Empirical Economics (2021) 61:417–441
https://doi.org/10.1007/s00181-020-01862-y

Higher frequency hedonic property price indices: a state-space approach

Robert J. Hill[1] · **Alicia N. Rambaldi**[2] · **Michael Scholz**[1]

Received: 16 June 2019 / Accepted: 17 March 2020 / Published online: 11 April 2020

Abstract
The hedonic imputation method allows characteristic shadow prices to evolve over time. These shadow prices are used to construct matched samples of predicted prices, which are inserted into standard price index formulas. We use a spatio-temporal model to improve the method's effectiveness on housing data at higher frequencies. The problem is that at higher frequencies, there may not be enough observations per period to reliably estimate the characteristic shadow prices. In such cases, the reliability of the hedonic imputation method is improved by using a state-space formulation which yields estimates of the shadow prices that are weighted sums of previous periods' information. In addition, the state-space representation of the model includes a geospatial spline surface which significantly reduces the number of parameters to be estimated when compared to the standard practice of including postcode dummies in the model. Empirically, using a novel criterion, we show that in higher frequency comparisons, our hedonic method outperforms competing alternatives.

Keywords Housing market · Hedonic imputation · State-space model · Geospatial data · Spline · Quality adjustment and matched sample

JEL C33 · C43 · R31

We acknowledge financial support for this project from the Austrian Research Promotion Agency (FFG), Grant #10991131, and we thank Australian Property Monitors for supplying the data. We are very thankful for the detailed and excellent comments provided by an anonymous referee and the handling editor.

✉ Michael Scholz
michael.scholz@uni-graz.at

Robert J. Hill
robert.hill@uni-graz.at

Alicia N. Rambaldi
a.rambaldi@uq.edu.au

[1] Department of Economics, University of Graz, Universitätsstr. 15/F4, 8010 Graz, Austria

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

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, and for the development of tradable derivatives (Bokhari and Geltner 2012). The increased availability of housing data and advances in computing power and econometric techniques is making it possible to deliver more timely indices to meet this demand.

Much progress has been made recently on computing higher frequency repeat-sales house price indices (Bokhari and Geltner 2012; Bollerslev et al. 2016; Bourassa and Hoesli 2016). However, less progress has been made on computing higher frequency hedonic indices.[1] It is such higher frequency hedonic indices that are the focus of our attention here.

Hedonic methods estimate the price of a product (here housing) as a function of a vector of explanatory characteristics. The hedonic imputation method, first proposed by Court (1939) and further developed by Griliches (1961), re-estimates the hedonic model each period. The hedonic model is then used to predict prices for matched samples, after which the overall price index can be computed using a standard price index formula. The method is more flexible and timely than other hedonic methods such as the time-dummy method in that the characteristic shadow prices are updated each period. This flexibility can be important, especially when structural changes occur in the market (Shimizu and Nishimura 2007). Unlike the time-dummy method, the hedonic imputation method also satisfies non-revisability (i.e. the indices once computed are never revised). Index users often find this useful. For example, Eurostat (2016) advises European countries to use a non-revisable hedonic method when constructing their official house price indices.

The hedonic imputation method, however, becomes problematic in a number of settings. von Auer (2007) raised the issue in the context of constructing price indices for information technology products. For the case of housing as in this study, the challenge arises when computing indices at higher frequencies. For example, the hedonic model typically includes dummies to control for location (e.g. using postcodes) in addition to other hedonic characteristics of the dwelling. At higher frequencies (e.g. weekly indices), 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

[1] Quality-adjusted indices are typically computed using either hedonic or repeat-sales methods. The latter are more common in the USA—the best-known example being the S&P CoreLogic Case-Shiller indices. In Europe, 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 USA. In Europe, 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. One advantage of hedonic methods is that they are less prone to sample selection bias issues. For example, Shimizu et al. (2010) find that repeat-sales indices for Tokyo fail to correctly measure the turning points in the housing market.

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).

The innovation in this paper is to use a spatio-temporal specification to improve the effectiveness of the hedonic imputation method at higher frequencies. As noted above, at higher frequencies, there may not be enough observations per period to reliably estimate the characteristic shadow prices. In such cases, the reliability of the hedonic imputation method is improved by using a state-space formulation which yields estimates of the shadow prices that are weighted sums of previous periods' information. The spatial component of the model is a locational price effect defined by a geospatial spline surface. This replaces the standard postcode dummies commonly used in the housing hedonic imputation literature and thus significantly reduces the number of parameters that need to be estimated. The spatio-temporal specification provides a unique form of accounting for spatial and temporal variability. This approach builds on the smoothed polynomial method proposed by von Auer (2007), and the state-space representation of hedonic functions proposed by von Auer and Trede (2012) when modelling the laser printer market, and Rambaldi and Fletcher (2014) in the housing context.

Like von Auer (2007), we use the double imputation hedonic method to construct the index, which partially controls for omitted variables (when they are reasonably stable over time as is typically the case in a housing context) (Hill 2013). Double imputation implies using predicted prices in both the numerator and denominator when computing each price relative to the price index formula. To see how double imputation controls for omitted variables, consider an example where a property is located next to a busy road. This means that its predicted price from the hedonic model in each period may tend to be too high (as the hedonic model will have difficulty fully capturing the effect of the road). The upward bias in the predicted prices will somewhat offset each other when we take price relatives (i.e. the price change) for this property.

We represent the hedonic pricing function using a spatio-temporal state-space specification. This model is used to produce matched predictions of the sale price of each house in the sample across two time periods, providing a model-based measure of price relatives which enter the computation of a superlative index formula (see Diewert 1976). We are not aware of any prior research in econometrics that incorporates spatial heterogeneity and variability, temporal dynamics and their interactions in the form we propose.

The estimation is a two-step approach. A nonparametric estimate of a locational price effect that varies over individual properties at any given time period is identified by the first stage. These estimates (and their statistical uncertainty) enter the second stage, a time-varying parameter model written as state-space. The estimation of the second stage is by the Kalman filter although the state-space model itself has a number of non-standard features. Lastly, the predictions from the model needed to provide the

correct matching as required by the Törnqvist price index formula (a superlative index based on the ratio of two predictions) are not from standard predictors. In summary, the model being estimated and its purpose, to provide a model based matching sample to construct a price index, are not a regular feature of either academic publications or standard practice by official or commercial providers of house price indices.

There are some similarities here with the literature on constructing monthly or weekly price indices for consumer goods using scanner data. In particular, Melser (2018) uses the hedonic imputation method to estimate superlative price indices and impose non-revisability using a rolling window method. He then endogenizes the window length and allows each period in the window to be weighted differently. We impose non-revisability using an approach that mimics an endogenous rolling window. However, there are important differences as well. We focus on the construction of house price indices, not consumer price indices. By representing the hedonic model as a state-space, estimation by the Kalman filter produces predictions that optimally weight prior periods' information (expressions for the weights can be found in Koopman and Harvey 2003).

In addition to evaluating how well each of our hedonic models predicts prices in each period, we also consider the performance of the Törnqvist indices. To do this, we use a recently developed criterion proposed by Hill and Scholz (2018) and apply it to data for Sydney (Australia) over the period 2003–2014. This criterion focuses on comparing the predicted price relatives of properties with those of actual observed repeat sales within our sample. The rationale behind the use of this measure is that predicted price relatives form the basic building blocks of the Törnqvist superlative price index, and thus, it is the ability of the model to predict price changes over time rather than the price level of each property that really matters. Based on this criterion, we find that our preferred index outperforms competing alternatives. Furthermore, we find that weekly indices are quite sensitive to the choice of method.

The remainder of this article is structured as follows. Section 2 provides an overview of the hedonic imputation method, the hedonic model, and the methods used to estimate the generalized additive and the spatio-temporal components of the model. The criterion used to compare the performance of competing hedonic imputed indices is also considered here. Section 3 presents our data set, the empirical study, and the results of our analysis. Section 4 concludes by summarizing our main findings.

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; Diewert and Shimizu 2015; Silver 2016). 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 (Silver and Heravi 2007). The hedonic imputation method uses the predictions from a hedonic model to predict prices over a matched sample which can then be inserted into a standard price index formula. Let $x'_{i,t}$ be a vector of characteristics associated with property i sold in period t, and $\hat{p}_{i,t+1}(x'_{i,t})$ as the predicted 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.[2] Taking a geometric mean of the Laspeyres and Paasche-type indices, we obtain a Törnqvist-type superlative 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.[3] This means that both prices in each price relative are predicted. For example, the double imputation Laspeyres (DIL), Paasche (DIP), and Törnqvist indices (DIT) are defined as follows:

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

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

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

where $i = 1, \ldots, N_t$ indexes the dwellings sold in period t, and $i = 1, \ldots, N_{t+1}$ indexes the dwellings sold in period $t+1$. The overall price index is then constructed by chaining together these bilateral comparisons between adjacent periods. As is 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] 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), Rambaldi and Rao (2011), and Rambaldi and Fletcher (2014) for a discussion on alternative weighting schemes.

[3] 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 predicted. The results are virtually indistinguishable. Hence, to save space, we focus here only on double imputation indices.

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 combines elements from the work of Wikle and Cressie (1999) and Rambaldi and Fletcher (2014). Wikle and Cressie (1999) provide a temporally dynamic and spatially descriptive model and an efficient estimation algorithm designed to deal with a large-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. Following Hill and Scholz (2018), the model incorporates an estimated locational price effect obtained by estimating a semi-parametric model using observed sales in each individual period. The periodwise estimation provides a required measure of spatial variability and identification of the parameters of the spatio-temporal model.

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

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

The random process ϵ_{it} is independent across location or time and captures overall measurement error. Thus, $E(\epsilon_t \epsilon_t') = \sigma^2_\epsilon I_{N_t}$, where N_t properties are transacted in period t.

At each time period $t = 1, 2, \ldots, T$ the multivariate process y^*_t is the sum of two components, one explained by temporal and spatial dynamics, $x^\dagger_t$, and a spatially descriptive random component, $v_{|t}$ (with the notation "$|t$" used to refer to random quantities that vary within each time period t) ,

$$y^*_{it} = x^\dagger_{it} + v_{i|t}; v_{|t} \sim N(0, V_{|t}) \tag{5}$$

where $v_{|t}$ is a random error that does not have a temporally dynamic structure but might have some spatial structure, and thus, the covariance $V_{|t}$ might not be diagonal. It is assumed that $E(v_{i|t}\epsilon_{jt}) = 0$ and $-\infty \leq t \leq \infty$.

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

$$x^\dagger_{it} = \mu_t + \sum_{k=1}^{K} \beta_{k,t} z_{k,it} + \gamma_t g_{i|t} \tag{6}$$

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).

Note that the vector $z_{k,t}$ contains the kth hedonic characteristic for the set of properties sold in period t, which is different to the set of properties sold in $t - s$ where $s \neq 0$, and thus, these are not trending variables.

$g_{i|t} := g_{i|t}(z_{\text{long},i}, z_{\text{lat},i})$ is a measure of the locational price effect of property i on a continuous surface defined by the N_t properties sold in period t. Thus, it is temporally uncorrelated as the set sold in period t is different from that sold in $t - 1$ or $t + 1$.[4]

$\beta_{k,t}$ are time-varying parameters associated with the hedonic characteristics and are to be estimated.

γ_t is a state parameter that carries the temporal information associated with location. At each time period, it interacts with the spline, $g_{i|t}$, to shift its position due to temporal information. The law of motion for this state parameter is presented below.

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

Putting together Eqs. (4), (5), and (6) gives the model to be estimated from where we then obtain the required predictions to construct the index [using the expressions in (1)–(3)],

$$y_{it} = \mu_t + \sum_{k=1}^{K} \beta_{k,t} z_{k,it} + \gamma_t g_{i|t} + v_{i|t} + \epsilon_{it} \tag{7}$$

The identification strategy and estimation of this model are presented in the next section.

2.2.1 Identification and estimation

Inspecting Eq. (7), it is clear that an identifying assumption is required to be able to estimate γ_t and compute covariances $V_{|t}$ and $H_t = \sigma_\epsilon^2 I_{N_t}$. We estimate a locational price effect, denoted by $\widehat{g}_{i|t}$, and the covariance $V_{|t}$ as a first step of the estimation using an auxiliary semi-parametric model of the form in (8) estimated at each time period t using only transacted properties from that period (details are provided in Appendix Sect. A.1).

$$y_{i|t} = \theta_{0|t} + z'_{i|t} \theta^{\dagger}_{|t} + g_{i|t} + v_{i|t}, \tag{8}$$

where $\theta^{\dagger}_{|t} = \{\theta_{1|t}, \ldots, \theta_{K|t}\}'$ and $g_{i|t}$ is centered around the (conditional) expectation of $y_{i|t}$.[5]

[4] The only overlap is that of properties that repeat sale. The hedonic imputation method treats these as independent sales. We use this feature of the approach to compute an index performance indicator as explained in Sect. 2.5.

[5] The shape restriction on the spline surface g is a standard identifiability constraint for generalized additive models.

The estimation of Eq. (8) gives predictions of (log) prices, $\hat{y}_{i|t}$, and of the location effect, $\hat{g}_{i|t}$, for each property $i = 1, \ldots, N_t$. The corresponding residuals, $\hat{v}_{i|t}$, are used to compute $\hat{V}_{|t}$.

With the above definitions, the model in Eq. (7) can be written in familiar state-space representation,

$$y_t = X_t \alpha_t + v_{|t} + \epsilon_t; \quad \epsilon_t \sim N(0, H_t) \tag{9}$$

$$\alpha_t = D\alpha_{t-1} + \eta_t; \quad \eta_t \sim N(0, Q) \tag{10}$$

where X_t is $N_t \times (K+2)$ and with the ith row being $x'_{it} = \{1, z_{1,it}, \ldots, z_{K,it}, \hat{g}_{i|t}\}$, y_t is the vector of log transformed observed prices of properties sold at t.
$H_t = \sigma^2_\epsilon I_{N_t}$
$\alpha_t = \{\mu_t, \beta_{1t}, \ldots, \beta_{K,t}, \gamma_t\}'$
$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & I_K & 0 \\ 0 & 0 & \rho \end{bmatrix}$; $0 \leq \rho \leq 1$; 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 in α_t.
$Q = \begin{bmatrix} \sigma^2_\mu & 0 & 0 \\ 0 & \sigma^2_\beta I_K & 0 \\ 0 & 0 & \sigma^2_\gamma \end{bmatrix}$

The estimate $\widehat{g_{i|t}}$ enters the spatio-temporal model as a generated regressor, and the parameter γ_t, in (9) and (10), 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 additional uncertainty induced by replacing $g_{i|t}$ by an estimate is captured by $v_{|t}$ in (9).

The combination of spatial and temporal information leads to two unconventional features of this model, compared to one in a standard setting (e.g. hedonic function with postcode dummies), with consequences for the Kalman filter algorithm as well as the price prediction to be used for the computation of the Törnqvist price index.

The estimation of the state vector via the Kalman filter in this model [Eq. (11)] differs from that in the standard case. First, as the model has a composite error term, the Kalman gain, G_t, is a function of the sum of the two covariances $(H_t + V_{|t})$ under the assumptions already stated [see (12)]. Second, the locational price effect must be consistent with the required prediction error (expression in curly brackets). The prediction error is the difference between the observed vector of log prices in the current time period t, y_t, and its conditional prediction. The conditional prediction is computed from the conditional state estimate, $\alpha_{t|t-1}$, and hedonic characteristics of the sold properties at t. For this model, the form of the matrix involved in this conditional prediction is $X^1_t = \{\mathbf{1}, z_{1,t}, \ldots, z_{K,t}, \hat{g}_{|t(t-1)}\}$.[6] That is, the hedonic characteristics in the z vectors are for properties sold in time t, and the prediction of the locational price

[6] $\mathbf{1}$ is an $N_t \times 1$ vector of ones.

effect is of the properties sold in time t computed using the model fitted at time $t-1$. To reflect this, we used the notation $\hat{g}_{|t(t-1)}$, which should be read as the estimated locational price effect of properties sold in time t as predicted by model (8) fitted in $t-1$.

$$\alpha_{t|t} = \alpha_{t|t-1} + G_t\{y_t - X_t^1\alpha_{t|t-1}\} \tag{11}$$

The mean square error matrix given information up to time period t is $P_{t|t} = P_{t|t-1} - G_t X_t P_{t|t-1}$, and the Kalman gain is given by,

$$G_t = P_{t|t-1}X_t'\{H_t + V_{|t} + X_t P_{t|t-1}X_t'\}^{-1} \tag{12}$$

The updating equations are of the standard form, $\alpha_{t|t-1} = D\alpha_{t-1|t-1}$ and $P_{t|t-1} = DP_{t-1|t-1}D' + Q$.

Estimates of the state (11), $\hat{\alpha}_{t|t}$, are obtained by replacing H_t, Q, D, and $V_{|t}$, by suitable estimates.

The model given by the state-space representation in Eqs. (9)–(10) with associated estimates from (11) and (12) will be referred to as SS+GAM.

2.3 Constructing the predictions

The computation of the index (3) depends crucially on the prediction of log price. Given the spatio-temporal features of the model, the prediction of the log price for property i is given by the natural predictor plus a correction term as follows,[7]

$$\widehat{y_{i,t|t}}^p = x_{i,t}'\alpha_{t|t} + c_{vt,i}'\Omega_{|t}^{-1}e_t \tag{13}$$

where $\widehat{y_{i,t|t}}^p$ is the predicted log price for property i and $\alpha_{t|t}$ is the state vector at period t conditional on information up to and including time period t; $\Omega_{|t} = \text{cov}\{y_t, y_t\}$; $c_{vt,i}' = E(v_{i|t}, v_{|t})$ is the row of $V_{|t}$ corresponding to property i and has elements $c_{vt,ij} \equiv E\{v_{i|t}v_{j|t}\} = \{c_v(i,j_1),\ldots,c_v(i,j_{N_t})\}'$ which could be equal to zero for $i \neq j$; $e_t = y_t - E(y_t)$.

To show this result, in addition to assumptions already stated, we assume $v_{i|t}$ and y_t have a joint multivariate normal distribution. Taking the characteristics and location of properties as given, the predictor is derived as follows,

$$\begin{aligned}
\widehat{y_{i,t|t}}^p &= E\{y_{it}^*|y_t, y_{t-1},\ldots,y_1\} \\
&= E\{x_{i,t}'\alpha_t + v_{i|t}|y_t, y_{t-1},\ldots,y_1\} \\
&= x_{i,t}'E\{\alpha_t|y_t, y_{t-1},\ldots,y_1\} + E\{v_{i|t}|y_t, y_{t-1},\ldots,y_1\} \\
&= x_{i,t}'\alpha_{t|t} + c_{vt,i}'\Omega_{|t}^{-1}e_t
\end{aligned}$$

The last term is of this form since $E\{v_{i|t}y_{jt}\} = c_{vt,ij}$.

[7] The term is similar to that derived by Goldberger (1962) for the first-order autoregressive model.

In this study, we implement this prediction by defining $\widehat{v_{i|t}} = y_{i|t} - \hat{y}_{i|t}$, which are the residuals from estimating (8) and $e_{it} = y_{it} - x'_{i,t}\hat{\alpha}_{t|t}$, which are the residuals from the estimated state-space model, where $x'_{i,t}\hat{\alpha}_{t|t}$ is the state-space prediction of the (log) price of property i at time t.

For the index calculation, predictions of the prices are needed. Replacing by suitable estimates on the right-hand side of the expressions and reverting the logarithmic transformation, the prediction of the price of property i sold in period $t = 1, \ldots, T$ is defined as

$$\hat{p}_{t,i}(z'_{i,t}, \hat{g}_{i|t(t)}) = \exp(x'_{i,t}\hat{\alpha}_{t|t} + \hat{c}'_{vt,i}\hat{\Omega}^{-1}_{|t}e_t), \tag{14}$$

and the prediction of the price of property i sold in period t for period $t-1$ is given by

$$\hat{p}_{t-1,i}(z'_{i,t}, \hat{g}_{i|t(t-1)}) = \exp({x^1_{i,t}}'\hat{\alpha}_{t-1|t-1} + \hat{c}'_{v(t-1),i}\hat{\Omega}^{-1}_{|t}e_{t(t-1)}) \tag{15}$$

The crucial point is that the constructed location effect and parameters need to be matched with the correct period for which the prediction is being made. In this case, $\hat{g}_{i|t(t-1)}$ enters in ${x^1_{i,t}}'$, $c'_{v(t-1),i}$ and together with $\hat{\alpha}_{t-1|t-1}$ in $e_{t(t-1)}$. Estimates of the location spline for $j = -1, 0, 1$, and $\hat{g}_{|t(t+j)}$, respectively, obtained from (8), are used to implement the predictions to construct the index.[8]

2.4 Specification and robustness

2.4.1 Postcodes versus spline surfaces

We use two alternative models from the literature to compare to the model presented in Sect. 2.2, SS+GAM. The first is the generalized additive hedonic model (proposed by Hill and Scholz 2018). The semi-parametric model (8) is estimated separately for each week.[9] The model (and corresponding index) will be referred to as GAM.

The second is a model where location is controlled by postcode dummies. As mentioned in Sect. 1, this is a common specification used in the price index literature to control for location. The model is given by

$$y_t = \mu_t + Z_t\beta_t + D_t\pi_t + \varepsilon_t \tag{16}$$

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

[8] The notation used in the subscript "$_{|t(t+j)}$" was explained in the paragraph above Eq. (11).

[9] A period-by-period estimation of the model is standard practice in the hedonic imputation literature.

Computing hedonic imputation price indices using period-by-period estimation with (16) 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 (details provided in Appendix A.3). We will refer to this model (and corresponding index) by SS+PC.

One interesting test to compare alternative model specifications is the non-nested test proposed by Goodman and Thibodeau (2003) to study submarkets. The Goodman–Thibodeau test derives from the J test of Davidson and MacKinnon (1981). This test is for linear models with fixed parameters estimated by least squares. However, in our setting, the models are semi-parametric (GAM), including a generated regressor (SS + GAM), and the parameters are time-varying (all three). These characteristics of our modelling imply the J test is not directly applicable, as its distribution in these settings is unknown.

For this reason, we follow a different approach here. We compare our three specifications (GAM, SS + GAM, SS + PC) by computing root-mean-square errors of log predictions (RMSPE) at different important geographical locations in the city and for the whole city, as follows:

$$\text{RMSPE} = \sqrt{\frac{1}{N_R}\sum_{i=1}^{N_R}(\ln \hat{p}_i - \ln p_i)^2}, \tag{17}$$

where N_R denotes the number of price observations within a particular region. We choose a given point (the entrance of Sydney harbour or Bondi beach in our empirical example) and find all observations in the sample that are located within a given radius distance from that point. These comparisons allow us to assess both the global fit of each model and the local fit in the vicinity of important boundaries such as beaches.

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 fit of the hedonic model and the performance of the resulting price index. Ultimately, it is the latter that matters more. Hence, performance criteria should focus more on the Törnqvist index defined in (3), rather than the within-period fit of the hedonic model itself. Guo et al. (2014) and Jiang et al. (2015) take a similar view. Guo et al. (2014) suggest criteria based on the autocorrelation and volatility of the index, and Jiang et al. (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 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 predicted 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 predicted 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 predicted price relatives. Following Hill and Scholz (2018), the key insight is that repeat-sales price relatives can be used as a benchmark for evaluating the predicted price relatives.[10] To ensure a large enough sample size, repeat-sales price relatives over any time horizon in our data set are compared to their predicted counterparts.

More formally, suppose property i sells in both periods t and $t+k$. For this property, therefore, we have an observed repeat-sales price relative: $p_{i,t+k}/p_{i,t}$. The corresponding predicted price relative is $\hat{p}_{i,t+k}/\hat{p}_{i,t}$. The subsample of properties that have repeat sales is indexed by $i = 1, \ldots, N_{RS}$. We can now define the ratio of predicted to actual price relative for house i as follows:

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

Our quality measure is given by the mean squared error (MSE) of the log predicted price relatives on the repeat-sales sample of each hedonic method:

$$\begin{aligned} \text{MSE(RS)} &= \left(\frac{1}{N_{\text{RS}}}\right) \sum_{i=1}^{N_{\text{RS}}} \left[\ln\left(\frac{\hat{p}_{i,t+k}}{\hat{p}_{i,t}}\right) - \ln\left(\frac{p_{i,t+k}}{p_{i,t}}\right) \right]^2 \\ &= \left(\frac{1}{N_{\text{RS}}}\right) \sum_{i=1}^{N_{\text{RS}}} [\ln(d_i)]^2, \end{aligned} \tag{19}$$

where the summation in (19) takes place across the whole repeat-sales sample.[11] We prefer whichever hedonic imputation model generates the smallest MSE, on the grounds that the resulting Törnqvist indices will be constructed from the most reliable predicted price relatives.

[10] Focusing on repeat sales potentially creates a sample selection problem. Starter homes typically transact more frequently than other properties, which could cause a *lemons bias* in a repeat-sales price index (see Clapp and Giaccotto 1992; Gatzlaff and Haurin 1997; and Shimizu et al. 2010). However, in our context, lemons bias is not so much of an issue since we are comparing matched actual and predicted price changes on the same properties. In other words, any lemons bias will apply equally to both the actual and predicted samples being compared.

[11] The reason for taking logs of the price relatives in (19) is so that over and under predictions of the price relatives are treated symmetrically. Ideally the predicted and actual price relatives should be the same, and hence $d_i = 1$. Suppose instead that $d_i = 1/d_j$. In this case, the predictions for properties i and j should be viewed as equally inaccurate. Taking logs before squaring ensures that $[\ln(d_i)]^2 = [\ln(d_j)]^2$.

Table 1 Summary of characteristics

	PRICE	AREA	LAT	LONG	FREQ.	BED	BATH
Minimum	56,500	100.0	−34.20	150.6	1:	1348	190,395
1st Quartile	420,000	461.0	−33.93	150.9	2:	38,578	174,161
Median	610,000	587.0	−33.84	151.0	3:	200,428	57,673
Mean	784,041	626.1	−33.85	151.0	4:	147,794	8835
3rd Quartile	900,000	720.0	−33.76	151.2	5:	38734	1746
Maximum	3,200,000	4998.0	−33.40	151.3	6:	6320	392
Minimum allowed	50,000	100.0	−34.20	150.60			
Maximum allowed	4,000,000	5000.0	−33.40	151.35			

Price is measured in Australian dollars. Area is land area measured in square meters. The last two rows show the thresholds that were applied to delete outliers. The last two columns show the frequency of observations under each size bedrooms and bathrooms

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.[12] 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 Table 1, and a plot of the number of sales per week is shown in Fig. 1. As can be seen from Fig. 1, the number of transactions falls very significantly each year during the summer holiday period from mid-December to late January. Any method for computing weekly indices needs to be able to handle such seasonal fluctuations in transactions volume.

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 also shown in Table 1. Complete data on all our hedonic characteristics are available for 433,202 observations. The quality of the data improves over time. In particular, missing characteristics are quite common in the first 2 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

[12] Indices are now increasingly also being computed using listing prices, which can often be scraped online. Listing price indices will tend to differ slightly from transaction-based indices (Haurin et al. 2010). Our spatio-temporal hedonic approach can equally well be applied to listing price data. However, our mean square error quality measure in (19) can only be applied to listing data when it is possible to identify repeat listings of the same property.

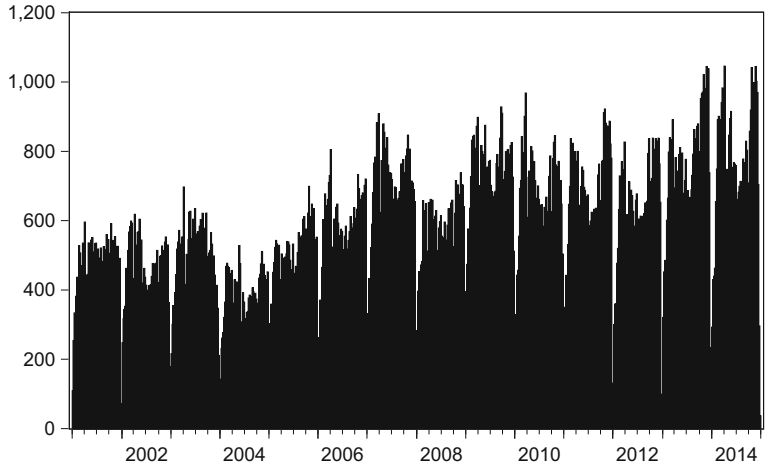

Fig. 1 Number of transactions per Week, 2001–2014

but compute the log likelihood function with all weeks in the 2003–2014 period (see Appendix Sects. A.2 and A.3).

3.2 Empirical model specification and parameter estimates

In all three models (GAM, SS+GAM, SS+PC), the dependent variable is the natural logarithmic transformation of the observed contract sale price. The hedonic characteristics included in all models are: $Z_t = (\ln Land, DBED3, DBED4, DBED5, DBATH2, DBATH3)$ where

Land: ln *Land* is the logarithmic transformation of the land area in sq. mts.

Bedrooms: DBED3 =1 if the number of bedrooms is equal to three; DBED4 =1 if the number of bedrooms equals to four; DBED5=1 if the number of bedrooms equals five or more. The models' intercepts capture houses with one or two bedrooms.

Bathrooms: DBATH2=1 if the number of bathrooms equals to two; DBATH3=1 if the number of bathrooms equals three or more. The models' intercept capture houses with one bathroom.

We compare the parameter estimates from the three models in Fig. 2 and Table 2.

The GAM estimates are much more volatile than their SS+GAM and SS+PC counterparts as expected. Estimating via a state-space representation provides a linking of the parameters overtime and reduces greatly the effect of the change in the composition of properties across periods. As transacted properties are not random samples of the market at each period, the effect of sales composition together with small samples in some periods can lead to this high volatility in the shadow price parameter estimates which should not be there in theory. We note that in a number of periods and across all the $\hat{\theta}^{\dagger}_{k|t}$, there are estimates that fall outside the 95% bound of $\hat{\alpha}_{k,t|t}$ (not shown but available from the authors). These can have potentially important implications for the index constructed using predictions from this model (presented in the next section).

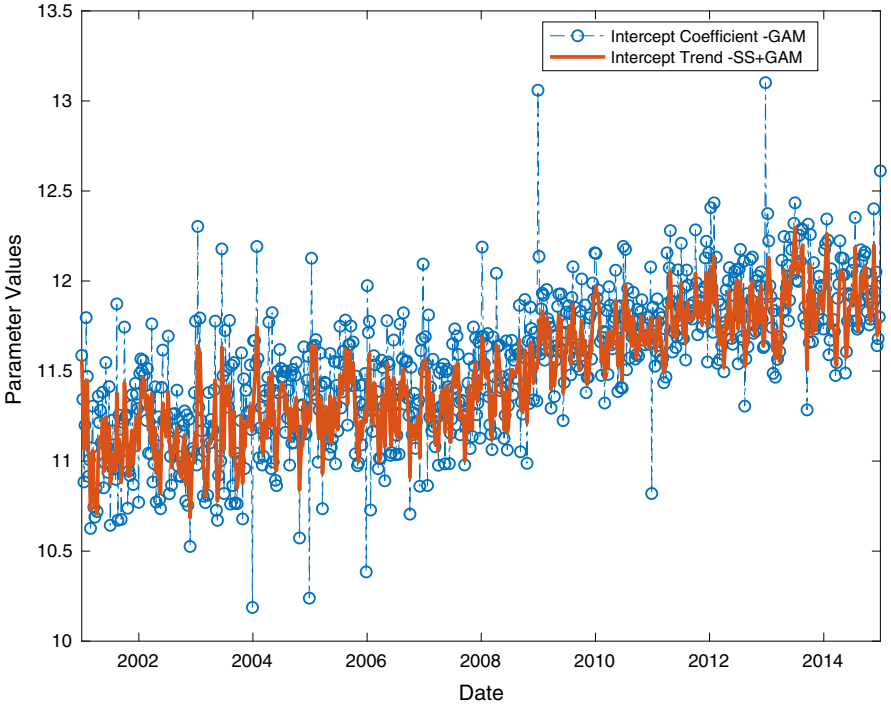

Fig. 2 Comparison of intercept trends (GAM and SS+GAM)

One additional parameter of the spatio-temporal representation is the process γ_t, which we modelled as possibly mean reverting unlike the shadow price parameters of hedonic characteristics. This process 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, and is mean reverting if the estimate of $|\rho| < 1$ [see Eq. (10)]. We found the estimate of ρ to be 0.4. Summary statistics of the estimates $\hat{\gamma}_t$ are shown in the last row of Table 2.

The RMSPE, as defined in (17), is shown in Table 3. The focus here is on assessing the extent to which our models capture locational effects. Each RMSPE is computed using the data from all time periods. The differences relate to geographical scope. The geographical scopes considered are: the whole city, within a 5km radius of the entrance of Sydney harbour, within a 2.5km radius of Bondi beach, and within a 30km radius of the Blue Mountains (while still being within Sydney). In all cases, SS+PC performs worst. Over the whole city, SS+GAM performs best. SS+GAM also performs best in the vicinity of Sydney harbor and in the vicinity of Bondi beach. GAM performs best in the vicinity of the Blue Mountains. So while over the whole city, SS+GAM generates the best price predictions, it does not dominate GAM in all locations.

Table 2 Comparison of time-varying parameter estimates across models

	Min	Max	Mean	Std_dev
Intercept_GAM	10.1872	13.1015	11.5346	0.4024
IntercTrend_SS+GAM	10.6850	12.2980	11.4770	0.3327
IntercTrend_SS+PC[a]	11.0230	12.8330	12.2250	0.3638
DBED3_GAM	−0.1238	0.5769	0.1040	0.0452
DBED3_SS+GAM	0.0070	0.1797	0.1105	0.0259
DBED3_SS+PC	−0.1057	0.1237	0.1023	0.0129
DBED4_GAM	0.0209	0.6095	0.2042	0.0501
DBED4_SS+GAM	0.1305	0.3058	0.2124	0.0301
DBED4_SS+PC	0.1013	0.2280	0.2014	0.0123
DBED5_GAM	−0.3828	0.7486	0.2741	0.0847
DBED5_SS+GAM	0.0062	0.4286	0.2806	0.0467
DBED5_SS+PC	0.0795	0.3053	0.2707	0.0178
DBATH2_GAM	0.0085	0.2854	0.1166	0.0299
DBATH2_SS+GAM	0.0063	0.1828	0.1098	0.0235
DBATH2_SS+PC	0.0727	0.1721	0.1063	0.0182
DBATH3_GAM	0.0911	0.8150	0.3065	0.0666
DBATH3_SS+GAM	0.1372	0.4424	0.2908	0.0488
DBATH3_SS+PC	0.1853	0.3729	0.2885	0.0419
lnLand_GAM	0.0012	0.4793	0.2469	0.0572
lnLand_SS+GAM	0.1462	0.3730	0.2558	0.0430
lnLand_SS+PC	0.1114	0.3143	0.2461	0.0300
$\hat{\gamma}_t$	0.4242	0.6150	0.5404	0.0083

[a] This trend is not comparable to that of the GAM or SS+GAM models as it captures the price movement of the market for houses with one or two bedrooms and one bathroom for the first postcode in the sample. The intercept parameter in the GAM and trend in SS+GAM capture the price movement of the market for houses with one or two bedrooms and one bathroom for the whole city

3.3 Property price indices

We construct hedonic price indices from the three models (GAM, SS+GAM, SS+PC), and in addition a repeat-sales index and a quality unadjusted index.

There are 80,060 repeat sales in our dataset. However, we 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 6 months on the grounds that this suggests that the first purchase was by a professional renovator.[13] Finally, for houses that sold more than twice during our sample period (2001–2011), we only include the two chronologically closest repeat sales (as long as these are more than six

[13] Exclusion of repeat sales within 6 months is standard practice in repeat-sales price indices such as the Standard and Poor's/Case-Shiller (SPCS) Home Price Index.

Table 3 Model prediction and index quality comparison

Radius	Model RMSPE				Index MSE(RS)	
	Sydney	Harbour 5 Km	Bondi beach 2.5 Km	Blue Mountains 30 Km	Weekly	Monthly
GAM	0.1857	0.3136	0.3008	0.1260	0.0233	0.0245
SS+GAM	0.1775	0.3067	0.2954	0.1315	0.0102	0.0112
SS+PC	0.2088	0.3518	0.3239	0.1540	0.0246	0.0264
Sample	433202	13222	6950	19089		

The mean square prediction error of prices (RMSPE) is uniformly higher for the model with postcodes across all geographical alternatives. Similarly, the mean square error of the prediction of price relatives (MSE(RS)) is higher at both the SS+PC at both weekly and monthly frequency. The RMSPE is lowest for the SS+GAM model except in one case (the Blue Mountains) when GAM is the lowest. The SS+GAM is uniformly the lowest in MSE(RS) for both weekly and monthly frequencies

months apart). This ensures that all repeat-sales houses exert equal influence on our results.

We compute all indices for the sample (2003–2014) although the state-space models are estimated for the full sample as indicated in the previous section. The cleaned repeat-sales sample for this period has 61 024 observations. Figure 3 shows the three hedonic indices (chained), the repeat-sales index calculated using the standard formula from Bailey et al. (1963), and the quality unadjusted price index computed from the median of the prices of observed sales in each week. The median index is both 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 may occur with the conventional hedonic imputation method when the market is thin as small samples and sales' composition in thin markets can affect the parameter estimates and lead to large changes in the price relatives. This is clearly the case in this instance as is discussed in Sect. 3.2. Chaining can then compound this drift. Rambaldi and Fletcher (2014) find chain drift occurs in monthly indices even when using a two-month rolling window to estimate the parameters of the 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 Fig. 3 are larger than one might expect to observe in hedonic indices computed at annual or quarterly frequency (Hill and Scholz 2018). To illustrate this point, we have estimated the three models and computed all the indices at a quarterly frequency. Figure 4 shows that at a quarterly frequency, our hedonic indices approximate each other quite closely. Hence, it can be seen that the choice of hedonic method is of greater importance when indices are computed at higher frequencies, such as weekly. We explore this point further in the

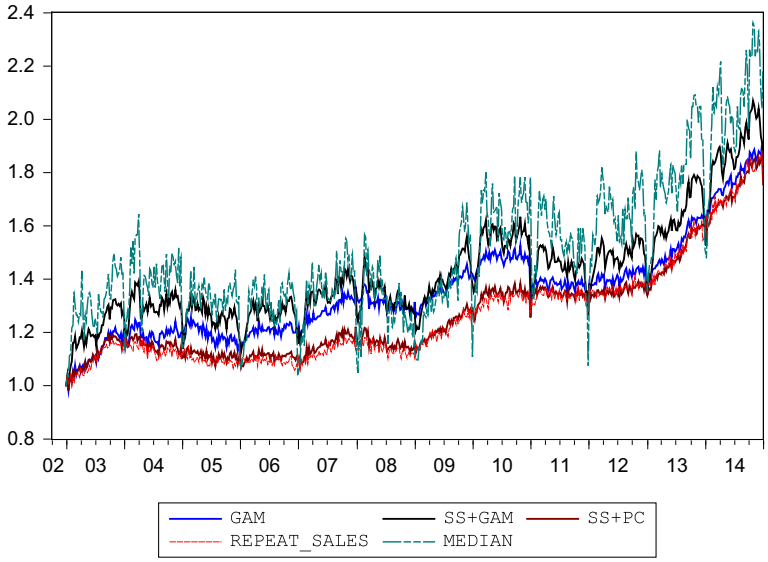

Fig. 3 Weekly property price indices from 2003 to 2014. *Note* GAM is based on periodwise estimation of model (8); SS+PC is the state-space model (16) with postcode dummies; SS+GAM is the spatio-temporal model; Repeat_Sales index is calculated using the Bailey et al. (1963) formula; Median is the usual median index computed at a weekly frequency. Base: Week starting 30/12/2002 = 1

next section by comparing the indices quality at the weekly and monthly frequencies using the MSE(RS) measure.

3.4 Comparing the quality of the indices

The performance of our three indices according to the mean square error of the predicted price relatives, MSE(RS), on the repeat-sales sample is shown in Table 3. We compute MSE(RS)'s based on weekly and monthly indices. In both cases, the ranking of methods is the same. 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 test whether the MSE(RS)s are significantly different across different hedonic models. It is clear from Eq. (19) that these measures are averages. Thus, the null hypothesis is that the true difference between two means is zero ($H_0 : \text{MSE(RS)}_{M1} - \text{MSE(RS)}_{M2} = 0$), where $M1$ and $M2$ denote two of the hedonic models (e.g. GAM and SS+GAM). To decide on a suitable statistic to conduct the test, we note that there is no dependence structure in the computed MSE(RS) values as we do not include repeat sales that are within 6 months of each other. Each pair in the sample is for a unique dwelling and for each pair the length between sales varies. Scatter plots of the $[\ln(d_i)]^2$ confirm there are no patterns.[14] Thus, based

[14] These scatter plots are available from the authors upon request.

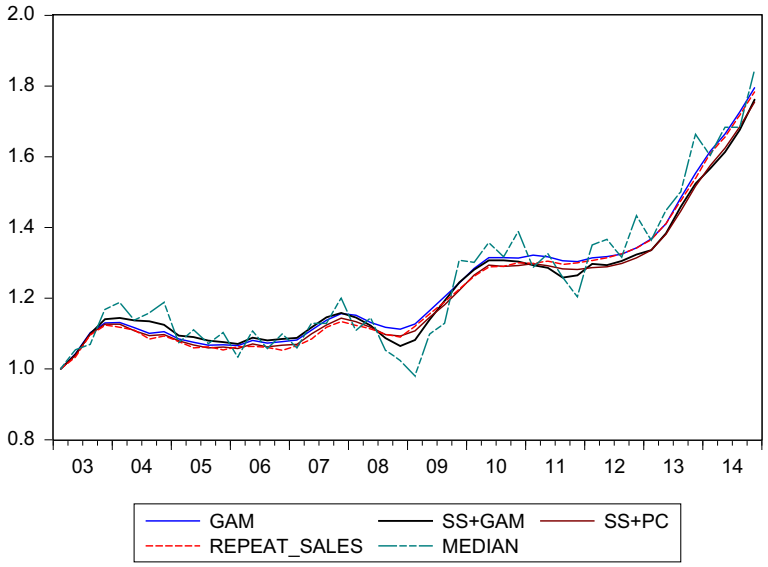

Fig. 4 Quarterly property price indices from 2003 to 2014. *Note*: GAM is based on periodwise estimation of model (8); SS+PC is the state-space model (16) with postcode dummies; SS+GAM is the spatio-temporal model; Repeat_Sales index is calculated using the Bailey et al. (1963) formula; Median is the usual median index computed at a quarterly frequency. Base: First Quarter 2003 = 1

Table 4 p values for H_0: $\text{MSE(RS)}_{M1} - \text{MSE(RS)}_{M2} = 0$

	Weekly	Monthly
SS+PC vs. SS+GAM	0.0000	0.0000
SS+PC vs. GAM	0.0483	0.0014
GAM vs. SS+GAM	0.0000	0.0000

These p values imply that SS+GAM is highly significantly different from both SS+PC and GAM at both the weekly and monthly frequencies

on the central limit theorem (see, for example, pp. 490–491 in Devore and Berk 2012),

$$\text{MSE(RS)}_{M1} - \text{MSE(RS)}_{M2} \sim \mathcal{N}\left(0, \frac{s_1^2 + s_2^2}{N_{\text{RS}}}\right),$$

where $s_j (j = 1, 2)$ is the sample standard deviation of the D_j for hedonic model j. The computed two-sided p values of this exercise are presented 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. The SS+GAM method generates the most accurate matched sample price relatives. The Törnqvist

price index as defined in (3) is computed by taking a geometric mean of these price relatives.

4 Conclusion

This article has focused 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 state-space model that controls for location with a geospatial spline surface. Estimation of the model requires a modified form of the Kalman filter. The geospatial spline surface replaces postcode dummies which are more commonly used to control for location in the hedonic price index literature. The use of the spline achieves greater precision and a large reduction in the dimensionality of the spatio-temporal model. Predicted prices are obtained using an adjusted form of the natural predictor. These predicted prices provide a matched sample, thus allowing the price index to be computed using the superlative Törnqvist price index formula. 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 lower frequencies such as quarterly. Hence, it is at these higher frequencies that the choice of hedonic method matters most. It is then shown, based on a criterion that compares the predicted price relatives with actually observed repeat-sales price relatives, that our preferred method for computing weekly indices outperforms alternative hedonic imputation methods.

Acknowledgements Open access funding provided by University of Graz.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

A Estimation and algorithms

A.1 Step I: Location spline and spatial uncertainty

This step provides estimates of $g_{|t}(.)$ and $V_{|t}$ using spatial only information given by the sample of N_t transacted houses at each period. A semi-parametric hedonic model

with the specification in (8) is 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. The estimates of the spline surface, $\hat{g}_{|t}$, enter the spatio-temporal model's X_t matrix; while the vector of N_t predictions from this model, $\hat{y}_{|t}$, provide residuals, $\widehat{v_{i|t}} = y_{it} - \hat{y}_{i|t}$, from which a sample estimate of $V_{|t}$, and thus $c'_{vt,i}$, makes operational the correction term in (14) and (15).

A.1.1 Generalized additive model: GAM implementation

The semi-parametric hedonic model in (8) is an example of a generalized additive model (GAM). Here, we follow Hill and Scholz (2018) when estimating the GAM (see Appendix A.3 of Hill and Scholz 2018, for further details). The GAM is estimated using the `mgcv` package (see Wood 2006) of the statistical software R 3.4.3 (R Core Team 2017). The values of two parameters k and n (where $k < n$) must be chosen by the researcher, where k is the rank approximation to the smoothing spline, and n is the number of observations over which the spline is fitted.

It is important that k is not too small; otherwise, it would force oversmoothing. On the other hand, using a too large value enormously increases the computational burden without necessarily improving the fit. Typically the number of covariate values used will be substantially smaller than the number of data points, and substantially larger than the basis dimension, k. For our house price data, we set $k = 600$ and a subsample of size $n = 2500$ when possible. For weeks when the number of observations is less than 2500, we use the available sample and choose k to be $n - 6$ (six is the number of parameters in our model). This approximation is required to reduce the computational burden to a manageable level. (The default choice of n in the `bam`-function is 2000 observations).

Hill and Scholz (2018) estimated a model at annual frequency and conducted an experiment to test the approximation. They found a strong decline in the mean square error up to $k = 600$ and a slight increase afterward. They also found an almost linear increase in computational time indicating that, for example, the use of $k = 900$ would give a similar mean square error as $k = 600$, but the computational cost would increase by 50 percent.

It is expected that for weeks when the sample is smaller than 1000 observations, the spline surface will be adjusted more by the γ_t parameter in step II.

A.2 Step II: Estimation of the state-space model for prediction

This step produces estimates of the state vector, α_t (and its mean squared prediction matrix, $P_{t|t}$) via the algorithm outlined in Sect. 2.2.1 [see Eqs. (11) and (12)]. The algorithm requires estimates of D, H_t, Q. These can be obtained by maximum likelihood estimation. Given y_t, $Z_t = \{z_{1t}, \ldots, z_{Kt}\}$, $\hat{g}_{|t}(\cdot)$ and $\hat{V}_{|t}$, the Kalman filter algorithm is run to evaluate the log likelihood, $\ln L$, in predictive form (details are provided next).

The model is given in Eqs. (9) and (10).

The estimator's algorithm is a function of a prediction error, $v_{t|t-1} = y_t - X_t^1 \hat{\alpha}_{t|t-1}$, and the Kalman Gain (12), which is a function of $F_t = E(v_{t|t-1} v'_{t|t-1}) = H_t + \hat{V}_{|t} + X_t P_{t|t-1} X'_t$.

Both $v_{t|t-1}$ and F_t are obtained by running the Kalman filter,

$$\begin{aligned}
\alpha_{t|t} &= \alpha_{t|t-1} + G_t\{y_t - X_t^1 \alpha_{t|t-1}\} \\
P_{t|t} &= P_{t|t-1} - G_t X_t P_{t|t-1} \\
G_t &= P_{t|t-1} X'_t \{H_t + \hat{V}_{|t} + X_t P_{t|t-1} X'_t\}^{-1} \\
\alpha_{t|t-1} &= D\alpha_{t-1|t-1}, \text{ and} \\
P_{t|t-1} &= D P_{t-1|t-1} D' + Q.
\end{aligned}$$

The log likelihood in prediction form is given by,

$$\begin{aligned}
&\ln L(\rho, \sigma_\epsilon^2, \sigma_\mu^2, \sigma_\beta^2, \sigma_\gamma^2; y_t, Y_{t-1}, Z_t, \hat{g}_t, \hat{g}_{t(t-1)}, V_{|t}) \\
&= -\frac{NT}{2} ln(2\pi) - \frac{1}{2} \sum_{t=d}^{T} ln|F_t| - \frac{1}{2} \sum_{t=d}^{T} v'_{t|t-1} F_t^{-1} v_{t|t-1}
\end{aligned}$$

where $Y_{t-1} = y_{t-1}, y_{t-2} \ldots$. We use a standard Newton–Raphson algorithm to estimate $\hat{\sigma}_\epsilon^2, \hat{\sigma}_\mu^2, \hat{\sigma}_\beta^2$ and $\hat{\sigma}_\gamma^2$ within a grid search for ρ in the range (0.1 to 1); $N = \sum_{t=d}^{T} N_t$; d is sufficiently large to avoid the log likelihood being dominated by the initial condition, $\alpha_0 \sim N(a_0, P_0)$. In the empirical implementation, we have 731 weeks and set $d = 105$ (with this choice, data from 2003 onward are used to estimate these hyperparameters). For details on estimation of state-space models, see Harvey (1989) or Durbin and Koopman (2012).

A.3 SS+PC state-space

The system is given in Eq. (16) with $H_t = \sigma_\epsilon^2 I_{N_t}$ and a set of transition equations

$$\alpha_t = \alpha_{t-1} + \eta_t$$

where $\alpha_t = \{\mu_t, \beta_{1t}, \ldots, \beta_{K,t}, \pi_2, \ldots, \pi_{N_{pc}}\}'$ and N_{pc} number of postcodes in the dataset

η_t has variance–covariance

$$Q = \begin{bmatrix} \sigma_\mu^2 & 0 & 0 \\ 0 & \sigma_\beta^2 I_K & 0 \\ 0 & 0 & \sigma_\pi^2 I_{N_{pc}} \end{bmatrix}$$

The Kalman filter in this case is of a standard form,

$$\begin{aligned}
\alpha_{t|t} &= \alpha_{t|t-1} + G_t\{y_t - X_t\alpha_{t|t-1}\} \\
P_{t|t} &= P_{t|t-1} - G_t X_t P_{t|t-1} \\
G_t &= P_{t|t-1}X_t'\{H_t + X_t P_{t|t-1}X_t'\}^{-1} \\
\alpha_{t|t-1} &= \alpha_{t-1|t-1}, \text{ and} \\
P_{t|t-1} &= P_{t-1|t-1} + Q.
\end{aligned}$$

The log likelihood in prediction form is given by,

$$\begin{aligned}
&\ln L(\sigma_\epsilon^2, \sigma_\mu^2, \sigma_\beta^2, \sigma_\pi^2; y_t, Y_{t-1}, Z_t) \\
&= -\frac{NT}{2}ln(2\pi) - \frac{1}{2}\sum_{t=d}^{T} ln|F_t| - \frac{1}{2}\sum_{t=d}^{T} v_{t|t-1}' F_t^{-1} v_{t|t-1} \qquad (20)
\end{aligned}$$

We use a standard Newton-Raphson algorithm to estimate $\hat{\sigma}_\epsilon^2$, $\hat{\sigma}_\mu^2$, $\hat{\sigma}_\beta^2$ and $\hat{\sigma}_\pi^2$; $N = \sum_{t=d}^{T} N_t$; d is sufficiently large to avoid the log likelihood being dominated by the initial condition, $\alpha_0 \sim N(a_0, P_0)$. In the empirical implementation, we have 731 weeks and set $d = 105$ (with this choice, data from 2003 onward is used to estimate these hyperparameters). 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.

B Model specification and functional form

The use of dummy variables across sizes of bedrooms and bathrooms provides flexibility and is much more informative on the size of the dwelling than using the number of bedrooms and number of bathrooms as single metric variables.

Nevertheless, we have tested this alternative specification using data for a year in fixed parameters postcodes dummy model and GAM model using relevant Wald tests. We strongly reject the specification using the bathroom and bedrooms in favour of the dummy specification. We also tested the functional form of land between a log linear and a log–log form and found support for the second, which is also the common choice in the literature.[15]

The choice of specifying location of the dwellings in the model (postcodes or spline) is dealt with by the computation of RSMPE and MSE(RS) which uniformly reject the postcode alternative (see Tables 3 and 4 in the body of the paper).

References

Bailey M, Muth R, Nourse R (1963) A regression method for real estate price index construction. J Am Stat Assoc 58:933–942

[15] The p values of the conducted tests are less than any standard level of significance.

Bokhari S, Geltner D (2012) Estimating real estate price movements for high frequency tradable indexes in a scarce data environment. J Real Estate Finance Econ 45(2):522–543
Bollerslev T, Patton AJ, Wang W (2016) Daily house price indices: construction, modeling, and longer-run predictions. J Appl Econ 31:1005–1025
Bourassa SC, Hoesli M (2016) High frequency house price indexes with scarce data. Swiss Finance Institute Research Paper Series, pp 16–27
Clapp JM, Giaccotto C (1992) Estimating price trends for residential property: a comparison of repeat sales and assessed value methods. J Real Estate Finance Econ 5(4):357–374
Core Team R (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/
Court AT (1939) Hedonic price indexes with automotive examples, in the dynamics of automobile demand. New York: The General Motors Corporation, pp 99–117
Davidson R, MacKinnon J (1981) Several tests for model specification in the presence of alternative hypotheses. Econometrica 49(3):781–793
de Haan J (2010) Hedonic price indexes: a comparison of imputation, time dummy and re-pricing methods. J Econ and Stat (Jahrbuecher für Nationalökonomie und Statistik) 230(6):772–791
Devore JL, Berk KN (2012) Modern mathematical statistics with applications, 2nd edn. Springer, New York
Diewert WE (2010) Alternative approaches to measuring house price inflation. Discussion Paper 10-10, Department of Economics, The University of British Columbia, Vancouver, Canada, V6T 1Z1, 2010
Diewert WE (1976) Exact and superlative index numbers. J Econ 4:115–145
Diewert WE, Shimizu C (2015) Residential property price indices for Tokyo. Macroecon Dyn 19:1659–1714
Durbin J, Koopman S (2012) Time series analysis by state space methods, 2nd edn. Oxford statistical science series. Oxford University Press, Oxford
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 DH, Haurin DR (1997) Sample selection bias and repeat-sales index estimates. J Real Estate Finance Econ 14:33–50
Geltner D, Ling D (2006) Considerations in the design and construction of investment real estate research indices. J Real Estate Res 28(4):411–444
Goldberger AS (1962) Best linear unbiased prediction in the generalized linear regression model. J Am Stat Assoc 57:369–375
Goodman AC, Thibodeau T (2003) Housing market segmentation and hedonic prediction accuracy. J Hous Econ 12(3):181–201
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
Guo X, Zheng S, Geltner D, Liu H (2014) A new approach for constructing home price indices: the pseudo repeat sales model and its application in China. J Hous Econ 25:20–38
Harvey A (1989) Forecasting, structural time series models and the kalman filter. Cambridge University, Cambridge
Haurin DR, Haurin JL, Nadauld T, Sanders A (2010) List prices, sale prices and marketing time: an application to U.S. housing markets. Real Estate Econ 38(4):659–685
Hill RJ (2013) Hedonic price indexes for housing: a survey, evaluation and taxonomy. J Econ Sur 27(5):879–914
Hill RJ, Melser D (2008) Hedonic imputation and the price index problem: an application to housing. Econ Inq 46(4):593–609
Hill RJ, Scholz M (2018) Incorporating geospatial data in house price indexes: a hedonic imputation approach with splines. Rev Income Wealth 64(4):737–756
Jiang L, Phillips PCB, Yu J (2015) New methodology for constructing real estate price indices applied to the Singapore residential market. J Bank Finance 61:S121–S131
Koopman S, Harvey A (2003) Computing observation weights for signal extraction and filtering. J Econ Dyn Control 27:1317–1333
Melser D (2018) Scanner data price indexes: addressing some unresolved issues. J Bus Econ Stat 36(3):516–522

Rambaldi AN, Rao DSP (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 AN, Fletcher CS (2014) Hedonic imputed property price indexes: the effects of econometric modeling choices. Rev Income Wealth 60:S423–S448

Shimizu C, Nishimura KG (2007) Pricing structure in Tokyo metropolitan land markets and its structural changes: pre-bubble, bubble, and post-bubble periods. J Real Estate Finance Econ 35(4):475–496

Shimizu C, Nishimura KG, Watanabe T (2010) Housing prices in Tokyo: a comparison of hedonic and repeat sales measures. J Econ Stat (Jahrbuecher für Nationalökonomie und Statistik) 230(6):792–813

Silver M (2016) How to better measure hedonic residential property price indexes. International Monetary Fund Working Paper 16/213

Silver M, Heravi S (2007) The difference between hedonic imputation indexes and time dummy hedonic indexes. J Bus Econ Stat 25:239–246

von Auer L (2007) Hedonic price measurement: the CCC approach. Empir Econ 33:289–311

von Auer L, Trede M (2012) The dynamics of brand equity: a hedonic regression approach to the laser printer market. J Oper Res Soc 63:1351–1362

Wikle CK, Cressie N (1999) A dimension-reduced approach to space-time Kalman filtering. Biometrika 86:815–829

Wood SN (2006) Generalized additive models: an introduction with R. Chapman, New York