A Multi-Criteria Evaluation of House Price Indexes

Andy Krause^ – Zillow Group Reid Johnson – Zillow Group

2024-11-16

Abstract

This work refines the current typology of house price indexes by dividing multiple popular and two novel methods into their fundamental trend extraction techniques. We then evaluate these methods by a collection of metrics aimed at measuring index volatility, revision and accuracy. We find that both at a county and at a smaller submarket level index performance varies widely depending on the extraction method and the evaluation metrics. Overall, a traditional hedonic time extraction approach offers the most consistent performance across all metrics, though rarely the best in any one.

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).

In the comparative work, methods are usually grouped by the type of data they use – repeat sales (cit), all sales with home attributes (cit) or tax value ratios (cit) – and/or by the algorithm used to fit the model – OLS vs xxx vs yyy (cit). We propose that grouping home price index by the fundamental method of trend extraction is a preferred categorization strategy because it most closely maps to the generalization ability of the index. More specifically, we split methods into aggregration (AGG), trained model extraction (TME) and imputation (IMP).

Additionally, research aimed at measuring the quality of house price indexes has received considerable less attention that the four main sets of literature noted above: 1) Comparison; 2) Issue identification; 3) Submarketing; and 4) New estimators. 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.

Additionally, we compare the ability of the test set of index methods to perform at smaller geographies; a common desire for many house price index (HPI) use cases. Finally, to the set of existing index methods we add two novel approaches – an TME index developed with a neural network and an imputation (IMP) method developed with a random forest.

Index Method Typology

The existing work tends to compare/contrast methods based on the specifics of the algorithm used. The most common are comparative studies between repeat sales and hedonic methods (cit) and/or between various implementation deviations of these two methods such as robust statistical modeling techniques (cit).

When looking at the possible approaches to creating house price indexes, we find that a two level, hierarchical taxonomy more appropriately captures explains the differences. In this taxonomy each house price index has a 1) Method; and 2) Implementation. The *Method* denotes the broad mathematical approach to deriving measures of price changes over time, the *Implentation* addresses the finer details of the approach such as the statistical technique, the hyperparameters and or the data used.

In reviewing the existing summarizing work (ex. Hill, cit), we find that there are three primaries methods to developing house price indexes: 1) Aggregate, Trained Model Extraction and Imputation.

- Aggregate (AGG): Aggregate price/value observations by time period and take simple distributional measures for each aggregated period
- Trained Model Extraction (TME): Train a stastistical or machine learning model on a price/value observations and extract from the trained model estimates of the model measures of the change in values/prices over the time aggregates.
- Imputation (IMP): Use a model or process to create hypothetical prices/values (impute), and then use an aggregation approach to extract changes over the temporal periods.

The divisions between these approaches may not always be perfect as, for example, one could argue that the imputation approach is simply another way of employing the aggregation method. We've separated them this way due to both the existing literature and the differences in what the various methods allow a user to generalize to, this second factor being the primary motivating factor in our taxonomy.

Example of each:

- Agg: Computing the median sale price per month and converting into an index.
- TME: Fitting a regression model with monthly time dummy variables and using the beta coefficients of the time variables to create an index
- Imp: Fitting the regression model from the TME example, but using it to predict the value of all homes in the area for each time period and then computing the median value of the entire universe each month.

Within these three there are many possible implementations given the variety of different choices that can be made about statistical and/or machine learning modeling techniques, hyperparameters and data. We do not attempt to categorize them all in this paper, but do develop a few examples within each method in the empirical section of this paper to highlight their differences.

Evaluating Indexes

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. To measure volatility we use a Seasonal and Trend decomposition using Loess (STL) approach (Cleveland et al 1990). An STL will extract a long term temporal trend, a seasonal factor and the residuals remaining from the trend and seasonal factors. We treat the residuals from the STL decomposition as the measure of volatility. We find this approach to be a compelling fit for house price indexes do to the prevalence of both non-linear trends over years and fairly consistent seasonal trends within years.

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.

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.

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 2018 through December 2023 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.

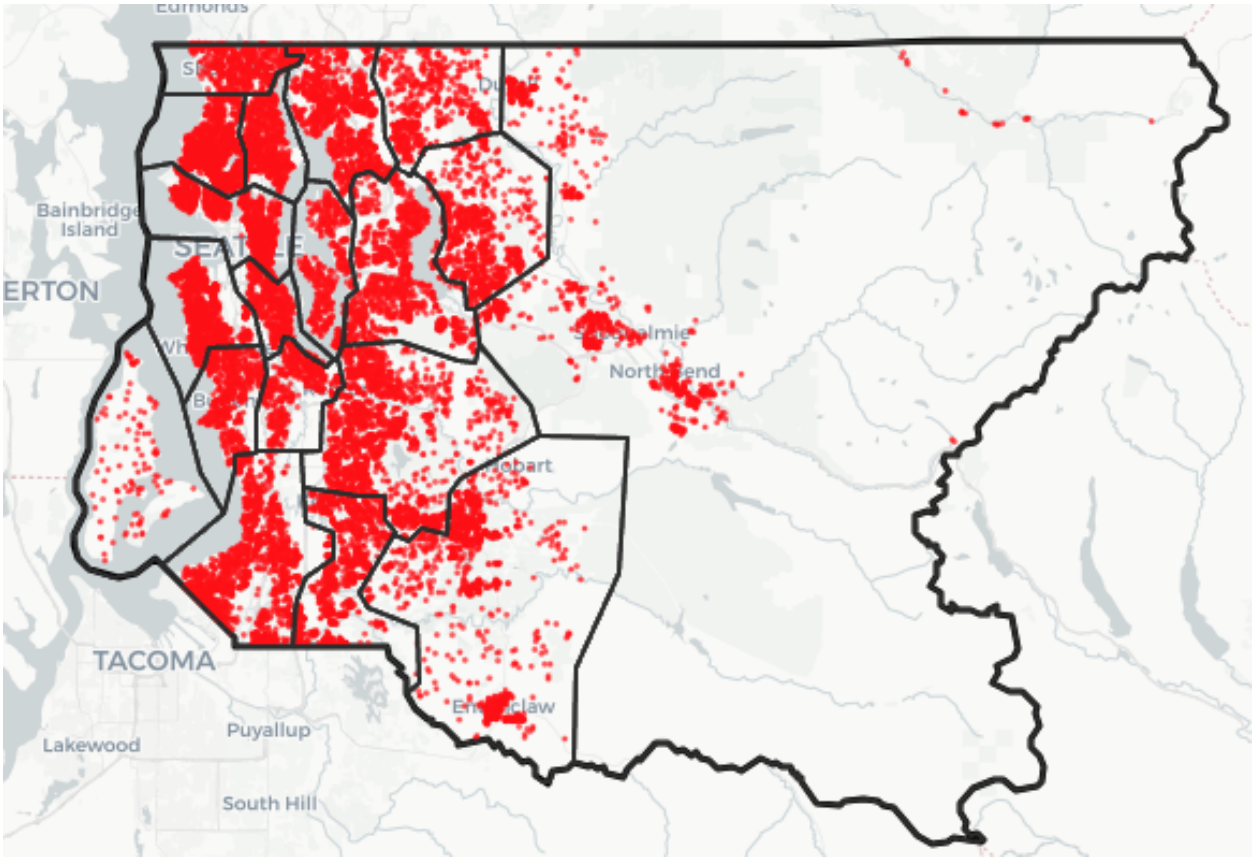
Table 1: Data Fields

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 1). 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 1: Study Area w/ Sales



Models

In this study we compare seven different approaches to creating house price indexes, two from the aggregate (AGG) class, three using trained model extraction (TME) and two with imputation (IMP). We evaluate each of these models on the four metric-based evaluation criteria discussed above, plus local specificity. A comparison of contructability is saved for the discussion section.

Aggregate (AGG) Implementations In the aggregate method we build and test two implementations; 1) Median Sales Price (MED) and 2) Median Sales Price per Square Foot (PSF). These models are as simple as they sound. The median sale price (MED) implementation simply takes the median of the sale price per period and convert this to an index by

« insert formula».

The second aggregation implementation conditions the sales price by home size, creating an index in the same manner, but instead of using raw home sale price, we use the price per square footage of the home.

Trained Model Extraction (TME) Implementations

- 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.

- OLS w/Time Dummies

Model specification is:

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

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).

- Neural Network

https://docs.google.com/document/d/1qpMNRNNT0n4G2py8rwA1XuPtGPgCg71q7OOFjct0o9Q/edit?tab=t.0

With the growing complexity and non-linear structure of real estate data, neural networks present a compelling approach for modeling. These machine learning techniques are adept at representing and capturing the structure and intricacies of real estate datasets, making them attractive for constructing home price indexes (HPIs). Our approach leverages the flexibility of neural networks to disentangle the effects of property-specific characteristics and shared market trends on home prices, aiming to produce HPIs that are highly accurate, granular, and timely.

Our model employs neural networks designed to output estimates of log property values based on input home details, with one or more market indexes being explicitly learned as a component of this process. This differs from traditional hedonic home price indexes that typically aggregate individual property estimates to compute indexes that reflect broader housing price movements over time. By contrast, the market index

effects produced by our neural network architecture are jointly learned with property-level effects and are a component of the property value estimates generated by the model. This architecture is motivated by the prevailing understanding of how property values are determined.

Property values are commonly reported as a product of price and quantity (Clapham), leading to a model formulation of log Vit = log Pit + log Qit + eit , where the log of a property's value at time t (log Vit) equals the sum of the log of the house price index (log Pit), the log of quantity (log Qit), and an error term (eit). Unlike a standard neural network that interconnects all input features, our model routes inputs through either a property-specific or a market-specific pathway. This bifurcation allows for the estimation of log property values by summing log quality and log index effects, thus explicitly distinguishing between these contributions to property valuation.

The joint modeling of property-level characteristics and shared market-level effects reflects the underlying house price dynamics, enabling an accurate estimation of the market-level effects. As the model is trained on property-level data, it can effectively make adjustments for property characteristics to provide a timely and granular measure of home price dynamics that capture average price shifts of properties within a given area, whether it be regional, national, or local.

Implementation Details

Our neural network approach combines an ensemble of quantile regression neural networks to estimate property-level values that factor in both property-level and market-level effects. The estimated index values from the trained model are extracted by exponentiating the output value of the market-level pathway for each desired time and region.

The model inputs are treated as raw numerical values, logarithmically transformed numerical values, categorical values, or ordinal values. The model performs additional transformations to temporal and spatial features. The specific input treatments are as follows: Numerical features: The input features "sqft", "lot sqft", and "age" are treated as numerical features that are logarithmically transformed before input to the model. Categorical features: The input features "present use" and "submarket" are treated as categorical features, with a learned embedding layer created for each feature before input to the model. In our experiments, we use an embedding size of 5 for all categorical features. Ordinal features: The input features "beds", "baths" and "grade" are treated as ordinal features, with an ordinal embedding layer created for each feature before input to the model. Unlike traditional embeddings, this learned ordinal embedding employs a "binary counting" or cumulative encoding strategy to represent ordinal values. Each ordinal level is encoded as a binary array, where the representation builds cumulatively with each increment in the ordinal value, with missing values represented by a zero array. The embedding values are created by passing this binary array through a linear layer. Temporal features: The sale date is decomposed into trend and seasonality features. The number of weeks as an integer counting from the earliest sale date is used as an ordinal input to the model to capture trend effects. The sine and cosine of the week of the year are used as numerical inputs to the model to capture seasonal effects. Spatial features: The latitude and longitude features are transformed into H3 grid cells (Brodsky) that are treated as categorical inputs. In our experiments, we use H3 cells of resolutions 6 and 7 to represent geography at multiple scales. The numerical features, categorical embedding outputs, and ordinal embedding outputs are catenated into a flattened array of values that constitute the input layer to the model.

For all of our experiments, the model uses two hidden dimensions of size 128 and 32 with ReLU activations for each pathway in the neural network. During model training, we employ simultaneous quantile regression (Tagsovska and Lopez-Paz), where a randomly sampled quantile is concatenated to the final hidden layer. To structurally constrain the monotonicity of the quantile predictions and prohibit crossing quantiles, we apply a monotonic constraint to the quantile inputs to the final layer (Runje and Shankaranarayana). A linear layer constitutes the output that is used to minimize the pinball loss.

The model uses the Adam optimizer (Kingma and Ba) with a learning rate of 0.001 and a batch size of 1,024. A dropout rate of 0.1 is applied to all inputs of the property-level pathway (Srivastava et al.). As model averaging has been proposed to improve the accuracy of neural network regression models (Ripley), we train an ensemble of 5 identical models, each using a different seed to initialize the model weights, and perform mean averaging of the outputs.

For our submarket experiments, we exclude the "latitude", "longitude", and "present use" features. Our experiments include two types of submarket models: a "global" approach that employs the neural network model trained on the full county dataset with individual indexes estimated for each submarket, and a "local" approach that employs separate neural network models trained for each submarket. A potential advantage of the global approach is the ability of a single model to learn more robust and granular estimates for regions with sparse records by leveraging correlations learned from other related regions.

The model is implemented in Python using the PyTorch framework (Paszke).

«HERE on 10/31»

Imputation (IMP) Implementations Imputation approaches use a trained model or other process to make repeated home value estimates at a given cadence and then convert those estimates into an index. The Zillow Home Value Index (ZHVI) (link) is one such example whereby at the end of each month an estimate of home value is extracted for every home in the national and then those time series and turned into indexes and aggregregated. Imputation approaches benefit from the fact that they can be applied to any sample of homes, not just those that have sold. In other words, while the underlying models that produce the predictions are still (usually) driven by price changes in sold homes, the imputation-based indexes can be expanded to represent the entire stock of housing.

- OLS
- RF

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).

Results

We begin by using the full period of data to create a 10-year home price index using the seven different methods. The full series is shown in Figure 2, with a zoomed in version for the 2021 through 2023 timeline shown in Figure 3. A few high level trends are obvious from a visual inspection of the candidate indexes.

First, the TME and IMP indexes generally move in tandem with within each method type, but, from 2020 onwards do exhibit differences from one another, with IMP indexes suggesting lower rates of appreciation in general post COVID. The differences are not surprising here as the IMP methods measure a perfectly consistent set of homes over the time period in question, while the TME approaches are subject to changes in sample over time, even though they attempt to control for known home attributes in the model.

Next, the two AGG indexes do not perform similarly at all. The uncontrolled mean sale price index offers a consistently and significantly lower estimate of cumulate appreciation for nearly the entire period, while controlling for home size (PSF), creates an index that is mostly in line with the TME approaches. This

speaks to the importance of controlling for some aspect of home quality (in this case, size) when estimating price index.

Figure 2: Comparison of 10-Year Indexes

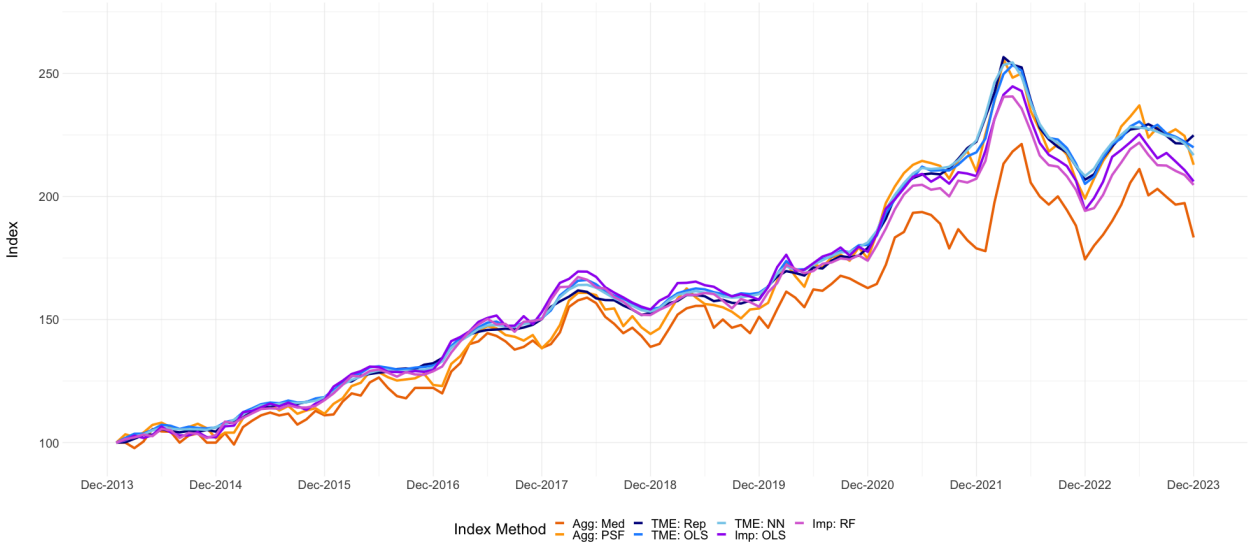
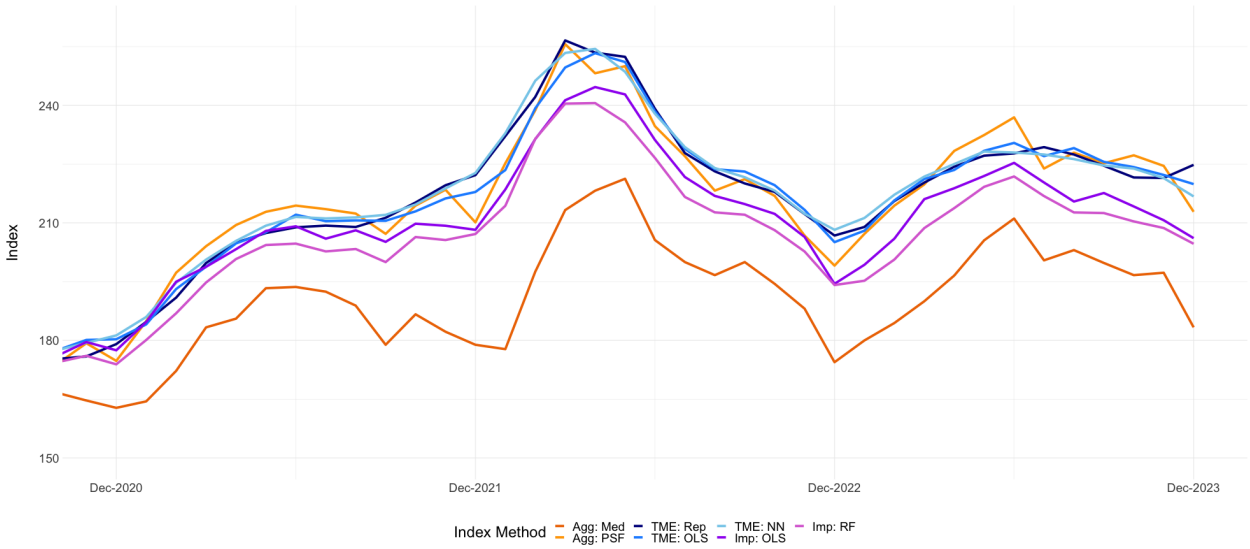


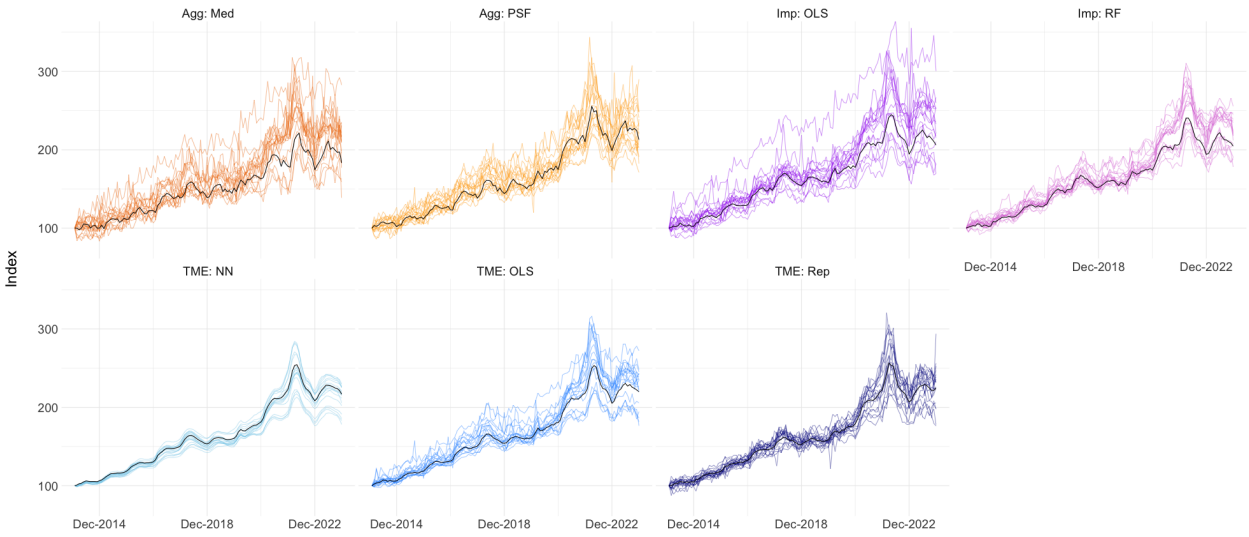
Figure 3: Comparison of 10-Year Indexes: 2020 to 2023



Next, we estimate the same ten year index for all 19 submarkets in our data. These indexes are shown in Figure 4, split by method with the county-wide index overlaid in black for each method. As expected there is a good degree of variability across the submarkets, confirming what others (CIT) have found – namely that there are localized differences in price movements.

Variability between submarkets are not equivalent across methods, though. The Median Price (AGG) and OLS (IMP) methods suggest a much high degree of difference between markets as opposed to the Random Forest, Repeast Sales and Neural Net approach.

Figure 4: Submarket Indexes by Method



Analysis

Differences in measure home price changes between the seven methods are evident from Figures 2 through 4. These changes range from fairly small at the national scale between different TME approaches, to quite large cross-method at the submarket level. However, none of these simple, visual comparisons speak to the quality of the index. In this section we evaluate each of the seven different approach on a set of metrics – Volatility, Accuracy and Revision.

Volatility Volatility is two-way street. Complete lack of volatilty (stability) would result in a perfectly flat index, which in the presence of true 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 true, and often 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.

We measure volatility with a Seasonal and Trend decomposition using Loess (STL) approach (Cleveland et al 1990). Figure 5 shows the overall trends of the seven indexes, with seasonality and noise removed. This visual offers a cleaner look at the overall movements that we see in Figure 2 and the takeaways are the same. The three TME indexes track closely with each other, as do the two IMP indexes (though with lower post COVID appreciation), while the pure Median index is a considerable outlier.

Looking at trends by submarkets (Figure 6) also tells a similar story. Some submarkets, B, M and E see very little differentiation between the methods, while other J, R and S offer marked differntly esitmates of price movements depending on which approach is taken.

Figure 5: STL Derived Trends

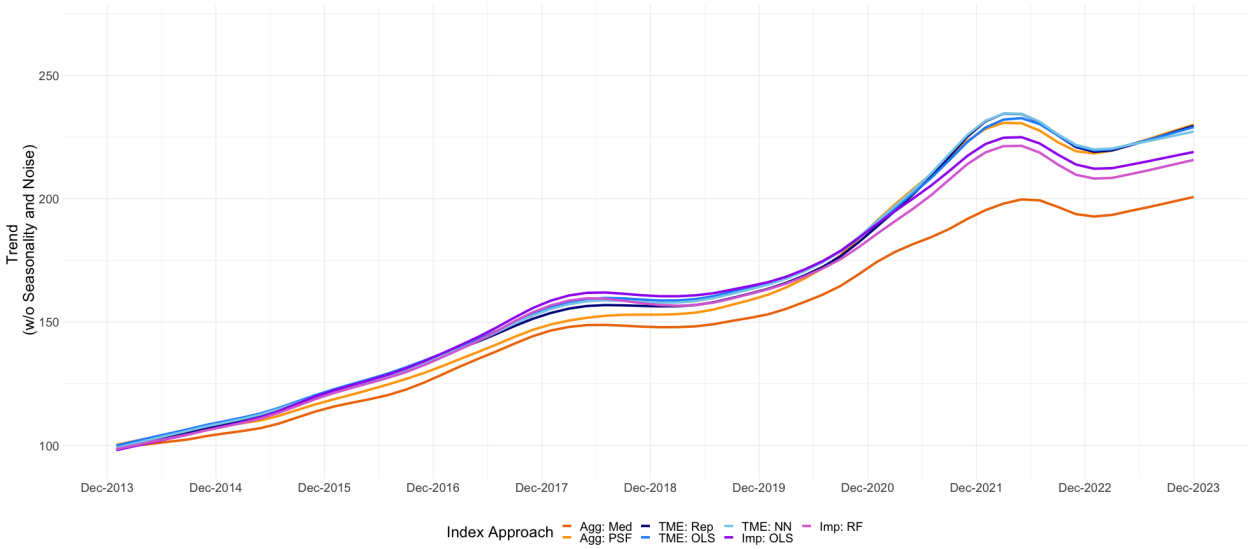
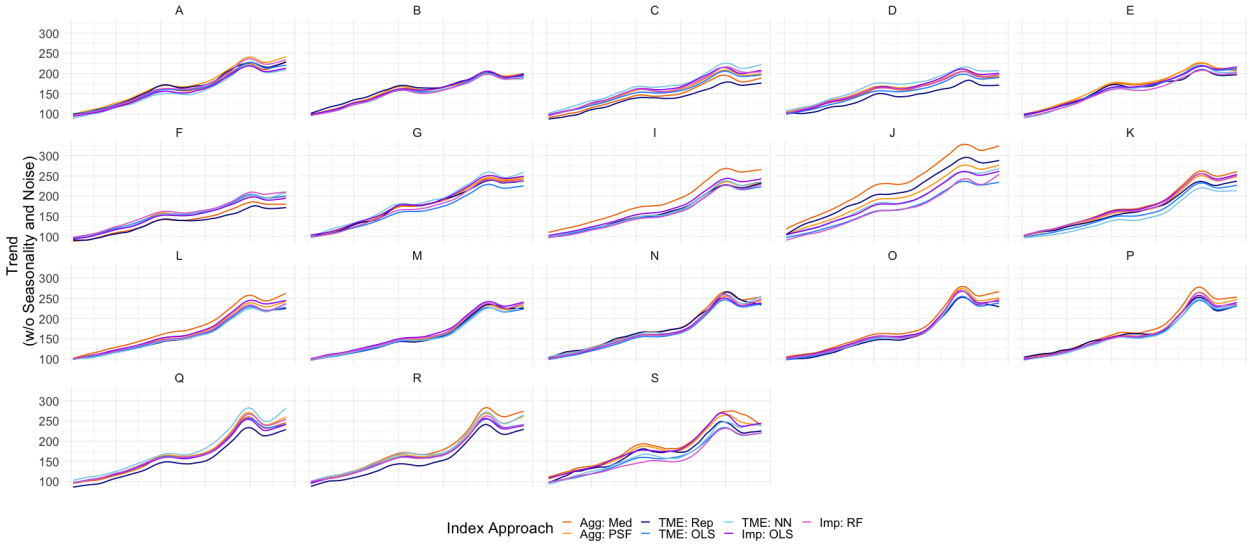


Figure 6: STL Derived Trends – Submarket Level

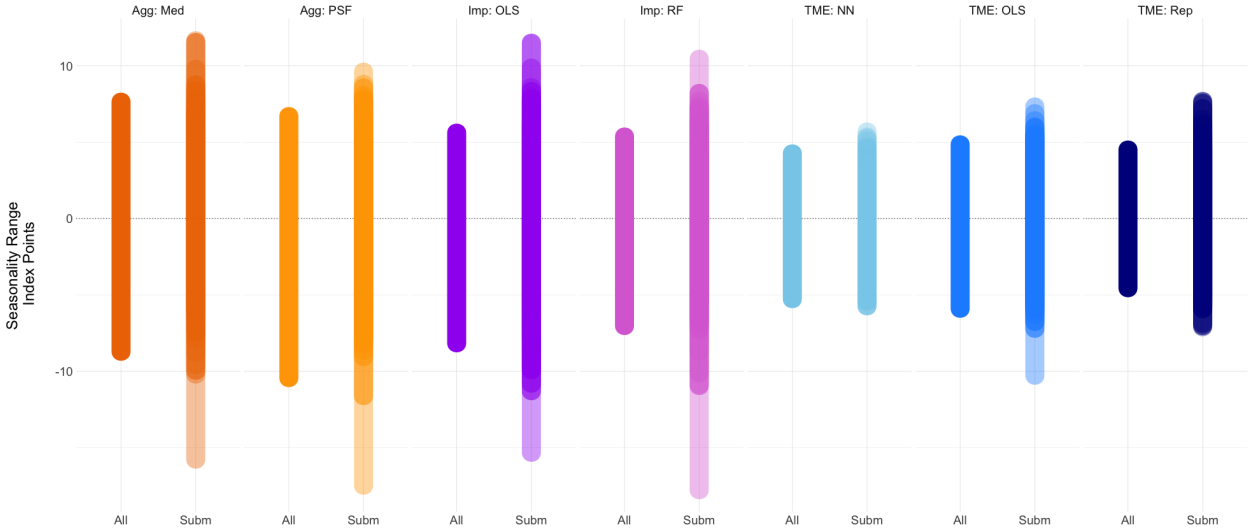


The trend view of the STL decomposition gives us a good overview of general trends, but it is the Seasonality and Noise components that we are more interested in when measuring volatility.

First we look at Seasonality. Figure 6 shows the range of seasonality for each method and by full county vs. submarket. This figure can be understood in the following way, using the Median (AGG) approach as an example. The 'All' column shows the seasonality from the STL decomposition for index for the entirety of King County. The seasonality here ranges from about -8 to +8, meaning that, about the trend line from Figure 5, this index varies, on average, 8 index points above during the strong summer season and, on average, 8 index points below. Moving to the Submarket bar immediately to its right, we have the same interpretation, but with the bar showing the distribution of seasonality over all the 19 submarkets. The lighter areas show levels that have fewer submarkets reaching that range.

We can extract a number of learnings from the Seasonality analysis. First, the TME approaches (blue) show less seasonality that the Imputation methods (purple), which, in turn, show less seasonality than the Aggregation methods. Next, the Submarkets analysis shows considerably more seasonality than the all-county indexes; however, the differences here vary widely by method. On the low end, the Neural Net indexes are only slightly more seasonal at the submarket level while for the imputation approaches we see nearly twice the seasonality for smaller regions.

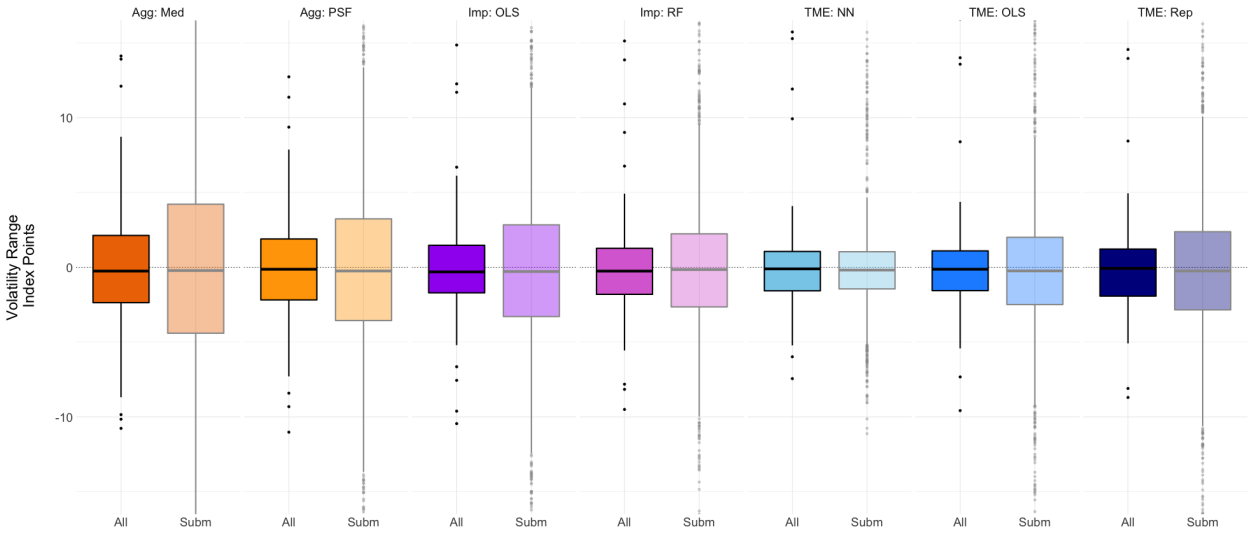
Figure 7: Seasonality



The second lens through which to evaluate volatility is that of the Noise component of the STL decomposition. Noise is the remainder or residual left over after the Trend and the Seasonal movements have been removed. As such, the Noise component represents the fine 'wiggle' around the general movement (trend) and the sinuous (seasonal) shifts.

The finding here – Figure 8 – are very similar to those of seasonality. The Aggregate methods shows the most volatility, the TME ones the least. The Neural Net approaches has the lowest overall volatility and its submarket volatily is only slightly more than the all-county variety. For the other 6 approaches, submarkets show considerably more volatility, like due to the smaller sample size.

Figure 8: STL Noise



Index Series

The visual comparisons (Figures 2-4) and the volatility analyses (Figures 5-8) can be performed on a single index created at a single point in time. Other measures of index quality – Revisions and Accuracy – require the creation of a series of indexes, each one period longer than the prior. Doing so allows for the

measurement of revision over time and the out of sample / out of time accuracy measures that are most reliable and generalizable to how an index is used in a real life situation.

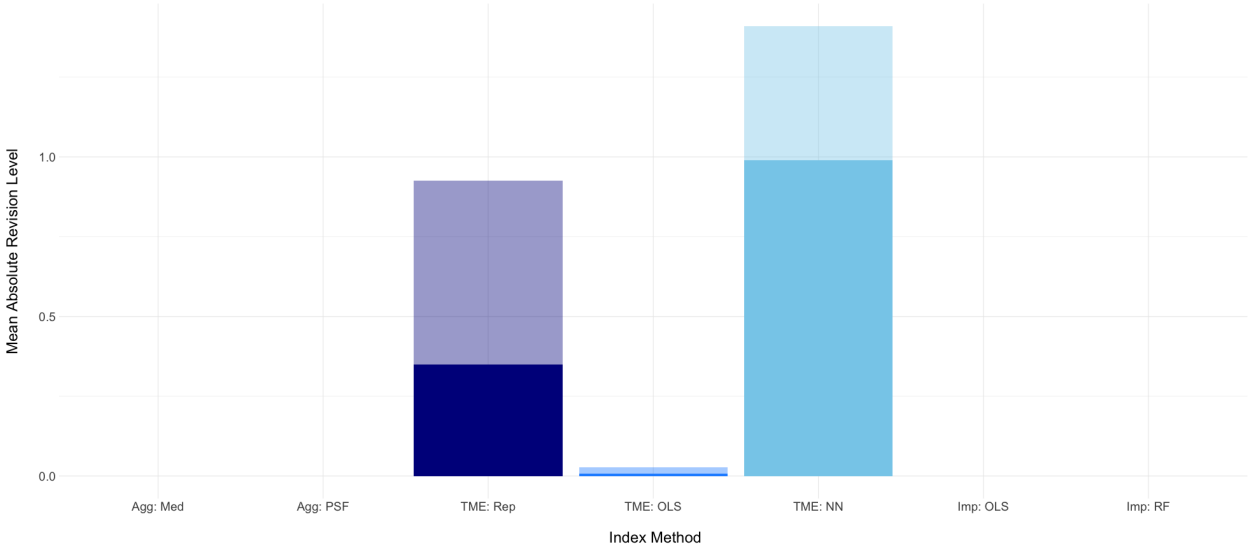
For the analyses that follow, we create index series for all of seven methods, full county and submarket. We begin with a 24 month index for each, then grow it by one month, evaluate revision and accuracy, then grow another month, re-evaluate, etc. up until the full 120 month period has been reached. As such, we then have 96 different measurements of revision and accuracy for our series of indexes.

Revision Revision to an index occur when the re-estimation of the index over time results in changes to prior estimates. In our case, this means measure how much prior periods change we add a new period of data and re-estimate the entire series again.

Figure 9 shows the revisions rates by method, with the full color representing the full county indexes and the lighter shades representing the submarket indexes (on average across the 19 submarkets). The immediate takeaway from our revision analysis is that the Aggregate and Imputation approaches do not, by design, allow for revision. This result will hold so long as their is no latency to the input data. Our study here, using public data collected historically, does not attempt to replicate the type of data latency that can be found in a real world case where data is received well after its sale date. In such a case, one would expect some measure of revision to impact all types of indexes.

Within the TME approaches, we also see considerable differences in revision rates. As the prior literature (CIT) has noted, repeat sales models are much more susceptible to revision than hedonic based measures (OLS). We see the same phenomenon in our data here as the OLS-based models show very little revision over time. Interestingly, the Neural Network models, which have the least amount of volatility, show the largest revisions by a long margin. This suggests that each month the index moves quite a bit historically, but in a smooth fashion. The county-wide index revises, on average, about 1 index point each month (in either direction).

Figure 9: Revisions



Accuracy (two kinds)

The above measures of volatility and revision talk to the shape and stability over time of the indexes. Neither; however, talks to the ability of the indexes to accurately track actual sales in the market. To do so, we employ two forms of measuring the accuracy of an index: 1) Predictive; and 2) Assistive.

Predictive accuracy measures the ability of an index to predict the second sale price of a sale-resale pair. This form of accuracy measure is common in the literature (CIT). However, it suffers from a key shortcomings in

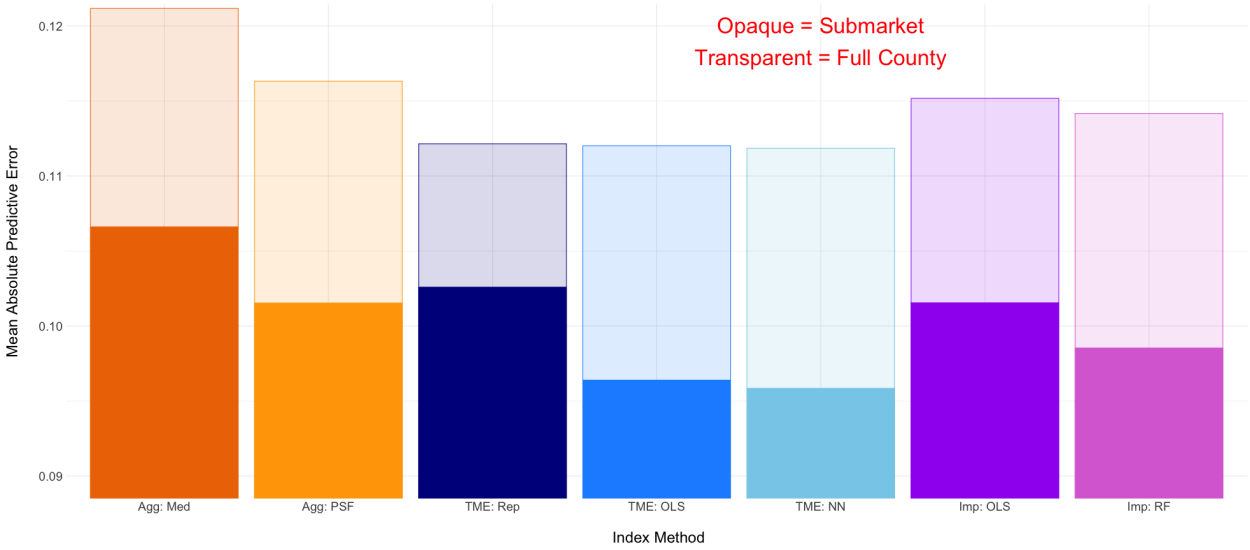
that it only measures homes that sell multiple times. Additionally, in a comparative study it also provides an advantage to a repeat sales approach which is trained specifically to minimize this error.

Assistive accuracy, instead, looks to measure the ability of an index to time adjust sale with the goal of making improvements to a predictive automated valuation model (AVM). While this may sounds somewhat unintuitive as a way to measure index quality, this is a common use case in the AVM industry and avoid the shortcoming of being biased toward any given methodology. Additionally, it allows for evaluation on all home sales, also avoiding sampling issues.

Figure 10 shows the predictive accuracy by method comparing both Full County (transparent) and Submarket (opaque). First off, it is clear from this analysis that the submarket specific indexes are more accurate than the full county ones, usually by 1 to 1.5% error points. This speaks to the local differences (as also shown by Figure 4) of home price movements and the importance of estimates indexes at the correct level. Some methods, however, offer much more gain from local indexes than others. Specifically, the Neural Net, OLS and RF approaches see big gains, while the Repeat Sales and Median Price methods offer more modest gains from local specificity.

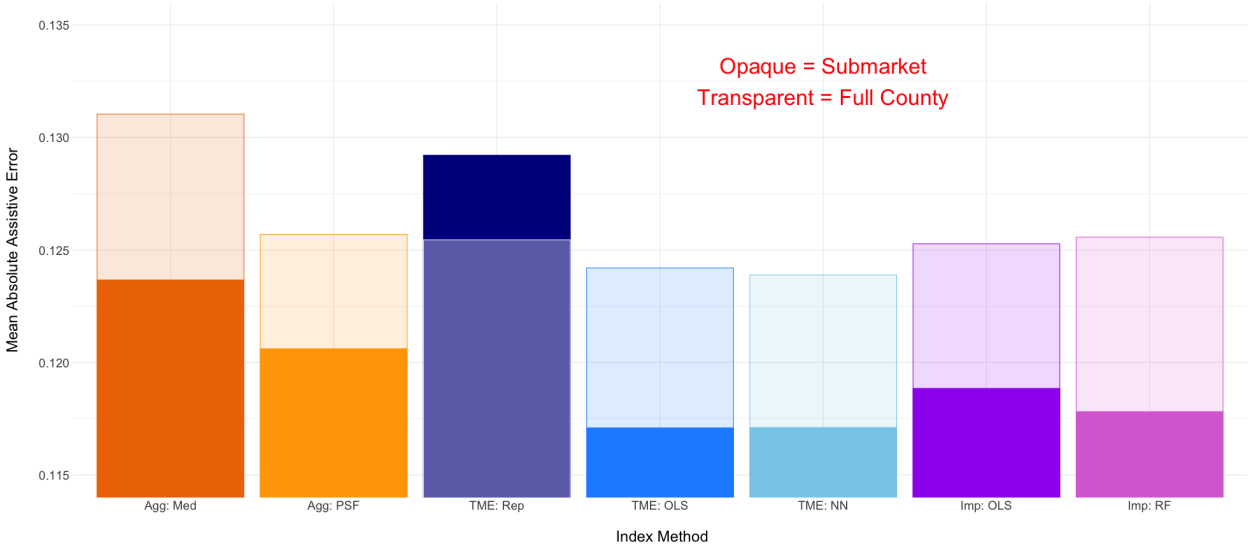
Comparing between approaches also shows some differences in accuracy. At the Full County level (transparent bars), these differences are most explained by the type of method, AGG vs TME vs IMP, with the TME all performing the best and similarly, with IMP next and AGG as the least accuracy. Moving to the submarket-specific approaches we see more differentiation. The OLS and NN models are by far the most accuracy, with RF imputation in third place. Median and Repeat Sales approaches remain the least accurate. For the Median approach this is expected, but it is a bit surprising for the repeat sales model for which the evaluation data is identical to the training data. We'll discuss possible reasons for this in the discussion section below.

Figure 10: Predictive Accuracy



Next, we do the same comparative analysis of Assistive Accuracy in Figure 11. With one exception, the same trend generally holds, that of the OLS and NN approaches being the most accuracy, Aggregate Median the least and local (submarket) indexes being more accurate than the full county. The one major exception here is the Repeat Sales approach. At the county level (transparent bar) it performs on par with PSF and the imputation approaches; however, down at the submarket level it is actually less predictive than when built county wide. This suggests that small sample sizes due to repeat sales requirements are hampering its performance for localized geographies.

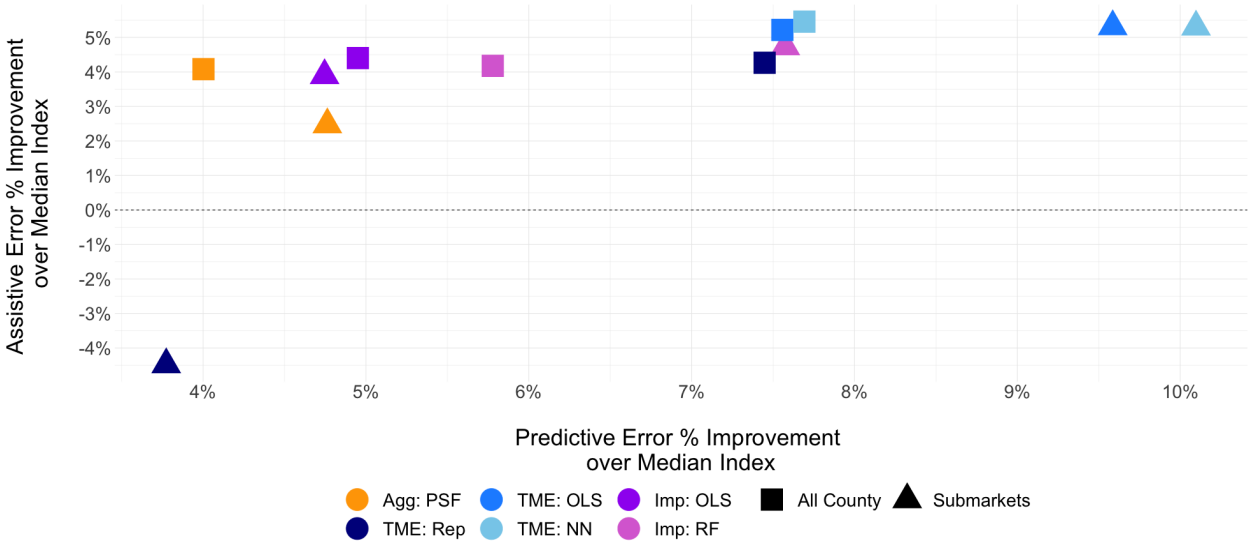
Figure 11: Assistive Accuracy



In Figure 12 we compare both types of accuracy – Predictive and Assistive – based on their relative gain over the simplest approach, the median sale price. This plot should be read as follows. On the y-axis, being above the dashed 0% line means that a given approach had greater Assistive Accuracy then the median sale approach, squares for all county, triangles for submarkets. Each geographic type is compared the same geographic type of the median sale price approach. On the x-axis, the farther to the right represents greater predictive accuracy as compared to the median sale approach, against within same geographic levels. As a result, being in the upper right are the indexes with the greater relative accuracy improvement of both types against the baseline.

The NN and OLS TME models are the clear winner in accuracy, putting up solid gains over the benchmark in both assistive and predictive acccuracy, at both the county and submarket levels. The RF imputation approach performs next best. The repeat sales approach has a very split performance. It is strong at the county level, but very weak at the submarket level.

Figure 12: Accuracy Comparison



Discussion

In this paper, we present two new methods for creating home price indexes – Random Forest Imputation and Neural Network Extraction. We also present two new measures for evaluating home price indexes: 1) Assistive Accuracy; and 2) Volatility. We add these two methods to the more common measures of Predictive 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 performance when we move to smaller submarkets.

Choosing which home price index to use for a given use case must also consider data availability and model complexity. In the first two columns of Table 1, we rate these for each of our seven methods.

For data needs: * Low = Only sale price and date needed * Moderate = Sale Price, Date and limited home facts * High = Sale Prie, Date and many home facts

For Model Complexity: * Low: Could be done in a spreadsheet * Moderate: Based on linear regression * High: Complex machine learning method

Users must consider the data available and the infrastructure and expertis before choosing an index.

We also summarize the relative performance of our results in Table 1.

Model	Data Needs	Model Complexity	Predictive Accuracy	Assistive Accuracy	Revision	Volatility	Locality
(AGG) Median	Low	Low	Worst	Worst	None	Very High	Improves
(AGG) Median PSF	Moderate	Low	Average	Average	None	Very High	Improves
(IMP) OLS	High	Moderate	Below Average	Average	None	High	Improves
(IMP) Random Forest	High	High	Average	Average	None	Average	Improves Greatly
(TME) Repeat Sales	Moderate	Moderate	Average	High	Almost None	Average	Deteriorates
(TME) Linear Model	High	Moderate	Best	Best	Almost None	Low	Improves Greatly
(TME) Neural Net	High	High	Best	Best	Very High	Very Low	Improves Greatly

The median sale price approach is the easiest to generate but performs the worst in terms of accuracy and volatility. By simply conditioning prices on home size – the PSF approach – accuracy is greatly improved, but volatility remains high.

Moving to either of the imputation approaches will reduce volatility but won't improve accuracy much over a price per square foot method, both at the cost great data needs and model complexity. Finally, the three trained model extraction approches over the overall best accuracy, however at the cost of potential revision. Within this group, the repeat sales approach differs from the OLS and NN. The repeat sales approach has lower data needs, but also has lower accuracy and higher volatility. Additionally, when moving to the more local submarkets the performance of the repeat sales model gets worse not better. When comparing the NN and OLS models, the overall accuracy is more or less the same – both county wide and in the submarkets – however the NN model offers smoother indexes, though at the cost of more revision and a much more complex model algorithm.

We argue that the choice of index should be driven by available data, developer expertise and the intended use of the index. Specifically:

- Large area estimates with limited data availability may want to use a repeat sales approach, in fact this is what metro-level estimates like Case-Shiller do
- If ease of modeling is the key criteria, a price per square foot approach will get average accuracy and no revision with very little modeling effort
- Use cases looking to generalize the index to a universe that is more than or different to the set of sold homes should use an imputation approach due to its flexibility; great accuracy and lower volatilty can be had with a random forest over a linear model, but at the cost of model complexity
- If an highly accuracy and highly smooth index is the goal, then the Neural Net approach offers the best choice, though revision will be an issue
- Overall, the TME OLS models performs at or near the top in all categories at both geographic levels. Provided that a rich set of home features is available, this approach is likely to be preferred for most use cases as the model complexity and revisions are significantly lower than the Neural Net.

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.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6(1), 3–33. http://bit.ly/stl1990

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