A Multi-Criteria Evaluation of House Price Indexes

Andy Krause^ – Zillow Group Reid Johnson – Zillow Group

2024-03-18

^Corresponding Author – andykr @ zillowgroup.com

Submitted to the ARES Annual Conference (March 20-23, 2024)

Abstract

House price indexes are a widely used tool for understanding the aggregate movements in housing markets. They play an important role in decision-making by government bodies, federal and local, as well as individual market actors. While a robust collection of prior research on methods for index creation exist, the set of work that addresses fundamental questions of measuring index quality is more limited. In this work, we extend the prior criteria and metrics to include assistive accuracy, volatility and local specificity. We illustrate our new criteria through an empirical example comparing a variety of common approaches – Aggregation, Repeat Sales, Hedonic Time Dummy and Imputation – as well as novel Neural Network Approach. We find that no index provides the best performance on all metrics and that users should consider the purpose of the index, the available data and end user requirements along with more objective model performance metrics when choosing an index.

Introduction

Since the seminal Bailey et al. (1963) study there has been considerable and sustained research effort put into comparing and improving competing methods for generating house price indexes. Published work in this sub-field of housing economics is generally focused on one or more of four aims: 1) Comparison of model differences (Case et al 1991; Crone & Voith 1992; Meese and Wallace 1997; Nagaraja et al 2014; Bourassa et al 2016); 2) Identification and correction of estimation issues or problems (Abraham & Schauman 1991; Haurin & Henderschott 1991; Clapp et al 1992; Case et al 1997; Steele & Goy 1997; Gatzlaff & Haurin 1997, 1998; Munneke & Slade 2000); 3) Creation of local or submarket indexes (Goodman 1978; Hill et al 1997; Gunterman et al 2016; Bogin et al 2019; Ahlfeldt et al 2023); and/or 4) Development of a new model or estimator (Case & Quigley 1991; Quigley 1995; Hill et al 1997; Englund et al. 1998, McMillen 2012; Bokhari & Geltner 2012; Bourassa et al 2016; Xu and Zhang 2022).

Research aimed at measuring the quality of house price indexes has received considerable less attention. Part of this lack of work likely stems from the fact that a house price index is usually created to measure a phenomena without an observable ground truth – the price or value movements of a given market – and therefore is, under any method or approach, a proxy at best. This inherent fact also means that any approach to measuring model quality necessarily comes with shortcomings and/or assumptions. In short, there is no perfect measuring stick for a house price index.

In this work, we evaluate a suite of different methods for generating house price indexes against two classes of criteria; Development and Output. Development criteria are those considerations that can and should be

addressed prior to building an index. They includes the data available, the explainability requirements and the generalization desires. Output criteria, on the the other hand, involve making quantitative assessment on the indexes and/or series of indexes that are created. Together, Development and Output criteria can assist an index creator make a determination on the best available approach in their particular situation.

In terms of house price index creation, traditionally, house price indexes have been derived from highly interpretable (statistical) modeling approaches such as linear models. Both the repeat sales and the hedonic time dummy approach – the two most commonly published approaches (Hill 2012; MacGuire et al 2013) – are generally estimated with linear regression-based models. Standard statistical models are a good fit for this task because the coefficient estimates are easily convertible into standardized price indexes. House price index generation is viewed as an statistical inference endeavor in which the estimation of market movements for a (sub) population is sought using some sample, rather than a pure prediction problem like predicting a single home value. As a result, many of the rapidly growing set of machine learning algorithms – e.g. support vector machines, random forests and neural networks – have not been used in the production of price indexes due to the fact that they do not directly and/or easily attribute price impacts to the variables or features in the model. However, with the rise of interpretability methods (Ribiero et al. 2016; Doshi-Velez and Kim 2017; Molnar 2019), these 'black-box' models can be made explainable and suitable for a more diverse set of tasks.

The remainder of this work is organized as follows: Section two provides a literature review of existing criteria for evaluating house price indexes, followed by an explanation of the novel metrics we add. Next, we offer a brief review of applying machine learning approaches to the task of house price indexing. Section three concludes with a discussion on using the interpretable random forest (IRF) approach to creating a house price index. In section four we discuss our three model specifications and the data – from King County, Washington, USA – that we use in our empirical analysis. We then cover our results in section five; concluding with the paper with recommendations and details on the reproducility of this work.

Previous Work

In the existing work, three general criteria have been applied to evaluating house price indexes; one based on inputs and estimation method and two on outputs. The input- and method- based criterion assesses the ease of construction. This question of the constructability of a house price index can be divided into two sub-questions or criteria: 1) Ease of data collection; and 2) Simplicity of Model or Estimator.

The first, ease of data, reduces to a question of the availability of granular sales and home attribute data. Under this criteria, the easiest index to create is one that relies only on aggregated prices, such as tracking the mean or the median of observed prices in a given market over a given time period. This results in simplistic mean or median value indexes; approaches that are often used as baselines in comparative work (ex. Goh et al 2012). Next are indexes based on repeat transaction of the same home. Repeat sales are appealing as they solve some of the constant quality issues that fully aggregated indexes cannot and they only require granular sales data – the sale price, the sale data and a unique home identifier such as address – and not home attributes. Finally, we have a wide variety of approaches, broadly characterized in the literature as hedonic methods (Hill 2013) that require both transaction and home attribute data for all transactions. These hedonic methods present the heaviest data collection burden of the three.

The second criteria under 'constructability' is the complexity of the model estimator itself. Again, a fully aggregated approach like a median or mean offers the lowest barrier as they include taking a mean or median. Traditionally, repeat sales models (Case and Shiller, 1989) utilized basic linear models (OLS) which are straight-forward to both create and understand. More recent work (Bourassa et al 2016) has developed more advanced or robust approaches to repeat sale estimation but they are still rooted in linear regressor. The broad class of 'hedonic' approach again offer the most complexity in estimation. Within this class of model, constructability can vary widely from simple in the case of a linear hedonic model with dummy estimators (ex. Hill and Trojanek 2022) to very complex method derives from neural network estimators (ex. Xu and Zhang 2023). Hedonic estimators also present complexity in the form of the choice of model specification (independent variables), structure and, in the case of machine learning-based approaches, the selection of

hyperparameters. Together, all of the these decision about the model combined with the potential for greatly increased compute times as the methodology becomes more specialized means that hedonic approach present the highest barriers to constructability.

Output-based metrics Index accuracy is the most common output metric discussed in the literature. As noted above, there is no exact ground truth of the movement of aggregate house prices in a region so any approach here is a proxy, with limitations. The favored approach in the existing literature is to test the ability of an index to predict the second sale of repeat transaction pair (Bogin et al 2019). This approach takes the first sale of a pair, indexes that price forward by the given index to the date of the repeat transaction and then computes an error measure where the error is the difference in the predicted value and the actual price of the second sale in the pair. When measuring the error – or the difference between predicted and actual second sale – we use log metrics due to their ability to avoid denominator bias and the skewness in possible error metrics that results (Tofallis 2015). The formula for calculating errors is:

$Error_i = log(price_{i,pred}) - log(price_{i,actual})$

This is an appealing approach as individual error metrics at the home level allow for computation of standard metrics like root mean squared error (RMSE), mean absolute percentage error (MAPE) and others that are common in the evaluation of Automated Valuation Models (AVMs), other hedonic pricing models and many/most regression-based machine learning tasks(Steurer et al 2021).

A downfall of this particular approach to accuracy; however, is that it relies solely on repeat transactions. Repeat transactions are not a random sample of all homes in the market nor are they are random sample even of those homes that sell (Hill 2013). Additionally, repeat transactions themselves can be subject to violations of constant quality in the case of renovation and/or depreciation (Steele and Gray 1997; Novak and Smith 2020).

A second approach to measure the objective qualities of an index looks at the revision of an index over time (Clapham et al 2006, Deng and Quigley 2008, Van de Minne et al 2020, Sayag et al 2022). Revisions occurs when the re-estimation of the index over time results in changes to prior estimates. For example, consider an index with a value of 134.2 in August of 2023. If when the index is re-estimated in September and the August value changes to 134.4, that is a revision of 0.2 points. Revision, particularly those that are substantial can create harmful impacts on users of indices in applied and/or policy perspectives. It is generally accepted that index revision should be minimized (Van de Minne et al 2020).

In summary, the existing work on measuring index quality uses three criteria to evaluation house price indexes: 1) a measure of the ease of constructability; 2) a measure of accuracy in the form of predicting the second of a repeat transaction; and 3) a measure of the extent to which the index values revise as a series of indexes are created over time. We use this existing framework as a basis for our extensions discussed in the section below.

Extending The Measure of Index Evaluation

One perspective missing from the current framework for index evaluation is that of the end use or user. Put another way, the purpose to which the index is being put should be considered. To bring this perspective we add three addition evaluation criteria to the existing three: 1) Relative predictive accuracy improvement; 2) Volatility; 3) Local specificity.

As noted above, the use of repeat transactions as the benchmark for predictive accuracy suffers from many of the same limitations as repeat transaction models – namely sample selection and constant quality issues stemming from renovation and/or depreciation. This repeat transactions-based approach also ignores one of the growing uses to which house price indexes are employed, that of supporting automated home pricing exercises. As an example, some approaches to the construction of automated valuation models (AVMs) will first time-adjust all sales in the training data (the target values) to the price as if the home sold at the time of the AVM training. By doing so a model can focus on distinguishing the cross-section variation between homes while relying on an outside model to make temporal adjustments.

Considering the growth of AVMs in valuing homes, we offer an alternative to the commonly used repeat transaction-based accuracy methods. In this alternative we compute the relative accuracy improvement of an AVM using an index for time adjustment over that of a baseline AVM that has no temporal adjustments or variables at all. This is a two-step process in which first a baseline AVM is estimated in which time is completely ignored; there are no temporal variables. We measure the predictive accuracy of this model as the baseline. We then time-adjusted all sale prices in the data to the valuation date of the AVM and re-estimate it, again measuring the predictive ability. The relative change in accuracy between the baseline model (no time controls) and the model with index-adjusted sale prices represents a comparative measure of the index's ability to improve a valuation model. This offers another measure of the index's accuracy where accuracy is a proxy for its ability to track the actual (unobserved) movements in the market. We term this 'Relative Accuracy' measurement as opposed to the 'Absolute Accuracy' measurement that the repeat-transaction based approach provides.

Next, to the extent that home price indexes are meant to represent an estimate of the aggregate movement of all home values in a region or market, a desirable characteristic is that the index is not overly noisy. 'Noisy' here means that that index 'chases' or is overly impacted by one or a single data points that are likely not representative of the underlying global phenomenon. In the machine learning domain this is often represented by a model that has great fit on in-sample data but has a large reduction in accuracy when applied to new, out of sample data. Likewise, with an index, a high noise-to-trend ratio could be indicative of the same problem.

In short, what is desirable is an index that tracks the market without fluctuating widely above and below the actual trend each period. In this paper, volatility is measured as the standard deviations of period-to-period changes in a rolling four-period time span.

$V = sd(D_{t,t+1,t+2})$ where $D = index_k - index_{k-1}$

This is an appealing metric as consistent, monotonic changes over a four month span – the three measures of period changes – will produce very low standard deviations. On the contrary, wildly fluctuating indexes with irregular directionally movements will produce high volatility measures.

In addition to the fundamental, estimator-based issues stemming from volatility, a highly noisy index is also hard to rationalize to users of indexes, particularly those looking to make important financial decisions based on the (perceived) market changes suggested by the index. Simply put, it is hard to build trust with users when indexes have a high noise-to-signal ratio.

Conversely, we shouldn't look to fully minimize volatility either as, taken to an extreme, a perfectly flat index would optimize volatility reduction but at the cost of being useful at all. It is a delicate balance. In this work, we measure volatility as the standard deviation of a rolling time window (3 months in this case), but look at the comparison only relatively. We balance off volatility against other measure in a subjective, criteria-ranking approach not as an objective measure to be optimized. As a result, it operates somewhat like the 'constructability' criteria – something that should be taken into account by developers and users, but not a criteria to necessarily be optimized against.

Finally, prior work has established the need to derive indexes for small geographic regions or other local definitions of submarkets (Haurin et al 1991; Ren et al 2017). Our third new measure of index quality is the relative change in the above metrics when the index is applied to smaller subregions within the larger market. We term this Local Specificity.

In the evaluations that follow we will focus on four objective measures of house price index performance: 1) Absolute accuracy via repeat transactions; 2) Relative accuracy via AVM improvement; 3) Revision; and 4) Volatility. With then look at the relative differences of accuracy measures at global and local (to our data) levels in order to measure the Local Specificity or ability of the index to discern local trends while also maintain overall predictive ability. In the discussion section we also layer in considerations of the 'constructability' dimension on indexes to highlight some of the tradeoffs between ease of creation and objective model results.

Emprical Tests

In this section, we describe the data used in the empirical tests that follow as well as the particular model specifications employed. As part of the data discussion, we describe the geographic subsetting employed to provide local tests as well as the county-wide global analyses.

Models

Three different models are compared in this work; 1) Interpretable random forest (IRF); 2) Hedonic price (HP); and 3) Repeat sales (RS). The particular model specifications, described in detail below, remain the same across the global and 19 local geographic areas. In all cases, indexes are estimated at a monthly frequency. All models and associated metrics and visualizations are computed in the R statistical language (R Core Team 2019). Details on particular package usage are contained in the discussion on each model.

This work develops and tests a framework for using random forest models combined with an interpretability layer to create a house price index. The discussion below will focus on the specific development of an Interpretable Random Forest (IRF) model. Readers interested in a broader coverage of approaches to and issues with existing house price index methods such as Repeat Sales and Hedonic Pricing Models are directed to the "Handbook on Residential Property Prices Indices"(Eurostat 2013).

Random Forests

The term 'machine learning' often conjures the pejorative term 'black box'. Or rather, a model for which predictions are given but for reasons unknown and, perhaps, unknowable, by humans. For use cases where a predicted outcome or response, in itself is all that is required the 'black box'-ness of a model or algorithm may not be an issue (Molnar 2019). However, in cases where model biases need to be diagnosed and/or individual feature or variable contributions are a key concern of the research or model application – such as for constructing house price indexes – machine learning models need to be extended with interpretabilty methods.

There are many options for the choice of machine learning model, though most all specific model classes fall into four generalized classes: 1) logical model (decision trees); 2) linear and linear combinations of trees or other features (random forests); 3) case-based reasoning (support vector machines); and 4) iterative summarization (neural networks) (Rudin and Carlson 2019). This paper uses random forests (Breiman 2001; Hastie et al 2008) as an example as they are a common modeling approach in the machine learning literature and in industry. Random forests create a large set of many decision trees, each based on a random set of the data. As each tree is grown, the partitions in the tree are limited to a random set of the variables (features) in the data. This set of (decision) trees 'grown' via randomness makes a random forest. To make a prediction, simply evaluate the subject instance (house in a real estate valuation context) in each tree – which gives a predicted value – and then combine all of these evaluations and take the mean (or some other measure of central tendency). The choice of the number of trees to use and the number of random variables to be considered at each partition step are (hyper) parameters that must be determined by the modeler.

Random forests, essentially bootstrapped submarketing routines, also have a natural link to real estate valuation via the selection of small subsets of like homes to drive predictions. Interestingly, random forests have been little used in academic real estate studies (see Mayer et al 2019 for an exception) and not at all in house price index creation (to the knowledge of the author). This lack of use can likely be explained by the fact that random forests are a 'black box' in that they do not directly create coefficient estimates as more traditional statistical models do and, therefore, do not offer a direct approach to create price indexes. A random forest model by itself will provide a predicted value but no direct explanation of how that prediction was generated. In short, they are not inherently interpretable.

Interpretability Methods As the use of machine learning models has grown, so too have methods to help increase the interpretability of these approaches (Slack et al 2019). One such set of enhancements are termed 'model-agnostic interpretability methods' (Molnar 2019). Model agnostic interpretability methods are post-hoc models that can be applied to any learner or model in order to provide a specific enhancement or extension in the overall interpretability of the model. Model agnostic interpretability methods can fall into a number of types or classes, some of which have varying aims. Some of the most common approaches are:

- **Simulated or counter-factual scoring**. In these approaches, machine learning models compare scored (predicted) values of counter-factual observations across a given variable(s) while holding all others constant. Individual conditional expectations (ICE) (Goldstein et al 2014) and partial dependence (PD) (Friedman 2001) are standard examples of this approach. Accumulated local effects (ALE) can also be used when extensive correlations exist in the independent variables of interest (Apley 2016). Often a goal of these approaches is to understand the marginal contribution of one or more features towards the predicted value.
- **Game Theory (Shapley Values)**. A game theory or bargaining approach where players (variables or features in the model) compete to determine the optimal payout (coefficients) for their contributions to each observed price (Cohen et al 2005; Molner 2019). Shapley values, like counter-factual scoring, seek to measure marginal contribution of specific features.
- **Global and local surrogates**. Surrogate interpretable models that roughly approximate a black box model can provide human-interpretable explanations. These surrogate models can be global – spanning all observations – or local – confined to a small subset of the data, such as location. The locally interpretable model explanation (LIME) method proposed by Ribiero et al (2016b) is the most widely known local surrogate approach. Local and global surrogates are usually used to more deeply understand the prediction of one or a few individual instances.
- **Feature importance via permutation**. Judging the importance of a particular feature or variable within a black box model can be estimated via a permutation method (Gregorutti et al 2017). This approach works by estimating a baseline model with all variables as is. For each feature (variable), permute or randomize the data for that feature and re-estimate the model. Do this for all features one at a time and measure the relative degradation of model performance when each feature is randomized. This provides a (relative) measure of which variables or features are the most important. Feature importance measures are used to identify which features in the model provided the biggest (relative) gains in model performance and aid in model specification tasks.

In this work, we use measures of individual conditional expectation (ICE) and partial dependence (PD) to extract interpretable insights on real estate market behavior over time. I have chosen this approach for two primary reasons. First, the ICE/PD approach - via counter-factual scoring across the variable of interest, time – conceptually mimics the basic questions that drive real estate price indexes, namely: What would this property/house have sold for across given intervals of time, had it sold repeated? In fact, this approach does exactly that by simulating a home sale for a given property at every time period in the study (ICE) and then combines those changes in price over time across all properties (PD).

Second, ICEs and PD are one of the easiest of the above methods to compute. It should be noted, partial dependence calculations are known to be potentially biased when the variable of interest is highly correlated with other independent variables (Molnar 2019). Most variables used in standard hedonic pricing models, such as bedrooms, bathrooms and home size are often highly correlated. Fortunately, for the purposes of house price index generation the variable of interest – time of sale – is generally highly orthogonal to other control variables making partial dependence an acceptable approach. This assumption could be violated if the quality or location of housing that transacts varies greatly over time. Practically though, this is only likely to occur in a relatively small geographic area that experienced significant new construction sales. The data in our empirical tests span a large, built-out urban municipality so this concern is minimized.

Partial dependence, and the individual conditional expectations that drive it, can be used to extract the marginal impact of each time period, conditionally, on the response or dependent variable: house prices in this case. The complexity of the resulting shape of the partial dependency – linear, monotonic, sinusoidal, spline-like, etc. – is entirely dependent on the flexibility of the underlying model being evaluated. Conceptually, an individual conditional expectation plot takes a single observation, X_i, and for one of the features or variables, X_s, simulates the predicted value of that observation under the hypothetical condition that this observation has the each individual unique value of X_s found in the entire dataset. By holding all other features constant, the marginal value of feature s on observation X_i can be simulated. This represents an Individual Conditional Expectation (ICE). Averaging across all X create a measure of partial dependency. Partial dependency is often illustrated by plotting, which is known as a partial dependency plot (Friedman 2001).

Converting this process to a real estate use for the purpose generating a house price index means valuing a given property (X_i) as if it had each unique value of time of sale (X_s) in the dataset. In other words, simulate the value of a property as if it had sold once in each time period. Do this for all properties in the dataset and average to get the full partial dependency of sale price on time of sale. A key point here is that any type or class of model could be used to simulate the series of value predictions; the approach is model agnostic.

An Example Figure 1 illustrates example plots of an individual condition expectation (left panel) and the overall partial dependency (right) derived from a random forest model. The left hand panel applies an ICE approach on top of a random forest model with time as the variable of interest. Each point on the line, 48 in total, represents the estimated price of the example property at each month over hypothetical four-year time frame. Applying this same approach to all homes in a dataset (695 in this example), provides the thin black lines in the right hand panel. Averaging the full set of ICEs results in the partial dependency, shown in thick red line. Note that the results are expressed in raw dollar values as the partial dependency still needs to be converted to an index.

Figure 1: Example of ICE and PD Plots

Conceptual Framework

In conceptualizing how an interpretable machine learning process could map onto the standard approach(es) for creating house price indexes, it is helpful to abstract the generic process. Broadly, estimating a house price index involves the following steps:

1) Choose a **model** and apply it to the data with the purpose of explaining house prices. The chosen model will need to have a specification that accounts for one or more temporal variables or features in order to allow the model to capture or express any impacts that time may be having on prices.

2) Subject the model results to an **interpretability method** to generate insight into the data generating process (DGP). For some models this is inherent (median by time period) and for others it is a standard output (regression beta coefficients). However, the output of many machine learning models will provide only predicted values. In these cases, a post-model interpretability method will need to be applied.

3) Take the inherent or derived **insights into the DGP** – the marginal contributions of each time period to price – and convert those into an index via one of a standard set of indexes procedures.

Figure 2: Conceptual Model

More simply, this can be mapped to three decisions or steps in the process. The table below maps the three steps to actual processes from a standard hedonic price model example.

Table 1: HPI Process

Step	Description
(1) Choose a model	Specify a hedonic regression model using control variables and some configuration of temporal variables
(2) Choose an interpretability method	Extract the coefficients on the temporal variables as the marginal contribution of each time period toward observed prices in the data
(3) Choose an indexing method	Convert these coefficients to an index via the Laspreyes approach

With this framework, we can now extend the creation of house price indexes to any class of model, machine learning or otherwise, provided that a sufficient interpretability method can be applied to extract or explain the marginal impact of time period on prices. In the interpretable random forest (IRF) example above (Figure 1), the partial dependence estimates provide the 'insight into the data generating process' – the impact of time on price – that is used to generate the house price index.

Interpretable Random Forest Model specification for a random forest is similar to those of standard hedonic price models. The dependent variable (response) is the price of the home (logged in this case) and the independent variables (features) are those factors that are believed to explain variance in the price:

$log(P) = f(S, L, T)$

where P is the sale price, S are structural features of the home (including lot size), L are locational features and T are temporal features. More specifically, the structural features, S include home size (sq.ft.), bedroom count, bathroom count, building quality and use type (SFR or townhome), locational features, L, are latitude and longitude and the temporal feature, T, is the month of sale.

Random forest models also require parameters to control how many trees are grown, how many variables are considered at each split ("mtry") and how small each final node of the tree can be. In each case here 500 trees are grown, using an "mtry" of 3 and a minimum node (or leaf) size of 5. The `ranger` R package is used to estimate the random forest models (Wright and Ziegler 2017).

Hedonic Model To keep the comparison as 'fair' as possible, the hedonic model uses the same set of independent variables as the random forest:

$log(P) = f(S, L, T)$

where P is the sale price, S are structural features – home size, lot size, bedrooms, baths, quality and use type – of the home, L are locational features – latitude and longitude – and T are temporal features. The temporal features in the hedonic model are treated as monthly dummy variables instead of a numeric vector as in the random forest. This allows the hedonic model to identify non-monotonic changes in prices over time – an ability that would not be possible if time were treated as an integer variable.

Following the advice of Bourassa et al (2016), we specify a robust regression to help minimize the impact of any outliers or data errors that have avoided filtering. Specifically, I use the `robustbase` R package to estimate a MM-estimator with a bi-square redescending score function (Maechler et al 2019).

Repeat Sales Many implementations of repeat sales models implement Case and Shiller's (1989) three stage weighted approach that provides greater weight to sale pairs with shorter holding periods. Work by Steele and Goy (1997) suggest that this may be a biasing factor as shorter holds are often less representative of standard home purchases and resales as the initial sale is more likely to be an opportune buyer. As a result of this and of work by Bourassa et al (2013), we do not weight direct by holding period length but, again, opt for a robust regression approach to help moderate any influence from outlying observation and/or changes to quality between sales that was not caught in the data preparation stage. Here, too, the `robustbase` R package is used with an MM-estimate with a bi-square redescending score function (Maechler et al 2019). The standard formulation of the repeat sales model with a logged dependent variable:

$$log(y_{it}) - log(y_{is}) = \delta_2(D_{2,it} - D_{2,is}) + ... + \delta_\tau(D_{\tau,it} - D_{\tau,is}) + u_{it} - u_{is}$$

where y_{it} is the resale, y_{is} is the initial sale and the $D_{\tau,i}$s are the temporal period dummies, -1 for the period of the first sale, 1 for the period of the second sale and 0 for all others.

Data

The data for this study originate from the King County (WA) Assessor. All transactions of single family and townhome properties within the county during the January 2017 through December 2022 period are included. The data are found in the `kingCoData` R package and can be freely downloaded at one of the author's Github pages as well as on Kaggle. The transactions were filtered to keep only arms-length transactions based on the County's instrument, sale reason and warning codes. Additionally, any sale that sold more than once and underwent a major renovation between sales was removed as these transactions violate the constant quality assumptions made in the repeat sales models estimated below. Finally, a very small number of outlying observations – those with sales under $150,000 and over $10,000,000 were removed. The data includes the following information for all 150,876 transactions remainder after the filtering applied above.

Table 2: Data Fields

Field Name	Type	Example	Description
pinx	chr	..0007600046	Tax assessor parcel identification number
sale_id	chr	2021..2621	Unique sale identifier
sale_price	integer	308900	Sale price
sale_date	Date	2021-02-22	Date of sale
use_type	factor	sfr	Structure type
area	factor	15	Tax assessor defined neighborhood or area
lot_sf	integer	5160	Size of lot in square feet
wfnt	binary	1	Is the property waterfront?
bldg_grade	integer	8	Structure building quality
tot_sf	integer	2200	Total finished square feet of the home
beds	integer	3	Number of bedrooms
baths	numeric	2.5	Number of bathrooms
age	integer	100	Age of home
eff_age	integer	12	Years since major renovation
longitude	numeric	-122.30254	Longitude
latitude	numeric	47.60391	Latitude

Within the data, there are 13,697 sale-resale pairs. This set of repeat transactions is limited to those which have at least a one-year span between the two sales. This constraint is applied to avoid potential home flips, which more often than not violate constant quality assumptions (Steele and Goy 1997; Clapp and Giacotto 1999).

In addition to comparison on performance at the global (King County) level, we also break the data into 19 submarkets (Figure 3). These submarkets are based the King County Assessor's 95 major residential tax assessment zones. Using combinations of tax assessment zones is preferable to common disaggregating regions such as Zip Codes as the tax assessment zones are relatively balanced in total housing unit counts and purposefully constructed to follow local housing submarket boundaries.

Figure 3: Study Area w/ Sales

Results

We begin by comparing the three approaches to index generation (Figure 4). The Hedonic and Repeat Sales approaches show very similar movements over time, with, perhaps, the only noticable systematic difference being slightly higher index values from the Hedonic index during the 2017 to 2020 period. The index derived from an Interpretable Random Forest (IRF) tells a much different story. The appreciation and corresponding depreciation rates (in late 2022) are considerably smaller in magnitude. Additionally, the overall month-over-month volatility is also much smaller as evidenced by the very smooth movements in prices suggested by this approach. Without even computing volatility it is evident that the IRF method is clearly the smoothest of the estimation approaches.

Figure 4: County Level Indexes

Interestingly, the IRF index falls below the other two by 7-10 index points after the first year and generally holds that lower level for most of the study period, including similarly tracking the high price appreciations of early 2021. During the next price runup (and subsequent fall) in 2022; however, the IRF method offers significantly reduced upward price trends. Finally, all three indexes end up nearly at the very same value for the end of 2022 – an index of approximately 153. Does this mean that the Hedonic and Repeat Sales approaches simply over-reacted in 2022, but not in 2021? To help answer that question we'll use the various criteria of index quality discussed above.

Accuracy

In the existing literature, the most used measure of accuracy is the ability of an index to accurate predict the second sale for a home sold twice during a study period. We've termed this Absolute Accuracy as the measure examines the absolute predictive error versus the known resale price of the home. This approach to accuracy benefits from the clarity of the method and the ease at which it can be employed. From a comparative perspective, however, it had three downfalls: 1) It only uses a small sample of the data; 2) Some of the sales likely suffer from violations of the assumption of constant quality between the repeat sales; and 3) Being evaluated only on repeat sales, it likely favors repeat sales models since their linear estimator is focused on estimating this sample of homes.

Errors in this case are the difference between actual resale price and the prior sale price indexed forward in time by the candidate indexing methods. We look at four different measures of accuracy, two measuring bias (distribution of the center of the errors) and two measure precision (width of the distribution of errors):

- MdPE: Median Percent Error
- MPE: Mean Percent Error
- MdAPE: Median Absolute Percent Error
- MAPE: Mean Absolute Percent Error

Table 3 shows the comparative results for Absolute Accuracy across the three models. From a statistical bias perspective, the Repeat Sales approach dominates the two other as it is nearly unbiased, while the Hedonic approach offers a 2-3% low bias and the IRF a ~5% low bias. Turning to precision, we see that the Repeat Sales and Hedonic approaches are roughly equivalent, with IRF being about 12% (4%) less precise in the median (mean).

Table 3: Absolute Accuracy

Model	MdPE (Bias)	MPE (Bias)	MdAPE (Precision)	MAPE (Precision)
Repeat Sales	0.004	0.001	0.086	0.120
Hedonic	-0.028	-0.025	0.087	0.120
IRF	-0.053	-0.053	0.097	0.125

The fact that the precision estimates are nearly identical between the Repeat Sales and Hedonic model approaches but the bias measures are quite different may hint that there are some upwardly biased repeat sales in the data. In Figure 4 we plot the median model error (bias) over time. From this plot we see a number of very low bias estimates for the non-Repeat Sales models in 2018 – a period where the majority of the repeat sales will be short-term holds, indicative of renovations not caught by our cleaning process.

Additionally, Figure 4 shows us that during the periods of the most rapid price inflation, first halves of 2021 and 2022, all approaches lag the market a bit, but the IRF method does lags by the largest amount. This presents some evidence to suggest that it, indeed, may be overly smooth or too slow to react to quick market changes.

Figure 5: Absolute Accuracy over Time

Relative Accuracy As discussed in the method section, there is a second and novel approach to measuring the accuracy of an index. This new approach involves the comparison of its ability to make relative accuracy improvements within an AVM; specifically when comparing an AVM with no time controls to one with sale prices time adjusted by the candidate index. Any AVM, no matter how simple, will do here since it is the relative accuracy gains by time adjusting that we are measuring.

Table 5 shows the same four accuracy measures, this time compared against a baseline, non-temporal AVM. The errors figures from Table 5 are not directly comparable with those from Table 4 as they are measuring errors on different set of data and in a different way. The important measure of comparison here is each method against each other and the baseline and not against other methods of measuring accuracy.

In terms of bias, all indexes are considerably better than the baseline. This is an expected finding given the large price movements in the Seattle region over this time period. Of the methods the Hedonic approach offers the least biased results, slightly better than the Repeat Sales and more so than the IRF. In terms of precision, the differences between the approaches are quite minimal, again with the Hedonic approach slightly edging out the Repeat Sales and IRF methods.

Table 4: Relative Accuracy

Model	MdPE (Bias)	MPE (Bias)	MdAPE (Precision)	MAPE (Precision)
Baseline (no index)	-0.128	-0.148	0.169	0.224
Repeat Sales	0.016	0.009	0.124	0.174
Hedonic	0.003	-0.006	0.121	0.172

Model	MdPE (Bias)	MPE (Bias)	MdAPE (Precision)	MAPE (Precision)
IRF	-0.018	-0.027	0.125	0.174

Examining the index biases over time, we see a similar trend to the Absolute Accuracy measures. All methods lag price appreciation during period of rapid price growth and, similar, lag price depreciation during periods of market decline. Of these, the IRF method present the biggest lags in either direction.

Figure 6: Relative Accuracy over Time

Revision Next, we look at the severity of index revisions across the three approaches. Index revisions occur when future data changes the estimator resulting in updates to prior estimates of market movements. These existing estimates then must be revised in the presence of the new data.

Within this work we measure period-wise revision as the individual mean of the revisions for each period in the index as it expands out to cover the entire time period. The revision for period k is: –>

$R_k = \sum (K_j - K_{j-1})/j$

where j is the new index being generated after each addition of data.

In the analysis analyzed below, we begin each with a 12-month period of training data to create the first index. WE then add in data from period 13 and recalculate the index, measuring the revision for periods 1 to 12. The same is then done for period 14 (measuring revision for periods 1 to 13) and on up through period 60. There is no revision number for period 60 as it is only estimated once.

Table 6 shows the median and the mean revisions, directionally, for the three different index creation approaches. As shown by others (Clapham et al 2006; Deng and Quigley 2008), we see considerable revisions for the Repeat Sales model. This is expected as when new resales enter the data all prior period estimates are influenced. What is most interesting about the Repeat Sales revisions is the directionality. The strong negative revisions suggest that early, short hold resales (like flips) may push index values up, only to be revise down later by longer hold resales.

The Hedonic approach offers revisions at magnitude of only 20% of the Repeat Sales approach. These revision average out to less than 1% and usually are in the upwards direction. Revisions in Hedonic approach are a product of the estimation strategy. In this paper we use a full period, dummy estimator method which through the changes in non-temporal beta coefficients over time can create revision values. As Hill (2013) and other note, using a rolling or chained approach can help mitigate the revisions in hedonic approaches.

Finally, with the IRF approach we see almost no revisions over time. This is expected given the branching technique that decisions trees use when building an estimator. If there is little to no market movement, time will not be a spitting feature and the estimates will remain static with respect to sale date. Conversely, if time trends are seen then the trees will likely split on the temporal feature, leaving earlier periods unaffected by the new data. It is not impossible under the current construction that we have for there to be revisions – in fact there are slight revisions, less than 0.1% – they are just likely to be very minimal given this type of algorithm.

Table 5: Revision

Model	Median	Mean
Repeat Sales	-0.054	-0.029
Hedonic	0.008	0.005
IRF	<0.000	<0.000

Volatility Volatiliy is two way street. Complete lack of volatilty would result in a complete flat index, on that, in the presence of market changes would prove inaccurate and mostly useless. On the other hand, two much volatility likely represent 'noise' in the data and does not represent the slow movements in the actual market. Real estate is not the stock market; the relative illiquidity of it usually means that markets are slow to react to all but the most severe exogenous signals.

As a result, we want to see index volatility that tracks the actual market but no more. To truly gauge the 'right' level of volatility in an index we must balance it off against predictive accuracy.

Beginning by looking at the raw volatility measures – the mean of the three month rolling standard deviations – we see that the Repeat Sales method is the most volatile and the IRF the least, by a long margin. The Hedonic method falls close to the Repeat Sales approach, but is slightly smoother. As we see in Figures 5 and 6, the IRF method struggles during period of high price movements in either direction suggesting that it is indeed too smooth

Table 6: Volatility

Model	Median	Mean
Repeat Sales	0.011	0.013
Hedonic	0.009	0.011
IRF	0.002	0.004

Local Specificity

The final perspective from which we will examine index quality is their ability to be applied to smaller submarkets. To do so, we've estimated each of the three indexes individually for the 19 different submarkets shown in Figure 3. We've then calculated both Absolute and Relative Errors at the submarket level and pooled those to create overall error measures for the entire county. By doing so we are able to measure if the index's predictive ability changes when we are applying it to a smaller region.

Here we will examine just a single metric of dispersion as the precision metrics more closely map the ability of a model to remain in high fidelity to differing local conditions. We choose to examine only the mean as the overall rank ordering in the median in the global county case was similar (Tables 3 and 4).

Table 7 shows the results for the Relative Accuracy approach. In the left most column are the results from the county-wide model (same as in Table 3). In the center column we see the pooled results from the individual local indexes and in the right most column the change in index quality when moving from a global to multiple local indices. Interesting, the repeat sales estimator gets significantly worse, moving from an average error in predicting the resale value of 12% up to nearly 16%, a decrease in quality of 32%. Conversely, the Hedonic model improved slightly by 2% and the IRF remained mostly unchanged, dropping just 1%.

Table 7: Local Absolute Accuracy

Model	MAPE (Global)	MAPE (Local)	Difference
Repeat Sales	0.120	0.159	-32%
Hedonic	0.119	0.117	+2%
IRF	0.125	0.126	-1%

Next, we do the same exercise for the relative accuracy measure. We see the same general stack ranking of improvement. The repeat sales model again gets worse at the local level, though down only 6% this

time. The hedonic and IRF model actually improve, 15% and 14% respectively, suggesting that these two approaches are much better able to capture differing local trends, when they exist.

As we saw in the global numbers, the IRF approach is competitive with the others in a Relative Accuracy framework, but not in the Absolute one, suggesting that there is something meaningfully different about the two different methods of measuring accuracy, particularly when using either the Repeat Sales or the IRF approach. The Hedonic model remains the top model – in mean dispersion – regardless of the method of measuring or the global or local nature of the analysis.

Table 8: Local Relative Accuracy

Model	MAPE (Global)	MAPE (Local)	Difference
Repeat Sales	0.174	0.184	-6%
Hedonic	0.172	0.146	+15%
IRF	0.174	0.149	+14%

Discussion

In this paper, we present two new measures for evaluating home price indexes: 1) Relative Accuracy; and 2) Volatility. We add these to the more common measures of Absolute Accuracy and Revision to create a four metric matrix across which to compare index performance. Additionally, we use our accuracy measures to create a direct comparison of the changes in index's predictive accuracy when we move to smaller submarkets. Finally, we also look more subjectively at each index type to gauge the ease of which the method can be employed.

Qualitatively our findings are summarized in Table 9. From an absolute accuracy perspective, the Repeat Sales and Hedonic approaches showed very similar performance, with the Repeat Sales sightly more performant, particularly in measure of bias. This should not be surprising as the absolute accuracy approach to measuring index quality uses repeat sales as the benchmark, effectively giving repeat sales based indexes an advantage in this regard.

The novel Relative Accuracy approach that we developed and test in this paper shows different results. With this measure of accuracy the Hedonic method is the more accurate and the IRF method looks competitive. While the Relative Accuracy approach may need more testing and validation before more widespread adoption, the marked differences in performance between the two approaches to measuring accuracy should give researchers and applied users pause when using the Absolute method, particularly if they are developing or using an index for the purpose of time adjustment in a valuation exercise.

Table 9: Index Quality Criteria Ratings

Model	Absolute Accuracy	Relative Accuracy	Revision	Volatility	Local	Constructability
Repeat Sales	Best	Near Best	High		Loss of Accuracy	Easy Data and Method
Hedonic	Near Best	Best	Very Low		Gains in Accuracy	Moderate Data and Easy Method
IRF	Worst	Near Best	Almost None		Gains in Accuracy	Moderate Data and Complex Method

Our findings on revision echo those of prior work (ex. Van de Minne 2020); namely that repeat sales approaches are often subject to high rates of revision, rates that are much higher than those from hedonic based methods. The novel IRF method that we explore in this paper is subject to almost no revision as we build out a series of increasing longer indexes over time. Like the repeat sales tendency for revision, the IRF's stability against it is a product of it method of construction. For use cases where one might be willing to trade some accuracy for reduced revision and IRF approach may be a good choice.

The above results all cover indexes at a relatively wide market scale – that of King County, Washington the heart of the Seattle MSA and home to over 2 million people. When we move to local indexes at about 1/20 the overall size of the county there are marked changes in the accuracy of the indexes. At this submarket size the repeat sales indexes become considerably less performant compared to the hedonic and IRF indexes. Given that repeat sales methods only use a small sample of the data – those homes that sell twice in the study period – reducing either the geographic or temporal sample can have much bigger impacts on repeat sales indexes that other than use all transactions from the market. Users looking to measure smaller region price movements should be cautious of employing repeat sales approaches, particularly over short time period such as five years.

Finally, we look at the constructability of each method. The repeat sales approach is the most easily built. It requires the simplest data – no facts about the home other than address or unique identifier – and, often, uses the simplest estimation models such as weighted linear models. Hedonic method require more complex data as they need information on the home attributes such as home size, age, etc. The estimation models can be relatively simple, however, with linear models and temporal dummy variables offering the baseline method here. The IRF method needs the same data as the Hedonic model but requires slightly more complex and harder to explain and replicate estimation method. The compute time is also generally longer, though it is heavily dependent on the hyperparameters that are used.

Overall, our recommendations are the following:

- A repeat sales approach is the preferred option (over a median sale price, for example) if you do not have home level attribute data. However, you should be wary of using this approach if your market is small and/or your use case is subject to negative effects from revisions.
- A hedonic model offers consistently high performance across all criteria, is relatively easy to estimate and can be used at more granular sub-market levels. If you have home level attribute data, you should start with this approach due to its flexibility.
- The novel IRF method struggles during times of quick price movements and should be used with caution. At this point in time we'd only recommend use this if low volatility and low revision are your primary criteria, particularly in use cases where it is supporting time adjustment in valuation cases like an AVM.

Reproducibility and Software

This work is completely reproducible. All raw data, code and general instructions to exactly recreate the analyses above is found at https://www.github.com/andykrause/irf. All code is written in the R statistical language. In addition to the `hpiR` package, which includes the custom functions for the IRF models and the wrapper functions that make for easy computation of accuracy, volatility and revision figures this work also directly uses the following R packages: `caret`(Kuhn 2019), `dplyr`(Wickham et al 2019), `forecast`(Hyndman et al 2019), `ggplot`(Wickham 2016), `imputeTS`(Moritz and Bartz-Beielstein 2017), `knitr`(Xie 2019), `lubridate`(Grolemund and Wickham 2011), `pdp`(Greenwell 2017), `purrr`(Henry and Wickham 2019), `ranger`(Wright and Ziegler 2017), `robustbase`(Maechler et al 2019), `tidyr`(Wickham and Henry 2019) and `zoo`(Zeileis and Grothendieck 2005).

References

Abraham, J. M., & Schauman, W. S. (1991). New evidence on home prices from Freddie Mac repeat sales. *Real Estate Economics*, 19(3), 333-352.

Ahlfeldt, G. M., Heblich, S., and Seidel, T. (2023) Micro-geographic property price and rent indices, *Regional Science and Urban Economics*, 98.

Apley, D. (2016). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. https://arxiv.org/abs/1612.08468

Bailey, M., Muth, R., & Nourse, H. (1963). A Regression Method for Real Estate Price Index Construction. *Journal of the American Statistical Association*, 58, 933-942.

Bogin, A. N., Doerner, W. M., Larson, W. D., & others. (2016). Local House Price Dynamics: New Indices and Stylized Facts *FHFA Working Paper*.

Bogin, A N., Doerner, W and Larson, W. (2019) Missing the mark: Mortgage valuation accuracy and credit modeling. Financial Analysts Journal 75.1.

Bokhari, S. & Geltner, D. J. (2012). Estimating Real Estate Price Movements for High Frequency Tradable Indexes in a Scarce Data Environment. *The Journal of Real Estate Finance and Economics* 45(2), 533-543.

Bourassa, S., Cantoni, E., & Hoesli, M. (2016). Robust hedonic price indexes. *International Journal of Housing Markets and Analysis*, 9(1), 47-65.

Breiman, L. (2001) Random Forests. *Machine Learning* 45(1), 5-32. doi:10.1023/A:1010933404324

Case, B., Pollakowski, H. O., & Wachter, S. M. (1991). On choosing among house price index methodologies. *Real Estate Economics*, 19(3), 286-307.

Case, B., Pollakowski, H. O., & Wachter, S. (1997). Frequency of transaction and house price modeling. *The Journal of Real Estate Finance and Economics*, 14(1), 173-187.

Case, B. & Quigley, J. M. (1991). The dynamics of real estate prices. *The Review of Economics and Statistics*, 50-58.

Case, K. & Shiller, R. (1987). Prices of Single Family Homes Since 1970: New Indexes for Four Cities. *New England Economic Review*, Sept/Oct, 45-56.

Case, K. & Shiller, R. (1989). The Efficiency of the Market for Single Family Homes. *The American Economic Review*, 79(1), 125-137.

Clapham, E., Englund, P., Quigley, J. M., & Redfearn, C. L. (2006). Revisiting the past and settling the score: index revision for house price derivatives. *Real Estate Economics*, 34(2), 275-302.

Clapp, J. M., & Giaccotto, C. (1999). Revisions in Repeat-Sales Price Indexes: Here Today, Gone Tomorrow? *Real Estate Economics*, 27(1), 79-104.

Clapp, J. M., Giaccotto, C., & Tirtiroglu, D. (1992). Repeat sales methodology for price trend estimation: an evaluation of sample selectivity. *Journal of Real Estate Finance and Economics*, 5(4), 357-374.

Cohen, SB, Dror, G & Ruppin, E (2005) Feature Selection Based on the Shapley Value. in *Proceedings of IJCAI*. pp. 1-6.

Crone, T. M., & Voith, R. (1992). Estimating house price appreciation: a comparison of methods. *Journal of Housing Economics*, 2(4), 324-338.

Deng, Y., & Quigley, J. M. (2008). Index revision, house price risk, and the market for house price derivatives. The Journal of Real Estate Finance and Economics, 37, 191-209

Doshi-Velez, F. and Kim, B. (2017) Toward a Rigorous Science of Interpretable Machine Learning. *arXiv::1702.08608*. https://arxiv.org/abs/1702.08608

Englund, P., Quigley, J. M., & Redfearn, C. L. (1999). The choice of methodology for computing housing price indexes: comparisons of temporal aggregation and sample definition. *The Journal of Real Estate Finance and Economics*, 19(2), 91-112.

Eurostat (2013) Handbook on Residential Property Prices Indices (RPPIs). *Eurostat: Methodologies and Working Papers* doi:10.2785/34007

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics* 1189-1232.

Gatzlaff, D. H., & Haurin, D. R. (1997). Sample Selection Bias and Repeat-Sales Index Estimates. *The Journal of Real Estate Finance and Economics*, 14, 33-50.

Gatzlaff, D. H., & Haurin, D. R. (1998). Sample Selection and Biases in Local House Value Indices. *Journal of Urban Economics*, 43, 199-222.

Goh, Y., Costello, G. and Schwann, G. (2012) Accuracy and Robustness of House Price Index Methods, Housing Studies, 27:5, 643-666

Goldstein, A., Kapelner, A., Bleich, J. and Pitkin, E. (2014) Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectiation. https://arxiv.org/pdf/1309.6392.pdf

Goodman, A. C. (1978). Hedonic prices, price indices and housing markets. *Journal of Urban Economics*, 5, 471-484.

Greenwell, B. (2017). pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal*, 9(1), 421–436. URL https://journal.r-project.org/archive/2017/RJ-2017-016/index.html.

Gregorutti, B., Bertrand, M., and Saint-Pierre, P. (2017) Correlation and variable importance in random forests. *Statistics and Computing*, 27(3), 659-678.

Grolemund, G. and Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, 40(3), 1-25. URL: http://www.jstatsoft.org/v40/i03/.

Guntermann, K. L., Liu, C., & Nowak, A. D. (2016). Price Indexes for Short Horizons, Thin Markets or Smaller Cities. *Journal of Real Estate Research*, 38(1), 93-127.

Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference and Predcition.* Springer.

Haurin, D. R., & Hendershott, P. H. (1991). House price indexes: issues and results. *Real Estate Economics*, 19(3), 259-269.

Henry, L. and Wickham, H. (2019). purrr: Functional Programming Tools. R package version 0.3.2. https://CRAN.R-project.org/package=purrr

Hill, R. (2013) Hedonic Price Indexes for Residential Housing: A Survey, Evaluation and Taxonomy. *Journal of Economic Surveys* 27(5), 879-914. https://doi.org/10.1111/j.1467-6419.2012.00731.x

Hill, R. C., Knight, J. R., & Sirmans, C. F. (1997). Estimating capital asset price indexes. *Review of Economics and Statistics*, 79(2), 226-233.

Hill, R. and Trojanek, R. (2022) An evaluation of competing methods for constructing house price indexes: The case of Warsaw. *Land Use Policy*, 120, 106226.

Hoesli, M., Giacotto, C., and Favarger, P. (1997) Three new real estate price indices for Geneva, Switzerland. *The Journal of Real Estate Finance and Economics*, 15(1), 93-109.

Hyndman, R., Akram, M. and Archibald, B. (2008) The admissible parameter space for exponential smoothing models. *Annals of the Institute of Statistical Mathematics*, 60(2), 407-426.

Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2019). *forecast: Forecasting functions for time series and linear models.* R package version 8.7, <URL: http://pkg.robjhyndman.com/forecast>.

Kuhn, M. (2019) caret: Classification and Regression Training. R Package. https://CRAN.R-project.org/package=caret

McMillen, D. (2012). Repeat sales as a matching estimator. *Real Estate Economics*, 40(4), 745-773.

Maechler, M., Rousseeuw, P. Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E., and di Palma, M.A. (2019). robustbase: Basic Robust Statistics R package version 0.93-5. http://CRAN.R-project.org/package=robustbase

Maguire, P., Miller, R., Moser, P. and Maguire, R. (2016) A robust house price index using sparse and frugal data, *Journal of Property Research*, 33:4, 293-308, https://doi.org/10.1080/09599916.2016.1258718

Mayer, M., Bourassa, S., Hoesli, M., and Scognamiglio, D. (2019). Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Research.* 12(1), 134-150. https://doi.org/10.1108/JERER-08-2018-0035.

Meese, R. A., & Wallace, N. (1997). The construction of residential housing price indices: a comparison of repeat-sales, hedonic-regression, and hybrid approaches. *The Journal of Real Estate Finance and Economics*, 14(1), 51-73.

Molnar, C. (2019) *Interpretable Machine Learning: A Guide for Making Black Box Model Explainable.* Leanpub. ISBN 978-0-244-76852-2. https://christophm.github.io/interpretable-ml-book/

Moritz S, Bartz-Beielstein T (2017). "imputeTS: Time Series Missing Value Imputation in R." *The R Journal*, *9*(1), 207-218. doi:10.32614/RJ-2017-009 (URL: https://doi.org/10.32614/RJ-2017-009).

Munneke, H. J., & Slade, B. A. (2000). An empirical study of sample-selection bias in indices of commercial real estate. *The Journal of Real Estate Finance and Economics*, 21(1), 45-64.

Nagaraja, C., Brown, L., & Wachter, S. (2014). Repeat sales house price index methodology. *Journal of Real Estate Literature*, 22(1), 23-46.

Nowak, Adam D., and Patrick S. Smith. (2020) "Quality-adjusted house price indexes." American Economic Review: Insights: 339-356.

Quigley, J. M. (1995). A simple hybrid model for estimating real estate price indexes. *Journal of Housing Economics*, 4(1), 1-12.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ren, Y., Fox, E. B., & Bruce, A. (2017). Clustering correlated, sparse data streams to estimate a localized housing price index. The Annals of Applied Statistics, 808-839.

Ribiero, M., Singh, S. and Guestrin, C. (2016) Model-agnostic Interpretability of Machine Learning. *arXiv::1606.05386.* https://arxiv.org/abs/1606.05386

Ribiero, M., Singh, S., and Guestrin, C. (2016b) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* pp 1135–1144, doi: 10.1145/2939672.2939778.

Rudin, C. and Carlson, D. (2019) The Secrets of Machine Learning: Ten Things You Wish You Had Known Earlier to Be More Effective at Data Analysis. *Tutorials in Operations Research* https://arxiv.org/pdf/1906.01998v1

Sayag, D., Ben-hur, D., & Pfeffermann, D. (2022). Reducing revisions in hedonic house price indices by the use of nowcasts. International Journal of Forecasting, 38(1), 253-266.

Slack, D., Friedler, S., Scheidegger, C. and Roy, C.D. (2019) Assessing the Local Interpretability of Machine Learning Models. https://arxiv.org/pdf/1902.03501.pdf

Steele, M., & Goy, R. (1997). Short holds, the distributions of first and second sales, and bias in the repeat-sales price index. *The Journal of Real Estate Finance and Economics*, 14(1), 133-154.

Steurer, M, Hill, R. and Pfeifer, N. (2021) Metrics for evaluating the performance of machine learning based automated valuation models. Journal of Property Research, 38.2, 99-129.

Tofallis, C (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, 66, 1352-1362. doi:10.1057/jors.2014.103

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wickham, H., Francois, R., Henry, L. and Muller, K. (2019). dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr

Wickham, H. and Henry, L. (2019). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.8.3. https://CRAN.R-project.org/package=tidyr

Wright, M. and Ziegler, A. (2017) ranger: A Fast Implmentation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 1-17.

Van de Minne, Alex et al. "Using Revisions as a Measure of Price Index Quality in Repeat-Sales Models." Journal of Real Estate Finance and Economics 60 (May 2020): 514–553

Xie, Y. (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.23.

Xu, X. and Zhang, Y. (2022) Second-hand house price index forecasting with neural networks, *Journal of Property Research*, 39:3, 215-236.

Zeileis, A. and Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software*, 14(6), 1-27. doi:10.18637/jss.v014.i06