ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

UK House Price Index (UK HPI) monthly imputation methods

Status: Final
Expected publication: Alongside minutes

Purpose

1. The UK House Price Index (UK HPI) is an Accredited Official Statistic and UK HPI statistics are used to estimate owner occupiers' housing costs within the Retail Price Index (RPI). For use in RPI, the current month's average house price is estimated by using the Nationwide's index to forecast forward the UK HPI first estimate (which is on a 1-month lag).
2. The UK HPI provisional early estimates for new builds tend to be initially over-estimated, leading to downward revisions in the new build and headline UK HPI estimates.
3. The April 2023 paper APCP-T(23)03 "Replatforming the UK House Price Index (UK HPI)" outlined several potential methodology improvements to UK HPI which APCP-T provided feedback on. This included recommending changes to the imputation approach in the UK HPI for handling missingness in price-determining variables, which tends to be higher for new build properties than for existing properties.
4. ONS has conducted preliminary exploratory work on the impact of improving the imputation approach currently used in the monthly UK HPI production for transaction data entering the monthly regression model.
5. Preliminary work suggests that improving the monthly imputation in UK HPI may reduce the size of revisions to the new build series and hence reduce revisions to UK HPI headline statistics.
6. ONS proposes to improve the imputation approach for the monthly sales data from a basic imputation (set missing "floor area" to zero) to imputing a non-zero value. Also to impute the "number of rooms" variable (currently "missing indicator" is used) due to the missingness rate correlation with "floor area", which comes from the same data source.
7. ONS intends to follow a two-step plan:
    a. Stage 1: By mid-2025, investigate and implement an improved monthly imputation (median or k-nearest neighbour), to increase the accuracy of UK HPI provisional estimates used in RPI production and to reduce subsequent revisions to UK HPI provisional estimates. This is intended for the "floor area" and "number of rooms" variables, but ONS will also explore applying this imputation to the other price-determining variables.
    b. Stage 2: Follow with a comprehensive imputation review within a wider UK HPI methods review.
8. This paper focuses on Stage 1 of this plan.

Actions

9. Members of the Panel are invited to provide feedback on:
    a. Does the Panel support ONS' proposal to improve the monthly imputation in the UK HPI, to use a median or k-nearest neighbour imputation?
    b. Does the Panel support ONS' two-step plan?

c. Feedback on which imputation methods to consider in a later imputation review (e.g. CANCEIS, which is used in the HPI annual imputation, or univariate decision tree, which is used in ONS' rents imputation).

Background

10. The UK HPI tends to observe downwards revisions to its provisional estimates, driven by downwards revisions in estimates for new builds.
11. Revisions to UK HPI estimates (particularly for new builds) are mainly driven by two sources:
    a. Time needed for sales transactions data to become available for inclusion in the UK HPI, with new builds taking much longer to be available. It can typically take 6-8 weeks for a transaction to be registered with the Land Registry for processing before becoming available for use in the UK HPI, which can increase significantly for new build property.
    b. Time needed for property attributes information to be updated, records for new builds generally taking longer to be created and added to the data (leading to higher missingness rates in early provisional estimates), than records for existing properties to be updated.
12. Under the current imputation method (particularly for the continuous variable "floor area"), ONS believes that the high missingness rate of England and Wales' new builds' property attributes in the months immediately following a new build sale leads to over-estimation of Great Britain's new build prices in early provisional estimates. This paper proposes an approach aimed at reducing revisions arising from this second source.
13. The April 2023 paper APCP-T(23)03 "Replatforming the UK House Price Index (UK HPI)" proposed methodology changes to the imputation approach in the UK HPI:
    a. Proposal 4 described a methodology change involving imputation of missing ("NULL") floor area in the data used for the monthly regression model.
    b. Proposal 5 suggested exploring alternatives to CANCEIS to use for imputation of missing values in the UK HPI.
14. The UK HPI Replatforming project launched in 2024 and ONS has conducted preliminary exploratory work on the impact of improving the imputation approach currently used in the monthly UK HPI production for transaction data entering the monthly regression model.
15. This paper outlines a proposal to improve this basic imputation routine, in line with the recommendations presented to APCP-T in April 2023, and supported by a theoretical explanation and indicative analysis.

Current situation

16. In the current HPI system for Great Britain, each April a fixed basket of properties is created using the previous year's property transactions data. CANCEIS (nearest neighbour imputation) is used to populate missing values in the fixed basket for each of the price-determining independent variables used in the monthly hedonic regression model (number of rooms, property type, local authority, ACORN group, floor area and new build flag). All properties in Scotland have floor area set to zero in the fixed basket, due to differences in measurement between Scotland and England & Wales.
17. Monthly, price paid transaction data for property sales in Great Britain is linked to property attributes data. Where values are unknown, "missing indicator" imputation is used for

categorical variables, while missing values on the continuous "floor area" variable are set to zero. For Scotland, floor area is set to zero for all properties, as in the fixed basket.

18. Focusing on the floor area and number of rooms variables (both of which have the highest missingness rate and are usually either both present or both missing), this means that the hedonic regression model is run on monthly data with "missing" present for rooms, and a floor area of 0 for some properties.
19. In the fixed basket, all missing values have been imputed so there are no properties with "missing" or "unknown" values (except for Scotland properties' floor area, which is set to zero). This means that while Scotland's "floor area" characteristics are consistent (both set to zero) in both the fixed basket and the monthly regression model data, England and Wales' "floor area" has different characteristics in these datasets, so the Great Britain regression model is being run on data with different characteristics than the fixed basket that the fit is applied to.
20. New build properties are more likely to have missing property characteristics than existing properties, so new builds are more likely to be impacted. In the recent months following a new build sale, that property is less likely to exist in the property attributes data, leading to a higher missingness rate for some characteristics. As months pass, new builds are added to the attributes data and the England & Wales missingness rate decreases significantly. The behaviour of missing property characteristics having a bias for new build properties is believed to be introducing an upwards bias into the UK HPI provisional early estimates for new builds.
21. Price tends to be higher for new builds than for existing builds of a similar type. This is known as a 'new build premium' and is measured using the 'new build flag' coefficient in the HPI regression model.
22. Where a new build property has missing floor area, the model will observe the relatively-high price for that new build property transaction (with missing floor area set to 0) compared with lower prices being paid for similar existing builds (with populated floor area for England and Wales, 0 for Scotland).
23. Some of the England and Wales new build property price which should have been attributed to the (missing) floor area size will instead be misattributed to the 'new build flag' regression coefficient.
24. Since new build transactions (with missings) account for only a small proportion of total transactions, this will have a large inflationary effect on the 'new build flag' coefficient and minimal effect on the floor area coefficient.
25. When the monthly model fit is applied to the Great Britain fixed basket of properties, this leads to over-estimation of predicted price for new builds in our provisional estimates. For example, a new build property in England and Wales in the fixed basket will have a contribution to predicted price from both floor area (as all missing values are imputed in the fixed basket) and from the inflated 'new build flag' coefficient, leading to over-estimation of predicted price for that new build.
26. For a new build property in Scotland in the fixed basket, although floor area will still be 0 (as in the monthly sales regression data), the inflated 'new build flag' coefficient will still lead to over-estimation of predicted price for that new build since the regression model is run for Great Britain as a whole. This means that the over-estimation of new build price (arising from different characteristics of the England & Wales data in the fixed basket compared with

the monthly data for the regression model) will influence Scotland's new build price estimates as well as England and Wales' estimates.

27. Property attributes data is updated each month, with England & Wales new build property information being increasingly available over time (illustrated by Figure 1). This means the missingness rates of property attributes (especially "floor area" and "number of rooms") in new build properties reduces over time, and so the regression model's over-estimation of the 'new build flag' coefficient also reduces as UK HPI estimates are revised in subsequent months. This leads to downward revisions in the predicted prices for new builds.

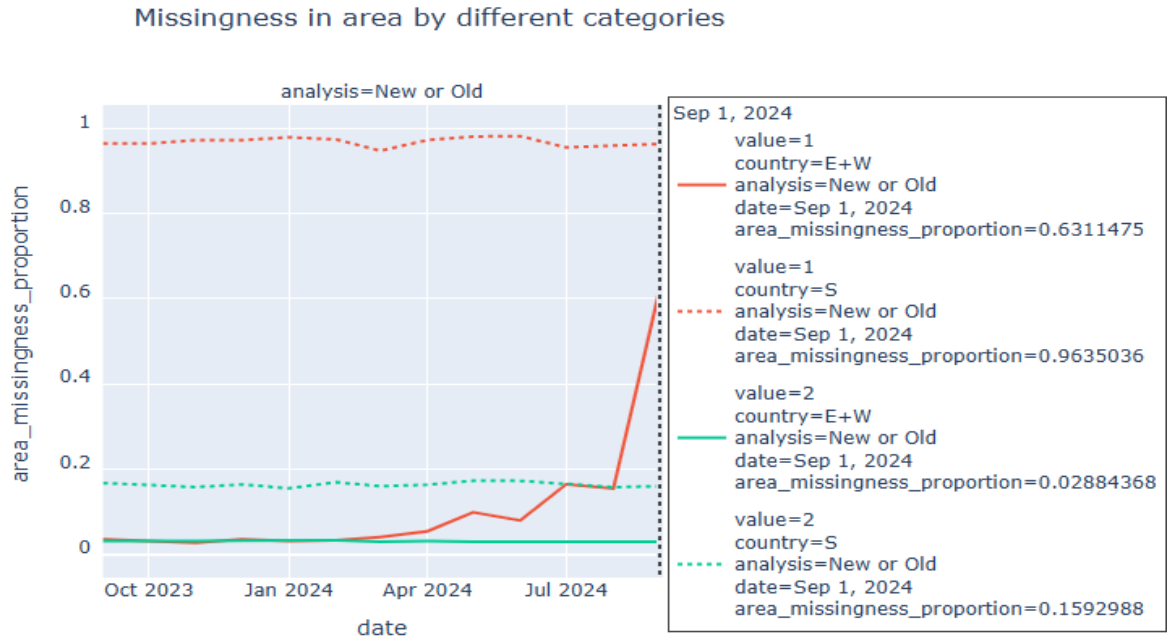

Figure 1: Proportion of properties with missing values for floor area, for England & Wales (E+W) and Scotland (S), for new builds (value=1) and existing builds (value=2). Data is provisional. Note that the Scotland missingness rates shown here are for the "floor area" raw data in Scottish Government's Energy Performance Certificate data, but floor area is set to zero for all Scotland properties in the current methodology so only the E+W solid lines reflect missingness rate observed by the model.

28. Although complicated by increasing transaction data volumes over time (as more new build data is processed each month and is incorporated into the UK HPI), this rationale fits with the published and revised UK HPI data, where the new build estimates tend to be revised downwards over time.

Stage 1: Proposed change and preliminary analysis

29. ONS proposes that the monthly imputation method be improved, for at least the variables "floor area" and "number of rooms", which are most impacted by this bias. The proposed improvement will reduce the differences in characteristics between the fixed basket data and monthly data for the regression model, reducing this upwards bias for new build prices.

Variable	Current monthly imputation	Proposed monthly imputation
Floor area	For Scotland, floor area (from EPC data) is set/imputed to zero for all properties.	For Scotland, floor area (from EPC data) is set/imputed to zero for all properties.

	For England & Wales, non-missing floor area data (from VOA data) is used and missing values for floor area are set/imputed to zero.	For England & Wales, non-missing floor area data (from VOA data) is used and missing values for floor area are imputed to a non-zero value using England & Wales properties for the imputation.
Number of rooms	For Scotland, non-missing number of rooms (from EPC data) is used, and missing values for number of rooms are set/imputed to a 'missing' category. For England & Wales, non-missing number of rooms (from VOA data) is used, and missing values for number of rooms are set to a 'missing' category.	For Scotland, non-missing number of rooms (from EPC data) is used, and missing values for number of rooms are imputed using Scotland properties for the imputation. For England & Wales, non-missing number of rooms (from VOA data) is used, and missing values for number of rooms are imputed using England & Wales properties for the imputation.

30. For stage 1 of the plan, ONS has conducted preliminary indicative analysis, intending to implement one of the following imputation methods on the monthly data:
    a. using the median (for continuous variables) and the mode (for categorical variables) to impute missings for each group (donors with matching local authority and new/old flag);
    b. using a k-nearest neighbours imputation, with k = 5, for each group (considers all other price-determining variables).
31. A k nearest-neighbours routine for monthly imputation was explored to indicate the impact of implementing the CANCEIS nearest neighbour routine already used for UK HPI annual imputation. However, these two approaches gave extremely similar results, such that the two outputs would not be visibly different in the charts below, so only results for the median/mode imputation approach (for "floor area" and "number of rooms") are displayed.
32. The analysis showed that:
    a. Average new build price is consistently lower if imputation is applied, compared with the current "no imputation" approach.
    b. Over-estimation for new builds is greater in the early provisional estimates (demonstrated by a larger gap between the blue and red lines), and reduces as older periods are revised.
33. This observed behaviour agrees with the theory presented above and demonstrates that improving the monthly imputation in UK HPI is likely to improve early provisional new build estimates and hence increase the quality of early provisional UK HPI estimates, which are used in RPI production.

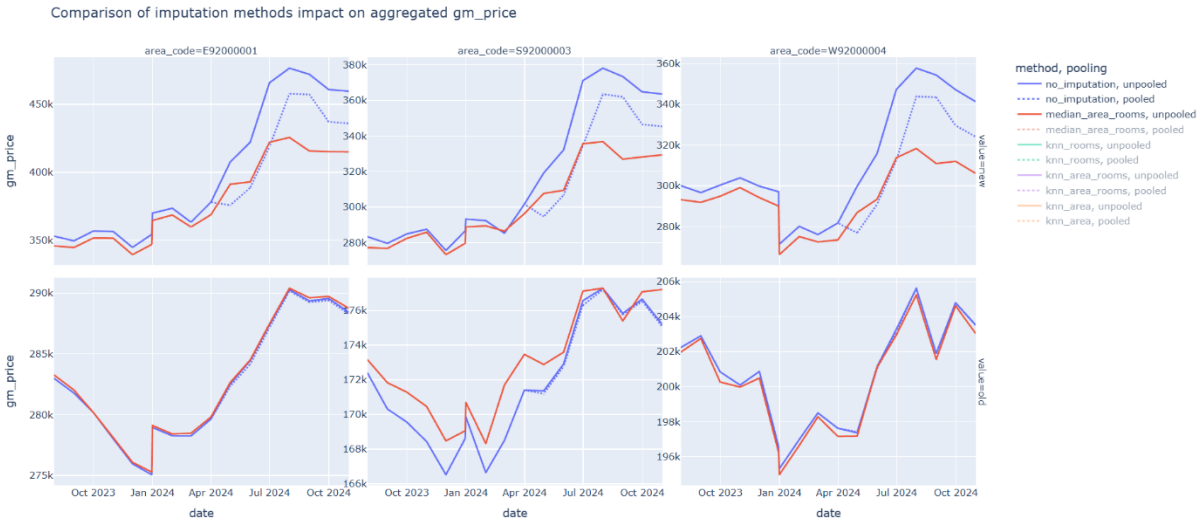

Figure 2: Geometric mean price (unchained, hence the discontinuity in January), for England, Scotland and Wales, Nov-23 to Nov-24, using a test system (not output directly from the UK HPI production system) to give indicative results. Top row represents new build price, bottom row represents existing 'old' build price. Data is provisional.

34. Figure 2 also illustrates that the over-estimation bias for new build price is mitigated better by applying imputation for missings, than by the current "pooling" approach.[1] Pooling generally reduces early provisional UK HPI estimates, with greatest impact on the 'oldest' estimates (where new build volumes are higher). In contrast, imputation has greatest impact on the 'youngest' estimates (where missingness rates are higher).
35. Therefore, 'imputation' is more effective than 'pooling' at reducing new build price over-estimation in the provisional 1st estimate, and hence is better at reducing revision size between the 1st and 13th (final) UK HPI estimates. The temporary "pooling" measure within the UK HPI is reviewed regularly and its implementation status will be reviewed following improvements to the monthly imputation.

[1] Since 2020, new build transactions for England and Wales have been pooled with the previous month's new build transactions for some months early in the UK HPI revision period. This increases new build data volumes and generally reduces the predicted price for new builds, reducing revision size and improving the accuracy of provisional UK HPI estimates. Currently, new build "pooling" is applied to the 2nd to 7th provisional UK HPI estimates. For example, in the latest UK HPI data up to Nov-24, new builds data was pooled for May-24 to Oct-24. "Pooling" is intended to be a temporary measure until provisional data volumes for new build transactions return to near historical levels. More information is published in the UK House Price Index reports page on GOV.UK

Figure 3: Geometric mean price (unchained), for England, Scotland and Wales, Nov-23 to Nov-24, using a test system (not output directly from the UK HPI production system) to give indicative results. Top row represents new build price, bottom row represents existing 'old' builds price. An "end_month" of 11 represents results from data available "as at UK HPI's publication for data up to Nov-24". Data is provisional.

36. Figure 3 provides indications of the effect on UK HPI revisions.
    a. The blue series represents UK HPI estimates with "floor area" and "number of rooms" imputation, and no pooling applied to new builds data.
    b. The green series represents UK HPI estimates produced under the current system (i.e. with no imputation and with pooling applied to new builds data for early estimates).
37. Figure 3 indicates that improving the monthly imputation in UK HPI may reduce the size of revisions to the new build series and hence reduce revisions to UK HPI headline statistics. These improvements could also reduce the need for pooling new build data for early estimates.
38. Prices Division has consulted ONS' internal Methodology and Quality Division (MQD) for advice. MQD have supported this proposed change and recommend that an imputation method that is both reliable and coherent with other methods used in HPI should be investigated and implemented.

Stage 2: Intentions for the next HPI methods review

39. For stage 2 of the plan, ONS intends to investigate other imputation methods in a later methods review, including:
    a. Using CANCEIS (the CANadian Census Edit and Imputation System, a k-nearest neighbour imputation (k=10) currently used in the UK HPI to impute missing values in the annual fixed basket) to also impute missing values in the monthly data for the regression model.

b. Approaches such as a multivariate nearest neighbours routine (e.g. MICE) or the univariate decision tree imputation used in ONS' Price Index of Private Rents (scikit-learn package in Python).

40. This intention was presented to APCP-T in the April 2023 paper, and APCP-T gave their support in April 2023 for this investigation.

Conclusion

41. ONS proposes improving the UK HPI monthly imputation for price-determining variables, at least for the 'floor area' and 'number of rooms' variables due to missingness in these characteristics having greatest potential to influence published UK HPI estimates.
42. ONS proposes to implement one of the two imputation approaches outlined above, in the UK HPI in 2025, subject to completion of further analysis and approval from the UK HPI Working Group.
43. Early evidence suggests that improving the monthly imputation in UK HPI is likely to reduce over-estimation in early provisional estimates for new build predicted prices, and hence improve the quality of provisional UK HPI estimates.
44. Since the RPI uses the provisional 1st estimates from UK HPI in its monthly calculation, commitment to accuracy and quality of RPI estimates is a key driver of this investigation, and therefore relevant for the Panel. However, it should also be noted that:
    a. The over-estimation primarily affects the provisional estimates for the new build breakdown, and (due to the small weight of new builds) the impact is much smaller on the Great Britain-level headline estimates used in RPI production.
    b. The provisional UK HPI 1st estimates used in the RPI are forecast forward a month using the Nationwide's index to account for the one month lag in UK HPI, reflecting the use of actual transaction prices.
    c. In monthly RPI production, the inflation rate is used, not the price level, reducing the impact of price-level overestimation on RPI calculations.
    d. In annual RPI production, UK HPI price level estimates are used, but RPI uses both revised and provisional estimates. Since new build price overestimation reduces significantly in early revisions, this further reduces the impact on RPI calculations.
45. Prices Division has support from ONS' Methodology experts on this proposal and now seeks feedback from the Panel on:
    a. The proposal to improve the monthly imputation in the UK HPI, using a median/mode or k-nearest neighbour imputation;
    b. The intention to implement a two-part plan;
    c. Feedback on which other imputation methods to consider in a later imputation review (e.g. CANCEIS, which is used in the HPI annual imputation, or univariate decision tree, which is used in ONS' rents imputation).

Aimee North
Head of Housing Market Indices
Prices Division, Office for National Statistics
January 2025