ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

Replatforming the UK House Price Index (UK HPI)

Status: Final
Expected publication: Alongside minutes

Purpose

1. Re-platforming the UK House Price Index (UK HPI) is essential to meet ONS' policy to reduce the use of proprietary tools and legacy systems, improve efficiency and ease of use in production, and thus reduce risk to monthly publication.
2. Re-platforming also provides the opportunity to review methods used in UK HPI production to ensure continued adherence to current best practice.
3. This paper presents the proposed methodology improvements to be undertaken during the re-platforming of the UK HPI. Upon launch, it was proposed that the UK HPI methodology should be reviewed periodically to consider utilisation of newly available data sources and address limitations where possible. Statistics produced by the UK HPI are used to estimate owner occupiers' housing costs within the Retail Price Index (RPI).

Actions

4. Members of the Panel are invited to raise any potential concerns regarding the proposed methodology improvements:
    a. Using an Ordinary Least Squares regression model
    b. Using the natural logarithm of 'floor area' in the regression model
    c. Using two 'floor area' variables in the regression model: one for England and Wales; one for Scotland
    d. Identifying outliers earlier in the pipeline
    e. Moving away from using a legacy imputation routine and using a nearest neighbours imputation or univariate decision tree imputation

Background

5. Currently, the Office for National Statistics (ONS) publishes the UK House Price Index (UK HPI) as a joint publication with HM Land Registry (HMLR), Registers of Scotland (RoS), and Land and Property Services Northern Ireland (LPS).
6. The UK HPI was introduced in 2016, bringing together data sources from across government to improve the official measure of house price inflation. Before this date, two official indices were published, one by the ONS that was based on a sample of mortgage data and the other by HMLR, using a repeat sales methodology that didn't capture new build property.
7. The UK HPI publication includes a monthly index, average price, and annual percentage change for residential property prices in the UK, its countries, English regions, and other administrative geographies. Breakdowns by property type, buyer status (first time buyer or former owner occupier), funding status (cash or mortgage), and property age (new build or existing resold property) are also published.
8. The current UK HPI methodology was developed in 2016 and relies on the legacy coding software SAS.

9. The Office for Statistics Regulation designated the UK HPI as a National Statistic in September 2018.

10. ONS also publishes the related House Price Statistics for Small Areas in England and Wales (HPSSA). This makes use of the same HMLR price paid data used in UK HPI and publishes at greater geographic granularity than the UK HPI, but these estimates are not comparable over time.

11. The objectives of the UK HPI replatforming project are:

    a. To move away from legacy systems, in line with the ONS' policy to reduce the use of proprietary tools, such as SAS. This includes replatforming all housing-related systems that use SAS, including UK HPI, HPSSA and the Quarterly and Annual Tables (which are based on Regulated Mortgage Survey data).

    b. To ensure the UK HPI (and related) production systems meet the requirements of a reproducible analytical pipeline (RAP) to ensure the production of statistics is efficient and sustainable.

    c. To review the existing UK HPI methodology and make improvements where necessary.

    d. To continue to update and review available data sources to identify new variables that are potentially suitable for inclusion in the regression model.

12. ONS methodology experts have been consulted and their recommendations for improvements to the UK HPI methods are outlined in the following section.

Proposed methodology improvements

13. The current UK HPI methodology is a double imputation hedonic index, based on a semi-log regression model. The index is based on the change in the predicted geometric average prices for an annual fixed basket of properties, and is chain linked between years to produce the full time series for publication.

14. Upon review, the ONS methodology experts did not recommend any major methodology changes that would affect the existing methodology at high level. However, several improvements were recommended, which we propose to implement as follows.

15. **Proposal 1: Use an ordinary least squares (OLS) regression model, rather than weighted least squares (WLS) as used in the current model.** This introduces two further differences:

    a. No observation weights need to be calculated, unlike the existing method. This reduces time resource burden on production.

    b. Heteroskedastic robust standard errors need to be calculated

16. This proposal simplifies the regression model being used, improving transparency and understanding for users, while also reducing processing time since the annual observation weights is data intensive and takes significant time to produce. Additionally, it is difficult to assess the quality of the output observation weights. OLS was also recommended, and is being implemented, in the redeveloped rents system (previously presented to the Panel).

17. **Proposal 2: Investigate the suitability of using the natural logarithm of 'floor area' as an explanatory variable in the regression model, rather than the non-transformed 'floor**

area'. Currently, using non-transformed 'floor area' means that the relationship between ln(price) and area is assumed to be linear. In other words, price is assumed to be proportional to the exponential of floor area. This means that even a small difference in floor area would lead to a large difference in the model-predicted price of a property. As floor area increases, this effect becomes exponentially larger. This may lead to increased spread for larger properties and extreme price outliers, driven by only small differences in floor area, which may only be a small percentage difference of the area total. Analysis by Statistics Netherlands showed that ln(area) is much more normally distributed than non-transformed area, and that the transformed area values into natural logs are more useful in the hedonic regression model when it comes to explaining sales price. Using the natural logarithm of area may be expected to improve model fit, especially for larger properties. This will be investigated and this proposal will be considered if evidence supports it. No change is recommended to the other existing explanatory variables: Acorn group, property type, number of rooms, new or existing, and local authority.

18. **Proposal 3: Use two explanatory 'floor area' variables in the regression model (one for England and Wales; one for Scotland), rather than one 'floor area' variable to account for differences in how floor area is measured in different data sources.** The data source for floor area for England and Wales the Valuation Office Agency (VOA) Council Tax data. For Scotland, the source is Energy Performance Certificate (EPC) data.

19. Due to the differing ways of measuring floor area between these data sources ONS methodology experts recommend that, since measurement of area differs for Scotland, two area variables should be used in the regression model: one for England and Wales (from VOA) and one for Scotland (from EPC).

20. The current (a) and proposed (b) regression models are given below:
    a. $\ln(price) = area + acorn\_group + property\_type + num\_rooms \ + new + la\_code$
    b. $\ln(price) = \ln(area_{ew}) + \ln\,(area_{scot}) + acorn\_group + property\_type + num\_rooms \ + new + la\_code$

    Where $area_{ew}$ and $area_{scot}$ are the floor areas for England and Wales (VOA), and Scotland (EPC) properties, respectively.

21. Northern Ireland calculate their own HPI independently to ONS. ONS will engage with Northern Ireland on the proposed methodology changes for Northern Ireland to consider potential application in the Northern Ireland methodology.

22. **Proposal 4: Identify and handle floor area outliers earlier in the system pipeline.** Although the property attributes data used in the UK HPI has been assessed for quality, outliers are present in the data. Each month, property price data is linked to property attributes data before running a regression model to quantify the relationship between explanatory variables and house price, which is used to estimate the price of a fixed basket of properties every month.

23. Currently, outliers are identified at the regression stage only, where properties are identified as outliers if the z score for the property is ≥ 5. It is recommended that, in addition to retaining the outlier quality assurance step after the regression stage, floor area outliers be

identified each month earlier in the pipeline, immediately after the data linking stage. The proposal is to do this by calculating the following ratio for each property:

$$R = \frac{\ln Area}{\ln Price}$$

The z score for each property would be calculated as:

$$Z = \frac{R - \bar{R}}{sd(R)}$$

Where $\bar{R}$ and $sd(R)$ are the mean and standard deviation of R, respectively. Properties for which $|Z| \geq 5$ would be marked as outliers, and their floor area is set to NULL. These NULL values will be imputed in the next stage, before the regression model is run.

24. **Proposal 5: Use multivariate nearest neighbours imputation (e.g. MICE) or a univariate decision tree imputation to populate missing data in the annual fixed basket of properties, rather than CANCEIS.**
25. Currently, the UK HPI uses CANCEIS to populate missing data in the fixed basket only, with separate runs for Scotland and for England and Wales. There is limited knowledge or support for CANCEIS methodology and so it is more difficult to identify and resolve issues. It is recommended to move to an alternative imputation method, such as MICE or utilisation of the imputation method being employed in the rents development system (univariate decision tree) to improve method transparency and reduce risk to publication from difficulties with CANCEIS.
26. It is proposed that this imputation process is used to populate missing data, both in the fixed basket (all explanatory variables) and the monthly dataset upon which the regression is run (for missing floor area only). No change is proposed to running separate imputations for Scotland, and England and Wales data, due to different property attributes data sources and differences between the UK nations.

27. It is proposed that the imputation procedure designed and accepted for use in the new rents system (previously presented to the Panel) be considered for use in the UK HPI. This imputation would use the package scikit-learn within a RAP, rather than a separate process run using CANCEIS.

Other considerations

28. A number of other improvements will be considered as part of the discovery phase, using the lessons learned and investigations already conducted during the redevelopment of private rental price statistics. The new Price Index for Private Rents (PIPR) uses a similar methodology to the UK HPI, and several of the same datasets. Lessons learned from redevelopment of the rents system will be considered during the replatforming of the UK HPI to maximise processing efficiency and safeguard the new UK HPI system for the future. One such example is investigating the use of Unique Property Reference Numbers to link price paid data with property attributes data, rather than matching by address information. This has been explored during development of PIPR.
29. Additionally, data pre-processing requirements will be considered during the Discovery Phase of UK HPI Re-platforming to ensure the data is appropriately prepared for use in the UK HPI system. For example, this includes:

a. Pre-processing data to contain two area columns: one for England and Wales, one for Scotland. This is because floor area has a different data source in these UK nations and they each measure floor area differently. Properties in England and Wales should have a non-zero value in the England and Wales area size column and a null value in the Scotland area size column. Scottish properties should have a non-zero value in the Scotland area size column while England and Wales properties should have a null value.
b. Before running the monthly regression model, all explanatory variables need to be populated. Imputation flags should also be used to indicate which property characteristics have been imputed. Checks should be automatically set up to check for missing or null values in the data.

c. Pre-processing checks should be set up to assess that input data contains all the expected variables and that they are populated as expected (e.g. string, numerical value, etc), and do not contain unrealistic values e.g. a house sale value of £1, which does not reflect market value.

30. Suitable diagnostic outputs will be considered during the Discovery Phase for quality assurance purposes. For example, in addition of monitoring the total number of transactions per month at UK nation level, we may consider monitoring transaction volumes at region level, or a property type breakdown to support quality assurance.

Timelines

31. By the end of Summer 2023, we aim to:

a. Engage with stakeholders (APCP-T, UK HPI Working Group) regarding the proposed methodology changes and gather feedback on any potential concerns about the proposals.

b. Engage with ONS methodology experts to finalise the methodology changes to be implemented in the UK HPI.

c. Present the finalised methodology changes to stakeholders (APCP-T, UK HPI Working Group).

Conclusion

32. We are confident that the five methodology change proposals represent improvements to the UK HPI, having consulted with methodology experts within ONS.

33. We invite the Panel to raise any potential concerns they may have with the above proposals, particularly regarding proposals 2 and 5.

Prices Division, Office for National Statistics
April 2023