Graz Economics Papers – GEP

GEP 2014–05

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

Robert J. Hill, Michael Scholz

September 2014

Department of Economics
Department of Public Economics
University of Graz

Incorporating Geospatial Data in 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

23 September 2014

Abstract:
The increasing availability of geospatial data (i.e., exact longitudes and latitudes for each house) has the potential to improve the quality of house price indexes. It is not clear though how best to use this information. We show how geospatial data can be included as a nonparametric spline surface in a hedonic model. The hedonic model itself is estimated separately for each period. Price indexes are then computed by inserting the imputed prices of houses obtained from the hedonic model into the Fisher price index formula. Using a data set consisting of about 450 thousand observations for Sydney, Australia over the period 2001-2011 we demonstrate the superiority of a geospatial spline over postcode dummies as a way of controlling for locational effects. While the difference in the resulting price indexes is not that large – since the postcodes in Sydney are quite narrowly defined – we nevertheless find evidence of a slight bias in the postcode based indexes. This can be attributed to systematic changes over time within each postcode in the locational quality of houses sold. (*JEL*. C43; E01; E31; R31)

Keywords: Housing market; Hedonic imputation; Price index; Geospatial spline; Quality adjustment

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), and the Society for Economic Measurement Conference at University of Chicago (August 2014).

1 Introduction

The slump in the US housing market that started in 2006 precipitated a global financial crisis. The crisis demonstrated the pivotal role played by the housing market in the broader economy. It is important therefore that governments, central banks and market participants are kept well informed of developments in housing markets. One potential source of confusion is the sensitivity of house price indexes to the method of construction (see Silver 2012). 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 has been recently recognized by the international community. Eurostat, the UN, ILO, OECD, World Bank and IMF together commissioned a Handbook on Residential Property Price Indices that was completed in 2013 (see de Haan and Diewert 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. 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).[1]

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. Hedonic indexes are often constructed using 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), most of the important information contained in geospatial data is lost when one simply uses the average longitude and latitude. In other words, the use of geospatial data requires a shift away from average-characteristics hedonic methods.

[1] A hedonic house price index should not be confused with an automated valuation model (AVM). The latter aims to impute prices for individual houses (see Thibodeau 2003). A price index measures changes in house prices over time. Also, the unit of comparison for an index is typically a region (e.g., a city) rather than an individual house.

The two main alternatives to the average-characteristics method that have been used in the hedonic price index literature are the time-dummy and hedonic imputation methods (see Diewert 2011, and Hill 2013). 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, hedonic imputation methods estimate the hedonic model separately for each period and then use the hedonic model to impute a price for each individual house. These imputed prices can then be inserted into a standard Fisher price index formula.

This is the approach we follow here. Our hedonic model is semilog in the physical characteristics with locational effects captured using a geospatial spline. We then compare this model with equivalent models that capture locational effects using postcode dummies or region dummies.

We apply our methods to a data set consisting of 454 507 observations for Sydney, Australia over the period 2001 to 2011. A serious problem with the data is that one or more of the characteristics are missing for a substantial portion of the houses. Simply excluding these observations can cause sample selection bias. We show that an additional benefit of the hedonic imputation method is that it can be modified to deal with this problem, and thus ensure that the price indexes are calculated using the full data set.

Our results clearly confirm the superiority of geospatial splines over postcodes, both in terms of the deviation between actual and imputed prices and in the case of repeat-sales between actual and imputed price relatives. Although the difference is small, our results indicate a slight downward bias in the postcode-based price indexes, which becomes more pronounced when postcode dummies are replaced with more aggregated Residex-region dummies.

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 data set 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. Details

regarding the estimation of the geospatial spline function are provided in an Appendix.

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 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 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.

Hedonic price indexes for housing are constructed in three main ways (see Diewert 2011 and Hill 2013). We briefly discuss each of these below as they are applied in a housing context.

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 of log prices p_h (i.e., $y_h = \ln p_h$), 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 and T denote respectively the number of houses, 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 after controlling for differences in the attributes of the houses. 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:[2]

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

Although it is the original hedonic method, the time-dummy method has not been used much by index providers. 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.[3] 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 (see Hill 2013).

2.3 Hedonic imputation methods

The hedonic imputation approach estimates a separate hedonic model for each period or a few adjacent periods:[4]

$$y_t = Z_t\beta_t + \varepsilon_t. \tag{3}$$

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

[2]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.

[3]By contrast, in an academic study where the results do not need to be continually updated, this lack of fixity in the results is not a problem. This probably explains why the time-dummy method is more widely used in academic research (see for example Bracke 2014).

[4]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 observations to estimate the model separately for each year.

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] ,[6]

$$\hat{p}_{t+1,h}(z_{t,h}) = \exp\left(\sum_{c=1}^{C} \hat{\beta}_{c,t+1} z_{c,t,h}\right). \tag{4}$$

These imputed prices can then be inserted into standard price index formulas. We will refer to a formula that focuses on the houses that sold in the earlier period t as Laspeyres-type, and a formula that focuses on the houses 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 house.[7] 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.

The indexes presented below are all of the single imputation variety. This means 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 for example Silver and Heravi 2001, de Haan 2004, and Hill and Melser 2008). Empirically we try both approaches. The resulting price indexes are virtually indistinguishable. Hence to simply the presentation we focus here only on single imputation price indexes.

The (single imputation) Paasche, Laspeyres and Fisher price indexes between periods

[5]See Silver and Heravi (2007) and Rambaldi and Rao (2013) for a discussion of some of the advantages of the hedonic imputation method. Rambaldi and Rao (2013) and Rambaldi and Fletcher (2014) show how the stability of the hedonic imputation method can be improved by using Kalman filters to link the hedonic models across time periods.

[6]Omitted variables are a potentially serious problem in hedonic models of the housing market (see Coulson 2008). The problem though is not as bad for price indexes as for AVMs. To see why, consider the hedonic imputation method. By construction about half the houses will perform better than average on the omitted variables. The imputed prices of these houses will be too low. Conversely, about half the houses will perform worse than average on the omitted variables. The imputed prices of these houses will be too high. In other words, the biases created by omitted variables will partially offset each other in a price index. The same is not true for an AVM.

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

t and $t+1$ are calculated as follows:

$$\begin{aligned}
\text{Paasche Imputation}: \; P_{t,t+1}^{PI} &= \prod_{h=1}^{H_{t+1}} \left[\left(\frac{p_{t+1,h}}{\hat{p}_{t,h}(z_{t+1,h})} \right)^{1/H_{t+1}} \right] \\
\text{Laspeyres Imputation}: \; P_{t,t+1}^{LI} &= \prod_{h=1}^{H_t} \left[\left(\frac{\hat{p}_{t+1,h}(z_{t,h})}{p_{t,h}} \right)^{1/H_t} \right] \\
\text{Fisher Imputation}: \; P_{t,t+1}^{FI} &= \sqrt{P_{t,t+1}^{PI} \times P_{t,t+1}^{LI}}
\end{aligned} \tag{5}$$

Hedonic imputation methods require reasonably large data sets. However, this is becoming less of a constraint, given the large increase in data availability. Hedonic 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 maintained indexes to use a hedonic imputation method as far as we are aware are the FNC Residential Price Index in the US (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, Feilmayr, and Wagner 2012).

2.4 Average-characteristics methods

Average-characteristics methods, like hedonic imputation methods, generally begin by estimating the hedonic model separately for each period as in (3). They also use standard price index formulas. The key difference is that average-characteristics methods typically construct an average house for each period, and then impute the price of this hypothetical house (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 in (3). A price index is obtained by taking the ratio of the imputed price of the same average house in two different periods. By construction the average-characteristics method must impute the price of the hypothetical average house in both periods.

Taking the semi-log hedonic model as our point of reference, a price index between periods t and $t+1$ can be calculated using the average house from either period (see

Diewert 2003). 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 house over time. Its biggest weakness is that it cannot easily be extended to incorporate geospatial data. This is because averaging longitudes and latitudes does not make much sense. For example, in the case of a city like Sydney built round a natural harbor it may be underwater!

It is perhaps unfortunate therefore that 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 (see Saarnio 2006), Statistics Norway (see Thomassen 2007) and INSEE (the national statistical office of France) (see Gouriéroux and Laferrère 2009) all use implicit Paasche price indexes to compute national house price indices. Statistics Sweden uses a variant on an implicit Laspeyres price index (see Ribe 2009).

2.5 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. Postcode dummies can be used in combination with the time-dummy or hedonic imputation methods. While with the average characteristics method it is in principle possible to fractionally allocate postcodes to the average house (e.g., postcode 1 gets a weight of 10 percent, postcode 2 a weight of 5 percent, etc.) such an approach fails to make use of the information contained in the spatial dependence structure of the data.

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 or imputation versions of the hedonic model (see for example Hill and Melser 2008).[8]

Using distances to amenities as characteristics is 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) 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. This last problem is potentially the easiest to solve, by using quadratics, cubics, splines, etc. 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

[8] A perhaps more informative alternative to distance is commuting time, which can be calculated by combining geospatial data with information on the public transport system in a city (see for example Shimizu 2014).

form (see for example Corrado and Fingleton 2012):

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

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

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.

Replacing a spatial weights matrix with a spatiotemporal weights matrix significantly increases the computational burden and complicates the derivation of price indexes (see for example Nappi-Choulet and Maury 2009). 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). Dorsey et al. 2010) and Rambaldi and Rao (2013) 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.

Nonparametric approaches

Nonparametric methods provide a different and potentially more flexible alternative to spatial autoregressive models for modeling spatial dependence. Nonparametric 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 nonparametric 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 nonparametric function $g_t(\cdot)$ defined on the geospatial data z_{lat}, z_{long} as follows:

$$y = Z\beta_t + g_t(z_{lat}, z_{long}) + \varepsilon, \tag{7}$$

where again $y = \ln p$.

Imputed prices for each house can then be obtained by inserting its particular mix of characteristics (including the longitude and latitude) into the estimated hedonic model. More specifically, consider the Fisher price index in (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 (7) 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 (2014). 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 (2014) 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 other nonparametric methods by amongst others Colwell (1998), Pavlov (2000), Fik, Ling and Mulligan (2003), Clapp (2004), Hardman (2011) and Knight (2014). Again

though – with the exception of Hardman (2011) – none of these authors combines a nonparametric treatment of geospatial data with the hedonic imputation method.[9]

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 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 house we have the following characteristics: the actual sale price, time of sale, postcode, property type (i.e., detached or semi), number of bedrooms, number of bathrooms, land area, exact address, longitude and latitude. (We exclude all townhouses from our analysis since the corresponding land area is for the whole strata and not for the individual townhouse itself.) Some summary statistics are provided in Table 1.

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 is a concentration of data entry errors in the tails, caused for example by the inclusion

[9]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.

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

While we deleted bedroom, bathroom and land area counts outside the allowed ranges, we retained the house itself in the data set as long as the price and longitude/latitude were available and within the allowed ranges as specified in Table 2. In total less than 1 percent of the houses were deleted. After deletions, our data set consists 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' data set. 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.

Clearly observations with missing characteristics are a serious problem in our data set. We explain in section 3.3 how we deal with this problem.

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
Complete	5886	6188	7759	8668	21662	26467	32936	28063	35703	32933	33877

3.2 Model estimation and performance

Here we compare the performance of three models:
(i) semilog in physical characteristics with a geospatial spline;
(ii) semilog in physical characteristics with postcode dummies;
(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 nonparametric model (see for example Stone 1986). Each model is estimated separately for each year $t = 2001, \ldots, 2011$. Model (i) takes the following form:

$$y = Z\beta_t + g_t(z_{lat}, z_{long}) + \varepsilon, \tag{8}$$

where $g_t(\cdot)$ now denotes a spline. In the case of (i) it is necessary to estimate the characteristic shadow price vector β_t and the spline surface $g_t(z_{lat}, z_{long})$. An example of one of our spline surfaces (for 2007) is provided in Figure 1.

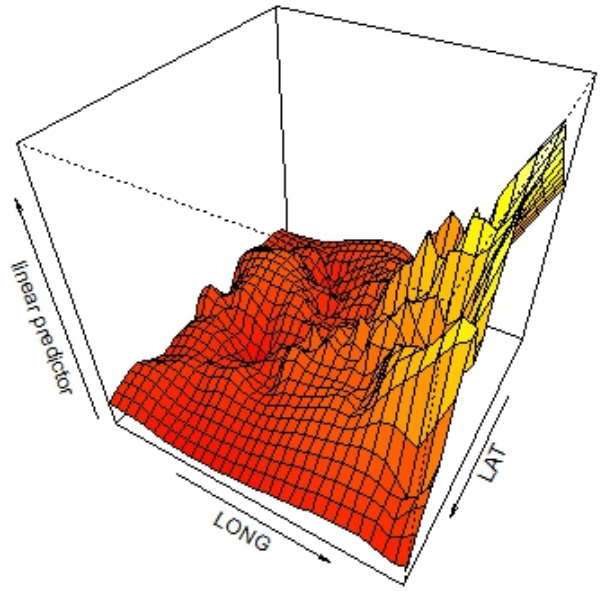
Models (ii) and (iii) take the following form:

$$y = Z\beta_t + L\lambda_t + \varepsilon. \tag{9}$$

In the case of models (ii) and (iii) it is necessary to estimate the characteristic shadow price vector β_t and the location dummy variable shadow price vector λ_t. The difference between (ii) and (iii) is that L and λ_t are defined over 242 postcodes for (ii) versus 16 Residex regions for (iii). Each Residex region therefore consists of about 15 postcodes. The 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), Penrith-Windsor (2740 to 2777).

The way in which we estimate the semiparametric model (i) is explained in the Appendix. Table 4 shows the values of the Akaike information criterion (AIC) for each

Figure 1: Spline Surface Based on Restricted Data Set for 2007



model in each year (for the restricted data set). Table 5 shows the average squared error of the log prices, C_t, defined as follows:

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

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 in both Tables 4 and 5, model (i) with its geospatial spline in (8) clearly outperforms its postcode/Residex-region based competitors (ii) and (iii) in (9) in terms of goodness-of-fit. As expected model (ii) with its finer postcode classification of regions likewise outperforms model (iii) with its broader Residex regions.

Table 4: Akaike information criterion (restricted data set)

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
(i)	-452	-203	-1448	-2471	-9638	-10644	-14052	-13980	-20436	-18857	-23659
(ii)	1321	1515	493	158	-4930	-4384	-6372	-8070	-12506	-11857	-16522
(iii)	3463	3807	3650	3634	4841	7970	11996	8583	14842	16980	5223

Note: Method (i) is the semiparametric model defined in (8). Methods (ii) and (iii) are both semilog models with location dummies as defined in (9). Method (ii) uses postcode dummies while method (iii) uses Residex dummies.

Table 5: Average squared error of the log prices (restricted data set)

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
(i)	0.048	0.050	0.044	0.040	0.036	0.038	0.037	0.034	0.032	0.032	0.028
(ii)	0.068	0.070	0.059	0.057	0.046	0.049	0.048	0.043	0.041	0.040	0.035
(iii)	0.105	0.108	0.093	0.089	0.073	0.079	0.084	0.079	0.089	0.098	0.068

Note: Methods (i), (ii) and (iii) are explained in the note in Table 4.

3.3 Missing characteristics

Excluding 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 data set (see Table 2). An alternative approach is to estimate 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.[10]

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

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

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)*

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

3.4 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 (see Reid 2007).

Suppose house h sells in both periods t and $t+k$. We therefore obtain the following repeat-sales price relative: $p_{t+k,h}/p_{t,h}$. Corresponding imputed price relatives for this repeat-sales house can be calculated as follows:

$$\text{Imputed Price Relative} : \sqrt{\frac{p_{t+k,h}}{\hat{p}_{t,h}} \times \frac{\hat{p}_{t+k,h}}{p_{t,h}}},$$

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

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

$$Z_h = \frac{p_{t+k,h}}{p_{t,h}} \Bigg/ \sqrt{\frac{p_{t+k,h}}{\hat{p}_{t,h}} \times \frac{\hat{p}_{t+k,h}}{p_{t,h}}} = \sqrt{\frac{p_{t+k,h}}{p_{t,h}} \Bigg/ \frac{\hat{p}_{t+k,h}}{\hat{p}_{t,h}}}. \tag{10}$$

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

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

We prefer whichever model has the smaller value of D (see Table 6).

Given that we use repeat-sales as a benchmark for our imputed price relatives, our intention is to exclude repeat sales where the house was renovated between sales. We attempt to identify such houses in two ways. First, we exclude repeat sales where one or more of the characteristics have changed between sales (for example a bathroom has been added). Second, we exclude repeat sales that occur within six months on the grounds that this suggests that the first purchase was by a professional renovator.[11] Finally, for houses that sold more than twice during our sample period (2001-2011), we only include the two chronologically closest repeat sales (as long as these are more than six months apart). This ensures that all repeat-sales houses exert equal influence on our results.

Initially in the full data set we started with 101 752 repeat-sales houses. As a result of the deletions explained above, the sample was reduced to 87 700 houses. For the restricted data set, the corresponding figures are 27 852 repeat sales of which 18 224 were left after deletions. Our results, shown in Table 6, again find that the hedonic model (i) including geospatial splines defined in (8) outperforms the postcode and Residex-region based models (ii) and (iii) defined in (9).

Table 6: Average squared error of the log price relatives D

Model	Restricted Data Set	Full Data Set
(i)	0.016802	0.036523
(ii)	0.016857	0.038863
(iii)	0.029087	0.052078

Note: Methods (i), (ii) and (iii) are explained in Table 4.

3.5 House price indexes

Here we focus on the Fisher price index formula in (5). The results for our restricted data set with no missing characteristics are shown in Table 7 and Figure 2. Corresponding

[11] 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.

price indexes obtained using the full data set using models HM1-HM8 are shown in Table 8 and Figure 3. In both cases, five sets of price indexes are presented. Methods (i), (ii) and (iii) are defined in (8) and (9). 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.

Table 7: House Price Indexes (Restricted Data Set)

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
(i)	1.0000	1.2085	1.3635	1.3822	1.3330	1.3283	1.3891	1.3877	1.4589	1.6080	1.6129
(ii)	1.0000	1.2081	1.3591	1.3792	1.3286	1.3230	1.3806	1.3782	1.4492	1.5981	1.6016
(iii)	1.0000	1.2138	1.3447	1.3516	1.2666	1.2525	1.3077	1.2954	1.3417	1.4822	1.4943
(iv)	1.0000	1.1820	1.3554	1.3865	1.3905	1.4056	1.4772	1.4809	1.5779	1.7294	1.7324
(v)	1.0000	1.190	1.2164	1.1642	0.8583	0.8284	0.8507	0.8194	0.8493	1.0075	0.9433

Note: Methods (i), (ii) and (iii) are explained in Table 4. Method (iv) is a median price index, and method (v) is a repeat-sales price index.

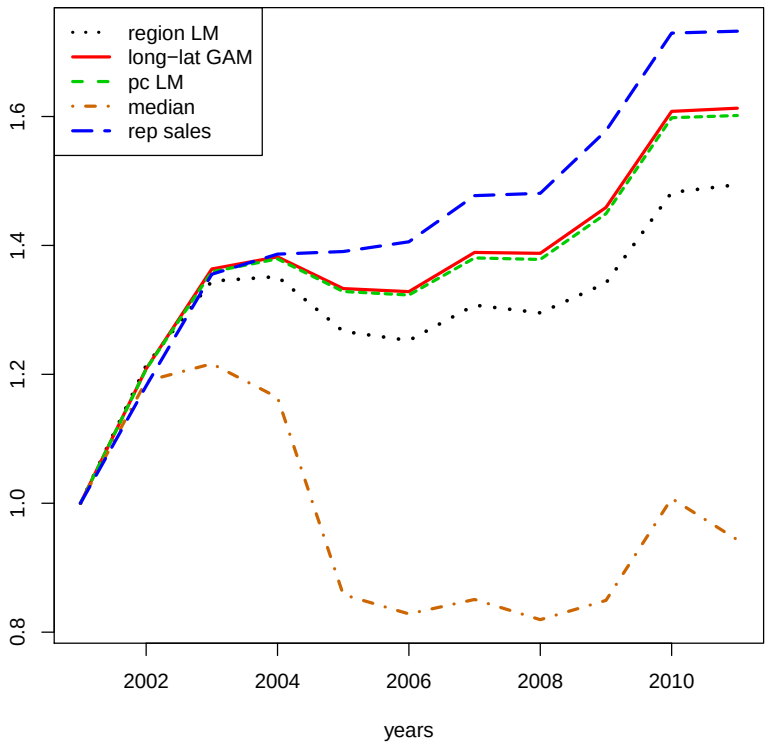
Table 8: House Price Indexes (Full Data Set)

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
(i)	1.0000	1.2298	1.4312	1.4891	1.4266	1.4173	1.4728	1.4733	1.5458	1.7059	1.7177
(ii)	1.0000	1.2293	1.4281	1.4857	1.4231	1.4136	1.4671	1.4664	1.5400	1.7005	1.7093
(iii)	1.0000	1.2308	1.4181	1.4704	1.3949	1.3816	1.4373	1.4239	1.4740	1.6320	1.6487
(iv)	1.0000	1.2304	1.4555	1.5215	1.4605	1.4517	1.5003	1.5010	1.5957	1.7502	1.7721
(v)	1.0000	1.2329	1.3720	1.4702	1.4247	1.4164	1.4932	1.4247	1.4795	1.7808	1.6986

Note: Methods (i), (ii) and (iii) are explained in Table 4, and methods (iv) and (v) in Table 7.

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 data set, house prices were lower in 2011 than in 2001. By contrast, based on the full data set, 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

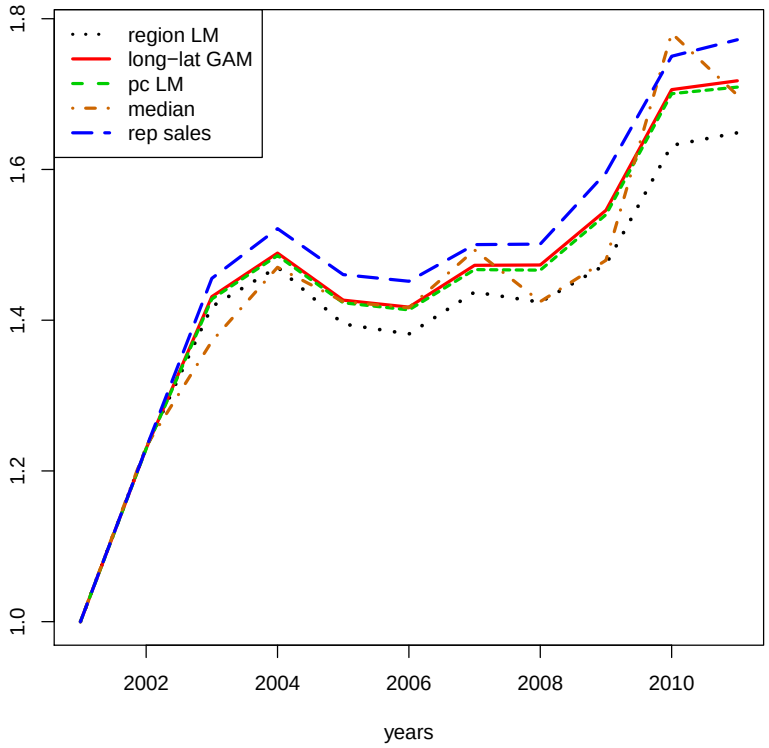
Figure 2: Price Indexes Calculated on the Restricted Data Set (with 2001 Normalized to 1)



predominantly in the early part of our data set. The three hedonic indexes in 2011 are also larger when calculated over the whole data set. For example, focusing on our preferred method (i), house prices are 72 percent higher in 2011 than in 2001 when calculated over the full data set, but only 61 percent higher in 2011 when calculated over the restricted data set. 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 data set. The gap though is rather larger when method (i) is compared with method (iii) (Residex re-

Figure 3: Price Indexes Calculated on the Full Data Set (with 2001 Normalized to 1)



gions), 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 will correct for this type of quality shift while the postcode and Residex-region based indexes will 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, etc.

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 Figures 4 and 5).

Most of the fall in the geometric means in Figures 4 and 5 occurs 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 are the geographical zones 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 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 ended in 2004. 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 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 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 nonparametric 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 spline allows locational effects to be captured more flexibly than in a fully parametric model that uses postcode dummies, while avoiding the curse of dimensionality that arises in a fully nonparametric 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 serious sample selection bias problem. It is important therefore that the full data set 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. However, its impact on the resulting price indexes is quite small, as compared with when postcode dummies are used. When Residex region dummies are used, the impact 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 systematic changes over time within each postcode in the locational quality of houses sold (when the housing market was booming). The downward bias is much more pronounced for a hedonic model that controls for locational effects using the more aggregated Residex-region dummies (where there are on average 15 postcodes in each Residex region).

An implication of this finding is that the benefit of using geospatial data in a house price index depends on how finely defined the identifiable locational zones are in a city. The postcodes in Sydney are sufficiently finely defined (on average they include 14,300 residents and cover 7.39 square kilometers) that a switch to using geospatial data has only a small impact on the resulting house price index. For cities with bigger locational zones, the benefit to using geospatial data will be larger.

A Appendix

A.1 Estimation of our semiparametric hedonic model

The semiparametric hedonic model in (8) is a simple example of a generalized additive model (GAM), a flexible model class that generalizes linear models with a linear predictor combined with a sum of smooth functions of covariates. The problem is to select the smooth functions and their degree of smoothness. Here, we use a penalized likelihood approach (see Wood 2006, and the references therein) based on a simple transformation and truncation of the basis that arises from the solution of the thin plate spline smoothing problem. This method is computationally efficient and avoids the problem of choosing the location of knots, known to be crucial for other basis functions. For example, consider the following function:

$$y = g(x) + \varepsilon, \tag{11}$$

where x is a d-vector ($d \leq n$), and n is the number of observations. A thin-plate spline smoothing function estimates g by finding the function $\hat{f}$ that minimizes

$$||\boldsymbol{y} - \boldsymbol{f}||^2 + \lambda J_{md}(f), \tag{12}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))^\top$, and $J_{md}(f)$ is a penalty function measuring the wiggliness of f with smoothing parameter λ, which controls the trade-off between the goodness of fit and smoothness of f.[12] Under suitable conditions it can be shown that the solution of (12) has the form,

$$\hat{f}(x) = \sum_{i=1}^{n} \delta_i \eta_{md}(||x - x_i||) + \sum_{j=1}^{M} \alpha_j \phi_j(x), \tag{13}$$

where δ_i and α_j are coefficients to be estimated, such that $\boldsymbol{T}^\top \boldsymbol{\delta} = \boldsymbol{0}$ with $T_{ij} = \phi_j(x_i)$. The $M = \binom{m+d-1}{d}$ functions ϕ_i are linearly independent polynomials spanning the space of polynomials in $\mathbb{R}^d$ of degree less than m, while the ϕ_i span the null space of J_{md}. Defining the matrix $\boldsymbol{E}$ by $E_{ij} = \eta_{md}(||x_i - x_j||)$, the thin plate spline fitting problem is

[12] For more details on J_{md} see Wood (2006). The order of the derivatives in the thin plate spline penalty term is specified by m. It is set to the smallest value that satisfies $2m > d + 1$ (in our case we have $d = m = 2$).

now the minimization of

$$||\boldsymbol{y} - \boldsymbol{E\delta} - \boldsymbol{T\alpha}||^2 + \lambda \boldsymbol{\delta}^\top \boldsymbol{E\delta} \qquad \text{s. t.} \qquad \boldsymbol{T}^\top \boldsymbol{\delta} = \boldsymbol{0}. \tag{14}$$

There are as many unknown parameters as there are data points. The computational cost of model estimation is proportional to the cube of the number of parameters. The computational burden of (14) can be reduced with the use of a low rank approximation. The basic idea of thin plate regression splines is now the truncation of the space of the wiggly components of the spline (with parameter $\boldsymbol{\delta}$), while leaving the $\boldsymbol{\alpha}$-components unchanged. For this let $\boldsymbol{E} = \boldsymbol{UDU}^\top$ be the eigen-decomposition of $\boldsymbol{E}$, such that $\boldsymbol{D}$ is the diagonal matrix of eigenvalues and the columns of $\boldsymbol{U}$ the corresponding eigenvectors. Also, $\boldsymbol{\delta}$ is restricted to the column space of $\boldsymbol{U}_k$, by writing $\boldsymbol{\delta} = \boldsymbol{U}_k \boldsymbol{\delta}_k$. Now with the choice of an appropriate submatrix $\boldsymbol{D}_k$ of $\boldsymbol{D}$ and $\boldsymbol{U}_k$, as the corresponding columns of $\boldsymbol{U}$, the minimization problem (14) becomes

$$\text{Min}_{\boldsymbol{\delta}_k, \boldsymbol{\alpha}} \left\{ ||\boldsymbol{y} - \boldsymbol{U}_k \boldsymbol{D_k} \boldsymbol{\delta_k} - \boldsymbol{T\alpha}||^2 + \lambda \boldsymbol{\delta_k}^\top \boldsymbol{D_k} \boldsymbol{\delta_k} \right\} \qquad \text{s. t.} \qquad \boldsymbol{T}^\top \boldsymbol{U}_k \boldsymbol{\delta_k} = \boldsymbol{0}. \tag{15}$$

Hence the computational cost is reduced from $O(n^3)$ to $O(k^3)$. The remaining problem is to find $\boldsymbol{U}_k$ and $\boldsymbol{D}_k$ sufficiently cheaply. Remember that a full eigen-decomposition requires $O(n^3)$ operations and thus is inappropriate. The use of the Lanczos method (cf. A.11 in Wood 2006) allows the calculation of $\boldsymbol{U}_k$ and $\boldsymbol{D}_k$ at the substantially lower cost of $O(n^2 k)$ operations.

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

A.2 Robustness checks on our semiparametric hedonic model

All computations are performed using R 2.15.3, R Core Team (2013). For the estimation of the GAMs we use the `mgcv` package,[13] in which the ideas discussed above are

[13] We apply the `bam`-function with the gaussian distribution and the identity link. `bam` was designed for very large datasets and is characterized by a much lower memory footprint and can be much faster than the also available `gam` function.

Table 9: Sum of squared errors of the price relatives and computational time for different basis dimensions

k	100	200	300	400	500	600	700	800	900
D	0.01805	0.01757	0.01729	0.01705	0.01690	0.01682	0.01699	0.01690	0.01677
time	1247	2491	3678	4656	5902	7148	8595	10135	11866

implemented. As previously mentioned, thin plate regression splines have the advantage of being a 'knot-free' method that is easily applied also in multivariate settings. As mentioned above, we construct them by starting with the basis and penalty for a full thin plate spline and then truncating this basis in an optimal manner. In other words, to obtain a low rank smoother we have to choose k, the dimension of the basis used to represent the smooth term. It is important to ensure that the basis dimension is not too small otherwise it would force oversmoothing. On the other hand using a too large value does not necessarily generate significant improvements in the goodness-of-fit while at the same time enormously increasing the computational burden. Furthermore, for large datasets the calculation of the thin plate basis can be time-consuming. As suggested in the `mgcv` package, the user can retain most of the advantages of the thin plate regression splines approach by supplying a reduced set of covariate values from which to obtain the basis - typically the number of covariate values used will be substantially smaller than the number of data points, and substantially larger than the basis dimension, k. For our house price data we set $k = 600$ and use only 2500 randomly chosen observations in each year for basis construction. To select these values, we first fixed the basis dimension at values $100, 200, \ldots, 900$, estimated each time the GAM for all 11 years and constructed the sum of squared errors of the price relatives D. Figure 6 and Table 9 show the results of this exercise. We find a strong decline of D up to $k = 600$ and a slight increase afterwards. We find also an almost linear increase in computational time indicating that, for example, the use of $k = 900$ would give a similar D value as for $k = 600$ but the computational cost would increase by 60 percent. Second, we checked the dependence of the estimation procedure with respect to the used observations for the basis construction. For this, we fixed $k = 300$ and

Table 10: AIC distribution and computational time for 100 repetitions of the fit for the year 2011 based on different number of randomly chosen observations.

nr. obs	1000	1500	2000	2500	3000	3500
AIC_{min}	-21569.03	-21611.09	-21597.29	-21709.60	-21531.56	-21411.30
AIC_{mean}	-21081.18	-21154.98	-21190.62	-21210.73	-21230.27	-21203.38
AIC_{max}	-20701.36	-20714.04	-20735.70	-20774.76	-20847.71	-20696.57
time	15581	22045	26636	34578	38375	44686

estimated for $1000, 1500, \ldots, 3500$ randomly chosen observations 100 times the GAM only for the year 2011. The distribution of the corresponding AIC-values as well as the computational time can be found in Figure 7 and Table 10. The results show that the minima and maxima of AIC are quite close. Thus the chosen number of observations for basis construction is not crucial at least when it substantially exceeds the number of basis dimensions. A growing number of observations only linearly increases the computational cost. The lowest AIC-value was observed for 2500 and for this reason we fix in our experiments the number of observations for basis construction to this value.

References

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), 487-507.

Bracke, P. (2014), "House Prices and Rents: Micro Evidence from a Matched Dataset in Central London," *Real Estate Economics*, forthcoming.

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

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

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 (2012), "Where Is the Economics in Spatial Econometrics?" *Journal of Regional Science* 52(2), 210-239.

Coulson, E. (2008), *Monograph on Hedonic Methods and Housing Markets*, Penn State University.

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

de Haan, J. and W. E. Diewert (eds.) (2013), *Handbook on Residential Property Price Indexes*, Luxembourg: Eurostat.

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

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

Diewert, W. E. and C. Shimizu (2014), "Residential Property Price Indexes for Tokyo," *Macroeconomic Dynamics*, forthcoming.

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.

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.

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

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

Knight, E. (2014), Australian Housing Market - Construction of Metropolitan Sydney House Price Indices Using Hedonic Estimation and Repeat Sales Method, PhD Thesis, University of Sydney, Australia.

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.

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.

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

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Rambaldi, A. N. and C. S. Fletcher (2014), "Hedonic Imputed Property Price Indexes: The Effects of Econometric Modeling Choices," *Review of Income and Wealth*, forthcoming.

Rambaldi, A. N. and D. S. P. Rao (2013), "Econometric Modeling and Estimation of Theoretically Consistent Housing Price Indexes," CEPA Working Papers Series WP042013, School of Economics, University of Queensland, Australia.

Reid, B. (2007). Hedonic Imputation House Price Indexes: Bias and Other Issues. Honours Thesis, School of Economics, University of New South Wales, Sydney, Australia.

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.

Shimizu, C. (2014), "Estimation of Hedonic Single-Family House Price Function Considering Neighborhood Effect Variables," *Sustainability* 6, 2946-2960.

Silver, M. (2012), "Why House Price Indexes Differ: Measurement and Analysis, IMF Working Paper, WP/12/125

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.

Silver, M. and S. Heravi (2007), "The Difference between Hedonic Imputation Indexes and Time Dummy Hedonic Indexes," *Journal of Business and Economic Statistics* 25(2), 239-246.

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

Thibodeau, T. G. (2003), "Marking Single-Family Property Values to Market," *Real Estate Economics* 31(1), 1-22.

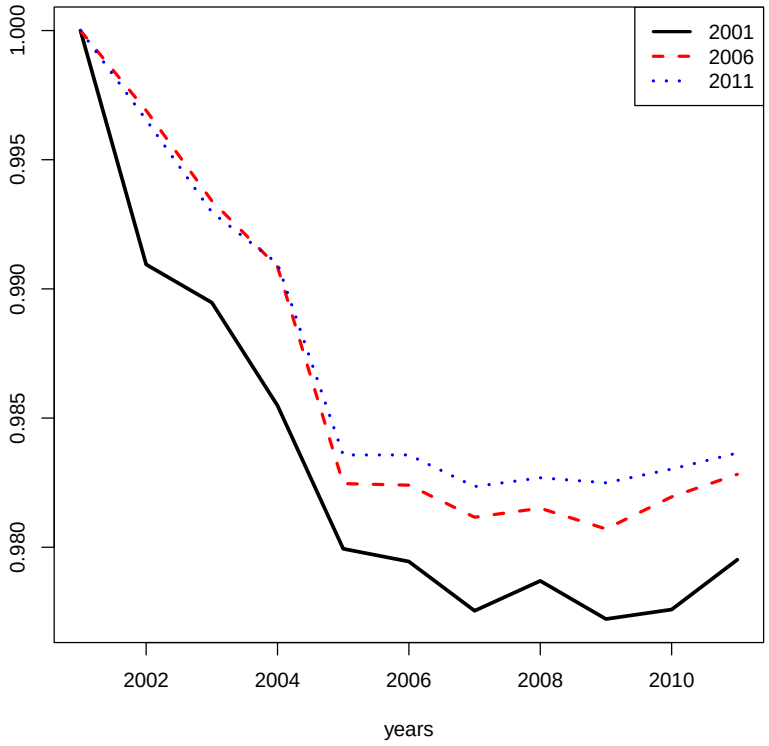
Thomassen, A. (2007), *Price Index for New Multidwelling Houses: Sources and Methods*, Statistics Norway/Department of Industry Statistics/Construction and Service Statistics, Document 2007/9.

Triplett, J. E. (2004), *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products*, STI Working Paper 2004/9, Directorate for Science, Technology and Industry, Organisation for Economic Co-operation and Development, Paris.

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

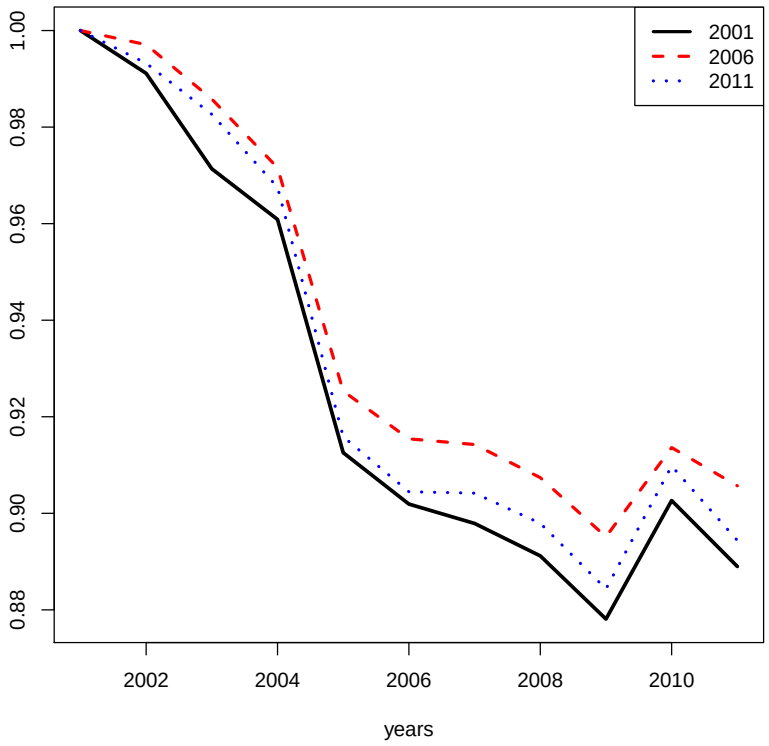
Wood, S. N. (2011), Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models, *Journal of the Royal Statistical Society B* 73(1), 3-36.

Figure 4: Evidence of Bias in the Postcode-Based Price Indexes (with 2001 Normalized to 1)



Note: Each curve measures the change in the value of the average location within postcodes 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 data set with no missing characteristics.

Figure 5: Evidence of Bias in the Residex-Region Based Price Indexes (with 2001 Normalized to 1)



Note: Each curve measures the change in the value of the average location of sold houses within regions over time, using a reference geospatial spline surface to make the comparison. Here the reference splines considered are those of 2001, 2006 and 2011. Again, irrespective of which spline is used, the value of the average location of sold houses declines over time. Again, to simplify matters we use only the splines derived from the restricted data set with no missing characteristics.

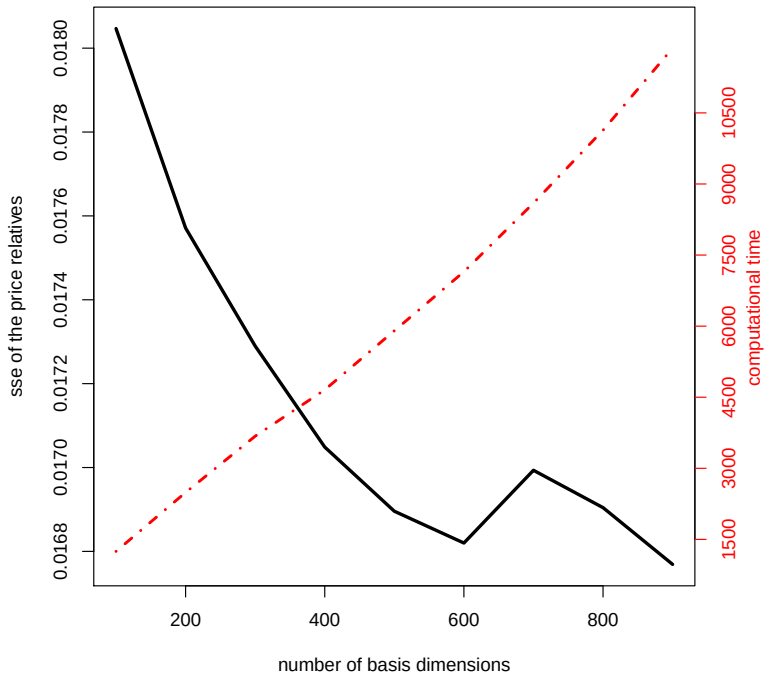


Figure 6: Sum of squared errors of the price relatives and computational time for different basis dimensions

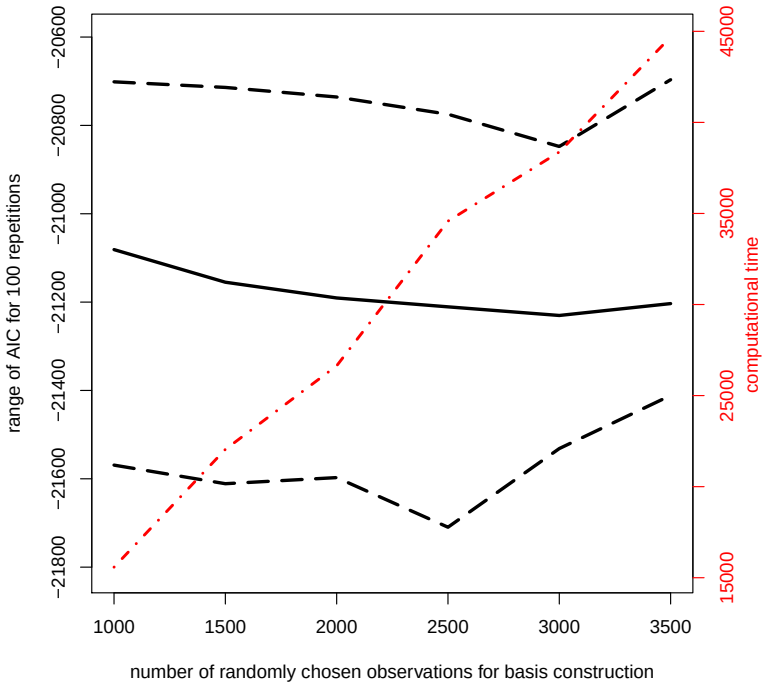


Figure 7: AIC distribution and computational time for 100 repetitions of the fit for the year 2011 based on different number of randomly chosen observations.

Graz Economics Papers

For full list see:
http://ideas.repec.org/s/grz/wpaper.html
Address: Department of Economics, University of Graz,
Universitätsstraße 15/F4, A-8010 Graz

07–2014 **Birgit Bednar–Friedl, Karl Farmer**: Existence and efficiency of stationary states in a renewable resource based OLG model with different harvest costs

06–2014 **Karl Farmer, Irina Ban**: Modeling financial integration, intra-EMU and Asian-US external imbalances

05–2014 **Robert J. Hill, Michael Scholz**: Incorporating Geospatial Data in House Price Indexes: A Hedonic Imputation Approach with Splines

04–2014 **Y. Hossein Farzin, Ronald Wendner**: The Time Path of the Saving Rate: Hyperbolic Discounting and Short-Term Planning

03–2014 **Robert J. Hill, Iqbal A. Syed**: Hedonic Price-Rent Ratios, User Cost, and Departures from Equilibrium in the Housing Market

02–2014 **Christian Gehrke**: Ricardo's Discovery of Comparative Advantage Revisited

01–2014 **Sabine Herrmann, Jörn Kleinert**: Lucas Paradox and Allocation Puzzle – Is the euro area different?

08–2013 **Christoph Zwick**: Current Account Adjustment in the Euro-Zone: Lessons from a Flexible-Price-Model

07–2013 **Karl Farmer**: Financial Integration and EMUs External Imbalances in a Two-Country OLG Model

06–2013 **Caroline Bayr, Miriam Steurer, Rose-Gerd Koboltschnig**: Scenario Planning for Cities using Cellular Automata Models: A Case Study

05–2013 **Y. Hossein Farzin, Ronald Wendner**: Saving Rate Dynamics in the Neoclassical Growth Model – Hyperbolic Discounting and Observational Equivalence

04–2013 **Maximilian Gödl, Jörn Kleinert**: Interest rate spreads in the Euro area: fundamentals or sentiments?

03–2013 **Christian Lininger**: Consumption-Based Approaches in International Climate Policy: An Analytical Evaluation of the Implications for Cost-Effectiveness, Carbon Leakage, and the International Income Distribution

02–2013 **Veronika Kulmer**: Promoting alternative, environmentally friendly passenger transport technologies: Directed technological change in a bottom-up/top-down CGE model

01–2013 **Paul Eckerstorfer, Ronald Wendner**: Asymmetric and Non-atmospheric Consumption Externalities, and Efficient Consumption Taxation

10–2012 **Michael Scholz, Stefan Sperlich, Jens Perch Nielsen**: Nonparametric prediction of stock returns with generated bond yields

09–2012 **Jörn Kleinert, Nico Zorell**: The export-magnification effect of offshoring

08–2012 **Robert J. Hill, Iqbal A. Syed**: Hedonic Price-Rent Ratios, User Cost, and Departures from Equilibrium in the Housing Market

07–2012 **Robert J. Hill, Iqbal A. Syed**: Accounting for Unrepresentative Products and Urban-Rural Price Differences in International Comparisons of Real Income: An Application to the Asia-Pacific Region

06–2012 **Karl Steininger, Christian Lininger, Susanne Droege, Dominic Roser, Luke Tomlinson**: Towards a Just and Cost-Effective Climate Policy: On the relevance and implications of deciding between a Production versus Consumption Based Approach

05–2012 **Miriam Steurer, Robert J. Hill, Markus Zahrnhofer, Christian Hartmann**: Modelling the Emergence of New Technologies using S-Curve Diffusion Models

04–2012 **Christian Groth, Ronald Wendner**: Embodied learning by investing and speed of convergence

03–2012 **Bettina Brüggemann, Jörn Kleinert, Esteban Prieto**: The Ideal Loan and the Patterns of Cross-Border Bank Lending

02–2012 **Michael Scholz, Jens Perch Nielsen, Stefan Sperlich**: Nonparametric prediction of stock returns guided by prior knowledge

01–2012 **Ronald Wendner**: Ramsey, Pigou, heterogenous agents, and non-atmospheric consumption externalities