A Local House Price Index for Aberdeen
Methodology Revision 2019

Rainer Schulz and Martin Wersing
University of Aberdeen Business School

January 2020

Contents

1 Introduction

Since 2012Q2, we publish the *Aberdeen Housing Market Report* (AHMR) jointly with the *Aberdeen Solicitors' Property Centre* (ASPC). The AHMR gives constant-quality price information for three property types and five local areas within the housing market of Aberdeen City and Aberdeenshire on a quarterly basis. The AHMR contains also additional information and is available from the ASPC Information Centre: House Prices.

In 2019, we examined our calculation methodology for the constant-quality prices, because we wanted to: *i*) use property characteristics that the ASPC had started to record from 2017 onwards; *ii*) use a more flexible econometric model for the calculation; *iii*) provide information on the relationship between sale and ask prices for successful listings. This examination resulted in a revision of the calculation methodology; it also led to an extension of the coverage of the AHMR. Starting with 2020Q2, the revised methodology is used to produce the price information contained in the AHMR. This technical report details the revised methodology. It replaces the previous technical report (Owusu-Ansah et al., 2013), which details the methodology that was used for the AHMR 2012Q2-2020Q1.

2 Hedonic regressions

2.1 Overview

The dependent variable in the regressions is either the log sale price p or the log sale to ask price ratio r. The respective dependent variable is regressed on a bundle of property characteristics collected in the vector $\mathbf{x}$. Some characteristics enhance pleasure (e.g., a good location), others reduce pleasure (e.g., a bad location). This should affect the price a property will fetch and motivates the expression *hedonic* regression.

We fit the regressions in each quarter t to data provided by the ASPC. The data

cover all transactions that took place in the last eight quarters, including quarter t (in total a period of two years). The regressions result in the estimated functions $\hat{p}_t(\mathbf{x})$ and $\hat{r}_t(\mathbf{x})$, respectively. As we will see later, the estimated functions allow the computation of prices, price indices, and ratios of sale to ask prices over time for notional properties with reference characteristics $\mathbf{x}_0$.

2.2 Regression equations

We assume that each of the unknown functions $p_t(\mathbf{x})$ and $r_t(\mathbf{x})$ follow a semi-parametric additive model

$$m(\mathbf{x}_i) = \boldsymbol{\iota}_i\boldsymbol{\gamma} + \mathbf{z}_i\boldsymbol{\Theta}_{j(i)} + f_1(FA_i) + f_2(LAT_i, LON_i) \tag{1}$$

The shape of the f_j functions and the coefficients collected in $\boldsymbol{\gamma}$ and $\boldsymbol{\Theta}$ can change over time. The functions and coefficients can also differ depending on the dependent variable. All we assume is that the functional *form* follows always the model in Eq. 1.[1] The regression equations are

$$p_i = p_t(\mathbf{x}_i) + \epsilon_i^p \tag{2a}$$

$$r_i = r_t(\mathbf{x}_i) + \epsilon_i^r \tag{2b}$$

where both ϵ_i^p and ϵ_i^r are noise terms and $p_t(\mathbf{x}_i)$ and $r_t(\mathbf{x}_i)$ follow the functional form from Eq. 1.

2.3 Data samples and characteristics

The estimation sample for quarter t results (e.g. 2021Q1) covers all transactions that took place in the current quarter t or in one of the previous seven quarters. Although the number N of observations can vary between estimation samples,

[1] We could have used more notation, such as sub- and superscripts, to make this explicit, but this would have cluttered the presentation.

we will not make this explicit below to keep the notation simple. For each property i in the estimation sample, we observe: p_i, r_i, property's type, the local area in which the property is located, and the bundle of characteristics $\mathbf{x}_i = (\boldsymbol{\iota}_i, \mathbf{z}_i, FA_i, LAT_i, LON_i)$.

The first element in $\mathbf{x}_i$ is the (1×7) vector $\boldsymbol{\iota}_i$. This vector indicates in which of the eight quarters from t back to $t-7$ property i has been transacted. If property i was transacted in $t-7$, $\boldsymbol{\iota}_i$ contains only zeros, if property i was transacted in $t-6$, the last entry of $\boldsymbol{\iota}_i$ is one, all other entries are zeros, and so on. The coefficients in the (7×1) vector $\boldsymbol{\gamma}$ control therefore for the average trend over the period covered by the estimation sample, see Eq. 1. The second element in $\mathbf{x}_i$ is the (1×18) vector $\mathbf{z}_i$. This vector contains a constant and binary indicators for the characteristics in Table 1. For instance, property i has either a garden or not and the binary indicator for this characteristic in $\mathbf{z}_i$ takes either the value one or zero. For the number of bedrooms, there are seven categories. For instance, property i could have three bedrooms, in which case the indicator for three bedrooms in $\mathbf{z}_i$ takes the value one, indicators for other possible number of bedrooms categories take the value of zero.[2] We allow the coefficients for the characteristics in $\mathbf{z}_i$ to differ with respect to property type (detached, semi-detached, flat) and the local area in which the property is located (one out of five).[3] In Eq. 1, the matrix $\boldsymbol{\Theta}$ has $3 \times 5 = 15$ columns, one for each possible type-area combination. The subscript $j(i)$ on the matrix means that the coefficient vector in column j is selected that corresponds to the type-area combination of property i.

The remaining three element in $\mathbf{x}_i$ are the floor area FA_i of property i in square metres, and the location coordinates of the property measured by latitude LAT_i and longitude LON_i. The impact of these continuous characteristics is considered by the smooth, but otherwise unspecified, functions $f_1(\cdot)$ and $f_2(\cdot)$.

[2] For each characteristic, one category is used as baseline and is considered indirectly. For instance, a property has no garage if the garage indicator is zero; a property has one bedroom if all the indicators for two or more bedrooms are zero.

[3] The semi-detached property type category includes terraced houses.

Table 1: Categorical property characteristics. Categorical property characteristics that are included with binary indicators in the vector $\mathbf{z}_i$.

Characteristic	Indicators
Garage	1 if property has a garage, 0 otherwise
Garden	1 if property has a garden, 0 otherwise
Bathrooms	1 if property has n bathrooms, 0 otherwise ($n = 2, \ldots, 5$)
Bedrooms	1 if property has n bedrooms, 0 otherwise ($n = 2, \ldots, 7$)
Public rooms	1 if property has n public rooms, 0 otherwise ($n = 1, \ldots, 5$)

2.4 Estimation

We model the function $f_1(FA_i)$ in Eq. 1 with the cubic spline basis

$$f_1(FA_i) = FA_i\beta_{11} + \sum_{k=2}^{K_1} |FA_i - FA_k^*|^3 \beta_{1k} \tag{3}$$

The $K_1 - 1$ knots $FA_2^*, \ldots, FA_{K_1}^*$ are placed at the $K_1 - 1$ equally spaced quantiles of the observed FA. We impose during estimation the two natural spline constraints $\sum_{k=2}^{K_j} \beta_{1k} = 0$ and $\sum_{k=2}^{K_1} \beta_{1k} FA_k^* = 0$, so that the second derivative of $f_1(\cdot)$ is zero outside $[FA_2^*, FA_{K_1}^*]$. This reduces the risk of extrapolation (Wood and Augustin, 2002, p. 160). As the constant is already considered in $\mathbf{z}_i$, there is none in Eq. 3.

We model the function $f_2(LAT_i, LON_i)$ in Eq. 1 with the thin plate spline basis

$$\begin{aligned} f_2(LAT_i, LON_i) &= LAT_i\beta_{21} + LON_i\beta_{22} \\ &\quad + \sum_{k=3}^{K_2} b_{2k}(||(LAT_i, LON_i) - (LAT_k^*, LON_k^*)||)\beta_{2k} \end{aligned} \tag{4}$$

which has again no constant term. In Eq. 4, $||\mathbf{u}|| = \sqrt{\mathbf{u}'\mathbf{u}}$ is the Eucledian norm and $b_{2k}(||\mathbf{u}||) = \frac{1}{8\pi}||\mathbf{u}||^2 \log ||\mathbf{u}||$ is a function of this norm (Wood and Augustin, 2002, p. 171). The $K_2 - 2$ location knots (LAT_k^*, LON_k^*) in Eq. 4 are determined from a random subset of the actual locations of the observations via the eigenvalue decomposition described in Wood (2003).

We estimate each of the two regression equations in Eq. 2 using the model given in Eq. 1 and the two basis functions Eq. 3 and Eq. 4 with the penalised least squares estimator

$$(5) \quad \left(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\beta}}\right) = \arg\min_{\boldsymbol{\gamma},\boldsymbol{\Theta},\boldsymbol{\beta}} \left[\sum_{i=1}^{N} \left\{y_i - m(\mathbf{x}_i; \boldsymbol{\gamma}, \boldsymbol{\Theta}_{j(i)}, \boldsymbol{\beta})\right\}^2 + \sum_{j=1}^{2} \lambda_j \boldsymbol{\beta}_j' \mathbf{D}_j \boldsymbol{\beta}_j\right]$$

where y_i is either p_i (when we fit Eq. 2a) or r_i (when we fit Eq. 2b). N is the number of observation in the estimation sample. The vector $\boldsymbol{\beta}_1$ collects all coefficients from the basis function Eq. 3, the vector $\boldsymbol{\beta}_2$ collects all coefficients from the basis functions Eq. 4 and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. The term $\boldsymbol{\beta}_j' \mathbf{D}_j \boldsymbol{\beta}_j$ evaluates $\int \left[f_j''(x)\right]^2 dx$ and becomes large if f_j is very wiggly and small if the function is fairly straight.[4]

The smoothing parameter λ_j determines the degree at which wiggliness of the estimate of f_j is penalised. We select $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ by minimizing the double cross-validation score (DCVS), i.e.,

$$(6) \quad \hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \frac{N \sum_{i=1}^{N} \left(y_i - \hat{y}_i(\boldsymbol{\lambda})\right)^2}{\left\{N - 1.5\text{tr}\left(\mathbf{H}(\boldsymbol{\lambda})\right)\right\}^2}$$

through an iterative procedure. In the expression for the DCVS on the right-hand side of Eq. 6, y_i and $\hat{y}_i(\boldsymbol{\lambda})$ are the price p_i and the predicted price $\hat{p}_i$ (the ratio r_i and the predicted ratio $\hat{r}_i$) when we fit Eq. 2a (Eq. 2b) for a given set of $\boldsymbol{\lambda}$ values. $\mathbf{H}$ is the *hat matrix* of the penalized least squares estimator in Eq. 5, see Wood (2017, pp. 249-50). The DCVS is a consistent estimator of the mean squared prediction error of the regression model and minimizing this score prevents excess smoothing (Wood, 2017, pp. 260-61).

[4] The elements of $\mathbf{D}_j$ are discussed in Wood (2017, Sec. 5.3 and 5.5).

3 Constant-quality information

3.1 Reference characteristics

Once the coefficients $\boldsymbol{\gamma}, \boldsymbol{\Theta}$, and $\boldsymbol{\beta}$ are estimated for quarter t and for both regressions in Eq. 2, we can predict the log price of a property given a set of reference characteristics $\mathbf{x}_0$ using $\hat{p}_t(\mathbf{x}_0)$. We can do the same for the log sale to ask ratio using $\hat{r}_t(\mathbf{x}_0)$. As the constant-quality information should be for the current quarter t, the vector $\boldsymbol{\iota}_0$ in $\mathbf{x}_0$ has as first element a one and all other elements are zero. This ensures that the estimated average price level for t is picked out of $\boldsymbol{\gamma}$. Of the remaining elements in $\mathbf{x}_0$, $\mathbf{z}_0$ contains always the entry for the constant

Table 2: Reference characteristics for properties. Gives the characteristics which are used in the type-area dependent reference vectors $\mathbf{x}_0$. The reference characteristics are based on the medians in the ASPC data over the period from 2015Q3 to 2019Q4. Floor area is measured in square metres. Location in Panel B is measured by the coordinates of the British National Grid.

Panel A. Property characteristics						
	Floor area	Bathrooms	Bedrooms	Public rooms	Garage	Garden
Detached	130	2	2	4	1	1
Flat	60	1	1	2	0	0
Semi-detached	86	1	1	3	0	1
Panel B. Property location						
	East	North	Area			
Aberdeen						
Detached	386800	807700	Kingswell			
Flat	392800	806400	Rosemount			
Semi-detached	392600	811500	Old Machar			
Country	369500	795700	Banchory			
Ellon	394700	831100	Raeburn Place			
Inverurie	376100	822300	Uryside			
Stonehaven	387400	785800	Old Town			

term (a one), but depends otherwise on the property type. The same applies to the floor area, see Panel A of Table 2. Intuitively, the location coordinates in $\mathbf{x}_0$ depend on the areas for which we provide constant-quality information, see Panel B of Table 2. For the City of Aberdeen, we allow further that the coordinates differ by property type.

3.2 Preparation of the information

As we are not really interested in predictions of the log price and the log sale to ask price ratio, we must re-transform the log predictions to the natural scale. We do this with

$$\hat{P}_t(\mathbf{x}_0) = \exp\{\hat{p}_t(\mathbf{x}_0)\} \cdot \left(\frac{1}{N} \sum_{i=1}^{N} \exp\{\hat{\epsilon}_i^p\} \right) \tag{7a}$$

$$\hat{R}_t(\mathbf{x}_0) = \exp\{\hat{r}_t(\mathbf{x}_0)\} \cdot \left(\frac{1}{N} \sum_{i=1}^{N} \exp\{\hat{\epsilon}_i^r\} \right) \tag{7b}$$

see Duan (1983). The residuals $\hat{\epsilon}_i^p$ in Eq. 7a come from the regression in Eq. 2a; the residuals $\hat{\epsilon}_i^r$ in Eq. 7b come from the regression in Eq. 2b.

Price indices are calculated as

$$I_t(\mathbf{x})_0 = \frac{\hat{P}_t(\mathbf{x}_0)}{\hat{P}_b(\mathbf{x}_0)} \tag{8}$$

where b indicates the base period.

3.3 Information reported in the AHMR

The headline figures in the AHMR on the constant-quality price trend are based on the price index $I_t(\mathbf{x}_0)$ for a semi-detached properties in Aberdeen City, see Table 1 for the characteristics $\mathbf{x}_0$ of this particular reference property. The price index reported in the figure in the AHMR is for the very same property.

The price changes for the five areas reported in the AHMR are all for semi-detached properties. The *quarterly* rate of price change $G_{t,t-1}(\mathbf{x}_0)$ is calculated

as

$$G_{t,t-1}(\mathbf{x}_0) = \frac{I_t(\mathbf{x}_0) - I_{t-1}(\mathbf{x}_0)}{I_{t-1}(\mathbf{x}_0)} = \frac{I_t(\mathbf{x}_0)}{I_{t-1}(\mathbf{x}_0)} - 1 \tag{9}$$

To understand the meaning of $G_{t,t-1}(\mathbf{x}_0)$ better, we rewrite Eq. 9 as

$$(1 + G_{t,t-1}(\mathbf{x}_0)) \cdot I_{t-1}(\mathbf{x}_0) = I_t(\mathbf{x}_0) \tag{10}$$

A positive (negative) rate of price change implies that prices as measured by the index have increased (decreased) during the quarter.

The *yearly* rate of price change $G_{t,t-4}(\mathbf{x}_0)$ is calculated as

$$G_{t,t-4}(\mathbf{x}_0) = \frac{I_t(\mathbf{x}_0) - I_{t-4}(\mathbf{x}_0)}{I_{t-4}(\mathbf{x}_0)} = \frac{I_t(\mathbf{x}_0)}{I_{t-4}(\mathbf{x}_0)} - 1 \tag{11}$$

Eq. 11 can be rewritten and understood similar to Eq. 9 and Eq. 10, but for a yearly frequency. As this is obvious, we do not do it here. The *annualised* rate of price change over *five years* is calculated as

$$G^a_{t,t-20} = \left(\frac{I_t(\mathbf{x}_0)}{I_{t-20}(\mathbf{x}_0)} \right)^{1/5} - 1 \tag{12}$$

Rewriting Eq. 12

$$(1 + G^a_{t,t-20})^5 \cdot I_{t-20}(\mathbf{x}_0) = I_t(\mathbf{x}_0) \tag{13}$$

shows that $G^a_{t,t-20}$ is the *notional* constant growth rate per annum at which the index has changed over the past five years. If this rate is positive, it gives the rate at which the index has increased (deceased) per year over the last five years.

The imputed prices in the AHMR for the current and the previous quarter for the three different property types in the five different local areas are calculated

with Eq. 7a and the respective reference characteristics from Table 2.

The mark-up reported in the AHMR is for a semi-detached property in Aberdeen City. The mark-up is calculated with Eq. 7b and the respective reference characteristics from Table 2.

Disclaimer

The disclaimer in the Aberdeen Housing Market Report extends to this technical report.

4 References

Duan, N.: 1983, Smearing estimate: A nonparametric retransformation method, *Journal of the American Statistical Association* **78**, 605–610. (Cited on page 9.)

Owusu-Ansah, A., Roberts, D., Schulz, R., and Wersing, M.: 2013, Developing a local house price index: The case of Aberdeen, Scotland, *Technical report*, University of Aberdeen Business School. (Cited on page 3.)

Wood, S. N.: 2003, Thin plate regression splines, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**, 95–114. (Cited on page 6.)

Wood, S. N.: 2017, *Generalized additive models. An introduction with R*, Texts in Statistical Science, 2 edn, CRC Press, Boca Raton. (Cited on page 7.)

Wood, S. N. and Augustin, N. H.: 2002, GAMs with integrated model selection using penalized regression splines and applications to environmental modelling, *Ecological Modelling* **157**, 157–177. (Cited on page 6.)