the review of income and wealth

Review of Income and Wealth
Series 60, Supplement Issue, November 2014
DOI: 10.1111/roiw.12143

HEDONIC IMPUTED PROPERTY PRICE INDEXES: THE EFFECTS OF ECONOMETRIC MODELING CHOICES

by Alicia N. Rambaldi*

University of Queensland

and

Cameron S. Fletcher

CSIRO Ecosystem Sciences

In this paper we consider how choices in the econometric approach to impute prices affect the Törnqvist and Jevons hedonic imputed indexes. We compare the rolling window approach to estimation by smoothing methods. The main difference between the rolling window and the smoothing methods is in the way information is weighted. We propose that the Kalman filter smoother is the most appropriate estimator for the task as it optimally weights current and past information. We show the rolling window approach does not produce estimates that are attenuated over time leading to chain drift in the index. We also compare two alternative specifications to model property location. The empirical section uses data from a small and homogeneous market in the state of Queensland, Australia. The Törnqvist and Jevons indexes differ in value during periods of market volatility. This seems expected given their different weighting of transactions and the likelihood that price movements of properties at the upper and lower end of the price distribution might differ during periods of market volatility.

JEL Codes: C33, C43, C50

Keywords: property price indexes, rolling windows, spatial errors, spatial regressors, state-space

1. Introduction

The provision of residential property price indexes (RPPIs) is based on combining suitable index numbers theory with available data. When the indexes are hedonic or repeat sales based, the mix also involves appropriate regression techniques. A recent and very comprehensive review of all alternative approaches to the computation of RPPI can be found in the *Handbook on Residential Property Price Indexes* (European Commission *et al*., 2013). The general recommendation of the Handbook is that hedonic imputed (HI) indexes are to be preferred when

Note: The authors acknowledge funding from the Australian Research Council (DP120102124). The data used were a subset of that from Fletcher *et al*. (2011), which were collected, cleaned, and checked with funding from the Department of Climate Change and Energy Efficiency, CSIRO Climate Adaptation Flagship, and the 2011/12 UQ Summer Scholarship Program funded by the School of Economics and The University of Queensland. Many thanks to Thi Thuy (Kelly) Dung Trinh and Kian Nam Loke for their work on the dataset, and to Ryan McAllister for data contributions. Two anonymous referees, and the editors of this special issue provided insightful comments that were very helpful.

*Correspondence to: Alicia N. Rambaldi, School of Economics, The University of Queensland, St Lucia, QLD 4072, Australia (a.rambaldi@uq.edu.au).

the required data are available (see chapter 12, p. 159 of the handbook, and also Silver and Heravi, 2007; Hill and Melser, 2008). The HI index construction method is a type of matching method in that to compute the index, each property is priced at two time periods (t and $t + s$, $s \neq 0$); however, unlike the repeat sales method, it does not require a matched sample. HI indexes do not assume that the hedonic characteristics of the property have remained constant across the two comparison periods, and thus all available transactions sales data can be used in computation of the index. This is an important statistical feature as it avoids having to discard large numbers of observations which can lead to the use of a sample that is not representative of the population, one of the drawbacks of the repeat sales method.

The computation of an HI index is based on a hedonic regression which is estimated to provide a prediction of the sale price of each property transacted at time t and an imputation of its sale price at the comparison period, $t + s$. The main objective of this paper is to study the robustness of the constructed HI indexes to choices made when specifying and estimating the hedonic regression. We compare the use of the popular rolling window approach (ROW) to using a smoothing based estimator (SM). It is shown that theoretically they differ in the way transactions are weighted to obtain parameter estimates and imputations. For the empirical part of the paper, we compare the indexes computed from the model estimated under alternative model specifications of property location, using spatial econometrics and/or spatial regressors.

The work of Triplett (2004) popularized the use of a "two-period adjacent" (or "rolling two-period") estimation of the hedonic model to impute prices for the computation of HI indexes in consumer goods. For the case of constructing RPPI, European Commission *et al.* (2013) discuss the use of M–period rolling windows, while Hill and Melser (2008) and Hill and Scholz (2013), estimate the model yearly without overlapping. For the purpose of this study we will use the terminology rolling window for all cases where the prediction of the price is based on the re-estimation of a regression model using two or more consecutive time periods and a rolling sample. ROW meets the requirement of not fixing the coefficients of the hedonic regression, leading to time-varying shadow prices, which is justified by economic theory (see Diewert, 2003). In addition, the resulting index does not need to be revised as new periods are added to the sample. However, the prediction performance of ROW, measured by the mean square prediction error, is dismal compared to that of a model with time-varying parameters estimated using an optimal statistical estimator such as the Kalman Smoother (Rambaldi and Rao, 2011, 2013). This paper shows how the ROW relates to the Kalman Filter Smoother (KF)[1] and the Kalman Smoother (KS). We will refer to the KF and KS as smoothing methods (SM) and in Section 4 we discuss the difference between them, and under what circumstances their application leads to revisions of the computed index.

The main difference between SM and ROW methods is how they weight the available sample information. This is a crucial issue because unlike consumer goods, the composition of sales (i.e., the composition of the observed sample) can

[1]This is the forward filter conditional on current and past information. See Section 4 for details.

vary greatly from one period to the next. How information is weighted by estimators is potentially crucial to the robustness of the computed index. The empirical results in this paper illustrate the occurrence of index chain drift when the ROW estimates are used to construct the index.

Even optimum estimators are affected by omitted variables in the model. Controlling for the location of the property is crucial when modeling property prices. The typical model used for the purpose of property price prediction includes characteristics of the property (size of the land, structure, number of bedrooms, etc.) as well as measures of its location. Property characteristics are dictated mainly by data availability; location on the other hand can be incorporated into the model in a number of ways. Two alternatives proposed in the RPPI literature are to use a spatial error covariance (Rambaldi and Rao, 2011) and to fit splines using the property coordinates (Hill and Scholz, 2013). A third possible alternative is to construct measures of distances to landmarks (e.g., distance to schools, train station, park) and then incorporate them as *spatial regressors* in the hedonic model. In this paper we compare the constructed indexes obtained from the model when location is specified under two alternatives: a spatial error (SEM); and the use of spatial regressors.

2. Hedonic Imputed RPPIs Used in This Study

Hill and Melser (2008), Hill (2011), and Rambaldi and Rao (2013) discuss a range of index number formulae that are based on different sets of weighting systems and on different sets of imputed prices. In this paper the general recommendations from these works are taken and two types of indexes are used, a Törnqvist and a Jevons index. These indexes "weight" information differently. A Jevons index equally weights transactions, as it uses geometric averages. A Törnqvist index, on the other hand, weights the relative value of each of the properties included in the sample.[2] The Törnqvist index is computed using *actual shares* based on actual prices as defined in (1), and *imputed* (or *predicted*) prices in both the base and current periods. This is known as a double imputation method in the price index literature. Single imputation methods combine the observed price with predictions used for the comparison period. The reader is referred to Hill (2011) for a comprehensive treatment.

Let P_t^h represent the sale price of house h in period t. Further, let w_t^h be the value share of house h defined as:

$$w_t^h = \frac{P_t^h}{\sum_{n=1}^{N_t} P_t^n} \tag{1}$$

where P_t^h is the observed sale price of house h and N_t is the number of houses sold in period t.

[2]Rambaldi and Rao (2011, 2013) labeled these "plutocratic and democratic" Törnqvist, respectively. However, here we use the more conventional index number definition. We thank the editor for pointing this out.

The hedonic Törnqvist index is defined as follows:

$$T_{t,t+s} = \sqrt{GL^{w}_{t,t+s} \times GP^{w}_{t,t+s}} \tag{2}$$

where $GL^{w}_{t,t+s}$ and $GP^{w}_{t,t+s}$ are the geometric Laspyeres and geometric Paasche indexes which are defined as:

$$GL^{w}_{t,t+s} = \prod_{h=1}^{N_t} \left[\frac{\hat{P}^{h}_{t+s}\left(x^{h}_{t}\right)}{\hat{P}^{h}_{t}\left(x^{h}_{t}\right)} \right]^{w^{h}_{t}} \tag{3}$$

$$GP^{w}_{t,t+s} = \prod_{h'=1}^{N_{t+s}} \left[\frac{\hat{P}^{h'}_{t+s}\left(x^{h'}_{t+s}\right)}{\hat{P}^{h'}_{t}\left(x^{h'}_{t+s}\right)} \right]^{w^{h'}_{t+s}} \tag{4}$$

where $\hat{P}^{h}_{i}\left(x^{h}_{t}\right)$ for $i = t$, $t + s$ is an imputation of the price of house h with vector of hedonic characteristics x^{h}_{t} at periods; $\hat{P}^{h'}_{i}\left(x^{h'}_{t+s}\right)$ is an imputation of the price of house h' with vector of hedonic characteristics $x^{h'}_{t+s}$.

N_t and N_{t+s} are the number of transacted houses in periods t and $t + s$, respectively. Although it is possible for a house to have sold in both periods, they are not overlapping samples in general.

These indexes might be influenced by properties with large price tags. Despite this, the index measures the changes in the housing stock value that can be attributed exclusively to price changes, and therefore provides useful information. Official consumer price indexes, for instance, are constructed with these types of weights because they track the change in the cost of what consumers in the aggregate are buying. If we want to track the change in the aggregate value of the housing stock for national accounts purposes these are appropriate weights.

The hedonic Jevons index consistent with the use of a log-price hedonic model is defined as:

$$J_{t,t+s} = \sqrt{GL_{t,t+s} \times GP_{t,t+s}} \tag{5}$$

$$GL_{t,t+s} = \prod_{h=1}^{N_t} \left(\left[\frac{\hat{P}^{h}_{t+s}\left(x^{h}_{t}\right)}{\hat{P}^{h}_{t}\left(x^{h}_{t}\right)} \right]^{\frac{1}{N_t}} \right) \tag{6}$$

$$GP_{t,t+s} = \prod_{h'=1}^{N_{t+s}} \left(\left[\frac{\hat{P}^{h'}_{t+s}\left(x^{h'}_{t+s}\right)}{\hat{P}^{h'}_{t}\left(x^{h'}_{t+s}\right)} \right]^{\frac{1}{N_{t+s}}} \right). \tag{7}$$

The form of the index recognizes the unequal number of properties sold in the two periods and defines a geometric mean of the geometric Laspyeres and Paasche indexes (see equations (6) and (7)).

We have labeled the two indexes as *hedonic Törnqvist/Jevons* above to emphasize these are not based on matching pairs as in the standard price index case. In both cases the use of a geometric mean is consistent with a general log-normal distribution of price relatives. The use of a Jevons index might be more appropriate if the principal aim is to generate a statistically sound estimator of the central tendency of the distribution of the change in property prices. Given that the expenditure weights used in hedonic imputed price indexes do not have the same theoretical basis as the expenditure shares used in the construction of the consumer price index, the choice between the Törnqvist and Jevons should be driven by the main objective behind the property price index construction. For instance, it might be reasonable to construct RPPIs for meaningful sub-regions (and even types of houses if the sample size permits) using the Jevons index and then aggregate using value weights.

Silver and Heravi (2007) and Hill and Melser (2008) discuss the importance of computing RPPIs using estimates of the parameters which vary over time and regions (see equation (9) and related discussion in Hill and Melser (2008)). This is a very important issue that deserves further research. However, the objectives of this paper relate to the commonality amongst these indexes, that is, they depend on a model prediction, $\hat{P}_i^h\left(x_t^h\right)$ $(\hat{P}_i^{h'}\left(x_{t+s}^{h'}\right))$ $i = t$, $t + s$. The next section discusses alternative modeling frameworks available to construct the imputation required to construct these indexes.

3. Hedonic Models of Property Prices

For the purpose of this discussion it will be convenient to set a general framework that can accommodate a number of alternative models and estimators that are candidates for computing $\hat{P}_i^h\left(x_t^h\right)$ $(\hat{P}_i^{h'}\left(x_{t+s}^{h'}\right))$ $i = t,t + s$. The alternative estimators are then discussed, with an emphasis on how they weight the information available, and which estimator leads to index revisions.

Consider the hedonic model for the logarithmic transformation of the sale price of properties[3] that includes a term that captures overall macroeconomic conditions in the market, μ_t, a term that captures the size and quality attributes of the properties, $X_t\beta_t$, and a random disturbance assumed normally and identically distributed (*NID*).

$$y_t = \mu_t + X_t\beta_t + \varepsilon_t \quad \varepsilon_t \sim NID(0, H_t) \tag{8}$$

where:

y_t —$N_t \times 1$ vector of observations of the dependent variable, typically the log of sale price (P_t), $y_t = \ln P_t$ for the N_t transactions observed in period t;

μ_t —captures overall macroeconomic trends;

β_t —$K \times 1$ vector of unknown slope parameters (shadow prices);

X_t —$N_t \times K$ matrix of independent variables, house and land attributes, which will typically include measures of spatial characteristics (such as distances to transport, schools, etc);

[3]The use of a semi-log model is standard across regression based methods (e.g., repeat sales, time-dummy, and hedonic imputed) used in the computation of RPPIs (see Hill, 2011).

ε_t—$N_t \times 1$ vector of random disturbances assumed to be normally distributed although not necessarily independent (this is to allow for spatial dependence); $E(\varepsilon_t \varepsilon_t') = H_t$ is the variance–covariance, $H_t = \sigma_\varepsilon^2 I_{N_t}$ if the errors are not spatially dependent. The specification in the case of spatial dependence is discussed shortly.

A number of assumptions can be made regarding the behavior of μ_t and β_t in (8) which have consequences for how the model is estimated. To discuss those it is convenient to start from a general specification,

$$\mu_t = \mu_{t-1} + \sigma_\mu \xi_t \tag{9}$$

$$\beta_{kt} = \beta_{kt-1} + \sigma_{\beta_k} \zeta_{kt} \tag{10}$$

where, $\xi_t \sim NIID(0, 1)$, $\zeta_{kt} \sim NIID(0, 1)$ $k = 1, \ldots, K$ are $K + 1$ normal, independent, and identically distributed (*NIID*) random variables. In the context of hedonic modeling for housing several special cases have been considered previously in the literature:

Case 1: Restrict $\sigma_{\beta_k} = 0, k = 1, \ldots, K$, which results in $\beta_{kt} = \beta_k$. This is a more flexible form of the time-dummy hedonic regression model, as the conventional fixed time effects (time-dummies) are modeled as a flexible stochastic trend μ_t instead.[4] Schwann (1998), Francke and de Vos (2000), and Francke and Vos (2004) used this model. Schwann (1998) labeled the estimated $exp(\hat{\mu}_t)$ as a "time-series price index." Like the time-dummy hedonic index, this approach provides a measure of price change which would be equivalent to an HI index if $\beta_{kt} = \beta_k$ is true for all t (this issue was studied in detail by Silver and Heravi (2007) and the interested reader is referred to their work).

Case 2: Restrict $\sigma_{\beta_k} = \sigma_\beta$, with $Q = Var(\sigma_\beta \zeta_{kt} I_K)$ a diagonal matrix. Rambaldi and Rao (2011, 2013) used this model to compute HI indexes. In this case the shadow price parameters, β_{kt}, are assumed to be stochastic processes subjected to independent shocks, ζ_{kt}, as in the general case in (10); however, the standard deviation of each of these shocks is assumed to be of size σ_β for all K. This restriction simplifies the model, but might be overly restrictive.

Case 3: Assume $\mu_t \neq \mu_{t-1}$ and $\beta_t \neq \beta_{t-1}$ and restrict $\sigma_\mu \to \infty$ and $\sigma_{\beta_k} \to \infty$.[5] This is the Rolling Window approach (Court, 1939; Griliches, 1961). The model is estimated over M–periods (such as the popular adjacent two-period rolling window, Triplett, 2004). We are aware that there is not a single approach to the implementation of this method. In this paper we will estimate the hedonic regression over two adjacent periods, $t - 1$ and t, to obtain the estimates of μ_t and β_t, roll the window and estimate the regression over t and $t + 1$ to obtain estimates of μ_{t+1} and β_{t+1}, and so on. The first set of estimates is used to price properties in period t and the second to price properties in period $t + 1$. These price predictions then enter the corresponding Laspyeres and Paasche formulae to form the Törnqvist or Jevons indexes as required.

[4]It is a well known result that fixed time effects are a restricted form of a stochastic trend (see Harvey, 2006).

[5]We thank an anonymous referee for pointing this out.

The use of hedonic regressions with time-varying parameters modeled as stochastic processes to compute and predict price movement in real estate markets is not new. They have been proposed for the repeat sales approach (Francke, 2010), the time-dummy hedonic indexes (Schwann, 1998; Francke and de Vos, 2000; Francke and Vos, 2004), and the HI approach (Rambaldi and Rao, 2011, 2013).[6] The estimation of the parameters as stochastic trends involves smoothing methods (which will be discussed in Section 4). Both Schwann (1998) and Francke (2010) highlight the suitability of this approach when dealing with thin markets. Their robustness is due to the weighting of past and current market information.[7] In Section 4 we provide expressions to show how this weighting occurs when using the KF, the KS, and ROW. In this paper we will not consider Case 1 for two reasons: first, econometrically it is a special case of Case 2; and if the data are consistent with Case 1 (that is, the shadow price coefficients are constant over time), estimation of Case 2 should lead to an estimate of σ_{β_k} which is statistically zero. Second, from an index construction perspective it assumes hedonic characteristics have fixed shadow price parameters over time.

3.1. *Controlling for Location*

There are a number of ways in which location can be incorporated in hedonic models. First we consider the specification of ε_t in (8) with spatial dependence. As many characteristics associated with the location of a property might not be measured through the conventional X_t factors (e.g., number of bedrooms and bathrooms, size of the land) an omitted bias problem arises. A well known option to account for the omitted bias is to use a spatial lag model in the error term. This is often referred to as a Spatial Error Model (SEM). The spatial error process has the form

$$\varepsilon_t = \rho W_t \varepsilon_t + u_t \quad u_t \sim NID\left(0, \sigma_u^2 I_{N_t}\right) \tag{11}$$

where:

ε_t —$N_t \times 1$ vector of spatially correlated errors with covariance H_t;

u_t —$N_t \times 1$ vector of uncorrelated errors (independent and identically distributed);

ρ —scalar spatial autocorrelation parameter, $|\rho| < 1$.

W_t —$N_t \times N_t$ matrix of spatial weights (with elements w_{ij}) with the following characteristics, $w_{ii} = 0$ for all i

$0 \leq w_{ij} \leq 1$ are weights representing the strength of the "neighbor relationship" of the i-th property with the j-th property sold at time t.

W_t is a row-stochastic matrix (i.e., rows sum to unity).

To compute each w_{ij}, the distance between property i and all other properties sold at time t, $j \neq i$, $j = 1, \ldots, (N_t - 1)$, must be measured. This can be easily computed using a triangulation algorithm and the unique coordinates (latitude,

[6]Knight *et al*. (1995) proposed a seemingly unrelated regression framework whereby annual regressions are jointly estimated.

[7]A feature recognized in the real estate modeling literature by the works of Quan and Quigley (1991).

longitude) of each property.[8] We use Delaunay triangulation[9] to generate a set of closest neighbors around each i (a detailed presentation is provided by LeSage and Pace, 2009), with weights which are inversely proportional to the distance.

In this case the form of H_t can easily be shown to be

$$H_t = \sigma_u^2 \left(I_{N_t} - \rho W_t\right)^{-1} \left(I_{N_t} - \rho W_t\right)^{-1'} \tag{12}$$

since $\varepsilon_t \sim N(0, H_t)$, and $\varepsilon_t = \left(I_{N_t} - \rho W_t\right)^{-1} u_t$, using (11). It is easy to verify that the errors of the model in (8) are homoskedastic and not spatially correlated if $\rho = 0$, in which case $H_t = \sigma_u^2 I_{N_t}$, as already indicated. The interested reader is referred to LeSage and Pace (2009) for an introductory, but comprehensive, treatment of spatial econometric models. One important feature of the specification in (12) is that W_t is changing over time. As will be shown in the next section, whether one estimates the model using a ROW approach or an SM approach, the estimator of μ_t and β_{kt} will be a function of a time-varying spatial structure if ε_t is assumed to follow (11). The formulation is spatially varying and attenuated over time, in the same spirit as that proposed by Gelfand *et al*. (2003; see model 4, p. 391).[10]

A second alternative specification of location is to include *spatial regressors* (i.e., as factors added to the X_t matrix). Spatial regressors are measures of the relative location of each property with respect to landmarks such as bus stops, schools, parks, and industry. They can be easily generated using the coordinates data and GIS software. It is also possible to have both the set of spatial regressors and the spatial error structure in the model. In the empirical section we include all these alternatives.

Finally, the recent work by Hill and Scholz (2013) has proposed modeling location by fitting a non-parametric surface to location coordinates and adding this constructed variable to the hedonic regression as a regressor. They then estimate the model yearly before computing the HI indexes. We do not consider this alternative in this paper. This is left for future research.

4. Estimation and Prediction: ROW vs SM

In this section we discuss the difference between a ROW approach and an SM approach with specific reference to how information is incorporated and weighted in the estimators of μ_t and $\beta_t = (\beta_{1t}, \ldots, \beta_{Kt})$. We discuss in detail when the estimation approach leads to revisions of the computed RPPI. Finally, we highlight the role of other parameters of the model (i.e., variance and covariance parameters), discuss their estimation and their role in inducing revisions of the indexes.

[8]Coordinates are readily available as they are routinely provided with transaction level data, unlike a number of hedonic characteristics which are often missing.

[9]A Delaunay triangulation for a set P of points in a plane is a triangulation such that no point in P is inside the circumcircle of any triangle.

[10]Their work uses dummy variables to model time; however, both their work and that of Rambaldi and Rao (2011), used here, are based on a spatially varying correlation which is function of unknown parameters (here σ_u^2, ρ) which are not time-varying. In the estimation section (see Section 4) we discuss the role of these parameters.

To aid the discussion we collect the general model in (8)–(11), and write it in the form (13) and (14)

$$y_t = Z_t\alpha_t + \varepsilon_t \tag{13}$$

$$\alpha_t = \alpha_{t-1} + \eta_t \tag{14}$$

where:

$Z_t = [1\ X_t]$ is an $N_t \times m$ matrix; where, and 1 is an $N_t \times 1$ vector of ones.

$\alpha_t = [\mu_t\ \beta_t]'$; where $\beta_t = [\beta_{1t} \cdots \beta_{Kt}]'$

$\eta_t = [\xi_t\ \zeta_t]'$ with $Q^* = E(\eta_t\eta_t'),\ Q^* \sim N\left(0, \begin{bmatrix} \sigma_\mu^2 & 0 \\ 0 & diag\left(\sigma_{\beta_1}^2, \ldots \sigma_{\beta_K}^2\right) \end{bmatrix}\right)$

$\alpha_0 \sim N(a_0, \Omega_0)$ is an initial condition.

The system in (13) and (14) is known as a state-space representation. One feature of state-space representations is that they separate the parameters of the original econometric model (in this case μ_t, β_t, $\sigma_u^2, \sigma_\mu^2, \sigma_{\beta_1}^2 \ldots \sigma_{\beta_K}^2$, and ρ) into two types for the purpose of estimation, namely, the parameters in α_t (known as the state-vector) and the rest (which are labeled as hyperparameters; see Durbin and Koopman (2012) for discussion of the term). The latter are the parameters that define what is referred to as *system matrices* in the state-space literature (see Harvey, 1989). In the above representation, the unknown *hyperparameters* are $\psi = \left[\sigma_u^2, \sigma_\mu^2, \sigma_{\beta_1}^2 \ldots \sigma_{\beta_K}^2, \rho\right]$ and they are in H_t and Q^*. If ψ is known or an estimate exists, the state-vector (α_t) can be estimated by SM.

These estimates are conditional on ψ. In order to understand how and when the use of these estimation approaches will lead to revisions, it will be convenient to present the KF and KS.[11]

4.1. *Kalman Filter Smoother Estimation*

Assuming first that ψ is known (see Section 4.5) and we have a sample from $t = 1, \ldots, T$, we start by considering the time period $t = \tau$. The KF estimator of μ_τ and β_τ is given by $a_{\tau|\tau} = E(\alpha_\tau|y_\tau, y_{\tau-1}, \ldots, Z_\tau, Z_{\tau-1}, \ldots, Z_1, a_0, \Omega_0)$ [12], which can be written as a recursive formula,

$$a_{\tau|\tau} = a_{\tau-1|\tau-1} + G_\tau v_\tau \tag{15}$$

where:

$v_\tau = y_\tau - Z_\tau a_{\tau-1|\tau-1}$ is the prediction error using the parameter estimates at $\tau - 1$,

$v_\tau \sim (0, F_\tau)$, with size $N_t \times 1$.

$F_\tau = Z_\tau\Omega_{\tau|\tau-1}Z_\tau' + H_\tau$ is the variance-covariance of the prediction error, v_τ.

[11]The presentation here is minimal and only for the purpose of explaining how the estimates compare to those from ROW. For derivations and more details, see Harvey (1989) or Durbin and Koopman (2012).

[12]We note here that we are defining the KF estimate as $a_{\tau|\tau}$ and not as $a_{\tau|\tau-1}$, and thus refer to it as a "Kalman Filter Smoother." This definition follows Harvey (1989) in that the KF estimate is conditional on the current time period as well as the information up to $\tau - 1$ (i.e., using both the prediction and updating equations of the Kalman filter).

$G_\tau = M_\tau F_\tau^{-1}$ is known as the Kalman gain and captures the information gain from $\tau - 1$ to τ.

$$M_\tau = \Omega_{\tau|\tau-1} Z'_\tau$$

$$\Omega_{\tau|\tau-1} = \Omega_{\tau-1|\tau-1} + Q^*$$

Equation (15) shows that at a *given time period*, τ, the estimates of $a_\tau = (\mu_\tau, \beta_\tau)'$ are equal to what they were at time $\tau - 1$, plus an adjustment given by a proportion (G_τ) of the error we would make in τ if we were using the parameter estimates from $\tau - 1$ to predict the prices of the properties sold in τ (see definition of v_τ). Therefore, the task in practice is to compute G_τ and v_τ when a new time period of data becomes available which then allows the implementation of the updating in (15).

An assumption about the initial condition (a_0, Ω_0) is necessary to start the recursion. The standard practice is to initialize the filter by setting a_0 to a fixed arbitrary value and Ω_0 as such that it has a diffuse prior density.[13] Standard errors for these estimates are obtained by computing the square root of the diagonal elements of the covariance matrix $\Omega_{\tau|\tau}$, as one would in any other estimation approach. The variance–covariance is given by

$$\Omega_{\tau|\tau} = \Omega_{\tau|\tau-1} - M_\tau F_\tau^{-1} M'_\tau. \tag{16}$$

The KF can be also written as a weighted sum of past and current information. Koopman and Harvey (2003) have shown that that the KF for time period τ is given by

$$a_{\tau|\tau} = \sum_{j=1}^{\tau} \omega_{j\tau} y_j. \tag{17}$$

They provide specific expressions for the $\omega_{j\tau}$ (which are functions of G_t and v_t, for $t = 1, \ldots, \tau$). What is important about this result is that the highest weight is at τ and these weights decrease towards zero for t further back in time from τ. The number of past periods with non-zero weights depends on the underlying econometric model and specific dataset.

4.2. *Kalman Smoother Estimation*

As new information becomes available, it is possible to revise the estimates $a_{\tau|\tau}$. The algorithm that allows this revision is the KS, which is effectively a revision of the past estimates taking into account the most current information, and its application is known as *state smoothing* or simply *smoothing*. The commonly known fixed interval smoothing is given by $E(\alpha_t | y_T, y_{T-1}, \ldots y_1, Z_T, Z_{T-1}, \ldots Z_1, a_0,$

[13]A diffuse prior corresponds to $\Omega_0 = \kappa I$ and letting $\kappa \to \infty$. In the empirical section we start the filter at $a_0 = 0_{m\times 1}$ and $\Omega_0 = 1E + 4 \times I_m$.

Ω_0). Consider the time period $t = \tau$ above, we can now use the KS to revise it. The KS can be written as follows:

$$a_{\tau|T} = a_{\tau|\tau} + \sum_{j=\tau+1}^{T} Cov(\alpha_\tau, v_j) F_J^{-1} v_j. \tag{18}$$

The expression (18) tells us that the KF estimate will be revised by bringing into the estimation the accumulated information from periods $\tau + 1$ to T (for a full discussion and details, see Durbin and Koopman, 2012, section 4.4). We can write (18) as a sum of the weighted information used as well to obtain

$$a_{\tau|T} = \sum_{j=1}^{T} \omega_{jT} y_j \tag{19}$$

where a plot of the weights in this case would show that the highest weight is given to $t = \tau$ as in the KF. However, *both past and future observations* in the proximity of τ will have have a non-zero weight (see Francke, 2010, figure 1, for an example obtained for his local linear repeat sales model). The previous works that have used SM and considered price indexes (specifically, Schwann, 1998; Francke and Vos, 2004; Francke, 2010; Rambaldi and Rao, 2011, 2013), have estimated the parameters using a KS (i.e., as in (19)).We return shortly to the differences created when KF and KS are used to impute the prices to construct the indexes.

4.3. *Rolling Window Estimation*

We now turn to the estimation under a ROW approach and then discuss and compare all these alternatives. In a ROW framework the estimation involves the use of OLS, or generalized least squares (GLS) if the variance–covariance of ε_t is not spherical. Considering the form of the GLS estimator for a two-adjacent period rolling regression (it is easy to see that (20) reduces to OLS if $\rho = 0$, as $H_t = \sigma_u^2 I$ in that case),

$$\hat{a}_t = \left(Z'_{(t-1,t)} H^{-1}_{(t-1,t)} Z_{(t-1,t)}\right)^{-1} Z'_{(t-1,t)} H^{-1}_{(t-1,t)} y_{(t-1,t)} \tag{20}$$

with variance–covariance,

$$\Omega^{GLS}_{(t-1,t)} = \left(Z'_{(t-1,t)} H^{-1}_{(t-1,t)} Z_{(t-1,t)}\right)^{-1}$$

where the subscript $(t - 1, t)$ is to indicate that the estimate $\hat{a}_t$ is using the current and immediately past period market information. The expression in (20) is also a weighting function of available information. To see this let $n = N_{\tau\text{-}1} + N_\tau$ be the number of observations used in the window to estimate the parameters for time τ, then the ROW estimator is given by

$$\hat{a}_\tau = C_\tau y_\tau \tag{21}$$

where $C_\tau = \left(Z'_{(m\times n)} H^{-1}_{(n\times n)} Z_{(n\times m)}\right)^{-1} Z'_{(m\times n)} H^{-1}_{(n)}$ is an $m \times n$ matrix and y_τ is an $n \times 1$ vector of log prices. The parameter estimates obtained from this weighting

function do not distinguish sales in period $\tau - 1$ from those in τ and thus prices are equally weighted in time. The k-th, $k = 1, \ldots, m$, parameter in $\hat{a}_\tau$ is the product of the k-th row of C_τ and y_τ, and therefore a weighed sum of log prices. The parameter estimates are spatially weighted functions of the observations in the window when H_t is not $\sigma_u^2 I$. It is now easy to see that a ROW approach using an arbitrary M–periods to compute the parameters will not perform an optimal weighting of past information. This issue has been raised in the literature before by Pace *et al.* (2000). They were concerned about how the information entered the estimation of the parameters in a time-dummy parameter model. Their proposal was to use a covariance structure that effectively accounted for the time and space sequence of sales in the estimation.

4.4. *Price Prediction*

To study how KF, KS, and ROW estimates enter an imputed index, we first consider the expression to impute prices. Consider the pricing of house h' sold in period $T - 1$ for the purpose of computing an index of the price change from $T - 1$ to T. As the model is log-linear, the imputation of the prices at periods $T - 1$ and T is obtained from the following expression:

$$\hat{P}_i^{h'}\left(x_{T-1}^{h'}\right) = \exp\left[\begin{pmatrix}1 & x_{T-1}^{h'}\end{pmatrix}\widehat{a_{i|\Im}}\right] \tag{22}$$

$$= \exp\left(\widehat{\mu_{i|\Im}} + x_{T-1}^{h'}\widehat{\beta_{i|\Im}}\right) \tag{23}$$

where $i = T, T - 1$, "$\widehat{\ }$" denotes they are estimates,[14] $\Im$ denotes the information set used, and thus distinguishes the estimator (KF, KS, or GLS). The information set is $\Im = y_i, y_{i-1}$ when the estimates are those obtained from ROW (GLS), $\Im = \{y_i, y_{i-1}, \ldots, y_1\}$ when the estimates are KF, and $\Im = \{y_T, y_{T-1}, \ldots, y_1\}$ when the estimates are KS.

Therefore, the price prediction for $T - 1$ uses $a_{T-1|T-1}$ when using the KF, $\hat{a}_{T-1}$ when using ROW, and $a_{T-1|T}$ when using the KS. An inconsistency arises when using KS in that the prediction is made using parameters that have been estimated using information from time period T which was unknown to the market at time $T - 1$. Thus, the choice is between the KF and ROW. Harvey (1989) shows the KF is an optimal estimator given the information known at that point in time. In addition, it is practically feasible as it only requires implementation of recursion (15).

[14]In the case of spatial errors, (22) is a truncated predictor (see Rambaldi and Rao (2013) for details and further references). It is well known that the form of the predictor used here (in (22)) is one alternative, with a second alternative being a "corrected" version given by $\hat{P}_i^{h'}\left(x_{T-1}^{h'}\right) \times \exp(\hat{\sigma}_\varepsilon/2)$ which is derived from the properties of the log-normal distribution. In our specification, "σ_ε" is a time-varying function of W_t and ρ when errors are spatially correlated. When there are no spatial errors, the correction term will be the same for numerator and denominator of the index and thus it cancels out. In addition, when this correction is merited depends on the normality assumption as well as the objective of the prediction (for further discussion, see, for instance, Greene, 2012, p. 123).

4.5. *Hyperparameters: Estimation and Their Role in Revisions*

In practice ψ is unknown and an estimate is needed. Inspection of the general model shows that if the estimates of the parameters $\psi = [\sigma_u^2, \sigma_\mu^2, \sigma_{\beta_1}^2 \ldots \sigma_{\beta_K}^2, \rho]$ change, estimates of H_t and Q^* change as well. When estimating the model under Case 3 (a ROW approach), all the unknown parameters (ψ and a_t) can be estimated by maximum likelihood. The most convenient form of the log-likelihood is given by

$$\ln L(y_n; \psi, \beta_n, \mu_n) = -(n/2)\ln(\pi\sigma_u^2) + \ln|I_n - \rho W_n| - \frac{e'e}{2\sigma_u^2}; \quad \rho \in (0,1] \tag{24}$$

where $n = N_{t-1} + N_t$ (if two adjacent periods are used), and $e = y_n - Z_n\hat{a}_n$. The reader is directed to LeSage and Pace (2009) for detailed presentation of maximum likelihood estimation of spatial models. When $\rho = 0$ (no spatial errors), maximization of (24) will yield the OLS estimator.

When estimating the general model under Case 1 or 2, the time series ordering of the data has to be taken into account and thus the appropriate form of the log-likelihood is in prediction error form (see Harvey (1989) for detailed presentation). The log-likelihood for a sample of $N = \sum_{t=1}^{T} N_t$ transactions over T time periods is given in this case by

$$\ln L(y_t; \psi, Y_{t-1}) = -\frac{1}{2}\sum_{t=1}^{T} N_t \ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T} \ln|F_t| - \frac{1}{2}\sum_{t=1}^{T} v_t' F_t^{-1} v_t. \tag{25}$$

This form is convenient when using the SM because it is written as a function of v_t and F_t which are computed by running the KF algorithm (see definitions below equation (15)). In practice there are a number of important results that allow efficient running of the algorithm for the purpose of computing v_t and F_t. For a more detailed exposition of the maximum likelihood estimation of hyperparameters in state-space models, see Harvey (1989) or Durbin and Koopman (2012). The main difference between (24) and (25) is that the first is based on the joint distribution of n observations assumed to be identically distributed (and independent if $\rho = 0$), while the second recognizes the time series nature of the observations as well, which are conditional on the information set at time $t-1$, Y_{t-1}.

To obtain estimates of ψ, a numerical maximization of the log-likelihood for the unknown parameters has to be carried out using Newton–Raphson type algorithms. That is, the estimates are given by

$$\hat{\psi} = argmax_{\psi} \ln L(y_t|\psi). \tag{26}$$

Our general model follows the standard state-space specification, and thus the parameters in ψ are not assumed to be time-varying; however, as in any other statistical estimation, sampling variation means the estimates will change depending on the sample size used to obtain them. As T becomes larger, or the number of transactions in each time period are larger, the estimates are expected

to settle and should not change significantly by the addition of just one or two time periods.

SM are time series based methods and thus a reasonable time series length is necessary for the KF to settle and to avoid the estimates being overly dominated by the initial condition ($\alpha_0 \sim (a_0, \Omega_0)$). For this reason, when implementing the numerical maximization of (25) for the estimation of ψ, it is advisable to construct the log-likelihood without the first d periods (in the empirical illustration below we set d to the first 12 months of the sample);[15] that is, the KF is run from $t = 1$, but the sum in (25) is constructed for the period $t = d, \ldots, T$. In practice, once the initial condition issue has been taken into account, if the construction of the index starts with a reasonable number of time periods there should not be a need to re-estimate ψ at every time period. However, when revised estimates of the ψ are obtained, the KF should be run to revise the estimated parameters $a_{t|t}$ given the updated H_t and Q^*. In the empirical section we study how the estimates of ψ change as more periods are added to the data.

In the case of ROW the parameters ψ are re-estimated within each new window of M–periods, and thus having more data in the time series dimension is of no benefit to reduce sampling variation. In the empirical section we study how the estimates change and how the sample size of the window influences the variation over windows.

5. Empirical Illustration

In this section we provide an illustration of the main concepts presented in the previous sections. Our contention is that the KF is theoretically preferred to a rolling window approach and is the most suitable estimator to compute (22). The KF is a time series method dependent on an initial condition, and thus in practice a reasonable number of time periods are required for the estimates of α_t not to be overly influenced by the initial condition. Similarly, a reasonable number of time periods is needed to obtain a reliable estimate of ψ initially. Re-estimation of these hyperparameters every time period might not be necessary, making the practical implementation of the KF less computationally intensive. We explore some of these issues next.

The data are from a coastal town in the state of Queensland, Australia. The advantage of this dataset is that the data are from a small and homogeneous region, constructed by merging transaction sales data from real estate transactions with local council records, and with the addition of GIS generated regressors. The disadvantage is that for the initial part of the sample period the number of transactions per month is not very large; however, this provides the opportunity to study how thin markets affect the volatility of estimates and the computed indexes.

5.1. *Data*

The data used were compiled, cleaned, and checked with funding from the Department of Climate Change and Energy Efficiency, and the CSIRO Climate

[15]See section 3.3.4 of Harvey (1989) or section 2.9 of Durbin and Koopman (2012) for a detailed presentation of initialization and convergence of the Kalman Filter.

TABLE 1
DESCRIPTION OF THE DATASET

	Min	Max	Mean	Median	S.D.
Sale price	15,000	1,250,000	226,352	210,000	130,820
Month of sale	1	12	6	6	3
Year of sale	1991	2010	2002	2003	5
Land characteristics					
Lot size (m^2)	261	10,220	933	630	1068
Small lot < 500 m^2	0	1	0	0	0.29
Large lot > 2000 m^2	0	1	0	0	0.28
Dist_Coast (m)	15.81	5749.44	1313.33	1283.20	903.67
Dist_Waterway (m)	5.00	863.54	263.53	245.20	157.31
Dist_Parks (m)	7.07	983.93	139.61	115.43	112.86
Dist_Schools (m)	10.00	6482.48	586.53	304.51	998.07
Dist_Shops (m)	14.14	4796.77	493.89	364.59	471.45
Dist_OffensiveIndust (m)	175.00	8378.48	2908.76	2260.27	1867.23
Dist_BusStop (m)	15.00	4349.01	419.68	212.66	729.45
Dist_RailStn (m)	2453.59	13,583.90	7067.73	6700.23	1765.23
Dist_BoatRamp (m)	55.90	6294.16	1940.04	1528.61	1537.51
Dist_PubsClubs (m)	14.14	5336.24	1346.46	1177.65	912.40
Dist_Hospitals (m)	10.00	13,255.00	3163.51	2310.68	2622.82
Structure characteristics					
Bedrooms	0	8	3	3	1
Bathrooms	1	4	1	1	1
Carspaces	0	12	2	2	1
Structure footprint (m^2)	35.82	920.90	220.00	205.24	89.43
Max_Height_Building (m)[a]	2.27	27.02	8.72	8.14	4.58
Age (years)	0	86	16	15	11
Number of transactions			9984.00		

[a]This is the height of the building in meters above sea level.

Adaptation Flagship, and are a subset of those used to produce some of the results presented in Fletcher *et al*. (2011). These data are sourced primarily from one of Australia's leading providers of real estate sales transaction data (RPData); a number of spatial hedonic characteristics, such as distances to landmarks (descriptive statistics appear in Table 1, and regression specification in Table 2), were added through GIS analysis. Further cleaning was performed as part of the 2011/12 UQ Summer Scholarship Program funded by the School of Economics and The University of Queensland. The dataset consists of individual transactions of family dwelling residential property (i.e., units, townhouses, and terraces are not included) for the period May 1991 to September 2010. Only sales of land with structure are included (that is, there are no transactions of vacant land). The number of transactions per month is presented in Figure 1.

The area in this study is in the south east corner of the state of Queensland (SEQ) and is a coastal area (Moreton Bay Regional Council, see Figure 2). The markets in the SEQ region went through a boom period between 2001 and 2005 and this is reflected in the number of transactions per month for that period. This market shows high activity concentrated in the 2001 to 2003 period. The 2008 global financial crisis is noticeable in that the number of transactions drops to levels similar to the early 1990s and has remained volatile since. The difference between the data used by Rambaldi and Rao (2011, 2013) and the data used here

TABLE 2

SENSITIVITY OF HYPERPARAMETERS ESTIMATES; OVERALL MODEL FIT

	No Spatial Errors				Spatial Errors			
	No Distance Regressors		Distance Regressors		No Distance Regressors		Distance Regressors	
	Fixed Parameter	State-Space	Fixed Parameter	State-Space	Fixed Parameter	State-Space	Fixed Parameter	State-Space
Sample: 1991:5–1995:12 (56 months)								
$\hat{\sigma}_u^2$	0.077	0.075	0.070	0.070	0.070	0.076	0.067	0.070
$\hat{\rho}$					0.408	0.375	0.258	0.380
$\hat{\sigma}_\eta^2$		0.010		0.010		0.010		0.010
$\hat{\sigma}_\beta^2$		1E – 12		1E – 10		1E – 10		1E – 10
m	14	10	23	19	14	10	23	19
ln$L$	1.68E + 03	1.73E + 03	1.82E + 03	1.82E + 03	358.88	1.74E + 03	402.57	1.79E + 03
BIC	−3.26E + 03	−3.39E + 03	−3.47E + 03	−3.50E + 03	−6.16E + 02	−3.41E + 03	−6.37E + 02	−3.44E + 03
N	1470	1470	1470	1470	1470	1470	1470	1470
Sample: 1991:5–1999:12 (104 months)								
$\hat{\sigma}_u^2$	0.079	0.080	0.071	0.070	0.073	0.070	0.069	0.070
$\hat{\rho}$					0.388	0.380	0.265	0.380
$\hat{\sigma}_\eta^2$		0.010		0.010		0.010		0.010
$\hat{\sigma}_\beta^2$		1E – 10		1E – 10		1E – 10		1E – 10
m	18	10	27	19	18	10	27	19
ln$L$	2.78E + 03	2.22E + 3	3.05E + 03	3.41E + 03	560.84	3.24E + 3	653.90	3.32E + 03
BIC	−5.42E + 03	−4.36E + 03	−5.89E + 03	−6.67E + 03	−9.81E + 02	−6.40E + 03	−1.10E + 03	−6.49E + 03
N	2488	2488	2488	2488	2488	2488	2488	2488
Sample: 1991:5–2010:9 (233 months)								
$\hat{\sigma}_u^2$	0.053	0.048	0.048	0.045	0.049	0.047	0.045	0.045
$\hat{\rho}$					0.370	0.374	0.379	0.379
$\hat{\sigma}_\eta^2$		0.009		0.010		0.010		0.010
$\hat{\sigma}_\beta^2$		1E – 10		1E – 12		1E – 10		1E – 12
m	29	10	38	19	29	10	38	19
ln$L$	1.52E + 04	1.91E + 04	1.62E + 04	2.03E + 4	4.22E + 03	1.94E + 4	4.68E + 3	2.02E + 4
BIC	−3.01E + 04	−3.81E + 04	−3.21E + 04	−4.04E + 04	−8.17E + 03	−3.87E + 04	−9.01E + 03	−4.02E + 04
N	9984	9984	9984	9984	9984	9984	9984	9984

Notes: m is the number of columns in Z_t (includes year dummies for the fixed parameter specifications); $BIC = -2\ln L + m^*\ln N$.
All regressions include the following hedonics: log(Landm2), smalllot, largelot, Beds, Baths, Cars, log(Structm2), Max_Height_Building, Age.
Spatial Regressors used: Dist_Coast, Dist_Waterway Dist_Parks, Dist_Schools, Dist_OffensiveIndust, Dist_RailStn Dist_BoatRamp, Dist_PubsClubs, Dist_Hosiptals.

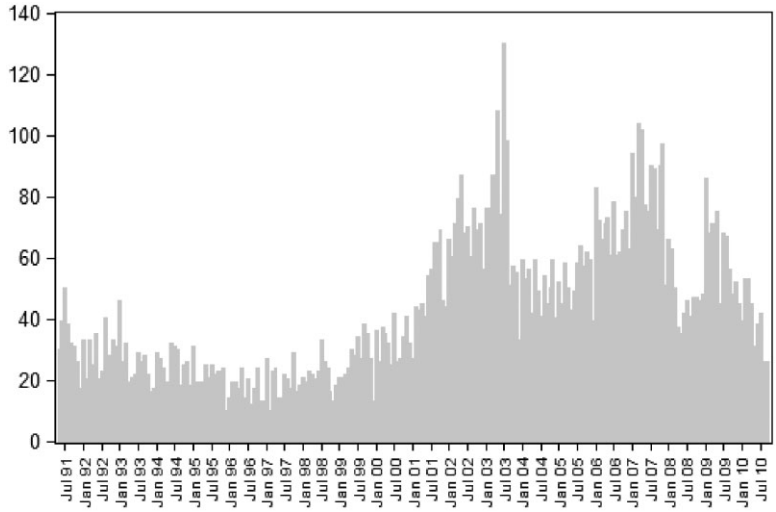

Figure 1. Number of Single Transactions Per Month

is that their data were for the city of Brisbane, the state capital of Queensland, with a population of over 1 million with heterogeneous submarkets (covering 1985–2005). The data used here are from a smaller market located north of the city of Brisbane in a few of the urban centers surrounding the Bay, which act as satellite suburbs to Brisbane in that a substantial proportion of the population travel into Brisbane every day to reach their employment location.

Table 1 presents the set of hedonic characteristics available for the study with their descriptive statistics. They have been divided into two groups, "land characteristics" and "structure characteristics." The first includes the size of the land and two dummy variables that identify small and large lots. The definitions of small and large lot are those published by the Council of the area under study. In addition, a number of measures are used to model the spatial location of each property. These are in the form of distance to amenities and landmarks. These were obtained through GIS analysis. The characteristics of the structure include the standard set available from RPData (bathrooms, bedrooms, car park spaces), the footprint and height of the structure which were obtained through GIS, and the age provided by the local Council.

5.2. *Estimation and Comparison of Predictions*

In this empirical exercise we compare Cases 2 and 3 of the general model[16] under four alternative specifications for location. The first is the model without controlling for location ($\rho = 0$ and there are no spatial regressors in the model). The second has $\rho = 0$ and, in addition to the land and structure regressors, spatial

[16]Allowing the standard deviation of the shock to shadow price parameters, σ_β, to vary over K or at least over subsets of the regressors is appealing; however, it is left for further research.

Figure 2. Polygon Marking Area Covered by Study (map sourced from the Moreton Bay Regional Council)

regressors are included to model location (see Table 1).[17] This specification will be labeled "No Spatial Err, Spatial Reg." The third has a spatial error, $\rho \neq 0$, but excludes all spatial regressors measures from the regression, and thus the location of the property is only controlled through the spatial error. This specification is labelled "Spatial Err, No Spatial Reg." The final specification combines both spatial regressors as well as spatial errors in the model specification ("Spatial Err,

[17]Dist_BusStop and Dist_Shops are not significant and so they are not included in the results presented here.

Spatial Reg"). The rationale behind these is to study the possible trade-off of modeling "location" by comparing the use of a number of spatial regressors as explanatory variables (eight in this case) to the alternative of a spatial error model. Both specifications require the location coordinates of the property (latitude and longitude). Distance measures are created through the use of coordinates, data layers describing the location of landscape features, and GIS software, while the construction of a spatial weights matrix (to use the a spatial error) only requires a triangulation algorithm readily available in Matlab. This latter approach could be appealing to statistical offices as the data manipulation requirements are lower.

In all cases, the models estimated are log-linear in all regressors except for lot size and house size which are also log-transformed.[18] Monthly predictions are produced in all cases. Although we could compute quarterly and annual indexes, constructing a monthly index is appealing at at least two levels. First, it is a good test for the methods as the sample sizes are small; second, it shows that it is possible to construct monthly indexes which is of interest to both public and private institutions (e.g., central banks track housing prices monthly as one important indicator of economic activity, investors use them to compare returns to other assets).

When errors are assumed to be spatially uncorrelated, the covariance matrix H_t, (12), reduces to $\sigma_u^2 I$ and $\psi = [\sigma_u^2, \sigma_\mu^2, \sigma_\beta^2]$. In this case the ROW estimation is based on a rolling window of two adjacent periods estimated by OLS. We start by studying the overall fit of the alternative specifications, and the sensitivity of hyperparameters estimates to the length of the sample used, the model specification, and the estimator.

5.2.1. Overall Model Fit and Sensitivity of Covariance (Hyperparameters) Estimates

Table 2 presents a comparison of the specifications by treatment of location estimated over three sample lengths. The purpose of presenting these results is two-fold. First, they provide diagnostics to choose the best fitting model for the data and evaluate the robustness of the estimation of ψ to sample size. Second, they allow us to show that the estimates of σ_μ^2 and ρ are very close in value, whether they are estimated running the Kalman filter algorithms in the state-space form of the model or OLS/GLS in a regression model of the full sample. The usefulness of this result is that we can use the latter estimates as initial values in the optimization routine to obtain the MLE of ψ (see equation (26)).

Table 2 presents estimates of ψ, the log-likelihood (lnL) value, and computed Bayesian Information Criteria (BIC) for all the model specifications estimated by the Kalman filtering algorithms (column labeled state-space)[19] or by running a conventional regression with year dummy intercepts (column labeled fixed

[18]Alternative hedonic specifications could be tried, such as use of splines for age, lot size, and measures of location to capture possible non-linearities. We have not explored these possibilities in this paper.

[19]The estimation code was written by the first author in Matlab and will be made available upon request.

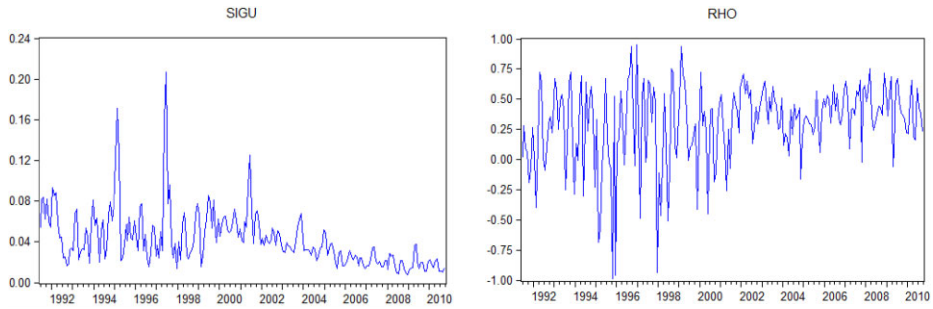

Figure 3. Rolling Windows Estimates of σ_u^2 and ρ; Model: Spatial Err, No Spatial Reg

parameter) for the corresponding sample length[20]. To illustrate how the ψ estimates change as more data become available over time, three samples have been used. The first covers the first 56 months of the sample, 1991:5–1995:12; the second covers the longer period of 104 months, 1991:1–1999:12; and the third covers the complete sample, 1991:5–2010:9 (233 months). The ROW approach estimates of ψ[21] cannot be presented in the table as ROW produces estimates of the ψ, log-likelihood, and BIC which are different in every window. To allow comparison of the estimates of ψ obtained using ROW, Figure 3 plots the estimates of the two covariance parameters, σ_u and ρ.

Inspection of the estimates across model specifications, sample sizes, and estimators indicates a remarkable similarity in the estimates of $\psi = [\sigma_u^2, \sigma_\mu^2, \sigma_\beta^2]$ across model specifications, especially within a particular sample size. The parameter σ_u^2 represents variance of the noise in the hedonic regression. It is estimated to be between 0.07 and 0.08 when the sample size used is 56 months (1470 transactions) or 104 months (2488 transactions), but decreases to 0.05 when the sample used is 233 months (9984). The estimate of ρ is smaller when location is measured by including both distance regressors and a spatial error structure in the model; otherwise it is around 0.4 across all models, estimators, and sample sizes. The estimates of σ_μ^2 and σ_β^2 are identical across models and sample sizes. The estimate of σ_β^2 is very small (order of 10E-10 to 10E-12). This could be due to the overly restrictive specification of a common σ_β (instead of σ_{β_k}), and the issue deserves further research. Notwithstanding the restriction, an advantage of the state-space specification is that it is flexible and thus the rate of variation of the state parameters, μ_t and β_t, is determined by the data for the market under study (see further discussion in Section 5.2.2).

Figure 3 shows the estimates from ROW. In this approach estimates of σ_u^2 and ρ vary over windows. When the number of transactions per month is small (as it is at the beginning of the sample), the variation over windows can be very substantial. The estimates are much more settled during the second part of the

[20]The estimation is by least squares if $\rho = 0$ and by maximum likelihood, using the *sem.m* routine of the Spatial Econometrics Toolbox in Matlab (http://www.spatial-econometrics.com/), if errors are assumed spatial.

[21]They are estimated using two periods of consecutive data by OLS or MLE if spatial errors are assumed.

sample (2000 onwards) due to the larger sample sizes. Although very volatile, the average estimate of ρ over the whole sample sits around 0.3 in the 90s and 0.4 in the 00s, which is roughly consistent with the estimate from the KF. Likewise, the estimates of σ_u^2 appear to be in the range 0.07–0.08, on average, in the first part of the sample and drop to around 0.04–0.05 on average for the second part of the sample, again consistent with the state-space and time-dummy specifications.

The overall fit of the models has been measured using the Bayesian Information Criteria. As expected the time-varying parameters models dominate the time-dummy regressions.[22] The overall conclusion for location is that the dominant model for this dataset is the one with spatial regressors and no spatial error, and this is independent of the sample size. The second preferred model is that with both spatial errors and regressors, the third is the model with spatial errors only, and the last model is the one when location is not modeled in any form. These results confirm the importance of controlling for location.

The results indicate that the estimates of the hyperparameters attached to the time-variation of μ_t and β_t (σ_μ^2 and σ_β^2) and the spatial correlation parameter, ρ, hardly (if at all) change as more time periods of data become available (see columns labeled State_Space as sample sizes increase). The parameter σ_u^2, which is attached to the overall noise in the hedonic model, is about the same size when using 56 or 104 months of data but reduces considerably for the sample of 233 months. This is expected as the overall noise reduces the longer the sample, leading to reductions in the standard errors of the estimates of μ_t and β_t. The practical implication of this result is that *re-estimation of the hyperparameters is not needed with every addition of a new time period*. Thus, the recursion in (15) can be computed, given $\hat{\psi}$, by simply plugging in prices and property characteristic data as they become available to obtain the next time period's estimate of a_t and compute the HI indexes. We do not suggest that the hyperparameters need never be re-estimated, but rather that this can be done sporadically to capture potential changes in these structural parameters which could occur over time.

5.2.2. Trend and Shadow Prices Parameter Estimates

We start by presenting a summary of how the estimates of μ_t and β_t vary between KF and ROW based estimates for the selected specification (No Spatial Err, Spatial Reg). We present the complete comparison between (No Spatial Err, Spatial Reg) and (Spatial Err, No Spatial Reg) in the Appendix. Figure 4 presents the estimates of four parameters of the model chosen to provide the illustration: μ_t, $\beta_{baths,t}$, $\beta_{lotsize,t}$, $\beta_{age,t}$.[23] The KF are the estimates from the Kalman filter smoother (15) with hyperparameters obtained using the first 104 months of data (see Table 2). In the case illustrated, the estimates of the hyperparameters have not been updated over the last ten years of the data. Although this update frequency might seem insufficient, re-running the analysis with the hyperparameters updated every three years yielded virtually unchanged estimates of the μ_t and β_t. The main difference is in the size of the standard errors, which become smaller.[24] The

[22]except in the case of $T = 104$ when location is omitted from the model.
[23]The complete set of estimates is available from the authors.
[24]Results available from the authors.

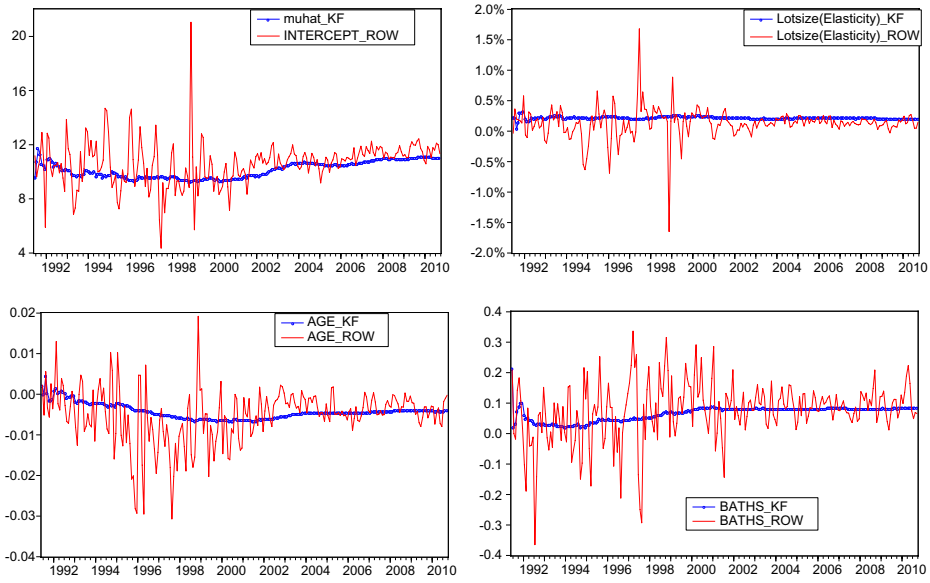

Figure 4. ROW and KF Estimates of Selected Parameters

estimates of these selected parameters with standard errors are also presented in the Appendix. Although the hyperparameters do need updating occasionally, they change slowly. Once they have been estimated with a reasonable sample size, the addition of one or two months of data leads to only minimal changes in the hyperparameter estimates, so updating every period is unnecessary. Estimates labeled ROW are obtained using the two-period rolling window approach ($t-1$ and t) in each estimation window.

Because the model is log-linear, the shadow prices for the parameters *bathrooms* and *Age* are given by $\beta_{baths,t} \times Price_t$ and $\beta_{age,t} \times Price_t$, respectively. These examples illustrate that the ROW approach produces unreasonably volatile estimates of the shadow prices for these relatively stable parameters compared to the state-space approach. With the exception of the first two years, during which it is settling, the estimates from the KF exhibit a slow time-varying trend as expected. In contrast, the ROW estimates are highly volatile, and often unstable. For the parameter *Age*, for instance, we can observe a considerable number of window estimates that are positive, indicating the structure appreciates with age, which is clearly not theoretically correct, except in the case of vintage effects. Morover, the volatility of ROW estimates is clearly related to the influence of the composition and size of the sample in each estimation window. The number of transactions available per window in the second half of the sample is much larger, as is obvious from Figure 1. This translates into much less volatile ROW estimates over the portion of the sample after the year 2002. In contrast, the state-space based estimates exhibit a consistent, relatively low volatility across the entire sample.

In general we find the rate of variation in the elements of β_t to be very slow for this market over the sample period. This may be due to a truly low level of

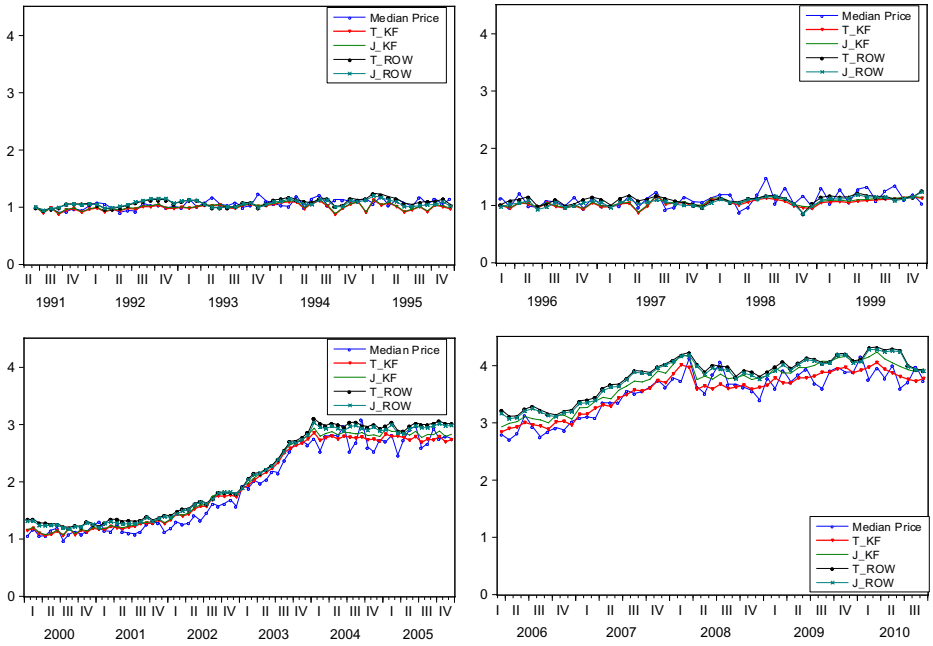

Figure 5. RPPIs Base 1991:6 Using Model with Spatial Regressors and No Spatial Error; Bottom Panels are Based on Hyperparameters Estimated with Sample 1991:6–1999:12

variability in the market studied, or due to the restrictions placed on σ_β as part of the model structure employed. In comparison to the current study, Rambaldi and Rao (2011, 2013) used the same model with data for the city of Brisbane over the period 1985–2005 and found the variation in some of the parameters to be larger than those found here. Without further analysis, however, it is not possible to conclude whether the low variation in β_t is intrinsic to the study market, or due to the restrictions placed on σ_β as part of the analysis.

5.3. *Computed Price Indexes*

In this section, we study the effect of using predictions based on ROW and KF. A comparison including estimators from alternative models of location is presented in the Appendix. A larger number of results and comparisons are available from the authors.

Figure 5 presents the computed Jevons (J) and Törnqvist (T) indexes using ROW and KF estimates, as well as the monthly median of the observed prices which has been converted into an index by setting the 1991:6 median price to one and referring each month's median to that base. This is shown to provide a quality unadjusted estimate of the evolution of prices as a comparison (labeled *median price*). There are periods when the median price is volatile, as is expected. The index's base is 1991:6 = 1; however, to facilitate visual depiction, the top panels show the evolution of the index from 1991:6–1995:12 and 1996:1–1999:12, and the bottom panels show the evolution from 2000:1–2005:12 and 2006:1–2010:09. At the beginning of the sample there is very little visible difference between the median price, the ROW and the KF based indexes. This is

expected because KF parameter estimates are volatile. From 1996 it becomes more apparent that the volatility of the median price is higher than that of the constructed indexes, and some separation between the ROW and KF based indexes is visible. It is clear there was not much movement in the market over this period. The indexes remained around one and only started to increase above one from around 1999. A period of rapid increase is observed between 2001 and 2003, prices plateaud in 2004–2005 and rose again until the global financial crises (GFC) where there was a sudden drop in early 2008. Prices remained stable to slightly rising until early 2010 where a decrease is observed again.

The period from 2000 onwards (two bottom panels) is where the bulk of transactions were observed and where the behavior of the different indexes and the median price differ substantially. Starting with the KF based indexes, we note that the J and T indexes are very similar in value until around 2004, where they diverge between 2004 and up to the GFC, at which point they converge for a couple of months, diverging again for the rest of the period. The movement of the T index is that of a smoothed trend of the median price, whereas the J index has a tendency to deviate more from the median. This separation arises from two possible sources interacting with the weighting of each index. The first source is the proportion of properties sold in the upper and lower part of the price distribution during heated and thinner markets. The second is a differential price change experienced by properties in the upper vs the lower part of the price distribution during heated and very slow market periods.

Consider a period of high demand, leading to rapidly rising prices, similar to that experienced in this market from mid 2001 to the end of 2003 (as evidenced by the large number of transactions illustrated in Figure 1). In this case all types of property might be rising in value similarly and thus the two indexes are very close. Consider now the period leading up to the global financial crises (2006–2007). Prices are increasing, but the market is not moving as rapidly (the number of transactions is more volatile; see Figure 1). In this case, the price changes of properties in the upper and lower parts of the price distribution might be different, and it seems reasonable to assume that the ratio of lower priced to upper priced houses in the market might be higher than in the previous situation. Given that the J index weights the transactions equally, the value of the index will be above that of the T index. In the early 2008 period when the market was hardly moving (note the significant drop in the number of transactions in Figure 1), the two indexes converge again as all types of property suffered similar price decreases during the GFC.

In contrast, none of these patterns are visible when the indexes are computed from ROW estimates. The J and T indexes are virtually indistinguishable over the whole sample. In general they are more volatile than their KF counterparts and higher in value. A large difference opens between the ROW and KF based indexes in January 2004 (see bottom left panel of Figure 5). The index values at this point are: median price 2.76, J-KF is 2.94, T-KF is 2.86, J-ROW is 3.03, and T-ROW is 3.10. This coincides with a sudden drop in the number of transactions which commences towards the end of 2003 and lasts over the 2004–2005 period, indicating a slower moving market over the period. It is clear that the small samples affect the ROW-based index greatly at this point. The level of the ROW based indexes persists above that of the KF based indexes and the median price for the rest of the sample period. This is an example of a chain drift in the hedonic index.

6. Conclusions

In this paper we consider the relationship between Törnqvist and Jevons hedonic imputed indexes and some of the choices of econometric estimation and model specification made to impute prices. First, we compare the rolling window estimation approach, which is popular in the price index literature, to estimation by smoothing methods. Second, we compare two alternative approaches to controlling for property location in the model.

The main difference between the rolling window and the smoothing methods is in the way information is weighted. We show that the Kalman filter is the most appropriate approach for the task as it optimally weights current and past market information when computing the indexes. The rolling window approach does not produce estimates that are attenuated over time, because in an M–period rolling window there is no recognition of the time series ordering of the transactions. The proposed Kalman filter approach is simple to compute and does not require frequent revisions of the published index as new data become available.

Using the data on the individual properties' latitude and longitude, two alternative approaches to control for property location are studied. The first is the use of a spatial lag model in the error term and the second is the use of spatial regressors generated using GIS software. For the market under consideration the model with spatial regressors and no spatial errors performs best (using BIC).

We construct monthly indexes. There is a substantial difference in the constructed Törnqvist and Jevons indexes based on the Kalman filter and the rolling window approach. The indexes based on the rolling window are more volatile and appear to be adversely affected by sharp turns in the market. This leads to chain drift in the computed indexes. This is because in thin markets the number of transactions drops and, unlike the Kalman filter, the rolling window approach does not link to previous transactions. In addition, the equal weighting of transactions within a window results in ROW being less sensitive to turns in the market which can result in over/under prediction of prices. The Törnqvist and Jevons indexes differ in value during periods of market volatility. This is expected given the different weighting of transactions between them and the likelihood that the rate of change in prices differs between properties at the high and low end of the price distribution during periods of volatile markets.

References

Court, A. T., "Hedonic Price Indexes with Automotive Examples," in C. F. Roos (ed.), *The Dynamics of Automobile Demand*, General Motors Corporation, New York, 99–117, 1939.

Diewert, E. W., "Hedonic Regressions: A Consumer Theory Approach," in *Scanner Data and Price Indexes, NBER Conference on Research in Income and Wealth, Studies in Income and Wealth*, Volume 64, University of Chicago Press, Chicago, IL, 318–48, 2003.

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

European Commission, Eurostat, OECD, and World Bank, *Handbook on Residential Property Price Indexes (RPPI)*, Publications Office of the European Union. Luxembourg, 2013.

Fletcher, C. S., R. R. J. McAllister, A. N. Rambaldi, and K. Collins, "The Economics of Adaptation to Protect Appreciating Assets from Coastal Inundation," Final Report F110609, CSIRO Climate Adaptation Flagship, 2011.

Francke, M. K., "Repeat Sales Index for Thin Markets," *Journal of Real Estate Finance and Economics*, 41, 24–52, 2010.

Francke, M. K. and A. F. de Vos, "Efficient Computation of Hierarchical Trends," *Journal of Business and Economic Statistics*, 18, 51–7, 2000.
Francke, M. K. and G. A. Vos, "The Hierarchical Trend Model for Property Valuation and Local Price Indices," *Journal of Real Estate Finance and Economics*, 28, 179–208, 2004.
Gelfand, A. E., H. Kim, C. F. Sirmans, and S. Banerjee, "Spatial Modeling with Spatially Varying Coefficient Processes," *Journal of the American Statistical Association*, 98, 387–96, 2003.
Greene, W., *Econometric Analysis*, 7th edition, Pearson Education, Boston, MA, 2012.
Griliches, Z., "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*, 1961.
Harvey, A. C., *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge, 1989.
———, "Forecasting with Unobserved Components Time Series Models," in G. Elliott, C. W. J. Granger, and A. Timmermann (eds), *Handbook of Economic Forecasting, Volume 1*, Elsevier Science (doi: 10.1016/S1574-0706(05)01007-4), 2006.
Hill, R., "Hedonic Price Indexes for Housing," OECD Statistics Working Papers, 2011/01, OECD Publishing (http://dx.doi.org/10.1787/5kghzxpt6g6f-en), 2011.
Hill, R. J. and D. Melser, "Hedonic Imputation and the Price Index Problem: An Application to Housing," *Economic Inquiry*, 46, 593–609, 2008.
Hill, R. and M. Scholz, "Incorporating Geospatial Data into House Price Indexes: A Hedonic Imputation Approach with Splines," Ottawa-Group, Statistics Denmark, Copenhagen, 2013.
Knight, J. R., J. Dombrow, and C. F. Sirmans, "A Varying Parameters Approach to Constructing House Price Indexes," *Real Estate Economics*, 23, 187–205, 1995.
Koopman, S. and A. Harvey, "Computing Observation Weights for Signal Extraction and Filtering," *Journal of Economic Dynamics and Control*, 27, 1317–33, 2003.
LeSage, J. P. and R. K. Pace, *Introduction to Spatial Econometrics*, Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton, FL, 2009.
Pace, R. K., R. Barry, O. W. Gilley, and C. F. Sirmans, "A Method for Spatial-Temporal Forecasting with an Application to Real Estate Prices," *International Journal of Forecasting*, 16, 229–46, 2000.
Quan, D. and J. Quigley, "Price Formation and the Appraisal Function in Real Estate Markets," *Journal of Real Estate Finance and Economics*, 4, 175–90, 1991.
Rambaldi, A. N. and D. S. P. Rao, "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, 2011.
———, "Econometric Modeling and Estimation of Theoretically Consistent Housing Price Indexes," No. WP04/2013 in CEPA Working Papers Series, School of Economics, University of Queensland, 2013.
Schwann, G. M., "A Real Estate Price Index for Thin Markets," *Journal of Real Estate Finance and Economics*, 16, 269–87, 1998.
Silver, M. and S. Heravi, "The Difference between Hedonic Imputation Indexes and Time Dummy Hedonic Indexes," *Journal of Business & Economic Statistics*, 25, 239–46, 2007.
Triplett, J., "Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products," OECD Science, Technology and Industry Working Papers, 2004/9, ISBN-92-64-02814-5 (http://www.oecd.org/dataoecd/37/31/33789552.pdf), 2004.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix A: Location modeling—Comparison of Parameter Estimates and Computed Indexes
Figure A1: Parameter Estimates Comparison with Alternative Modeling of Location
Figure A2: Parameter Estimates Comparison with Alternative Modeling of Location
Figure A3: Hedonic Imputed Price Indexes—base 1991:6. Effect of Different Specification of Location
Appendix B: Parameter Estimates with Standard Errors
Figure B1: KF Estimates and Two-Standard Error Bound shown. Model with No Spatial Errors and Spatial Regressors (Hyperparameters estimated with sample from 1991:5–1999:12