Weekly Hedonic House Price Indexes: An Imputation Approach with Geospatial Splines and Kalman Filters

Michael Scholz (University of Graz, Austria)
Robert J. Hill (University of Graz , Austria)
Alicia Rambaldi (University of Queensland, Australia)

Paper prepared for the 34th IARIW General Conference

Dresden, Germany, August 21-27, 2016

Session 6A: Accounting for Real Estate

Time: Thursday, August 25, 2016 [Afternoon]

Weekly Hedonic House Price Indexes: An Imputation Approach with Geospatial Splines and Kalman Filters

Robert J. Hill[1,*], Alicia N. Rambaldi[2], and Michael Scholz[1]

[1] Department of Economics, University of Graz, Universitätsstr. 15/F4, 8010 Graz. Austria

[2] School of Economics, The University of Queensland, St Lucia, QLD 4072. Australia

Preliminary version: July 29, 2016

Abstract:

The hedonic imputation method provides a flexible way of constructing quality-adjusted house price indexes. However, the method becomes unreliable at higher frequencies (e.g., for weekly indexes), since then the underlying price trend will be close to zero and even in large data sets there may not be enough price observations in each period. As a consequence computational and statistical problems occur (e.g., no observations for some postcodes, a loss in degrees of freedom, or an increased variance of estimated parameters). We show how the reliability of weekly indexes can be improved by replacing postcode dummies by a geospatial spline and then using a Kalman filter. This approach has two advantages. First, the dimensionality of the model is reduced. Replacing postcode dummies by values from the geospatial spline function at each location in the data set very significantly reduces the number of parameters that need to be estimated, and the number of covariance restrictions that must be imposed to make the Kalman filter operational. Second, the small number of observations in each period causes larger variability in estimated parameters (shadow prices) which should not change that much from one week to the next. Estimation of a dynamic linear model with the Kalman filter interconnects those parameters over time. Applying this hedonic geospatial spline/Kalman filter approach to data for Sydney (Australia) we show that it outperforms competing alternatives for computing house price indexes at a weekly frequency. (*JEL*. C32; C43; E01; E31; R31)

Keywords: Housing market; House Price index; Hedonic imputation; Geospatial data; Spline; Quality adjustment; Kalman-Filter; State Space Models

*Corresponding author: R. Hill (+43(0)316 380 3442, robert.hill@uni-graz.at), This project has benefited from funding from the Austrian National Bank (Jubiläumsfondsprojekt 14947). We thank Australian Property Monitors for supplying the data.

1 Introduction

The hedonic imputation method provides a flexible way of constructing quality-adjusted house price indexes. The interplay of this well known method with the increasing number of recorded residential property transaction prices and the advances in computing power and econometric techniques offers new opportunities in constructing higher frequency indexes (at the weekly or even daily level) and in deepening the knowledge about the real estate asset class. Bokhari and Geltner (2012) give further reasons for the usefulness of higher frequency indexes:

> "[T]he greater utility of higher frequency indexes has recently come to the fore with the advent of tradable derivatives based on real estate price indexes. Tradability increases the value of frequent, up-to-date information about market movements, because the lower transactions and management costs of synthetic investment via index derivatives compared to direct cash investment in physical property allows profit to be made at higher frequency based on the market movements tracked by the index. Higher-frequency indexes also allow more frequent "marking" of the value of derivatives contracts, which in turn allows smaller margin requirements, which increases the utility of the derivatives."

However, the hedonic imputation method becomes unreliable at higher frequencies (e.g., for weekly indexes), since then even in large data sets there may not be enough price observations in each period. As a consequence computational and statistical problems occur (e.g., no observations for some postcodes, a loss in degrees of freedom, or an increased variance of estimated parameters). Geltner and Ling (2006) describe the trade-off between statistical quality per period and the frequency of index reporting,

holding constant the overall quantity and quality of raw valuation data and index construction methodology. They conclude that the usefulness of an index for research purposes clearly increases the greater the frequency of reporting, holding statistical quality constant (per period) (Bokhari and Geltner, 2012).

Only in recent years have researchers in the housing field started to construct high-frequency indexes. For example, Bokhari and Geltner (2012) propose a two-step procedure based on a generalised inverse estimator that improves the accuracy of high-frequency indexes in scarce data environment (in an application to commercial property repeat-sales data). In recent work, Bourassa and Hoesli (2016) apply the procedure of Bokhari and Geltner (2012) and construct high frequency house price indexes for both cities and submarkets within cities. Bollerslev, Patton, and Wang (2015) develop daily house price indexes for 10 major US metropolitan areas. Their calculations are based on a database of several million residential property transactions and a standard repeat-sales method that closely mimics the methodology of the monthly Case-Shiller house price index. Bollerslev, Patton, and Wang (2015) apply a multivariate time series model for the daily house price index returns, explicitly allowing for commonalities across cities and GARCH effects.

In this article we show how the reliability of weekly hedonic indexes can be improved by replacing postcode dummies by a geospatial spline and then using a Kalman filter. This approach has two advantages. First, the dimensionality of the model is reduced. Replacing postcode dummies by values from the geospatial spline function at each location in the data set very significantly reduces the number of parameters that need to be estimated, and the number of covariance restrictions that must be imposed to make the Kalman filter operational. Second, the small number of observations in each period causes larger variability in estimated parameters (shadow prices) which should not change too much from one week to another. Estimation of a dynamic linear model with the Kalman filter interconnects those parameters over time.

Applying this hedonic geospatial spline/Kalman filter approach to data for Sydney (Australia) over the period 2001–2014 we show that it outperforms competing alternatives for computing house price indexes at a weekly frequency. We evaluate the different indexes and judge the index quality with the measures proposed in Guo *et al.* (2014) where formal tests of the quality of an index in terms of precision and reliability, and a smoothness test against random noise were introduced. In addition, we use the criterion proposed by Hill and Scholz (2014) which is based on the repeat-sales price relatives in our data set.

The remainder of this paper is structured as follows. Section 2 provides an overview of the applied econometric methods for estimation of the hedonic model (a generalized additive model and the Kalman-Filter), the hedonic price index construction, and discusses ways of measuring the quality of an index. Section 3 presents our data set, the empirical study and the results of our analysis. Section 4 concludes by considering some implications of our findings and gives a short outlook for further research. Technical details regarding the estimation procedures are postponed to appendix.

2 Hedonic Imputation and Index Quality

Hedonic price indexes for housing are constructed in three main ways: time-dummy methods, hedonic imputation, and average characteristic methods (Diewert, 2010; Hill, 2013). All of them have in common that in a hedonic model the price of a product is regressed 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.

2.1 Estimation of the hedonic model

In this paper, we use the hedonic imputation method and model the hedonic equation in two different ways: (i) in a semi-parametric generalised additive model (GAM) and (ii) a linear varying coefficient (dynamic) model. Our main goal is to extend the work of Hill and Scholz (2014), who used (i) for annual data, to the weekly frequency and combine it with (ii). The idea behind this is as follows. Hill and Scholz (2014) estimate in each period t the semilog GAM model with a parametric part based on the physical characteristics Z (including an intercept) and a fully nonparametric function $g_t(\cdot)$ defined on the geospatial data (latitudes z_{lat} and longitudes z_{long}):

$$y = Z\beta_t + g_t(z_{lat}, z_{long}) + \varepsilon, \tag{1}$$

where $y = \ln p$. An increase in the estimation frequency (from annual to weekly) will drastically reduce the number of observations in each period. As a result one expects larger variability in estimated parameters. But the shadow prices $\hat{\beta}_t$ and the price surface $\hat{g}_t$ should not change too much from one week to another. One possibility to combine the shadow prices $\hat{\beta}_t$ over time would be the estimation of a dynamic linear model with the Kalman-filter. For example, Rambaldi and Rao (2011) use time-varying hedonic models for the construction of house price indexes. The problem with such an approach is that it is not clear ad-hoc, how to incorporate the fully nonparametric part (the unknown function g_t) in such a setting[1]. It is well known that the information of the location of the individual dwellings is a main explanatory variable and price driver in a hedonic context. As an alternative to a function $g_t(\cdot)$ defined on longitudes and latitudes often the use of postcode dummies is proposed:

$$y = Z\beta_t + D\delta_t + \varepsilon, \tag{2}$$

[1]We are only aware of approximative interpolation methods (kriging) which are combined with the state-space model, known as the kriged Kalman filter (Mardia *et al.*, 1998; Cressie and Wikle, 2002).

where D is a matrix of postcode dummies containing the location information and δ is the vector of corresponding shadow prices. The drawback with (2) in our weekly estimation context is now that the number of degrees of freedom is extremely reduced due to data sparsity. It can even happen that for some postcodes we have no observations in our records causing not only statistical problems but also computational problems especially in the hedonic prediction step.

To overcome the disadvantages mentioned above we propose the combination of both methods in the following way. First, we estimate (1) for the weekly frequency and extract each time the estimated locational component $\hat{g}_t$. In a second step we include this estimate of the spline function in (2) as a constructed variable replacing all the postcode dummies

$$y = Z\beta_t + \hat{g}_t\gamma_t + \varepsilon. \tag{3}$$

Note that γ is a scalar coefficient that can now vary over time and thus shift the whole spline surface up or down. But even more important, the Kalman filter can now interrelate the spline surfaces over time. In addition, we have replaced a large number of parameters (we have around 250 postcodes in our data set) by a single parameter, what is a great gain in parsimony. In other words, the number of parameters to be estimated and covariance restrictions that must be imposed to make the Kalman filter operational have been heavily reduced.

2.2 Index construction

As usual in the hedonic imputation literature, we use the estimated hedonic models (1)-(3) and impute prices which can 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 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.[2] 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 and is consistent with a log-price hedonic model.

The indexes presented below are all of the double imputation type.[3] This means that both prices in each price relative are imputed. For example, for model (1) and (3) the double imputation Paasche index (DIP), Laspeyres index (DIL), and Fisher index (DIF) are defined as follows:

$$P_{t,t+1}^{DIP} = \prod_{h=1}^{H_{t+1}} \left[\left(\frac{\hat{p}_{t+1,h}(z_{t+1,h}, \hat{g}_{t+1,h})}{\hat{p}_{t,h}(z_{t+1,h}, \hat{g}_{t+1,h})} \right)^{1/H_{t+1}} \right], \tag{4}$$

$$P_{t,t+1}^{DIL} = \prod_{h=1}^{H_t} \left[\left(\frac{\hat{p}_{t+1,h}(z_{t,h}, \hat{g}_{t,h})}{\hat{p}_{t,h}(z_{t,h}, \hat{g}_{t,h})} \right)^{1/H_t} \right], \tag{5}$$

$$P_{t,t+1}^{DIF} = \sqrt{P_{t,t+1}^{DIP} \times P_{t,t+1}^{DIL}} \tag{6}$$

In the index calculation for the hedonic model (3) where we apply the Kalman filter one has to be careful to compute the correct predictions $\hat{p}_{t,h}(z_{t+1,h}, \hat{g}_{t+1,h})$ and $\hat{p}_{t+1,h}(z_{t,h}, \hat{g}_{t,h})$. The crucial point is that the constructed location effect needs to be matched with the correct period. For example, $\hat{p}_{t,h}(z_{t+1,h}, \hat{g}_{t+1,h})$ imputes prices for houses sold in period $t+1$ based on the parameters estimated with houses sold in period t. In this case, $\hat{g}_{t+1,h}$ has to be the estimated location effect for house h sold in period $t+1$ extracted from the GAM of period t, i.e. from $g_t(z_{lat}, z_{long})$. In other words, $\hat{p}_{t,h}(z_{t+1,h}, \hat{g}_{t+1,h}) = \exp\left(z_{t+1,h}^{\top}\hat{\beta}_t + \hat{g}_{t,h}(z_{t+1,h,lat}, z_{t+1,h,long})\hat{\gamma}_t\right)$. Thus, standard prediction commands in software packages such as R usually give wrong results.

[2]This democratic weighting structure is in our opinion more appropriate in a housing context than weighting each house by its expenditure share. See de Haan (2010) for a discussion on alternative weighting schemes.

[3]We also calculated single imputation indexes where only one price of the price relatives is imputed. The results are virtually indistinguishable. Hence to safe space we focus here only on the double imputation type.

2.3 Measuring the quality of the index

The constructed indexes should be useful instruments for policymakers or market participants. But this claim requires measures for the quality of the proposed indexes. In this subsection we present four different criteria for judging index quality.

Guo *et al.* (2014) distinguish between two types of errors: (i) systematic effects and (ii) random errors. An example of the former is omitted variables bias which might cause a systematic difference in the long-term growth rate of the indexes. Others are sample selection bias, unbalanced data, or lagging. An example of the latter is the noise in the index arising from coefficient estimation in the hedonic model. This source of error has a larger impact on high-frequency indexes than on low-frequency indexes where the underlying signal is stronger and the number of data points per period is higher.

Guo *et al.* (2014) propose the use of the following three *signal-to-noise* measures for index returns $r_t = \ln(P_{t+1}/P_t)$, where P_t is the level of the price index in period t:

(i) The volatility (VOL) measure is the standard deviation of r_t.

(ii) The first-order autocorrelation (AC1) measure is $\hat{\beta}$ from the following OLS regression:

$$r_{t+1} = \beta r_t + \varepsilon_{t+1}.$$

(iii) The deviation from a smoothed Hodrick-Prescott (HP) filter representation. This deviation is calculated as follows:

$$HP = \sum_{t=1}^{T-1} \left[\ln\left(\frac{P_{t+1}}{P_t}\right) - \ln\left(\frac{HP_{t+1}}{HP_t}\right) \right]^2,$$

where HP_t is the smoothed price index calculated using the HP filter.

A smaller value of VOL and HP, and a higher value of AC1 indicate a smoother and hence better performing index.

Our fourth measure uses repeat sales as a benchmark against which to measure index performance (see Hill and Scholz (2014)).[4] Since the price relatives are the building blocks in our price index formulas (4) and (5), what most matters is the quality of the estimated price relatives $\hat{p}_{t+1,h}/\hat{p}_{t,h}$. Suppose house h sells in both periods t and $t+k$. For this house therefore we have a repeat-sales price relative: $p_{t+k,h}/p_{t,h}$, and can also provide an imputed price relative $\hat{p}_{t+k,h}/\hat{p}_{t,h}$. Therewith we can define V_h as the ratio of the actual to imputed price relative for house h:

$$V_h = \frac{p_{t+k,h}}{p_{t,h}} \Big/ \frac{\hat{p}_{t+k,h}}{\hat{p}_{t,h}}. \tag{7}$$

Our quality measure is then the average squared error of the log price relatives of each hedonic method:

$$D = \left(\frac{1}{H}\right) \sum_{h=1}^{H} [\ln(V_h)]^2, \tag{8}$$

where the summation in (8) takes place across the whole repeat-sales sample. We prefer whichever model has the smaller value of D.

3 Empirical application

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 2001–2014. 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 the Appendix in Table 2.

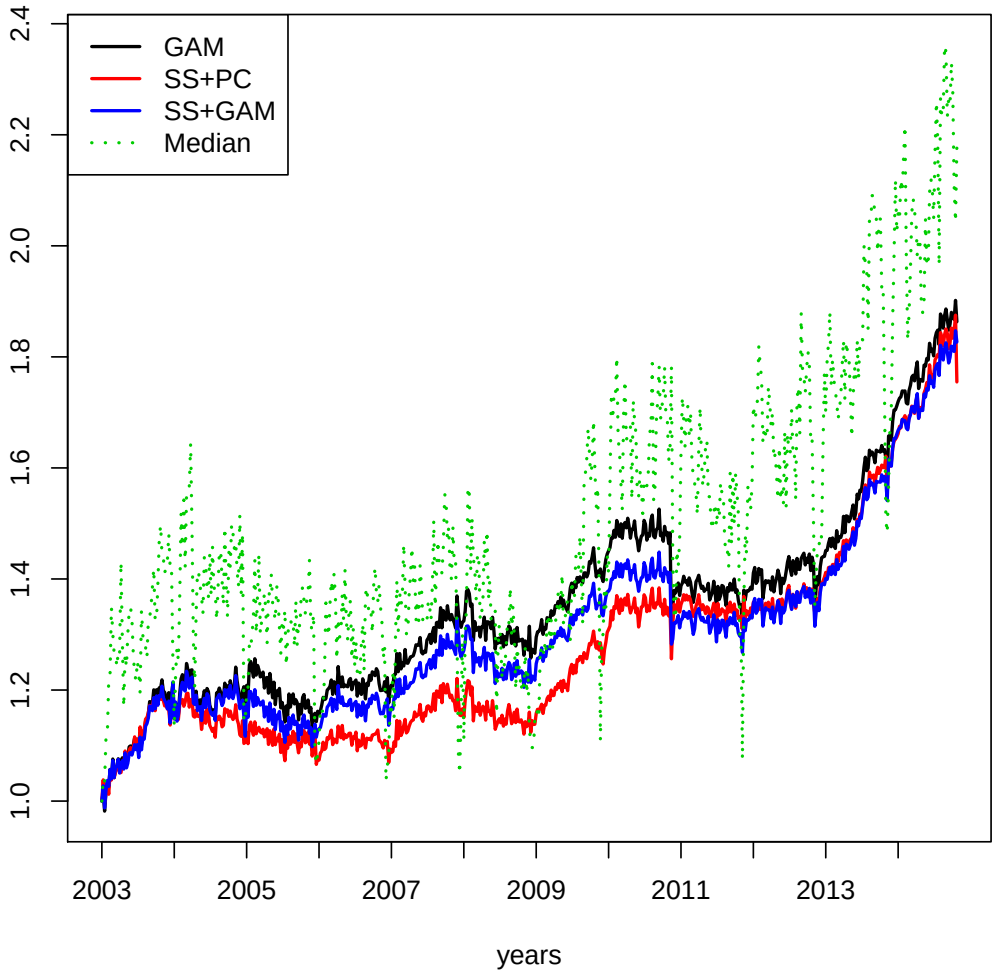
[4] For an exact definition of repeat-sales and a broader discussion see Hill and Scholz (2014).

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 of erroneous extra zeroes. These extreme observations can distort the results. The exclusion criteria we applied are shown in the Appendix in Table 3. Complete data on all our hedonic characteristics are available for 433 202 observations. To simplify the computations we also merged the number of bathrooms and number of bedrooms to broader groups (one, two, and three or more bathrooms; one or two, three, four, five or more bedrooms). From earlier studies with the same data set we know that the quality of the data improves over time. Especially the data of the first two years seems to be poor. Thus we decided to present the hedonic indexes starting in 2003. Nevertheless, we use the full sample period for estimation of the state space model. An other reason for discarding the first two years when it comes to construct an index based on the Kalman filter is that we have to assure that the filter has settled. Starting the index in 2003 we can be confident that the initial condition is not affecting the index base.

3.2 House price indexes

House price indexes for Sydney generated using the GAM method (i.e., hedonic imputation with a spline but no state space model), SS+PC (i.e., hedonic imputation with postcodes and a state space model), and SS+GAM (i.e., hedonic imputation with a spline and state space model) are shown in Figure 1. Also shown is a median price index. The median index is extremely volatile, thus demonstrating the need for quality adjustment to generate an economically meaningful index. The three hedonic indexes while broadly comparable, exhibit significant differences particularly in the middle part of the data set.

Figure 1: Weekly House Price Indexes from 2003 to 2014



Note: GAM is based on periodwise estimation of model (1); SS+PC is the state space model (2) with postcode dummies; SS+GAM is the state space model (3) with the spline component; Median is the usual median index on a weekly frequency.

3.3 Comparing the quality of the indexes

The performance of the three hedonic methods, a median index, and a repeat-sales index are compared in Table 1. SS+PC performs best according to VOL followed by SS+GAM. The AC1 results are striking in that all the coefficients are negative, unlike in Guo *et al.* (2014). These authors show that the AC1 measure is a function of the following components (none of which can be directly observed):

$$AC1 = \frac{\rho_r \sigma_r^2 - \sigma_\eta^2/2}{\sigma_r^2 + \sigma_\eta^2},$$

where ρ_r is the autocorrelation coefficient of the true return, σ_r^2 is the variance of the true return, and σ_η^2 is the noise variance. Guo *et al.* (2014) use monthly data. Moving from monthly to weekly data acts to reduce ρ_r and σ_r^2 while increasing σ_η^2. This can explain why our AC1 coefficients are all negative. Surprisingly though the median index performs best according to AC1. This finding draws into question the usefulness of this criterion in the context of weekly data. The HP results are similar to the VOL results, and in particular generate the same ranking of methods. One problem with the HP method is that the smoothed index is not very smooth. Figure 2 shows this finding exemplarily for the GAM method. This suggests that the HP filter is not able to smooth out all the volatility in the weekly indexes. A filter with a higher degree of smoothness would work better here.

The SS+GAM method outperforms the SS+PC method according to the D criterion that compares imputed price relatives with their repeat-sales counterparts. However the GAM method without a state space model performs even better. These computations though are preliminary. In particular the first years of the data set which constitute the burning period of the state space model should be excluded when comparing methods.

Figure 2: Log Weekly Returns and HP Weekly Returns for the GAM Method

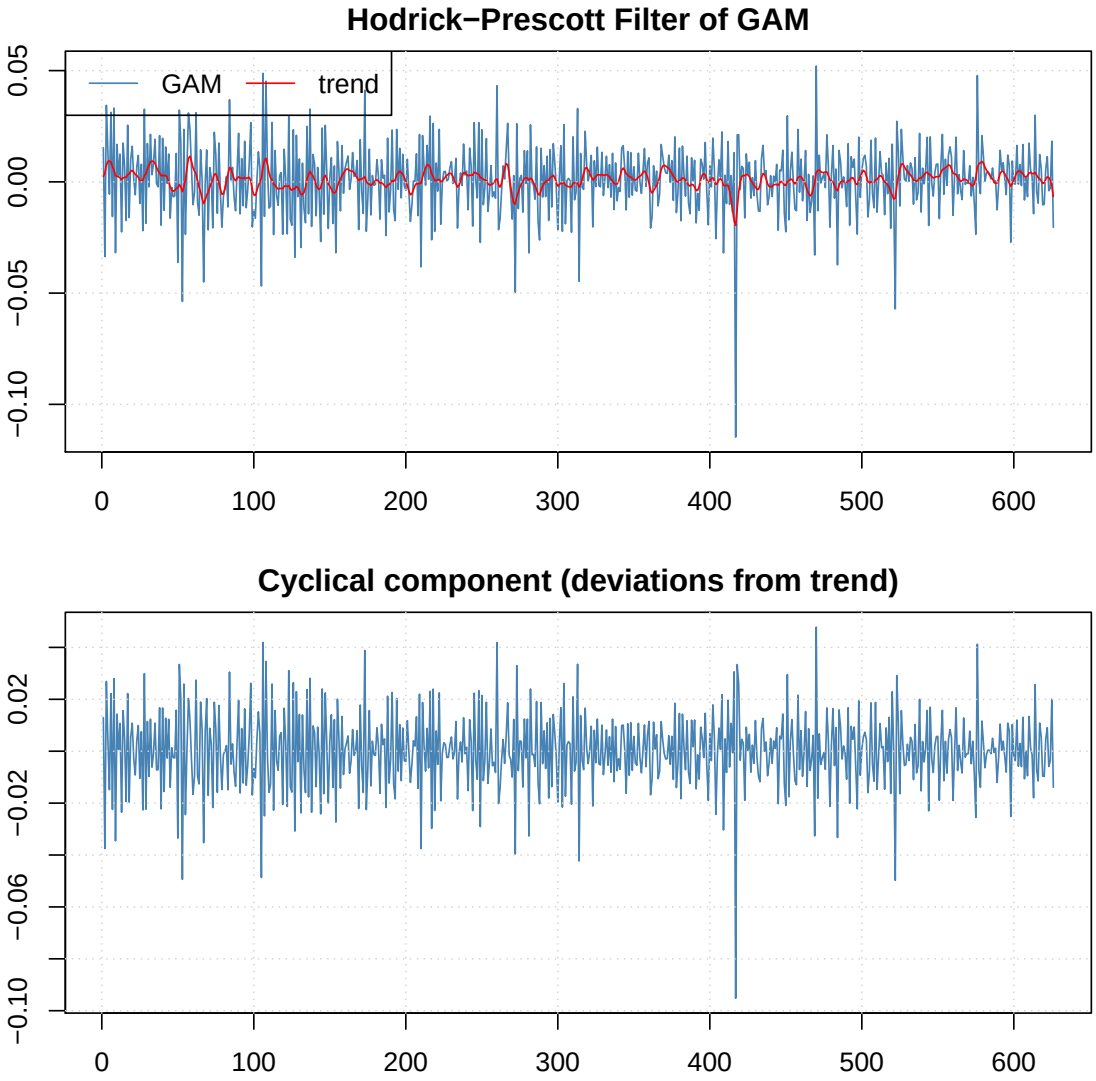


Table 1: Index quality criteria

	VOL	AC1	HP	D
GAM	0.0161	-0.4312	0.1415	0.1008
SS+PC	0.0148	-0.4673	0.1203	0.1068
SS+GAM	0.0156	-0.4376	0.1339	0.1020
Median	0.0567	-0.1819	1.5845	–
RS	0.0170	-0.5182	0.1631	–

Note: GAM is based on periodwise estimation of model (1); SS+PC is the state space model (2) with postcode dummies; SS+GAM is the state space model (3) with the spline component; Median is the usual median index, RS is the repeat-sales index. All indexes based on a weekly frequency.

4 Conclusion

Our results are still very preliminary. Our performance criteria are still being fine tuned, and we may still also include more criteria.

The extent of the short-term volatility in our weekly hedonic indexes though is surprising. It remains to be seen how much of this volatility is genuine and how much reflects measurement problems. It is also perhaps surprising that the use of state space models does little to reduce the volatility of our price indexes, and it remains to be seen whether it can really be argued that the use of state space models improves the quality of our weekly indexes.

We intend to increase the flexibility of the state space model by including Residex region dummies (there are 16 Residex regions in Sydney). This will allow the state space model to shift up or down different parts of the spline surface by differing amounts. It remains to be seen what impact this increased flexibility will have on SS+GAM. Another issue we are considering is to try extend the analysis to daily indexes.

References

Bokhari, S. and D. Geltner (2012). Estimating Real Estate Price Movements for High Frequency Tradable Indexes in a Scarce Data Environment. *Journal of Real Estate Finance and Economics* **45**(2), 522–543.

Bollerslev, T., A.J. Patton, and W. Wang (2015). Daily House Price Indices: Construction, Modeling, and Longer-Run Predictions. *Journal of Applied Econometrics* forthcoming.

Bourassa, S.C. and M. Hoesli (2016). High Frequency House Price Indexes with Scarce Data. *Swiss Finance Institute Research Paper Series* **16-27**.

Cressie, N. and C.K. Wikle (2002). Space-time Kalman filter. *Encyclopedia of Environmetrics* **4**, 2045–2049.

de Haan, J. (2010). Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and Re-Pricing Methods. *Journal of Economics and Statistics* **230**(6), 772–791.

Diewert, W. E. (2010). Alternative Approaches to Measuring House Price Inflation. Discussion Paper 1010, Department of Economics, The University of British Columbia, Vancouver, Canada, V6T 1Z1, 2010.

Geltner, D. and D. Ling (2006). Considerations in the Design and Construction of Investment Real Estate Research Indices. *Journal of Real Estate Research* **28**(4), 411-444.

Guo, X., S. Zheng, D. Geltner, and H. Liu (2014). A new approach for constructing home price indices: The pseudo repeat sales model and its application in China. *Journal of Housing Economics* **25**, 20-38.

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 M. Scholz (2014). Incorporating Geospatial Data in House Price Indexes: A Hedonic Imputation Approach with Splines. *Graz Economics Papers* **2014-05**.

Mardia, K.V., C. Goodall, E.J. Redfern, and F.J. Alonso (1998). The Kriged Kalman Filter. *Test* **7**(2), 217–285.

Rambaldi, A.N. and D.S.Pr. Rao (2011). Hedonic Predicted House Price Indices Using Time-Varying Hedonic Models with Spatial Autocorrelation. School of Economics Discussion Paper 432, School of Economics, University of Queensland.

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.

Appendix

A1. Estimation of the semiparametric hedonic model

The semiparametric hedonic model in (1) is an 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 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:

$$y = g(x) + \varepsilon, \tag{9}$$

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

$$||\boldsymbol{y} - \boldsymbol{f}||^2 + \lambda J_{md}(f), \tag{10}$$

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.[5] Under suitable conditions it can be shown that the solution of (10) has the form,

$$\hat{f}(x) = \sum_{i=1}^{n} \delta_i \eta_{md}(||x - x_i||) + \sum_{j=1}^{M} \alpha_j \phi_j(x), \tag{11}$$

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}.

[5] 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 $d = m = 2$).

Defining the matrix $\boldsymbol{E}$ by $E_{ij} = \eta_{md}(||x_i - x_j||)$, the thin plate spline fitting problem is now the minimization of

$$||\boldsymbol{y} - \boldsymbol{E\delta} - \boldsymbol{T\alpha}||^2 + \lambda \boldsymbol{\delta}^\top \boldsymbol{E\delta} \qquad \text{s. t.} \qquad \boldsymbol{T}^\top \boldsymbol{\delta} = \boldsymbol{0}. \tag{12}$$

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 (12) 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 (12) becomes

$$\text{Min}_{\boldsymbol{\delta}_k, \boldsymbol{\alpha}} \left\{ ||\boldsymbol{y} - \boldsymbol{U}_k \boldsymbol{D_k} \boldsymbol{\delta_k} - \boldsymbol{T\alpha}||^2 + \lambda \boldsymbol{\delta_k}^\top \boldsymbol{D_k} \boldsymbol{\delta_k} \right\} \qquad \text{s. t.} \qquad \boldsymbol{T}^\top \boldsymbol{U}_k \boldsymbol{\delta_k} = \boldsymbol{0}. \tag{13}$$

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 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).

A2. Estimation of the time-varying hedonic model

The time-varying hedonic model (3) is estimated in the following way.[6] Remember that the parameters β and γ will be interconnected over time:

$$\beta_t = \beta_{t-1} + \eta_\beta \quad \text{and} \quad \gamma_t = \gamma_{t-1} + \eta_\gamma \tag{14}$$

where $\eta_\beta \sim N(0, \sigma^2_\beta)$, $\eta_\gamma \sim N(0, \sigma^2_\gamma)$. For the error terms in (3) we also assume $\varepsilon \sim N(0, \sigma^2_\varepsilon)$. Note that the intercept in Z could be interpreted as a local trend for period t (which could also be expanded to cover a seasonal component). Define $\alpha_t = (\beta_t, \gamma_t)^\top$ as the vector of time-varying parameters to be estimated and $E(\alpha_t \alpha_t' | \Im) = \Omega_{t|\Im}$ as the variance-covariance matrix of parameters given available information. We require $\Omega_{t|t-1}$ for estimation and $\Omega_{t|t}$ to construct standard errors and confidence intervals.

Given y_t, Z_t, $\hat{g}_t$, $\hat{\sigma}^2_\varepsilon$, $\hat{\sigma}^2_\beta$ and $\hat{\sigma}^2_\gamma$, estimates of $\alpha_{t|t}$ are obtained using the Kalman filter estimator of α_t given information up to and including week t,

$$a_{t|t} = a_{t-1|t-1} + G_t \nu_{t|t-1}, \tag{15}$$

where $G_t = \Omega_{t|t-1} Z_t' F_t^{-1}$ is the Kalman Gain, $F_t = E(\nu_{t|t-1} \nu_{t|t-1}')$, $\nu_{t|t-1} = y_t - Z_t \hat{\beta}_{t-1} - \hat{g}_{t-1|t} \gamma_{t-1}$, and $\hat{g}_{t-1|t}$ is the estimated non-parametric surface for properties sold in period t evaluated at time $t-1$.

The variance-covariance matrix $\Omega_{t|t-1}$ is a function of $\hat{\sigma}^2_\beta$ and $\hat{\sigma}^2_\gamma$, and F_t is a function of $\hat{\sigma}^2_\varepsilon$. To compute (15) for week $t = \tau$, the Kalman filter algorithm is run for period $t = 1, \ldots, \tau$ to provide the estimate $a_{\tau|\tau}$. Estimates $\hat{\sigma}^2_\varepsilon$, $\hat{\sigma}^2_\beta$ and $\hat{\sigma}^2_\gamma$ are given by maximizing the log-likelihood $\ln L$ in predictive form

$$lnL(\sigma^2_\varepsilon, \sigma^2_\beta, \hat{\sigma}^2_\gamma; y_t, Z_t, \hat{g}_t) = -\frac{NT}{2} ln(2\pi) - \frac{1}{2} \sum_{t=1}^{T} ln|F_t| - \frac{1}{2} \sum_{t=1}^{T} \nu_{t|t-1}' F_t^{-1} \nu_{t|t-1},$$

where $N = \sum_{t=d}^{T} N_t$; d is sufficiently large to avoid the log-likelihood being dominated by the initial condition, $\alpha_0 \sim N(a_0, \Omega_0)$.

[6] As mentioned before, model (2) has more parameters involved but the estimation is quite similar. Thus we only present the setting for model (3).

A3. Further information on the data set

Some summary statistics for our data set are provided in Table 2.

Table 2: Summary of characteristics

	PRICE ($)		BED		BATH	AREA	LAT	LONG
Minimum	56500	1:	1348	1:	190395	100.0	-34.20	150.6
1st Quartile	420000	2:	38578	2:	174161	461.0	-33.93	150.9
Median	610000	3:	200428	3:	57673	587.0	-33.84	151.0
Mean	784041	4:	147794	4:	8835	626.1	-33.85	151.0
3rd Quartile	900000	5:	38734	5:	1746	720.0	-33.76	151.2
Maximum	3200000	6:	6320	6:	392	4998.0	-33.40	151.3

For a robust analysis it was necessary to remove some outliers. The exclusion criteria we applied are shown in Table 3.

Table 3: Criteria for removing outliers

	PRICE	BED	BATH	AREA	LAT	LONG
Minimum Allowed	50000	1.000	1.000	100.0	-34.20	150.60
Maximum Allowed	4000000	6.000	6.000	5000.0	-33.40	151.35