

**AN ASYMMETRIC LOCALLY WEIGHTED REGRESSION
SMOOTHER FOR HOUSING SALES PRICE DATA**

Methodology report prepared for John D. Wood and Co.

Dr. Stephen Gibbons* and Professor John Muellbauer**

August 2008

* Department of Geography and Environment, Spatial Economics Research Centre and Centre for Economic Performance, London School of Economics

**Nuffield College, University of Oxford

1. Aims and Introduction

This report sets out a methodology for estimation of a monthly sales price ‘index’ from retrospective time series data on housing transactions. The aim is to provide clients John D. Wood and Co. with a simple methodology for construction of an index of monthly sales prices based on their transactional data, in such a way that the index is invariant to additions to the series in subsequent months. The primary proposed index is based on smoothing raw data on sales price per square foot using an asymmetric LOWESS (locally weighted regression) methodology. An alternative index based on the simple average of the past three months of data is also discussed. Both smoothers can be easily implemented in Microsoft Excel, avoiding the need for specialist software.

2. Specification of the problem and description of the data

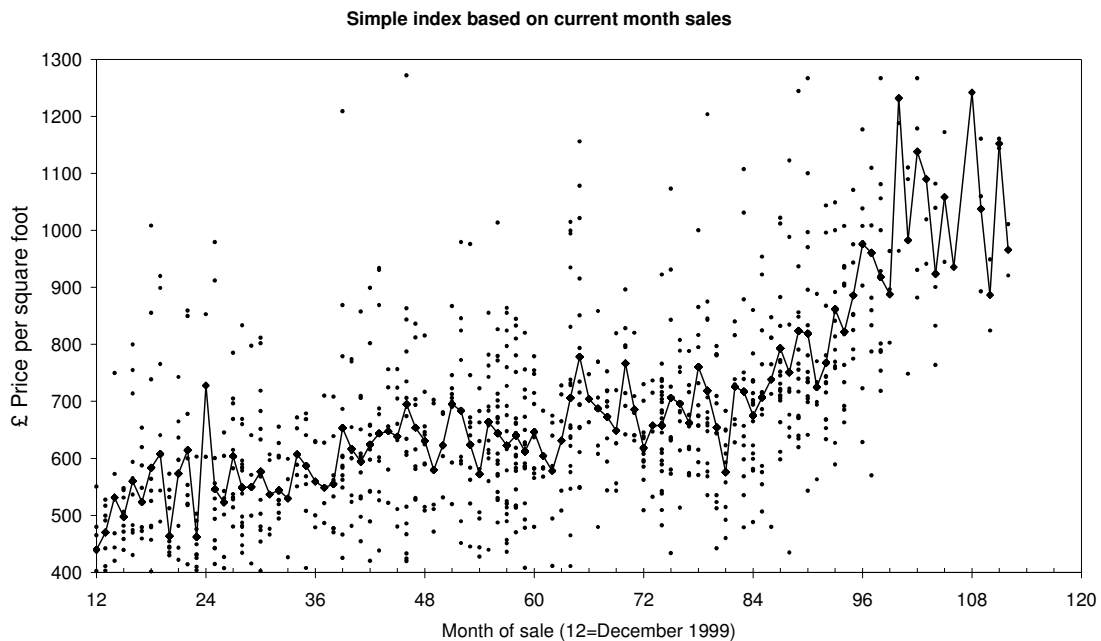
John D. Wood and Co. collect data on their housing transactions in the central London area. These data provide information on sales price and other characteristics, including floor area and property type (house, flat etc.), address postcode and date of sale. The objective is to provide an index of price per square foot on a monthly basis, for predefined groups of housing sales grouped by property type (e.g. flats) and geographical area (e.g. postcode districts). A specific requirement of the project brief is that the sales price index is based only on price per square foot of floor area and date of sale.

An essential requirement of the proposed index is that it provides an estimate of mean transaction prices in a given month s that is insensitive to the addition of data in subsequent months ($s+1, s+2, \dots$). A trivial solution is the simple monthly mean of price per square foot, which provides unbiased estimates of the market price of properties of the type sold in that month (‘unbiased’ means that the estimate equals, on average, the true unobservable market

price). However, this index will have a high variance when the number of sales in a given month is low, resulting in a ‘noisy’ and erratic series.

The problem is illustrated for example data in Figure 1. In the sample data provided, the number of recorded flat sales per month in postcode district SW3 varies between 1 and 23, with a mean of 9.26 and a standard deviation of 5, giving rise to a noisy monthly mean series, as shown in Figure 1. The aim of the methodology described here is to smooth the data further to reveal the underlying trend more effectively, and to provide lower variance estimates of expected market prices in each month.

Figure 1: Raw data points and monthly mean index in sample data 2000-2008



3. Smoothing methodology for the John D. Wood London index

A range of standard techniques are available for ‘smoothing’ data series to produce estimates of underlying trends. These ‘smoothers’ are commonly implemented in applications like Excel, SPSS and JMP7.0. The assumption in all these techniques is that the data is generated by an underlying deterministic process (the housing market in the case of property transactions), but that the

underlying process is obscured in the observable data by random fluctuations, measurement error and variability in factors that are not of interest (for example, unusually high or low price agreements between buyer and seller).

A standard approach to obtaining an estimate of the underlying process in this kind of data is to 'fit' a non-linear trend line through the entire data series by estimating a least-squares regression with polynomial terms in time (squared, cubic etc. terms) as explanatory variables. The predictions from this regression model are then used as the values for the trend line. However, the trend lines produced by this approach are sensitive to observations throughout the entire data set, so will change as new data are added to the series. Polynomial trends fitted throughout a long data series will also be incapable of revealing short term market fluctuations, such as seasonal patterns.

A second class of smoothers, referred to as moving averages, is based on 'local' subsets of the full set of data. Moving averages are simple averages of the data points that are close in time to each time period at which an estimate of the underlying expected price is required. Typical implementations make use of both past and future price observations. Hence, to estimate the expected price in a series of transactions y at time period s a typical symmetric moving average smoother uses the mean or median of the values of y that fall within a predefined window either side of period s ¹. Estimates at time s based on a symmetric moving average will be sensitive to new data that falls within the predefined time window. A moving average can be arranged to use data points prior to the required period only, which will mean that new data does not affect past estimates. This type of smoother is also discussed below. However, an asymmetric moving average is likely to be biased upwards when prices are falling (because past prices are above current prices), and biased downwards when prices are rising (because past prices are below current prices), i.e. the estimate will not, on average equal the actual unobservable market price during rising or falling markets.

A third class of smoothers combines these two ideas, using a 'local' subset of the data (as in the moving average) to estimate a polynomial regression. These smoothers are known as Locally Weighted Regressions (LOWESS, or LOESS), and are attributed originally to Cleveland (1979). Cleveland and Levin (1988) and other papers discuss the properties of these smoothers and the conditions under which the procedure provides an unbiased estimate of the underlying trend in the data.

The method applied in the current application is a LOWESS smoother, adapted for the specific needs of John Wood and Co that are outlined in Section 2. The technical Appendix provides full details. The main modification to the methodology described in Cleveland (1979) is to use an asymmetric regression weighting function. This means that, for an estimate at time period s , positive weights are applied to data observations (transactions) prior to period s and zero weights to data added after time period s . A polynomial regression model is then fitted using this weighted data. In other words, the prediction for the market price at a period s is derived from a polynomial regression of price on time, fitted through the data for a number of transactions prior to and including period s . Observations after period s are ignored in the estimation so period s predictions are unaffected by new data added after period s . A separate regression is estimated for each transaction and provides a predicted price for each transaction. The mean of predicted prices of transactions within any period s are then averaged to provide the market price index for period s . The procedure is thus, in effect, a moving average based on past data, but adjusted for local rising and falling trends in the data.

For comparison, we also calculate and present some simple asymmetric Moving Average smoothers. The first calculates the mean of the past 32 sales for each transaction, then averages these by month to give a monthly index. The second estimates the price at month s from the mean of the three past months' prices (i.e. for periods $s, s-1, s-2$).

3.1. Choice of bandwidth and regression polynomial in LOWESS

The two factors that must be chosen in any LOWESS smoother are (i) the order of the polynomial regression (constant only, linear, quadratic, cubic, etc.) and (ii) the weighting system.

A first feature of the weighting system is its 'bandwidth', which in this application is simply the number of observations used for each regression and index point. A second requirement of the weighting function applied here is that it is asymmetric, such that only past observations are used in estimation. For simplicity in computation, equal weights are applied to all observations within the chosen bandwidth². Bandwidth choice in the current context is inherently subjective, and must balance the desire to achieve a predicted price index that follows short run fluctuations in the market (e.g. seasonal variation) against the desire to have a smooth trend that removes 'noise' in the price series due to atypical sales.

Choice of polynomial has similar implications. For a given bandwidth, a lower order polynomial (e.g. a linear regression of price on time) results in a 'smoother' index plot than a higher order polynomial (e.g. with linear, quadratic and cubic time variables as explanatory variables)³. Provided that the bandwidth is sufficiently narrow (based on a short period in the market history) there is little to be gained from using a high order polynomial. The only requirement for unbiased predictions is that regression line can adequately follow the general underlying time trends within the period spanned by the bandwidth, i.e. rising trends, falling trends, plus turnaround points at peaks and troughs in the market. A quadratic function can capture these features adequately, and is the polynomial function adopted here⁴.

Given this choice of quadratic functional form, the appropriate bandwidth can be chosen by experimentation, and will depend to some extent on the characteristics of the data. Choice depends largely on a) the capability of the smoother to follow the underlying data b) the effectiveness of the smoother in eradicating random variation. One way of assessing the

smoother on the first criterion is to estimate the correlation of the predicted prices with the raw data. For the second criterion, we can consider the minimising the standard deviation of the smoother, within a chosen time frame.

Table 1 below shows the correlation with price, and the average within-year standard deviation for the raw data, a simple monthly average, a Moving Average smoother based on the past 32 sales (3 months on average), a Moving Average based on the past three months of sales, and the quadratic LOWESS smoother with various bandwidth choices. As can be seen from the table, all the smoothers provide a dramatic reduction in within-year standard deviation, and there is a trade off between the degree of smoothing and the correlation with the raw price data. For the quadratic smoother, a bandwidth of 90 sales performs best in terms of balancing predictive power with low within-year variance and a similar choice of bandwidth is reached by visual inspection of the plots (see below). On this basis, the optimal bandwidth choice based on this sample data of flats in SW3 sold between 2001 and 2008 is around 90, corresponding to approximately nine months of historical data. This bandwidth can be easily adjusted by the end-user to achieve an index with more, or less, variability.

Table 1: Evaluation of effects of bandwidth choice

Smoother	Bandwidth	Within-year standard dev.	Correlation with price
Raw price	-	150.0	1.000
Monthly mean	12 sales (mean)	52.3	0.656
Past 32 sales MA	32 sales	32.0	0.600
3 month MA	32 sales (mean)	34.4	0.617
Quadratic	30 sales	52.7	0.644
Quadratic	60 sales	48.8	0.623
Quadratic	80 sales	42.0	0.620
Quadratic	90 sales	40.2	0.621
Quadratic	100 sales	40.7	0.617
Quadratic	120 sales	41.8	0.613
Quadratic	150 sales	42.5	0.606

Bias is estimated relative to simple monthly mean

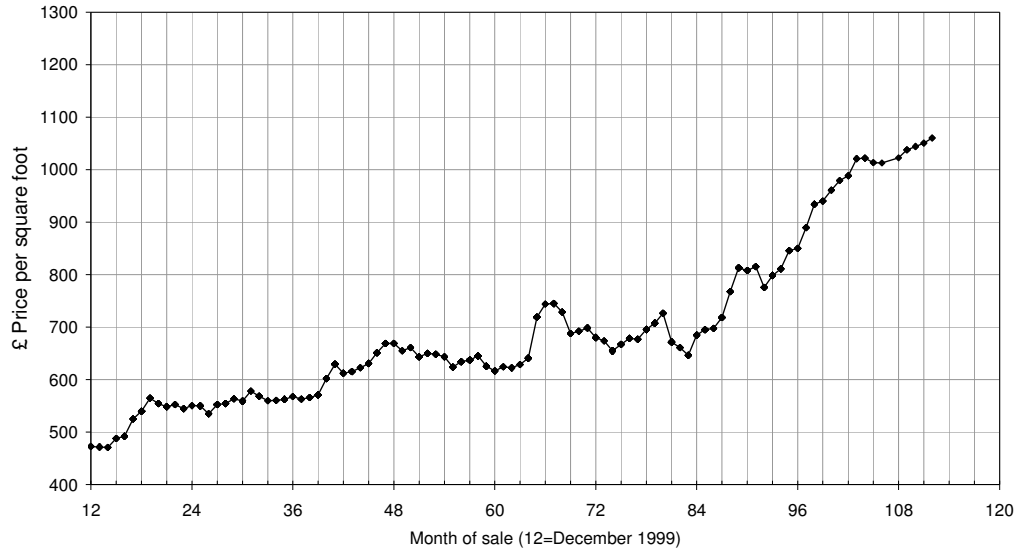
MSE is total variance + bias-squared

The simple past 32 sales Moving Average has a low-within-year variance, but performs relatively poorly in terms of explanatory power (correlation). The three month Moving Average based on the past three months of sales (an average of 32 sales) also provides a high degree of within-year smoothing but performs well in terms of correlation with the raw prices. This smoothing method, provides an alternative simple solution, but, as discussed above, may under predict at the peak of fast rising markets and over predict at the bottom of periods of rapidly falling markets, because the trend over the past three months is not taken into account in estimating the market price in a given month. This is evident from visual inspection of the plots of these smoothers on the sample data as discussed below.

Figure 2 plots the past 32 sales and 3 Month Moving Average smoother. Figure 3 presents the results of the LOWESS smoothing methodology for three bandwidth choices using a local quadratic regression: 30 transactions, 90 transactions and 150 transactions. Comparing the plots, it is evident that the 32 sales Moving Average creates a highly smoothed plot that eliminates many short run features of the data. The Three Month Moving Average is better at detecting peaks and troughs, but flattens out short run variation relative to the LOWESS smoother at moderate bandwidths. It is noticeable that the peaks in the market e.g. at 66 months and 103 months are attenuated by the Moving Average smoother, relative to the LOWESS smoother.

Figure 2: Asymmetric Moving Averages

Monthly mean of last 32 sales



Mean of last 3 monthly means

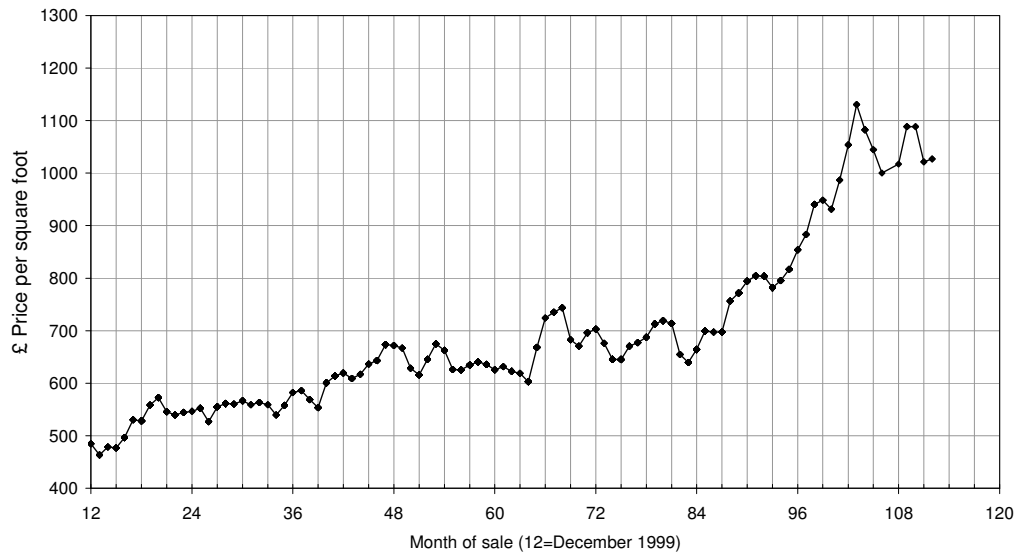
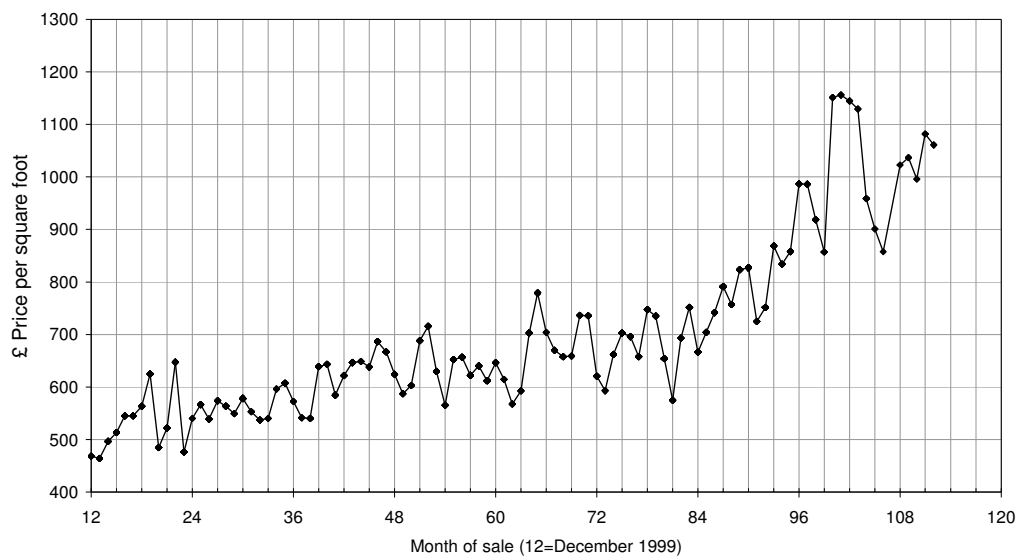
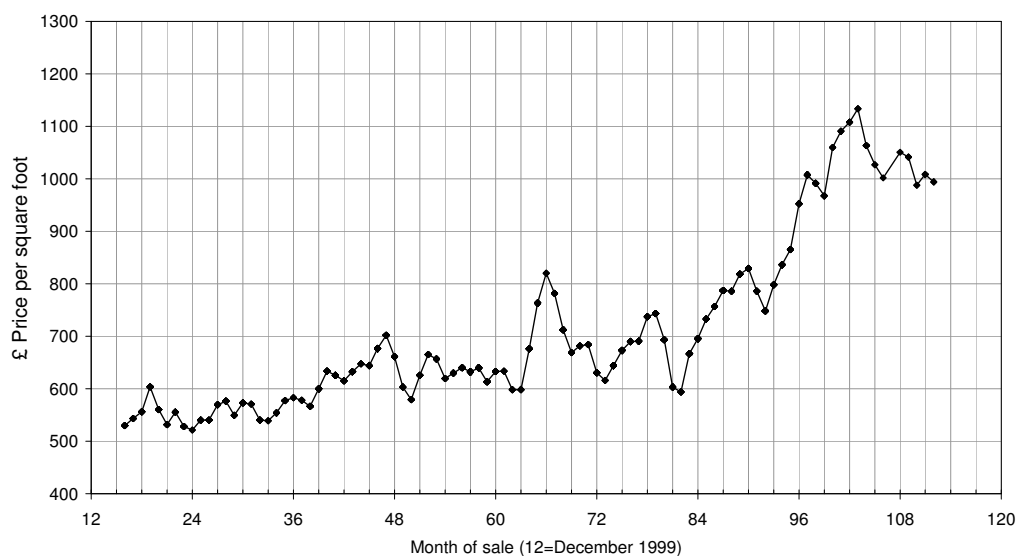


Figure 3: Quadratic asymmetric LOWESS smoother plots for various bandwidths

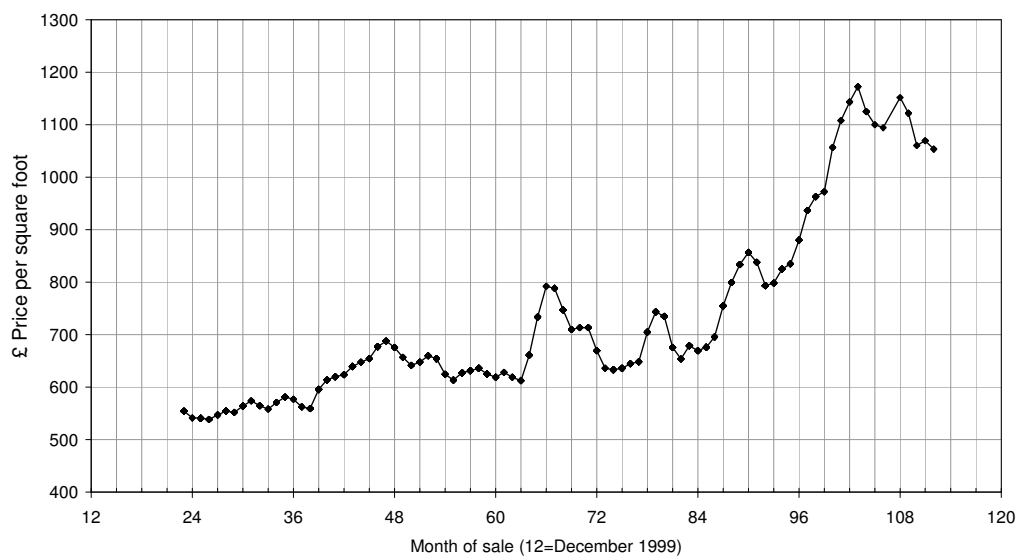
Bandwidth
= 30 sales



Bandwidth
= 90 sales



Bandwidth
= 150
sales



3.2. Limitations and health warnings

It should be recognised that a price index constructed using this simple smoothing approach will only provide an estimate the expected market price of properties of the type sold during the period over which the index is estimated, and on which sales price and floor area are included in the underlying data. The index is not adjusted for variation in the mix of properties sold in any period in terms of type, amenities and other aspects of quality, except through the choice of property sales included in the data. If the smoother is estimated using data on a wide range of property types and locations, it can not properly be interpreted as a market price index for a sale of a representative property without first adjusting the sales prices for period to period variation in the quality, type and location of properties sold (e.g. using 'hedonic' methods, see Gibbons and Machin 2008)). A second caveat is that, as shown in Section 3.1, the smoother provides an estimate of the underlying price trends which will vary according to the choice of bandwidth (and other parameters of the LOWESS smoother). Hence, the method is appropriate only for comparing prices in different periods where the index is estimated using the same methodology on the same types of properties in the same local market.

One further point that should be taken into account when interpreting an index based on this methodology is the limitation imposed by the constraints of asymmetry in estimation (only past data is used) and quadratic smoothing. Because the smoother makes use of past data only, the current price prediction is determined only by past price trends, and is not subsequently adjusted for new information that becomes available in later periods. Given quadratic smoothing, the index can only follow price trends that follow a local parabolic pattern within the chosen data bandwidth. Hence, there are situations where the smoother may over-predict or under-predict the market price within a given month because of very sharp changes in direction in price trends, which the quadratic is unable to follow.

4. References

Cleveland, William (1979) Robust Locally Weighted Regression and Smoothing Scatterplots, *Journal of the American Statistical Association*, 74 (368): 829-836

Cleveland, William and Susan J. Devlin (1988) Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, *Journal of the American Statistical Association*, 83 (403): 596-610

Gibbons, Stephen and Stephen Machin (2008) Valuing school quality, better transport and lower crime: evidence from house prices, *Oxford Review of Economic Policy*, 24 (1) 99-119

5. Technical Appendix

The smoothing procedure assumes that prices are generated by a function $g(s)$ of unknown functional form, with an additive error term

$$p_s = g(s) + \varepsilon_s \quad (1)$$

Where p_s is a sales price (per square foot) on day s , running from $s=1$ to $s=T$. The aim of the smoothing procedure is to estimate the function $g(s)$. The LOWESS smoother with a quadratic polynomial is a weighted regression of price on time (s) and time-squared, of the form

$$w_{st} p_t = \alpha_s w_{st} + \beta_{1s} w_{st} s_t + \beta_{2s} w_{st} s_t^2 + v_t \quad (2)$$

Where t indexes data observations used in the estimate of α, β_1, β_2 at period s , and the weights w_{st} decrease as the time interval between period s and t increases. In the methodology described in this report, we define:

$$w_{st} = \begin{cases} 1 & \text{if } s-t < k \\ 0 & \text{if } s-t < 0 \text{ or } s-t > k \end{cases} \quad (3)$$

Where k is a bandwidth parameter that defines the number of observations used in estimating each LOWESS regression. The model specified in (2) is estimated by ordinary least squares. The parameter estimates are then used to provide a prediction for p_s

$$\hat{p}_s = \hat{\alpha}_s + \hat{\beta}_{1s}s + \hat{\beta}_{2s}s^2$$

A price index for any longer time period (e.g. months, quarters, years) for which there are multiple estimates p_s can be obtained from the mean of these predictions within each encompassing period. For example, to construct a monthly index for any month m from estimates p_s that fall within that month:

$$\bar{p}_m = \frac{1}{N_M} \sum_{t \in M} \hat{p}_t$$

Where M is the set of observations falling within month m and N_m is the number of observations in this set.

Endnotes

¹ For example a moving average of y at period s could be the mean

$$\hat{y}_s = \text{mean}(y_{s-k}, \dots, y_{s-2}, y_{s-1}, y_s, y_{s+1}, y_{s+2}, \dots, y_{s+k})$$

² More general weighting systems can be used, e.g. with lower weights for observations further away from the prediction period, but are more difficult to implement and are unlikely to offer any advantages in the current application.

³ Note that a zero order polynomial implies a simple mean of the data within the specified bandwidth, and is thus equivalent to a moving average.

⁴ Note that if the bandwidth spans only two periods, the quadratic reduces to linear regression. If the bandwidth spans only one month index reduces to a simple mean of the monthly sales prices.