July 2019

Home Price, Return and Volatility Indices

Brodie Gay, VP of Research

unison®
Investment Management

Table of Contents

1 Introduction and Motivation

An investor bought a home in Seattle in 2010 for $400k. How much is it worth today? How well has it performed as an investment? Investments in cash, stocks, and bonds have the virtue of liquidity. For example, shares of AAPL (the stock ticker for Apple Inc.) are fungible and trade millions of times a day within a narrow range of prices. When exchanges are open, an owner of AAPL stock knows exactly how much their shares can be sold for. In contrast, homeowners do not enjoy a comparable market mechanism: the investor's home is unique, has not transacted in over eight years, and probably does not have a line-up of bidding buyers.

A few alternatives exist to solve this pricing problem. A traditional approach is to order an appraisal. However, since this process involves dispatching a professional to the physical property, an appraisal can cost between $500 and $1000, and the resulting price estimate is prone to human bias. In contrast, automated valuation models (AVMs) estimate prices using a purely data-driven procedure. Companies such as Attom, Black Knight, Clear Capital, CoreLogic, First American, House Canary, Redfin, Trulia, and Zillow provide inexpensive AVMs with expected median absolute errors within a range of 5 to 20%. This level of precision is adequate, considering that the AVMs are either free or cost less than $5 an estimate.

Pricing is not the only metric of interest. In order to integrate a home into a modern investment portfolio, reliable expectations of return and risk are paramount. For the past 50 years, the National Association of Realtors (NAR) and the Bureau of Labor Statistics (BLS) have published median home transaction price indices to serve as a proxy for a return benchmark index. More recently, Bailey, Muth and Nourse (1963), Webb (1981), and Case and Shiller (1987) published a series of papers improving estimates of home returns by fitting a return series to the actual historical performance of individual homes with at least two sales. This class of models is known as the repeat-sales models. This repeat-sales framework has solved a major issue with inferring expected home returns from an index of transaction prices, namely that the sample of transacting homes may vary in quality over time. Presently, both the Federal Home Finance Agency (FHFA) and Standard & Poor's (S&P) rely on repeat-sales models to produce monthly home price return series. A major drawback of their methodologies is that the estimates are extremely sensitive to outliers and corrupt transaction data. To circumvent this problem, the institutions perform significant but obscure data cleaning operations, making the final estimates impossible to replicate, even with an identical data set.

The final metric we discuss is home price volatility. To understand a home's position in a modern portfolio, the dynamics of its risk need to be understood and quantified. Within the Case-Shiller repeat-sale framework, idiosyncratic volatility is assumed to be a static constant and is computed in an ad-hoc fashion during the construction of the index. The FHFA publishes these static volatility estimates online. However, Miller and Peng (2006) detect a statistically significant time-varying component of volatility. A separate but related effect is that the contributions of systematic (correlated) and idiosyncratic components of total volatility change over time. To date, very little research can be found quantifying these effects.

As stated by Karl Case and Robert Shiller (1987) in their groundbreaking paper, "65 percent of all households owned their homes, and for most of those households the net equity in their homes represents the bulk of their net worth." As of 2018, homeownership rates are roughly unchanged. An amazing amount of diligence is applied to portfolio allocations spanning equities, fixed income securities, and alternative investments. Yet, for a typical household, while home equity dwarfs the remaining financial portfolio, it is rarely incorporated in the optimization. If better benchmarks for long run home price returns and volatility could be developed, it would be possible to produce holistic financial advice for homeowners.

Institutions investing in portfolios of homes are similarly concerned with return and risk. Homes have a large idiosyncratic risk component. Thus, substantial diversification benefits arise from pooling real estate investments. However, when leverage is applied to one of the largest and least liquid asset classes in the world, underestimation of correlation–and therefore systematic risk–has the propensity to fracture a global financial network.

The objective of this white paper is to propose a model and an associated estimation procedure that produce better benchmark indices for each of residential real estate **prices**, **returns**, and **volatility**. **Chapter 2** provides a detailed summary of key results. In **Chapter 3**, we investigate the transaction price series. We use this simple setting to introduce the concept of robust and fragile estimates, taking note of the benefits of using a median (robust) price index versus a mean (fragile) price index. In **Chapter 4**, a repeat-sales home price return model is described and a method for computing a mean home price return index similar to the Case-Shiller type index is described. Two key innovations are presented: a sub-sample aggregation technique that provides a significant improvement in computing speed, and a robust alternative to the mean home price return index, namely the median home price return index. **Chapter 5** focuses on the generation of robust estimates for a home price volatility index. The time-varying components of volatility are investigated, and the total volatility of an individual home is decomposed into idiosyncratic and systematic (correlated) components.

In the spirit of practicality, each of the above-mentioned chapters begins with an example showcasing the model's handle on the Seattle home performance problem. Thereafter, that specific example is nested within a more general model of home price dynamics. Ultimately, the best procedure for estimating model parameters is selected on the basis of the following criteria:

- Accuracy: Index estimates converge to their true values, providing accurate expectations for homeowners.
- Robustness: Estimates do not break down in the presence of outliers and corrupt data points.
- Transparency: Given an identical data set, results are perfectly reproducible using the methods described herein. Data-cleaning and opaque black box transformations are completely avoided.
- Speed: The solution algorithms are fast even when applied to hundreds of millions of home transactions spanning many time periods.

To conclude, **Chapter 6** presents variations of the previously mentioned models. Each of the variations targets a specific applied problem associated with benchmarking home performance.

2 Summary of Key Results

This chapter provides an overview of the results obtained in the following three chapters, which address–in turn–the construction of a price index, a return index, and a volatility index. A summary of the additional tools, with examples, is also provided.

Chapter 3 begins by establishing a home price index. The distribution of transaction prices for homes in Seattle is investigated. We note that a significant number of outlier data points exist in the transaction price database. In light of this problem, estimates such as mean and standard deviation are **fragile**: they will break down when large outliers are present in the sample. A better alternative is to use their **robust** counterparts, namely the median and median absolute deviation (MAD) or interquartile range (IQR), because these metrics are much less susceptible to the influence of large outliers.

Consider the following toy example which compares the performance of fragile and robust estimates with and without an outlier (shown in bold in data set 2).

1a. Data Set (clean):

$0.8mm	$0.9mm	$1.0mm	$1.1mm	$1.2mm

1b. Summary Statistics (clean):

	Fragile		Robust	
Location	Mean	$1mm	Median	$1mm
Scale	Std. Deviation	$0.16mm	Interquartile Range	$0.20mm

2a. Data Set (corrupt):

$0.8mm	$0.9mm	$1.0mm	$1.1mm	**$12mm**

2b. Summary Statistics (corrupt):

	Fragile		Robust	
Location	Mean	**$1.76mm**	Median	$1mm
Scale	Std. Deviation	**$1.81mm**	Interquartile Range	$0.20mm

Note that by replacing the $1.2mm data point with a large $12mm outlier both the mean and standard deviation estimates increase substantially. However, the median and interquartile range estimates are left unchanged.

When applied to home price data series, the median price index, shown in **Figure 1**, is simple, quick to compute, and incredibly robust to outliers.

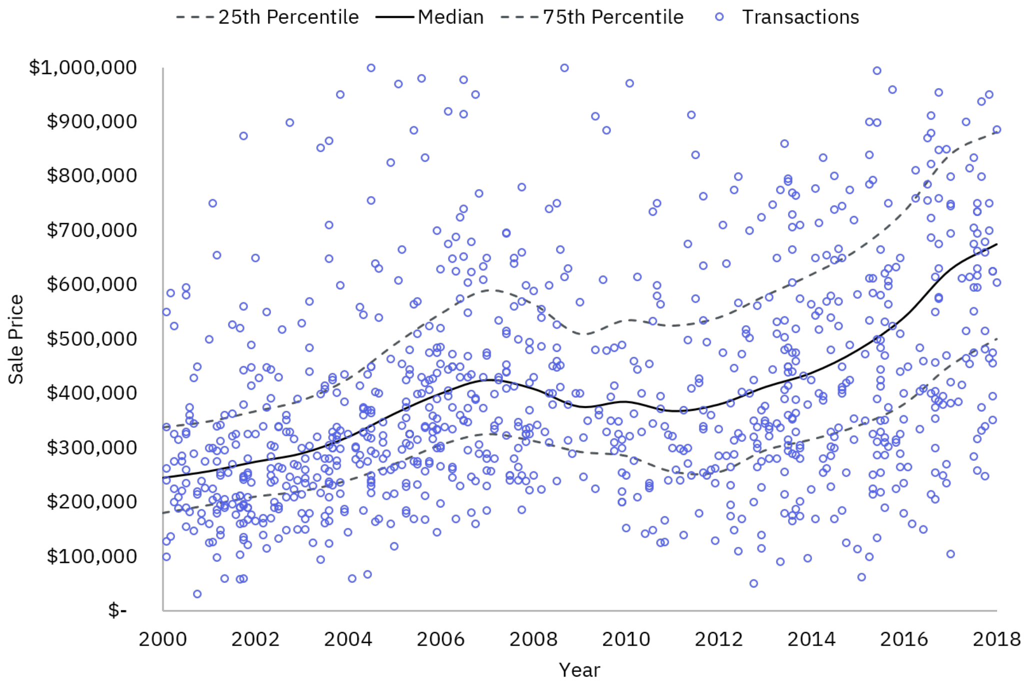

Figure 1: Home price index for home transactions in Seattle between 2000 and 2018 showing trend lines representing the median and the first and third quartiles of prices each year.

Even considering these advantages, some issues remain. For example, an increase in the median of transaction prices does not necessarily imply that homes are appreciating in value. It is possible that a large number of high quality, newly constructed properties have hit the market. Hence, we must face the first major challenge, namely controlling for quality in calculating returns.

This leads us to **Chapter 4**, where we focus on a better approach to estimating returns than the transaction price index. Our strategy is to focus on homes which have sold at least twice, thereby producing an observable investment return. Since each return is associated with a specific house, we have done a better job of controlling for shifts in quality. As an example, repeat-sales returns for the slice of homes which were purchased in 2010 and sold thereafter are shown in **Figure 2**.

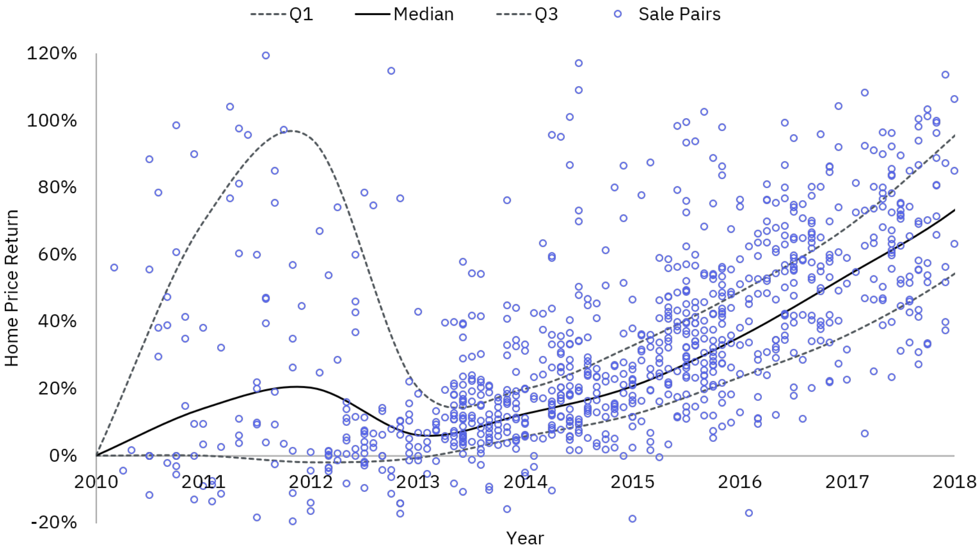

Figure 2: Total returns of homes purchased in 2010 and sold in subsequent years in Seattle are shown in a scatter plot. Trend lines representing the median as well as the first and third quartiles are shown.

It isn't sufficient to generate a return index for a home purchased in 2010 and sold thereafter. We require an index which best explains all of the returns in our Seattle data set, including all combinations of purchase and sale dates. Noting the large number of outliers and respecting a reluctance to manipulate the data set, we apply a quantile regression to obtain a robust return index estimate. This intuitive index, shown in **Figure 3**, represents the median performance of homes in Seattle.

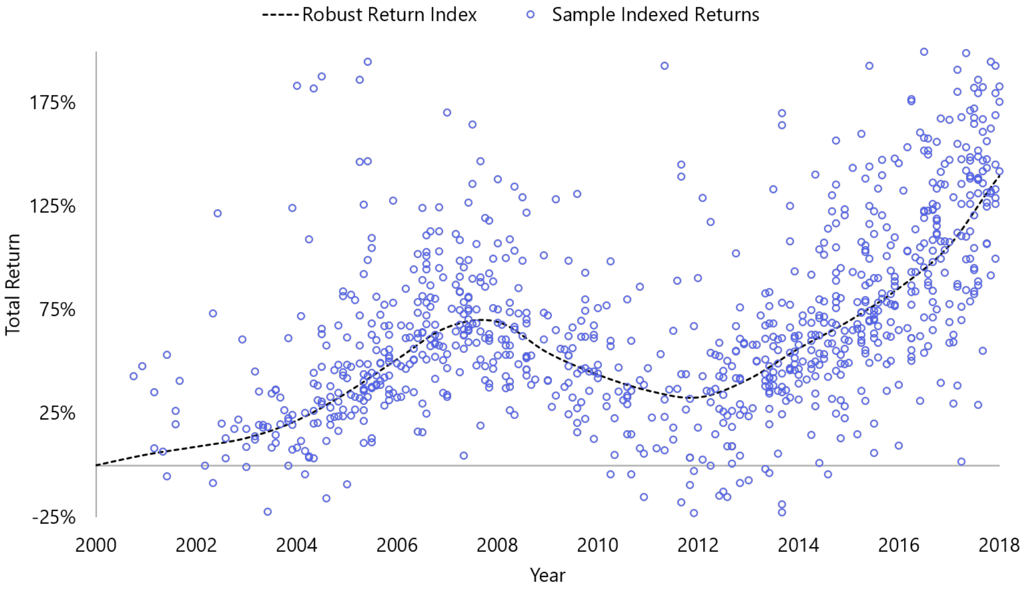

Figure 3: The median return index represents the median return performance of homes in Seattle. This method produces accurate results without any data cleansing.

We contrast our approach to the Case-Shiller methodology, a mean return index which is very sensitive to outliers and thereby requires substantial data cleansing. We argue that the median return index is a more intuitive measure of expected performance for a typical homeowner. Without any additional information, it is equally likely that a homeowner's investment in their home outperforms rather than underperforms the median home return index. On the other hand, a mean home return index can outperform a majority of homes if there is significant positive skew in the underlying distribution of individual home price returns.

Robust methods, such as the metrics we are introducing, are typically more computationally expensive than their fragile counterparts (e.g., the Case-Shiller methodology) and often cannot handle as large a quantity of data points. To solve this problem, we present in **Chapter 4** a novel and fast algorithm that scales handily.

In **Chapter 5**, we move on to estimating a robust volatility index, quantifying the magnitude of individual home returns relative to the return index. **Figure 4** illustrates how a fragile measure of volatility tends to overestimate the magnitude of deviations of individual home returns from the index. Our proposed robust index, based on the median absolute deviation, provides a more intuitive and better calibrated estimate for the magnitude of dispersion of individual home returns in excess of the median return index.

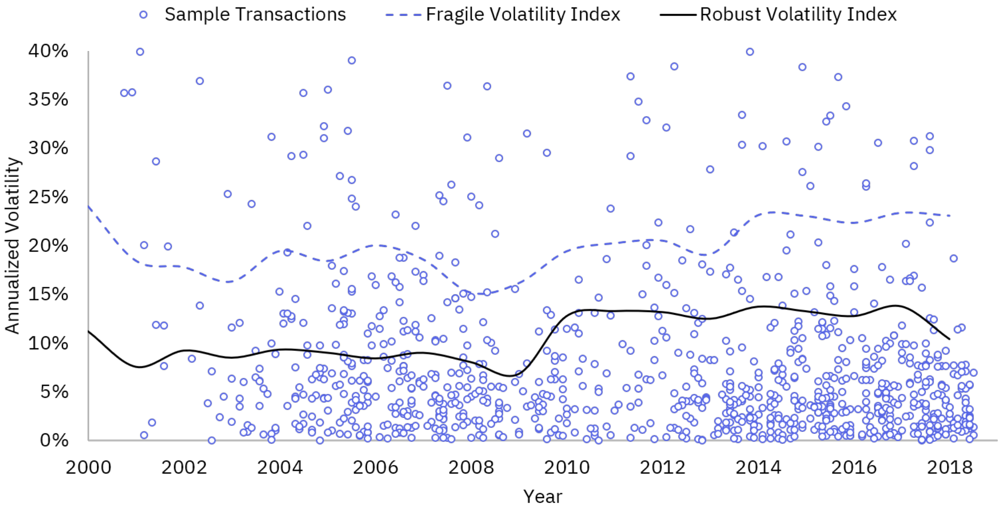

Figure 4: Robust and fragile volatility indices are shown. The annualized absolute deviations are overlaid.

Tying together our return and volatility indices, we can decompose a home's risk into idiosyncratic and systematic (correlated) components, as shown in **Figure 5**.

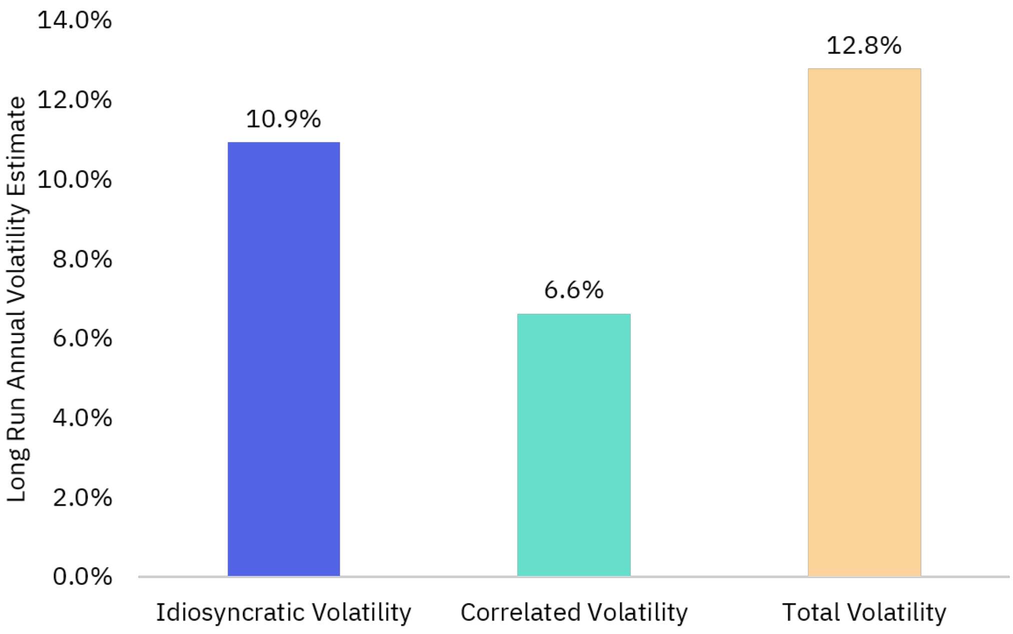

Figure 5: Long run expected values for idiosyncratic, correlated, and total annualized volatility.

An insight that follows from the decomposition shown in **Figure 5** is that institutions can diversify the risks of homes much more efficiently than individual homeowners can. **Figure 6** illustrates the total portfolio volatility as the number of homes in the portfolio grows.

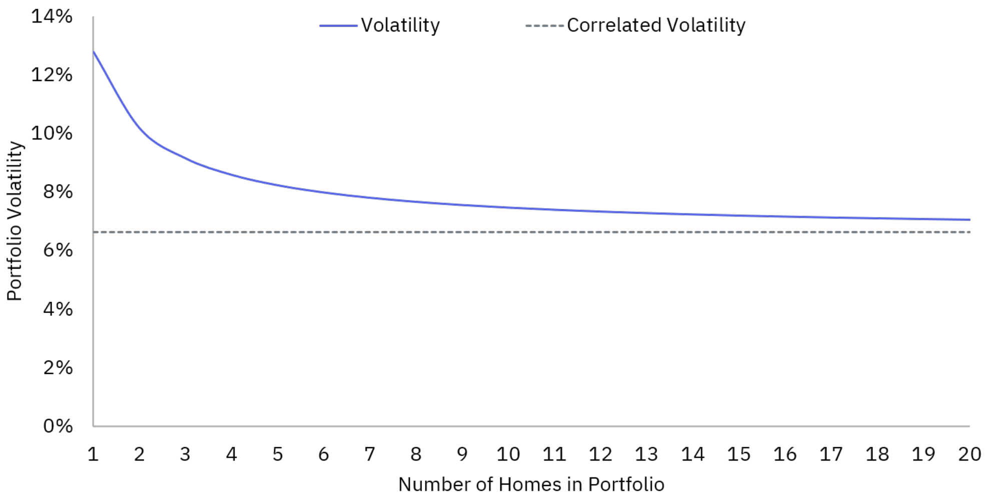

Figure 6: Total volatility as a function of the number of (equally weighted) homes in the portfolio.

We have developed fast and robust indices describing the evolution of home prices, the expectation of home price returns, and the volatility of returns for an individual home or portfolio of homes. Hence, the main objective of the paper is complete. However, before concluding, **Chapter 6** presents a few variations to the indices. Each variation addresses a specific applied problem:

2 Summary of Key Results

1. The first problem we address is that of time filtration. So far, we have generated the entire index with all transaction data between 2000 and 2018. Realistically, we will need to generate estimates each year using only information available at that time. This was not a problem when we generated a transaction price index, since future data will not affect past estimates. However, we show that this is not the case when using sale-pair data spanning multiple years. **Figure 7** illustrates the difference between a fully informed (i.e., simultaneous) series and one that generates each year's return with only information available at that time (i.e., sequential). A hybrid of the two methods, which uses a simultaneous solution up until 2010 and then a sequential solution thereafter, is also provided.

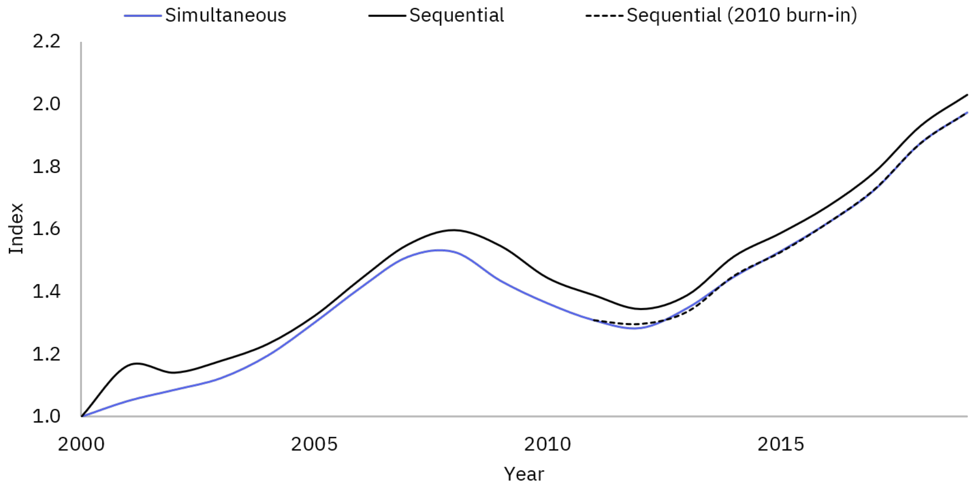

Figure 7: Integrated series for simultaneous and sequential solutions. Starting in 2010, a sequential solution with a 10-year burn-in period is provided.

2. The next problem we address is whether a practitioner should value weight or count weight the data points in our indices. The resulting differences in return behavior across price tiers are investigated. In Seattle, higher-priced homes exhibit lower long run mean returns and lower systematic (correlated) volatility. **Figure 8** illustrates the robust return indices for four price tiers in Seattle.

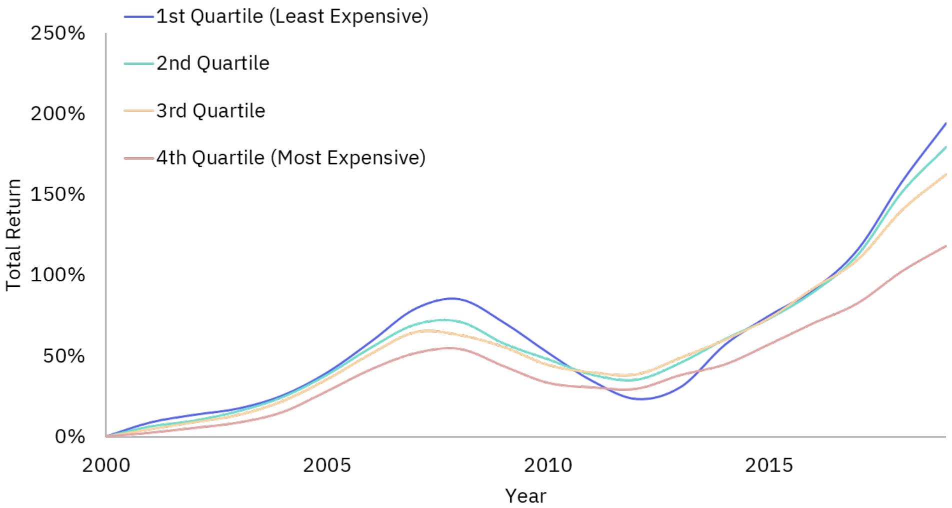

Figure 8: Robust return indices are provided for four price tiers of homes in Seattle.

We investigate the effects of value weighting in light of price tier differences. Specifically, we note that value weighting will dampen the effect of high risk, high return homes from the least expensive tier, as shown in **Figure 9**.

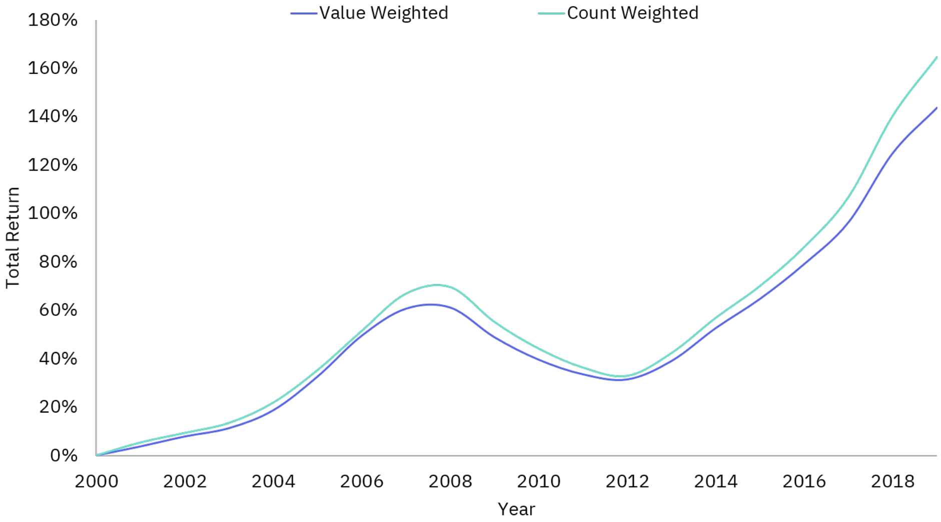

Figure 9: Value weighted and count weighted, fast robust home return indices in Seattle.

3. Finally, we note that a higher frequency of sub-sampling (i.e., using monthly rather than annual time slices) results in considerable estimation noise. We explore methods to smooth and seasonally adjust higher frequency estimates of the indices using a simple moving average, as shown in **Figure 10**.

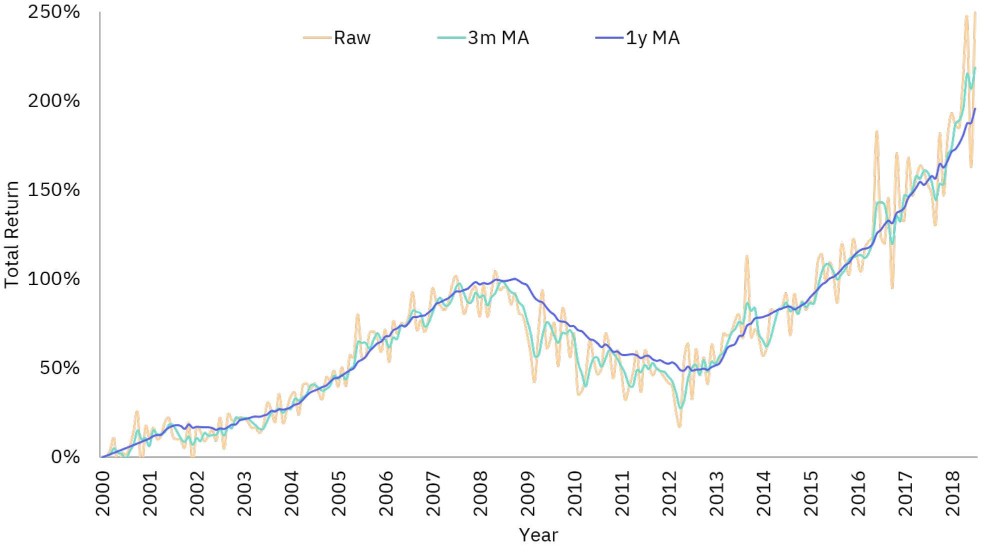

Figure 10: Smoothed (3m MA) and seasonally adjusted (12m MA) monthly return indices for Seattle homes.

3 Prices

We begin our model development by investigating home price series. **Figure 11** illustrates a series of transaction prices in Seattle. This scatter plot is an intuitive way to visualize the evolution of home prices. Referring to our stylized example, **Figure 11** reveals the following features:

1. **Trend**: The overall trend shows that transaction prices tend to increase over time, but not necessarily monotonically.
2. **Outliers**: A significant number of transactions occur far from the median.
3. **Time-varying Dispersion**: The size of the spread of prices around the median trend changes over time.

Our goal in this chapter is to produce a good model to capture these effects. However, before doing so in earnest, we can intuit a first guess for our example home's current value by examining the chart.

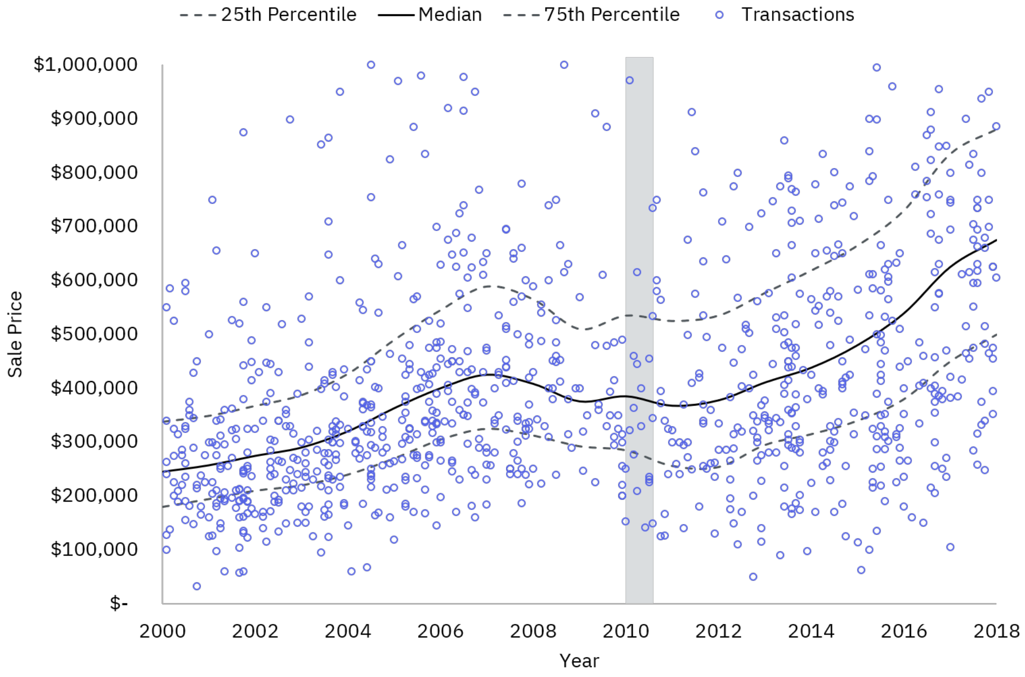

Figure 11: Blue dots represent individual transactions in Seattle; trend lines show the evolution of median and quartiles of transaction prices. The example home was purchased in 2010 (shaded in gray).

Example: Seattle Home Price

Recall that the home from our example in **Chapter 1** was purchased in 2010. Focusing on this slice of time, we can determine how its sale price compares to others sold that same year. **Figure 12** is a histogram of transactions by price in 2010 and 2018.

The home was more expensive than 55% of homes sold that year. Assuming its rank does not change over time, a good guess for the home price would be the 55th quantile of homes sold in 2018, which is $680k.

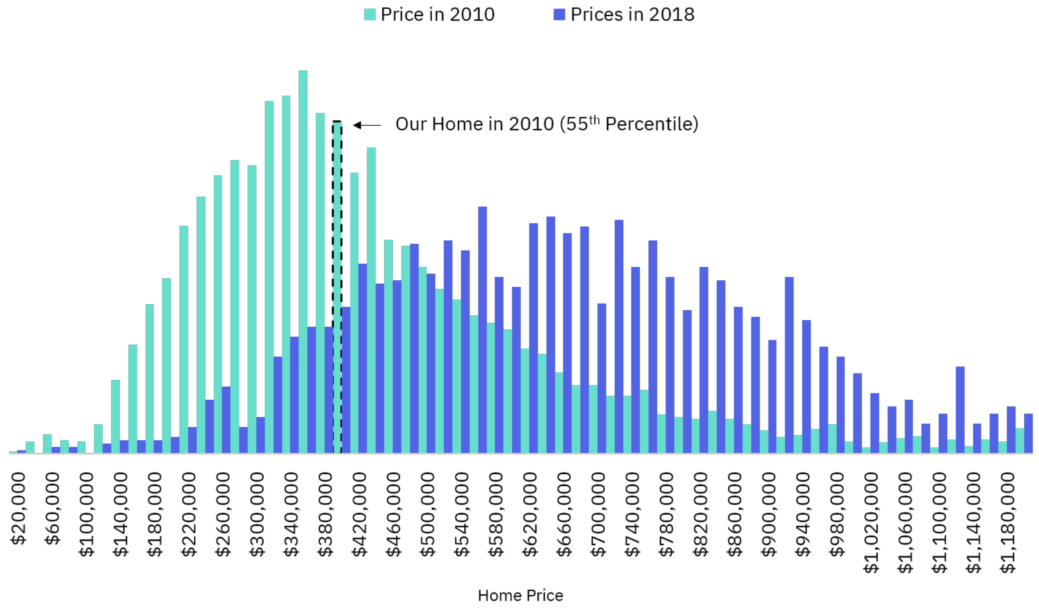

Figure 12: The light blue bar chart shows the distribution of prices of homes sold in 2010. The dark blue bar chart shows the corresponding distribution for homes sold in 2018. The outlined bar identifies the bin comprising the $400k home.

Price Index

We can generalize this simple and powerful approach for analyzing home prices as follows:

Home Price Model

$$p_{h,\tau} = \pi_\tau + \sigma_\tau \epsilon_{h,\tau} \tag{1}$$

$$\epsilon_{h,\tau} \sim \mathcal{D}_\tau(0,1) \tag{2}$$

where:

- $p_{h,\tau}$ is the price of transaction h in year τ (e.g., $p_{1,2010} = \$400k$).
- π_τ is the general price level of homes in year τ, and captures shifts in the entire distribution of home prices over time.
- σ_τ is a time-varying scaling factor representing the width of the distribution of home prices around the general price level π_τ.
- $\epsilon_{h,\tau}$ is the idiosyncratic deviation of a home from the general price level π_τ scaled by a factor σ_τ.
- $\epsilon_{h,\tau} \sim \mathcal{D}_\tau$ means that ϵ is drawn from a distribution $\mathcal{D}_\tau$ of possible deviations from the general price level. $\mathcal{D}_\tau(0,1)$ indicates a standardized random variable (i.e., it has an expectation of 0 and has an expected squared deviation of 1).

At each time period, a particular home's price lies somewhere on the distribution $\mathcal{D}_\tau$. Since we are concerned with how prices evolve over time, the representation derived here conveniently separates two characteristics of price changes, namely:

1. shifts of the center π_τ of the distribution, and
2. expansions (or contractions) of the scaling factor σ_τ applied to the standardized distribution $\mathcal{D}_\tau$ of individual home price deviations $\epsilon_{h,\tau}$.

This representation appeals to our intuition and lends structure to the next phase of our modelling exercise, which is to describe the effects listed earlier: **Trend, Outliers, and Time-varying Dispersion**.

Identifying the Trend

A mean price index is the most obvious method for estimating the trend.

Mean Price Index

$$\hat{\pi}_\tau = Mean(p_{h,\tau}) \tag{3}$$

$$= \frac{1}{N_\tau} \sum_{\forall h:t=\tau} p_{h,t} \tag{4}$$

$$= \underset{\pi^*_\tau}{\text{arg min}} \sum_{\forall h:t=\tau} |p_{h,t} - \pi^*_\tau|^2 \tag{5}$$

where:

- $\hat{\pi}_\tau$ is the sample mean price level in year τ.
- N_τ is the total number of transactions that occur during time τ.[a]
- $p_{h,t}$ is the transaction price of home h in year t.

[a] The *circumflex* ˆ above $\hat{\pi}_\tau$ signifies that this is an estimate, which distinguishes it from the true (but unobserved) value π_τ.

Figure 13 shows how the estimated mean price index behaves using Seattle transaction prices over the past 18 years. The drawbacks are apparent. First, a single large outlier in 2012 pulls the mean up substantially. Second, almost 70% of prices are below the mean since the distribution of prices is skewed upwards.

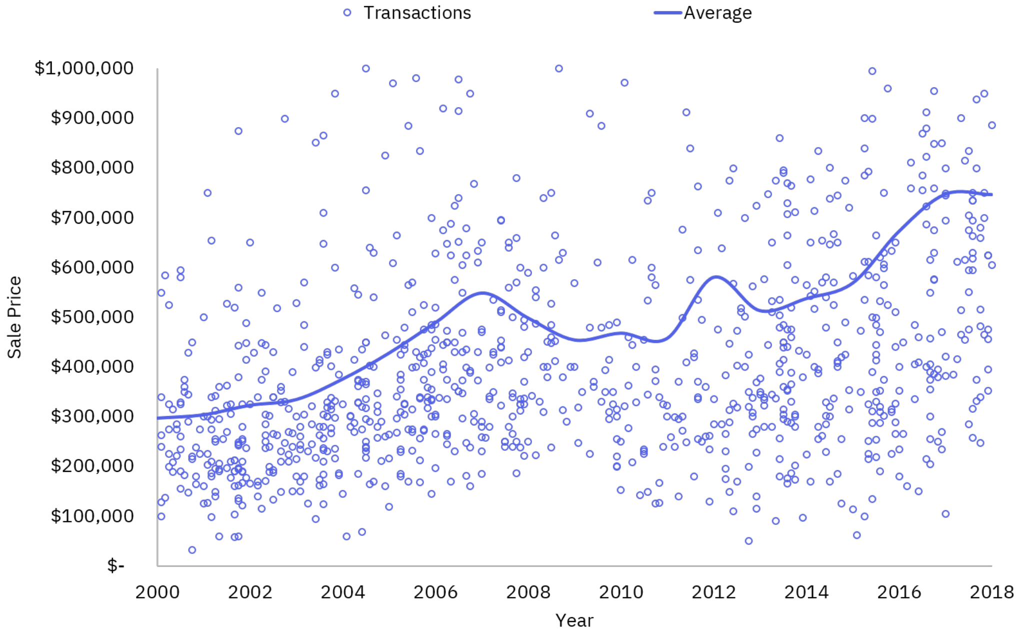

Figure 13: The mean price index for Seattle. Blue dots represent individual transactions.

Figure 14 compares a theoretical log-normal distribution with a histogram of actual transaction prices. The prevalence of outliers is apparent.

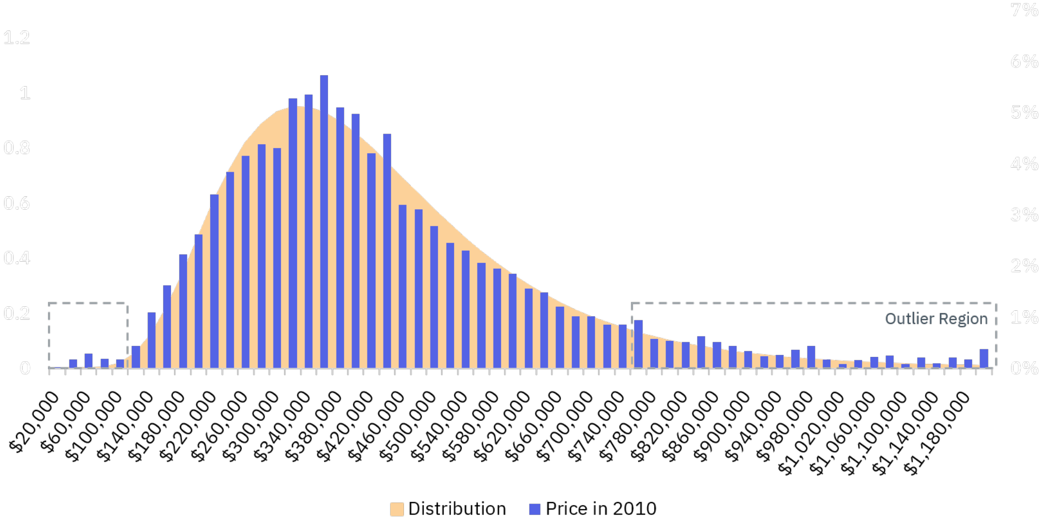

Figure 14: A fitted log-normal distribution is compared to the true distribution of transaction prices for homes sold in 2010. Many outliers are visible.

Data Cleaning

Fitting a model involves minimizing a measure of error between a model prediction and the actual underlying data. In most cases, this error is the sum of squared errors for each data point. Throughout the paper, we refer to the minimization of a squared error function as the “fragile” framework. For example, the mean minimizes the squared error function of a one-dimensional list of numbers. This procedure will normally guarantee a unique and closed-form solution, making it a very convenient choice. However, the squared error can unduly weight extreme outliers, leading to a breakdown in the model estimate. To avoid this effect, the data set needs to be cleaned thoroughly. Popular methods for cleaning real estate transaction data include:

- Removing or down-weighting data points that are many standard deviations from the mean.
- Removing distressed sales (resulting from auction, foreclosure, REO transactions, etc.) and other non-arm’s length sales (e.g., between family members).
- Identifying sellers who are known developers or home flippers and removing transactions associated with these market participants.

It is prohibitively expensive to produce a clean stream of real estate finance data. The county record system used in most states is an archipelago of disparate data sets with inconsistent formatting and imperfect data quality. The MLS (Multiple Listing Service) does a better job of harvesting the data upstream, but enjoys a monopoly,[a] making it difficult and expensive to access their data. Providers such as First American and CoreLogic aggregate and resell historical transaction data, also at high price.

Another problem is that, even if data are widely available, real estate data providers such as the FHFA and S&P still publish different indices, because data-cleaning is prevalent and the methods used to remove or reweight data points are not thoroughly documented. The methodologies disclosed in their white papers do not provide sufficient information for a practitioner to reproduce the published results, even with an identical data set.

The first of these problems relates to the national data infrastructure. Hopefully real estate finance data will become more accessible to the public in the near future. However, our real goal is to solve the second problem (and circumvent the first) by developing methods which require absolutely no data cleaning and are therefore reproducible.

[a] The National Association of Realtors, the stakeholders of the MLS, are on pace to spend about $500mm in total lobbying expenses to the federal government and Congress between 1998 and 2018. Only one other organization has spent more, namely the United States Chamber of Commerce.

Dealing with Outliers

Instead of cleaning, we can apply estimators that work even when our data are corrupted. Specifically, we propose to estimate the median price index.

Median Price Index

$$\hat{\pi}_\tau = Median(p_{h,\tau}) \tag{6}$$

$$= \underset{\pi^*_\tau}{\text{arg min}} \sum_{\forall h:t=\tau} |p_{h,t} - \pi^*_\tau| \tag{7}$$

where:

- $\hat{\pi}_\tau$ is the sample median price level in year τ.
- N_τ is the total number of transactions that occur during time τ.
- $p_{h,t}$ is the transaction price of home h in year t.

The median price index, shown in **Figure 15**, is the gold standard of off-the-shelf price index models: it requires no cleaning, is very intuitive (represents the 50th percentile of prices), and delivers on its promise of expressing trend.

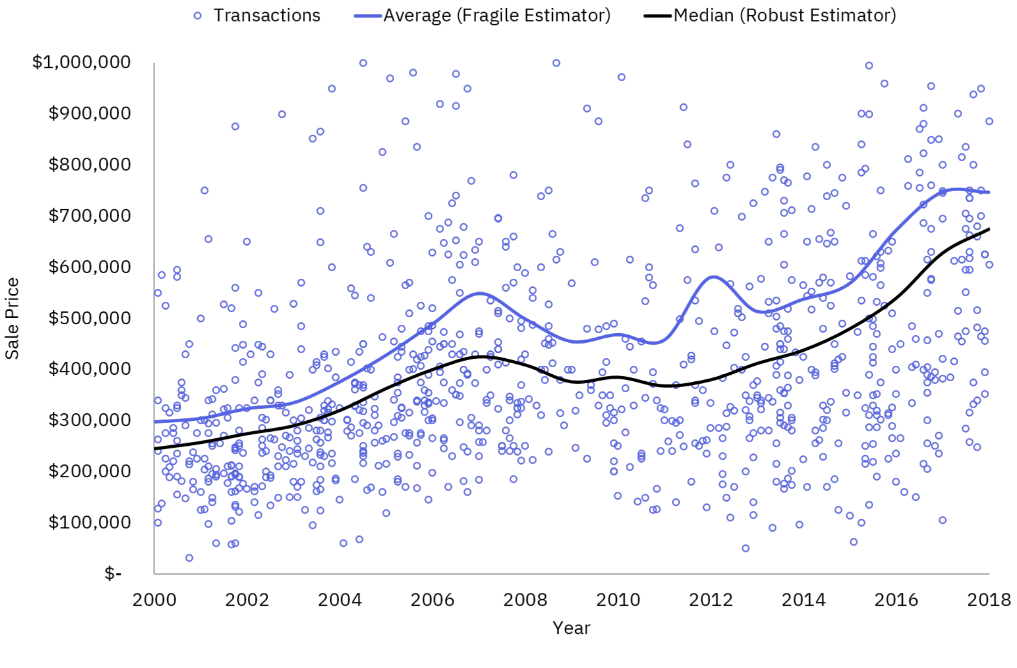

Figure 15: The mean (blue) and median (black) price indices are shown for Seattle. The median index is robust to outliers.

The reason robust estimates work so well in light of outliers and data corruption is that they tend to have a higher breakdown point than their fragile counterparts. Roughly speaking, a breakdown point is the fraction of data points that can be corrupted while still producing a relatively stable statistic. For example, the sample median has a breakdown point of 50%. Half of the data can be randomly replaced by extremely large outliers without producing a large error in the estimated median. On the other hand, the sample mean has a breakdown point of 0%. A single very large outlier can throw off the estimate.

An example of how fragile and robust estimates perform with clean and corrupt data is summarized below.

Summary of Robust vs. Fragile Estimators

Consider the following stream of home prices:

$0.8mm	$0.9mm	$1.0mm	$1.1mm	$1.2mm

We can compute (1) fragile and (2) robust measures of (A) location and (B) scale.

	Fragile		Robust	
Location	Mean	$1mm	Median	$1mm
Scale	Std. Deviation	$0.16mm	Interquartile Range	$0.20mm

What happens when we replace the $1.2mm price with a large $12mm outlier.

$0.8mm	$0.9mm	$1.0mm	$1.1mm	**$12mm**

Robust estimates remain unchanged, whereas the fragile ones break down.

	Fragile		Robust	
Location	Mean	**$1.76mm**	Median	$1mm
Scale	Std. Deviation	**$1.81mm**	Interquartile Range	$0.20mm

Without a perfectly clean data sample, the robust class of estimators comprises a much better choice than its fragile counterparts. The trade-off is that algorithms are often more complex and computation time is expanded.

Time-varying Dispersion

We will follow a similar thought process for estimating the shape parameter σ_τ which scales the standardized distribution $\mathcal{D}_\tau(0,1)$ of home prices. First, we will begin with the popular (fragile) sample standard deviation of transaction prices for each year.

Standard Deviation of Prices

$$\hat{\sigma}_\tau^2 = Var(p_{h,\tau}) \tag{8}$$

$$= \frac{1}{N_\tau} \sum_{\forall h:t=\tau} p_{h,t}^2 - \Big(\frac{1}{N_\tau} \sum_{\forall h:t=\tau} p_{h,t}\Big)^2 \tag{9}$$

where:

- $\hat{\sigma}_\tau$ is an estimate of the standard deviation of prices observed in year τ.[a]
- N_τ is the total number of transactions that occur during time τ.
- $p_{h,t}$ is the transaction price of home h in year t.

[a] A small sample bias correction factor of $\frac{N}{N-1}$ can be applied to obtain an unbiased sample standard deviation

Much like the mean price index, this is a very good estimator if there are no outliers. However, the following graph shows how sensitive these estimates can be to noisy real estate transaction data.

The blue dotted lines in **Figure 16** show just how much the sample standard deviation changes over time. At the extreme in 2012, very few transaction prices lie outside the range of values within this band, implying the bands are far too large to be meaningfully descriptive of most of the data.[1]

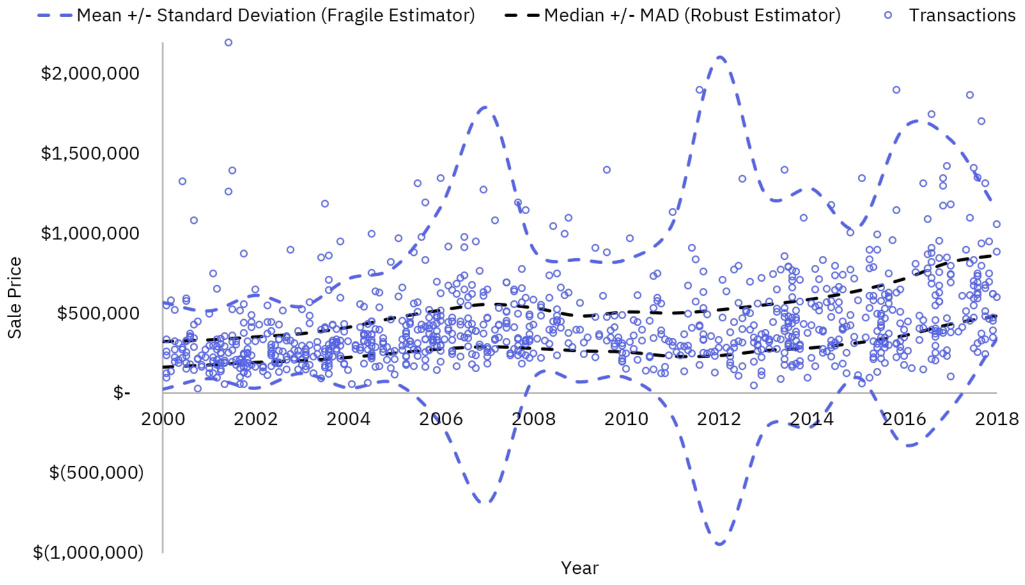

Figure 16: Standard deviation bands around the mean index (blue) and median absolute deviation bands around the median index (black) are shown for Seattle transaction prices over time.

[1] Note that by fitting our model to raw prices, it is possible to obtain deviation bands that reach negative values. If we instead transform prices into log prices $log(p_{h,\tau})$ before fitting the model, our outputs can be restricted to positive values.

The median absolute deviation range is a better alternative:

Median Absolute Deviation of Prices

$$\hat{\sigma}_\tau = MAD(p_{h,\tau}) \tag{10}$$

$$= Median(|p_{h,\tau} - Median(p_{h,\tau})|) \tag{11}$$

where:

- $\hat{\sigma}_\tau$ is an estimate of the median absolute deviation of prices observed in year τ.
- N_τ is the total number of transactions that occur during time τ.
- $p_{h,\tau}$ is the transaction price of home h in year τ.

The black dotted lines in **Figure 16** demonstrate much better behavior under the stress of outliers and a skewed and kurtotic distribution of prices.

So far, our measures of scale are agnostic to skewness in the underlying data. It is more informative for our purposes to break symmetry and deal with up side and down side deviations separately. One method that can be applied to do so is to estimate the upper band and lower bands separately. This involves computing $\hat{\sigma}^+$ by only looking at the sum of absolute deviations above the trend and $\hat{\sigma}^-$ with those below the trend line (i.e., we are simply computing the distance between the median and the 75th and 25th percentiles separately, as shown in **Figure 17**). For most of our data, $\hat{\sigma}^+ > \hat{\sigma}^-$ indicates positive skew in the distribution of home prices.

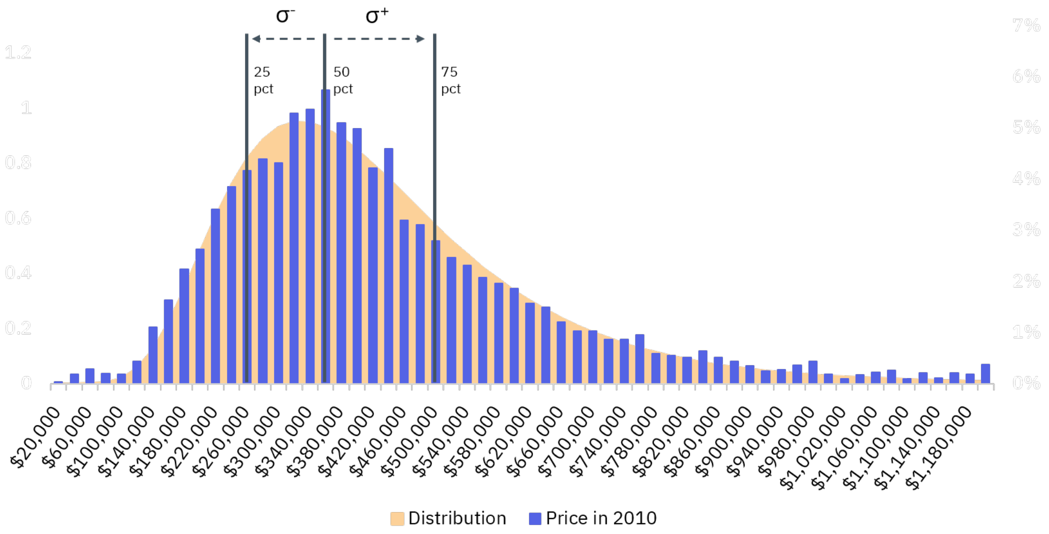

Figure 17: The median positive deviation is larger than the median negative deviation.

Overlaying the 75th and 25th percentile estimates onto the median price index produces the informative and concise price index shown in **Figure 18**.

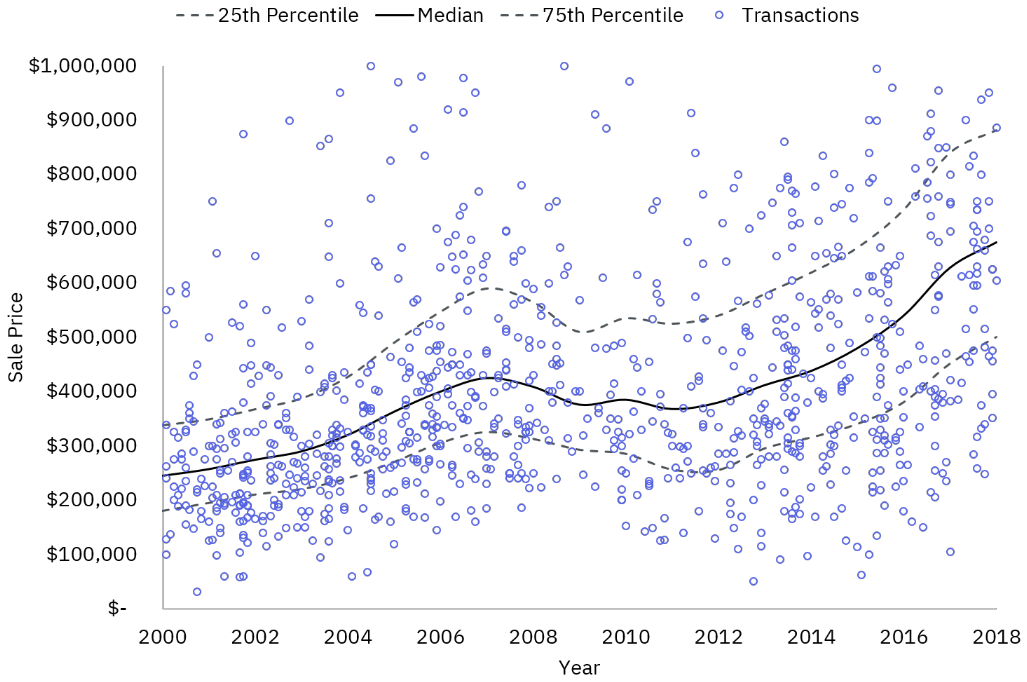

Figure 18: Median and quartile price indices.

4 Returns

In the previous chapter, we focused on transaction prices to infer how the example home in Seattle has performed. The price index analysis suggests a price estimate of $680k in 2018. This estimate was deduced by first noting that the home was in the 55th price percentile when it was purchased in 2010, and then assuming this ranking is stable over time. However, a number of factors may have affected its relative ranking. For instance, a home may have depreciated, or newly constructed homes may have entered the market: the relative quality of the home, compared to homes selling on the market, is subject to change.

Bailey, Muth, and Nourse (1963) suggest that we circumvent this problem by modelling the change in prices of individual homes. So long as the home is in roughly the same condition at purchase and sale, we have established some quality control. This style of index is referred to as a repeat-sales index, since it is fit to homes which have both a purchase and sale price, thereby inferring an individual return.

Figure 19 provides some interesting insights about the resale behavior of homeowners who purchased in 2010.

- Very few homeowners sold in the first 3 years, but those who did experienced large returns (with a huge spread of outcomes around the median). Short holding periods are generally indicative of home flipping and reconstruction activity, which breaks our constant-quality assumptions and may negatively impact our return series estimates.
- After 3 years, normal resale activity picks up. Returns increase with holding period and the variance of outcomes around the median grows over time.

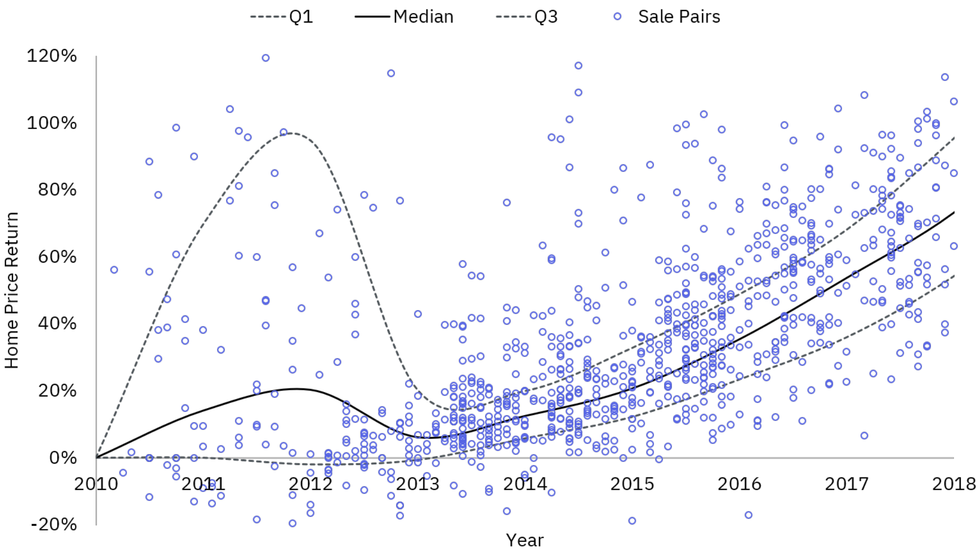

Figure 19: Return series for Seattle homes purchased in 2010 and sold in a subsequent year.

We will want to generalize the representation shown in **Figure 19**, which only focuses on homes purchased in 2010, by consolidating information about returns of homes from all purchase and sale year combinations. A first attempt at consolidating all of these combinations is shown in **Figure 20**.

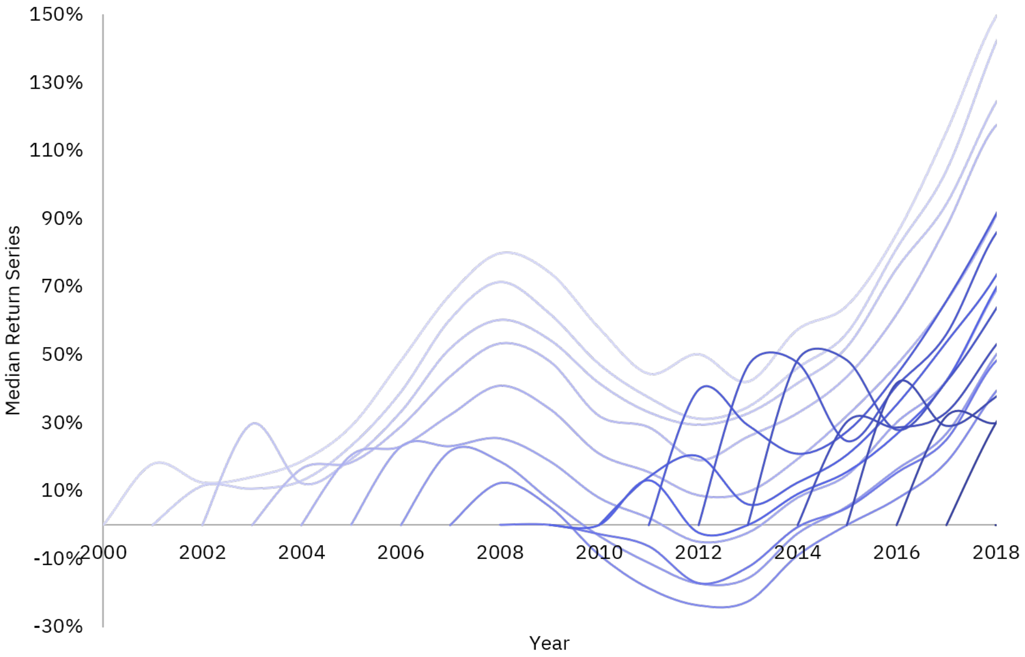

Figure 20: Median return series for homes purchased in each year between 2000 and 2018.

While containing most of the information we would want in a return index, **Figure 20** is horribly crowded. Ideally, we should consolidate all of this information into a single overall home return index. The goal of this chapter is to construct such an index while upholding the ideal benchmark criteria; namely, our index should:

1. Describe the trend: the estimates should aim to best inform the overall trend in home returns over a given time period.
2. Be robust: the estimates are resilient to a small number of corrupt data points and extreme outliers.
3. Be transparent: to maintain reproducibility, opaque data cleaning is prohibited.
4. Be computationally efficient: the solution algorithm is fast, even when handling millions of sale pairs over many time periods.
5. Minimize model assumptions: we should avoid making assumptions about the functional form of nuisance parameters (e.g., variance of returns).

Before beginning our modelling exercise, we provide a quick example of how a return index can be used to price our Seattle home.

Example: Seattle Home Return

Figure 21 is a histogram of returns for Seattle homes purchased in 2010 and sold in 2018.

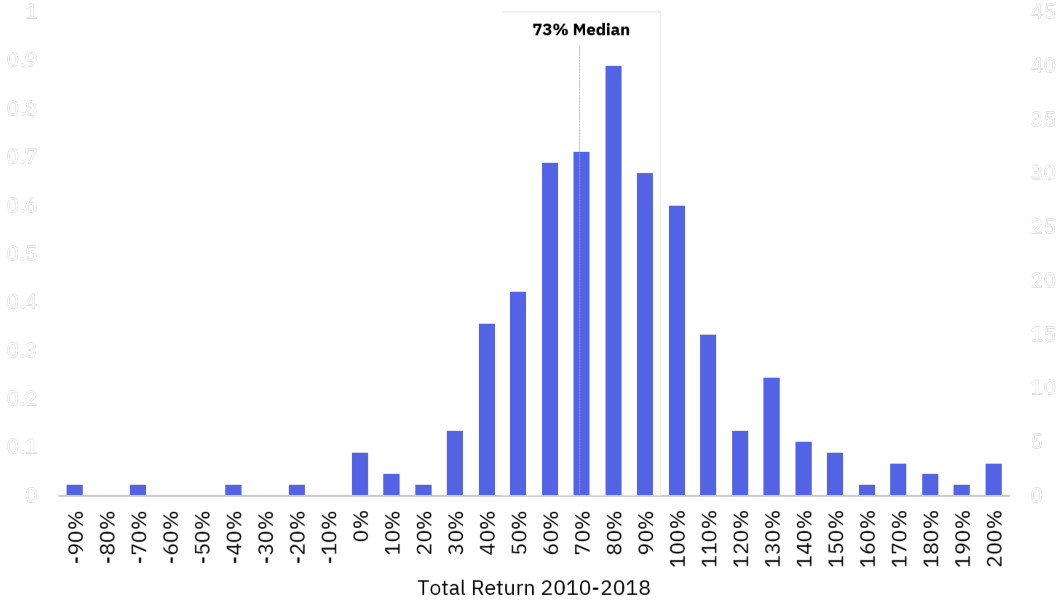

Figure 21: Median total return between 2010 and 2018 is 73%.

Recall that our price index analysis in the previous chapter suggested that the example home was worth approximately $680k, but it did not account for relative changes in the quality of the home compared to those used to generate the distribution (i.e., the market of homes sold in 2018). By taking the median of all homes' total returns between 2010 and 2018, which was 73%, we achieve a robust measure of mean performance for individual homes. The home was purchased for $400k so we estimate a 2018 value of $692k.

Return Index

The general model for home price returns is defined below.

Repeat Sale Model

$$p_{h,T} = p_{h,t}(1 + R_{h,t})(1 + R_{h,t+1})...(1 + R_{h,T-1}) = p_{h,t} \prod_{\tau=t}^{T-1} (1 + R_{h,\tau}) \qquad (12)$$

where:

- $p_{h,t}$ is the purchase price of home h in year t.
- $p_{h,T}$ is the sale price of home h in year $T \geq t$.
- $R_{h,\tau}$ is the total return in year τ.

By taking a log transformation of this equation we obtain a linear sum of log-returns, which will simplify the estimation procedure without any loss of information.

$$ln\left(\frac{p_{h,T}}{p_{h,t}}\right) = r_{h,t} + r_{h,t+1} + ... + r_{h,T-1} = \sum_{\tau=t}^{T-1} r_{h,\tau} \tag{13}$$

$$r_{h,\tau} = ln(1 + R_{h,\tau}) \sim \mathcal{D}_\tau(\mu_\tau, \sigma_\tau) \tag{14}$$

where:

- $r_{h,\tau}$ is the log return in year τ.
- $\mathcal{D}_\tau(\mu_\tau, \sigma_\tau)$ represents some distribution $\mathcal{D}_\tau$ with location parameter (center) μ_τ and scale parameter (spread) σ_τ.

Identifying the Trend

When dealing with transaction prices, each data point corresponds to a specific time period. However, returns span a series of time periods. For example, a home purchased in year 2000 and sold in 2002 can inform our estimates of returns in years 2000, 2001, and 2002.

The following table shows different possible combinations of purchase and sale events between 2000 and 2002 and their associated index returns:

Time Period	**2000**	**2001**	**2002**	Expected Return
2000 – 2000	Purchased/Sold			μ_{00}
2000 – 2001	Purchased	Sold		$\mu_{00} + \mu_{01}$
2001 – 2001		Purchased/Sold		μ_{01}
2000 – 2002	Purchased		Sold	$\mu_{00} + \mu_{01} + \mu_{02}$
2001 – 2002		Purchased	Sold	$\mu_{01} + \mu_{02}$
2002 – 2002			Purchased/Sold	μ_{02}
Index	μ_{00}	μ_{01}	μ_{02}	

Table 1: Horizontal: Purchase and Sale event combinations for associated holding period. Vertical: Holding periods and the corresponding index parameter μ_τ. Colors: Lightest blue boxes represent information revealed in year 2000. Boxes are progressively darker as they become revealed in time, as new sales occur.

We will begin with the simplest of all return index models, the mean return index:

Mean Return Index

Recall from our general model:

$$ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) = r_{h,t} + r_{h,t+1} + ... + r_{h,T-1} = \sum_{\tau=t}^{T-1} r_{h,\tau} \tag{15}$$

Each return $r_{h,\tau}$ in the series is a random variable.

$$r_{h,\tau} = \mu_\tau + \sigma_\tau \epsilon_{h,\tau} \tag{16}$$

The total expected log-return is just the sum of expected individual returns.

$$E\Big[ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big)\Big] = \mu_t + \mu_{t+1} + ... + \mu_{T-1} = \sum_{\tau=t}^{T-1} \mu_\tau \tag{17}$$

Similarly, if periodic returns are assumed to be independent and identically distributed, then variance is additive.

$$Var\Big[ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big)\Big] = \sigma_t^2 + \sigma_{t+1}^2 + ... + \sigma_{T-1}^2 = \sum_{\tau=t}^{T-1} \sigma_\tau^2 \tag{18}$$

With an assumption that log-returns are normally distributed,

$$ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) \xrightarrow{d} \mathcal{N}\Big(\sum_{\tau=t}^{T-1} \mu_\tau, \sum_{\tau=t}^{T-1} \sigma_\tau^2\Big) \tag{19}$$

we can solve for our index parameters with the following optimization program.

$$\hat{\mu}_\tau = \underset{\mu_\tau^*}{\text{arg min}} \sum_{\forall h} w_{t,T} \Big| ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) - \sum_{\tau=t}^{T-1} \mu_\tau^* \Big|^2 \tag{20}$$

We scale each estimate by a precision factor of $w_{t,T}$. If we assume variance is constant over time, then the precision factor can be simplified to a function of holding period $T - t$

$$w_{t,T} = \frac{1}{\sum_{\tau=t}^{T-1} \sigma_\tau^2} = \frac{1}{\sigma^2(T-t)} \propto \frac{1}{T-t} \tag{21}$$

where:

- $p_{h,t}$ is the purchase price of home h in year t.
- $p_{h,T}$ is the sale price of home h in year $T \geq t$.
- $r_{h,\tau}$ is the log return $ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big)$ in year τ with expected value μ_τ and variance σ_τ^2.
- $w_{t,T}$ is a precision factor which assigns a higher weight to data points with lower variance.
- $\hat{\mu}_\tau$ is a vector of estimates of our return index.

Figure 22 illustrates the mean return index for Seattle between 2000 and 2018, compressing the mean hairlines into a consolidated index.

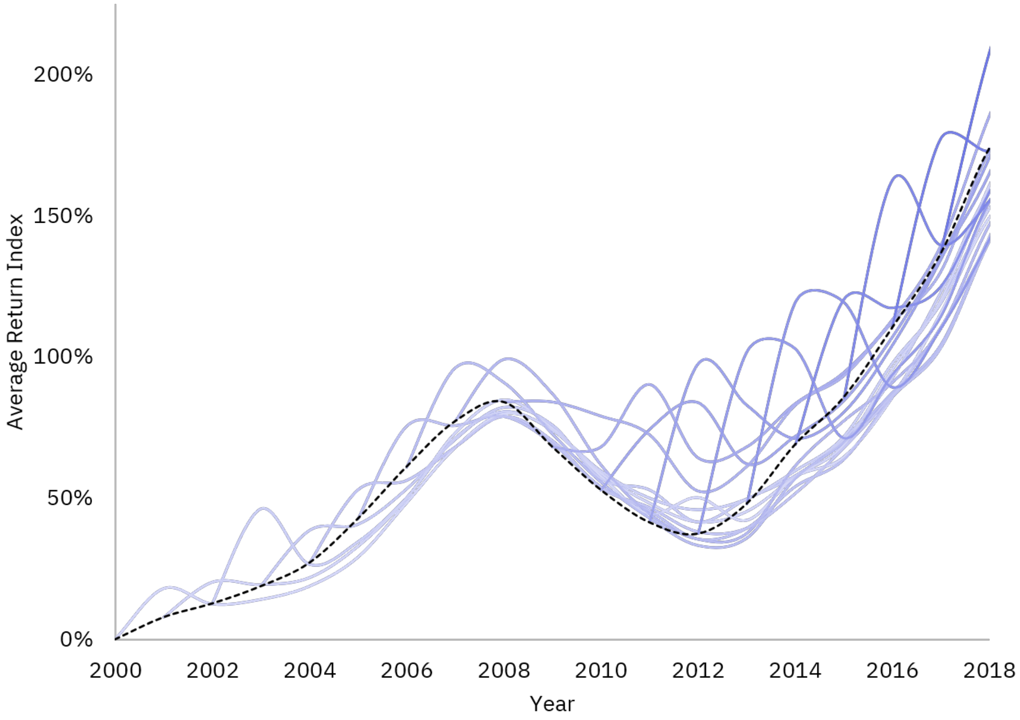

Figure 22: Mean return index represents the overall trend of home returns. Hairlines growing out of the index display dispersion not explained by the trend.

Brief History of Variance Functional Forms

Assuming that variance of log-returns is (1) constant (homoskedastic) $Var\left[ln\left(\frac{p_{h,T}}{p_{h,t}}\right)\right] = constant$ (Bailey, Muth and Nourse 1963) or (2) proportional to the holding period $Var\left[ln\left(\frac{p_{h,T}}{p_{h,t}}\right)\right] \propto (T-t)$ (Webb 1981), then our current minimization program is a straightforward convex optimization with a closed-form solution. This framework has been extended by Webb and subsequently Case, Shiller and Weiss by developing increasingly sophisticated models for variance.

Variance Model $Var\left[ln\left(\frac{p_{h,T}}{p_{h,t}}\right)\right]$	Model
$constant$	Bailey, Muth, and Nourse (1963)
$\sigma_d^2(T-t)$	Webb (1981)
$2\sigma_j^2 + \sigma_d^2(T-t)$	Case-Shiller (1987)
$2\sigma_j^2 + \sigma_d^2(T-t) + \sigma_q^2(T-t)^2$	FHFA

where,

- σ_d^2 is the return variance associated with holding period drift.
- σ_j^2 is the return variance associated with a transaction shock (note that a multiple of 2 is appended to represent the fact that there are two transactions, a purchase and a sale).
- σ_q^2 is a second order quadratic term associated with the holding period drift.

Within the Case-Shiller (1987) regression framework, a two-step procedure is required to estimate the variance (and then the correct weight) for each data point. The first stage is identical to the previously mentioned regression assuming a constant variance. Then the squared magnitude of errors is regressed against the holding period to obtain the σ_j^2 and σ_d^2 parameter estimates. The second stage involves computing the fitted variance estimates of each sale pair (using the parameter estimates σ_j^2 and σ_d^2 and the known holding period) in order to determine the correct weight for each data point. The FHFA model is very similar, but an additional term σ_q^2 is included in the squared error regression.

As the model for variance increases in sophistication, the solution's stability decreases. A single large outlier can produce senseless and even negative estimates of variance during the squared error regression. Our current project is to estimate the trend and not the functional form of variance, meaning it is just a nuisance parameter. For now, some thoughtful aggregation and knowledge of the asymptotic properties of sample statistics will allow us to circumvent the estimation of a variance model while enjoying a significant improvement in speed.

Boosting for Speed

Rather than working directly with all of the individual returns, we will define some summary statistics for each pair of distinct purchase and sale years.

Sub-sample Count:

$$n_{t,T} = \sum_{h \in (t,T)} 1$$

Where $h \in (t, T)$ refers to all homes purchased in time period t and sold in time period T.

Sub-sample Mean:

$$m_{t,T} = \frac{\sum_{h \in (t,T)} ln\left(\frac{p_{h,T}}{p_{h,t}}\right)}{n_{t,T}}$$

Sub-sample Variance:

$$s^2_{t,T} = \frac{\sum_{h \in (t,T)} \left(ln\left(\frac{p_{h,T}}{p_{h,t}}\right) - m_{t,T}\right)^2}{n_{t,T} - 1}$$

Table 2 shows, for example, the parameters corresponding to the expected returns illustrated in **Table 1**.

Time Period t, T	**Sub-Sample Count**	**Sub-Sample Mean**	**Sub-Sample Variance**
2000, 2000	$n_{00,00}$	$m_{00,00}$	$s^2_{00,00}$
2000, 2001	$n_{00,01}$	$m_{00,01}$	$s^2_{00,01}$
2001, 2001	$n_{01,01}$	$m_{01,01}$	$s^2_{01,01}$
⋮	⋮	⋮	⋮

Table 2: Summary statistics for each purchase and sale year combination.

The Central Limit Theorem (CLT) reveals that the distributions of our sub-sample means $m_{t,T}$ converge to the true index values $\sum_{\tau=t}^{T-1} \mu_\tau$ for large $n_{t,T}$. The CLT also parameterizes the rate of convergence between the sample and true values as a function of the sample count, variance $(n_{t,T} - 1)/s^2_{t,T}$ and sample deviation from the mean.

Fast Mean Return Index

Recall from our general model:

$$ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) = r_{h,t} + r_{h,t+1} + ... + r_{h,T-1} = \sum_{\tau=t}^{T-1} r_{h,\tau} \tag{22}$$

with,

$$r_{h,\tau} = \mu_\tau + \sigma_\tau \epsilon_{h,\tau} \tag{23}$$

In expectation these returns converge to their true values.

$$E\Big[ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big)\Big] = \mu_t + \mu_{t+1} + ... + \mu_{T-1} = \sum_{\tau=t}^{T-1} \mu_\tau \tag{24}$$

with a variance,

$$Var\Big[ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big)\Big] = \sigma_t^2 + \sigma_{t+1}^2 + ... + \sigma_{T-1}^2 = \sum_{\tau=t}^{T-1} \sigma_\tau^2 \tag{25}$$

The Central Limit Theorem provides us with the limiting distribution of our sub-sample summary statistics.

$$\hat{m}_{t,T} \xrightarrow{d} \mathcal{N}\Big(\sum_{\tau=t}^{T-1} \mu_\tau, \frac{\sum_{\tau=t}^{T-1} \sigma_\tau^2}{n_{t,T}}\Big) \tag{26}$$

We can solve for our index parameters with the following optimization program:

$$\hat{\mu}_\tau = \underset{\mu^*_\tau}{\text{arg min}} \sum_{\forall t,T} w_{t,T} \Big| \hat{m}_{t,T} - \sum_{\tau=t}^{T-1} \mu^*_\tau \Big|^2 \tag{27}$$

We scale each estimate by a precision factor of $w_{t,T}$. This gives sample mean estimates $\hat{m}_{t,T}$ with smaller standard errors a higher influence, since they contain more information.

$$w_{t,T} = \frac{n_{t,T}}{\sum_{\tau=t}^{T-1} \sigma_\tau^2} \rightarrow \frac{n_{t,T} - 1}{\hat{s}^2_{t,T} + (\hat{m}_{t,T} - \sum_{\tau=t}^{T-1} \mu_\tau)^2} \approx \frac{n_{t,T} - 1}{\hat{s}^2_{t,T} + \hat{m}^2_{t,T}} \tag{28}$$

where:

- $p_{h,t}$ is the purchase price of home h in year t.
- $p_{h,T}$ is the sale price of home h in year $T \geq t$.
- $r_{h,\tau}$ is the log return $ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big)$ in year τ with expected value μ_τ and variance σ_τ^2.
- $\hat{m}_{t,T}$ is the sample mean and $\hat{s}^2_{t,T}$ is the sample variance of the sub-sample of homes purchased in year t and sold in year $T \geq t$.
- $w_{t,T}$ is a precision factor which assigns a higher weight to sub-samples with a larger number of data-points $n_{t,T}$ and lower variance.
- $\hat{\mu}_\tau$ is a vector of estimates of our return index.

Note that this optimization program has far fewer terms in the summation than in the previous model. Rather than minimizing the error between our prediction $\sum_{\tau=t}^{T-1}\mu_{h,\tau}^{*}$ and every single return $ln\left(\frac{p_{h,T}}{p_{h,t}}\right)$ in the sample, we are only fitting the prediction to the sub-sample means $m_{t,T}$. As a result, the algorithm is substantially faster.[2]

We also enjoy the added benefit of using a sample variance $s_{t,T}^2$ in place of a variance function $\sigma_{t,T}^2+\hat{m}_{t,T}^2$ for our weights $w_{t,T}$, obviating the need to impose a functional form on the variance parameter. This step also better handles the first three years of resales shown in **Figure 19**, since these time periods are characterized by a low count of data points and very high variance, and thus will be automatically assigned a very low weighting.

Figure 23 shows that the two methods, namely the mean return index and the fast mean return index, produce identical results. A mathematical proof is provided in **Appendix A**.

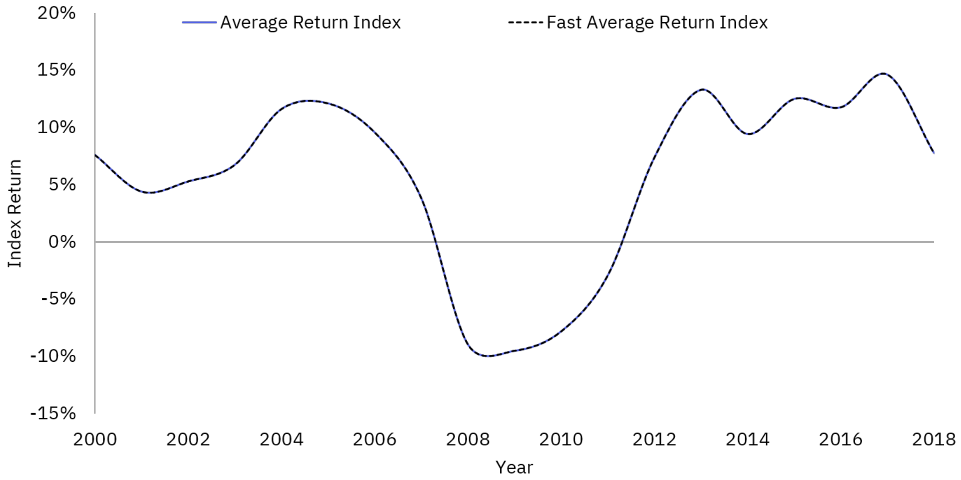

Figure 23: The mean and fast mean return indices are identical if we use the same weighting scheme.

Dealing with Outliers

Our estimates are still very fragile to corrupt data and outliers. We will now introduce a robust analog. We can do this by changing our loss function from the squared deviation function $|.|^2$ to the absolute deviation $|.|$, like we did with the median price index series.

Robust Return Index

Recall from our general model:

$$ln\left(\frac{p_{h,T}}{p_{h,t}}\right) = r_{h,t} + r_{h,t+1} + ... + r_{h,T-1} = \sum_{\tau=t}^{T-1} r_{h,\tau} \tag{29}$$

Rather than minimizing the squared error, we can minimize the absolute error in the following optimization program to obtain a robust return index.

[2] The fast regression can perform 10^2–10^3 times faster depending on if we are using annual or monthly time buckets. The complexity of the simple regression is approximately $O(np^2)$ where n is the number of sale pairs and p is the total number of time slices. The complexity of the fast regression is the greater of the following two steps. The first step, estimating sub-sample statistics, is $O(n)$ (and even faster if each sub-sample is aggregated and sample statistics are estimated in parallel). The second step is $O(p^4)$.

$$\hat{\mu}_\tau = \underset{\mu^*_\tau}{\text{arg min}} \sum_{\forall h} w_{t,T} \left| ln\left(\frac{p_{h,T}}{p_{h,t}}\right) - \sum_{\tau=t}^{T-1} \mu^*_\tau \right| \qquad (30)$$

We scale each estimate by a precision factor of $w_{t,T}$.

$$w_{t,T} = \frac{1}{\sqrt{\sum_{\tau=t}^{T-1} \sigma^2_\tau}} \propto \frac{1}{\sqrt{T-t}} \qquad (31)$$

where:

- $p_{h,t}$ is the purchase price of home h in year t.
- $p_{h,T}$ is the sale price of home h in year $T \geq t$.
- $r_{h,\tau}$ is the log return $ln\left(\frac{p_{h,T}}{p_{h,t}}\right)$ in year τ with expected value μ_τ and variance σ^2_τ.
- $w_{t,T}$ is a precision factor which assigns a higher weight to data points with lower volatility.
- $\hat{\mu}_\tau$ is a vector of estimates of our return index.

While this optimization program is convex, it is discontinuous, has no closed form solution, and is computationally expensive when applied to large amounts of data. Fortunately, servers are cheap and algorithms to handle these types of programs have been developed[3], but they are not currently able to reasonably handle millions of data points. The robust return index for Seattle is shown in **Figure 24**.

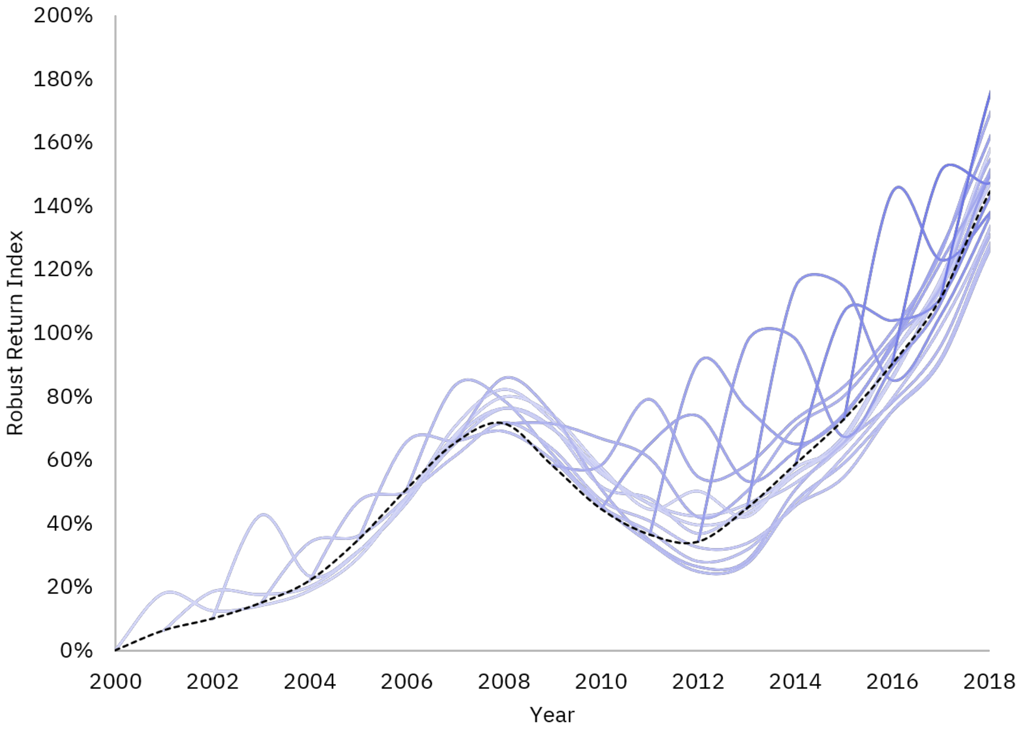

Figure 24: Robust return index for Seattle.

[3] Interior point and simplex algorithms are the preferred algorithms.

Putting It All Together

We want to repeat the process of sub-sample aggregation. However, instead of estimating the sample mean and sample variance, we will use their robust counterparts, namely the median $med_{t,T}$ and the median absolute deviation (MAD) $mad_{t,T}$.

Count:

$$n_{t,T} = \sum_{h \in (t,T)} 1$$

Sample Median:

$$med_{t,T} = Median\Big(ln(\frac{p_{h,T}}{p_{h,t}})\Big)$$

Sample Median Absolute Deviation:

$$mad_{t,T} = Median\Big(\Big|ln(\frac{p_{h,T}}{p_{h,t}}) - med_{t,T}\Big|\Big)$$

The asymptotic distributions of the sample medians $med_{t,T}$ also converge to their respective index values $\sum_{\tau=t}^{T-1} \mu_\tau$ at a rate approximately proportional to $n_{t,T}/mad^2_{t,T}$. A mathematical proof is provided in **Appendix B**.

Fast Robust Return Index

Recall from our general model:

$$ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) = r_{h,t} + r_{h,t+1} + ... + r_{h,T-1} = \sum_{\tau=t}^{T-1} r_{h,\tau} \tag{32}$$

The medians of these returns converge to their true values.

$$E\Big[\hat{med}_{t,T}\Big] = \mu_t + \mu_{t+1} + ... + \mu_{T-1} = \sum_{\tau=t}^{T-1} \mu_\tau \tag{33}$$

The asymptotic distribution of the median can be represented by:

$$\hat{med}_{t,T} \xrightarrow{a} \mathcal{N}\Big(\sum_{\tau=t}^{T-1} \mu_\tau, \frac{1}{4n_{t,T}f(x)^2}\Big) \approx \mathcal{N}\Big(\sum_{\tau=t}^{T-1} \mu_\tau, \frac{\pi\kappa^2}{2}\frac{mad^2_{t,T}}{n_{t,T}}\Big) \tag{34}$$

where κ is a constant which depends on the distribution and relates the median absolute deviation to the standard deviation ($\kappa \approx 1.4826$ for a normally distributed random variable), and $f(x)$ is the probability density function of the distribution, of returns in that sub-sample. We can solve for our index parameters with the following optimization program:

$$\hat{\mu}_\tau = \underset{\mu^*_\tau}{\text{arg min}} \sum_{\forall t,T} w_{t,T}\Big|\hat{med}_{t,T} - \sum_{\tau=t}^{T-1} \mu^*_\tau\Big|^2 \tag{35}$$

As in the fragile case, we scale each estimate by a precision factor $w_{t,T}$, which gives more weight to higher precision, more informative sub-samples.

$$w_{t,T} = \frac{n_{t,T}}{\kappa^2 mad^2_{t,T} + (\hat{med}_{t,T} - \sum_{\tau=t}^{T-1} \mu_\tau)^2} \approx \frac{n_{t,T}}{\kappa^2 mad^2_{t,T} + \hat{med}^2_{t,T}} \tag{36}$$

where:

- $p_{h,t}$ is the purchase price of home h in year t.
- $p_{h,T}$ is the sale price of home h in year $T \geq t$.
- $r_{h,\tau}$ is the log return $ln\left(\frac{p_{h,T}}{p_{h,t}}\right)$ in year τ with expected value μ_τ and variance σ_τ^2.
- $m\hat{e}d_{t,T}$ is the sample median and $m\hat{a}d_{t,T}^2$ is the sample median absolute deviation of the sub-sample of homes purchased in year t and sold in year $T \geq t$.
- $w_{t,T}$ is a precision factor which assigns a higher weight to sub-samples with a larger number of data points $n_{t,T}$ and lower variance.
- κ is a scalar which relates the median absolute deviation to the standard deviation.
- $f(x)$ is the probability density function of the distribution (in this case the normal distribution).
- $\hat{\mu}_\tau$ is a vector of estimates of our return index.

The fast robust return index gives us the best of both worlds. It is robust, extremely fast to compute and does not require us to impose a functional form on the variance parameter. **Figure 25** illustrates the fast median return index for Seattle between 2000 and 2018, compressing the median hairlines into a consolidated index.

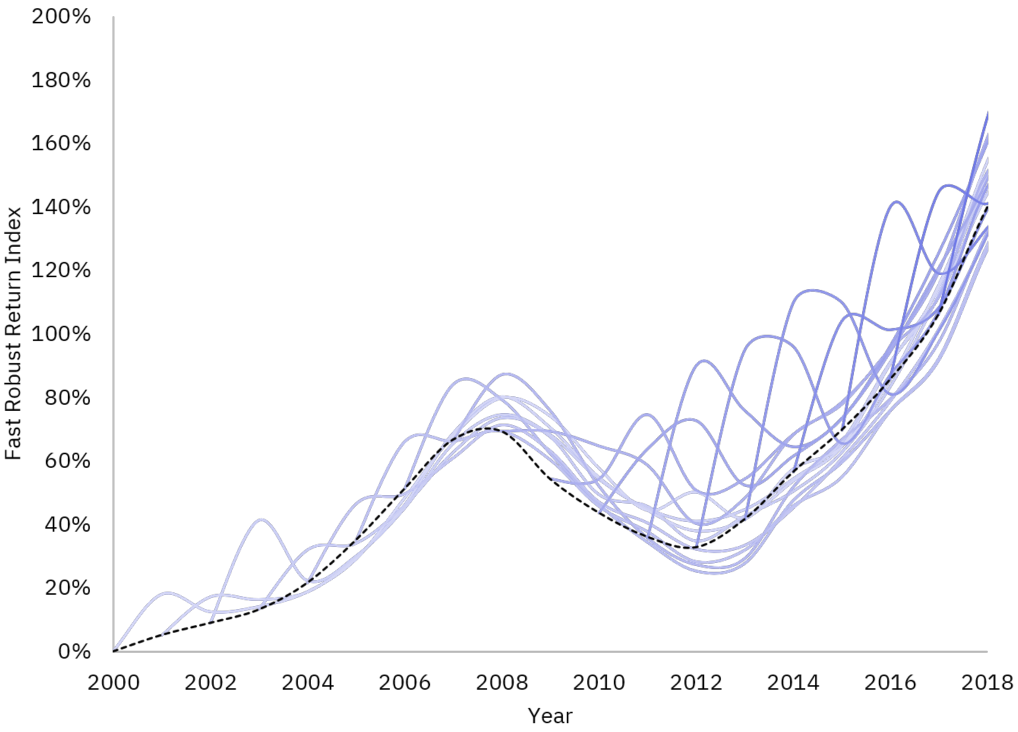

Figure 25: Fast robust return index for Seattle.

Figure 26 compares the results of the robust and the fast robust return indices. Though they are not identical, they converge asymptotically if log-returns are normally distributed.

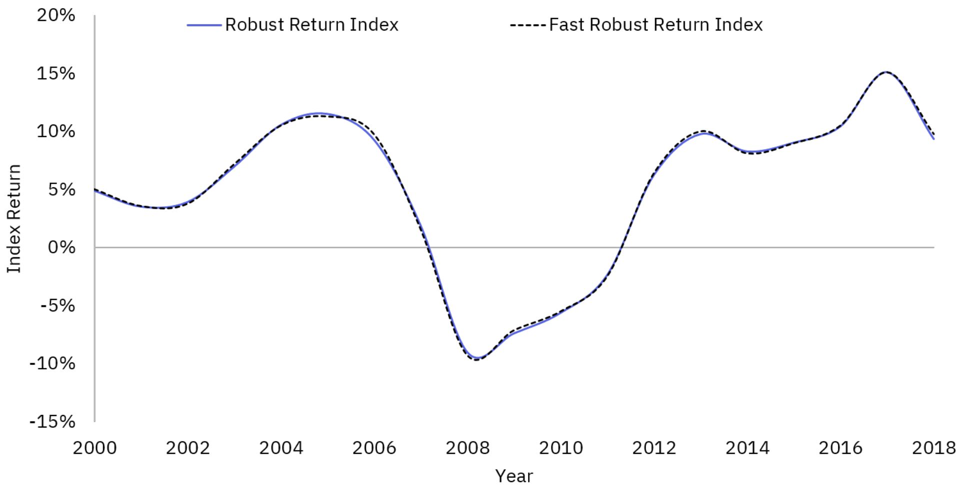

Figure 26: If we use the same weighting scheme, the robust and fast robust return indices are almost identical.

Finally, **Figure 27** compares our median (robust) and mean (fragile) return index frameworks.

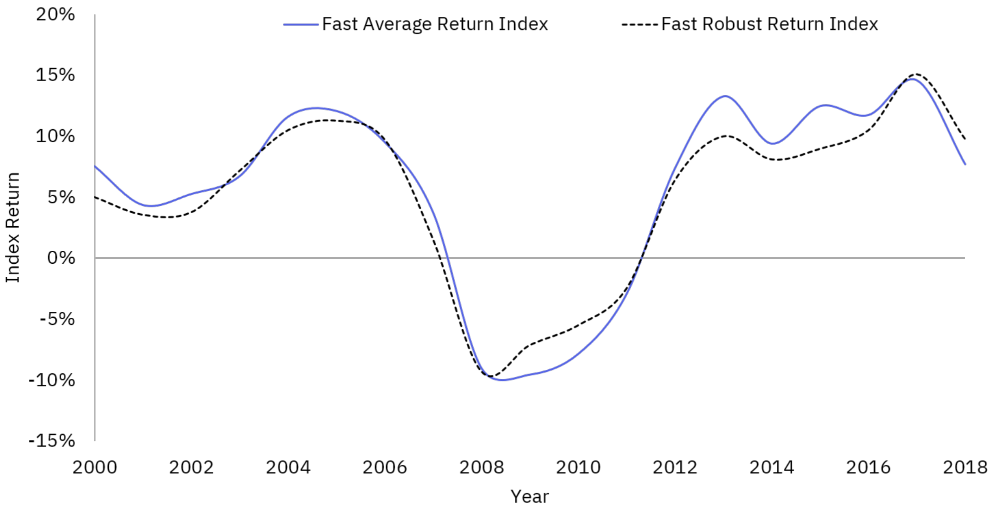

Figure 27: Comparison between the median (robust) and mean (fragile) return indices.

5 Volatility

In addition to our home return index, which expresses an expectation of home performance, we may also be interested in quantifying the uncertainty of an individual home's return around that expectation. **Figure 28** illustrates the magnitude of dispersion remaining after accounting for the trend.

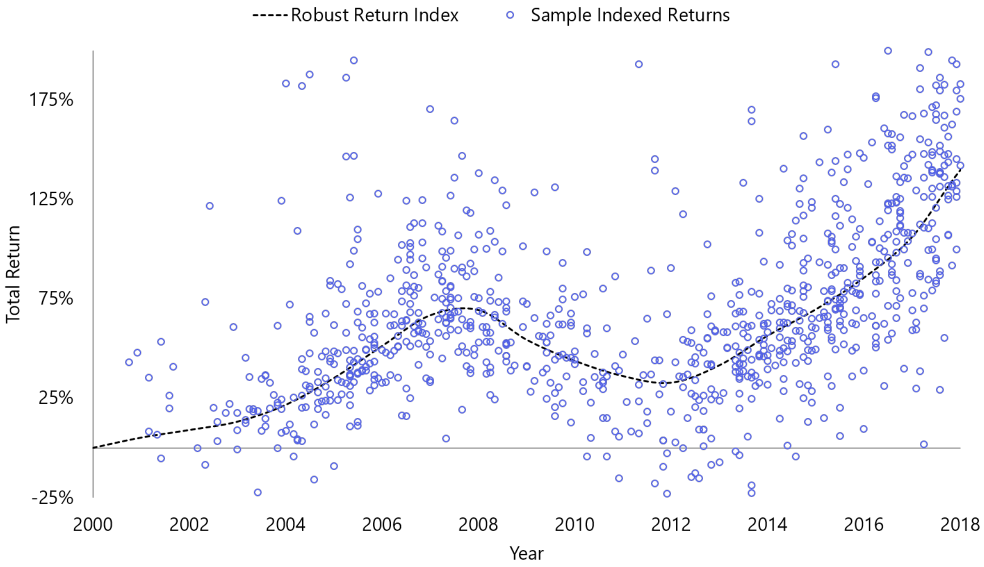

Figure 28: Indexed returns show a large spread around the median price index line.

Figure 29 comprises hairlines representing the median absolute deviations from the robust return index, organized by purchase and sale year. Note the large number of outliers in the first few years. As we saw in the section on return series, these short holding periods are highly volatile, but have a very low volume of transactions.

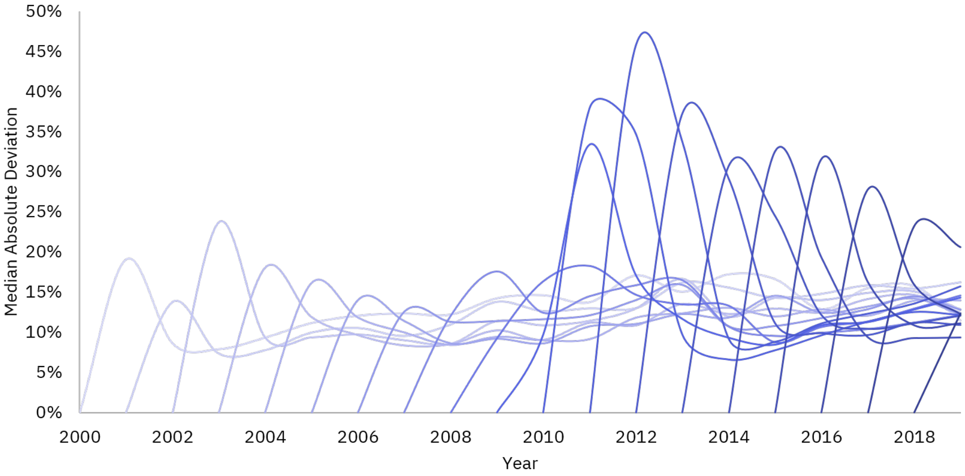

Figure 29: Median absolute deviations from the robust index are shown. Hairlines represent homes purchased in the same year.

The goal of this chapter is to produce a volatility index for home returns. We aim to uphold the same ideals as we did in the previous chapter, namely the index should identify the trend (i.e., how the amount of dispersion varies over time), be robust of outliers, be transparent, be fast to compute, limit backward revision, and minimize model assumptions. We will also uphold our tradition of working through an example before beginning the modelling exercise.

Example: Seattle Home Volatility

Figure 30 is a histogram of returns for Seattle homes purchased in 2010 and sold in 2018. It is the same chart as was shown in the previous chapter, but with a focus on dispersion instead of central tendency.

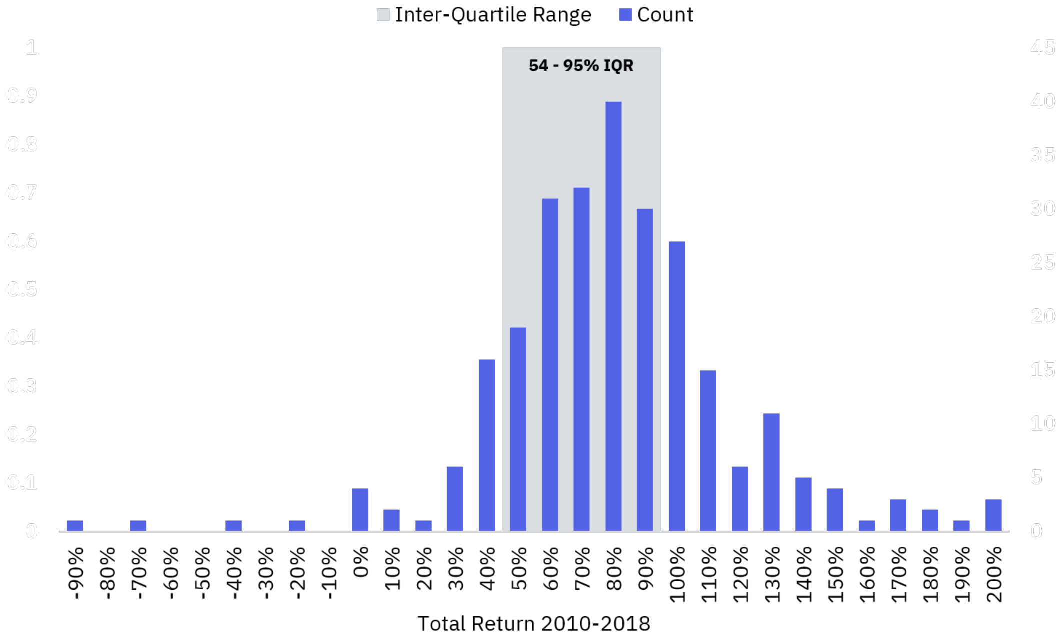

Figure 30: Interquartile range of returns is 54 to 95% for Seattle homes purchased in 2010 and sold in 2018.

Our previous empirical analysis provided a median expectation of $692k for the example home price. We are now interested in the range of probable outcomes. The interquartile range of returns provides us with an expected range of prices between $616k and $780k.

With this distribution of returns, we can look at the distribution of terminal prices for a $400k home. The example home started in the 55th percentile of prices. **Figure 31** suggests that, in 2018, there is a 50% chance that our home moves outside the 42nd to 71st percentile range, validating our intuition that a home can move around quite a bit in ranking over 8 years.

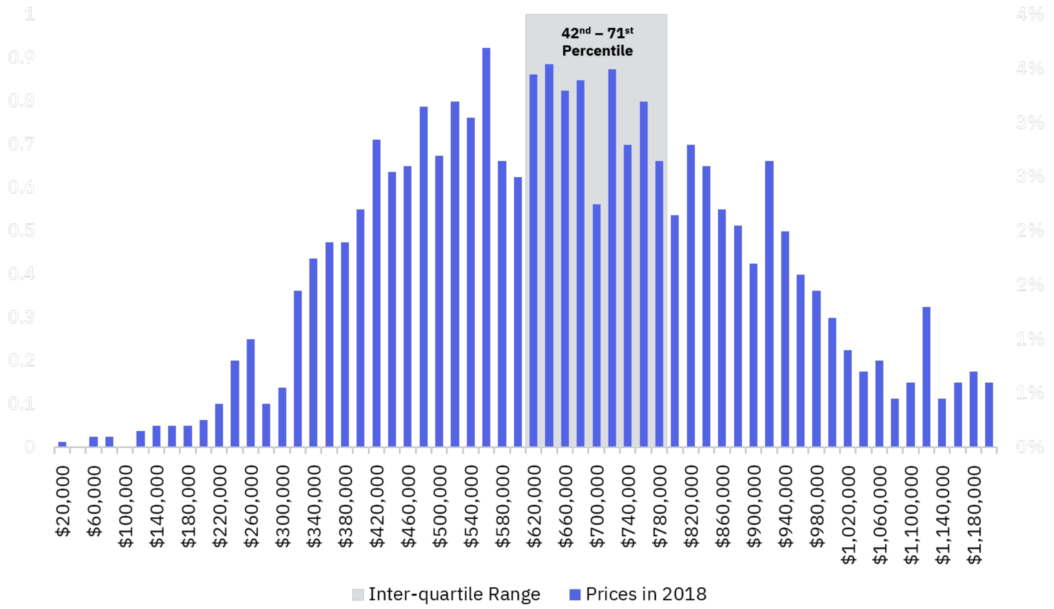

Figure 31: Interquartile range of outcomes for $400k home purchased in Seattle in 2010.

Volatility Index

Volatility Model

Recall our return series model:

$$ln\left(\frac{p_{h,T}}{p_{h,t}}\right) = r_{h,t} + r_{h,t+1} + ... + r_{h,T-1} = \sum_{\tau=t}^{T-1} r_{h,\tau} \tag{37}$$

$$r_{h,\tau} = ln(R_{h,\tau}) \sim \mathcal{D}_\tau(\mu_\tau, \sigma_\tau^2) \tag{38}$$

We separate the trend terms μ_τ from the dispersion terms σ_τ^2.

$$ln\left(\frac{p_{h,T}}{p_{h,t}}\right) = \mu_t + \mu_{t+1} + ... + \mu_{T-1} + \left(\sigma_t^2 + \sigma_{t+1}^2 + ... + \sigma_{T-1}^2\right)^{\frac{1}{2}} \epsilon_h \tag{39}$$

$$= \sum_{\tau=t}^{T-1} \mu_\tau + \left(\sum_{\tau=t}^{T-1} \sigma_\tau^2\right)^{\frac{1}{2}} \epsilon_h \tag{40}$$

$$\epsilon_h \sim \mathcal{D}_\tau(0, 1) \tag{41}$$

where:

- p_t is the purchase price in year t.
- p_T is the sale price in year $T \geq t$.
- R_t in is the return in year t.
- $r_{h,\tau}$ is the log return in year τ
- $\mathcal{D}_\tau(\mu_\tau, \sigma_\tau^2)$ represents some distribution $\mathcal{D}_\tau$ with location parameter (center) μ_τ and scale parameter (spread) σ_τ.

We already have parameters μ_τ. Our focus herein is to estimate the series of variance terms σ_τ^2.

Fragile Volatility Index

Volatility Index

Starting with our general model:

$$ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) = \sum_{\tau=t}^{T-1} \mu_\tau + \Big(\sum_{\tau=t}^{T-1} \sigma_\tau^2\Big)^{\frac{1}{2}} \epsilon_h \tag{42}$$

$$\epsilon_h \sim \mathcal{D}_\tau(0,1) \tag{43}$$

Assuming a normally distributed random variable $\mathcal{D}_\tau(0,1)$.

$$ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) \sim \mathcal{N}\Big(\sum_{\tau=t}^{T-1} \mu_\tau, \Big(\sum_{\tau=t}^{T-1} \sigma_\tau^2\Big)^{\frac{1}{2}}\Big) \tag{44}$$

We can express the following likelihood function as a function of our σ_τ^2 parameter.

$$ln\mathcal{L}(\sigma_\tau^2) = -\frac{1}{2} \sum_{\forall h} \Big(ln\big(\sum_{\tau=t}^{T-1} \sigma_\tau^2\big) + \frac{\Big| ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) - \sum_{\tau=t}^{T-1} \mu_\tau \Big|^2}{\sum_{\tau=t}^{T-1} \sigma_\tau^2} \Big) \tag{45}$$

Minimizing the negative likelihood function provides an estimate of the volatility index.

$$\hat{\sigma}_\tau^2 = \underset{\sigma_\tau^{2*}}{\text{arg min}} \sum_{\forall h} \Big(ln\big(\sum_{\tau=t}^{T-1} \sigma_\tau^{2*}\big) + \frac{\Big| ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) - \sum_{\tau=t}^{T-1} \mu_\tau \Big|^2}{\sum_{\tau=t}^{T-1} \sigma_\tau^{2*}} \Big) \tag{46}$$

where:

- $p_{h,t}$ is the purchase price of home h in year t.
- $p_{h,T}$ is the sale price of home h in year $T \geq t$.
- $r_{h,\tau}$ is the log return $ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big)$ in year τ with expected value μ_τ and variance σ_τ^2.
- $ln\mathcal{L}(\sigma_\tau^2)$ is the loglikelihood function. It is a function that expresses the likelihood of generating the data in our sample for a given value of our estimator σ_τ^2.
- $\hat{\sigma}_\tau$ is a vector of estimates of our volatility index.

Just like we did with the return index, we can improve the performance of our algorithm by pre-computing some summary statistics for each pair of purchase and sale years.

Count:

$$n_{t,T} = \sum_{h \in (t,T)} 1$$

Sum of Squared Residuals:

$$SSE_{t,T} = \sum_{h \in (t,T)} \left| ln(\frac{p_{h,T}}{p_{h,t}}) - \sum_{\tau=t}^{T-1} \mu_{h,\tau} \right|^2$$

Table 3 shows, for example, the sum of squared errors associated with the returns illustrated in **Table 1**.

Time Period t, T	**Sample Count**	**Sum of Squared Residuals**
2000, 2000	$n_{00,00}$	$SSE_{00,00}$
2000, 2001	$n_{00,01}$	$SSE_{00,01}$
2001, 2001	$n_{01,01}$	$SSE_{01,01}$
$\vdots$	$\vdots$	$\vdots$

Table 3: Summary statistics for each purchase and sale year combination.

Fast Volatility Index

Armed with our summary statistics, we can simplify the maximization program.

$$ln\mathcal{L}(\sigma_\tau^2) = -\frac{1}{2}\sum_{\forall t,T}\Big(n_{t,T}ln(\sum_{\tau=t}^{T-1}\sigma_\tau^2) + \frac{SSE_{t,T} + n_{t,T}(m_{t,T} - \sum_{\tau=t}^{T-1}\mu_\tau)^2}{\sum_{\tau=t}^{T-1}\sigma_\tau^2}\Big) \tag{47}$$

Minimizing the negative of this likelihood function provides a fast estimate of the volatility index.

$$\hat{\sigma}_\tau^2 = \underset{\sigma_\tau^{2*}}{\text{arg min}} \sum_{\forall t,T}\Big(n_{t,T}ln(\sum_{\tau=t}^{T-1}\sigma_\tau^{2*}) + \frac{SSE_{t,T} + n_{t,T}(m_{t,T} - \sum_{\tau=t}^{T-1}\mu_\tau)^2}{\sum_{\tau=t}^{T-1}\sigma_\tau^{2*}}\Big) \tag{48}$$

Where:

- $m_{t,T}$ is the sub-sample mean and $\sum_{\tau=t}^{T-1}\mu_\tau$ is the predicted sub-sample mean.
- $n_{t,T}$ is the count of homes in the sub-sample of homes purchased in year t and sold in year $T \geq t$.
- σ_τ^2 is the variance of home returns during period τ.
- $SSE_{t,T}$ is the sum of squared errors between the model prediction and the actual sale pairs in the sub-sample of homes purchased in year t and sold in year $T \geq t$.
- $\hat{\sigma}_\tau$ is a vector of estimates of our volatility index.

A proof for the fast volatility index is provided in **Appendix B**. Additionally, the optimization program, including both the gradient and Hessian of the minimization program, are provided in **Appendix C**.

Robust Volatility Index

In order to increase to robustness of our estimates, we will look at the median absolute deviation of returns around the return index as a measure of dispersion.

Count:

$$n_{t,T} = \sum_{h \in (t,T)} 1$$

Sample Median Absolute Deviation:

$$mad_{t,T} = Median\Big(\Big|ln(\frac{p_{h,T}}{p_{h,t}}) - \sum_{\tau=t}^{T-1} \mu_\tau\Big|\Big)$$

Fast Robust Volatility Index

First, recall that the median absolute deviation has a known relationship to the standard deviation for the majority of distributions, often differing by a scaling factor κ. For normally distributed data $\kappa \approx 1.4826$.

$$\sigma_{t,T} = \kappa \cdot mad_{t,T} \tag{49}$$

Assuming our data is normally distributed with some corrupt data, we should replace the sum of squared errors $SSE_{t,T} \approx \sum_{\tau=t}^{T-1} \sigma^2_\tau$ with sum of squared median absolute errors $n_{t,T} mad^2_{t,T}$ scaled by the appropriate factor κ^2.

$$ln\mathcal{L}(\sigma^2_\tau) = -\frac{1}{2}\sum_{\forall t,T}\Big(n_{t,T} ln(\sum_{\tau=t}^{T-1}\sigma^2_\tau) + \frac{n_{t,T}\kappa^2 mad^2_{t,T} + n_{t,T}(med_{t,T} - \sum_{\tau=t}^{T-1}\mu_\tau)^2}{\sum_{\tau=t}^{T-1}\sigma^2_\tau}\Big) \tag{50}$$

Minimizing the negative of this likelihood function provides a fast estimate of the volatility index.

$$\hat{\sigma}^2_\tau = \underset{\sigma^{2*}_\tau}{\text{arg min}} \sum_{\forall t,T}\Big(n_{t,T} ln(\sum_{\tau=t}^{T-1}\sigma^{2*}_\tau) + \frac{n_{t,T}\kappa^2 mad^2_{t,T} + n_{t,T}(med_{t,T} - \sum_{\tau=t}^{T-1}\mu_\tau)^2}{\sum_{\tau=t}^{T-1}\sigma^{2*}_\tau}\Big) \tag{51}$$

where:

- $med_{t,T}$ is the sub-sample median and $\sum_{\tau=t}^{T-1} \mu_\tau$ is the predicted sub-sample mean.
- $n_{t,T}$ is the count of returns in the sub-sample of homes purchased in year t and sold in year $T \geq t$.
- σ^2_τ is the variance of home returns during period τ.
- $SSE_{t,T}$ is the sum of squared errors between the model prediction and the actual sale pairs in the sub-sample of homes purchased in year t and sold in year $T \geq t$.
- $m\hat{a}d^2_{t,T}$ is the sample median absolute deviation of the sub-sample of homes purchased in year t and sold in year $T \geq t$.
- κ is a scalar which relates the median absolute deviation to the standard deviation.
- $\hat{\sigma}_\tau$ is a vector of estimates of our volatility index.

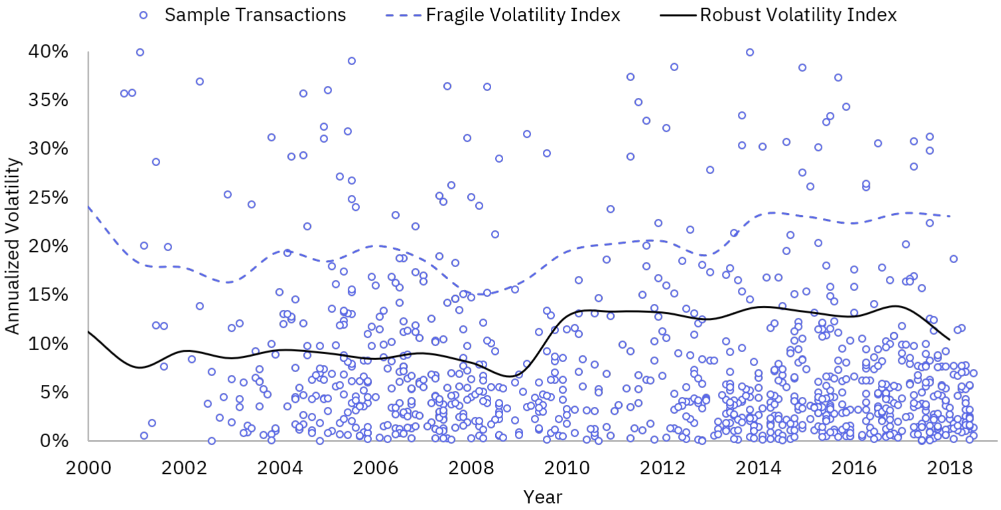

Figure 32: Robust and fragile volatility indices are shown. The annualized absolute deviations are included.

Figure 32 showcases the performance of our volatility indices. As a heuristic, if our returns are normally distributed we should expect roughly 68% of our annualized transactions to fall below the index[4]. Taking a random sample of transactions, 75% fall below the **robust volatility index** and 90% fall below the **fragile volatility index**, suggesting that the fragile index is overestimating dispersion and overfitting the outliers, assuming normally distributed returns.

Correlation

In this chapter, we focused on the idiosyncratic volatility of individual homes. This was achieved by subtracting the overall market trend from each home's return before quantifying the residual returns. However, a homeowner is primarily interested in the total volatility of the home, which also includes the volatility of the overall market trend (correlated). Conversely, an institutional investor with a large enough portfolio of homes can diversify away the idiosyncratic components, which leaves only the market trend (correlated).

Volatility Decomposition

Total variance (square of volatility) is the sum of the correlated and idiosyncratic components.

$$\sigma^2_{tot} = \sigma^2_{corr} + \sigma^2_{idio} \tag{52}$$

From our general stochastic model:

$$ln\Big(\frac{p_{h,\tau+1}}{p_{h,\tau}}\Big) = \mu_\tau + \sigma_\tau \epsilon_h \tag{53}$$

We can obtain these values from our previously generated estimates of return and volatility indices.

$$\sigma^2_{tot} = Var\Big[ln\Big(\frac{p_{h,\tau+1}}{p_{h,\tau}}\Big)\Big] \tag{54}$$

$$\sigma^2_{corr} = Var\Big[\mu_\tau\Big] \tag{55}$$

[4] Theoretically, 68% of normally distributed random variables should exhibit an absolute deviation less than one standard deviation.

$$\sigma_{tot}^2 = Var\Big[\sigma_\tau \epsilon_h\Big] = E\Big[\sigma_\tau^2\Big] \tag{56}$$

Where:

- p_τ is the home price at time τ.
- μ_τ is the index return at time τ.
- σ_τ^2 is the square of the volatility index at time τ.
- ϵ_h is the idiosyncratic error of home h.
- σ_{tot}^2 is total variance of a home.
- σ_{corr}^2 is the variance of the index (i.e., a very large portfolio) of homes.
- σ_{idio}^2 is the residual variance of a home in excess of the index.

Figure 33 presents a time series decomposition of the idiosyncratic and correlated components of variance. The area charts add up to the total variance.

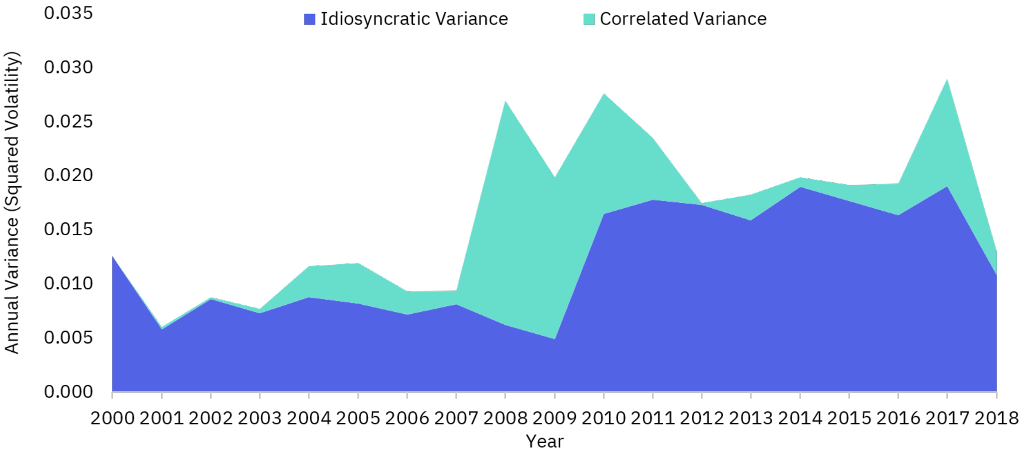

Figure 33: Area chart of the magnitude of idiosyncratic and correlated components of annual variance of home returns in Seattle between 2000 and 2018. The area charts add up to the total annual variance of home returns.

There are a few features worth noting. First, total variance (square of volatility) increases during the 2008–2010 period of home price depreciation. Second, the majority of variance during that period is correlated. This is a well-documented feature of financial markets, wherein asset volatility and correlation spike during periods of stress.

Figure 34 summarizes the long run annual volatility components (square root of variance is shown) of home returns in Seattle. We can apply these decomposed values to determine the volatility of a portfolio of homes, thereby quantifying the benefits of diversification.

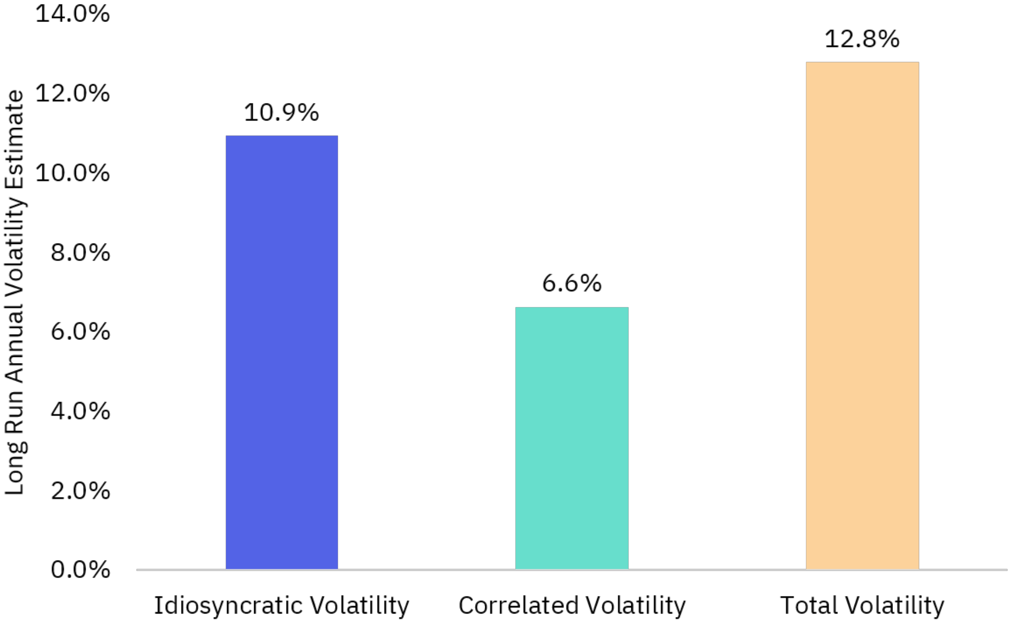

Figure 34: Long run expected values for idiosyncratic, correlated, and total volatility.

Portfolio Volatility

According to modern portfolio theory, portfolio variance can be computed for a two asset portfolio:

$$\sigma_p^2 = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\sigma_1\sigma_2\rho_{12} \tag{57}$$

We can extend this formula to a portfolio of many assets:

$$\sigma_p^2 = \sum_i \sum_j w_i w_j \sigma_i \sigma_j \rho_{ij} \tag{58}$$

Using this formula, we can determine the total volatility of a portfolio of N equally weighted homes?

$$w_i = \frac{1}{N} \tag{59}$$

$$\sigma_i\sigma_j\rho_{ij} = \begin{cases} \sigma_{corr}^2 + \sigma_{idio}^2 & \text{if } i = j \\ \sigma_{corr}^2 & \text{if } i \neq j \end{cases} \tag{60}$$

This results in the following simplified formula:

$$\sigma_p^2 = \sigma_{corr}^2 + \frac{\sigma_{idio}^2}{N} \tag{61}$$

Where:

- σ_p^2 is the portfolio variance.

- σ_i^2 is the variance of asset i.
- ρ_{ij} is the correlation between assets i and j.
- w_i is the portfolio weight associated with asset i.
- N is the number of assets in the portfolio.
- σ_{corr}^2 is the correlated component of a home's variance.
- σ_{idio}^2 is the idiosyncratic component of a home's variance.

Figure 35 illustrates how much more risk an individual homeowner is taking when compared to an institutional investor with a portfolio of homes.

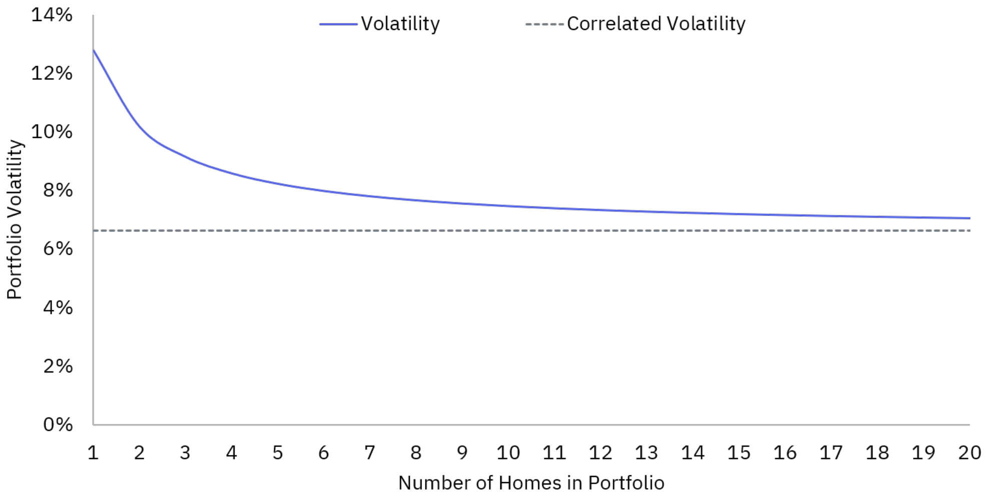

Figure 35: Total volatility as a function of the number of (equally weighted) homes in the portfolio.

6 Additional Tools

Before concluding, there are some additional tools worth mentioning. Until **Chapter 6.1**, the issue of backward revision has yet to be addressed. So far, when estimating our indices, we used all data available between 2000 and 2018 for Seattle homes. However, once 2019 transactions emerge, we may be tempted to revise past estimates of these series if new sales had long holding periods spanning the past two decades. In the following section, we will introduce the concept of time filtration and discuss how our estimates behave if we allow them to only consume data available at the time of estimation. In **Chapter 6.2**, we will investigate the effects of value weighting data points, a popular technique in index generation. Finally, in **Chapter 6.3**, smoothing techniques will be introduced to deal with the noisy estimates associated with higher frequency indexing.

Time Filtration

Unlike the price series, as time passes new transactions provide us with new information about past estimates of return. For example, in the year 2000 we only have one year of returns to inform our index value for that year. In 2001, homes purchased in 2000 then sold in 2001 may update our previously generated index value for the year 2000. **Table 4** shows that future transactions may "reach back" and influence past index series values.

Time Period	**2000**	**2001**	**2002**
2000 – 2000	Purchased/Sold		
2000 – 2001	Purchased	Sold	
2001 – 2001		Purchased/Sold	
2000 – 2002	Purchased		Sold
2001 – 2002		Purchased	Sold
2002 – 2002			Purchased/Sold

Table 4: Sub-sample groupings. Lightest blue boxes represent information available in year 2000. Boxes become progressively darker as they are revealed in time.

The emergence of new returns may compel us to retroactively restate index values. This is the result of **simultaneously** estimating the entire index each reporting period.

If we want to uphold the ideals of a benchmark index, specifically eliminating backward revision, we must lock-in estimates at the end of each reporting period, thereby estimating the index **sequentially**. At the expense of some accuracy, our benchmark will be consistent.

Simultaneous Solution

By the end of year 2000, we only have returns from the period 2000–2000 (lightest blue cells in **Table 4**). We can estimate $\hat{\mu}_{00}$:

Step 1: Year 2000

$$\hat{\mu}_{00} = \operatorname*{arg\,min}_{\mu^*_{00}} w_{00,00}|m_{00,00} - \mu^*_{00}|^2$$

where:

- $w_{t,T}$ is the precision factor of our estimate.

$$w_{00,00} = \frac{n_{00,00} - 1}{s^2_{00,00}}$$

- $\hat{\mu}_{00}$ is the estimate of our return index.
- $m_{00,00}$ is the sub-sample mean return for homes purchased in 2000 and sold in 2000.

This program has the trivial solution:

$$\hat{\mu}_{00} = m_{00,00}$$

Step 2: Year 2001 Things get more complicated when estimating μ_{01}. Homes that were purchased in 2000 and sold in 2001 (medium blue cells in **Table 4**) are providing new information about returns in 2000. The most tempting solution is to retrain $\hat{\mu}_{00}$ and $\hat{\mu}_{01}$ **simultaneously** from scratch at each time period with all available information:

$$\hat{\mu}_{00}, \hat{\mu}_{01} = \operatorname*{arg\,min}_{\mu^*_{00},\mu^*_{01}} \Big[w_{00,00}|m_{00,00} - \mu^*_{00}|^2 \tag{62}$$

$$+ w_{00,01}|m_{00,01} - \mu^*_{00} - \mu^*_{01}|^2 \tag{63}$$

$$+ w_{01,01}|m_{01,01} - \mu^*_{01}|^2 \Big] \tag{64}$$

Taking first order conditions, we obtain a vector values of $m\hat{u}_{t,T}$ which solves the following linear equation in standard $Ax = b$ format:

$$\begin{bmatrix} w_{00,00} + w_{00,01} & w_{00,01} \\ w_{00,01} & w_{00,01} + w_{01,01} \end{bmatrix} \begin{bmatrix} \hat{\mu}_{00} \\ \hat{\mu}_{01} \end{bmatrix} = \begin{bmatrix} w_{00,00}m_{00,00} + w_{00,01}m_{00,01} \\ w_{01,01}m_{01,01} + w_{00,01}m_{00,01} \end{bmatrix} \tag{65}$$

Each year τ, we obtain all new $\hat{\mu}_\tau$ values.

Sequential Solution

The simultaneous solution certainly improves estimate precision. However, an index that constantly undergoes backward revision makes for an inconsistent benchmark. A slick solution is to simply lock-in estimates each year and only estimate the latest parameters **sequentially**. **Step 1: Year 2000** is identical for the sequential solution.

$$\hat{\mu}_{00} = m_{00,00}$$

When we get to **Step 2: Year 2001**, we will instead only solve for $\hat{\mu}_{01}$ and implant our locked-in value of $\hat{\mu}_{00}$ from **Step 1**.

$$\hat{\mu}_{01} = \underset{\mu^*_{01}}{\text{arg min}} \Bigg[w_{00,00} |m_{00,00} - \boldsymbol{\hat{\mu}_{00}}|^2 \tag{66}$$

$$+ w_{00,01} |m_{00,01} - \boldsymbol{\hat{\mu}_{00}} - \mu^*_{01}|^2 \tag{67}$$

$$+ w_{01,01} |m_{01,01} - \mu^*_{01}|^2 \Bigg] \tag{68}$$

With the solution:

$$\hat{\mu}_{00,01} = \frac{w_{01,01} m_{01,01} + w_{00,01}(m_{00,01} - \hat{\mu}_{00})}{w_{00,01} + w_{01,01}} \tag{69}$$

Both steps can be combined into a lower-triangular matrix linear equation in standard $Ax = b$ format, which can be solved efficiently row by row:

$$\begin{bmatrix} w_{00,00} & 0 \\ w_{00,01} & w_{00,01} + w_{01,01} \end{bmatrix} \begin{bmatrix} \hat{\mu}_{00} \\ \hat{\mu}_{01} \end{bmatrix} = \begin{bmatrix} w_{00,00} m_{00,00} \\ w_{01,01} m_{01,01} + w_{00,01} m_{00,01} \end{bmatrix} \tag{70}$$

Figure 36 demonstrates how the simultaneous and sequential solutions differ. Note that the sequential solution is very volatile at the beginning of the period of estimation. This is due to the fact that by 2002 our largest holding period is only two years long and short holding periods have very noisy return signals. Eventually, the two return series synchronize.

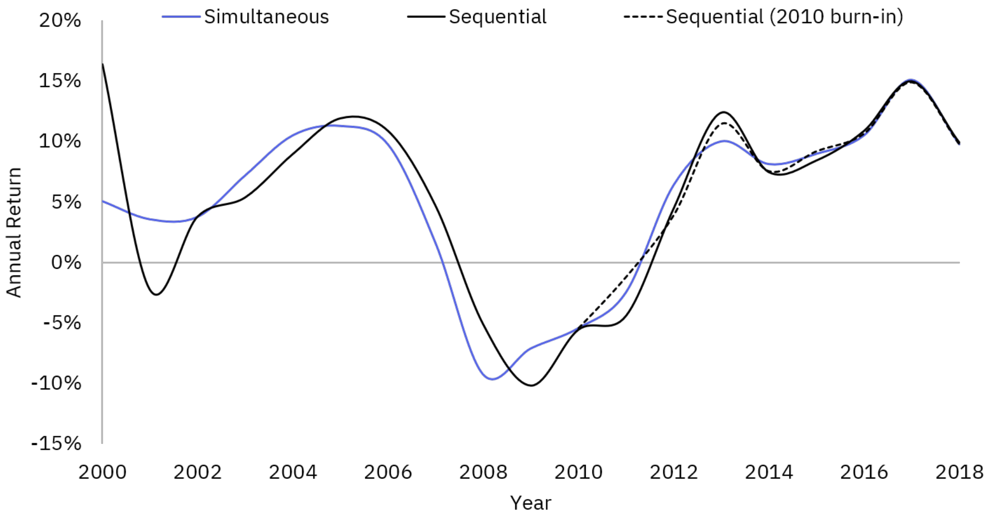

Figure 36: Annual return estimates for simultaneous and sequential solutions. Starting in 2010, a sequential solution with a 10-year burn-in period is provided.

In addition to the simultaneous and sequential series, a third series is presented. In this case, we use a 10-year burn-in period which solves the simultaneous solution between 2000 and 2010 and then proceeds using the sequential solution thereafter. With this method, we avoid the large degree of noise at the beginning of the estimation period, while producing benchmark index values which do not use future information.

Figure 37 depicts the same series but in integrated form. The sequential solution with a 10-year burn-in produces a much smaller tracking error than the full sequential solution, when compared to the fully informed simultaneous solution.

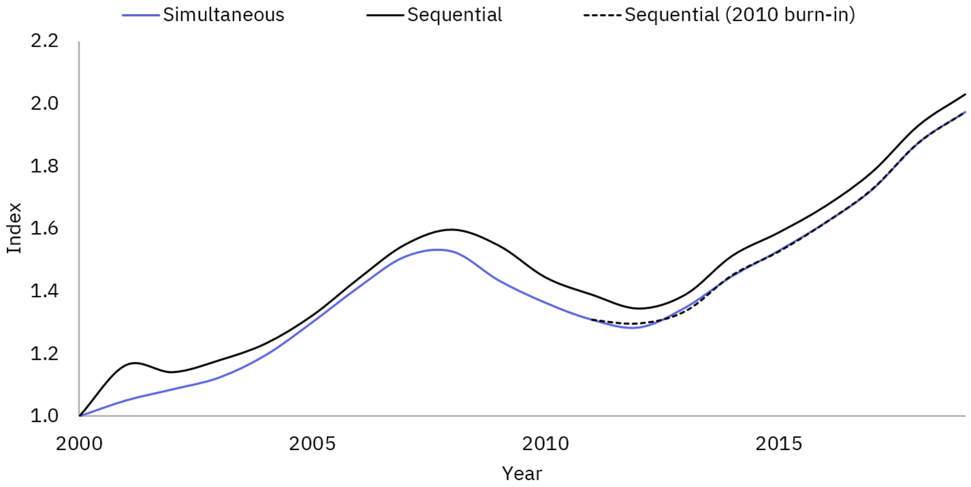

Figure 37: Integrated series for simultaneous and sequential solutions. Starting in 2010, a sequential solution with a 10-year burn-in period is provided.

Value Weighting

Another popular technique in index construction is value weighting. This technique applies a weight to each home's return proportional to its purchase price, thereby magnifying the impact of high-priced homes. In order for this technique to yield any impact on our estimates, we should detect a difference in behavior between the least and most expensive homes.

Figure 38 suggests that the least expensive homes have the highest correlated volatility and long run return.

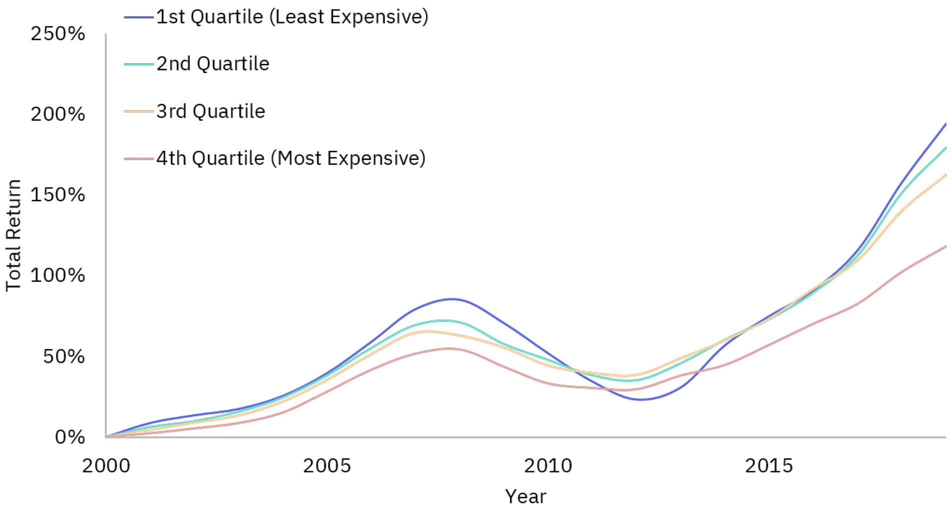

Figure 38: Robust return indices are provided for four price tiers of homes in Seattle.

Figure 39 suggests that this relationship extends to idiosyncratic volatility as well as its correlated counterpart.

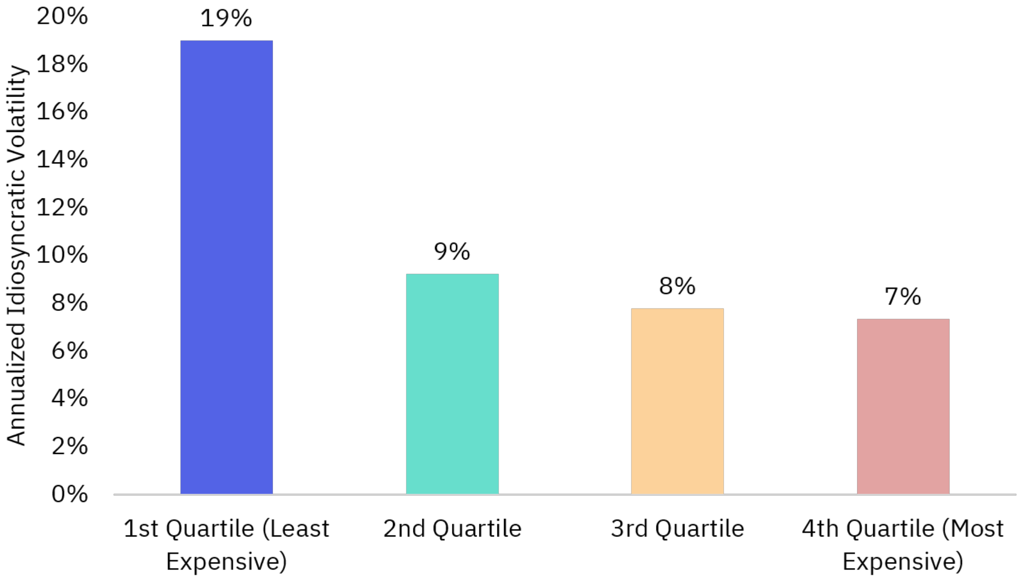

Figure 39: Idiosyncratic volatility for four price tiers of homes in Seattle.

It is straightforward to apply value weights to our fast robust indices. All changes occur during the first step of generating sub-sample summary statistics.

First, we replace the count of homes with the sum of purchase prices for homes in the sub-sample:

$$n_{t,T} = \sum_{h \in (t,T)} p_t$$

Next, we replace the median with the weighted median operator. Rather than dividing the series into two groups with the same number of terms, we aim to divide the series into two groups with equal total dollar values:

$$med_{t,T} = W.Median\Big(ln(\frac{p_{h,T}}{p_{h,t}}), weights = p_t\Big)$$

Finally, the median absolute deviation operator is replaced with the weighted median absolute deviation operator:

$$mad_{t,T} = W.Median\Big(\Big|ln(\frac{p_{h,T}}{p_{h,t}}) - med_{t,T}\Big|, weights = p_t\Big)$$

Figure 40 illustrates the annual return estimates for both value and count weighted return indices. As expected, the value weighted index expresses a reduced mean and variance of the return series.

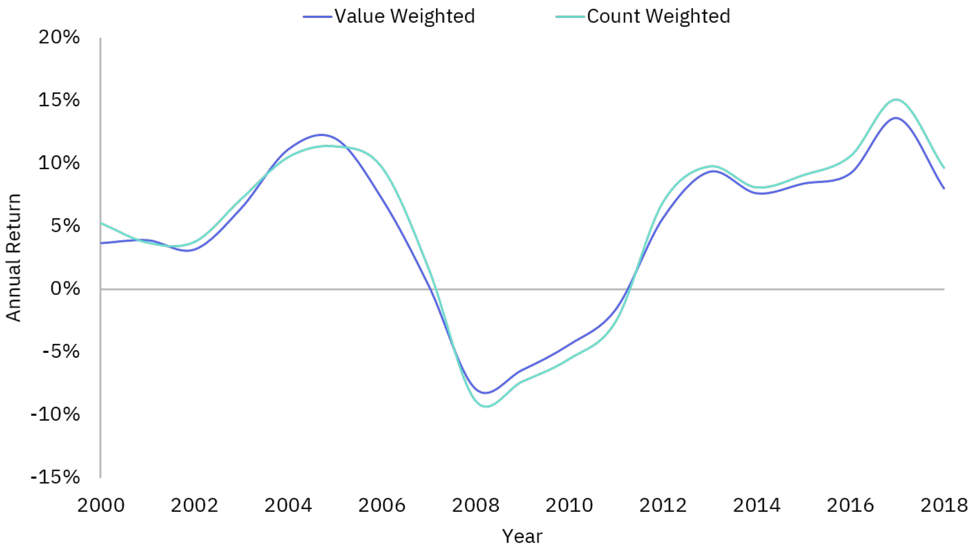

Figure 40: Annual returns of value weighted and count weighted fast robust home return indices.

The same series is shown in integrated form in **Figure 41**.

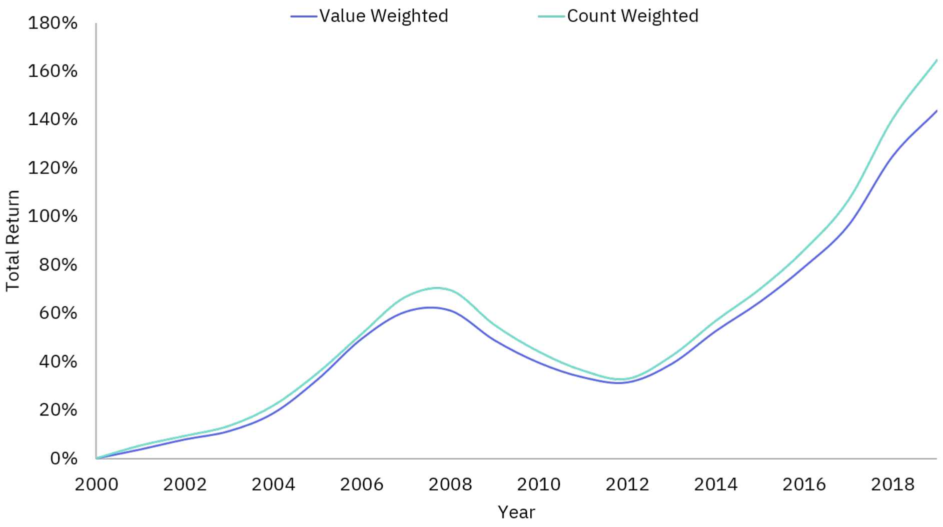

Figure 41: Total return of value weighted and count weighted fast robust home return indices.

Value weighting is appropriate when benchmarking the performance of a portfolio, since the portfolio's weights will likely be proportional to price. However, if we are looking at an individual home, the count weighted version–especially in the home's associated price tier–provides a better characterization.

Regularization and Seasonality

For the sake of simplicity, we have focused primarily on annual time periods. However, we may also be interested in a more frequent look at real estate returns.

The trade-off for increasing resolution is that we have fewer data points in each sub-sample of purchase and sale dates. This amounts to an increase in estimation noise, and therefore lower precision. **Figure 42** illustrates the magnitude of index return noise.

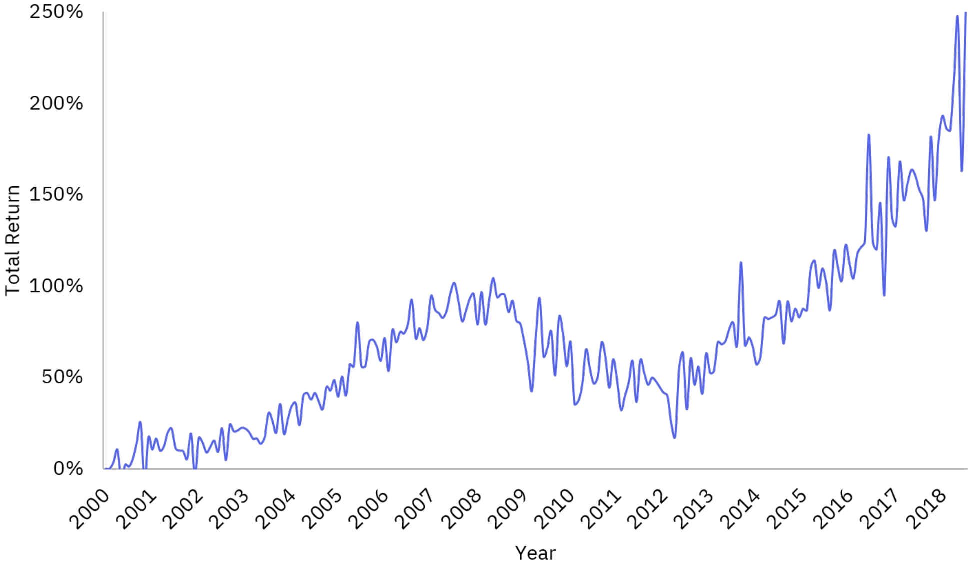

Figure 42: Raw monthly return index for Seattle homes.

A simple method to reduce noise is to apply a moving average (smoothing) of the past three months of returns. However, this reveals a seasonal component which aligns with a well-documented increase in prices during spring and summer, and a drop in prices during fall and winter.[5] To remove the seasonal component we can take a 12-month moving average. **Figure 43** illustrates the smoothed and seasonally adjusted indices.

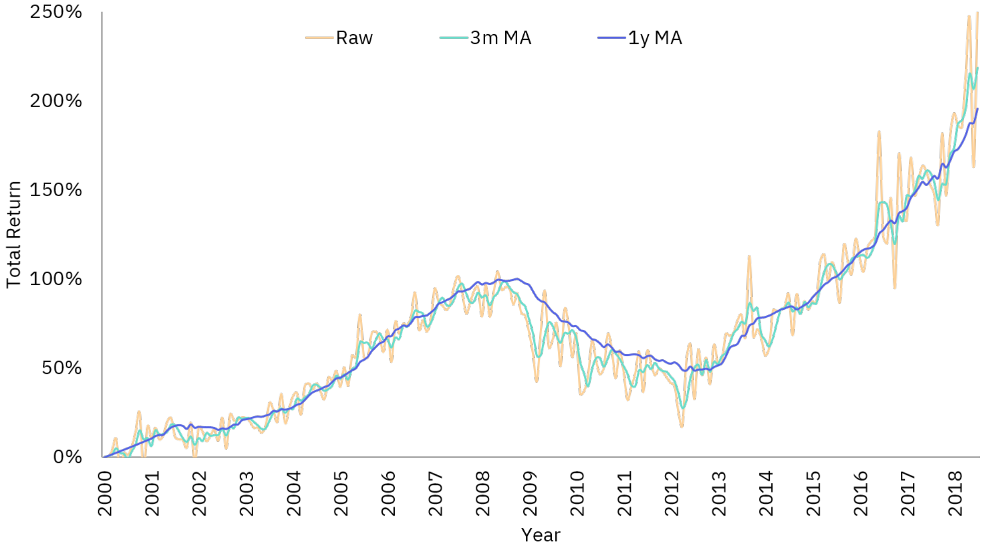

Figure 43: Smoothed (3m MA) and seasonally adjusted (12m MA) monthly return indices for Seattle homes.

[5] Case, K.E., Shiller, R.J. (1989). "The Efficiency of the Market for Single Family Homes"

A Appendix: Proof of Fast Mean Algorithm

The mean price index is given by the following optimization program.

$$\hat{\mu}_\tau = \underset{\mu^*_\tau}{\text{arg min}} \sum_{\forall h} \frac{1}{\sigma^2_{t,T}} \left| ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) - \sum_{\tau=t}^{T-1} \mu^*_\tau \right|^2 \tag{71}$$

It will be sufficient to show that the fast mean index estimates equal the mean index estimates $\mu^f_\tau = \mu_\tau$. The fast mean index estimates are given by:

$$\hat{\mu}^f_\tau = \underset{\mu^*_\tau}{\text{arg min}} \sum_{\forall t,T} \frac{n_{t,T}}{\sigma^2_{t,T}} \left| \hat{m}_{t,T} - \sum_{\tau=t}^{T-1} \mu^*_\tau \right|^2 \tag{72}$$

Note that μ_τ and μ^f_τ are vectors comprising maximum likelihood estimates of our return index. In order to solve the minimization programs, we can set the follow first order conditions (FOC).

$$0 = -\sum_{\forall h} \frac{2}{\sigma^2_{t,T}} \left(ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) - \sum_{\tau=t}^{T-1} \hat{\mu}_\tau \right) \tag{73}$$

By grouping all homes $h_{t,T}$ into sub-samples that share the same purchase year t and sale year $T \geq t$, we can rewrite this as a nested sum.

$$0 = -\sum_{\forall t,T} \sum_{\forall h_{t,T}} \frac{2}{\sigma^2_{t,T}} \left(ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) - \sum_{\tau=t}^{T-1} \hat{\mu}_\tau \right) \tag{74}$$

Noting that,

$$\hat{m}_{t,T} = \frac{1}{n_{t,T}} \sum_{\forall h_{t,T}} ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) \tag{75}$$

The inner sum can be simplified to a sum of the means,

$$\sum_{\forall h_{t,T}} \frac{2}{\sigma^2_{t,T}} \left(ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) - \sum_{\tau=t}^{T-1} \hat{\mu}_\tau \right) = \frac{2n_{t,T}}{\sigma^2_{t,T}} \left(\hat{m}_{t,T} - \sum_{\tau=t}^{T-1} \hat{\mu}_\tau \right) \tag{76}$$

The original FOC is rewritten and shown to equal the FOC of the fast mean optimization program.

$$0 = -\sum_{\forall t,T} \frac{2n_{t,T}}{\sigma^2_{t,T}} \left(\hat{m}_{t,T} - \sum_{\tau=t}^{T-1} \hat{\mu}_\tau \right) = -\sum_{\forall t,T} \frac{2n_{t,T}}{\sigma^2_{t,T}} \left(\hat{m}_{t,T} - \sum_{\tau=t}^{T-1} \hat{\mu}^f_\tau \right) \tag{77}$$

B Appendix: Proof of Robust Return Index

The limiting distribution of the sample median of normally distributed random variables is well known to itself be normally distributed.

$$\hat{med}_{t,T} \xrightarrow{a} \mathcal{N}\Big(\sum_{\tau=t}^{T-1}\mu_\tau, \frac{1}{4n_{t,T}f(\sum_{\tau=t}^{T-1}\mu_\tau)^2}\Big) \tag{78}$$

Where $f(x)$ refers to the value of the probability density function of the random variable $ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big)$ in the sub-sample $h \in (t,T)$. Since we are assuming a normally distributed variable, we can use its well-known density function.

$$f(x|M_{t,T},S_{t,T}) = \frac{1}{\sqrt{2\pi S_{t,T}^2}}exp\Big\{-\frac{(x-M_{t,T})^2}{2S_{t,T}^2}\Big\} \tag{79}$$

With,

$$M_{t,T} = \sum_{\tau=t}^{T-1}\mu_\tau \tag{80}$$

And S^2 is the variance, which can be estimated as the sample variance $s_{t,T}^2$ of the sub-sample of homes which were purchased in year t and sold in year $T \geq t$. For $x = M_{t,T}$ we can simplify the value of our function.

$$f(M_{t,T}|M_{t,T},S_{t,T}) = \frac{1}{\sqrt{2\pi S_{t,T}^2}} \tag{81}$$

The asymptotic distribution is now given by the following expression.

$$\hat{med}_{t,T} \xrightarrow{a} \mathcal{N}\Big(\sum_{\tau=t}^{T-1}\mu_\tau, \frac{\pi s_{t,T}^2}{2n_{t,T}}\Big) \tag{82}$$

We can now consider the likelihood that the parameter μ_τ takes on the value μ_τ^*.

$$\mathcal{L}\Big(\mu_\tau^*|\hat{med}_{t,T}\Big) = C_{t,T}exp\Big\{-\frac{n_{t,T}(\hat{med}_{t,T} - \sum_{\tau=t}^{T-1}\mu_\tau^*)^2}{\pi s_{t,T}^2}\Big\} \tag{83}$$

$C_{t,T}$ is simply a normalizing constant. If we take the joint likelihood function across all sub-sample medians.

$$\mathcal{L}\Big(\mu_\tau^*|\forall\hat{med}_{t,T}\Big) = \prod_{\forall(t,T)} C_{t,T}exp\Big\{-\frac{n_{t,T}(\hat{med}_{t,T} - \sum_{\tau=t}^{T-1}\mu_\tau^*)^2}{\pi s_{t,T}^2}\Big\} \tag{84}$$

The log-likelihood function that we wish to maximize is simply the natural logarithm of this express.

$$ln\mathcal{L}\Big(\mu_\tau^*|\forall\hat{med}_{t,T}\Big) = \sum_{\forall(t,T)} ln(C_{t,T}) + -\frac{n_{t,T}(\hat{med}_{t,T} - \sum_{\tau=t}^{T-1}\mu_\tau^*)^2}{\pi s_{t,T}^2} \tag{85}$$

Which can be simplified when considering only the maximization of μ_τ^*.

$$\underset{\mu_\tau^*}{\text{arg max}} \sum_{\forall(t,T)} -\frac{n_{t,T}(\hat{med}_{t,T} - \sum_{\tau=t}^{T-1}\mu_\tau^*)^2}{s_{t,T}^2} \tag{86}$$

C Appendix: Algorithm for Volatility Index Estimation

Starting with the log-likelihood function of our (squared) volatility index parameters σ_τ^{2*}

$$\hat{\sigma}_\tau^2 = \underset{\sigma_\tau^{2*}}{\text{arg min}} \sum_{\forall h} \Big(ln(\sum_{\tau=t}^{T-1} \sigma_\tau^{2*}) + \frac{\left| ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) - \sum_{\tau=t}^{T-1} \mu_\tau \right|^2}{\sum_{\tau=t}^{T-1} \sigma_\tau^{2*}} \Big) \tag{87}$$

We can replace this with a nested sum for sub-samples with the same purchase year t and sale year $T \geq t$.

$$\hat{\sigma}_\tau^2 = \underset{\sigma_\tau^{2*}}{\text{arg min}} \sum_{\forall (t,T)} \sum_{\forall h_{t,T}} \Big(ln(\sum_{\tau=t}^{T-1} \sigma_\tau^{2*}) + \frac{\left| ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) - \sum_{\tau=t}^{T-1} \mu_\tau \right|^2}{\sum_{\tau=t}^{T-1} \sigma_\tau^{2*}} \Big) \tag{88}$$

The inner sum can be simplified to a sum of the average variance $s_{t,T}^2$ within sub-samples and the variance between sub-samples and the fitted index $(m_{t,T} - \sum_{\tau=t}^{T-1} \mu_\tau)^2$.

$$\sum_{\forall h_{t,T}} \Big(ln(\sum_{\tau=t}^{T-1} \sigma_\tau^{2*}) + \frac{\left| ln\Big(\frac{p_{h,T}}{p_{h,t}}\Big) - \sum_{\tau=t}^{T-1} \mu_\tau \right|^2}{\sum_{\tau=t}^{T-1} \sigma_\tau^{2*}} \Big) = n_{t,T} ln(\sum_{\tau=t}^{T-1} \sigma_\tau^{2*}) + \frac{n_{t,T}(s_{t,T}^2 + (m_{t,T} - \sum_{\tau=t}^{T-1} \mu_\tau)^2)}{\sum_{\tau=t}^{T-1} \sigma_\tau^{2*}} \tag{89}$$

Therefore our optimization program can be rewritten as follows.

$$v_{t,T}^2 = s_{t,T}^2 + (m_{t,T} - \sum_{\tau=t}^{T-1} \mu_\tau)^2 \tag{90}$$

$$\hat{\sigma}_\tau^2 = \underset{\sigma_\tau^{2*}}{\text{arg min}} \sum_{\forall (t,T)} n_{t,T} ln(\sum_{\tau=t}^{T-1} \sigma_\tau^{2*}) + \frac{n_{t,T} v_{t,T}^2}{\sum_{\tau=t}^{T-1} \sigma_\tau^{2*}} \tag{91}$$

For the sake of clarity, we will replace the sum of variance terms $\sum_{\tau=t}^{T-1} \sigma_\tau^2$ by a dot product between:

1. an identifier vector $\vec{x}_{t,T}$ which takes a value of 1 if home is held during that period and 0 if it is not; and
2. a vector with all of the variance terms $\vec{\beta} = [\sigma_0^2, \sigma_1^2, \ldots, \sigma_T^2]^T$.

$$\sum_{\tau=t}^{T-1} \sigma_\tau^2 = \vec{x}_{t,T}^T \vec{\beta} \tag{92}$$

$$\vec{\beta} = \underset{\sigma_\tau^{2*}}{\text{arg min}} \sum_{\forall (t,T)} n_{t,T} ln(\vec{x}_{t,T}^T \vec{\beta}^*) + \frac{n_{t,T} v_{t,T}^2}{\vec{x}_{t,T}^T \vec{\beta}^*} \tag{93}$$

In order to improve the speed of convergence of this algorithm, it will help to solve the gradient $\nabla f(x)$ and the Hessian $H(x)$.

$$\nabla f(\vec{\beta}^*) = \sum_{\forall (t,T)} \frac{n_{t,T}}{\vec{x}_{t,T}^T \vec{\beta}^*} \Big(1 - \frac{v_{t,T}^2}{\vec{x}_{t,T}^T \vec{\beta}^*} \Big) \vec{x}_{t,T} \tag{94}$$

$$H(\vec{\beta}^*) = \sum_{\forall (t,T)} \frac{n_{t,T}}{(\vec{x}_{t,T}^T \vec{\beta}^*)^2} \Big(2 \frac{v_{t,T}^2}{\vec{x}_{t,T}^T \vec{\beta}^*} - 1 \Big) \vec{x}_{t,T} \vec{x}_{t,T}^T \tag{95}$$

C Appendix: Algorithm for Volatility Index Estimation

With a suitable initial guess of $\vec{\beta}_0$ and a constraint that all values of beta are positive, the following algorithm efficiently converges to the optimal solution.

$$\vec{\beta}_{n+1} = gain * \delta + \vec{\beta}_n \tag{96}$$

Where δ is the solution to the following linear equation, and the gain parameter can be used to reduce the magnitude of the update in case a β with negative terms is generated.

$$H(\vec{\beta}_n)\delta = \nabla f(\vec{\beta}_n) \tag{97}$$

$$\delta = H^{-1}(\vec{\beta}_n)\nabla f(\vec{\beta}_n) \tag{98}$$

So long as $H(.)$ is invertible and $s_{t,T}^2$ can be estimated, this method will converge to the optimal solution.

Acknowledgements

I would like to acknowledge the significant contributions of the following persons:

- Eric Reiner for performing an incredibly thorough review of the technical aspects and style of this paper.
- Laurent El Ghaoui for introducing me to the application of robust methods in financial engineering during my graduate studies, and for his review of the technical aspects of this paper.
- Terrance Odean for the long, thought-provoking discussions that seeded the ideas presented herein.

It isn't a coincidence that all three of these advisors associate with the Master of Financial Engineering program at the Haas School of Business, University of California, Berkeley. Linda Kreitzman, executive director and assistant dean, and her wonderful team have nurtured a rich network and made it readily available to the students and alumni of the program.

I must also thank the Unison Investment Management team members Rayan Rafay, Thomas Sponholtz, Brad Lookabaugh, Andrew Toby, and Scott Thompson for having the patience and motivation to invest in long-term research projects, which promote greater understanding of single-family home real estate investments. This white paper would not have been possible without their support and contributions.

Finally, I want to thank Winfield Xu, Andrew Gierke, Payton Bush and Nathan Neeteson for reviewing the paper and providing valuable feedback on its style and substance.

Any errors are my own and should not be associated with any of the above-mentioned persons.

References

Bailey, M.J., Muth, R.F., Nourse, H.O. (1963). “A Regression Method for Real Estate Price Index Construction.” Journal of the American Statistical Association, 58, 933–942.

Brown, L.D., Nagaraja, C.H., Wachter, S.M. (2010). “House Price Index Methodology.” Department of Statistics, The Wharton School, University of Pennsylvania. https://repository.upenn.edu/statistics_papers/145

Calhoun, C. (1996). “FHFA House Price Indices: HPI Technical Description.”
https://www.fhfa.gov/PolicyProgramsResearch/Research/Pages/HPI-Technical-Description.aspx

Case, K.E., Shiller, R.J. (1987). “Prices of Single-Family Homes Since 1970: New Indexes for Four Cities.” New England Economic Review, September/October 45–56.

Case, K.E., Shiller, R.J. (1989). “The Efficiency of the Market for Single Family Homes.” The American Economic Review, Vol. 79, No. 1, pp. 125-137.

First American. Deeds, Mortgages and Foreclosures data set.

Huber, Peter J. (1981). “Robust statistics.” New York: John Wiley & Sons, Inc.

Miller, Norm G. and Peng, Liang (2006). “Exploring Metropolitan Housing Price Volatility.” Journal of Real Estate Finance and Economics, 33:1, 5–18.

S&P CoreLogic Case-Shiller Home Price Indices (2007).
https://us.spindices.com/index-family/real-estate/sp-corelogic-case-shiller

Webb, Cary. (1981). “The Expected Accuracy of a Price Index for Discontinuous Markets.”