Using Kalman Filters and Geospatial Splines to Improve House Price Indexes: A Hedonic Imputation Approach

Michael Scholz
University of Graz, michael.scholz@uni-graz.at

Robert J. Hill
University of Graz, robert.hill@uni-graz.at

Alicia Rambaldi
University of Queensland, a.rambaldi@uq.edu.au

Follow this and additional works at: http://repository.cmu.edu/sem_conf

Part of the Economics Commons
Using Kalman Filters and Geospatial Splines to Improve House Price Indexes: A Hedonic Imputation Approach

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

\textsuperscript{1}Department of Economics
University of Graz, Austria

\textsuperscript{2}School of Economics
UQ, Australia

SEM2015
Paris, July 2015

* Supported by funds of the Oesterreichische Nationalbank
(Anniversary Fund, project number: 14947)
Overview of the talk:

1. Methodology for Residential Property Price Indexes (House Price Indexes – HPI):
   - Geospatial Data and Hedonic Imputation
   - Repeat-Sales as a (Bias-Corrected) Performance Measure

2. Model Setup:
   - Period-wise GAM with Spline vs. OLS with Postcode Dummies
   - Improved HPIs Due to Interrelation Over Time: Kalman-Filter

3. Empirical Study: Data Set and Main Findings

4. Conclusion and Outlook
Motivation I

- **Sensitivity** of HPIs to the method of construction: A potential source of confusion.

Known **difficulties**: HPIs have to be compiled from *infrequent* transactions on *heterogeneous* properties (*physical characteristics* and *location*).

- A key determinant of house prices is *location*. It means that houses which are identical in their structural characteristics vary in value due to their location.

Methods for the incorporation of location in HPIs:

- **Distance to amenities** (city center or nearest train station, etc.) and **Dummy-Variables** (postcodes or school districts, etc.) for each house.
- **Spatial autoregressive models**
- A nonparametric function (e.g. a *spline function* on *longitudes* and *latitudes*)
Hedonic imputation approach for computing the HPI:

- **Estimate period-wise the hedonic regression**
  \[
  \log(p_{t,h}) = f_t(z_{t,h}) + \varepsilon_{t,h}
  \] (1)

- **Impute prices** for each house: Insert its particular mix of characteristics $z_{t,h}$ into the estimated hedonic model (1).

- **Double imputation** Fisher price index formula gives the HPI:
  \[
  P_{t,t+1}^{PDI} = \prod_{h=1}^{H_{t+1}} \left[ \left( \frac{\hat{p}_{t+1,h}(z_{t+1,h})}{\hat{p}_{t,h}(z_{t+1,h})} \right)^{1/H_{t+1}} \right] \\
  \text{and} \\
  P_{t,t+1}^{LDI} = \prod_{h=1}^{H_t} \left[ \left( \frac{\hat{p}_{t+1,h}(z_{t,h})}{\hat{p}_{t,h}(z_{t,h})} \right)^{1/H_t} \right] \\
  \Rightarrow \\
  P_{t,t+1}^{FDI} = \sqrt{P_{t,t+1}^{PDI} \times P_{t,t+1}^{LDI}}
  \] (2)
We consider two different period-wise hedonic regressions.

- **Model 1 (M1):** Linear Model with Postcode-Dummies:
  \[
  \log(p) = Z\beta_t + D\delta_t + \varepsilon
  \]  
  **Advantage:** Easily estimated with standard software (OLS)  
  **Problems:** Can produce downward biased HPIs (see, e.g., Hill and Scholz (2015)). Computational Problems for shorter periods.

- **Model 2 (M2):** Generalized Additive Model (GAM)
  \[
  \log(p) = Z\beta_t + g_t(z_{lat}, z_{long}) + \varepsilon
  \]  
  **Advantage:** Fully flexible nonparametric part (unknown function \(g_t\)) defined on the geospatial data (latitude \(z_{lat}\) and longitude \(z_{long}\))

For both models (especially for shorter periods): **Interrelation over time?**  
⇒ State space models and the **Kalman-Filter**
• Period-wise estimated surfaces of **M2** for the median house on quarterly data (left: 2005Q1, right: 2011Q4)

• Analog Figures for **M1** possible (no smooth curves but bars defined over postcodes instead).
Our Models I

- **M1**: Linear Model with Postcode-Dummies:

\[
\log(p) = Z\beta_t + D\delta_t + \varepsilon
\]  

(5)

- **M2**: Generalized Additive Model (GAM)

\[
\log(p) = Z\beta_t + g_t(z_{lat}, z_{long}) + \varepsilon
\]  

(6)

- Components:
  - \(\log(p)\) is a \((H_t \times 1)\)-vector of log-prices
  - \(Z\) is an \((H_t \times C)\)-matrix of physical characteristics (including a constant)
  - \(D\) is a matrix of postcode dummies
  - \(g_t(z_{lat}, z_{long})\) is the geospatial spline function
  - \(\beta_t\) is a \((C \times 1)\)-vector of characteristic shadow prices
  - \(\delta_t\) is a \((B \times 1)\)-vector of postcode shadow prices
Our Models II

Estimation:

- Model **M1** is estimated using **OLS** (**M1(OLS)**) and the **Kalman-Filter** (**M1(KF)**).
- For the Kalman-Filter, we use **Matlab** for **Maximum-Likelihood** estimation of the unknown parameters together with **Newton-Raphson** type algorithms.
- Model **M2** is estimated using **Penalized Least Squares** (**M2(PLS)**). The application of the Kalman-Filter for this case remains for further research.
- The semiparametric model is estimated using the **bam**-function (Big Additive Models) in the **mgcv** package in **R**.
- For the nonparametric part we use an **approximation to a thin plate spline** (TPS) (developed by S. Wood, 2003), so called **thin plate regression splines** (TPRS), defined on the longitudes and latitudes.
- The smoothing parameters are estimated via **Restricted Maximum Likelihood** (REML).
The data

- Sydney, Australia from 2001 to 2011 (but provide indexes starting Jan2005).
- Our **characteristics** are:
  - Transaction price + exact date of sale
  - **Physical**: Number of bedrooms, Number of bathrooms, Land area
  - **Location**: Residex-Region (16), Postcode (242), Longitude + Latitude
- Some (physical) 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 454,567 transactions.
- All characteristics are available for only 240,142 of these transactions.
Price Indexes Calculated on the Restricted Data Set I

The graph shows the calculated price indexes over time. The following models are represented:
- M1 (OLS)
- M1 (KF)
- M2 (PLS)
- Repeat-Sales
- Median

The data is plotted from January 2005 to January 2012.
Price Indexes Calculated on the Restricted Data Set II

M1(OLS)  M1(KF)  M2(PLS)  Repeat-Sales

Jan05  Jan06  Jan07  Jan08  Jan09  Jan10  Jan11  Jan12
Different performance measures:

- **Sum of squared log errors**

\[
SSLE = \frac{1}{H} \sum_{h=1}^{H} \left[ \ln(\hat{p}_h/p_h) \right]^2
\]

- **Repeat-Sales as a benchmark**

\[
V_h = \frac{p_{t+k,h}}{p_{th}} \sqrt{\frac{p_{t+k,h}}{\hat{p}_{th}} \times \frac{\hat{p}_{t+k,h}}{p_{th}}} = \sqrt{\frac{p_{t+k,h}}{p_{th}} \left[ \frac{\hat{p}_{t+k,h}}{\hat{p}_{th}} \right]}
\]

\[
D = \frac{1}{H} \sum_{h=1}^{H} \left[ \ln(V_h) \right]^2
\]
Potential problem:

- A repeat-sales sample may have a **lemons bias**.
- Starter homes sell more frequently as a result of people upgrading as their wealth rises.

Bias correction:

\[
V_h^* = \left[ \left( \frac{P_{t+k}^{Hed}}{P_t^{Hed}} \right) \left( \frac{P_{t+k}^{RSI}}{P_t^{RSI}} \right) \right]^{1/2} \cdot V_h
\]

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

where \( \frac{P_{t+k}^{Hed}}{P_t^{Hed}} \) denotes the **change in the hedonic index** between the periods \( t \) and \( t + k \), while \( \frac{P_{t+k}^{RSI}}{P_t^{RSI}} \) is the **change in the repeat-sales index**.
### Table: Performance measures

<table>
<thead>
<tr>
<th>Model</th>
<th>SSLE</th>
<th>D</th>
<th>$D^*$ based on Hedonic index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M1(OLS)</td>
<td>0.036540</td>
<td>0.004146</td>
</tr>
<tr>
<td></td>
<td>M1(KF)</td>
<td>0.041124</td>
<td>0.003552</td>
</tr>
<tr>
<td></td>
<td>M2(PLS)</td>
<td>0.029043</td>
<td>0.003627</td>
</tr>
</tbody>
</table>
Main Findings

- The **median index is extremely volatile** and cannot capture any quality changes.

- The **price index rises less** when locational effects are captured using a geospatial spline, rather than postcode dummies.

- The gap between the spline and
  - postcode based indexes is about 2.0 percent (over 7 years).
  - Kalman-Filter version of model M1 is about 2.2 percent (over 7 years).
  - repeat-sales indexes is larger (about 3.5 percent over 7 years).

- In terms of sum of squared log errors the M2(PLS) performs best.

- While the M1(KF) is favored by the repeat-sales benchmark.
Conclusions

- Splines (or some other nonparametric method), when combined with the hedonic imputation method, provide a **flexible way** of incorporating geospatial data into a HPI.

- For a city with postcodes as finely defined as Sydney (about 14,300 residents and 7.39 square kilometers per postcode), postcode dummies do a **good job** of controlling for locational effects.

- **Improved HPIs** are possible when an over time interrelated parameters (or surfaces) are estimated with the Kalman-Filter.

- In our data set the repeat-sales sample seems to be **biased**. This bias has to be taken into account for an adequate performance measure.
Work in progress! Some open points:

- **Combining the Kalman-Filter with PLS** (maybe with an Taylor approximation).

- It is important to **use the full data set**, and not just observations with no missing characteristics. The Kalman-Filter is also an adequate tool for this kind of problem.

- What is the difference between the proposed methods when the periods are even shorter than months? Construction of **weekly indexes**.

- Controlling for **seasonality**.

- Estimating a **spatial-temporal** model.
Thank you for your attention!

Merci pour leur attention!
GAM

- We estimate the **Generalized Additive Model**

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

with a **thin plate regression spline** \( tprs \) → Wood (2003)

- an optimal low rank approximation of a **thin plate spline**

- A \( tprs \) uses far fewer coefficients than a full spline:
  - **computationally efficient**, while losing little statistical performance.
  - **Big Additive Models** \( \text{bam}-\text{function from the R package mgcv} \).

- We can **avoid certain problems** in spline smoothing:
  - Choice of knot location and basis
  - Smooths of more than one predictor
Thin plate spline smoothing problem → Duchon (1977):

Estimate the smooth function $g$ with $d$-vector $\mathbf{x}$ from $n$ observations s.t.

$$y_i = g(x_i) + \epsilon_i$$

by finding the function $\hat{f}$ that minimizes penalized problem

$$\|\mathbf{y} - \mathbf{f}\|^2 + \lambda J_{md}(f),$$  \hspace{1cm} (7)

- $\mathbf{y} = (y_1, \ldots, y_n)^T$,
- $\mathbf{f} = (f(x_1), \ldots, f(x_n))^T$,
- $\lambda$ is a smoothing parameter,
- $J_{md}(f)$ is a penalty function measuring the wiggliness of $f$

$$J_{md} = \int \cdots \int \sum_{\nu_1 + \cdots + \nu_d = m} \frac{m!}{\nu_1! \cdots \nu_d!} \left( \frac{\partial^m f}{\partial x_1^{\nu_1} \cdots \partial x_d^{\nu_d}} \right)^2 dx_1 \cdots dx_d$$
It can be shown that the solution of (7) has the form,

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

where

- $\delta_i$ and $\alpha_j$ are coefficients to be estimated,
- $\delta_i$ such that $T^T \delta = 0$ with $T_{ij} = \phi_j(x_i)$.

The $M = \binom{m+d-1}{d}$ functions $\phi_j$ are linearly independent polynomials spanning the space of polynomials in $\mathbb{R}^d$ of degree less than $m$ (i.e. the null space of $J_{md}$)

$$\eta_{md}(r) = \begin{cases} 
\frac{(-1)^{m+1+d/2}}{2^{2m-1} \pi^{d/2} (m-1)! (m-d/2)!} r^{2m-d} \log(r) & d \text{ even} \\
\frac{\Gamma(d/2-m)}{2^{2m} \pi^{d/2} (m-1)!} r^{2m-d} & d \text{ odd}
\end{cases}$$
With $E$ by $E_{ij} = \eta_{md}(\|x_i - x_j\|)$, the thin plate spline fitting problem is now the minimization of

$$\|y - E\delta - T\alpha\|^2 + \lambda\delta^T E\delta \quad \text{s.t.} \quad T^T\delta = 0. \quad (9)$$

with respect to $\delta$ and $\alpha$

- **Ideal smoother**: Exact weight to the conflicting goals of matching the data and making $f$ smooth

- **Disadvantage**: As many parameters as there are data, i.e. $O(n^3)$ calculations

- More on thin plate splines: Wahba (1990) or Green and Silverman (1994)
The computational burden can be reduced with the use of a **low rank approximation**, Wood (2003).

A parameter space basis that perturbs (9) as little as possible.

The basic idea: **Truncation of the space** of the wiggly components of the spline (with parameter $\delta$), while leaving the $\alpha$-components unchanged.

$E = U D U^T$ eigen-decomposition of $E$

Appropriate submatrix $D_k$ of $D$ and corresponding $U_k$, restricting $\delta$ to the column space of $U_k$, i.e. $\delta = U_k \delta_k$,

$$||y - U_k D_k \delta_k - T \alpha||^2 + \lambda \delta_k^T D_k \delta_k \text{ s.t. } T^T U_k \delta_k = 0,$$

(10)

with respect to $\delta_k$ and $\alpha$. 

• The **computational cost is reduced** from $O(n^3)$ to $O(k^3)$

• Remaining problem: Find $U_k$ and $D_k$ sufficiently cheaply.

• Wood (2003) proposes the use of the Lanczos method (Demmel (1997)) which allows the calculation at the substantially lower cost of $O(n^2k)$ operations.

• **Smoothing parameter selection**: Laplace approximation to obtain an approximate REML which is suitable for efficient direct optimization and computationally stable, Wood (2011)
We can formulate the model of interest in a **state space** formulation

\[
y_t = \log(p_t) = X_t \alpha_t + \varepsilon_t
\]

\[
\alpha_t = \alpha_{t-1} + \eta_t
\]

where

- \( X_t = \begin{pmatrix} Z_t & D_t \end{pmatrix} \) is an \((H_t \times m)\)-matrix of characteristics,
- \( \alpha_t = \begin{pmatrix} \beta_t & \delta_t \end{pmatrix}^T \) is a \((m \times 1)\)-vector of shadow prices,
- \( \varepsilon_t \sim NID(0, R_t) \), where \( R = \mathbb{E}(\varepsilon_t \varepsilon_t^T) = \sigma^2_{\varepsilon} I_{H_t} \) is the variance-covariance matrix,
- \( \eta_t \sim NID(0, Q) \), where \( Q = \mathbb{E}(\eta_t \eta_t^T) = \text{diag}(\sigma^2_{\beta} I_C, \sigma^2_{\delta} I_B) \) is the variance-covariance matrix, and
- \( \Psi = \begin{pmatrix} \sigma^2_{\varepsilon} & \sigma^2_{\beta} & \sigma^2_{\delta} \end{pmatrix}^T \) are some hyper-parameters.
The Kalman-Filter estimator is given by

\[
\hat{\alpha}_{\tau|\tau} = \mathbb{E}(\alpha_{\tau}| y_\tau, y_{\tau-1}, \ldots, X_\tau, X_{\tau-1}, \ldots, a_0, \Omega_0)
\]

and set up recursively

\[
\hat{\alpha}_{\tau|\tau} = \hat{\alpha}_{\tau-1|\tau-1} + G_\tau \nu_\tau \tag{11}
\]

where,

- \(\nu_\tau = y_\tau - X_\tau \hat{\alpha}_{\tau-1|\tau-1}\) is the prediction error using the parameter estimates at \(\tau - 1\) and \(\nu_\tau \sim (0, F_\tau)\), with size \(H_t \times 1\).
- \(F_\tau = X_\tau \Omega_{\tau|\tau-1} X_\tau^T + R_\tau\) is the variance-covariance of the prediction error, \(\nu_\tau\).
- \(G_\tau = M_\tau F_\tau^{-1}\) is known as the Kalman gain and captures the information gain from \(\tau - 1\) to \(\tau\),
- \(M_\tau = \Omega_{\tau|\tau-1} X_\tau^T\), and
- \(\Omega_{\tau|\tau-1} = \Omega_{\tau-1|\tau-1} + Q\)
To obtain estimates of $R_t$ and $Q$, we use **maximum likelihood estimation**.

The log-likelihood for a sample of $H = \sum_{t=1}^{T} H_t$ transactions over $T$ time periods is given in this case by,

$$
\ln L(y_t; \psi) = -\frac{1}{2} \sum_{t=1}^{T} N_t \ln(2\pi) - \frac{1}{2} \sum_{t=1}^{T} \ln |F_t| - \frac{1}{2} \sum_{t=1}^{T} \nu_t^T F_t^{-1} \nu_t
$$  \hspace{1cm} (12)

To obtain estimates of $\psi$, a numerical maximization of the log-likelihood for the unknown parameters is carried out using **Newton-Raphson type algorithms**. The estimates are given by,

$$
\hat{\psi} = \arg \max_\psi \ln L(y_t|\psi)
$$  \hspace{1cm} (13)

Once $\hat{\psi}$ are available, $R_t$ and $Q$ are replaced by $\hat{R}_t$ and $\hat{Q}$ in the expressions to implement (11).