ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

# Inflation for household groups: calculation of weights

## Purpose

1. This paper reviews the methodology underlying the calculation of weights that is necessary in order to produce inflation indices for household groups[1].

## Actions

2. Members of the Panel are invited to:
   a) comment on the existing methods and results of analysis to date
   b) advise on any improvements that could be made to existing methodology
   c) advise on the vintage of LCF data that should be used in calculating expenditure estimates for household groups
   d) advise on any alternative methods and techniques that ONS should consider as part of this work

## Background

3. Survey data from the Living Costs and Food survey (LCF) can be used to derive expenditure estimates for different household groups, but there are many differences between the data source used for CPIH weights (mainly derived from Household Final Consumption Expenditure – HHFCE) and the LCF.

4. In "Variation in the inflation experience of UK households" (ONS, 2014), a method for reconciling the different data sources was developed, to ensure the expenditure weights for CPI household groups summed to the aggregate expenditure totals used in the calculation of CPI. In turn, this ensured that the indices produced for different household groups were comparable to the aggregate CPI index.

5. Recent work has started to review the reconciliation method and to expand on previous work to enable the creation of inflation indices for CPIH household groups. This requires the computation of CPIH consistent expenditure for individual households. **Annex A** presents the work that has been carried out to date. It is expected that this work will go towards an ONS publication in July 2017 showing the impact of different weighting methods on CPIH, which also requires these expenditure estimates at the individual household level.

6. There are 3 areas of methodology that are considered with regards to producing weights for CPIH household groups. These are:

   A. Reconciling LCF data with CPIH expenditure totals

   B. Vintage of LCF data used to reapportion CPIH expenditure totals

   C. Measuring owner occupier housing costs for individual households

[1] Note: This paper will focus on CPIH sub-groups, but similar methodology is considered for the Household Costs Indices (HCIs) – excluding the imputed rental components.

## Reconciling LCF data with CPIH expenditure totals

7. The method used to reconcile LCF data with CPIH expenditure totals divides reported aggregate CPIH expenditure on each COICOP class among the individual households observed in the LCF in proportion to their observed spending on that class-level category.
8. There are some instances where it is not advisable to use straight forward reconciliation methodology because the CPIH expenditure differs drastically to reported LCF expenditure, or a very small number of households report expenditure within a particular class. In these instances, a proxy is used where expenditure on a COICOP class is reapportioned to households using a higher aggregate (e.g. group or division).
9. The rules for this method are arbitrary (where CPIH expenditure totals are more than double the LCF expenditure totals and fewer than 20% households report spending) and classes identified using this method account for 4-5% of the CPIH basket.
10. Alternative methods considered are nearest neighbour imputation and a two-step regression model.

## Vintage of LCF data used to reapportion CPIH expenditure totals

11. The vintage of LCF data used can have a significant impact on the resulting indices. There are 3 vintages of LCF data currently in consideration:
    - an annual LCF dataset that covers the same year as the HHFCE data that is used to calculate CPIH expenditure totals (i.e. a 2 year lag)
    - the latest version of LCF data available at the time of calculating weights (i.e. a 6 month lag)
    - a pooled dataset of 3 years of LCF data
12. As the LCF is used to calculate HHFCE it makes intuitive sense to use the vintage of LCF data that corresponds to the same year of HHFCE data that is used to calculate CPIH weights. However, the latest version of LCF data would ensure that the household mix is more in line with the current household mix for that year.
13. Concerns have been expressed over the LCF sample size, and as such it is also considered that pooled expenditure estimates from a given number of years (e.g. 3) could be used. These could be centred on the year of HHFCE that is used to calculate the CPIH weights, or it could be weighted towards the most recent period.
14. A decision is also needed regarding whether the vintage of LCF data chosen would need to be price uprated to the base period to match the existing price uprated CPIH expenditure totals for a given year. Alternatively, non price uprated CPIH expenditure could be used to reconcile the LCF and CPIH expenditure totals, and then the LCF data could be price uprated to the base period (when aggregated, this would be consistent with the price uprated CPIH expenditure totals for a given year).

## Measuring owner occupiers' housing costs for individual households

15. A two-stage Heckman selection model has been developed to calculate imputed rents for individual households. The model explains log rent paid as a function of the characteristics of the house rented and the household living there, while partially accounting for selection into rented accommodation. The model accounts for a large amount of the variance in price (r-sq = 0.8) and the rental prices predicted for owner occupiers make intuitive sense.

**Robert Bucknall**
**Methodology, ONS**
**May, 2017**

## List of Annexes

| | |
|---|---|
| **Annex A** | Calculation of weights for CPIH household groups |

## Annex A – Calculation of weights for CPIH household groups

### Background

1. The CPIH is a measure of UK consumer price inflation that includes owner occupier's housing costs (OOH). These are the costs of housing services associated with owning, maintaining and living in one's own home.

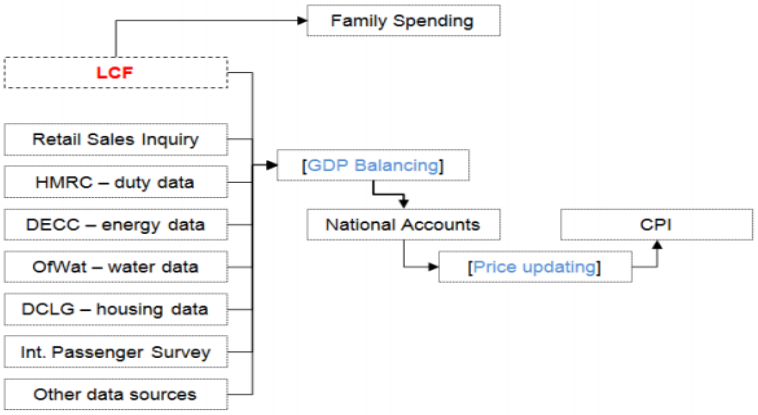
2. The OOH component of CPIH is calculated using a method called rental equivalence. This approach involves using data on the housing costs of actual renters to estimate the price owner occupiers would have had to pay to rent their own home on the market. This concept – known as 'imputed rentals' – captures the implied costs of owner occupation.

3. Work is currently in progress to create inflation indices for household groups on a CPIH consistent basis. This involves producing a unique set of weights for each specified household group. This paper reviews some of the methodology previously developed in "Variations in the inflation experience of UK households" (ONS, 2014), and expands the methods to include an estimate of owner occupier housing costs for individual households.

### A - Reconciling LCF data with CPIH expenditure totals

4. Survey data from the Living Costs and Food survey (LCF) can be used to derive expenditure for household groups but there are many differences between the data source used for CPIH weights (Household Final Consumption Expenditure – HHFCE) and the LCF. HHFCE uses administrative and other data sources to increase the accuracy of its expenditure estimates, and to modify coverage to domestic expenditure[2]. HHFCE also makes adjustments for under and over reporting in the LCF and goes through a balancing process. Figure 1 shows the underlying data sources for the compilation of CPI expenditure weights. As well as this data, the CPIH includes additional data from the VOA and DCLG to estimate expenditure on the owner occupiers' housing component of CPIH.

**Figure 1**: Sources used in the compilation of CPI




**Source:** ONS
**Note(s):** (1) Figure shows a number of the sources and processes used in the compilation of the CPI. LCF is the Living Costs and Food Survey, HMRC is Her Majesty's Revenue and Customs, DECC is the Department of Energy and Climate Change, OfWat is the water regulator, DCLG is the Department for Communities and Local Government, Int. Passenger Survey is the International Passenger Survey.

[2] The LCF is based on a national concept and only captures spending by UK private households.

5. As the data sources are different, a reconciliation process is applied to ensure that the expenditure for CPIH subgroups is consistent with aggregate CPIH expenditure, so that the subgroup indices produced are comparable to the headline CPIH index.

6. The method to reconcile LCF and CPIH expenditure totals divides reported total CPIH expenditure on each COICOP class among the individual households observed in the LCF in proportion to their observed spending on that class-level category.

$$e_{h,i,t}^{CPI} = \frac{e_{h,i,t}^{LCF}}{\sum_{h} e_{h,i,t}^{LCF} * w_{h,i,t}^{LCF}} . e_{i,t}^{CPI}$$

Where:

$e_{h,i,t}^{CPI}$ is the level of expenditure consistent with the CPIH for household *h*, in COICOP class *i* at time *t*

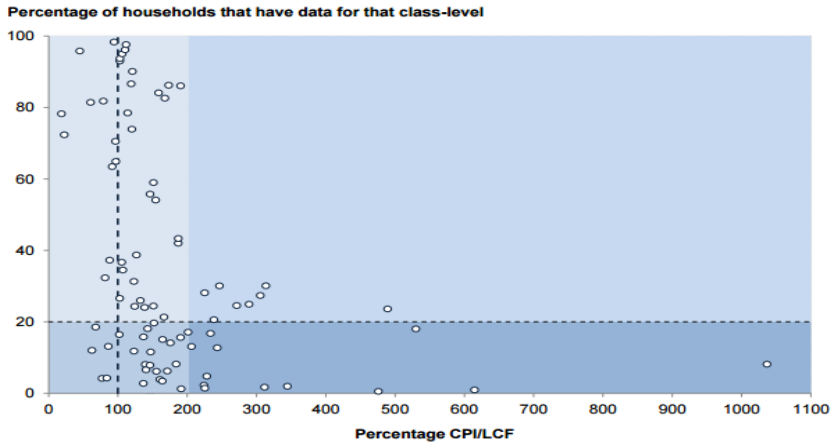
$e_{h,i,t}^{LCF}$ is the level of expenditure consistent with the LCF for household *h*, in COICOP class *i* at time *t*

$w_{h,i,t}^{LCF}$ is the weight of household *h*, in COICOP class *i* at time *t*

7. Analysis in the 'Variation of inflation experiences of UK households' (2014) showed that where the expenditure weight in the CPI is based on data other than the LCF, the differences between the LCF and CPI expenditure totals can be extremely large (for example, CPI expenditure on medical and paramedic services is 230 times more than expenditure reported through the LCF).

8. There are also some instances where a large amount of expenditure is allocated to a small number of households, as only few surveyed households' may report expenditure within a particular class. Figure 2 displays the percentage of households that have data for each class against the percentage difference between LCF and the CPI expenditure totals.

**Figure 2**: LCF and CPI expenditure by COICOP class




**Source:** ONS calculations
**Note(s):** (1) Figure is a scatter plot showing the 85 class-level categories of the CPI. The vertical axis plots the percentage of surveyed households who report some expenditure on that class-level item. The horizontal axis shows the percentage difference between the expenditure total in the CPI and LCF.
(2) The Figure excludes class-level category 06.3.1: Medical & paramedic services as the CPI expenditure total is around 230 times bigger than in the LCF.

9. The horizontal axis shows the difference between the LCF and the CPI expenditure total as a proportion of the LCF total. Values close to 100 indicate close correspondence between the expenditure estimates from each data source, while observations further from 100 indicate greater differences. The vertical axis plots the number of households who report positive expenditure. Each point is a single COICOP class, representing expenditure on a given set of products.

10. The chart is divided into four zones. In the top left are COICOP classes in which the LCF and CPI expenditure totals are of broadly similar orders of magnitude (CPI expenditure less than double the LCF total) and where the number of households reporting positive expenditure is relatively high (above 20%). In this segment are many products for which the LCF is taken as the basis for the CPI weights, for example food. We can be relatively confident that the reconciliation method will be suitable for these classes.

11. In the bottom left quadrant are instances where the number of households reporting strictly positive expenditure is relatively small, but where the CPI and LCF expenditure totals are similar. Points in the two left quadrants will introduce the least potential bias.

12. Points in the top right hand quadrant represent classes where CPI expenditure is high relative to the LCF total, but where a relatively large number of households have reported positive spending. In these cases the potential for bias is also limited as a large proportion of households will be affected by the micro-level attribution mechanism.

13. It is points in the bottom right hand quadrant that present the most difficulty: these are COICOP classes in which the CPI expenditure total is more than double the LCF total, and in which fewer than 20% of households report spending. Medical & Paramedic Services remains an outlier, with just 66 households reporting spending on this COICOP class-level over the eleven years of available data.

14. A fix was implemented in the 2014 paper by applying some additional methodology, and the same methodology is proposed to calculate CPIH consistent subgroup measures of inflation.

15. The first step was to identify COICOP classes that had the biggest impact on the potential distortion that was being caused. These classes were identified as those where the:

    - average ratio of CPI(CPIH) to LCF expenditure over the previous 10 years is greater than 2
    - average percentage of households that report spending in that COICOP class over the previous 10 years is less than 20%

16. The classes that are identified using this method generally account for 8-9% of the overall CPI (4-5% of CPIH) each year.

17. For these classes, spending on the class is allocated using the reported proportion of household expenditure on a higher aggregate (group if available, or division). This ensures that the methodology does not allocate very high levels of expenditure to a relatively small number of households. For example, expenditure on ‘medical and paramedic services’ may be allocated to households based on their expenditure on ‘health’.

18. An issue with this method is that households with no expenditure on a particular class can be allocated expenditure, based on their level of expenditure on a higher aggregate.

19. The parameters used in the identification process were originally selected arbitrarily and can be tested in detail along with whether the use of a 10 year rolling average is appropriate. Analysis already completed indicates that a 5-year rolling average would usually identify the same classes as the 10-year average.

20. Alternative methods have been considered for future development. These include:
    a. Nearest neighbour imputation - expenditure in the class is imputed based on households with similar characteristics
    b. A two-step model:
        i. First, estimating the likelihood of expenditure – this can be done using logistic regression where the dependant is 1 if expenditure>0 or 0 if expenditure=0, the explanatory variables are characteristics of the households
        ii. Second, estimating the extent of that expenditure

## B - Vintage of LCF data used to reapportion CPIH expenditure totals

21. As noted in the Section A, LCF data is used to calculate CPIH consistent expenditure at an individual household level. A decision needs to be made on which vintage of the LCF data to use, as different vintages may lead to different results. There are three vintages of LCF data under consideration, including:

    - an annual LCF dataset which covers the same year of HHCFE that is used to calculate the CPIH weights (i.e. a 2 year lag)
    - the latest version of the LCF available at the time of calculating the weights (i.e. a 6 month lag)
    - a pooled dataset of at least 3 years worth of LCF

22. Another question is around whether the vintage of LCF data chosen would need to be price uprated to the base period to match the existing price uprated CPIH expenditure totals for a given year. Alternatively, non price uprated CPIH expenditure could be used to reconcile the LCF and CPIH expenditure totals and then the LCF data could be price uprated to the base period (when aggregated, this should still be consistent with the price uprated CPIH expenditure totals for a given year).

23. There are advantages and disadvantages to each approach, which need to considered when using the LCF in this way.

24. For LCF covering the same year as the HHFCE, the main advantage is that expenditure recorded in the LCF refers to the same period HHFCE used in the Blue Book. Since LCF makes up a significant proportion of the HHFCE, the COICOP categories for which the LCF is the sole source in the HHFCE the expenditures would be similar, subject to National Accounts balancing procedures, and the expenditures for the other COICOP categories should be proportional. With this approach, it would make sense to use non price uprated CPIH

expenditure to reconcile the LCF and CPIH expenditure totals, and then the LCF data could be price uprated to the base period.

25. One disadvantage of using this vintage of the LCF is that the mix of households in the LCF for this year may not reflect the mix of households in the year for which the CPIH subgroups would be calculated. Using the latest version of the LCF dataset available at the time of calculating the weights would result in different expenditures but the household mix would be the closest to the current mix. However, it may cause inconsistencies when considering what price uprating should be used.

26. The sample size of the LCF is also an issue, especially when stratifying by subgroups as even though the sample size of the LCF is approximately 5000 households, after non-response, stratifying cuts the sample size to only those with the desired characteristic. A possible solution to this is to pool the dataset over a given number of years; this increases the sample size of households available for the stratification and has the possibility of improving the accuracy of the estimates. Pooling also smoothes out volatile expenditure in certain COICOP categories, which may help reduce the problematic classes discussed in Section A.

27. One issue with using pooled data is that the average expenditure in the pooled estimate does not match up with the HHFCE, though taking an average centred at the HHFCE year would reduce this mismatch.

28. Another issue is that the household mix over the years could vary and therefore using the households in the pooled dataset may not be representative of the households in the current year. This is more apparent if the subgroups are to be calculated by fixed expenditure or income deciles over the three year period, as households are more likely to change deciles than they are from one demographic group (e.g. non-retired) to another (retired). For example, take three years of LCF data with the HHFCE corresponding to the second year. For each of the years in the pooled dataset the sample size is 5000, then if the first year 20% of the households are in the fixed first income decile, in the second year only 5% of households and in the third year 10% of households, then in the pooled dataset 11.7% of households are in the first income decile. This would then over represent the households in this income decile when calculating the subgroups.

## C - Measuring Owner Occupiers' Housing Costs (OOH)

29. The main data source for our estimates of household level imputed rentals is the LCF, which includes information on the level of rent, any housing benefit received by households and the characteristics of the house and household.

30. To deliver CPIH consistent inflation rates for sub-groups of households, it is necessary to estimate owner occupier housing costs for each individual household shown to be an owner occupier (ie those who don't already show expenditure on actual rent). These estimates take the form of expenditure weights in our sub-group indices. In overview, we:

    i. Assemble data on the level of rent (including the value of any housing benefit paid) and the characteristics of both rented properties and renting households.

ii. Estimate a two-stage Heckman model which explains log rent paid as a function of the characteristics of the house rented and the household living there, while partially accounting for selection into rented accommodation.

iii. Use the coefficients from this model to estimate the level of imputed rent for all owner-occupiers.

31. The Heckman selection model assumes that there exists an underlying regression relationship. The dependent variable, however, is not always observed and standard regression techniques can potentially yield biased results. The two stage Heckman model provides consistent, asymptotically efficient estimates for all the parameters. Households choose whether to rent, and thus from our point of view, whether we observe their rent in our data. If households made this decision randomly, we could ignore that not all rents are observed and use ordinary regression to fit a rents model. Such an assumption of random participation, however, is unlikely to be true; households which would have low income may be more likely to choose to rent, and thus the sample of observed rent would be biased.

**Existing Model**

32. A model was developed in a previous iteration of this work (not published). Under the existing model the following variables were included in both the selection and outcome equations of the Heckman model:

- Housing Type
- Year
- Number of Rooms
- Region
- Expenditure Percentile
- Expenditure Percentile Squared
- Tenure Type
- Council Tax Band

33. The selection stage is not properly identified because of a lack of an appropriate exclusion restriction (i.e. a variable which influences the probability of renting, but not the level of rents). Consequently, the estimated coefficients are unlikely to have been purged of selection bias. As a result, the imputed value of rentals may differ from the unobserved 'true' value depending on whether higher or lower quality properties tend to be rented.

34. Using the coefficients from the model to estimate the level of imputed rent for all owner-occupier households, these are then constrained to Valuation Office Agency (VOA) aggregates at year and region level as follows:

$$R_{h,t} = \frac{\hat{R}_{h,t,LCF}}{\overline{R}_{t,r,LCF}} . \overline{R}_{t,r,VOA}$$

Where:

$\hat{R}_{h,t,LCF}$ is the modelled imputed rent for owner occupiers in household *h*, year *t* on the *LCF* dataset

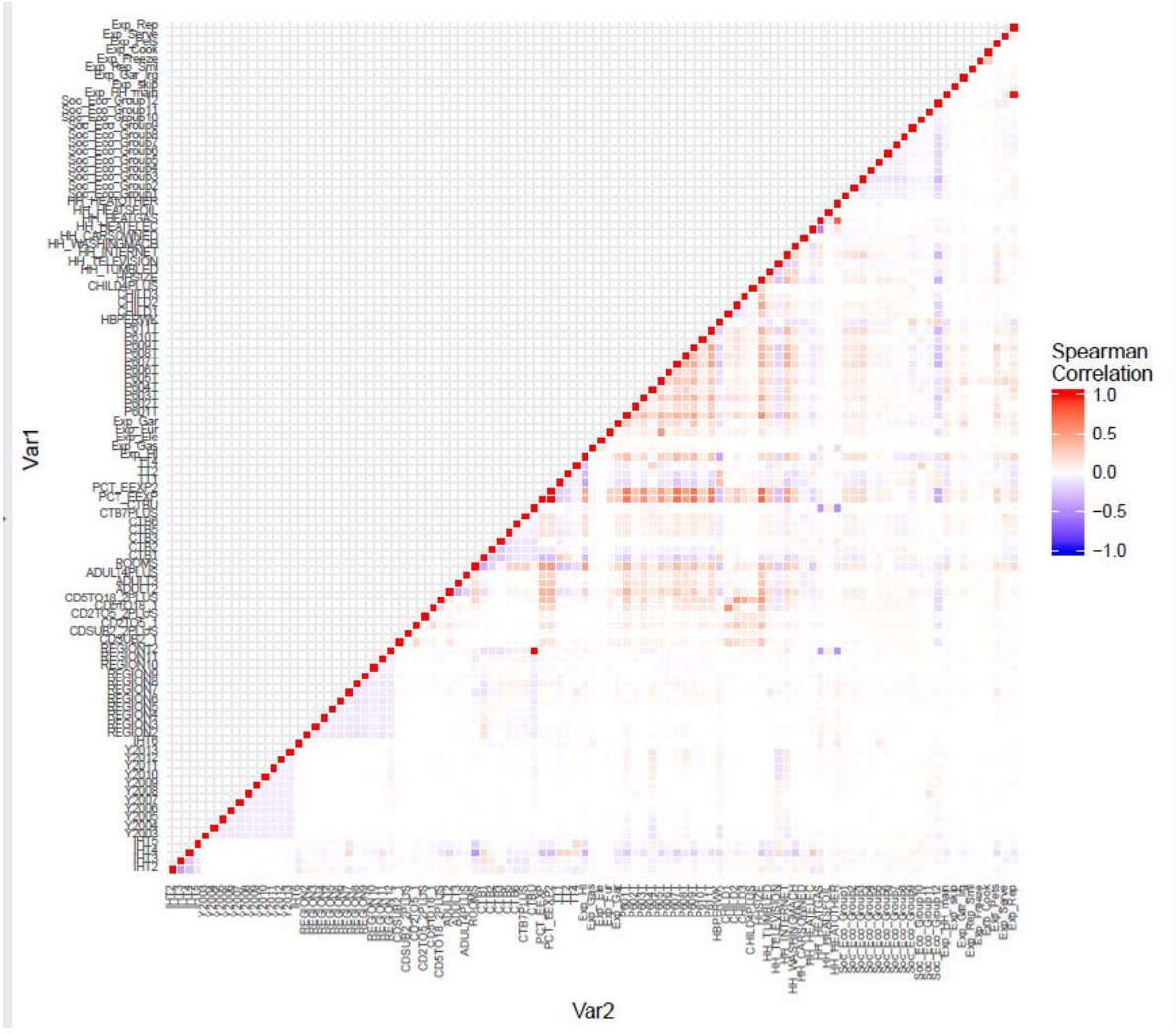
$\overline{R}_{t,r,LCF}$ is the average rent for renters in year *t*, region *r* on the *LCF* dataset

$\overline{R}_{t,r,VOA}$ is the average rent for renters in year *t*, region *r* on the *VOA* dataset

**New model**

35. Figure 3 displays a correlation heat map, which is a comprehensible representation of the data set, which allowed identification of significant correlations to determine which variables would be ineffective should they be paired in the model.

**Figure 3**: Correlation of variables considered for new model




36. The selection stage of the existing model is not properly identified because of a lack of an appropriate exclusion restriction. Characteristics of renters and owner occupiers are different in some ways, so to try and minimise selection bias, extra variables have been added to the model which influence the probability of renting (not necessarily the level of

rent). The variables included in the selection and outcome equation of the Heckman model are:

- Housing Type
- Year
- Number of Rooms
- Region
- Number of Adults
- Number of Children
- Council Tax Band
- Expenditure Percentile
- Expenditure Percentile Squared
- Tenure Type
- Expenditure on Housing Fuel and Power
- Expenditure on Education
- Housing Benefit per Week
- Social Economic Group
- Expenditure on repairs
- Expenditure on Pets

37. Using the coefficients from the model to estimate the level of imputed rent for all owner-occupier households, these are then constrained to Valuation Office Agency (VOA) aggregates at year, region and dwelling type level as follows:

$$R_{h,t} = \frac{\hat{R}_{h,t,LCF}}{\overline{R}_{t,r,d,LCF}} . \overline{R}_{t,r,d,VOA}$$

Where:

$\hat{R}_{h,t,LCF}$ is the modeled imputed rent for owner occupiers in household *h*, year *t* on the *LCF* dataset

$\overline{R}_{t,r,d,LCF}$ is the average rent for renters in year *t*, region *r*, dwelling type *d* on the *LCF* dataset

$\overline{R}_{t,r,d,VOA}$ is the average rent for renters in year *t*, region *r,* dwelling type *d* on the *VOA* dataset

**Analysis**

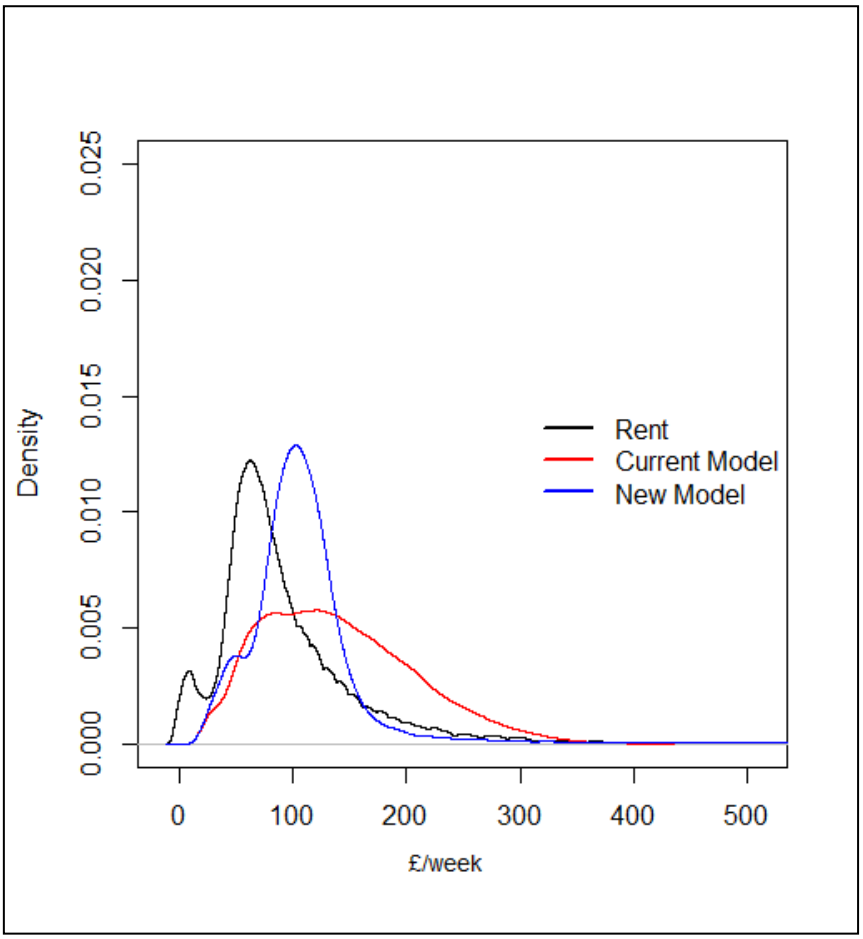
38. The model diagnostics for the existing and new model are presented in Table 1. The R-squared for the new model is higher than that for the existing model, indicating that the new model is a better fit for rents. The Adjusted R-squared and Predicted R-squared were checked to prevent over-fitting of the model. The Akaike information criterion (AIC) measures the quality of statistical models relative to other models. It offers a relative estimate of the information lost when a given model is used to represent the process that generates the data. In doing so, it deals with the trade-off between the goodness of fit of the model and the complexity of the model. The lower the value of AIC, the better the model. AIC is lower for the new model, indicating a better fit while also considering the complexity of the model.

**Table 1**: Model diagnostics for the existing and new model

| Model | R sqd | Adj R sqd | Pred R sqd | AIC |
|---|---|---|---|---|
| Existing | 62.95 | 62.83 | 62.68 | 15581.13 |
| New | 80.25 | 80.17 | 80.02 | 6470.00 |

39. The results of the existing and new model are presented in Figure 4. The distribution plot illustrates that for both models the average imputed rentals of owner occupied housing tends to be above those of the actual rentals. To some extent, this reflects differences in the quality of the properties that are offered for rent relative to those which are owner-occupied.

**Figure 4**: Distribution of imputed and actual rental values, density by £/wk bracket




40. Figure 5 illustrates that the average imputed rentals of owner occupied housing tend to be above those of the actual rentals. This is a plot of average rent (renters) and average imputed rent (owner occupiers) by housing type.

**Figure 5**: Average rent (renters) and average imputed rent (owner occupiers) in pounds (£) per week, by housing type