Journal of the Operational Research Society (2009) 60, 339–347

www.palgrave-journals.com/jors/

Predicting a house's selling price through inflating its previous selling price

A Brint*

The University of Sheffield, Sheffield, UK

This paper considers how accurately inflating the previous selling price of a modern property predicts its selling price. Predicting a house's value is an important capability as it allows how the asking price affects the time to sale and the price achieved, to be modelled. The analysis is carried out on a data set of 105 pairs of earlier and later selling prices for UK properties constructed since January 1999. As an alternative to using published house price indices for inflating the prices, a novel approach for modifying the published house price indices through the use of observed repeat sales of properties is put forward and analysed. Using the best published index gives an average predictive error of 10.9% while using the published index modified by repeat-sales information, gives an average predictive error of 8.4%.

Journal of the Operational Research Society (2009) **60,** 339–347. doi:10.1057/palgrave.jors.2602567
Published online 6 February 2008

Keywords: econometrics; forecasting; markov processes; statistics; urban studies

Introduction

The main issue investigated by this paper is:

> How accurate a guide to a house's selling price is its previous selling price inflated by the house price index for the period between the sales?

This is an important question, as inflating the previous selling price provides a zero cost method of immediately estimating the value of a house. Being able to estimate a house's value is vital when analysing the decision problem of choosing an asking price for a house and how this asking price affects the price achieved and the time to sale. Additionally, if the estimates of the later selling price are sufficiently good, then it could be of interest to non-professionals involved in house sales, that is, the vendor and prospective purchasers, as a way of reducing the information asymmetry that they face.

This paper focuses on the housing market in England and Wales where a relatively recent change (April 2000) has led to the selling price of all transactions since then being publicly available. Besides assessing the accuracy of inflating the previous selling price, this paper also considers using local house sales information to try to improve the house price index that is used to inflate the selling prices.

There are several factors that led to variations in house prices between sales besides general house price inflation, for example, changes to a house such as an extension, the increasing age of the house, and differing vendor circumstances such as pressure for a quick sale. This paper reduces some of this variability by restricting attention to recently constructed houses, that is, those whose first sale was in 1999 or later, and for which there are no discernible material changes to the house as described in the two sets of advertising material.

*Correspondence: A Brint, The Management School, Sheffield University, 9 Mappin Street, Sheffield S1 4DT, UK.
E-mail: A.Brint@sheffield.ac.uk

Overview of house price indices

Considerable attention has been given to the problem of producing house price inflation indices. The main approaches have centred on summary methods, hedonic regression, repeated sales and hybrid schemes.

Summary methods

The simplest approach is to form a central, summary measure of the houses sold in each time period, for example the median selling price, and then to use these to derive a house price index (Gatzlaff and Ling, 1994; Wang and Zorn, 1997). As the characteristics of the houses sold in a time period will fluctuate randomly, for example one time period may have seen the sale of more larger houses, very large data sets are needed to reduce this effect. Despite this drawback, the straightforward nature of the approach coupled with the lack of information needed about each sale, means that it has often been used in practice.

Hedonic regression

A natural progression from the summary methods is to try to account for the different types of property that are being sold. In hedonic regression, the property's attributes such as its location and physical size are recorded along with its selling time and price. A regression model of the form (see, eg Englund

et al, 1999)

$$\log(P_{it}) = \sum_j (\beta_j A_{ijt}) + \sum_q (\lambda_q D_{iq}) + v_{it} \qquad (1)$$

is then fitted to the data. Here P_{it} is the price house i sold for at time t; i identifies the house; t the time of sale; A_{ijt} the value of the j^{th} attribute of house i at time t, for example, the size of the living area; β_j the weight assigned to attribute j by the regression model; D_{iq} is 1 if house i's sale was in time interval q, and 0 otherwise. The time intervals used in this paper are quarter years, hence the choice of q; λ_q is assigned by the regression model—the inflation index is derived from it; and v_{it} is an error term, typically taken to follow an autoregressive process.

Hence besides estimating the inflation index, the model also estimates the effect on a house's price of its individual attributes, and so it allows individual properties to be priced (Dodgson and Topham, 1990).

Repeat sales

Weaknesses of the hedonic regression approach are that neither the functional form of the relationship between the attributes nor what attributes should be used in the model are known with certainty (Quigley, 1995). Additionally a relationship or attributes that are suitable for one market may not be appropriate for another (Iversen, 2001). The repeat-sales approach reduces these weaknesses by only considering houses that have sold at least twice, then subtracting Equation (1) for the earlier sale from Equation (1) for the later sale gives

$$\log(P_{iT_{2i}}) - \log(P_{iT_{1i}}) = \sum_j (\beta_j (A_{ijT_{2i}} - A_{ijT_{1i}})) + \sum_q (\lambda_q D'_{iq}) + v_{iT_{2i}} - v_{iT_{1i}}$$

where T_{1i} is the time of the first sale of house i, T_{2i} the time of the second sale of house i, D'_{iq} is 1 if house i's second sale was in time period q, -1 if house i's first sale was in time period q and 0 otherwise.

If neither the attributes nor their weightings change over time, then this reduces to (Hill *et al*, 1999)

$$\log(P_{iT_{2i}}) - \log(P_{iT_{1i}}) = \sum_q (\lambda_q D'_{iq}) + e_i \qquad (2)$$

where e_i is the error or variation term for house i. Hill *et al* (1997) argue that this term should be modelled as a normal variable with mean 0 and variance $2\sigma_i^2(1 - \rho^{T_{2i}-T_{1i}})/1 - \rho^2$ where ρ and σ_i are parameters with the assumption that $|\rho| < 1$. This error formulation follows on from assuming that the v_{it} error in Equation (1) follows a first-order autoregressive process. However, it should be noted that other forms have been put forward, see for example Goetzmann and Spiegel (1995), and Dreiman and Pennington-Cross (2004).

The assumption that the attributes do not change over time ignores the fact that one of the attributes is usually the property's age. This attribute is commonly separated out from the other attributes in Equations (1) and (2), but as Bailey *et al* (1963) discuss in the original description of the repeat-sales method, the effect of depreciation with age cannot be separated out from house price inflation in the pure repeat-sales method. (Cannaday *et al* (2005) propose a modification to the method that controls for age depreciation.) Hence we have dropped the depreciation term from Equation (2).

Wang and Zorn (1997) considered what the target of the standard repeat-sales approach is. Perhaps it is not surprising given the form of Equation (2) that they found that it averages the logarithms of the individual property indices, rather than being the logarithm of the average of the individual property indices.

Shortage of data has meant that there has been little analysis of house price indices in the UK using the repeat-sales method. Leishman and Watkins (2002) used data for four Scottish cities that had been extracted from paper and electronic records, to compare different formulations of the repeat-sales approach. However, data were not available at that time for England and Wales.

The relative merits of the different approaches and hybrid schemes

Each of the two main ways of estimating house price inflation have an obvious weakness—hedonic regression is not invariate to the choice of hedonic indices, and the repeat-sales approach is wasteful of data as properties that sold only once are ignored (Goetzmann and Spiegel, 1997). Additionally, the repeat-sales method suffers from the assumptions that all homes appreciate at the same rate, that house quality is constant and concerns that the method can be biased as it is unlikely that often transacted properties are an unbiased sample of all properties (see eg Gatzlaff and Haurin, 1997). Therefore a number of hybrid schemes have been put forward. Quigley (1995) uses the properties sold more than once to estimate the variance–covariance matrix for his model of the error (disturbance) term v_{it} in Equation (1). The entire data set is then used to estimate the coefficients of Equation (1) by generalized least squares. A similar idea is put forward by Hill *et al* (1997), but here the use of repeat sales to estimate the error model is carried out using maximum likelihood at the same time as the hedonic regression equation is fitted.

Local house price inflation indices

Clearly one approach to estimating local house price indices is to restrict the data set to the locality, that is, to ignore sales from outside this area, but the resulting reduction in the data set size has led to other approaches being developed. Goetzmann and Spiegel (1997) introduce the distance-weighted repeat-sales method that regards the inflation elements in Equation (2) as being dependent on location. The

inflation rate is split up into an average rate and a location-specific rate. If $d()$ is the distance measure function, then the covariance of the log of the location-specific inflation rates for regions l and m is taken to be $\sigma_r^2 e^{-d(l,m)}$, where σ_r^2 is the variance of the log of the location-specific inflation rates. This approach can be used to estimate the inflation rates for different types of house in the same locality by defining $d()$ to be a similarity measure between the houses.

Iversen (2001) puts forward a Markov Random Field approach. (A Markov Random Field is a two-dimensional Markov chain, that is, no extra information about the inflation at a specified location and time is given by knowing the inflation at points further away if the inflation rate is known at the spatial and temporal neighbours (Ishikawa, 2003).) The inflation rate R_{jt} of the logged prices for area j at time t is related to the inflation rate in a neighbouring area k by

$$(R_{kt} - R_{jt}) \sim N(0, \theta^2) \tag{3}$$

and to the logged inflation rate in the preceding time period by

$$(R_{jt} - R_{j(t-1)}) \sim N(0, \tau^2) \tag{4}$$

where θ and τ are constants.

By considering all the links between neighbours in space and time, Equations (3) and (4) lead to a multivariate normal density that is proportional to

$$\exp\left(-\frac{1}{2}\left[\frac{1}{\theta^2}\sum_{\text{spatial neighbours } k}(R_{jt} - R_{kt})^2 + \frac{1}{\tau^2}(R_{jt} - R_{j(t-1)})^2\right]\right) \tag{5}$$

Equation (5) forms the prior distribution for the inflation indices R_{jt} for the logged prices. The likelihood comes from modelling the difference between the observed changes in the log of the sale price and the inflation indices for the logged prices as a normal distribution. A Markov Chain Monte Carlo approach is then used to estimate the posterior distribution.

Data

Available inflation indices

The main house price inflation indices for England and Wales are published by Halifax plc/HBOS, the Nationwide Building Society and the Land Registry/Communities and Local Government (CLG). Each of the three institutions publish a number of indices to cover different regions, property and purchaser types. For example, the CLG data include among other things, indices for first time buyers in the North West of England, and bungalows in the East Midlands area of England. All three sources of house price inflation indices primarily use hedonic regression methods.

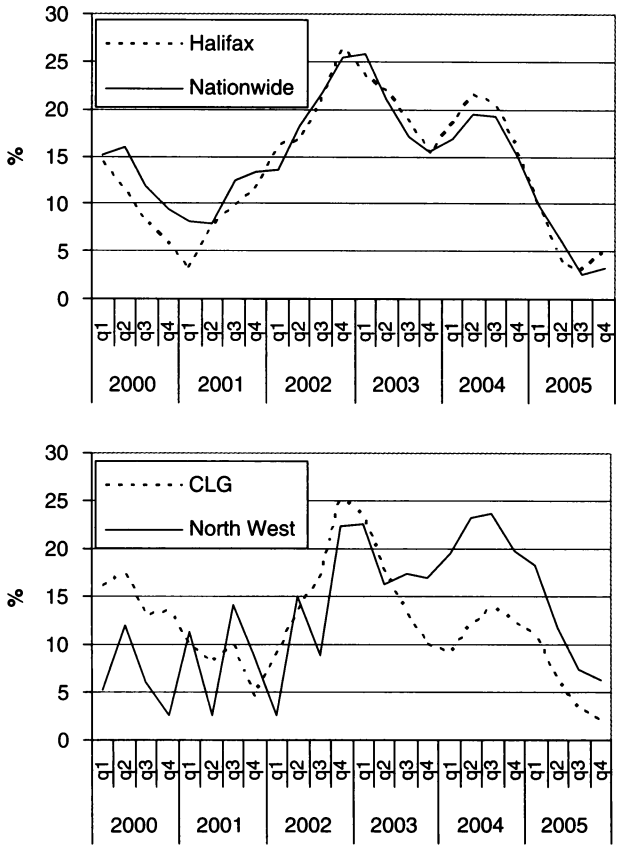

Figure 1 The published house price indices.

The published indices used in this paper are the England and Wales inflation rates from the Halifax, Nationwide and CLG as gathered together in CLG's Table 521, along with the CLG's inflation rates for the region under consideration (the North West of England) from CLG's Table 591. These four indices were selected as being the three principal national indices plus the most relevant CLG regional index. The indices for the period of the study are shown in Figure 1.

House sales data

Data were collected on houses for sale in two neighbouring towns (approximately 6 miles apart) in North West England. Houses that were for sale were identified from the estate agency advertisements in the local newspapers. These houses were then filtered so that only houses that met the following criteria were added to the data set:

- The property had to have been recently constructed—this was defined to have been occupied for the first time since the start of 1999.
- The property has at least two sale prices publicly available from the Land Registry at the end of the data collection period (March 2006), that is, it has been sold at least twice since the start of April 2000.

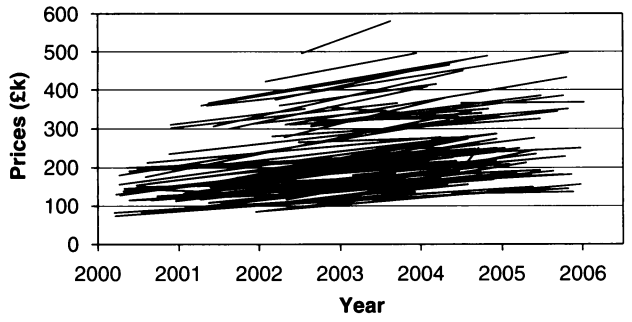

Figure 2 The first and second sale prices for the 105 property pairs.

- There were no significant changes to the house between its sales dates that could be discerned from the sales particulars, for example, houses where a garage had been converted into a study were ruled out.
- There were no major concerns about the accuracy of the data—just one property was ruled out on these grounds as it had been marketed at £360, 000 but its sale price was recorded as £500, 000.

The identification of potential data set elements through their advertisements in local newspapers had two purposes. Firstly, as a way of reducing any selection bias in the data set, that is, there was a clearly defined criterion for identifying potential data points. Secondly, it provided a first advertised date for use when modelling the relationship between the asking price, the achieved price and the time to sale.

Note that some houses sell without ever making it to a newspaper advertisement, and so the results of this paper do not apply to them, that is, we cannot say how accurate inflating the previous sale price is for these properties.

The reason for only considering relatively recently constructed properties was to try to reduce some of the problems that the repeat-sales method has with houses changing condition, for example being refurbished, between sales. As the houses are modern, it is assumed that any refurbishment will be largely cosmetic. Any resulting price effect will just add to variation between the predicted and actual sale prices.

Additionally as all the houses are approximately the same age, any inaccuracy caused by properties of different ages depreciating at different rates is reduced.

Data were collected for 105 earlier–later house sale pairs meeting the above criteria. These 105 pairs include three relating to one property, that is, it was sold four times during the period, and four resulting from two other properties, that is, each property was sold three times (the other 98 pairs came from different properties). Of the 105 pairs, 45 came from town A and 60 from town B. The first and second prices along with their selling (completion) times are shown in Figure 2. Additionally, the advertisements allowed the collection of hedonic variables such as whether the property was terraced, semi-detached or detached, the number of reception rooms, etc, but no use is made of this extra information in this paper.

Model and method

Price prediction model using published inflation indices

We follow the standard repeat-sales model and predict the second selling price to be

$$\hat{P}_{iT_{2i}} = P_{iT_{1i}} \times \prod_{q=T_S}^{T_E} r_{iq}^{\chi_{iq}} \tag{6}$$

where $\hat{P}_{iT_{2i}}$ is the predicted second selling price of house i; T_S and T_E are the starting and ending quarters of the data collection period, that is, all the first and second selling times lie within the period starting at the start of quarter T_S and ending at the end of quarter T_E; q identifies which quarter is under consideration; χ_{iq} takes the value 1 if house i was for sale for the whole of quarter q, 0 if it was for sale for none of the quarter, and otherwise its value is the fraction of time it was for sale during the quarter; r_{iq} is the appropriate inflation rate for quarter q, that is if i is in town A then it is a_q, the inflation rate for town A, while if i is in town B, then it is b_q.

Note that we have chosen to model growth rates in Equation (6) rather than the more usual price indices that appear in Equation (2). The two formulations are equivalent with the main difference being that χ_{iq} is used instead of D'_{iq} (Wang and Zorn, 1997).

Taking the natural logarithm of Equation (6) gives

$$\log(\hat{P}_{iT_{2i}}) = \log(P_{iT_{1i}}) + \sum_{q=T_S}^{T_E} (\chi_{iq} \times \log(r_{iq})) \tag{7}$$

In line with Hill *et al* (1999), we assume that the errors between the predicted second selling price, $\hat{P}_{iT_{2i}}$, and the observed selling price, $P_{iT_{2i}}$, are modelled as

$$\log(\hat{P}_{iT_{2i}}) - \log(P_{iT_{2i}}) = \varepsilon_i$$

that is,

$$\log(P_{iT_{1i}}) + \sum_{q=T_S}^{T_E} (\chi_{iq} \times \log(r_{iq})) - \log(P_{iT_{2i}}) = \varepsilon_i \tag{8}$$

where ε_i is normally distributed with mean zero and variance

$$\zeta_i^2 = \frac{2\sigma^2(1-\rho^{T_{2i}-T_{1i}})}{1-\rho^2}$$

Here σ and ρ are constants modelling the variability in the data, with ρ allowing the accuracy of the prediction to alter depending on the time between the sale dates.

Unfortunately we do not have enough data to determine reliable house price indices for towns A and B using Equation (7)—the study covers 24 quarters (from April 2000 to March 2006), hence there are 48 price indices to be estimated (plus

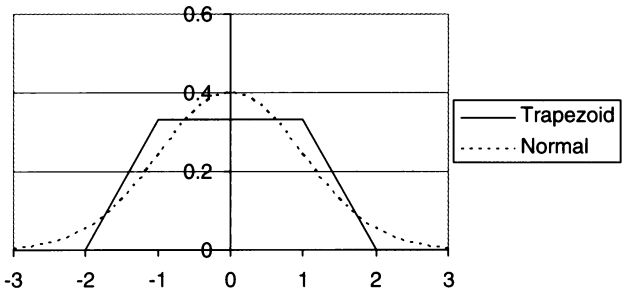

Figure 3 The shape of the trapezoidal prior considered as an alternative to the standard normal distribution.

σ and ρ) from the 105 data pairs. Therefore the method put forward in this paper bases the local house price indices on the published indices and then seeks to adjust the local indices so as to improve the accuracy of the prediction of the second sale price, $\hat{P}_{iT_{2i}}$. For example, if a nearby house on the same development that originally sold at the same time as the house in question, has just been sold again, then this may give a more accurate indication of the selling price of the house in question than using the published house price inflation rate.

Geltner and Goetzmann (2000) used ridge regression for periods when repeat-sales data are scarce, so as to reduce the problems of the regression being badly conditioned. However, we prefer to adopt a Bayesian approach to ill-posed problems (Kaipio and Somersalo, 2005).

The adopted approach is similar to that of Iversen (2001) in that it favours some values of the local price indices over others through specifying a similarity function that is based on each local price index and its spatial and temporal neighbours.

$$\prod_{q=T_S}^{T_E} \exp\left(-\frac{1}{2}\left(\frac{(a_q-h_q)^2}{\tau^2}+\frac{(b_q-h_q)^2}{\tau^2}+\frac{(a_q-b_q)^2}{\theta^2}+\frac{(a_q-a_{q-1})^2}{\psi^2}+\frac{(b_q-b_{q-1})^2}{\psi^2}\right)\right) \quad (9)$$

The formula we use is shown in Equation (9) where h_q is the published house price index for time q, τ is a constant that penalizes local house indices being different from the published house price indices for the same time period, θ is a constant that penalizes differences between the local house price indices between the neighbouring towns, ψ is a constant that penalizes temporal changes in the local house price indices.

Equation (9) differs from Iversen's formulation (Equation (5)) in that it includes terms involving the published house price index, h_q.

The similarity function defined by Equation (9) is essentially the product of several normal distributions. As an alternative, the normal distributions were replaced by a regular trapezoid (trapezium) as illustrated in Figure 3 for the case where the mean is 0 and the variance is 1. The consequences of using the trapezoid are that there is a limit to how far the variable can differ from its expected value, for example the variation of the local inflation rate from the published inflation rate is bounded, and the region around the expected value is just as likely as the expected value, for example if the published rate is 4.5% the local rate could just as likely be 4.3% as 4.5%. However, the results of using regular trapezoids were sufficiently similar to using normal distributions that only the normal distribution results are reported in this paper.

The selection of the local house price indices that best meet Equations (8) and (9) is done by forming the product of the probability of observing Equation (8)'s ε_i for each house. This product is then multiplied by Equation (9), and the local house price indices giving the highest final product are chosen. This corresponds to a maximum *a posteriori* estimate where the prior is proportional to Equation (9) and the likelihood to Equation (8) that is, for the posterior

$$\prod_i \left(\exp\left(-\frac{\varepsilon_i^2}{2\zeta_i^2}\right)\right) \times \prod_{q=T_S}^{T_E} \exp\left(-\frac{1}{2}\left(\frac{(a_q-h_q)^2}{\tau^2}+\frac{(b_q-h_q)^2}{\tau^2}+\frac{(a_q-b_q)^2}{\theta^2}+\frac{(a_q-a_{q-1})^2}{\psi^2}+\frac{(b_q-b_{q-1})^2}{\psi^2}\right)\right) \quad (10)$$

Determining the unknown parameters

Maximizing Equation (10) involves assigning values to the unknown parameters. These unknown parameters are divided into two groups, firstly the prior parameters $\{\sigma, \rho, \tau, \theta, \psi\}$ and secondly the unknown inflation rates a_q and b_q. If values have been assigned to all the unknown parameters, then the estimate of the second sale price, $\hat{P}_{iT_{2i}}$, of property i can be determined from Equation (6). How good this prediction is, can be assessed by comparing how far the predicted price $\hat{P}_{iT_{2i}}$ formed using the modified price indices a_q and b_q is from the observed price, $P_{iT_{2i}}$, against how far $\hat{P}_{iT_{2i}}$ formed using the published price indices h_q is from the observed price $P_{iT_{2i}}$. Letting i range over a number of values, for example from 22 to 105, then there are various ways to assess the set of pairs of $\hat{P}_{iT_{2i}}-P_{iT_{2i}}$ resulting from using $\{a_q, b_q\}$ and h_q. The one used in this paper is Wilcoxon's signed rank test as we wish to avoid making assumptions about the distribution of these prediction errors, and so we cannot use Student's t-test. Therefore a way of assigning values to $\{\sigma, \rho, \tau, \theta, \psi\}$ and $\{a_q, b_q\}$ is needed.

Consider the situation where values have been assigned to $\{\sigma, \rho, \tau, \theta, \psi\}$ but not to $\{a_q, b_q\}$, and that we want to estimate the second sale price of property i. Then we can arbitrarily assign values to $\{a_q, b_q\}$ and work out Wilcoxon's signed rank statistic for properties 1 to $i-1$ for these values of $\{a_q, b_q\}$. Next we can vary the values assigned to $\{a_q, b_q\}$ so as to find

SUB ADJUST_INFLATION_TO_GIVE_THE_BEST_RANK_VALUE

Inputs: τ, θ, ψ, σ, ρ and i

Outputs: Best signed rank value,and the predicted house price $\hat{P}_{iT_{2i}}$

For the given τ, θ, ψ, σ and ρ

For the given house i

Find the $\{a_q, b_q\}$ which give the best signed rank value for the set {1,...,i-1}

i.e. alter $\{a_q, b_q\}$ so that $\{\hat{P}_{1T_{21}},\ldots,\hat{P}_{(i-1)T_{2(i-1)}}\}$ give the best signed rank value when compared with the predictions using $\{h_q\}$

For the chosen $\{a_q, b_q\}$ predict $\hat{P}_{iT_{2i}}$

Figure 4 The procedure for predicting the price of house i given the model parameters τ, θ, ψ, σ and ρ. The returned best signed rank value is used in the procedure of Figure 5, while the predicted price is used in the procedure of Figure 6.

SUB DETERMINING_THE_VALUES_OF_THE_PARAMETERS

Vary τ, θ, ψ, σ and ρ

Call ADJUST_INFLATION_TO_GIVE_THE_BEST_RANK_VALUE for house i=21

Return the τ, θ, ψ, σ and ρ that give the best signed rank value

Figure 5 The procedure for determining the model parameters τ, θ, ψ, σ and ρ, using the first 21 houses as the fitting set.

the values that give the greatest value of Wilcoxon's statistic. These optimal values of $\{a_q, b_q\}$ are then used to predict $\hat{P}_{iT_{2i}}$. The outline procedure for this approach is given in Figure 4.

Note that when it comes to predicting $\hat{P}_{(i+1)T_{2(i+1)}}$ the optimal values of $\{a_q, b_q\}$ are likely to be different and so they have revision volatility (Wang and Zorn, 1997). This is to be expected of house price indices as later sales provide information about earlier inflation either directly as in repeat sales, or indirectly through the changing of the regression coefficients in hedonic regression.

Consequently given a set of properties, for example properties 1–21, and the values of $\{\sigma, \rho, \tau, \theta, \psi\}$, then Wilcoxon's signed rank statistic can be calculated by using the above approach to predict for $\hat{P}_{1T_{21}}$, $\hat{P}_{2T_{22}}$, $\hat{P}_{3T_{23}}$, etc. Hence $\{\sigma, \rho, \tau, \theta, \psi\}$ can be varied until the maximum value of Wilcoxon's signed rank statistic is found for this set of properties, and so we have a way of determining the values of $\{\sigma, \rho, \tau, \theta, \psi\}$. The outline of this approach is given in Figure 5.

Two modifications need to be made to the above approach before we can assess whether using $\{a_q, b_q\}$ is more beneficial than using h_q.

SUB PREDICT_THE_HOUSE_PRICES

Inputs: τ, θ, ψ, σ and ρ

For i=22 to 105

Call ADJUST_INFLATION_TO_GIVE_THE_BEST_RANK_VALUE for house i

Store prediction $\hat{P}_{iT_{2i}}$

Figure 6 The procedure for predicting the house prices for the assessment set.

Firstly, when predicting the price of house i, we only know the published house price indices up to the end of the previous quarter, that is, we do not know the current quarter's index. Therefore, if the current quarter is T, then we assume that h_T has the same value as h_{T-1} when applying Equation (9) for the $\{a_q, b_q\}$ prediction and for Equation (6) when applying the prediction based on the published house price indices. However, there could now be an information bias in that the $\{a_q, b_q\}$ prediction could have access to repeat-sales information where the second sale could be in the current quarter T, that is, after the available information for h_q. Therefore, these repeat sales are ignored when calculating $\{a_q, b_q\}$, that is, only second sales before the start of the current quarter are visible.

Secondly, when assessing the approach using $\{a_q, b_q\}$ against the approach using h_q, it is not valid to include the properties that were used to select the values of $\{\sigma, \rho, \tau, \theta, \psi\}$ (as these values were chosen so that the $\{a_q, b_q\}$ approach fitted these properties well). Therefore, the data set is divided into a set used to assign values to $\{\sigma, \rho, \tau, \theta, \psi\}$, and a set used to compare the model's predictions with those using the published house price indices. These will be referred to, respectively, as the *fitting set* and the *assessment set*. The fitting set comprised the 21 properties with the earliest second sale dates, and the assessment set was the remaining 84 properties. The assessment set was chosen to be much larger as the main interest was in whether the new approach gave better predictions, rather than ensuring that the fitted values were very close to the optimal values. (Clearly though there is a trade-off in that making the fitting set too small would lead to a bad model and hence the benefit of the larger assessment set size would be nullified.)

The above approach involves two optimization steps: choosing the best values of $\{\sigma, \rho, \tau, \theta, \psi\}$ using the fitting set, and choosing the best values of $\{a_q, b_q\}$ for the property set under consideration for the chosen values of $\{\sigma, \rho, \tau, \theta, \psi\}$. The outline of these steps are shown in Figures 4–6.

Unfortunately, these are not straightforward optimizations as there are a plethora of local maxima. Ishikawa (2003), and Kolmogorov and Zabih (2004) have considered ways to solve Markov Random Fields through transforming the problem to that of finding minimum graph cuts. Unfortunately, these

methods are quite restrictive on the form of the prior and the likelihood functions, for example they often need to be convex, and the form of Equation (10) does not meet these requirements. Therefore, as our interest is in price prediction rather than a detailed analysis of, for example, the spatial variation of house prices, we employed a grid search to find reasonable values to assign to $\{\sigma, \rho, \tau, \theta, \psi\}$, that is, the situation corresponding to Figure 5.

Finding good values for $\{a_q, b_q\}$ was more important as they feed directly into the sale price prediction through Equation (6). The adopted approach was to start off with values of $\{\sigma, \rho, \tau, \theta, \psi\}$ that allowed $\{a_q, b_q\}$ to vary widely in time and space. Optimal values for $\{a_t, b_t\}$ for these relaxed values of $\{\sigma, \rho, \tau, \theta, \psi\}$ were then found using a gradient descent approach. The values of $\{\sigma, \rho, \tau, \theta, \psi\}$ were then moved towards the original values of $\{\sigma, \rho, \tau, \theta, \psi\}$ and the optimal values of $\{a_q, b_q\}$ were recalculated using the previous iteration's final values of $\{a_q, b_q\}$ as the starting point. The process iterated in this way until the original values of $\{\sigma, \rho, \tau, \theta, \psi\}$ were reached. Although this approach is unlikely to find the best values of $\{a_q, b_q\}$ it has the merits that it is simple to implement and the solutions for $\{a_q, b_q\}$ compared favourably with those from optimizations starting from a variety of different starting points for$\{a_q, b_q\}$, for example randomly perturbing the published house price indices.

Results

The second selling prices for properties 2–105 were predicted using Equation (6) and r_q being in turn each of the four published house price indices mentioned earlier, that is, the Halifax, Nationwide and CLG indices in CLG Table 521, and the North West index in CLG Table 591 (and taking the final quarter's r_q to be the same as the preceding quarter's r_q for each property). For each prediction, its absolute percentage difference from the actual selling price was formed, that is

$$\frac{abs(\hat{P}_{iT_{2i}} - P_{iT_{2i}})}{P_{iT_{2i}}} \times 100 \qquad (11)$$

The resulting means (and standard errors of the mean in brackets) were CLG 11.0% (0.9), Halifax 15.6% (1.3), Nationwide 16.1% (1.3) and North West 14.0% (1.2). Following on from these results, the CLG index was chosen as the h_q index, that is, the published house price index to bench mark the local house price index against, as the predictions from the CLG index are the hardest to improve on for this data set. As a precaution, the Halifax, Nationwide and North West indices were tried as the h_q index, and in each case the local house price index gave a larger improvement over h_q than in the case when the CLG index was used as h_q.

Using the CLG index as h_q and properties 1–21 as the fitting set, the best values for $\{\sigma, \rho, \tau, \theta, \psi\}$ were determined by the approach of Figure 5. The values of $\{\sigma, \rho, \tau, \theta, \psi\}$ that led to the largest Wilcoxon signed rank statistic were $\sigma = 0.2$, $\rho = 0.0$, $\tau = 0.03$, $\theta = 0.015$ and $\psi = 0.02$. It should be noted

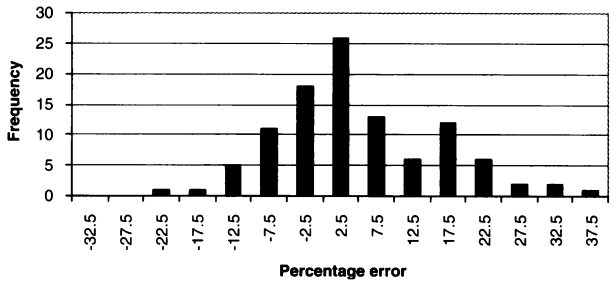

Figure 7 Percentage prediction errors from using the local house price indices $\{a_q, b_q\}$.

though that other parameter values gave statistic values that were very close to the one given by these values. The purpose of the analysis was to find reasonable values for $\{\sigma, \rho, \tau, \theta, \psi\}$ for use in determining $\{a_q, b_q\}$ when predicting the selling prices for properties 22–105, rather than to make findings about whether, for example, ρ is zero or not.

Using these values for $\{\sigma, \rho, \tau, \theta, \psi\}$, the selling prices for properties 22–105 (the assessment set) were predicted using the approach of Figure 6. The difference between the predicted selling price and the observed selling price when using $\{a_q, b_q\}$ was compared with the difference between the predicted selling price and the observed selling price when using h_q, by forming the Wilcoxon signed rank statistic. The resulting P-value from applying the one-sided test to the null hypothesis that the result from using $\{a_q, b_q\}$ will be no better than from using h_q is 0.011. Therefore, there is reasonable evidence to conclude that using information from other local repeat sales improves the prediction. What is of interest is how great this improvement is. The average percentage difference of Equation (11) when the prediction used $\{a_q, b_q\}$ was 8.4% with a standard error of 0.7, while the average when h_q was used was 10.9% with a standard error of 1.1. Additionally, the difference in the accuracy of the predictions for each house was formed by subtracting Equation (11) when $\hat{P}_{iT_{2i}}$ is estimated using $\{a_q, b_q\}$ from Equation (11) when $\hat{P}_{iT_{2i}}$ is estimated using h_q. The average difference in the accuracy of the two predictions was 2.5% with a standard error of 0.8, again indicating that there is evidence that the predictions using $\{a_q, b_q\}$ are better than those using just h_q.

The distribution of the percentage prediction errors (ie Equation (11) without the absolute function) when using $\{a_q, b_q\}$ is shown in Figure 7. The modal column shows that the most common result was a prediction 0 to 5% more than the actual selling price. Figure 7 also shows that there is a slight tendency for the prediction to be higher than the selling price, and this is reflected in the mean of the percentage prediction errors being 0.9%.

Figure 8 investigates whether the percentage prediction errors when using $\{a_q, b_q\}$ are related to the time between a property's two selling times. There seems to be no discernible relationship and calculating the correlation coefficient for the 84 predicted properties gives –0.1.

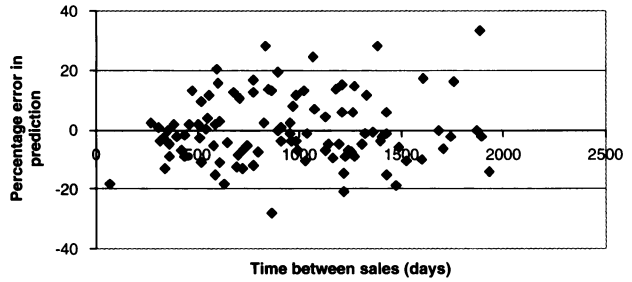

Figure 8 How the prediction errors from using the local house price indices $\{a_q, b_q\}$ relate to the time between sales.

Concluding remarks

This paper has looked at how well inflating previous selling prices predicts current selling prices. For the analysed data set, the average absolute percentage prediction error was 10.9% when using published house price indices. By incorporating local repeat sales into the model, this figure could be improved to 8.4%.

The data set used was restricted to recently constructed properties. It is expected that the predictive capability will be considerably worse on older properties as they are likely to be less uniform and have more potential to be materially altered between sales, for example renovated.

Costello and Watkins (2002) argue that collecting information on physical attributes would give little additional benefit when compared with using repeat sales on Land Registry data when calculating local house price indices. However, incorporating hedonic factors into Equation (10) needs to be explored as different categories of houses may behave differently.

Glossary of symbols

The following symbols appear at more than one place in the paper—symbols that only appear in one location are just defined at that place.

A_{ijt}	is the value of the jth attribute of house i at time t, for example the size of the living area
a_q	the inflation rate for town A for quarter q
b_q	the inflation rate for town B for quarter q
D_{iq}	is 1 if house i's sale was in time period q, and 0 otherwise
D'_{iq}	is 1 if house i's second sale was in time period q, -1 if house i's first sale was in time period q and 0 otherwise
e_i	is house i's error or variation term for the repeat-sales model of Equation (2).
h_q	is the published house price index for quarter q
i	identifies the house
q	identifies the time quarter being considered
P_{it}	is the price of house i sold for at time t
$\hat{P}_{it}$	is the prediction of the price of house i sold for at time t
r_{iq}	is the appropriate inflation rate used for property i for quarter q, that is if i is in town A then it is a_q, while if i is in town B then it is b_q
R_{jq}	is the inflation rate of the logged prices for area j for quarter q as used in Iversen's (2001) local price index model
T_S and T_E	are the starting and ending quarters of the data collection period, that is, all the first and second selling times lie within the period starting at the start of quarter T_S and ending at the end of quarter T_E
T_{1i}	is the time of the first sale for house i
T_{2i}	is the time of the second sale for house i
t	is the time of sale
β_j	is the weight assigned to attribute j by the hedonic regression model of Equation (1)
χ_{iq}	takes the value 1 if house i was for sale for the whole of quarter q, 0 if it was for sale for none of the quarter, and otherwise its value is the fraction of time it was for sale during the quarter
λ_q	is assigned by the hedonic regression model of Equation (1)–the inflation index is derived from it
v_{it}	is the error term for the hedonic regression model of Equation (1). It is typically taken to follow an autoregressive process
ζ_i^2	is defined to be $\frac{2\sigma^2(1-\rho^{T_{2i}-T_{1i}})}{1-\rho^2}$, the variance of the error term ε_i in Equation (8)
ρ	models how the repeat sales error term of Equation (2) is affected by the time between the first and second sales
σ	models the variance of the repeat sales error term of Equation (2) if the time between sales factor is ignored
ψ	is a constant that penalizes local house indices being different from the published house price indices for the same time period
θ	is a constant that penalizes differences between the local house price indices between the neighbouring towns
τ	is a constant that penalizes temporal changes in the local house price indices

Acknowledgements—I am grateful to the Land Registry for the use of the house sales information and published house price indices. The paper was improved by discussions with Victoria Bradley and the comments of the referees.

References

Bailey MJ, Muth RF and Nourse HO (1963). A regression method for real estate price index construction. *Am Stat Assoc J* **58**: 933–942.

Cannaday RE, Munneke HJ and Yang TT (2005). A multivariate repeat-sales model for estimating house price indices. *J Urban Econ* **57**: 320–342.

Costello G and Watkins C (2002). Towards a system of local house price indices. *Housing Stud* **17**(6): 857–873.

Dodgson JS and Topham N (1990). Valuing residential properties with the hedonic method: a comparison with the results of professional valuations. *Housing Stud* **5**(3): 209–213.

Dreiman MH and Pennington-Cross A (2004). Alternative methods of increasing the precision of weighted repeat sales house prices indices. *J Real Estate Finance Econ* **28**(4): 299–317.

Englund P, Gordon TM and Quigley JM (1999). The valuation of real capital: a random walk down Kungsgatan. *J Housing Econ* **8**: 205–216.

Gatzlaff DH and Haurin DR (1997). Sample selection bias and repeat-sales index estimates. *J Real Estate Econ* **14**(1–2): 33–50.

Gatzlaff DH and Ling DC (1994). Measuring changes in local house prices: an empirical investigation of alternative methodologies. *J Urban Econ* **35**: 221–244.

Geltner D and Goetzmann W (2000). Two decades of commercial property returns: a repeated-measures regression-based version of the NCREIF index. *J Real Estate Finance Econ* **21**(1): 5–21.

Goetzmann WN and Spiegel M (1995). Non-temporal components of residential real estate appreciation. *Rev Econ and Stat* **77**(1): 199–206.

Goetzmann WN and Spiegel M (1997). A spatial model of housing returns and neighbourhood substitutability. *J Real Estate Econ* **14**(1–2): 11–31.

Hill RC, Knight JR and Sirmans CF (1997). Estimating capital asset price indexes. *Rev Econ Stat* **79**(2): 226–233.

Hill RC, Sirmans CF and Knight JR (1999). A random walk down main street? *Regional Sci Urban Econ* **19**: 89–103.

Ishikawa H (2003). Exact optimization for Markov Random Fields with convex priors. *IEEE Trans Pattern Anal Machine Intell* **25**(10): 1333–1336.

Iversen ES (2001). Spatially disaggregated real estate indices. *J Bus Econ Stat* **19**(3): 341–357.

Kaipio J and Somersalo E (2005). *Statistical and Computational Inverse Problems*. Springer: New York.

Kolmogorov V and Zabih R (2004). What energy functions can be minimized via graph cuts? *IEEE Trans Pattern Anal Machine Intell* **26**(2): 147–159.

Leishman C and Watkins C (2002). Estimating local repeat sales house price indices for British cities. *J Prop Invest Finance* **20**(1): 35–58.

Quigley JM (1995). A simple hybrid model for estimating real estate price indexes. *J Housing Econ* **4**: 1–12.

Wang FT and Zorn PM (1997). Estimating house price growth with repeat sales data: What's the aim of the game? *J Housing Econ* **6**: 93–118.

Received June 2006;
accepted November 2007 after one revision