ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

Redeveloping Private Rental Market Price Statistics

Status: Final

Expected publication: For publication when no longer market sensitive

Purpose

1. This paper outlines the proposed methodology for the new private rental market price statistics. These statistics are expected to be used in future to produce the owner occupiers' housing costs (OOH) element of Consumer Prices Index including OOH (CPIH), as well as the private rent element of the suite of Consumer and Retail Price Indices (Annex A). The OOH measure alone accounts for 18.5% of CPIH in 2021, ONS's headline measure of inflation. Separate analysis on the impact on consumer price indices will be circulated at a later date.

Actions

2. Members of the Panel are invited to:
    - agree the proposed methodology for the new private rental market price statistics
    - comment on any fundamental issues with the proposed methodology

Key points

3. Throughout this paper we will consider the most suitable methodology for the new private rental market price statistics. Key areas that we explore include:
    - The type of model used in our hedonic regression methodology: we propose to use an ordinary least squares regression
    - The transparency of the model
    - The use of Acorn as an independent variable: we propose to include Acorn as an independent variable
    - The use of interaction terms in the model: we propose not to use interaction terms
4. The methodology for the new private rental price statistics measure will be reviewed every 5 years to ensure it is still the best method for the data we are receiving.

Background

5. Currently, the Office for National Statistics (ONS) publish two private rental prices statistical outputs: the UK Index of Private Housing Rental Prices (IPHRP), and Private rental market summary statistics in England (PRMS).
6. IPHRP measures the change in price tenants face when renting residential property from private landlords. It includes an index of private rental prices and annual percentage change of the index for the UK, its countries, and the English regions.

7. PRMS are point-in-time rental price estimates for England, the English regions, and English local authorities. Current methodology limitations prevent compositional changes from being considered, so it is not appropriate to compare the estimates year-on-year to infer trends in the rental market, and a price index cannot be produced.
8. In late 2019, the ONS gained access to the Valuation Office Agency's (VOA) lettings database at a microdata level, this meant we could develop our methodology to better suit user needs. Prior to this, we received aggregated data that was calculated using a matched pairs approach.
9. We aim to unify private rental price statistics by replacing the IPHRP and PRMS with a new, single, monthly publication. This new publication will use the latest available data sources to publish private rental prices statistics comparable over time and to lower geographic levels.
10. The new publication will contain:
    a) an index of private rental prices
    b) annual and monthly percentage change
    c) private rental price levels
    d) a breakdown by geography (UK, its countries, English regions, and local authorities/broad rental market areas) and bedroom category (studio, one bedroom, two bedrooms, three bedrooms, and four or more bedrooms)

Proposed methodology

11. Annually, over 450,000 private rents prices are collected in England (by the Valuation Office Agency), 30,000 in Wales (by the Welsh Government), 25,000 in Scotland (by the Scottish Government) and 15,000 in Northern Ireland (by the Northern Ireland Housing Executive). Data are collected using a convenience sample, more information on the sampling methodology can be found in Section 6 of the IPHRP Quality and Methodology Information.
12. We need to control for compositional and quality changes that take place in property markets. International best practice suggests using a hedonic regression to control for these. Hedonic regression uses a regression model to estimate the influence that various price determining characteristics have on the price of a rental property.
13. In line with international practice, we propose to calculate our new rental price statistics using hedonic regression to provide estimated rental prices that a tenant would face when renting from a private landlord.
14. The proposed hedonic regression methodology is based on the approach used to calculate the UK House Price Index; however, the exact detail is tailored to suit the rental data.
15. A summary of the proposed methodology is shown below (Figure 1). The key stages are:
    a) Link the property-level rental price data with property attributes and location data
    b) Automatic and manual data cleaning exercises are carried out, as specified in Annex A
    c) A fixed basket of rental properties is updated on an annual basis, this accounts for compositional changes in the sample.
    d) Populate any missing data in both the fixed basked and monthly dataset of prices using the most appropriate imputation methods
    e) Model dependence of rent price on property characteristics using a hedonic ordinary least squares (OLS) regression model. The hedonic regression is run each month on the

latest rolling 14-month dataset of prices (Annex B). The price-determining characteristics that we propose to use are:

i) Number of bedrooms
ii) Floor area (in m^2)
iii) Property type (Flat/Maisonette, Detached, Semi-detached, Terraced)
iv) Furnished status
v) ACORN group classification
vi) Local authority
vii) Property age bracket

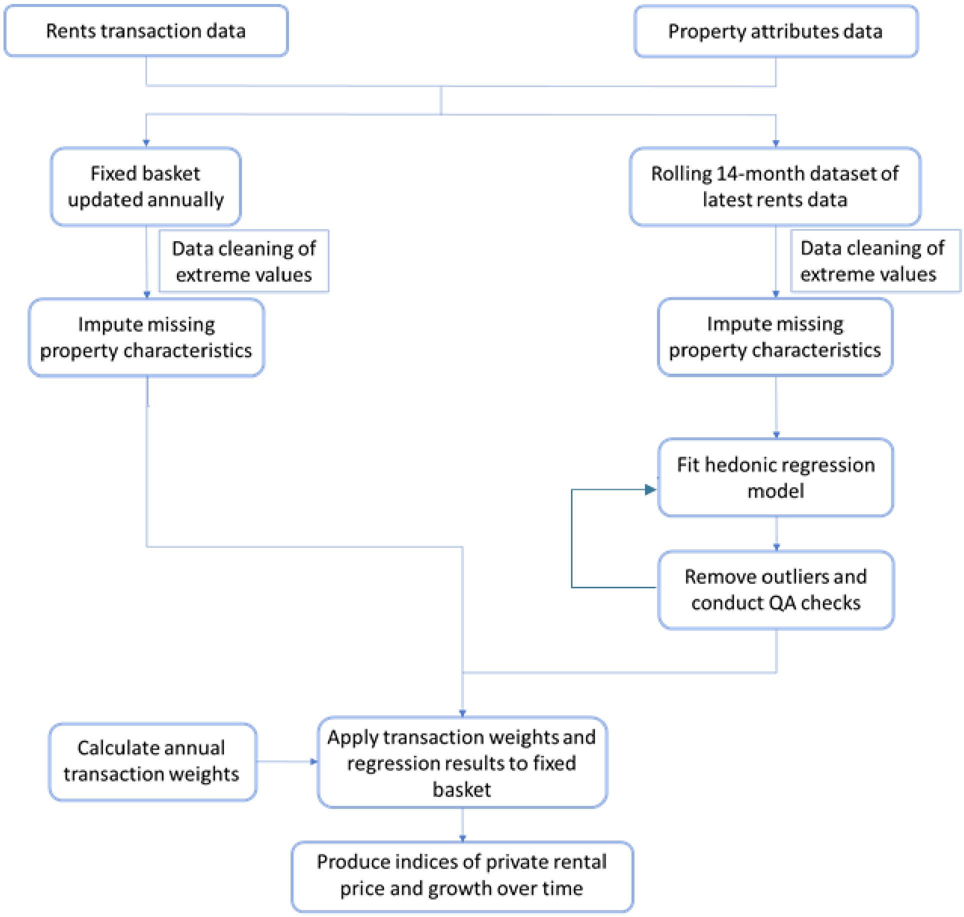
f) The coefficients from the latest month are applied to the fixed basket of rental properties and predicted prices are calculated.

g) Elementary aggregates are produced at a local authority level by taking the ratio of the geometric means of the predicted prices (from the hedonic equation) in the base month and the current month.

h) Elementary aggregates are weighted together (Lowe index) and then chain-linked annually to produce a rental price index series over time for the UK, its countries, English regions and local authorities/broad rental market areas. Expenditure weights are calculated by using dwelling stock estimates at a local authority level (from the ONS subnational dwelling stock estimates, StatsWales and Scottish Government) and average rental prices from the rental price data. Proportions for the type of property (detached, semi-detached, terraced, flat) and furnished status are also used when calculating expenditure shares.

16. The corresponding average rental price series is derived by applying the index to a base set of rental prices from the reference period. For example, if the average rental price in the reference period (currently 2015) was £500, and the index in the current month was 110.0, a 10% growth would be applied to the reference period average rental price. So, the average rental price in the current period would be estimated at £550. This ensures the price series is consistent with the published index, which is a key requirement of this development work.

Figure 1: Proposed Rents Development methodology


Rents transaction data
Property attributes data
Fixed basket updated annually
Rolling 14-month dataset of latest rents data
Data cleaning of extreme values
Data cleaning of extreme values
Impute missing property characteristics
Impute missing property characteristics
Fit hedonic regression model
Remove outliers and conduct QA checks
Calculate annual transaction weights
Apply transaction weights and regression results to fixed basket
Produce indices of private rental price and growth over time

Analysis to support the proposed methodology

17. The analysis in this paper considers the England and Wales data, which cover approximately 90% of the rental market. Scotland and Northern Ireland data have not been used because they are not yet complete. We are waiting for an updated Energy Performance Certificate dataset to link property attributes to the Scotland rental prices, this should be received in October. For the Northern Ireland data, we are in conversations with the Northern Ireland Housing Executive about what data we can use.
18. To provide the highest possible quality statistics, we have developed and compared four different modelling approaches. These are, in order of increasing complexity:
    a) Ordinary Least Squares
    b) Weighted Least Squares without interaction terms
    c) Weighted Least Squares allowing interaction terms
    d) Random Forest including pruning
19. All approaches use the same input data and we have worked diligently to ensure that the comparison metrics are as meaningful as possible across them. To assess the impact of each model we provide 10-fold cross validation root mean squared errors (RMSE) for the datasets created by these models.

20. As well as assessing the statistical accuracy and performance of each model, we also considered other quality aspects such as transparency and timeliness.
21. We tested whether the use of Acorn as an independent variable in our model was appropriate using the generalized variance inflation factor (GVIF) and 10-fold cross validation root mean squared errors (RMSE).

Models considered

Ordinary Least Squares

22. Ordinary Least Squares (OLS) is used to estimate the coefficients of a linear regression model by minimising the sum of squares in the difference between the observed and predicted values of rental prices.
23. OLS is the most accessible of our four modelling approaches. Each collected rental price is deemed to be equally as important as all the others and no explicit weighting of them is enforced.
24. We discussed our proposed methodology with internationally recognised price index experts, by correspondence via Economic Statistics Centre of Excellence (ESCoE), their feedback can be found in Annex B. These experts suggested to compare the results of WLS to those of the less complex OLS.
25. The inclusion of this model acts as a baseline. It allows us to understand the impact of attributing weights to features, based on our quality assessments.

Weighted Least Squares

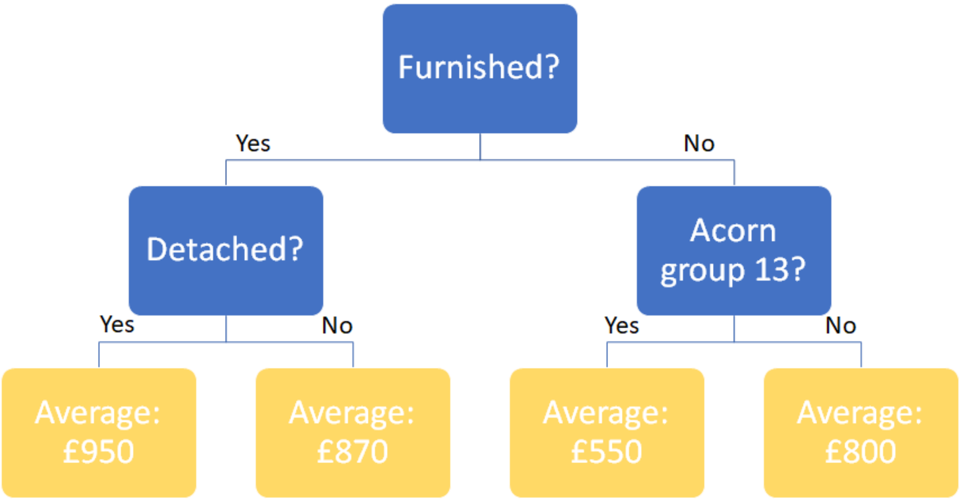
26. Weighted Least Squares (WLS) is an extension of Ordinary Least Squares regression. When the WLS model is fitted, the model puts less weight on observations which are considered less reliable. The model predictions are therefore more heavily influenced by the higher quality data.
27. Observation weights are calculated for the WLS model by understanding the impact of missing data on the robustness of an observation. The weight is equal to the inverse ratio of Mean Squared Error from the OLS model with property attribute variables omitted, to the OLS model with all property attribute variables included. Smaller weights are applied to those observations with a higher number of missing property attributes.

Random Forests

28. Random Forests are an ensemble supervised machine learning algorithm built from an often large number of decision trees. In our case we will use them for regression, although they are also often used for classification.
29. A decision tree uses sequential questions which send you down a particular route in the tree. The root of a decision tree is at the top and splits into branches and leaves as you travel down the tree. In Figure 2, the blue shaded rectangles represent conditions, based on which the tree splits into branches. The end of a branch with no subsequent splits is the leaf, the decision reached (yellow shaded rectangles). The questions are fitted to the training data, such that questions at each level are chosen to reduce price variance in subsequent nodes the most, measured by the mean squared error. When the fitted tree is then applied to new data, the

questions are first applied to reach a leaf and the decision tree prediction is an average of training data cases from that leaf.

Figure 2: Example of a Decision Tree (average prices aren't correct)



30. In a Random Forest, each decision tree is trained independently using a subset of training data. If 200 decision trees are trained in the forest, then applying this model results in 200 different rental value predictions being made (one from each tree). The final random forest prediction averages across the tree predictions. By aggregating in this way, a random forest can turn decision trees, which are generally individually weak learners, into a powerful regression algorithm.
31. Each time a random forest is trained it can provide slightly different models because there is a stochastic aspect to the training. This can be controlled using a random seed to make the splitting of the training data deterministic. Once the model is trained, the outputted data are deterministic.
32. The appropriate parameters used in our random forest model were determined by using the train/test split approach. For example, when deciding on the maximum depth of the trees, if the performance on the training data was far better than the testing data this would indicate overfitting and that the trees were too deep and so a smaller value would be tried until the difference between the train and test became small. The chosen parameters are specified to be as shown in Table 1.

Table 1: Parameters used in the random forest

The number of trees in the forest	200
The maximum depth of the tree	20
The minimum number of samples required to split an internal node	20
The minimum number of samples required to be at a leaf node	10

Whether to use out-of-bag samples to estimate the generalisation score	True
Reuse the solution of the previous call to fit and add more estimators to the ensemble	True

Results

33. Figures 3 and 4 show the indices and growth rates of each model for England and Wales from 2015 onwards. IPHRP has also been provided for comparison.

Figure 3: Comparing the index of the three model options for England and Wales

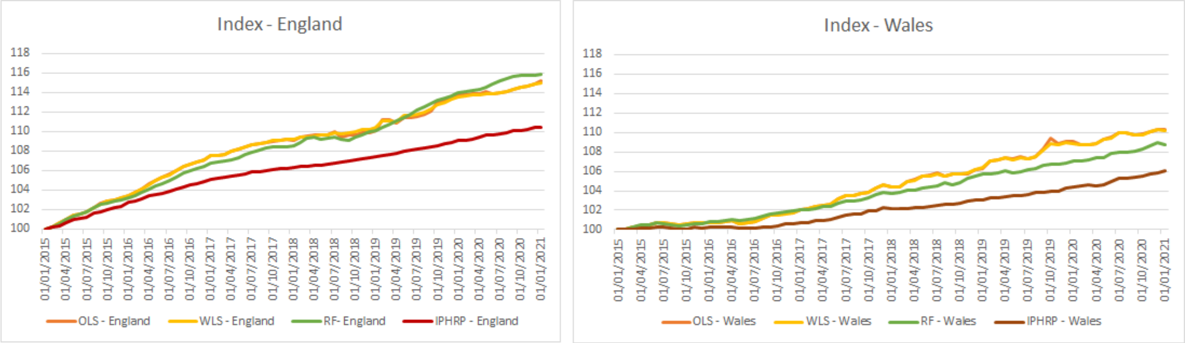
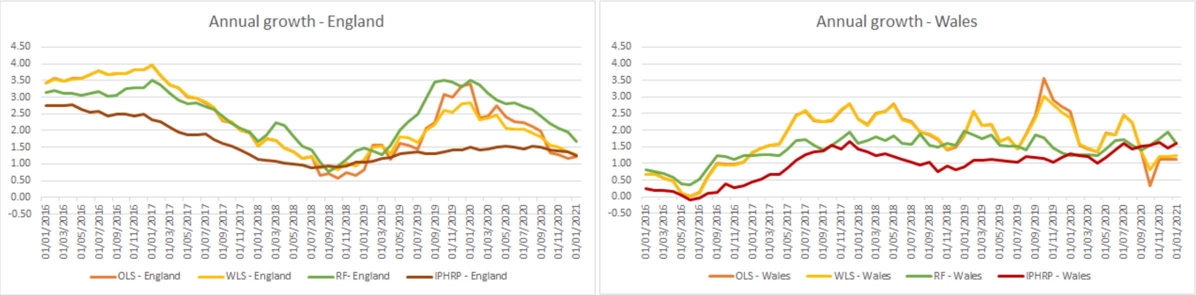


Figure 4: Comparing the growth rate of the three model options for England and Wales

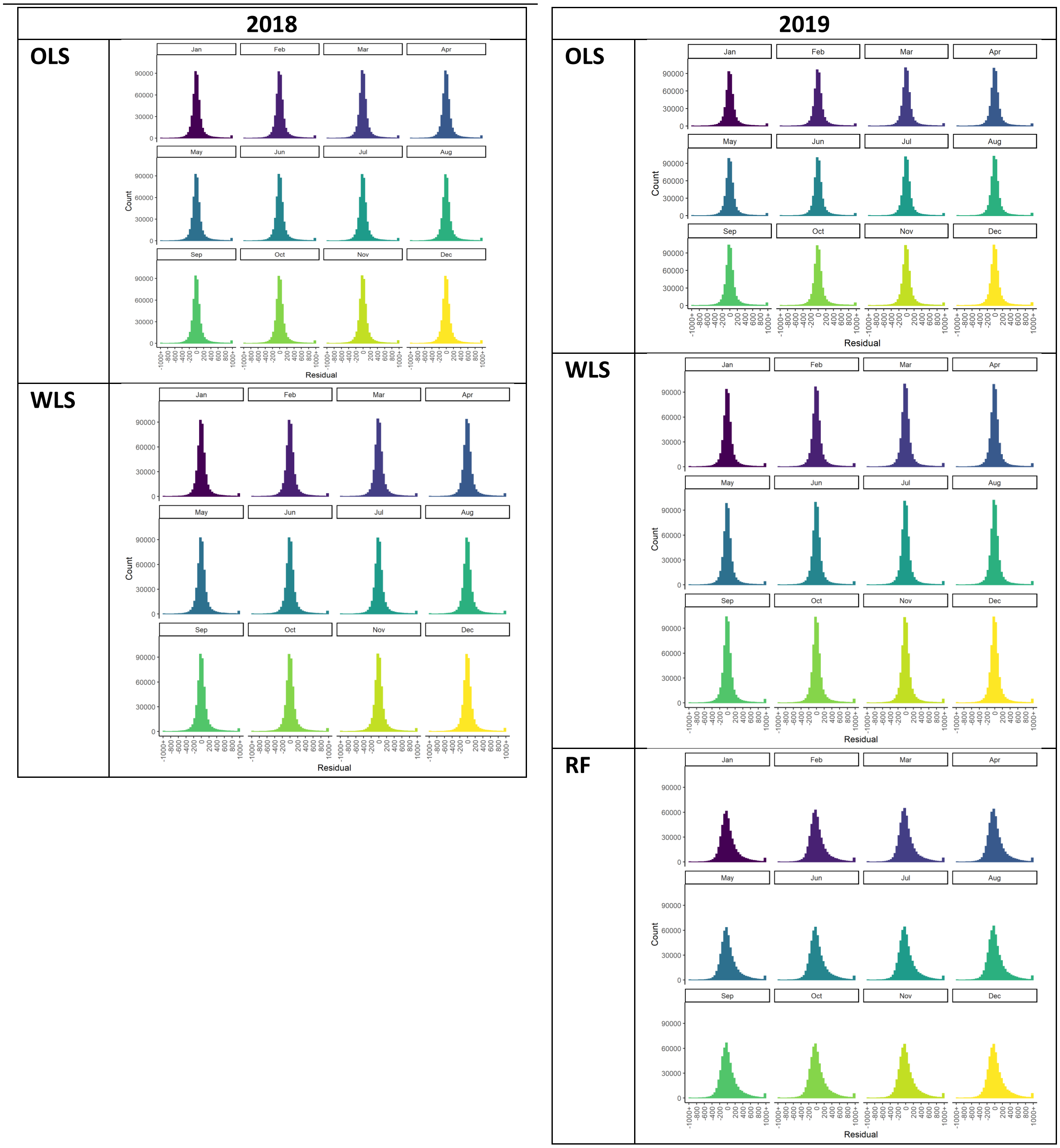


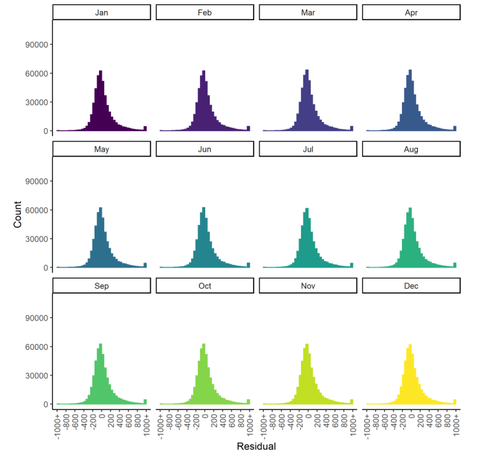
Comparison of models: K-fold cross-validation

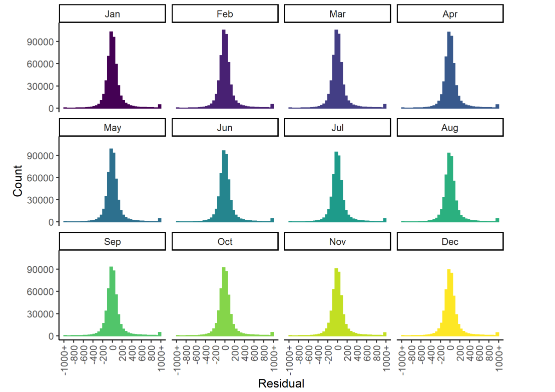
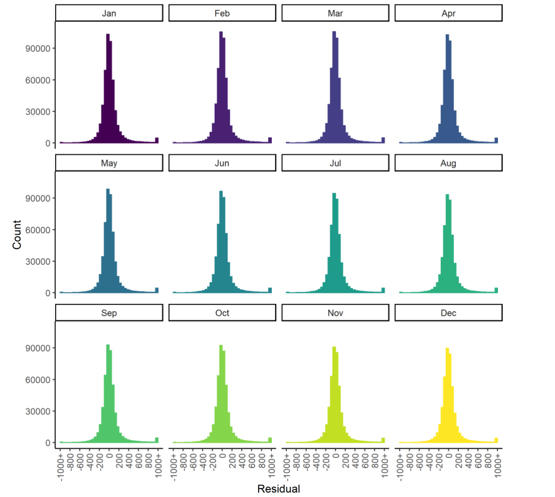
34. To compare the models' ability to predict the rental-prices based on unseen data we used K-fold cross-validation. K-fold cross-validation is a process for testing the model where the data are divided randomly into several 'folds' (K), the model is fitted on the data in K-1 folds and tested on the remaining 1, and this process is repeated until every fold has been used for testing. For each test fold we computed the root mean-squared error (RMSE), coefficient of determination (R^2 statistic), and standardized residuals. For this analysis, we chose K = 10, this is the most popular in machine learning research.
35. This test was performed on OLS, WLS and random forest using England and Wales data. Scotland and Northern Ireland data are not included in the analysis because the data are not currently available.

36. The distribution of residuals for each model across the years are shown in the table below. If the residuals are centred around zero then the predicted rent is close to the actual rent. It is evident that the majority of residuals for OLS and WLS are centred around or close to zero. The distribution of residuals for the random forest is mainly centred around or close to zero, however it is slightly positively skewed. For all models there is a cluster of residuals on the right tail and closer examination of observed rent prices and the corresponding predicted rent showed that the models tend to underestimate very high rents (approximately 0.8% to 1.3% of the data falls within the highest bin, see Table 2 for a detailed breakdown).

Figure 3: Distribution of residuals (also found in Annex E)



RF	

2020	
OLS	
WLS	

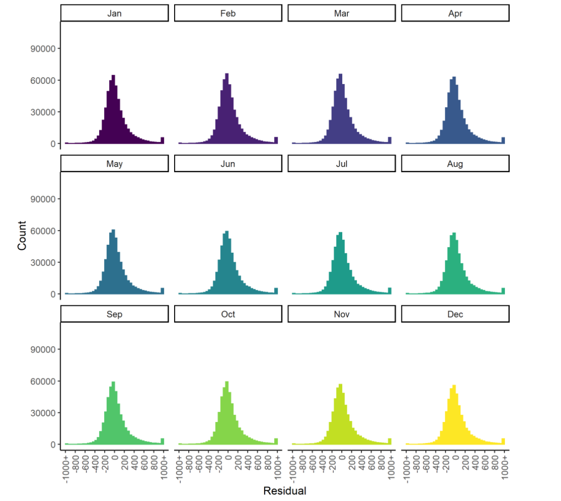
RF	

Table 2: Percentage of data falling into the highest residual bin (1000+)

Percentage of data falling within highest residual bin (1000+)									
	2018			2019			2020		
Month	OLS	WLS	RF	OLS	WLS	RF	OLS	WLS	RF
January	0.86	0.86	1.16	0.97	0.96	1.19	1.01	1.02	1.28
February	0.9	0.90	1.19	0.96	0.96	1.21	1.03	1.03	1.29
March	0.87	0.87	1.17	0.96	0.96	1.18	1.03	1.03	1.30
April	0.87	0.87	1.17	0.95	0.95	1.17	1.04	1.04	1.30
May	0.88	0.88	1.18	0.96	0.95	1.19	1.04	1.04	1.30
June	0.89	0.89	1.18	0.95	0.95	1.18	1.05	1.05	1.32
July	0.89	0.89	1.19	0.95	0.94	1.17	1.05	1.05	1.34
August	0.91	0.91	1.19	0.96	0.96	1.21	1.05	1.05	1.34
September	0.9	0.90	1.19	0.99	0.98	1.25	1.06	1.05	1.36
October	0.93	0.91	1.21	1.02	1.04	1.24	1.05	1.05	1.35
November	0.95	0.94	1.21	1.01	1.01	1.26	1.05	1.05	1.36
December	0.97	0.96	1.21	1.03	1.04	1.28	1.06	1.06	1.37

37. The Root Mean Squared Error (RMSE) measures the prediction error made by the model. The lower the RMSE, the better the model is at predicting outcomes. WLS and OLS have a consistently lower average RMSE than Random Forest, but the difference isn't substantial. As RMSE is the mean of the squared error, it is particularly sensitive to large errors. The numbers in brackets show the standard deviation of the RMSE, which is typically smaller for the Random Forest model.
38. OLS and WLS perform better than Random Forest when considering the average RMSE across 10 folds, however the standard deviations are usually larger in OLS and WLS than Random Forest.

Table 3: Average RMSE (and its standard deviation) across 10 folds

Average RMSE across 10 folds (standard deviation in brackets)

	2018			2019			2020		
Month	OLS	WLS	RF	OLS	WLS	RF	OLS	WLS	RF
January	15.580 (0.672)	15.581 (0.672)	17.012 (0.545)	15.648 (0.454)	15.639 (0.452)	16.880 (0.319)	15.907 (0.233)	15.903 (0.389)	17.246 (0.334)
February	15.942 (0.611)	15.944 (0.611)	17.288 (0.489)	15.565 (0.472)	15.566 (0.471)	16.916 (0.302)	16.024 (0.253)	16.020 (0.328)	17.327 (0.318)
March	15.764 (0.709)	15.766 (0.709)	17.163 (0.574)	15.574 (0.587)	15.575 (0.588)	16.884 (0.384)	16.060 (0.523)	16.065 (0.448)	17.381 (0.341)
April	15.603 (0.859)	15.607 (0.860)	17.032 (0.754)	15.470 (0.490)	15.466 (0.491)	16.838 (0.347)	16.134 (0.271)	16.129 (0.385)	17.431 (0.303)
May	15.687 (0.625)	15.689 (0.625)	17.077 (0.486)	15.581 (0.391)	15.577 (0.393)	16.943 (0.292)	16.147 (0.402)	16.147 (0.391)	17.433 (0.289)
June	15.714 (0.619)	15.718 (0.620)	17.042 (0.494)	15.624 (0.357)	15.619 (0.357)	16.982 (0.344)	16.073 (0.407)	16.075 (0.304)	17.380 (0.262)
July	15.740 (0.625)	15.743 (0.625)	17.047 (0.395)	15.592 (0.309)	15.588 (0.310)	16.988 (0.256)	16.042 (0.357)	16.041 (0.308)	17.393 (0.308)
August	15.677 (0.607)	15.671 (0.605)	16.963 (0.480)	15.668 (0.368)	15.663 (0.368)	17.076 (0.331)	16.040 (0.312)	16.034 (0.345)	17.386 (0.349)
September	15.517 (0.405)	15.509 (0.405)	16.878 (0.365)	15.742 (0.502)	15.733 (0.551)	17.121 (0.466)	16.072 (0.308)	16.064 (0.328)	17.399 (0.266)
October	15.463 (0.706)	15.454 (0.704)	16.819 (0.536)	15.915 (0.570)	15.936 (0.439)	17.150 (0.227)	15.983 (0.225)	15.978 (0.291)	17.374 (0.193)
November	15.589 (0.472)	15.580 (0.470)	16.880 (0.342)	15.883 (0.503)	15.884 (0.507)	17.214 (0.406)	15.994 (0.353)	15.985 (0.470)	17.420 (0.434)
December	15.659 (0.483)	15.651 (0.482)	16.923 (0.382)	15.874 (0.485)	15.883 (0.409)	17.207 (0.328)	16.027 (0.287)	16.025 (0.316)	17.455 (0.241)

39. R-squared (R^2), which indicates the percentage of the variance in the dependent variable that the independent variables explain collectively, is very consistent between folds and between months and years. The R^2 between WLS and OLS is virtually identical (Table 3), indicating that the models are able to explain similar amounts of the variation in prices observed. There is a lower R^2 for random forest, along with a slightly higher standard deviation, suggesting the model accounts for less of the variation in prices observed than WLS or OLS.

Table 4: Average R^2 (and its standard deviation) across 10 folds

AVERAGE R^2 ACROSS 10 FOLDS									
	2018			2019			2020		
MONTH	OLS	WLS	RF	OLS	WLS	RF	OLS	WLS	RF
January	0.839 (0.000)	0.838 (0.000)	0.627 (0.006)	0.833 (0.000)	0.835 (0.000)	0.631 (0.008)	0.839 (0.000)	0.842 (0.000)	0.621 (0.008)
February	0.840 (0.000)	0.840 (0.000)	0.631 (0.007)	0.840 (0.000)	0.839 (0.000)	0.629 (0.008)	0.847 (0.000)	0.847 (0.000)	0.627 (0.006)
March	0.839 (0.000)	0.839 (0.000)	0.629 (0.007)	0.840 (0.000)	0.839 (0.000)	0.629 (0.006)	0.848 (0.000)	0.848 (0.000)	0.628 (0.006)
April	0.840 (0.000)	0.840 (0.000)	0.632 (0.006)	0.839 (0.000)	0.839 (0.000)	0.627 (0.007)	0.850 (0.000)	0.849 (0.000)	0.631 (0.007)

May	0.841 (0.000)	0.840 (0.000)	0.634 (0.007)	0.839 (0.000)	0.839 (0.000)	0.622 (0.004)	0.850 (0.000)	0.850 (0.000)	0.632 (0.006)
June	0.841 (0.000)	0.841 (0.000)	0.635 (0.005)	0.840 (0.000)	0.840 (0.000)	0.622 (0.005)	0.848 (0.000)	0.849 (0.000)	0.631 (0.008)
July	0.841 (0.000)	0.841 (0.000)	0.633 (0.009)	0.841 (0.000)	0.841 (0.000)	0.621 (0.006)	0.847 (0.000)	0.849 (0.000)	0.628 (0.007)
August	0.835 (0.000)	0.837 (0.000)	0.632 (0.006)	0.840 (0.000)	0.841 (0.000)	0.618 (0.007)	0.846 (0.000)	0.847 (0.000)	0.628 (0.006)
September	0.835 (0.000)	0.837 (0.000)	0.628 (0.007)	0.839 (0.000)	0.840 (0.000)	0.620 (0.007)	0.845 (0.000)	0.847 (0.000)	0.628 (0.005)
October	0.835 (0.000)	0.837 (0.000)	0.632 (0.007)	0.825 (0.000)	0.833 (0.000)	0.619 (0.007)	0.845 (0.000)	0.847 (0.000)	0.625 (0.008)
November	0.834 (0.000)	0.837 (0.000)	0.634 (0.008)	0.837 (0.000)	0.840 (0.000)	0.622 (0.004)	0.846 (0.000)	0.848 (0.000)	0.624 (0.008)
December	0.833 (0.000)	0.836 (0.000)	0.633 (0.008)	0.838 (0.000)	0.841 (0.000)	0.621 (0.004)	0.847 (0.000)	0.848 (0.000)	0.623 (0.007)

Other considerations to make when choosing a model

40. When considering models for underlying the production of official statistics, further quality dimensions must be considered:
    a) Accessibility and clarity: The behaviour of some models is easy to describe, understand and extend. In the case of the models presented here, those based on linear models are well established in existing literature, while random forest models currently require specialised expertise in machine learning.
    b) Coherence and comparability: The House Price Index uses a Weighted Least Squares method, and so although we propose to use OLS for the rental prices, this does have internal similarity with other housing statistics produced by the ONS. StatCan recently moved from a matched pairs model to a hedonic model using OLS when calculating their rental prices.
    c) All models presented here can be used for extrapolating their predictions to new datasets. However, the random forests are relatively large and require a deployment to be of use, while the behaviour and predictions of the linear models can be shared as a set of coefficients.
    d) Timeliness and punctuality: Models that are easier to troubleshoot and computationally less intensive tend to be easier to implement in the production of official statistics. The monthly production round provides us with around one week to produce, quality assure and evaluate our statistics. More complex models can take more time to run and evaluate.
    e) Longer term support: It is usually easier to find support for, and to maintain, simpler, well understood models.

Proposed model

41. We propose to use an ordinary least squares model in our hedonic regression. Both OLS and WLS perform similarly in our analysis presented above and both outperform Random Forest. As OLS is a more accessible model, we propose to use this to reduce complexities and run time.

Investigation into the effect of interaction terms

42. Interaction effects occur when the effect of one variable depends on the value of another variable. The Random Forest model automatically accounts for interaction effects in the model, whereas OLS and WLS don't.
43. We considered using interaction effects in our model because the rental market can be complex, and prices can not only depend on whether it is a detached house, for example, but it may also depend on whether it is a detached house in an affluent area.
44. The following analysis considers an WLS model including the following interactions:
    - $\text{Acorn} \times (\text{type} + \text{beds} + \text{ln\_area})$
45. This analysis is demonstrated using a WLS model. Given the similarities between the OLS and WLS models, the OLS results would be comparable.
46. The differences between the models excluding and including interaction terms are minimal (Figure 4).
47. In Figures 4 and 5, the solid line shows the results of the model without interactions and the dotted line shows the results of the model with interactions.

Figure 4: Rental price index for England and Wales for WLS models with and without interactions

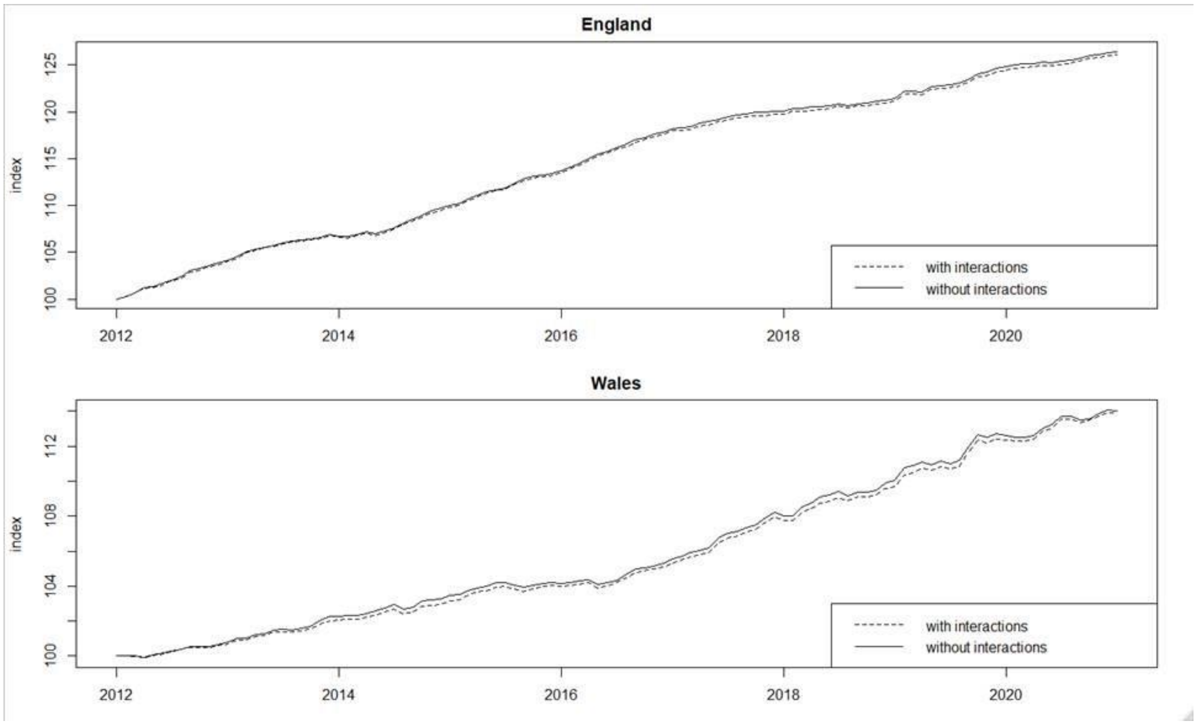
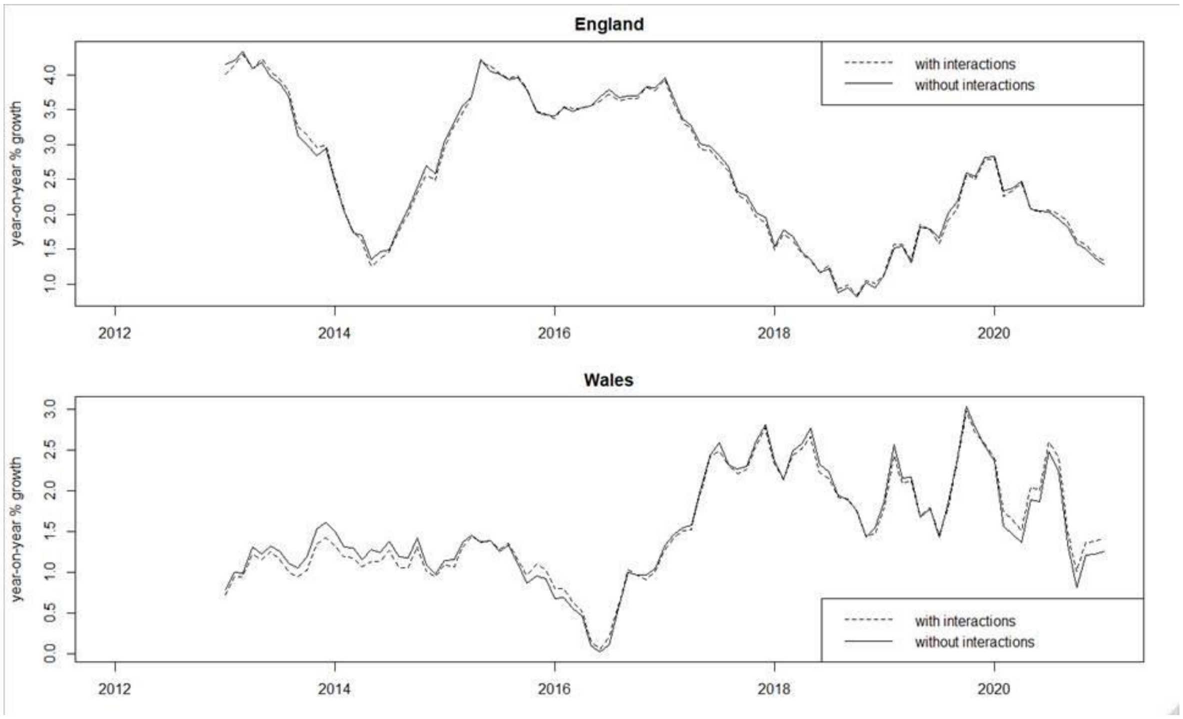


Figure 5: Annual rental price growth for England and Wales for WLS models with and without interactions



48. Table 5 shows a summary of differences between the results of the models with and without interactions for each country and each English region.

Table 5: Differences between models with interactions and without interactions

	Differences in indices (with interactions minus without interactions), index points			Differences in year-on-year % growth (with interactions minus without interactions), percentage points		
Country/Region	minimum	mean	maximum	minimum	mean	maximum
England	-0.38	-0.23	0.00	-0.14	-0.01	0.15
Wales	-0.39	-0.21	0.00	-0.19	-0.01	0.20
North East	-0.77	-0.52	0.00	-0.27	-0.06	0.12
North West	-0.27	-0.17	0.01	-0.18	-0.02	0.07
Yorkshire and The Humber	-0.75	0.01	0.53	-0.90	0.10	0.96
East Midlands	-0.45	-0.25	0.00	-0.21	-0.01	0.15
West Midlands	-0.35	-0.18	0.05	-0.29	-0.01	0.14
East	-0.33	-0.19	0.00	-0.18	-0.01	0.16
London	-0.82	-0.33	0.13	-0.44	-0.03	0.55
South East	-0.36	-0.19	0.07	-0.27	-0.01	0.18
South West	-0.36	-0.16	0.04	-0.22	0.00	0.18

49. Including all available 2-way interaction terms in the model increases the R^2 value slightly. Using no interactions has an average R-squared value of 0.84 in 2019, including all 2-way interactions

has an average R-squared value of 0.85 in 2019. The difference in the average R-squared value is minimal.

50. The models are showing little difference in the indices and growth rates when interactions with Acorn are included. This suggests a less complex model, without interaction terms would be suitable.
51. A secondary benefit of using a less complex model is that the computation time is faster than when running it with interactions. When considering this as a method to use in monthly production it is preferable to minimise the computation time and reduce method complexity where this is not detrimental to the quality of the results.
52. Following our analysis, we therefore propose not to include interaction terms in the model.

Acorn

53. Acorn is a segmentation tool which categorises the UK's neighbourhoods and postcodes into demographic types. For the purpose of this work, the Acorn group is used to classify property according to the postcode where it is situated, for example, a property (based on the postcode) could be classified in Acorn category "lavish lifestyles" through to category "difficult circumstances". The Acorn variable level used is the Acorn group which is a categorical variable with 17 different classes. The Acorn group does not enter the regression as a single variable, but essentially as 16 (n-1) separate binary variables.
54. We proposed to use Acorn as one of our independent variables, however, concerns were raised in previous APCP-T meetings over the use of Acorn. This is due to the potential endogeneity of Acorn with rental prices across local authorities.
55. Suggestions from the panel were to test for endogeneity in Acorn with a Hausman specification test. The Durbin-Wu-Hausman test works by comparing the regression coefficients of the ACORN variables under OLS with those on the fitted values under 2-stage least squares (2SLS). If Acorn is endogenous, then it will be biased and differ in a significant way from the robust/unbiased coefficients calculated under 2SLS.
56. Performing this test with Acorn is difficult because the Acorn variables are binary (0,1) and ln(rents) (our dependent variable) is continuous. Therefore, stage 1 in 2SLS would need to be estimated using a non-linear model – in this case a Probit model. Whereas the second stage where ln(rents) is the dependent variable is estimated by OLS.
57. The Durbin-Wu-Hausman test doesn't apply to cases where a continuous variable (ln(rents)) and a set of binary variables are being tested. Hausman describes this as the 'forbidden regression' problem. An ONS economist has tried to solve this problem but did not find an easy solution.
58. On top of this, it would be hard to apply an appropriate instrument to use for the Acorn variable in the test because the instrument needs to be something directly causal to Acorn but not ln(rents). If we use an instrumental variable(s) without these properties then the 2SLS estimates will suffer the same biases as OLS/WLS, and the Durbin-Wu-Hausman test will fail to find endogeneity even if it exists. Not using Acorn may be inserting more omitted variable bias into the regression than the potential endogeneity bias being corrected for.

Generalised variance inflation factor (GVIF)

59. The variance inflation factor may be used to detect collinearity in the explanatory variables of a linear regression model. In the case of the models discussed in this document, most of the explanatory variables are categorical, and so they are represented by multiple dummy variables in the model design matrices. To overcome this obstacle in the use of variance inflation factors, we use the generalized variance inflation factor (GVIF) as introduced by John Fox and Georges Monette (Fox & Monette, 1992). This supplies a single measure for each variable as opposed to one for each (non-reference) level of each factor. To make these measures comparable, we use Fox and Monette's suggested adjustment: $GVIF^{1/(2df)}$. For variables with one degree of freedom, this is equal to the square root of the variance inflation factor.
60. A higher GVIF means that the variable is more correlated with others, it is suggested that a GVIF of less than square root of 5 is acceptable, but it is important to note that we shouldn't use rules of thumb. Tables 6 and 7 show a summary of the adjusted GVIF for each variable in our models, as calculated for every month from 2012 to 2020. The values for the models are consistent across the period and show no cause for concern.

Table 6: Adjusted GVIF for Ordinary Least Squares

Variable	Minimum adjusted GVIF	Mean adjusted GVIF	Maximum adjusted GVIF
Local authority	1.0028	1.0030	1.0032
ACORN group	1.0667	1.0720	1.0743
Property type	1.2919	1.3060	1.3145
Property age	1.0470	1.0486	1.0509
Number of bedrooms	1.1552	1.1574	1.1595
Furnished status	1.0675	1.0864	1.1068
Natural log of floor area	2.1616	2.1938	2.2379

Table 7: Adjusted GVIF for Weighted Least Squares

Variable	Minimum adjusted GVIF	Mean adjusted GVIF	Maximum adjusted GVIF
Local authority	1.0028	1.0030	1.0031
ACORN group	1.0665	1.0717	1.0742
Property type	1.2919	1.3059	1.3142
Property age	1.0472	1.0485	1.0502
Number of bedrooms	1.1551	1.1572	1.1593
Furnished status	1.0672	1.0860	1.1048
Natural log of floor area	2.1587	2.1916	2.2327

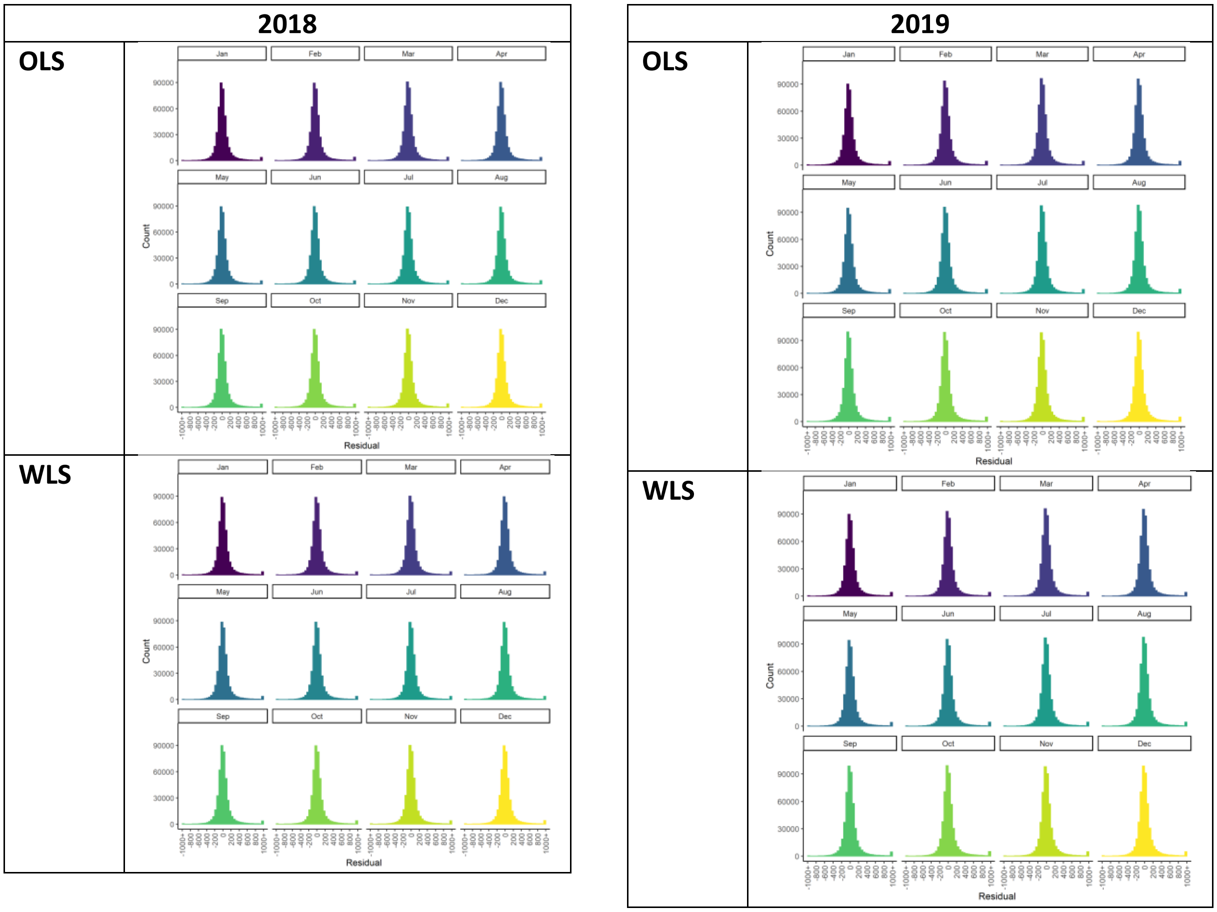
K-fold cross-validation without Acorn

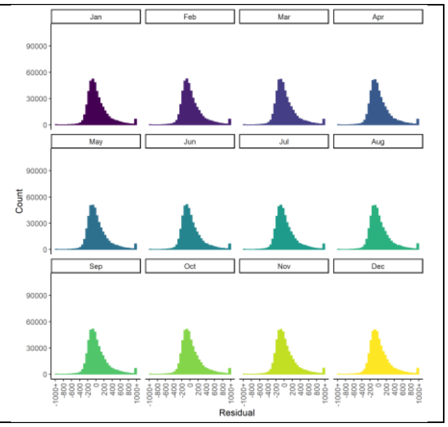
61. To further consider whether Acorn should be used in our model, we performed a K-fold cross validation with and without Acorn. The results with Acorn included can be found in Figure 3,

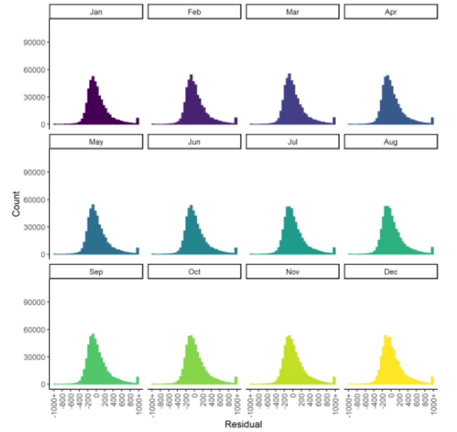
Table 3 and Table 4. The results for without Acorn included can be found in Figure 6, Table 9 and Table 10.

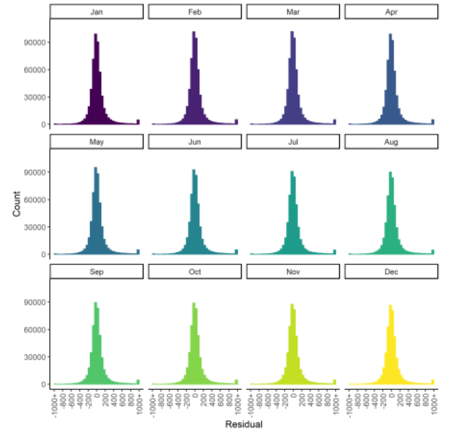
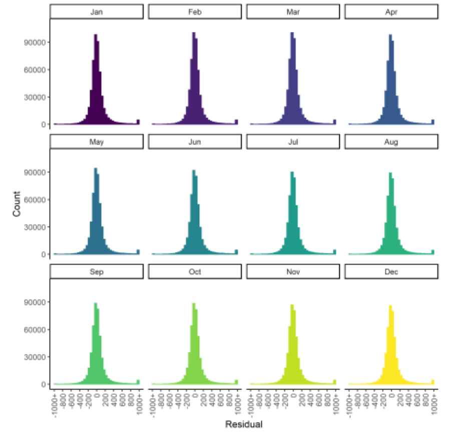
62. The distribution of residuals for each model across the years is similar to the distribution with Acorn included. Like with Acorn included, it is evident that the majority of residuals for OLS and WLS are centred around or close to 0. The distribution of residuals for the random forest is mainly centred around or close to 0 however it is slightly positively skewed. For all models there is a cluster of residuals on the right tail and closer examination of observed rent prices and the corresponding predicted rent showed that the models tend to underestimate very high rents (approximately 1% to 2% of the data falls within the highest bin, see Table 8 for a detailed breakdown).

Figure 6: Distribution of residuals



RF	

RF	

2020	
OLS	
WLS	

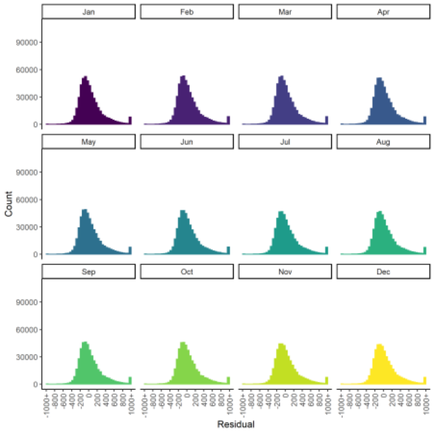
RF	

Table 8: Percentage of data falling into the highest residual bin (1000+) (without Acorn)

Percentage of data falling within highest residual bin (1000+) (without Acorn)									
	2018			2019			2020		
Month	OLS	WLS	RF	OLS	WLS	RF	OLS	WLS	RF
January	0.94	0.93	1.68	1.03	1.02	1.73	1.08	1.07	1.83
February	0.97	0.96	1.71	1.03	1.02	1.78	1.10	1.08	1.90
March	0.95	0.94	1.68	1.03	1.02	1.75	1.10	1.07	1.90
April	0.95	0.94	1.69	1.01	1.00	1.71	1.10	1.08	1.92
May	0.95	0.94	1.69	1.02	1.01	1.69	1.11	1.08	1.91
June	0.96	0.95	1.67	1.02	1.01	1.72	1.12	1.09	1.94
July	0.96	0.95	1.69	1.01	1.00	1.74	1.12	1.09	1.94
August	0.98	0.97	1.74	1.03	1.02	1.77	1.13	1.09	1.94
September	0.96	0.95	1.70	1.06	1.05	1.80	1.13	1.10	1.94
October	0.99	0.96	1.71	1.12	1.12	1.82	1.12	1.10	1.95
November	1.01	0.99	1.75	1.08	1.07	1.83	1.12	1.09	1.95
December	1.03	1.02	1.76	1.10	1.08	1.84	1.12	1.09	2.01

63. As was the case with the Acorn variable included, WLS and OLS have a lower mean RMSE than Random Forest, but the difference isn't substantial. Not having the Acorn variable present results in a slightly larger RMSE across all models.

Table 9: Average RMSE and its standard deviation across 10 folds (without Acorn)

AVERAGE RMSE ACROSS 10 FOLDS (without acorn) with standard deviation in brackets									
	2018			2019			2020		
MONTH	OLS	WLS	RF	OLS	WLS	RF	OLS	WLS	RF
January	15.949 (0.668)	15.802 (0.681)	18.136 (0.553)	15.999 (0.468)	15.951 (0.680)	18.088 (0.450)	16.244 (0.222)	16.149 (0.421)	18.391 (0.344)

February	16.335 (0.612)	16.199 (0.647)	18.402 (0.493)	15.905 (0.480)	15.883 (0.413)	18.131 (0.239)	16.344 (0.230)	16.253 (0.398)	18.533 (0.270)
March	16.144 (0.684)	16.024 (0.566)	18.336 (0.456)	15.918 (0.591)	15.903 (0.484)	18.077 (0.263)	16.384 (0.499)	16.317 (0.272)	18.598 (0.220)
April	15.986 (0.857)	15.882 (0.625)	18.216 (0.385)	15.776 (0.455)	15.751 (0.426)	18.047 (0.321)	16.450 (0.277)	16.364 (0.328)	18.602 (0.275)
May	16.072 (0.617)	15.932 (0.677)	18.234 (0.571)	15.892 (0.373)	15.863 (0.440)	18.078 (0.336)	16.461 (0.391)	16.364 (0.363)	18.683 (0.224)
June	16.072 (0.638)	15.956 (0.479)	18.221 (0.265)	15.928 (0.365)	15.907 (0.341)	18.134 (0.264)	16.403 (0.411)	16.288 (0.436)	18.590 (0.248)
July	16.106 (0.647)	15.979 (0.600)	18.279 (0.476)	15.890 (0.303)	15.865 (0.308)	18.193 (0.303)	16.371 (0.349)	16.251 (0.358)	18.602 (0.244)
August	16.056 (0.596)	15.892 (0.770)	18.148 (0.581)	15.978 (0.352)	15.934 (0.238)	18.209 (0.234)	16.372 (0.343)	16.260 (0.162)	18.588 (0.159)
September	15.880 (0.422)	15.728 (0.634)	17.951 (0.552)	16.058 (0.482)	15.994 (0.529)	18.278 (0.363)	16.399 (0.332)	16.285 (0.251)	18.647 (0.192)
October	15.815 (0.720)	15.802 (0.556)	18.007 (0.468)	16.265 (0.545)	16.209 (0.473)	18.330 (0.260)	16.308 (0.229)	16.208 (0.387)	18.605 (0.267)
November	15.949 (0.463)	15.910 (0.550)	18.084 (0.295)	16.222 (0.483)	16.175 (0.273)	18.361 (0.108)	16.311 (0.349)	16.203 (0.367)	18.572 (0.288)
December	16.018 (0.501)	15.973 (0.682)	18.107 (0.432)	16.216 (0.474)	16.147 (0.362)	18.324 (0.351)	16.325 (0.279)	16.234 (0.305)	18.659 (0.255)

64. R-squared (R^2) is very consistent between folds and between months and years. R^2 between WLS and OLS is virtually identical. The R^2 for random forest is smaller, along with a slightly higher standard deviation.
65. Comparing the R^2 values for the models without Acorn to the R^2 values to the models with Acorn, the model without Acorn performs slightly worse than with Acorn. This suggests including Acorn in the model would be appropriate.

Table 10: Average R^2 across 10 folds (without Acorn)

AVERAGE R^2 ACROSS 10 FOLDS (without acorn) with the standard deviation in brackets									
	2018			2019			2020		
MONTH	OLS	WLS	RF	OLS	WLS	RF	OLS	WLS	RF
January	0.820 (0.000)	0.820 (0.000)	0.493 (0.006)	0.814 (0.000)	0.817 (0.000)	0.494 (0.006)	0.821 (0.000)	0.824 (0.000)	0.471 (0.005)
February	0.822 (0.000)	0.821 (0.000)	0.497 (0.005)	0.822 (0.000)	0.821 (0.000)	0.49 (0.006)	0.829 (0.000)	0.829 (0.000)	0.473 (0.008)
March	0.821 (0.000)	0.820 (0.000)	0.491 (0.005)	0.822 (0.000)	0.821 (0.000)	0.486 (0.008)	0.831 (0.000)	0.830 (0.000)	0.473 (0.005)
April	0.823 (0.000)	0.822 (0.000)	0.492 (0.005)	0.821 (0.000)	0.821 (0.000)	0.485 (0.005)	0.833 (0.000)	0.832 (0.000)	0.472 (0.007)
May	0.823 (0.000)	0.822 (0.000)	0.493 (0.005)	0.821 (0.000)	0.821 (0.000)	0.49 (0.009)	0.833 (0.000)	0.833 (0.000)	0.477 (0.006)

June	0.824 (0.000)	0.822 (0.000)	0.497 (0.006)	0.822 (0.000)	0.821 (0.000)	0.481 (0.006)	0.831 (0.000)	0.832 (0.000)	0.481 (0.005)
July	0.823 (0.000)	0.822 (0.000)	0.495 (0.007)	0.823 (0.000)	0.823 (0.000)	0.475 (0.004)	0.830 (0.000)	0.831 (0.000)	0.477 (0.006)
August	0.816 (0.000)	0.819 (0.000)	0.493 (0.005)	0.822 (0.000)	0.822 (0.000)	0.473 (0.007)	0.829 (0.000)	0.830 (0.000)	0.479 (0.006)
September	0.816 (0.000)	0.818 (0.000)	0.493 (0.007)	0.821 (0.000)	0.821 (0.000)	0.475 (0.005)	0.828 (0.000)	0.830 (0.000)	0.472 (0.004)
October	0.816 (0.000)	0.819 (0.000)	0.497 (0.005)	0.805 (0.000)	0.813 (0.000)	0.471 (0.006)	0.828 (0.000)	0.830 (0.000)	0.472 (0.005)
November	0.816 (0.000)	0.819 (0.000)	0.495 (0.005)	0.818 (0.000)	0.821 (0.000)	0.474 (0.006)	0.830 (0.000)	0.832 (0.000)	0.477 (0.003)
December	0.815 (0.000)	0.817 (0.000)	0.496 (0.006)	0.819 (0.000)	0.822 (0.000)	0.475 (0.005)	0.831 (0.000)	0.832 (0.000)	0.468 (0.008)

Conclusions

66. Following the comprehensive analysis presented above, we propose to use a simple Ordinary Least Squares model with no interaction terms:

    $\log(p_i) = k + \sum_j \beta_j x_j^i + e_j$, where:

    - p_i is the rental price of property i
    - k is a constant
    - x_j^i indicates whether property i has the characteristic j (such as detached property). If so, it takes the value 1, otherwise it takes the value 0 (except for floor area where it takes the floor area in m^2
    - β_j is the coefficient associated with characteristic j
    - e_j is the statistical error term

67. In the case of rents development, we propose to include the following the price-determining characteristics in the model:

    - Number of bedrooms
    - Floor area (in m^2)
    - Property type (Flat/Maisonette, Detached, Semi-detached, Terraced)
    - Furnished status
    - ACORN group classification
    - Local authority
    - Property age bracket

68. The methodology for the new rental price statistics measure will be reviewed every 5 years to ensure it is still the best method for the data we are receiving.

References

Diewert E, Shimizu C, (2021) *Consumer Prices Index Theory, CHAPTER 10: THE TREATMENT OF DURABLE GOODS AND HOUSING* https://econ2017.sites.olt.ubc.ca/files/2021/05/pdf_paper_diewert-erwin_IMFCPIChapter10.pdf

Fox, J., & Monette, G. (1992). Generalized Collinearity Diagnostics. Journal of the American Statistical Association, 87(417), 178-183. doi:10.2307/2290467

Groenwald, R.H.H, et. al. CMAJ. *Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis*, 2012 Aug 7; 184(11): 1265–1269.

Harrell, Frank E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer, 2001. Print.

Authors

Natalie Jones, Waliur Rahman, Joni Karanka, Laura Dimond, Annabel Summerfield, Helen Verey

Prices Division and ONS Methodology, October 2021

List of Annexes

Annex A	Proportion of the CPIH, CPI and RPI basket impacted
Annex B	Advice from ESCoE received in April 2021
Annex C	Data cleaning methodology
Annex D	14-month rolling window
Annex E	Comparison to the IPHRP
Annex F	Distribution of residuals for OLS, WLS and Random Forest with Acorn included

Annex A: Proportion of the CPIH, CPI and RPI basket impacted

Inflation measure	Proportion of the 2021 basket impacted
CPIH	OOH: 18.5%, Actual rentals: 7.4%
CPI	9.4% (actual rents)
RPI	7.9% (actual rents)

Annex B: Advice from ESCoE received in April 2021

- The ONS Rental Price methodology is far from being a standard one. The description of the data set and the method of index construction are very difficult to follow.
- We would not use observations that had imputed values for characteristics.
- We would not use the Acorn variable. This variable does not describe a characteristic of the property; it describes a characteristic of a tenant. It is a black box variable.
- It is not necessary to use weighted least squares in the context of property hedonic regressions. Each property is equally important and has a weight of unity. Weighting is very useful in

constructing indexes using scanner data since quantity weights can vary tremendously over the products in scope.

- On the issue of using spatial coordinates versus local area dummy variables: our experience is that it is not necessary to use spatial coordinate techniques. They are much more complex to implement and in the end, simple local area dummy variables will tend to generate very similar indexes; See Diewert and Shimizu (2021).
- We agree with the ONS that it is not necessary to use Ridge regressions in the property context.
- We agree with the ONS in rejecting random forest models. These models have a black box character and are difficult to explain to the public. And as noted by the ONS, they can give counterintuitive results.
- On the issue of using interaction terms for multiple explanatory variables that use discrete categories. If our advice is followed and all available rental data are used (instead of just a sample of it), there should be ample degrees of freedom to introduce interaction terms. A multicollinearity problem can occur if there are too many explanatory variables. This matters since it is important that the model give “reasonable” estimates for the monthly or quarterly structure depreciation rate. Thus our advice is to start out with a relatively simple model with relatively few explanatory variables, check the resulting estimates for depreciation rates and add additional explanatory variables until a reasonably high R square is obtained and hopefully, with a resulting reasonable depreciation rate.
- A key recommendation is that the ONS should endeavour to obtain land plot areas for at least a sample of their property data base. In our view, the main explanatory variables which explain rents are: (i) the location of the property (a local area dummy variable is required); (ii) the type of property (detached, row, and so on); (iii) the floor space of the structure; (iv) the land area of the property; (v) the age of the structure (this can be approximate but of course, once a property is in the sample, it ages one month or one quarter for each additional period that it remains in the sample of properties) and (vi) the type of construction of the structure. Of course, there can be a host of other characteristics that help explain property prices but our experience is that typically, the addition of these additional characteristics will not materially affect the resulting price indexes.

Annex C: Data cleaning

69. Automatic and manual data cleaning exercises are carried out monthly on the live dataset, and annually on the fixed basket. These checks remove extreme observations and ensure that reported values are reasonable. For example, a check that the floor area is greater than or equal to the legal minimum threshold for bedroom floor area multiplied by the number of bedrooms in the property is performed.
70. A check is performed on any flats that have a high number of bedrooms. This identifies where flats have been matched to the main property in the property attributes dataset, rather than the flat at the same address. The analyst inspects the property attributes dataset to see if it contains an appropriate property match. If it does, this is flagged as a “mismatch” and the property characteristics are corrected. If there is no matching property in the admin data, then the property characteristics that would have been obtained from the admin data are set to missing and imputed in the following step before a hedonic regression model is fitted to the data.

Annex D: 14-month rolling window

71. The aim of the rental prices development work is to develop a stock-based measure. This represents what all tenants are paying for their privately rented property.

72. When a rental price is collected, it is assumed to be valid for 14 months from its entry date into the system, or until an update is received. A 14-month validity period is used as it balances typical contract lengths (which tend to be either 6, 12, 18 or 24 months) against operational practices. In particular, the time it takes rent officers to follow up the same property. There is an emphasis on following up properties between 12 and 14 months since they were last collected, and therefore there are methodological benefits (in terms of significantly improving the number of property updates) to using a 14-monthly validity period over using a 6-month period (which would only capture a very small portion of updates) or a 12-month period (which would capture around half of the updates). Moreover, the 14-month validity period has the additional advantage of mitigating for properties that were originally rented some time ago at much lower prices, as well as limiting the effect of depreciation.

Annex E: Comparison to IPHRP

73. The rents development work is showing greater growth than the IPHRP over the time period

Figure A: Comparison of rents development index to IPHRP

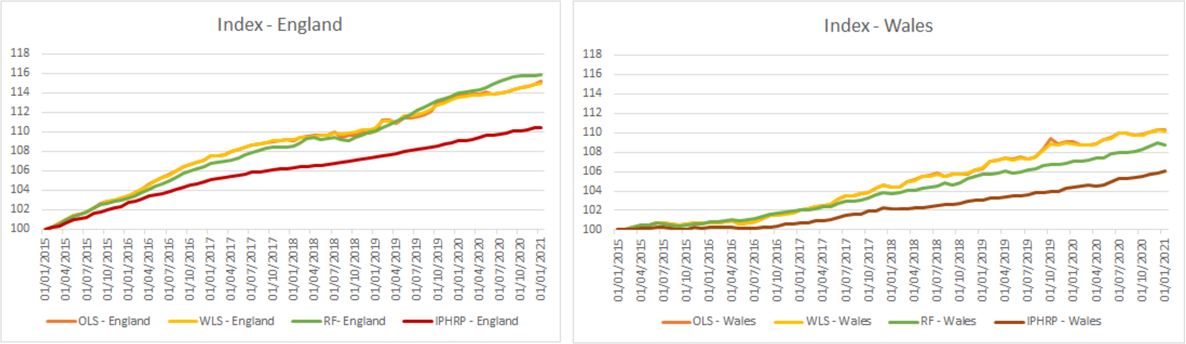
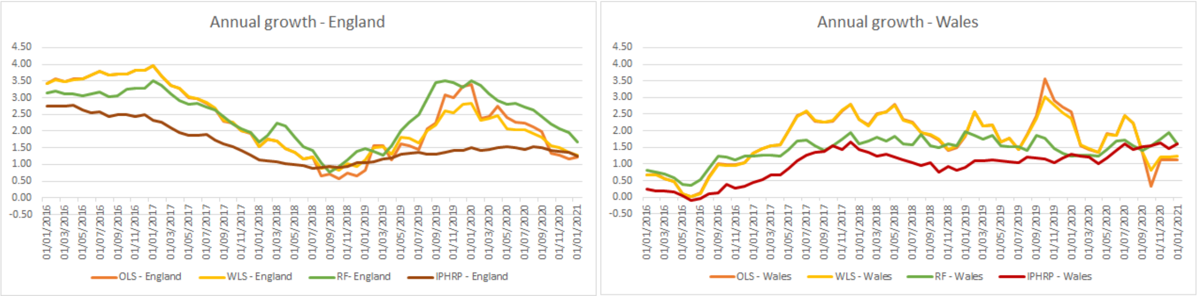


Figure B: Comparison of rents development growth rate to IPHRP



74. There are many differences in methodology that could be impacting this:

- We are now utilising all available data

- Weighting at local authority
- Change of methodology from matched pairs to hedonic regression

75. Further work is required to understand these differences.

Utilising all available data

76. The matched pairs approach randomly splits the rental prices collected in to two datasets on a yearly basis, 50% goes into the sample and 50% goes into the substitution pool. The sample is what gets used to calculate the IPHRP index.
77. The length of time that a price for a selected property remains in the sample and substitution pool is monitored. They are valid for a maximum of 14 months from entry onto the system (if no updated price is collected).
78. In the matched pairs approach, data collected in the month are matched to records that are currently in the sample. Any unmatched properties are moved to the substitution pool. The price of an existing property in the sample is updated when a match is made, once this has updated the price is valid for up to 14 months again.
79. Properties that are outside the validity period (>14 months) are removed from the sample and are replaced with a comparable replacement from the substitution pool, this property maintains its entry date on the system when moved into the sample.
80. In the matched pairs approach, we will only see a newly collected rental price entering the sample at two points:
    - An updated rent is collected for the existing tenant (contract renewal) – this could be the same price
    - A new rent is collected as the property has been re-let to a new tenant since the last visit (this price could have been agreed at any time)
81. Any new rental that is collected that isn’t already in the sample is put into the substitution pool
82. In the rents development work, we still consider the 14-month period, and update any rental prices with newly collected data, however all data collected are used in the model, so we will now be including any new properties collected by rent officers that weren’t previously collected. These may represent a flow (a new rental on the market) or an existing rental that has been newly collected
83. The rents development model is using all data available in the latest month, and so there are more opportunities to pick up the dynamic price changes taking place.

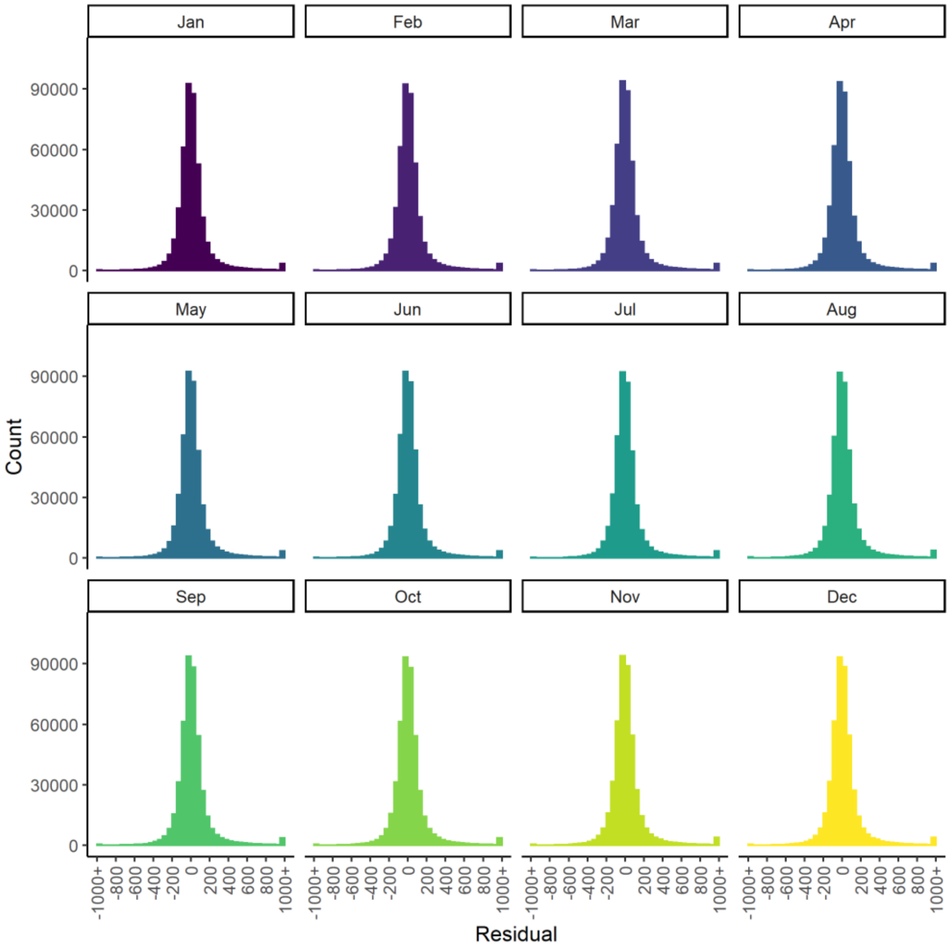
Weighting at local authority

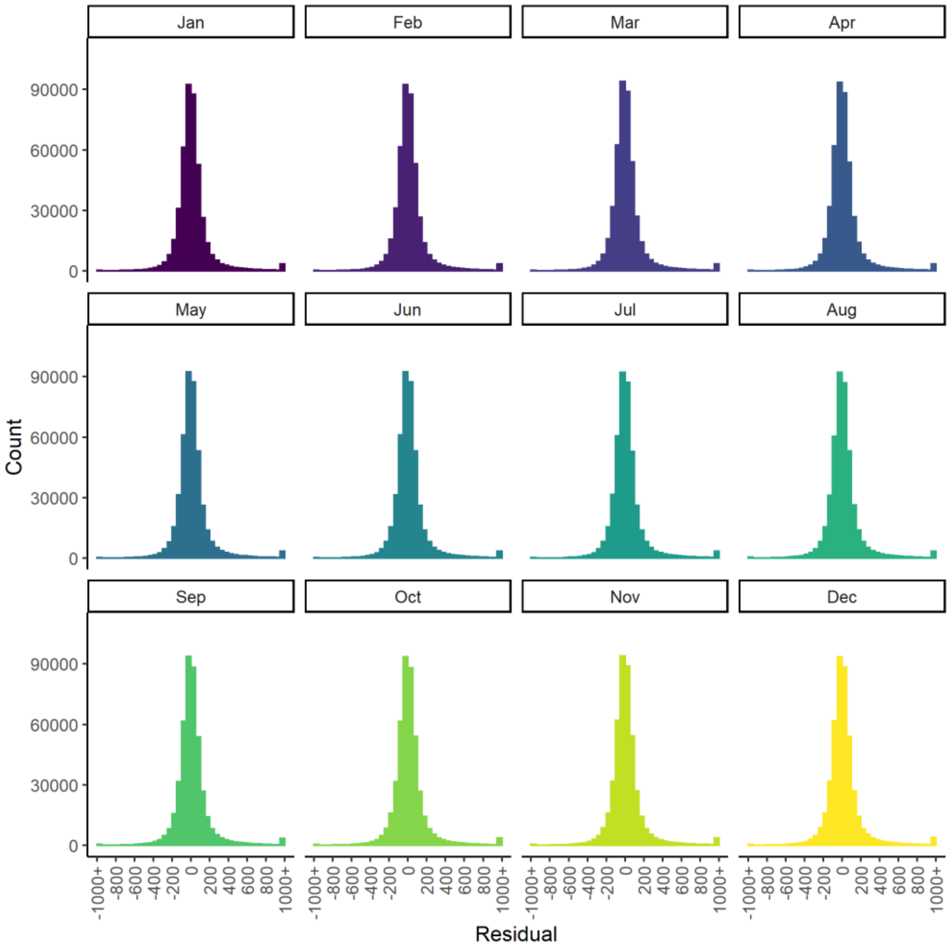
84. In the IPHRP, we weight our rental prices at a regional level, however we can now weight at a local authority level. The benefit of this is that we can account for any under- or over-sampling.
85. This means that if a local authority that is under-sampled is showing greater price increases across the time period on average than what is being seen at a regional level, the increases in price will be greater in the rents development.

Annex F: Distribution of residuals for OLS, WLS and Random Forest with Acorn included

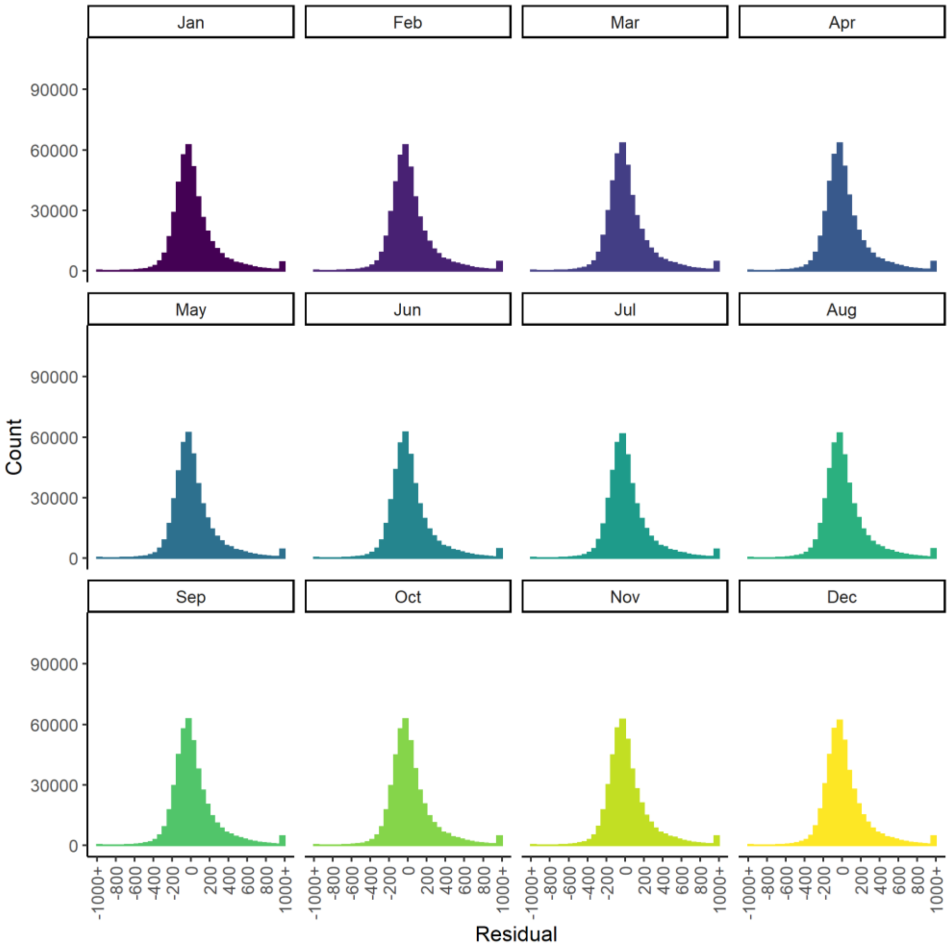
A larger image of the distribution of residuals for OLS, WLS and Random Forest with Acorn included can be found in Figure 3.

Figure 3: Distribution of residuals

	2018
OLS	

WLS	

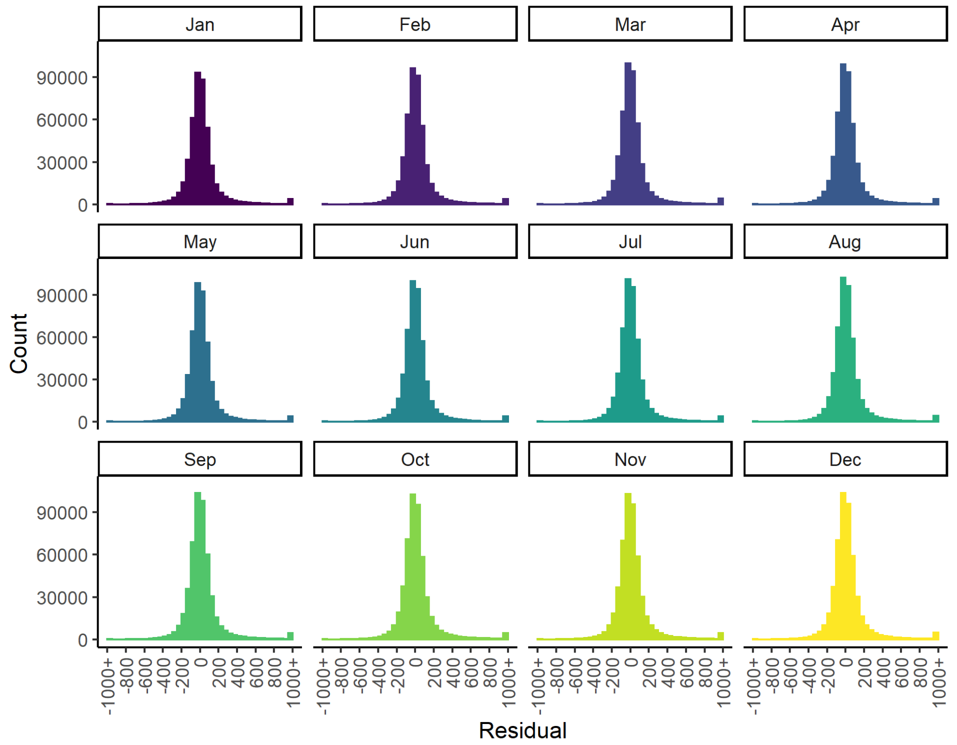
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Count
90000
60000
30000
0
-1000+
-800
-600
-400
-200
0
200
400
600
800
1000+
Residual

RF	

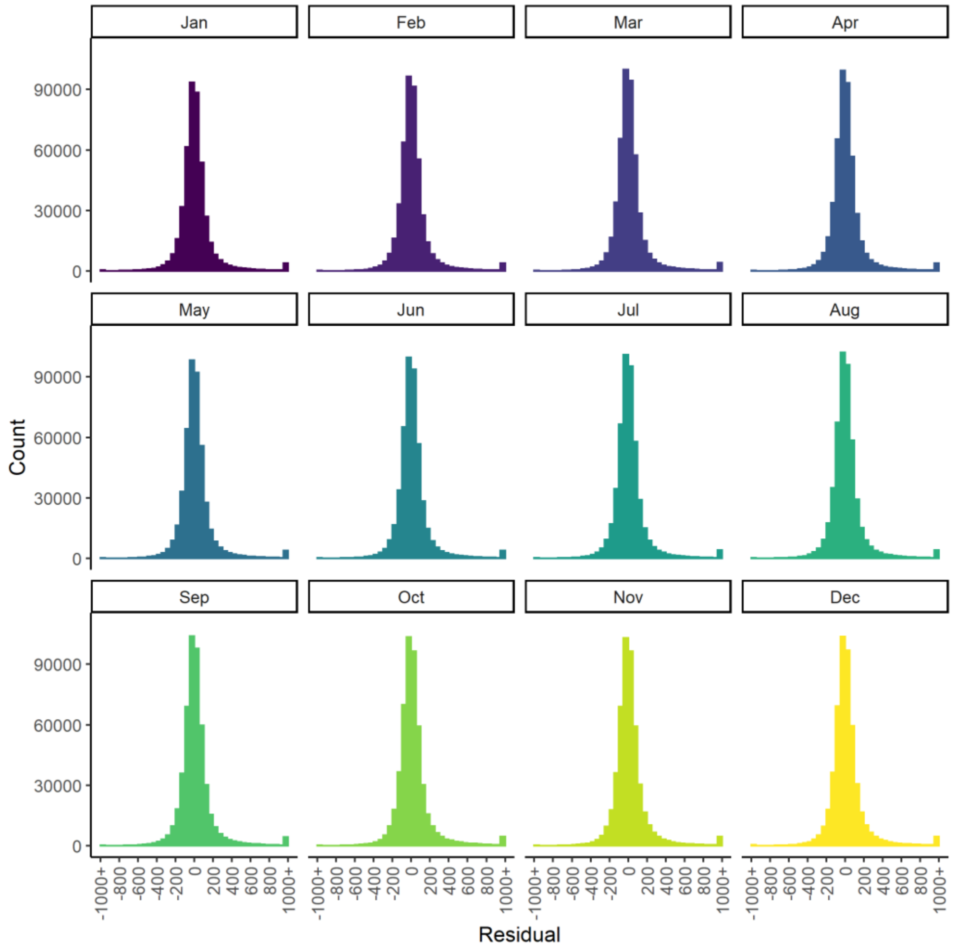
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
90000
60000
30000
0
Count
-1000-
-800
-600
-400
-200
0
200
400
600
800
1000+
Residual

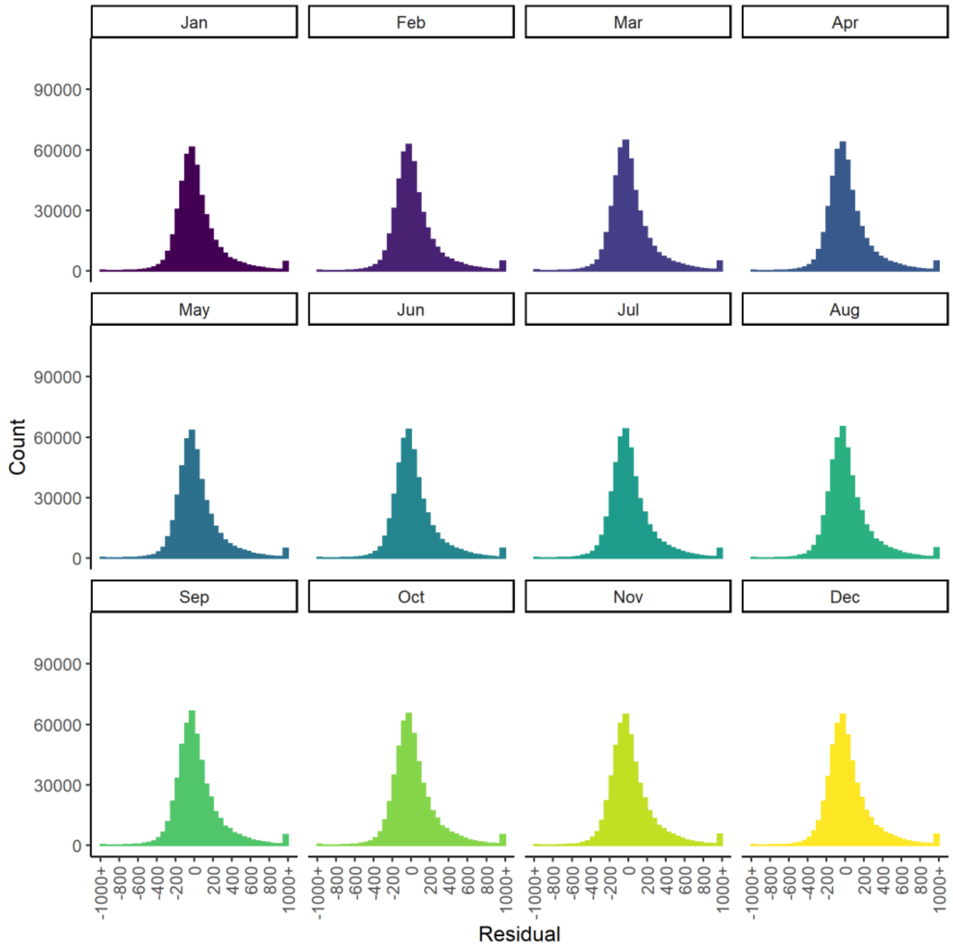
2019

OLS


Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
90000
60000
30000
0
Count
-1000+
-800
-600
-400
-200
0
200
400
600
800
1000+
Residual

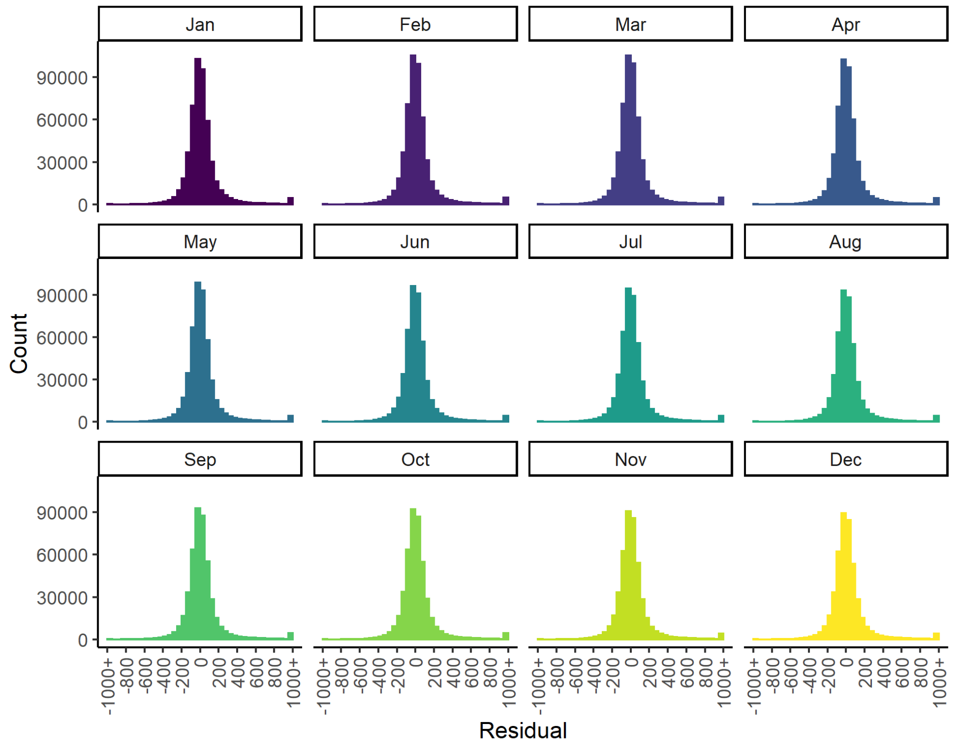
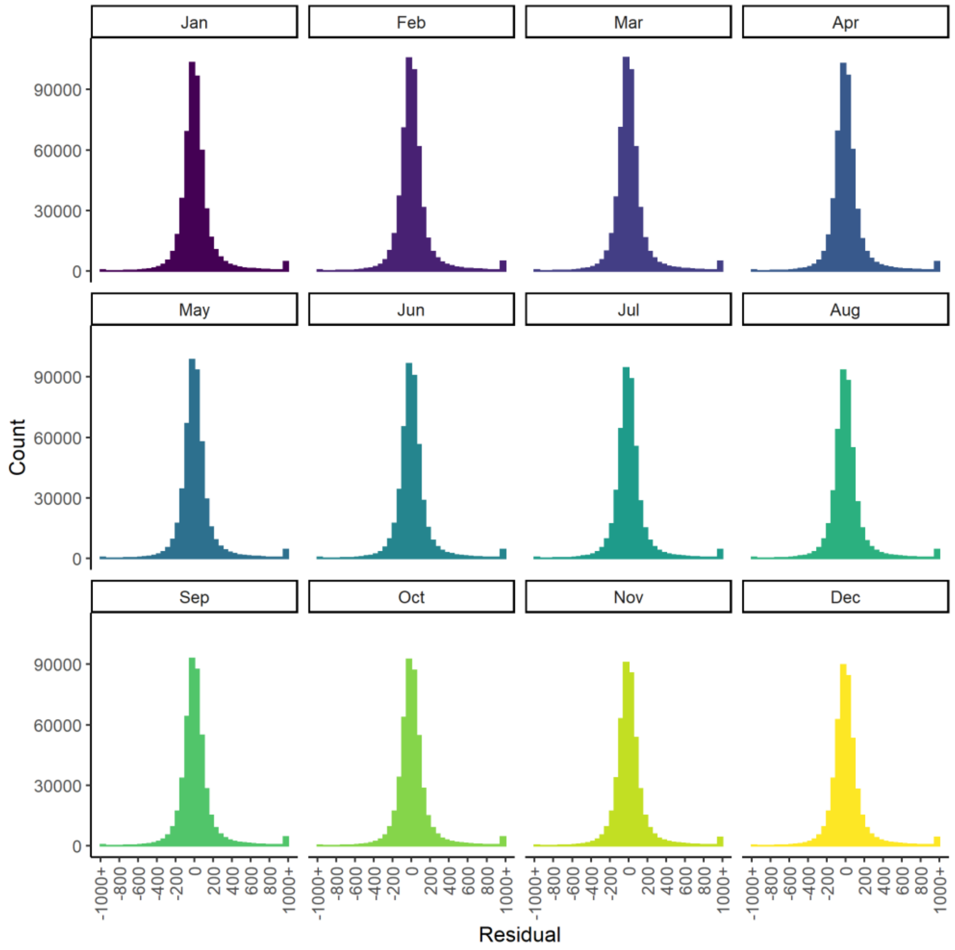
WLS


Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
90000
60000
30000
0
Count
-1000+
-800
-600
-400
-200
0
200
400
600
800
1000+
Residual

RF	

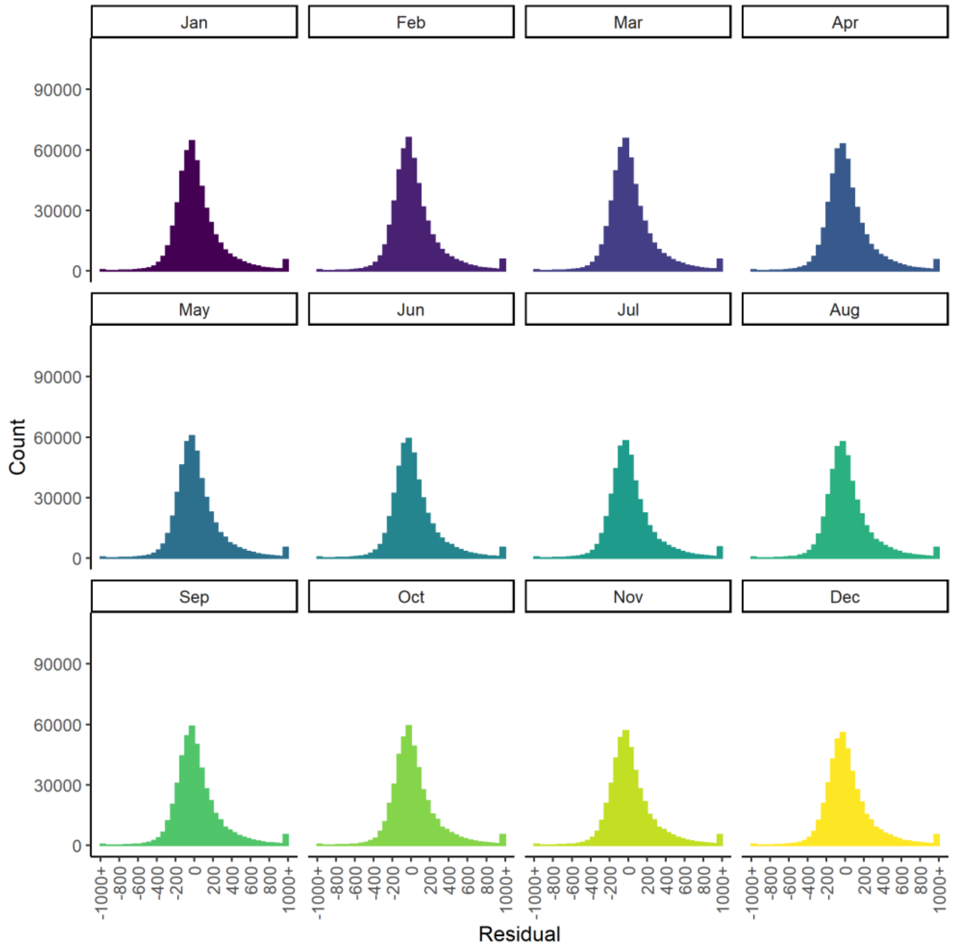
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Count
90000
60000
30000
0
-1000+
-800
-600
-400
-200
0
200
400
600
800
1000+
Residual

2020

OLS	
WLS	

Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Count
90000
60000
30000
0
-1000+
-800
-600
-400
-200
0
200
400
600
800
1000+
Residual

Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Count
90000
60000
30000
0
-1000+
-800
-600
-400
-200
0
200
400
600
800
1000+
Residual

RF	Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Count 90000 60000 30000 0 -1000+ -800 -600 -400 -200 0 200 400 600 800 1000+ Residual