ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

UK House Price Index (UK HPI) monthly imputation improvement

Status: Final
Expected publication: Alongside minutes

Summary

1. The UK House Price Index (UK HPI) is a critical statistical output that tracks changes in residential property prices across the United Kingdom. Due to the complex nature of property transactions and delays in registration, revisions are an integral part of the monthly publication cycle.
2. The January 2025 paper APCP-T(25)01 "UK HPI monthly imputation methods" proposed improving HPI's monthly imputation for Great Britain (GB) and presented preliminary analysis indicating that improvements to the monthly imputation could reduce the over-estimation of new build prices observed in provisional early UK HPI estimates.
3. Following APCP-T and the UK HPI Working Group's support for this proposed improvement, ONS has considered APCP-T's January 2025 feedback and conducted further analysis, investigating several imputation options.
4. The January 2025 paper APCP-T(25)01 "UK HPI monthly imputation methods" outlined ONS' two-step plan, of which Stage 1 was to investigate and implement an improved monthly imputation in the UK HPI by mid-2025.
5. This paper presents the outcome of Stage 1, presenting ONS' recommended imputation improvement and impact analysis, which demonstrates the value that improving imputation has on the accuracy of provisional UK HPI estimates and reduction in revision size.
6. The analysis indicates that implementation of improved imputation (K nearest-neighbour imputation using one-hot-key embedding and postcode coordinates), to the number of rooms (Great Britain) and floor area (England and Wales) variables in the Great Britain HPI model will:
    a. Improve the quality of provisional UK HPI estimates by reducing over-estimation in early provisional estimates of new build price in Great Britain;
    b. Significantly reduce revisions in the new build and headline UK HPI estimates;
    c. Enable discontinuation of the temporary 'new build pooling' approach introduced during the COVID-19 pandemic;
    d. Address much of the uncertainty in new build price estimates in the current methodology, by targeting and mitigating the impact of one of the causes: higher missingness rates for new build property characteristics in early provisional estimates.

Actions

7. Members of the Panel are invited to provide feedback on:
    a. ONS' selected model for improving the monthly imputation in the UK HPI: imputing floor area and number of rooms using a K nearest-neighbour imputation (K = 10) with one-hot-key embedding and postcode coordinates to optimise donor selection.

   b. ONS' intention to simultaneously discontinue the temporary 'new build pooling' approach introduced during the COVID-19 pandemic, due to this improved imputation method being more effective than this temporary approach at improving the quality of early provisional estimates.

8. Prices division previously consulted its Methodology experts who identified several areas for potential methodology improvement in the UK HPI. The April 2023 paper APCP-T(23)03 "Replatforming the UK House Price Index (UK HPI)" outlined these, which included Methodology's recommendation to explore alternative imputation methods to CanCEIS in UK HPI's existing imputation process. Consultation with ONS Methodology experts in 2025 reiterated their support for exploring alternatives to CanCEIS for imputation in the UK HPI, such as K nearest-neighbours. The Panel is invited to comment on:
   a. ONS' plan to review the annual imputation method in UK HPI and consider replacing CanCEIS imputation, investigating implementation of the same imputation method proposed in this paper, amongst other options.

Background

9. The UK House Price Index (UK HPI) is used to measure mortgage interest payments and depreciation within the Retail Price Index (RPI). For RPI purposes, the current month's average house price is estimated using Nationwide's index to project the UK HPI's first estimate (which is on a one-month lag) forward to the current month.
10. Therefore, improvements to the accuracy of UK HPI's provisional estimates, used in RPI production, will improve the quality of RPI estimates. Impact on RPI is explained in point 44 of the January 2025 paper.
11. The UK HPI has a 12-month revision period policy, so each new monthly release includes revisions to data from the preceding 12 months. These revisions apply to average property prices, index values, annual and monthly inflation rates and sales volumes.
12. In normal monthly production, revisions are primarily due to the inherent time lag in data availability for:
    a. **Property sales being registered and processed.** A property transaction can take several months to be registered and then processed to become available for inclusion in UK HPI calculations, with new build transactions often taking longer due to increased complexity. Our target is for the first provisional UK HPI estimate for a given month to be based on 40% of total transactions for Great Britain that will ultimately get registered for that month, rising each month until the end of the 12-month revision period. Additionally, previous investigation found that, in England and Wales, more expensive properties tend to be registered quicker, leading to downwards revisions to provisional price estimates in the UK HPI.[1]
    b. **Property attributes data, particularly for new builds.** Records for new builds generally take longer to become available in property attributes data for England and Wales, than existing properties. New build properties also have a higher missing rate of property attributes for the provisional estimates.
13. Otherwise, the most common source of revisions in the UK HPI are from implementation of methodology improvements (which may be applied to historical data), application of the 5-

[1] https://www.gov.uk/government/publications/about-the-uk-house-price-index/quality-and-methodology

yearly re-referencing process, revisions to historical price input data (e.g. from delayed registration of sales price data), or correction of historical errors.

14. It was previously identified that new build prices tend to be over-estimated in early provisional UK HPI estimates.[2] Several strategies are currently used to improve the accuracy of UK HPI's early provisional estimates, attempting to optimise provisional estimates and mitigate against observed over-estimation of new build prices in early provisional estimates:
    a. Temporary **pooling** of new build transactions with the previous months' new build transactions to boost the number of new build transactions available for the regression model for a given month, for early provisional estimates (2nd to 7th estimate only). This temporary pooling approach was introduced during the COVID-19 pandemic for England and Wales only, and is under monthly review intending to be discontinued when possible.[3]
    b. Use of a **dampening factor** (applied to predicted prices for new builds in the 2nd estimate only) was introduced in 2017 to reduce the over-estimation of new build predicted prices in early provisional estimates (2nd estimate only).[4]
    c. Use of a **reduced estimation model** for the 1st estimate; using the ratio of the 1st estimate of predicted price for month m and m-1 (from the reduced regression model), multiplied by the 2nd estimate of predicted price (from the full regression model) for month m-1 (1st estimate only).[5]
    d. A price **ratio** is calculated to cap the monthly inflation in the predicted price for a given property in the fixed basket at +50% or -34% to mitigate against high monthly inflation volatility at an individual property level, applied to the most recent monthly inflation (1st estimate only).
15. Missingness rates tend to be highest when calculating early provisional UK HPI estimates (especially for new builds) and decrease over time. The January 2025 paper explained how missing property attributes data (and initial low data volumes) for GB new builds drives over-estimation of early price estimates for new builds, resulting in subsequent downward revisions in both the new build and overall UK HPI estimates.

Current situation

16. Monthly transactions price data are linked to property attributes data. If values are missing, categorical variables are set to a "missing" label, and missing floor area is set to zero. The regression model uses this monthly data, thus, the regression model is run on data with missingness.
17. In the fixed basket, all missing values are imputed using CanCEIS (nearest-neighbour hot-deck imputation), so all categorical variables are populated and all England and Wales properties have a non-zero floor area value.[6]

[2] https://www.gov.uk/government/publications/about-the-uk-house-price-index/quality-and-methodology
[3] https://www.gov.uk/government/collections/uk-house-price-index-reports
[4] https://www.gov.uk/government/publications/about-the-uk-house-price-index/quality-and-methodology#how-we-present-the-data
[5] Equation 2 in https://www.gov.uk/government/publications/about-the-uk-house-price-index/quality-and-methodology#how-we-present-the-data
[6] Scotland properties have floor area set to zero in both the fixed basket and monthly data for the regression model

18. This means the model is applied to data with different characteristics than the data it was fitted on.
19. If a new build property in England or Wales has a floor area of 0 in the monthly data, then the model will attribute some of its price contribution from its true floor area size to the other property characteristics. Since floor area is more likely to be missing for new builds (and new builds are a small fraction of all transactions), the floor area price contribution will be attributed mostly to the new build flag coefficient, with little effect on other regression coefficients.
20. When predicting the price of a new build in England and Wales in the fixed basket, the price contribution of the new build characteristic (inflated due to the floor area being set to 0 in the monthly data) will be added, but the floor area is now non-zero so the price contribution from floor area will also be added. This causes over-estimation of new build price.
21. Although Scotland properties' floor area is set to zero in both the fixed basket and monthly regression data, the GB model means that the over-inflated new build regression coefficient is applied to Scotland properties, affecting Scotland new build prices too.
22. Therefore, steps taken to reduce over-inflation of the new build regression coefficient will have a positive impact on over-estimation of new build price in Scotland as well as England and Wales.

Variables for which to improve imputation

23. Currently, CanCEIS (nearest-neighbour hot-deck imputation, with "stage control" set to a pool of 10 donors from which a donor record is randomly selected) is used in UK HPI's annual imputation process for Great Britain. In the monthly process, a 'missing indicator' approach is used for categorical variables and floor area is set to zero for missing values.
24. The January 2025 APCP-T(25)01 paper reported ONS' intention and justification for focusing on the floor area and number of rooms variables for improvements to imputation:
    a. These variables come from linked property attributes data and have much higher missingness rates than any other variable, and missingness is not at random which leads to the observed over-estimation of new build price.
    b. Floor area and number of rooms have almost perfectly correlated missingness. If floor area is missing, number of rooms is almost always missing as well.
    c. Missingness rates for these variables in Great Britain data tends to be highest for early provisional HPI estimates, but rapidly reduce as the months pass, so improvements to imputation has greatest benefit on early provisional estimates, with impact successively reducing for later estimates as floor area and number of rooms data becomes available in the attributes data.
25. In January 2025, APCP-T supported ONS' proposal for improving monthly imputation methods and focusing on floor area and number of rooms, and expressed preference for K nearest-neighbour (KNN) imputation.
26. KNN relies on distance metrics, such as Euclidean distance, to identify nearest neighbours. These metrics are naturally suited for continuous numerical data. Applying them directly to categorical variables can be problematic because it requires converting categories into

numerical values, which can introduce arbitrary relationships that do not reflect the true nature of the data.[7,8,9,10,11]

27. The National Statistics Postcode Lookup (NSPL) is used in the UK HPI annual and monthly process to assign property transactions to the corresponding local authority. This NSPL dataset also contains property coordinates which represent the centroid location of each postcode. For this analysis, ONS has utilised this additional location data to improve the identification of 'nearby' donors for imputation. ONS has also applied one-hot-key encoding (see Annex A for details).
28. ONS has considered several imputation approaches (see Figure 1) and proposes to improve the current monthly imputation approach by implementing ***K nearest-neighbour imputation using one-hot-key embedding and postcode coordinates*** to impute missing values of area and rooms (with K=10 based on the existing annual imputation set up of selecting a donor from a pool of 10 nearest neighbours).
    a. This is in line with APCP-T's feedback in January 2025, supported by a theoretical explanation and data analysis.
    b. This is more similar to HPI's existing annual imputation approach than other methods considered (see below).
    c. This targets the two variables with the highest missingness rate (~90% of missing values are found in these two variables).
    d. Analysis shows that implementation of this recommended imputation improvement should reduce the magnitude of revisions in GB new build price significantly.
29. We also considered proposing changes for the other categorical variables used in the Great Britain HPI model. We concluded to retain the 'missing indicator' approach currently used for the other categorical variables because:
    a. Floor area and number of rooms have much higher missingness rates than the other variables, so imputation for the other variables will have a much smaller impact for the other variables;
    b. Acorn captures socioeconomic groupings. Imputing this categorical variable may result in assigning groups that are not meaningful or likely within the context of the data, thereby compromising the statistical integrity of the categorical information.

[7] A new approach to K-nearest neighbors distance metrics on sovereign country credit rating
[8] Scaling and Normalization: Standardizing Numerical Data
[9] Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications - Journal of Big Data
[10] Ordinal and One-Hot Encodings for Categorical Data
[11] A comparison of imputation methods for categorical data

Investigation of imputation methods

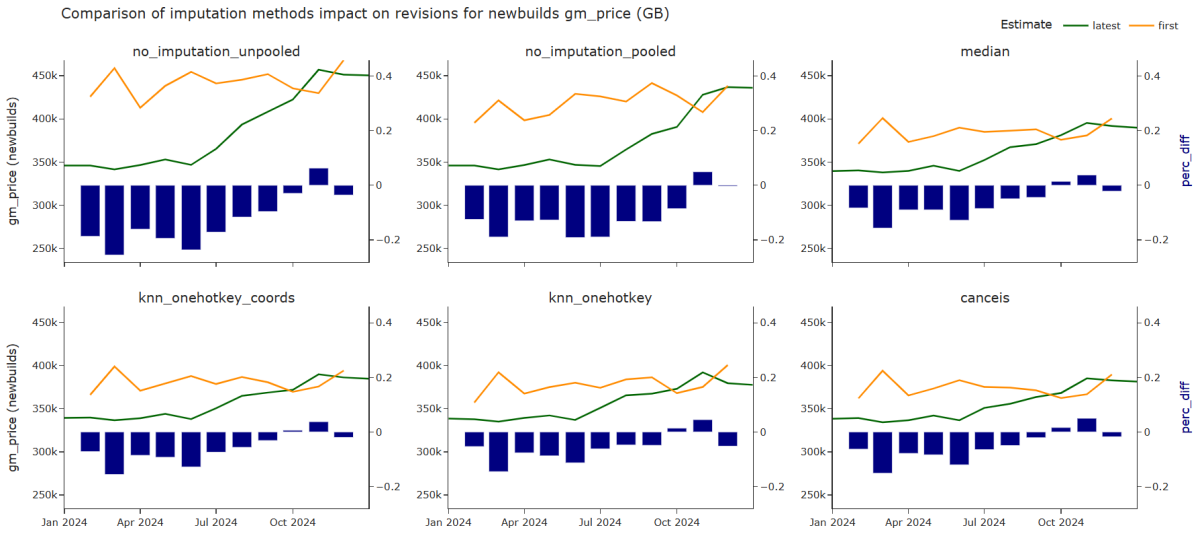

Figure 1: Average new build price for Great Britain, first estimate compared with the latest (revised) estimate for each month, by imputation method

Footnotes:

1. *The top left trace shows the latest price estimate for Jul-24 is ~17% lower than the 1st estimate for Jul-24. This is represented by the dark blue bar, showing -0.17 on the y-axis.*
2. *The percentage difference between the first and latest price estimate for a given month, averaged over all months, was used as a metric assessing the size of revisions to new build GB price.*
3. *The traces are ordered from highest to lowest, by "average price percentage difference between first and latest estimates", so the top left trace has the largest average price percentage difference between the first and latest estimates, while the bottom right trace has the smallest.*
4. *Price data in this figure have been produced by a test system, not by the HPI system used in normal monthly production, and reflect the within-year unchained geometric mean price for Great Britain new builds. In contrast, current UK HPI price estimates are chained geometric mean prices, using January 2023 as the reference period. Before February 2025, the reference period was January 2015. This means the price levels in this Figure are not directly comparable with the published UK HPI price levels for Great Britain new build prices.*

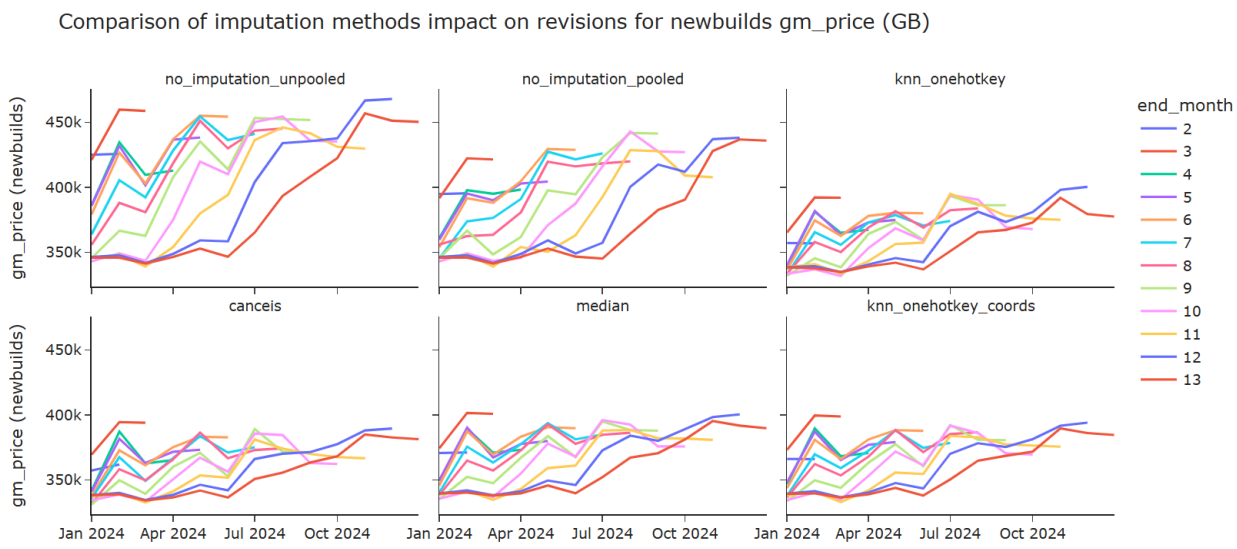

Figure 2: Average new build price for Great Britain, successive estimates, by imputation method

Footnotes:

1. *Each trace shows the price predicted for each month, with each line showing the price estimates at that point in time, labelled by the most recent month available at that point. E.g. The line with end_month = 7 shows the price estimates reported at the time when the 1st estimate for July 2024 was available (2nd estimate for June 2024, etc). The line end_month = 13 represents the data for where January 2025 was the 1st estimate.*
2. *The traces are ordered from highest to lowest, by "average price difference between successive monthly estimates, over all months", so the top left trace has the largest average price difference month to month, while the bottom right trace has the smallest. This price percentage difference should be as low as possible for every end month, to reflect smaller differences between successive monthly estimates and price is not extremely varying or being overestimated.*
3. *Price data in this figure have been produced by a test system, not by the HPI system used in normal monthly production, and reflect the within-year unchained geometric mean price for Great Britain new builds. In contrast, current UK HPI price estimates are chained geometric mean prices, using January 2023 as the reference period. Before February 2025, the reference period was January 2015. This means the price levels in this Figure are not directly comparable with the published UK HPI price levels for Great Britain new build prices.*

30. Techniques KNN and median were investigated, with a CanCEIS comparator included. We also examined the relative impact of the temporary 'new build pooling' approach. KNN imputation with and without the use of NSPL coordinates (as opposed to local authority to determine 'nearby' donors) were also investigated.
31. The charts labelled "no_imputation_pooled" represent the approach UK HPI currently uses for publication. Comparing "unpooled" with "pooled" scenarios:
    a. Figure 1 shows that the mean price percentage difference between the first and latest estimates is slightly higher for the "unpooled" scenario than "pooled" scenario, but the difference is within the error bars, showing that 'pooling' has little effect on reducing revisions from 1st to last estimate.

   b. Figure 2 shows that, while 'pooling' generally reduces the over-estimation of early provisional prices for new builds slightly, compared with 'unpooled', its effect is greatest for the oldest month pooling is applied for (when volumes are highest within the six-month pooling period) and has its effect on reducing over-estimation in early provisional estimates is much smaller than for the improved imputation scenarios investigated.

32. Due to its limited impact on reducing early over-estimation, the remaining scenarios were conducted without using the temporary 'new build pooling' measure introduced during the COVID-19 pandemic, to explore the relative impact of improving imputation methods compared with the temporary 'pooling' measure. The UK HPI reports inform users that the temporary 'pooling' measure is under continuous review, aiming to return UK HPI calculations to the standard "no pooling" approach, which is the standard UK HPI method described in the Quality and Methodology Information.
33. Although the UK HPI regression model uses local authority as the categorical location variable, postcode centroid coordinates are also available in the National Summary Postcode Lookup (NSPL) dataset. Due to KNN's affinity for continuous variables, and the increased granularity of postcode coordinates, imputation scenarios were tested using NSPL coordinates (in addition to local authority and the embedding and similarity metrics) to identify 'nearest neighbours'.
    a. Figure 1 shows the mean percentage difference between the first and latest estimates was similar for the "knn_onehotkey_coords" scenario (KNN using postcode coordinates) than the "knn_onehotkey" scenario (KNN without using postcode coordinates).
    b. Figure 2 shows the mean price difference over all months (not just the first vs latest estimate) was smaller for KNN using postcode coordinates than for KNN without using postcode coordinates. Use of postcode centroid coordinates was more effective at reducing revisions.
34. Next, scenarios of KNN, median and CanCEIS were compared using the same metrics. The average percentage difference between the first and latest estimate was similar (within uncertainty range) for all three scenarios, while the average price difference over all months was slightly lower for KNN with postcode coordinates than for the other scenarios.
35. In all scenarios, the revision metrics were around half the size of those for the 'pooling' and 'no pooling' approaches using the current basic imputation approach. This demonstrates that significant benefit to revisions is realised from improvements to monthly imputation, independent of which of these three methods is selected.
36. KNN produced the smallest downward revision in price over all months. This (combined with the theoretical discussion outlined above and in the January 2025 paper, APCP-T's previous feedback, and ONS Methodology's encouragement for KNN), meant KNN with use of postcode coordinates became the preferred approach.

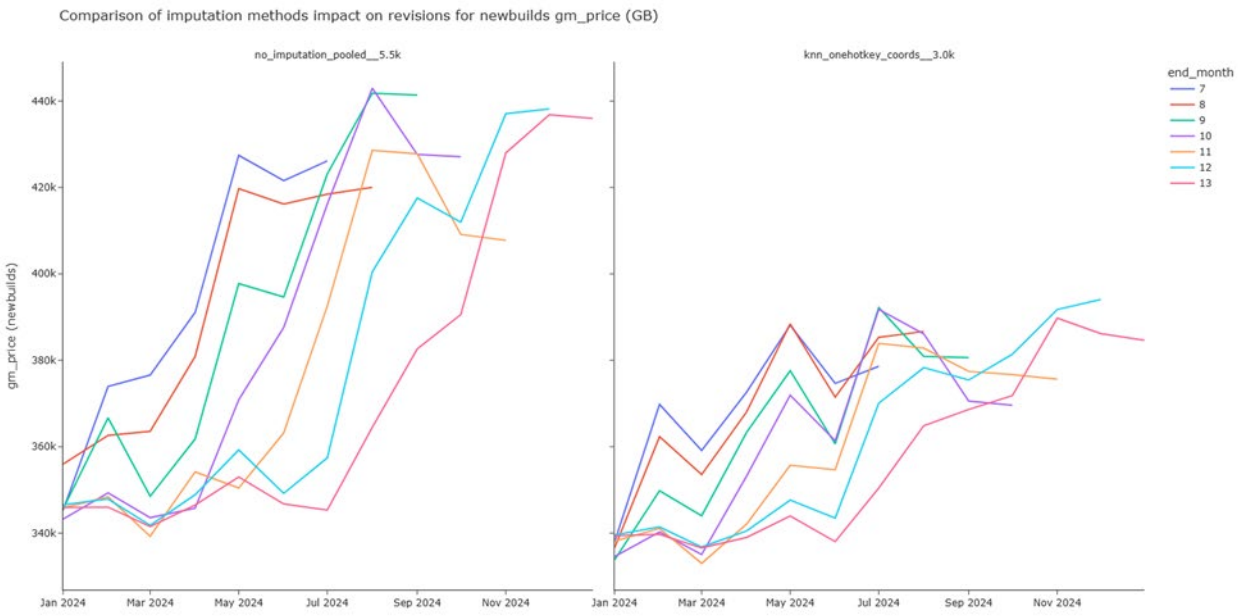

Figure 3: Great Britain new build price, "current" versus "proposed" methods

Footnotes:

1. *Price data in this figure have been produced by a test system, not by the HPI system used in normal monthly production, and reflect the within-year unchained geometric mean price for Great Britain new builds. In contrast, current UK HPI price estimates are chained geometric mean prices, using January 2023 as the reference period. Before February 2025, the reference period was January 2015. This means the price levels in this Figure are not directly comparable with the published UK HPI price levels for Great Britain new build prices.*

37. 'New build pooling' generally reduces over-estimation of new build price in early provisional UK HPI estimates, with greatest impact on the 'oldest' estimates for which pooling is applied (where new build volumes are higher). In contrast, improved imputation has greatest impact on the 'youngest' estimates (where missingness rates are higher), targeting the early provisional estimates where over-estimation is generally observed.
38. Figure 3 shows that the proposed improvements to the imputation approach is more effective than the current 'pooling' approach at reducing revision size between the 1st and 13th (final) UK HPI estimates. Under the proposed methodology, early provisional estimates for new build prices are significantly lower and the monthly variability in price is reduced.
39. Comparing the average of monthly price differences for all months for the 13-month production runs ending Jul-24 to Jan-25, the average mean price difference for GB new build price was almost twice as large for the current UK HPI published outputs (£5,500), compared with outputs under the proposed methodology (£3,000).
40. The magnitude of revisions to GB new build price from first estimate to latest estimate is also consistently lower under the proposed methodology. The lines for the latest and first estimates are closer together, indicating higher stability in price.
41. This reduction in revisions indicates that improving the monthly imputation increases the accuracy of initial estimates and reduces the necessity for the temporary 'new build pooling'

measure previously introduced to mitigate over-estimation of early provisional new build prices.

42. While implementation of this improved imputation method substantially reduces over-estimation of new build price in early provisional estimates, Figure 3 shows that new build price continues to generally observe downwards revisions between the first and last estimate. This is likely to be driven by the previously-observed trend of more expensive property transactions being registered more quickly than cheaper transactions, leading to inherent downwards revisions. Due to this, ONS is not currently recommending to discontinue the 'dampening factor' measure introduced in 2017, but will keep it under review.

Conclusion

43. Overall, the analysis demonstrated that any of the tested imputation approaches would improve the accuracy of early provisional new build price estimates, and reduce revision size by a similar magnitude.
44. The analysis showed that all tested approaches, including the proposed improved imputation approach, are more effective at reducing over-estimation of new build price in early provisional estimates than the temporary 'new build pooling' approach currently used. Improved imputation is therefore more effective than 'pooling' for reducing revision size between the 1st and 13th (final) UK HPI estimates.
45. The analysis evidenced that KNN imputation with one-hot-key encoding and use of postcode coordinates was the most effective scenario tested for reducing the observed over-estimation of new build price in early provisional estimates for Great Britain, and hence most effective for improving the quality of provisional UK HPI estimates while simultaneously significantly reducing the magnitude of revisions.
46. Therefore, ONS' recommended approach is:
    a. To use a K nearest-neighbour imputation with one-hot-key embedding and postcode coordinates, using 10 nearest neighbours (K=10), to improve monthly imputation of the variables number of rooms (Great Britain) and floor area (England and Wales).
    b. To continue using 'missing indicator' approach for other categorical variables.
    c. To continue setting floor area to zero for all properties in Scotland, to retain consistency in property characteristics between the monthly regression data and the fixed basket for properties in Scotland.[12]
    d. To discontinue the temporary 'new build pooling' (2nd estimates to 7th estimates) approach introduced during the COVID-19 pandemic, since improving the imputation approach is more effective than, and reduces the benefit derived from, the temporary 'pooling' approach.
47. The analysis indicates that this proposed approach will:
    a. Improve the quality of provisional UK HPI estimates by reducing the over-estimation of new build prices in early provisional estimates, resulting in a lower average price difference across all months and a smaller mean percentage difference between the first and latest estimates;

[12] APCP-T2501-HPI-Imputation.pdf explains why Scotland's floor area is set to zero for all properties in the annual and monthly data

b. Significantly reduce the size of revisions in GB new build price;
c. Produce more stable and accurate early provisional price estimates, which are closer to the final estimate;
d. Address the uncertainty in new build price estimates in the current methodology, by targeting and mitigating the impact of one of the causes: higher missingness rates for new build property characteristics in early provisional estimates.

48. ONS recommends this option above other considered imputation options because:
    a. This method preserves information without assuming any ordinal relationship between categories, thus avoiding biases.
    b. It makes better use of available data (currently unused postcode coordinates data) to optimise accurate identification of 'nearest neighbour' donors.
    c. The flexibility of the function has been utilised to enhances the model's ability to find 'similar' donors.
    d. The approach is computationally efficient, scalable and fast, suitable for use in tight monthly production.
    e. Reviews of methodology to identify improvements to maximise the accuracy of UK HPI statistics aligns with Code of Practice for Statistics standards.
    f. Consultation with ONS Methodology experts confirmed their continuing support for considering alternative imputation methods to CanCEIS within the UK HPI, such as K nearest-neighbour, which independently aligns with our selected imputation approach from this analysis.
49. Table 1 summarises the 'current' and 'proposed' monthly imputation applied to variables with missing values, for the Great Britain HPI system.

Variable with missing values	Current HPI monthly imputation	Proposed HPI monthly imputation
 Floor area	England and Wales: Set to zero Scotland: Set to zero	England and Wales: Imputed using 10 nearest neighbours with matching property characteristics. Scotland: Set to zero
 Number of rooms	Assigned "missing label"	Imputed using 10 nearest neighbours with matching property characteristics
 Property type, Acorn	Assigned "missing label"	Assigned "missing label"
 Local authority code, New/Old	Never missing	Never missing

Table 1: Improvements to HPI monthly imputation for Great Britain

50. A detailed description of the KNN imputation specification is provided in Annex A.

Aimee North and Malik Khalid
Housing Market Indices
Prices Division, Office for National Statistics
April 2025

Annex A: KNN imputation implementation

51. The KNN OneHotKey imputation method was motivated by its similarity to CanCEIS' nearest neighbour hot-deck imputation. It demonstrates significant advantages over the temporary 'new build pooling' measure introduced during the COVID-19 pandemic. This method enhances the accuracy and consistency of the initial estimates, reducing the frequency of large revisions.
52. ONS' selected KNN imputation model uses the scikit-learn python package and is set up to
    a. Embed the observation using OneHotKey encoder for categorical variables and geographical coordinates for postcode.
    b. Train the KNN model on the annual fixed basket (one-hot encoded, with NSPL postcode coordinates).
    c. Identify the 10 'nearest neighbours' for each record with missing values.
    d. Impute missing values using the median of the 10 nearest neighbours.
53. Fitting the KNN Model:
    a. **Initialisation:**
        i. The script uses OneHotEncoder from scikit-learn to perform the one-hot encoding.
        ii. This encoder is initialized in the fit_knn method when the embedding is set to onehotkey.
    b. **Fitting the encoder:**
        i. The encoder is fitted on the annual fixed basket transactions data, which contains the categorical columns defined in similarity columns (LA code, new/old, property type, Acorn, rooms).
        ii. The fitting process involves learning the unique categories in each column and creating binary vectors for each category.
    c. **Transforming data:**
        i. Once fitted, the encoder transforms the categorical data into one-hot encoded vectors. This transformation is applied to both the basket data (donors) and the data to be imputed (receivers).
        ii. For example, the property_type column has four unique values (e.g. flat, terraced, semi-detached, detached), and will be transformed into four binary columns: property_type_flat, property_type_terraced, etc.
    d. **Using coordinates:**
        i. The method includes geographical postcode coordinates (from the National Summary Postcode Lookup, NSPL) in the KNN model.
        ii. A function assigns NSPL coordinates (*northing* and *easting*) to each record based on postcode data.
        iii. If coordinates are missing, it approximates them using LA averages.
        iv. The coordinates are combined with the one-hot encoded data to form the final donor array and scaled down, because we then use Euclidean distance on the embedding so that a difference in one of the categorical variables would correspond to 1-2km geographical distance.
54. Training the KNN Model:
    a. **Initialization of the KNN model:**

i. The script uses the NearestNeighbors class from scikit-learn to create the KNN model.
ii. The n_neighbors parameter is set to 10 by default, which means the model will consider the 10 nearest neighbours for each record with missingness during imputation.

b. **Training process:**
i. The fit method of the NearestNeighbors model is called with the donor array as input. It creates a smart tree representation to allow a fast search. The optimized KNN searches the space cleverly and it doesn't need all pairwise distances to do so.
ii. The distances are calculated using a distance metric (Euclidean distance), which measures the similarity between records based on their feature vectors.
iii. The training process involves storing these distances and the indices of the nearest neighbours for each record.

c. **Distance metric:** The distance metric used is Euclidean distance.

d. **Efficiency considerations:**
i. The NearestNeighbors model is optimised for efficiency, allowing it to handle large datasets.
ii. It uses data structures like Ball-trees to speed up the nearest neighbour search, especially in high-dimensional spaces.

55. This detailed process ensures that the KNN model is well-trained to find the most similar records based on both categorical and geographical information. The trained model is then used to impute missing values by leveraging the information from the nearest neighbours.
56. Predict KNN values:
    a. **Identifying missing values:**
    i. The method first identifies which records have missing values in floor area and number of rooms.
    ii. It creates masks to flag records with missing values. These masks are used to filter the records that need imputation.
    b. **Receivers:** It embeds them in exactly the same way as the donors. So they live in the same Euclidean space, and we can use that distance to find the neighbours.
    c. **Finding 10 nearest neighbours:**
    i. The KNN model, which was trained earlier, is used to find the nearest neighbours for each record in the receiver's array.
    ii. The k-neighbours method of the NearestNeighbors model is called with the receiver's array as input. This method returns:
    iii. The distances to the nearest neighbours.
    iv. The indices of the nearest neighbours in the donor array.
    d. **Imputing missing values:** For each record with missing value of area and rooms, the method imputes the values using the median of the nearest neighbours.
    e. **Assign imputed values:** The imputed values are assigned back to the original dataframe, replacing the missing values.