radius of each property. Considering the commonly used distances mentioned above and the example of a person wanting to go jogging in the park, a seemingly adequate distance of 2000 metres was chosen as the buffer radius for the variable of park area. Subsequently, the amount of square metres of park was calculated for each property within its respective buffer zone. Taking a look at the first ten properties in the data set, the prior mentioned procedures results in figure 28 in the appendix for the first 10 properties of the data set, outlining the procedure visually as well as serving as control mechanism to check the correctness of the measurements.

Quality check. To check the quality of the data delivered by the OpenStreetMap API in general, amenity locations and number thereof were cross-checked with results obtained from Google Places API. Although the Google Places API did offer slightly more returns for amenities such as restaurants around selected locations, which seem to exist based on Google Street View, OpenStreetMap data does correctly portray the general spatial trend. Although Google Places API would theoretically be preferred due to the seemingly more accurate coverage, and in particularly for the added information on the quality of the amenities, measured through the vast number of user ratings, the API has considerable drawbacks and limitations. First of all, in contrast to the open source OpenStreetMap API, the usage of the Google Places API is costly. Despite the first $200 of usage being free each month, the cost of obtaining the desired data for the data set employed in this thesis is estimated to be at least $6766[27] for the 34'172 observations and the corresponding 9 variables for which data could

[27] The estimate was calculated using the place details query with atmosphere data for the ratings from

be provided. A second limitation is that one cannot obtain the entire directory of amenities and their coordinates with one query for each variable, as it possible with OpenStreetMap, but rather is forced to use a so-called "Nearby Search"[28] query for each location and each variable. Moreover, the query results are limited to a maximum number of 20, meaning additional queries would have to be run for those places with, for instance, more than 20 restaurants within a 400 metres radius. Hence, the above estimate is considered a lower bound, rather than a precise estimate. One possible solution would be to work with dummy variables, categorising properties into bins of number of restaurants in close proximity with the last bin reflecting all properties with 20 or more restaurants. However, this approach may distort the distribution, leading to possible subsequent econometric problems. Lastly, to my knowledge, the Places API is not able to return polygon data, that was, for example required to calculate the park areas within the given radius of each property, as outlined above. Also, line data representing the road, highway and railway networks does not seem to be obtainable from Google Places.

The unavailability of detailed housing characteristics is one downside of the New York data set employed in this thesis. The structural variables taken into account are floor area, building type, and age. Despite the meagre selection of structural variables, one may remark that many other structural variables can be expected to be correlated with these three core variables, and thus do not add significantly to the predictive power of the model. For example, adding the number of rooms as a predictor variable will usually not greatly affect prediction power of the model, given the existing inclusion of floor area (Pace et al., 2002, p. 141). Furthermore, the inclusion of both floor area and number of bedrooms may lead to perverse coefficient signs, given the interpretation of *ceteris paribus*, such that an increase in the number of bedrooms holding all else constant would only be achievable in reducing the size of rooms, which is generally not desired. Other variables, such as the presence of a swimming pool, may not be predicted by the three available structural variables while adding significant value to the property. However, not all variables add significant value. For example, the presence of a spa is argued not to materially affect the prediction accuracy of the model, given that one can obtain a spa on the home improvement market for a moderate amount - an argument in line with arbitrage theory (Pace et al., 2002, p. 142).

---

https://cloud.google.com/maps-platform/pricing

[28] More information under https://developers.google.com/places/web-service/search#PlaceSearchRequests

5.2 Model specification

Multicollinearity. After gathering and cleaning the relevant housing data for New York, the issue of multicollinearity is addressed. Table 10 in the appendix lists the 20 most highly correlated variables in absolute terms, and it comes as no surprise that variables measuring the same effect are extremely highly correlated. Specifically, these include the three measures of air quality, NO_x, $PM_{2.5}$, SO_2, and the three measures of school quality, *maths*, *reading*, *writing*. Further troublesome variables are the set of road variables, where *d_highway_link*, measuring the distance to the nearest highway link, and *d_highway*, measuring the distance to the nearest highway are highly correlated. While the intent of using both variables was to capture the amenity of accessibility with the highway link variable and the disamenity of noise with the distance highway variable, the attempt failed due to the highway links being too numerous and thus highly correlated with the distance to highway. Furthermore, the distance to trunk roads variable *d_trunkroad* showcases a high correlation with both distance to airport and distance to subways. This is mainly due to the fact that the number of trunk roads are sparse and mostly located near the airport, as seen in figure 11 in the appendix. Also, it comes as no surprise that the variables *n_restaurants*, *n_cafes*, and *n_bars*, which measure the number of restaurants, cafes, or bars within five minutes walking distance respectively, are highly correlated, given that such amenities tend to require a significant population density to be profitable. Similarly, population density is likely to significantly affect the number of polluting vehicles in the neighbourhood and hence the number of polluting particles in the air. Consequently, including both measures of air quality and measures of accessibility to restaurants, cafes, and bars will raise issues of multicollinearity. Interestingly, but not unexpectedly, the noise complaints and the number of shootings within a given radius (as a measure of crime), showcase significant correlation and consequently also pose an issue[29]. The last concerning variable pairing is that of distance to the nearest airport and distance to the nearest subway station. Upon inspection of the scatterplot (see figure 16 in the appendix), it appears that the two variables are correlated due to the observations in Staten Island, where there are no subway stations or airports.

As a remedy for the last mentioned variable pairing, the observations in Staten Island were removed, as opposed to dropping either *d_airport* or *d_subway*, due to the theoretical importance of both variables. For the unresolved variable groupings, both principle compo-

[29] Upon inspection of the predictions for the individual variables, shootings obtained a perverse predictions, and hence was deemed to suffer from multicollinearity, despite its rather low correlation coefficient with the noise complaints variable

nent analysis and simple aggregation methods were used and compared. For the five variable groupings all primary principal components were able to explain more than 95 percent of the respective variation, and hence were able to alleviate the issues of multicollinearity successfully. Due to this success and the lack of an adequate aggregation method for shootings (crime) and noise complaints, the two variables were replaced by their primary principal component. For the remaining variable groupings, adequate aggregation methods were available and hence preferred due to the retained practical meaning. For the variables *n_restaurants*, *n_cafes*, and *n_bars* the simple sum was deemed adequate, while dropping all but *d_highway* for measures of highway accessibility and noise, and all but *maths* for measures of school quality removed any remaining concerns of multicollinearity.

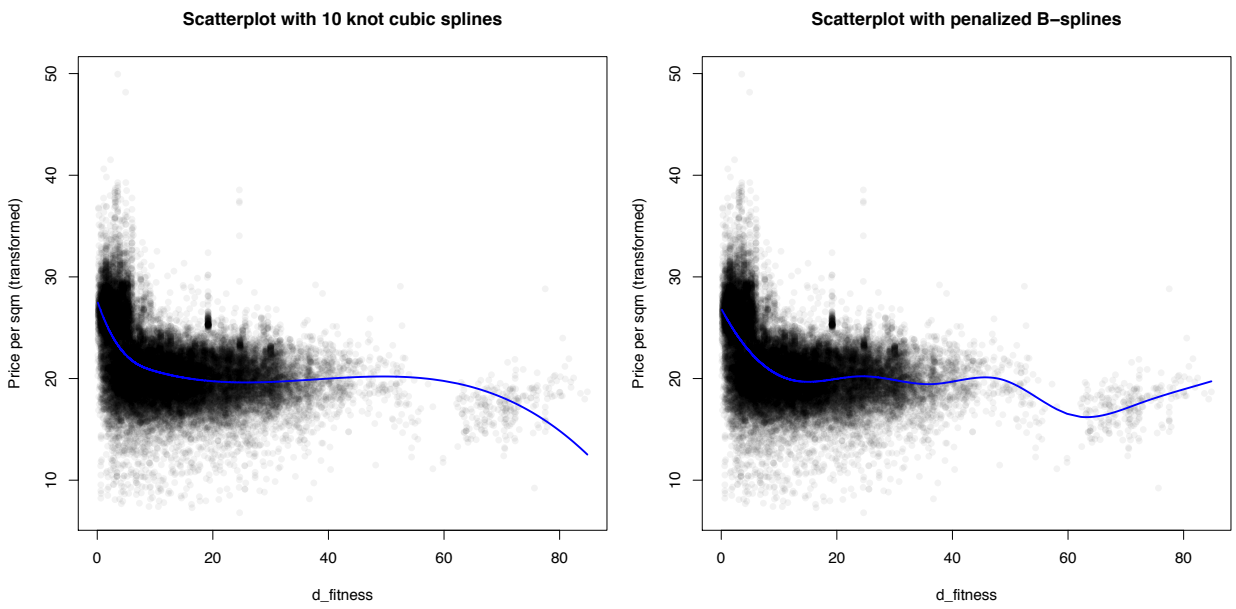
Heteroskedasticity & non-normality. Taking a look at the residuals plot for the graphical detection of heteroskedasticity (see figure 17 in the appendix), and the *Normal Q-Q* and histogram plots for the graphical detection of non-normality (see figure 18, 19 and 20 in the appendix), it is evident that both issues are present. This conclusion is further solidified after applying the *Jarque-Bera* and *Anderson-Darling* tests for the detection of non-normality and the *Breusch-Pagan* and *score* tests for the detection of heteroskedasticity (see `?ncvTest` in *R* for the latter). As a means of addressing both issues simultanseously, the variance stabilising Box-Cox transformation of the dependent variable was applied, with a suggested λ of 0.18. Comparing the before and after plots and statistics[30], significant improvements can be observed (see figures 17, 18, 19, and 20 in the appendix).

Nonlinearity. In order to detect the issue of nonlinearity, the pairwise scatterplots of the dependent and all independent variables respectively were analysed (see figure D.1). In addition to the graphical inspection, the *White Neural Network Test* indicated significant neglected nonlinearity, suggesting the need to transform the independent variables adequately (see `?white.test()` in the *tseries* package in *R*).

Due to the lack of theoretical guidance, assistance was sought from data-driven smoothing functions, such as penalised B-splines and cubic regression splines (Wood and Augustin, 2002; Wood, 2003). While the knot selection is automated in the penalised splines approach, cubic regression splines require the researcher to define an adequate number of splines, where too few may inadequately represent nonlinear relationships, while too many are likely to result in severe overfitting, and consequently bad out-of-sample performance. Using a trial-and-error

[30] Given the lack of expressiveness of test statistics when applied to large data samples (see section B.2 in the appendix), the change in values is of interest, as opposed to the reported p-values.

method, the number of knots was limited to 10 for the cubic regression splines approach, as the corresponding smooth functions were deemed reasonable and theoretically explainable. For example, taking a look at the cubic spline for the variable `d_fitness` in figure 1, one can see that the value of proximity to fitness centres levels off drastically after around a kilometre distance (the units are in 100m). Furthermore, potential issues of overfitting are already discernible in the penalised spline approach. The resulting outputs of both cubic regression splines and penalised splines for all variables can be seen in D.1, which serve as a basis for determining adequate polynomial degrees for alternative parametric transformations, which are latter made use of in the model comparison section.



(a) Cubic spline

(b) Penalised spline

Figure 1: Smoothing functions depicting the nonlinearity of the distance to fitness centres.

5.3 Descriptive statistics

Table 1 reports the descriptive statistics for the processed data set used to analyse the pricing implications of housing location in New York City. The data set consists of 28'826 observations located throughout the four boroughs of Manhattan, Queens, Brooklyn, and the Bronx. With prices per square metre ranging from $82 up until $297'180, there is a very wide spectrum of housing quality present in the data set, with the *normal Q-Q* plot indicating heavy tails (see figure 18 in the appendix).

Table 1: Descriptive statistics of the numerical variables included in the processed data set used to analyse the pricing implications of housing location for New York City. In addition, there is one categorical variable, *Building class*, with 4 levels.

	Obs.	**Minimum**	**Maximum**	**Median**	**Mean**	**Standard deviation**
price_m2	28826	82.01	297179.73	217805788.96	5276.47	7388.24
age	28826	1.00	219.00	1987660.00	84.00	37.86
math	28826	317.00	735.00	12337440.00	414.00	65.49
n_crime	28826	0.00	32.00	76906.00	1.00	4.38
n_noise	28826	0.00	756.00	1613078.00	35.00	62.77
sum_n_rcb	28826	0.00	172.00	362830.00	1.00	25.20
d_subway	28826	0.04	80.94	353091.81	5.83	14.74
d_railway	28826	0.00	92.56	369716.27	9.16	12.16
d_highway	28826	0.00	71.76	329730.83	8.91	10.24
d_mainroad	28826	0.00	51.72	167309.16	4.01	5.81
d_exchange	28826	0.44	265.57	3823377.86	126.21	64.85
d_highschool	28826	0.31	49.14	272993.27	7.77	6.58
d_supermarket	28826	0.10	46.58	169908.74	4.23	5.53
d_convenience	28826	0.11	48.45	236476.18	6.07	7.40
d_fitness	28826	0.10	84.80	373747.43	10.21	11.08
d_cinema	28826	0.11	186.83	1297239.52	33.78	38.19
d_playground	28826	0.04	29.01	168845.52	4.85	4.14
a_park	28826	0.00	4.92	24353.53	0.49	0.86
d_waterfront	28826	0.00	67.20	585154.63	15.40	15.70
d_shopping	28826	0.05	126.82	552366.57	15.26	16.93

The analysis of spatial data calls for the creation of a spatial weights matrix. Given the nature of point data, distance based weights matrices offer themselves as the obvious candidate, where the initial distance threshold was chosen as the maximum distance of all nearest neighbour distances in the data set. Consequently, each observation is attributed at least one neighbour, alleviating the issue of so-called pseudo-islands. Furthermore, graphical inspection of the distance threshold deemed it to be reasonable from a theoretical point of view (see figure 7). Due to the lack of theoretical guidance regarding the weights specification, four common specifications were calculated in order to ensure robustness of the results. The specifications include inverse distance, negative exponential distance, and the two kernel functions Epanechnikov and Gaussian, and have 28'826 regions, 2'446'388 nonzero links, 29.4 percent nonzero weights and an average of 84.9 links.

The importance of neighbourhood and locational amenities provide theoretical reasoning for the existence of spatial autocorrelation among house prices. Given the immobility of

housing, neighbouring properties are bound to exhibit similar locational characteristics and hence receive comparable locational valuations by prospective home buyers. However, the strength of the spatial autocorrelation is likely to be determined by the relative importance of locational amenities to New Yorkers, and thus motivates the calculation of the global spatial autocorrelation statistics of *Moran's I* and *Geary's C*. Table 2 reports the statistics for the four different spatial weights matrices which all indicate the presence of significant spatial autocorrelation in the prices per square metre throughout New York, suggesting that location plays an important role in the New York housing market. Furthermore, these statistics will be used to showcase how well different models explain the spatial autocorrelation in the model comparison section.

Table 2: Global spatial statistics

Statistic	**Moran's I**	**Geary's C**
Inverse Distance	0.61	0.37
Negative Exponential	0.59	0.34
Epanechnikov	0.60	0.37
Gaussian	0.61	0.36

The presence of spatial autocorrelation is further underpinned by the graphical depiction of local Moran's I in figure 2. In particular large parts of Manhattan exhibit strong spatial homogeneity, meaning that the house price per square metre of each observation is likely to be very similar to the weighted average house price per square metre of their respective neighbours. Moreover, most observations in Manhattan exhibit significant spatial autocorrelation at the one percent significance level, despite using the conservative *Bonferroni* correction, with scattered observations throughout the Bronx, Queens, and Brooklyn also showing strongly significant levels of spatial autocorrelation. Without the adjustment of the p-value, considerably more observations become significant at the one percent level, suggesting that the Bonferroni adjustment may indeed be too conservative, as Anselin (1988) mentioned (for depiction of the significance levels see figures 12a and 12b in the appendix).

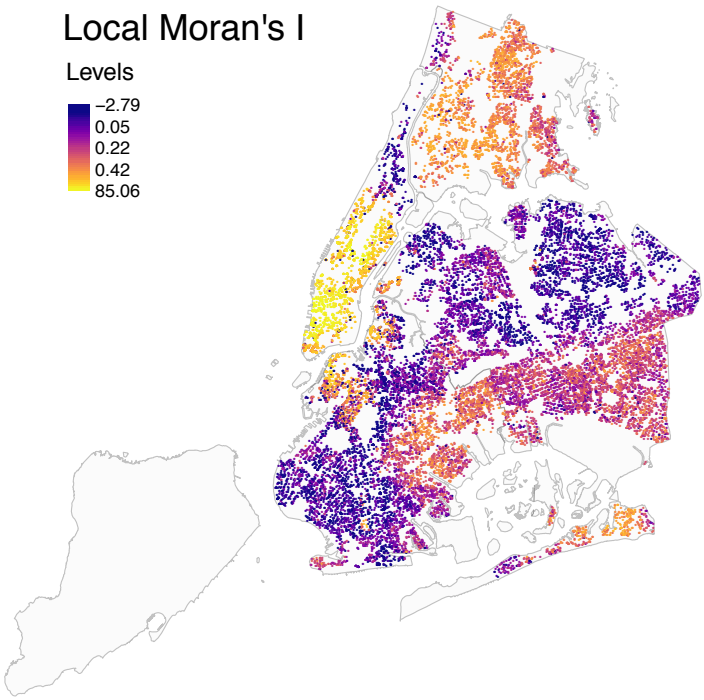


Figure 2: Local Moran's I depicted as a choropleth map for New York indicating strong spatial autocorrelation in Manhattan, and moderate spatial autocorrelation in the Bronx, the neighbourhood of Brooklyn Heights, and some parts of Queens and Brooklyn.

5.4 Model comparison

Given the presence of both nonlinearity and spatial autocorrelation, several advanced semi-parametric and/or spatial hedonic models were compared against each other in order to determine the most adequate and best performing model. The comparison was made on a subsample of the entire data set, namely for the 3'460 observations in the Bronx. The specification of the polynomial degrees were derived from the data-driven cubic regression splines with ten knots for the numerical variables, as seen in section D.1 in the appendix. As an extension to the GAM, the geoadditive models include a smoothing term over both latitude and longitude `s(X,Y)`. For each model, Moran tests using the four different spatial weights specifications (Epanechnikov, Gaussian, inverse distance, and negative exponential) were conducted on the residuals to detect how well each model captures the spatial autocorrelation present in the data. Looking at table 3, it is evident that all models, including the a-spatial OLS, do a good job in accounting for the spatial dependence, given the little spatial autocorrelation left in the residuals. This result suggests that the plethora of loca-

tional variables included in the model are capable of accounting for the majority of spatial dependence in house prices per square metre.[31]

Table 3: As a means of identifying how well the different models capture the spatial effects in the data, Moran's I was calculated on the residuals of each model using four different spatial weights specifications.

Method	**Epanechnikov**	**Exponential**	**Gaussian**	**Inverse**
GAM (ps)	0.0066	0.0387	0.0131	0.0143
GAM (cs)	0.0073	0.0372	0.0133	0.0140
polyMESS	0.0074	0.0019	0.0039	0.0123
MESS	0.0080	0.0027	0.0040	0.0124
geoGAM (ps)	0.0135	0.0336	0.0193	0.0080
polyOLS	0.0143	0.0729	0.0006	0.0343
OLS	0.0159	0.0793	0.0013	0.0365
polySTSLS	0.0164	0.0936	0.0020	0.0408
STSLS	0.0173	0.0921	0.0022	0.0406
geoGAM (cs)	0.0178	0.0253	0.0222	0.0036
ESF	0.0235	0.0200	0.0269	0.0042

As a measure of accuracy, the median absolute percentage error (MAPE) was calculated in the untransformed variable space. Additionally, test statistics for heteroskedasticity and normality are reported where possible, using the *Breusch-Pagan* (BP) and *Anderson-Darling* (AD) tests respectively, where the interest lies in comparing the statistics across the models, rather than analysing their respective significance levels (see section B.2 in the appendix as to why). As seen in table 4, the advanced spatial hedonic techniques do not add significant accuracy to the predictive model compared to an ordinary OLS, while adding considerable complexity in interpretation. Apart from eigenvector spatial filtering, all models exhibit similar levels of heteroskedasticity in their residuals. Due to the similarity in performance, but considerably greater simplicity in value attribution to the respective variable categories, the GAM with cubic regression splines using ten knots is considered to be the model of choice. Despite its slightly greater inaccuracy compared to its counterpart, the cubic regression spline approach was preferred due to the sound theoretical interpretation of the resulting smooths and its decreased likelihood of succumbing to overfitting.

[31] For comparison, the spatial autocorrelation of house prices per square metre in the Bronx lies between 0.13 and 0.18 depending on the weights specification.

Table 4: A comparison of different model specifications measured by the median absolute percentage accuracy, along with the corresponding Breusch-Pagen and Anderson-Darling test statistics.

Model	**MAPE**	**BP**	**AD**
ESF	0.2051	188.3065	33.1881
geoGAM (ps)	0.2055	93.0758	36.8090
geoGAM (cs)	0.2064	93.0758	35.4878
GAM (ps)	0.2079	88.7198	36.3714
GAM (cs)	0.2111	88.7198	35.9691
polyMESS	0.2146	98.2239	34.2744
MESS	0.2153	85.2229	35.8551
polySTSLS	0.2163	NA	33.6322
polyOLS	0.2168	99.8716	33.8473
OLS	0.2173	88.7198	35.3637
STSLS	0.2183	NA	35.2107

5.5 Results

A major peculiarity when working with distance based explanatory variables is the interpretation of the intercept. While the intercept can be interpreted as the expected value of the dependent variable when all numerical explanatory variables are zero, zero distance to the nearest locational amenity is physically not possible, rendering the interpretation useless. Alternatively, the distance based variables can be transformed to reflect the distance from a specified reference value, usually a measure of the average. Importantly, centring the explanatory variables changes the interpretation of the intercept, but not that of the regression coefficients (see section C.1 in the appendix for an econometric exposition). Regarding the measure of average in the presence of skewed data, such as is the case with most variables measuring the distance to the nearest amenity, it is often recommended to employ the median rather than the mean, as the former is less susceptible to extreme values. Furthermore, the median is often what is referred to when colloquially saying, for example, "the average person lives x metres away from a subway station". Consequently, all numerical variables were centred prior to the analysis. For the categorical variable *Building class*, the reference value was chosen as that level with the most observations, namely *One family houses*.

The exhaustive list of variables used in the generalised additive model is shown in table 5, where noise complaints and crime were represented by their corresponding primary prin-

cipal component. For all numerical variables, cubic splines with ten knots were specified. The concluding model was then estimated using generalised cross-validation[32] and the `gam` function in the `mgcv` package.

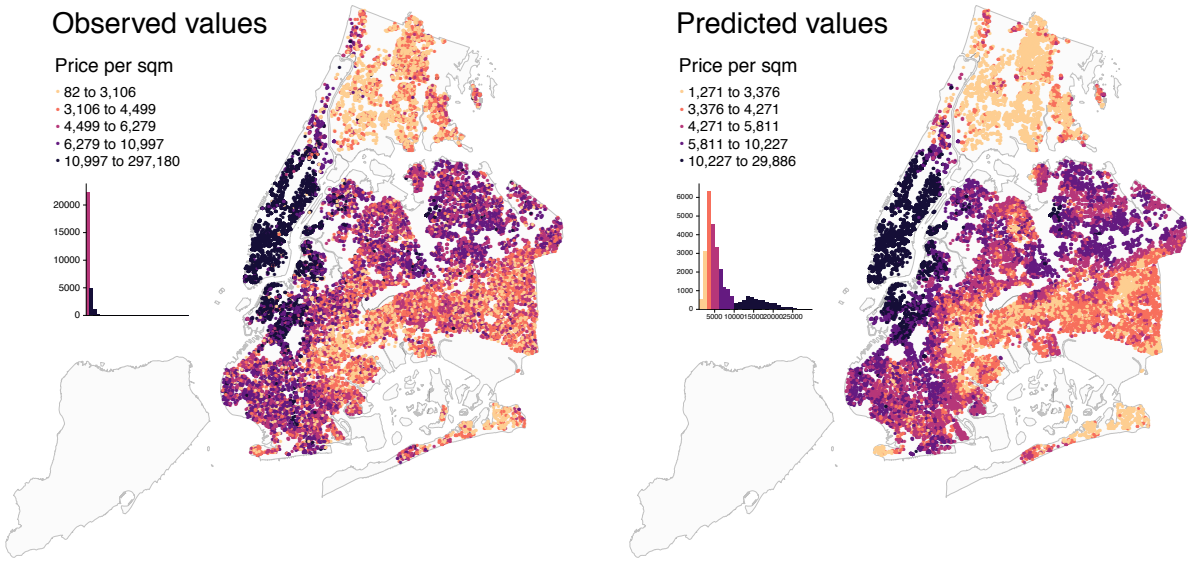
Table 5: An overview of all the variables initially defined for the model.

Locational variables	**Structural variables**	**Neighbourhood variables**
Highschool quality	Age	Noise (complaints) (n)
Restaurants, cafes, bars (n)	Building class	Crime (shootings) (n)
Parks (a)		
Highschool distance (d)		
Airports (d)		
Fitness centres (d)		
Playgrounds (d)		
Railways (d)		
Shopping centres (d)		
Supermarket (d)		
Convenience stores (d)		
Stock exchange (d)		
Roads (d)		
Waterfront access (d)		
Cinema (d)		

Note: (a) = area, (d) = distance, (n) = number

The model has a median absolute percentage error of 20 percent on the entire data set containing the four boroughs of Manhattan, the Bronx, Brooklyn, and Queens, while the *Moran's I* for the residuals corresponds to 0.1 for the Epanechnikov spatial weights matrix. This result is quite remarkable given the lack of any neighbourhood dummy variable or spatial smoothing parameter and the *Moran's I* of 0.60 in the dependent variable. A graphical depiction of the model accuracy may be obtained by comparing the observed and fitted values, displayed in figure 3, and by viewing the spatial distribution of residuals, displayed in the appendix in figure 24.

[32] A comparison with *restricted maximum likelihood (REML)* only showed marginal improvements with substantially longer estimation times.



(a) Observed Values (b) Predicted Values

Figure 3: Two maps portraying the observed values and predicted values.

It is evident that the model is able to capture the spatial patterns of housing prices fairly well, especially considerin the lack of spatial pattern remaining in the residuals map. In order to analyse the pricing implications of housing location, the price per square metre for each observation was estimated by terms (for each variable individually), such that all individual locational variable values could be aggregated to form an overall locational value, with the same being done for the structural and neighbourhood variables. An excerpt of the predicted values in the transformed variable space for each explanatory variable showing the first five observations can be found in table 11 in the appendix, while the summary statistics of the aggregated values can be seen in table 6.

Table 6: Summary output of the estimated locational, structural, and neighbourhood values of an average one family house in Queens, along with the residual values.

Locational value	**Neighbourhood value**	**Structural value**	**Residual value**
Min.: 1997	Min.: 5231	Min.: 3060	Min.: -22117.2
1st Qu.: 4112	1st Qu.: 5748	1st Qu.: 5164	1st Qu.: -980.3
Median: 5372	Median: 6135	Median: 5950	Median: 155.2
Mean: 6789	Mean: 6055	Mean: 6009	Mean: 499.9
3rd Qu.: 7573	3rd Qu.: 6391	3rd Qu.: 6691	3rd Qu.: 1344.4
Max.: 24331	Max.: 10787	Max.: 9662	Max.: 271761.3

The values in table 6 refer to the respective estimated values of houses of otherwise average houses.[33] For instance, the minimum neighbourhood value represents the value of a house in the worst neighbourhood with average (median) locational and structural characteristics. As such, we find that the best location in New York is 3.5 times more expensive than the average location, while the locational value in the 75th percentile demands a 41 percent premium. On the flip side, the worst location is estimated to be compensated with a 63 percent discount compared to the average house, while the locational value in the 25th percentile still receives an estimated 23 percent discount. The spread of neighbourhood and structural value is estimated to be much smaller, with an estimated maximum 75 percent and 62 percent neighbourhood premium and structural premium, respectively. However, the respective 75th percentiles only command an estimated 4 and 12 percent neighbourhood and structural premium. Similar values are estimated for the respective discounts, where the worst houses obtain a 15 and 49 percent neighbourhood and structural discount, respectively, while the 75th percentile estimates merely enjoy a 6 and 13 percent discount, respectively. Furthermore, an impressive 58 percent of variation can be explained by the structural value alone using a simple OLS regression.

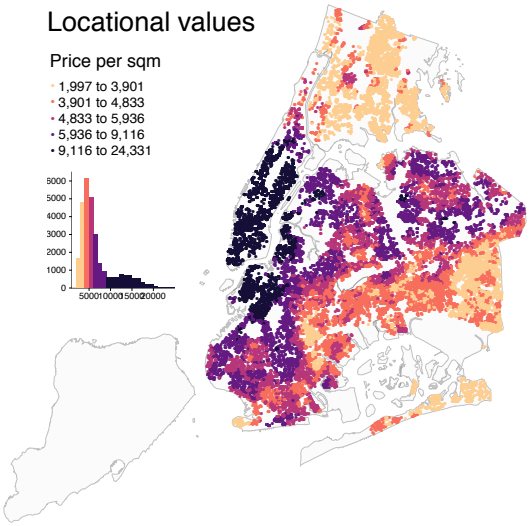


Figure 4: A map with the predicted locational value for each observation in the data set.

Figure 4 depicts the spatial distribution of estimated locational value, showcasing a desir-

[33] This reporting approach follows Linneman (1980), who calculated the difference of housing value between two hypothetical structurally equivalent houses with neighbourhood qualities at either side of the spectrum.

able amount of spatial autocorrelation, given that neighbouring observations should exhibit similar locational values. Maps displaying the spatial distribution of neighbourhood value and structural value can be found in the appendix in figures 22 and 23, respectively.

6 Discussion

Picking up on the result maps depicted in the figures 21, 22, 23 and 24 in the appendix, the structural values seem to exhibit the least amount of spatial data, which is in line with theoretical considerations. Nevertheless, one may state that properties in Manhattan are of considerably higher structural quality than in the remaining boroughs, while Brooklyn showcases the greatest variety. Another interesting aspect is that buildings in Queens are of rather a consistently decent build, showing notably more spatial homogeneity compared to both the Bronx and Brooklyn, which may be due to historical developments. Furthermore, there is a slight tendency of housing being of below average structural quality in the crime ridden neighbourhoods of the Bronx (around Harlem and Melrose) and Brooklyn (around Brownsville and Bushwick). The neighbourhood value strongly reflects the spatial distribution of shootings and noise complaints, due to the limited availability of further neighbourhood values. Nevertheless, the resulting map highlights the above average neighbourhood quality in Queens, in particular around the neighbourhoods of Murray Hill in the north and St. Albans in the south. Furthermore, the neighbourhood values reflect the rather noisy and tourist ridden Manhattan. The locational values exhibit strong positive tendencies towards Manhattan and the New York stock exchange, suggesting that the distance to work and availability of "urban" amenities is highly valued by New Yorkers. While the maximum locational value premium of 350 percent does seem rather generous, it may very well be adequate for cities with significant concentrations of billionaires, like New York. Furthermore, the sheer size of the city would suggest greater locational value premiums compared to smaller cities, like Zurich, alone due to the reduction in commuting time.

In order to underpin the validity of the results, the most accurate model in the model comparison, ESF, was also applied to the entire data set. Compared to the GAM model with cubic regression splines, the ESF is only able to report accuracy gains of 1.2 percent, resulting in a MAPE of 18.8 percent. As such, it is found that inclusion of spatial econometric techniques beyond nonlinear modelling does not significantly increase the predictive power of a model. These results are shared by Bourassa et al. (2007) and Hill and Scholz (2018), who also did not find significant improvements using spatial econometric techniques in their

predictive studies of house prices.

6.1 Limitations

One major drawback of the model pertains to the unavailability of quality measures and views. The valuation of very expensive properties, listed in table 7, are expected to be driven by these two characteristics in particular. Consequently, the model employed in this thesis is expected to fail at explaining such exorbitant prices, as well as the unrealistically low prices. This limitation is also alluded to in the diagnostic plots, seen in figure 14, where the residual distribution remains tail heavy. Further analysing this peculiarity, outliers were removed from the data set, where outliers were classified as such if the *Cook's distance* was greater the $\frac{4}{28'826}$. However, the spatial distribution of the outliers, as depicted in figure 13 in the appendix, suggests that views may not have such great predictive power of whether an observation is an outlier or not after all. Running the same GAM on the filtered data set resulted in an reduction in mean absolute percentage error (not median) from 43 percent to 26 percent, and significant improvements in the diagnostic plots (as seen in figure 15 in the appendix).

However, the removal of extreme values only offers a fractional improvement of the median percentage error, namely from 20 percent to 18.9 percent, suggesting that a considerable amount of variation remains unexplained. One possible reason for the remaining variation, in addition to the absence of quality and view measures, may be that the distance based measures of proximity are insufficient proxies for travelling time. Likewise, measurement errors may pertain to noise levels or air quality, to name the most likely victims. However, more accurate measures of variables are not expected to substantially increase the predictive power of the model, given the modest improvement attained by the nonparametric model specification compared to OLS. A further reason may be linked to lack of freedom in locational choice and timing. Bayer et al. (2009) show that when moving between housing locations is costly, the variation in housing prices across locations may no longer reflect the value of difference in local amenities. Lastly, the residual noise may be suggestive of an inefficient real estate market, where prospective homebuyers are unaware of the "true" market value of a house.

To further the literature of housing demand, future research may address the above mentioned issues of variable availability in creating a more comprehensive data set than was available in this thesis. Especially in regard to quality measurements of amenities, Google

Maps ratings offer a promising solution to obtain vast amounts of data with reasonably impartial assessments. On the econometric front, further research is required to better understand the role of spatial autocorrelation in predictive models.

7 Conclusion

This thesis analyses the pricing implications of housing location, leveraging a data set containing both the prices and structural characteristics of properties sold in New York City throughout 2018, which is augmented with locational data obtained from OpenStreetMaps as well as data available from New York City's open data portal. To account for spatial autocorrelation and nonlinearity, a generalised additive model was employed and compared against a battery of alternative spatial and a-spatial estimation techniques, including Moran eigenvector spatial filtering, geoadditive models, spatial two stage least squares and a matrix exponential spatial specification. The model, which includes 15 locational variables, two structural variables and two neighbourhood variables estimates results which are in support of the common belief that location is the most important factor in house prices, as it was found that 58 percent of house price variation can be explained by the locational value alone. As such, this thesis capitalises on an extensive set of locational characteristics obtained from openly available sources to further the understanding of location and its importance for house prices, and applies several econometric advances to address the shortcomings of the traditional hedonic approach.

8 Acknowledgments

The empirical analysis was conducted using the *R* open source software (R Core Team, 2013) with a plethora of packages being used. The exhaustive list of packages employed consists of Dowle and Srinivasan (2019), Zhu (2019), Pebesma (2018), Tennekes (2018), Padgham, Rudis, Lovelace, and Salmon (2017), Bivand et al. (2013), Murakami (2019), Wickham (2017), Papadakis, Tsagris, Dimitriadis, Fafalios, Tsamardinos, Fasiolo, Borboudakis, Burkardt, Zou, Lakiotaki, and Chatzipantsiou. (2019), Bivand, Keitt, and Rowlingson (2019), Strimas-Mackey (2018), Wickham (2019), Izrailev (2019), Fox and Weisberg (2019a), Fox and Weisberg (2019b), Gavrilov and Pusev (2014), Gross and Ligges (2015), Trapletti and Hornik (2019), Dorman (2019), Grosjean and Ibanez (2018), Xie (2019a),

Zeileis and Hothorn (2002), and Xie (2019b).

```
sf.playground <- c(sf.playground_points, sf.playground_centroids) %>%
  st_transform(32118) %>%
  st_combine()
rm(list = c("sf.playground_centroids", "sf.playground_points"))

# Calculate distance to closest playground
sf.nyc$d_playground <- st_distance(sf.nyc, sf.playground) %>%
  as.numeric() %>%
  `/`(100)
sf.nyc$d_playground_unit <- "100m"
```

Listing 2: Appropriate combination of polygon and point data.

References

Ahlfeldt, G. M., Maennig, W., 2010. Impact of sports arenas on land values: evidence from berlin. The Annals of Regional Science 44, 205–227.

Alonso, W., et al., 1964. Location and land use. toward a general theory of land rent. Location and land use. Toward a general theory of land rent. .

Andersson, H., Jonsson, L., Ögren, M., 2010. Property prices and exposure to multiple noise sources: Hedonic regression with road and railway noise. Environmental and resource economics 45, 73–89.

Andreyeva, E., Patrick, C., 2017. Paying for priority in school choice: Capitalization effects of charter school admission zones. Journal of Urban Economics 100, 19–32.

Anselin, L., 1988. Spatial econometrics : methods and models. Studies in operational regional science, Kluwer, Dordrecht.

Anselin, L., 1995. Local indicators of spatial association—lisa. Geographical analysis 27, 93–115.

Anselin, L., 2010. Thirty years of spatial econometrics. Papers in regional science 89, 3–25.

Anselin, L., Gallo, J. L., 2006. Interpolation of air quality measures in hedonic house price models: Spatial aspects. Spatial Economic Analysis 1, 31–52.

Anselin, L., Lozano-Gracia, N., 2008. Errors in variables and spatial effects in hedonic house price models of ambient air quality. Empirical economics 34, 5–34.

Anselin, L., Lozano-Gracia, N., 2009. Spatial hedonic models. In: *Palgrave handbook of econometrics*, Springer, pp. 1213–1250.

Anselin, L., Lozano-Gracia, N., Deichmann, U., Lall, S., 2010. Valuing access to water—a spatial hedonic approach, with an application to bangalore, india. Spatial Economic Analysis 5, 161–179.

Arguea, N. M., Hsiao, C., 2000. Market values of environmental amenities: A latent variable approach. Journal of Housing Economics 9, 104–126.

Armstrong, R., Rodríguez, D., 2006. An evaluation of the accessibility benefits of commuter rail in eastern massachusetts using spatial hedonic price functions. Transportation 33, 21–43.

Atack, J., Margo, R. A., 1998. "location, location, location!" the price gradient for vacant urban land: New york, 1835 to 1900. The Journal of Real Estate Finance and Economics 16, 151–172.

Bajari, P., Benkard, C. L., 2005. Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach. Journal of political economy 113, 1239–1276.

Baltagi, B., 2008. Econometric analysis of panel data. John Wiley & Sons.

Baltagi, B. H., Song, S. H., Koh, W., 2003. Testing panel data regression models with spatial error correlation. Journal of Econometrics 117, 123 – 150.

Basile, R., Durbán, M., Mínguez, R., Montero, J. M., Mur, J., 2014. Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. Journal of Economic Dynamics and Control 48, 229–245.

Bateman, I., Day, B., Lake, I., Lovett, A., 2001. The effect of road traffic on residential property values: a literature review and hedonic pricing study. Prepared for Scottish Executive and The Stationary Office, Edinburgh, Scotland .

Bayer, P., Keohane, N., Timmins, C., 2009. Migration and hedonic valuation: The case of air quality. Journal of Environmental Economics and Management 58, 1–14.

Benson, E. D., Hansen, J. L., Schwartz, A. L., Smersh, G. T., 1998. Pricing residential amenities: the value of a view. The Journal of Real Estate Finance and Economics 16, 55–73.

Besley, T., Mueller, H., 2012. Estimating the peace dividend: The impact of violence on house prices in northern ireland. American Economic Review 102, 810–33.

Bin, O., 2005. A semiparametric hedonic model for valuing wetlands. Applied Economics Letters 12, 597–601.

Bivand, R., Keitt, T., Rowlingson, B., 2019. rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.4-6.

Bivand, R., Wong, D. W. S., 2018. Comparing implementations of global and local indicators of spatial association. TEST 27, 716–748.

Bivand, R. S., Pebesma, E., Gomez-Rubio, V., 2013. Applied spatial data analysis with R, Second edition. Springer, NY.

Black, D., et al., 1958. The theory of committees and elections. Springer.

Bond, M. T., Seiler, M. J., 1996. School Quality and Housing Values. Ohio State University. Center for Real Estate Education and Research, Max M ....

Bontemps, C., Simioni, M., Surry, Y., 2008. Semiparametric hedonic price models: assessing the effects of agricultural nonpoint source pollution. Journal of Applied Econometrics 23, 825–842.

Bourassa, S. C., Cantoni, E., Hoesli, M., 2007. Spatial dependence, housing submarkets, and house price prediction. The Journal of Real Estate Finance and Economics 35, 143–160.

Bowen, H. R., 1943. The interpretation of voting in the allocation of economic resources. The Quarterly Journal of Economics 58, 27–48.

Bowen, W. M., Mikelbank, B. A., Prestegaard, D. M., 2001. Theoretical and empirical considerations regarding space in hedonic housing price model applications. Growth and change 32, 466–490.

Box, G. E., Tidwell, P. W., 1962. Transformation of the independent variables. Technometrics 4, 531–550.

Box, G. E. P., Cox, D. R., 1964. An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological) 26, 211–252.

Brasington, D. M., Hite, D., 2005. Demand for environmental quality: a spatial hedonic analysis. Regional science and urban economics 35, 57–82.

Breunig, R., Hasan, S., Whiteoak, K., 2018. Impact of playgrounds on property prices: evidence from australia. In: *Australian Conference of Economists*.

Brown, G. M., Pollakowski, H. O., 1977. Economic valuation of shoreline. The review of Economics and Statistics pp. 272–278.

Calhoun, C. A., 2001. Property valuation methods and data in the united states. Housing Finance International 16, 12–23.

Cameron, T. A., 2006. Directional heterogeneity in distance profiles in hedonic property value models. Journal of Environmental Economics and Management 51, 26–45.

Can, A., 1990. The measurement of neighborhood dynamics in urban house prices. Economic geography 66, 254–272.

Case, K. E., Quigley, J. M., Shiller, R. J., 2005. Comparing wealth effects: the stock market versus the housing market. Advances in macroeconomics 5.

Cassel, E., Mendelsohn, R., 1985. The choice of functional forms for hedonic price equations: comment. Journal of Urban Economics 18, 135–142.

Chattopadhyay, S., 1999. Estimating the demand for air quality: new evidence based on the chicago housing market. Land Economics pp. 22–38.

Chattopadhyay, S., 2000. The effectiveness of mcfaddens's nested logit model in valuing amenity improvement. Regional Science and Urban Economics 30, 23–43.

Chau, K. W., Chin, T., 2003. A critical review of literature on the hedonic price model. International Journal for Housing Science and Its Applications 27, 145–165.

Chay, K. Y., Greenstone, M., 2005. Does air quality matter? evidence from the housing market. Journal of political Economy 113, 376–424.

Chiang, Y.-H., Peng, T.-C., Chang, C.-O., 2015. The nonlinear effect of convenience stores on residential property prices: A case study of taipei, taiwan. Habitat International 46, 82–90.

Chiu, T. Y., Leonard, T., Tsui, K.-W., 1996. The matrix-logarithmic covariance model. Journal of the American Statistical Association 91, 198–210.

Cho, S.-H., Bowker, J. M., Park, W. M., 2006. Measuring the contribution of water and green space amenities to housing values: An application and comparison of spatially weighted hedonic models. Journal of agricultural and resource economics pp. 485–507.

Choy, L. H., Mak, S. W., Ho, W. K., 2007. Modeling hong kong real estate prices. Journal of Housing and the Built Environment 22, 359–368.

Chun, Y., 2014. Analyzing space–time crime incidents using eigenvector spatial filtering: an application to vehicle burglary. Geographical Analysis 46, 165–184.

Clark, D. E., Herrin, W. E., 2000. The impact of public school attributes on home sale prices in california. Growth and change 31, 385–407.

Clauretie, T. M., Neill, H. R., 2000. Year-round school schedules and residential property values. The Journal of Real Estate Finance and Economics 20, 311–322.

Cli, A., Ord, J. K., 1981. Spatial processes: models and applications. Pion, London .

Cliff, A., Ord, K., 1972. Testing for spatial autocorrelation among regression residuals. Geographical analysis 4, 267–284.

Cohen, J. P., Coughlin, C. C., 2008. Spatial hedonic models of airport noise, proximity, and housing prices. Journal of regional science 48, 859–878.

Correll, M. R., Lillydahl, J. H., Singell, L. D., 1978. The effects of greenbelts on residential property values: some findings on the political economy of open space. Land economics 54, 207–217.

Crompton, J. L., 2001. The impact of parks on property values: A review of the empirical evidence. Journal of Leisure Research 33, 1–31.

Cropper, M. L., Deck, L. B., Mcconnell, K. E., 1988. On the choice of funtional form for hedonic price functions. The Review of Economics and Statistics 70, 668–675.

Danton, J., Himbert, A., 2018. Residential vertical rent curves. Journal of Urban Economics 107, 89–100.

Darling, A. H., 1973. Measuring benefits generated by urban water parks. Land economics 49, 22–34.

Davidson, R., MacKinnon, J. G., 1985. Testing linear and loglinear regressions against box-cox alternatives. Canadian Journal of Economics pp. 499–517.

De Jong, P., Sprenger, C., Van Veen, F., 1984. On extreme values of moran's i and geary's c. Geographical Analysis 16, 17–24.

Del Giudice, V., Manganelli, B., De Paola, P., 2015. Spline smoothing for estimating hedonic housing price models. In: Gervasi, O., Murgante, B., Misra, S., Gavrilova, M. L., Rocha, A. M. A. C., Torre, C., Taniar, D., Apduhan, B. O. (eds.), *Computational Science and Its Applications – ICCSA 2015*, Springer International Publishing, Cham, pp. 210–219.

Des Rosiers, F., Lagana, A., Thériault, M., Beaudoin, M., 1996. Shopping centres and house values: an empirical investigation. Journal of Property Valuation and Investment 14, 41–62.

DiPasquale, D., Wheaton, W. C., 1996. Urban economics and real estate markets, vol. 23. Prentice Hall Englewood Cliffs, NJ.

Doling, J., 1973. A two-stage model of tenure choice in the housing market. Urban Studies 10, 199–211.

Dorman, M., 2019. nngeo: k-Nearest Neighbor Join for Spatial Data. R package version 0.3.0.

Dowle, M., Srinivasan, A., 2019. data.table: Extension of 'data.frame'. R package version 1.12.6.

Downes, T. A., Zabel, J. E., 2002. The impact of school characteristics on house prices: Chicago 1987-1991. Journal of urban economics 52, 1–25.

Downs, A., 1957. An economic theory of political action in a democracy. Journal of political economy 65, 135–150.

Dubin, R. A., 1988. Estimation of regression coefficients in the presence of spatially autocorrelated error terms. The Review of Economics and Statistics pp. 466–474.

Dubin, R. A., Sung, C.-H., 1990. Specification of hedonic regressions: non-nested tests on measures of neighborhood quality. Journal of Urban Economics 27, 97–110.

Earnhart, D., 2001. Combining revealed and stated preference methods to value environmental amenities at residential locations. Land economics 77, 12–29.

El-Geneidy, A., Grimsrud, M., Wasfi, R., Tétreault, P., Surprenant-Legault, J., 2014. New evidence on walking distances to transit stops: identifying redundancies and gaps using variable service areas. Transportation 41, 193–210.

Ellickson, B., 1981. An alternative test of the hedonic theory of housing markets. Journal of Urban Economics 9, 56–79.

Epple, D., 1987. Hedonic prices and implicit markets: estimating demand and supply functions for differentiated products. Journal of political economy 95, 59–80.

epsg, 2020. Epsg:32118 - nad83 / new york long island. https://epsg.io/32118.

Espey, M., Owusu-Edusei, K., 2001. Neighborhood parks and residential property values in greenville, south carolina .

Fan, G.-Z., Ong, S. E., Koh, H. C., 2006. Determinants of house price: A decision tree approach. Urban Studies 43, 2301–2315.

Federal Statistical Office, 2019. Detaillierte haushaltsausgaben sämtlicher haushalte nach jahr. https://www.bfs.admin.ch/bfs/en/home/statistics/economic-social-situation-population/income-consumption-wealth/household-budget/household-expenditure.assetdetail.10867225.html.

Feitelson, E. I., Hurd, R. E., Mudge, R. R., 1996. The impact of airport noise on willingness to pay for residences. Transportation Research Part D: Transport and Environment 1, 1–14.

Feng, X., Humphreys, B., 2018. Assessing the economic impact of sports facilities on residential property values: A spatial hedonic approach. Journal of Sports Economics 19, 188–210.

Fischer, M., Füss, R., Stehle, S., 2020. Local house price comovements, working paper.

Fischer, M. M., 2014. Handbook of regional science : 1, vol. 1. Springer, Berlin.

Follain, J. R., Jimenez, E., 1985. Estimating the demand for housing characteristics: a survey and critique. Regional science and urban economics 15, 77–107.

Forrest, D., Glen, J., Ward, R., 1996. The impact of a light rail system on the structure of house prices: a hedonic longitudinal study. Journal of Transport Economics and Policy pp. 15–29.

Fox, J., Weisberg, S., 2019a. An R Companion to Applied Regression. Sage, Thousand Oaks CA, third ed.

Fox, J., Weisberg, S., 2019b. An R Companion to Applied Regression. Sage, Thousand Oaks CA, third ed.

Fox, T., 1990. Urban open space: An investment that pays. A Monograph series (USA) .

Freeman, A. M., 1981. Hedonic prices, property values and measuring environmental benefits: a survey of the issues. In: *Measurement in public choice*, Springer, pp. 13–32.

Friedman, J., 1981. A conditional logit model of the role of local public services in residential choice. Urban Studies 18, 347–358.

Galilea, P., de Dios Ortuzar, J., 2005. Valuing noise level reductions in a residential location context. Transportation Research Part D: Transport and Environment 10, 305–322.

Gallimore, P., Fletcher, M., Carter, M., 1996. Modelling the influence of location on value. Journal of Property Valuation and Investment 14, 6–19.

Garrod, G. D., Willis, K. G., 1992. Valuing goods' characteristics: an application of the hedonic price method to environmental attributes. Journal of Environmental management 34, 59–76.

Gavrilov, I., Pusev, R., 2014. normtest: Tests for Normality. R package version 1.1.

Geary, R. C., 1954. The contiguity ratio and statistical mapping. The incorporated statistician 5, 115–146.

Geniaux, G., Napoléone, C., 2008. Semi-parametric tools for spatial hedonic models: an introduction to mixed geographically weighted regression and geoadditive models. In: *Hedonic methods in housing markets*, Springer, pp. 101–127.

Getis, A., Ord, K., 1992. The analysis of spatial association by use of distance statistics. Geographical Analysis 24, 189 – 206.

Gibbons, S., Overman, H. G., 2012. Mostly pointless spatial econometrics? Journal of Regional Science 52, 172–191.

Gillard, Q., 1981. The effect of environmental amenities on house values: the example of a view lot. The Professional Geographer 33, 216–220.

Glaeser, E. L., Gyourko, J., Morales, E., Nathanson, C. G., 2014. Housing dynamics: An urban approach. Journal of Urban Economics 81, 45–56.

Goodman, A. C., 1986. Neighborhood impacts on housing prices. In: *Urban Neighborhoods, Research and Policy*, Praeger New York.

Griffith, D., Chun, Y., 2014. Spatial autocorrelation and spatial filtering. Handbook of regional science pp. 1477–1507.

Griffith, D., Chun, Y., 2016. Evaluating eigenvector spatial filter corrections for omitted georeferenced variables. Econometrics 4, 29.

Griffith, D. A., 1989. Economic Geography 65, 160–162.

Griffith, D. A., 2017. Some robustness assessments of moran eigenvector spatial filtering. Spatial Statistics 22, 155 – 179.

Griffith, D. A., 2019. Negative spatial autocorrelation: One of the most neglected concepts in spatial statistics. Stats 2, 388–415.

Griffith, D. A., Paelinck, J. H. P., 2011. Non-standard spatial statistics and spatial econometrics. Advances in geographic information science, Springer, Berlin.

Griliches, Z., 1961. Hedonic price indexes for automobiles: An econometric of quality change. In: *The price statistics of the federal goverment*, NBER, pp. 173–196.

Griliches, Z., 1971. Price indexes and quality change : studies in new methods of measurement. Harvard Univ. Press, Cambridge, Mass.

Grosjean, P., Ibanez, F., 2018. pastecs: Package for Analysis of Space-Time Ecological Series. R package version 1.3.21.

Gross, J., Ligges, U., 2015. nortest: Tests for Normality. R package version 1.0-4.

Halvorsen, R., Pollakowski, H. O., 1981. Choice of functional form for hedonic price equations. Journal of urban economics 10, 37–49.

Hamilton, B. W., Röell, A., 1982. Wasteful commuting. Journal of political economy 90, 1035–1053.

Harrison Jr, D., Rubinfeld, D. L., 1978. Hedonic housing prices and the demand for clean air. Journal of environmental economics and management 5, 81–102.

Hastie, T., 1984. Principal curves and surfaces. Tech. rep., STANFORD UNIV CA LAB FOR COMPUTATIONALSTATISTICS.

Hastie, T., Tibshirani, R., 1986. Generalized additive models. Statist. Sci. 1, 297–310.

Hastie, T., Tibshirani, R., 1987. Generalized additive models: some applications. Journal of the American Statistical Association 82, 371–386.

Haurin, D. R., Brasington, D. M., et al., 1996. The impact of school quality on real house prices: Interjurisdictional effects. Journal of Housing Economics 5, 351–368.

Helbich, M., Brunauer, W., Vaz, E., Nijkamp, P., 2014. Spatial heterogeneity in hedonic house price models: The case of austria. Urban Studies 51, 390–411.

Hill, R. J., Scholz, M., 2018. Can geospatial data improve house price indexes? a hedonic imputation approach with splines. Review of Income and Wealth 64, 737–756.

Holcombe, R. G., 1989. The median voter model in public choice theory. Public Choice 61, 115–125.

Horn, R., Johnson, C., 1985. Matrix Analysis. Cambridge University Press, second ed.

Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28, 321–377.

Izrailev, S., 2019. tictoc: Functions for timing R scripts, as well as implementations of Stack and List structures. R package version 1.0.

Jud, G., 1985. A further note on schools and housing values. AREUEA Journal 13, 452.

Jud, G. D., Watts, J. M., 1981. Schools and housing values. Land Economics 57, 459–470.

Judge, G. G., 1982. Introduction to the theory and practice of econometrics. Wiley series in probability and mathematical statistics : Applied probablity and statistics, Wiley, New York, NY.

Kain, J. F., Quigley, J. M., 1970. Measuring the value of housing quality. Journal of the American statistical association 65, 532–548.

Kammann, E. E., Wand, M. P., 2003. Geoadditive models. Journal of the Royal Statistical Society: Series C (Applied Statistics) 52, 1–18.

Kelejian, H. H., Prucha, I. R., 1998. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. The Journal of Real Estate Finance and Economics 17, 99–121.

Kelejian, H. H., Prucha, I. R., 1999. A generalized moments estimator for the autoregressive parameter in a spatial model. International economic review 40, 509–533.

Kelejian, H. H., Prucha, I. R., 2007. Hac estimation in a spatial framework. Journal of Econometrics 140, 131 – 154, analysis of spatially dependent data.

Kennedy, M., Kopp, S., et al., 2000. Understanding map projections. Esri Redlands, CA.

Kim, C. W., Phipps, T. T., Anselin, L., 2003. Measuring the benefits of air quality improvement: a spatial hedonic approach. Journal of environmental economics and management 45, 24–39.

Kim, S., Lee, K. O., 2018. Potential crime risk and housing market responses. Journal of Urban Economics 108, 1–17.

Kuang, C., 2017. Does quality matter in local consumption amenities? an empirical investigation with yelp. Journal of Urban Economics 100, 1–18.

Kwong, L. M., Cyr, D., Kushner, J., Ogwang, T., 2011. A semiparametric hedonic pricing model of ontario wines. Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie 59, 361–381.

Lambert, D. M., Xu, W., Florax, R. J., 2014. Partial adjustment analysis of income and jobs, and growth regimes in the appalachian region with smooth transition spatial process models. International Regional Science Review 37, 328–364.

Lancaster, K. J., 1966. A new approach to consumer theory. Journal of political economy 74, 132–157.

Lee, L.-F., 2004. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. Econometrica 72, 1899–1925.

Lee, L.-F., 2007. Gmm and 2sls estimation of mixed regressive, spatial autoregressive models. Journal of Econometrics 137, 489 – 514.

Lehner, M., 2011. Modelling housing prices in singapore applying spatial hedonic regression. Institute for Transport Planning and Systems (IVT), ETH Zurich, Zurich .

Lerman, S. R., 1979. Neighborhood choice and transportation services. In: *The economics of neighborhood*, Elsevier, pp. 83–118.

LeSage, J., Pace, R. K., 2009. Introduction to spatial econometrics. Chapman and Hall/CRC.

LeSage, J. P., Pace, R. K., 2007. A matrix exponential spatial specification. Journal of Econometrics 140, 190–214.

Li, M. M., Brown, H. J., 1980. Micro-neighborhood externalities and hedonic housing prices. Land economics 56.

Liao, W.-C., Wang, X., 2012. Hedonic house prices and spatial quantile regression. Journal of Housing Economics 21, 16–27.

Limsombunchai, V., 2004. House price prediction: hedonic price model vs. artificial neural network. In: *New Zealand Agricultural and Resource Economics Society Conference*, pp. 25–26.

Linneman, P., 1980. Some empirical results on the nature of the hedonic price function for the urban housing market. Journal of urban economics 8, 47–68.

Linneman, P., 1985. An economic analysis of the homeownership decision. Journal of Urban Economics 17, 230 – 246.

Liu, C. H., Rosenthal, S. S., Strange, W. C., 2018. The vertical city: Rent gradients, spatial structure, and agglomeration economies. Journal of Urban Economics 106, 101–122.

Livy, M. R., 2017. The effect of local amenities on house price appreciation amid market shocks: The case of school quality. Journal of Housing Economics 36, 62–72.

Luttik, J., 2000. The value of trees, water and open space as reflected by house prices in the netherlands. Landscape and urban planning 48, 161–167.

Lynch, A. K., Rasmussen, D. W., 2001. Measuring the impact of crime on house prices. Applied Economics 33, 1981–1989.

Mahan, B. L., Polasky, S., Adams, R. M., 2000. Valuing urban wetlands: a property price approach. Land economics pp. 100–113.

Mark, J., Goldberg, M. A., 1988. Multiple regression analysis and mass assessment: A review of the issues. Appraisal Journal 56, 89.

Marra, G., Radice, R., 2010. Penalised regression splines: theory and application to medical research. Statistical Methods in Medical Research 19, 107–125.

Masiero, L., Heo, C. Y., Pan, B., 2015. Determining guests' willingness to pay for hotel room attributes with a discrete choice model. International Journal of Hospitality Management 49, 117–124.

Mayer, T., Trevien, C., 2017. The impact of urban public transportation evidence from the paris region. Journal of Urban Economics 102, 1–21.

McCord, M., McCord, J., Davis, P., Haran, M., Bidanset, P., 2019. House price estimation using an eigenvector spatial filtering approach. International Journal of Housing Markets and Analysis .

McFadden, D., 1978. Modeling the choice of residential location. Transportation Research Record .

McMillen, D. P., 2010. Issues in spatial data analysis. Journal of Regional Science 50, 119–141.

McMillen, D. P., 2012. Perspectives on spatial econometrics: linear smoothing with structured models. Journal of Regional Science 52, 192–209.

McMillen, D. P., Jarmin, R., Thorsnes, P., 1992. Selection bias and land development in the monocentric city model. Journal of Urban Economics 31, 273–284.

McMillen, D. P., McDonald, J., 2004. Reaction of house prices to a new rapid transit line: Chicago's midway line, 1983–1999. Real Estate Economics 32, 463–486.

Mehaffy, M. W., Porta, S., Romice, O., 2015. The "neighborhood unit" on trial: a case study in the impacts of urban morphology. Journal of Urbanism: International Research on Placemaking and Urban Sustainability 8, 199–217.

Millo, G., Piras, G., 2012. splm: Spatial panel data models in R. Journal of Statistical Software 47, 1–38.

Mills, E. S., 1972. Studies in the structure of the urban economy. .

Mills, E. S., 1990. Housing tenure choice. The Journal of Real Estate Finance and Economics 3, 323–331.

Mills, T., Patterson, K., 2009. Palgrave Handbook of Econometrics: Volume 2: Applied Econometrics, vol. 2. Springer.

Mok, H. M., Chan, P. P., Cho, Y.-S., 1995. A hedonic price model for private properties in hong kong. The Journal of Real Estate Finance and Economics 10, 37–48.

Molin, E. J., 2011. Conjoint analysis. In: *The measurement and analysis of housing preference and choice*, Springer, Dordrecht, pp. 127–155.

Murakami, D., 2019. spmoran: Moran Eigenvector-Based Spatial Regression Models. R package version 0.1.7.1.

Muth, R., 1969. Cities and Housing: The Spatial Pattern of Urban Residential Land Use. Graduate School of Business, University of Chicago. Third series: Studies in business and society, University of Chicago Press.

Nechyba, T. J., Strauss, R. P., 1998. Community choice and local public services: A discrete choice approach. Regional Science and Urban Economics 28, 51–73.

Neill, H. R., Hassenzahl, D. M., Assane, D. D., 2007. Estimating the effect of air quality: spatial versus traditional hedonic price models. Southern Economic Journal pp. 1088–1111.

Ng, C. F., 2008. Commuting distances in a household location choice model with amenities. Journal of Urban Economics 63, 116–129.

Nguyen-Hoang, P., Yinger, J., 2011. The capitalization of school quality into house values: A review. Journal of Housing Economics 20, 30 – 48.

NYC Department of Finance, 2020. Property tax highlights. https://www1.nyc.gov/site/finance/taxes/property.page.

Oates, W. E., 1969. The effects of property taxes and local public spending on property values: An empirical study of tax capitalization and the tiebout hypothesis. Journal of political economy 77, 957–971.

Osborne, J. W., 2010. Improving your data transformations: Applying the box-cox transformation. Practical Assessment, Research & Evaluation 15, 1–9.

Pace, K., 1998. Appraisal using generalized additive models. Journal of Real Estate Research 15, 77–99.

Pace, R., LeSage, J. P., 2004. Chebyshev approximation of log-determinants of spatial weight matrices. Computational Statistics & Data Analysis 45, 179 – 196.

Pace, R. K., Barry, R., Slawson, V. C., Sirmans, C., 2004. Simultaneous spatial and functional form transformations. In: *Advances in Spatial Econometrics*, Springer, pp. 197–224.

Pace, R. K., LeSage, J. P., 2010. Omitted variable biases of ols and spatial lag models. In: *Progress in Spatial Analysis*, Springer, pp. 17–28.

Pace, R. K., LeSage, J. P., Zhu, S., 2013. Interpretation and computation of estimates from regression models using spatial filtering. Spatial Economic Analysis 8, 352–369.

Pace, R. K., Sirmans, C., Slawson, V. C., 2002. Automated valuation models. In: *Real Estate Valuation Theory*, Springer, pp. 133–156.

Padgham, M., Rudis, B., Lovelace, R., Salmon, M., 2017. osmdata. The Journal of Open Source Software 2.

Palmquist, R. B., 1992. Valuing localized externalities. Journal of Urban Economics 31, 59–68.

Palmquist, R. B., Israngkura, A., 1999. Valuing air quality with hedonic and discrete choice models. American journal of agricultural economics 81, 1128–1133.

Papadakis, M., Tsagris, M., Dimitriadis, M., Fafalios, S., Tsamardinos, I., Fasiolo, M., Borboudakis, G., Burkardt, J., Zou, C., Lakiotaki, K., Chatzipantsiou., C., 2019. Rfast: A Collection of Efficient and Extremely Fast R Functions. R package version 1.9.8.

Pearson, K., 1901. Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2, 559–572.

Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10, 439–446.

Plattner, R. H., Campbell, T. J., 1978. A study of the effect of water view on site value. Appraisal Journal 46, 20–25.

Polinsky, A. M., Shavell, S., 1976. Amenities and property values in a model of an urban area. Journal of Public Economics 5, 119–129.

Quigley, J. M., 1985. Consumer choice of dwelling, neighborhood and public services. Regional Science and Urban Economics 15, 41–63.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramsey, J. B., 1969. Tests for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society: Series B (Methodological) 31, 350–371.

Rasmussen, D. W., Zuehlke, T. W., 1990. On the choice of functional form for hedonic price functions. Applied Economics 22, 431–438.

Redfearn, C. L., 2009. How informative are average effects? hedonic regression and amenity capitalization in complex urban housing markets. Regional Science and Urban Economics 39, 297 – 306.

Ridker, R. G., Henning, J. A., 1967. The determinants of residential property values with special reference to air pollution. The Review of Economics and Statistics pp. 246–257.

Roback, J., 1982. Wages, rents, and the quality of life. Journal of political Economy 90, 1257–1278.

Rodriguez, M., Sirmans, C., 1994. Quantifying the value of a view in single-family housing markets. Appraisal Journal 62, 600–600.

Rosen, H. S., Fullerton, D. J., 1977. A note on local tax rates, public benefit levels, and property values. Journal of Political Economy 85, 433–440.

Rosen, S., 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy 82, 34–55.

Rosen, S., 1979. Wage-based indexes of urban quality of life. Current issues in urban economics pp. 74–104.

Rothenberg, J., 1991. The maze of urban housing markets : theory, evidence, and policy. Univ. of Chicago Press, Chicago, Ill., first ed.

Salvi, M., 2008. Spatial estimation of the impact of airport noise on residential housing prices. Swiss journal of economics and statistics 144, 577–606.

Schmenner, R. W., 1981. The rent gradient for manufacturing. Journal of Urban Economics 9, 90–96.

Schulz, R., Werwatz, A., 2004. A state space model for berlin house prices: Estimation and economic interpretation. The Journal of Real Estate Finance and Economics 28, 37–57.

Selim, H., 2009. Determinants of house prices in turkey: Hedonic regression versus artificial neural network. Expert Systems with Applications 36, 2843–2852.

Sheppard, S., 1999. Hedonic analysis of housing markets. Handbook of regional and urban economics 3, 1595–1635.

Shukur, G., Mantalos, P., 2004. Size and power of the reset test as applied to systems of equations: A bootstrap approach. Journal of Modern Applied Statistical Methods 3, 10.

Šidák, Z., 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62, 626–633.

Sirpal, R., 1994. Empirical modeling of the relative impacts of various sizes of shopping centers on the values of surrounding residential properties. Journal of Real Estate Research 9, 487–505.

Smith, V., Poulos, C., Kim, H., 2002. Treating open space as an urban amenity. Resource and Energy Economics 24, 107–129.

So, H. M., Tse, R. Y., Ganesan, S., 1997. Estimating the influence of transport on house prices: evidence from hong kong. Journal of Property Valuation and Investment 15, 40–47.

Söderberg, B., Janssen, C., 2001. Estimating distance gradients for apartment properties. Urban Studies 38, 61–79.

Stadelmann, D., 2010. Which factors capitalize into house prices? a bayesian averaging approach. Journal of Housing Economics 19, 180 – 204.

Strimas-Mackey, M., 2018. smoothr: Smooth and Tidy Spatial Features. R package version 0.1.1.

Su, L., Jin, S., 2010. Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. Journal of Econometrics 157, 18–33.

Tandon, S. B., 2012. Measuring Spillo Measuring Spillover Effects of Residential Amenity Impr ects of Residential Amenity Improvements Using Spatial Hedonic Approach. Ph.D. thesis, University of Nevada, Las Vegas.

Tennekes, M., 2018. tmap: Thematic maps in R. Journal of Statistical Software 84, 1–39.

Teugels, J. L., Vanroelen, G., 2004. Box-cox transformations and heavy-tailed distributions. Journal of Applied Probability 41, 213–227.

Teulings, C. N., Ossokina, I. V., de Groot, H. L., 2018. Land use, worker heterogeneity and welfare benefits of public goods. Journal of Urban Economics 103, 67–82.

Theebe, M. A., 2004. Planes, trains, and automobiles: the impact of traffic noise on house prices. The Journal of Real Estate Finance and Economics 28, 209–234.

Tobler, W. R., 1970. A computer movie simulating urban growth in the detroit region. Economic Geography 46, 234–240.

Tomkins, J., Topham, N., Twomey, J., Ward, R., 1998. Noise versus access: the impact of an airport in an urban property market. Urban studies 35, 243–258.

Trapletti, A., Hornik, K., 2019. tseries: Time Series Analysis and Computational Finance. R package version 0.10-47.

Tyrväinen, L., 1997. The amenity value of the urban forest: an application of the hedonic pricing method. Landscape and Urban planning 37, 211–222.

United States census Bureau, 2020. Quickfacts - new york city. https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork,bronxcountybronxboroughnewyork,kingscountybrooklynboroughnewyork , newyorkcountymanhattanboroughnewyork , queenscountyqueensboroughnewyork,richmondcountystatenislandboroughnewyork/PST045219.

Uyar, B., Brown, K. H., 2005. Impact of local public services and taxes on dwelling choice within a single taxing jurisdiction: A discrete choice model. Journal of Real Estate Research 27, 427–443.

Walsh, P., 2009. Hedonic property value modeling of water quality, lake proximity, and spatial dependence in central Florida. Ph.D. thesis, University of Central Florida.

Waugh, F. V., 1928. Quality factors influencing vegetable prices. Journal of farm economics 10, 185–196.

Wickham, H., 2017. tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1.

Wickham, H., 2019. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0.

Wilkinson, R., 1973. House prices and the measurement of externalities. The Economic Journal 83, 72–86.

Wong, K. C., So, A. T. P., Hung, Y. C., 2002. Neural network vs. hedonic price model: Appraisal of high-density condominiums. In: *Real estate valuation theory*, Springer, pp. 181–198.

Wood, S. N., 2003. Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65, 95–114.

Wood, S. N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99, 673–686.

Wood, S. N., 2006. On confidence intervals for generalized additive models based on penalized regression splines. Australian & New Zealand Journal of Statistics 48, 445–464.

Wood, S. N., 2008. Fast stable direct fitting and smoothness selection for generalized additive models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70, 495–518.

Wood, S. N., 2017. Generalized additive models: an introduction with R. Chapman and Hall/CRC.

Wood, S. N., Augustin, N. H., 2002. Gams with integrated model selection using penalized regression splines and applications to environmental modelling. Ecological Modelling 157, 157 – 177.

Wood, S. N., Goude, Y., Shaw, S., 2015. Generalized additive models for large data sets. Journal of the Royal Statistical Society: Series C (Applied Statistics) 64, 139–155.

Wu, C., Ye, X., Du, Q., Luo, P., 2017. Spatial effects of accessibility to parks on housing prices in shenzhen, china. Habitat International 63, 45–54.

Xiao, Y., Chen, X., Li, Q., Yu, X., Chen, J., Guo, J., 2017. Exploring determinants of housing prices in beijing: an enhanced hedonic regression with open access poi data. ISPRS International Journal of Geo-Information 6, 358.

Xie, Y., 2019a. knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.25.

Xie, Y., 2019b. testit: A Simple Package for Testing R Packages. R package version 0.11.

Yinger, J., 2015. Hedonic markets and sorting equilibria: Bid-function envelopes for public services and neighborhood amenities. Journal of Urban Economics 86, 9–25.

Young, F. W., Takane, Y., de Leeuw, J., 1978. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. Psychometrika 43, 279–281.

Zabel, J. E., 2004. The demand for housing services. Journal of Housing Economics 13, 16–35.

Zabel, J. E., Kiel, K. A., 2000. Estimating the demand for air quality in four us cities. Land Economics pp. 174–194.

Zeileis, A., Hothorn, T., 2002. Diagnostic checking in regression relationships. R News 2, 7–10.

Zhu, H., 2019. kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.1.0.

Zillow, 2020. New york home prices & values. https://www.zillow.com/new-york-ny/home-values/.