The Effect of Hedonic Modelling and Index Weights on Hedonic Imputation Indices.

Alicia N. Rambaldi and D.S. Prasada Rao*

March, 2011

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

Abstract

Housing price indexes are generally computed using variants of hedonic housing price models. The commonly used methods are the time-dummy method, the hedonic imputation method and the rolling window hedonic models. In addition, hedonic models that explicitly account for spatial correlation in prices reflecting the price determining role of locational characteristics have been recently developed. The main objective of the paper is to develop a class of hedonic models which explicitly account for time-varying nature of the coefficients of the hedonic model as well as for the presence of spatially correlated errors and to provide an assessment of the predictive performance of various alternatives currently available. Construction of housing price index series with alternative weighting systems, plutocratic versus democratic, is also considered in the paper. Seasonality in the house sales data is considered in constructing monthly chained indices and annual chained indices based on averages of year-on-year monthly indexes. The empirical results presented in the paper make use of residential property sales data for Brisbane over the period 1985 to 2005. On the basis of the Root Mean Square Prediction Error criterion the time-varying parameter model with spatial errors is found to be the best performing model and the rolling-window model to be the worst performing model. The results indicate the presence of three episodes of housing price escalation during the study period.

1 Introduction

Compilation and publication of housing price index numbers is considered critical in the assessment of the general economy and is an important input into monetary policy including the setting of interest rates. In the past median

*Corresponding author: p.rao@economics.uq.edu.au

house prices were used in measuring price changes but it is generally recognised that median prices are unduly affected by the mix of houses sold and can exhibit strong patterns of seasonality. Over the last two decades, the use of hedonic models of house prices has become more prevalent. Hill and Melser (2008)-HM provides an excellent overview of the hedonic methods used in the construction of housing price index numbers. Theoretical foundations that underpin the use of hedonic models as well as the multitude of index number formulae that are available for the construction of housing price index numbers are discussed in some detail by HM. In contrast to the focus on alternative varieties of hedonic price index numbers in HM, the main objective of this paper is on the econometric aspects of the specification and efficient estimation of parameters of the hedonic models. Hedonic housing price indices were traditionally computed by maintaining the hedonic parameters fixed and adding intercept time-dummies, leading to the "time-dummy method" (TDH). Recently, Hedonic Imputation indices (HI) have been advocated (Triplett (2004); Silver (2007)) where the price imputations are essentially computed using the predictions of prices from estimated hedonic models. HM show that TDH is biased and that HI suffer from bias unless the hedonic parameters are allowed to vary over time and over heterogeneous regions. HM label this "substitution bias"[1] and advocate the estimation of a separate regression for each time period and region. A particular case of this approach which uses an adjacent-period regression, which consists of estimating a fixed parameter model over a two-period "rolling window" means that the hedonic coefficients are only held constant for two periods, as opposed to the entire sample period. Triplett (2004) argues that this is a more "benign constraint" because coefficients would usually change less between two adjacent periods than over extended intervals, and hence labels the adjacent-period approach as best practice among TDH.

The use of rolling-window and the time-varying hedonic regression models considered in this study allow the hedonic coefficients to be completely independent across different time periods. If the hedonic regressions have a theoretical foundation along the lines discussed in Diewert (2001) and HM then one would expect the hedonic regression coefficeints to evolve over time. Formalising this notion of smooth evolution of parameters, Cominos et al (2007)-CRR proposed the use of a time-varying hedonic regression model where the vector of hedonic regression parameters are assumed to follow a random-walk process. In addition, the standard hedonic regression models discussed in HM and Triplett (2004) also make the implicit assumption[2] that prices of houses sold are independent of each other and depend only on the hedonic characteristics. This assumption is not consistent with the popular notion that when it comes to sales prices of houses location is a major factor. It is, therefore, important to reflect this important feature of location on the random disturbance term and postulate that house prices in the same neighbourhood or spatial location may be moving together. CRR and Svetchnikova et al (2008) propose hedonic regression models with spatially correlated errors in their analysis of house sale prices in Brisbane.

The main objective of the paper is to provide a comparative assessment of the most popular of the hedonic

[1] It is difficult to interpret the bias induced by ignoring the time-varying nature of the hedonic regression parameters.

[2] The assumption is implicit in the specification that the random disturbance terms in the housing price regression model are independently and identically distributed.

regression models used in the analysis of housing price data. In particular the paper focuses on the performance of the time-dummy method (TDH), rolling window model (RWE), time-varying parameter models (TVE) and TVE with spatial errors (TVE_SEM) based on the predictions (imputations) of prices generated using alternative specifications. The paper provides details of the estimation procedure used for the time-varying parameter model with spatial errors. As the hedonic regressions are an intermediate step in the computation of housing price index numbers, the second part of the paper is devoted to the compilation of housing price index numbers. The paper draws on the main recommendations of HM (2008) and focuses mainly on the Fisher and Tornqvist index number formulae. The paper significantly deviates from the HM approach and considers both plutocratic weights based on value shares of houses sold and democratic weights which are based on the number of houses sold. Making explicit recognition of the difference between the housing price index numbers and the standard cost-of-living index numbers, the paper argues for the use of both types of weights. An important feature of the housing price sales is the presence of seasonality in the mix of houses sold and its influence on the median house prices. Recognising the presence of seasonality, the paper constructs year-on-year monthly price index numbers using hedonic imputations and compares these with annual hedonic price index numbers.

The paper is organised as follows. Section 2 describes the housing sales price data for the Brisbane metropolitan area used in the study. Various models considered in the study are presented in Section 3. Details of the econometric estimation and hedonic imputation for the substance of Section 4. The root mean squared prediction error used in assessing the performance of the hedonic models is also discussed in the section. Section 5 presents estimates of hedonic models with special focus on the temporal movements of the hedonic coefficients relating to land, number of bedrooms and the number of bathrooms. Hedonic imputed indices for housing are presented in Section 6. A few concluding observations are made in Section 7.

2 Data

The data used in this study are multiple single transactions of residential property sales in the Brisbane (Australia) metropolitan area for the period 1985:1 to 2005:12. The data are from one of the leading providers of property information services in Australia, 'RP Data Ltd' (`www.rpdata.com`). These data were first collected by Cominos (2006) and used in Cominos et al (2007). Further filtering of the data was conducted by Svetchnikova (2007) and the resulting data set, which is used in this study, was also used by Svetchnikova et al (2008), where detailed descriptive statistics are presented. The empirical work for the study is limited to price data for residential houses on blocks of land and excludes units, terraces, townhouses and duplexes.[3]

Preparation of data for analysis was undertaken in three steps.

First, a decision on the variables for inclusion and exclusion was made. There is a big trade-off between the number of included attributes and the sample size with these type of data as there were many observations with

[3] It is therefore necessary to be cautious in generalising these results to all types of dwellings.

incomplete data on property attributes. As a result of this trade-off, the dataset used in the study contains 65,239 single transactions over the sample period. Each data point (transaction) includes, the date (month and year) of sale, sale price, geocode (latitude, longitude), the postcode, the size of the land (lot) in m^2(AREA), the number of bedrooms (BED), the number of Bathrooms (BATH), the number of car spaces (lock-up garages and carports) combined into one series (CARLUG).

Second, the dataset needed to be cleaned for errors, incomplete observations, significant outliers and obvious errors in transcription. A number of simple checks were developed and all the observations that failed these checks were dropped. For example, all properties with sale prices below $1000 and above $30 million were dropped. Further details of the checks performed can be found in Cominos, Rambaldi and Rao (2007). The initial information on 316,359 transactions of house sales was downloaded for the Brisbane city area spanning the period from early 1950's to the end of 2005. Recognising the fact that a lot of sales data prior to January, 1985 was missing, the empirical results reported in this paper are based on data for the period 1985 to 2005. The final analysis is restricted to 65,239 sale price observations that met all the considerations discussed here.

Finally, the address of each house was geocoded to provide a latitude/longitude coordinate for each observations using the Geodetic Datum of Australia 194 using MapInfo Professional. Over 90 percent of the original records were successfully geocoded.

The distribution of transactions over the sample period is important as it might have an impact on the accuracy of some of the results. Figure 1 plots the number of transactions per month in the dataset. The number of recorded transactions has risen substantially since the mid 1990s. While the actual number of transactions is likely to have risen due to a very high population growth in the city of Brisbane in the last 20 years, it is also the case that the market for electronic databases was not established in the earlier part of the period, and therefore it is possible that some non-trivial number of transactions were never included in the electronic database for the earlier period.

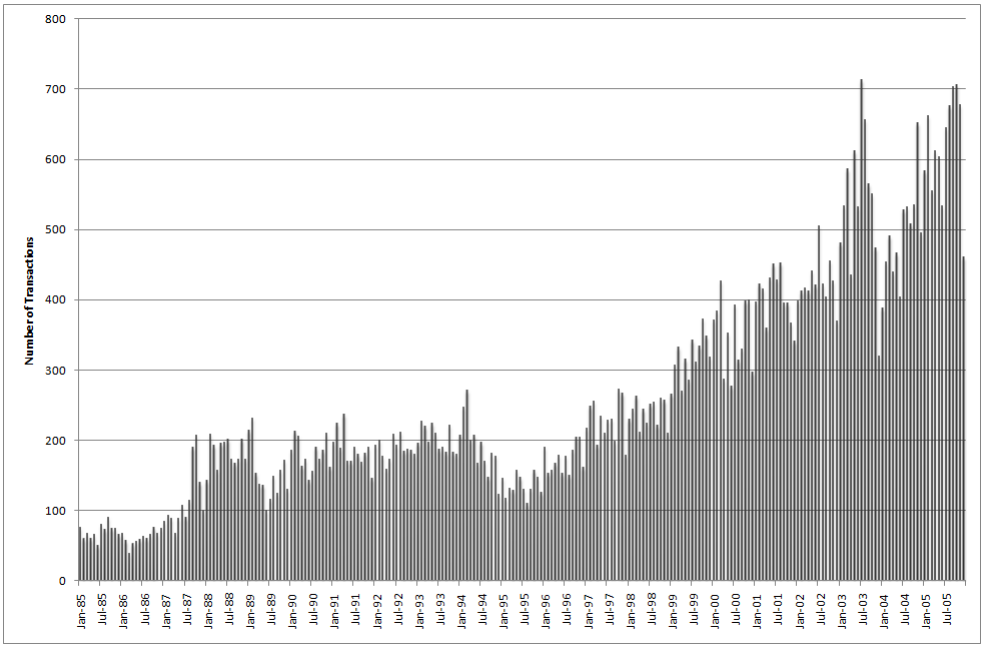

Figure 1: Number of transactions per month in the dataset

3 Econometric Modelling

The housing price indexes reported in the paper are based on suitably specified hedonic regression models. As one of the objectives of the paper is to evaluate the in-sample prediction performance of these models, a comprehensive range of hedonic models are considered in the study. We start with basic *time dummy hedonic model* (TDH) which is the most commonly used model in the construction of housing price index numbers. The TDH model includes dummy variables representing time and assumes constancy of parameters over time. The TDH model is generalised to accommodate the presence of spatially correlated disturbance term. . We will denote the time dummy hedonic model with spatial errors, Spatial Error Model, denoted by TDH-SEM. The TDH and TDH-SEM are essentially fixed parameter models and we consider several extensions to these. The main extension of the TDH model is to allow hedonic parameters to vary over time through a specific stochastic process. The model proposed here is slightly different from the time-varying parameter models considered by Cominos et al. (2007) and Svetchnikova et al. (2008) in that it allows for a different stochastic process for the time-varying intercept parameter compared to that used to model the slope coefficients. The basic idea here is that movements in constant term represent secular movements in prices independent of the movements in the hedonic regression coefficients. In practice, trends in the constant term tend to dominate the house price movements over time. We denote this generalised model using the extension, TV. The TV model is extended to allow for the presence of spatial correlation resulting in a new new model denoted by TV_SEM. The spatial correlation parameter is assumed to be fixed over time and space. In addition to these general classes of hedonic models we also consider the *rolling window* models that are essentially TDH models applied to housing price data from two adjacent time period. There has been some discussion and

support in the housing price indices literature for the use of adjacent period regressions (see Triplett (2004)). In this method the regression is estimated with data from two adjacent time periods, and the estimation is a "rolling window" with the end result that parameters are allowed to vary over time. The shortcoming of this approach is that it is not possible to obtain standard errors for the estimates. We include a two-period adjacent period regression with a model model that includes spatial errors. This model is denoted by RWE-SEM.

We present formal specifications of the models used in the study.

3.1 Time Dummy Hedonic Model (TDH)

The TDH model is a multiple regression model where the dependent variable is typically the log of the sale price and the explanatory variables are hedonic characteristics (attributes) of the houses in the sample. The model includes time dummy intercepts which under the assumption of fixed hedonic attributes are proper price indices.

$$y = \gamma D + X\beta + \varepsilon \tag{1}$$

where,

N- Number of observations in the sample, that is, the total number of houses sold over the sample period;

y-$N \times 1$ vector of observations of the dependent variable, typically the log of sale price (p), $y = \ln p$;

β-$K \times 1$ vector of unknown parameters;

X-$N \times K$ matrix of independent variables which include house attributes as well as time dummy variables to estimate the price indices;

D- is a $N \times (T-1)$ matrix of $T-1$ year time-dummy variables;and

ε-$N \times 1$ vector of random errors.

This is traditional model used in the literature where year time dummy variables are included. The model has essentially $T-1+K$ parameter to be estimated. As the set of K parameters includes a constant term, we introduce only $T-1$ year time dummy variables.

3.1.1 Time Dummy Hedonic Model with Spatial Errors (SEM)

An extension of (1) to include a spatial correlation structure through the error term is given by:

$$\begin{aligned} y &= \gamma D + X\beta + \varepsilon \\ \varepsilon &= \rho W\varepsilon + u \end{aligned} \tag{2}$$

where,

u-$N \times 1$ vector of uncorrelated errors;

W-$N \times N$ matrix of spatial weights (that is, it is only a function of distance between houses in the sample);

ε-$N \times 1$ vector of correlated errors;

ρ-*scalar* spatial autocorrelation parameter, $|\rho| < 1$.

This model is an extension of (1) where the error structure follows a spatial errors model. Inclusion of spatial errors is designed to take explicit account of the role of locational characteristics in determining house prices. This model is particularly useful when the hedonic model does not include location in the regression model.

The matrix W has the following characteristics

- $w_{ii} = 0$ for al i
- w_{ij} weight representing the 'neighbour strength of the ith house with the jth house.
- **W** is a row-stochastic matrix, it has row sums of unity.

In this study we assume that housing prices are influenced by the prices of the nearest neighbours. The year of sale does not enter the construction of W. In order to identify the nearest neighbours, we make use of the Delauney triangulation method and the geocoded information on the latitude and longitude of the houses sold.

3.2 Rolling Window Spatial Errors Model (RW-SEM)

This is a regression model with fixed parameters and a spatial error structure. However, the parameters are allowed to vary over time through the re-estimation of the model over time. We pooled the data over two consecutive periods, and roll the sample. This is often referred to as the *rolling window* (RW) model[4].

$$\begin{aligned} y_\tau &= X_\tau \beta_\tau + \varepsilon_\tau \\ \varepsilon_\tau &= \rho W_\tau \varepsilon_\tau + u_\tau \end{aligned} \tag{3}$$

where,

$\tau = t + (t+1)$- two consecutive years of pooled observations of houses sold;

y-$(N_t + N_{t+1}) \times 1$ vector of observations of the dependent variable, typically the log of sale price (p), $y = \ln p$;

β_τ-$K \times 1$ vector of unknown parameters;

X_τ-$(N_t + N_{t+1}) \times K$ matrix of independent variables which include house attributes as well as time dummy variables to estimate the price indices;

ε_τ-$(N_t + N_{t+1}) \times 1$ vector of random errors.

[4]Though this model is intuitive and practical and a method recommended by Tripplet (2004), there is a logical inconsistency in the approach in that if parameters are the same for periods t and $t+1$ and then for periods $t+1$ to $t+2$ it should then imply that parameters in periods t and $t+2$ are identical and following this argument should lead to a TDH model. Notwithstanding this problem, we simply follow the literature and implement the RWE model.

We use the SEM specification accounting for possible spatially correlated errors. This flexible form of (3) is given by estimating (3) through an rolling window. For instance, in a two-adjacent periods overlapping window, estimates for a pooled sample of the first two periods is obtained first, periods two and three are then pooled together, three and fourth and so on. Two estimates of each period (except for the initial and last periods) are obtained through this procedure. In this paper we present the average value between the two estimates of each time period. As mentioned, a drawback of this approach is that we are unable to obtain standard errors for the estimates.

3.3 Time Varying Parameter Models (TV)

Now, we consider a more general specification where parameters are allowed to vary over time. If all the parameters are allowed to vary without a structure, the model is underidentified as there will be more parameters than the observations. Further, it is intuitive to consider the case when parameters move through time in a systematic manner and we use a random walk model where the parameters in period are a small (random) perturbation from the parameter values of the previous period. In the specification of the model, we make a distinction between the intercept and the slope parameters. The model is specified as:,

$$y_t = \mu_t + \sum_{k=1}^{K} X_{kt}\beta_{kt} + \epsilon_t, \;\; \epsilon_t \sim NID(0, \sigma_\epsilon^2 I_t) \tag{4}$$

$$\begin{aligned} \beta_t &= \beta_{t-1} + \eta_t \qquad \eta_t \sim N\left(0, \sigma_\eta^2 I_k\right) && (5)\\ \mu_t &= \mu_{t-1} + \xi_t, \;\; \xi_t \sim NID(0, \sigma_\xi^2) && (6)\\ E\left(\epsilon_t \eta_t\right) &= 0 && (7) \end{aligned}$$

where,

$t = 1, 2, ..., \tau$

N_t number of houses sold at time t.

$N = \Sigma_{t=1}^{\tau} N_t$

$\mathbf{X}_t$ is a (Kx1) vector of hedonic characteristics

The intercept level, μ_t, follows a separate stochastic process from the slope parameters (hedonic attribute parameters, β_t). This model has been denoted as a local level model with explanatory variables (Commandeur and Koopman (2007)). We denote this model by TV and note that it is in the form of a state-space model with state vector $\alpha_t = [\mu_t, \beta_t]$. Therefore the estimation of this model is straightforward using Kalman filtering and smoothing algorithms.

3.3.1 Time Varying Hedonic Model with Spatial Errors (TV_SEM)

A variation of the time-varying parameter model is the model where errors are assumed to be spatially correlated.

$$y_t = \mu_t + \sum_{k=1}^{K} X_{kt}\beta_{kt} + \epsilon_t, \;\; \epsilon_t \sim NID(0, \sigma_\epsilon^2 \Omega_t) \tag{8}$$

$$\begin{aligned} \varepsilon_t &= \rho W_t \varepsilon_t + u_t \qquad u_t \sim N\left(0, \sigma_u^2 I_{N_t}\right) && (9)\\ \beta_t &= \beta_{t-1} + \eta_t \qquad \eta_t \sim N\left(0, \sigma_\eta^2 I_k\right) && (10)\\ \mu_t &= \mu_{t-1} + \xi_t, \;\; \xi_t \sim NID(0, \sigma_\xi^2) && (11) \end{aligned}$$

where,

u_t-$N_t \times 1$ vector of uncorrelated errors;

W_t-$N_t \times N_t$ matrix of spatial weights (that is, it is only a function of distance between houses in the sample in period t);

ϵ_t-$N \times 1$ vector of correlated errors;

ρ-*scalar* spatial autocorrelation parameter, $|\rho| < 1$.

μ_t is the intercept level process

β_t is the vector of time-varying hedonic characteristics

We note here that the parameter ρ is assumed to be the same for all time periods, t. Similar to the case in (4), (5) and (6), this is also a state-space model. Although the error term of the measurement equation (8), ε_t, is spatially correlated, it is assumed to be uncorrelated over time and Gaussian and therefore satisfies the assumptions necessary to use the Kalman algorithms. That is, the measurement and state equations both have linear Gaussian forms. To show this, we incorporate equation (9) into the Kalman algorithms. We can transform the measurement equation (8) to

$$\begin{aligned} Y_t - X_t\alpha_t &= \rho W_t \left(Y_t - X_t\alpha_t\right) + u_t \\ \left(I_{N_t} - \rho W_t\right) Y_t &= \left(I_{N_t} - \rho W_t\right) X_t\alpha_t + u_t \\ \widetilde{Y}_t &= \widetilde{X}_t\alpha_t + u_t \end{aligned} \tag{12}$$

where $\widetilde{Y}_t = \left(I_{N_t} - \rho W_t\right) Y_t$, $\widetilde{X}_t = \left(I_{N_t} - \rho W_t\right) X_t$ and $\alpha_t = [\mu_t, \beta_t]$, and we obtain (12), a linear Gaussian form since $u_t \sim N\left(0, \sigma_u^2 I_{N_t}\right)$. Second, setting $\varepsilon_t = \left(I_{N_t} - \rho W_t\right)^{-1} u_t$, we see (8) has linear Gaussian form since $\varepsilon_t \sim N\left(0, H_t\right)$ where

$$H_t = \sigma_u^2 \left(I_{N_t} - \rho W_t\right)^{-1} \left(I_{N_t} - \rho W_t\right)^{-1\prime}. \tag{13}$$

The state-space model is then given by (14) and (15):

$$\widetilde{Y}_t = \widetilde{X}_t \alpha_t + u_t \tag{14}$$

$$\alpha_t = \alpha_{t-1} + \zeta_t \tag{15}$$

where,

$$\zeta_t = \begin{bmatrix} \epsilon_t \\ \eta_t \end{bmatrix}$$

3.3 Econometric Estimation

The estimation of all the models in this paper were performed using Matlab. The TDH model is a simple least squares model. The TDH_SEM model was estimated using the maximum likelihood based routines in the Spatial Econometrics Matlab Toolbox. The TV and TV_SEM models were estimated using code specially written by the first author. In both cases hyperparameters (that is, constants of proportionality,σ_i^2-$i = \epsilon, \eta, ..$, and the spatial parameter ρ) were estimated using the standard approach for the estimation of state-space models based on numerical maximization of the conditional likelihood function (see Harvey (1989), Chapter 3 for example) and the use of the Kalman filter and smoothing algorithms. The estimates of ρ from the TDH_SEM and TV_SEM are very close, $\hat{\rho} = 0.46$ for the first and $\hat{\rho} = 0.48$ for the TV_SEM. This is very reassuring as the spatial correlation parameter is assumed to be constant across time and space. In the next section the performance of the models are compared using a mean squared prediction approach.

3.4 Model Results

For the estimation of DH and TDH-SEM the sample is pooled (1985-2005). These models are not expected to perform well in prediction; however, they will serve as base models for the purpose of comparison. To study their performance we compute the Root Mean Square Prediction Error (Root MSPE) of each alternative model in their prediction of individual log transform of sale price. For the RWE-SEM we pool two years and overlap one year as we move through the sample (that is 1985 and 1986, 1986 and 1987,..). The TV and TV-SEM are estimated with monthly transactions ($\tau = 252$ for the period 1985:1 to 2005:12).

3.4.1 Comparative Performance of Alternative Models

From the fitted models we compute the Root Mean Square Prediction Error (RMSPE), and the results are in presented in Table 1.

Table 1: Root Mean Square Prediction Error - Prediction of ln(Sale Price)

Model	No Spatial Effects		Spatial Effects		
	Time Dummy (TDH)	Time Varying (TV)	Time Dummy (TDH-SEM)	Rolling window Model (RW-SEM)	Time Varying (TV-SEM)
Root MSPE	0.4224	0.4153	0.4214	0.4263	0.3780
reduction/ increase from base model	base	-1.7%	base	1.2%	-11.5%

In the table we present the models in two groups depending on whether spatial effects were considered in the modelling and estimation. The TDH is the base model for the no spatial effects case, and we see that allowing the hedonic parameters to vary over time results in a reduction in RMSPE of 1.7%. For the model with spatial effects, the base model is SEM. We note that SEM has a lower RMSPE than TDH. However, a surprising finding is that relaxing the fixed parameters assumptions by implementing a adjacent period rolling window (RWE_SEM) results in an increase and not a decrease in RMSPE. The compounding effect of spatial errors and time-varying hedonic parameters in TV-SEM results in a large reduction in RMSPE (11.5%) over the SEM model's performance.

These results indicate there are gains to be made by using time-varying parameters; however, using adjacent periods regression might not result in any gains in predictions. The case for our data is that the use of a rolling window increases the prediction error.

3.4.2 Coefficient Estimates from TV-SEM and Rolling Window (two years) SEM Model

In this section we compare the estimates of hedonic coefficients associated with number of bathrooms, bedrooms and land. We also present estimates of the constant term in the time-varying SEM as well as the rolling window SEM. The general observation is that the estimated coefficients from the RW-SEM do not perform well. We present our estimates from these two models in the following figures.

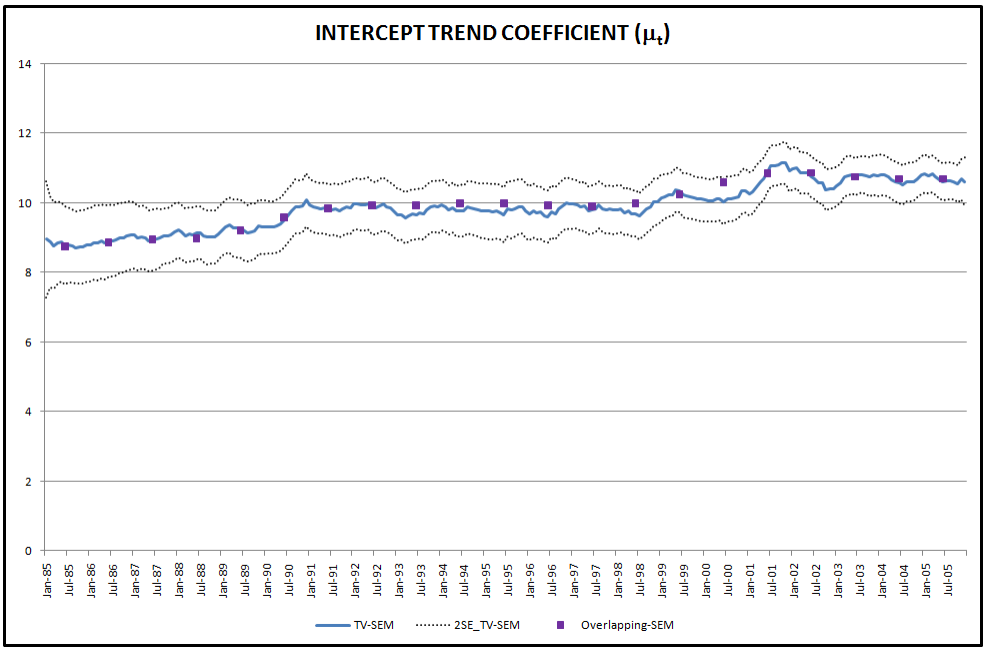

Figure 2: TV-SEM. Intercept Coefficient

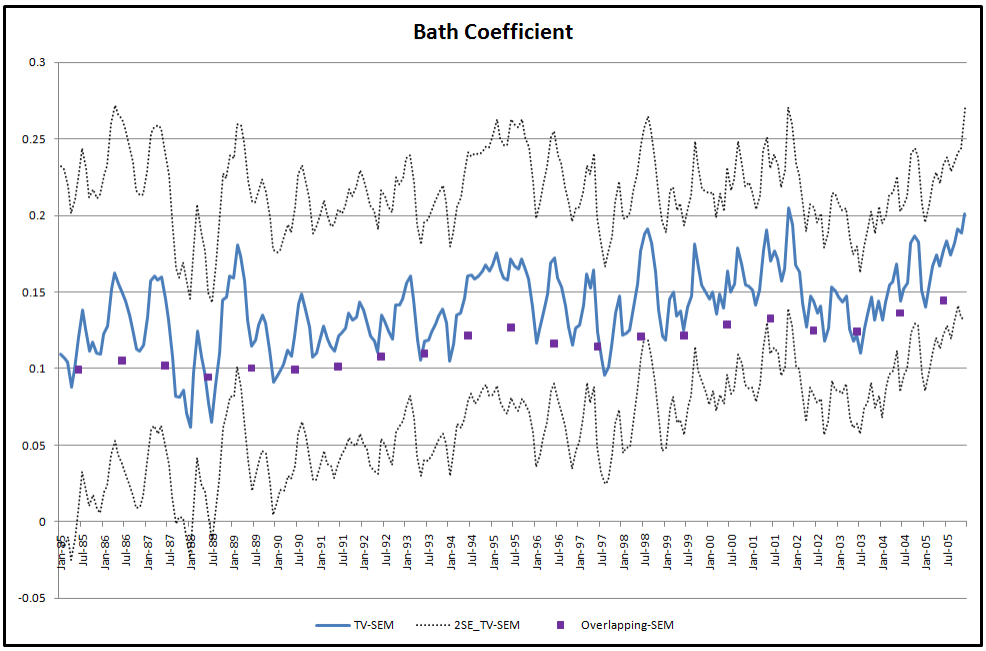

Figure 3: TV-SEM. BATH Coefficient

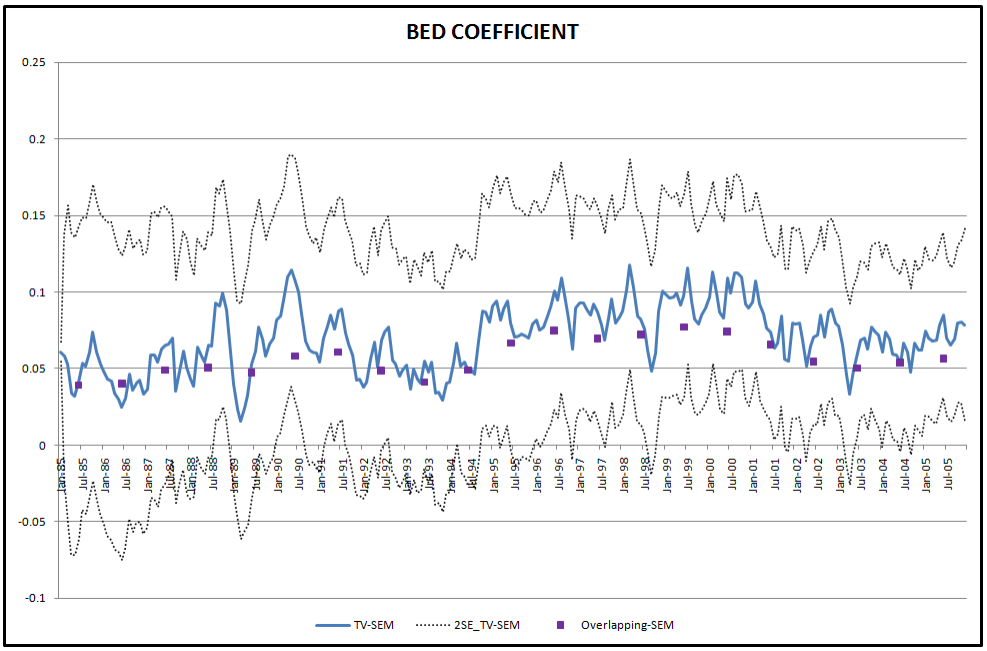

Figure 4: TV-SEM. BED Coefficient

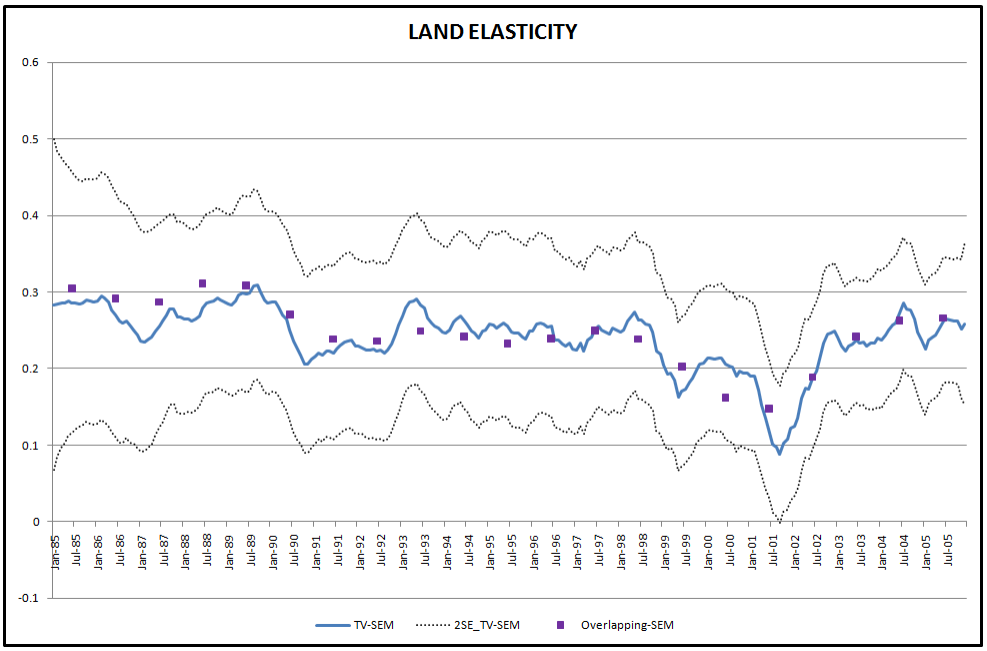

Figure 5: TV-SEM. Land Coefficient

From all the figures it is clear that the RW-SEM parameter estimates lie within the 2-standard error interval around the TV-SEM model. However, we find serious discrepancies between the two sets of estimates. In Figure 2 we have the intercept trend in the log-linear model. While the RW-SEM estimates of the intercept appear to track the more general and conceptually superior TV-SEM model there are certain periods (July 93 to July 2000) when the RW-SEM intercepts are higher than those derived using the TV-SEM model. Given the relative magnitude of

the intercept coefficient, it is clear that these differences will have a significant impact on the predictive performance of these two models. From Table 1 we can see that that RW-SEM produces predictions with the highest root mean square prediction error. The poor performance of the RW-SEM may also be attributable to estimates of parameters of the slope coefficients from the RW-SEM. Estimates of hedonic coefficients for the bathrooms and bedrooms, in Figures 3 and 4 respectively, from the TV-SEM model are a lot more volatile than the coefficients from the RW-SEM which is consistent with the RW-SEM approach which implies relative constancy of parameters through time. From Figure 5 we find that house prices are relatively inelastic with respect to the land size. This is a somewhat surprising result as the cost of land is a major component of the price of the house. A careful examinations shows that a possible reason for this result is the lack of variability in the size of land. In fact, most of the blocks of land on which dwellings in Brisbane are found are around 670 square meters. With a few exceptions, the RW-SEM appears to perform reasonably well with respect to the land coefficient for most of the periods.

3.5 The Evolution of Prices

The model with the lowest RMSPE is the TV_SEM. We produce house price predictions for the houses sold in a particular period using the model:

$$\widehat{\ln p_t} = \hat{\mu}_t + \hat{\beta}_{1t} \ln land_t + \hat{\beta}_2 BED_t + \hat{\beta}_3 BATH_t + \hat{\beta}_4 CARLUP_t + \hat{\rho} W_t \hat{\epsilon}_t \tag{16}$$

where,

$\hat{\epsilon}_t$ is the GLS residual from the estimated model TV_SEM (in Section 3.3.1).

Due to the log-log nature of the model in terms of the land area, it was not possible to predict the "land component" of the dwelling price. Therefore we focus on the predicted price of the whole house inclusive of the land component. In Figure 6 we presents an estimate of the median monthly price for the sample period. The estimates are obtained as follows:

$$\hat{p}_t^m = median(exp(\widehat{\ln p_t}))$$

For each period, we compute the median of the predicted prices of all the houses sold in that period computed using our preferred TV-SEM. This slightly different, but conceptually superior, to the normal practice of computing predicted price of a house using median values of hedonic characteristics (land, bedrooms and bathrooms) in different months. As expected, the use of median values of characteristics produced a much more volatile series of median prices. Given the superior predictive performance of the TV-SEM, the observed and predicted median house prices are closely aligned over the period.

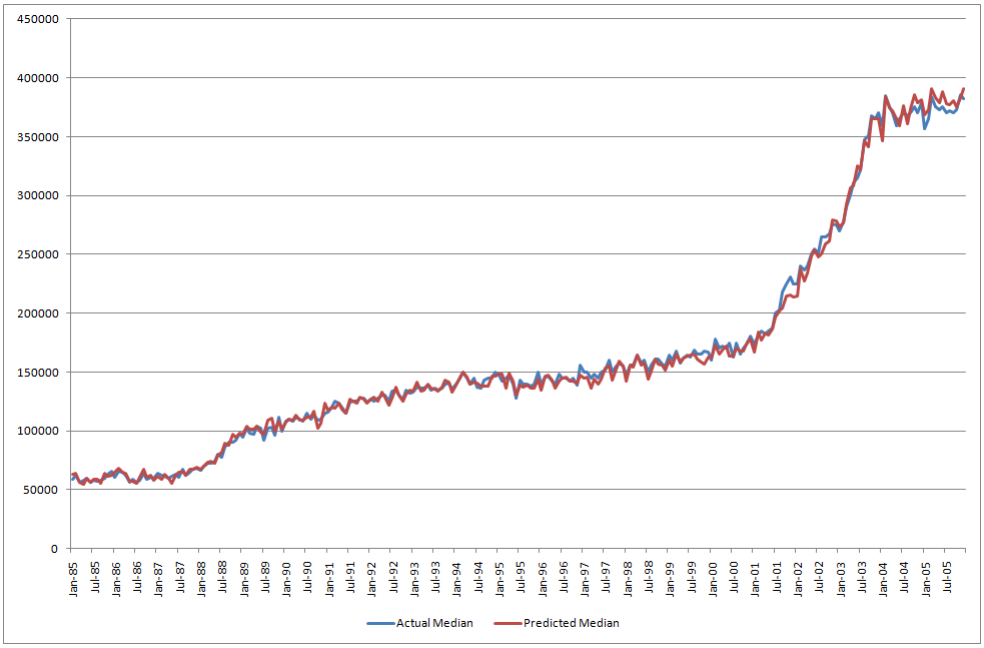

Figure 6: Prediction of median sale prices over the sample period

Figure 6 provides an interesting profile of prices of houses sold in Brisbane over the last two decades. Treating the median prices as an observed time series from 1985 to 2005, we can see that there are several structural breaks in the price series. After a relative stable period until July 1988, a surge in house prices is evident over the two year period until July 1989. Over the next decade from January 1990 until January 2001 there has been a steady increase in median prices from just above 100,000 dollars to 175,000 dollars. There has been a sharp rise in house prices from January/July 2001 until January 2004 where the prices had more than doubled. We have conducted formal time-series tests for structural breaks and the visual trends are strongly supported by the econometric results. Joint tests for two structural breaks in early 2001 and late 2004 are significant at the 5% level.

4 Hedonic Imputed Indices for Housing

In this paper we report several sets of hedonic imputed price index numbers for housing. HM provide an extensive discussion of a range of index number formulae that are based on different sets of weighting systems and using different sets of imputed prices. The general conclusion by HM is that it is best if imputed prices are used for both current and base periods instead of using imputed prices only for the current period. In addition they recommend the use of value shares should be based on actual sale prices instead of imputed prices. We basically follow these recommendations.

As a deviation from the general practice in this area, we construct price index numbers with *plutocratic and democratic weights.* Plutocratic weights reflect the prices of different houses and higher priced houses are accorded higher weights in the index construction. These are essentially value shares of different houses sold at a given point

of time. The use of plutocratic weights[5] along with a Laspeyres type index (as in equation 19) measures the price change by comparing the total value of the housing stock in the base period and current period using hedonic imputations. Similarly the Paasche index compares the housing stock of the current period at the base and current period prices. However, the geometric indices like the Tornqvist indices cannot be interpreted along the same lines. All the variants discussed in HM are essentially plutocratic indices.

In contrast, the democratic weighting system gives the same weight for each house sold in the market at any given point of time. Therefore, the use of democratic weights leads to unweighted arithmetic or geometric averages of imputed price relatives. The use of democratic weights essentially stems from the use of a stochastic approach where the houses sold in in any given area is taken as a random sample and therefore the price observations are assumed to have the same variance. However, when the geometric Tornqvist index is computed, we explicitly recognise the unequal numbers of houses sold in the two years and defined a weighted geometric mean of the geometric Laspeyres and Paasche indices (see equation 24).[6] The use of democratic weights is appropriate if the principal aim is to generate a statistically sound estimator of the central tendency of the distribution of houses. Given that that the expenditure weights used in hedonic imputed price indexes do not have the same theoretical basis as the expenditure shares used in the construction of the consumer price index, the choice between the plutocratic and democratic weights really depends upon the main objective behind the housing price index construction.

4.1 Index Number Formulae Used

Let $\hat{P}_t^h$ represent the imputed price of house h in period t. Further, let $\hat{w}_t^h$ be the value share of the house h defined as:

$$\hat{w}_t^h = \frac{\hat{P}_t^h}{\sum_{n=1}^{H_t} \hat{P}_t^n} \tag{17}$$

where,

$\hat{P}_t^h$ is the imputed price. Typically in our case t refers to a particular month as we are making use of monthly sales data. Construction of annual indices is considered in Section 4.2.

We define the following types of indices used in the study.

PLUTOCRATIC INDICES: These indices are weighted indices where weights represent the relative value of each of the houses included in the sample. In the paper we use the Fisher and Tornqvist variants of this index. These indices are computed using:

(i) *actual shares* instead of shares based on imputed prices; and

[5] This type of interpretation holds only when the expenditure share weights are also based on imputed prices.

[6] It is possible to consider a more sophisticated approach after stratifying the sample into different regions and by the type of houses. However, we are yet to implement the stratified sampling approach.

(ii) imputed prices in both the base and current periods.

Hence, the Hedonic Imputed price indices for period t with period s used in our study are defined as follows.

The **Fisher index** (F) is defined as:

$$F_{s,t}^{P} = \sqrt{L_{s,t}\, P_{s,t}} \tag{18}$$

where $L_{s,t}$ is the Laspeyres index and $P_{s,t}$ is the Paasche index with definitions:

$$L_{s,t} = \sum_{h=1}^{H_s} w_s^h \left(\frac{\hat{P}_t^h(x_s^h)}{\hat{P}_s^h(x_s^h)} \right) \tag{19}$$

$$P_{s,t} = \left[\sum_{h=1}^{H_t} w_t^h \left(\frac{\hat{P}_s^h(x_t^h)}{\hat{P}_t^h(x_t^h)} \right) \right]^{-1} \tag{20}$$

with the value shares defined as in (17).

We note here that if the shares are based on predicted prices then the Laspeyres and Paasche indices defined in (19) and (20) simply turn out to be *ratios of the value of the stock of houses* in periods t and s respectively evaluated at the hedonic price models in these periods. Thus the index in (18) simply measures the change in the value of the housing stock due to changes in prices as reflected in the hedonic model of prices.

The **Tornqvist indices** are defined similar to equations (18), (19) and (20). Following HM, we define these indices as follows:

$$T_{s,t}^{P} = \sqrt{GL_{s,t} \cdot GP_{s,t}} \tag{21}$$

where GL and GP are the geometric Laspyeres and geometric Paasche indices which are defined as:

$$GL_{s,t}^{P} = \prod_{h=1}^{H_s} \left[\frac{\hat{P}_t^h}{\hat{P}_s^h} \right]^{w_s^h} \tag{22}$$

$$GP_{s,t}^{P} = \prod_{h=1}^{H_t} \left[\frac{\hat{P}_t^h}{\hat{P}_s^h} \right]^{w_t^h} \tag{23}$$

These indices are "plutocratic" and are influenced by houses with large price tags. Despite this, the Fisher and Tornqvist indices in (18) and (21) measure the changes in the housing stock values that can be attributable to price changes.

We now deviate from the HM approach and define democratic indices which are statistically more meaningful measures of price changes.

DEMOCRATIC INDICES: Consistent with the use of a log-price hedonic model, we focus on the geometric

Laspeyres, Paasche and Tornqvist indices. These are defined as:

$$T^D_{s,t} = \sqrt{GL^D_{s,t}\, GP^D_{s,t}} = \sqrt{\left[\prod_{h=1}^{N_s}\left(\left[\frac{\hat{P}^h_t(x^h_s)}{\hat{P}^h_s(x^h_s)}\right]^{\frac{1}{N_s}}\right)\right]\left[\prod_{h=1}^{N_t}\left(\left[\frac{\hat{P}^h_t(x^h_t)}{\hat{P}^h_s(x^h_t)}\right]^{\frac{1}{N_t}}\right)\right]} \tag{24}$$

The democratic index provides a better measure of price change that is consistent with the distribution of price relatives. The distribution of the prices is likely to be skewed and the use of geometric mean is consistent with a general log-normal distribution of price relatives.

4.2 Annual Chained indices

In the presence of seasonality we want to consider how to construct chained indices. From an index number perspective, chaining may be undesirable when it leads to index drift. Szulc (1983) made the point that when prices or quantities oscillate ('bounce'), chaining can lead to considerable index drift: that is, if after several periods of bouncing, prices and quantities return to their original levels, a chained index will not normally return to unity. Hence, the use of chained indices for noisy monthly or quarterly series is not recommended.

In view of the drift caused by chaining in the presence of oscillations and in view of the presence of in the sales of houses and the types of houses sold it may be better if we compute month-on-month housing price indexes and combine them to yield a year-on-year indexes. The following methods are drawn from chapter 22 of the ECE-ILO Manual on the Consumer Price Index (ILO, 2006).

4.2.1 Yule (1921)'s method (page 8, Chapter 22, ILO, 2006)

Step 1: Compute the year-over-year monthly index for each month using a standard index number formula. In our case we can use Fisher and Tornqvist indexes with plutocratic and democratic weights.

Step 2: The year-on-year index is then computed as a simple unweighted geometric mean of the month-on-month index

4.2.2 Stone (1965)'s index (pp. 15-16, Chapter 22, ILO, 2006)

Step 1: Compute the year-over-year monthly indexes using standard index number formulae.

Step 2: Compute the year-on-year annual indices as follows:

$$L_{t,t+12} = \sum_{m=1}^{12} \sigma^t_m L_{t,t+12,m} \tag{25}$$

$$P_{t,t+12} = \sum_{m=1}^{12} \sigma^{t+1}_m P_{t,t+12,m} \tag{26}$$

$$F_{t,t+12} = \sqrt{L_{t,t+12} P_{t,t+12}} \tag{27}$$

where,

$$\sigma_m^s = \frac{\sum_{h \in H_m^s} p_h^{s,m} \cdot q_h^{s,m}}{\sum_{m=1}^{12} \sum_{h \in H_m^s} p_h^{s,m} \cdot q_h^{s,m}} \quad \text{with} s = t, t+12$$

are the value shares of houses sold in different months.

Step 3: Step 2 provides plutocratic indices. Weights in the Laspeyres and Paasche indices can be replaced by the number of houses sold in different months.

Step 4: We can use geometric versions of these formulae leading to Tornq*v̈*ist indices.

4.2.3 Annual Housing Price Indices

In this section we present annual housing price index which provide a measure of changes in housing price indexes from one year to the next starting from 1985. We present indices based on the application of the *time dummy hedonic* (TDH) model as well as its extension that accounts for the presence of spatial correlation of errors, TDH-SEM, generated through locational characteristics. Even though the *rolling window* (RWE_SEM) method is quite popular in the hedonic price index literature, we found the RWE_SEM method to be the least performing model in terms of its predictive power within the sample. As hedonic price indexes depend upon imputed prices of houses sold in different time periods, the use of RWE method could introduce serious biases. As the main focus of the paper is on *time-varying* (TV) *parameter hedon*ic models we present several indices based on the TV model and the TV-SEM which accounts for spatially correlated errors. We note here that the annual price indices from the TDH and TDH-SEM models are automatically given by the estimates of the parameters of the dummy variables contained therein, there is no need to use any specific index number formula. In contrast, when the TV and TV-SEM models are used we need to decide on whether a Fisher or Tornqvist index number formula is used and whether we compute these two indices using plutocratic or democratic weights.

In Figure 7, we present chained annual housing price index numbers from different models and computed using different index number formulae. We also present chained annual index computed using the median price observed in each year. The median price index provides a frame of reference. All the indices are computed using 1985 as the base year. A striking feature of the indices in the figure is that the TDH and TDH-SEM model based indices are smooth concave functions of time without any points of inflexion. This is in contrast to all other indices which clearly show several phases in the acceleration of the price indexes. We note a clear acceleration of prices over the short period 1986 to 1990 and a smooth increase until 1994 followed by a stable period until 2000 where a noticeable acceleration has taken place until 2004. Therefore, the use of time dummy hedonic approach is likely to provide an indication of the general trends over a long period but appears to mask more interesting trends over sub-periods. As the Fisher and Tornqvist indices are both superlative [7]and in most empirical studies tend to be numerically close we expect the same result in our case. From Figures 7, 8 and 9 we find that the Fisher and Tornqvist plutocratic

[7]See Diewert (1976) for more details on *exact* and *superlative indices*.

indices are almost identical consistent with out expectations. Another interesting feature of the results is that the Tornqvist indices based on plutocratic and democratic weights appear to differ thereby indicating the need to be clear about the meaning and interpretation of the index generated. All the chained indexes are significantly below the median-based chain index of housing prices indicating an upward bias in the median price indexes normally reported in the popular press.

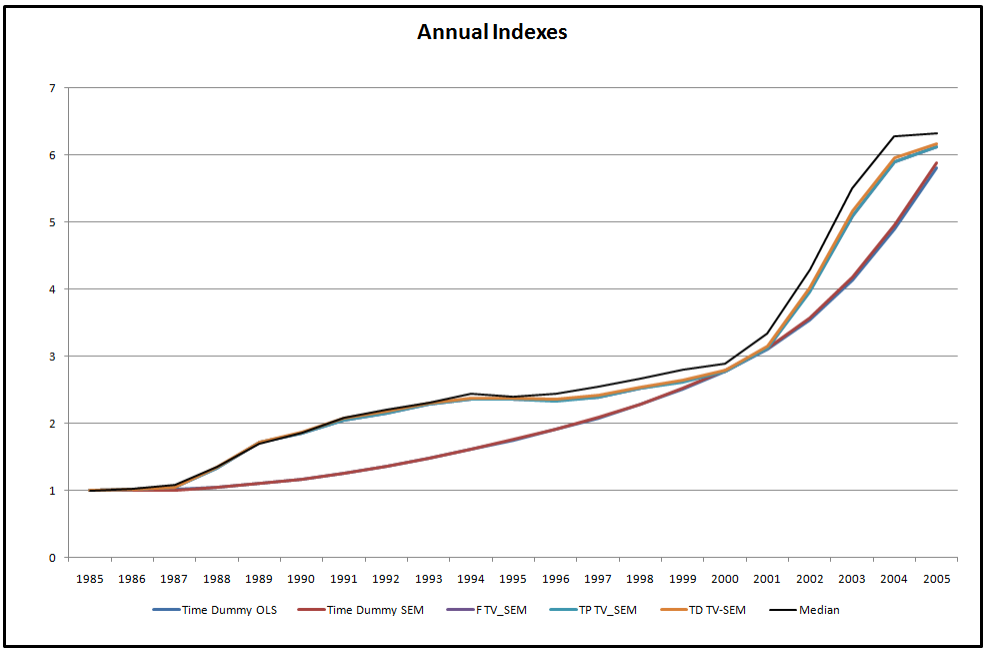

Figure 7: Annual Indices - Chained

4.3 Monthly Chained Indices from TV-SEM Model

From the annual price indexes we now turn to chained prices indexes constructed using month-to-month price indexes. In this section we mainly focus on the plutocratic and democratic weighted price indexes computed using the Fisher and Tornqvist index numbers. The Median housing price indexes are also presented. Based on the results reported in Figure 7, we do not present price indexes based on time dummy methods which implicitly assume constancy of parameters of the hedonic model. As the *time-varying parameter model with spatial errors* (TV-SEM) has the best predictive power within the sample period, we present only results based on the TV-SEM model.

Figure 8 presents chained monthly indices with January, 1985 as the base computed using the plutocratic Fisher and Tornqvist indices and the democratic weighted Tornqvist indices. The median housing price index is also presented. By the end of the study period, there has been a significant difference between the median and the hedonic price index numbers and of the magnitude of 20 to 30 percent higher when median is used relative to the Fisher and Tornqvist indices. We also note that the democratic weighted Tornqvist index is uniformly higher than the plutocratic weighted index but the percentage difference is much smaller compared to the media based price

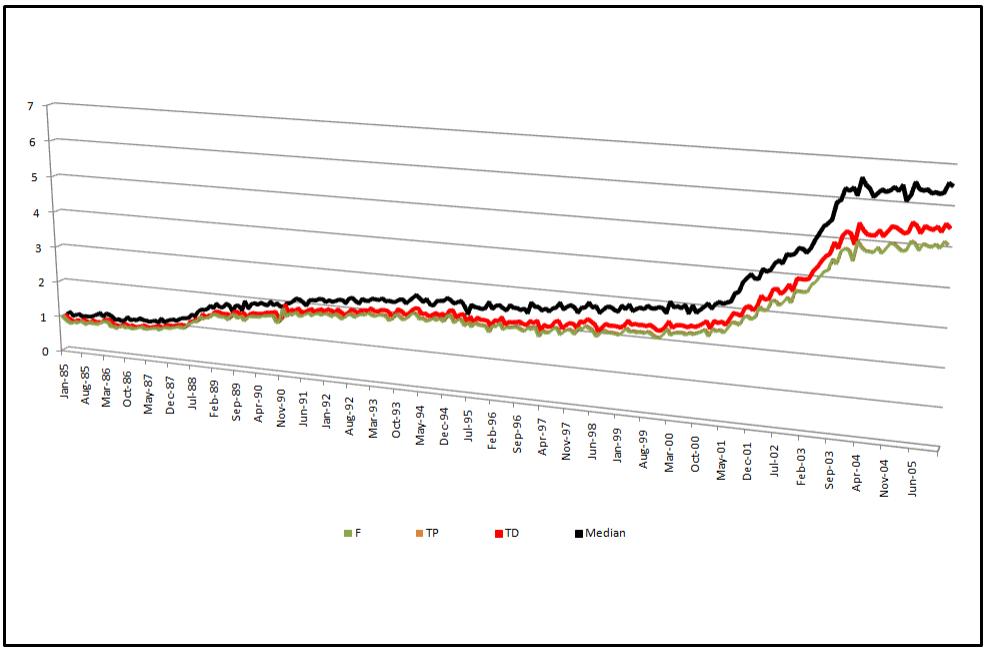

Figure 8: Chained Monthly Indices. F:Fisher Plutocratic, TP: Tornqvist Plutocratic, TD: Tornqvist Democratic for the period 1985:1 to 2005:12

index.

As with the annual chained price indices, we note the presence of three episodes of acceleration in the housing prices. However, there is evidence of seasonal fluctuations in the indices but we do not notice any major drift in the indexes. In order to facilitate visual examination of the differences, we split the period into two periods, 1985 to 1995 and from 2001 to 2005 and present the indices in Figures 9 and 10 separately for the two periods. We found these periods to represent periods of accelerated increases in prices. From Figure 9 we observe that there are several periods during which the trends in the index of median prices and the hedonic index are in the opposite direction which is a clear indication of the influence of the effect of the mix of houses sold in different periods. However, all the indices are much more closely aligned during the period 2001 to 2005 the differences between the median and hedonic price indexes and this is in sharp contrast to the significant deviations between these two sets of indices observed during 1985 to 1995 period. The close alignment observed during 2001 to 2005 when the house prices experienced significant rises is that the housing price boom during this period was uniform across all types of houses sold which in turn implies that the mix of houses sold will not significantly affect the housing price indexes. This is an aspect that requires further analysis[8].

[8]Tests for stationarity in the presence of structural breaks were not conducted for this version of the paper but will be included in the next draft

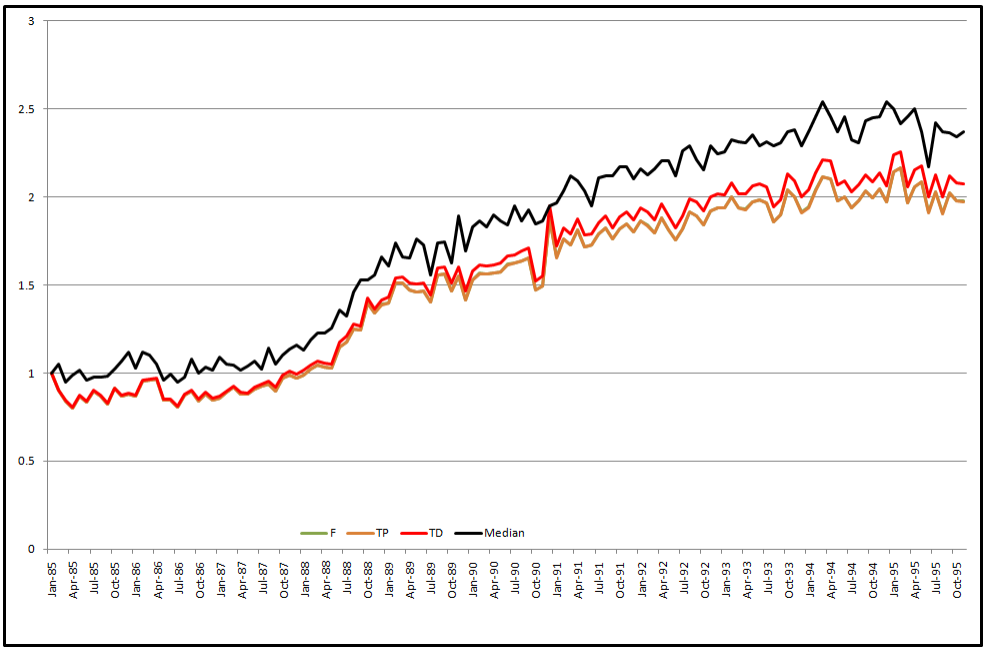

Figure 9: Chained Monthly Indices. F:Fisher Plutocratic, TP: Tornqvist Plutocratic, TD: Tornqvist Democratic for the period 1985:1 to 1995:12

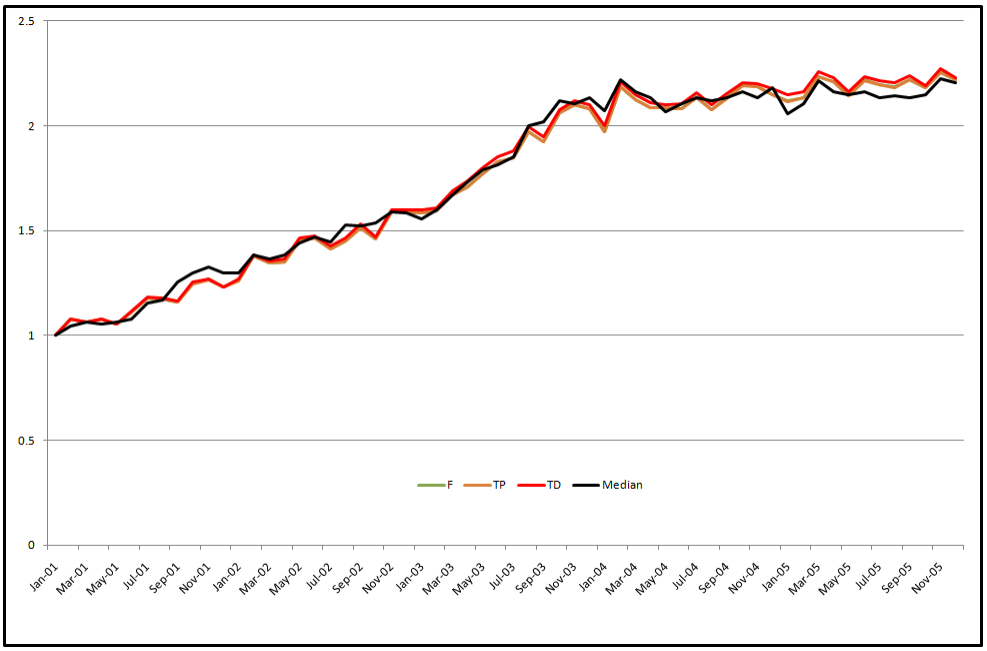

Figure 10: Chained Monthly Indices. F:Fisher Plutocratic, TP: Tornqvist Plutocratic, TD: Tornqvist Democratic for the period 2000:1 to 2005:12

5 Conclusions

The paper has focused on several important issues relating to hedonic modelling of housing prices and their use in the construction of housing price index numbers. First, the paper focuses on the issue of econometric specification and highlights the need to model the time-varying nature of the hedonic coefficients and also the importance of making optimal use of the information on the influence of locational characteristics available in the form of spatially autocorrelated errors. A related issue is the problem of choosing the best specification. We make use of the root mean squared error of prediction of houses sold in different periods. The second objective of the paper is to examine the effect of using various hedonic models on the housing price index numbers. We also focus on the influence of the weights, plutocratic versus democratic weights, and on the chained annual indexes. Finally, we examine the month-to-month housing price indexes based on time-varying hedonic regression models to examine the general trends, seasonality, stationarity and the presence of structural breaks in the index series. The empirical analysis of the paper is based on housing price data from the city of Brisbane in Australia for the period 1985 to 2005. The analysis clearly demonstrates the predictive power of the time-varying hedonic model with spatially correlated errors and we also show that the worst performing model is the rolling window approach recommended in the hedonic price index literature. We also find that the use of time dummy approach is likely to mask important underlying movements and features of the hedonic price index numbers. It is not clear if this applies mainly to our Brisbane sample but it is likely that this is an intrinsic feature of the time-dummy hedonic price index numbers. In general, the median price index provides an upper-bound and clearly well above the price indexes from all the other approaches with the possible exception of periods of rapid and uniform price increases. We find that the median housing price index significantly diverges during the period 1985 to 1995 but seems to align quite well with the hedonic price indexes during the period 2001 2005. This result is particularly interesting as the housing market experienced a price boom during this period. We attribute this feature to the possibility that house price increases were uniform across different types of dwellings and in different locations. Trends in the chained annual price indexes as well as chained monthly price indexes clearly show three phases of housing price acceleration during the study period. These periods are consistent with the anecdotal evidence on house prices in Brisbane during this period.

References

Cominos, H. (2006), "Estimation of House Prices and the Construction of House Price Index Numbers: A new methodology applied to the Brisbane Metropolitan Area," University of Queensland, A Thesis submitted to the School of Economics in partial fulfillment for the Degree of Bachelor of Economics (Honours), in the field of Econometrics.

Cominos H., A.N. Rambaldi, and D.S. Prasada Rao (2007), "Hedonic Imputed Housing Price Indices from a Model with Dynamic Shadow Prices Incorporating Nearest Neighbour Information," in ESAM07, 2007 Australasian Meeting of the Econometric Society. `https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=ESAM07&paper_id=181`.

Commandeur J.J.F and S. J. Koopman (2007), An Introduction to State Space Time Series Analysis. Oxford University Press. New York.

Diewert, W.E. (1976), "Exact and Superlative Index Numbers", *Journal of Econometrics*, 4, 114-145.

Diewert, W.E. (2001), "Hedonic Regressions: A consumer Theory Approach", Discussion Paper 01-12, Department of Economics, University of British Columbia, Vancouver, Canada.

Harvey, A. C (1989), Forecasting, Structural Time Series Models and the Kalman Filter, Cambridge.

Hill, R. J. and D. Melser (2008), "Hedonic Imputation and the Price Index Problem: An Application to Housing," *Economic Inquiry*, 46:4, 593–609.

Silver, M. (2007), "The Difference Between Hedonic Imputation Indexes and Time Dummy Hedonic Indexes", *Journal of Business & Economic Statistics*, 2007, 25, 239-246.

Syed,I., R.J. Hill and D. Melser (2008), "Flexible Spatial and Temporal Hedonic Price Indexes for Housing in the Presence of Missing Data," School of Economics Discussion Paper: 2008/14. Australian School of Business. University of New South Wales.

Svetchnikova, D. (2007), Spatial-Temporal Modeling of the Real Estate Market, University of Queensland, A Thesis submitted to the School of Economics in partial fulfillment for the Degree of Bachelor of Economics (Honours), in the field of Econometrics.

Svetchnikova, D., A. N., Rambaldi, R. Strachan, "A Comparison Of Methods For Spatial-Temporal Forecasting With An Application To Real Estate Prices," in ESAM08, 2008 Australasian Meeting of the Econometric Society. `http://nzae.org.nz/conferences/2008/110708/nr1215386413.pdf`

Szulc, B. (1983), "Linking Price Index Numbers" in Diewert and Montmarquette (eds.) Price Level Measurement, Statistics Canada, Ottawa, Canada.

Triplett, J. (2004), "Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products", OECD Science, Technology and Industry Working Papers, 2004/9, OECD Publishing. doi:10.1787/643587187107