A Multi-Criteria Evaluation of House Price Indexes

Andy Krause^ – Zillow Group Reid Johnson – Zillow Group

2024-10-30

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.

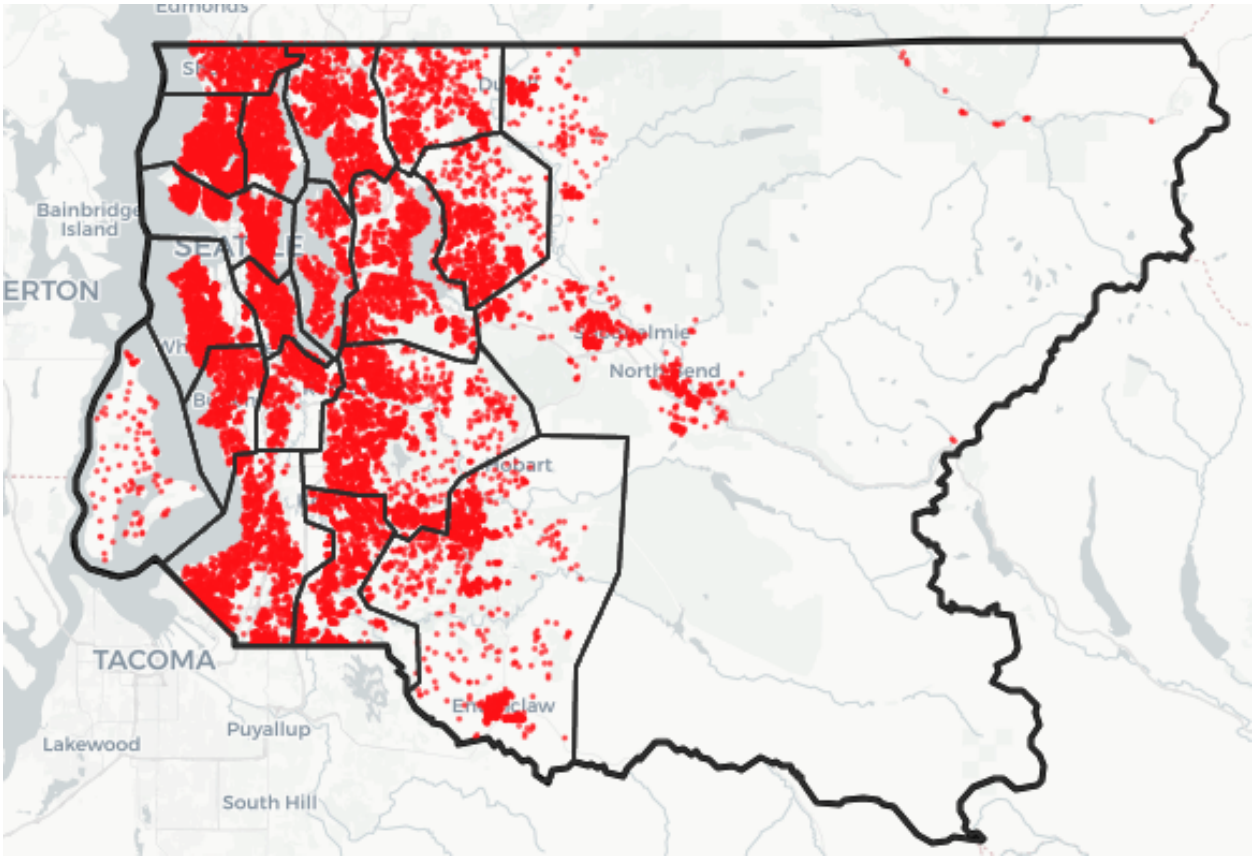
Table 1: Data Fields

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

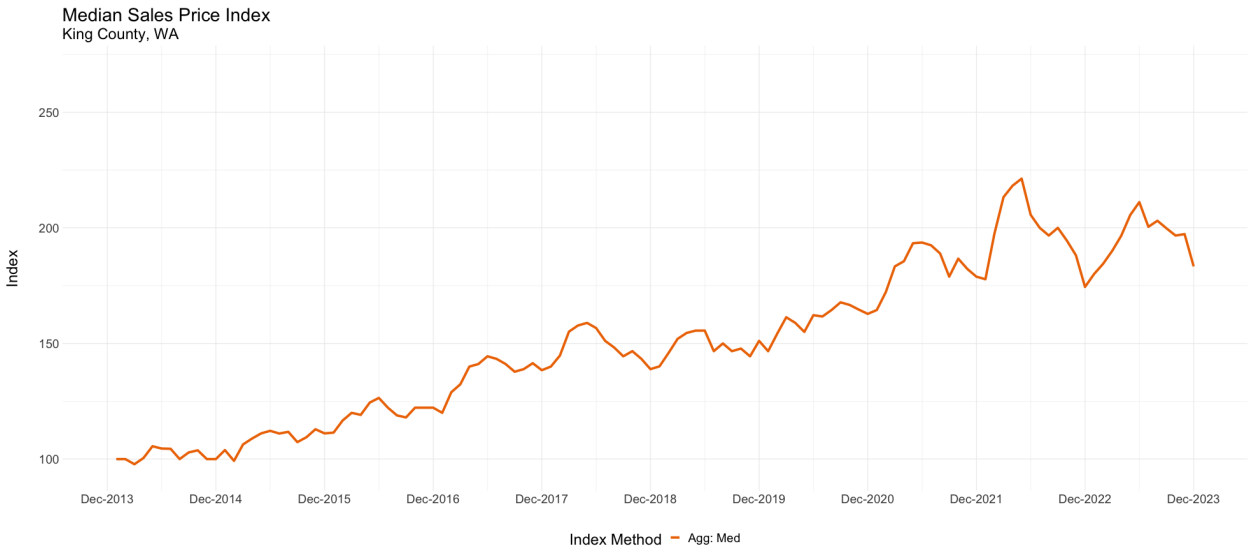


Models

- Agg (2)
- TME (3)
- Imp (4)

Results

Figure 2: Median Price (Agg) Index



Discussion

Goals of this paper:

- Refine the current typology of HPIs based on their fundamental trend extraction technique
- Test different classes, including by sample
- Add some ML options to the mix
- Create multi-criteria evaluation framework
- Present data and code (open source) for researchers

Each has an implementation method as well. * Chaining * Index creation * Statistical estimators – base, robust, QR * Algorithmic choice – LM vs NN for Math, Any AVM for Imputation * Coefficient extraction – Chaining, Single statistical model, Post Hoc explainable method for ML

Outline

Show aggregate index on median prices.

Why do we create indexes?
* Primary: We don't believe that aggregate indexes control for the sample of homes that sell. * Secondary: There is noise in the observed data, how do we better generalize (bigger problem for small samples)

What questions are we trying to answer?

- What was the median home sale price? – agg
- How have prices/values moved for any given home? – index —- But which sample of homes? —— All sold homes? – making comparable adjustments —— All homes? – tracking stock —— Some special subset of homes – tracking a portfolio —— A specific, individual home — a porfolio of 1

To the extent that samples differ between what sells and what exists, the choice of HPI method should be driven by the question you are attemping to answer

Continuum

- Pure observed sold price changes – Agg
- Observed with controls – $ normalized by some sample features (ex. PPSF)
- All homes that sold twice – RT – Also looks to control for possible unobserved features in the data
- All homes that sold – In the respective periods: Hed Ind/Chained, NN – During the entire time period: Hed Imp w/sales
- All homes – Hedonic Imputation w/Universe

Lit Review

- Taxonomy
- Comparisons / Evaluations * Are they different? (Samples – RT vs Uni…H&T) * Volatility * Robustness (Hill and Trojanek 2022) * Revisions (Silverstein 2014; VanDeMinner 2020) * Accuracy * Abs (Nagaraja) * Rel (new)

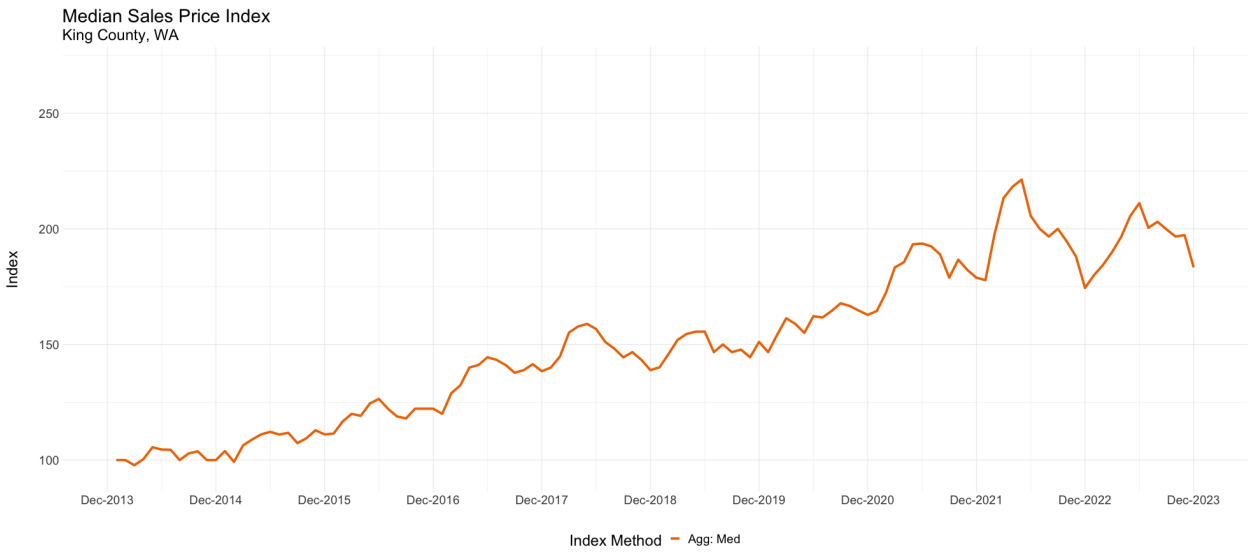
Methods Tests

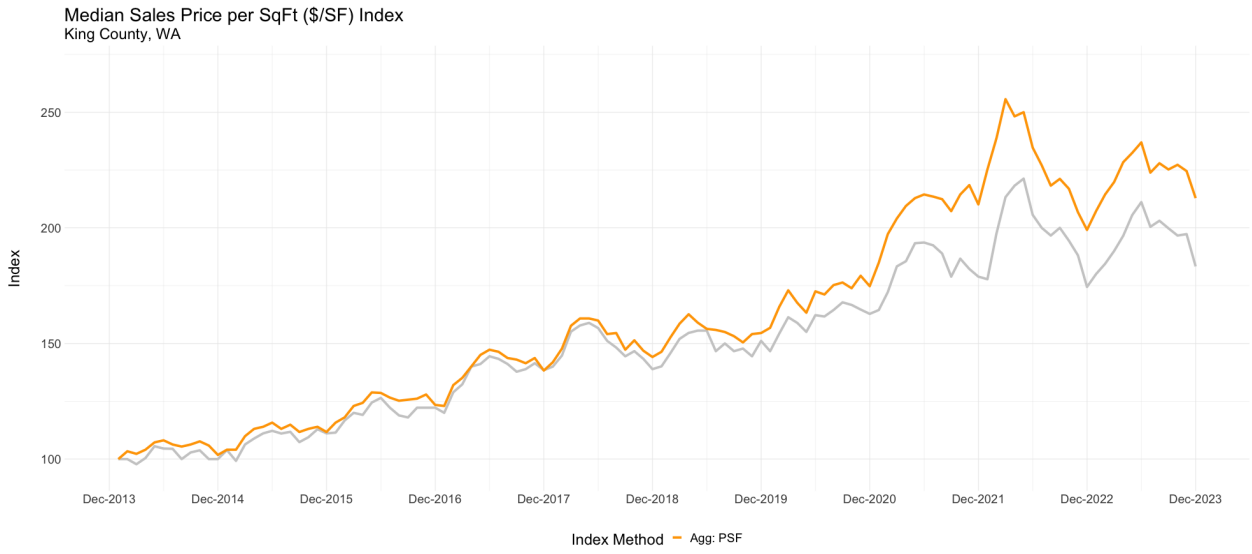
Eval Methods

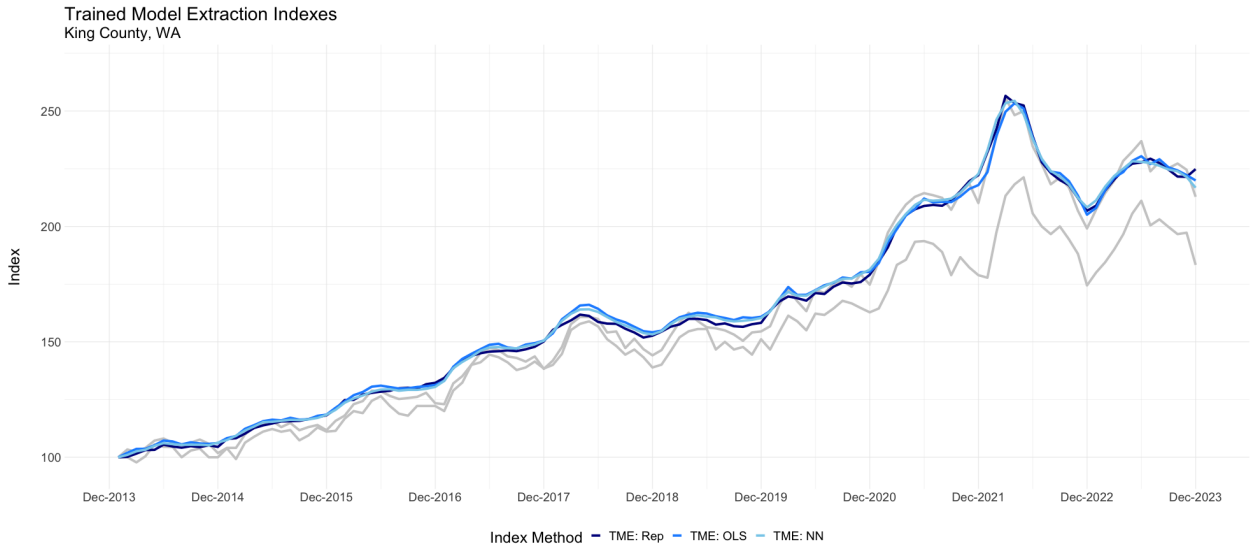
Analysis

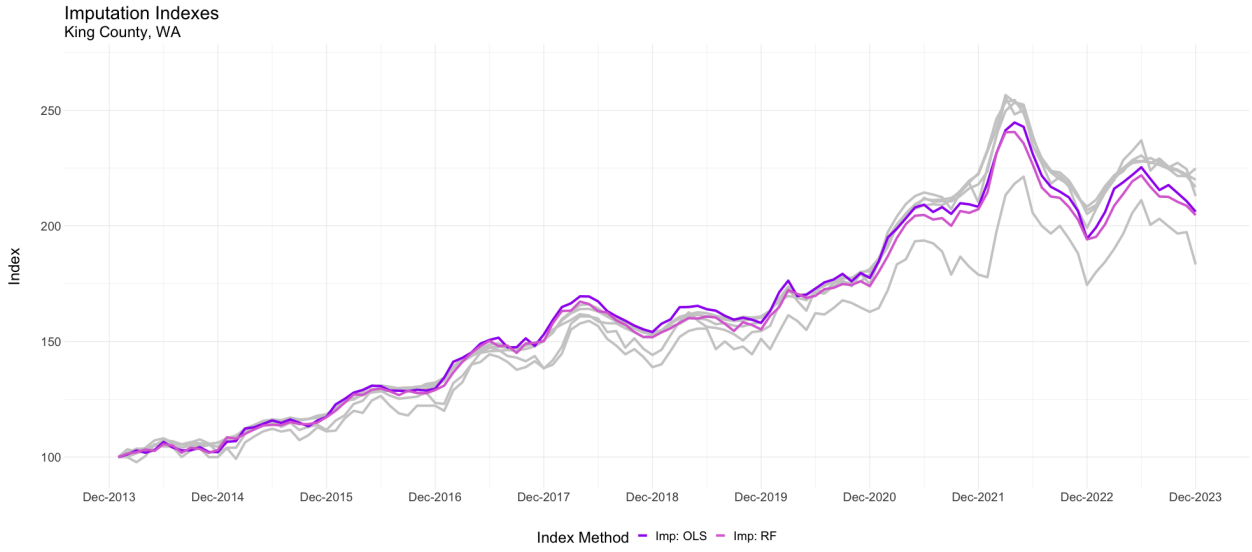
Index comparison

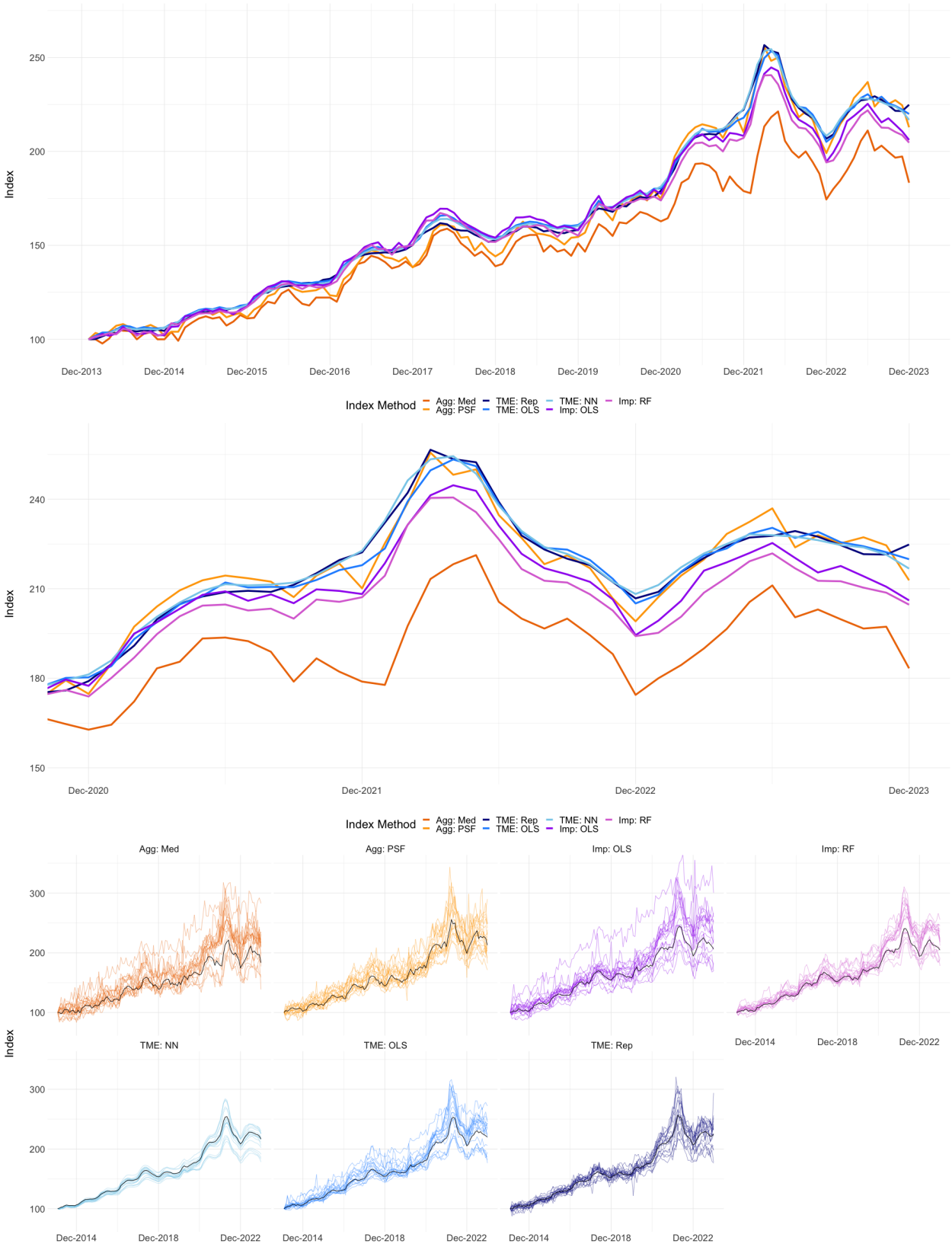
- Start with 10 year index comparison of the approaches. Talk through combination of sample and algorithm on results.



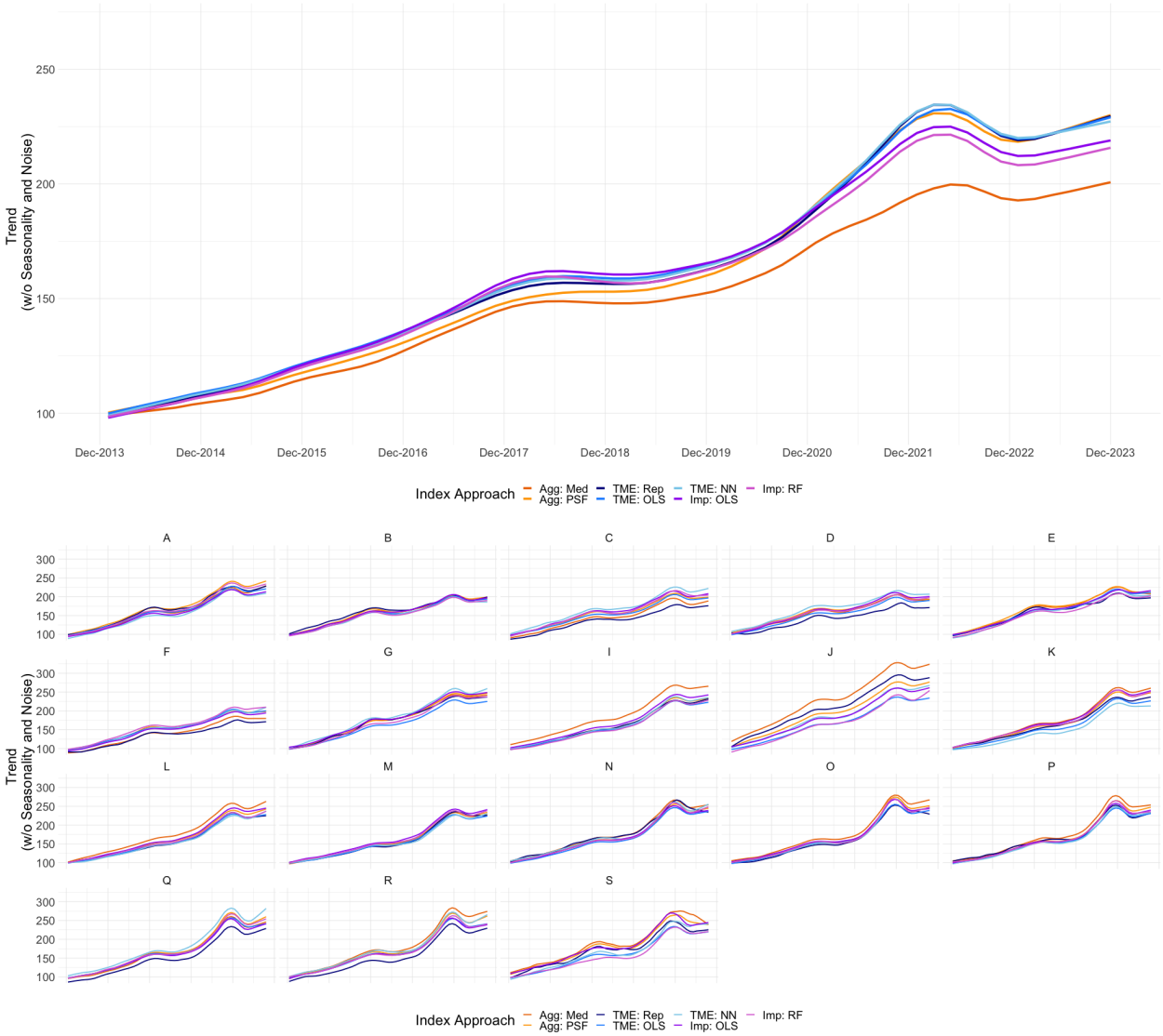

Median Sales Price per SqFt ($/SF) Index
King County, WA
Index
250
200
150
100
Dec-2013
Dec-2014
Dec-2015
Dec-2016
Dec-2017
Dec-2018
Dec-2019
Dec-2020
Dec-2021
Dec-2022
Dec-2023
Index Method
Agg: PSF

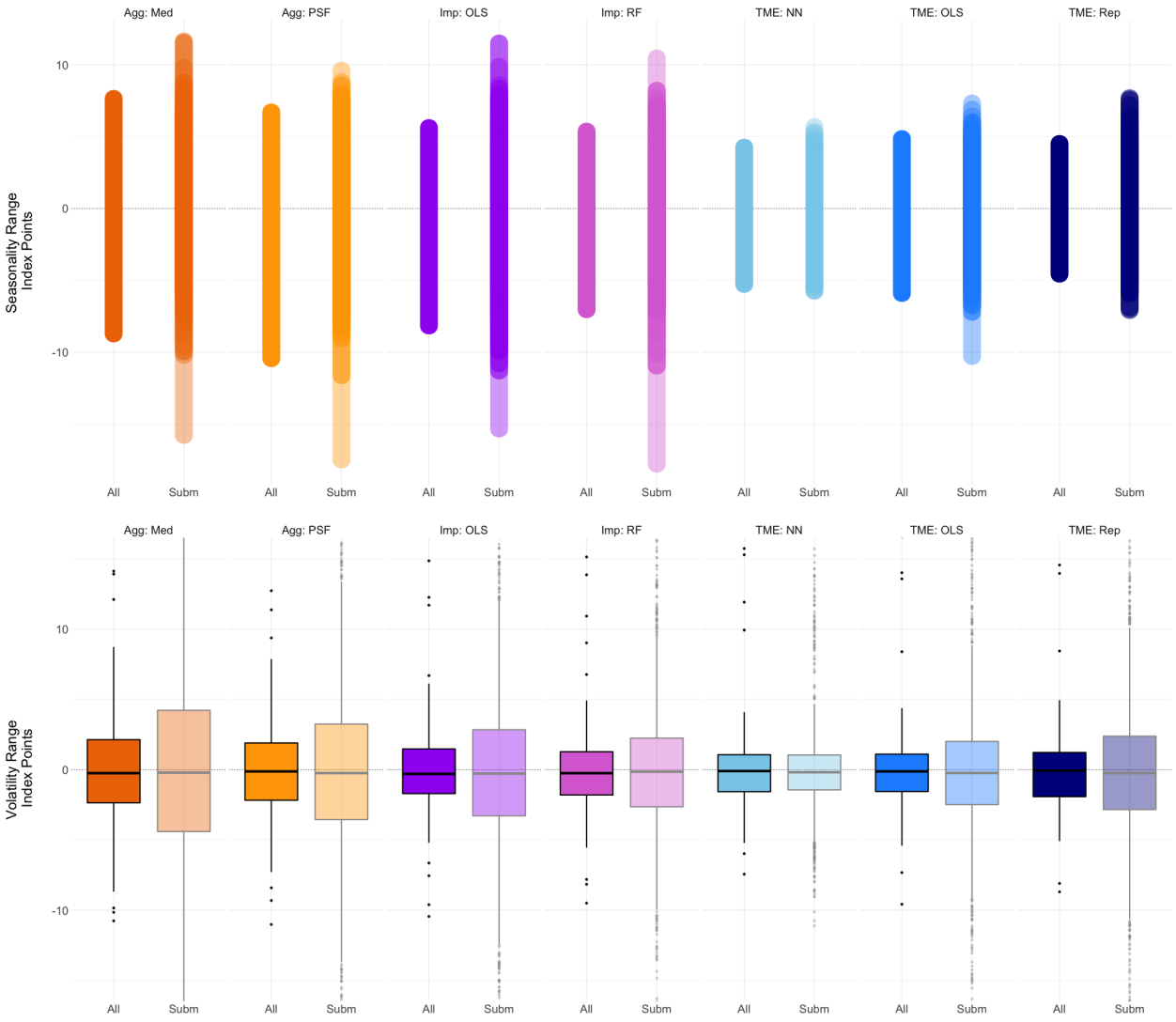

Trained Model Extraction Indexes
King County, WA
Index
250
200
150
100
Dec-2013
Dec-2014
Dec-2015
Dec-2016
Dec-2017
Dec-2018
Dec-2019
Dec-2020
Dec-2021
Dec-2022
Dec-2023
Index Method
TME: Rep
TME: OLS
TME: NN


Imputation Indexes
King County, WA
Index
250
200
150
100
Dec-2013
Dec-2014
Dec-2015
Dec-2016
Dec-2017
Dec-2018
Dec-2019
Dec-2020
Dec-2021
Dec-2022
Dec-2023
Index Method
Imp: OLS
Imp: RF


Index
250
200
150
100
Dec-2013
Dec-2014
Dec-2015
Dec-2016
Dec-2017
Dec-2018
Dec-2019
Dec-2020
Dec-2021
Dec-2022
Dec-2023
Index Method
Agg: Med
Agg: PSF
TME: Rep
TME: OLS
TME: NN
Imp: OLS
Imp: RF
Index
240
210
180
150
Dec-2020
Dec-2021
Dec-2022
Dec-2023
Index Method
Agg: Med
Agg: PSF
TME: Rep
TME: OLS
TME: NN
Imp: OLS
Imp: RF
Agg: Med
Agg: PSF
Imp: OLS
Imp: RF
TME: NN
TME: OLS
TME: Rep
Index
300
200
100
Dec-2014
Dec-2018
Dec-2022

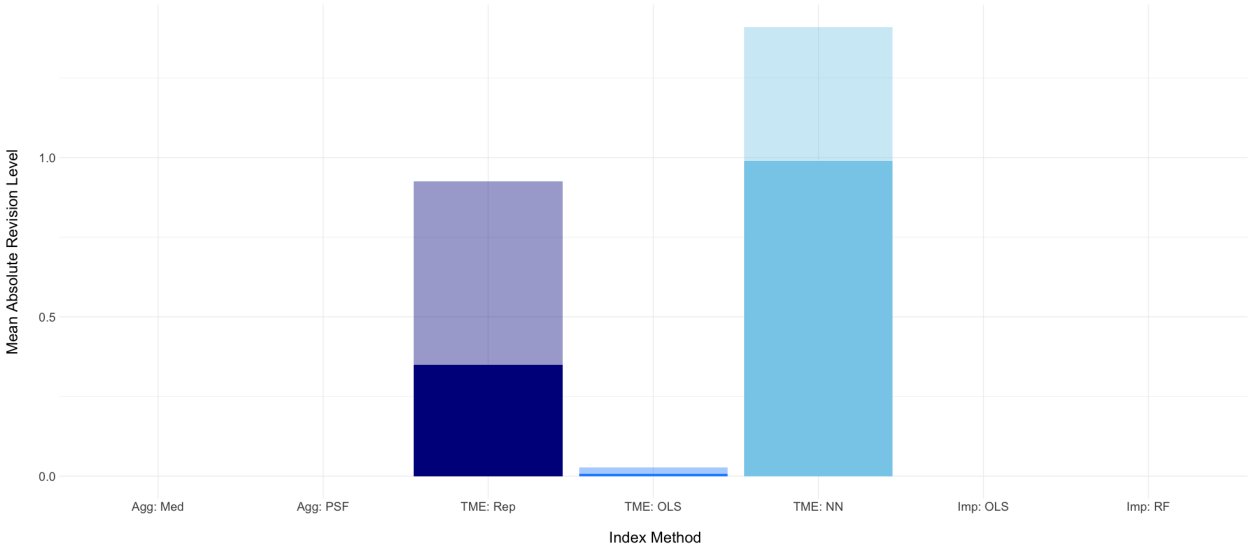
Volatility


Trend
(w/o Seasonality and Noise)
250
200
150
100
Dec-2013
Dec-2014
Dec-2015
Dec-2016
Dec-2017
Dec-2018
Dec-2019
Dec-2020
Dec-2021
Dec-2022
Dec-2023
Index Approach
Agg: Med
Agg: PSF
TME: Rep
TME: OLS
TME: NN
Imp: OLS
Imp: RF
A
B
C
D
E
F
G
I
J
K
L
M
N
O
P
Q
R
S
300
250
200
150
100

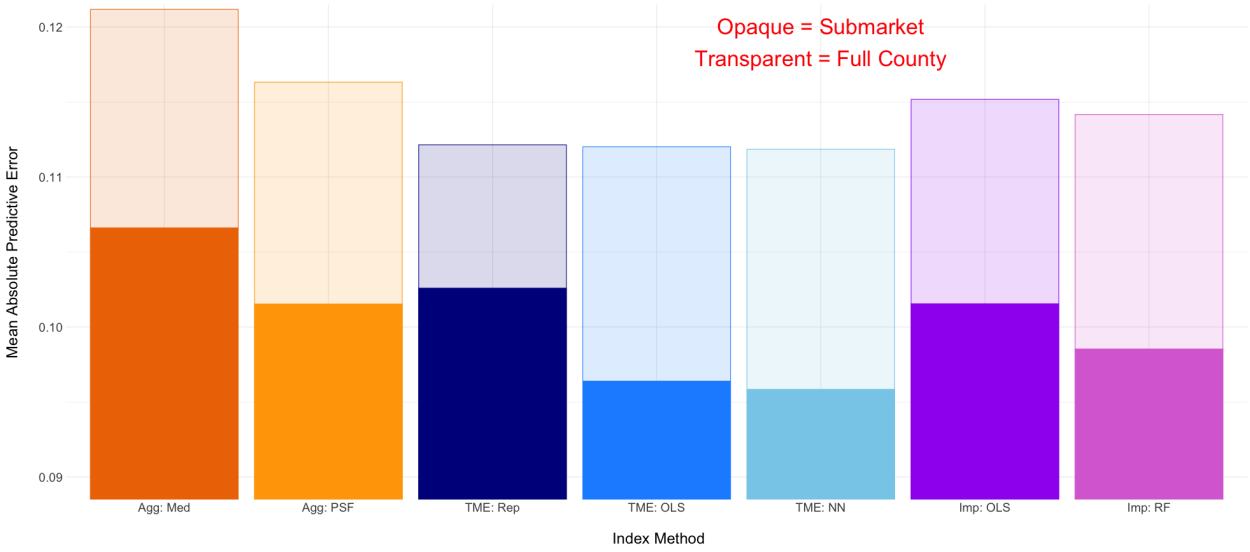

Agg: Med
Agg: PSF
Imp: OLS
Imp: RF
TME: NN
TME: OLS
TME: Rep
Seasonality Range
Index Points
Volatility Range
Index Points
10
0
-10
All
Subm

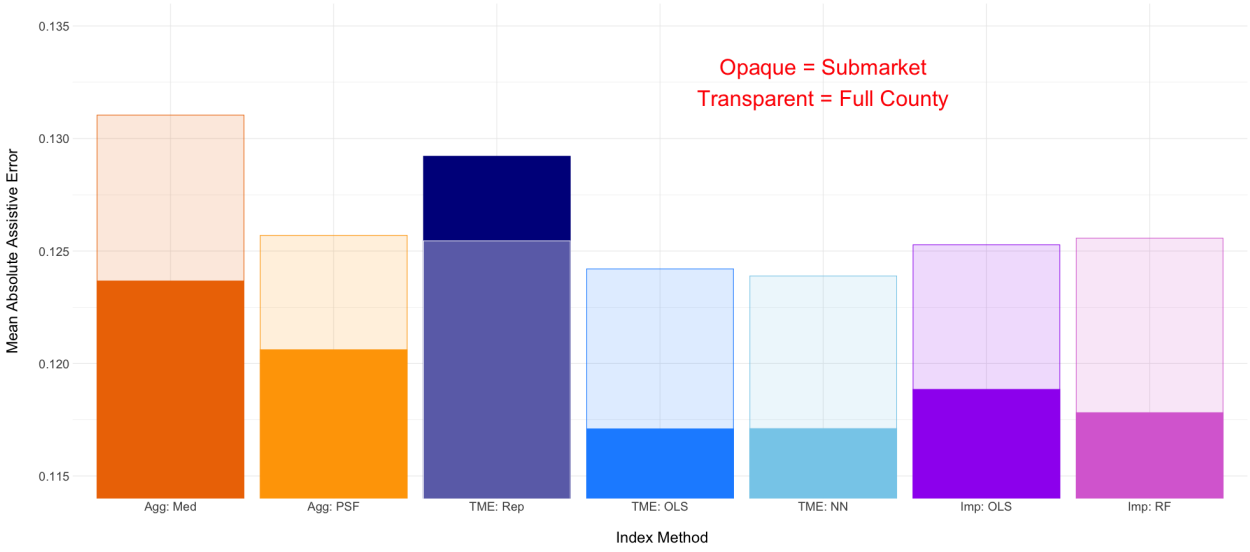
Concept of a Series

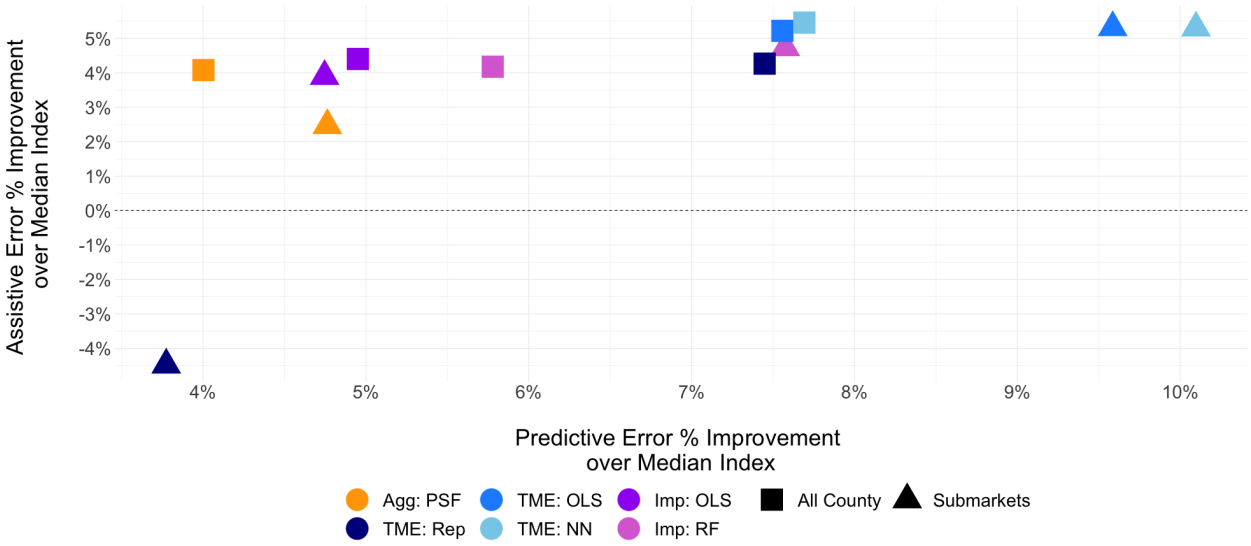
Revision



Accuracy (two kinds)







Robustness checks

Time 5 and 20 year indices

Space Submarkets and Areas

Conclusion

- Evaluate indexes based on your goal
-

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