



**Centre for Efficiency and Productivity Analysis**

**Working Paper Series  
No. 01/2007**

**Title**  
**Hedonic Imputed Housing Price Indices from a Model with Dynamic Shadow Prices  
Incorporating Nearest Neighbour Information**  
**Authors**  
**Harry Cominos, Alicia Rambaldi and D.S. Prasada Rao**

**School of Economics  
University of Queensland  
St. Lucia, Qld. 4072  
Australia**

ISSN No. 1832 - 4398

**Hedonic Imputed Housing Price Indices from a Model with Dynamic Shadow Prices  
Incorporating Nearest Neighbour Information<sup>†</sup>**

**Harry Cominos, Alicia Rambaldi\* and D.S. Prasada Rao**  
School of Economics, University of Queensland  
Brisbane 4072, Australia

*This paper has been presented at:*

*2006 Measurement Workshop, UNSW (December 2006)*  
*Reserve Bank of Australia, Seminar (February 2007)*

*Copyright 2006 by H. Cominos, A. N. Rambaldi and D.S.P. Rao. All rights reserved.  
Readers may make verbatim copies of this document for non-commercial purposes by  
any means, provided that this copyright notice appears on all such copies.*

<sup>†</sup> *The empirical results for the SSSEM model in Section 6 are only preliminary. The  
authors gratefully acknowledge funding provided by UQ Research Development Grant,  
“A Unified Approach to the Computation of Housing Price Indices: Hedonic Imputation  
Indexes vs Time Dummy Hedonic Indexes”*

\* Corresponding author: [a.rambaldi@economics.uq.edu.au](mailto:a.rambaldi@economics.uq.edu.au)

# **Hedonic Imputed Housing Price Indices from a Model with Dynamic Shadow Prices Incorporating Nearest Neighbour Information**

*Harry Cominos, Alicia Rambaldi and D.S. Prasada Rao*

## **Abstract**

The main objective of this paper is to propose an improvement to existing methods of house price index construction by addressing three important oversights in the literature. Firstly, it is plausible that the shadow prices of property attributes evolve slowly over time, in line with the nature of consumer preferences. Existing methods either assume (explicitly or implicitly) that shadow prices are constant or change in a haphazard fashion from one period to the next. Secondly, the price of a house is spatially correlated with the price of houses that are in close proximity. Thirdly, there has been little research into the accurate specification of index number formulae particular to the housing case, which is partly due to existing methods implicitly defining an index within their formulation. However, improvements in the comprehensiveness of housing data allows for methodologies that consider a wider spectrum of index formulae.

This paper specifies a hedonic model with smoothly time-varying parameters which is estimated using a state-space formulation with spatially autocorrelated errors. Subsequently, the estimated model is used in constructing a housing price index using hedonic imputation (HI) methods. The formulated HI indexes are also applied to the traditional spatial errors model, which is new to the literature. The methods proposed in this paper are illustrated using housing price data for the Brisbane metropolitan area, compiled from the property information service, *RP Data*.

*Keywords:* Housing; Price Index; Hedonic Regression; Spatial Errors Model; State-Space Model;

*JEL Classification:* C43, E31, O47, R31, Y40

## 1. Introduction

The paper deals with the construction and estimation of house price index numbers with an application to housing data for the Brisbane Metropolitan Area. A general/common approach is to use median house prices to measure price changes. Three main issues can be identified with this approach, and there are a few methodologies developed in the literature to resolve these issues. Firstly, the composition of the transacted houses may be substantially different to the composition of the total housing stock (the ‘compositional change’ problem). This occurs because the sample of houses sold in a given time period is often not indicative of the population of houses. Secondly, there are quality changes in houses over time (the ‘quality change’ problem) due to refurbishments, renovations and newly built houses, as well as depreciation effects due to wear and tear. Among existing house price index methodologies, the hedonic approach is ‘preferred’ because both the aforementioned issues can be accounted for. The trade-off of this approach is that it is more data intensive, ideally requiring information on property attributes and neighbourhood variables (such as quality of schooling etc). This draws attention to the third issue relevant for Australia – the lack of (available) accurate and comprehensive data on housing. This last issue is of particular concern because resulting index numbers will be biased regardless of the methodology used if the data are unsuitable.

The current literature on the hedonic approach mainly focuses on the specification of the hedonic function and the application of the estimated hedonic function to construct house price index numbers, ignoring the (potentially) spatially correlated nature of house prices. Likewise the preferences for various property attributes and other relevant housing variables have typically been modeled as constant over time or varying in a haphazard fashion. In this paper we consider the computation of housing price indexes based on the traditional time dummy hedonic model as well as several models with time varying parameters and spatially correlated errors. Adjusting for the spatial correlation of house prices is particularly advantageous when data are not available to specify and fit a hedonic model sufficiently well (as is the case for housing in Australia). In this paper a *spatial errors model* (SEM) is specified and estimated for every time period separately.

In addition, the paper then attempts to impose a time-varying structure upon the hedonic coefficients using a state space formulation of the hedonic function. This combined model (the SSSEM) incorporates prior information regarding the preferences of consumers in the housing market – namely, that preferences are dynamic and evolve slowly over a relatively lengthy period of time.

The data used in the empirical analysis are from the property information service ‘RP Data’ for the Brisbane Metropolitan area. The dataset spans the period 1975:1 – 2005:12 and covers 65 postcodes around Brisbane. Geocoding was used to translate the address of a house into latitude/longitude coordinates, which were needed to construct weight matrices (to account for spatial autocorrelation) used in the SEM and SSSEM. Time-dummy, adjacent-period and hedonic imputation indexes (computed from the SEM and SSSEM regressions) are presented.

The remainder of the paper is organized as follows: Section 2 briefly introduces existing house price index methodologies before outlining and contrasting hedonic methods for housing price indices. Section 3 outlines a hedonic specification, the SEM, which incorporates spatially autocorrelated residuals. Section 4 casts the SEM in state space form to create a hedonic specification (the SSSEM) with dynamic shadow prices of property attributes. Section 5 discusses the Brisbane housing data specifically collected for this study. Section 6 provides index number series for the Brisbane metropolitan area estimated using a range of methodologies. Section 7 provides concluding remarks and possibilities for further research.

## **2. Hedonic Methods for Housing Price Indices**

Houses are sold at infrequent intervals, and consequently, constructing a house price index over time is dependent on a small sample of the total stock of houses. Consequently, the composition of the transacted houses may be substantially different to the composition of the total housing stock. Furthermore, there are quality changes in houses over time. Hence, measuring the change in house prices directly between two

periods from the raw data can lead to a biased and misleading estimate. As a result, a number of methods are outlined in the literature to account for these problems.

A *mix-adjusted measure* is the methodology used by the Australian Bureau of Statistics (ABS) in its indices for established house prices (ABS, 2005). The mix-adjusted measure of house price changes stratifies the sample into sub-groups based on a common characteristic (typically location) and the median price of each subgroup is aggregated (using a weighted average for instance) to determine overall price change. This method is designed to control for the compositional change problem but does not address the quality change problem. Furthermore, the mix-adjusted method does not directly compare ‘like with like’ because different houses are sold in different time periods. Prasad and Richards (2006) recently proposed an improvement to the mix-adjusted measure by combining geographic stratification with “price-based stratification”.

A popular method (particularly with the real estate industry) of estimating house price changes is the *Repeat Sales Measure*, originally devised by Bailey, Muth and Nourse (1963). The repeat sales method uses a regression framework to compare only repeat sale properties over time. This method is data friendly only requiring information on sale price, sale date and address. This allows for ‘matched product’ comparisons to be made, by matching the same houses across time periods. Unfortunately, this has the repercussion of dramatically reducing (the already small) sample size. Moreover, like the mix-adjusted method, it does not account for quality change over time.

As a result, using a hedonic method is preferred over other alternatives, because both compositional and quality change can be accounted for. The essential idea behind hedonic measures in the housing context is to use a regression based approach to explain the price of a house in each transaction using a range of characteristics (in this case property attributes) such as number of bedrooms, size and location. A drawback of the approach is that it is more data intensive, and this has previously restricted its use in Australia. The hedonic approach has led to the *time-dummy method* (DTH), the *hedonic imputation method* (HI) and the *characteristics price method* (CP).

The interested reader is referred to Cominos (2006) and Hansen (2006) for a more detailed introduction to price index methodologies other than hedonic methods.

*(i) The Time-Dummy Hedonic Method*

The generic hedonic regression for the DTH method can be specified as follows

$$\ln P_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{d}_{it}'\boldsymbol{\alpha} + v_{it} \quad (1)$$

where  $P_{it}$  is the price of house  $i$  ( $i = 1, \dots, N$ ) at time  $t$  ( $t=1, \dots, T$ ).  $\mathbf{x}_{it}$  is a ( $K \times 1$ ) vector of household characteristics,  $\mathbf{d}_{it}$  is a ( $T \times 1$ ) vector of dummy variables  $D_{jit}$  ( $j=1, \dots, T$ ) which takes the value 1 if house  $i$  is sold in period  $j$  and 0 otherwise.<sup>1,2</sup> With this functional form,  $\boldsymbol{\alpha}$  is a vector of price index parameters that capture the pure effect of price changes over time. Note that the regression in (1) is a pooled rather than panel data regression.

Much of the popularity of the time-dummy method probably lies in the fact that the price index can be obtained directly from the parameter vector,  $\boldsymbol{\alpha}$ . If we assume that the errors in equation (1) are normal, then the estimator of the price index in time period  $t$  (with period 1 as a base) is as follows:

$$\hat{I}_t = \exp(\hat{\alpha}_t - \text{var}(\hat{\alpha}_t)/2 - \hat{\alpha}_1) \quad (2)$$

---

<sup>1</sup> Alternatively, the model can be specified with T-1 dummy variables and a constant term. However, the Price Index formula in equation (2) will become:  $\hat{I}_t = \exp(\hat{\alpha}_t - \text{var}(\hat{\alpha}_t)/2)$  for  $t = 2, \dots, T$ , with  $\hat{I}_1 = 1$ .

<sup>2</sup> A slight digression is in order at this point regarding the semi-log specification in (1). Diewert (2003a) outlines that an advantage of the semi-log model in this context is that it can deal with situations in which one or more characteristics are equal to zero, whereas the log-log model cannot. This is an important consideration in the housing case, since the regression may include a variable like the number of car spaces (for example), but some inner city homes may not have a car space. For a more detailed assessment of the advantages of the semi-log model, the reader is referred to Diewert (2003b).

where  $\hat{\alpha}_t$  denotes the OLS estimator of  $\alpha_t$ . Hill and Melser (2005, henceforth HM) note that in most cases,  $\text{var}(\hat{\alpha}_t)/2$  is negligible for large samples, and  $\hat{I}_t \approx \exp(\hat{\alpha}_t - \hat{\alpha}_1)$ . The inclusion of half the variance in (2) is known as the ‘Goldberger correction’.

As noted by HM, the main problem with the conventional time-dummy variable method is that it assumes the same hedonic model and characteristics for each period, and the shadow prices of the characteristics do not change over time. Goodman (1978) computes house price indices for fifteen submarkets within New Haven, Connecticut, USA, using Box-Cox transformations. He observes that estimated hedonic coefficients are not constant either across submarkets or over time. Triplett (2004) proposes the use of the *adjacent period* approach. In this approach equation (1) is estimated for every pair of periods. Namely, the data are pooled for periods  $t$  and  $t+1$ , and then equation (1) is estimated to get an estimate of  $\alpha_{t+1}$ . The multi-period price index (with the first period as a base) is obtained by chaining together (antilog of) the adjacent year estimates of  $\alpha_t$ . The adjacent-period alternative (henceforth AP-DTH) is still a pooled regression, but it pools the minimum necessary to implement the dummy variable method.

Using the adjacent-period regression 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 dummy variable indices. Unfortunately, this ‘rolling window’ method brings a new problem to the table, in that if something unusual happens in a given period (e.g the sale of multiple acreage properties), it is not smoothed over time. That is, using the adjacent-period method, it is possible to have shadow prices that change dramatically from one time period to the next. In the housing context, it is more likely that preferences would change slowly over time. Consequently, large jumps in shadow prices between consecutive months or years are more likely to be attributable to data inaccuracies and/or misspecified hedonic functions, which is also attributable to the lack of accurate data on housing characteristics (particularly in Australia, see Section

5). If the Hedonic function is misspecified, shadow price parameters may incorporate information from omitted variables. As a result, this paper develops a methodology (outlined in Section 4) which allows the shadow prices of attributes to evolve smoothly over time.

*(ii) The Hedonic Imputation Method*

The Hedonic Imputation (HI) method constructs a price index using the imputed prices for products that are missing in a particular time period. This is of particular use in the housing context as houses are sold infrequently, and using the HI method allows matched price indices to be computed, that are not restricted to repeat sales.

Consider the following form of the hedonic model:

$$\ln P_t^h = \sum_{c=1}^C x_{c,t}^h \beta_{c,t} + v_t^h \quad h = 1, \dots, H_t; c = 1, \dots, C; t = 1, \dots, T \quad (3)$$

where  $P_t^h$  is the sale price of house  $h$  in time period  $t$ ,  $x_{c,t}^h$  is the value of the  $c^{\text{th}}$  characteristic for house  $h$  in time period  $t$ ,  $\beta_{c,t}$  is the hedonic coefficient of characteristic  $c$  in time period  $t$  and  $v_t^h$  is the disturbance term.  $H_t$  refers to the number of houses sold in time period  $t$ , and can differ between time periods. A further point to note is that, for example, the first house at  $t=1$  is different from the first house at  $t=2$ . That is, each time period consists of a different set of houses. There is a distinct difference between the hedonic equation in (1) and (3). The hedonic coefficients of the characteristics vary over time in (3). This is possible because the regression is run for all houses in each time period (as opposed to the pooled model in the time-dummy hedonic specification).

An important assumption of this method is that the included characteristics do not change over time. For the case of housing, this is satisfied, since the same property attributes can be applied to all houses across all time periods (ie. Number of bedrooms, number of

bathrooms etc). This means that an imputed price for a house  $h$  actually sold in period  $t$  can be computed for any other time period by estimating (3) using the property attributes of house  $h$ . For example, for a house  $h$  actually sold in period  $t$  with a vector of characteristics  $x_{c,t}^h$  ( $c = 1, \dots, C$ ), the imputed price for that house in period  $s$  would be computed as:

$$\hat{P}_s^h(x_t^h) = \exp\left(\sum_{c=1}^C x_{c,t}^h \hat{\beta}_{c,s}\right) \quad (4)$$

where  $\hat{P}_s^h(x_t^h)$  denotes the imputed price in period  $s$  of house  $h$  (actually sold in period  $t$ ) and  $\hat{\beta}_{c,s}$  denotes the OLS estimator of  $\beta_{c,s}$  in (3). Triplett (2004) notes that because of the semi-log specification, the regression prediction in (4) is biased as an estimate of the predicted price, and that the adjustment required will depend upon the price index formula being used as well as the assumptions regarding the distribution of errors. In practice, the bias often proves to be insignificant.

Using the imputed prices defined in (4) it is possible to compute matched price indices. A variety of price index formulae for the HI method are suggested by Hill and Melser (2005). With so many possible price index specifications, there needs to be a method of choosing the ‘best’ specification. Luckily, a clear consensus has emerged in the price index literature that Fisher and Törnqvist should be preferred to Paasche, Laspeyres, geometric-Paasche and geometric-Laspeyres. Diewert (2004) shows that Fisher and Törnqvist have superior axiomatic properties, while Diewert (1976) shows that they have superior economic properties (i.e., they are superlative). One class of Törnqvist indexes is listed below:

Geometric Paasche (GP):

$$I_{s,t}^{GP} = \prod_{h=1}^{N_t} \left( \left[ \frac{\hat{P}_t^h(x_t^h)}{\hat{P}_s^h(x_t^h)} \right]^{\frac{1}{N_t}} \right) \quad (5)$$

Geometric Laspeyres (GL):

$$I_{s,t}^{GL} = \prod_{h=1}^{N_s} \left( \left[ \frac{\hat{P}_t^h(x_s^h)}{\hat{P}_s^h(x_s^h)} \right]^{\frac{1}{N_s}} \right) \quad (6)$$

Törnqvist 1 (T1):

$$I_{s,t}^{T1} = \sqrt{I_{s,t}^{GP} \times I_{s,t}^{GL}}$$

$$= \left[ \prod_{h=1}^{N_t} \left( \left[ \frac{\hat{P}_t^h(x_t^h)}{\hat{P}_s^h(x_t^h)} \right]^{\frac{1}{N_t}} \right) \prod_{h=1}^{N_s} \left( \left[ \frac{\hat{P}_t^h(x_s^h)}{\hat{P}_s^h(x_s^h)} \right]^{\frac{1}{N_s}} \right) \right]^{1/2} \quad (7)$$

$$= \left( \prod_{h=1}^{N_t} \frac{\hat{P}_t^h(x_t^h)}{\hat{P}_s^h(x_t^h)} \right)^{\frac{1}{2N_t}} \left( \prod_{h=1}^{N_s} \frac{\hat{P}_t^h(x_s^h)}{\hat{P}_s^h(x_s^h)} \right)^{\frac{1}{2N_s}}$$

It is important to note that the index formulae in (5) through (7) use imputed prices in both the base and current periods, regardless of whether actual prices are available or not. Silver and Heravi (2001) show that the use of actual prices in the base period can introduce distortions into the price relatives. As a result, imputed prices are used.

Note that the Törnqvist index in (7) is different from any Törnqvist HI indexes specified by HM. Namely, this specification of Törnqvist weights all houses equally in the price index, whereas HM use expenditure share weights. Expenditure share weights can be criticized because they overly weight more expensive houses in the sample, however, price changes amongst more expensive houses is not any more indicative of market conditions than relatively cheaper houses. Further, the expenditure shares used in HM do not have the same interpretation of a budget share in the consumption context, where expenditure shares indicate the importance of a commodity in a consumer's budget. In fact, if we were trying to calculate the price index of a 'typical house', then expensive houses may even be weighted lower than their cheaper counterparts. However, the purpose is to calculate a house price index which is indicative of a region as a whole. Hence, equal weights are justified. HM fail to address this issue and do not substantiate the choice of expenditure share weights.

Alternatively, another Törnqvist-type index could be defined as follows:

$$I_{s,t}^{T2} = \left( I_{s,t}^{GP} \right)^{\frac{N_t}{N_t+N_s}} \times \left( I_{s,t}^{GL} \right)^{\frac{N_s}{N_t+N_s}}$$

Törnqvist 2 (T2):

$$= \left[ \left( \prod_{h=1}^{N_t} \frac{\hat{P}_t^h(x_t^h)}{\hat{P}_s^h(x_t^h)} \right)^{\frac{1}{N_t}} \right]^{\frac{N_t}{N_t+N_s}} \left[ \left( \prod_{h=1}^{N_s} \frac{\hat{P}_t^h(x_s^h)}{\hat{P}_s^h(x_s^h)} \right)^{\frac{1}{N_s}} \right]^{\frac{N_s}{N_t+N_s}} \quad (8)$$

Note that (8) is equivalent to (7) when  $N_t = N_s$ , because  $\frac{N_t}{N_t+N_s} = \frac{1}{2}$ . However, in the more typical event that  $N_t \neq N_s$ , the index in (8) gives a greater weight to  $I_{s,t}^{GP}$  if  $N_t > N_s$  or a greater weight to  $I_{s,t}^{GL}$  if  $N_t < N_s$ . For instance, if the year  $t$  was to be compared against  $s$ , and more observations are available for  $s$ , the sample of houses in  $s$  is likely to be more representative of the total population of houses than those in  $t$ . Hence, the sampling error attributed to an index resulting from the  $t$  data ( $I_{t,s}^{GL}$ ) will probably be larger than the sampling error attributed to the index from the  $s$  data ( $I_{t,s}^{GP}$ ). In general, as reporting requirements have improved over time, so has the accuracy and density of data available for econometric estimation of Hedonic models. Consequently, it makes sense to weight the sample which is more indicative of the population. In this sense, the Geometric-Paasche index will almost always be weighted more heavily than the Geometric-Laspeyres, because GP is based on current period data, while GL is contingent on base period data. This type of weighting is similar to that in the more recent work of Hill and Timmer (2006).

Recall that the imputed prices, as derived in (4), have an upward bias. Appendix A derives this bias and consequently the formula for calculating unbiased imputed prices.

*(iii) The Characteristics Price Index Method*

The characteristics price index method (CP method) uses the shadow prices of the characteristics (i.e. the regression coefficients in the hedonic function) in a weighted index number formula. The interested reader is again referred to HM for a discussion of this method in the housing context. This discussion is omitted here due to the useful result that some HI indices have CP counterparts. For instance, HM show that a particular variety of HI Törnqvist index is equivalent to a Fisher index constructed via the CP method.

*(iv) Comparison of the Hedonic Approaches*

HI and CP allow for time-varying parameters (the shadow prices of characteristics) in the hedonic model, while the standard DTH model constrains them to be the same. Although the adjacent-period DTH method only constrains the parameters to be equal across two time periods, the HI and CP methods make maximum use of the available data by imputing sale price values in order to compare ‘like with like’. Furthermore, they allow flexibility in the choice of price index formula, whereas, price index numbers are obtained directly from the regression for DTH methods. Given that HM have shown that some variants of HI indexes have direct counterparts using CP indexes, HI is preferred overall simply because the resulting price index formulae are often easier to derive.<sup>3</sup>

Given that HI is preferred, there is still the matter of specifying the hedonic regression as accurately as possible, which turns the spotlight onto two remaining issues. Firstly, how best to model the movement of parameters to vary over time. Secondly, how to incorporate the spatial correlation of house prices in the hedonic model.

Whenever the hedonic approach has been used to construct house price index numbers in the literature<sup>4</sup>, the hedonic function is usually assumed to have a white noise error term.

---

<sup>3</sup> See Silver and Heravi (2006) and Triplett (2004) for a more detailed comparison of the hedonic methods.

<sup>4</sup> See Hansen (2006), Costello (1997) and Flaherty (2004).

An exception to this is in Hill, Knight and Sirmans (1997a), where the errors are assumed to be serially correlated and heteroskedastic. However, spatial correlation, which occurs when population members are related through their geographic location, has been ignored in the literature on house price indexes. In contrast, literature concerned with the prediction of house prices has paid greater attention to the relevance of spatial correlation in the specification of a hedonic function.

### 3. Hedonic Specification with Spatial Autocorrelation

Basu and Thibodeau (1998) nominate two reasons why house prices should be spatially autocorrelated. First, neighborhoods tend to be developed at the same time, so neighborhood properties have similar structural characteristics such as block size, age, number of bedrooms and house size. The second reason why house prices should be correlated is that properties share location amenities (such as supermarkets, schools and public services) and socioeconomic variables specific to a particular neighbourhood, including local crime rates, wealth levels and racial composition. It should be noted at this point that a fully specified hedonic regression (in theory) accounts for all significant explanatory variables and consequently the residuals would not exhibit spatial autocorrelation. However, in practice many significant variables are omitted due to data constraints (particularly on socioeconomic and neighbourhood variables). Hence, there is usually spatial correlation left unaccounted for when using the hedonic method.

For the following discussion we write the hedonic model as follows

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (9a)$$

where  $\mathbf{y}$  is a  $n \times 1$  vector of observations on the log of house sale prices,  $\mathbf{X}$  is the  $n \times k$  matrix of explanatory variables (hedonics),  $\boldsymbol{\beta}$  is the  $k \times 1$  vector of unknown parameters associated with the exogenous variables and  $\boldsymbol{\varepsilon}$  represents the  $n \times 1$  vector of unknown disturbances.

The *spatial error model*<sup>5</sup> (SEM) is defined when (9a) is combined with (9b) as follows:

$$\boldsymbol{\varepsilon} = \rho \mathbf{W}\boldsymbol{\varepsilon} + \mathbf{u}, \quad (9b)$$

where  $\mathbf{W}$  is the  $n \times n$  spatial weight matrix and  $\boldsymbol{\varepsilon}$  is the  $n \times 1$  vector containing the error terms, with  $E(\boldsymbol{\varepsilon}) = 0$  and  $\text{Cov}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\Psi}$ . Furthermore,  $\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ , with  $\mathbf{I}$  the  $n \times n$  identity matrix. In this model,  $\rho$  is the parameter that captures the magnitude of the spatial autocorrelation, with  $0 \leq |\rho| \leq 1$ . Note that (9b) can be expressed as:

$$\boldsymbol{\varepsilon} = (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{u}, \quad (10)$$

and therefore (9) can be expressed in an equivalent way as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{u}, \quad (11)$$

from which it is easily seen that the covariance matrix of the error terms is given by

$$\boldsymbol{\Psi} = \sigma^2 (\mathbf{I} - \rho \mathbf{W})^{-1} (\mathbf{I} - \rho \mathbf{W}')^{-1} \quad (12)$$

This model can be estimated using generalized least squares (GLS) or maximum likelihood methods (ML). For further details on alternative spatial models and estimation methods the reader is referred to Anselin (1988).

#### (i) *Spatial Weight Matrix*

These spatial regression models require the specification of a spatial weight matrix, previously specified as  $\mathbf{W}$ , with elements  $w_{ij}$  representing the spatial relationship between units  $i$  and  $j$ . Dubin (1998) states that “the most common practice is to treat  $\mathbf{W}$  as nonstochastic; that is, the researcher takes  $\mathbf{W}$  as known *a priori*, and therefore, all results

---

<sup>5</sup> See Anselin (1999) p. 12 – 13.

are conditional upon the specification of  $\mathbf{W}$ .” There are many different ways to specify  $\mathbf{W}$ , however, there are some common properties that all weight matrices must share:

- a)  $\mathbf{W}$  is non-negative
- b)  $w_{ii} = 0$  (ie. it is supposed that an observation does not affect its own prediction)

Pace and Gilley (1998) weight each house depending on their proximity to other houses. Data on the latitude and longitude of each house in the sample was used to calculate the Euclidean distance between each house in the sample and each observation  $j$  was weighted by its distance  $d_{ij}$  from observation  $i$ . Hence, the weight between two houses decreases the further apart the houses are by measure of Euclidean distance, and for all pairs of houses which are more than  $d_{\max}$  apart, the corresponding weight is zero.

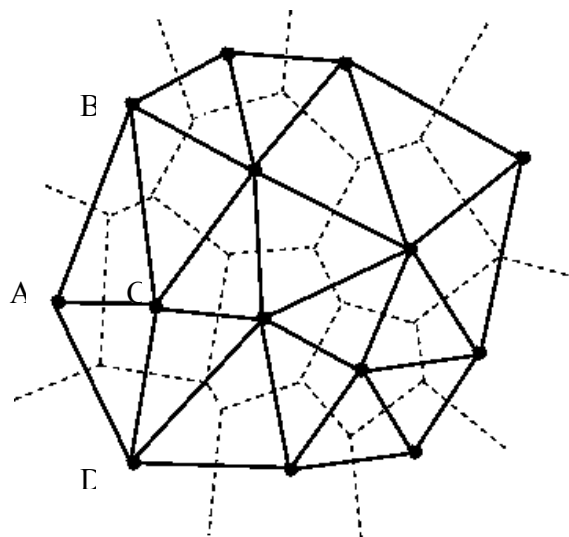
Another common method of forming  $\mathbf{W}$  is to use ‘nearest neighbours’. Under this scheme  $w_{ij} = 1$  if  $i$  and  $j$  are such that there is no observation closer to each  $i$  or  $j$ . In this specification  $\mathbf{W}$  is already row normalized, since there is only one non-zero observation per row (which takes the value 1). In the housing case, it is restrictive to assume that each house is only correlated with its immediate nearest neighbour; however, this scheme can easily be extended to  $n$  nearest neighbours in two ways. Firstly, a popular approach outlined in Dubin (1998) and used in Militino et al. (2004), is to set  $w_{ij} = 1$  if  $i$  and  $j$  are separated by a distance less than some prespecified limit (1 km for example). For this case, a house in a densely populated area would be correlated with many other houses, while a house in an acreage area may not be correlated with any other houses (ie. the number of nearest neighbours ( $n$ ) will change for each house). Secondly, the number of nearest neighbours ( $n$ ) can be selected *a priori* and then computed for each house in the sample. In this method each house is forced to be correlated with the same number of nearest neighbours, regardless of the neighbourhood density.

Alternatively, neighbours can be defined as contiguous<sup>6</sup> observations. That is, if two spatial units have a common border of non-zero length, they are considered to be contiguous (Anselin, 1988). Consider the following diagram:

---

<sup>6</sup> The American Heritage Dictionary defines contiguous as “sharing an edge or boundary; touching.”

**Figure 1**



In Figure 1, the points A and B are contiguous because they share a common edge. Contiguity<sup>7</sup> can be constructed artificially using a Delaunay triangle algorithm, which subdivides an area into triangles. The Delaunay triangulation of a point set is a collection of edges satisfying an "empty circle" property: for each edge a circle can be found containing the edge's endpoints but not containing any other points. For instance, in Figure 1, a circle can be found which includes the points A and D, to the exclusion of all other points. In this case, A and D are contiguous observations. Hence, if the points in Figure 1 represent houses, then house A has neighbours B, C and D, because it shares an edge with these houses. Note that house A only has 3 neighbours because it is near a boundary, whereas house C has five neighbours, because it is more central. This is consistent with houses in densely populated areas having more neighbours (with which they are correlated) than houses in less dense areas. Once an algorithm such as Delaunay triangulation<sup>8</sup> has been applied to determine the neighbours to each house, a weight matrix can be defined in the same fashion as the nearest neighbours scheme outlined

---

<sup>7</sup> The notion of contiguity referred to here is 'first order' contiguity. See Anselin (1988, p. 10) for higher orders of contiguity.

<sup>8</sup> MATLAB has an inbuilt Delaunay triangulation function. See the MATLAB help menu for more information.

earlier. Namely, for all houses  $i$  and  $j$ ,  $w_{ij} = 1$  if  $i$  and  $j$  are contiguous observations, and zero otherwise. The resulting weight matrix  $\mathbf{W}$  is then row normalized in the usual fashion.

R. Kelley Pace has used this form of weight matrix in much of his work on spatial statistics for real estate (for example, see Lesage and Pace, 2004), and includes a code to create it (*FDELW2.m*) in his *Spatial Statistics Toolbox 2.0* for Matlab, which can be downloaded for free from [www.spatial-statistics.com](http://www.spatial-statistics.com). In order to construct  $\mathbf{W}$  in this manner, the latitude and longitude coordinate of each house must be known, allowing the *FDELW2* Matlab function to be used to convert the Delaunay algorithm results into a contiguity matrix. This function constructs a weight matrix which sets  $w_{ij} = 1$  if  $i$  and  $j$  are contiguous observations, for all houses  $i$  and  $j$ , and zero otherwise. This is the approach followed to estimate  $\mathbf{W}$  in the present paper.

#### **4. Hedonic Specification with Dynamic Shadow Prices**

In the housing market, as in most goods markets, preferences for housing characteristics will change over time. Hence, it is intuitive that the coefficients in a hedonic regression vary from one period to the next. However, in the current literature the hedonic coefficients have either been constrained to be constant over the sample period of interest (ie. the DTH method) or have been allowed to vary in a sporadic fashion over. In the latter case, the shadow prices vary sporadically because the hedonic regression is re-estimated in every time period (or every second period for the adjacent period method). This can be problematic if something unusual happens in a particular time period or there is an anomaly in the data, as the resulting house price index may be inaccurate.

Alternatively, a viable formulation of a hedonic model can be obtained by assuming that the vector of parameters is generated by a stochastic process. There are three obvious classes of model which can be used to enforce this structure:

Firstly, the parameters can be assumed to vary randomly about a fixed, but unknown, mean. This is known as a *random coefficient* model. However, this will allow the parameters to change in a haphazard fashion. Whereas, in many cases, if parameters are to be regarded as stochastic, it seems reasonable to suppose that they will change gradually over time. This suggests a framework in which the parameters are generated by a multivariate ARMA process. However, in the housing context it is possible that the stochastic process is non-stationary. An important example of this kind of behaviour arises when the parameters follow a multivariate random walk. Because the parameters are no longer constrained to have a fixed mean, the model can gradually evolve over time. This specification allows, if needed, the values at the end of the sample period to be very different from those at the beginning. This specification is warranted in the housing context (and adopted here) as there is no reason to expect the marginal valuations of the characteristics to be stationary, particularly over long periods of time.

The spatial errors model (SEM) in (9) can be specified with time varying coefficients following a random walk process as follows:

$$\mathbf{y}_t = \mathbf{X}_t \boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t \quad (t=1 \dots T) \quad (13a)$$

$$\boldsymbol{\varepsilon}_t = \rho \mathbf{W}_t \boldsymbol{\varepsilon}_t + \mathbf{u}_t, \quad (13b)$$

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t \quad (13c)$$

where:

$\mathbf{y}_t$  is an  $N_t \times 1$  vector of (log) house prices and  $N_t$  is the number of houses sold in time period  $t$ ;

$\mathbf{X}_t$  is an  $N_t \times k$  matrix of hedonic characteristics and a constant term;

$\boldsymbol{\beta}_t$  is a  $k \times 1$  vector of unknown parameters and  $\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t \sim (0, \boldsymbol{\Sigma}_t)$  ;

$\mathbf{u}_t \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_t)$ , with  $\mathbf{I}_t$  the  $N_t \times N_t$  identity matrix;

$\mathbf{W}_t$  is the  $N_t \times N_t$  spatial weight matrix;

$\boldsymbol{\varepsilon}_t$  is the  $N_t \times 1$  vector containing the error terms, with  $E(\boldsymbol{\varepsilon}_t) = 0$  and  $\text{Cov}(\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_t') = \boldsymbol{\Psi}_t$  ;

From these definitions we obtain:

$$\Psi_t = \sigma_\varepsilon^2 (\mathbf{I}_t - \rho \mathbf{W}_t)^{-1} (\mathbf{I}_t - \rho \mathbf{W}_t')^{-1};$$

$\rho$  is the parameter that captures the magnitude of the spatial correlation ( $0 \leq \rho \leq 1$ );

$$\eta_t \sim (0, \sigma_\eta^2 \mathbf{I}_k);$$

$$E(\varepsilon_t \eta_t') = \mathbf{0}$$

The model in (13) can be recognized as a state space representation of the Hedonic SEM model and opens the way for the application of the Kalman filter and smoother algorithms, which gives the optimal estimator of the state ( $\beta_t$ ) based on all sample information. Furthermore, the unknown parameters  $\sigma_\varepsilon^2$ ,  $\sigma_\eta^2$  and  $\rho$ , can be estimated using MLE. Initialisation of the Kalman filter requires estimates of the initial state ( $\beta_0$ ) and its variance-covariance matrix ( $\Sigma_0$ ). In this paper,  $\beta_0$  is started at zero and assumed to have a diffuse distribution, with  $\Sigma_0 = \kappa \mathbf{I}$ . As previously mentioned,  $\mathbf{W}_t$  is constructed using the FDELW2 Matlab function.<sup>9</sup>

The number of regressors,  $k$ , are constant over all time periods, meaning that all houses in the sample are a function of the same hedonic characteristics. This may be a restrictive assumption in other empirical applications (e.g. computers), however, is justifiable in the housing context.

The next section discusses the data specifically collected for this study.

## 5. Data

The data used in this study are those accumulated by property information services, which aim their products at the Real Estate Industry and other paying customers. One leading provider of property information services is 'RP Data', accessible via the internet

---

<sup>9</sup> See Section 3 (i).

at [www.rpdata.com](http://www.rpdata.com).<sup>10</sup> The information provided by RP Data is designed for the consumer (or agencies assisting the consumer), and as such is much like a search engine, whereby one can search for properties based on a set of criteria such as location, price, sale date, zoning restrictions etc. Furthermore, they provide suburb profiles, street sales history reports, investment reports and the like. A brief description of retrieving data for RP Data is provided below. For full details the reader is referred to Cominos (2006).

The data preparation can be segregated into three distinct steps. Firstly, once the raw data were obtained from RP Data, the variables (ie. property attributes) to include and exclude were chosen. Secondly, the address of each house was geocoded to provide a latitude/longitude coordinate for each observation. Thirdly, the data were filtered of outliers, errors and incomplete observations.

Initially information on 316,359 house sales was downloaded for the Brisbane City Council Area, which spanned a period backdating from 31/12/2005 to the early 1950s.<sup>11</sup> However, most of these observations did not contain information on property attributes, which are needed to use a Hedonic Model. Furthermore, the data needed to be cleaned for errors. In order to do this the raw data were organized into the following fields: address, postcode, sale price, sale date, map reference, Area (m<sup>2</sup>), number of bedrooms, number of bathrooms, number of car spaces and number of lock up garages. Some houses contained data on additional property attributes; however, these were excluded from the analysis mainly because there were too many blank fields resulting in a highly reduced sample.<sup>12</sup> Thus, we are left with a trade-off between the number of included attributes and the sample size. As a result of this trade-off, it was decided only to include AREA, BED, BATH, number of car spaces ('CAR') and number of lock up garages ('LUG'), where CAR and LUG were combined into one series (henceforth 'CARLUG') as follows:

- If CAR was specified, but LUG was not, then take the value specified by CAR.
- If LUG was specified, but CAR was not, then take the value specified by LUG.

---

<sup>10</sup> To access the information provided by RP Data, a (rather expensive) license is required.

<sup>11</sup> Although, data pre 1970 was extremely sparse.

<sup>12</sup> See Cominos (2006) for a full list of excluded property attributes, and associated discussion.

- If both were specified, take the value specified by CAR. This is justified because the number of car spaces reported is always greater than or equal to the number of garage spaces (for overlapping reports), which implies that the car variable includes car spaces and lock-up garages.

The usual measure of location was available from RP Data, ie the address of each house. However, to use spatially correlated models, a measure of distance between houses (by some metric) needs to be computed to construct a spatially correlated covariance matrix. The procedure for taking the addresses of houses and converting them into latitude/longitude coordinates (which in turn can be used to calculate the distance between houses) is known as *geocoding*. In this paper, the address fields were converted to latitude, longitude using the Geodetic Datum of Australia 1994 using MapInfo Professional. The reference data layers used for the geocoding process were RoadNet Comprehensive and Australian Postcodes from Map Data Sciences. Over 90% of the original records were successfully geocoded.

With the variables specified, the dataset needed to be cleaned of errors, incomplete observations and significant outliers. Hence, observations satisfying the following criteria were removed<sup>13</sup>:

- Sale Prices less than \$1000 or greater than \$30,000,000;
- Incorrectly specified address (ie. Missing postcode, missing house number etc.);
- Missing sale date or sale Date outside the range 01/01/1971 – 31/12/2005;
- Missing map reference;
- Area less than 100 m<sup>2</sup>;
- Missing number of bedrooms, 0 bedrooms or greater than 9 bedrooms;
- Missing number of bathrooms, 0 bathrooms or greater than 9 bathrooms;
- Missing number of car spaces or greater than 9 car spaces;

---

<sup>13</sup> While the process of filtering the data was laborious, it was relatively simple. The same cannot be said of the process of organizing the data into columns, which was quite difficult. Hence, anyone attempting to replicate this procedure may contact the author for a Microsoft excel template.

- Outside the Brisbane Metropolitan Area;
- Property types other than residential houses.
- Observations for which a latitude/longitude co-ordinate was unable to be generated.

The dataset was chosen to include only those observations post 01/01/1971. This was not chosen arbitrarily, with earlier data being sporadic and incomplete. Hence, the period pre-1971 contains too few observations to be considered. In fact, it seems that a default date of sale is 01/01/1930 – the sale date entered by agents when the true sale date is unknown.

At the end of this procedure a dataset containing 71,583 total observations was generated for the Brisbane City Council Area, covering 65 postcodes. See Appendix B for a list of the postcodes (and associated suburbs) included in the dataset. Table 1 contains a summary of the data in each time period. Median measures of the property attributes are provided in order to observe the ‘typical’ house in any particular time period, and changes in the ‘typical’ house over time.

**Table 1 Summary of the data over time**

YEAR	PRICE (\$)			AREA (m <sup>2</sup> )			BED			BATH			CARLUG			No. Obs
	Min	Median	Max	Min	Median	Max	Min	Median	Max	Min	Median	Max	Min	Median	Max	
1971	2500	11000	26500	372	615	23900	1	3	7	1	1	3	1	1	5	79
1972	1850	13500	38700	286	607	20200	1	3	6	1	1	5	1	2	4	131
1973	4000	18700	73000	263	610	10180	2	3	6	1	1	4	1	2	5	181
1974	2600	23000	92000	341	607	2023	2	3	6	1	1	3	1	2	4	150
1975	4000	24275	470000	354	612	42000	1	3	6	1	1	3	1	2	6	226
1976	2796	26000	170000	304	607	10390	1	3	7	1	1	4	1	2	7	265
1977	6095	27875	96500	235	607	28500	1	3	6	1	1	4	1	2	4	238
1978	3000	29000	176000	218	610	10800	1	3	6	1	1	5	1	2	4	275
1979	6200	31000	97500	202	607	18400	1	3	6	1	1	4	1	2	5	326
1980	9000	34000	195000	202	610	20300	1	3	6	1	1	4	1	2	7	495
1981	4750	44000	210000	202	610	20200	1	3	7	1	1	4	1	2	6	524
1982	15000	52000	160000	304	607	15860	1	3	7	1	1	5	1	2	4	391
1983	6000	54000	625000	126	607	18710	1	3	9	1	1	9	0	2	7	645
1984	15000	58000	275000	152	607	51900	1	3	7	1	1	5	1	2	7	791
1985	4720	59000	325000	202	607	21000	1	3	6	1	1	5	1	2	8	843
1986	15000	60000	435000	202	607	26100	1	3	6	1	1	4	1	2	8	740
1987	5000	64000	560000	172	607	67200	1	3	7	1	1	5	1	2	6	1387
1988	10000	80000	750000	169	607	60400	1	3	8	1	1	5	0	2	9	2212
1989	3500	102000	1800000	169	607	35900	1	3	7	1	1	5	1	2	8	1861
1990	10000	110000	4800000	169	607	34000	1	3	7	1	1	7	1	2	8	2198
1991	8500	123000	1070000	169	607	40000	1	3	6	1	1	5	0	2	6	2251
1992	3000	130000	770000	152	607	99100	1	3	8	1	1	5	1	2	8	2288
1993	3000	137500	2300000	146	607	45300	1	3	8	1	1	6	1	2	9	2462
1994	3996	145000	980000	152	607	66000	1	3	7	1	1	5	0	2	8	2332
1995	2000	142000	1790000	152	607	60400	1	3	8	1	1	6	1	2	7	1660
1996	5000	144900	1400000	126	607	31800	1	3	7	1	1	5	1	2	6	2107
1997	3000	152000	1410000	120	607	40500	1	3	9	1	1	5	1	2	8	2793
1998	2000	158500	2475000	106	607	98100	1	3	8	1	1	6	0	2	9	2941
1999	1400	165000	3000000	143	607	65000	1	3	8	1	1	9	1	2	8	3892
2000	1090	172300	2750000	120	607	163700	1	3	8	1	1	6	1	2	7	4323
2001	1210	200000	5350000	106	607	100000	1	3	8	1	1	5	1	2	8	4954
2002	1610	255000	5900000	101	607	198000	1	3	8	1	1	7	0	2	9	5240
2003	1860	330000	8200000	113	607	99100	1	3	8	1	1	9	0	2	8	6672
2004	2004	375000	6000000	101	607	114500	1	3	9	1	2	7	0	2	8	6097
2005	1111	375000	7000000	107	607	114500	1	3	9	1	1	6	0	2	8	7613

As expected, the median house price tends to rise over time. An exception to this is the case of 1995, when the median house price fell slightly from \$145,000 to \$142,000. Interestingly, the median area is either 607 m<sup>2</sup> or a value within 8 m<sup>2</sup> of this amount. Initially this seemed most unusual; however, this amount is exactly 24 perches, a

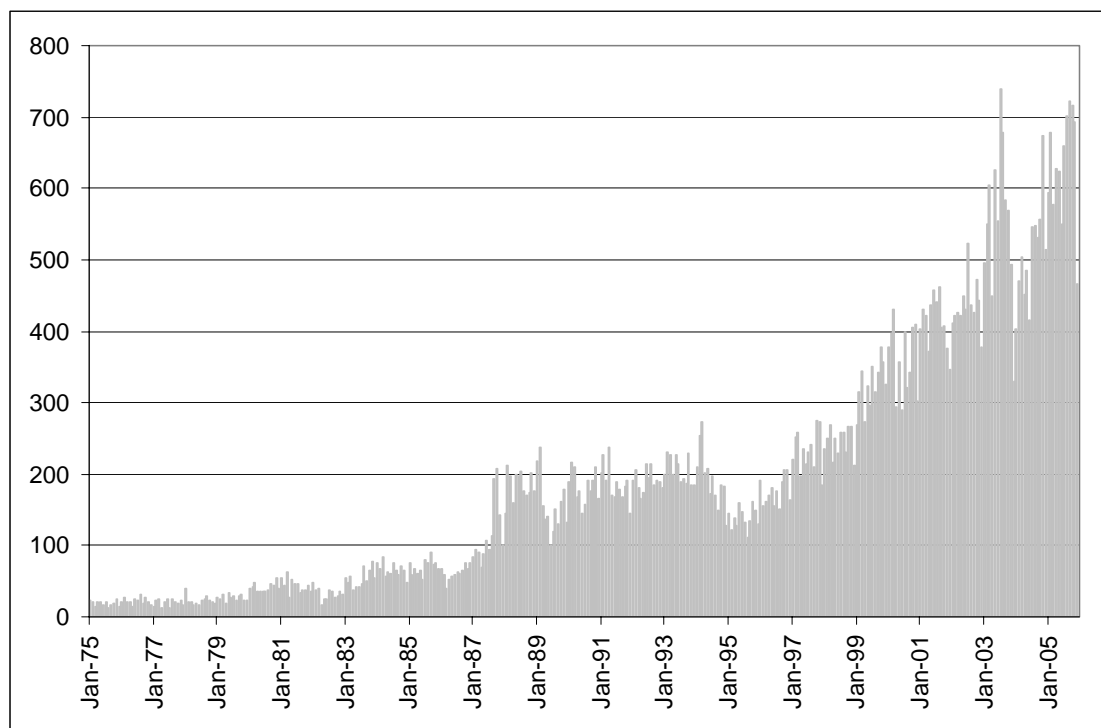
'standard' block of land for the Brisbane Metropolitan Area. Similarly, the median 'number of bedrooms', 'number of bathrooms' and 'number of car spots' are remarkably constant over time. This implies that with respect to the above property attributes, the typical house has not changed over time. While this would be expected over a short time horizon (because new houses and renovations to existing houses comprise a small proportion of the overall housing stock), this is a remarkable result over a period of thirty-five years. A more intuitive result is that the minimum number of car spots is zero in later years, while it takes the value one for earlier years. This may capture the relationship that houses built in recent years that are close to the CBD have no need for car spaces, whereas, in earlier time periods, the city was less populated and house blocks were generally bigger. This notion is supported by the minimum area seemingly decreasing over time. Namely, the minimum size for a block of land was much larger in the 1970s than in more recent years. Alternatively, later time periods have a larger sample of data and therefore are more likely to include extremes in the minimum and maximum values of the characteristics.

The fact that the medians are so resilient in the data probably reflects the fact that the housing stock is slow to change over time, with newly built homes only capturing a fraction of the total housing stock in any period. Hence, even if preferences change (which are reflected in newly built houses) it is unlikely to alter the characteristics of a typical house.

Another crucial point is that the number of observations typically increases over time, with a sample size of only 79 for the year 1971, increasing to a sample size of 7613 for the year 2005. This is probably due to the combination of two factors. Firstly, that the Brisbane metropolitan area (and associated population) has grown over the sample period of interest; and secondly, that the data has become more comprehensive over time.

In the next section the SSSEM is estimated using monthly data ranging from 1975:1 to 2005:12. The years 1971-1974 were dropped because these had less than 10 observations per month. The number of observations per month is shown below in Graph 1.

**Graph 1 – Number of Observations per month**



As expected, the number of observations per month increases with time. Furthermore, there appears to be some seasonality in the number of houses sold per month. For instance, December typically has fewer sales than November and January.

While the dataset has been filtered and cleaned of errors, the dataset may not be representative of the Brisbane Metropolitan area and/or inaccurate as data could not be downloaded for all suburbs across all time periods of interest (see Cominos (2006) Table A1.1, Appendix 1). Furthermore, the majority of the raw data obtained from RP data were not useful since they did not contain information on property attributes. This is a problem in that it reduces the sample size, however, further bias may be introduced into the resulting price index if there is some systematic difference between the houses that report attributes and the houses that do not. If, for instance, Real Estate Agents are more likely to record the property attributes of a house for more expensive properties, then the resulting price index will be biased upwards.

Much still remains to be done regarding housing data in Australia. While improvements in methodology can adjust for other sample biases, inaccuracies in the data or incomplete data are problems in themselves. In addition, incoming data need to be recorded with greater accuracy and comprehensiveness. Secondly, the data need to be compiled in a form that is easy to download, such as a workable spreadsheet format. For this to be achieved, either a government institution needs to include it in their existing list of responsibilities, or it can be outsourced to a private property information service such as RP Data for a fee. Outsourcing the work seems more plausible, especially considering that providing the information in spreadsheet format would be little hassle for a large information service. As if to service this point, the RBA recently published a discussion paper comparing the Hedonic and Repeat Sales measures (Hansen, 2006) for Australian capital cities that used Australian Property Monitors (APM) and the Real Estate Institute of Victoria (REIV) to prepare the data. HM (2005) also used APM data to construct Hedonic House Price Indices for Sydney over the years 2001, 2002 and 2003. As the demand for such data grows, the responsibility falls upon notable institutions such as the RBA or ABS, for which such data are especially important.

## **6. Empirical Results**

The data for the Brisbane metropolitan area are used to compute house price indexes derived from the State-Space Spatial Errors Model (SSSEM) method; the time-dummy (DTH) and adjacent-period methods (AP-DTH); and the spatial errors model (SEM).

Note that all results reported are for the period 1990:1 – 2005:12, however, the SSSEM model was estimated for the period to 1975:1 – 2005:1. The years 1971:1974 were excluded from this analysis because the SSSEM model was estimated using monthly data and the number of observations per month was too sparse during these initial years. The period 1990 to 2005 was chosen for the computation of the indices for two reasons. Firstly, because it required less computation in the case of the SEM and adjacent-period models, which needed to be re-estimated in every time period. Secondly and more

prominently, the majority of the data fall in the period 1990:2005. To be exact, 59823 observations out of the total 71583 observations (83.57 per cent) are from the period 1990:2005.

The traditional dummy-time hedonic (DTH) model<sup>14</sup> is used to construct an annual house price index for the Brisbane area for the period 1990 to 2005. The model is a pooled model consisting of 59823 observations. The independent variables included were AREA, BED, BATH, CARLUG, intercept and time-dummy variables for the years 1991:2005.

The DTH method holds fixed the hedonic coefficients for all the 16 years, which has been criticized in the literature. A preferred alternative is the ‘adjacent period’ approach. The specification remains the same, except that data from two adjacent periods are pooled, so the regression is re-estimated for every pair of adjacent time periods. Note that in this case the resulting index numbers satisfy temporal fixity since the computation of a new index number requires the estimation of a new regression, so that existing index numbers are not revised retrospectively. Consequently, unlike the pooled DTH approach, the index numbers remain the same regardless of whether the entire sample is used or not. The resulting set of bilateral index numbers from the adjacent period regressions can be easily linked through successive chaining.

The spatial errors model (9) was estimated annually over the period 1990:2005 with the same hedonics as the DTH model.<sup>15</sup> Note that this is not a time series model; rather, the regression in (9) was re-estimated for every individual time period. Similar to the AP-DTH method, the resulting index numbers satisfy temporal fixity since (9) is estimated for every time period. Therefore, estimation of a price index series over the full sample does not change the results. The SEM was computed in Matlab using the generic code *sem.m* available in James LeSage’s Econometric Toolbox.<sup>16</sup> The code makes use of sparse matrix routines, developed by Pace and Barry (1997) and Pace and Lesage (2004)

---

<sup>14</sup> See Section 2 for specification of the DTH regression.

<sup>15</sup> The SEM is defined in Section 3.

<sup>16</sup> LeSage’s Econometric Toolbox is freely available for download from [www.spatial-econometrics.com](http://www.spatial-econometrics.com).

to minimise the computational burden of matrix operations on  $N_t \times N_t$  matrices. For further information on the toolbox and its functions, see LeSage (1998).<sup>17</sup>

To ensure that the SEM is justified, it is necessary to check that the residuals are in fact spatially autocorrelated. Two common alternatives are the Moran  $I$  statistic and LR-test. The Moran  $I$  statistic requires the prior estimation of  $\mathbf{W}_t$  in every time period, while the LR-test requires OLS and the spatial errors model to be estimated.<sup>18</sup> Table 2 reports these statistics.

**Table 2 – Tests for the presence of Spatial Autocorrelation**

Year	Moran $I$	Moran $Z_t$	LR stat
1990	0.1464	11.9155	127.1655
1991	0.2563	21.0753	334.7955
1992	0.2156	17.8345	254.9895
1993	0.2031	17.4664	247.0573
1994	0.2081	17.4404	232.4361
1995	0.1982	13.9541	161.8869
1996	0.1655	13.1358	141.6757
1997	0.2379	21.8062	369.8247
1998	0.2311	21.7087	362.4950
1999	0.2779	29.9798	688.2507
2000	0.3392	38.5383	1053.7
2001	0.3392	41.4676	1157.9
2002	0.3718	46.6997	1431.0
2003	-	-	1598.8
2004	-	-	1187.6
2005	-	-	2611.7

Note:  $Z_t$  could not be estimated for the years 2003:2005 because of computational problems in Matlab due to large sample sizes.

The null hypothesis is the same for both tests and is as follows:

$H_0$ : No spatial autocorrelation ( $\rho = 0$ )

$H_1$ :  $H_0$  is false

$Z_t$  is defined in Appendix C and is asymptotically normally distributed. Hence, it can be compared with a critical value of 2.33 at the 1 per cent level of significance. The null

<sup>17</sup> Appendix 11, Cominos (2006), includes the code used to compute the SEM models, spatial autocorrelation tests and imputed prices for index construction.

<sup>18</sup> Appendix C outlines the Moran and LR tests in more detail.

hypothesis is easily rejected for every year that it was computed, confirming that the residuals are spatially correlated. Alternatively, the LR is asymptotically distributed with a  $\chi^2(1)$  distribution, which has a critical value (at the 1 per cent level of significance) of 6.635. Again, the null hypothesis is easily rejected for all years.

It is not surprising that the Moran and LR statistics increase over time, because the number of observations typically increases over time.<sup>19</sup> The effect of a greater number of observations is the presence of closer neighbours to each house in the sample (reflected in the **W** matrix) which may translate into higher correlation amongst those house prices. Furthermore, the incorporation of spatially correlated errors provides a better fit, by measure of  $R^2$ , relative to the AP-DTH method.

Once the model is estimated in every time period, hedonic imputation indexes of the Geometric-Paasche (GP), Geometric-Laaspeyres (GL), Törnqvist 1 (T1) and Törnqvist 2 (T2) variety can be calculated (see Section 2). Imputed prices are computed with and without the bias correction (see Appendix A).

The SSSEM model was estimated with monthly data from 1975:1 to 2005:12. State space models are time series models, and therefore they assume a reasonably lengthy time series. The number of monthly sales in the Brisbane data set is relatively large especially in the latter periods of the samples, giving the opportunity to compute both monthly and annual indices. Unfortunately, computational restrictions (on  $N_t \times N_t$  matrices) in GAUSS prohibited the estimation of the SSSEM annually, however, R. K Pace and J. LeSage<sup>20</sup> have created sparse matrix routines in MATLAB to allow operations on matrices of larger dimension.<sup>21</sup>

The data set contained a number of “acreage” properties (ie houses located on land with an area considerably larger than the standard size block). It was the case for some of the

---

<sup>19</sup> See Graph 1.

<sup>20</sup> See Pace and Barry (1997) and Pace and LeSage (2004).

<sup>21</sup> The SEM model was able to be estimated annually because it was estimated in Matlab using LeSage's Econometric toolbox (as previously noted), which makes use of sparse matrix functions.

earlier months in the sample, that the monthly index was distorted by the presence of some of these atypical observations. This was due to earlier months having relatively few observations, and therefore the inclusion of some acreage properties significantly affected the regression coefficients, and the consequent index. This was not a problem in the other models as they were estimated annually. Thus, an extra variable was included in the SSSEM specification; namely, a dummy variable with the value of one when AREA was larger than 2000 m<sup>2</sup> and zero otherwise.<sup>22</sup>

The monthly  $\mathbf{W}_t$  were computed using Kelley Pace's FDELW2 Matlab function, and the output exported to be used in the SSSEM which was coded in GAUSS<sup>23</sup>. The unknown parameters  $\sigma_\varepsilon^2, \sigma_\eta^2$  and  $\rho$  were estimated using MLE. The estimated values were  $\hat{\sigma}_\varepsilon^2 = 0.4$ ,  $\hat{\sigma}_\eta^2 = 0.2$  and  $\hat{\rho} = 0.5$ . The estimated autocorrelation parameter,  $\hat{\rho}$ , is consistent with the values obtained for the SEM model estimates. There is scope for improvement on this optimization procedure by concentrating one of the unbounded parameters (either  $\sigma_\varepsilon^2$  or  $\sigma_\eta^2$ ) out of the likelihood function, however, is left to future research. The reader is referred to Harvey (1990 p. 133) for more details.

We now summarise the main results. Detailed results are available from the authors.

*(i) Estimated Hedonic Coefficients from different models*

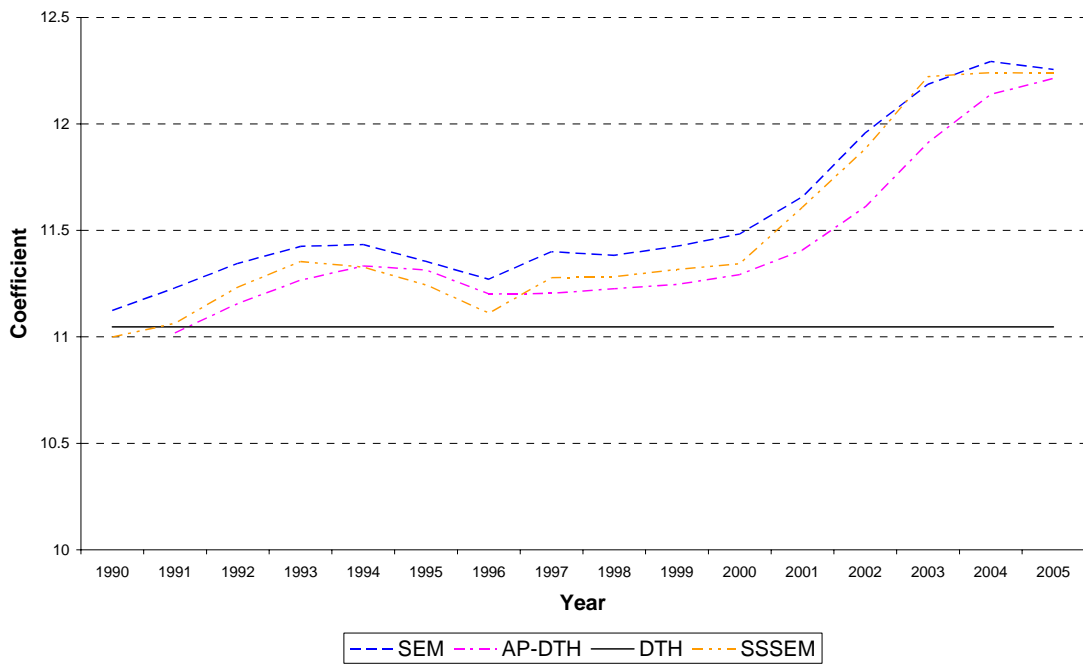
The regression output for the models is not provided in this paper due to the vast amount, particularly for the SEM and AP-DTH which are re-estimated every year, and the parameters of the SSSEM need to be re-estimated each month. The reader is referred to Cominos (2006) for the output. Instead, Graphs 2 to 6 present a comparison of the estimated parameters from all the models. Note that in order to compare the coefficients from the SSSEM with the other models, an average of the monthly parameter values was taken for each year.

---

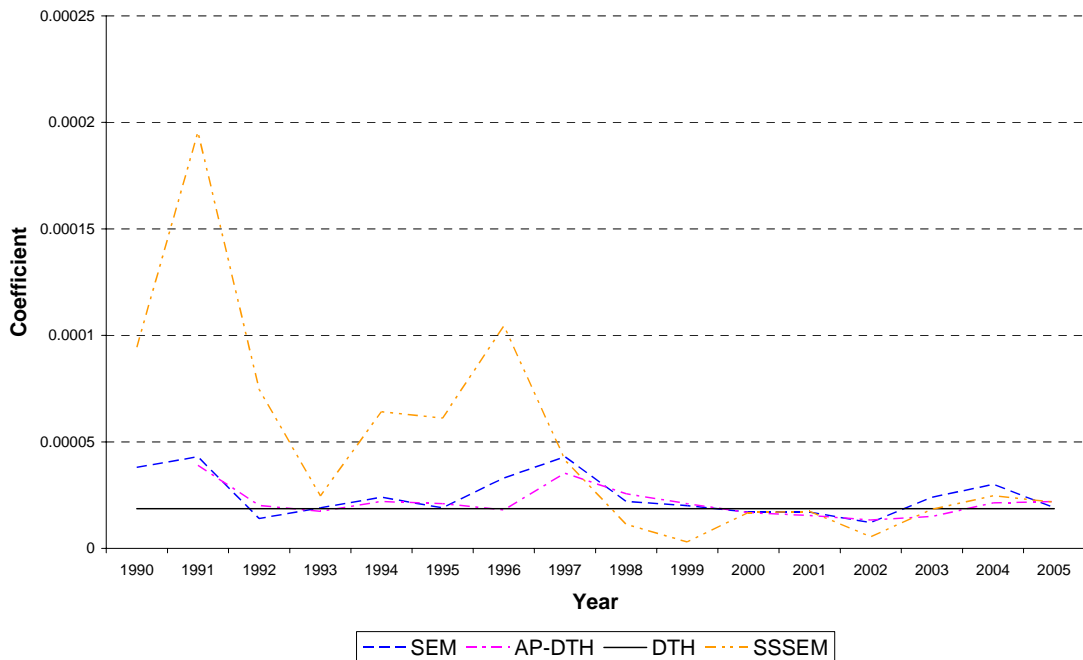
<sup>22</sup> In particular, the imputed prices were inflated substantially for some observations, which affected the resulting index numbers. Inclusion of the dummy variable lessened, but did not remove, this effect.

<sup>23</sup> There is no current equivalent in GAUSS to the Delauney function coded in Matlab.

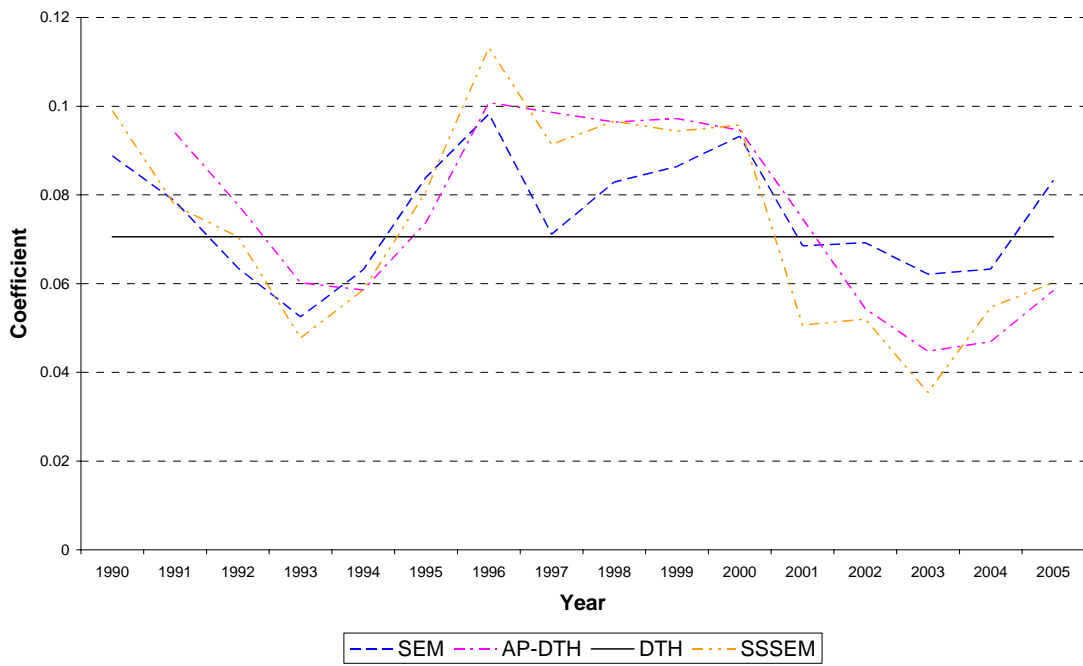
**Graph 2 – Intercept coefficient over time**



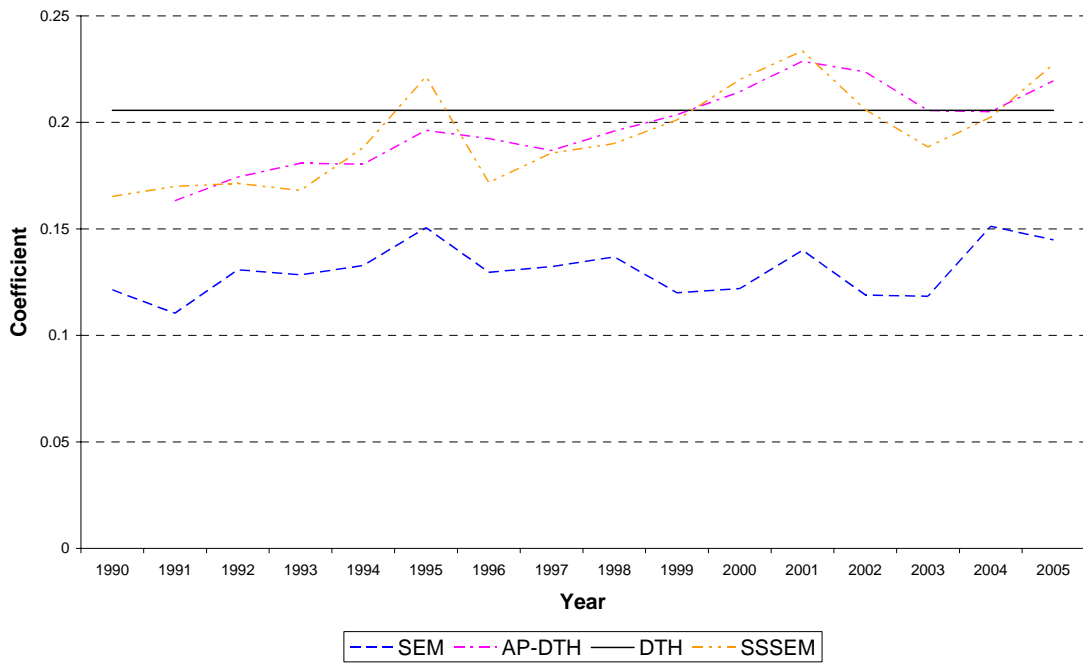
**Graph 3 – AREA Coefficient over time**



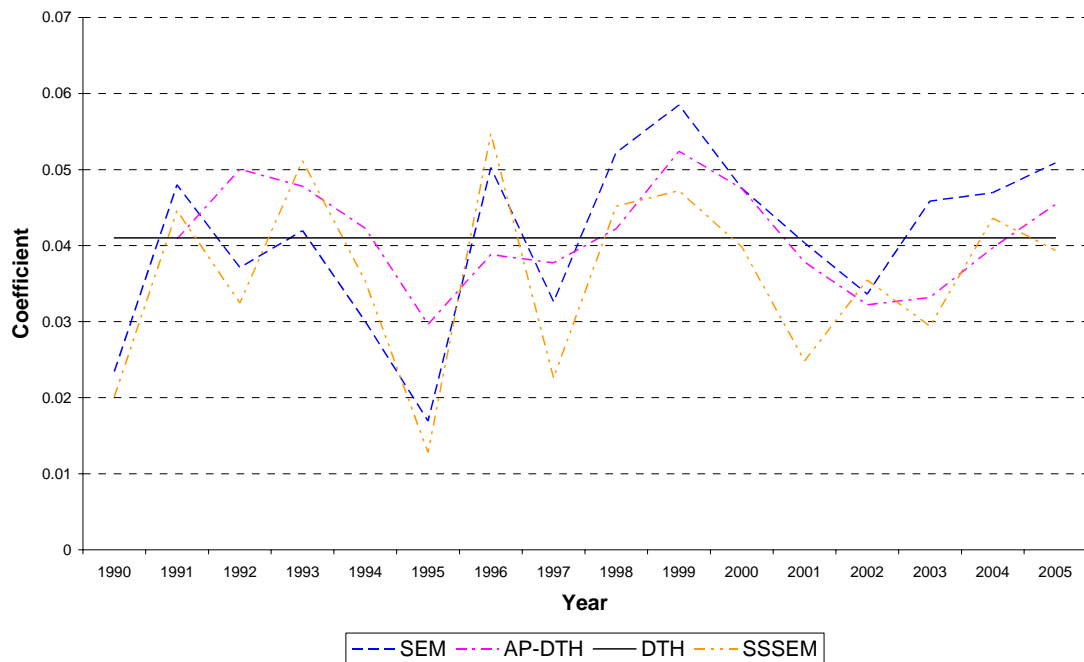
**Graph 4 – BED coefficient over time**



**Graph 5 – BATH coefficient over time**



**Graph 6 – CARLUG coefficient over time**



The coefficients for the BED and CARLUG variables for all models vary over time but appear to do so around a constant mean – that is, the series seem to be stationary. Given that the DTH model is a pooled regression constraining the coefficients to be constant, it is not surprising that the associated coefficient value appears to be the mean level.

The same interpretation can be applied to the AREA parameter, except that the coefficient series resulting from the SSSEM is significantly different (ie larger) during the early 90s before converging towards the other series later in the sample. This is due to earlier months having relatively few observations, and therefore the inclusion of some acreage properties significantly affected the regression coefficients in some of these months (in turn affecting the average for each year). This point was made previously, and the inclusion of a dummy variable in the regression for large acreage properties did not seem to completely remove this effect.

The BATH coefficients are particularly intriguing. Graph 5 shows that the bathroom coefficient for the SEM is significantly smaller across all time periods compared to the

other models – ie when spatial autocorrelation is accounted for. Given that the SSSEM also accounts for spatial autocorrelation in a similar fashion, it is interesting that its BATH coefficient series does not track that of the SEM. However, it is important to note that the SSSEM has a slightly different hedonic functional form than the other models – namely, it includes a dummy variable for acreage houses. In fact, when this variable is removed, the coefficient on the BATH variable closely tracks that of the SEM. However, when the dummy is inserted, it more closely tracks that of the AP-DTH (as shown). Furthermore, notice that the intercept term is systematically higher for the SEM model than for the others, which may help explain the similarities in the resulting index series (shown in the next sub-section). While the SEM bathroom coefficient series is systematically lower, the SEM intercept series is systematically higher, leaving the imputed prices somewhat unaffected.

The BATH coefficient series also trends slightly upwards for the AP-DTH and SSSEM series, implying that the value attributed to every bathroom by consumers becomes successively larger through time. The SEM model does not exhibit this trend.

The coefficients attributed to the AP-DTH regressions have the smoothest coefficients (not including the DTH regression for which they are constant). This is not surprising given that pair-wise time periods are pooled meaning a large proportion of the same observations are used in consecutive regressions. It is interesting, however, that the SSSEM model appears to vary in the most haphazard fashion. One might have expected the SEM coefficients to vary more haphazardly given that the regression is re-estimated each year, while the SSSEM imposes a random walk process upon coefficient transitions.

Given that a semi-logarithmic functional form is used for the hedonic regressions, the above coefficients are hard to interpret. To transform the above coefficients into shadow prices (which are in dollars), they need only to be multiplied by the average price of the

houses in each regression (ie.  $\beta\bar{y}$ ).<sup>24</sup> Table 3 reports the shadow prices of the characteristics in (Australian) dollars resulting from the AP-DTH regressions:

**Table 3 – AP-DTH Shadow Prices of Characteristics (\$AUS)**

Regression	Transformed coefficients			
	Area	Bed	Bath	Carlug
90/91	5.33	12855.58	22306.20	5587.05
91/92	2.84	10966.54	24570.32	7053.14
92/93	2.62	9081.11	27282.22	7209.75
93/94	3.50	9329.99	28732.39	6731.67
94/95	3.43	12062.30	32089.59	4847.16
95/96	2.97	16602.70	31690.66	6394.21
96/97	6.02	16844.34	31908.81	6447.75
97/98	4.61	17363.85	35278.57	7599.45
98/99	3.98	18457.41	38647.37	9952.09
99/00	3.31	18703.66	42399.85	9389.46
00/01	3.35	16193.85	49608.79	8206.20
01/02	3.50	14310.56	58905.83	8483.93
02/03	5.11	15261.88	70111.51	11317.29
03/04	8.73	19214.66	83927.50	16257.60
04/05	9.76	25989.98	97631.61	20202.24

A point to note from Table 3 is that the AREA variable, while significant in the above regressions, seems to contribute little to the overall price. For example, for the period 2004/2005, a house on 500 m<sup>2</sup> of land implies that the size of the block only contributes \$4880 (=500\*9.76) to the overall house price. This may indicate that the size of the land is not as important as the location of the land. Alternatively, the constant term may be capturing much of the effect attributable to AREA. This has intuitive value if you consider that every house has a certain ‘base value’, which is the price of the block of land itself (without any house). The price then appreciates as structural features are added such as a house, tennis court, garage etc. In this case, the effects of the AREA variable and the intercept term may not be able to be disentangled, in which case an identification problem would exist.<sup>25</sup>

<sup>24</sup> See Hill et al. (2001) page 130.

<sup>25</sup> Note that colinearity does not affect prediction.

The magnitude of the shadow prices associated with the BATH property attribute are also worthy of discussion. In 2005, each extra bathroom adds almost \$100,000 to the overall sale price of a house. It is likely that the variable is capturing the effects of other omitted variables, ie. the number of bathrooms is potentially highly correlated with the splendor and luxury of a house. For instance, a house with three or four bathrooms is likely to have multiple stories and living areas. Alternatively and/or simultaneously, the explanatory variables may not be independent, given that the greater the number of bathrooms, the larger the house (usually), the larger the block of land and more bedrooms and carspaces. That is, the usual assumption that explanatory variables are independent of each other may be violated.<sup>26</sup>

*(ii) Computed House Price Indexes*

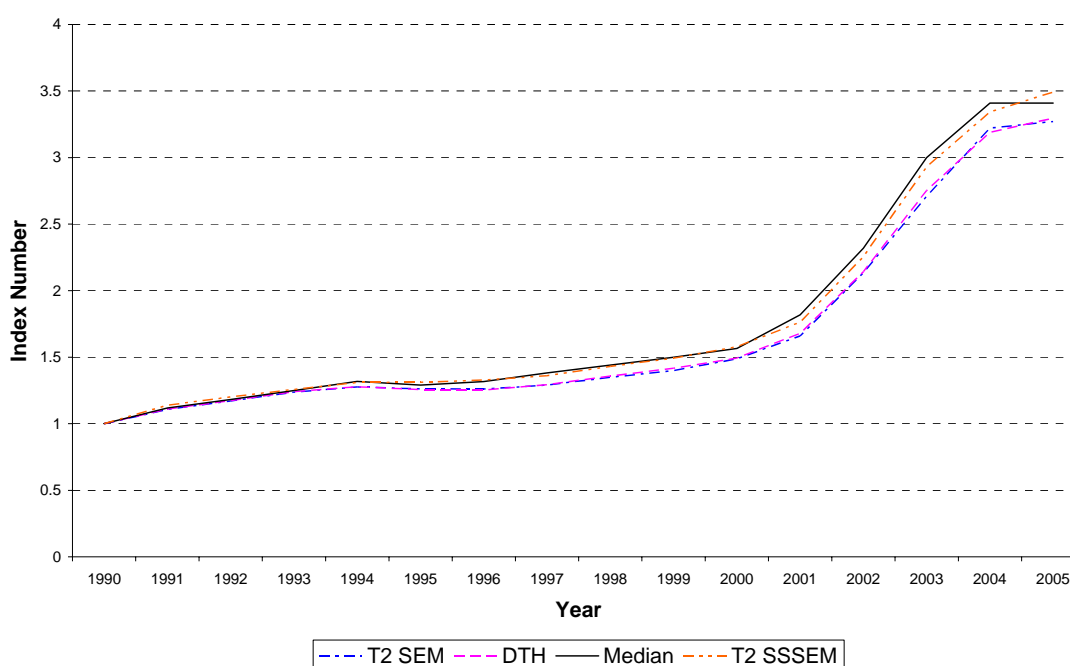
Graph 7 presents the computed annual indices for the DTH, SEM and SSSEM compared against an index of median house price changes. The SEM and SSSEM indices are computed using the T2 formula in (8). In order to compute a comparable annual index from the SSSEM (which was estimated monthly), ‘same month’ index numbers were computed (ie January 1990 to 1991, February 1990 to 1991 etc) and weighted geometric averages of these index numbers were computed for each year. The weighting on each index number was the ratio of the number of observations in that month with the total number of observations for the particular year.<sup>27</sup>

---

<sup>26</sup> The bathroom shadow prices for the SEM are more plausible but still quite outlandish. For instance, the estimated shadow price of each extra bathroom in 2005 is \$64,456.

<sup>27</sup> This weighted geometric average is preferable to an unweighted geometric average of the twelve ‘same month’ index numbers because it weights each house equally in the corresponding yearly index number. If, for instance, fewer houses were sold in December than in other months, an unweighted geometric average would overly weight the December observations in the yearly index number. Incidentally, both averages were computed to construct the yearly indices and only a marginal difference was found.

**Graph 7 – Annual Index Number Series**



The AP-DTH index is not presented because it is identical to the DTH index to two decimal places. As a result, the DTH index series is used for comparison purposes henceforth. Note that it was practically insignificant whether the approximate imputed prices given by (4) were used, or whether the unbiased imputed price formula in Appendix A was used for the SEM and SSSEM hedonic imputation indices. This paper endorsed T2 (see equation 8) as a particularly desirable index number formula for the housing case, however, Table 4 outlines that there is little practical difference between different index number formulae, as illustrated for the SEM.

**Table 4 – Chained Index Numbers (SEM)**

Period	Geometric Paasche	Geometric Laspeyres	Törnqvist Type 1	Törnqvist Type 2
1990	1	1	1	1
1991	1.1071	1.1082	1.1077	1.1076
1992	1.1684	1.1749	1.1716	1.1716
1993	1.2335	1.2402	1.2368	1.2368
1994	1.2739	1.2807	1.2773	1.2773
1995	1.2588	1.2655	1.2621	1.2621
1996	1.2564	1.2659	1.2611	1.2609

1997	1.2849	1.2959	1.2904	1.2901
1998	1.3413	1.3561	1.3487	1.3483
1999	1.3922	1.4071	1.3996	1.3993
2000	1.4800	1.4958	1.4879	1.4875
2001	1.6511	1.6672	1.6591	1.6588
2002	2.1223	2.1459	2.1341	2.1336
2003	2.6935	2.7231	2.7083	2.7077
2004	3.2057	3.2374	3.2215	3.2207
2005	3.2556	3.2871	3.2713	3.2705

The Geometric Paasche (GP) and Geometric Laspeyres (GL) indexes are remarkably similar, and given that the Törnqvist indexes are simply a geometric average of the GP and GL, they are also remarkably similar. It is pointless graphing these series as they lie virtually on top of one another. Of course, the differentiation between T1 and T2 (see Section 2) is redundant in this case, because the weights attributed to the GP and GL components are insignificant given the similarity between GP and GL. Furthermore, if  $N_t$  and  $N_{t+1}$  are quite similar then the difference between T1 and T2 will be less again.

Graph 7 illustrates that the SEM and DTH indices are remarkably similar, while the SSSEM index is systematically higher (more inflated) being closer to the median-based index, which seems to be an upper bound to all the indices. In theory, all the hedonic methods shown are an improvement on a simple median measure of house price change because they account for both compositional and quality changes. It is plausible that the median measure overstates the pure price changes because it fails to account for the rising quality of housing over time, due to renovations and the construction of new houses (assuming that this effect outweighs the depreciation effect). On the other hand, it is less likely that the difference is due to compositional changes, because it is consistently above the other index number series. That is, compositional changes (such as a greater proportion of sales occurring in expensive houses in a particular time period) are more likely to introduce volatility into the estimates over the short term. The annual median-based index series presented does not seem to exhibit volatility, however, a monthly index may be more susceptible to compositional changes, if it is subject to seasonality. Prasad and Richards (2006) did not find significant evidence of compositional change

effects for Brisbane, however, it would be interesting to test for such effects using the dataset specified.

It is initially surprising that the SEM index is closer to the DTH than the SSSEM. All the models have the same hedonics, however, one would expect the SEM and SSSEM indices to be similar given that the same index number formula is used and the models are both based on the same spatial autocorrelation structure. Table 5 shows the difference between the index numbers for the SSSEM and DTH model. A positive difference is associated with a higher SEM index number estimate, while a negative difference implies a higher DTH index number estimate. Not including 1990, the SSSEM estimates a higher index number than DTH for 10 out of the 15 years, however, in half the cases the difference is marginal. For the other half, the difference is one per cent or greater. Of particular influence is the difference in estimates for 1991, with the SSSEM index number 2.9% higher than the DTH counterpart. Given that Graph 7 shows the cumulative indices, this effect is accumulated through time. In fact, if the indices are rebased to 1991 (which removes the impact of this difference), the resulting index series is closer to the SEM and DTH series than it is to the Median.

**Table 5 – Difference between SSSEM and DTH Bilateral Index Numbers**

Year	SSSEM	DTH	Difference
1990	1.000	1.000	0
1991	1.139	1.110	0.029
1992	1.054	1.058	-0.004
1993	1.047	1.057	-0.010
1994	1.042	1.030	0.012
1995	1.000	0.983	0.018
1996	1.013	0.998	0.014
1997	1.025	1.032	-0.006
1998	1.051	1.050	0.001
1999	1.044	1.044	0.000
2000	1.057	1.053	0.004
2001	1.117	1.125	-0.007
2002	1.279	1.277	0.003
2003	1.299	1.285	0.014
2004	1.141	1.159	-0.018
2005	1.045	1.033	0.011

It is important to note at this point that the SSSEM produced here is only a preliminary effort, and future research needs to focus on adjusting monthly or quarterly indexes for seasonal influences<sup>28</sup>, and a control used for dealing with different types of housing (particularly acreage properties) when the sample size is small. It seems that the systematic overstating of index numbers from the SSSEM is particularly related to this latter point, with the inclusion of the dummy variable only partially controlling for over-inflated imputed prices on some observations.<sup>29</sup>

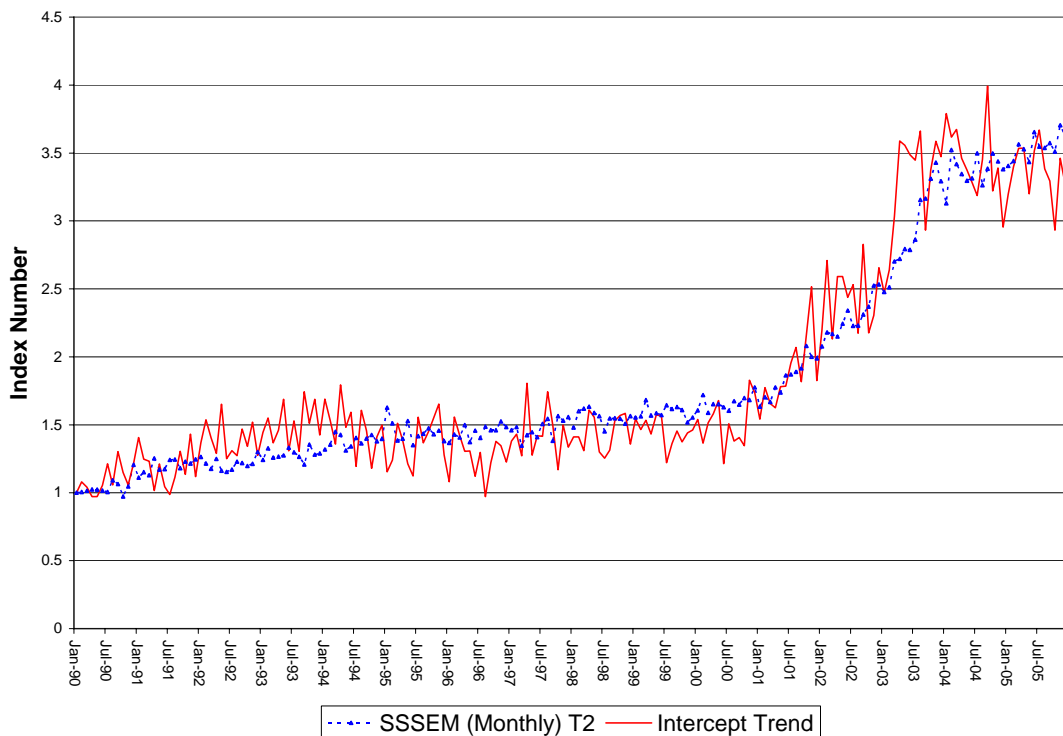
It was somewhat surprising that the DTH, AP-DTH and SEM hedonic imputation indexes were all so similar. On one hand there was strong evidence of spatial autocorrelation (as evidenced by the Moran and LR statistics in Table 2), but on the other, the resulting price indexes did not seem to be affected regardless of whether it was accounted for or not. Graph 2 implies that the indices may be remarkably similar because they are being driven by the magnitude of the intercept terms. Not only are the intercepts much larger in magnitude than the other coefficients, and consequently have the most effect on the imputed prices, but the series clearly tracks the computed indices. This is more clearly seen in Graph 8, which shows the case of the SSSEM. Notice the slow appreciation reflected in the intercept series through the 90s, and the sudden appreciation from 2000-2004. This implies that the pure price change effect is being captured by the intercept term, along with other omitted information from the hedonic functional form.

---

<sup>28</sup> The monthly index series seemed to exhibit some seasonality, however, was not examined in this research.

<sup>29</sup> It is unclear whether the acreage properties should be removed from the sample, or dealt with in some other robust manner – an issue to be addressed in future research.

**Graph 8 – SSSEM Monthly T2 Index vs Intercept trend**



Given that the hedonic regression contains only four hedonics, it is not surprising (in retrospect) that the resulting indices are so close. Namely, the difference between the hedonic methods (aside from spatially autocorrelated residuals) is the way in which the shadow prices are dealt with. However, given that the significance of each of the four hedonics to the overall house price is modest with the intercept providing the largest contribution, variations in the coefficients across models are not having a substantial effect on the resulting imputed prices and consequent index numbers.

A further explanation for the similarity between the models that incorporate spatially autocorrelated errors (SEM and SSSEM) and those that do not (DTH and AP-DTH) lies in the fact that hedonic imputation index formulae contain a ratio of imputed prices. Specifically, even though the estimated coefficients may be different in the SEM model compared with the AP-DTH model (for instance), so long as the difference was systematic in some way, then the difference incorporated into the imputed prices would

occur in both the numerator and denominator and consequently would not affect the resulting index number. This may suggest that adjusting for spatial autocorrelation is more important in a predictive context than for constructing index numbers.

Finally, it is interesting that Hansen (2006) and Prasad and Richards (2006), in their complementary studies for the RBA, also found that their favoured specifications for the time-dummy hedonic model, repeat sales model and mix-adjusted measure provided remarkably similar price index series for a selection of Australian cities over the period 1993:2 to 2005:3.

## **7. Conclusions**

The paper has focused on the construction of housing price index numbers using hedonic methods with particular focus on improvements to the specification and estimation of the hedonic function. Given the importance location plays in the determination of house prices, an attempt is made here is to incorporate the location effects through the specification of a spatially correlated structure for the disturbances in the hedonic model (SEM). Further, as the hedonic model is estimated using data on houses sold in different years and as the current study spans a long period since 1970, an attempt is made to accommodate the time-varying nature of hedonic coefficients using a random walk model. This feature of smoothly evolving hedonic coefficients with spatially autocorrelated disturbances are incorporated in the SSSEM considered in the paper. While the SEM model is estimated using maximum likelihood methods, the SSSEM has been estimated using the Kalman Filter.

The empirical part of the paper utilized data on housing sales, along with property attributes, from the Brisbane Metropolitan area. In order to incorporate spatial effects, geographical coordinates of each house in the sample are first identified. These data are used in modeling geographical contiguity using the Dealuny Triangulation method. An important feature of the data set is that the number of observations are significantly

higher over the later period of our sample with data for the years 1990 to 2005 accounting for more than 85% of the houses in the sample.

The price indexes computed using various approaches, DTH, AP-DTH, SEM and SSSEM models appear to yield fairly similar results. While the estimated hedonic coefficients from various models appear to differ significantly, their effect on the price indexes seems to be negligible. A major finding here is that the price index is largely determined by the movements in the constant term – the trend in the constant term is consistent with the reported increases in land prices in the Brisbane metropolitan area over the study period.

There is scope for further refinements and improvements in the methodology used in the paper. For instance, there is no clear consensus regarding the choice of weight matrix used in SEM. Geostatistical models are a promising alternative to the specification used here as these models do not require the specification of a weighting matrix. The problem of weighting estimates of price changes for individual households needs further consideration. In this paper simple unweighted geometric averages are used in measuring price changes. If sufficient data are available it may be possible to stratify the measures and employ strata weights in deriving the final index numbers.

## REFERENCES

- ABS (2005), 'Renovating the Established House Price Index,' *Information Paper No. 6417.0*
- Anselin, L. (1988). "Spatial Econometrics: Methods and Models." Dordrecht, Kluwer.
- Anselin, L. (1999). "Spatial Econometrics." (Available from [www.csiss.org/learning\\_resources/content/papers/baltchap.pdf](http://www.csiss.org/learning_resources/content/papers/baltchap.pdf))
- Bailey, M. J., R. F. Muth and H. O. Nourse (1963), "A regression method for Real Estate Price Index Construction," *Journal of American Statistical Association*, 58, pp 933-942.
- Basu, S. and Thibodeau, T. G. (1998), "Analysis of Spatial Autocorrelation," *Journal of Real Estate Finance and Economics*, 17:1, pp 61-95.
- 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.*
- Costello, G. (1997), "Transaction based index methods for housing market Analysis," *Australian Land Economics Review*, 3(2), pp 19–27.
- Diewert, W. E. (1976), "Exact and superlative index numbers," *Journal of Econometrics*, 4, 115-145.
- Diewert, W. E. (2003a), "Hedonic Regressions: A Review of Some Unresolved Issues," available from [siteresources.worldbank.org/ICPINT/Resources/Hedonics.doc](http://siteresources.worldbank.org/ICPINT/Resources/Hedonics.doc)
- Diewert, W. E. (2003b), "Hedonic Regressions: A Consumer Theory Approach," (Chapter 10) in Feenstra, R. C and M. D. Shapiro, *Scanner Data and Price Indexes*, Studies in Income and Wealth, 64, The University of Chicago Press.
- Diewert, W. E. (2004), "A New Axiomatic Approach to Index Number Theory," *Discussion Paper 04-05*, Department of Economics, University of British Columbia.
- Dubin, R. A. (1998), "Spatial Autocorrelation: A Primer," *Journal of Housing Economics*, 7, pp 304-327.
- Flaherty, J. (2004), "Measuring and evaluating changes in returns for residential Property," paper presented at The Tenth Annual Pacific Rim Real Estate Society Conference, Bangkok, 25–28 January.
- Goodman, A. C., and Thibodeau T. G. (1995), "Age-Related Heteroskedasticity in Hedonic House Price Equations," *Journal of Housing Research* 6(3), pp 25-42.
- Hansen, J. (2006), "Australian House Prices: A comparison of Hedonic and Repeat Sales Measures," *Reserve Bank of Australia Discussion paper*, 2006-03.
- Harvey, A. C. (1990), *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press.
- Hill, R. C., J. R. Knight, and C. F. Sirmans (1997a), "Estimating Capital Asset Price Indexes," *The Review of Economics and Statistics*, 79, pp 226-233.
- Hill, R. C., W. E. Griffiths and G. G. Judge (2001), *Undergraduate Econometrics*, John Wiley and Sons, Inc.
- Hill, R. J. and D. Melser (2005), "Constructing Panel Price Indexes using Hedonic Methods: The Case of House Prices in Sydney," *Working paper presented at NZAE conference*, (available from [www.nzae.org.nz/conferences/2005/62-HILLandMelser.pdf](http://www.nzae.org.nz/conferences/2005/62-HILLandMelser.pdf))
- Hill, R. J. and M. P. Timmer (2006), "Standard errors as weights in multilateral price indexes," *Journal of Business and Economic Statistics*, 24:3, pp 366-377.

- LeSage, J. P. (1998), "Spatial Econometrics", (Available from <http://www.spatial-econometrics.com>)
- LeSage, J. P. and R. K Pace (2004), "Models for Spatially Dependent Missing Data", *Journal of Real Estate Finance and Economics*, 29:2, pp 233-254.
- Militino, A. F., M. D. Ugarte and L. Garcia-Reinaldos (2004), "Alternative Models for Describing Spatial Dependence among Dwelling Selling Prices," *Journal of Real Estate Finance and Economics*, 29:2, pp 193-209.
- Pace, R. K., and J. LeSage. (2004). "Chebyshev Approximation of Log-determinants of Spatial Weight Matrices," *Computational Statistics and Data Analysis*, 45, pp 179-196.
- Pace, R. K., and R. Barry (1997). "Quick Computation of Spatial Autoregressive Estimators," *Geographical Analysis*, 29, pp 232-246.
- Pace, R. K. and O. W. Gilley (1998), "Generalising the OLS and Grid Estimators," *Journal of Real Estate Economics*, Vol. 26:2, pp 331-347.
- Prasad, N. and Richards, A. (2006), "Measuring House Price Growth – Using Stratification to Improve Median based Measure," *Reserve Bank of Australia Discussion Paper*, 2006-04.
- Silver, M. and S. Heravi (2006), "The Difference Between Hedonic Imputation Indexes and Time Dummy Hedonic Indexes," *IMF Working Paper*, WP/06/181.
- 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

## APPENDIX A:

### Bias in imputed prices arising from Semi-log Hedonic Functions

Consider the following hedonic model:

$$\mathbf{y}_t = \mathbf{X}_t \boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t \quad t = 1, \dots, T \quad (\text{A.1})$$

where:

$\mathbf{y}_t$  is an  $N_t \times 1$  vector of (log) house prices and  $N_t$  is the number of houses sold in time period  $t$ ;

$\mathbf{X}_t$  is an  $N_t \times k$  matrix of hedonic characteristics and a constant term;

$\boldsymbol{\beta}_t$  is a  $k \times 1$  vector of unknown parameters;

$\boldsymbol{\varepsilon}_t$  is the  $N_t \times 1$  vector of residuals and  $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Omega}_t)$ ;

Imputed prices which are estimated using (A.1) are biased as a result of the impact of ‘undoing’ the logarithmic transformation.

$$\hat{P}_s^h(\mathbf{x}_t^h) = \exp\left(\sum_{c=1}^k x_{c,t}^h \hat{\beta}_{c,s}\right) \quad (\text{A.2})$$

where (A.2) is equivalent to equation (4).  $\hat{P}_s^h(\mathbf{x}_t^h)$  represents the imputed price of house  $h$  in period  $t$  using the hedonic coefficients from the period  $s$  regression.

Consider the regression estimates  $\hat{\boldsymbol{\beta}}_t$  from (A.1), under the assumption that  $\hat{\boldsymbol{\beta}}_t \sim N\left(\boldsymbol{\beta}_t, \boldsymbol{\sigma}_\varepsilon^2 (\mathbf{X}_t' \boldsymbol{\Omega}_t^{-1} \mathbf{X}_t)^{-1}\right)$ . Given the assumption of normality, the bias can be determined as follows:

$$\begin{aligned}
\hat{\beta}_t &\sim N\left(\beta_t, \sigma_\varepsilon^2 (\mathbf{X}_t' \boldsymbol{\Omega}_t^{-1} \mathbf{X}_t)^{-1}\right) \\
\therefore \mathbf{x}_t' \hat{\beta}_t &\sim N\left(\mathbf{x}_t' \beta_t, \sigma_\varepsilon^2 \mathbf{x}_t' (\mathbf{X}_t' \boldsymbol{\Omega}_t^{-1} \mathbf{X}_t)^{-1} \mathbf{x}_t\right) \\
\therefore \exp\left(\mathbf{x}_t' \hat{\beta}_t\right) &\sim \text{LN}\left(\mathbf{x}_t' \beta_t, \sigma_\varepsilon^2 \mathbf{x}_t' (\mathbf{X}_t' \boldsymbol{\Omega}_t^{-1} \mathbf{X}_t)^{-1} \mathbf{x}_t\right) \\
\therefore E\left[\exp\left(\mathbf{x}_t^{h'} \hat{\beta}_t\right)\right] &= \exp\left(\mathbf{x}_t^{h'} \beta_t + \frac{1}{2} \sigma_\varepsilon^2 \mathbf{x}_t^{h'} (\mathbf{X}_t' \boldsymbol{\Omega}_t^{-1} \mathbf{X}_t)^{-1} \mathbf{x}_t^h\right) \\
&> \exp\left(\mathbf{x}_t^{h'} \beta\right)
\end{aligned}$$

where  $\mathbf{x}_t^h$  is a  $(k \times 1)$  vector that is the transpose of the row of  $\mathbf{X}_t$  corresponding to house  $h$ .

Note that using (A.2) leads to an upward bias because  $(\mathbf{X}_t' \boldsymbol{\Omega}_t^{-1} \mathbf{X}_t)^{-1}$  is a positive definite matrix, and consequently  $\mathbf{x}_t' (\mathbf{X}_t' \boldsymbol{\Omega}_t^{-1} \mathbf{X}_t)^{-1} \mathbf{x}_t > 0$ .

Hence, unbiased imputed prices can be calculated as follows:

$$\hat{P}_s^h(\mathbf{x}_t^h) = \exp\left(\mathbf{x}_t^{h'} \hat{\beta} - \frac{1}{2} \hat{\sigma}_\varepsilon^2 \mathbf{x}_t^{h'} (\mathbf{X}' \hat{\boldsymbol{\Omega}}^{-1} \mathbf{X})^{-1} \mathbf{x}_t^h\right) \quad (\text{A.3})$$

**Appendix B:**  
**List of postcodes and (a selection of) suburbs covered by the dataset**

Postcode	Suburb
4000	Brisbane Central, Spring Hill
4005	Merthyr, Teneriffe
4006	Bowen Hills, Exhibition, Fortitude Valley, Herston, Mayne, Newstead
4007	Ascot, Doomben, Hamilton
4008	Meeandah, Myrtle town, Pinkenba
4010	Albion, Breakfast Creek
4011	Clayfield, Eagle Junction, Hendra
4012	Nundah, Toombul, Wavell Heights
4013	Northgate
4014	Banyo, Virginia
4017	Bracken Ridge, Brighton, Deagon, Sandgate
4030	Kalinga, Lutwyche, Windsor, Woolloowin
4034	Aspley, Boondall, Carseldine, Geebung, Zillmere
4035	Albany Creek, Bridgeman Downs
4036	Bald Hills
4051	Alderley, Enoggera, Gaythorne, Grange, Newmarket, Wilston
4053	Everton Hills, Everton Park, McDowall, Mitchelton, Stafford
4054	Arana Hills, Grovely, Keperra
4055	Ferny Grove, Ferny Hills, Upper Kedron
4059	Kelvin Grove, Red Hill
4060	Ashgrove, Dorrington
4061	The Gap
4064	Milton, Paddington, Rosalie
4065	Bardon, Mt. Cootha, Ramworth
4066	Auchenflower, Toowong, Torwood
4067	St. Lucia
4068	Chelmer, Indooroopilly, Taringa
4069	Brookfield, Fig Tree Pocket, Kenmore, Kenmore Hills, Pinjarra Hills
4070	Anstead, Beebowrie, Moggill
4073	Seventeen Mile Rocks, Sinnamon Park
4074	Jamboree Heights, Jindalee, Middle Park, River Hill, Westlake
4075	Corinda, Graceville, Oxley, Sherwood
4076	Darra
4077	Doolandella, Durack, Inala, Richlands
4101	Highgate Hill, South Brisbane, Westlake
4102	Dutton Park, Woolloongabba
4103	Annerley, Fairfield, Fruitgrove
4104	Yeronga
4105	Moorooka, Tennyson, Yeerongpilly
4108	Archerfield, Coopers Plains
4110	Acacia Ridge, Heathwood, Larapinta
4113	Eightmile Plains, Runcorn

4115	Algester
4116	Calamvale, Drewvale
4117	Berrinba, Karawatha
4120	Greenslopes, Stones Corner
4121	Holland Park, Holland Park West, Tarragindi, Wellers Hill
4122	Mansfield, Mt. Gravatt, Mt. Gravatt East, Wishart
4123	Rochedale
4151	Cooparoo
4152	Camp Hill, Carina, Carindale, Whites Hill
4153	Belmont
4154	Gumdale, Ransome
4155	Chandler
4156	Burbank, MacKenzie
4157	Capalaba
4169	Chapel Hill, Kangaroo Pt.
4170	Cannon Hill, Morningside, Norman Park, Seven Hills
4171	Balmoral, Bulimba, Hawthorne
4172	Murarrie
4173	Tingalpa
4178	Lytton
4179	Lota, Manly, Manly West
4300	Bellbird Park, Camira, Carole Park
4306	Benarkin, Amberley, Banks Creek (and many others)

Due to the downloading problems outlined in Cominos (2006) Appendix 1, with only partial downloads in the local authorities of Kedron, Sherwood, Tingalpa and Yeerongpilly, a selection of postcodes (and corresponding suburbs) have been omitted from the database. Some of these are contained in the following table:

#### **Omitted Postcodes and Suburbs**

<b>Postcode</b>	<b>Suburb</b>
4018	Taigum
4019	Margate
4021	Kippa-ring
4031	Kedron
4032	Chermside
4037	Eatons Hill
4107	Salisbury
4109	McGregor, Sunnybank, Sunnybank Hills
4114	Logan Central
4118	Heritage Park, Hillcrest
4129	Loganholme
4131	Loganlea, Meadowbrook

## APPENDIX C:

### Moran, LR and LM Tests for Spatial Autocorrelation

The Moran I statistic and Likelihood-ratio test are common measures used to test for the presence of spatial autocorrelation. The null hypothesis for both tests is as follows:

$H_0$ : No spatial autocorrelation ( $\rho = 0$ )

The rejection of the null hypothesis implies that spatial autocorrelation is present in the data and that OLS will lead to spatially autocorrelated residuals. Hence, one of the models outlined in Section 3.2.1 or 3.2.2 is needed.

Moran's I Statistic (see Anselin, 1988) is given by:

$$I = \left( \frac{n}{S} \right) \frac{\boldsymbol{\varepsilon}' \mathbf{W} \boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon}}, \quad (\text{C.1})$$

where  $n$  is the number of observations,  $\boldsymbol{\varepsilon}$  is the  $n \times 1$  vector of OLS residuals,  $\mathbf{W}$  is the spatial weight matrix and  $S = \sum_i \sum_j w_{ij}$  (ie. the sum of all the elements of  $\mathbf{W}$ ). Note that if  $\mathbf{W}$  is row-normalised then the statistic reduces to:

$$I = \frac{\boldsymbol{\varepsilon}' \mathbf{W} \boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon}} \quad (\text{C.2})$$

The asymptotic distribution for Moran's I based on least-squares residuals corresponds to a standard normal distribution after adjusting the I-statistic by subtracting the mean and dividing by the standard deviation of the statistic. Again assuming that  $\mathbf{W}$  is standardized, then the adjustment is of the following form:

$$E(I) = \text{tr}(\mathbf{M}\mathbf{W}) / (n-k)$$

$$V(I) = \left[ \text{tr}(\mathbf{M}\mathbf{W}\mathbf{M}\mathbf{W}') + \text{tr}(\mathbf{M}\mathbf{W})^2 + (\text{tr}(\mathbf{M}\mathbf{W}))^2 \right] / d - E(I)^2$$

$$d = (n-k)(n-k+2)$$

$$Z_I = [I - E(I)] / V(I)^{1/2}$$

$$\text{where } \mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Anselin (1999) states that

“Moran’s I test has been shown to be locally best invariant [King (1981)] and consistently outperforms other tests in terms of power in simulation experiments [see, e.g., Bartels and Hordijk (1977), Anselin and Rey (1991), Anselin and Florax (1995b), Kelejian and Robinson (1998)].”

Cressie (1993) provides the (modified) log-likelihood test statistic as follows:

$$LR = 2 \left( \frac{n-p-r}{n} \right) (L_p - L_{p+r}), \quad (\text{C.3})$$

where  $L_p$  is the negative log-likelihood for the restricted model with  $p$  parameters and  $L_{p+r}$  is the negative log-likelihood for the unrestricted model with  $p+r$  parameters. In general, the LR statistic is asymptotically distributed as  $\chi^2$  with  $r$  degrees of freedom.