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

1 May 2013

Presentation to the Ottawa Group
Houses differ both in their physical characteristics and location

Exact longitude and latitude of each house are now increasingly included as variables 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

\[ y = Z\beta + D\delta + \varepsilon \quad P_t = \exp(\hat{\delta}_t) \]

where \( Z \) is a matrix of characteristics and \( D \) is a matrix of dummy variables.
Average characteristics method

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

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

where \( \bar{z}_{c,t} = \frac{1}{H_t} \sum_{h=1}^{H_t} z_{c,t,h} \) 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.
Imputation method

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

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

Fisher Single Imputation: $P_{t,t+1}^{FSI} = \sqrt{P_{t,t+1}^{PSI} \times P_{t,t+1}^{LSI}}$
Distance to Amenities as Additional Characteristics

- Throws away a lot of potentially useful information
- Distance from an amenity may impact on price in a nonmonotonic way
- Direction may matter as well (e.g., do you live under the flight path of an airport)?
Spatial autoregressive models

The SARAR(1,1) model takes the following form:

\[ y = \rho S y + X\beta + u, \]
\[ u = \lambda S u + \varepsilon, \]

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

The impact of location on house prices is captured by the parameters \( \rho \) and \( \lambda \).

SARAR models can be combined with either the time-dummy or hedonic imputation methods.
The limitations of the SAR(1) model are endless. These include: (1) the implausible and unnecessary normality assumption, (2) the fact that if $y_i$ depends on spatially lagged $y$s, it may also depend on spatially lagged $x$s, which potentially generates reflection-problem endogeneity concerns . . ., (3) the fact that the relationship may not be linear, and (4) the rather likely possibility that $u$ and $X$ are dependent because of, e.g., endogeneity and/or heteroskedasticity. Even if one were to leave aside all of these concerns, there remains the laughable notion that one can somehow know the entire spatial dependence structure up to a single unknown multiplicative coefficient [two unknown coefficients in the case of SARAR(1,1)]. (Pinkse and Slade 2010, p. 106 - text in square brackets added by the authors)
Our Models (estimated separately for each year)

(i) generalized additive model (GAM) with a geospatial spline

\[ y = c_1 + D\delta_1 + \sum_{c=1}^{C} f_{1,c}(z_c) + g_1(z_{lat}, z_{long}) + \varepsilon_1 \]

(ii) GAM with postcode dummies

\[ y = c_2 + D\delta_2 + \sum_{c=1}^{C} f_{2,c}(z_c) + m_2(z_{pc}) + \varepsilon_2 \]
(iii) semilog with geospatial spline

\[ y = c_3 + D\delta_3 + \sum_{c=1}^{C} z_c\beta_{3,c} + g_3(z_{lat}, z_{long}) + \varepsilon_3 \]

(iv) semilog with postcode dummies

\[ y = c_4 + D\delta_4 + \sum_{c=1}^{C} z_c\beta_{4,c} + \sum_{pc=1}^{250} z_{pc}m_{4,pc} + \varepsilon_4 \]
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.
We impute the price of each house from the model below that has exactly the same mix of characteristics.

(HM1): \( \ln \text{price} = f(\text{quarter dummy, land area, num bedrooms, num bathrooms, postcode}) \)
(HM2): \( \ln \text{price} = f(\text{quarter dummy, num bedrooms, num bathrooms, postcode}) \)
(HM3): \( \ln \text{price} = f(\text{quarter dummy, land area, num bathrooms, postcode}) \)
(HM4): \( \ln \text{price} = f(\text{quarter dummy, land area, num bedrooms, postcode}) \)
(HM5): \( \ln \text{price} = f(\text{quarter dummy, num bathrooms, postcode}) \)
(HM6): \( \ln \text{price} = f(\text{quarter dummy, num bedrooms, postcode}) \)
(HM7): \( \ln \text{price} = f(\text{quarter dummy, land area, postcode}) \)
(HM8): \( \ln \text{price} = f(\text{quarter dummy, postcode}) \)
Comparing the Performance of Our Models

Table 1: Akaike information criterion for models 1-4

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>416</td>
<td>89</td>
<td>-778</td>
<td>-1599</td>
<td>-7290</td>
<td>-6417</td>
<td>-8544</td>
<td>-10271</td>
<td>-14059</td>
<td>-14953</td>
<td>-18493</td>
</tr>
<tr>
<td>2</td>
<td>4888</td>
<td>5456</td>
<td>5780</td>
<td>5598</td>
<td>8635</td>
<td>11678</td>
<td>16233</td>
<td>11652</td>
<td>12819</td>
<td>12313</td>
<td>8696</td>
</tr>
<tr>
<td>3</td>
<td>-55</td>
<td>-85</td>
<td>-1093</td>
<td>-1571</td>
<td>-7192</td>
<td>-6199</td>
<td>-8917</td>
<td>-10286</td>
<td>-15529</td>
<td>-14649</td>
<td>-18520</td>
</tr>
<tr>
<td>4</td>
<td>4730</td>
<td>5337</td>
<td>5677</td>
<td>5571</td>
<td>8630</td>
<td>11677</td>
<td>16009</td>
<td>11564</td>
<td>12086</td>
<td>12307</td>
<td>8662</td>
</tr>
</tbody>
</table>

Table 2: Sum of squared log errors for models 1-4

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.061</td>
<td>0.057</td>
<td>0.051</td>
<td>0.047</td>
<td>0.041</td>
<td>0.046</td>
<td>0.045</td>
<td>0.040</td>
<td>0.039</td>
<td>0.037</td>
<td>0.034</td>
</tr>
<tr>
<td>2</td>
<td>0.133</td>
<td>0.140</td>
<td>0.123</td>
<td>0.111</td>
<td>0.087</td>
<td>0.091</td>
<td>0.096</td>
<td>0.089</td>
<td>0.084</td>
<td>0.085</td>
<td>0.076</td>
</tr>
<tr>
<td>3</td>
<td>0.056</td>
<td>0.056</td>
<td>0.049</td>
<td>0.048</td>
<td>0.042</td>
<td>0.046</td>
<td>0.044</td>
<td>0.040</td>
<td>0.038</td>
<td>0.037</td>
<td>0.034</td>
</tr>
<tr>
<td>4</td>
<td>0.130</td>
<td>0.138</td>
<td>0.121</td>
<td>0.111</td>
<td>0.087</td>
<td>0.091</td>
<td>0.095</td>
<td>0.088</td>
<td>0.082</td>
<td>0.085</td>
<td>0.075</td>
</tr>
</tbody>
</table>

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 models significantly outperform their postcode counterparts.
▶ The GAM outperforms its semilog counterpart

Repeat-Sales as a Benchmark

\[ Z_{h}^{SI} = \frac{p_{t+k,h}}{p_{th}} \times \frac{\hat{p}_{t+k,h}}{p_{th}} = \sqrt{\frac{p_{t+k,h}}{p_{th}} / \frac{\hat{p}_{t+k,h}}{\hat{p}_{th}}} \]

\[ Z_{h}^{SI} = \text{Actual Price Relative} / \text{Imputed Price Relative} \]
\[ D^{SI} = \left( \frac{1}{H} \right) \mathop{\sum}_{h=1}^{H} [\ln(Z_{h}^{SI})]^2. \]

Table 3: Sum of squared log price relative errors for models 1-4

<table>
<thead>
<tr>
<th>Model</th>
<th>(D^{SI})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-GAM spline</td>
<td>0.017467</td>
</tr>
<tr>
<td>2-GAM postcode</td>
<td>0.020900</td>
</tr>
<tr>
<td>3-semilog spline</td>
<td>0.016927</td>
</tr>
<tr>
<td>4-semilog postcode</td>
<td>0.036040</td>
</tr>
</tbody>
</table>

Spline outperforms postcodes.
Surprisingly, semilog spline outperforms GAM spline.
Price Indexes

- Restricted data set with no missing characteristics: Figures 1 and 2
- Full data set: Figures 3 and 4

Main Findings

- The mean and median indexes are dramatically different when the full data set is used.
- Prices rise more when geospatial data is used instead of postcodes.
- The gap is slightly smaller when the full data set is used. It is also smaller for GAM than for semilog.
Figure 1: GAM on restricted data set

SIF for post code and long/lat

- post code
- long/lat
- median price
- mean price

years

Figure 2: Semilog on restricted data set

SIF for post code and long/lat partlin

years

SIF

post code
long/lat
median price
mean price
Figure 3: GAM on full data set

SIF for post code and long/lat

years

post code
long/lat
median price
mean price


Hill and Scholz
Ottawa Group 2013
**Figure 4**: Semilog on full data set

SIF for post code and long/lat

```
years
```

(source: Hill and Scholz, Ottawa Group 2013)
Are Postcode Based Indexes Downward Biased?

A downward bias can arise when the locations of sold houses in a postcode get worse over time. We test for this as follows:

- Choose a postcode
- Calculate the mean number of bedrooms, bathrooms, land area and quarter of sale over the 11 years for that postcode.
- Impute using the semilog model with spline of year \( t \) (where \( t \) could be 2001, \ldots, 2011) the price of this average house in every location in which a house actually sold in 2001, \ldots, 2011 in that postcode.
- Take the geometric mean of these imputed prices for each year.
- Repeat for another postcode.
- Take the geometric mean across postcodes in each year.
Questions:

- Does this geometric mean rise or fall over time?
- How much difference does it make which year’s semilog model is used to impute prices?
Conclusions

- Splines (or some other nonparametric method) provide a flexible way of incorporating geospatial data into a house price index.
- Switching from postcodes to geospatial data can have a big impact. Between 2001 and 2011 house prices rose by 60 percent based on geospatial data as compared with only 40 percent based on postcodes.
- In our data set postcode 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.