Weekly Hedonic House Price Indices: An Imputation Approach from a Spatio-Temporal Model

Robert J. Hill\textsuperscript{1}, Alicia N. Rambaldi\textsuperscript{2} and Michael Scholz\textsuperscript{1}

\textsuperscript{1}University of Graz, Austria, \textsuperscript{2}The University of Queensland, Australia

15th Meeting of the Ottawa Group 2017. Deutsche Bundesbank
Outline

Introduction and Background

Hedonic Imputation

Our Work
  Model
  Estimation and Prediction
  Quality of the Index

Empirical Example

Conclusions

References
Residential Property Price Indices

- **Repeat Sales**: Assume hedonics are constant over time - Change in log price of repeat sales pair depends on dummy. Parameters of dummies give index
  - Standard and Poor’s/Case-Shiller Home Price Indices in the US

- **Hedonic Based**
  - **Time-Dummy Method**: Assume hedonics are constant over time - log-linear model with time dummies. Index is given by exponentiation of time dummy parameters
  - **Hedonic Imputation Method**: Hedonics can change over time - predictions from model provide imputed price relatives to enter index formula
  - Most European Countries use hedonic methods ((EuroSTAT, 2016))

- **Hybrid**: Assume hedonics are constant over time. Combines Repeat Sales and Time-Dummy Method

- **Others**: Stratification or Mix Adjustment, Appraisal based (SPAR)

- **Recent Summary of all methods**:
    DOI:10.1787/9789264197183-en
  - Hill, R.J. (2013) in *Journal of Economic Surveys*
Brief and Incomplete Literature

- **Repeat Sales:**
  - Bailey, Muth and Nourse (1963), generalisation of Wyngarden (1927) and Wenzlick (1952), Case and Shiller (1987; 1989)

- **Hedonic**
    - Silver and Heravi (2007) *JBES* derive the formal difference between TD and HI and show HI are grounded in index number theory and preferred over the constrained TD

**Hedonic Imputation Indices**

- Double imputation: Laspeyres index (DIL), Paasche index (DIP), and Törnqvist index (DIT) are defined as follows:

  \[
  P_{DIL}^{t,t+1} = \prod_{i=1}^{N_t} \left[ \left( \frac{\hat{p}_{i,t+1}(x'_{i,t})}{\hat{p}_{i,t}(x_{i,t})} \right) \right]^{1/N_t}
  \]

  \[
  P_{DIP}^{t,t+1} = \prod_{i=1}^{N_{t+1}} \left[ \left( \frac{\hat{p}_{i,t+1}(x'_{i,t+1})}{\hat{p}_{i,t}(x'_{i,t+1})} \right) \right]^{1/N_{t+1}}
  \]

  \[
  P_{DIT}^{t,t+1} = \sqrt{P_{DIP}^{t,t+1} \times P_{DIL}^{t,t+1}}
  \]

  \[i = 1, \ldots, N_t\] indices the dwellings sold in period \(t\), \(i = 1, \ldots, N_{t+1}\) indices the dwellings sold in period \(t+1\).

  The overall price index is then constructed by chaining together these bilateral comparisons between adjacent periods.

- Single imputation uses \(p_{i,t}(x'_{i,t})\) and \(p_{i,t+1}(x'_{i,t+1})\) instead of predicted, \(\hat{p}_{i,t}(x_{i,t})\) and \(\hat{p}_{i,t+1}(x_{i,t+1})\), in the DIL and DIP formulae.

- A model is required to provide the predictions and imputations to construct the matching sample.
HI Index Frequency and Modelling

- HI indices at **annual or quarterly frequency** are typically constructed using hedonic models estimated period-by-period (mostly by OLS)
  - Controls for characteristics (land and structure) and location are included
  - Hill and Scholz (2017) using a Generalised Additive Model (semi-parametric) - annual

- HI indices at **monthly frequency**
  - Thin market periods can lead to index chain drift (small sample and composition of sales influence parameter estimates)
  - Rambaldi and Fletcher (2014) find evidence of chain drift when comparing the indices from a model estimated using two-adjacent period (two months) rolling window to one using filter estimates of the parameters from a state-space model.

- **This paper: HI index at weekly frequency**
  - Builds from the work of Wikle and Cressie (1999) and Rambaldi and Fletcher (2014)
We develop a spatio-temporal model to obtain the imputed prices.

- Advantage: Link the parameters over time without leading to index revision.

- A geospatial spline surface controls for location and is obtained using only current period information
  - is embedded in a state-space formulation that controls for trends and property quality.

- The spatio-temporal specification leads to:
  - a modified form of the Kalman filter, and
  - a Goldberger’s adjusted form of the predictor to obtain the imputations.

- Use a criterion based on price relatives to evaluate the index against two competing hedonic imputation methods and the repeat-sales method.
The model

- The objective:
  - estimate $y_{it}^*$, a smoother and quality adjusted, but unobservable, $y_{it} = \ln \text{price}_{it}$ of property $i$.
  - At any $t$ $N_t$ properties are sold, $t = 1, \ldots, T$, $\sum_{t=1}^{T} N_t = N$
  - We write this model as
    \[ y_{it} = y_{it}^* + \epsilon_{it}; \epsilon_{it} \sim N(0, \sigma^2) \] (1)
  - $\epsilon_{it}$ is not correlated across location or time and captures overall measurement error.
  - At (any) given time period $\tau$, the vector with elements $y_{i\tau}^*$ is given by
    \[ y_{\tau}^* = x_{\tau}^\dagger + v_{\tau}; v_{\tau} \sim N(0, V_{\tau}) \]
  - where, $v_{\tau}$ is a (vector) random error that does not have a temporally dynamic structure but might have some spatial structure and thus $V_{\tau}$ might not be diagonal. It is assumed that $E(v_{i\tau} \epsilon_{j}) = 0$ for all $i, j = 1, \ldots, N$ and $-\infty \leq t \leq \infty$. 

The model (cont)

- $x_{it}^+$ is assumed to evolve according to three components, trend, property quality and location,

$$x_{it}^+ = \mu_t + \sum_{k=1}^{K} \beta_{k,t} z_{k, it} + \gamma_t g_{it}(z_{\text{long}}, z_{\text{lat}})$$

where,

- $\mu_t$ is a trend component common to all $i$ in period $t$ and captures overall macroeconomic conditions that affect all locations in the market under study;
- $z_{k, it}$ is the $k\text{th}$ hedonic characteristic from a set of $K$ providing information on the type/quality of the property (e.g., number of bedrooms, bathrooms, size of the lot). These are not trending variables.
- $g_{it}(z_{\text{long}}, z_{\text{lat}})$ is a measure of the location of property $i$ defined on a continuous surface at time period $t$. It is not a function of time.
- $\beta_{k,t}$ and $\gamma_t$ are parameters to be estimated
- $E(z_k v_t) = 0, E(z_k \epsilon_t) = 0$ for all $k = 1, \ldots, K$, $E(g_{it} v_{jt}) = 0, E(g_{it} \epsilon_{jt}) = 0$, for all $i, j$. 


A few key points

- \( \hat{g}_{it}(z_{\text{long}}, z_{\text{lat}}) \) is obtained at each time period from those properties that have sold that period.

- \( \gamma_t \), in (2), provides flexibility. \( \gamma_t \neq 1 \rightarrow \hat{g}_{it}(z_{\text{long}}, z_{\text{lat}}) \) will be shifted by temporal market information up to time \( t \).

- The combination of spatial and temporal information leads to two unconventional features of this model:
  - The error has two components, \( \epsilon_t \), the overall measurement error, and \( v_t \) arising from predicting the (log) sale price using only the spatial variability within each time period.
  - \( \hat{g}_{it}() \) for property \( i \) sold in period \( t \) will not be identical in value if property \( i \) is priced in a different time period.
    - \( \hat{g}_{t(t)}(z_{\text{long}}, z_{\text{lat}}) \) the vector of spline values for properties sold and priced in period \( t \)
    - \( \hat{g}_{t(t-1)}(z_{\text{long}}, z_{\text{lat}}) \) the vector of the set of properties sold in \( t \) when priced in \( t - 1 \).
State-Space Form

\[ y_t = X_t \alpha_t + v_t + \epsilon_t; \epsilon_t \sim N(0, H) \quad (3) \]
\[ \alpha_t = D \alpha_{t-1} + \eta_t; \eta_t \sim N(0, Q) \quad (4) \]

- \( X_t \) is \( N_t \times (K + 2) \) and with the \( i \)th row being
  \[ x_{it}' = \{1, z_{1, it}, \ldots, z_{K, it}, g_{it}(z_{long}, z_{lat})\} \]

- \( y_t = \ln(\text{price}_t)\) \( i \) sold in \( t \).

- \( H = \sigma_\epsilon^2 I_{N_t} \)

- \( \alpha_t = \{\mu_t, \beta_{1t}, \ldots, \beta_{K, t}, \gamma_t\}' \)

- \( D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & I_K & 0 \\ 0 & 0 & \rho \end{bmatrix}; 0 \leq \rho \leq 1; \)
  - If \( \rho < 1 \) \( \gamma_t \) is mean reverting.
  - If \( \rho = 1 \), \( \gamma_t \) evolves as a random walk as do the other state parameters in the model.

- \( Q = \begin{bmatrix} \sigma_\mu^2 & 0 & 0 \\ 0 & \sigma_\beta^2 I_K & 0 \\ 0 & 0 & \sigma_g^2 \end{bmatrix} \)
Estimator of $\alpha_{t|t}$ (estimates of quantities in red required)

- The state at time $t$ given information up to and including

$$\hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + G_t \{y_t - X_t^1 \hat{\alpha}_{t|t-1}\}$$ (5)

$$P_{t|t} = P_{t|t-1} - G_t X_t P_{t|t-1}$$

$X_t^1$ which is the $X_t$ matrix with the $\hat{g}_{i,t}(t)$ replaced by $\hat{g}_{i,t(t-1)}(z_{long}, z_{lat})$.

$P_{t|t}$ is the mean square error matrix given information up to time period $t$.

- The Kalman gain under the assumptions already stated

$$G_t = P_{t|t-1} X_t' \{H + V_t + X_t P_{t|t-1} X_t'\}^{-1}$$

- The updating equations are given by

$$\hat{\alpha}_{t|t-1} = D \hat{\alpha}_{t-1|t-1}$$ (6)

$$P_{t|t-1} = D P_{t-1|t-1} D' + Q$$ (7)
\( \hat{g}_t() \) and \( \hat{V}_t \)

- Hill and Scholz (2017) period-by-period semi-parametric model (Generalised Additive Model (GAM))

\[
y_{it} = \theta_{0t} + z'_{it} \theta^\dagger_t + g_{i,t}(z_{long}, z_{lat}) + v_{it}
\]  

(8)

- \( \theta^\dagger_t = \{\theta_{1t}, \ldots, \theta_{K,t}\}' \)

- predicted (log) prices, \( \hat{x}^\dagger_t \), is obtained from (8) based on observed \( z_k \), \( k = 1, \ldots, K \), and

- estimates of \( g_{i,t}(z_{long}, z_{lat}) \) and \( \theta^\dagger_t \)

- estimate of \( \nu_{it} \), \( \hat{\nu}_{it} = y_t - \hat{x}^\dagger_t \to \hat{V}_t = \frac{1}{N_t} \sum_i \hat{\nu}_{it}^2 \)

- Estimator (penalized likelihood approach (see Wood 2006 and the references therein))
  
  - based on a transformation and truncation of the basis that arises from the solution of the thin plate spline smoothing problem.
  - is computationally efficient and avoids the problem of choosing the location of knots
  - the penalized likelihood maximization problem is solved by Penalized Iteratively Reweighted Least Squares (P-IRLS) - Wood (2011)
\[ \hat{H}, \hat{D}, \hat{Q} \text{ and } \hat{\alpha}_{t\mid t} \]

- \( \hat{\rho} \) enters \( D \), \( \hat{\sigma}_\varepsilon^2 \) defines \( H \), and \( \hat{\sigma}_\mu^2, \hat{\sigma}_\beta^2 \) and \( \hat{\sigma}_\gamma^2 \) enter \( Q \), once known the Kalman filter algorithm gives \( \hat{\alpha}_{t\mid t} \)

- under assumptions stated the log-likelihood \( \ln L \) in predictive form:

\[
\ln L(\rho, \sigma_\varepsilon^2, \sigma_\beta^2, \sigma_\gamma^2; y_t, Y_{t-1}, Z_t, \hat{g}_{t(t-1)}) =
\]

\[
- \frac{N}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=d}^T \ln |F_t| - \frac{1}{2} \sum_{t=d}^T \nu'_t|t-1 F_t^{-1} \nu_t|t-1
\]

- \( N = \sum_{t=d}^T N_t \); \( d \) is sufficiently large to avoid the log-likelihood being dominated by the initial condition, \( \alpha_0 \sim N(a_0, P_0) \)

- \( \nu_{t\mid t-1} = y_t - X_t^1 \hat{\alpha}_{t\mid t-1} \), and its variance-covariance,

\[
F_t = E(\nu_t|t-1 \nu'_t|t-1) = H + V_t + X_t P_{t|t-1} X'_t \text{ outputs of running the Kalman Filter}
\]

- We use grid search over \( \rho \) and Newton-Raphson algorithm over the other four parameters
Prediction

- given assumptions already stated plus $v_{it}$ and $y_t$ have a joint multivariate normal distribution, the prediction of the log price for property $h$,

$$\hat{y}_{t|t,h} = x'_{t,h} \alpha_{t|t} + c'_{vt,h} \Omega^{-1} e_t$$

- $\Omega = \text{cov}\{y_t, y_t\}$
- $c'_{vt,h} = E(v_{ht}, v_t)$ is the row of $V_t$ corresponding to property $h$ and has elements $c_{v,hj} \equiv E\{v_{ht}v_{jt}\}$ which could be equal to zero for $h \neq j$.
- $e_t = y_t - E(y_t)$; $\hat{e}_t = y_t - X_t \hat{\alpha}_{t|t}$;

- The prediction of the price of property $h$ sold in period $t$ for period $t$

$$\hat{p}_{t,h}(z'_{t,h}, \hat{g}_{h,t(t)}) = \exp(\hat{y}_{t|t,h})$$ (9)

- The imputation of the price of property $h$ sold in period $t$ for period $t-1$ is given by

$$\hat{p}_{t-1,h}(z'_{t,h}, \hat{g}_{h,t(t-1)}) = \exp(x'_{t,h} \alpha_{t-1|t-1} + c'_{(t-1),h} \Omega^{-1} e_{t(t-1)})$$ (10)

- plugging in estimates of the $\alpha_{t|t}$, $\Omega$, $c'_{vt}$ and $e_t$ allows implementation.
Measuring the quality of the index

- The building blocks of the index
  - Laspeyres-type index are the imputed price relatives \( \hat{p}_{i,t+1}(x'_{i,t})/\hat{p}_{i,t}(x'_{i,t}) \).
  - Paasche-type index are the imputed price relatives \( \hat{p}_{i,t+1}(x'_{i,t+1})/\hat{p}_{i,t}(x'_{i,t+1}) \).
- Hence the performance of the index depends on the quality of these imputed price relatives.

- Sample of repeat-sale dwellings are indexed by \( i = 1, \ldots, H_{RS} \).
- Define the ratio of imputed to actual price relative for house \( i \):

\[
V_i = \frac{\hat{p}_{i,t+k}}{\hat{p}_{i,t}} \frac{p_{i,t+k}}{p_{i,t}}
\]

(11)

- Our quality measure is

\[
D = \left( \frac{1}{H_{RS}} \right)^{H_{RS}} \sum_{i=1}^{H_{RS}} [\ln(V_i)]^2
\]

(12)

- where the summation in (12) takes place across the whole repeat-sales sample.
Measuring the quality of the index (cont.)

- To avoid "lemon" bias: starter homes sell more frequently as people upgrade as their wealth rises (Clapp and Giaccotto (1992), Gatzlaff and Haurin (1997), and Shimizu, Nishimura and Watanabe (Shimizu et al. (2010)))

- Adjust

\[
V_{i}^{adj} = V_{i} \left[ \left( \frac{P_{t+k}^{RS}}{P_{t}^{RS}} \right) \right] / \left( \frac{P_{t+k}^{Hed}}{P_{t}^{Hed}} \right)
\]

\[
D^{adj} = \left( \frac{1}{H_{RS}} \right) \sum_{i=1}^{H_{RS}} [\ln(V_{i}^{adj})]^2.
\]

- \(P_{t+k}^{RS}/P_{t}^{RS}\) change in the repeat-sales price index between \(t\) and \(t + k\)
- \(P_{t+k}^{Hed}/P_{t}^{Hed}\) change in a HI price index between \(t\) and \(t + k\)
The Data

- Sydney (Australia) for the years 2001- 2014.
- Hedonic 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.
- The quality of the data improves over time. In particular, missing characteristics are quite common in the first two years (i.e., 2001 and 2002).
  - We use the full sample period for estimation of the state space model.
  - We present the hedonic indices starting in 2003.
- Hedonic Imputed Indices Computed
  - GAM is based on periodwise estimation of the semiparametric model;
  - SS+GAM based on the spatio-temporal model;
  - SS+PC based on a semilog hedonic model with postcodes dummies estimated as a state-space.
- Other Indices
  - Repeat Sales: Bailey, Muth, and Nourse (1963) formula
  - Median: From observed prices
Number of Transactions per Week

![Graph showing the number of transactions per week from 2002 to 2014. The x-axis represents the years, and the y-axis represents the number of transactions ranging from 0 to 1,200. The graph shows fluctuations in transaction numbers over the years.]
Results (cont.)

Table: Index quality based on $D$ and $D^{adj}$ criteria (2003-2014)

<table>
<thead>
<tr>
<th></th>
<th>$D$</th>
<th>$D_{GAM}^{adj}$</th>
<th>$D_{SS+GAM}^{adj}$</th>
<th>$D_{SS+PC}^{adj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAM</td>
<td>0.0233</td>
<td>0.0272</td>
<td>0.0313</td>
<td>0.0230</td>
</tr>
<tr>
<td>SS+GAM</td>
<td><strong>0.0102</strong></td>
<td><strong>0.0096</strong></td>
<td><strong>0.0099</strong></td>
<td><strong>0.0133</strong></td>
</tr>
<tr>
<td>SS+PC</td>
<td>0.0246</td>
<td>0.0279</td>
<td>0.0320</td>
<td>0.0240</td>
</tr>
</tbody>
</table>

GAM is based on periodwise estimation of the semiparametric model;
SS+GAM is the spatio-temporal model;
SS+PC is the state space model applied to the semilog model with location effects captured using postcodes.

$D_{GAM}^{adj}$ refers to the adjusted $D$ criteria with lemons bias corrected for using the GAM hedonic price index as the adjustment factor.

$D_{SS+GAM}^{adj}$ and $D_{SS+PC}^{adj}$ use the SS+GAM and SS+PC hedonic price indices, respectively as the adjustment factors.

▶ The differences are statistically significant
Conclusions

- The hedonic imputation method provides a flexible way of constructing quality-adjusted house price indices using a matching sample approach.
- We develop a spatio-temporal model to obtain the imputed prices.
- A geospatial spline surface controls for location and is embedded in a state-space formulation that controls for trends and property quality.
- The spatio-temporal specification leads to a modified form of the Kalman filter and a Goldberger’s adjusted form of the predictor to obtain the imputations.
- Using a criterion proposed by HS it is shown that embedding a semi-parametric model with geospatial spline surface in a state-space model generates house price indices that outperform two competing hedonic imputation methods and the repeat-sales method.


Subcommittee on Economic Statistics of the Joint Economic Committee, 87th Congress.


