the review of income and wealth

Review of Income and Wealth
Series 64, Number. 4, December 2018
DOI: 10.1111/roiw.12303

CAN GEOSPATIAL DATA IMPROVE HOUSE PRICE INDEXES? A HEDONIC IMPUTATION APPROACH WITH SPLINES

by Robert J. Hill* and Michael Scholz

Department of Economics, University of Graz

Determining how and when to use geospatial data (i.e. longitudes and latitudes for each house) is probably the most pressing open question in the house price index literature. This issue is particularly timely for national statistical institutes (NSIs) in the European Union, which are now required by Eurostat to produce official house price indexes. Our solution combines the hedonic imputation method with a flexible hedonic model that captures geospatial data using a non-parametric spline surface. For Sydney, Australia, we find that the extra precision provided by geospatial data as compared with postcode dummies has only a marginal impact on the resulting hedonic price index. This is good news for resource-stretched NSIs. At least for Sydney, postcodes seem to be sufficient to control for locational effects in a hedonic house price index.

JEL Codes: C43, E31, R31

Keywords: housing market, house price measurement, geospatial spline surface, quality adjustment, semiparametric hedonic model

1. Introduction

The importance of housing to the broader economy has been demonstrated by the global financial crisis of 2007–11, which began in the U.S. housing market. It is essential therefore that governments, central banks, and market participants are kept well informed of trends in house prices.

House price indexes, however, can be highly sensitive to the method of construction, and this sensitivity can be a source of confusion amongst users (see Silver, 2011). The problem is that every house is different both in terms of its physical characteristics and its location. House price indexes need to take account of these quality differences. Otherwise, the price index will confound price changes and quality differences. The importance of these measurement problems

Note: This project has benefited from funding from the Austrian National Bank (Jubiläumsfondsprojekt 14947). We thank Australian Property Monitors for supplying the data. We also thank participants at the following conferences for their comments: the ICP/PPP workshop at Princeton University (May 2013), the Ottawa Group meeting in Copenhagen (May 2013), the UNSW Economic Measurement Group (EMG) workshop in Sydney (November 2013), the University of Queensland workshop on Housing Markets and Residential Property Price Indexes in Brisbane (December 2013), the OECD workshop on House Price Statistics in Paris (March 2014), the Society for Economic Measurement conference at University of Chicago (August 2014), the Austrian National Bank workshop "Are House Prices Endangering Financial Stability? If So, How Can We Counteract This?" in Vienna (October 2014), and the 23rd European Real Estate Society Annual Conference in Regensburg (June 2016).

*Correspondence to: Robert J. Hill, Department of Economics, University of Graz, Universitätsstrasse 15/F4, 8010 Graz, Austria (robert.hill@uni-graz.at).

has been recently recognized by the international community. The European Commission, Eurostat, the United Nations, the International Labour Organization (ILO), the Organisation for Economic Co-operation and Development (OECD), the World Bank, and the International Monetary Fund (IMF) together commissioned a *Handbook on Residential Property Price Indexes* that was completed in 2013 (see European Commission *et al*., 2013).

Hedonic methods—which express house prices as a function of a vector of characteristics—are ideally suited for constructing quality-adjusted house price indexes (see Hill, 2013).[1] Most hedonic indexes at present adjust for locational effects using region (e.g. postcode) dummy variables. A key recent development is the increased availability of geospatial data (i.e. exact longitudes and latitudes for each house), which may allow locational effects to be captured in a more precise way.

Determining how and when to use geospatial data is probably the most pressing open question in the house price index literature. This issue is particularly timely for national statistical institutes (NSIs) in the European Union, given that Eurostat now requires that all Member States produce official house price indexes (see Eurostat, 2015).[2]

In this paper, we make four main contributions to the house price index literature. First, we develop a method for constructing hedonic house price indexes that incorporates geospatial data. We estimate a hedonic model for each period that includes geospatial data as a non-parametric spline surface, and then impute a price for each house from the hedonic model. These imputed prices from adjacent periods are then inserted into a Törnqvist-type price index formula. Finally, these Törnqvist-type indexes are chained to obtain the overall house price index.

Second, we consider the problem of how to compare the performance of alternative versions of the hedonic imputation method. Rather than focusing on the fit of the hedonic model itself (as is standard in the literature), we focus on the imputed price relatives that are used to compute the price index. The quality of the imputed price indexes can be evaluated using repeat-sales price relatives as a benchmark. We propose new performance metrics based on this approach.

Third, using a dataset consisting of 454 507 actual housing transactions in Sydney, Australia over the period 2001–11, we show that the extra precision in the imputed price relatives as measured by our performance metrics provided by geospatial data as compared with postcode dummies has only a marginal impact on the resulting index. This is good news for resource-stretched statistical institutes (NSIs) and other index providers. At least in the case of Sydney, postcode dummies are for most purposes sufficient for constructing quality-adjusted house price indexes.

[1]Hedonic house price indexes should not be confused with automated valuation models (AVMs). The latter aim to impute prices for individual houses. A price index measures changes in house prices over time. Also, the unit of comparison for a house price index is typically a city or country rather than an individual house.

[2]Geospatial data are not the only new type of data presenting a challenge to NSIs and other price index providers. For example, web-scraping and scanner data have the potential to hugely increase the number of price quotes used in the consumer price index (CPI) (see Cavallo, 2013; de Haan and Krsinich, 2014).

Fourth, we show that this result relies on the postcodes being quite narrowly defined. When we use broader Residex regions rather than postcodes as locational dummy variables, the difference with our preferred index is no longer marginal. Furthermore, we find evidence of a downward bias in the index when Residex-region dummies are used. This bias can be attributed to a systematic decline over time within Residex regions in the locational quality of houses sold. This trend may itself be a natural consequence of the long housing boom in Sydney, which started in 1993 and has continued ever since (except for a flat period between 2004 and 2008). An analogous bias can even be discerned in an index that uses postcode dummies, although its magnitude there is much smaller. To fully eliminate such biases, geospatial data are required.

One complication with our dataset is that one or more of the characteristics are missing for many of the housing transactions. Following Hill and Syed (2016), we deal with this problem by estimating multiple versions of our hedonic model. Each version contains a different mix of characteristics. The price of each dwelling is then imputed from whichever hedonic model includes exactly the same mix of characteristics as are available for that particular dwelling.

With regard to our hedonic model, a spline is one of a number of alternative techniques that could be used to fit a geospatial non-parametric surface. Other possibilities include partial linear models, locally weighted regression, and kriging. The reason we chose splines is that efficient and robust fitting methods for generalized additive models (even for large datasets) are available in off-the-shelf statistical software packages, such as R. Splines have also been extensively used in diverse scientific fields (e.g. biology, medicine, and environmental sciences) for many years, where they are state of the art (see, e.g., Wood, 2011). Hedonic models that include geospatial data non-parametrically (although not using splines) have been estimated previously by, amongst others, Colwell (1998), Fik *et al.* (2003), and Clapp (2004). However, while these authors include geospatial data in a hedonic model, they do not consider the problem of how to include geospatial data in a house price index, or how replacing postcodes with geospatial data affects a house price index.

The remainder of this paper is structured as follows. Section 2 provides an overview of the hedonic price index literature, and discusses ways of incorporating location into a hedonic house price index. Section 3 presents our dataset and hedonic models, compares the performance of these models, derives the resulting hedonic price indexes, and explores the apparent downward bias in the postcode- and Residex-region–based indexes. Section 4 concludes by considering some implications of our findings. More information on our dataset and details regarding the estimation of the geospatial spline function in R using methods developed by Wood (2006, 2011) are provided in the Appendix (in the Online Supporting Information).

2. Hedonic Price Indexes for Housing

2.1. *An Overview*

A hedonic model regresses the price of a product on a vector of characteristics (the prices of which are not independently observed). The hedonic equation is

a reduced form that is determined by the interaction of supply and demand. Hedonic models are used to construct quality-adjusted price indexes in markets (such as computers) where the products available differ significantly from one period to the next. Housing is an extreme case in that every house is different.

One can distinguish between a house's physical and locational attributes. Examples of the former include the number of bedrooms and the 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.[3]

2.2. *The Hedonic Imputation Method*

Here, we focus on the hedonic imputation method. Other ways of computing hedonic price indexes, such as the time-dummy method, adjacent period, and the average characteristics method, are discussed in Diewert (2011) and Hill (2013). The hedonic imputation approach estimates a separate hedonic model for each period or a few adjacent periods:[4]

$$y_t = Z_t \beta_t + \varepsilon_t. \tag{1}$$

The hedonic model is then used to impute prices for individual houses. For example, let $\hat{p}_{t+1,h}(z_{t,h})$ denote the imputed price in period $t+1$ of a house sold in period t. This price is imputed by substituting the characteristics of house h sold in period t, $z_{t,h}$, into the estimated hedonic model of period $t+1$ as follows:[5]

$$\hat{p}_{t+1,h}(z_{t,h}) = \exp\left(\sum_{c=1}^{C} \hat{\beta}_{c,t+1} z_{c,t,h}\right). \tag{2}$$

These imputed prices can then be inserted into standard price index formulas as follows:

$$\text{Paasche-type imputation}: P^{PI}_{t,t+1} = \prod_{h=1}^{H_{t+1}} \left[\left(\frac{p_{t+1,h}}{\hat{p}_{t,h}(z_{t+1,h})}\right)^{1/H_{t+1}}\right], \tag{3}$$

$$\text{Laspeyres-type imputation}: P^{LI}_{t,t+1} = \prod_{h=1}^{H_t} \left[\left(\frac{\hat{p}_{t+1,h}(z_{t,h})}{p_{t,h}}\right)^{1/H_t}\right], \tag{4}$$

[3]Omitted variables are a potentially serious problem in hedonic models of the housing market. The problem is worse, though, if the objective is to construct an AVM, than if the objective is to construct a hedonic price index. This is because with the latter the effects of omitted variables tend to partially offset each other (see Hill, 2013).

[4]The appropriate time horizon for each model depends partly on the size of the dataset. For example, for our Sydney data, there are enough observations to estimate the model separately for each year.

[5]For a discussion of some of the advantages of the hedonic imputation method, see Silver and Heravi (2007), Diewert *et al.* (2009), and Rambaldi and Rao (2013).

$$\text{(5)} \qquad \text{Törnqvist-type imputation}: P^{TI}_{t,t+1} = \sqrt{P^{PI}_{t,t+1} \times P^{LI}_{t,t+1}}.$$

In a comparison between periods t and $t+1$, the Laspeyres-type index focuses on houses that sold in the earlier period t, while the Paasche-type index focuses on houses that sold in the later period $t+1$. These price indexes give equal weight to each house sold.[6] By taking the geometric mean of Paasche and Laspeyres, the Törnqvist-type index gives equal weight to both periods. The Paasche-, Laspeyres-, and Törnqvist-type indexes above are all of the single-imputation variety, meaning that only one of the prices in each price relative is imputed. A double-imputation approach, by contrast, imputes both prices. There has been some discussion in the literature over the relative merits of the two approaches (see, e.g. Hill and Melser, 2008). Empirically, we try both approaches. The resulting price indexes are virtually indistinguishable. Hence to simplify the presentation, we focus here only on single-imputation price indexes. The hedonic imputation method is flexible in that it allows the characteristic shadow prices to evolve over time. The hedonic imputation method is used for example by the FNC Residential Price Index in the United States (see Dorsey *et al*., 2010), some indexes produced by RPData-Rismark in Australia (see Hardman, 2011), and the Bank Austria/Austrian National Bank Residential Property Price Index in Austria (see Brunauer *et al*., 2012).

2.3. *Methods for Incorporating Location into House Price Indexes*

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 house in the hedonic model. However, given the increasing availability of geospatial data, it should be possible to adjust more precisely for locational effects.

Distances to Amenities

Given the availability of geospatial data, the distance of each house 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, adjacent period, or imputation versions of the hedonic model (see, e.g. Hill and Melser, 2008; Rambaldi and Fletcher, 2014).[7]

The hedonic model in this case takes the following form:

[6]This democratic weighting structure is, in our opinion, more appropriate in a housing context than weighting each house by its expenditure share. For a discussion on alternative weighting schemes, see de Haan (2010).

[7]In some cases, a more informative alternative to distance may be travelling time (see, e.g. Shimizu, 2014).

$$y = Z\beta + \sum_{k=1}^{K} D_k(z_{\text{lat}}, z_{\text{long}})\delta_k + \varepsilon, \tag{6}$$

where z_{lat} and z_{long} denote the longitude and latitude of each dwelling, and D_k is the distance to amenity k. Again, $y = \ln p$, and to simply the notation the time subscript t is suppressed.

The use of distances to amenities as characteristics can be problematic in hedonic models for a few reasons. First, and most importantly, 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) often 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, and so on. This last problem is potentially the easiest to solve, by using quadratics, cubics, splines, and so on to model the impact of distance.

Spatial-Autoregressive Models

Locational effects can be captured more effectively by a spatial autoregressive model. A first-order spatial autoregressive model with autoregressive errors takes the following form (see, e.g. Corrado and Fingleton, 2012):

$$\begin{aligned} y &= \rho Sy + Z\beta + u, \\ u &= \lambda Su + \varepsilon, \end{aligned} \tag{7}$$

where y is the vector of log prices, (i.e. each element $y_h = \ln p_h$), Z is the matrix of characteristics, S is a spatial weights matrix that is calculated from the geospatial data, and ρ and λ are scalars that are estimated simultaneously with the β vector of characteristic shadow prices.

Price indexes can be obtained from a spatial autoregressive hedonic model by simply including quarter or year dummies in the Z characteristics matrix, and then by exponentiating the estimated parameters on these dummy variables. One problem with this approach 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. That is, the magnitude of the dependence between observations depends inversely on both their spatial and temporal separation.

The replacement of a spatial weights matrix with a spatiotemporal weights matrix significantly increases the computational burden and complicates the derivation of price indexes (see, e.g. Nappi-Choulet and Maury, 2009). One response to this problem is to use the adjacent-period method. In this case, the temporal separation between observations never exceeds one period 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 *et al*. (2009). Dorsey *et al*. (2010), Rambaldi and Rao (2013), and Rambaldi and Fletcher

(2014) combine a rolling-period spatial autoregressive model with the hedonic imputations method.

The main problem with spatial autoregressive models is that they impose a lot of prior structure on the spatial dependence. This may explain why Rambaldi and Fletcher (2014) find that including distances to 11 amenities (e.g. parks, schools, and shops) as a way of controlling for location effects outperforms a spatial error model (a simplified version of equation (7) where $\rho = 0$).

Semiparametric Approaches

Semiparametric methods provide a different and potentially more flexible alternative to spatial autoregressive models for modeling spatial dependence. Non-parametric methods can be used to construct a 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 non-parametric hedonic model defined over the physical characteristics. For example, a semilog model for period t defined on the physical characteristics Z could be combined with a non-parametric function $g(\cdot)$ defined on the geospatial data $z_{\text{lat}}, z_{\text{long}}$ as follows:

$$y = Z\beta + g(z_{\text{lat}}, z_{\text{long}}) + \varepsilon. \tag{8}$$

Comparing equation (8) with equation (6) it can be seen that this non-parametric approach can be viewed as an extension of the distance-to-amenities method.

Imputed prices for each house can be obtained by inserting its particular mix of characteristics (including the longitude and latitude) into the estimated hedonic model of equation (8). More specifically, consider the Törnqvist price index in equation (5). Imputed prices in period t of houses actually sold in period $t + 1$, denoted by $\hat{p}_{t,h}(z_{t+1,h})$ (where $z_{t+1,h}$ here consists of both the physical and geospatial characteristics), can be derived from the hedonic model of period t. That is, one can take the physical characteristics and longitude/latitude of house h sold in period $t + 1$ and insert them into the hedonic model of period t in equation (8) to obtain an imputed price of this same house h in period t. 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 $t + 1$. 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 $t + 1$.

Spline components have been included in semiparametric hedonic models by Bao and Wan (2004) and Diewert and Shimizu (2015). However, our approach differs from theirs in two important respects. First, their splines are not defined on longitudes and latitudes. Bao and Wan (2004) estimate a three-dimensional spline defined over floor space, garage space, and age of the dwelling, while Diewert and Shimizu (2015) estimate one-dimensional splines defined on land area and age, respectively. Second, we combine our semiparametric model with the hedonic imputation method to compute price indexes. By contrast, Bao and Wan do not compute price indexes, while Diewert and Shimizu use a different price index methodology that attempts to separate the prices of land and structures.

Geospatial data, however, have been included in hedonic models previously using various non-parametric methods by, amongst others, Colwell (1998), Fik *et al*. (2003), Clapp (2004), Hardman (2011), Knight (2015), and Schäfer and Hirsch (2016). Of these, only Schäfer and Hirsch (2016) use splines. With the exception of Hardman (2011), though, none of these authors combines a non-parametric treatment of geospatial data with the hedonic imputation method.[8]

In the next section, we illustrate our approach using data for Sydney, Australia. First, we estimate a semiparametric hedonic model that includes a geospatial spline. We then combine it with the hedonic imputation method to compute price indexes.

3. Empirical Strategy

3.1. *The Dataset*

We use a dataset obtained from Australian Property Monitors that consists of prices and characteristics of houses sold in Sydney (Australia) for the years 2001–11. For each house, we have the following characteristics: the actual sale price, time of sale, postcode, property type (i.e. detached or semi), number of bedrooms, number of bathrooms, land area, exact address, longitude, and latitude. (We exclude all townhouses from our analysis, since the corresponding land area is for the whole strata and not for the individual townhouse itself.) Some summary statistics are provided in Table A.1.

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

While we deleted bedroom, bathroom, and land area counts outside the allowed ranges, we retained the house itself in the dataset as long as the price and longitude/latitude were available and within the allowed ranges as specified in Table A.2. In total, less than 1 percent of the houses were deleted. After deletions, our dataset consisted of 454 507 house sales. Complete data on all our hedonic characteristics are available for 240 142 observations. This is what we refer to as the "restricted" dataset. Table A.3 shows the distribution of houses with missing characteristics. It can be seen from Table A.3 that the quality of the data improves over time. We explain in Section 3.3 how we deal with the missing characteristics problem so that we are then able to compute hedonic price indexes for the full dataset.

3.2. *Model Estimation*

Here, we estimate the following three models:

(i) semilog in physical characteristics with a geospatial spline;
(ii) semilog in physical characteristics with postcode dummies; and

[8]Hardman, who describes the method used to compute the RPData-Rismark's Daily Home Value Index, does not provide enough detail to allow one to determine exactly how the RPData-Rismark index is constructed.

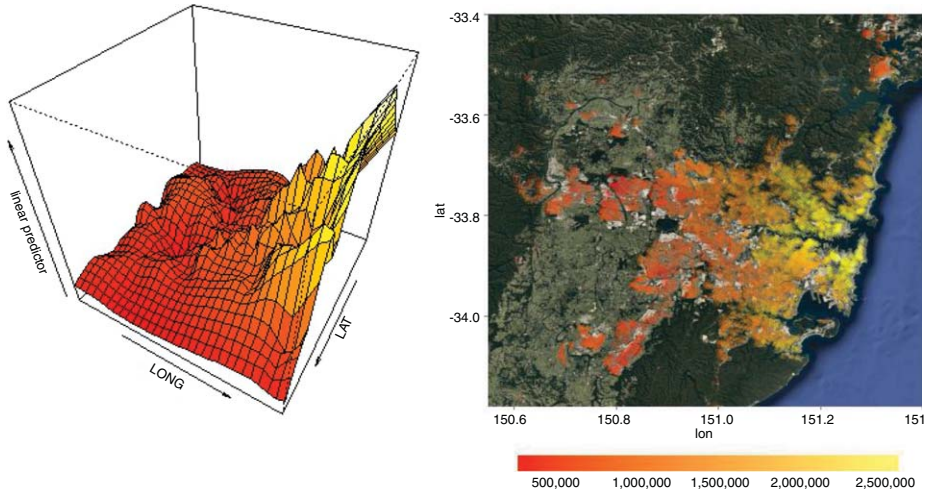

Figure 1. The Spline Surface and Contour Plot Based on the Restricted Dataset for 2007

Note: Prices on the contour map are predicted for the median house with three bedrooms, two bathrooms, and an area of 590 m^2, sold in the third quarter of 2007. The predicted prices are projected onto a map provided by Google, TerraMetrics 2016. [Colour figure can be viewed at wileyonlinelibrary.com]

(iii) semilog in physical characteristics with Residex-region dummies.

The semiparametric formulation in model (i) is more flexible than the fully parametric semilog formulations in (ii) and (iii). At the same time, model (i) avoids the curse of dimensionality problem that arises in a fully non-parametric model (see, e.g., Stone, 1986).[9] Each model is estimated separately for each year $t=2001,\ldots,2011$. Model (i) takes the following form:

$$y=Z\beta+g(z_{\text{lat}},z_{\text{long}})+\varepsilon, \tag{9}$$

where $g(\cdot)$ now denotes a spline. Again, the time subscript t is suppressed. The way in which we estimate the semiparametric model (i) in R (see R Core Team, 2013) using methods developed by Wood (2006, 2011) is explained in the Appendix. It is necessary here to estimate the characteristic shadow price vector β and the spline surface $g(z_{\text{lat}},z_{\text{long}})$. An example of one of our estimated $g(z_{\text{lat}},z_{\text{long}})$ surfaces (for 2007) and a corresponding colored contour map are provided in Figure 1. The contour map shows how the price of the median house in terms of its physical characteristics (i.e. three bedrooms, two bathrooms, and 590 m^2) varies at different locations in Sydney in 2007. It provides an intuitive way of interpreting the spline surface. The contour map accords well with the conventional wisdom (see, e.g. the "property heatmap for Sydney produced by AMP).[10]

[9]Nevertheless, it would be possible here to estimate the whole hedonic model non-parametrically. However, we do not explore this option in this paper.

[10]The spline surface, however, cannot easily capture the impact of fine details such as busy roads. Hence there is scope to further refine the hedonic model by including additional locational variables such as dummy variables to indicate that the dwelling is located next to a busy road.

Models (ii) and (iii) take the following form (with the time subscript t again suppressed):

$$y = Z\beta + L\lambda + \varepsilon. \tag{10}$$

In the case of models (ii) and (iii), it is necessary to estimate the characteristic shadow price vector β and the location dummy variable shadow price vector λ. The difference between (ii) and (iii) is that L and λ are defined over 242 postcodes for (ii) versus 16 Residex regions for (iii). Each Residex region therefore consists of about 15 postcodes.[11]

3.3. *Missing Characteristics*

The exclusion of houses with one or more missing characteristics may cause sample selection bias, particularly since missing characteristics occur more frequently for cheaper houses in the earlier part of the dataset (see Table A.3). Following Hill and Syed (2016), we deal with this problem by estimating eight versions of our hedonic model, each containing a different mix of characteristics. Here, we focus on the following three characteristics: land area, number of bedrooms, and number of bathrooms. This yields eight possible combinations of characteristics. None could be missing (HM1), one could be missing (HM2, HM3, and HM4), two could be missing (HM5, HM6, HM7), or all three could be missing (HM8). 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.[12]

(HM1): *ln price = f(quarter dummy, house type, land area, num bedrooms, num bathrooms, location)*
(HM2): *ln price = f(quarter dummy, house type, num bedrooms, num bathrooms, location)*
(HM3): *ln price = f(quarter dummy, house type, land area, num bathrooms, location)*
(HM4): *ln price = f(quarter dummy, house type, land area, num bedrooms, location)*
(HM5): *ln price = f(quarter dummy, house type, num bathrooms, location)*
(HM6): *ln price = f(quarter dummy, house type, num bedrooms, location)*
(HM7): *ln price = f(quarter dummy, house type, land area, location)*
(HM8): *ln price = f(quarter dummy, house type, location)*

[11]The Residex regions (with their constituent postcodes listed in brackets) are as follows: Inner Sydney (2000 to 2020), Eastern Suburbs (2021 to 2036), Inner West (2037 to 2059), Lower North Shore (2060 to 2069), Upper North Shore (2070 to 2087), Mosman–Cremorne (2088 to 2091), Manly–Warringah (2092 to 2109), North Western (2110 to 2126), Western Suburbs (2127 to 2145), Parramatta Hills (2146 to 2159), Fairfield–Liverpool (2160 to 2189), Canterbury–Bankstown (2190 to 2200), St George (2201 to 2223), Cronulla–Sutherland (2224 to 2249), Campbelltown (2552 to 2570), and Penrith–Windsor (2740 to 2777).

[12]While other methods exist for dealing with the problem of missing characteristics (such as multiple imputation or including "missing" dummy variables), the method used here exploits the underlying structure of the hedonic imputation method.

TABLE 1
THE AVERAGE SQUARED ERROR OF THE LOG PRICES (C_T) (FULL DATASET)

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
(i)	0.052	0.049	0.045	0.044	0.043	0.044	0.045	0.042	0.040	0.039	0.034
(ii)	0.066	0.063	0.057	0.056	0.053	0.056	0.057	0.052	0.049	0.048	0.042
(iii)	0.101	0.098	0.088	0.083	0.082	0.087	0.095	0.089	0.097	0.107	0.077

Note: Model (i) is the semiparametric model that includes a geospatial spline defined in equation (9). Models (ii) and (iii) are both semilog models with location dummies as defined in equation (10). Model (ii) uses postcode dummies while model (iii) uses Residex dummies.

In Section 3.5, we compute hedonic price indexes based on the restricted dataset with no missing characteristics (i.e. using only HM1) and on the full dataset (i.e. using all eight models HM1–HM8). Our results in Section 3.5 indicate a strong sample selection bias in the restricted dataset.

3.4. *Comparing the Performance of Our Hedonic Models*

The fit of a hedonic model can be evaluated by comparing imputed prices $\hat{p}_{th}$ with their actual counterparts p_{th}. Table 1 shows the average squared error of the log prices, C_t, defined as follows:

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

A lower value of C_t in Table 1 implies a better fit. It can be seen in Table 1 that model (i) with its geospatial spline in equation (9) outperforms its postcode-/Residex-region–based competitors (ii) and (iii) in equation (10). As expected, model (ii) with its finer postcode classification of regions likewise outperforms model (iii) with its broader Residex regions. The results in both Table 1 and Table 2 are computed using the full dataset. The results are ordinally equivalent if the comparison is made over the restricted dataset (where there are no missing characteristics).

In Table A.4, we also compute Akaike information criterion (AIC) values. The AIC values are computed only for the restricted dataset. A lower AIC also

TABLE 2
THE AVERAGE SQUARED ERROR OF THE LOG PRICE RELATIVES D AND D^{adj} (FULL DATASET)

Model	D	D^{adj}(i)	D^{adj}(ii)	D^{adj}(iii)
(i)	0.00913	0.00905	0.00905	0.00913
(ii)	0.00972	0.00962	0.00961	0.00967
(iii)	0.01302	0.01290	0.01289	0.01290

Note: Model (i) is the semiparametric model that includes a geospatial spline defined in equation (9). Models (ii) and (iii) are both semilog models with location dummies as defined in equation (10). Model (ii) uses postcode dummies, while model (iii) uses Residex dummies. The D^{adj} results include a correction for the "lemons" bias in the repeat-sales price relatives. The correction is made by comparing the change in a repeat-sales index with the corresponding change in a reference hedonic index. The D^{adj}(i) results are calculated using the hedonic price index derived from model (i) as the reference hedonic index. Similarly, D^{adj}(ii) and D^{adj}(iii) use the hedonic price indexes derived from models (ii) and (iii), respectively, as the reference.

implies a better fit (where AIC can be negative). The AIC results likewise show that model (i) is best followed by (ii) and then (iii).

Our ultimate objective here is the price indexes that are derived from the hedonic models, rather than the within-sample fit of the hedonic models. In this sense, what matters most is the quality of our estimated price relatives $p_{t+1,h}/\hat{p}_{t,h}$ and $\hat{p}_{t+1,h}/p_{t,h}$. This is because—as can be seen from the Paasche- and Laspeyres-type formulas in equations (3) and (4)—the price relatives 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 dataset that can be used as a benchmark (Reid, 2007).

Suppose that house h sells in both periods t and $t+k$. For this house, therefore, we have a repeat-sales price relative: $p_{t+k,h}/p_{t,h}$. Taking the geometric mean of the corresponding Paasche- and Laspeyres-type imputed price relatives, $p_{t+1,h}/\hat{p}_{t,h}$ and $\hat{p}_{t+1,h}/p_{t,h}$, we obtain the following

$$\text{Imputed price relative}: \quad \sqrt{\frac{p_{t+k,h}}{\hat{p}_{t,h}} \times \frac{\hat{p}_{t+k,h}}{p_{t,h}}},$$

where $p_{t,h}$ again denotes an actual price and $\hat{p}_{t,h}$ an imputed price.

Now define V_h as the ratio of the actual to imputed price relative for house h:

$$V_h = \frac{p_{t+k,h}}{p_{t,h}} \Bigg/ \sqrt{\frac{p_{t+k,h}}{\hat{p}_{t,h}} \times \frac{\hat{p}_{t+k,h}}{p_{t,h}}} = \sqrt{\frac{p_{t+k,h}}{p_{t,h}} \Bigg/ \frac{\hat{p}_{t+k,h}}{\hat{p}_{t,h}}}. \tag{11}$$

The average squared error of the log price relatives of each hedonic method is given by:

$$D = \left(\frac{1}{H}\right) \sum_{h=1}^{H} [\ln (V_h)]^2, \tag{12}$$

where the summation in equation (12) takes place across the whole repeat-sales sample. We prefer whichever model has the smaller value of D (see Table 2).

Given that we use repeat sales as a benchmark for our imputed price relatives, our intention is to exclude repeat sales where the house was renovated between sales. We attempt to identify such houses in two ways. First, we exclude repeat sales where one or more of the characteristics have changed between sales (e.g. 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.[13] Finally, for houses that sold more than twice during our sample period (2001–11), 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 houses exert equal influence on our results.

[13]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.

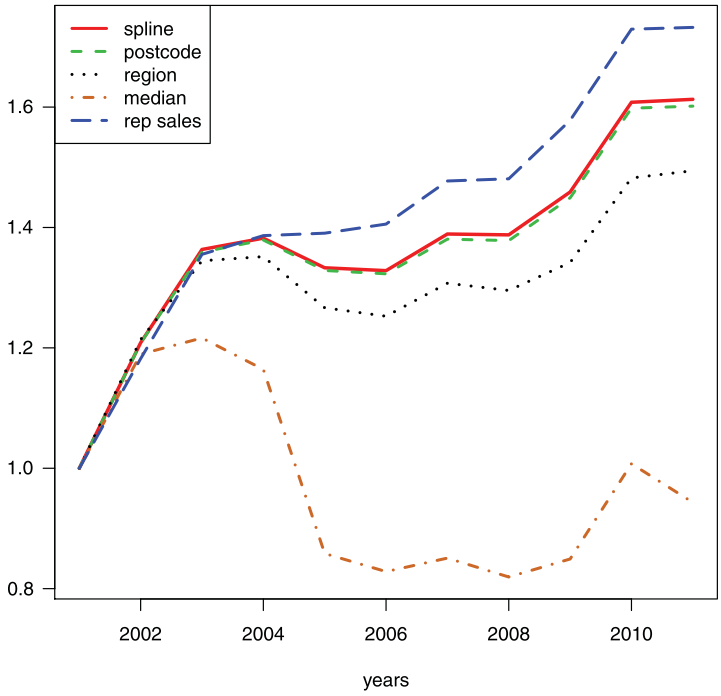

Figure 2. Price Indexes Calculated on the Restricted Dataset (with 2001 Normalized to 1)

Note: Here, *spline*, *postcode*, and *region* denote, respectively, the price indexes generated by the hedonic models with (i) geospatial splines, (ii) postcode dummies, and (iii) Residex-region dummies: *median* and *rep sales* denote the median and repeat-sales price indexes. [Colour figure can be viewed at wileyonlinelibrary.com]

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

In our dataset, Figures 2 and 3 indicate that there is an upward bias in the repeat-sales price relatives. We correct for this bias by adjusting the repeat-sales price relatives $p_{t+k,h}/p_{t,h}$ as follows:

$$\left(\frac{p_{t+k,h}}{p_{t,h}}\right)^{\text{adj}} = \left[\left(\frac{P^{\text{Hed}}_{t+k}}{P^{\text{Hed}}_{t}}\right) \Big/ \left(\frac{P^{\text{RS}}_{t+k}}{P^{\text{RS}}_{t}}\right)\right]\left(\frac{p_{t+k,h}}{p_{t,h}}\right), \tag{13}$$

where $P^{\text{RS}}_{t+k}/P^{\text{RS}}_{t}$ denotes the change in the repeat-sales price index between periods t and $t+k$, while $P^{\text{Hed}}_{t+k}/P^{\text{Hed}}_{t}$ is the change in a reference hedonic index, calculated using the Törnqvist formula in equation (5) over the same time

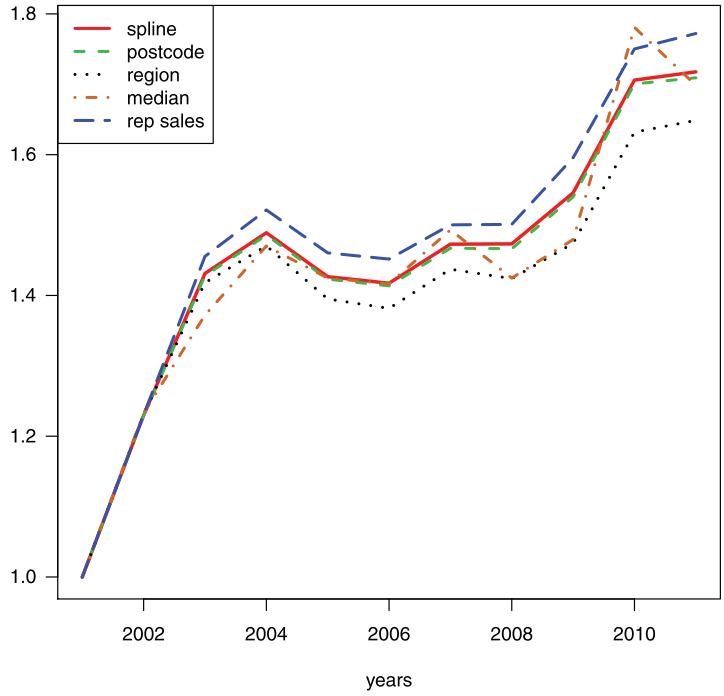

Figure 3. Price Indexes Calculated on the Full Dataset (with 2001 Normalized to 1)

Note: Here, *spline*, *postcode*, and *region* denote, respectively, the price indexes generated by the hedonic models with (i) geospatial splines, (ii) postcode dummies, and (iii) Residex-region dummies: *median* and *rep sales* denote the median and repeat-sales price indexes. [Colour figure can be viewed at wileyonlinelibrary.com]

(a) Postcode-Based Price Indexes

(b) Residex-Region–Based Price Indexes

interval. Hence the ratios of actual to imputed price relatives are adjusted as follows:

$$V_h^{\text{adj}} = \sqrt{\left(\frac{p_{t+k,h}}{p_{t,h}}\right)^{\text{adj}} \Big/ \frac{\hat{p}_{t+k,h}}{\hat{p}_{t,h}}} = V_h \sqrt{\left[\left(\frac{P_{t+k}^{\text{Hed}}}{P_t^{\text{Hed}}}\right) \Big/ \left(\frac{P_{t+k}^{\text{RS}}}{P_t^{\text{RS}}}\right)\right]}. \tag{14}$$

Bias corrected D coefficients, denoted by D^{adj} in Table 2, are then calculated as follows:

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

When calculating V_h^{adj}, we have three sets of hedonic results—corresponding to models (i), (ii), and (iii)—that could be used in equation (14) when making the bias correction. Hence in Table 2, we use each hedonic index in turn as the reference to calculate the D^{adj} coefficients. This allows us to check the robustness of our correction.

There are 101 752 repeat-sales houses in the full dataset. As a result of the deletions explained above, the sample was reduced to 87 700 houses. Our results are shown in Table 2. The average squared error of the log price relatives D and D^{adj} are lowest for model (i), with its geospatial splines as defined in equation (9). The ranking of models is the same for D and all three versions of D^{adj}. Hence our findings are robust to treatment of the "lemons" bias.

Hence irrespective of whether we correct for the upward bias in the repeat-sales price relatives, and in the case where a correction is made irrespective of which hedonic model is used as the benchmark, we always obtain the same ranking of methods. The best performing model is (i), followed by postcode- and Residex-region–based models (ii) and (iii) as defined in equation (10).

Although the differences reported in Tables 1 and 2 in the C and D statistics of models (i), (ii), and (iii) are small, they are nevertheless statistically significant. To show this, we apply the following hypothesis test based on the central limit theorem (see, e.g. Devore and Berk, 2012, pp. 490–1). Both criteria, C and D, are of the form:

$$\bar{X} = \frac{1}{H}\sum_{h=1}^{H} u_h^2,$$

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

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

where s_i (i=1,2) is the sample standard deviation of u_h^2 of the hedonic model i. The test statistic and corresponding two-sided p-values of this exercise are shown in Table A.7.

3.5. *House Price Indexes*

Here, we focus on the Törnqvist price index formula in equation (5). The results for our restricted dataset with no missing characteristics are shown in Figure 2. Price indexes for the full dataset (where prices are imputed from models HM1–HM8) are shown in Figure 3. Tables corresponding to these figures are provided in the Appendix (see Tables A.5 and A.6).

Five price index series are presented in Figures 2 and 3. Methods (i), (ii), and (iii) are derived from the hedonic models defined in equations (9) and (10). Method (iv) is a median index and Method (v) is a repeat-sales index (where all repeat sales are given equal weight irrespective of the time interval between sales). 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.

Two main themes emerge from these results. First, the exclusion of houses with missing characteristics has a big impact. For the case of the median index, the impact is dramatic. According to the median index calculated on the restricted dataset in Figure 2, house prices were lower in 2011 than in 2001. By contrast, based on the full

dataset in Figure 3, house prices were 70 percent higher in 2011 than in 2001. The explanation for this result is that the houses with missing characteristics tend to be cheap and are concentrated predominantly in the early part of our dataset. The three hedonic indexes in 2011 are also larger when calculated over the whole dataset. For example, focusing on our preferred Method (i), house prices are 72 percent higher in 2011 than in 2001 when calculated over the full dataset, but only 61 percent higher in 2011 when calculated over the restricted dataset. These results emphasize the importance of addressing the missing characteristics problem.

The second main theme is that the price indexes derived using geospatial splines in both Figures 2 and 3 rise faster than their postcode- or Residex-region–based counterparts. The gap between the spline- and postcode-based indexes is small. Prices rose from 2001 to 2011 by 71.8 percent according to Method (i) (geospatial spline), and by 70.9 percent according to Method (ii) (postcodes), based on the full dataset. The gap, though, is rather larger when Method (i) is compared with Method (iii) (Residex regions), according to which prices rose by 64.9 percent from 2001 to 2011. One possible explanation for these findings is that the average locational quality of the houses sold within a postcode and Residex region gets worse over time. Our geospatial spline-based indexes correct for this type of quality shift while the postcode- and Residex-region–based indexes do not. Also, if shifts in locational quality occur, they should be more pronounced in the geographically larger Residex regions than in postcodes, thus potentially explaining why the gap is bigger for Method (iii) (Residex regions) than for Method (ii) (postcodes).

We can check whether these kinds of declines in the quality of the locations of sold houses within postcodes and regions occur using the following algorithm:

1. Choose a postcode.
2. Calculate the mean number of bedrooms, bathrooms, land area, and quarter of sale over the 11 years for that postcode.
3. Impute the price of this average house in every location in which a house actually sold in 2001, …, 2011 in that postcode using the semilog model with spline of year 2001.
4. Take the geometric mean of these imputed prices for each year.
5. Repeat for another postcode.
6. Take the geometric mean across postcodes in each year.
7. Repeat steps 3–6 using the spline of year 2002, and then the spline of 2003, and so on.

If our hypothesis is correct, then irrespective of which year's spline is used as the reference, the geometric means from step 6 should fall over time. This is indeed what we observe for both the postcodes and regions (see Figure 4).

Most of the fall in the geometric means in Figure 4 occur in the first half of the sample. Also, the fall is much larger for the Residex regions than for the postcodes. This indicates that the extent of the downward bias depends on how fine the geographical zones are over which the locational dummies are defined. Smaller zones generate smaller biases.

There remains the question of why the average quality of houses sold within postcodes and regions deteriorated over our sample period. One possible

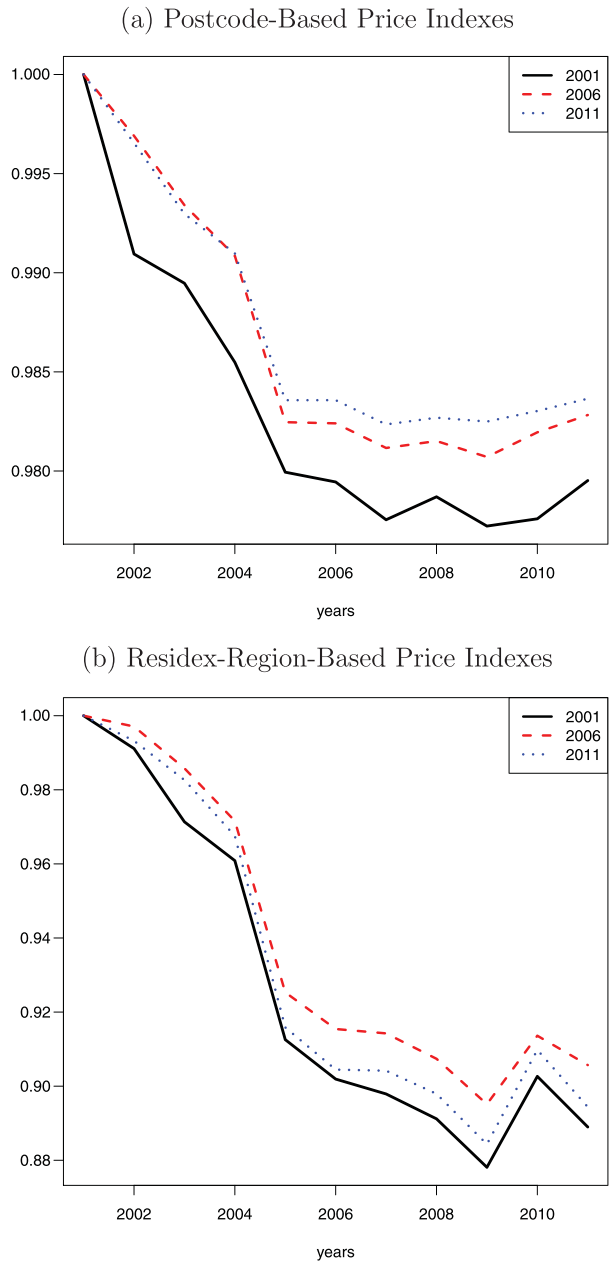

Figure 4. Evidence of Bias in the Postcode-Based/Residex-Region–Based Price Indexes (with 2001 Normalized to 1)

Note: Each curve measures the change in the value of the average location within (a) postcodes/(b) Residex regions of sold houses over time, using a reference geospatial spline surface to make the comparison. Here, the reference splines considered are those of 2001, 2006, and 2011. Irrespective of which spline is used, the value of the average location of sold houses declines over time. To simplify matters, we use only the splines derived from the restricted dataset with no missing characteristics. [Colour figure can be viewed at wileyonlinelibrary.com]

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 and has continued ever since (except for a flat period between 2004 and 2008). In a hot market it may be that "beggars (i.e. buyers) can't be choosers" and hence must settle for progressively worse locations in addition to paying higher prices.

4. Conclusion

The increasing availability of geospatial data could potentially lead to improvements in 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 using a two-step approach. First, a hedonic model is estimated that consists of a parametric part defined on the physical characteristics of houses and a non-parametric spline function defined on the longitudes and latitudes of the houses. Second, the price indexes are then calculated from the hedonic model using the hedonic imputation method. The use of a geospatial spline allows locational effects to be captured more precisely than in a fully parametric model that uses postcode dummies, while avoiding the curse of dimensionality that arises in a fully non-parametric model.

Applying our semiparametric approach to data for Sydney, Australia, three main results emerge. First, restricting the comparison to houses for which we have a full set of characteristics causes a sample selection bias problem. It is important, therefore, that the full dataset is used. The hedonic imputation method is well suited to resolving this problem, since it allows each house price to be imputed from a hedonic model with exactly the same mix of characteristics.

Second, the inclusion of a geospatial spline clearly improves the performance of the hedonic model as measured by the D statistic. However, its impact on the resulting price indexes is quite small, as compared with when postcode dummies are used. When the alternative is Residex-region dummies, the impact of using a geospatial spline is much larger.

Third, although the difference is small, our results indicate a slight downward bias in the price index when postcodes are used. This can be attributed to a systematic decline over time within each postcode in the locational quality of houses sold. This trend may itself be a natural consequence of the long housing boom in Sydney that started in 1993. The downward bias is much more pronounced for a hedonic model that controls for locational effects using the more aggregated Residex-region dummies (there are on average 15 postcodes in each Residex region).

The main implication of our findings is that the benefit of using geospatial data in a house price index depends on how finely defined the identifiable locational zones in a city are. The postcodes in Sydney are sufficiently compact (on average during the 2001–11 period, the number of residents per postcode was about 16,000) that a switch to using geospatial data has only a marginal impact on the resulting house price index. This is good news for NSIs. At least in Sydney, postcodes seem to be sufficient to control for locational effects in a hedonic house

price index. It remains to be seen how representative Sydney is of other cities in this regard.[14]

References

Bao, H. X. H. and A. T. K. Wan, "On the Use of Spline Smoothing in Estimating Hedonic Housing Price Models: Empirical Evidence Using Hong Kong Data," *Real Estate Economics*, 32(3), 487–507, 2004.

Brunauer, W. A., W. Feilmayr, and K. Wagner, "A New Residential Property Price Index for Austria," *Statistiken Daten & Analysen*, Q3/12, 90–102, 2012.

Cavallo, A., "Online and Official Price Indexes: Measuring Argentina's Inflation," *Journal of Monetary Economics*, 60(2), 152–65, 2013.

Clapp, J. M., "A Semiparametric Method for Estimating Local House Price Indices," *Real Estate Economics*, 32(1), 127–60, 2004.

Clapp, J. M. and C. Giaccotto, "Estimating Price Trends for Residential Property: A Comparison of Repeat Sales and Assessed Value Methods," *Journal of Real Estate Finance and Economics*, 5(4), 357–74, 1992.

Colwell, P. F., "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, 1998.

Corrado, L. and B. Fingleton, "Where Is the Economics in Spatial Econometrics?" *Journal of Regional Science*, 52(2), 210–39, 2012.

de Haan, J., "Direct and Indirect Time Dummy Approaches to Hedonic Price Measurement," *Journal of Economic and Social Measurement*, 29(4), 427–43, 2004.

———, "Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and Re-Pricing Methods," *Journal of Economics and Statistics (Jahrbücher für Nationalökonomie und Statistik)*, 230(6), 772–91, 2010.

de Haan, J. and F. Krsinich, "Scanner Data and the Treatment of Quality Change in Non-Revisable Price Indexes," *Journal of Business and Economic Statistics*, 32(3), 341–58, 2014.

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

Diewert, W. E., "Alternative Approaches to Measuring House Price Inflation," Economics Working Paper 2011-1, Vancouver School of Economics, 2011.

Diewert, W. E. and C. Shimizu, "Residential Property Price Indexes for Tokyo," *Macroeconomic Dynamics*, 19(8), 1659–714, 2015.

Diewert, W. E., S. Heravi, and M. Silver, "Hedonic Imputation versus Time Dummy Hedonic Indexes," in W. E. Diewert, J. Greenlees, and C. Hulten (eds), *Price Index Concepts and Measurement*, NBER Studies in Income and Wealth, University of Chicago Press, Chicago, 161–96, 2009.

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

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

Eurostat (2015), *Detailed Technical Manual on Owner-Occupied Housing for Harmonised Index of Consumer Prices* (June 2015 Draft), Eurostat, Luxembourg.

Fik, T. J., D. C. Ling and G. F. Mulligan, "Modeling Spatial Variation in Housing Prices: A Variable Interaction Approach," *Real Estate Economics*, 31(4), 623–46, 2003.

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

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

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

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

Hill, R. J. and I. Syed, "Hedonic Price-to-Rent Ratios, User Cost, and the Detection of Departures from Equilibrium in the Housing Market, *Regional Science and Urban Economics*, 56, 60–72, 2016.

[14]The postcodes are quite finely defined in Sydney. However, the sensitivity of prices to location also matters. It is probable that prices in Sydney are particularly sensitive to location, due to the presence of a harbor and beaches.

Hill, R. J., D. Melser and I. Syed, "Measuring a Boom and Bust: The Sydney Housing Market 2001–2006," *Journal of Housing Economics*, 18(3), 193–205, 2009.
Knight, E., "Australian Housing Market—Construction of Metropolitan Sydney House Price Indices Using Hedonic Estimation and Repeat Sales Method," Ph.D. thesis, University of Sydney, Australia, 2015.
Nappi-Choulet, I. and T. Maury, "A Spatiotemporal Autoregressive Price Index for the Paris Office Property Market," *Real Estate Economics*, 37(2), 305–40, 2009.
R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/, 2013.
Rambaldi, A. N. and C. S. Fletcher, "Hedonic Imputed Property Price Indexes: The Effects of Econometric Modeling Choices," *Review of Income and Wealth*, 60, Supplementary Issue, S423–48, 2014.
Rambaldi, A. N. and D. S. P. Rao, "Econometric Modeling and Estimation of Theoretically Consistent Housing Price Indexes," CEPA Working Papers Series WP042013, School of Economics, University of Queensland, Australia, 2013.
Reid, B., "Hedonic Imputation House Price Indexes: Bias and Other Issues," Honours thesis, School of Economics, University of New South Wales, Sydney, Australia, 2007.
Schäfer, P. and J. Hirsch, "Do Urban Tourism Hotpots Affect Berlin Housing Rents?" Mimeo, 2016.
Shimizu, C., "Estimation of Hedonic Single-Family House Price Function Considering Neighborhood Effect Variables," *Sustainability*, 6, 2946–60, 2014.
Shimizu, C., K. G. Nishimura and T. Watanabe, "Housing Prices in Tokyo: A Comparison of Hedonic and Repeat Sales Measures," *Journal of Economics and Statistics (Jahrbücher für Nationalökonomie und Statistik)*, 230(6), 792–813, 2010.
Silver, M., "House Price Indices: Does Measurement Matter?" *World Economics*, 12(3), 69–86, 2011.
Silver, M. and S. Heravi, "The Difference Between Hedonic Imputation Indexes and Time Dummy Hedonic Indexes," *Journal of Business and Economic Statistics*, 25(2), 239–46, 2007.
Stone, C. J., "The Dimensionality Reduction Principle for Generalized Additive Models," *Annals of Statistics*, 14(2), 590–606, 1986.
Wood, S. N., *Generalized Additive Models: An introduction with R*, Chapman & Hall/CRC, Boca Raton, FL, 2006.
———, "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, 2011.

Supporting Information

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

Appendix

A.1: Further Information on the Data Set
A.2: Testing Whether the *C* and *D* Statistics are Significantly Different across Models
Table A.1: Summary of Characteristics
Table A.2: Criteria for Removing Outliers
Table A.3: Number of Observations per Year with Missing Characteristics
Table A.4: Akaike Information Criterion (Restricted Data Set)
Table A.5: House Price Indexes (Restricted Data Set)
Table A.6: House Price Indexes (Full Data Set)
Table A.7: Test Statistic and Two-Sided *p*-Values (in Brackets) of Significance Tests
A.3: Estimation of our Semiparametric Hedonic Model and Robustness Checks
Table A.8: Sum of Squared Errors of the Price Relatives and Computational Time for Different Basis Dimensions
Figure A.1: Sum of Squared Errors of the Price Relatives and Computational Time for Different Basis Dimensions
Table A.9: AIC Distribution and Computational Time for 100 Repetitions of the Fit for the Year 2011 Based on Different Number of Randomly Chosen Observations
Figure A.2: AIC Distribution and Computational Time for 100 Repetitions of the Fit for the Year 2011 Based on Different Number of Randomly Chosen Observations