Incorporating Geospatial Data into House Price Indexes: A Hedonic Imputation Approach with Splines

Robert J. Hill and Michael Scholz
Department of Economics
University of Graz
Universitätsstrasse 15/F4
8010 Graz, Austria
robert.hill@uni-graz.at, michael.scholz@uni-graz.at

Preliminary Draft to Be Presented at Princeton Conference on ICP/PPP
(May 23-24, 2013)

Abstract:
We estimate a hedonic model of the housing market that includes a spline surface defined on geospatial data (i.e., the longitudes and latitudes of individual dwellings). House price indexes are then obtained by imputing prices for individual dwellings from the hedonic model and then inserting them into the Fisher price index formula. Using data for Sydney, Australia we compare the performance of four models: (i) generalized additive model (GAM) with a geospatial spline, (ii) GAM with postcode dummies, (iii) semilog with geospatial spline, (iv) semilog with postcode dummies. Our results clearly confirm the superiority of geospatial-splines, both in terms of the deviation between actual and imputed prices and in the case of repeat-sales between actual and imputed price relatives. Furthermore, the use of geospatial splines significantly affects the resulting house price indexes. The cumulative increase in our Fisher price indexes is between 8 and 30 percent higher (depending on the functional form and data sample) over the 2001 to 2011 period when a geospatial spline is used. This difference can be attributed to the failure of postcode dummies to fully adjust for omitted locational characteristics. Price indexes generated using geospatial splines are also more robust to changes in the functional form of the hedonics model and relatively immune to sample selection bias. (*JEL*. C43; E01; E31; R31)
Keywords: Housing market; Hedonic model; Price index; Spline; Geospatial data; Generalized additive model

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

1 Introduction

Every house is different both in terms of its physical characteristics and its location. It is important that house price indexes take account of these quality differences. Otherwise the price index will confound price changes and quality differences.

Hedonic methods which express house prices as a function of a vector of characteristics are potentially useful for this purpose. In recent years, the increased availability of housing data and improvements in computing power have together led to a surge in the number of providers of hedonic house price indexes around the world (see Hill 2013).

Most hedonic indexes at present adjust for location using postcode dummy variables. The increased availability of geospatial data (i.e., longitudes and latitudes), however, means that a more sophisticated approach is possible. Geospatial data have thus far been largely ignored for two reasons. First, the use of geospatial data clearly complicates the index computation process. Second, most existing hedonic indexes use the average-characteristics method, which defines an average house and then measures how the price of this hypothetical average house changes over time. While it may be meaningful to average the physical characteristics of the houses (such as land area and number of bedrooms), it makes no sense to average their longitudes and latitudes. In other words, the inclusion of geospatial data will require a shift away from average-characteristics methods.

The two main alternatives to the average-characteristics method are the time-dummy and imputation methods. Time-dummy methods include a dummy variable for each period. The price index for that period is then obtained directly from the estimated coefficient on the dummy variables (the coefficient must be exponentiated when the semilog model is used). By contrast, imputation methods use the hedonic model to impute a price for each individual house. These imputed prices can then be fed into a standard price index formula.

One way of including geospatial data in a hedonic model is through a spatial-autoregressive specification. The spatial dependence is captured through a distance weights matrix that is itself derived from the geospatial data. However, the focus of these papers is typically on the estimation of a hedonic model for a single period rather than the computation of hedonic price indexes.

There are a few examples where spatial-autoregressive models have been combined with time-dummy methods. In our opinion, however, this is not the best way of using geospatial data. The main problem with spatial-autoregressive models is that they impose a lot of prior structure (that may not be valid) on the spatial dependence of house prices.

A more flexible way of incorporating geospatial data is to include it as a nonparametric surface such as a spline in the hedonic model, and then compute the price indexes using an imputations method. This is the approach we follow here. We implement this approach in two different ways. First, we combine a geospatial spline with a generalized additive model (GAM) defined on the physical characteristics. Second, we combine a geospatial spline with a semilog model on the physical characteristics. We then compare these two models with equivalent GAM and semilog models that use postcode dummies instead of geospatial splines.

Applying our methods to data for Sydney, Australia covering the period 2001 to 2011, our results clearly confirm the superiority of geospatial-splines, both in terms of the deviation between actual and imputed prices and in the case of repeat-sales between actual and imputed price relatives.

An interesting feature of our data set is that all our hedonic indexes rise faster than simple median or mean indexes. This indicates that the average quality of houses sold deteriorates during our sample period. Furthermore, we find that our indexes that use geospatial data rise faster than our indexes that use postcode dummies. This is because part of this quality shift is from better to worse locations, and this shift can be discerned even within postcodes. A hedonic index that uses postcode dummies therefore fails to fully quality adjust for this shift. In the case of Sydney, as a result of the underlying dynamics in the market, this failure to fully quality adjust causes a downward bias in the price index (by between 8 and 30 percent over 11 years depending on the functional form and data sample used). Turning this result around, our results also imply that house prices in Sydney have risen rather more than previously realized over this period.

The remainder of this paper is structured as follows. Section 2 provides an overview of the hedonic price index literature. Section 3 discusses ways of incorporating location into a hedonic house price index. Section 4 presents our data set and hedonic models, compares the performance of these models, and then derives the resulting hedonic price indexes. Section 5 concludes the paper.

2 Hedonic Price Indexes for Housing

2.1 An overview

A hedonic model regresses the price of a product on a vector of characteristics (whose prices are not independently observed). The hedonic equation is a reduced form equation that is determined by the interaction of supply and demand. One use of hedonic models is for constructing quality-adjusted price indexes. The majority of research in this field has focused on products subject to rapid technological change, such as computers (see e.g., Dulberger, 1989, and Berndt et al., 1995).

Hedonic methods can also be used to construct quality-adjusted price indexes for differentiated products. Housing is an extreme case of a differentiated product in the sense that every house is different. One can distinguish between a house's physical and locational attributes. Examples of the former include the number of bedrooms and land area, while examples of the latter include the exact longitude and latitude of a house, and the distance to local amenities such as a shopping center, park or school.

The hedonic approach can be implemented in three main ways (see Hill 2013). We briefly discuss each of these below.

2.2 Time-dummy methods

The time-dummy method is the original hedonic method. It typically uses the semi-log functional form – see Diewert (2003) and Malpezzi (2003) for a discussion of some of the advantages of the semi-log model in this context. A standard semi-log formulation is as follows:

$$y = Z\beta + D\delta + \varepsilon, \tag{1}$$

where y is an $H \times 1$ vector with elements $y_h = \ln p_h$, Z is an $H \times C$ matrix of characteristics (some of which may be dummy variables), β is a $C \times 1$ vector of characteristic shadow prices, D is an $H \times (T-1)$ matrix of period dummy variables, δ is a $(T-1) \times 1$ vector of period prices (with the base period price index normalized to 1), and ε is an $H \times 1$ vector of random errors. Finally, H, $C-1$ and T denote respectively the number of dwelling, characteristics and time periods in the data set. The first column in Z consists of ones, and hence the first element of β is an intercept. It is possible also to include functions of characteristics (for example land size entering the model as a quadratic function), and interaction terms between characteristics.

When the objective of the exercise is to construct a quality-adjusted price index, the primary interest lies in the δ parameters which measure the period-specific fixed effects on the logarithms of the price level after controlling for the effects of the differences in the attributes of the dwellings. One attraction of the semi-log time-dummy model is that the price index P_t for period t is derived by simply exponentiating the estimated coefficient $\hat{\delta}_t$ obtained from the hedonic model:[1]

$$\hat{P}_t = \exp(\hat{\delta}_t). \tag{2}$$

Although it is the original hedonic method and is widely used in other contexts, the time-dummy method has received little attention in a housing context. This is perhaps due its lack of flexibility, in that the shadow prices cannot evolve over time and because each time a new period is added to the data set all the results need to be recomputed. A more flexible version of the method only compares adjacent periods. A longer time series is then obtained by chaining these bilateral comparisons together. The adjacent period (AP) version of the method is used by RPData-Rismark in Australia and Informations und Ausbildungszentrum für Immobilien in Switzerland.

2.3 Imputation methods

The hedonic imputation approach estimates a separate hedonic model for each period or a few adjacent periods.[2] The hedonic model is then used to impute prices for individual dwellings. For example, let $\hat{p}_{t+1,h}(z_{t,h})$ denote the imputed price in period $t+1$ of a dwelling sold in period t. This price is imputed by substituting the characteristics of dwelling h sold in period t into the estimated hedonic model of period $t+1$ as follows:

$$\hat{p}_{t+1,h}(z_{t,h}) = \exp\left(\sum_{c=1}^{C} \hat{\beta}_{c,t+1} z_{c,t,h}\right). \tag{3}$$

[1] While $\hat{P}_t$ is a biased estimator of P_t, Hill, Melser and Syed (2009) show that at least for the Sydney data set used here the bias is so small it can be ignored.

[2] The appropriate time horizon for each model depends partly on the size of the data set. For example, for our Sydney data, there are enough data to estimate the model separately for each year. However, even when the focus is on quarterly indexes, we would not recommend estimating the model separately for each quarter. We recommend estimating the model on an annual basis and including quarterly dummy variables.

These imputed price indexes can then be inserted into standard price index formulas. We will refer to a formula that focuses on the dwellings that sold in the earlier period t as Laspeyres-type, and a formula that focuses on the dwellings that sold in the later period $t+1$ as Paasche-type. Our price indexes are constructed by taking the geometric mean of the price relatives, giving equal weight to each dwelling.[3] Taking a geometric mean of Laspeyres and Paasche type indexes, we obtain a Fisher-type index that has the advantage that it treats both periods symmetrically.

In a hedonic setting a further complication is that while the counterfactual prices (i.e., the prices in period $t+1$ for Laspeyres and in period t for Paasche) must be imputed, we have a choice whether or not to impute prices in period t for Laspeyres and in period $t+1$ for Paasche. A *single imputation* Laspeyres or Paasche price index imputes in only one period, while a *double imputation* index imputes in both periods (see Silver and Heravi 2001, Pakes 2003, de Haan 2004, and Hill and Melser 2008).

The price index between periods t and $t+1$ is calculated as follows:

$$\begin{aligned}
&\text{Paasche Single Imputation}: \ P^{PSI}_{t,t+1} = \prod_{h=1}^{H_{t+1}} \left[\left(\frac{p_{t+1,h}}{\hat{p}_{t,h}(z_{t+1,h})} \right)^{1/H_{t+1}} \right] \\
&\text{Laspeyres Single Imputation}: \ P^{LSI}_{t,t+1} = \prod_{h=1}^{H_{t}} \left[\left(\frac{\hat{p}_{t+1,h}(z_{t,h})}{p_{t,h}} \right)^{1/H_{t}} \right] \\
&\text{Fisher Single Imputation}: \ P^{FSI}_{t,t+1} = \sqrt{P^{PSI}_{t,t+1} \times P^{LSI}_{t,t+1}} \\
&\qquad = \sqrt{\prod_{h=1}^{H_{t+1}} \left[\left(\frac{p_{t+1,h}}{\hat{p}_{t,h}(z_{t+1,h})} \right)^{1/H_{t+1}} \right] \times \prod_{h=1}^{H_{t}} \left[\left(\frac{\hat{p}_{t+1,h}(z_{t,h})}{p_{t,h}} \right)^{1/H_{t}} \right]}
\end{aligned} \tag{4}$$

$$\begin{aligned}
&\text{Paasche Double Imputation}: \ P^{PDI}_{st} = \prod_{h=1}^{H_{t+1}} \left[\left(\frac{\hat{p}_{t+1,h}(z_{t+1,h})}{\hat{p}_{sh}(z_{t+1,h})} \right)^{1/H_{t+1}} \right] \\
&\text{Laspeyres Double Imputation}: \ P^{LDI}_{st} = \prod_{h=1}^{H_{t}} \left[\left(\frac{\hat{p}_{t+1,h}(z_{t,h})}{\hat{p}_{t,h}(z_{t,h})} \right)^{1/H_{t}} \right] \\
&\text{Fisher Double Imputation}: \ P^{FDI}_{t,t+1} = \sqrt{P^{PDI}_{t,t+1} \times P^{LDI}_{t,t+1}} \\
&\qquad = \sqrt{\prod_{h=1}^{H_{t+1}} \left[\left(\frac{\hat{p}_{t+1,h}(z_{t+1,h})}{\hat{p}_{t,h}(z_{t+1,h}} \right)^{1/H_{t+1}} \right] \times \prod_{h=1}^{H_{t}} \left[\left(\frac{\hat{p}_{t+1,h}(z_{t,h})}{\hat{p}_{t,h}(z_{t,h})} \right)^{1/H_{t}} \right]}
\end{aligned} \tag{5}$$

[3] This democratic weighting structure is in our opinion more appropriate in a housing context than weighting each dwelling by its expenditure share.

Imputation methods require reasonably large data sets. However, this is becoming less of a constraint than it used to be, given the large increase in data availability. Imputation methods are flexible in that they allow the characteristic shadow prices to evolve over time. Even so they have not been used much. This may be because they are conceptually more complicated than time-dummy and average-characteristics methods. The only index providers to use an imputation method as far as we are aware are the FNC Residential Price Index in the US and RPData-Rismark in Australia. The FNC index described in Dorsey et al. (2010) uses the double imputation Laspeyres formula in the context of a SASAR(1,1) model (see section 3.3), while RPData-Rismark use a nonparametric method (see Hardman 2011 - although not enough information is provided to know how the method works).

2.4 Average-characteristics methods

Average-characteristics methods, like imputation methods, generally estimate the hedonic model separately for each period. They also use standard price index formulas. The key difference is that an average-characteristics price index is defined in characteristics space. Average-characteristics methods typically construct an average dwelling for each period, and then impute the price of this hypothetical dwelling (which for example may have two and a half bedrooms) as a function of its characteristics using the shadow prices derived from the hedonic model. A price index is obtained by taking the ratio of the imputed price of the same average dwelling in two different periods. By construction the average-characteristics method uses double imputation since the average dwelling is hypothetical rather than an actual dwelling.

Taking the semi-log hedonic model as our point of reference, a price index between periods s and t can be calculated using the average dwelling from either period (see Dulberger 1989 and Diewert 2001). In this way we obtain Laspeyres, Paasche and Fisher-type indexes.

$$\text{Laspeyres}: \ P^L_{t,t+1} = \hat{p}_{t+1}(\bar{z}_t)/\hat{p}_t(\bar{z}_t) = \exp\left[\sum_{c=1}^{C}(\hat{\beta}_{c,t+1} - \hat{\beta}_{c,t})\bar{z}_{c,t}\right],$$

$$\text{Paasche}: \ P^P_{t,t+1} = \hat{p}_{t+1}(\bar{z}_{t+1})/\hat{p}_t(\bar{z}_{t+1}) = \exp\left[\sum_{c=1}^{C}(\hat{\beta}_{c,t+1} - \hat{\beta}_{c,t})\bar{z}_{c,t+1}\right],$$

$$\text{Fisher}: \ P^F_{t,t+1} = \sqrt{P^L_{t,t+1} \times P^P_{t,t+1}} = \exp\left[\frac{1}{2}\sum_{c=1}^{C}(\hat{\beta}_{c,t+1} - \hat{\beta}_{c,t})(\bar{z}_{c,t} + \bar{z}_{c,t+1})\right], \quad (6)$$

$$\text{where } \bar{z}_{c,t} = \frac{1}{H_t}\sum_{h=1}^{H_t} z_{c,t,h} \quad \text{and } \bar{z}_{c,t+1} = \frac{1}{H_{t+1}}\sum_{h=1}^{H_{t+1}} z_{c,t+1,h}.$$

The main strength of the average-characteristics method is its intuitive interpretation as measuring the change in the price of the average dwelling over time. Its biggest weakness is that it cannot easily be extended to incorporate geospatial data. We return to this issue in the next section.

The average-characteristics method in its various guises has proved to be by far the most popular for computing hedonic house price indexes. The New House Price Index computed by the Census Bureau in the US, the Halifax and Nationwide indexes in the UK, and the permanent tsb index in Ireland are calculated using the Laspeyres version of the characteristics method with a semi-log functional form for the hedonic equation (see US Census Bureau undated, and Fleming and Nellis 1985). Statistics Finland uses an implicit Paasche price index (see Saarnio 2006) to compute house price indices for Finland. Statistics Norway uses the same method as Statistics Finland except that it calculates its hedonic model using the previous five years of data and chains the index on an annual basis (see Thomassen 2007). Statistics Sweden also uses a variant on an implicit Laspeyres price index (see Ribe 2009), although the exact details of the method are not provided. Closely related to these Nordic methods is the Conseil Supérieur du Notariat (CSN) and INSEE method used to compute hedonic indexes for regions in France (see Gouriéroux and Laferrère 2009).

3 Methods for Incorporating Location into House Price Indexes

3.1 Postcode dummy variables

One of the key determinants of house prices is location. The explanatory power of the hedonic model can therefore be significantly improved by exploiting information on the location of each property. Probably the simplest way to do this is to include postcode identifiers for each dwelling in the hedonic model. Hill, Melser and Syed (2009) find that the inclusion of postcode dummy variables increases their R-squared coefficients from about 0.56 to 0.76. Postcode dummies can be used in combination with any of the time-dummy, imputation, and average-characteristics methods.

3.2 Distances to amenities

Given the availability of geospatial data, the distance of each dwelling to landmarks such as the city center, airport, nearest train station, or nearest beach can be measured. These distances (or some function of them) can then be included as additional characteristics in the time-dummy or imputation versions of the hedonic model (see for example Hill and Melser 2008). Such an approach, however, does not work with the characteristics method. Averaging longitudes and latitudes throws away the underlying spatial dependence. Also, the average geospatial location may anyway not make much sense. For example, in the case of a city like Sydney built round a natural harbor it may be underwater!

Using distances to amenities as characteristics is problematic even for the time-dummy and imputation methods for a few reasons. First, it makes only limited use of the available geospatial data, and hence throws away a lot of potentially useful information. Second, direction (i.e, north, south, east or west) matters as well as distance. For example, in the case of an airport, a house's position relative to the flight path is at least as important as the actual distance from the airport. Third, the impact of distance from an amenity on the price of a house may be quite complicated and not necessarily monotonic. For example, one may want to live not too close and not too far from the city center, airport, etc.

3.3 Spatial-autoregressive models

The inclusion of postcode dummies or distance-to-amenities characteristics will only partially capture the effect of locational omitted variables. Locational effects can be captured more effectively by a spatial autoregressive model. For example, the (first-order) autoregressive spatial model with (first-order) autoregressive errors, referred to henceforth as the SARAR(1,1) model, has been widely used for this purpose (see for example Anselin 1988 and Corrado and Fingleton 2011).

$$y = \rho S y + X\beta + u,$$

$$u = \lambda M u + \varepsilon,$$

where y is the vector of log prices, (i.e., each element $y_h = \ln p_h$), and S and M are spatial weights matrix that are calculated from the geospatial data. Often S and M

are the same.

A simplified version of the SARAR(1,1) model where λ is set to zero (typically referred to as the spatial lag model) is used in the construction of the FNC Residential Price Index (see Dorsey et al. 2010), and by Ord (1975), Can and Megbolugbe (1997), and Kim, Phipps and Anselin (2003), and others. It can be rationalized by buyers and sellers treating the price at which nearby dwellings sell as a signal of value. Alternatively, sometimes ρ is set to zero. LeSage and Pace (2009) provide an externality motivation for this version of the model (typically referred to as the spatial error model) where the quality of nearby dwellings directly influences the price of a particular dwelling. This version of the SAR model is used for example by Cliff and Ord (1973), Pace and Gilley (1997), Bell and Bockstael (2000), and Hill, Melser and Syed (2009).

Methods used to estimate the β vector and ρ and λ scalars of the SARAR(1,1) model include maximum likelihood (see Anselin 1988, and Pace and Barry 1997), two-stage-least squares (2SLS) (see Anselin 1988, Kelejian and Prucha 1998 and Lee 2003), and generalized method of moments (GMM) (see Lee 2007, Kelejian and Prucha 2010, and Liu, Lee and Bollinger 2010). 2SLS and GMM estimators, while generally less efficient than ML, have the advantage of relying on weaker assumptions and being computationally simpler (see Lee 2007).

Most of this literature, however, is concerned with the estimation of a hedonic model at a point in time rather than the construction of house price indexes. In principle, such indexes can be easily obtained by simply including quarter or year dummies in the X characteristics matrix, and then by exponentiating the estimated parameters on these dummy variables. The problem is that when the model is estimated over a number of years of data the spatial weights matrix S should be replaced by a spatiotemporal weights matrix W. That is, the magnitude of the dependence between observations depends inversely on both their spatial and temporal separation. One response to this problem is to use the adjacent-period (AP) version of the time-dummy method. In this case the temporal separation between observations never gets that large and hence it is more defensible to use a spatial weights matrix instead of the theoretically preferred spatiotemporal weights matrix. This is the approach followed by Hill, Melser and Syed (2009), and the FNC Residential Price Index, which is computed on a monthly basis using a 12-month moving window, using the spatial lag version of the SARAR(1,1) model (see Dorsey et al. 2010). Rambaldi and Rao (2011) also use a variant on this approach.

An alternative approach is to actually compute a spatiotemporal weights matrix. The literature on this topic is thin. The main references are Pace, Barry, Clapp and Rodriques (1998), Tu, Yu and Sun (2004), Sun, Tu and Yu (2005) and Nappi-Choulet and Maury (2009).

The main problem with spatial autoregressive models is that they impose a lot of prior structure on the spatial dependence. This point is made forcefully by Pinkse and Slade (2010). They criticize SAR(1) models, and by extension models where the errors also "have some simple spatial dependence relationship". In other words, their criticisms apply to the SARAR(1,1) model as well.

> The limitations of the SAR(1) model are endless. These include: (1) the implausible and unnecessary normality assumption, (2) the fact that if y_i depends on spatially lagged ys, it may also depend on spatially lagged xs, which potentially generates reflection-problem endogeneity concerns . . ., (3) the fact that the relationship may not be linear, and (4) the rather likely possibility that u and X are dependent because of, e.g., endogeneity and/or heteroskedasticity.
>
> Even if one were to leave aside all of these concerns, there remains the laughable notion that one can somehow know the entire spatial dependence structure up to a single unknown multiplicative coefficient [*two unknown coefficients in the case of SARAR(1,1)*]. (Pinkse and Slade 2010, p. 106 - text in italics added by the authors)

3.4 Nonparametric approaches

Nonparametric methods provide an alternative to parametric modeling of spatial dependence that largely avoid the problems highlighted by Pinkse and Slade (2010). Nonparametric methods can be used to construct a flexible topographical surface describing how price varies by location (measured by longitude and latitude) holding the other characteristics fixed. Such a surface can then be added to a parametric or nonparametric hedonic model defined over the physical characteristics. Examples of this type of approach include Colwell (1998), Pavlov (2000), Clapp, Kim and Gelfand (2002), Fik, Ling and Mulligan (2003), Clapp (2003, 2004), McMillen and Redfearn (2010), Brunauer et al. (2010), and Hardman (2011). What all these papers lack, with the exception of Hardman (2011), is a method for obtaining a house price index from the estimated hedonic model. The imputation method is the natural choice for this task.[4]

More specifically, consider the single imputation Fisher price index in (4). Imputed prices in period t of houses actually sold in period $t+1$, denoted by $\hat{p}_{t,h}(z_{t+1,h})$, can be derived from the hedonic model of period $t+1$. That is, one can take the physical characteristics and longitude/latitude of house h and insert them into the hedonic model of period $t+1$ to obtain an imputed price. Similarly, imputed prices in period $t+1$ of houses actually sold in period t, denoted by $\hat{p}_{t+1,h}(z_{t,h})$, can be derived from the hedonic model of period t. This is all the hedonic model is required for, to make sure that prices are available for each house included in the price index formula in both period t and $t+1$.

The hedonic imputation method can therefore be applied to any hedonic function that provides an imputed price, irrespective of whether the functional form is parametric or nonparametric. In this sense the imputation method is very flexible and provides a natural way of incorporating geospatial data in the form of a nonparametric surface such as a spline into the index calculation.[5]

4 Empirical Strategy

4.1 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 2001–2011. For each dwelling we have the following characteristics: the actual sale price, time of sale, postcode, property type (i.e., detached house or semi), number of bedrooms, number of bathrooms and land area. In addition, we have the exact address and longitude and

[4] Hardman (2011), which describes the method used to compute the Daily Home Value Index of RPData-Rismark, however, is rather vague and lacking in detail.

[5] As far as we are aware the only papers to use splines in the estimation of hedonic models of the housing market are Bao and Wan (2004) and Brunauer, et al. (2010). However, Bao and Wan do not have geospatial data. They fit a spline to the physical characteristics of houses. Also, neither Bao and Wan nor Brunauer and et al. construct house price indexes.

latitude of each dwelling. Some summary statistics are provided in Table 1.

Table 1: Summary of characteristics

	PRICE	BED	BATH	AREA	LAT	LONG
Minimum	100000	1.000	1.000	100.0	-34.20	150.6
1st Quartile	426000	3.000	1.000	461.0	-33.92	150.9
Median	615000	3.000	2.000	590.0	-33.83	151.1
Mean	758311	3.454	1.744	631.8	-33.84	151.1
3rd Quartile	885000	4.000	2.000	725.0	-33.75	151.2
Maximum	4000000	6.000	6.000	9994.0	-33.40	151.3

For a robust analysis it was necessary to remove some outliers. This is because there are a much higher proportion of data entry errors in the tails, caused for example by the inclusion of erroneous extra zeroes. These extreme observations can distort the results. The exclusion criteria we applied are shown in Table 2.

Table 2: Criteria for removing outliers

	PRICE	BED	BATH	AREA	LAT	LONG
Minimum Allowed	100000	1.000	1.000	100.0	-34.20	150.60
Maximum Allowed	4000000	6.000	6.000	10000.0	-33.40	151.35

We also exclude all townhouses from our analysis since the corresponding land area is for the whole strata. To keep the estimation procedure as simple as possible we also excluded from the raw data in a prior step all observations missing one or more of our characteristics.[6] Our data set consists in total of 454 507 house sales. Complete data on all our hedonic characteristics are available for 240 142 observations. Table 3 shows the distribution of houses with missing characteristics per year. It can be seen from Table 3 that the quality of the data improves over time. For this reason the fit of our hedonic models also improves in later years.

[6] For example, if the number of bedrooms is missing for a particular house, we exclude it. We intend to revisit this issue and redo our results including these houses.

Table 3: Number of observations per year with missing characteristics

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Total	51885	47351	47374	34734	34361	37072	42938	34601	44791	40114	39286
Missing											
-price	116	135	110	74	204	194	292	206	264	404	215
-long	2589	1994	1718	1194	1308	1347	1789	1807	4533	5027	5093
-lat	2589	1994	1718	1194	1308	1347	1789	1807	4533	5027	5093
-bed	34355	31294	29000	17382	9754	8747	8921	5978	8471	6512	480
-bath	45834	40987	39435	25871	12314	10143	9404	6053	8566	6613	484
-area	582	547	450	353	399	462	605	488	466	500	494
Included	5886	6188	7759	8668	21662	26467	32936	28063	35703	32933	33877

4.2 Model estimation and performance

Here we compare the performance of four models:
(i) generalized additive model (GAM) with a geospatial spline;
(ii) GAM with postcode dummies;
(iii) semilog with geospatial spline;
(iv) semilog with postcode dummies.

Model (i) is nonparametric, (ii) and (iii) are semiparametric, and (iv) is parametric. In (iii) the geospatial data are modelled nonparametrically, while in (ii) it is the physical characteristics that are modelled nonparametrically. A GAM as used in (i) and (ii) has the advantage of being more flexible than semilog while avoiding the curse of dimensionality that arises in a fully nonparametric model (see for example Stone 1986).[7] Furthermore, it is relatively straightforward to include a bivariate function of the longitude and latitude in the modelling process and to account in this way for topographical (locational) effects in house prices. For an overview of non- and semi-parametric models, their properties and estimation, see Härdle et al. (2004).

Models (i) and (ii), which are estimated separately for each year $t = 2001, \ldots, 2011$,

[7] To be more precise, we use the Gaussian family with the identity link function. In future work we may explore alternative distributions of log-prices.

take the following form:

$$
\begin{aligned}
y &= c_1 + D\delta_1 + \sum_{c=1}^{C} f_{1,c}(z_c) + g_1(z_{lat}, z_{long}) + \varepsilon_1, &(7)\\
y &= c_2 + D\delta_2 + \sum_{c=1}^{C} f_{2,c}(z_c) + m_2(z_{pc}) + \varepsilon_2, &(8)
\end{aligned}
$$

where y is a $H \times 1$ vector of log-prices, c_i is a $H \times 1$ vector of constants, D is a $H \times 3$ matrix of quarterly dummy variables, δ_i is a 3×1 vector of parameters, $f_{i,c}$ an unknown function for characteristic z_c (in our case number of bathrooms, number of bedrooms and land area), g_i is a unknown bivariate function of latitude and longitude, m_i is an unknown function of the postcode, and ε_i an $H \times 1$ vector of error terms. In the semilog versions we replace the unknown functions f_c by their linear analogs:

$$
\begin{aligned}
y &= c_3 + D\delta_3 + \sum_{c=1}^{C} z_c \beta_{3,c} + g_3(z_{lat}, z_{long}) + \varepsilon_3, &(9)\\
y &= c_4 + D\delta_4 + \sum_{c=1}^{C} z_c \beta_{4,c} + \sum_{pc=1}^{250} z_{pc} m_{4,pc} + \varepsilon_4, &(10)
\end{aligned}
$$

where in (10) Z_{pc} is a $H \times 250$ matrix of postcode dummy variables (where 250 is the number of postcodes included in our data set), and m_4 is a 250×1 vector of parameters.

We estimate the splines in models (7) and (9) with 2500 randomly selected knots. The smoothing parameter is selected using Restricted Maximum Likelihood (REML).[8] Table 4 shows the values of the Akaike information criterion (AIC) for each model in each year, and Table 5 the sum of squared log errors, C_t, defined as follows:

$$
C_t = \left(\frac{1}{H_t}\right) \sum_{h=1}^{H_t} [\ln(\hat{p}_{th}/p_{th})]^2.
$$

In both Tables 4 and 5, a lower value implies a better fit. The C_t coefficients are bounded from below by zero, while the AIC can be negative.

We find that the models including geospatial splines, (7) and (9), dramatically outperforms their postcode-based competitors, (8) and (10), in terms of goodness-of-fit. Also, the GAM version of each model, (7) and (8), outperforms its semilog counterpart, (9) and (10), although in this case the differences are not so large.

[8] We use the implementation provided in the mgcv package of the R-Language. For details see for example Wood (2011).

Table 4: Akaike information criterion for models (7)-(10)

Model	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
(7)	416	89	-778	-1599	-7290	-6417	-8544	-10271	-14059	-14953	-18493
(8)	4888	5456	5780	5598	8635	11678	16233	11652	12819	12313	8696
(9)	-55	-85	-1093	-1571	-7192	-6199	-8917	-10286	-15529	-14649	-18520
(10)	4730	5337	5677	5571	8630	11677	16009	11564	12086	12307	8662

Table 5: Sum of squared log errors for models (7)-(10)

Model	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
(7)	0.061	0.057	0.051	0.047	0.041	0.046	0.045	0.040	0.039	0.037	0.034
(8)	0.133	0.140	0.123	0.111	0.087	0.091	0.096	0.089	0.084	0.085	0.076
(9)	0.056	0.056	0.049	0.048	0.042	0.046	0.044	0.040	0.038	0.037	0.034
(10)	0.130	0.138	0.121	0.111	0.087	0.091	0.095	0.088	0.082	0.085	0.075

4.3 Missing characteristics

One problem with our data set is that some of the characteristics are missing for some houses. For example, the number of bedrooms may be missing. One way of dealing with this problem is to restrict the comparison to houses with full sets of characteristics. However, this may cause sample selection bias, particularly since missing characteristics occur more frequently for cheaper houses in the earlier part of our data set. An alternative approach is to estimate eight versions of our hedonic model, each containing a different mix of characteristics. The price for a particular house is then imputed from whichever model has exactly the same mix of characteristics. For example, the price of a house missing the number of bedrooms is imputed from HM3.

(HM1): *ln price = f(quarter dummy, land area, num bedrooms, num bathrooms, location)*

(HM2): *ln price = f(quarter dummy, num bedrooms, num bathrooms, location)*

(HM3): *ln price = f(quarter dummy, land area, num bathrooms, location)*

(HM4): *ln price = f(quarter dummy, land area, num bedrooms, location)*

(HM5): *ln price = f(quarter dummy, num bathrooms, location)*

(HM6): *ln price = f(quarter dummy, num bedrooms, location)*

(HM7): *ln price = f(quarter dummy, land area, location)*

(HM8): *ln price = f(quarter dummy, location)*

4.4 Using repeat-sales as a benchmark

Our ultimate objective here is the construction of price indexes. In this sense, what matters most is the quality of our estimated price relatives $p_{t+1,h}/p_{t,h}$, since they are the building blocks from which our price indexes are computed. While in general we do not observe both $p_{t,h}$ and $p_{t+1,h}$, we do have some repeat-sales observations in our data set that can be used as a benchmark.

Let $p_{t+k,h}/p_{th}$ denote a repeat-sale price relative for house h. A corresponding imputed price relative for these repeat-sales dwellings can be calculated as follows:

$$\text{Double imputation}: \frac{\hat{p}_{t+k,h}}{\hat{p}_{th}},$$

$$\text{Single imputation}: \sqrt{\frac{p_{t+k,h}}{\hat{p}_{th}} \times \frac{\hat{p}_{t+k,h}}{p_{th}}},$$

where p_{th} denotes an actual price and $\hat{p}_{th}$ an imputed price.

Now define Z_h as the ratio of the actual to imputed price relative for dwelling h:

$$Z_h^{DI} = \frac{p_{t+k,h}}{p_{th}} \bigg/ \frac{\hat{p}_{t+k,h}}{\hat{p}_{th}}, \tag{11}$$

$$Z_h^{SI} = \frac{p_{t+k,h}}{p_{th}} \bigg/ \sqrt{\frac{p_{t+k,h}}{\hat{p}_{th}} \times \frac{\hat{p}_{t+k,h}}{p_{th}}} = \sqrt{\frac{p_{t+k,h}}{p_{th}} \bigg/ \frac{\hat{p}_{t+k,h}}{\hat{p}_{th}}} = \sqrt{Z_h^{DI}}. \tag{12}$$

We can now calculate the sum of squared errors of the price relatives of each hedonic method:

$$D^{DI} = \left(\frac{1}{H}\right) \sum_{h=1}^{H} [\ln(Z_h^{DI})]^2,$$

$$D^{SI} = \left(\frac{1}{H}\right) \sum_{h=1}^{H} [\ln(Z_h^{SI})]^2.$$

It follows from (12) that $D^{SI} = D^{DI}/4$. Hence D^{SI} and D^{DI} will generate identical rankings of methods. Here we focus on D^{SI}. We prefer whichever model has the smaller value of D^{SI}.

Given that we use repeat-sales as a benchmark for our imputed price relatives, our intention is to exclude repeat sales where the dwelling was renovated between sales. We attempt to identify such dwellings in two ways. First, we exclude repeat sales where one or more of the characteristics have changed between sales (for example a bathroom has been added). Second, we exclude repeat sales that occur within six months on the grounds that this suggests that the first purchase was by a professional renovator.[9] Finally, for dwellings 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 months apart). This ensures that all repeat-sales dwellings exert equal influence on our results.

Initially we started with 27 852 repeat-sales dwellings. As a result of our deletions this reduced our sample to 18 224 dwellings (or 7.6 percent of the dwellings in our complete data set).

Our results are shown in Table 6. Again we find that the hedonic models including geospatial splines (7) and (9) outperform their postcode-based competitors (8) and (10). Interestingly the best performing model according to D^{SI} is the semilog model with geospatial spline in (9). This may be because the lower flexibility of the semilog model acts to stabilize the imputed prices from one year to the next, and that this gain in stability outweighs the loss of flexibility. Given our overall objective of estimating the price-relatives as accurately as possible, a simple semilog model with a geospatial spline seems worthy of serious consideration.

Table 6: Sum of squared log price relative errors for models (7)-(10)

Model	D^{SI}
(7)	0.017467
(8)	0.020900
(9)	0.016927
(10)	0.036040

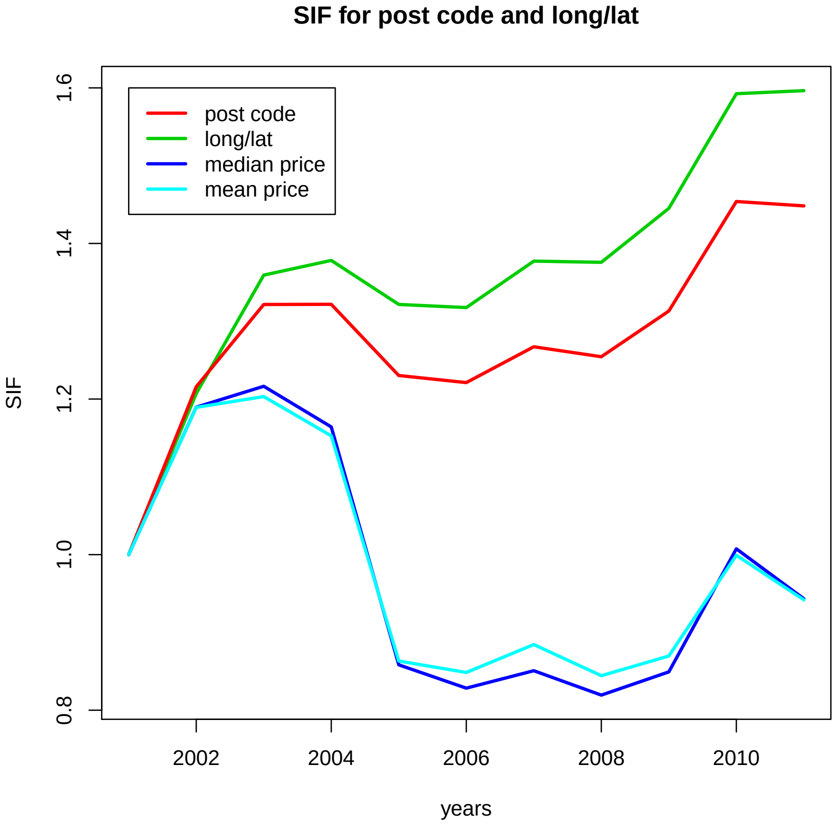
[9] Exclusion of repeat-sales within six months is standard practice in repeat-sales price indexes such as the Standard and Poor's/Case-Shiller (SPCS) Home Price Index.

4.5 House price indexes

Here we focus on the single-imputation Fisher price index formula in (4). The results obtained using the double-imputation Fisher price index formula in (5) are almost indistinguishable.

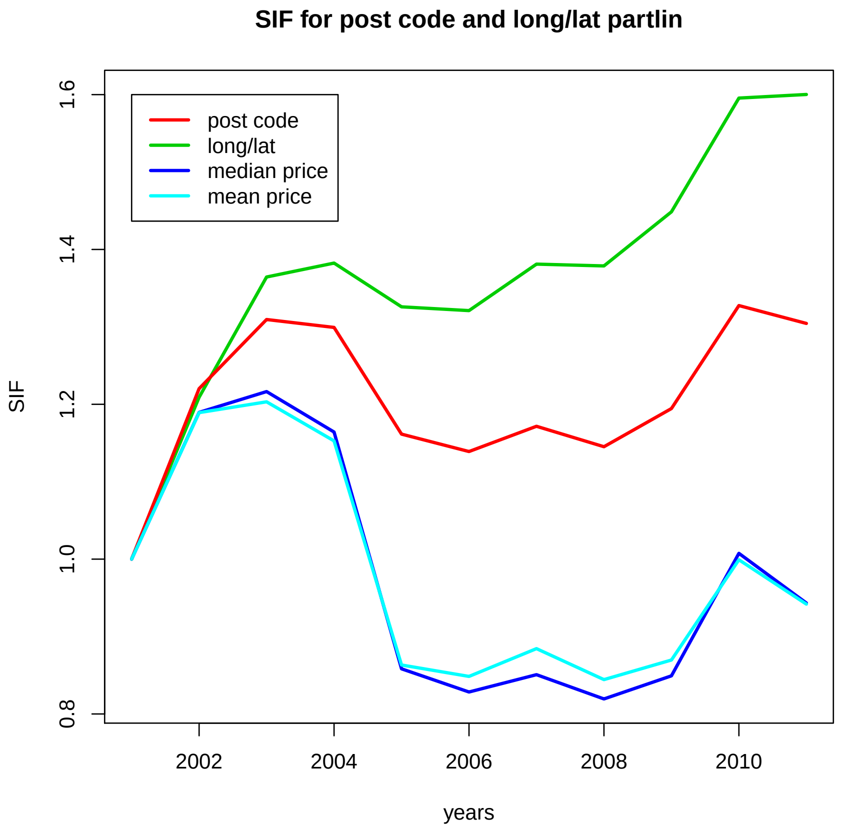
The results for our restricted data set with no missing characteristics are shown in Figures 1 and 2. The price indexes obtained from the GAM with geospatial spline in (7) and with postcodes in (8) are graphed in Figure 1. Similarly, the price indexes obtained from the semilog model with geospatial spline in (9) and with postcodes in (10) are graphed in Figure 2. Also shown in Figures 1 and 2 are simple median and mean indexes. In all cases, the price index is normalized to 1 in 2001. The index value for all other years measures the cumulative price change since 2001. Corresponding price indexes obtained using the full data set usign models HM1-HM8 are shown in Figures 3 and 4.

Figure 1: GAM on restricted data set



The results in Figures 1-4 are striking in two respects. First, the price indexes derived

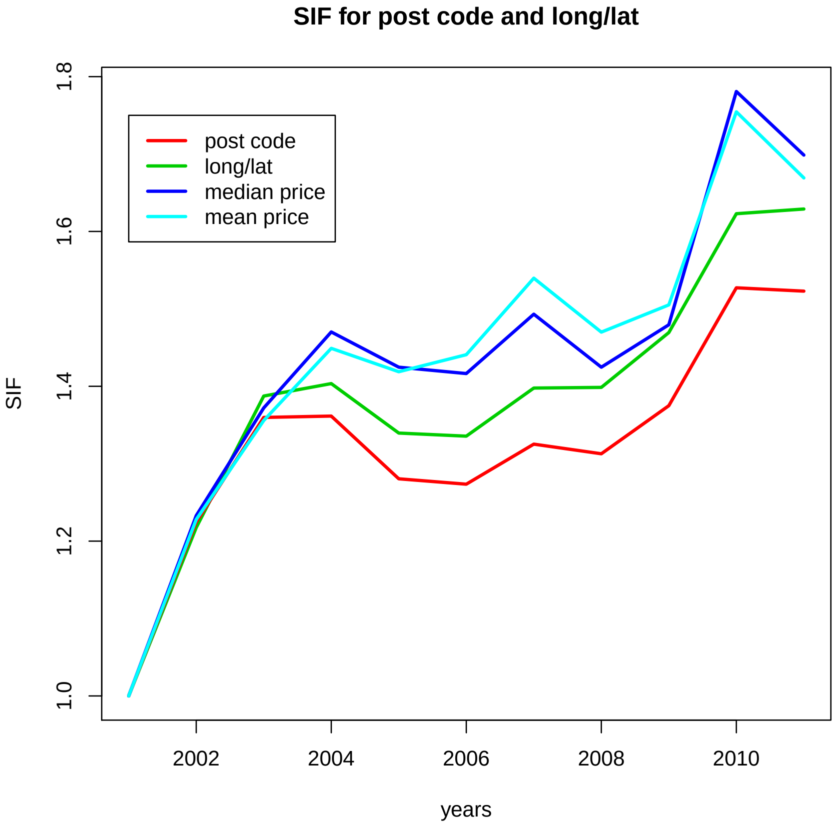
Figure 2: Semilog on restricted data set



using geospatial splines (green lines) in all four Figures rise faster than their postcode based counterparts (red lines). The gap ranges from about 8 percent in Figures 1 and 3 to 20 percent in Figure 4 and 30 percent in Figure 2 over the 2001-2011 period. One explanation for this finding is that the average locational quality of the houses sold within a postcode is getting worse over time. Our geospatial spline based indexes correct for this quality shift while the postcode based indexes do not. We are still in the process of investigating whether we can find clear evidence of such a shift towards worse locations within postcodes in our data set.

Second, the geospatial spline based price indexes are almost identical in all four Figures. This again suggests that there is not much to be gained from using a GAM in preference to the semilog model when a geospatial spline is included. Furthermore, it also indicates that sample selection bias is not really a problem when the hedonic model includes a geospatial spline. Omitting houses with missing characteristics, however, seems to cause a slight downward bias in the postcode based price indexes. This

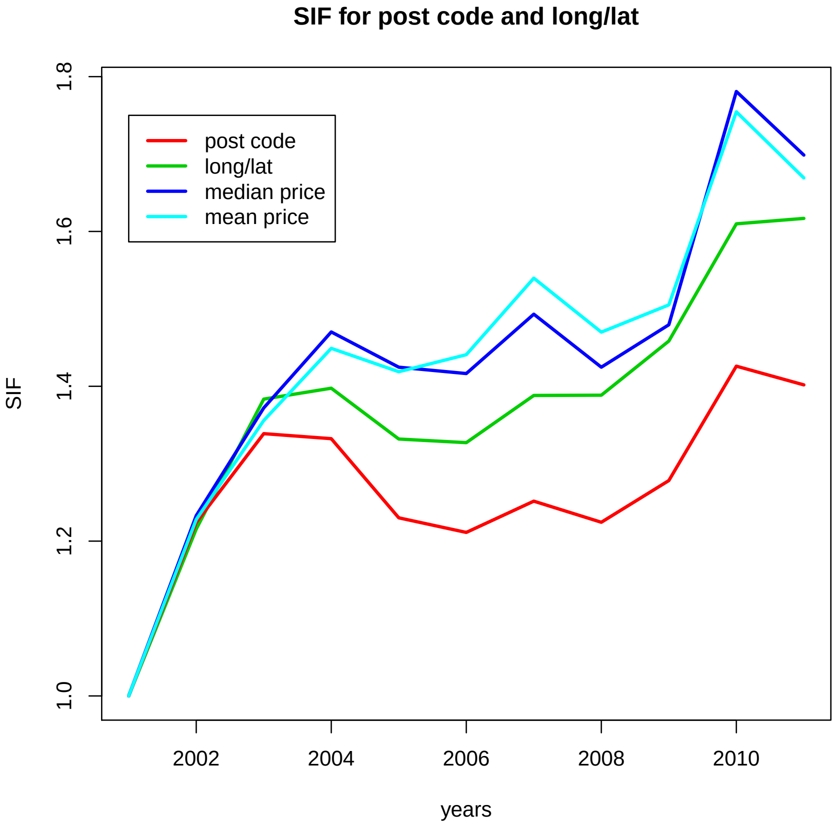
Figure 3: GAM on full data set



downward bias is far more pronounced in the median and mean indexes. Based on the restricted data set where there are no missing characteristics (see Figures 1 and 2), the mean and median indexes house price indexes are lower in 2011 than in 2001. By contrast, over the full data set (see Figures 3 and 4), the mean and median indexes house price indexes in 2011 are about 70 percent higher than in 2001. The explanation for this result is that as was noted earlier, the houses with missing characteristics tend to be cheap and are concentrated predominantly in the early part of our data set.

There remains the question of why the average quality of houses sold within postcodes deteriorated over our sample period. One possible explanation is that this is a general phenomenon that is observed in "hot" housing markets. The Sydney market experienced a long boom that started in about 1993. As can be seen from the Figures (green lines), although there was a slight correction in 2004-6, the boom is still going in 2011. In a normal or falling market, better located dwellings may sell relatively more frequently than their worse located counterparts, as compared with in a hot market. In a hot

Figure 4: Semilog on full data set



market “beggars (i.e., buyers) can’t be choosers”.

5 Conclusion

The increasing availability of geospatial data has the potential to significantly improve the quality of house price indexes. Thus far, however, no consensus has emerged in the literature as to how geospatial data can best be used. We have shown here how geospatial data can be incorporated into house price indexes by combining a hedonic model that includes a nonparametric geospatial spline function with the hedonic imputation method. The use of a spline allows locational effects to be modelled more flexibly than in a fully parametric model such as a spatial autoregressive model. Applying this approach to data for Sydney, Australia we find that using geospatial splines instead of postcode dummy variables revises upwards the cumulative price change from 2001 to 2011 by between 8 and 30 percent, depending on how postcodes are included in the competing model and on the data sample used. This difference can be attributed to a failure of postcode dummies to fully capture changes over time in the locational quality of houses sold. We also find that price indexes generated using geospatial splines are robust to the functional form of the hedonic model and are relatively immune to sample selection bias. It is desirable therefore that index providers start using geospatial data in their house price indexes.

References

Anselin L. (1988), *Spatial Econometrics: Methods and Models.* Dordrecht: Kluwar Academic Publishers.

Bao H. X. H. and A. T. K. Wan (2004), “On the Use of Spline Smoothing in Estimating Hedonic Housing Price Models: Empirical Evidence Using Hong Kong Data,” *Real Estate Economics* 32(3), 487507.

Bell K. P. and N. E. Bockstael (2000), “Applying the Generalized Moments Estimation Approach to Spatial Problems Involving Microlevel Data,” *Review of Economics and Statistics* 82, 72-82.

Berndt E. R., Z. Griliches, and N. J. Rappaport (1995), Econometric Estimates of Price Indexes for Personal Computers in the 1990s, *Journal of Econometrics* 68, 243268.

Brunauer W. A., S. Lang, P. Wechselberger and S. Bienert (2010), “Additive Hedonic Regression Models with Spatial Scaling Factors: An Application for Rents in Vienna,” *Journal of Real Estate Finance and Economics* 41, 390-411.

Can A. and I. Megbolugbe (1997), “Spatial Dependence and House Price Index Construction,” *Journal of Real Estate Finance and Economics* 14, 203222.

Clapp J. M. (2003), “A Semiparametric Method for Valuing Residential Locations: Application to Automated Valuation,” *Journal of Real Estate Finance and Economics* 27(3), 303-320.

Clapp J. M. (2004), “A Semiparametric Method for Estimating Local House Price Indices,” *Real Estate Economics* 32(1), 127-160

Clapp J. M., H. J. Kim and A. E. Gelfand (2002), “Predicting Spatial Patterns of House Prices Using LPR and Bayesian Smoothing,” *Real Estate Economics* 30(4), 505-532.

Cliff A. and J. K. Ord (1973), *Spatial Autocorrelation.* London: Pion Publishing.

Colwell P. F. (1998), "A Primer on Piecewise Parabolic Multiple Regression Analysis via Estimations of Chicago CBD Land Prices," *Journal of Real Estate Finance and Economics* 17(1), 87-97.

Corrado L. and B. Fingleton (2011), "Where Is the Economics in Spatial Econometrics?" *Journal of Regional Science*, forthcoming.

de Haan J. (2004), "Direct and Indirect Time Dummy Approaches to Hedonic Price Measurement," *Journal of Economic and Social Measurement* 29(4), 427-443.

Diewert W. E. (2001), "Hedonic Regressions: A Consumer Theory Approach," Discussion Paper 01-12, Department of Economics, University of British Columbia.

Diewert W. E. (2003), "Hedonic Regressions: A Review of Some Unresolved Issues," Mimeo.

Dorsey R. E., H. Hu, W. J. Mayer and H. C. Wang (2010), "Hedonic Versus Repeat-Sales Housing Price Indexes for Measuring the Recent Boom-Bust Cycle," *Journal of Housing Economics* 19, 75-93.

Dulberger E. R. (1989), "The Application of a Hedonic Model to a Quality-Adjusted Price Index for Computer Processors," in D. W. Jorgenson and R. Landau (eds.), *Technology and Capital Formation.* Cambridge, MA: MIT Press, 37-75.

Fik T. J., D. C. Ling and G. F. Mulligan (2003), "Modeling Spatial Variation in Housing Prices: A Variable Interaction Approach," *Real Estate Economics* 31(4), 623-646.

Fleming M. C. and J. G. Nellis (1985), "The Application of Hedonic Indexing Methods: A Study of House Prices in the United Kingdom," *Statistical Journal of the United Nations Economic Commission for Europe* 3, 249-270.

Gouriéroux C. and A. Laferrère (2009), "Managing Hedonic Housing Price Indexes: The French Experience," *Journal of Housing Economics*, 206-213.

Hardman M. (2011), "Calculating High Frequency Australian Residential Property Price Indices," Rismark Technical Paper, Rismark International.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004), *Nonparametric and Semiparametric Models*, Springer.

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

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

Hill R. J., D. Melser and I. Syed (2009), "Measuring a Boom and Bust: The Sydney Housing Market 2001-2006," *Journal of Housing Economics* 18(3), 193-205.

Kelejian H. H. and I. R. Prucha (1998), "A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances," *Journal of Real Estate Finance and Economics* 17, 99-121.

Kelejian H. H. and I. R. Prucha (2010), "Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances," *Journal of Econometrics* 157, 53-67.

Kim C. W., T. T. Phipps, and L. Anselin (2003), "Measuring the Benefits of Air Quality Improvement: A Spatial Hedonic Approach," *Journal of Environmental Economics and Management* 45, 2439.

Lee L. F. (2003), "Best Spatial Two-Stage Least Squares Estimators for a Spatial Autoregressive Model with Autoregressive Disturbances," *Econometric Reviews* 22, 307335.

Lee L. F. (2007), "GMM and 2SLS Estimation of Mixed Regressive, Spatial Autoregressive Models," *Journal of Econometrics* 137, 489514.

LeSage J. P. and R. K. Pace (2009), *Introduction to Spatial Econometrics*. New York: CRC Press.

Liu X., L. F. Lee and C. R. Bollinger (2010), An Efficient GMM Estimator of Spatial Autoregressive Models," *Journal of Econometrics* 159, 303-319.

Malpezzi S. (2003), "Hedonic Pricing Models: A Selective and Applied Review," in A. O'Sullivan and K. Gibb (eds.) *Housing Economics: Essays in Honor of Duncan Maclennan*, 67-89. Blackwell: Malder MA.

McMillen D. and C. L. Redfearn (2010), "Estimation and Hypothesis Testing for Nonparametric Hedonic House Price Functions," *Journal of Regional Science* 50(3), 712-733.

Nappi-Choulet I. and T. Maury (2009), "A Spatiotemporal Autoregressive Price Index for the Paris Office Property Market," *Real Estate Economics* 37(2), 305-340.

Ord J. K. (1975), "Estimation Methods for Models of Spatial Interaction," *Journal of the American Statistical Association* 70, 120-126.

Pace R. K. and R. Barry (1997), "Quick Computation of Spatial Autoregressive Estimators," *Geographical Analysis* 29, 232246.

Pace R. K., R. Barry, J. M. Clapp and M. Rodriguez (1998), "Spatiotemporal Autoregressive Models of Neighborhood Effects," *Journal of Real Estate Finance and Economics* 17(1), 15-33.

Pace R. K. and O. Gilley (1997), "Using Spatial Configuration of the Data to Improve Estimation," *Journal of Real Estate Finance and Economics* 14(3), 333-340.

Pakes A. (2003), "A Reconsideration of Hedonic Price Indices with an Application to PC's," *American Economic Review* 93(5), 1578-1596.

Pavlov A. D. (2000), "Space-Varying Regression Coefficients: A Semi-Parametric Approach Applied to Real Estate Markets," *Real Estate Economics* 28(2), 249-283.

Pinkse J. and M. E. Slade (2010), "The Future of Spatial Econometrics," *Journal of Regional Science* 50(1), 103117.

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

Ribe M. (2009), *House Prices in a Swedish CPI Perspective*, Statistics Sweden. Paper Presented at the 11th Ottawa Group Meeting in Neuchâtel, 27-29 May, 2009.

Saarnio M. (2006), "Housing Price Statistics at Statistics Finland," Paper presented at the OECD-IMF Workshop on Real Estate Price Indexes, Paris, 6-7 November 2006.

Silver M. and S. Heravi (2001), "Quality Adjustment, Sample Rotation and CPI Practice: An Experiment," Presented at the Sixth Meeting of the International Working Group on Price Indices, Canberra, Australia, April 2-6.

Stone, C. J. (1986), The Dimensionality Reduction Principle for Generalized Additive Models, *Annals of Statistics* 14(2), 590-606.

Sun H., Y. Tu and S. Yu (2005), "A Spatio-Temporal Autoregressive Model for Multi-Unit Residential Market Analysis," *Journal of Real Estate Finance and Economics* 31(2), 155187.

Thomassen A. (2007), *Price Index for New Multidwelling Houses: Sources and Methods*, Statistics Norway/Department of Industry Statistics/Construction and Service Statistics, Document 2007/9.

Tu Y., S. Yu and H. Sun (2004), “Transaction-Based Office Price Indexes: A Spatiotemporal Modeling Approach,” *Real Estate Economics* 32(2), 297328.

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