Weekly Hedonic House Price Indices: An Imputation Approach from a Spatio-Temporal Model

Robert J. Hill*

Department of Economics, University of Graz, Universittsstr. 15/F4, 8010 Graz. Austria

Alicia N. Rambaldi†

School of Economics, The University of Queensland, St Lucia, QLD 4072. Australia

Michael Scholz

Department of Economics, University of Graz, Universittsstr. 15/F4, 8010 Graz. Austria

DRAFT VERSION: 27 April 2017.

Abstract

Since the global financial crisis there is an increased demand for timely house price indices. The aim of this paper is to develop a method for computing house price indices at a weekly frequency using the hedonic imputation method. The hedonic imputation method provides a flexible way of constructing quality-adjusted house price indices using a matching sample approach. At annual frequencies the implementation of the hedonic imputation approach typically entails estimating the hedonic model period-by-period and then using the parameter estimates (i.e., characteristics shadow prices) to obtain the required imputed house prices. Once these imputed prices are available for a matched sample, standard price index formulas (e.g., Laspeyres, Fisher or Törnqvist) can be used to compute the overall price index. A common approach to control for location in hedonic models has been to include postcode dummies. This may not be feasible at higher frequencies as there may not be enough observations for each postcode and small samples might cause large variability in the shadow price parameters when estimated period-by-period. We develop a spatio-temporal model to obtain the

---

*This project has benefited from funding from the Austrian National Bank (Jubilumsfondsprojekt 14947). We thank Australian Property Monitors for supplying the data.

†Corresponding author: A.N.Rambaldi (+61(0)7 3365 6576), a.rambaldi@uq.edu.au

imputed prices. A geospatial spline surface controls for location and is embedded in a state-space formulation that controls for trends and property quality. The advantage is that the model is parsimonious and shadow price parameters are connected over time while retaining the property that values are not revised as new time periods are added to the data set. We show the spatio-temporal specification leads to a modified form of the Kalman filter and a Goldberger's adjusted form of the predictor to obtain the imputations. Using a recently developed measure of index performance and applying this hedonic geospatial spline/Kalman filter approach to data for Sydney (Australia) we show that it outperforms competing alternatives for computing house price indices at a weekly frequency. Furthermore, we show that weekly house price indices are much more sensitive than annual or quarterly indices to the choice of hedonic method. Hence the choice of hedonic method is of greater practical significance for weekly indices. (***JEL***. C33; C43; E01; E31; R31)

Keywords: Housing market; House Price index; Hedonic imputation; Geospatial data; Spline; Quality adjustment; State Space Models.

1 Introduction

Since the global financial crisis there is an increased awareness of the importance of the housing market to the broader economy. Hence there is a growing demand from central banks, governments, banks, real estate developers, and households for reliable and more timely house price indices. Silver (2011) shows that house price indices however can be quite sensitive to how they are constructed. This is especially true for higher frequency (e.g., weekly) indices. It is essential that weekly house price indices are quality adjusted, since differences in the sample composition each week will cause a simple median or mean index to be highly volatile. Quality-adjusted indices are typically computed using either hedonic or repeat-sales methods. The latter are more common in the US – the best known example being the Standard and Poors' Case-Shiller (SPCS) indices. In Europe, by contrast, hedonic methods are more widely used. For example the national statistical institutes (NSIs) of most member countries of the European Union now compute an official House Price Index (HPI) at a quarterly frequency using hedonic methods (Eurostat, 2016). One reason for this difference is that repeat-sales methods tend to work better when the frequency of transactions (i.e., turnover) is high as it is in the US. In Europe by contrast turnover is generally much lower. Elsewhere in the world, it is less clear which approach is preferred. CoreLogic for example computes both hedonic and repeat-sales indices for Australian cities.

The increased availability of housing data and advances in computing power and econometric techniques offer new opportunities for constructing higher frequency quality-adjusted indices, and for deepening our knowledge of the real estate asset class. Bokhari and Geltner (2012) give further reasons for the usefulness of higher frequency indices:

> "[T]he greater utility of higher frequency indices has recently come to the fore with the advent of tradable derivatives based on real estate price indices. Tradability increases the value of frequent, up-to-date information about market movements, because the lower transactions and management costs of synthetic investment via index derivatives compared to direct cash investment in physical property allows profit to be made at higher frequency based on the market movements tracked by the index. Higher-frequency indices also allow more frequent 'marking' of the value of derivatives contracts, which in turn allows smaller margin requirements, which increases the utility of the derivatives."

Both hedonic and repeat-sales indices however become more problematic at higher frequencies. The construction of higher frequency repeat-sales indices is considered by Bollerslev, Patton, and Wang (2015), and Bourassa and Hoesli (2016). Bokhari and Geltner (2012) propose a two-step procedure based on a generalised inverse estimator that improves the accuracy of high-frequency indices in scarce data environment (in an application to commercial property repeat-sales data). In recent work, Bourassa and Hoesli (2016) apply the procedure of Bokhari and Geltner (2012) and construct high frequency house price indices for both cities and submarkets within cities. Bollerslev, Patton, and Wang (2015) develop daily house price indices for 10 major US metropolitan areas. Their calculations are based on a database of several million residential property transactions and a standard repeat-sales method that closely mimics the methodology of the monthly SPCS house price index. Bollerslev, Patton, and Wang (2015) use a multivariate time series model to compute daily house price index returns, explicitly allowing for commonalities across cities and GARCH effects.

Here we focus on hedonic indices, since they have the potential to especially benefit from improvements in housing data and computing power. More specifically we focus here on the hedonic imputation method (see Hill (2013) for a taxonomy of hedonic methods for computing house price indices). The hedonic imputation method provides a flexible way of constructing quality-adjusted house price indices. An estimated hedonic model is used to obtain predictions of the sale price for each dwelling at two comparison periods which provide imputed price relatives that enter the index formula. These are then averaged to obtain the overall price index (this is formally presented in Section 2).

Hedonic imputation was first proposed by Court (1939), more formally developed as a method by Griliches (1961, 1971), and it is recommended as the superior method to construct price indices for residential property in the Handbook on Residential Property Price Indices, which was written as a joint project (see European Commission, Eurostat, OECD, and World Bank (2013)). In practice it is typically implemented by estimating a separate hedonic model for each period which includes dummies to control for location (e.g., using zip or post codes) in addition to other hedonic characteristics of the dwelling. Using the estimated hedonic models price relatives are imputed for each dwelling. A problem with such an approach is that the method can become unreliable at higher frequencies (e.g., weekly indices), since then even in large data sets there may not be enough price observations in each period to satisfactorily estimate the hedonic model. As a consequence computational and statistical problems occur (e.g., no observations for some postcodes, a loss in degrees of freedom, or an increased variance of estimated parameters). Geltner and Ling (2006) describe the trade-off between statistical quality per period and the frequency of index reporting, holding constant the overall quantity and quality of raw valuation data and index construction methodology. They conclude that the usefulness of an index for research purposes clearly increases the greater the frequency of reporting, holding statistical quality (per period) constant (Bokhari and Geltner, 2012).

In this article we show how the reliability of weekly hedonic indices can be improved by replacing postcode dummies by a geospatial spline and then using time-varying hedonic model in state-space form. This approach has two advantages. First, the dimensionality of the model is reduced. Replacing postcode dummies by values from the geospatial spline function at each location in the data set very significantly reduces the number of parameters that need to be estimated. Second, the small number of observations in each period causes larger variability in the estimated parameters (shadow prices) obtained from the weekly hedonic model. Estimation of a state space model with the Kalman filter interconnects those parameters over time and optimally weighs the past market information as shown by Rambaldi and Fletcher (2014).

We consider a basic hedonic imputation method that uses a geospatial spline to control for locational effects but no state-space model, with a state-space model that controls for location using postcodes, and a state-space model that incorporate a geospatial spline.

Using a recently developed criterion proposed by Hill and Scholz (2017) we compare the performance of our indices data for Sydney (Australia) over the period 2003–2014. This criterion focuses on comparing the imputed price relatives that form the basic building blocks of the hedonic imputation price index with their corresponding actual repeat-sales price relatives. Based on this criterion, we find that the index obtained by combining a state-space model with a geospatial spline outperforms the indices obtained by the simpler hedonic models. Furthermore, we find that the results are quite sensitive to the choice of method, far more than they would be if the indices were computed at an annual or quarterly frequency.

The remainder of this paper is structured as follows. Section 2 provides an overview of the hedonic imputation method, the econometric model and the methods for estimation of the models (a generalized additive model and the Kalman-Filter), and the criterion used to compare the performance of competing hedonic imputed indices. Section 3 presents our data set, the empirical study and the results of our analysis. Section 4 concludes by considering some implications of our findings and gives a short outlook for further research. Some technical details regarding the estimation procedures and the data set are discussed in the Appendix.

2 Hedonic Imputation and Index Quality

2.1 Index Definition

Hedonic price indices for housing are typically constructed using one of the time-dummy, hedonic imputation, and average characteristic methods (Diewert, 2010; Hill, 2013; European Commission, Eurostat, OECD, and World Bank, 2013). All of them have in common that in a hedonic model the price of a product is regressed on a vector of characteristics (whose prices are not independently observed). The hedonic equation is a reduced form that is determined by the interaction of supply and demand. Hedonic models are used to construct quality-adjusted price indices in markets (such as computers) where the products available differ significantly from one period to the next. Housing is an extreme case in that every house is different.

Here we focus on the hedonic imputation method since it is more flexible than either the time-dummy or average characteristics methods. The hedonic imputation method uses the predictions from a hedonic model to impute prices which can be inserted into standard price index formulas. Let $x'_{t,h}$ be a vector of characteristics associated with property h sold in period t, and $\hat{p}_{t+1,h}(x'_{t,h})$ as the imputed price for that property had it sold in period $t+1$. The model used in this study to produce these predictions is presented in the next section. To obtain a hedonic imputed price index comparing periods t and $t+1$, we use a *Laspeyres*-type formula that focuses on the properties sold in the earlier period t, and a *Paasche*-type formula that focuses on the properties sold in the later period $t+1$. Our price indices are constructed by taking the geometric mean of the price relatives, giving equal weight to each house.[1] Taking a geometric mean of the Laspeyres and Paasche-type indices, we obtain a Törnqvist-type index, that has the advantage that it treats both periods symmetrically and is consistent with a log-price hedonic model (Hill and Melser, 2008).

The indices presented below are all of the double imputation type.[2] This means that both prices in each price relative are imputed. For example, the double imputation Laspeyres index (DIL), Paasche index (DIP), and Törnqvist index (DIT) are defined as follows:

$$P_{t,t+1}^{DIL} = \prod_{i=1}^{N_t} \left[\left(\frac{\hat{p}_{i,t+1}(x'_{i,t})}{\hat{p}_{i,t}(x'_{i,t})} \right)^{1/N_t} \right], \tag{1}$$

$$P_{t,t+1}^{DIP} = \prod_{h=1}^{N_{t+1}} \left[\left(\frac{\hat{p}_{i,t+1}(x'_{i,t+1})}{\hat{p}_{i,t}(x'_{i,t+1})} \right)^{1/N_{t+1}} \right], \tag{2}$$

$$P_{t,t+1}^{DIT} = \sqrt{P_{t,t+1}^{DIP} \times P_{t,t+1}^{DIL}} \tag{3}$$

where $i = 1, \ldots, N_t$ indices the dwellings sold in period t, and $i = 1, \ldots, N_{t+1}$ indices the dwellings sold in period $t+1$. The overall price index is then constructed by chaining together these bilateral comparisons between adjacent periods. As it will be discussed in the next section, the predictions used to compute the bilateral indices must take into account the spatio-temporal nature of our modelling approach.

2.2 The Model

The objective of the hedonic model is to provide predictions of the prices of properties included in the Törnqvist index calculation. The econometric model is a spatio-temporal hedonic model that combines elements from the work of Wikle and Cressie (1999)-WC and Rambaldi and Fletcher (2014)-RF. WC provide a temporally dynamic and spatially descriptive model and an efficient estimation algorithm designed to deal with a large

---

[1] This democratic weighting structure is in our opinion more appropriate in a housing context than weighting each house by its expenditure share. See Hill and Melser (2008), de Haan (2010) and Rambaldi and Fletcher (2014) for a discussion on alternative weighting schemes.

[2] Double imputation indices tend to be slightly more robust to omitted variables bias (Hill and Melser, 2008). We also calculated single imputation indices where only one price in each price relative is imputed. The results are virtually indistinguishable. Hence to save space we focus here only on double imputation indices.

scale spatio-temporal dataset. We adopt a similar modelling approach in that measurement error, location, property quality components, and a term that captures small scale spatial variability are incorporated. This term conceptually extends the spatio-temporal models proposed by Rambaldi and Fletcher (2014), where two parametric alternatives to model location are used. The model incorporates Hill and Scholz (2017)'s measure of location, obtained by estimating a geospatial spline surface within a semi-parametric framework using observed sales in each individual period. The periodwise estimation also provides a required measure of spatial variability.

We denote the observed (log transformed) price by $y_{it} = \ln price_{it}$. The objective is to predict y^*_{it}, a smoother but unobservable (log) price of property i in period t, for i in any location and over all time periods t, regardless of when and where the data are observed.

We write this model as

$$y_{it} \quad = \quad y^*_{it} + \epsilon_{it}; \epsilon_{it} \sim N(0, \sigma^2_\epsilon). \tag{4}$$

The random process ϵ_{it} is not correlated across location or time and captures overall measurement error.

At a given time period, τ, N_τ properties are sold, and y^*_τ is given by

$$y^*_\tau = x^\dagger_\tau + v_\tau; v_\tau \sim N(0, V_\tau)$$

where, v_τ is a random error that does not have a temporally dynamic structure but might have some spatial structure and thus V_τ might not be diagonal. It is assumed that $E(v_{i\tau}\epsilon_{jt}) = 0$ for all $i, j = 1, \ldots, N$ and $-\infty \leq t \leq \infty$.

$x^\dagger_t$ is assumed to evolve according to three components, trend, property quality and location,

$$x^\dagger_{it} = \mu_t + \sum_{k=1}^{K} \beta_{k,t} z_{k,it} + \gamma_t g_{it}(z_{long}, z_{lat})$$

where,

μ_t is a trend component common to all i in period t and captures overall macroeconomic conditions that affect all locations in the market under study;

$z_{k,it}$ is the kth hedonic characteristic from a set of K providing information on the type/quality of the property (e.g., number of bedrooms, bathrooms, size of the lot). These are not trending variables.

$g_{it}(z_{long}, z_{lat})$ is a measure of the location of property i defined on a continuous surface at time period t. It is not a trending function of time.

$\beta_{k,t}$ and γ_t are time-varying parameters to be estimated.

$E(z_k v_t) = 0$, $E(z_k \epsilon_t) = 0$ for all $k = 1, \ldots, K$, $E(g_{it} v_{jt}) = 0$, $E(g_{it}\epsilon_{jt}) = 0$, for all i, j.

Assuming an estimate of location, denoted by $\widehat{g}_{it}(z_{long}, z_{lat})$, is available (estimation is discussed in Section 2.4) then the model in (4) with above definitions can be written in familiar state-space representation

$$\begin{aligned} y_t &= X_t\alpha_t + v_t + \epsilon_t; \epsilon_t \sim N(0, H) \qquad (5) \\ \alpha_t &= D\alpha_{t-1} + \eta_t; \eta_t \sim N(0, Q) \qquad (6) \end{aligned}$$

where,

X_t is $N_t \times (K + 2)$ and with the ith row being $x'_{it} = \{1, z_{1,it}, \ldots, z_{K,it}, \hat{g}_{it}(z_{long}, z_{lat})\}$

y_t is the vector of log transformed observed prices of properties sold at t.

$H = \sigma^2_\epsilon I_{N_t}$

$\alpha_t = \{\mu_t, \beta_{1t}, \ldots, \beta_{K,t}, \gamma_t\}'$

$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & I_K & 0 \\ 0 & 0 & \rho \end{bmatrix}$; $0 \leq \rho \leq 1$; If $\rho < 1$ the estimate of γ_t is mean reverting. If $\rho = 1$, γ_t evolves as a random walk as do the other state parameters α_t.

$Q = \begin{bmatrix} \sigma^2_\mu & 0 & 0 \\ 0 & \sigma^2_\beta I_K & 0 \\ 0 & 0 & \sigma^2_g \end{bmatrix}$

The estimate of the location spline surface for property i sold in period t, $\hat{g}_{it}(z_{long}, z_{lat})$ is obtained non-parametrically at *each time period* using only those properties that have sold that period. This estimate enters the spatio-temporal model as a generated regressor and the parameter γ_t, in (5) and (6), provides the flexibility for the vector of location spline estimates of properties sold in period t, , $i = 1, \ldots, N_t$, to be shifted by temporal market information up to time t. The combination of spatial and temporal information leads to two unconventional features of this model, compared to one in a standard time-series setting, with consequences for the form of the Kalman filter algorithm as well as the price prediction to be used for the computation of the Törnqvist price index. First the error has two components, ϵ_t, the conventional overall measurement error, and v_t arising from predicting the (log) sale price using only the spatial variability within each time period. This results in the Kalman gain, G_t, which is a function of the sum of the two covariances $(H + V_t)$ under the assumptions already stated. The second is that the value of the location spline for property i sold in period t will not be identical in value if property i is priced in a different time period. That is, a given property has fixed location coordinates and hedonic characteristics; however, its location spline value, unlike the size of the land, will differ between period t and period $t+1$. We denote by $\hat{g}_{t(t)}(z_{long}, z_{lat})$ the vector of spline values for properties sold and priced in period t, and by $\hat{g}_{t(t-1)}(z_{long}, z_{lat})$ the vector of spline values for the set of properties sold in t when priced in $t-1$. The implications for the form of the Kalman filter algorithm are presented next.

Using the innovation form of the filter, the state at time t is given by

$$\alpha_{t|t} = \alpha_{t|t-1} + G_t\{y_t - X_t^1 \alpha_{t|t-1}\} \tag{7}$$

where, the prediction step of the Kalman filter uses X_t^1 which is the X_t matrix with the $\hat{g}_{i,t(t)}$ replaced by $\hat{g}_{i,t(t-1)}(z_{long}, z_{lat})$. This is necessary to obtain the conditional prediction error from a conditional prediction of the state. The mean square error matrix given information up to time period t is $P_{t|t}$,

$$P_{t|t} = P_{t|t-1} - G_t X_t P_{t|t-1} \tag{8}$$

The Kalman gain takes the form

$$G_t = P_{t|t-1} X_t' \{H + V_t + X_t P_{t|t-1} X_t'\}^{-1} \tag{9}$$

The updating equations are given by

$$\alpha_{t|t-1} = D\alpha_{t-1|t-1} \tag{10}$$

$$P_{t|t-1} = DP_{t-1|t-1}D' + Q \tag{11}$$

Estimates of the state (7), $\hat{\alpha}_{t|t}$, are obtained by replacing H, Q, D, and V_t, by suitable estimates.

Hill and Scholz (2017) use a period-by-period semi-parametric model to construct price indices. This model can provide two key estimates to our state-space model, the estimate of location, $\hat{g}_{t(t+j)}(z_{lat}, z_{long})$ for $j = -1, 0, 1$, and a prediction of the (log) of price for each property based only on the spatial information of properties sold in a given time period which will allow us to obtain an estimate of V_t.

For any period $\tau \in t = -\infty, \ldots, +\infty$ the semi-parametric model is given by

$$y_{i\tau} = \theta_{0\tau} + z_{it}'\theta_{\tau}^{\dagger} + g_{i,\tau}(z_{long}, z_{lat}) + v_{i\tau} \tag{12}$$

where,
$\theta_{\tau}^{\dagger} = \{\theta_{1\tau}, \ldots, \theta_{K,\tau}\}'$

Estimating (12) predictions of (log) prices, denoted by $\hat{x}_t^{\dagger}$, estimates of the parameters, $\hat{\theta}_t^{\dagger}$, and of the location spline $\hat{g}_t(z_{lat}, z_{long})$ can be obtained for each property in each period of the sample, $t = 1, \ldots, T$. Residuals and estimates of the location spline are required for for $j = -1, 0, 1$, $\hat{v}_{t(t+j)}$ and $\hat{g}_{t(t+j)}(z_{lat}, z_{long})$, respectively, to implement the predictions/imputations to construct the index (discussed in the next section). Details of the estimation of the semi-parametric model are provided in section 2.4.

2.3 Constructing the Predictions

The computation of the index, (3), depends crucially on the prediction of log price. In Appendix A1 we show that given the state vector at period t conditional on information up to time period t, the prediction of the log price for property h is given by the natural predictor plus a Goldberger's correction term (Goldberger (1962)) as follows,

$$\widehat{y^*_{t|t,h}} = x_{t,h}'\hat{\alpha}_{t|t} + c_{vt,h}'\Omega^{-1}e_t \tag{13}$$

where,
$\Omega = cov\{y_t, y_t\}$
$c_{vt,h}' = E(v_{ht}, v_t)$ is the row of V_t corresponding to property h and has elements $c_{v,hj} \equiv E\{v_{ht}v_{jt}\}$ which could be equal to zero for $h \neq j$.
$e_t = y_t - E(y_t)$

In this study we implement this prediction by defining $\widehat{v_t} = y_t - \hat{x}_t^{\dagger}$ and $e_t = y_t - \widehat{y_{t|t}}$, where $\widehat{y_{t|t}}$ is the state-space prediction of the (log) price of property h at time t, $\widehat{y_{t|t}} = X_t\hat{\alpha}_{t|t}$.

For the index calculation predictions and imputations are needed. The prediction of the price of property h sold in period $t = 1, \ldots, T$ is defined as

$$\hat{p}_{t,h}(z_{t,h}', \hat{g}_{h,t(t)}) = \exp(\widehat{y^*_{t|t,h}}) \tag{14}$$

The imputation of the price of property h sold in period t for period $t-1$ is given by

$$\hat{p}_{t-1,h}(z_{t,h}', \hat{g}_{h,t(t-1)}) = \exp(x_{t,h}^{1}{}'\hat{\alpha}_{t-1|t-1} + c_{v(t-1),h}'\Omega^{-1}e_{t(t-1)}) \tag{15}$$

The crucial point is that the constructed location effect and parameters need to be matched with the correct period that is being imputed. In this case, $\hat{g}_{t(t-1),h}$ enters in ${x^1_{t,h}}'$, $c'_{v(t-1),h}$ and together with $\hat{\alpha}_{t-1|t-1}$, $e_{t(t-1)}$.

2.4 Estimation

Estimation of $g_t(.)$ and V_t

A semi-parametric hedonic model with the specification in (12) can be implemented as a generalized additive model (GAM) – a flexible model class that generalizes linear models with a linear predictor combined with a sum of smooth functions of covariates estimated period by period using the available sample data on prices, characteristics and location coordinates.

$$
\begin{aligned}
y_{it} &= x'_{it}\theta_t + v_{it} && (16)\\
&= z'_{it}\theta^{\dagger}_t + g(z_{long}, z_{lat}) + v_{it} && (17)
\end{aligned}
$$

The estimates of the spline surface, $\hat{g}_t(.)$ enter the spatio-temporal model's X_t matrix; while the predictions from this model, $\hat{x}^{\dagger}_t$, provide $\widehat{v_t} = y_t - \hat{x}^{\dagger}_t$, from which a sample estimate of V_t, and thus $c'_{vt,h}$, make operational the correction term in (13) (see section 2.3).

To estimate (17) the problem is to select the smooth functions and their degree of smoothness. Here, we use a penalized likelihood approach (see Wood 2006 and the references therein) based on a transformation and truncation of the basis that arises from the solution of the thin plate spline smoothing problem. This method is computationally efficient and avoids the problem of choosing the location of knots, known to be crucial for other basis functions.

For the selection of the smoothing parameter we refer to Wood (2011), who proposes a Laplace approximation to obtain an approximate restricted maximum likelihood (REML) estimate which is suitable for efficient direct optimization and computationally stable. The REML criterion requires that a Newton-Raphson approach is used in model fitting, rather than a Fisher scoring. The penalized likelihood maximization problem is solved by Penalized Iteratively Reweighted Least Squares (P-IRLS).

The semi-parametric model is estimated using the *mgcv* package of the statistical software R 2.15.3 (R Core Team 2013). The same basis dimension and sample size are used as in Hill and Scholz (2017).[3]

Estimation of D, H, Q and α_t

Given y_t, $Z_t = \{z_{1t}, \ldots, z_{Kt}\}$, $\hat{g}_t(\cdot)$ and $\hat{V}_t$, estimates of $\hat{\rho}$, $\hat{\sigma}^2_\epsilon$, $\hat{\sigma}^2_\mu$, $\hat{\sigma}^2_\beta$ and $\hat{\sigma}^2_\gamma$ are required to obtain the estimated state-vector $\hat{\alpha}_{t|t}$ (and mean squared prediction matrix) using the Kalman filter recursion given by (7) which optimally weighs information up to and including week t. These parameters define the system matrices D, H and Q. The estimator's algorithm is a function of a prediction error, $\nu_{t|t-1} = y_t - X^1_t \hat{\alpha}_{t|t-1}$, and the Kalman Gain (9), which is a function of $F_t = E(\nu_{t|t-1}\nu'_{t|t-1}) = H + V_t + X_t P_{t|t-1} X'_t$. The Kalman filter algorithm is run to evaluate the log-likelihood $\ln L$ in predictive form.

$$lnL(\rho, \sigma^2_\epsilon, \sigma^2_\mu, \sigma^2_\beta, \sigma^2_\gamma; y_t, Y_{t-1}, Z_t, \hat{g}_{t|t-1}) = -\frac{NT}{2} ln(2\pi) - \frac{1}{2}\sum_{t=d}^{T} ln|F_t| - \frac{1}{2}\sum_{t=d}^{T} \nu'_{t|t-1} F_t^{-1} \nu_{t|t-1}$$

We use a standard Newton-Raphson algorithm to estimate $\hat{\sigma}^2_\epsilon$, $\hat{\sigma}^2_\mu$, $\hat{\sigma}^2_\beta$ and $\hat{\sigma}^2_\gamma$ within a grid search for ρ in the range (0.1 to 1). $Y_{t-1} = y_{t-1}, y_{t-2} \ldots$ $N = \sum_{t=d}^{T} N_t$; d is sufficiently large to avoid the log-likelihood being dominated by the initial condition, $\alpha_0 \sim N(a_0, \Omega_0)$. In the empirical implementation we have 731 weeks and set $d = 105$ (this choice is explained in the empirical section). For details on estimation of state-space models see Harvey (1989) or Durbin and Koopman (2012). The estimation of the model and computation of indices were coded by the authors.

2.5 Measuring the quality of the index

The constructed indices should be useful instruments for policymakers and market participants. A criterion is needed therefore to evaluate the quality of the proposed indices. An important distinction can be made here between the hedonic model and the resulting price index. What matters is the performance of the index. Hence performance criteria should focus on the Törnqvist index defined in (3), and not the within-period fit of the hedonic model itself. Guo, Zheng and Geltner (2014) and Jiang, Phillips and Yu (2015) take a similar view. Guo, Zheng and Geltner (2014) suggest criteria based on the autocorrelation and volatility of the index, and Jiang, Phillips and Yu (2015) create a testing sample which is used for out-of-sample evaluation of the model's fit. We follow a more direct approach here that makes use of the underlying structure of

---

[3] It is important that the sample size each period exceeds the basis dimension. Hill and Scholz select these values by comparing computing time and model fit as measured by the Akaike Information Criterion (AIC).

our hedonic imputation price indices.

The Törnqvist index is the geometric mean of the Laspeyres and Paasche-type price index formulas (1) and (2). From inspection of (1) and (2) it can be seen that the building blocks of the Laspeyres-type index are the imputed price relatives $\hat{p}_{i,t+1}(x'_{i,t})/\hat{p}_{i,t}(x'_{i,t})$, while the building blocks of the Paasche-type index are the imputed price relatives $\hat{p}_{i,t+1}(x'_{i,t+1})/\hat{p}_{i,t}(x'_{i,t+1})$. Hence the performance of the index depends on the quality of these imputed price relatives. Following Hill and Scholz (2017), the key insight is that for some houses in our data set repeat sales are available. These actual repeat sales price relatives can be used as a benchmark for evaluating the imputed price relatives. To ensure a large enough sample size, repeat-sales price relatives over any time horizon in our data set are with their imputed counterparts, and not just in adjacent periods.

More formally, suppose proprerty i sells in both periods t and $t+k$. For this property therefore we have a repeat-sales price relative: $p_{i,t+k}/p_{i,t}$. The corresponding imputed price relative is $\hat{p}_{i,t+k}/\hat{p}_{i,t}$. The sample of repeat-sale dwellings are indexed by $i = 1, \ldots, H_{RS}$. We can now define the ratio of imputed to actual price relative for house i as follows:

$$V_i = \frac{\hat{p}_{i,t+k}}{\hat{p}_{i,t}} \bigg/ \frac{p_{i,t+k}}{p_{i,t}}. \tag{18}$$

Our quality measure is then the average squared error of the log price relatives of each hedonic method:

$$D = \left(\frac{1}{H_{RS}}\right) \sum_{i=1}^{H_{RS}} [\ln(V_i)]^2, \tag{19}$$

where the summation in (19) takes place across the whole repeat-sales sample. We prefer whichever hedonic imputation model generates the smallest value of D, on the grounds that the resulting Törnqvist index will be constructed from the most reliable imputed price relatives.

Given that we use repeat-sales as a benchmark for our imputed price relatives, our intention is to exclude repeat sales where the house was renovated between sales. We attempt to identify such houses in two ways. First, we exclude repeat sales where one or more of the characteristics have changed between sales (for example a bathroom has been added). Second, we exclude repeat sales that occur within six months on the grounds that this suggests that the first purchase was by a professional renovator.[4] Finally, for houses that sold more than twice during our sample period (2003-2014), we only include the two chronologically closest repeat sales (as long as these are more than

---

[4] Exclusion of repeat-sales within six months is standard practice in repeat-sales price indices such as the Standard and Poor's/Case-Shiller (SPCS) Home Price Index.

six months apart). This ensures that all repeat-sales houses exert equal influence on our results. There are 83 258 repeat-sales houses in the full data set. As a result of the deletions explained above, the sample was reduced to 61 024 houses.

One potential problem with using repeat-sales as a benchmark is that a repeat-sales sample may have a "lemons" bias, since starter homes sell more frequently as people upgrade as their wealth rises. This lemons bias has been documented by, amongst others, Clapp and Giaccotto (1992), Gatzlaff and Haurin (1997), and Shimizu, Nishimura and Watanabe (2010). The quality of the house between repeat sales may also decline due to depreciation or it could improve due to renovations and repairs. If over the whole data set one of these effects dominates the other, then the repeat-sales index will not be fully quality adjusted.

We correct for any such bias by adjusting the repeat-sales price relatives $p_{t+k,h}/p_{t,h}$ as follows:

$$\left(\frac{p_{i,t+k}}{p_{i,t}}\right)^{adj} = \left[\left(\frac{P^{RS}_{t+k}}{P^{RS}_{t}}\right) \Big/ \left(\frac{P^{Hed}_{t+k}}{P^{Hed}_{t}}\right)\right]\left(\frac{p_{i,t+k}}{p_{i,t}}\right), \tag{20}$$

where P^{RS}_{t+k}/P^{RS}_{t} denotes the change in the repeat-sales price index between periods t and $t+k$, while $P^{Hed}_{t+k}/P^{Hed}_{t}$ is the change in a reference hedonic index, calculated using the Törnqvist formula in (3) over the same time interval. Hence the ratios of actual to imputed price relatives are adjusted as follows:

$$V_i^{adj} = V_i\left[\left(\frac{P^{RS}_{t+k}}{P^{RS}_{t}}\right) \Big/ \left(\frac{P^{Hed}_{t+k}}{P^{Hed}_{t}}\right)\right]. \tag{21}$$

Bias corrected D coefficients, denoted by D^{adj} in Table 1, are then calculated as follows:

$$D^{adj} = \left(\frac{1}{H_{RS}}\right)\sum_{i=1}^{H_{RS}}[\ln(V_i^{adj})]^2.$$

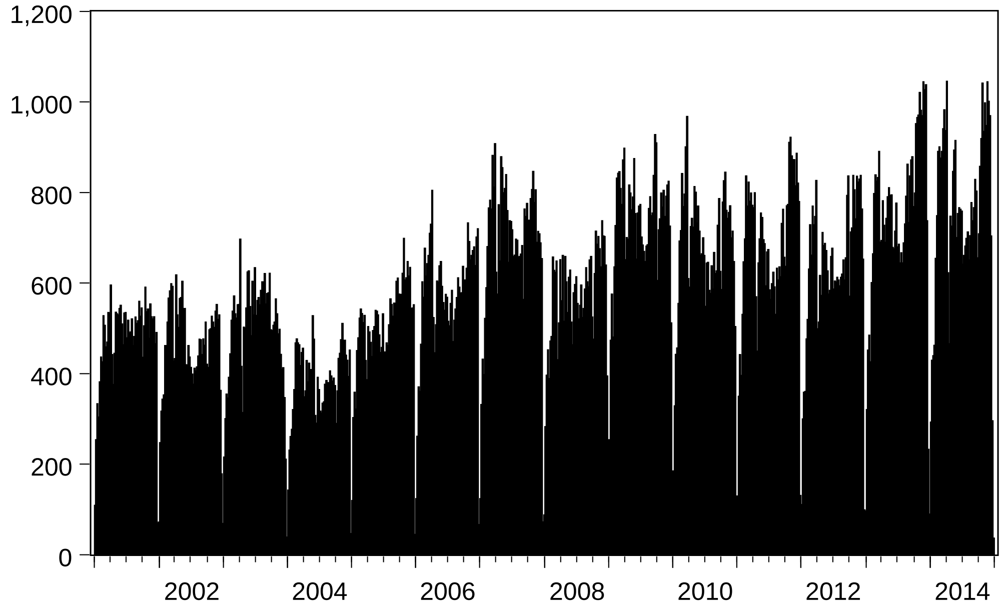
There remains the question of which set of hedonic price indices should be used to make the lemons bias correction when computing (20) and (21). As a robustness check we take each of three indices (details in the next section) in turn as the reference method when making the bias correction. Hence in Table 1 we present three alternative D^{adj} coefficients. In all cases the ranking of methods is the same. Hence our findings are robust to the treatment of lemons bias.

3 Empirical application

3.1 The data set

We use a data set obtained from Australian Property Monitors that consists of prices and characteristics of houses sold in Sydney (Australia) for the years 2001–2014. For each house we have the following characteristics: the actual sale price, time of sale, postcode, property type (i.e., detached or semi), number of bedrooms, number of bathrooms, land area, exact address, longitude and latitude. (We exclude all townhouses from our analysis since the corresponding land area is for the whole strata and not for the individual townhouse itself.) Some summary statistics are provided in the Appendix in Table 2, and a plot of the number of sales per week is shown in Figure 1.

Figure 1: Number of Transactions per Week, 2001-2014



For a robust analysis it was necessary to remove some outliers. This is because there is a concentration of data entry errors in the tails, caused for example by the inclusion of erroneous extra zeroes. These extreme observations can distort the results. The exclusion criteria we applied are shown in the Appendix in Table 3. Complete data on all our hedonic characteristics are available for 433 202 observations. To simplify the computations we also merged the number of bathrooms and number of bedrooms to broader groups (one, two, and three or more bathrooms; one or two, three, four, five or more bedrooms). The quality of the data improves over time. In particular, missing characteristics are quite common in the first two years (i.e., 2001 and 2002). Thus we present the hedonic indices starting in 2003. Nevertheless, we use the full sample period to run the Kalman filter algorithm but compute the log-likelihood function (see section 2.4) with all weeks in the 2003-2014 period.

3.2 Property price indices

We construct three hedonic price indices. A basic index is computed from the semi-parametric hedonic model in (17) estimated separately each week. This index is referred to as **GAM**. The second is based on the spatio-temporal hedonic model presented in section 2 and is referred to as **SS+GAM**. As discussed in section 1, a frequent control for location used in hedonic imputation is the addition of postcode dummies. Thus, a simpler alternative to (17) is

$$y_t = \mu_t + Z_t\beta_t + D_t\pi_t + \varepsilon_t \tag{22}$$

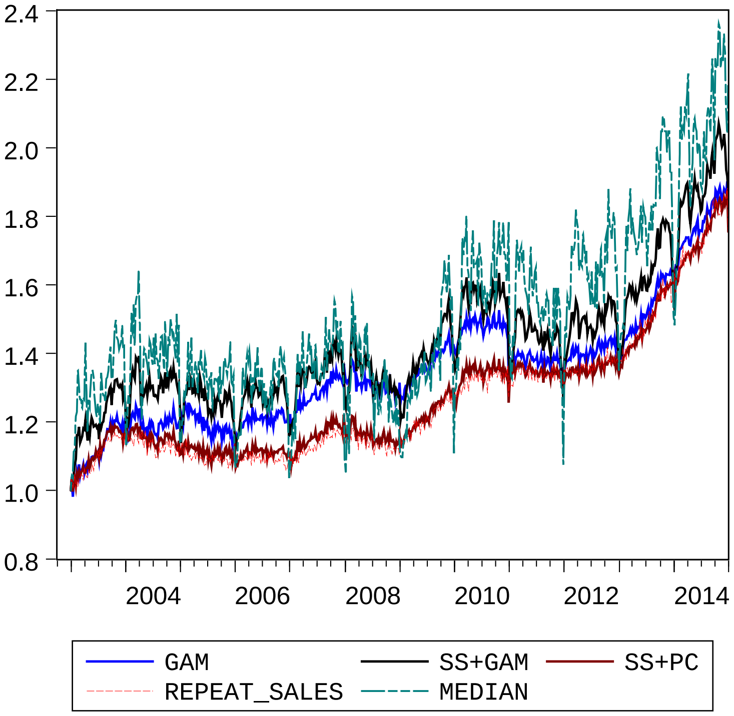
where μ_t is local level trend, Z_t hedonic characteristics, D_t is a matrix of postcode dummies containing the location information and π_t is the vector of corresponding shadow prices.

Computing hedonic imputation price indices using period-by-period estimation with (22) is not feasible in a weekly context. It happens that for some postcodes we have no observations in some weeks causing both statistical and computational problems, especially in the hedonic prediction step. However, it can be estimated as a regression with time-varying parameters by setting it up as a state-space model. The index obtained from this model is referred to here as **SS+PC**. Imputed price relatives from the three models are inserted into the Törnqvist formula to generate the respective price index.

Figure 2 shows the three hedonic indices (chained), a repeat sales index calculated using the standard Bailey, Muth, and Nourse (1963) formula, and a median price index computed from the median of the prices of observed sales in each week. The median index is a quality and location unadjusted index. It is extremely volatile, thus demonstrating the need for quality adjustment to generate an economically meaningful index. All indices except for SS+GAM lie below the median price index for most of the sample period. The GAM index appears to suffer from some chain drift. Prior to 2011 the index is closer to the median and the SS+GAM; however, it drifts down to the SS+PC and repeat-sales indices after 2011. Index drift is likely to occur in the conventional approach to hedonic imputation when the market is thin. Small samples and sales' composition in thin markets can affect the parameter estimates and lead to large changes in the price relatives. Chaining then compounds the effect. Rambaldi and Fletcher (2014) find chain drift occurs in monthly indices even when using a two-months rolling window to estimate the parameters of the model, while this is not the case when the imputation is obtained from a state-space model. The SS+PC and repeat-sales indices are uniformly below the median and virtually indistinguishable from each other.

The differences between the hedonic indices in Figure 2 are surprisingly large and far larger than one would expect to observe in hedonic indices computed at annual or quarterly frequency (Hill and Scholz, 2017). The results therefore demonstrate the importance of the choice of hedonic method for indices computed at lower frequencies, such as weekly.

Figure 2: Weekly Property Price Indices from 2003 to 2014



Note: GAM is based on periodwise estimation of model (12); SS+PC is the state space model (22) with postcode dummies; SS+GAM is the spatio-temporal model; Repeat_Sales index is calculated using the Bailey, Muth, and Nourse (1963) formula; Median is the usual median index on a weekly frequency. Base:Week starting 30/12/2002 = 1

The median is a unbiased estimator of the central tendency in log-normal data, such as property prices, and thus the large deviation of these two indices from the median over the period would seem to indicate a systematic bias. The next section formally evaluates the quality of these indices.

3.3 Comparing the quality of the indices

The performance of our three indices according to the D and D^{adj} criteria is shown in Table 1.

Table 1: Index quality based on D and D^{adj} criteria (2003-2014)

	D	D^{adj}_{GAM}	D^{adj}_{SS+GAM}	D^{adj}_{SS+PC}
GAM	0.0233	0.0272	0.0313	0.0230
SS+GAM	0.0102	0.0096	0.0099	0.0133
SS+PC	0.0246	0.0279	0.0320	0.0240

Note: GAM is based on periodwise estimation of the semiparametric model (17) with a geospatial spline; SS+GAM is the spatio-temporal model; SS+PC is the state space model applied to the semilog model in (22) with location effects captured using postcodes. D^{adj}_{GAM} refers to the adjusted D criteria with lemons bias corrected for using the GAM hedonic price index as the adjustment factor. Similarly, D^{adj}_{SS+GAM} and D^{adj}_{SS+PC} use the SS+GAM and SS+PC hedonic price indexers, respectively as the adjustment factors.

All our criteria (D, D^{adj}_{GAM}, D^{adj}_{SS+GAM}, D^{adj}_{SS+PC}) generate the same ranking of hedonic methods. In all cases, the SS+GAM model performs best followed by GAM, with SS+PC performing worst.

Furthermore, the superior performance of SS+GAM is highly statistically significant. To show this we apply the following hypothesis test based on the Central Limit Theorem (see, for example, pages 490-491 in Devore and Berk, 2012). The D and D^{adj} criteria are of the form:

$$\bar{X} := \frac{1}{H_{RS}} \sum_{i=1}^{H_{RS}} u_i^2,$$

with the prediction errors u_i equal to $\ln\big(\hat{p}_i/p_i\big)$ or $\ln\big(V_i\big)$, respectively. Now we want to test whether $\bar{X}_1$ and $\bar{X}_2$ are significantly different, where $\bar{X}_1$ and $\bar{X}_2$ are the results (criteria) of different hedonic models. To test the null hypothesis that the true difference is zero ($H_0: \ \bar{X}_1 - \bar{X}_2 = 0$), assume

$$\bar{X}_1 - \bar{X}_2 \ \sim \ \mathcal{N}\left(0, \frac{s_1^2 + s_2^2}{H}\right),$$

where s_j (j=1,2) is the sample standard deviation of u_h^2 of the hedonic model j. The test-statistic and corresponding two-sided p-values of this exercise are shown in the Appendix in Table 4.

These results therefore show the importance of correctly modelling space and time in a unified framework which can account for all sources of error.

4 Conclusion

This article focuses on the construction of weekly house price indices using the hedonic imputation method. The hedonic imputation method provides a flexible way of constructing quality-adjusted house price indices using a matching sample approach. We develop a spatio-temporal model to obtain the imputed prices. A geospatial spline surface controls for location and is embedded in a state-space formulation that controls for trends and property quality. We show the spatio-temporal specification leads to a modified form of the Kalman filter and a Goldberger's adjusted form of the predictor to obtain the imputations.

The paper makes three main contributions to the hedonic literature. First, it shows how flexible and robust hedonic indices can be estimated by embedding a geospatial spline surface in a state-space framework. Second, using a data set for Sydney (Australia) weekly hedonic indices are shown to be far more sensitive to the method of construction than indices computed at an annual or quarterly frequency. Hence it is at these higher frequencies that the choice of hedonic method matters most. Third, using a criterion proposed by Hill and Scholz (2017) it is shown that embedding a semi-parametric model with geospatial spline surface in a state-space model generates house price indices that outperform two competing hedonic imputation methods and the repeat-sales method.

References

Bailey, M., R. Muth, and R. Nourse (1963), A Regression Method for Real Estate Price Index Construction. *Journal of the American Statistical Association* **58**, 933-942.

Bokhari, S. and D. Geltner (2012). Estimating Real Estate Price Movements for High Frequency Tradable Indexes in a Scarce Data Environment. *Journal of Real Estate Finance and Economics* **45**(2), 522–543.

Bollerslev, T., A.J. Patton, and W. Wang (2015). Daily House Price Indices: Construction, Modeling, and Longer-Run Predictions. *Journal of Applied Econometrics* forthcoming.

Bourassa, S.C. and M. Hoesli (2016). High Frequency House Price Indexes with Scarce Data. *Swiss Finance Institute Research Paper Series* **16-27**.

Clapp, J.M. and C. Giaccotto (1992). Estimating Price Trends for Residential Property: A Comparison of Repeat Sales and Assessed Value Methods. *Journal of Real Estate Finance and Economics* **5**(4), 357–374.

Cressie, N. and C.K. Wikle (2002). Space-time Kalman filter. *Encyclopedia of Environmetrics* **4**, 2045–2049.

Court, A. T. (1939). The Dynamics of Automovile Demand, New York, N.Y.: General Motors Corporation, chap. Hedonic Price Indexes with Automotive Examples, 99?117.

de Haan, J. (2010). Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and Re-Pricing Methods. *Journal of Economics and Statistics* **230**(6), 772–791.

Devore, J.L., and K.N. Berk (2012). *Modern Mathematical Statistics with Applications.* Springer, New York, 2nd edition.

Diewert, W. E. (2010). Alternative Approaches to Measuring House Price Inflation. Discussion Paper 1010, Department of Economics, The University of British Columbia, Vancouver, Canada, V6T 1Z1, 2010.

Durbin, J. and Koopman, S. (2012). *Time Series Analysis by State Space Methods.* Oxford Statistical Science Series. Oxford University Press, 2nd edition.

European Commission, Eurostat, OECD, and World Bank (2013): Handbook on Residential Property Price indices (RPPIs). Eurostat: Luxembourg.

Eurostat (2016). Detailed Technical Manual on Owner-Occupied Housing for Harmonised Index of Consumer Prices. Eurostat: Luxembourg.

Gatzlaff, D.H. and D.R. Haurin (1997). Sample Selection Bias and Repeat-Sales Index Estimates. *Journal of Real Estate Finance and Economics* **14**, 33–50.

Geltner, D. and D. Ling (2006). Considerations in the Design and Construction of Investment Real Estate Research Indices. *Journal of Real Estate Research* **28**(4), 411-444.

Griliches, Z. (1961). Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change, in *Government Price Statistics, Hearings Before the Subcommittee on Economic Statistics of the Joint Economic Committee*, 87th Congress.

Griliches, Z. (1971). Price Indexes and Quality Change, Cambridge MA: Harvard University Press., chap. Introduction: Hedonic Price Indexes Revisited.

Goldberger, A.S. (1962). Best Linear Unbiased Prediction in the Generalized Linear Regression Model. *Journal of the American Statistical Association* **57**, 298, 369 –375.

Guo, X., S. Zheng, D. Geltner, and H. Liu (2014). A new approach for constructing home price indices: The pseudo repeat sales model and its application in China. *Journal of Housing Economics* **25**, 20-38.

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

Hill, R.J. (2013). Hedonic Price Indexes for Housing: A Survey, Evaluation and Taxonomy. *Journal of Economic Surveys* **27**(5), 879–914.

Hill, R.J. and D. Melser (2008). Hedonic Imputation and the Price Index Problem: An Application to Housing. *Economic Inquiry* **46**(4), October 2008, 593–609.

Hill, R.J. and M. Scholz (2017). Incorporating Geospatial Data in House Price Indexes: A Hedonic Imputation Approach with Splines. *Review of Income and Wealth*, forthcoming.

Jiang, L., P.C.B. Phillips, and J. Yu. New Methodology for Constructing Real Estate Price Indices Applied to the Singapore Residential Market. *Journal of Banking and Finance* **61**, S121–S131.

Mardia, K.V., C. Goodall, E.J. Redfern, and F.J. Alonso (1998). The Krigged Kalman Filter. *Test* **7**(2), 217–285.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/.

Rambaldi, A.N. and D.S.Pr. Rao (2011). Hedonic Predicted House Price Indices Using Time-Varying Hedonic Models with Spatial Autocorrelation. School of Economics Discussion Paper 432, School of Economics, University of Queensland.

Rambaldi, A.N. and C.S. Fletcher (2014). Hedonic Imputed Property Price Indexes: The Effects Of Econometric Modeling Choices.*Review of Income and Wealth* **60**, Supplement Issue, S423-S448. DOI:`10.1111/roiw.12143`.

Shimizu, C., K.G. Nishimura and T. Watanabe (2010). Housing Prices in Tokyo: A Comparison of Hedonic and Repeat Sales Measures. *Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik)* **230**(6), 792–813.

Silver, M. (2011). House Price Indices: Does Measurement Matter? *World Economics* **12**(3), 69–86.

Wikle C.K. and N. Cressie (1999). A Dimension-Reduced Approach to Space-Time Kalman Filtering. *Biometrika* **86**, 815–829.

Wood, S.N. (2006). *Generalized Additive Models: An introduction with R*, Chapman & Hall/CRC.

Wood, S.N. (2011). Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society B* **73**(1), 3–36.

Appendix

A1. Proof of prediction expression (13)

$$\widehat{y^*_{t|t,h}} = x'_{t,h}\alpha_{t|t} + c'_{vt,h}\Omega^{-1}e_t \tag{23}$$

In addition to assumptions already stated, we assume v_{it} and y_t have a joint multivariate normal distribution. Taking the characteristics and location of properties as given, the predictor is derived as follows,

$$\begin{aligned}\widehat{y^*_{i,t|t}} &= E\{y^*_{it}|y_t, y_{t-1}, \dots, y_1\} \\ &= E\{X_{it}\alpha_t + v_{it}|y_t, y_{t-1}, \dots, y_1\} \\ &= X_{it}E\{\alpha_t|y_t, y_{t-1}, \dots, y_1\} + E\{v_{it}|y_t, y_{t-1}, \dots, y_1\} \\ &= X_{it}\alpha_{t|t} + c'_{vt,h}\Omega^{-1}e_t\end{aligned}$$

The last term is of this form since $E\{v_{it}y_{jt}\} = c_{v,ij}$; $c_{v,ij} \equiv E\{v_{it}v_{jt}\}$, and $c'_{v,it} = E\{v_{it}, v_t) = \{c_v(i, j_1), \dots, c_v(i, j_{N_t})\}'$

A2. Further information on the data set

Some summary statistics for our data set are provided in Table 2.

Table 2: Summary of characteristics

	PRICE ($)	BED		BATH		AREA	LAT	LONG
Minimum	56500	1:	1348	1:	190395	100.0	-34.20	150.6
1st Quartile	420000	2:	38578	2:	174161	461.0	-33.93	150.9
Median	610000	3:	200428	3:	57673	587.0	-33.84	151.0
Mean	784041	4:	147794	4:	8835	626.1	-33.85	151.0
3rd Quartile	900000	5:	38734	5:	1746	720.0	-33.76	151.2
Maximum	3200000	6:	6320	6:	392	4998.0	-33.40	151.3

For a robust analysis it was necessary to remove some outliers. The exclusion criteria we applied are shown in Table 3.

Table 3: Criteria for removing outliers

	PRICE	BED	BATH	AREA	LAT	LONG
Minimum Allowed	50000	1.000	1.000	100.0	-34.20	150.60
Maximum Allowed	4000000	6.000	6.000	5000.0	-33.40	151.35

A3. Hypothesis tests to show that the D and D^{adj} criteria are significantly different

The p-values for the hypothesis tests (with null hypothesis of equality of the D criteria across hedonic methods are as follows:

Table 4: p-values for hypothesis tests

	D	D^{adj}_{SS+PC}	D^{adj}_{SS+GAM}	D^{adj}_{GAM}
SS+PC vs. SS+GAM	0.00000000	0.0000000	0.0000000	0.0000000
SS+PC vs. GAM	0.04830767	0.1004416	0.2572753	0.2732328
GAM vs. SS+GAM	0.00000000	0.0000000	0.0000000	0.0000000