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

Robert J. Hill and Michael Scholz
University of Graz
Austria
robert.hill@uni-graz.at michael.scholz@uni-graz.at

28-29 November 2013

EMG Workshop - Sydney

Introduction

- Houses differ both in their physical characteristics and location
- Exact longitude and latitude of each house are now increasingly available in housing data sets
- How can we incorporate geospatial data (i.e., longitudes and latitudes) in a hedonic model of the housing market?
    1. Distance to amenities (including the city center, nearest train station and shopping center, etc.) as additional characteristics.
    2. Spatial autoregressive models
    3. A spline function (or some other nonparametric function)

A Taxonomy of Methods for Computing Hedonic House Price Indexes

- Time dummy method

$$y = Z\beta + D\delta + \varepsilon \qquad P_t = \exp(\hat{\delta}_t)$$

where Z is a matrix of characteristics and D is a matrix of dummy variables.

- Average characteristics method

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

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

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

Average characteristics methods cannot use geospatial data, since averaging longitudes and latitudes makes no sense.

- Hedonic imputation method

$$\text{Paasche Single Imputation}: \; P_{t,t+1}^{PSI} = \prod_{h=1}^{H_{t+1}} \left[\left(\frac{p_{t+1,h}}{\hat{p}_{t,h}(z_{t+1,h})} \right)^{1/H_{t+1}} \right]$$

$$\text{Laspeyres Single Imputation}: \; P_{t,t+1}^{LSI} = \prod_{h=1}^{H_{t}} \left[\left(\frac{\hat{p}_{t+1,h}(z_{t,h})}{p_{t,h}} \right)^{1/H_{t}} \right]$$

$$\text{Fisher Single Imputation}: \; P_{t,t+1}^{FSI} = \sqrt{P_{t,t+1}^{PSI} \times P_{t,t+1}^{LSI}}$$

Our Models

(i) semilog with geospatial spline

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

(ii) semilog with postcode or region dummies

$$y = Z\beta + D\delta + \varepsilon$$

- y is a $H \times 1$ vector of log-prices
- Z is an $H \times C$ matrix of physical characteristics (including a constant and quarterly dummies)
- $g(z_{lat}, z_{long})$ is the geospatial spline function defined on the longitudes and latitudes
- D is a matrix of postcode or region dummies.

- The parameters to be estimated in (i) are the $C \times 1$ vector of characteristic shadow prices β, and the geospatial spline surface $g(z_{lat}, z_{long})$.
- The parameters to be estimated in (ii) are the $C \times 1$ vector of characteristic shadow prices β, and the $B \times 1$ vector of postcode or region shadow prices δ.
- We consider 15 regions and about 250 postcodes. So on average there are 16-17 postcodes in a region.

The Spline Function

- The **G**eneralized **A**dditive **M**odel (i) is estimated with a `thin plate regression spline`, an optimal low rank eigen approximation of a thin plate spline (which has *n* unknown parameters)
- A thin plate regression spline uses far fewer coefficients than a full spline, and is therefore computationally efficient, while losing little statistical performance.
- We can avoid certain problems in spline smoothing:
  - Choice of knot location
  - Basis functions useful for representing smooths of one predictor
  - It is not clear to what extent the bases are better or worse than any other basis that might be used.

The Spline Function (continued)

- For the approximation:
  - Choose degree of approximation (tradeoff: oversmoothing vs. computational burden)
  - Construct tprs basis from a randomly chosen sample of 2500 data points. The locations of these observations depend on the locational distribution of house sales in that period.
- The smoothing parameter is selected using Restricted Maximum Likelihood (REML).
- The spline is estimated using the **B**ig **A**dditive **M**odels function from the R package `mgcv`.
- Imputed prices at the boundary of the spline are a problem. They will tend to be too low (due to the beach premium).

Our Data Set

Sydney, Australia from 2001 to 2011.
Our characteristics are:

- Transaction price
- Exact date of sale
- Number of bedrooms
- Number of bathrooms
- Land area
- Postcode
- Longitude
- Latitude

Our Data Set (continued)

- Some characteristics are missing for some houses.
- There are more gaps in the data in the earlier years in our sample.
- We have a total of 454567 transactions.
- All characteristics are available for only 240142 of these transactions.

Dealing with Missing Characteristics

We impute the price of each house from the model below that has exactly the same mix of characteristics.

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

Comparing the Performance of Our Models

Table 1: Akaike information criterion (splines versus postcodes)

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
(i)	-55	-85	-1093	-1571	-7192	-6199	-8917	-10286	-15529	-14649	-18520
(ii)	4730	5337	5677	5571	8630	11677	16009	11564	12086	12307	8662

Table 2: Sum of squared log errors (splines versus postcodes)

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
(i)	0.056	0.056	0.049	0.048	0.042	0.046	0.044	0.040	0.038	0.037	0.034
(ii)	0.130	0.138	0.121	0.111	0.087	0.091	0.095	0.088	0.082	0.085	0.075

The sum of squared log errors is calculated as follows:

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

Results (continued)

The spline model significantly outperforms its postcode counterpart.

Repeat-Sales as a Benchmark

$$Z_h^{SI} = \text{Actual Price Relative} / \text{Imputed Price Relative}$$

$$Z_h^{SI} = \frac{p_{t+k,h}}{p_{th}} \Bigg/ \sqrt{\frac{p_{t+k,h}}{\hat{p}_{th}} \times \frac{\hat{p}_{t+k,h}}{p_{th}}} = \sqrt{\frac{p_{t+k,h}}{p_{th}} \Bigg/ \frac{\hat{p}_{t+k,h}}{\hat{p}_{th}}}$$

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

Results (continued)

Table 3: Sum of squared log price relative errors (spline versus postcodes)

Model	D^{SI}
(i)	0.016927
(ii)	0.036040

Spline again outperforms postcodes.

Price Indexes

- Restricted data set with no missing characteristics: Figure 1
- Full data set: Figure 2

Main Findings

- Prices rise most when locational effects are captured using a geospatial spline. While the gap between the spline and postcode based indexes is small, the gap is larger when region dummies are used.
- The median index is dramatically different when the full data set is used.
- The gap between spline and postcode/region hedonic price indexes is slightly smaller when the full data set is used.

Figure 1: Price Indexes Calculated on the Restricted Data Set

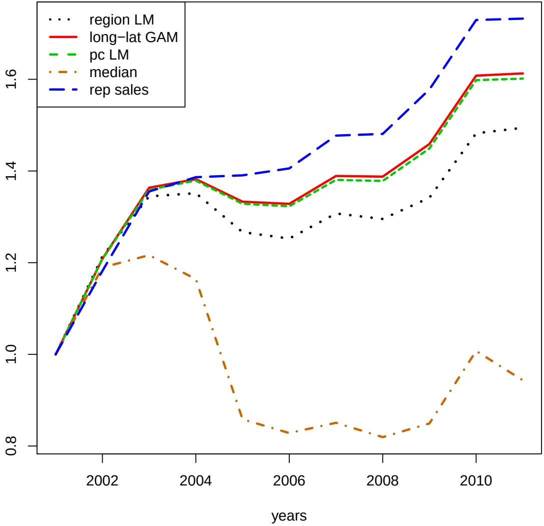
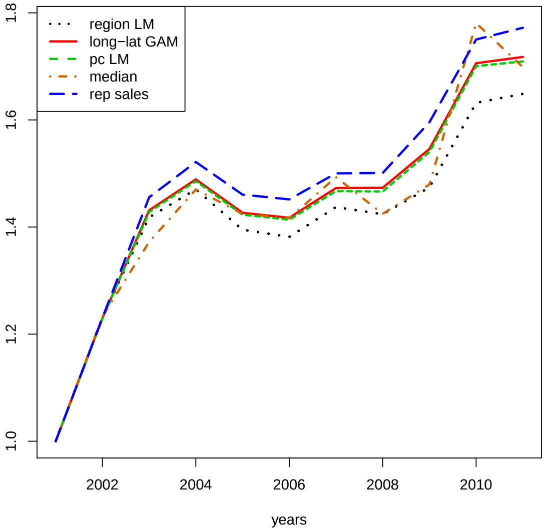


Figure 2: Price Indexes Calculated on the Full Data Set



Are Postcode/Region Based Indexes Downward Biased?

A downward bias can arise when the locations of sold houses in a postcode or region get worse over time.

We test for this as follows:

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 using the semilog model with spline of year 2001 the price of this average house in every location in which a house actually sold in 2001,. . .,2011 in that postcode
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.

Are Postcode Based Indexes Downward Biased? (continued)

Findings:

- The geometric means from step 6 fall over time irrespective of which year's spline is used as the reference.
- Most of the fall occurs in the first half of the sample.
- The fall is bigger for regions than postcodes.

Figure 3: Evidence of Bias in the Postcode-Based Price Indexes

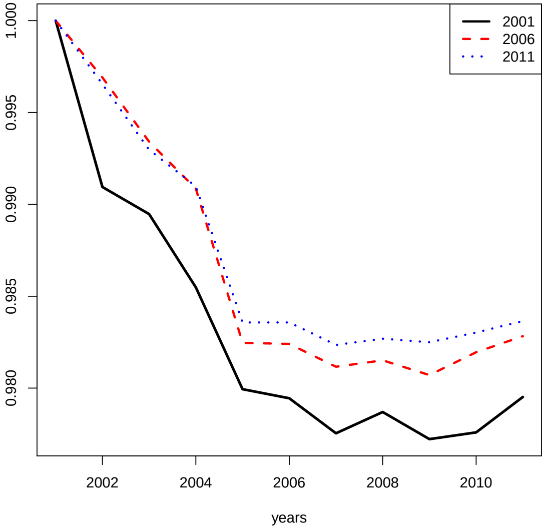
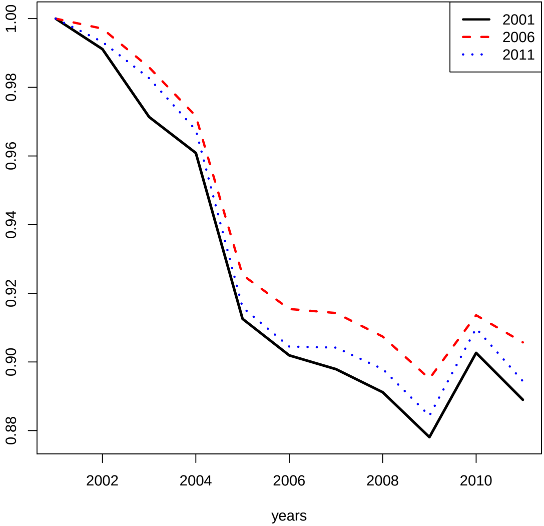


Figure 4: Evidence of Bias in the Region-Based Price Indexes



Conclusions

- Splines (or some other nonparametric method), when combined with the hedonic imputation method, provide a flexible way of incorporating geospatial data into a house price index
- In our data set postcode/region based indexes seem to have a downward bias since they fail to account for a general shift over time in houses sold to worse locations in each postcode/region. The bias is bigger for regions.