Modelled rental prices and their effect on Price Indices

Paper prepared for the 18th Meeting of the Ottawa Group, Ottawa, 13th-15th May 2024

Authors: Tony Liu & Ben Hillman
Affiliation: UK Office for National Statistics, Methodology and Quality Directorate

Abstract: This paper evaluates the performance of the double imputation hedonic approach in the context of rental prices. We approach this by constructing test sets of matched properties from a large English rental dataset, allowing predicted hedonic price relatives and elementary aggregates to be compared with real price data over the test set. We test the effectiveness of a straightforward countrywide regression model and two stratified regression variants of it at producing Carli and Jevons elementary aggregates under two practically usable elementary aggregate grouping schemes.

After compensating for missing variables bias in the dataset using a proxy variable, we find that despite the inability of the hedonic models to replicate price relatives at the property level, the bias of Jevons elementary aggregates is very low for all tested models. Stratified regression specifications produce tighter error spreads (i.e., lower error variance), with greater error spread reduction for lower-level elementary aggregate stratification schemes. Carli elementary aggregates show larger signs of bias under lower-level elementary aggregate grouping schemes and are thus not recommended for these. This result is linked to the hedonic price relatives having lower dispersion than real data and the Carli's higher sensitivity to price relative dispersion.

Introduction

A widely used and internationally recognised approach in the field of rental and house price indices is the double imputation hedonic regression index, described in (Handbook on Residential Property Prices Indices (RPPIs), 2013). In this, hedonic models trained on price and property characteristic data in each period are used to impute the prices of a fixed set of properties with known characteristics. Indices are derived from these imputed prices in different periods. This is often combined with stratification of the set of properties imputed over, with properties of a similar type grouped together into strata. Unweighted elementary aggregates are produced for these strata individually. Such an approach allows weights associated with the strata to be used in aggregation. This can compensate for unrepresentativeness in the sampling, enabling us to construct accurate indices even in the absence of a census.

The effectiveness of the double imputation hedonic approach is not straightforward to test, however. Tests of the out of sample accuracy of the imputed prices generated by the hedonic model are not sufficient to determine the accuracy of any indices generated. This is because any indices are a ratio of two imputed prices, meaning biases or inaccuracies in the model have the potential of cancelling out in the index (Hill, 2011).

We therefore need a direct test of the accuracy of the indices and price relatives generated by the double imputation approach. This is possible in the domain of rental price indices, where it is

practical to repeatedly update the prices of a large swathe of properties over a relatively short span of time. With such a dataset, it is possible to calculate indices and price relatives from real data and compare them to price relatives and indices derived from imputed prices generated by the double imputation approach. We adopt a test/train approach common in the machine learning field. This uses subsets of the data to train hedonic models, which are then used to impute prices and calculate price relatives for properties not used in training. This gives a direct out-of-sample test of accuracy for double imputation indices and price relatives.

There have been some evaluations of the accuracy of hedonic indices, most notably (Hoffman & Kurz, 2002). This was done using a survey of approximately 6,000 West German rental households in years 1984-1999. The analysis compared direct matched model indices to those generated by a time dummy hedonic and a characteristics-prices approach from the same dataset. The analysis did not test the double imputation hedonic method. In this analysis, a much larger dataset was used to directly examine the performance of the double imputation hedonic method.

In addition, we tested the performance of the method when calculating elementary aggregates in several practical stratification schemes. This is a much lower level of aggregation than was previously examined in (Hoffman & Kurz, 2002), where only performance at a national level was considered. An index aggregated using stratification weights will be constrained by the accuracy of the strata elementary aggregates, so assessing performance at strata level is critical.

In this investigation, test datasets containing matched pairs of properties over a 15-month window were used to evaluate the performance of the double imputation hedonic approach. In addition, we explored selection bias from unavailable property characteristics, and possible mitigation efforts.

After adjusting for selection bias, we tested a simple country-wide hedonic regression and two stratified regression specifications of the model at two levels of grouping for defining elementary aggregates schemes that group together properties with similar key characteristics[1]. We first examine the property-by-property price relative distributions produced by the hedonic method, and then look at the average bias of the hedonic elementary aggregates and the standard deviation (or spread) of the errors, using both a Carli and Jevons index formula, under three different model specifications. From these results, we draw conclusions on the choice of hedonic model, index formula and the choice of elementary aggregate stratification scheme.

Methodology

Dataset Description

The data that we used for this research are from the Valuation Office Agency lettings information and property attributes administrative data (see Quality assurance of administrative data used in the Price Index of Private Rents - Office for National Statistics (ons.gov.uk)). This includes information on monthly rental price and property characteristics. Every month, newly surveyed properties are added to the existing dataset; if the property already exists within the dataset, the record is updated. A rental price of a property is assumed to remain valid for 14 months from collection; any property not re-collected within 14-months is dropped from the dataset. We hereafter refer to this as the monthly dataset (see also Price Index of Private Rents QMI - Office for National Statistics (ons.gov.uk), Creation of monthly dataset).

[1] We refer to the level of stratification of the model, by which we mean the geographical level the model(s) are trained at, as 'stratification', and the stratification of the elementary aggregate, by which we mean the level at which we calculate the first index, as 'grouping', to avoid confusion.

Matching

We chose to examine 15-month apart pairs of this monthly dataset. The set of matched properties are the properties that stay within the monthly dataset throughout the 15-month window. We hereafter refer to this as the matched pool. This definition removes cases where there are no observed price changes over the window from the test set, since all properties that stay in the sample over the window must be re-surveyed at least once after the start of the window. This is because collected prices are rolled forward up to 14 months, so if a property is not re-surveyed at least once in the 15-months in question, it will be dropped from the monthly dataset by the end of the window.

The matched pool is created by using matching on complete address data over the monthly datasets within the 15-month window. The result is a set of properties with a price quote at the beginning of the 15-month window and a newer second price quote at the end of the window (a set of matched pairs). Some duplicate and special case removal is applied before the matching operation to prevent erroneous matches and reduce outliers. These affect only a small fraction of cases.

The properties that do not stay in the monthly dataset throughout the 15-month window are hereafter referred to as the unmatched pool. Seven 15-month time windows are considered in total (Jan 2015-April 2016, Jan 2016 – April 2017 etc. up until Jan 2022 – April 2023, excluding Jan 2021 – April 2022).

The Test/Train Approach

This paper's core methodology is based on the concept of a 'train and test split' in a dataset. This involves splitting a dataset at random into two elements; one a 'training dataset' used to train a model, the other is a 'test dataset' to which the model is applied to test its performance. Our approach is to use random samples of the matched pool as our test data, while the rest of the matched pool and the full unmatched pool are used to train the model. Since properties in the matched pool have real price data at the beginning and end of the 15-month window, these can be used to create benchmark bilateral price relatives and elementary aggregates over this window at various levels of grouping using different index methods (e.g., Jevons). This provides a direct out-of-sample performance test that compares collected price relatives to imputed price relatives generated by the double imputation hedonic method.

Missing Variables and Selection Bias

The test set can only contain properties in the matched pool, since it is not possible to calculate price relatives based on real data for properties in the unmatched pool over the 15-month window. We require real price relatives for the test set if we are to construct a performance metric for hedonic models.

However, this method of testing will be less effective if there are missing characteristics that are correlated with whether a property is in the matched or unmatched pool that also affect rental price inflation. This causes a selection bias in the test set that must be controlled for. For example, this has been observed in the matched pairs analysis of washing machines (Silver & Heravi, 2001) where the pool of resold washing machines (the matched pairs) exhibits different price behaviour from the pool of washing machines that are not resold. This is also analogous to attempting to measure the effect of hospital treatment on patient health without controlling for sicker individuals being more likely to attend hospital in the first place (Mostly Harmless Econometrics: An Empiricist's Companion, 2009). In the context of rental prices, an important missing characteristic that could be correlated with the matched/unmatched status is tenancy length. Depending on how the data is collected, it is possible

that longer term tenants are more likely to be in the matched pool. This will cause different inflation behaviours in the matched and unmatched pools since previous research (Hoffman & Kurz, 2002) shows that there are real tenancy discounts.

If this is observed in the data, the use of properties from both the matched and unmatched pool in the training set for the hedonic model will produce erroneous results, given the test set is constrained to be only properties from the matched pool, and the hedonic model will perform poorly in calculating elementary aggregates on our imbalanced test data.

To test for this, we conduct an exercise similar to that in (Hoffman & Kurz, 2002), as follows:

- For each of the 15-month windows analysed, the datasets are divided into a matched and unmatched pool (as described in the Matching section).
- For each 15-month window, a stratified time dummy hedonic regression is run on the matched and unmatched pool separately, with the strata being the nine regions of England.

A time dummy hedonic model controls for the effect of property characteristics on price over different time periods, with any residual price changes being attributed to inflation. If there are no missing characteristics that affect inflation and are also correlated with matched and unmatched status, we would expect the inflation rates of both pools to be similar.

The specification of the time-dummy hedonic model for each region r and pair (base, end) of 15-month apart periods is:

Equation 1

$$ln(p_{it}) = \alpha_r + \delta_r u_{it} + \sum_k \beta_{kr} x_{ikt} + \varepsilon_{it}$$

Where $ln(p_{it})$ is the natural logarithm of the rental price of property i at time t. The properties i in region r at the start and end of each 15-month window are pooled together into a single set, with the time dummy for a property i at time t, u_{it}, distinguishing property records between the start and end periods. This dummy is 1 if the record is in the "end" period of the 15-month window and 0 if it is in the "base" period of the 15-month window. The coefficient on this is δ_r, which varies by region. The set of characteristics is indexed by k, and the coefficient for that characteristic in region r is β_{kr}. The characteristic k for property i at time t is x_{ikt}, and the error term for that property at time t is ε_{it}.

The set of characteristics k are:

- The natural logarithm of property floor area
- The property type (flat, detached, semi-detached etc.)
- The Postcode District (i.e. remove the last three characters of the postcode) - see Postal geographies - Office for National Statistics (ons.gov.uk)
- The number of bedrooms (capped to 5)
- Furnished status
- ACORN variable – or geo-demographic segmentation (Acorn | Geodemographic Segmentation | Acorn Data | CACI)
- Property age

All variables are categorical except the log of floor area. This is a straightforward linear hedonic model on log price, with coefficients computed by ordinary least squares (OLS). A price index for region r over the 15-month window is given by e^{δ_r} for that region's regression.

Table 1 below shows summarised results for the time dummy hedonic test for all nine regions in England, averaged across all seven 15-month windows mentioned in the Matching section. Inflation is defined in the following way: 1.01 is 1% positive inflation, 0.99 is 1% negative inflation and so on.

Table 1 – Inflation for matched and unmatched data by region, averaged across all seven time windows

Region	Matched average inflation	Unmatched average inflation
North East	1.005185	1.040346
North West	1.009079	1.061424
Yorkshire and The Humber	1.013892	1.052859
East Midlands	1.019779	1.060663
West Midlands	1.014163	1.055131
East of England	1.021236	1.056201
London	1.019574	1.036518
South East	1.017549	1.046935
South West	1.019483	1.049241

The results show that, for all nine English regions and within all seven 15-month time windows examined, the unmatched pool has higher average inflation than the matched pool. No evidence of compositional differences in the data were found that could explain these results. Therefore, this suggests that properties that remain in the sample for longer than 14 months (i.e., those in the matched pool) have different characteristics or behaviours that lower rental price inflation, compared to those who drop out of the sample before the start of a new 14-month validity period. This can be explained by an effect of "good tenant" behaviour, or tenancy length. This is in line with previous results, such as in (Hoffman & Kurz, 2002).

Missing Variables Proxy

The above results suggest that there are likely missing characteristics to control for when applying the hedonic method. To do this, one approach could be to restrict both the test and training sets to contain only data from the matched pool. However, this would discard all the unmatched properties. It would be better if these could be used in the training set at least, to ensure there is enough data to test more detailed hedonic models with stratified regressions.

The apparent high correlation between the missing characteristics and membership of the matched and unmatched pools means we can use unmatched and matched status as a proxy for the missing variables. This means that for each 15-month window, we construct a dummy variable that indicates if a property is a member of the unmatched or matched pool and add this variable in the specification of hedonic OLS models.

The specification we begin with is the following:

Equation 2

$$\ln(p_{it}) = \alpha_t + \rho_t s_{it} + \sum_k \beta_{kt} x_{ikt} + \varepsilon_{it}$$

To simulate the double imputation hedonic approach, a different model is fitted for every time period (each window consists of a pair of 15-month apart periods), which gives rise to the time index t. The set of characteristics k are the same as those described in Equation 1, with the only difference being the use of the local authority (LA) code instead of the postcode district. This is of reduced granularity than the postcode district. s_{it} is a dummy variable indicating whether the property is matched or unmatched, ρ_t is the coefficient on this dummy. The model is fit by OLS on the training set. The dummy variable s_{it} is used as a training variable for all tested hedonic models (including the one shown later) to enable control for the missing characteristics.

We randomly sample 30% of the matched pool for every 15-month window and use this as the test set, with the other 70% of the matched pool and all unmatched properties used as training data for the model specified in Equation 2. These models (one for the "base" and one for the "end" period of each time window) were used to calculate a predicted price relative for each property in the corresponding test set, which allows us to construct a hedonic Carli and Jevons index for every 15-month time window for the entire test set. This was compared to a Carli and Jevons index for the test set calculated using the observed prices at the start and the end of the relevant window, providing us with a measure of the model performance.

Since the results are based on random samples taken from the matched pool, for every time window we repeat the exercise 10 times and calculate an average error across the 10 replicates. This is shown in Table 2 below, as the average difference between the hedonic-predicted and observed Carli and Jevons indices for each 15-month time window considered. A difference of 0.01 means that the hedonic index is 1 index point higher than the true (i.e., observed) index. The mean differences are much smaller than that, being at most 0.2 index points, which is a very small difference.

Table 2 – Average difference between predicted and observed Carli and Jevons indices when using a missing variable proxy

Time Window	Carli difference	Jevons difference
Jan 2015 - Apr 2016	-0.000646	-0.000057
Jan 2016 - Apr 2017	-0.000478	0.000230
Jan 2017 - Apr 2018	-0.000696	-0.000091
Jan 2018 - Apr 2019	-0.000418	-0.000017
Jan 2019 - Apr 2020	-0.000267	0.000222
Jan 2020 - Apr 2021	-0.000066	0.000346
Jan 2022 - Apr 2023	-0.001283	0.000517

We repeated the same exercise after excluding the proxy variable from the models in Equation 2 and using only data from the matched pool to create both the training and test sets. The results are shown in **Error! Reference source not found.** below:

Table 3 - Average difference between predicted and observed Carli and Jevons indices without missing variable proxy and matched data only

Time Window	Carli difference	Jevons difference
Jan 2015 - Apr 2016	-0.000790	0.000042
Jan 2016 - Apr 2017	-0.001157	-0.000178
Jan 2017 - Apr 2018	-0.000794	-0.000017

Jan 2018 - Apr 2019	-0.000378	0.000183
Jan 2019 - Apr 2020	-0.000575	0.000056
Jan 2020 - Apr 2021	-0.000491	0.000139
Jan 2022 - Apr 2023	-0.001749	0.000360

The results from the simple country-wide hedonic model with a missing variable proxy are very close to those from the same model applied without the proxy and using the matched data only. This provides evidence that using this proxy is effective in controlling for the missing characteristics, and allows us to use all the unmatched data in the training data without generating inaccuracies.

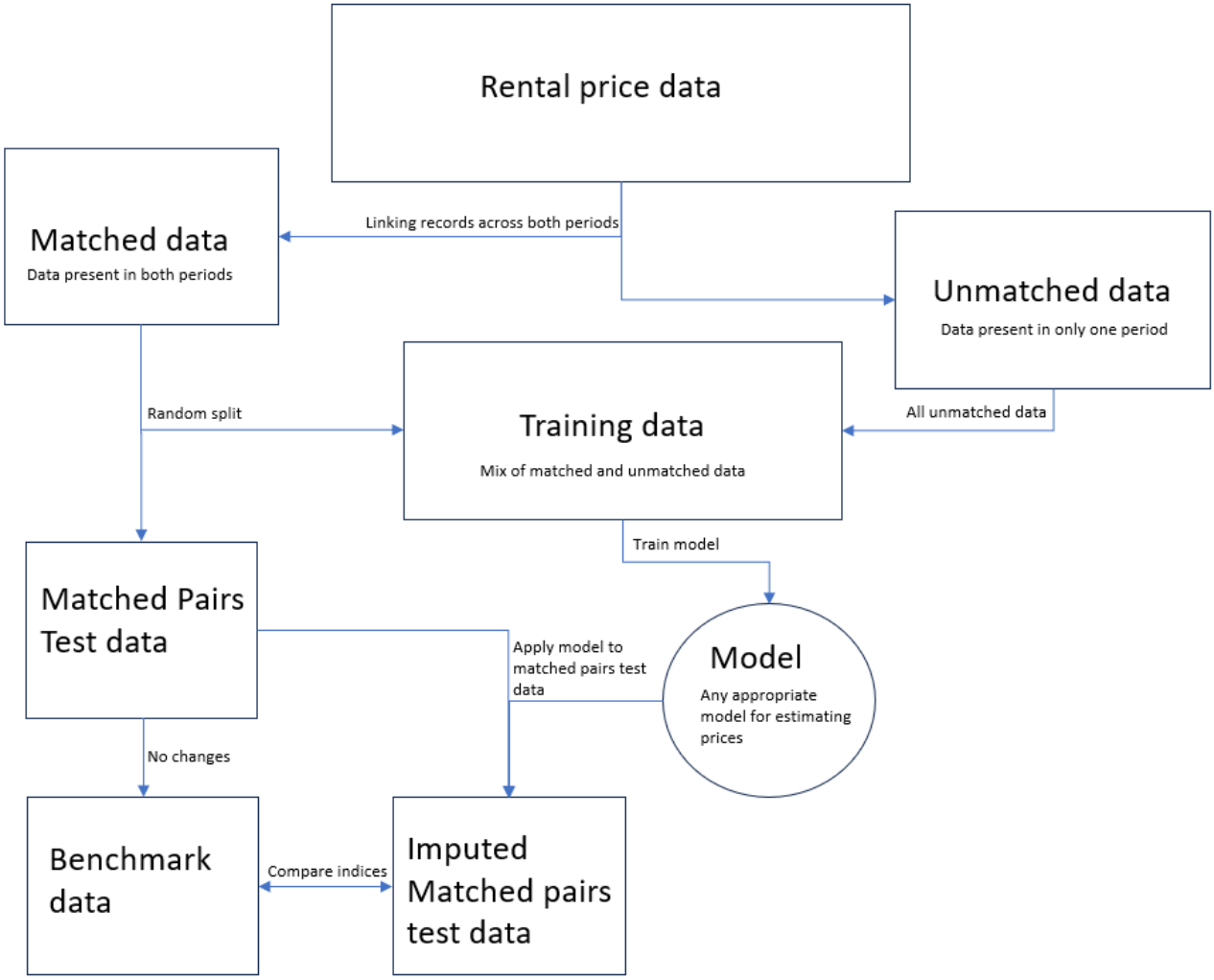
Final Methodology Training/Test Split

For the results of the paper, we use the same approach as that described in the previous section, but with a test set built from random samples of 50% of the matched pool for each 15-month window. Again, seven 15-month time windows are considered in total (Jan 2015-April 2016, Jan 2016 – April 2017, until Jan 2022 – April 2023, excluding Jan 2021 – April 2022). The aim is to measure the performance of the double imputation hedonic index under two grouping schemes for elementary aggregates, and for the Carli and Jevons index.

The overall method is shown in

Figure 1.

Figure 1 – Train / test method overview



We test three different models to explore the interaction between model choice, elementary aggregate grouping scheme, and index formula. The first is the model specified in Equation 2. The second model is a stratified version of the first model, with local authorities as the strata. This is described by the following equation:

Equation 3

$$ln(p_{it}) = \alpha_{Lt} + \rho_{Lt} s_{it} + \sum_{k} \beta_{Lkt} x_{ikt} + \varepsilon_{it}$$

Where the subscript L denotes a given stratum (or LA code) and the regression run by strata, so only including properties that are in stratum L. Thus, the estimated coefficients vary by LA code, rather than using LA code as an explanatory variable (or set of characteristics). The final version of the model we test is stratified at the regional level, a higher level of stratification compared to LA code, and uses the first section of the postcode (or postcode district) for detailed locational information in the regression rather than the LA code:

Equation 4

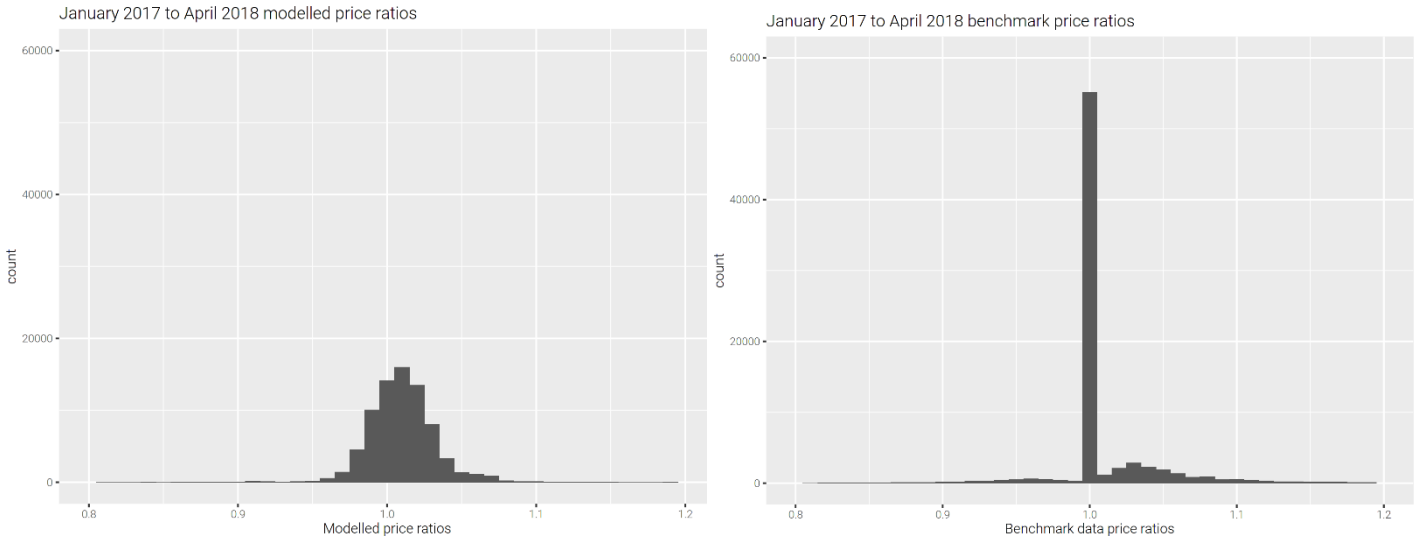
$$\ln(p_{it}) = \alpha_{Rt} + \rho_{Rt} s_{it} + \sum_k \beta_{Rkt} x_{ikt} + \varepsilon_{it}$$

The level of stratification is higher, but the locational information used is substantially finer than the LA code. We test stratified variants of the simple model specified in Equation 2 since they could potentially be better hedonic models that don't require a dataset with more variables.

Results and Discussion

We begin by examining price relatives' distributions produced by the simple countrywide hedonic model specified in Equation 2 and comparing with the benchmark distribution of real price relatives over the same test dataset of matched properties. Price relatives are defined as the ratio of prices between the end and beginning of the window. The two shown below in Figure 2 are for all of England (~75,000 properties) in 2017-2018, and use the same y-axis.

Figure 2 – Distribution of modelled, benchmark, and stratified modelled prices for England, January 2017 to April 2018



The real data (right) exhibits a pronounced spike at 1, showing that a large fraction of properties in the test set (matched properties) do not change their prices over a 15-month window. Due to the definition of the matched pool, this cannot be due to properties not being resurveyed: all properties in the test set must have had their rental prices updated at least once in the window. The large spike at 1 is indicative of contract stickiness in the matched pool. This is not replicated in the hedonic price relatives. The tails of the real price relatives also appear to be fatter than that of the modelled price relative distribution. This is reflected in the aggregate price relative standard deviations being higher in the real data compared to any of the tested hedonic models. This is shown below in

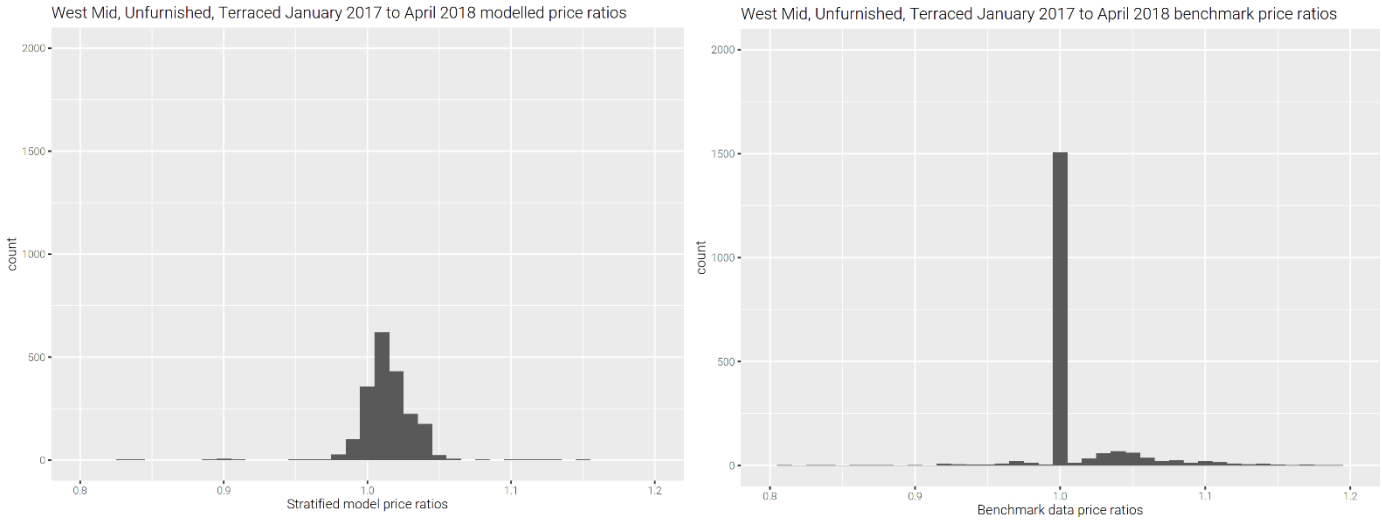
Table 4.

Table 4 – Standard deviation of the price relatives for the base data and for various model options

PR standard deviation, data	PR standard deviation, Basic Model	PR standard deviation, LA Stratified Model	PR standard deviation, Region Stratified Model
0.048796	0.027952	0.045029	0.033739

This pattern of differences in price relative distributions is observed at lower stratification levels too. An example of unfurnished terraced properties in the West Midlands region of England in 2017-2018 (~2000 properties) is shown in Figure 3 below.

Figure 3 - Distribution of modelled and benchmark prices for unfurnished, terraced properties in the West Midlands, January 2017 to April 2018



This general pattern appears to hold at all levels of stratification that still contain a reasonably large number of properties in the test set.

With the data available to us, hedonic models do not accurately predict price relatives for individual properties. More information and more complex models are likely required, such as data on tenancy length and some consideration of contract stickiness.

When the aim is to calculate index numbers, however, the goal is to have a good prediction of the elementary aggregate using an unweighted index method at some low level of stratification. We therefore only need the hedonic methods to perform well in aggregate (rather than at the individual/household level) to generate good indices. To test this, we consider some potential elementary aggregate grouping schemes that might be used practically and evaluate the performance of the three hedonic methods specified in Equation 2 to 4 within them.

The main grouping scheme investigated is combining region code, property type and furnished status to derive a group. This grouping scheme was chosen because we expect properties with similar characteristics to have co-moving prices. Some of the groups, however, will have very low property counts, and this will naturally force errors to be large due to a lack of data in both the training set and test set. Given the size of the nine English regions, we do not expect regional level groups to have very poor counts. Therefore, runs with fewer than 500 properties in the test set were removed from the results and only the groups with a full set of 10 reruns were retained. This leaves us with approximately half of the original number of groups.

Firstly, the 10 repeat runs for each group in each time window were grouped and averaged to obtain an estimate for the error for each group and period, and then overall average errors and error standard deviations were calculated from these. **Error! Reference source not found.**5 below shows aggregate performance for the three tested models for a Jevons elementary aggregate at the group level, averaged over all group and time period pairs after low count removal, i.e. treating each group

and time period as equal in weight. After removing runs with fewer than 500 properties and fewer than 10 runs in the test set, there were 233 combinations of times and strata with 10 reruns each for a total of 2330 simulations. The results are expressed as raw differences in elementary aggregates (a difference of 0.01 meaning the hedonic index is 1 basis point higher than the true index).

Table 5 – Average error by model type for the Jevons index, regional grouping scheme

Jevons error - basic model	Jevons error - LA stratified model	Jevons error – regionally stratified model
0.000290	0.000494	0.000287

Performance for all three double imputation hedonic models averaged over all groups and time windows is good, with a very low mean bias. Indices over 15-month windows have a very low average error (less than 0.05 index points). The simple country-wide regression model's average error does marginally better than a hedonic model stratified to LA code level. However, the error standard deviations for the simple country-wide model are significantly larger than the stratified model at the LA code level, while the average error is only negligibly smaller. The simple country-wide model has an error standard deviation almost double that of the LA code stratified model. The regionally stratified model is between the two in terms of error standard deviation. This indicates a tighter spread of errors for the stratified models, and larger maximum error sizes for the simpler country-wide model.

Table 6 – Average error standard deviation by model type for the Jevons index, regional grouping scheme

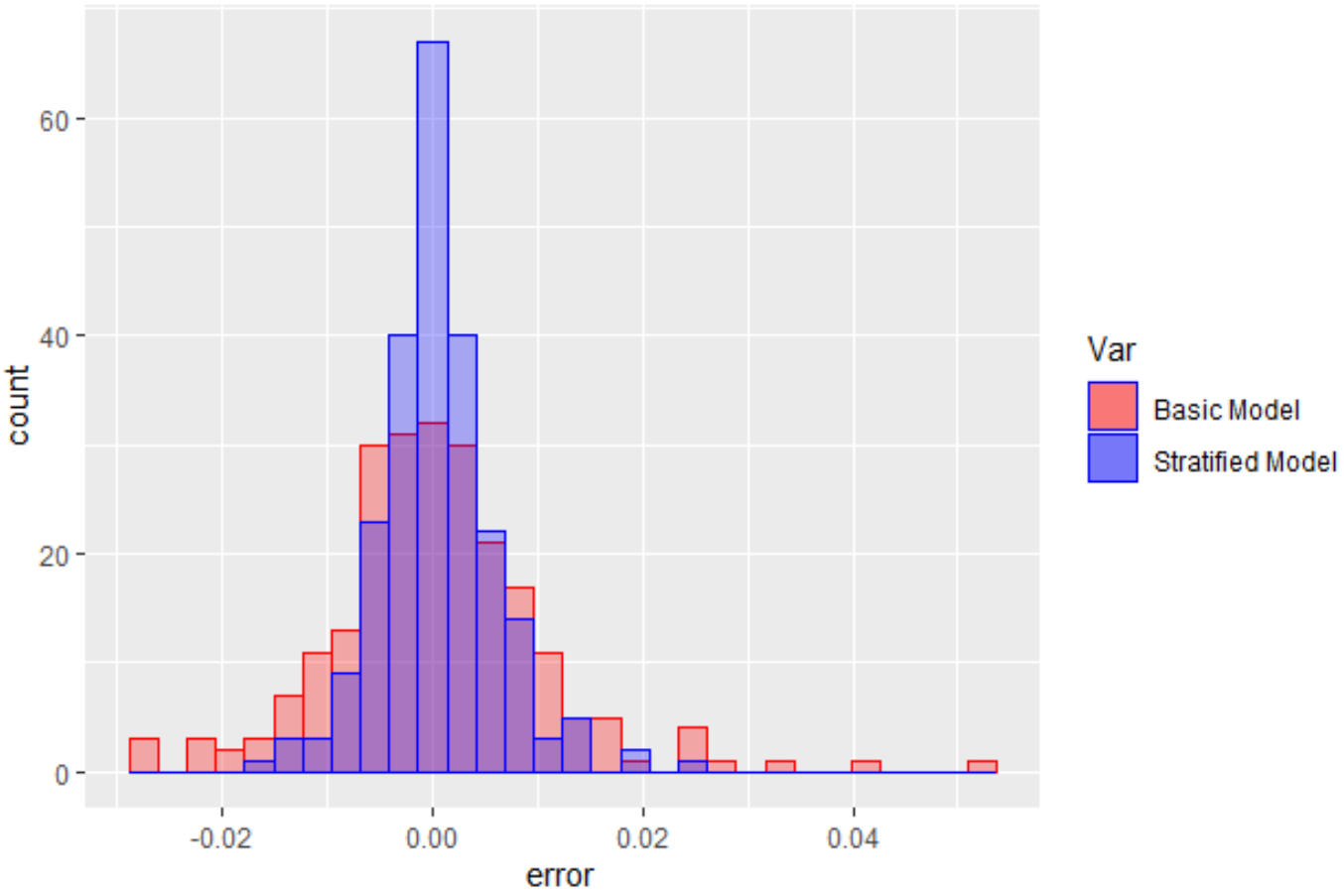
Jevons error standard deviation-basic model	Jevons error standard deviation - LA stratified model	Jevons error standard deviation - regionally stratified model
0.010302	0.005518	0.006397

To illustrate this further, we can compare the distribution of errors between the simple country-wide model and the LA level stratified model for Jevons elementary aggregates over all simulated instances. Figure 4 – Distributions of LA-stratified and basic model Jevons index errors, regional elementary aggregate grouping

below plots the histogram of errors for both models over all considered elementary aggregate groups and time windows, with the error being the difference between hedonic imputed and real data Jevons indices. The blue/purple areas are the error distribution of the stratified model, and the red/purple areas are the error distribution of the simple countrywide model. We observe that:

- Both error distributions seem reasonably evenly spread around zero error – there does not seem to be a consistent bias in one direction or the other. This is consistent with the low average error for all models.
- The error distribution of the stratified model is clearly more tightly dispersed, with fewer cases exhibiting large error magnitudes.
- The overall slightly better average error for the simple model is likely due to a marginally more even dispersion around the zero-error point.

Figure 4 – Distributions of LA-stratified and basic model Jevons index errors, regional elementary aggregate grouping



We can further quantify the lower error spread for the stratified models. Out of 233 total group and period combinations, 61 have error magnitudes below 0.25 index points per year for the basic model, while 111 have error magnitudes below this same threshold for the LA stratified model and 112 are within this threshold for the regionally stratified model.

Largely similar results for the elementary aggregate grouping are observed when using the Carli index formula, rather than Jevons, to calculate the unweighted index:

Table 7 - Average error by model type for the Carli index, regional grouping scheme

Carli error - basic model	Carli error - LA stratified model	Carli error – regionally stratified model
-0.000475	0.000359	-0.000297

Table 8 - Average error standard deviation by model type for the Carli index, regional grouping scheme

Carli error standard deviation - basic model	Carli error standard deviation - LA stratified model	Carli error standard deviation - regionally stratified model
0.01034	0.005647	0.006429

Average performance seems largely comparable to the Jevons case, with the stratified models still showing a narrower spread of errors. These results do not show hedonic methods being significantly better at calculating a Carli over a Jevons (or vice-versa) using the regional level (region/property type/furnished) elementary aggregate grouping scheme.

We also examined the impact of a finer (more detailed) grouping scheme for elementary aggregates, using a combination of LA code, property type and furnished status as the group designation. However, we were limited by very low property counts in the test and training sets for some groups. If we only consider cases where we achieved 10 reruns with at least 50 properties in the test set, the following average performance metrics were obtained for a Jevons index:

Table 9 – Average error by model type for the Jevons index at a finer elementary aggregate grouping

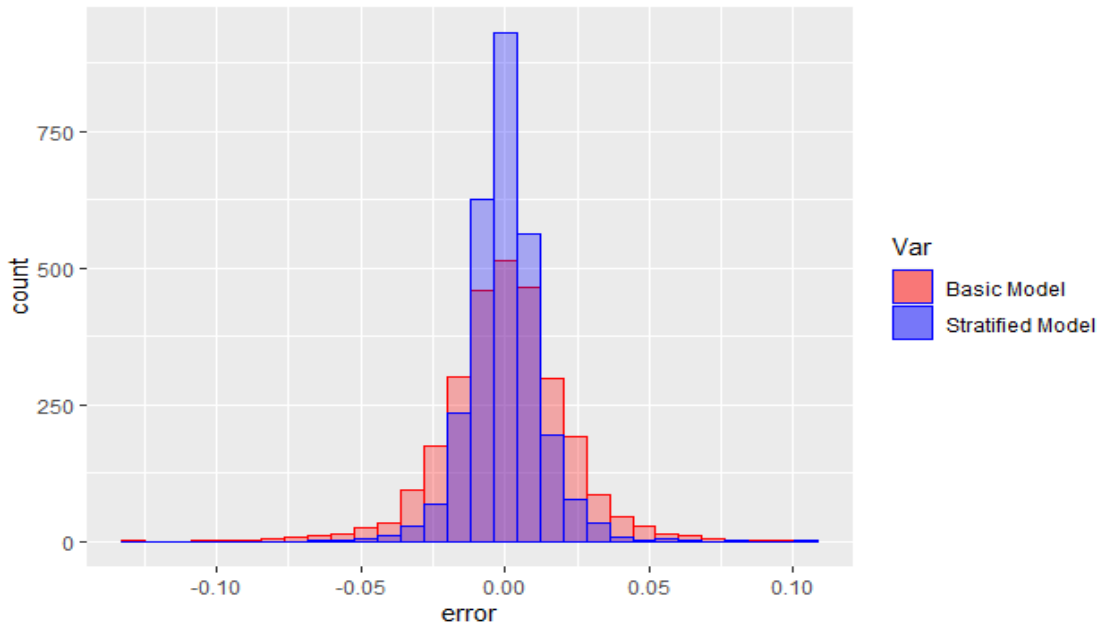
Jevons error, finer aggregation grouping - basic model	Jevons error, finer aggregation grouping - LA stratified model	Jevons error, finer aggregation grouping – regionally stratified model
-0.000164	-0.000114	-0.000100

Table 10 – Average error standard deviation by model type for the Jevons index at a finer elementary aggregate grouping

Jevons error standard deviation, finer aggregation grouping -basic model	Jevons error standard deviation, finer aggregation grouping - LA stratified model	Jevons error standard deviation, finer aggregation grouping - regionally stratified model
0.021134	0.012174	0.016887

2793 group and period combinations were in the experiment, with 10 reruns for each. The results are largely in line with the results for the higher-level grouping scheme. Mean accuracy seems very high still, with low overall bias and with negligible differences in performance between the three models. Error spread (standard deviation) is much larger for all models, however. The LA level stratified model still has the lowest error spread (about half that of the simple countrywide model) while the regionally stratified model's performance falls between the two. An error plot of the LA stratified model and the simple countrywide model in **Error! Reference source not found.** largely confirm the same results as those in Figure 4.

Figure 5 - Distributions of LA-stratified and basic model Jevons index errors at LA (finer) elementary aggregate grouping



At this level of grouping however, there begins to be a divergence in bias when using Carli index formula compared to a Jevons index. The LA level stratified regression has a significantly lower bias on average than the others when it comes to calculating a Carli index, and bias starts becoming noticeable for the simplest country-wide model (and order of magnitude larger than the bias for a Jevons index using the same model). The results are shown in the tables below:

Table 11 - Average error by model type for the Carli index at a finer elementary aggregate grouping

Carli error, finer grouping – basic model	Carli error, finer grouping – LA stratified model	Carli error, finer grouping – regionally stratified model
-0.001059	-0.000490	-0.000800

Table 12 - Average error standard deviation by model type for the Carli index at a finer elementary aggregate grouping

Carli error standard deviation, finer grouping - basic model	Carli error standard deviation, finer grouping – LA stratified model	Carli error standard deviation, finer grouping – regionally stratified model
0.021246	0.012314	0.016998

From these results, it seems that that a more detailed stratified regression model is best at calculating a Carli index at this fine level of stratification in terms of bias and error spread.

Overall, these preliminary results suggest that these simple hedonic methods give low biases (average error) for Jevons elementary aggregates when applied to a range of potential elementary aggregate grouping schemes, despite their low accuracy at predicting individual property rents. They seem to capture price movements well overall at elementary aggregate level. The low average errors in elementary aggregates means that errors in individual groupings are more likely to cancel out when aggregated. There appears to be no increase in overall bias as the stratification level becomes finer.

Stratified regression specifications that effectively allow for interaction between the LA code (or region) and other characteristics seem to perform better at narrowing error spreads at the elementary aggregate level. This is especially the case as the grouping of the elementary aggregates gets finer, although the mean error for an elementary aggregate selected at random is still very small. This is expected – the stratified models have more degrees of freedom to capture fine details in price behaviour at lower levels. This is supported by the results on price relative standard deviation shown earlier in Table 4 – Standard deviation of the price relatives for the base data and for various model options, which show the LA level stratified model generating price relative dispersion that come closest to that of the real data from matched properties. It follows that using more detailed stratified regression methods has greater value when elementary aggregates are at finer levels, since the benefit to shrinking error magnitudes (by about half) is larger when all error magnitudes are higher.

Hedonic methods seem to have more difficulty replicating Carli elementary aggregates in the finer stratification scheme, with the simple countrywide model generating elementary aggregates with a notable level of bias. This is likely to be also linked to the better replication of price relative dispersion by the stratified regression models with more degrees of freedom, since Carli indices are more sensitive to price relative dispersion. This is because they are an arithmetic mean of price relatives, which does not dampen the effect of extreme values (unlike the Jevons, which is a geometric mean of price relatives). This potential difficulty in replicating the Carli at finer elementary groupings suggests that a Jevons index is a better choice for finer grouping schemes. If a Carli must

be calculated, a stratified regression model capable of coming closer to the price relative dispersion observed in real data should be used.

There are two significant caveats to these results, however. Firstly, by definition, we can only test properties that are in the matched pool. The unmatched pool shows significantly different price behaviour. Although we controlled for the missing variables responsible for this with some degree of success, we cannot be certain that the results extend to unmatched properties - we do not know their price relatives. To improve our understanding of how well the double imputation approach predicts elementary aggregates over the unmatched pool, further work is required to understand the nature and effects of the missing variables as well as the effectiveness of the proxy approach. More detailed data would be required to do this further analysis, ideally with tenancy data and where the matched and unmatched status is decoupled from the tenancy data.

Secondly, the estimates of error dispersion will not apply in general to all hedonic models. More sophisticated hedonic regressions with more data will likely achieve tighter error spreads, although they will probably not reach noticeably better bias levels since even simple regression models show very low bias in elementary aggregates.

Within this testing framework, it would also be interesting to see how machine learning models which do not impose assumptions on functional form perform, and whether they do better or worse in terms of bias and error dispersion than a simple regression model.

Conclusions

This paper applied the test/train approach to a large dataset of both unmatched and matched rental properties, allowing a direct examination of the accuracy of indices produced by the double imputation hedonic approach at an elementary aggregate level. We did this by testing against the observed indices produced from a subset of the matched pool of properties. An initial exploration of the dataset showed some evidence of potential missing variables that are highly correlated with matched and unmatched status, causing the unmatched pool to display higher inflation rates.

After adding a proxy for these missing variables to allow for the use of unmatched properties in the training set, we demonstrated that using a relatively simple set of variables in a country-wide regression model produces elementary aggregates in practically usable stratification schemes that have very low bias (average error). This is despite noticeable differences between the imputed and real price relative distributions for individual properties, with real data showing a large concentration of properties with no price change that is not seen in the modelled data and also having a larger dispersion of price relatives.

More sophisticated stratified regressions using the same set of variables do not attain appreciably lower levels of bias when a Jevons index is used for the elementary aggregates. However, these stratified regressions give tighter error spreads, that is, lower standard deviations for elementary aggregate errors. This is likely because of these models having more degrees of freedom, allowing them to capture lower-level details in price behaviour.

At finer levels of elementary aggregate grouping, error spreads for all models increase. This means that stratified regressions offer larger performance increases in these scenarios since the decrease in elementary aggregate error spread they give is larger for finer grouping schemes. Jevons indices do not show any overall increase in bias as the grouping level becomes finer, but the Carli does, likely because of Carli's higher sensitivity to price dispersion in the price relatives. This makes the Jevons a better choice of index formula for finer grouping schemes.

This is a preliminary study, however. We can only test matched properties by definition and our investigation indicated that unmatched properties behave differently. Further work with more detailed datasets and further analysis on the missing variables would be needed to confirm these results.

References

(2009). In J. Angrist, & J. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (p. Chapter 1). Princeton University Press.

Handbook on Residential Property Prices Indices (RPPIs). (2013). *Eurostat Methodologies and Working Papers*, p. 54.

Hill, R. (2011). Hedonic Price Indexes for Housing. *OECD Statistics Working Paper* .

Hoffman, J., & Kurz, C. (2002). Rent Indices for Housing in West Germany 1985 to 1998. *European Central Bank Working Paper Series*.

Silver, M., & Heravi, S. (2001). Why the CPI Matched-models Method May Fail Us: Results from a Hedonic and Matched Experiment Using Scanner Data. *CEPR/ECB Workshop "Issues in the Measurement of Price Indices".*