Modeling unobserved heterogeneity in hedonic price models

Francke, M.; van de Minne, A.

DOI
10.1111/1540-6229.12320

Publication date
2021

Document Version
Final published version

Published in
Real Estate Economics

License
CC BY-NC-ND

Link to publication

Citation for published version (APA):
Francke, M., & van de Minne, A. (2021). Modeling unobserved heterogeneity in hedonic price models. *Real Estate Economics*, *49*(4), 1315-1339. https://doi.org/10.1111/1540-6229.12320

DOI: 10.1111/1540-6229.12320

ORIGINAL ARTICLE

Modeling unobserved heterogeneity in hedonic price models

Marc Francke[1] | Alex Van de Minne[2]

[1]Faculty of Economics and Business, University of Amsterdam, Amsterdam, The Netherlands

[2]Center for Real Estate and Urban Economic Studies, University of Connecticut, Connecticut

Correspondence
Marc Francke, Faculty of Economics and Business, University of Amsterdam, Plantage Muidergracht 12, 1018 TV Amsterdam and Ortec Finance, Naritaweg 51, 1043 BP Amsterdam, The Netherlands.
Email: M.k.francke@uva.nl

Abstract

This paper studies unobserved heterogeneity in hedonic price models, arising from missing property and locational characteristics. Specifically, 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.

1 | INTRODUCTION

This paper studies unobserved heterogeneity in hedonic price models (*HPM*s). These models are widely used, for example, to create price indexes (and concomitant deprecation) for cars (Berndt, Griliches, & Rappaport, 1995), computers (Reis & Santos Silva, 2006), and residential housing (Hill, 2012), among many other types of goods. The number of applications within real estate is large. The *HPM* has, for example, been used to value residential housing (Francke & De Vos, 2000; Sirmans, MacDonald, Macpherson, & Zietz, 2006), commercial real estate (Bokhari & Geltner, 2011), (residential) land

(Diewert, de Haan, & Hendriks, 2015), and to estimate the depreciation rate of houses (Francke & van de Minne, 2017b; Knight & Sirmans, 1996).

Rosen (1974) explicated the formal microeconomic theory underlying *HPM*s, although the technique has older roots in consumer and marketing empirical analytics practice (Court, 1939). The basic idea is that heterogeneous goods can be described by their attributes (de Haan & Diewert, 2013). In other words, a good is a bundle of characteristics. In the case of real estate properties, the relevant bundle contains attributes of the building structure and 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 the nearest subway station. There is no market for the individual 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. Regression-based techniques are typically used to estimate these marginal value contributions.

*HPM*s for residential and in particular commercial real estate properties are in practice hard to develop. First, properties are heterogeneous in nature, implying many potential value drivers. Second, the property turnover rate, and so the number of transactions, is relatively low. Third, the number of recorded property characteristics is in most real estate databases quite limited: many value drivers are missing. And when they are sufficiently available, there is the risk of misspecification and overfitting.

This paper focuses on unobserved heterogeneity in *HPM*s, due to missing property and locational characteristics, at least from the perspective of the econometrician. Unobserved heterogeneity in *HPM*s arises when available characteristics do not fully account for the observed heterogeneity in sale prices of individual properties. If characteristics that are related to sale prices are missing, the estimated coefficients of the included characteristics will be biased (Arellano, 2003). The focus of this paper is on out-of-sample sale price prediction accuracy.

This paper has broad relevance, but real estate is a particularly important area (see Gormley & Matsa, 2013, for other applications). Real estate is characterized by very long-lived goods that therefore often transact more than once, and also by the importance of spatial location. With this in mind, we control for time invariant property-related unobserved heterogeneity by adding mutual independent property-level random effects to the *HPM*, taking advantage of the fact that some properties transact more than once. Moreover, we add time invariant spatial random effects to deal with spatial dependencies. Spatial dependencies exist because nearby properties often have similar characteristics and also share locational amenities (Basu & Thibodeau, 1998). It is expected that the inclusion of property and spatial random effects in the *HPM* captures time invariant unobserved heterogeneity, and increases out-of-sample prediction performance, in particular for repeat sales.

The property random effects *HPM* is related to the hybrid hedonic-repeat sales model (Case & Quigley, 1991). The main difference is that the former includes random effects for all properties, where the latter inconsistently includes fixed effects for repeat sales only, and not for one only sales.

We use two different specifications to model 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. An obvious disadvantage of the spatial random walk is that we reduce a two-dimensional plane into a one-dimensional line, at the risk of ignoring important information. An important advantage of the spatial random walk over the general Besag model is its relative ease of estimation; it considerably reduces computation time, especially in a large data environment.

We evaluate out-of-sample predictions for seven *HPM*s: the standard *HPM*, the hybrid model, the property random effects *HPM* (all three including location fixed effects), and two spatial models (Besag and spatial random walk) for the hybrid model and property random effects *HPM* (all four excluding location fixed effects). We perform leave-one-out (*LOO*) cross-validation to measure the out-of-sample prediction performance for the seven *HPM*s, so we can check whether adding property and spatial random effects helps to increase out-of-sample prediction accuracy, and whether the spatial random walk model performs similar to the Besag model. We use an efficient Bayesian estimation procedure, Integrated Nested Laplace Approximation (*INLA*, see Rue, Martino, and Chopin, 2009), for all *HPM*s, as *LOO* analysis is computational expensive.

To illustrate the different methods, we estimate the seven *HPM*s on multifamily housing (income generating properties) 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 *HPM* improves the prediction accuracy, more than in the hybrid model. The standard deviation of the *LOO* residuals is reduced by approximately 5% in both markets. Adding property and spatial random 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 .93 and .99 in Los Angeles, and between .88 and .99 in Heemstede.

When having only one sale per property, the property random effects *HPM* including spatial effects performs better than the model excluding 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 and spatial random effects *HPM* outperforms more standard *HPM*s even after excluding important characteristics.

The contribution of the paper is threefold. First, we add to the literature a property random effects HPM that controls for unobserved heterogeneity, a consistent model for one only and repeat sales, unlike the hybrid model, and in our applications better performing than the hybrid model. Second, we specify spatial random effects by a spatial random walk model, having similar performance as the well-established Besag model, however, having a large computational advantage. Third, we systematically perform LOO cross-validation to analyze out-of-sample performance for the seven *HPM*s. In many other studies, in-sample fit statistics are mainly being used.

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 | Hedonic price model

The *HPM* is widely used for modeling and tracking the prices of heterogeneous goods (including real estate). The *HPM* can be expressed as

$$y_p = x_p\beta + \epsilon_p, \quad \epsilon_p \sim \mathcal{N}\left(0, \sigma_\epsilon^2 I_{n_p}\right), \quad p = 1, \dots, P, \tag{1}$$

where the dependent variable y_p is a $(n_p \times 1)$ vector of log prices for property p, and n_p is the number of transactions for property p. For one only sales, it holds that $n_p = 1$ and for repeat sales, $n_p > 1$. P is the number of properties, and $N = \sum_{p=1}^{P} n_p$ is the total number of transactions over all properties.[1]

The $(n_p \times K)$ matrix x_p represents observable hedonic characteristics with corresponding coefficient vector β, constant over time. Apart from property characteristics and a constant, the matrix x_p could include location and time fixed effects. Note that we allow for changes over time in the characteristics of the same property. The error term ϵ_p is assumed to be normally and independently distributed with zero mean and variance $\sigma_\epsilon^2 I_{n_p}$, where I denotes the identity matrix.

The *HPM* is typically estimated by ordinary least squares (*OLS*). The estimated coefficient vector $\hat{\beta}$ 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, and hence, can be used to produce a price index (pooled *HPM*).

In this paper, we take specific interest in how to cope with time invariant unobserved heterogeneity. Unobserved heterogeneity is reflected in omitted variable bias, lower model fit, and out-of-sample prediction performance. Unobserved heterogeneity is in specific a problem for commercial real estate (Francke & 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 to model unobserved heterogeneity in the *HPM* is to include mutually independent property random effects ϕ_p in Equation (1), taking advantage of the fact that some properties transact more than once, leading to the following model,

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

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

where j is a vector of ones. The property random effects ϕ_p absorb time invariant omitted variables and model misspecification, and the error terms ϵ_p represent transaction noise, the difference between

[1] Note that the HPM can be expressed simpler in terms of individual transactions. However, we specify the HPM per property, to keep our notation consistent with the other models presented in the remainder of this section.

the market value ($x_p\beta + j_{n_p}\phi_p$), and the actual transaction price (y_p). Note that we allow for changes over time in the characteristics of the same property (x_p), as in the standard *HPM*.

$y = (y_1', \ldots, y_P')'$ and $X = (x_1', \ldots, x_P')'$. Conditional on variance parameters σ_ϵ^2 and σ_ϕ^2 estimates of β and ϕ_p in the property random effects *HPM* (Equations (2) and (3)) are provided by

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

$$\hat{\beta} = \mathrm{Var}\left(\hat{\beta}\right) \sum_{p=1}^{P} \left(x_p' \Omega_p^{-1} y_p\right), \ \mathrm{Var}\left(\hat{\beta}\right) = \left(\sum_{p=1}^{P} \left(x_p' \Omega_p^{-1} x_p\right)\right)^{-1},$$

$$\phi_p | y, X, \sigma_\epsilon^2, \sigma_\phi^2 \sim \mathcal{N}\left(\hat{\phi}_p, \mathrm{Var}\left(\hat{\phi}_p\right)\right), \tag{5}$$

$$\hat{\phi}_p = \omega_p n_p \left(\bar{y}_p - \bar{x}_p \hat{\beta}\right), \ \mathrm{Var}\left(\hat{\phi}_p\right) = \sigma_\epsilon^2 \omega_p + \left(n_p \omega_p\right)^2 \bar{x}_p \mathrm{Var}\left(\hat{\beta}\right) \bar{x}_p',$$

where $\Omega_p = \sigma_\epsilon^2 I_{n_p} + \sigma_\phi^2 j_{n_p} j_{n_p}'$, $\Omega_p^{-1} = \sigma_\epsilon^{-2}\left(I_{n_p} - \omega_p j_{n_p} j_{n_p}'\right)$, $\omega_p = \sigma_\phi^2 / \left(\sigma_\epsilon^2 + n_p \sigma_\phi^2\right)$, and $\bar{y}_p$ and $\bar{x}_p$ are the averages of transactions prices and characteristics of property p. A derivation of Equations (4) and (5) is provided in the Appendix.

The part ($\omega_p n_p$) 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, see Equation (5): The larger n_p and the smaller the ratio of $\sigma_\epsilon^2/\sigma_\phi^2$ is, the larger this part is. When $\sigma_\phi^2 \to \infty$, corresponding to property fixed effects, $\hat{\phi}_p$ is equal to ($\bar{y}_p - \bar{x}_p\hat{\beta}$), the fixed effect estimator. The shrinkage is identical for all properties with equal number of transactions. Clapp and Zhou (2019) allow the shrinkage to depend on property characteristics (aptypicality). Note that also for one only sales ($n_p = 1$), it is possible to estimate the random effect ϕ_p.

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 $\hat{y}_p = x_p\hat{\beta} + j_{n_p}\hat{\phi}_p$.

A necessary condition for identification of ($\sigma_\epsilon^2, \sigma_\phi^2$)—in the absence of prior information for these variance parameters—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 in the sample period.

To provide some intuition on the identification of ($\sigma_\epsilon^2, \sigma_\phi^2$), we split Equation (2) in two parts: group (property) means and deviations from the means. The estimate of σ_ϵ^2 is primarily based on the observations in deviation from the group means, canceling out the property random effects: $\tilde{y}_p = \tilde{x}\beta + \tilde{\epsilon}_p$, $\mathrm{Var}(\tilde{\epsilon}_p) = (I_{n_p} - 1/n_p \times j_{n_p} j_{n_p}')\sigma_\epsilon^2$, where the tilde denotes observations in deviation from their group means, so with elements $\tilde{y}_{ip} = y_{ip} - \bar{y}_p$, where i indicates an individual transaction. Note that the deviations from the means equation need some properties to have more than one transaction, so repeat sales. Given an estimate of σ_ϵ^2 from the observations in deviation from their property mean, an estimate of σ_ϕ^2 can be derived from the property means equation: $\bar{y}_p = \bar{x}_p\beta + \phi_p + \bar{\epsilon}_p$, $\mathrm{Var}(\phi_p + \bar{\epsilon}_p) = \sigma_\phi^2 + \sigma_\epsilon^2/n_p$. For more details, see, for example, Greene (2008, Chapter 13).

More formally, estimation of ($\sigma_\epsilon^2, \sigma_\phi^2$) can be done by likelihood-based methods (without the need for splitting the observations in means and deviations from means). The loglikelihood function $\ell(y|\sigma_\epsilon^2, \sigma_\phi^2)$ for Equations (2)–(3) is proportional to

$$-\frac{1}{2}\sum_{p=1}^{P}\left(n_p \ln \sigma_\epsilon^2 + \ln\left(1 + n_p \sigma_\psi^2/\sigma_\epsilon^2\right) + \left(y_p - x_p\hat{\beta}\right)' \Omega_p^{-1}\left(y_p - x_p\hat{\beta}\right)\right).$$

The loglikelihood function can be maximized with respect to $(\sigma_\epsilon^2, \sigma_\phi^2)$. In this paper, we apply a Bayesian estimation method, combining the likelihood function with uninformative priors for $(\beta, \sigma_\epsilon^2, \sigma_\phi^2)$.[2]

Note that it is practically infeasible to replace the random effects by fixed effects: By including property fixed effects, one *effectively* excludes all one only 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. We will not apply the Hausman test, because the fixed effects estimator does not allow for time invariant characteristics, and, in practice, most of the characteristics do not change between the date of buying and selling.

In this paper, we will focus on out-of-sample cross-validation to compare model performance, see Section 2.5 for more details. In an *HPM* 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.

2.3 | Hybrid model

The property random effects *HPM* is related to the hybrid hedonic-repeat sales 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 one only sales y^S $(n_p = 1)$ and repeat sales y^R $(n_p > 1)$, and provide different specifications for both subsamples. The hybrid model is given by

$$\begin{pmatrix} y_p^S \\ \Delta y_p^R \end{pmatrix} = \begin{pmatrix} x_p^S \\ \Delta x_p^R \end{pmatrix} \beta + \begin{pmatrix} \epsilon_p^S \\ \Delta \epsilon_p^R \end{pmatrix}, \forall p : \begin{pmatrix} n_p = 1 \\ n_p > 1 \end{pmatrix}, \tag{6}$$

where Δy_p^R is a $(n_p - 1)$ vector of “first differences” of log prices of the same property (p) with elements $(y_{p,t}^R - y_{p,s}^R)$, the difference in the log price at the time of selling t and the time of buying s. The hybrid model can be estimated by *OLS*.

A statistical equivalent representation of the repeat sales part $(n_p > 1)$ of the hybrid model is in levels, provided by $y_p^R = x_p^R \beta + j_{n_p} \phi_p^{\text{FE}} + \epsilon_p^R$, including property fixed effects ϕ_p^{FE}. This shows that the hybrid model is inconsistent by specifying property fixed effects for repeat sales only. Therefore, we propose as an alternative the property random effects *HPM* to model unobserved heterogeneity, a consistent model for one only and repeat sales.

2.4 | Spatial dependencies

The property random effects 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 structure characteristics and share location characteristics and

[2] In order to be consistent, we estimate all models within this paper by a computational efficient Bayesian method, see Section 2.5 for more details; the more complicated models in Section 2.4 are hard to estimate by maximum likelihood. In our applications, the differences between the maximum likelihood estimators and the Bayesian posterior means are small.

amenities (Basu & Thibodeau, 1998). We add spatial property effects θ to the property random effects *HPM*, leading to

$$y_p = x_p\beta + j_{n_p}\phi_p + j_{n_p}\theta_p + \epsilon_p,\ \epsilon_p \sim \mathcal{N}\left(0, \sigma_\epsilon^2 I_{n_p}\right). \tag{7}$$

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 log sale prices including property and spatial random 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 this section and do not, for example, consider the widely used spatial (spatiotemporal) autoregressive models (Pace, Clapp, & Rodriquez, 1998; Pace, Sirmans, & Slawson, 2002). Spatial-temporal autoregressive models have been used in recent literature, but most applications have been on residential properties. Some commercial real estate examples are Tu, Yu, and Sun (2004), Nappi-Choulet and Maury (2009), and Chegut, Eichholtz, and Rodrigues (2015), all focusing on price indexes. For an extensive overview of spatial *HPM*s, see Anselin and Lozano-Gracia (2009).

2.4.1 | Besag model

Intrinsic and conditional autoregressions were introduced by Besag (1974), and later extended by Besag, York, and Mollié (1991) and Besag and Kooperberg (1995). These models are examples of Gaussian Markov random fields (Lindgren, Rue, & Lindström, 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.[3] 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

$$\theta_p | \theta_{-p}, \sigma_\theta^2 \sim \mathcal{N}\left(\frac{\sum_{q\in\partial_p} w_{p,q}\theta_q}{m_p}, \frac{\sigma_\theta^2}{m_p}\right), \tag{8}$$

where θ_{-p} is the vector of spatial property effects excluding property p. From the right-hand side of Equation (8), 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.

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

Note that the unconditional joint distribution of θ is not proper, even in case $m_p > 0$, the rank of the precision matrix is only positive semidefinite (see Gelfand & Vounatsou, 2003). A proper specification is obtained by adding a positive parameter d to the denominator, giving

$$\theta_p|\theta_{-p}, \sigma_\theta^2, d \sim \mathcal{N}\left(\frac{\sum_{q\in\partial_p} w_{p,q}\theta_q}{d + m_p}, \frac{\sigma_\theta^2}{d + m_p}\right). \tag{9}$$

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

2.4.2 | 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 algorithms solving the *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 (Gutin, Yeo, & Zverovich, 2002; Lawler et al., 1985). There are multiple *TSP*-algorithms to be found in the literature. We use eight different versions: (a) nearest neighbor algorithm, (b) insertion algorithm, (c) nearest insertion, (d) farthest insertion, (e) cheapest insertion, (f) arbitrary insertion, (g) k-opt heuristics, and (h) the Lin–Kernighan heuristic. For more information on these different *TSP*-algorithms, see Lawler et al. (1985) and Hahsler and Hornik (2007). Subsequently, we pick the version that 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 should not affect the results too much, as the *route* remains similar, irrespective of the starting point.

In our application, see Section 3, we "only" have about 2,000 observations for each of our two markets. Solving all of the above-mentioned *TSP*-algorithms is therefore not a computational issue. However, for large dat sets, the time required to solve some *TSP*-algorithms becomes infeasible. In our study, we find that the nearest neighbor and arbitrary insertion algorithms finish within a second. The other algorithms take between 30 and 60 s. Bentley (1992) gives computation times for some *TSP*-algorithms for large data problems. Even in this relatively old paper, Bentley solves the nearest neighbor algorithm for 1M observations within 10 min.

In the second step, we use a structural time series specification for 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 in a spatial setting. In this application, we keep the specification simple, and use a random walk, 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 (Van de Minne, Francke, Geltner, & White, 2020), taking into account distances between properties, could also be applied, potentially improving model fit, but we leave this for future research.

The spatial random walk specification is given by

$$\theta_{(p)} \sim \mathcal{N}\left(\theta_{(p-1)}, \sigma_\theta^2\right), \tag{10}$$

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 Equation (8). In the spatial random walk, the neighbors of property p—denoted by ∂_p in Equation (8)—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 (as is the case in this paper), we find that estimation time itself is reduced considerably as well, especially when using Markov chained Monte Carlo (*MCMC*) algorithms. In fact, we find that estimating the spatial random walk instead of the Besag model using our relative small data set with *MCMC* procedures decreases computation time 20-fold.[4] The gain in computation time is less profound when using Laplace approximation. However, we still find a 25% computation time decrease when using the spatial random walk. Computation time differences become larger when the number of observations increases. Also note that the spatial random walk can be estimated using the Kalman filter, reducing computation time even more. An obvious disadvantage of the two-step approach is that we reduce a two-dimensional plane into a one-dimensional line, at the risk of ignoring important information.

2.5 | Leave-one-out cross-validation and estimation

We do a full *LOO* analysis 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 the *HPM*s in Table 2. We redo this analysis for every observation, so N times. By simply subtracting our predicted value from the actual log sale 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 Bayesian estimation procedure, the Integrated Nested Laplace Approximation (*INLA*, Rue et al., 2009) for all *HPM*s.[5] In essence, *INLA* computes an approximation to the posterior marginal distribution of the hyperparameters. Operationally, *INLA* proceeds by first exploring the marginal joint posterior for the hyperparameters 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 hyperparameter, 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 specify noninformative (flat) priors for all hyperparameters.

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

[5] The *HPM*s excluding spatial random effects could be estimated by less sophisticated methods, but for consistency, we estimate all *HPM*s by the same method.

TABLE 1 Descriptive statistics

	mean	sd	min	max
	Los Angeles			
Sales price ($)[a]	6,389,494	6,460,729	1,550,000	64,250,000
Net Operating Income ($)	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 (€)[a]	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
Semidetached (1)	0.03		0	1
Semidetached (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			

Note. *R* gives the reference categories in our model. Semidetached (1) are the properties that are connected via a garage, and semidetached (2) are the properties that are connected wall-to-wall.

[a] Estimates for Moran's I (sales prices) for Los Angeles and Heemstede are, respectively, +0.04 and +0.23.

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, and captures approximately 90% of all commercial property transactions in the United States over $2.5 million. The database contains 2,263 prefiltered transactions, of which 1,936 are unique properties, in the period 2001–2017. The annual number of transactions is 140 transactions on average. 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 1 provides some descriptive statistics.

TABLE 2 Overview of model specifications

Model	Specification	Spatial effect (θ_p)
Standard	Equation (1): $y_p = x_p^*\beta + \epsilon_p$	
Hybrid	Equation (6): $\begin{pmatrix} y_p^S \\ \Delta y_p^R \end{pmatrix} = \begin{pmatrix} x_p^{*S} \\ \Delta x_p^{*R} \end{pmatrix}\beta + \begin{pmatrix} \epsilon_p^S \\ \Delta\epsilon_p^R \end{pmatrix}, \forall p : \begin{pmatrix} n_p = 1 \\ n_p > 1 \end{pmatrix}$	
RE	Equation (2): $y_p = x_p^*\beta + j_{n_p}\phi_p + \epsilon_p$	
Besag(Hybrid)	$\begin{pmatrix} y_p^S \\ \Delta y_p^R \end{pmatrix} = \begin{pmatrix} x_p^S \\ \Delta x_p^R \end{pmatrix}\beta + \begin{pmatrix} \theta_p \\ 0 \end{pmatrix} + \begin{pmatrix} \epsilon_p^S \\ \Delta\epsilon_p^R \end{pmatrix}, \forall p : \begin{pmatrix} n_p = 1 \\ n_p > 1 \end{pmatrix}$	Equation (9)
Besag(*RE*)	Equation (7): $y_p = x_p\beta + j_{n_p}\phi_p + j_{n_p}\theta_p + \epsilon_p$	Equation (9)
***SRW*(Hybrid)**	$\begin{pmatrix} y_p^S \\ \Delta y_p^R \end{pmatrix} = \begin{pmatrix} x_p^S \\ \Delta x_p^R \end{pmatrix}\beta + \begin{pmatrix} \theta_p \\ 0 \end{pmatrix} + \begin{pmatrix} \epsilon_p^S \\ \Delta\epsilon_p^R \end{pmatrix}, \forall p : \begin{pmatrix} n_p = 1 \\ n_p > 1 \end{pmatrix}$	Equation (10)
***SRW* (*RE*)**	Equation (7): $y_p = x_p\beta + j_{n_p}\phi_p + j_{n_p}\theta_p + \epsilon_p$	Equation (10)

Note. The models including property random effects—***RE***, **Besag(*RE*)** and ***SRW*(*RE*)**—all have the same property random effect specification given by $\phi \sim \mathcal{N}(0, \sigma_\phi^2 I_P)$, see Equation (3).

The spatial effect from the Besag model is provided by $\theta_p|\theta_{-p} \sim \mathcal{N}\left(\frac{\sum_{q\in\partial_p} w_{p,q}\theta_q}{d+m_p}, \frac{\sigma_\theta^2}{d+m_p}\right)$, see Equation (9).

The spatial effect from the spatial random walk model is provided by $\theta_{(p)} \sim \mathcal{N}(\theta_{(p-1)}, \sigma_\theta^2)$, see Equation (10).

The matrix x consists of property characteristics and time fixed effects.

The models **Standard**, **Hybrid**, and ***RE*** have also location fixed effects. All other models do not. The location fixed effects have been added to x, denoted by x^*.

In the models **Besag(Hybrid)** and ***SRW*(Hybrid)**, spatial random effects have been added to the one only sales in the hybrid specification (Equation (6)).

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.

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 are affiliated to the *NVM*. The database contains residential real estate 2,262 transactions, of which 2,065 are unique properties, in the period 2001 to 2017 for the Dutch city of Heemstede. The majority of the transactions is one only sales (69%). The annual number of transactions is approximately 145 transactions on average. We observe the property subtype (row houses, corner house, two types of semidetached homes and detached), the age and size of the structure, the maintenance level (three groups from bad to good), the presence of a yard, latitude and longitude, and the transaction price. The lower panel of Table 1 provides some descriptive statistics. The average transaction price is about €485,000, the average size 151 m^2 (1,625 square feet), and the average age is 65 years. The largest numbers of properties are row houses (44%). The *NVM* distinguishes between two types of semidetached homes: (a) two properties are connected via a garage and (b) two properties are connected wall-to-wall. Most of the semidetached 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.[6] Almost all properties have a yard, and in only 6% of the transactions, this is not the case.

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

4 | RESULTS

In this section, we provide estimation results for seven different HPM specifications. All models have the natural logarithm of the transaction price as a dependent variable, and the natural logarithm of the size as one of the independent variables. In addition, we use the natural logarithm of the NOI per square foot as an 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 seven *HPM* specifications are

1. **Standard**: The standard *HPM* including location fixed effects: Equation (1).
2. **Hybrid**: The hybrid model including location fixed effects: Equation (6).
3. ***RE***: The property random effects *HPM* including location fixed effects: Equations (2) and (3).
4. **Besag(Hybrid)**: The hybrid model including spatial property effects, specified by the Besag model: Equations (6) and (9).
5. **Besag(*RE*)**: The property and spatial random effects *HPM*, where spatial effects are specified by the Besag model: Equations (3), (7), and (9).
6. ***SRW***(Hybrid): The hybrid model including spatial effects, specified by the spatial random walk model: Equations (6) and (10).
7. ***SRW***(***RE***): The property and spatial random effects *HPM*, where spatial effects are specified by the spatial random walk model: Equations (3), (7), and (10).

An overview of the model specifications is provided in Table 2. Note that only the standard, hybrid, and *RE* model include location fixed effects.[7] For Los Angeles, we have six 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 four locations, defined by the first four digits of the ZIP codes.

The remainder of this section is organized as follows. In the next subsection, we discusses the main estimation results. The following subsection provides summary statistics for the *LOO* cross-validation. The third subsection discusses the spatial effects, and finally, the fourth subsection gives some robustness checks.

4.1 | Estimation results

Tables 3 and 4 provide the posterior means of the coefficients and significance levels for Los Angeles and Heemstede, respectively.[8] 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 *NOI* in the model, which picks up a large part of the time variation (or the macroeconomic cycle). We first discuss the results for Los Angeles, and subsequently the results for Heemstede.

4.1.1 | 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

[7] Bourassa, Cantoni, and Hoesli (2007) advocate to use submarket fixed effects, defined by real estate agents.

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

TABLE 3 Posterior means and in-sample fit statistics for Los Angeles

				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***
$\ln\left(\frac{NOI}{\text{Size}}\right)$	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

Note. The model specifications are provided in Table 2. The omitted dummy variable is garden apartment (for property subtype) and 2001 (for time of sale). *NOI* stands for net operating income.

Moran's I is a measure for spatial autocorrelation. *DIC* denotes deviance information criterion, and *WAIC* Watanabe information criterion.

***means the parameter is statistically significantly different from 0 at the 1% level, **at the 5% level, and *at the 10% level.

estate studies find this "law of diminishing returns" (Bokhari & Geltner, 2018). 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* models, the coefficient becomes statistically significant and negative, which might indicate an interaction between property type and location, which the location fixed effects 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 coefficient

TABLE 4 Posterior means and in-sample fit statistics for Heemstede

				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***
Semidetached (1)	0.148***	0.162***	0.137***	0.083***	0.074***	0.088***	0.073***
Semidetached (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

Note. The model specifications are provided in Table 2. The omitted dummy variables are row house (for property subtype), maintenance [bad], and having no yard. Semidetached (1) are houses connected by a garage, and semidetached (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 statistically significantly different from 0 at the 1% level, **at the 5% level, and *at the 10% level.

and the square of age a positive coefficient. This confirms expectations, as depreciation is fastest when a property is young, see Bokhari and Geltner (2018). 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 & Geltner, 2018; Geltner & van de Minne, 2017). As such, most depreciation is "captured" by the *NOI* variable.

The year 2001 is the omitted category, and therefore, the reference year. The estimates of the year fixed effects for the standard, hybrid, and *RE* models are always smaller than that of the other models (especially compared to the *SRW* models). This is explained by the differences in the estimate for *NOI* per square foot, which considerably differ between models. Given that *NOI* also "captures" changes in the macroeconomic environment, this is 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 fixed effects and vice versa. Note that the crisis and subsequent recovery are still clearly visible in all models. However, the timing differs. The trough of the time fixed effects is in 2010 for the standard, hybrid, *RE*, and Besag models. The trough is a full year earlier for the other models.[9]

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 models, and it reduces to 0.04 in the *SRW* models. 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 *SRW* models, 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 nonspatial models: The standard, hybrid, and *RE* model. The *WAIC*[10] of the Hybrid model is actually higher compared to 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 and spatial random have the lowest *WAIC*. The best performing model is *SRW*(*RE*).

4.1.2 | Heemstede

Compared to row houses, detached houses are valued the highest, followed by semidetached 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 .69 and .90, depending on the 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 80 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, e.g., Coulson & McMillen, 2008; Francke & van de Minne, 2017b; Wilhelmsson, 2008), 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 *SRW* models (however, the age squared term is significant).

[9] Although it should be noted that 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.

[10] The Watanabe–Akaike or widely applicable information criterion (*WAIC*, Watanabe, 2010) is based on the series expansion of LOO cross-validation. *WAIC* can be viewed as an improvement of the deviance information criterion (*DIC*, Spiegelhalter, Best, Carlin, & van der Linde, 2002). Lower values indicate a better model-fit.

TABLE 5 *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

Note. The table provides summary statistics of the leave-one-out residuals. The model specifications are provided in Table 2 .

Since we have no variables that move with the economic cycle (like *NOI*) in Heemstede, the coefficients of the 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 effects have the lowest *WAIC*. The *WAIC* of the best performing model, Besag(*RE*), is 643 points lower compared to the *RE* model.

4.2 | Leave-one-out cross-validation

Table 5 provides the results of the LOO 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 random effects 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 random effects 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.

TABLE 6 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

Note. The model specifications are provided in Table 2.

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 *WAIC*s over different data sets.

Table 6 provides the absolute mean of the *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 *HPM*, 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 one sale, to 0.139, when having three sales per property (−6.1%), and in Heemstede from 0.139, when having one sale, to 0.120, when having three sales per property (−13.8%).

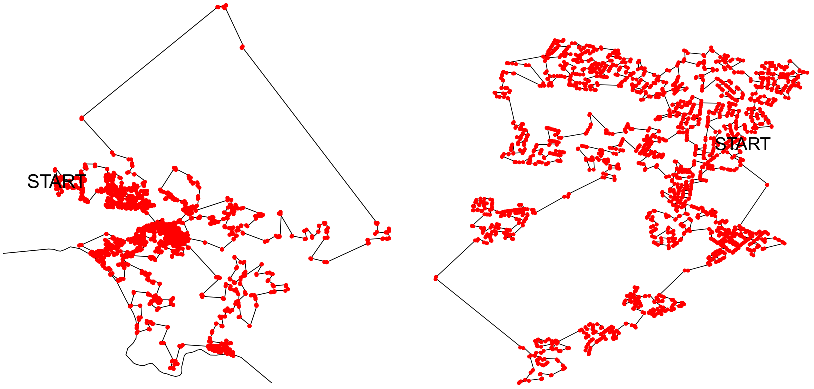
Note that the reduction is much higher for *RE* models. In Los Angeles, it goes down from 0.149, when having only one sale, to 0.112, when having three sales per property (−24.8%), and in Heemstede from 0.139, when having one sale, to 0.091, when having three sales per property (−35.0%). The property random effects *HPM* 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 two. In fact, the hybrid model performs equal to the standard model with just two sales. This is expected, given that the hybrid model can only get property-level estimates if the number of sales is three or more (because one observation is lost in the *LOO* analysis).

When having only one sale per property (zero during the *LOO* analysis), the property random effects *HPM* including spatial effects performs better than the model excluding the spatial effects. The difference in performance 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 random effects.

4.3 | Spatial effects

Figure 1 gives the TSP routes for both Los Angeles and Heemstede. The (random) starting point is given in the figure as well. In both cases, the TSP route goes clockwise. Figure 2 provides the size of the spatial effects θ along this route for the Besag and *SRW* models. Higher (lower) values for the spatial random effect means that the property has a higher (lower) value than can be explained by just



(a) Los Angeles.

(b) Heemstede.

FIGURE 1 Estimated *TSP*-route [Color figure can be viewed at wileyonlinelibrary.com]
Note. The dots represents individual properties in the data. From the starting point (START), follow the lines clockwise for the actual *TSP*-route. North is up, and note that the Pacific coast line is visible in the South in Figure 1a. The model specifications are provided in Table 2.

the covariates. It is already clear from Figure 2 that there is substantial variability in the spatial random effects. For more details on the spatial distribution of these random effects, see the online Appendix.

Table 7 provides 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 property, after correction for differences in property characteristics and *NOI*. In Heemstede, the difference between the 2.5% and 97.5% percentile of θ has similar magnitude, 0.615, corresponding to a 85% difference between the cheapest and most expensive property. The estimated spatial effects θ are positively correlated among the models. Correlations range between .93 and .99 in Los Angeles, and between .88 and .99 in Heemstede.

4.4 | Robustness check

In this subsection, we perform a simple robustness check. In Heemstede, we omit the level of maintenance and the property-type dummy variables as explanatory variables and rerun both the standard *HPM* and the spatial random walk with property random effects, the *SRW*(*RE*) model. Our basic interest is to compare the *SRW*(*RE*) model on the reduced data set with the standard *HPM* with all variables included. This can learn us something on how effectively the spatial and property random effects control for omitted variables/unobserved heterogeneity.

We do something similar for the commercial properties. It is well known that *NOI* explains a large part of prices, where higher *NOI* results in higher prices (Kok, Koponen, & Martínez-Barbosa, 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 *HPM* and repeat the estimation and *LOO* analysis. A summary of robustness checks is given in Tables 8 and 9.

Overall, the results are in line with our earlier findings. The *SRW*(*RE*) model outperforms the standard *HPM* to a large extent on the same set of characteristics, and in the *SRW*(*RE*) model, the fit is better for properties that transacted more often. Also unsurprisingly is that omitting explanatory variables deteriorates the model fit considerably. The standard deviation of the *LOO* residual increases

TABLE 7 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)		.988	.935	.933
Besag(*RE*)			.942	.948
SRW(Hybrid)				.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)		.975	.896	.880
Besag(*RE*)			.904	.908
SRW(Hybrid)				.991

Note. The model specifications are provided in Table 2.

TABLE 8 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

Note. The results with all variables can also be found in Tables 3 and 4. For Los Angeles, the reduced data set does not include (log of) net operating income per square foot. To create the reduced data set for Heemstede, we omit the variables on property types and maintenance levels.

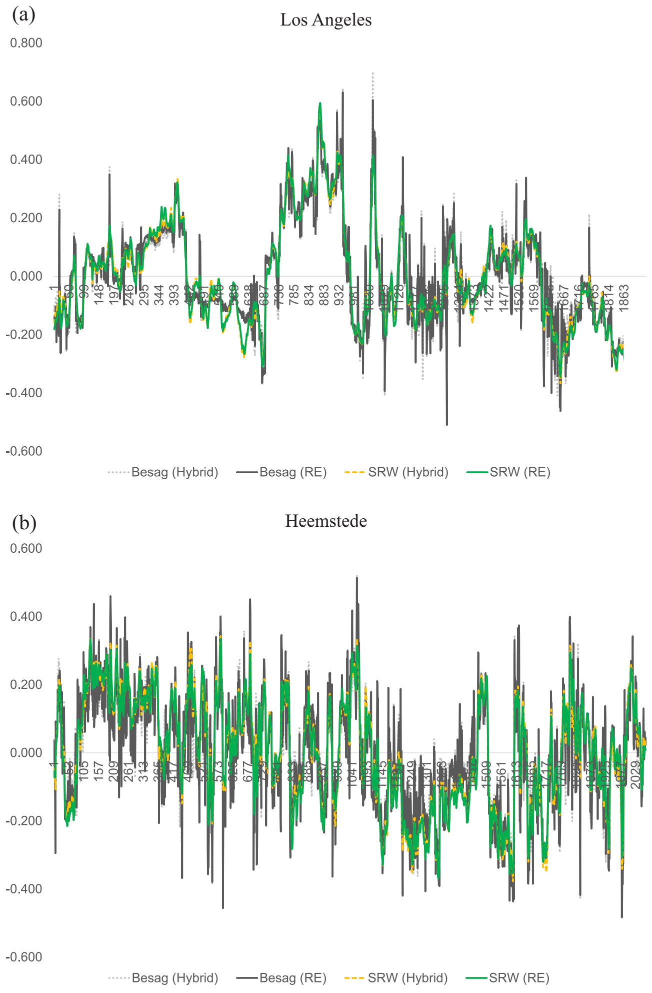


FIGURE 2 Spatial random effects θ values over the *TSP*-route [Color figure can be viewed at wileyonlinelibrary.com]

Note. The estimated *TSP*-route can be found in Figure 1. The horizontal axis gives the numbered property (p) along the *TSP*-route. The model specifications are provided in Table 2.

with almost 20% in Heemstede and even with 50% in Los Angeles, for both the standard and the *SRW*(*RE*) model.

However, our main interest is in comparing the performance of the *SRW*(*RE*) model on the reduced data set to the standard *HPM* 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 8 as the *LOO* residuals in Table 9 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 *HPM* including *NOI* as an 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 8) 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

TABLE 9 Absolute mean of *LOO* residuals as a function of the number of sales per property for the robustness check

# Sales	All variables		Reduced data set		
per property	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	322
3	0.120	0.093	0.130	0.087	39
Total	0.135	0.104	0.159	0.122	2,468

Note. The results with all variables can also be found in Tables 3 and 4 . For Los Angeles, the reduced data set does not include (log of) net operating income per square foot. To create the reduced data set for Heemstede, we omit the variables on property types and maintenance levels.

standard model over the *SRW*(*RE*) model. The average absolute *LOO* residuals in Table 9 also give an inconsistent picture. For properties that sold only once, the standard model including *NOI* as an 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 property and spatial random effects *HPM* excluding *NOI* data performs.

5 | CONCLUSION

This paper studies unobserved heterogeneity in HPMs, 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 an HPM 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 HPM. The standard HPM 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 HPM including property random effects also outperforms the related hybrid hedonic-repeat sales 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 and new 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 two 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.

ACKNOWLEDGMENTS

We would like to thank David Geltner, Martijn Dröes, two anonymous referees, and participants of the 2018 International AREUEA conference, the May 2018 research seminar of the Weimer School of Advanced Studies in Real Estate and Land Economics, the October 2018 research seminar at Laval University and the December 2019 research seminar at Maastricht University, for their valuable input. All remaining errors are ours. We also thank Real Capital Analytics Inc and the Dutch Association of Real Estate Brokers and Real Estate Experts for providing us with the data that made this research possible.

ORCID

Marc Francke https://orcid.org/0000-0002-7239-4868

REFERENCES

Anselin, L., & Lozano-Gracia, N. (2009). Spatial hedonic models. In T. C. Mills and K. Patterson (Eds.), *Palgrave handbook of econometrics*, *volume 2, applied econometrics* (pp. 1213–1250). London: Palgrave MacMillan.

Arellano, M. (2003). *Panel data econometrics*. Advanced Texts in Econometrics. Oxford: Oxford University Press.

Basu, S., & Thibodeau, T. G. (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*(4), 387–411.

Berndt, E. R., Griliches, Z., & Rappaport, N. J. (1995). Econometric estimates of price indexes for personal computers in the 1990's. *Journal of Econometrics*, *68*(1), 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., & Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, *82*(4), 733–746.

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

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

Bokhari, S., & Geltner, D. M. (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., & Geltner, D. M. (2018). Characteristics of depreciation in commercial and multifamily property: An investment perspective. *Real Estate Economics*, *46*(4), 745–782.

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

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

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

Clapp, J. M., & Zhou, T. (2019). *Controlling unobserved heterogeneity in repeat sales models: Application to anchoring to purchase price*. Technical report, University of Connecticut and Florida State University, https://ssrn.com/abstract=3358401.

Coulson, E. N., & McMillen, D. (2008). Estimating time, age and vintage effects in housing prices. *Journal of Housing Economics*, *17*(2), 138–151.

Court, A. T. (1939). Hedonic price indexes with automotive examples. In *The dynamics of automobile demand* (pp. 99–117). New York: General Motors Corporation.

de Haan, J., & Diewert, W. E. (2013). Hedonic regression methods. In *Balk, B. and de Haan, J. and Diewert, E. Handbook on residential property price indexes* (pp. 50–64). Luxembourg: Eurostat.

Diewert, W. E., de Haan, J., & Hendriks, R. (2015). Hedonic regressions and the decomposition of a house price index into land and structure components. *Econometric Reviews*, *34*(1–2), 106–126.

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

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

Gelfand, A. E., & Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, *4*(1), 11–25.

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

Gormley, T. A., & Matsa, D. A. (2013). Common errors: How to (and not to) control for unobserved heterogeneity. *The Review of Financial Studies*, *27*(2), 617–661.

Greene, W. H. (2008). *Econometric analysis, 6/E*. Upper Saddle River, NJ: Prentice Hall.

Gutin, G., Yeo, A., & Zverovich, A. (2002). Traveling salesman should not be greedy: Domination analysis of greedy-type heuristics for the TSP. *Discrete Applied Mathematics*, *117*(1), 81–86.

Hahsler, M., & Hornik, K. (2007). TSP-infrastructure for the traveling salesperson problem. *Journal of Statistical Software*, *23*(2), 1–21.

Harvey, A. (1989). *Forecasting structural time series models and the Kalman filter*. Cambridge: Cambridge University Press.

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., & Gelman, A. (2014). The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hwang, M., & Quigley, J. M. (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., & Sirmans, C. F. (1996). Depreciation, maintenance, and housing prices. *Journal of Housing Economics*, *5*(4), 369–389.

Kok, N., Koponen, E., & Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. *The Journal of Portfolio Management*, *43*(6), 202–211.

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

Lindgren, F., Rue, H., & Lindström, J. (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*(4), 423–498.

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

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

Pace, R. K., Sirmans, C. F., & Slawson Jr, V. C. (2002). Automated valution models. In K. Wang and M. L. Wolverton (Eds.), *Real estate valuation theory, research issues in real estate* (Vol. 8, pp. 133–156). Boston: 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), 1–12.

Reis, H. J., & Santos Silva, J. M. C. (2006). Hedonic prices indexes for new passenger cars in Portugal (1997–2001). *Economic Modelling*, *23*(6), 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., Martino, S., & Chopin, N. (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*(2), 319–392.

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

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

Van de Minne, A. M., Francke, M. K., Geltner, D. M., & White, R. (2020). Using revisions as a measure of price index quality in repeat-sales models. *The Journal of Real Estate Finance and Economics*, *60*(4), 514–553.

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*(1), 88–101.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Francke M, Van de Minne A. Modeling unobserved heterogeneity in hedonic price models. *Real Estate Econ*. 2021;49:1315–1339. https://doi.org/10.1111/1540-6229.12320

APPENDIX: ESTIMATION OF THE RANDOM EFFECTS HEDONIC PRICE MODEL

This appendix provides a derivation of the estimation of β and ϕ in the property random effects HPM, given by of Equations (2) and (3). A stacked representation is provided by

$$y = X\beta + D\phi + \epsilon, \epsilon \sim \mathcal{N}(0, \Sigma), \tag{A1}$$

$$\phi \sim \mathcal{N}(0, \Phi) \tag{A2}$$

where D is a selection matrix, where each row contains exactly one 1 to select the appropriate property, the remaining row elements are 0. Equation (A2) can be seen as a prior distribution of the coefficients

ϕ. Φ and Σ are nonsingular variance matrices. In Equations (2) and (3), it holds that both variance matrices are a multiple of the identity matrix, $\Sigma = \sigma_\epsilon^2 I_N$ and $\Phi = \sigma_\phi^2 I_P$.

Conditional on (Σ, Φ) estimates of (ϕ, β) are given by

$$\begin{pmatrix} \hat{\phi} \\ \hat{\beta} \end{pmatrix} = \begin{bmatrix} D'\Sigma^{-1}D + \Psi^{-1} & D'\Sigma^{-1}X \\ X'\Sigma^{-1}D & X'\Sigma^{-1}X \end{bmatrix}^{-1} \begin{pmatrix} D'\Sigma^{-1}y \\ D'\Sigma^{-1}X \end{pmatrix}. \tag{A3}$$

By using the matrix inverse lemma for block matrices, one can derive from Equation (A3) that

$$\hat{\beta} = \mathrm{Var}(\hat{\beta})X'(D\Phi D' + \Sigma)^{-1}y, \tag{A4}$$

$$\mathrm{Var}(\hat{\beta}) = (X'(D\Phi D' + \Sigma)^{-1}X)^{-1}, \tag{A5}$$

where $D\Phi D' + \Sigma$ is the variance of $(D\phi + \epsilon)$.

An expression for $\hat{\phi}$ can be derived by observing that (by premultiplying the left- and right-hand side in Equation (A3) with the matrix between square brackets) $(D'\Sigma^{-1}D + \Psi^{-1})\hat{\phi} + D'\Sigma^{-1}X\hat{\beta} = D'\Sigma^{-1}y$, leading to

$$\hat{\phi} = (D'\Sigma^{-1}D + \Psi^{-1})^{-1}(D'\Sigma^{-1}y - D'\Sigma^{-1}X\hat{\beta}). \tag{A6}$$

$\mathrm{Var}(\hat{\phi})$ can also be derived by applying the matrix inverse lemma for block matrices in Equation (A3):

$$\mathrm{Var}(\hat{\phi}) = \Lambda^{-1} + \Lambda^{-1}(D'\Sigma^{-1}X)\mathrm{Var}(\hat{\beta})(X'\Sigma^{-1}D)\Lambda^{-1}, \tag{A7}$$

where $\Lambda = D'\Sigma^{-1}D + \Psi^{-1}$ is the upper left matrix in Equation (A3).

Substituting Σ by $\sigma_\epsilon^2 I_N$ and Φ by $\sigma_\phi^2 I_P$ in Equations (A4)–(A7) gives the estimates provided in Equations (4) and (5).