An improved National Price Index using Land Registry Data

June 22, 2007

Abstract

The methods of constructing house price indices in the UK have lagged behind those employed in some Western countries. The deficiencies in these methods weaken the decision-making ability of investors, stakeholders and policy makers.

This paper overviews existing methodologies and introduces the repeat sales regression (RSR) method in a UK context, utilising Land Registry data. Particularly, we build upon the argument presented by Leishman & Watkins (2002) that some deficiencies of the existing indices can be remedied in part through the application of the repeat sales regression method using Land Registry Data.

We initially summarise evidence and conclusions drawn from previous studies in order to demonstrate through application that repeat sales regression analysis on the recently released Land Registry 'price paid' dataset, is the preferable methodology for the measurement of house price movements in England & Wales; addressing some of the shortcomings associated with the current methodologies.

We present empirical research conducted by the author, illustrating the first application of repeat sales regression for England & Wales using Land Registry data.

Keywords: *house price index, repeat sales regression, land registry.*

REVISED DECEMBER 2006

Contents

1 Introduction **3**

2 Current Indices and the Land Registry dataset **4**

3 Quality Adjustment Methodologies **6**
3.1 Mix-adjustment method (weighted averages) 6
3.2 Hedonic methods 7
3.3 Repeat sales regression 8
3.3.1 Methodology 8
3.3.2 Model Estimation 10

4 Practical and theoretical limitations of the various methodologies **10**
4.1 Mix-adjustment and Hedonic Models 11
4.2 Repeat sales regression 13
4.2.1 Heteroskedasticity 13
4.2.2 Sample Selection 13
4.2.3 Inconstancy of attribute appreciation 14
4.2.4 Multicollinearity 14
4.2.5 House Improvement Adjustment 15
4.2.6 Inefficiency 15

5 Comparison of indices produced using the Land Registry dataset **16**
5.1 National indices compared 18
5.2 RSR vs Hedonic 19
5.3 RSR vs. Mix-adjustment 19

6 Concluding remarks **21**

1 Introduction

A dependable and precise source of house price information is required for a variety of purposes. Changes in house prices can have a significant effect on consumer spending and saving patterns, for the majority of the population housing represents the most valuable asset they are ever likely to own. Moreover, information on house price levels and growth rates form the basis of key decisions made by a wide variety of decision makers, for example; home buyers, home sellers, mortgage lenders, valuation surveyors, and house builders. Accurate price series on a large number of assets, such as equities and bonds, is an essential feature of financial market research. Analogous information for local property markets (e.g. cities and towns) would be useful not only to researchers but also to town officials and homeowners.

Furthermore, economic and social policy, on a national and local level, require reliable house price information. House prices feed directly into the RPI, through measured physical housing depreciation and mortgage interest payments: physical housing depreciation represented approximately one third of annual RPI-X inflation in December 2002. This in turn plays a significant role in the setting of interest rates by the Monetary Policy Committee. House prices are an important consideration in assessing macroeconomic developments.

The measurement of house price movements is by no means an easy task. The principal problem is due to the fact that houses are heterogeneous goods. No two houses are exactly the same. Secondly, houses are sold infrequently; between 3% and 7% of all houses transact each year. Thirdly, as prices are negotiated, particular circumstances for individual buyers and sellers can lead to the situation that even extremely similar houses sell for very different amounts.

Given the nature of individual houses and the property market, it is convenient and useful to capture the overall average price trends followed by a group of houses. In such a situation, it is common practice in economics to propose a single price index, e.g. consumer price index. A house price index is simply "one of many plausible measures of the central tendency of house price appreciation for a particular group of properties" (Araham and Schauman, 1991).

Current problems regarding evaluation of the property market in the UK, are exacerbated by the lack of local or city level price indices. This fact gives grounds for several indices to be produced. There is a clear argument for the need to produce multiple indices.

The only official national dataset involving each completed property transaction is provided by Land Registry and is described in the next section. In section 3 we overview the widely used Mix-adjustment and Hedonic methodologies. In the same section we present a detailed description of the RSR method. Section 4 discusses the relative merits and shortcomings of each method, both from a theoretical and a practical level. The implementation of RSR using the Land Registry dataset is discussed in section 5. Direct comparisons are made with the other methodologies as a means of highlighting the significant advantages brought by a RSR based price index using Land Registry data. Finally, section 6 summarises and discusses our

findings.

2 Current Indices and the Land Registry dataset

Commentators including the Office of National Statistics acknowledge that none of the current house price indices completely meets user requirements. The current most popular measurements are those published by Halifax, DCLG and Nationwide using proprietary datasets. For Halifax and Nationwide, the recorded price refers to the negotiated price at time of approval of application for a mortgage at each lender. All cash purchases are excluded. It is worth noting that the indices produced are based on approved applications which may not always go through completion. Also, for applications that do proceed to completion, the price of some transactions may be re-negotiated prior to completion. Therefore, the Halifax and Nationwide indices may not be a true reflection of prices actually paid. DCLG calculates indices based on prices at completion of a mortgage.

The lack of reliable house price indicators weakens the decision-making ability of policy makers and investors. Nicol (1996) claims that such misinformation created problems for the Monetary Policy Committee of the Bank of England and exacerbated the problems of the last housing market recession. Leishman and Watkins (2002) claim that this has had significant implications for a range of principals in the housing market including private investors and policy makers. Neither the DCLG, Halifax nor Nationwide house price indices provide information to a geographical resolution greater than Government Office Regions.

The methods used by the three main house price indices (DCLG, Halifax and Nationwide) are variants of the Hedonic regression model but comparisons between the indices usually send conflicting messages. Leishman & Watkins (2002) express unease about the use of these house price indices in the Retail Price Index for the calculation of mortgage interest payments, depreciation and other costs borne by owner-occupiers.

The problems generally stem from the limitations in the volume of data that is being analysed. A lack of sample size introduces potential statistical bias and unreliability, particularly on a monthly basis while examining smaller geographical areas. Of the three main indices in use, the new DCLG house price index has the largest sample size. Since September 2005 the new mix-adjusted house price index is based on an enlarged sample of completions data (about 45,000 per month) from about 50 mortgage lenders who supply data through the Regulated Mortgage Survey (RMS) of the Council of Mortgage Lenders (CML)/BankSearch. Prior to this date the index was based on the Survey of Mortgage Lenders (SML) (about 25,000 completions per month). The number of cases received will also be affected by the total number of mortgages that have been completed. Table 1 presents the house price indices produced by several institutions based on monthly observations.
The new Land Registry 'price paid' dataset which was recently released (Jan 1^{st} 2005) contains approximately 100,000 transactions per month ($\sim$ 100% of housing completions) and incorporates data from Apr 1^{st} 2000 onwards. It includes all UK

	Source data	Observations (approx.)	Quality adj. method	Local Indices	Frequency	Price Observations
New Land Reg.	Land Registry	100k	RSR	Yes	Monthly	Actual price paid
DCLG	RMS data	45k	MA	Regional only	Monthly	Mortgage Completions
Halifax	Loan Approvals	12k	HR	Regional only	Monthly	Valuation at time of Loan approval
Nationwide	Loan Approvals	¡12k	HR	Regional only	Monthly	Valuation at time of Loan approval
Old ODPM	SML Data	¡5k	MA	No	Quarterly	Mortgage Completions
Hometrack	Estate Agents	¡4k	MA	No	Monthly	Estate agents reports
Calnea/HomeCo	Asking Prices online	100k	MA	Regional only	Monthly	Asking Price
Rightmove	Asking prices online	undisclosed	MA	Regional only	Montly	Asking Price
FT/Acadametrics	Land Registry	100k	MA	No	Monthly	Actual Price paid
Old Land Reg.	Land Registry	100k	None	No	Quarterly	Actual price paid

Table 1: Comparison of house price indices produced by several institutions. RSR: Repeat Sales Regression; HR: Hedonic Regression; MA: Mixed adjustment.

registered residential property transactions and it provides unambiguous, accurate information on the Exact Address, Postcode, Date and Price Paid for each transaction. Information provided on Property Type is acknowledged by Land Registry as slightly less reliable. The dataset incorporates no information on any other qualitative or quantitative characteristics of a house, like number of bedrooms, toilets, living rooms, number of garages and garage spaces etc. Leishman, Watkins and Fraser (2002) lamented the lack of methodologically consistent local housing market price indices and suggested that such indices be constructed by Land Registry.

The details held by Land Registry are updated following the receipt of an application at the appropriate local office. For guidance, on average this is a couple months after a property transaction. For a given month around 25% of the actual number of transactions are recorded at the end of the month, 80% by the end of the following month and 90% by the end of the second month. This time lag is reducing due to Land Registry improvements and the introduction of e-conveyancing is expected to have a further positive effect.

3 Quality Adjustment Methodologies

Price indices can be created using simple averages; the current quarterly Land Registry reports are examples of this. However, as different properties are sold in each period, comparisons of simple average prices are prone to error. As previously noted, houses are heterogeneous goods as opposed to commodities. The volatility of simple average prices, as argued later in this paper,is primarily due to the lack of any quality adjustment.

Reliable indicators require a system of measurement which adequately allows for differences in the sample of houses traded; in other words the sample data used in index calculation must be quality adjusted. Aside from the quality of the underlying data sample, the method chosen to adjust for heterogeneity in the data sample is the primary factor in determining the value of a particular system of measurement. This paper discusses the three main types of quality adjustment methods:

1. The Mix-adjustment Method
    - An approach based on weighted averages.
2. The Hedonic Method
    - An approach based on Hedonic Regression.
3. The RSR Method
    - An approach based on Repeat Sales Regression.

3.1 Mix-adjustment method (weighted averages)

The most commonly used method of quality adjustment is the Mix-adjustment method, occasionally called the 'matrix' or 'weighted averages' approach. The popularity of this method is due to the ease of its implementation. It can improve on the

reliability of an index calculated using simple averages by applying weightings to the constituents of the averages.

Examples of this method include the old ODPM (DETR) Index, the FT House Price Index produced by Acadametrics, the Rightmove Asking Price Index, and the Hometrack Index.

In this method a matrix is constructed dividing house price observations into groups or 'cells' of observations depending on various property characteristics. Examples of characteristics collected by Land Registry that can be used to define these cells are:

- Property Type (i.e. Detached, Semi-Detached, Terrace or Flat)
- Region (e.g. governmental office regions)

3.2 Hedonic methods

A Hedonic econometric model is one where the independent variables are related to quality; e.g. the quality of a product that one might buy or the quality of a job one might take. For example, a Hedonic model of wages might correspond to the idea that there are compensating differentials - that workers would get higher wages for jobs that were more unpleasant. The Hedonic methodology was devised in the context of measuring price changes for goods which consist of a wide range of characteristics.

The idea behind the Hedonic model of Lancaster (1966), in the context of constructing house price indices, is that the price of a house can be accurately estimated from its individual qualitative and quantitative properties or characteristics. The obvious problem is that when a property is sold, the selling price corresponds to the worth of the house as a unity, as a bundle of characteristics; the contribution of each characteristic to the total price cannot be discerned. Given sufficient information and using multivariate regression, the value that the market attributes to each of those characteristics can be estimated. In particular, Hedonic regression uses a sample of house prices and by considering their individual characteristics it calculates the underlying market price for the unit of each characteristic. These values are frequently referred to as "characteristics prices". Therefore a property is valued according to the price of the locational and physical attributes it possesses.

It is then possible to estimate the change in average price from one period to another by holding the set of characteristics constant (standardisation). A standardised index is then obtained as the ratio of the price at a time period t to the price at a reference point in time (price at the base period). It is sensible that when the original standard is no longer representative, then a new starting point in time is defined and considered as the basis.

Instead of delivering a simple summary of either raw growth rates or prices, Hedonic methods are trying to uncover the functionally correct mathematical model of house prices. In other words the models make explicit assumptions that both the functional form of the regression equation and the various parameters' values are correct.

Hedonic regression is essentially a parametric method. Parametric models are those

where the model structure is specified *a priori*. The term parametric is meant to imply that the number and nature of the parameters is fixed in advance. The Hedonic method was originally introduced in 1939; Rosen (1974) established its theoretical foundation while other early studies were conducted by Fleming and Nelis (1974); more recent references include Case et al. (1991), and Meese and Wallace (1991, 1997). For a more authoratative review of the hedonic literature, the reader could refer to Malpezzi, S (2003) Hedonic pricing models: a selective and applied review, in O'Sullivan, A and Gibb, K (eds) Housing Economics and Public Policy, Blackwells, Oxford.

3.3 Repeat sales regression

In this paper we suggest that the RSR method is a solution to the quality adjustment problems faced by the Land Registry dataset (Leishman, Watkins & Fraser, 2002; Costello & Watkins, 2002; Thwaite & Woods, 2003).

Repeat sales methodologies focus on price changes rather than prices themselves, directly measuring these changes by examining only properties that have been sold at least twice. These measurements are combined in fairly intuitive ways to form estimates of the price index or growth rate in any particular time period. By using only properties that have been sold at least twice, contributing factors to variation in price growth are controlled.

The RSR method of estimating house price indices was first introduced by Bailey, Muth and Nourse (1963). The main idea behind RSR is that the market-wide growth rate for a given period is reflected in averaging the observed individual growth rates of all properties that were transacted twice in that time period (Leishman, 2000). Wherever the necessary data has been available, RSR has gained popularity in economic application. RSR is now widely adopted by a number of large private, state and federal organizations in the United States, for example the Housing Economics and Financial Research Department at Freddie Mac (Federal Home Loan Mortgage Corporation). RSR is also a tool widely used in the study of other markets characterised by infrequent trading, such as the fine art market (Goetzmann, 1992).

3.3.1 Methodology

The model underlying the Bailey et al (1963) method can be written as follows, using appropriate notation. Any property n that has been sold twice satisfies the following equation,

$$R_{n,t_1,t_2} = \frac{P_{n,t_2}}{P_{n,t_1}} = \frac{I_{t_2}}{I_{t_1}} \times U_{n,t_1,t_2},$$

where P_{n,t_1} and P_{n,t_2} are the prices at which property n was sold at time periods t_1 and t_2 respectively, I_{t_1} and I_{t_2} correspond to the unknown indices at the times mentioned above and finally U_{n,t_1,t_2} is an idiosyncratic error term; $t_1 < t_2$, for $t_1 = 0, 1, \ldots, T-1$, $t_2 = 1, \ldots, T$.

The model means that the ratio of the final sales price in period t_2 to initial sales price in period t_1 for the n^{th} property, which is defined as R_{n,t_1,t_2}, is equal to the ratio

of the (unknown) indices of the corresponding two periods with a property-specific noise term.

The intuition behind the model is obvious. A pair of sales prices of a given property contains information on house price appreciation happening in the market it belongs to within the periods between the first and second sale (Case & Shiller, 1987). Therefore, the observed price appreciation between the two sales of this given property can be attributed to two factors: 1) the general trend of appreciation of the housing market this property belongs to, and 2) some property-specific elements that drive its house price to deviate from the overall trend of the housing market. The first factor is represented in the index ratio $\frac{I_{t_2}}{I_{t_1}}$, while the second factor is captured by the term U_{n,t_1,t_2}, in the above model.

To make the Bailey model more practical, one can transform the model into a linear form by taking logarithm of both sides of the equation,

$$\log(R_{n,t_1,t_2}) = -\log(I_{t_1}) + \log(I_{t_2}) + \log(U_{n,t_1,t_2}) \text{ or}$$

$$r_{n,t_1,t_2} = -i_{t_1} + i_{t_2} + u_{n,t_1,t_2},$$

where lower case letters stand for the logarithms of the corresponding capital letters.

In the Bailey model, it is assumed that the error terms u_{n,t_1,t_2} have a mean of zero, constant variance, and are uncorrelated with each other and any i_t.

Recall the goal is to estimate I_t, or, equivalently, i_t, for $t = 0, 1, \ldots, T$. If one introduces $T + 1$ dummy variables x_t, for $t = 0, 1, \ldots, T$, and rewrites the model above as

$$r_{n,t_1,t_2} = \sum_{t=0}^{T} i_t x_t + u_{n,t_1,t_2}$$

where, if we denote initial sale and final sale periods as t_1 and t_2 respectively,

$$x_t = \begin{cases} -1, & \text{if the period } t = t_1 \text{ is an initial sale;} \\ +1, & \text{if the period } t = t_2 \text{ is a final sale;} \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, the model becomes a multiple linear regression model with T as the dummy independent variable. In matrix notation, the model is

$$Y = X\beta + \epsilon,$$

where Y is a column vector containing the log of relative prices for each property

$$Y = \begin{pmatrix} r_{1,t_1,t_2} \\ r_{2,t_1,t_2} \\ r_{3,t_1,t_2} \\ \ldots\ldots\ldots \\ r_{n,t_1,t_2} \end{pmatrix},$$

If there are altogether M pairs of repeat sales in the sample, then X is an $M \times (T+1)$ matrix and at each row the t^{th} component is -1 if $t = t_1$, 1 if $t = t_2$ and 0 otherwise as the variables X_t was defined before. Thus the matrix X is of the form

$$X = \begin{pmatrix} 0 & -1 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \\ .. & .. & .. & .. & .. \\ -1 & 0 & 1 & \dots & 0 \end{pmatrix}$$

Finally, β is a $(T+1) \times 1$ column vector of the form

$$\beta = \begin{pmatrix} i_0 \\ i_1 \\ \dots \\ i_T \end{pmatrix}.$$

3.3.2 Model Estimation

The model can be estimated by ordinary least squared (OLS) method, which can be implemented by most statistical software packages. The regression outputs are the estimated parameters $\widehat{i_t}$, for $t = 0, 1, \dots, T$, which in matrix form is

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

However, this is still in logarithm form. In addition, it is convention to have an index of 100 for base period $t = 0$, i.e. $\bar{I}_0 = 100$. If we denote the indices estimated from OLS as $\widehat{I}_t$, and indices after rebasing as $\bar{I}_t$, for $t = 0, 1, \dots, T$, the relationship is the following:

$$\frac{\widehat{I}_t}{\widehat{I}_0} = \frac{\bar{I}_t}{\bar{I}_0},$$

which, by taking natural logarithms, in both sides of the above equation is equivalent to

$$\log(\bar{I}_t) = \log(\widehat{I}_t) - \log(\bar{I}_0) + \log(\bar{I}_0) = \widehat{i_t} - \widehat{i_0} + \log 100.$$

Therefore, for $t = 0, 1, \dots, T$,

$$\bar{I}_t = \exp(\widehat{i_t} - \widehat{i_0} + \log 100).$$

We hence get house price indices I_t, for $t = 0, 1, \dots, T$, after rebasing $\bar{I}_0 = 100$.

4 Practical and theoretical limitations of the various methodologies

A large number of academic papers have been written comparing the relative merits of the various methodologies and the overall consensus seems to be that given idealised data samples, there was little to distinguish between the theoretical merits of

the Hedonic and RSR methods. In 1992 Crone and Voith compared Hedonic and RSR methods, although their findings were vague, they favoured RSR, mainly because RSR was the method least affected by reductions in sample size. In 1991 Hosios and Pesando found in favour of RSR methods. In 1994 Gatzlaff and Ling compared a variety of RSR and Hedonic methods. They found that both of these methods produced precise estimates of the index and growth rates. In 2000, Leishman summarised the research claiming US studies show that the RSR and Hedonic methods are on par when given similar data sample sizes. The Mix-adjustment method is generally considered inferior to both Hedonic and RSR methods when constructing house price indices due its incomplete quality adjustment procedure.

4.1 Mix-adjustment and Hedonic Models

The Bank of Englands Structural Economic Analysis Division implies, the Mix-adjustment method when using Land Registry data is unreliable. The Land Registry 'price paid dataset does not record bedroom numbers, square footage, number of bathrooms and other key attributes. Both the Mix-adjustment and Hedonic regression approaches require a large number of dwelling characteristics to be recorded if they are to be reliable (Thwaites & Wood, 2003). The volatility of the Mix-adjustment method is empirically demonstrated in the later sections of this paper.

Overall, Mix-adjustment has been largely ignored by the academic literature that compares the various quality adjustment methods. Academics in the field have chosen to concentrate instead on comparing the two more advanced approaches to quality adjustment the Hedonic and RSR methods. The clear limitation with Mix-adjustment is the specification of the characteristics defining the cells. In practice, the cells cannot be defined sufficiently to remove sources of statistical bias. Compositional changes will invariably lead to volatility in the final measurement. For example the common specifications for property type (e.g. Detached, Semi-Detached, Terrace or Flat) can lead to a situation whereby a 20-room mansion is placed in the same category as a 3-room cottage.

An approach to address this limitation is to increase the number of characteristics used to describe a house price observation. The outcome of this, is a method that resembles the Hedonic model. The Hedonic method is essentially a more robust and advanced parametric form of Mix-adjustment. Indeed, Mix-adjustment and Hedonic regression can give very similar results if they adjust for the same property characteristics.

The Hedonic method is designed for datasets containing detailed property information. The Land Registry dataset does not lend itself to Hedonic analysis as it does not capture detailed property characteristic information. In cases where such detail is captured, for example in both the Halifax and Nationwide data, the Hedonic method can be employed. Even with such data there are still limitations to the method.

A significant difficulty the Hedonic method faces is the fact that it relies heavily on the correct specification of both the functional form of the model and the set of property characteristics (Meese and Wallace, 1997). Case and Quigley (1991) illustrated

the Hedonic model in a general form, $P_t = f(x, t)$, i.e. house price is a function of time t and the vector of all physical and locational characteristics x. This requires f to be correctly specified, and the vector x to be correctly chosen and accurately measured; none of these can be guaranteed. Incorrect specification of f and vector x introduces what is known as 'misspecification bias (Bailey, Muth and Nourse, 1963; Case and Shiller, 1987).

The most serious theoretical drawback of both Hedonic and Mix-adjusted methods concerns ignorance of the appropriate set of house attributes to include in the analysis. This can lead to inconsistent estimates of the implicit prices of the characteristics. Consistent estimates of implicit Hedonic prices will rely on the bold assumption that all omitted variables are orthogonal to those included in the analysis.

Omitted unobserved characteristics correlated with those included can severely bias the Hedonic estimates and create index inaccuracies. This is a particularly acute problem for goods like properties. For example, it is widely accepted that the location of a property can significantly affect its price. Even for properties located in the same building, qualitative characteristics like "view" or "aspect" can cause important difference in the price. Therefore, no matter how detailed the housing characteristics collected, no Hedonic housing equation could practically observe precise location. With as many different locations as different properties, Hedonic equations simply cannot capture all location data. The already very complex Hedonic equations deal with 10 UK regions (Halifax & DCLG equations). Conversely, mainly due to the fact that RSR utilises a larger sample size, while simultaneously observing the precise location, RSR is able to produce indices down to the region level.

In addition, if certain unobserved attributes were more common in houses sold at certain phases of the cycle (e.g. if higher quality properties transacted relatively more during booms) then the amplitude of the house price index fluctuations may be underestimated or overestimated. The crucial question for Hedonic and Mix-adjustment procedures, is whether the chosen characteristics used for adjustment are the main determinants of price differences. While some of these are easy to measure (e.g. number of rooms), other important factors (e.g. vista) are often difficult to capture.

The adoption of Hedonic methods requires a considerable data-collection effort as information is needed not only on product prices but also on their related characteristics. The most comprehensive single dataset containing sufficient characteristics for reliable analysis is the Halifaxs proprietary dataset. However capturing $\sim 12,000$ transactions per month, this dataset falls short of the $\sim 100,000$ transactions per month that is captured by Land Registry.

Whilst the Hedonic and Mix-adjustment methods cannot employ Land Registry data reliably (due to lack of property characteristic information), the Land Registry dataset can be reliably quality adjusted through the use of the RSR method. The lack of a sufficiently large dataset containing sufficient attribute information prevents the Hedonic method from providing a system of local house price indices for the United Kingdom. This is one of the drawbacks in the implementation of the Hedonic method and the reason why the RSR method should be examined as a

potential improvement method in the UK context.

4.2 Repeat sales regression

There are a number of technical issues that arise from implementation of the RSR method. It must be recognized that no index estimation method is perfect, and the RSR method, while we believe it to be extremely robust and value adding, is not by nature bias free. A description of various technical issues follows below.

4.2.1 Heteroskedasticity

Recall in the model of Bailey, Muth and Nourse (1963), the error terms u_{n,t_1,t_2} are by convention assumed to have constant variances. Case and Shiller (1987, 1989) argued that u_{n,t_1,t_2} has varying variances, and this is called heteroskedasticity in econometric literature. The solution Case and Shiller (1987, 1989) provided was a weighted repeat sales model. However, it has been argued that the effect of the model is ambiguous. Leishman and Watkins (2002) using Scottish data, applied both the normal RS method and weighted RS method and concluded that the normal RS method was preferred.

4.2.2 Sample Selection

An important feature of the RSR method is that the sample used in RS regression only includes houses that have been sold more than once, and therefore suffers from 'sample selection bias (Case, Pollakowski and Wachter, 1991 and 1997; Cho, 1996; Gatzlaff and Haurin, 1997; Meese and Wallace, 1997; Steele and Goy, 1997; Hwang and Quigley,2004). The empirical study of Clapp, Giacotto and Tirtiroglu (1991) found no systematic differences between the RS sample and the full sample of all transactions over the long run. They argued that arbitrage typically forces prices for the repeat sample to grow at the same rate as those for the full sample. Another study of Wallace and Meese (1997) also arrived at the conclusion that the sub-sample of RSR is actually representative of all home sales during the period under consideration.

Pryce and Mason (2006) found evidence that the proportion of the housing stock which trades in a given period varies non-randomly across space. This issue raises the important question about what the underlying target of measurement is. Much of the debate over index methodology can be distilled to largely unrecognised disagreement over the desired target or intended application (Wang and Zorn, 1999). The RSR index is naturally more reflective of properties that transact more frequently. In so far as a differential in price appreciation exists between properties based on the relative frequency of transactions, the RSR measure will be naturally weighted towards the more frequently transacting subset of properties.

There are a variety of reasons why the holding duration of properties might be unevenly distributed. The increase in transaction costs for more expensive properties due to stamp duty may result in a decreased turnover of more expensive homes. Life-cycle theories on property holding period posit that less expensive properties

are traded more frequently - when people move up the property ladder they tend to move home less often. In addition the Buy-to-Let market is more active in the lower price brackets. Policy-makers need to be aware of the price appreciation differentials between submarkets, especially when there is systematic variation in the frequency of transactions between these submarkets.

Hwang and Quigley (2004) conducted a study using comprehensive data from the Stockholm housing market during 1981-1999 to explore the effects of RSR sample selection bias and constant quality assumptions. They strongly concluded that these issues represented shortcomings in the RSR method and that their hybrid model yields significantly improved estimates. Unfortunately due to limitations in data availability, the Hwang and Quigley study and model cannot be replicated in the United Kingdom. The availability of housing characteristic data in Sweden, where even the type of roof tiles are recorded, is not comparable to characteristic data availability in the United Kingdom.

4.2.3 Inconstancy of attribute appreciation

If one takes a Hedonic perspective to consider a house as a bundle of separate attributes, both qualitative ones and quantitative ones, the setting of the RSR method implicitly assumes that the prices of all these attributes move at the same rate over time, which may not be the case (Case, Pollakowski and Watcher, 1991).

4.2.4 Multicollinearity

Multicollinearity refers to situations where there is an approximate linear relationship among independent variables (Kennedy, 2003). This is not a rare phenomenon in econometrics. Although the Gauss-Markov Theorem still ensures a best linear unbiased estimator, some problems can be caused in applied research.

- Small changes in the data produce wide swings in the parameter estimates.
- Coefficients may have very high standard errors and low significance levels even though they are jointly significant and the R^2 for the regression is quite high.
- Coefficients may have the wrong sign or implausible magnitudes. Unlike other model specification problems, the problem of multicollinearity is caused by the specific sample used in the regression (Kennedy, 2003).

Cho (1996) pointed out that this problem tends to arise with small sample sizes. With only a small percentage of transactions, two columns of the data matrix are by construction similar and hence highly correlated. However, this problem does not present itself materially in the UK housing market. The housing market liquidity in England & Wales is greater than most other countries. With $\sim 100,000$ residential property transactions per month and close to a 70% rate of home-ownership, the national sample data does not suffer from material multicollinearity.

4.2.5 House Improvement Adjustment

RSR requires that the property has undergone neither a significant enhancement in value, such as remodelling, nor substantial physical deterioration (Araham and Schauman, 1991), so that the single property price appreciation can be attributed solely to trends of market price movement. It is clear that all properties experience physical depreciation and in addition many properties are improved prior to sale. There are two contrasting approaches in the academic literature. The first approach advocates not adjusting for this issue. It holds that in the long-run the value of house improvements will equate to the value of depreciation such that this factor will hold constant. If it is viewed that the main component of value is space, i.e. square footage then the argument regarding the irrelevance of depreciation or improvements gains strength.

The alternative approach is to make an adjustment to the index to reflect the average value of improvements minus depreciation. Araham and Schauman (1991) believe that it is possible to correct the index directly from knowledge of the value of improvements nationwide. A perfect RSR model would require the absence of systematic property deterioration or improvement across the sample. It is worth noting that this same bias affects Hedonic model variants in so far as the home improvement or deterioration is not perfectly captured by the Hedonic variables. It is the view of the author that in so far as data on systematic property deterioration or improvement is available, this information can and should be practically incorporated in the model. One valuable source of such information is the English House Condition Survey (EHCS) annually conducted by the Department of Communities and Local Government.

4.2.6 Inefficiency

One critisism about the RSR method is that it only uses a portion of the transaction dataset (i.e. it only uses matched pairs and ignores other transactions) therefore suffering from inefficiency. Such comments often fail to note that the explanatory and informative power of the remaining portion of data (i.e. the matched pairs) is likely to be superior to a similar sized dataset used by Hedonic methods.

To fully understand the difference between the number of transactions and amount of information utilised it is important to understand how the transaction information is being used by the different methodologies. The characteristic based methods do not utilise exact address information - the only information on each transaction that is used are the characteristics analysed. RSR on the other hand utilises exact address information. Contained within exact address are *all* the characteristics that make an individual property unique. The information efficiency advantage of RSR stems from the fact that there is less chance of omitted variables, i.e. the method uses the information that is contained within the exact address.

Additionally, it is important to point out that this already valuable dataset is improving over time at an increasing rate. As the time-span of data collection increases, an increasing proportion of transactions within the dataset will be able to be matched

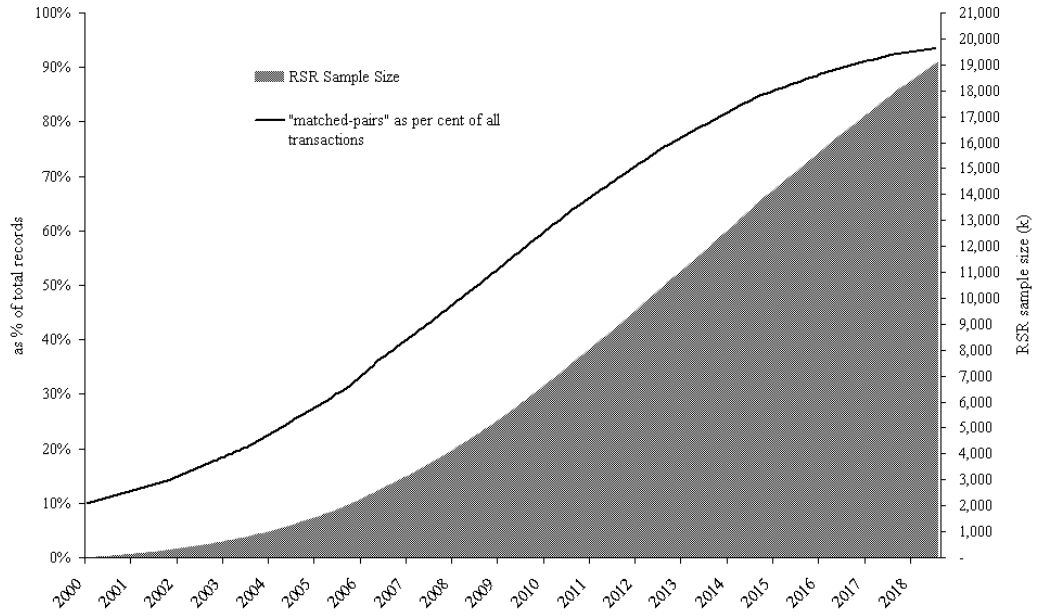

Figure 1: The number of transactions, which have identifiable "matched pairs" increases with time.

with other transactions, leaving fewer and fewer transactions left unmatched. Figure 1 shows these projections and their implications for the number of price observations available to RSR.

5 Comparison of indices produced using the Land Registry dataset

While there is little to separate RSR from the Hedonic method in theory, in practice the RSR method is preferable in the creation of UK house price indices. This is primarily because the less onerous data requirements enable RSR to utilise the significantly larger Land Registry dataset. The Land Registrys price paid dataset contains details on every residential home purchase in England & Wales.

As mentioned above, while Hedonic method is a valid technique, its heavy data requirements limit the datasets that can be used. To employ either the mix-adjusted or Hedonic approach with confidence, a detailed data sample needs to be collected. These data requirements are far stricter than the repeat sales methodology. Data on many of the attributes that are important determinants of the price of a property, particularly qualitative attributes are not collected nationally by any institution in the UK.

Although large databases containing information on both quantitative and qualitative attributes do exist (e.g. Nationwide, Halifax, and DCLG), the databases are still

significantly smaller than that held by Land Registry. It is clearly a major advantage to have the largest database, as for a given level of aggregation, more data means tighter standard errors about the estimated mean of the price series (Araham and Schauman, 1991). Given a much larger data set, RSR yields less volatile results on a national and local scale.

This paper is not the first report to note that the Land Registry data is best suited to the RSR method. In a 2002 paper, Leishman, Watkins and Fraser claimed that the data compiled by the Land Registry lends itself better to repeat sales methods rather than the Hedonic approach. In fact, the limitations in the type of data available in the Land Registry dataset effectively prevent its use by any Hedonic method.

A limitation of the Halifax, Nationwide and DCLG reports has been the lack of indices for geographical areas finer than broad government office regions; this is because necessary data do not exist in sufficient detail and volume for reliable local indices to be created. The size of the Land Registry dataset gives RSR the ability to create a system of indices for not only finer geographic areas but for other segmentation such as price and property type. Work carried out by the author, shows that reliable indices can be created to the postcode area level. These have been found to have an equivalent level of volatility as a national level mix-adjusted index using the same Land Registry dataset. With the ability to identify matched pairs from within the Land Registry dataset, over time these indices will continue to improve.

As frequently cited, one of the main problems with the existing indices is their lack of geographic focus. Munro & Maclennan (1986) point to the need to examine house price appreciation rates at neighbourhood level and caution against making assumptions about the aggregate nature and behaviour of markets. Costello and Watkins (2002), and Leishman Watkins and Fraser (2002) all argue strongly for the construction of a system of local house price indices for British cities using RSR, citing numerous decision-making benefits.

The author has carried out an empirical estimation of indices using the methodologies described above, utilising Land Registry data. In the following sections we present the key findings of this analysis - findings that support the arguments presented throughout this paper. Given the superior sample size opportunity presented by Land Registry data, the methodology that can best utilise this data could be seen as the logical choice for house price measurement in the UK.

The relatively recent introduction of RSR was due to the historically limited availability of data - the restoration of 'price paid' to the Land Register for England and Wales only occurred in April 2000. The subsequent public release of this data occurred in January 2005. Leishman (2000), Leishman and Watkins (2002), and Leishman, Watkins and Fraser (2002) were the first to conduct research in applying RSR in the UK; however due to data constraints, their work was confined to Scotland. Calnea Analytics produces monthly RSR, Mix-adjustment and simple average indices based on Land Registry data. These indices, combined with index data produced by Halifax, DCLG and Nationwide are compared side-by-side in the discussion that follows. Empirical results display the dominance of the RSR method in terms of low index volatility.

5.1 National indices compared

It is plausible to argue that relative Index reliability and accuracy can be inferred from the relative volatility or 'statistical noise' in the index results. Statistical noise caused by insufficiencies in either the data sample or the quality adjustment procedures tends to produce index volatility. Index volatility can be displayed visually and measured numerically.

We should note that volatility may be due to genuine changes of the house market rather than statistical noise. However, econometric literature suggests that high volatility on such an aggregated level is unlikely to exist; therefore an observer could infer that the more reliable indices are those which display less erratic movements, i.e. fewer monthly changes that differ from the underlying trend.

Numerical measurement of volatility can be performed. Common methods include measuring the standard deviation of monthly price changes. The stability of the RSR method using Land Registry data is consistently shown in all the empirical results in this paper. Relative comparisons of RSR with all the alternative methods on national, regional and segmental levels show RSR as the least volatile measure of price change. For the purpose of consistent and unbiased comparisons, all indices in this report have been observed prior to any application of seasonal adjustment or smoothing.

	RSR	Halifax	Nationwide	DCLG	Mix adjustment	Simple average
Total increase	109.4%	117.3%	109.7%	nc	92.5 %	93.3%
Standard deviation	0.7%	1.5%	1.1%	nc	1.7%	2.0%

Table 2: Performance of the indices over the last 79 months.

	RSR	Halifax	Nationwide	DCLG	Mix adjustment	Simple average
Total increase	10.2%	15.4%	10.5%	9.4%	9%	11.1%
Standard deviation	0.4%	1.1%	0.7%	0.9%	1.3%	1.8%

Table 3: Performance of the indices over the last 24 months.

A less volatile index will tend to have lower Standard deviation of growth rate. The results in Tables 2 and 3 and also in Figures 2 and 4, rank (in terms of index volatility) the RSR method ahead of the Hedonic methods which in turn just outperform Mix-adjustment. As expected, the worst performing price index according to Standard deviation is the Simple average index. The Simple average index is a series based on

monthly arithmetic means, without using any quality adjustment procedure.

We should stress that the above measure of volatility is rather subjective since the underlying growth rate of the market is not actually known. However, intuitive knowledge of the market's movements suggests that it is unlikely that the market shows erratic swings in monthly growth; therefore one might conclude that such growth rate swings should be attributed to imprecise estimation.

5.2 RSR vs Hedonic

The key feature of both these methods is that they offer a robust approach to quality adjustment. The differences in empirical results are almost wholly determined by the quality of the underlying dataset.

The importance of the data sample is illustrated by direct comparisons between the Halifax and Nationwide indices. These indices often give different monthly results. In an effort to uncover the causes of the discrepancy both indices were adjusted for a variety of possible effects. For example, the supposedly northern-biased Halifax was adjusted for its regional mix. None of these adjustments had a material effect. Independent analysis by the Royal Statistical Association found no methodological reason why the price indices varied. Martin Ellis of Nationwide said, "The only conclusion we can come to is that the raw data, the samples which are used to calculate the indices, make the figures diverge. There's no other factor which can be responsible for the variations".

The Halifax and Nationwide use property datasets that are significantly smaller than that held by the Land Registry. For example, the Halifax claims their database covers 12,000 mortgage approvals per month, the Land Registry records approximately 100,000 home purchase transactions per month. Of the 7.5 million price observations recorded by the Land Registry since April 2000, around 3 million are multiple transactions of the same property. Already this number is roughly twice the number of price observations available to Hedonic methods. Therefore, not only is there more information, but the type of information RSR processes holds more informative power.

The conclusions are corroborated in the graphs shown. Figure 1 illustrates the high historical correlation between indices in the long-run, however limited the short-term corroboration, while Figure 2 verifies that the volatility of RSR is lower than the Hedonic methods, as the Standard deviations of growth rates in Table 2 suggest.

5.3 RSR vs. Mix-adjustment

Mix-adjustment can be performed on the Land Registry dataset. However, the lack of characteristic data means that the Mix-adjustment method offers minimal reductions in index volatility. Compared to Hedonic methods, Mix-adjustment is more volatile. When comparing Mix-adjustment to RSR, which can utilise the same dataset, RSR offers much lower volatility. It would seem that the advantages RSR has in terms of quality adjustment vastly outweigh the Mix-adjustment methods ability to include more transactions in its calculation of averages.

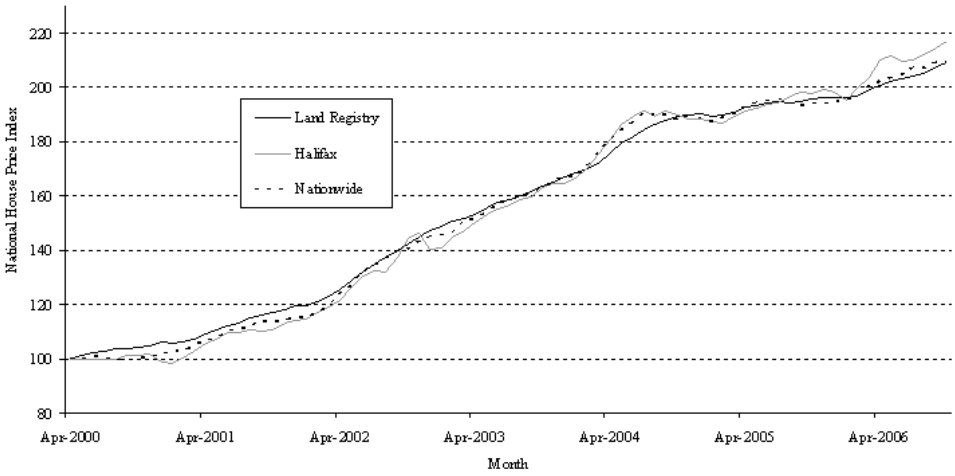

Figure 1: Monthly House Price Index

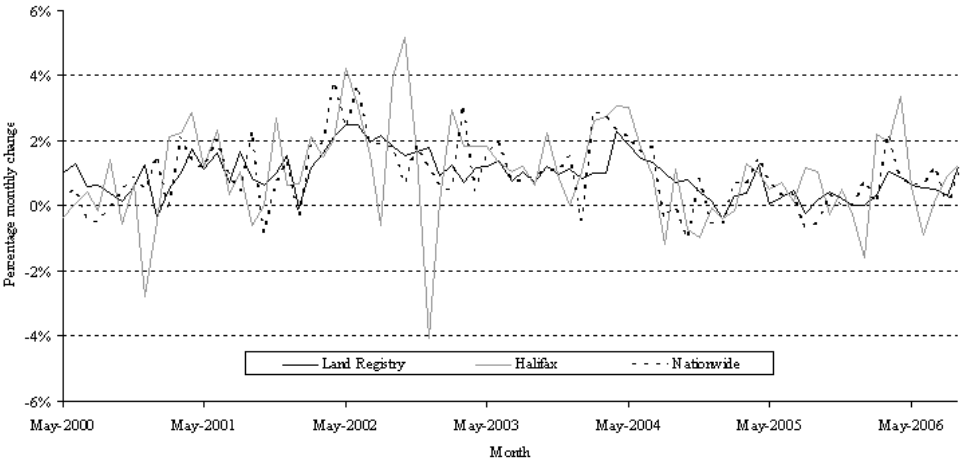

Figure 2: Monthly Price Inflation

Figure 3 verifies once more the pattern of long term correlation between indices, but limited short term corroboration. The Simple Average and Mix-adjustment methods are inherently biased towards the growth patterns of more expensive properties. Over the last four years, the boom in the buy-to-let and first-time buyer markets have caused less expensive properties to appreciate at a faster rate than more expensive properties. The lower growth rates of the Simple Average and Mix-adjustment methods in the graph above are caused by this phenomenon. This understatement is due to the fact that these indices are based on weighted averages of prices of 'cells' rather than weighted averages of growth of 'cells'. If the Mix-adjustment method

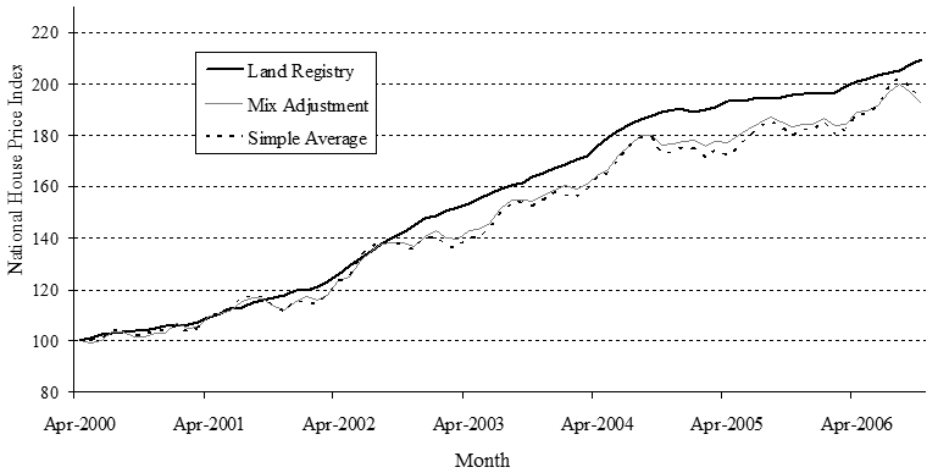

Figure 3: Monthly house Price Index

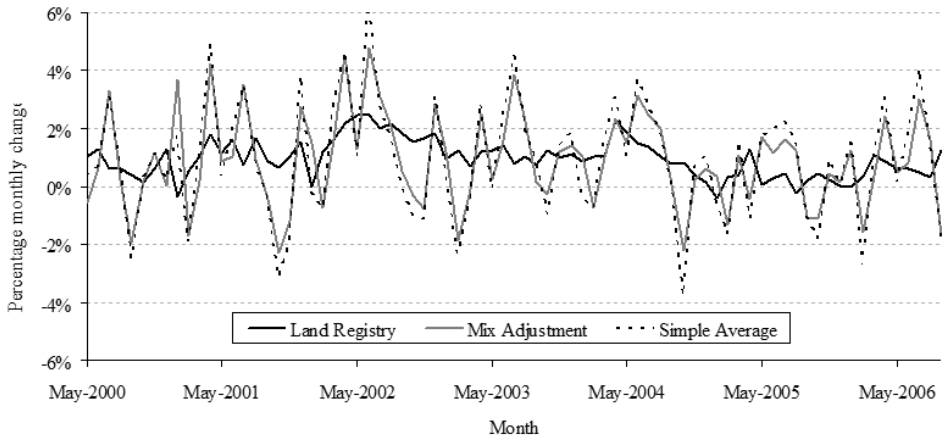

Figure 4: Monthly Price Inflation

is altered such that the weighted averages are applied to growth rates rather than prices, this bias is corrected.

6 Concluding remarks

Methodologically-consistent local and national housing market price indices created using the RSR method will undoubtedly be of use to a wide range of policy makers, investors and other stake-holders.

The purpose of this research has been to build upon the valuable work of Leishman, Watkins & Fraser (2002) amongst many others, in order to investigate the advantages brought by the RSR method in relation to its practical application using Land

Registry data.

The advantages and disadvantages of RSR have been well documented. Summarising we should stress that the greatest strength of RSR is its method of separating quality from price (Araham and Schauman, 1991). As mentioned earlier in this paper, the difficulty the Hedonic and Mix-adjustment methods face is the fact that they rely heavily on the correct specification of both the functional form of the model and the set of property characteristics (Meese and Wallace, 1997). Case and Quigley (1991) illustrated the Hedonic model in a general form, $P_t = f(x, t)$, i.e., house price is a function of time and the vector of all physical and location characteristics x. This requires f to be correctly specified, and the vector x to be correctly chosen and accurately measured, neither of these can be guaranteed. Errors here introduce what is known as misspecification bias (Bailey, Muth and Nourse, 1963; Case and Shiller, 1987).

For the RSR method, researchers control for Hedonic characteristics by examining only those properties that have been sold more than once during the period under consideration. Case and Quigley (1991) provided a general expression of the RSR model as: $P_{t_2}/P_{t_1} = g(t_1, t_2)$, which obviously highlights the feature of RSR that it only depends on price data and transaction dates, both of which can be measured accurately. The functional form g is also unique to the property, which is a clear advantage over Hedonic regression and Mix-adjustment.

Another clear strength of the method is the less strict data requirement. RSR requires only data on transaction prices and dates of two consecutive transactions, and does not require data on physical attributes. Thus, unlike other methodologies, RSR can perform extensive quality adjustment using the attribute - lacking Land Registry dataset. There are currently 7.5 million price observations recorded by Land Registry, at present approximately 38% of these (around 3 million) are multiple transactions of the same property. It is misleading to make direct comparisons of sample size due to the additional informative power that RSR extracts from matched pairs. However, any direct comparison will still show that there are already far more price observations available to RSR from the Land Registry price paid dataset than to alternative datasets.

The empirical research conducted by the author and presented in this paper suggests that not only can the Land Registry price paid dataset be usefully deployed in the construction of national, local and segmental price indices but that these RSR based indices display a significant stability advantage over other methodologies.

It is widely accepted that there is no perfect house price index. However, the RSR method can be seen as a potential solution to the quality adjustment problems faced by index creation using the Land Registry dataset (Thwaites & Woods, 2003; Leishman, Watkins & Fraser, 2002; Costello & Watkins, 2002). The RSR index provides a true constant quality house price series for existing houses that can be calculated with reliability down to the postcode area level. It can provide hitherto unavailable insights to aid the decision-making ability of policy makers and investors operating in specific housing markets. Furthermore as the time-span of the Land Registry dataset increases, the reliability of the RSR index will increase. At the time of writing there

remains the possibility that Land Registry will make available price paid data back to 1995. Alonger time span will improve index reliability and could be incorporated as a simple recalibration of the index.

Acknowledgments

We are grateful for the feedback from both the anonymous referees.

References

[1] Araham, J.M, David J., Schauman, W.S. (1991) "New evidence on house pries from Freddie Mac repeat sales", *AREUEA Journal* **19**(3), 333-352.

[2] Bailey, M. Muth,R and Nurse, H.(1963), "A regression method for real estate price index construction", *Journal of the American Statistical Association* **58**(304), 933-942.

[3] Baumohl, B. (2005), "The Secrets of Economic Indicators", *Pearson Education.*

[4] Case, B., Pollakowski, H. and Wachter, S. (1991), "On choosing among house price index methodologies", *AREUEA Journal*, **19**(3), 286-307.

[5] Case, B., Pollakowski, H. and Wachter, S. (1997), "Frequency of Transaction and House Price Modeling", *The Journal of Real Estate Finance and Economics*, **14**(1-2), 173-187.

[6] Case, B. and Quigley, J.(1991), "The dynamics of real estate prices", *Review of Economics and Statistics*, **73**(1), 50-58.

[7] Case, K. and Shiller, R. (1987), "Prices of single-family homes since 1970: new indexes for four cities", *New England Economic Review*, 45-56.

[8] Cho, M. (1996), "House price dynamics: a survey of theoretical and empirical issues", *Journal of Housing Research*, **7**(2), 145-172.

[9] Clapp, J. and Giaccotto, C. (1992), "Estimating price indices for residential property: a comparison of repeat sales and assessed value methods", *Journal of the American Statistical Association*, **87**, 300-306.

[10] Clapp, J. and Giaccotto, C. (1999) "Revisions in repeat sales price indices: here today, gone tomorrow?", *Real Estate Economics*, **27**(1), 79-104.

[11] Clapp, J., Giacotto, C. and Tirtiroglu , D. (1991), "Housing price indices based on all transactions compared to repeat subsamples", *AREUEA Journal*, **19**(3), 270-285.

[12] Costello, G. and Watkins, C. (2002), "Towards a system of local house price indices", *Housing Studies*, **17**(6), 857-873.

[13] Crone, T.M. and Voith, R.P. (1992), "Estimating House Price Appreciation: A Comparison of Methods", *Journal of Housing Economics*, **2**(4), 324-338.

[14] Fleming, M.C. and Nellis, J.G. (1994), "The measurement of UK house prices: a review and appraisal of the principal sources", *Housing Finance*, **24**.

[15] Gatzlaff, D.H. and Ling, D. (1994), "Measuring Changes in Local House Prices: An empirical Investigation of Alternative Methodologies", *Journal of Urban Economics*, **35**(2), 221-244.

[16] Gatzlaff, D.H. and Haurin, D.R. (1997), "Sample selection bias and repeat-sales index estimates", *Journal of Real Estate Finance and Economics*, **14**, 33-50.

[17] Goetzmann, W. (1992), "The accuracy of real estate indices: repeat sales estimators", *Journal of Real Estate Finance and Economics*, **5**, 5-53.

[18] Goetzmann, W. and Peng, L. (2002), "The bias of RSR estimator and the accuracy of some alternatives", *Real Estate Economics*, **30**(1), 13-39.

[19] Greene, W.H. (2003), "Econometric Analysis (5th Edition)", *Blackwell Publishing*, 56-59.

[20] Hosios, A.J. and Pesando, E.J. (1991), "Measuring Prices in Resale Housing Markets in Canada: Evidence and Implications", *Journal of Housing Economics*, **1**(4), 303-317.

[21] Hwang, M. and Quigley, J.M. (2004), "Selectivity, Quality Adjustment and Mean Reversion in the Measurement of House Values", *Kluwer Academic Publishers*, **28**(2/3), 161-178.

[22] Kennedy, P. (2003), "A Guide to Econometrics (5th Edition)", *Blackwell Publishing*, 205-206.

[23] Lancaster, K.J. (1966), "A new approach to consumer theory", *The Journal of Political Economy*, **74**(3), 132-157.

[24] Leishman, C. (2000), "Estimating local housing market price indices using Land Registry data", *Housing Finance*, **47**(August), 55-60.

[25] Leishman, C., Watkins, C. (2002), "Estimating local repeat sales house price indices for British cities", *JPIF*, **20**(1), 36-58.

[26] Leishman, C., Watkins, C. and Fraser, W.D. (2002), "The estimation of house price indices based on repeat sales regression using Land Registry data", *Report to RICS Education Trust*, London.

[27] Meese, R. and Wallace, N. (1997), "The construction of residential housing price indices: a comparison of repeat sales, Hedonic regression, and hybrid approaches", *Journal of Real Estate Finance and Economics*, **14**(1/2), 51-74.

[28] Munro, M. and Maclennan, D. (1986), "Intra-urban changes in housing prices: Glasgow 1972-1983", *Housing Studies*, **2**, 65-81.

[29] Nicol, C. (1996), "Interpretation and comparability of house price series", *Environment and planning A*, **28**, 119-133.

[30] Pryce, G. and Mason, P. (2006), "Which house price? : finding the right measure of house price inflation for housing policy : technical report", *Office of the Deputy Prime Minister*, April.

[31] Rosen, S. (1974), "Hedonic Price and Implicit Markets: Product Differentiation in Pure Competition", *Journal of Political Economy*, **82**, 34-55.

[32] Steele, M. and Goy, R. (1997), "Short holds, the distribution of first and second sales, and bias in the repeat-sales price index", *Journal of Real Estate Finance and Economics*, **14**, 133-154.

[33] Shiller, R.J. (1994), "Macro Markets: Creating Institutions for Managing Societys Largest Economic Risks (Chapter 8)", *Oxford University Press.*

[34] Thwaites, G. and Wood, R. (2003), "The Measurement of House Prices, *Bank of England Quarterly Bulletin*, 38-45.

[35] Wang, T. and Zorn, P.M. (1997), "Estimating house price growth with repeat sales data: Whats the aim of the game?", *Journal of Housing Economics*, **6**, 93-118.