Dealing with Unobserved Heterogeneity in Hedonic Price Models[☆,☆☆]

Marc Francke[b,c], Alex van de Minne[a,*]

[a] *MIT Center for Real Estate, 105 Massachusetts Ave, Cambridge, MA 02139.*
[b] *Faculty of Economics and Business, University of Amsterdam, Plantage Muidergracht 12, 1001 NL Amsterdam.*
[c] *Ortec Finance, Naritaweg 51, 1043 BP Amsterdam.*

Abstract

This paper deals with unobserved heterogeneity in hedonic price models, arising from missing property and locational characteristics. In specific, commercial real estate is very heterogeneous, and data on detailed property characteristics are often lacking. We show that adding mutually independent property random effects to a hedonic price model results in more precise out-of-sample price predictions, both for commercial multifamily housing in Los Angeles and owner-occupied single family housing in Heemstede, the Netherlands. The standard hedonic price model does not take advantage of the fact that some properties sell more than once. We subsequently show that adding spatial random effects leads to an additional increase in prediction accuracy. The increase is highest for properties without prior sales.

Keywords: Bayesian Inference , Besag model , Commercial real estate , INLA , Leave-on-out-cross validation , Travelings Salesperson Problem.
JEL-codes: R32 , C01.

[☆]MIT/CRE Price Dynamics Research Platform Working Paper
[☆☆]We would like to thank David Geltner for his valuable input. We also thank Real Capital Analytics Inc for providing us with the data that made this research possible.
[*]Corresponding author
Email addresses: m.k.francke@uva.nl (Marc Francke), avdminne@mit.edu (Alex van de Minne)

1. Introduction

This paper deals with unobserved heterogeneity in hedonic price models. Hedonic price models are widely used, for example to create price indexes (and concomitant deprecation) for cars (Berndt et al., 1995), computers (Reis and Santos Silva, 2006), and residential housing (Hill, 2012), among many other types of goods. The number of applications within real estate is large. The hedonic price model has for example been used to value residential housing (Francke and De Vos, 2000; Sirmans et al., 2006), commercial real estate (Bokhari and Geltner, 2011) and (residential) land (Diewert et al., 2015), and to estimate the depreciation rate of houses (Knight and Sirmans, 1996; Francke and van de Minne, 2017b).

Rosen (1974) explicated the formal microeconomic theory underlying the hedonic price models, though the technique has older roots in consumer and marketing empirical analytics practice (Court, 1939). It is based on the idea that heterogeneous goods can be described by their attributes (de Haan and Diewert, 2011). In other words, a good is a bundle of (performance) characteristics. In the case of real estate properties, the relevant bundle may contain attributes of both the building structure and the location site of the property. For example, attributes might include the size, age, and type of building, and the distance of the site from downtown or the airport or nearest subway station. There is no market for the characteristics as such, since they cannot be sold separately. In the market for property occupancy, demand and supply in the market for built space (the rental market) determine the characteristics' marginal contributions to the total value of the bundle. Statistical regression based techniques are typically used to estimate these marginal value contributions.

Hedonic price models for residential and in particular commercial real estate are in practice hard to develop. First, properties are very heterogeneous in nature, implying many value drivers. Second, the property turnover rate, and so the number of transactions, is relatively low. Third, the number of registered property characteristics is in most databases quite limited: many value drivers are missing. And when they are sufficiently available, there is the risk of misspecification and over-fitting.

This paper deals with the modeling of unobserved heterogeneity, arising from missing property and locational characteristics. And we do so with a view toward data scarce application environments. The paper has broad relevance, but real estate is a particularly important subject. Real estate is

characterized by very long-lived goods which therefore often transact more than once, and also by the importance of spatial location. With this in mind, we deal with property related unobserved heterogeneity by adding mutual independent property level random effects to the standard hedonic price model, taking advantage of the fact that some properties transact more than once. Moreover, we add spatial random effects to deal with spatial dependencies. Spatial dependencies exist because nearby properties often have similar characteristics and also share locational amenities (Basu and Thibodeau, 1998).

The property random effects hedonic price model is related to the hybrid model (Quigley, 1995). The main difference is that we include random effects for all properties, and the hybrid model includes fixed effects for the repeat sales only.

We use two different specifications to model the spatial property effects. The first one is a Besag type model (Besag, 1974), and the second one a newly proposed spatial random walk model. Both models have in common that the spatial effect for each property depends on its neighbors.

The spatial random walk can be viewed as a special case of the Besag model, where neighbors are defined by the Travelings Salesperson Problem (TSP) route, the shortest route visiting every property only once, and returning to the starting point. The shortest route is calculated by algorithms solving the TSP. Using the TSP route to define neighbors, restricts each property to have at most two neighbors, the preceding and subsequent property on the TSP-route. We keep the model structure simple, and apply a random walk model on the ordered properties, even without taking into account the distance between the properties on the TSP route.

We compare the outcomes of 7 different hedonic price models: a standard hedonic price model, a hybrid model, a hedonic price model including property random effects (all three including location fixed effects), and two spatial models (Besag and spatial random walk) for the hybrid and property random effects hedonic price model (all four excluding location fixed effects). We perform a leave-one-out (LOO) cross validation to measure the out-of-sample performance for the various models, so we can check whether adding property random and spatial effects helps to increase prediction accuracy. As LOO analysis is computational expensive, we use an efficient Bayesian estimation procedure, Integrated Nested Laplace Approximation (INLA), see Rue et al. (2009).

We apply our model to multifamily housing (income generating proper-

ties) in Los Angeles and single family housing (owner-occupied) in Heemstede, a city close to Amsterdam in the Netherlands. Both data sets cover the period from 2001 up to 2017. In both Los Angeles and Heemstede, approximately 30% of the transactions are repeat sales.

The results are in line with expectations. Adding property random effects to the standard hedonic price model improves the prediction accuracy, more than in the hybrid model. The standard deviation of the LOO residuals reduce with approximately 5% in both markets. Adding random property effects and spatial effects reduces the standard deviation of the LOO residuals by 23% and 24% in Los Angeles and Heemstede, respectively. The differences in prediction accuracy between the Besag and spatial random walk model are small, so using a restricted version of the Besag model – having at most two neighbors, the preceding and subsequent property on the TSP-route – does not lead to a loss in prediction accuracy. However, the spatial random walk model is computationally much more efficient. In Los Angeles the spatial random walk model is the best performing one, in Heemstede the Besag. The estimated spatial effects are correlated among the models. Correlations range between 0.93 and 0.99 in Los Angeles, and between 0.88 and 0.99 in Heemstede.

When having only 1 sale per property, the property random effects hedonic price model including spatial effects performs better than the model excluding the spatial effects. The difference in performances becomes smaller when the number of sales per property increases; then the property random effects pick up most of the unobserved heterogeneity, and there is less additional gain from the spatial structure. Finally, the property random effects hedonic price model including spatial effects outperforms more standard hedonic models even after excluding important characteristics.

The paper proceeds as follows. Section 2 gives the methodology. Section 3 provides a data description. Section 4 gives the estimation results, and finally, Section 5 concludes.

2. Methodology and Estimation

2.1. Pooled Hedonic Price model

For modeling and tracking the prices of heterogeneous goods (including real estate), a widely-used type of hedonic price model is the so-called *pooled*

model, given by

$$y_i = x_i\beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2), \quad i = 1, \ldots, N, \tag{1}$$

where the dependent variable y_i is the log price for transaction i, and N is the number of transactions. The $(1 \times K)$ vector x_i represents the observable hedonic characteristics with corresponding coefficient vector β, constant over time. Apart from property characteristics and a constant, the row vector x_i could include location fixed effects. In the pooled hedonic price model x_i also includes time fixed effects. The error term ϵ_i is assumed to be normally and independently distributed with zero mean and variance σ_ϵ^2.

The pooled hedonic price model is typically estimated by ordinary least squares (OLS). The estimated coefficient $\hat{\beta}$ represents the marginal value contributions, and can subsequently be used to predict the value of all properties – including the ones that were not sold – as long as we observe the hedonic attributes x. The estimated coefficients of the time fixed effects reflect longitudinal changes in the market, and can be interpreted directly as a time trend in the central tendency of market values, hence, can be used to produce a price index.

In the present paper we take specific interest in how to cope with the unobserved heterogeneity. Unobserved heterogeneity is reflected in omitted variable bias, lower model fit, and lower performance on out-of-sample prediction. Unobserved heterogeneity is in specific a problem for commercial real estate (Francke and van de Minne, 2017a), as properties are very heterogeneous, transaction prices are scarce, and detailed property characteristics are often lacking.

2.2. Property Random Effects

A way of dealing with unobserved heterogeneity in the hedonic price model is to include mutually independent property random effects in Eq. (1), taking advantage of the fact that some properties transact more than once, leading to the following model

$$y_p = x_p\beta + j_{n_p}\phi_p + \epsilon_p, \; \epsilon_p \sim N(0, \sigma_\epsilon^2 I_{n_p}), \tag{2}$$

$$\phi \sim N(0, \sigma_\phi^2 I_P), \tag{3}$$

where subscript $p = 1, \ldots, P$ indicates an individual property, P is the total number of properties, j is a vector of ones, and I is the identity matrix.

The number of transactions for property p is n_p; for single sales $n_p = 1$, and for repeat sales $n_p > 1$. The total number of transactions is given by $N = \sum_{p=1}^{P} n_p$. Note that y_p is now a $(n_p \times 1)$ vector of log transaction prices. The property random effects ϕ_p reflect omitted variables and model misspecification, and the error term ϵ transaction noise, the difference between the market value and the transaction price.

Conditional on the variance parameters σ_ϵ^2 and σ_ϕ^2 the hedonic price model including random property effects Eqs. (2)–(3) can be estimated by generalized least squares, giving

$$\beta | y, X, \sigma_\epsilon^2, \sigma_\phi^2 \sim N(\hat{\beta}, \mathrm{Var}(\hat{\beta})), \tag{4}$$

where $\mathrm{Var}(\hat{\beta}) = (\sum_{p=1}^{P} (x_p' \Omega_p^{-1} x_p))^{-1}$, $\hat{\beta} = \mathrm{Var}(\hat{\beta}) \sum_{p=1}^{P} (x_p' \Omega_p^{-1} y_p)$, and $\Omega_p = \sigma_\epsilon^2 I_{n_p} + \sigma_\phi^2 j_{n_p} j_{n_p}'$. Conditional on σ_ϵ^2 and σ_ϕ^2 estimates of the property random effects are given by

$$\hat{\phi}_p = \frac{n_p \sigma_\phi^2}{\sigma_\epsilon^2 + n_p \sigma_\phi^2} (\bar{y}_p - \bar{x}_p \hat{\beta}), \tag{5}$$

where $\bar{y}_p$ and $\bar{x}_p$ are the average of the transactions prices and the characteristics of property p. The part of the average residual $\bar{y}_p - \bar{x}_p \hat{\beta}$ that is attributed to the property random effect thus depends on the ratio of the variance parameters and the number of transactions for property p: The larger n_p and the smaller the ratio of $\sigma_\epsilon^2 / \sigma_\phi^2$ is, the larger this part is. The predicted values for property p – conditional on the ratio of the variance parameters σ_ϵ^2 and σ_ϕ^2 – can subsequently be expressed as $\hat{y}_p = x_p \hat{\beta} + j_{n_p} \hat{\phi}_p$.

The variance parameters σ_ϵ^2 and σ_ϕ^2 can be estimated by (restricted) maximum likelihood or by Bayesian methods. In absence of prior information on these variance parameters a necessary condition is that the number of transactions N must be larger than the number of properties P, so $N > P$. In other words, some – not all – properties need to transact more than once over the sample period.

Please note that it is practically infeasible to replace the random effects by fixed effects: By including property fixed effects one effectively excludes all single sales, which are in many applications the majority of the transactions. One could formally test whether the β coefficients are different in the fixed and random effects model by the Hausman test. An important reason why the two estimators could be different is the existence of correlation between X and ϕ, although other sorts of misspecification may also lead to rejection

of the null hypothesis of no difference in the β estimates. In our applications it is evident that the property fixed effects models are misspecified – the number of repeat sales is only a very small fraction of all sales, and for the repeat sales most of the characteristics do not change between the date of buying and selling – and therefore we will not apply the Hausman test.

In this paper we will focus on out-of-sample cross validation to test for potential over-fitting, see Section 2.4 for more details. In a hedonic price model with property fixed effects out-of-sample prediction is not possible, unless another sale of the same property has been included in the estimation. This drawback does not hold for random effects models, although the random effect will be zero when the property has not been included in the model estimation.

The hedonic price model with property random effects is related to the hybrid model as proposed by Case and Quigley (1991), although the focus in hybrid models is primarily on price indexes, see also Quigley (1995) and Hwang and Quigley (2004). They split the sample in two parts representing single and repeat sales, and provide different model specifications for both subsamples. The single sales y^S and the repeat sales y^R are modeled by

$$y_p^S = x_p^S \beta + \epsilon_p^S, \quad \epsilon_p^S \sim N(0, \sigma_\epsilon^2), \quad \forall p : n_p = 1, \tag{6}$$

$$y_{p,t}^R - y_{p,s}^R = (x_{p,t}^R - x_{p,s}^R)\beta + \epsilon_{p,t}^R - \epsilon_{p,s}^R, \quad \forall p : n_p > 1, \tag{7}$$

where $y_{p,t}^R - y_{p,s}^R$ is the difference in the log price at the time of selling t and the time of buying s. By leaving out all components of $(x_{p,t}^R - x_{p,s}^R)\beta$ except for the time fixed effects, one gets the repeat sales model (Bailey et al., 1963; Case and Shiller, 1987), which is widely used to estimate property price indexes. Eqs. (6)–(7) are simultaneously estimated by OLS. The repeat sales model (7) can equivalently be written in levels by adding property fixed effects, giving

$$y_p^R = x_p^R \beta + j_{n_p} \phi_p^{\mathrm{FE}} + \epsilon_p^R, \tag{8}$$

where ϕ_p^{FE} is the fixed effect for property p.

The combined model for single and repeat sales can subsequently be expressed as

$$\begin{pmatrix} y^S \\ y^R \end{pmatrix} = \begin{pmatrix} X^S & 0 \\ X^R & D^{\mathrm{FE}} \end{pmatrix} \begin{pmatrix} \beta \\ \phi^{\mathrm{FE}} \end{pmatrix} + \begin{pmatrix} \epsilon^S \\ \epsilon^R \end{pmatrix}, \tag{9}$$

where D^{FE} is the selection matrix to select the appropriate property. The hybrid model is somewhat inconsistent by specifying property fixed effects only for the repeat sales. Note that in hybrid models – in contrast to property fixed effects hedonic price models – out-of-sample prediction is also possible for single sales.

2.3. Spatial dependencies

The property random effects so far have been specified as mutually independent, $\text{Cov}(\phi_p, \phi_q) = 0$ for $p \neq q$, so spatial dependencies have not been explicitly taken into account. Spatial dependencies exist because nearby properties often have similar building characteristics and also share locational characteristics/amenities (Basu and Thibodeau, 1998). We add spatial property effects θ to the property random effects hedonic price model, leading to

$$y_p = x_p\beta + j_{n_p}\phi_p + j_{n_p}\theta_p + \epsilon_p,\ \epsilon_p \sim N(0, \sigma^2_\epsilon I_{n_p}). \tag{10}$$

The spatial property effect requires having latitude and longitude coordinates for all properties, which in most cases are easy to obtain.

We use two different specifications for the spatial property effects θ. The first one is a Besag type model (Besag, 1974), and the second one a newly proposed spatial random walk model, which can be seen as a special case of the Besag model. Both models have in common that the spatial property effect for property p depends on its neighbors, although the spatial dependence structure is different. The next subsections provide more details on both models.

We are interested in the estimates of β, ϕ, and θ, and predictions of property log prices including the property random and spatial effects, $\hat{y}_p = x_p\hat{\beta} + j_{n_p}\hat{\phi}_p + j_{n_p}\hat{\theta}_p$. For this reason we restrict ourselves to a specific class of spatial random effects models described in Section 2.3.1, and do not for example consider the widely used spatial (spatio-temporal) autoregressive models (Pace et al., 1998, 2002). Spatial-temporal autoregressive models have been used in previous literature, but most applications have been on residential properties. Some commercial real estate examples are Tu et al. (2004), Nappi-Choulet and Maury (2009), and Chegut et al. (2015), all focusing on price indexes. For an extensive overview of spatial hedonic price models, see Anselin and Lozano-Gracia (2009).

2.3.1. Besag model

Intrinsic and conditional autoregressions were introduced by Besag (1974), and later extended by Besag et al. (1991) and Besag and Kooperberg (1995). These models are examples of Gaussian Markov random fields (Lindgren et al., 2011), which are specified through the set of conditional distributions of one component (θ_p) given all the others (θ_{-p}).

Let $w_{p,q}$ denote a symmetric proximity measure for properties p and q. It is nonnegative when $p \neq q$, and 0 otherwise. In our application we use $w_{p,q} = 1$ if the distance between the properties is smaller than a predefined threshold, and 0 otherwise.[1] Let ∂_p denote all m_p neighbors of property p; all properties q for which it holds that $w_{p,q} \neq 0$. The conditional distribution of θ_p is given by

$$\theta_p|\theta_{-p}, \sigma_\theta^2 \sim N\left(\frac{\sum_{q\in\partial_p} w_{p,q}\theta_q}{m_p}, \frac{\sigma_\theta^2}{m_p}\right), \tag{11}$$

where θ_{-p} is the vector of spatial property effects excluding property p. From the right-hand-side of Eq. (11) is it clear that the spatial effect for property p is directly inferred from its neighbors only. In case $w_{p,q} = 1$ for neighboring properties, the conditional mean is simply the mean of the spatial effects of neighboring properties, and the conditional variance inversely related to the number of neighboring properties.

Note that the unconditional joint distribution of θ is not proper, the rank of the precision matrix is only positive semidefinite. A proper specification is obtained by adding a positive parameter d to the denominator, giving

$$\theta_p|\theta_{-p}, \sigma_\theta^2, d \sim N\left(\frac{\sum_{q\in\partial_p} w_{p,q}\theta_q}{d + m_p}, \frac{\sigma_\theta^2}{d + m_p}\right). \tag{12}$$

This model is sometimes referred to as a proper Besag model (Blangiardo and Cameletti, 2015). The parameter d will be estimated from the data.

[1] We use a maximum distance of 770m and 35m for Los Angeles and Heemstede respectively, resulting in at least 1 neighboring property for each property.

2.3.2. Spatial random walk

In this section we present a new two-step method to model spatial property effects, closely related to the Besag model. In the first step we calculate the shortest route visiting every property only once, and returning to the starting point. The shortest route is calculated by using algorithms solving the Travelings Salesperson Problem (TSP). This gives an ordering of the properties and distances between the ordered properties. The TSP is a well known and important combinatorial optimization problem (Lawler et al., 1985; Gutin et al., 2002). There are multiple TSP-algorithms to be found in literature. We use 8 different versions: (1) Nearest neighbor algorithm, (2) Insertion algorithm, (3) Nearest insertion, (4) Farthest insertion, (5) Cheapest insertion, (6) Arbitrary insertion, (7) k-Opt heuristics, and (8) the Lin-Kernighan heuristic. For more information on these different TSP-algorithms, please see Lawler et al. (1985) and Hahsler and Hornik (2007). Subsequently, we pick the version which renders the shortest route. Except for showing the shortest route, we do not give any statistics on this first step. Note that most software packages will pick a random starting point. This shouldn't affect the results too much, as the *route* remains similar, irrespective of the starting point.

In our application, see Section 3, we 'only' have approximately 2,000 observations for each of our two markets. Computing all of the above-mentioned TSP algorithms is therefore not a computational issue. However, with large data-sets (housing data can have 100,000s of transactions for example), the time required to solve some TSP algorithms might become infeasible. In our study we found that both the nearest neighbor and arbitrary insertion algorithms finished within a second. The other algorithms would take between 30 seconds to 1 minute to solve for the shortest route. Bentley (1992) gives the computational time for different TSP algorithms for large data problems. Even in this relatively old paper, Bentley (1992) solved the nearest neighbor algorithm for 1M observations within 10 minutes.

In the second step we use a structural time series specification to specify the value profile over the TSP route. Structural time series models have been widely and successfully applied in the last few decades (Harvey, 1989), but not so much as a spatial application. In this application we keep the model structure simple, and use a random walk specification, even without taking into account the distance between the ordered properties on the TSP route. More complex structural time series models, like local linear trend

and autoregressive representations (Francke et al., 2017), taking into account distances between properties, could also be applied, possibly improving model fit, but we leave this for future research.

The spatial random walk specification is given by

$$\theta_{(p)} \sim N(\theta_{(p-1)}, \sigma_\theta^2), \tag{13}$$

where subscripts (p) denote properties ordered by the TSP route. For identification purposes we will impose the restriction that the sum of the value profiles over all properties is zero, $\sum_{p=1}^{P} \theta_p = 0$.

Note that the spatial random walk is a special case of the Besag model in Eq. (11). In the spatial random walk the neighbors of property p – denoted by ∂_p in Eq. (11) – are defined by the TSP route. The TSP route restricts all properties – except the first and the last – to have at most two neighbors, the preceding and subsequent property on the TSP-route.

The advantage of this specification over the Besag model is its relative ease of estimation, especially in a large data environment. Besag models need a large $P \times P$ (sparse) matrix of zeros and ones identifying neighbors. Given that it is not uncommon to have large P, especially with housing data, this can result in computational issues. The spatial random walk only needs a vector $(1 \times P)$ indicating the ordering of the TSP route, thus reducing the size issue considerably. However, even with sparse data (which is the case in the current paper), we found that the estimation time itself is reduced considerably as well, especially when using Markov chained Monte Carlo (MCMC) algorithms. In fact, we found that estimating the spatial random walk instead of the Besag model using our relative small data set (see Section 3) with MCMC procedures decreased computational time 20-fold.[2] The estimation time difference is reduced considerably when using Laplace approximation. However, we still found a 25% computational time decrease after using the spatial random walk. Estimation time differences are larger if the data becomes larger. Also note that the spatial random walk can be estimated using the Kalman filter, reducing computational time even more.

An obvious disadvantage of the two-step approach is that we reduce a

[2]In an earlier version we ran our models using the the No-U-Turn-Sampler (Hoffman and Gelman, 2014). Even after very efficient re-parametrization of the models, the Besag models would take over 24h to estimate, compared to less than an hour for the spatial random walk.

two dimensional plane into an one dimensional line, at the risk of ignoring important information. For that reason we will perform a leave-one-out cross validation for the spatial random walk model and other (spatial) models.

2.4. Leave-one-out cross validation and estimation

We do a full Leave-One-Out analysis (LOO) to compare out-of-sample model performance. More specifically, we leave one observation (i) out of the data, and predict the value for this observation y_i, the posterior mean $E[y_i|y_{-i}]$, based on the remaining $N-1$ observations y_{-i} for different models as defined in Section 2. We redo this analysis for every observation, so N times. By simply subtracting our predicted value from the actual log transaction price, we get the LOO residual, which is essentially an out-of-sample prediction error. Subsequently, we use the LOO residuals to calculate out-of-sample performance statistics, such as the mean, the absolute mean, and the standard deviation.

As the LOO analysis is computationally expensive we use an efficient estimation procedure, the Integrated Nested Laplace Approximation (INLA, Rue et al., 2009). In essence, INLA computes an approximation to the posterior marginal distribution of the hyper-parameters. Operationally, INLA proceeds by first exploring the marginal joint posterior for the hyper-parameters in order to locate the mode, a grid search is then performed and produces a set of "relevant" points together with a corresponding set of weights, to give the approximation of the distributions. Each marginal posterior can be obtained using interpolation based on the computed values and correcting for (probably) skewness, by using log-splines. For each hyper-parameter, the conditional posteriors are then evaluated on a grid of selected values for the prior and the marginal posteriors are obtained by numerical integration. In this paper we have a flat prior for all hyper-parameters.

3. Data and Descriptive Statistics

We use two different data sources, commercial multifamily real estate (income generating properties) in the city of Los Angeles and single family housing (owner-occupied) in Heemstede, a city relatively close to Amsterdam in the Netherlands.

The first database is provided by Real Capital Analytics (RCA), and captures approximately ninety percent of all commercial property transactions

in the US over $2.5 million. The database contains 2,263 pre-filtered transactions, of which 1,936 are unique properties, in the period 2001 to 2017. The annual number of transactions is given in Figure A.1a, about 140 transactions on average. The highest transaction volume was realized before the crisis. Even in the most recent periods, the amount of transactions never reached pre-crisis levels. The relatively low number in 2017 is due to the fact that we do not observe all transactions in 2017.

We observe the Net Operating Income (NOI), property subtype (garden versus mid/highrise), the age and size of the structure (in square feet), latitude and longitude, and the transaction price. The upper panel of Table B.1 provides some descriptive statistics. The average transaction price is about $6.4 million, the average size is about 32,000 square feet, and the average age is 45 years. Most properties are designated garden.

[Place Figure A.1 about here]

The second database is provided by the Dutch Association of Real Estate Brokers and Real Estate Experts (NVM), the largest brokers organization in the Netherlands. About 70% of all real estate brokers in the Netherlands is affiliated to the NVM. The database contains 2,262 transactions, of which 2,065 are unique properties, in the period 2001 to 2017. The majority of the transactions is single sales (69%). The annual number of transactions is given in Figure A.1b, about 145 transactions on average. The transaction volume dropped by 50% during the crisis.

We observe the property subtype (row houses, corner house, 2 types of semi detached homes and detached), the age and size of the structure, the maintenance level (3 groups from bad to good), the presence of a yard, latitude and longitude, and the transaction price. The lower panel of Table B.1 provides some descriptive statistics. The average transaction price is about €485 thousand, the average size 151 square meters (1,625 square feet), and the average age is 65 years. The largest number of properties are row houses (44%). The NVM distinguishes between two types of semi detached homes; (1) two properties are connected via a garage, and (2) two properties are connected wall-to-wall. Most of the semi detached properties fall in the second category, 24% of the observations. More than half of the properties have an average maintenance level at the time of listing, compared to 18% badly

maintained and 23% well maintained.[3] Almost all properties have a yard, in only 6% of the transactions this is not the case.

[Place Table B.1 about here]

4. Results

In this section we provide estimation results for 7 different model specifications. All models have the log of the transaction price as dependent variable, and the log of the size as one of the independent variables. In addition, we use the log of the NOI per square foot as independent variable in the Los Angeles model. Property age is entered in a quadratic way. All other variables have been entered as dummy variables, including the annual time fixed effects. The 7 model specifications are

1. **Standard**: The standard hedonic price model including location fixed effects, given by Eq. (1).
2. **Hybrid**: The hybrid model including location fixed effects, given by Eq. (9).
3. **RE**: The property random effects hedonic price model including location fixed effects, given by Eqs. (2) and (3).
4. **Besag(Hybrid)**: The hybrid model, Eq. (9), where spatial property effects have been added by the Besag model, Eq. (12).
5. **Besag(RE)**: The property random effects hedonic price model, where spatial property effects have been added by the Besag model, given by Eqs. (3), (10), and (12).
6. **SRW(Hybrid)**: The hybrid model, Eq. (9), where spatial property effects have been added by the spatial random walk model, Eq. (13).
7. **SRW(RE)**: The property random effects hedonic price model, where spatial property effects have been added by the spatial random walk model, given by Eqs. (3), (10), (13).

[3] See Francke and van de Minne (2017b) for a discussion on how the maintenance data in the NVM data is compiled.

Note that only the Standard, Hybrid, and RE model include location fixed effects.[4] For Los Angeles we have 6 locations, defined by RCA: East LA/Long Beach, Hollywood/Santa Monica, Los Angeles - CBD, North LA County, Valley/Tri-Cities and West Covina/Diamond Bar. For Heemstede we have 4 locations, defined by the first 4 digits of the ZIP codes.

The remainder of this section is organized as follows. Subsection 4.1 discusses estimation results. Subsection 4.2 provides summary statistics for the leave-one-out cross validation. Subsection 4.3 discusses spatial effects, and finally subsection 4.4 gives some robustness checks.

4.1. Estimation results

Tables B.2 – B.3 provides the posterior means of the coefficients and significance levels for Los Angeles and Heemstede respectively.[5] The estimates of the time dummies can be interpreted as a log price index for Heemstede. In Los Angeles the interpretation is less straightforward, given that we also include the Net Operating Income in the model, which picks up a large part of the time variation (or the macro-economic cycle). We first discuss the results for Los Angeles, and then the results for Heemstede.

Los Angeles

The estimated elasticity for NOI per square foot on prices is about 0.7 on average over all models. The coefficient for size is slightly less than 1, indicating that prices increase less than proportional to property size. If the property doubles in size, the price increases with 95% on average. Most real estate studies find this law of diminishing returns (Bokhari and Geltner, 2016). The coefficient for Mid/Highrise properties in Los Angeles is positive but insignificant for the Standard, Hybrid, and RE model. For the Besag and SRW model the coefficient becomes statistically significant and negative, which might indicate an interaction between property type and location, which the location dummies in the standard, hybrid, and RE model do not pick-up. Also, *ceteris paribus*, one would expect that lowrise housing would be more popular compared to highrise housing. Age has a negative

[4] Bourassa et al. (2007) advocate to use submarket fixed effects, defined by real estate agents.

[5] The highest posterior density intervals are not shown for the sake of brevity. They are available on request.

coefficient and the square of age a positive coefficient. This confirms expectations, as depreciation is fastest when a property is young, see Bokhari and Geltner (2016). Here the estimated coefficient is also lower compared to other studies because of the inclusion of NOI (and age squared is not always significantly different from zero). It is well known that depreciation results in lower NOI, and not so much in higher cap rates (Bokhari and Geltner, 2016; Geltner and van de Minne, 2017). As such, most depreciation is 'captured' by the NOI variable.

The year 2001 is the omitted time dummy variable, and is therefore the reference group. The point estimates of the year dummies for the standard, hybrid and random effects model, are always smaller than that of the other models. (Especially compared to the spatial random walk models.) This is explained by the differences in the estimate for NOI per square foot, which is considerably different for the different models. Given that NOI also 'captures' changes in the macro-economic environment, this was expected. In other words, models with a high estimate for NOI per square foot (like the standard model) will result in less variation in the time dummy estimates and vice verse. Note that the crisis and subsequent recovery are still clearly visible in all models. However, the timing is slightly different. The trough of the time dummies is 2010 for the standard, hybrid and random effects and hybrid Besag model. The trough is a full year earlier for the other models. [6]

The residual standard error is highest in the Standard model, $\sigma_\epsilon = 0.19$. In other words, the model-fit is quite low. The standard error reduces to 0.18 in the Hybrid model, and is around 0.13 for the other models. The standard error of the property random effects σ_ϕ is 0.15 in the RE and Besag model, and it reduces to 0.04 in the SRW model. The standard errors of the spatial property effects σ_θ in the Besag and SRW are difficult to compare, because the underlying models are different. Note that in the spatial random walk model σ_θ is much lower in the RE model than in the Hybrid model. The Moran I statistic suggests that there is some spatial autocorrelation left in the residuals only for the non-spatial models: The Standard, Hybrid and RE model. The WAIC[7] of the Hybrid model is actually higher compared to

[6] Although it should be noted that the the different index levels do not differ from each other significantly between 2009 and 2010. This is not shown here, but is available upon request.

[7] The Watanabe-Akaike or widely applicable information criterion (WAIC, Watanabe, 2010) is based on the series expansion of leave-one-out cross-validation. WAIC can be

the Standard model, meaning worse model fit. The WAIC for the RE does improve considerably over the standard model with 820 points. The models including both property random and spatial effects have the lowest WAIC. The best performing model is SRW(RE).

[Place Table B.2 about here]

Heemstede
Compared to row houses, detached house are valued the highest, followed by semi detached and corner houses. Compared to poorly maintained houses, average and good maintained houses sell at a premium of 14% and 23%, respectively. The estimated premium for a yard sits at 2% on average, however is statistically insignificant different from zero for most of our models. The coefficient for size varies between 0.69 and 0.90, depending on the model specification, indicating that prices increase less than proportional to property size. Age has a positive coefficient and the square of age a negative coefficient. In other words, older houses have higher values. An eighty years old house – built in the thirties – has a 16% premium compared to a new house. This has most likely to do with vintage effects, see for example (Coulson and McMillen, 2008; Wilhelmsson, 2008; Francke and van de Minne, 2017b), combined with the fact that we hold constant for physical deterioration by controlling for maintenance. Interestingly, the effect of age on house prices is statistically insignificant for the spatial random walk models (however, the age squared term is significant).

Since we have no variables that move with the economic cycle in Heemstede, the coefficients of time dummy variables can be interpreted as a log price index. Between 2001 and 2008 prices increased by 32% – 34%. Subsequently, prices dropped between 2008 and 2013 by 13% – 16%. Note that the crisis took relatively long in the Netherlands. From 2013 to 2017 prices increased by 33% – 36%. The difference between the models is negligible.

The Moran I statistic suggests that there is some spatial autocorrelation left in the residuals only for the Standard model. The WAIC of the Hybrid model is almost similar to the Standard model. The RE model performs better, the WAIC of the RE model is 1,261 points lower compared to the Hybrid model. The models including both property random and spatial

viewed as an improvement of the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002). Lower values indicate a better model-fit.

effects have the lowest WAIC. The WAIC of the best performing model, Besag(RE), is 643 points lower compared to the RE model.

[Place Table B.3 about here]

4.2. Leave-one-out cross validation

Table B.4 provides the results of the leave-one-out cross validation, the upper part for Los Angeles, and the lower part for Heemstede. In general, the out-of-sample model fit is slightly better for single family in Heemstede compared to multifamily housing in Los Angeles.

In Los Angeles the standard deviation of the LOO residuals is similar for the Standard and Hybrid models, about 0.190. In the RE model the standard deviation is 0.184, which is still not a big improvement over the Standard model. The main reason for this small reduction is the relative small portion of repeat sales. Adding spatial structures reduces the standard deviation considerably though, to 0.146 in the best performing model SRW(RE), a reduction of 24% (23%) compared to the Standard (Hybrid) model. The spatial models including property random effects perform better than the hybrid spatial models, although the differences seem small.

In Heemstede the standard deviation of the LOO residuals is similar for the Standard and Hybrid model, both 0.172. In the RE model the standard deviation is 0.164, a small reduction of 4.7% compared to the Standard model. Adding spatial structures reduces the standard deviation even more, to 0.130 in the best performing model Besag(RE), a reduction of 24% compared to the Standard and Hybrid model. The spatial models including property random effects perform better than the hybrid spatial models.

The best performing models measured by the standard deviation of the LOO residuals coincide with the best ones measured by the WAIC criterion. Unlike LOO statistics one cannot compare WAICs over different data-sets.

[Place Table B.4 about here]

Table B.5 provides the absolute mean LOO residuals as a function of the number of sales per property (the first column). The final column gives the corresponding number of properties. Note that when the number of sales per property is n, in the leave-on-out analysis $n-1$ sales of the property have been used to estimate the model.

In the Standard model the-out-of-sample model fit increases when the number of sales per property increases, although the gain is relatively small. In Los Angeles it goes down from 0.148, when having only 1 sale, to 0.139, when having 3 sales of the property (-6.1%), and in Heemstede from 0.139, when having 1 sale, to 0.120, when having 3 sales of the property (-13.8%). Note that the reduction is much higher for the RE models. In Los Angeles it goes down from 0.149, when having only 1 sale, to 0.112, when having 3 sales of the property (-24.8%), and in Heemstede from 0.139, when having 1 sale, to 0.091, when having 3 sales of the property (-35.0%). The property random effects hedonic price model clearly takes advantage of the fact that some properties transact more than once. Note that the Hybrid model performs less than the RE model in particular when the number of sales per property is 2. In fact, the Hybrid model performs equal to the standard model with just 2 sales. This was expected, given that the Hybrid model can only get property level estimates if the property was sold three times or more (because we lose one observation in the leave-on-out analysis).

When having only 1 sale per property (or zero during the leave-on-out analysis), the property random effects hedonic price model including spatial effects performs better than the model excluding the spatial effects. The difference in performances becomes smaller when the number of sales per property increases, then the property random effects pick up most of the unobserved heterogeneity, and there is almost no additional gain from the spatial structure.

[Place Table B.5 about here]

4.3. Spatial effects

Figure A.2 gives the TSP routes for both Los Angeles and Heemstede. Figure A.3 provides the spatial effects θ along this route for the Besag and SRW models, and Figures A.4 and A.5 give heat maps for Los Angeles and Heemstede for the same models.

The heat maps of Los Angeles, Figure A.4, give a clear picture. The highest values of the spatial random effects are in the CBD area and Hollywood/Santa Monica, and generally speaking along the coast. Lower values are found in the North and the East of Los Angeles. Note that this after holding the model constant for NOI, which should also vary over space. The heat maps of Heemstede, Figure A.5 , give a less clear picture. This corresponds to the erratic pattern of the spatial effects over the TSP route, see

lower panel of Figure A.3. Although clearly the north-east (south) of the map is dominated by high (low) values of the spatial effect θ. It should be stressed though, that these heat maps represent the value of the spatial effect θ, and not the total property values per se (or square meter values).

Table B.6 gives some descriptive statistics on the spatial effects θ. In Los Angeles the difference between the 2.5% and 97.5% percentile of θ is about 0.644, corresponding to a 90% difference between the cheapest and most expensive location, after correction for differences in property characteristics and NOI. In Heemstede the difference between the 2.5% and 97.5% percentile of θ is similar with 0.615, corresponding to a 85% difference between the cheapest and most expensive location. The estimated spatial effects θ are positively correlated among the models. Correlations range between 0.93 and 0.99 in Los Angeles, and between 0.88 and 0.99 in Heemstede.

[Place Figure A.2 about here]

[Place Figure A.3 about here]

[Place Figure A.4 about here]

[Place Figure A.5 about here]

[Place Table B.6 about here]

4.4. Robustness check

In this Section we perform a simple robustness check. In Heemstede we omit the level of maintenance and the property type dummy variables as explanatory variables and re-run both the standard hedonic price model and the spatial random walk with property random effects (SRW (RE)) model. Our basic interest is to compare the SRW (RE) model on the reduced dataset with the standard hedonic price model with all variables included, see previous Sections. This can learn us something on how effectively the spatial and property random effects deal with omitted variables / unobserved heterogeneity.

We do something similar for the commercial properties. It is well known that Net-Operating-Income (NOI) or rents explain a large part of prices, where higher rents result in higher prices (Kok et al., 2017). For example, Geltner and van de Minne (2017) show that the (cross-sectional) variation in NOI is much higher compared to capitalization rates, using the same RCA data. For Los Angeles we therefore omit NOI per square foot from the regression and re-run the models and the leave-on-out analysis. A summary of the robustness checks is given in Tables B.7 – B.8.

[Place Table B.7 about here]

[Place Table B.8 about here]

Overall, the results are in line with expectations. The SRW (RE) model outperforms the standard hedonic price model to a large extent on the same set of characteristics, and in case of the SRW (RE) model, the fit is better for properties that transacted more often. Also unsurprisingly is that omitting maintenance/property types and NOI in the price model in respectively Heemstede and Los Angeles, deteriorates the model fit considerably. The standard deviation of the LOO residual increases with almost 20% in Heemstede and even 50% in Los Angeles after omitting our selection of characteristics. (For both the Standard and the SRW (RE) model.)

However, as noted earlier, our main interest is in comparing the fit of the SRW (RE) model on the reduced data-set with the standard hedonic price model using all variables. In Heemstede the SRW (RE) model on the reduced data-set clearly outperforms the standard model on the full data-set. Both 'traditional' metrics in Table B.7 as the LOO residuals in Table B.8 are better for the first over the latter. The average absolute LOO residual is 0.122 for the SRW (RE) on the reduced data, compared to 0.135 for the standard model on the full data. For properties that sold more than once, the relative gain is even bigger.

In Los Angeles the standard model including NOI as explanatory variable actually performs better than the SRW (RE) model excluding NOI on some metrics, but not on others. For example, the 'noise' (σ_ϵ in Table B.7) is considerably lower for the SRW (RE) model excluding NOI compared to the standard model including NOI and the WAIC also improves. However, the DIC is 'better' for the standard model over the SRW (RE) model. The

average absolute LOO residuals in Table B.8 also give an inconsistent picture. For properties that sold only once the standard model including NOI as explanatory variable outperforms the SRW (RE) model excluding NOI. More specifically, the average absolute LOO residual is 0.148 (0.172) for the standard model including NOI (SRW (RE) excluding NOI). However, for properties that sold multiple times, the SRW (RE) model results in a better model-fit. Given that we do not have that many repeat sales in Los Angeles, the mean absolute LOO residuals are lower for the standard model including NOI data overall. Still, given how much of the variance of property prices is explained by NOI, it is impressive how well the spatial random effects model excluding NOI data performs.

5. Conclusion

This paper deals with unobserved heterogeneity in hedonic price models, arising from missing property and locational characteristics. In specific commercial real estate is very heterogeneous, and detailed property characteristics are often missing.

We show that adding mutually independent property random effects to a hedonic price model results in more precise out-of-sample price predictions, both for commercial multifamily housing in Los Angeles and owner-occupied single family housing in Heemstede. The larger the share of repeat sales, the higher the increase in prediction accuracy is. Put differently, having more (previous) sales, reduces the prediction error for a property when property random effects are included in the hedonic price model. The standard hedonic price model does not take advantage of the fact that some properties sell more than once, and so the prediction accuracy only marginally improves when having previous sales. The hedonic price model including property random effects also outperforms the related hybrid model, including property fixed effects for repeat sales only.

We subsequently show that adding spatial effects leads to an additional increase in prediction accuracy. The increase in prediction accuracy is highest for properties without prior sales. When for a property a prior sale is available, the unobserved heterogeneity is already captured in the property random effect, and there is almost no additional gain from the spatial structure.

We use two different specifications for the spatial effects. The first specification is a Besag model where a neighbor is defined by properties within a

specific radius from the subject property. The second specification is a spatial random walk, a restricted Besag model, where neighbors are defined by the preceding and subsequent property on the TSP-route, so having at most 2 neighbors. The out-of-sample prediction results for both models are comparable, so the reduction of a two-dimensional plane to a one-dimensional line does not lead to a lower performance in our applications, and the correlations between the estimated spatial effects in both models are high. Moreover, the spatial random walk model is computationally much more efficient.

Note that we use a simple time series structure, a random walk, to model the spatial effects. More complex structural time series models, taking into account distances between properties, could also be applied, possibly improving model fit, but we leave this for future research.

References

Anselin, L., and N. Lozano-Gracia, 2009, Spatial hedonic models, in T. C. Mills, and K. Patterson, eds., *Palgrave Handbook of Econometrics, Volume 2, Applied Econometrics*, 1213–1250 (Palgrave MacMillan).

Bailey, M. J., R. F. Muth, and H. O Nourse, 1963, A regression method for real estate price index construction, *Journal of the American Statistical Association* 58, 933–942.

Basu, S., and T. G. Thibodeau, 1998, Analysis of spatial autocorrelation in house prices, *Journal of Real Estate Finance and Economics* 17(1), 61–85.

Bentley, J. J., 1992, Fast algorithms for geometric traveling salesman problems, *ORSA Journal on computing* 4, 387–411.

Berndt, E. R., Z. Griliches, and N. J. Rappaport, 1995, Econometric estimates of price indexes for personal computers in the 1990's, *Journal of Econometrics* 68, 243–268.

Besag, J., 1974, Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society. Series B (Methodological)* 36(2), 192–236.

Besag, J., and C. Kooperberg, 1995, On conditional and intrinsic autoregressions, *Biometrika* 82, 733–746.

Besag, J., J. York, and A. Mollié, 1991, Bayesian image restoration, with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics* 43, 1–20.

Blangiardo, M., and M. Cameletti, 2015, *Spatial and spatio-temporal Bayesian models with R - INLA* (John Wiley & Sons, Ltd).

Bokhari, S., and D. M. Geltner, 2011, Loss aversion and anchoring in commercial real estate pricing: Empirical evidence and price index implications, *Real Estate Economics* 39(4), 635–670.

Bokhari, S., and D. M. Geltner, 2016, Characteristics of depreciation in commercial and multifamily property: An investment perspective, *Real Estate Economics* online, 1–38.

Bourassa, S. C., E. Cantoni, and M. Hoesli, 2007, Spatial dependence, housing markets, and house price prediction, *Journal of Real Estate Finance and Economics* 35(2), 143–160.

Case, B., and J. M. Quigley, 1991, The dynamics of real estate prices, *Review of Economics and Statistics* 73, 50–58.

Case, K. E, and R. J. Shiller, 1987, Prices of single family homes since 1970: New indexes for four cities, *New England Economic Review* 45–56.

Chegut, A. M., P. M. A. Eichholtz, and P. J. M. Rodrigues, 2015, Spatial dependence in international office markets, *Journal of Real Estate Finance and Economics* 51(2), 317–350.

Coulson, E. N., and D. McMillen, 2008, Estimating time, age and vintage effects in housing prices, *Journal of Housing Economics* 17, 138–151.

Court, A. T., 1939, Hedonic price indexes with automotive examples., *The Dynamics of Automobile Demand,* General Motors Corporation, New York, 99–117.

de Haan, J., and W. E. Diewert, 2011, *Handbook on residential property price indexes*, chapter Hedonic regression methods, 50–64 (Eurostat Methodologies & Working papers).

Diewert, W. E., J. de Haan, and R. Hendriks, 2015, Hedonic regressions and the decomposition of a house price index into land and structure components, *Econometric Reviews* 34, 106–126.

Francke, M. K., and A. F. De Vos, 2000, Efficient computation of hierarchical trends, *Journal of Business and Economic Statistics* 18, 51–57.

Francke, M. K., D. M. Geltner, A. M. Van de Minne, and R. White, 2017, Real estate index revisions in thin markets, Technical report, MIT Center for Real Estate.

Francke, M. K., and A. M. van de Minne, 2017a, The hierarchical repeat sales model for granular commercial real estate and residential price indices, *The Journal of Real Estate Finance and Economics* 55(4), 511–532.

Francke, M. K., and A. M. van de Minne, 2017b, Land, structure and depreciation, *Real Estate Economics* 45(2), 415–451.

Geltner, D. M., and A. M. van de Minne, 2017, Age, productivity & property investment performance: A new Bayesian hybrid approach for Los Angeles., Technical report, MIT Center for Real Estate.

Gutin, G., A. Yeo, and A. Zverovich, 2002, Traveling salesman should not be greedy: domination analysis of greedy-type heuristics for the TSP, *Discrete Applied Mathematics* 117, 81–86.

Hahsler, M., and K. Hornik, 2007, TSP-infrastructure for the traveling salesperson problem, *Journal of Statistical Software* 23, 1–21.

Harvey, A., 1989, *Forecasting Structural Time Series Models and the Kalman Filter* (Cambridge University Press, Cambridge).

Hill, R. J, 2012, Hedonic price indexes for residential housing: A survey, evaluation and taxonomy, *Journal of Economic Surveys* 27(5), 879–914.

Hoffman, M. D., and A. Gelman, 2014, The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo., *Journal of Machine Learning Research* 15, 1593–1623.

Hwang, M., and J. M. Quigley, 2004, Selectivity, quality adjustment and mean reversion in the measurement of house values, *Journal of Real Estate Finance and Economics* 28(2–3), 161–178.

Knight, J.R., and C. F. Sirmans, 1996, Depreciation, maintenance, and housing prices, *Journal of Housing Economics* 5, 369–389.

Kok, N., E. Koponen, and C. A. Martínez-Barbosa, 2017, Big data in real estate? From manual appraisal to automated valuation, *The Journal of Portfolio Management* 43, 202–211.

Lawler, E. L., J. K. Lenstra, A. H. G. Rinnooy Kan, D. B. Shmoys, et al., 1985, *The traveling salesman problem: a guided tour of combinatorial optimization*, volume 3 (Wiley New York).

Lindgren, F., H. Rue, and J. Lindström, 2011, An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73, 423–498.

Nappi-Choulet, I., and T. Maury, 2009, A spatiotemporal autoregressive price index for the Paris office property market, *Real Estate Economics* 37(2), 305–340.

Pace, R. K., Barry R., J. M. Clapp, and M. Rodriquez, 1998, Spatiotemporal autoregressive models of neighborhood effects, *Journal of Real Estate Finance and Economics* 17(1), 15–33.

Pace, R. K., C. F. Sirmans, and V. C. Slawson Jr, 2002, Automated valution models, in K. Wang, and M. L. Wolverton, eds., *Real Estate Valuation Theory, Research Issues in Real Estate Volume 8*, 133–156 (Appraisal Institute and American Real Estate Society, Kluwer Academic Publishers).

Quigley, J. M., 1995, A simple hybrid model for estimating real estate price indexes, *Journal of Housing Economics* 4, 1 – 12.

Reis, H. J., and J. M. C. Santos Silva, 2006, Hedonic prices indexes for new passenger cars in Portugal (1997–2001), *Economic Modelling* 23, 890–908.

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

Rue, H., S. Martino, and N. Chopin, 2009, Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71, 319–392.

Sirmans, S. G., L. MacDonald, D. A. Macpherson, and E. N. Zietz, 2006, The value of housing characteristics: a meta analysis, *The Journal of Real Estate Finance and Economics* 33(3), 215–240.

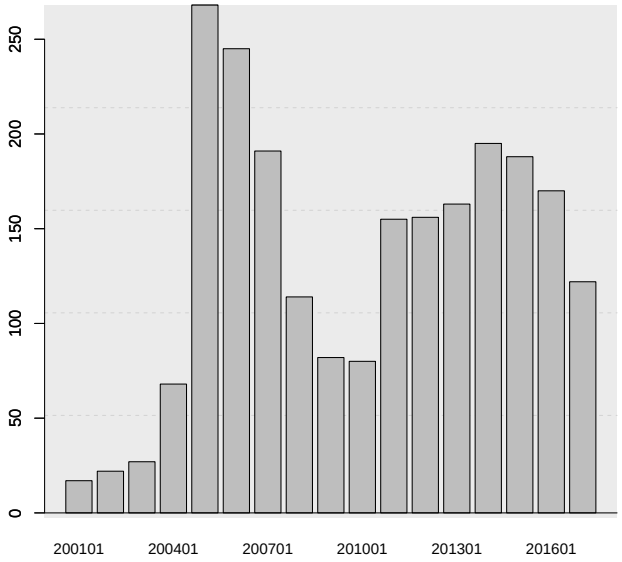
Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde, 2002, Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64, 583–639.

Tu, Y., S. Yu, and H. Sun, 2004, Transaction-based office price indexes: A spatiotemporal modeling approach, *Real Estate Economics* 32(2), 297–328.

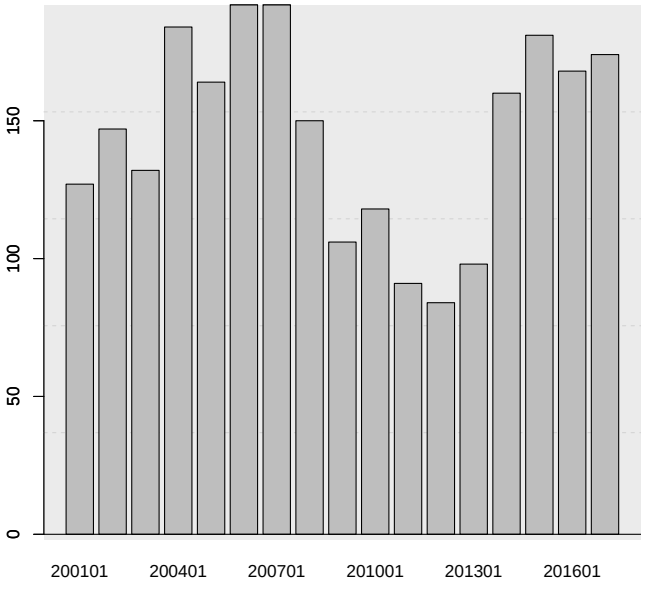
Watanabe, S., 2010, Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research* 11, 3571–3594.

Wilhelmsson, M., 2008, House price depreciation rates and level of maintenance, *Journal of Housing Economics* 17, 88–101.

Appendix A. Figures

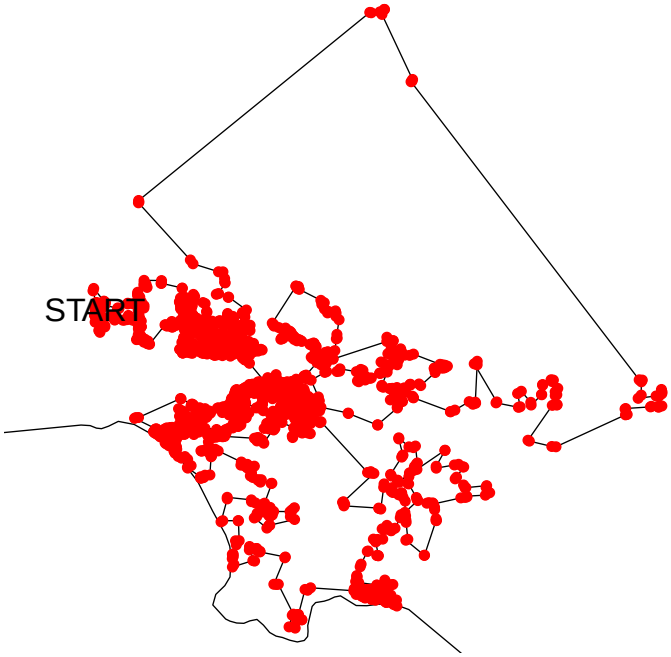


(a) Los Angeles.

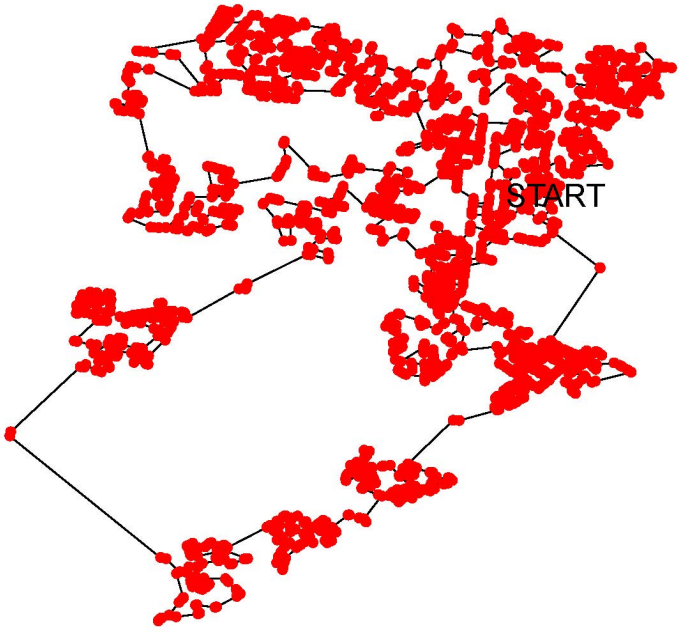


(b) Heemstede.

Figure A.1: Annual number of sales.

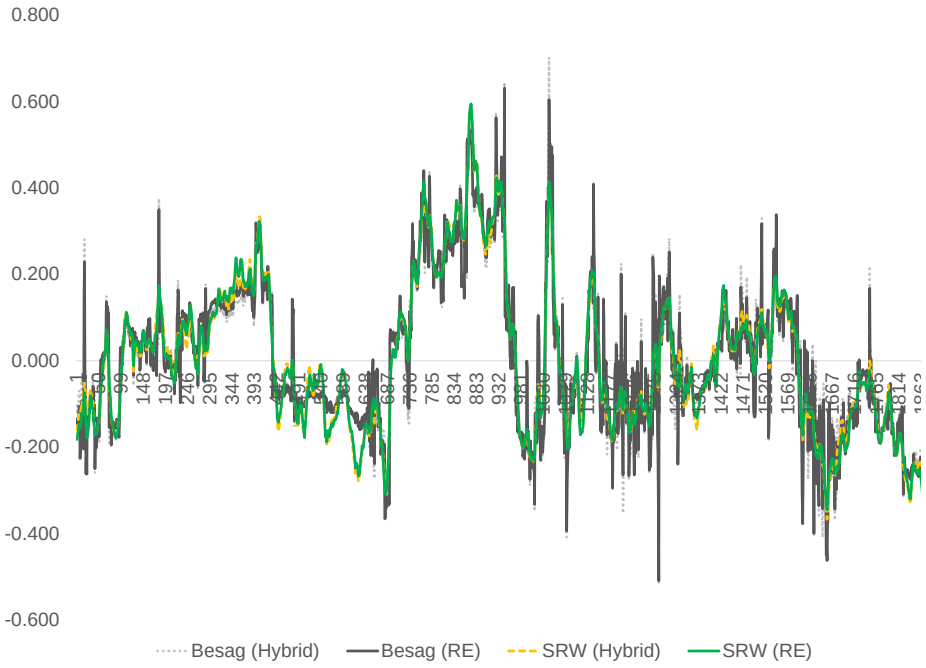


(a) Los Angeles.

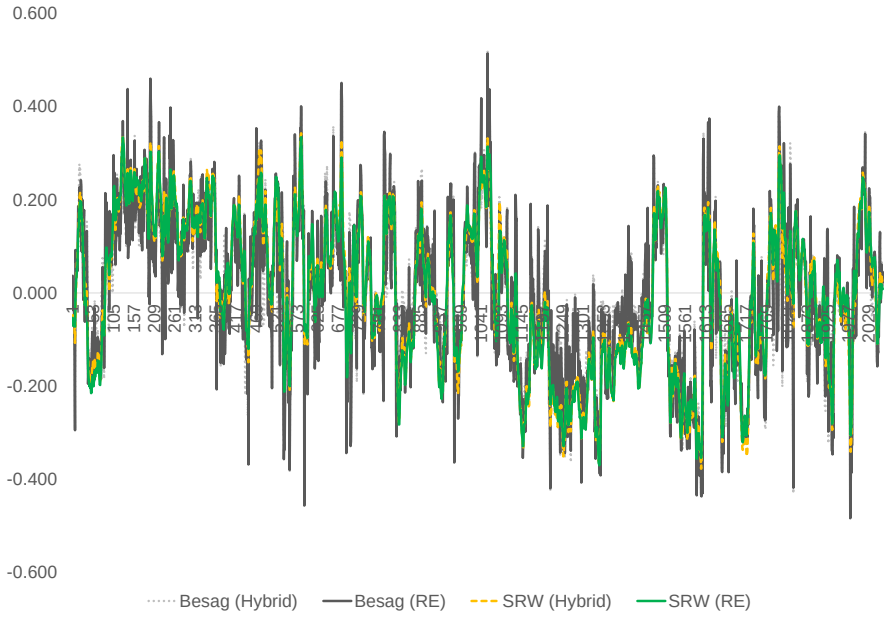


(b) Heemstede.

Figure A.23: TSP route.

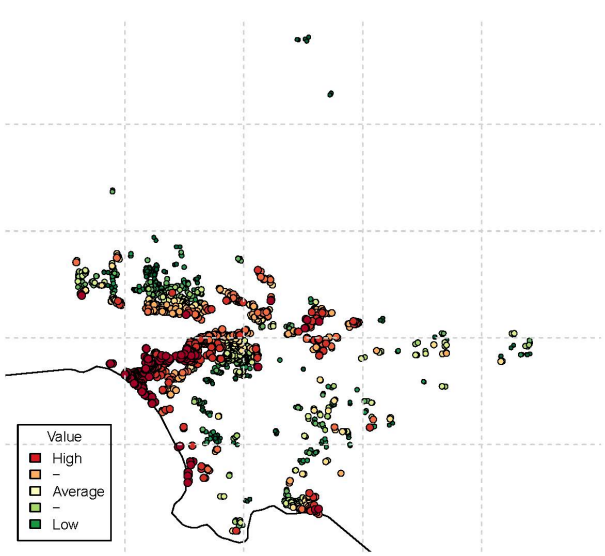


(a) Los Angeles.

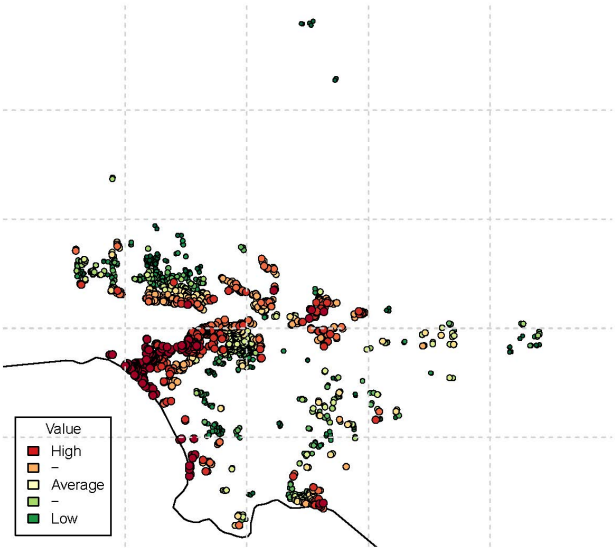


(b) Heemstede.

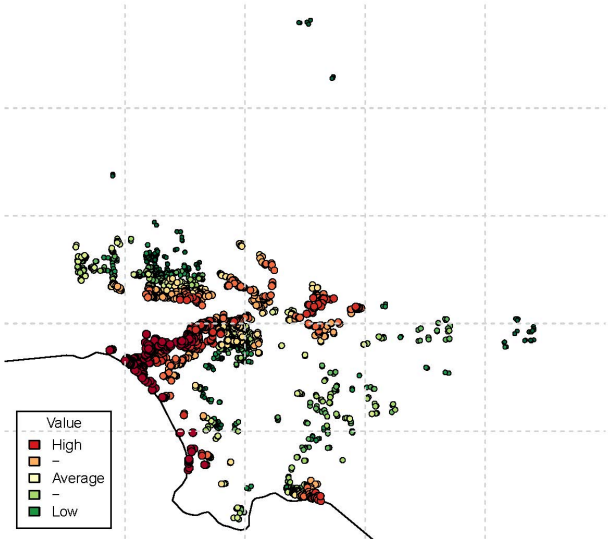
Figure A.3: Spatial effects θ values over TSP route.



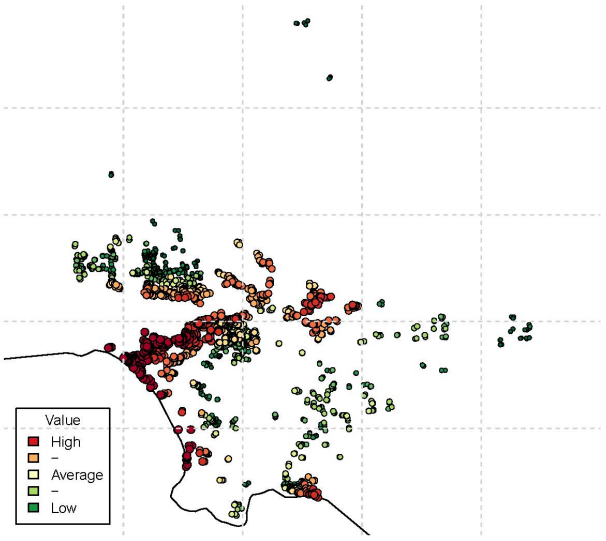
(a) Besag (Hybrid).



(b) Besag (RE).

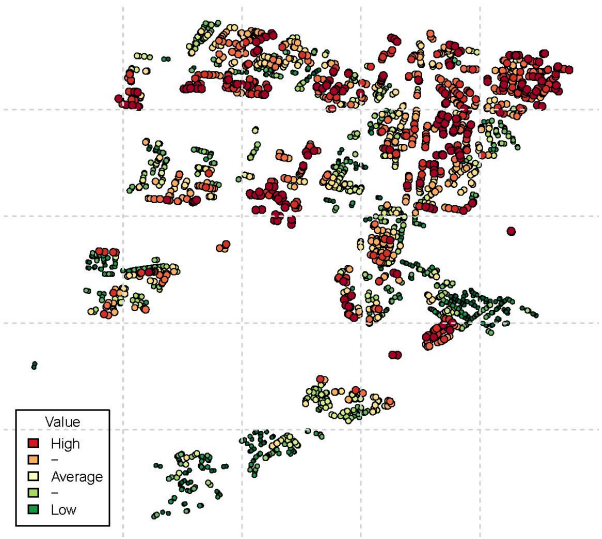


(c) SRW (Hybrid).

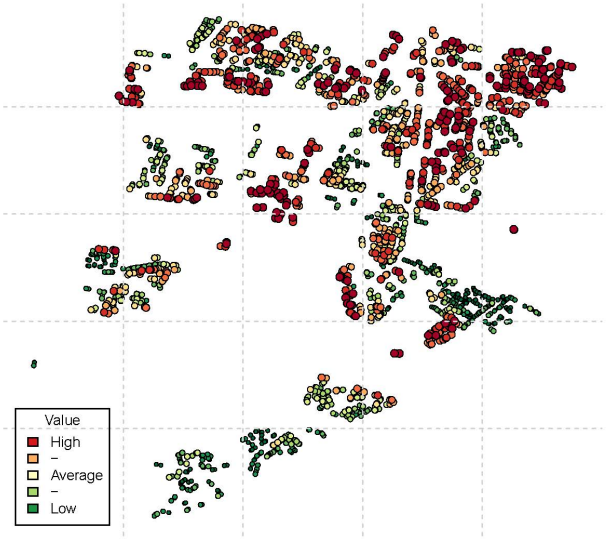


(d) SRW (RE).

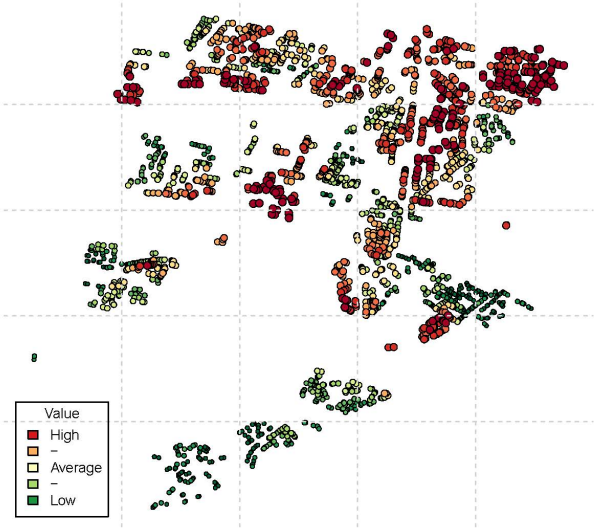
Figure A.4: Heat map of spatial effects for Los Angeles.



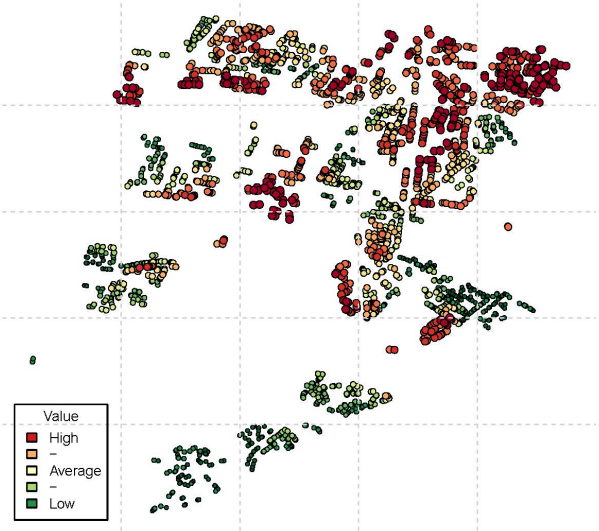
(a) Besag (Hybrid).



(b) Besag (RE).



(c) SRW (Hybrid).



(d) SRW (RE).

Figure A.5: Heat map of spatial effects for Heemstede.

Appendix B. Tables

Table B.1: Descriptive statistics.

	mean	**sd**	**min**	**max**
	Los Angeles			
Sales price ($)[1]	6,389,494	6,460,729	1,550,000	64,250,000
Net Operating Income ($)	325,223	353,270	67,200	3,600,000
Age (Years)	45	21	2	97
Size (SqFt)	31,718	31,972	5,964	271,757
Years between sales	4.61	2.87	0.17	12.00
Garden (R)	0.87		0	1
Mid/Highrise	0.13		0	1
Observations	2,263			
Unique properties	1,936			
	Heemstede			
Sales price (€)[1]	484,612	189,106	200,000	1,195,000
Age (Years)	65	22	15	106
Size (SqMt)	151	39	82	288
Years between sales	7.07	3.67	0.42	16.25
Maintenance [bad] (R)	0.18		0	1
Maintenance [average]	0.59		0	1
Maintenance [good]	0.24		0	1
Row house (R)	0.44		0	1
Semi detached (1)	0.03		0	1
Semi detached (2)	0.23		0	1
Corner home	0.25		0	1
Detached	0.05		0	1
Yard (yes)	0.94		0	1
Observations	2,468			
Unique properties	2,065			

R gives the reference categories in our model. Semi detached (1) are properties that are connected via a garage, and Semi detached (2) are properties that are connected wall-to-wall.
[1] Estimate for Moran's I (sales prices) for Los Angeles and Heemstede are respectively +0.04 and+0.23.

Table B.2: Los Angeles estimation results: posterior means and in-sample fit statistics.

				Besag		SRW	
	Standard	**Hybrid**	**RE**	**Hybrid**	**RE**	**Hybrid**	**RE**
(Intercept)	3.940***	4.091***	4.053***	4.135***	4.096***	4.029***	4.033***
ln Size	0.932***	0.924***	0.928***	0.945***	0.952***	0.956***	0.961***
ln $\left(\frac{\text{NOI}}{\text{Size}}\right)$	0.735***	0.708***	0.700***	0.613***	0.609***	0.603***	0.602***
Age	-0.001***	-0.001***	-0.001***	-0.003***	-0.004***	-0.004***	-0.004***
Age^2	0.000***	0.000	0.000	0.000	0.000*	0.000	0.000***
Mid/Highrise	0.018	0.014	0.019	-0.003***	-0.002***	-0.004***	-0.005***
2002	0.060	0.014	0.032	0.054	0.048	0.089*	0.094**
2003	0.170***	0.153*	0.140*	0.192***	0.184***	0.216***	0.202***
2004	0.275***	0.272***	0.264***	0.308***	0.293***	0.334***	0.326***
2005	0.318***	0.328***	0.336***	0.409***	0.397***	0.439***	0.437***
2006	0.344***	0.350***	0.362***	0.433***	0.421***	0.464***	0.461***
2007	0.309***	0.321***	0.333***	0.416***	0.398***	0.446***	0.437***
2008	0.318***	0.315***	0.329***	0.410***	0.402***	0.444***	0.440***
2009	0.217***	0.220***	0.213***	0.287***	0.264***	0.312***	0.297***
2010	0.192***	0.188***	0.203***	0.281***	0.273***	0.311***	0.309***
2011	0.253***	0.247***	0.258***	0.324***	0.317***	0.365***	0.359***
2012	0.287***	0.292***	0.294***	0.364***	0.356***	0.398***	0.392***
2013	0.320***	0.324***	0.330***	0.454***	0.450***	0.478***	0.475***
2014	0.411***	0.428***	0.439***	0.558***	0.547***	0.592***	0.585***
2015	0.528***	0.545***	0.554***	0.664***	0.656***	0.693***	0.691***
2016	0.628***	0.621***	0.641***	0.762***	0.762***	0.791***	0.796***
2017	0.622***	0.639***	0.648***	0.775***	0.767***	0.814***	0.811***
Location	FE	FE	FE				
σ_ϵ	0.190	0.184	0.124	0.126	0.123	0.137	0.120
σ_ϕ			0.146		0.149		0.040
σ_θ				0.033	0.029	0.011	0.064
$\sigma_\theta/\sqrt{d+\bar{w}_{p+}}$				0.010	0.008		
Moran I	0.112	0.079	0.084	0.000	0.003	-0.003	-0.002
DIC	-1,072.1	-919.8	-1,850.6	-2,132.5	-2,412.0	-2,035.4	-2,502.4
WAIC	-1,070.6	-1,031.9	-1,891.2	-2,220.0	-2,443.5	-2,074.4	-2,459.2

The omitted dummy variable is garden apartment (for property subtype) and 2001 (for time of sale). Moran's I is a measure for spatial autocorrelation and NOI stands for Net Operating Income. DIC denotes Deviance Information Criterion, and WAIC Watanabe Information Criterion.
*** means the parameter is significantly different from 0 at the 1% level, ** at the 5% level and * at the 10% level.

Table B.3: Heemstede estimation results: posterior means and in-sample fit statistics.

				Besag		SRW	
	Standard	**Hybrid**	**RE**	**Hybrid**	**RE**	**Hybrid**	**RE**
(Intercept)	8.010***	8.060***	8.135***	8.836***	8.891***	9.127***	9.095***
ln Size	0.898***	0.889***	0.876***	0.723***	0.714***	0.685***	0.694***
Age	0.004***	0.002**	0.003***	0.003***	0.003***	0.001	0.001
Age^2	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***
Semi detached (1)	0.148***	0.162***	0.137***	0.083***	0.074***	0.088***	0.073***
Semi detached (2)	0.199***	0.214***	0.205***	0.167***	0.152***	0.174***	0.163***
Corner Home	0.094***	0.117***	0.104***	0.100***	0.087***	0.109***	0.094***
Detached	0.337***	0.357***	0.347***	0.307***	0.292***	0.301***	0.286***
Maintenance [average]	0.128***	0.127***	0.133***	0.129***	0.130***	0.123***	0.128***
Maintenance [good]	0.216***	0.209***	0.208***	0.204***	0.208***	0.206***	0.207***
Yard	0.025	0.031*	0.025	0.013	0.014	0.015	0.016
2002	0.047**	0.041*	0.036*	0.031*	0.034**	0.032**	0.031**
2003	0.039*	0.036	0.027	0.026	0.029*	0.038*	0.038**
2004	0.078***	0.084***	0.068***	0.071***	0.071***	0.075***	0.068***
2005	0.141***	0.133***	0.134***	0.128***	0.130***	0.140***	0.139***
2006	0.171***	0.180***	0.170***	0.168***	0.171***	0.185***	0.181***
2007	0.261***	0.261***	0.252***	0.247***	0.253***	0.270***	0.272***
2008	0.280***	0.290***	0.278***	0.284***	0.283***	0.295***	0.286***
2009	0.244***	0.232***	0.232***	0.225***	0.227***	0.239***	0.241***
2010	0.224***	0.213***	0.216***	0.204***	0.213***	0.228***	0.228***
2011	0.224***	0.212***	0.223***	0.216***	0.224***	0.228***	0.238***
2012	0.135***	0.132***	0.133***	0.115***	0.116***	0.132***	0.133***
2013	0.135***	0.136***	0.135***	0.118***	0.113***	0.123***	0.121***
2014	0.176***	0.173***	0.185***	0.158***	0.161***	0.177***	0.184***
2015	0.216***	0.207***	0.230***	0.221***	0.229***	0.232***	0.241***
2016	0.344***	0.340***	0.354***	0.334***	0.343***	0.358***	0.365***
2017	0.420***	0.415***	0.434***	0.403***	0.408***	0.420***	0.427***
Location	FE	FE	FE				
σ_ϵ	0.171	0.164	0.102	0.096	0.095	0.116	0.098
σ_ϕ			0.139		0.011		0.070
σ_θ				0.119	0.118	0.057	0.053
$\sigma_\theta/\sqrt{d+\bar{w}_{p+}}$				0.049	0.048		
Moran's I	0.089	0.037	0.019	0.005	-0.021	-0.017	-0.011
DIC	-1,671.4	-1,517.1	-2,835.4	-2,581.9	-3,538.2	-2,813.2	-3,395.0
WAIC	-1,670.3	-1,664.6	-2,925.9	-2,700.3	-3,569.0	-2,834.2	-3,384.3

The omitted dummy variables are row house (for property subtype), maintenance [bad], and having no yard. Semi Detached (1) are houses connected by a garage, and Semi Detached (2) are houses that are connected wall-to-wall and 2001 (for time of sale).
Moran's I is a measure for spatial autocorrelation. DIC denotes Deviance Information Criterion, and WAIC Watanabe Information Criterion.
*** means the parameter is significantly different from 0 at the 1% level, ** at the 5% level and * at the 10% level.

Table B.4: LOO cross validation.

				Besag		SRW	
	Standard	**Hybrid**	**RE**	**Hybrid**	**RE**	**Hybrid**	**RE**
	Los Angeles						
Mean	0.000	-0.005	0.001	-0.001	0.000	-0.001	0.000
|Mean|	0.146	0.146	0.140	0.111	0.108	0.111	0.109
Standard deviation	0.191	0.190	0.184	0.150	0.147	0.148	0.146
Minimum	-0.777	-0.769	-0.759	-0.814	-0.830	-0.850	-0.846
Maximum	1.022	1.005	1.003	0.747	0.755	0.772	0.775
	Heemstede						
Mean	0.000	0.004	-0.001	0.001	0.001	0.004	-0.001
|Mean|	0.137	0.137	0.133	0.108	0.105	0.109	0.106
Standard deviation	0.172	0.172	0.164	0.173	0.130	0.136	0.133
Minimum	-0.656	-0.639	-0.663	-0.719	-0.547	-0.634	-0.653
Maximum	0.463	0.462	0.467	0.491	0.457	0.418	0.426

Table B.5: Absolute mean of LOO residuals as a function of the number of sales per property.

# Sales				**Besag**		**SRW**		
per property	**Standard**	**Hybrid**	**RE**	**Hybrid**	**RE**	**Hybrid**	**RE**	**Prop.**
	Los Angeles							
1	0.148	0.149	0.149	0.115	0.112	0.115	0.114	1,643
2	0.138	0.141	0.115	0.098	0.096	0.098	0.095	261
3	0.139	0.121	0.112	0.112	0.101	0.110	0.102	30
Total	0.146	0.146	0.140	0.111	0.108	0.111	0.109	2,263
	Heemstede							
1	0.139	0.139	0.139	0.112	0.108	0.112	0.110	1,703
2	0.129	0.130	0.102	0.093	0.090	0.094	0.091	644
3	0.120	0.100	0.091	0.088	0.087	0.088	0.083	117
Total	0.137	0.137	0.133	0.108	0.105	0.109	0.106	2,468

Table B.6: Summary statistics of spatial effects θ.

	Besag (Hybrid)	**Besag (RE)**	**SRW (Hybrid)**	**SRW (RE)**
		Los Angeles		
Mean	0.000	0.000	0.000	0.000
|Mean|	0.124	0.156	0.159	0.182
Standard deviation	0.168	0.171	0.169	0.171
Minimum	-0.515	-0.509	-0.367	-0.345
2.5%-perc	-0.277	-0.269	-0.259	-0.262
97.5%-perc	0.367	0.380	0.368	0.381
Maximum	0.702	0.631	0.593	0.595
		Correlations		
Besag (Hybrid)		0.988	0.935	0.933
Besag (RE)			0.942	0.948
SRW (Hybrid)				0.996
		Heemstede		
Mean	0.000	0.000	0.000	0.000
|Mean|	0.033	0.037	0.070	0.069
Standard deviation	0.161	0.166	0.162	0.161
Minimum	-0.486	-0.483	-0.377	-0.370
2.5%-perc.	-0.319	-0.326	-0.303	-0.299
97.5%-perc.	0.296	0.299	0.266	0.259
Maximum	0.521	0.514	0.343	0.334
		Correlations		
Besag (Hybrid)		0.975	0.896	0.880
Besag (RE)			0.904	0.908
SRW (Hybrid)				0.991

Table B.7: Standard metrics for the robustness check.

	All Variables		Reduced Data-set	
	Standard	**SRW (RE)**	**Standard**	**SRW (RE)**
	Los Angeles			
σ_ϵ	0.190	0.120	0.289	0.139
DIC	-1,072.1	-2,502.4	825.1	-1,243.8
WAIC	-1,070.6	-2,459.2	831.0	-1,302.6
	Heemstede			
σ_ϵ	0.171	0.098	0.204	0.118
DIC	-1,671.4	-3,395.0	-824.2	-2,531.8
WAIC	-1,670.3	-3,384.3	-823.2	-2,510.8

The results with all variables, can also be found in Tables B.2 – B.3. For Los Angeles the reduced data-set does not include (log of) Net Operating Income per square foot. In order to create the reduced data-set for Heemstede, we omit the variables on property types and maintenance levels.

Table B.8: Absolute mean of LOO residuals as a function of the number of sales per property for the robustness check.

# Sales per property	**All Variables**		**Reduced Data-set**		
	Standard	**SRW (RE)**	**Standard**	**SRW (RE)**	**Prop.**
	Los Angeles				
1	0.148	0.114	0.222	0.172	1,643
2	0.138	0.095	0.219	0.125	261
3	0.139	0.102	0.248	0.124	30
Total	0.146	0.109	0.222	0.159	2,263
	Heemstede				
1	0.139	0.110	0.164	0.130	1,703
2	0.129	0.091	0.151	0.107	644
3	0.120	0.093	0.130	0.087	117
Total	0.135	0.104	0.159	0.122	2,468

The results with all variables, can also be found in Tables B.2 – B.3. For Los Angeles the reduced data-set does not include (log of) Net Operating Income per square foot. In order to create the reduced data-set for Heemstede, we omit the variables on property types and maintenance levels.