FOI Reference: FOI/2021/3043

Under the Freedom of Information Act 2000, I am writing to request you make publicly available in electronic format the following information regarding the ONS House Price Index ("HPI") methodology.

1. In the standard, geometric version of the HPI, which is published publicly, a semi-log model is used, where the price is log-transformed prior to fitting the regression, and then transformed back to linear form when the imputed prices for each cell are being produced. Is this log-transform of the price still used in the arithmetic version of the HPI (produced exclusively for the monthly Retail Price Index depreciation series [CHOO]), or is the regression fit on the untransformed (level) prices in the arithmetic version of the HPI?
2. If a log-transform **is** still used for the arithmetic version of the HPI, as per question 1, where in the arithmetic HPI model pipeline is the transformation (via exponentiation) back to linear/level scale done:
    a. on the imputed prices in the fixed basket of properties, before aggregation into cells (cells are as defined in ONS methodology)
    b. on the aggregated average prices determined for each cell
    c. elsewhere (please specify).
3. In the document entitled Development of a single Official House Price Index published in 2016 (https://www.ons.gov.uk/economy/inflationandpriceindices/methodologies/developmentofasingleofficialhousepriceindex) the following is stated regarding the geometric mean:

   The price determining characteristics can then be combined together to give a predicted price for each property in the fixed basket. These predicted prices are then averaged using a geometric mean, which involves multiplying the 'n' predicted prices together, and then taking the nth root. (In practice, the geometric mean is calculated by summing the logs of the predicted prices, dividing by n, and taking the exponential of the result.)

   After producing the imputed prices for each property in the fixed basket (in the geometric version), are the imputed prices converted back to linear form (using exponentiation, with bias adjustment term), then transformed back to log form for the purposes of the geometric mean calculation from the excerpt above, or is the geometric average **directly** taken as the arithmetic mean of the log-scale imputed prices (which is then exponentiated to linear scale)?

   In the arithmetic version, are the imputed prices for the properties in the fixed basket exponentiated back to level/linear form, with the arithmetic mean of these linear-form prices then being taken? If not, how are the imputed prices converted to an average price in the arithmetic version of the HPI?

4. In the geometric version of the HPI, a fixed basket of properties is used for the imputation process. We have been informed by ONS previously that this fixed basket of properties consists of the prior calendar year's entire set of property sale transactions. For example, the house price index produced in 2021 would use all sale transactions from 2020 as its fixed basket.

   In the arithmetic version of the HPI, are there any differences in the composition or selection criteria for the fixed basket of properties, when compared to the selection process described above for the geometric version of the HPI?

5. The fixed basket of properties for a given calendar year's HPI is set in each January of said year, using all property transactions from the prior year, as described in question 4. However, the Price Paid Dataset (PPD), from which these transactions are drawn, revises each month with new transactions which were closed in prior months, but had not yet been processed by HM Land Registry. These revisions are typically quite large, in terms of the number of transactions added, for the first two to three revisions for a given month.

   As a result, when the fixed basket of properties is set in January of each year, the property sale transactions which were closed in the latter months of the prior year will not yet be completely available, as these months will be revised with added transactions in the Price Paid Data releases in February, March, April, etc. How do you handle these property sale transactions from the prior year which are added in PPD revisions **after** the fixed basket has been set in January?

6. Are there any changes to the property attributes (e.g. room variable, Acorn variable etc.) in the arithmetic version of the HPI, or is the exact same set of attributes used in both the arithmetic and geometric versions of the HPI, with the same configuration?

7. When you are creating your model each month for the HPI, are the geometric and arithmetic versions of the HPI drawn from precisely the same model, or are two separate hedonic regressions produced: one for the geometric index and one for the arithmetic index?

   a. If they are using **the same** model, could you specify the precise point in the methodology where the arithmetic version first deviates from the geometric version? (i.e. up to what point of the process do they share the exact same steps, before diverging)

   b. If they are using **different** models, could you give the primary reasoning behind this and also an average of the R-squared goodness-of-fit value which the hedonic regression used in the arithmetic version of the HPI achieves? I know that the geometric version generally fits with an R-squared of a little above 0.8, as per ONS methodology documentation.

8. What is the cap on the number of the habitable rooms for both the England/Wales habitable rooms and Scotland habitable rooms variables which are used in both the geometric and arithmetic version of the HPI? When a property exceeds this cap, you would presumably discard the record, would that be accurate?
9. Would you please share the weight penalties which you currently apply to a property transaction missing:
    a. the habitable room variable
    b. the floor area variable

when fitting the regression?

Thank you for your request.

Question 1:

The arithmetic mean version of the HPI that’s produced for the monthly Retail Price Index depreciation series is calculated using the same semi-log model as the geometric mean version of the HPI.

Questions 2 and 3:

The transformation (via exponentiation) back to linear/level scale is done on the predicted prices in the fixed basket. The aggregated data for the geometric mean version of the HPI is calculated using the following formula for each stratum:

$$\exp\left(\frac{\text{sum of log form of predicted prices}}{\text{number of observations in fixed basket}}\right)$$

This is mathematically equivalent to calculating the geometric mean of the predicted prices.

The aggregated data for the arithmetic mean version of the HPI is calculated using the following formula for each stratum:

$$\frac{\text{sum of predicted prices}}{\text{number of observations in fixed basket}}$$

Questions 4 and 5:

The fixed basket for both the geometric and arithmetic mean versions of the HPI is identical. To take account of the registration lag experienced by our data providers, we use transactions from Quarter 4 (y-2) to Quarter 3 (y-1).

Question 6:

The exact same property attributes are used in the arithmetic mean version of the HPI to the geometric mean version of the HPI.

Question 7:

In order to provide an accurate response to this question, we would need to conduct in depth research and analysis into the relevant codes and methodologies. We are not obligated to conduct such research or create analysis under FOI in order to fulfil requests and so we consider this to be information not held.

Question 8:

The cap on the number of rooms is 8. For any house with greater than 8 rooms, the number of rooms would be set to 8.

Question 9:

When running the hedonic regression model, some properties may be missing one or more of their price determining characteristics – for instance, floor area may not be available. These properties are still used in the regressions but are given less weight in the calculations depending on the importance of the missing variable as a price determinant. For instance, floor area is found to have more of a bearing on price than whether the property is new or old and existing property, so a property with missing floor area will have a lower weight than one missing the new or old existing property indicator.

- In 2021, a transaction that is only missing habitable room would have a weight of 0.853 (rounded to three decimal places).
- In 2021, a transaction that is only missing floor area would have a weight of 0.958 (rounded to three decimal places).