



Current AVM Testing Methodologies versus our  
new Predictive Testing Methodology (PTM™)

*OR*

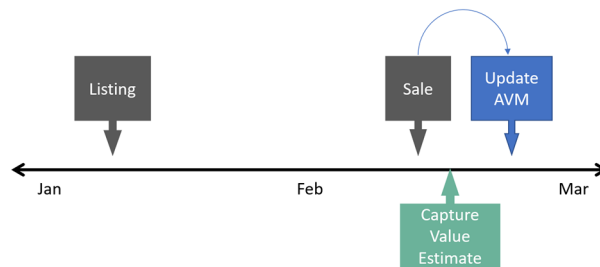
How Listing Prices Made Current AVM Testing  
Obsolete and How to Fix It

**Lee Kennedy,**

**President and Founder, AVMetrics, LLC.**

## Overview

AVMetrics has developed a new approach to testing AVM accuracy. AVM testing has evolved through three basic eras. In the 2000s, data updates and calculations were slow and expensive, everything was done in batches, and value estimates were calculated infrequently. Testing was done by comparing AVM value estimates to closed transactions without a lot of procedures to ensure that the AVM had not already incorporated the transaction value, because incorporating recent sales transactions into an estimation algorithm took time.



*Figure 1: Original Testing Method: Sale then Value Estimate and then AVM Update*

However, in the mid-2010s, data had become more available and much closer to real time, and computer processing power improved, allowing models to recalculate much more frequently. Multiple protocols were used to ensure that models' estimates were true estimates and not yet informed by the transaction price. Testing timing was compressed, and models were required to self-report the last known sales price, among other things.

In recent years these precautions have been made obsolete by the use of Multiple Listing Services' listing prices in the value estimation process by AVMs. This innovation presented a whole new challenge, because MLS data<sup>a</sup> may be available to models long before transactions are posted.

This phenomenon can be observed in figure 2, which depicts all participating models reacting in a particular market to listing prices in a particular month. By amalgamating all the predictions in a particular county (Maricopa, for this example), we can show how they react to listing prices. In the month of August in Maricopa, there were 333 properties for which we have listing price data and a history of price predictions for every model. Properties were listed at an average premium of 13.1% to the average predicted price, and in the next month, predicted prices on average jumped 11.3%. This example shows anchoring in action.

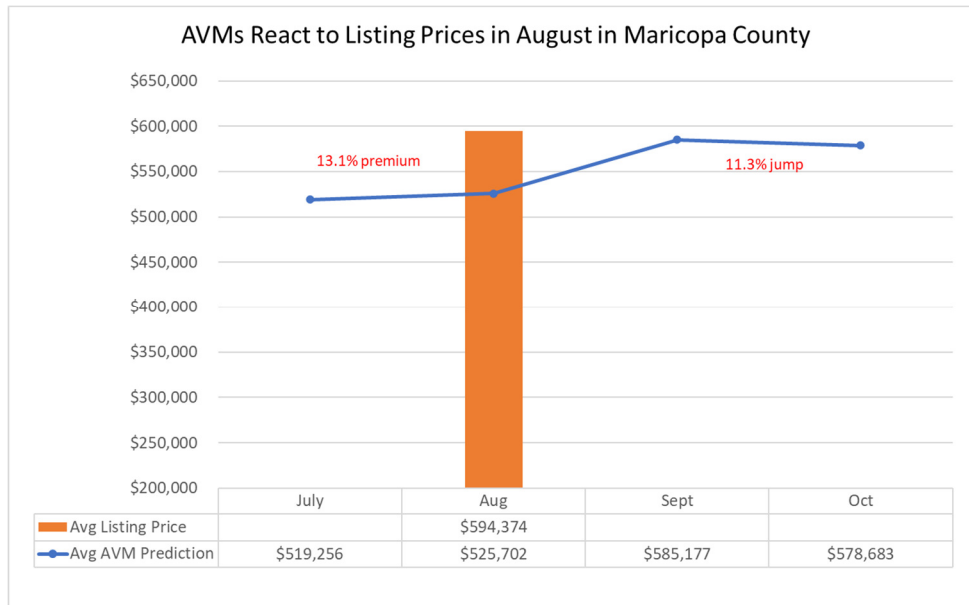


Figure 2 AVMs' average predictions vs listing price in August (Maricopa County example)

AVMetrics' solution is to capture the value estimate before the listing date, as illustrated in figure 3.

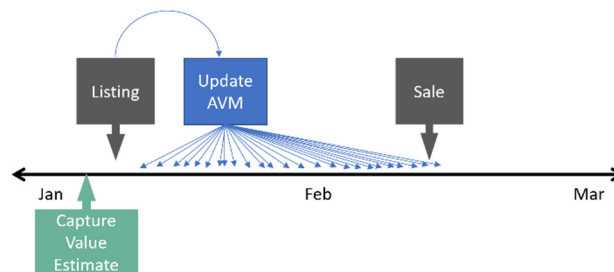


Figure 3: PTM™ Testing Method: Capture Value Estimates first, prior to Listings and AVM Updates. Sales, which are the basis for evaluating the estimates, occur last.

The challenge with this approach is knowing which homes will be listed and sold ahead of time. To circumvent this problem, we capture the predicted value of every residential property in the United States from every model in our testing on a monthly basis and have been doing so since 2021. This Model Repository Database (MRD™) lays the foundation for our Predictive Testing Methodology (PTM™).

## AVM Testing Requirements

A fundamental principle of testing is that ***the conditions of testing should match the conditions of use in the real world***. Currently, AVM testing conditions almost always include a listing price that may be known by the models. However, the real-world uses of AVMs often include valuing collateral for HELOCS and refinance or supporting quality assurance on appraisals of properties with no or dated MLS data. To be useful for those applications, an AVM needs to be accurate in the absence of recent MLS listing prices. An AVM that performs well only after it is informed of a property's listing price is not as useful as an AVM that can perform in the absence of such leading data.

Dodd-Frank, through FIRREA section 1125, requires regulated lenders to adhere to five principles of AVM use, including ensuring confidence in AVMs and performing random testing. But the evolution of the industry has made prior methods of testing fundamentally flawed. Our two studies demonstrate the extent to which the old methods are failing the industry.

## AVMetrics' Limited Study and (New) National Footprint Study on Anchoring

In 2022, AVMetrics conducted a study limited to seven valuation models and properties in California. That study came to the same conclusions, but to test the robustness of those findings, we have now expanded to 15 models in 2023 and expanded from a single state (California) to a national footprint with 50 States and 2,893 counties. We relied on about 3 million benchmark transactions throughout 2023, including a special analysis of the fourth quarter of 2023.

In this study, we have performed a comparative side-by-side analysis of both our Legacy and our new PTM™ testing systems. We used an outcomes analysis to see how the models were anchoring their valuation estimates after a listing price was published. We obtained value estimates for the same properties for use in both testing methodologies to ensure that the sampled properties were identical. This gave us a measure of the difference in the model(s) testing outcomes between the two methodologies. **We believe that this new testing method establishes a new “normal” for the prediction accuracy of models uninfluenced by listing price testing outcomes.**

Not surprisingly, as in our original POC, the data clearly shows that models do respond quite unmistakably to listing prices. As before, the time series data show that when a listing price becomes available and “informs” the model(s), the prediction of value adjusts to the new information. There is a discrete “jump” in the value estimate that can be seen in the data. This behavior reflects the models' response to being “informed” of the listing price and subsequently “anchoring” on that number instead of relying solely on their core estimation techniques.

In this expanded National analysis of our new PTM™ testing system, we focused the analysis on the National footprint and then drilled down on the states of California, Arizona and Florida and counties of Los Angeles, Maricopa and Miami-Dade as examples. In the four-quarter trend analysis the sample size was Q1~550,000, Q2 ~797,000, Q3~787,000 and Q4 at ~757,000 for a testing set of +/- 3 million properties.

| Location                | 2023Q1 | 2023Q2  | 2023Q3 | 2023Q4 |
|-------------------------|--------|---------|--------|--------|
| Arizona state           | 21,850 | 32,141  | 25,599 | 22,859 |
| Maricopa County, AZ.    | 14,454 | 21,028  | 15,695 | 14,236 |
| California state        | 51,908 | 75,421  | 82,782 | 69,120 |
| Los Angeles County, CA. | 9,716  | 13,158  | 12,549 | 12,617 |
| Florida state           | 74,944 | 101,930 | 82,735 | 88,157 |
| Miami-Dade County, FL.  | 5,116  | 5,341   | 6,404  | 6,559  |

Table 1- Record counts from 3 states and 3 counties for all four quarters of 2023.

To conduct the analysis, a common set of benchmarks from both testing platforms was created. This “filtered<sup>b</sup>” data set was ~636,000 records in Q4 of 2023, which is about 84% of the gross sample size. That ratio is representative of the other quarters. The advantage of having such enormous sample sizes in all of our testing means that we have the luxury of paring down each data set until we have the exact same benchmarks being used in two different methodologies and yet still maintain a dataset of ~636,000 records for Q4, which is comfortably capable of nationwide analysis, state level analysis and county level analysis in hundreds of the most populous counties in the country.<sup>c</sup>

The selected fifteen (15) models were part of our Legacy testing process in 2023. These models also have a full set of 2023 data in our MRD™ database covering the entire nation. (There are now over 20 models in the MRD™ to support the PTM™ going forward into 2024.) The comparative tests confirmed that, when comparing model performance before and after the listing price is accessible, ***the difference in accuracy and error is significant.***

The following analysis is displayed as both a four-quarter trend analysis and an individual model analysis focused on three key metrics. Those key metrics are

- **PPE** - the percentage of properties for which the AVM predicts selling prices to within +/- a given percentage. The complement of the Failure Rate.
- **MAE** (mean absolute error) - the absolute value of the average percentage error that the model exhibits when evaluating the properties in the test sample.
- **RT20** (or Right Tailed error greater than 20%) - the percentage of properties whose AVM valuation exceeds its corresponding selling price by more than 20 percent. RT20 indicates a potential overvaluation that can expose a lender to significant risk.

## The Analysis

Figure 4 shows the overall performance of the 15 models throughout 2023 based on the PTM™ dataset using only the uninformed value estimates. The “Within PPE 10%” trend for example, ranges from ~56% to ~62% on average without the listing price to anchor to. The best model each quarter performed ~12 to ~13 percentage points better than the average.

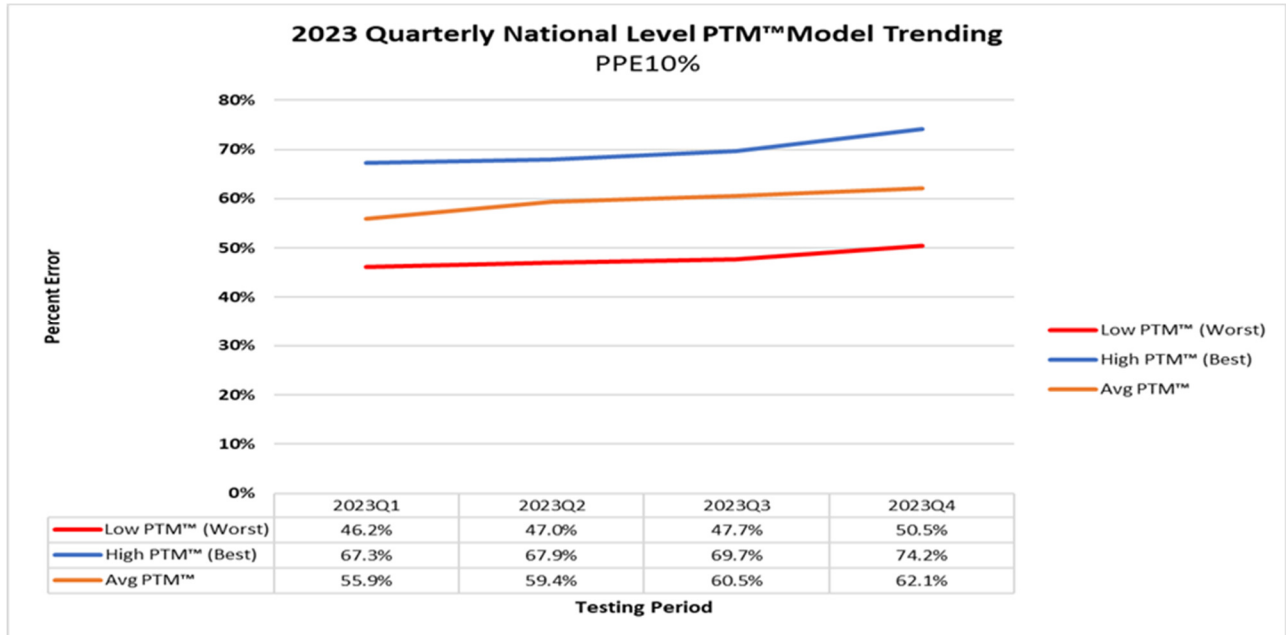


Figure 4 - 2023 overall model performance trend using PPE10.

Figure 5 below details the performance of each of the 15 models in the fourth quarter of 2023. The blue bars represent PPE10 error percentages when models are informed, and the red bars represent performance when models are uninformed. A few conclusions are immediately apparent<sup>d</sup>:

- There is a lot of variance in the baseline performance of the 15 subject models.
- Every model performs better when informed of listing prices.
- The improvement from exposure to listing prices is very significant.

For example, Model A improved its PPE10 metric by almost 25 percentage points by relying on listing prices (from 51.9% to 76.8%). Overall, the group improved their PPE10 performance by 22 percentage points through the use of listing prices (from 61.8% to 83.8%).

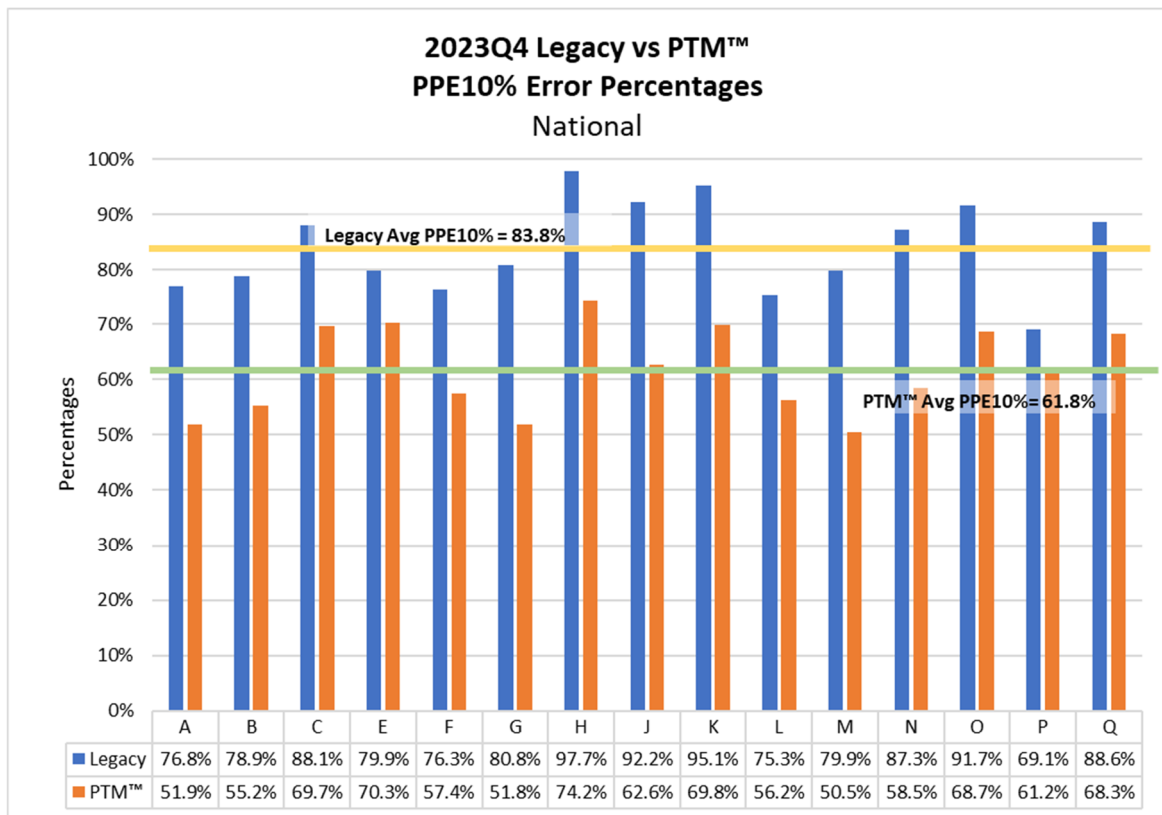


Figure 5 - Q4 Performance of each model, with and without anchoring, across all 50 states.

Figure 6 is analogous to figure 5, but performance metrics were limited to transactions in Q4 in California. The same conclusions are apparent. A few models gain only a very slight advantage by using listing prices (models E, F and P). Other models improve over 25 percentage points to a phenomenal PPE10 of 97.7% (model H).

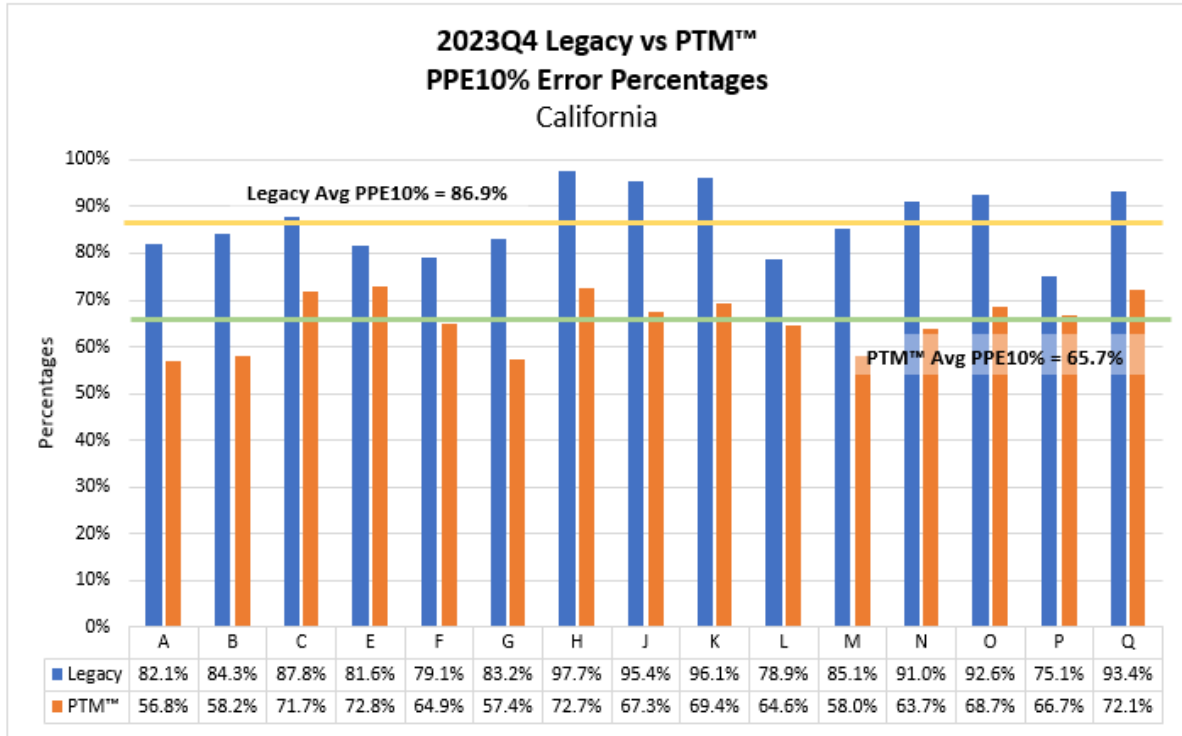


Figure 6 - Q4 Performance of each model, with and without anchoring, in California only.

Figure 7 is again analogous to the previous pictures, but it reflects an analysis that has been limited to Q4 in Los Angeles County, CA. Overall, the models improved their PPE10 performance by 21.7 percentage points (from 64% to 85.7%).

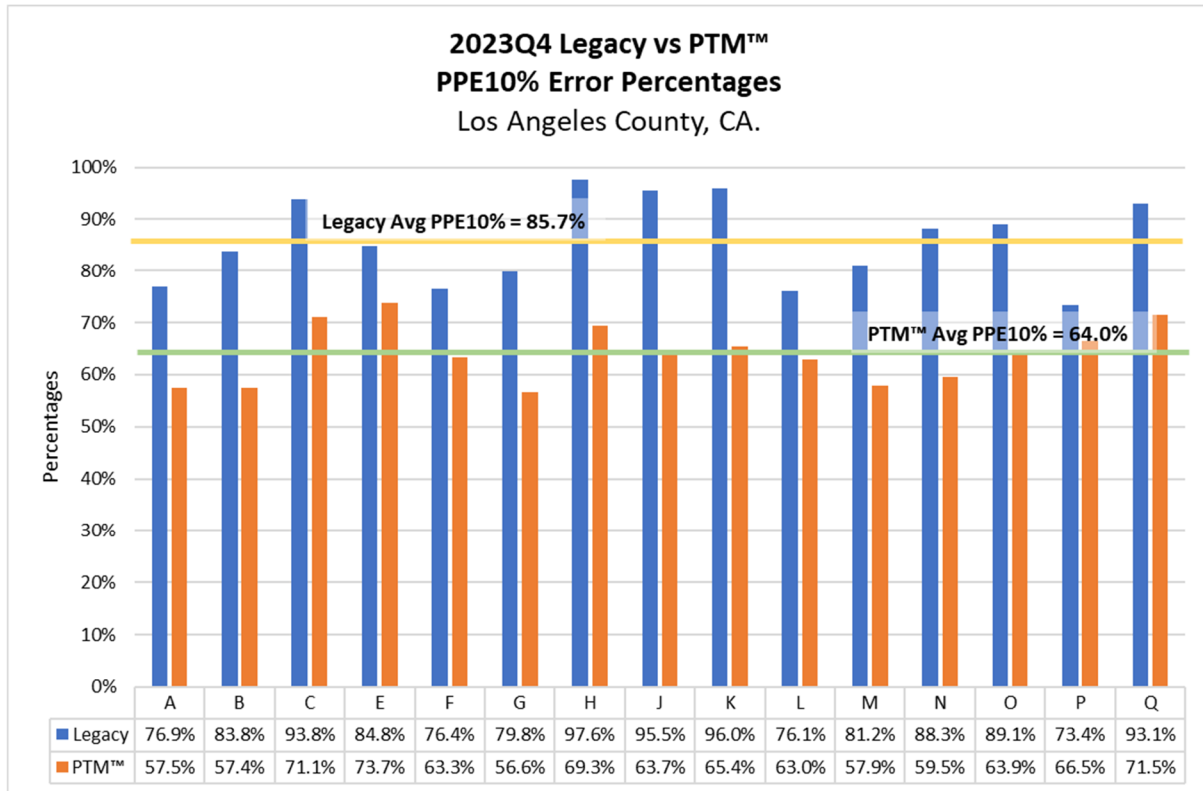


Figure 7 - Q4 Performance of each model, with and without anchoring, in Los Angeles County only.

A different perspective on the national trend analysis is available via the average mean absolute variance (MAE). Whereas a higher PPE10 is better, the opposite is true of MAE. As a result, the downward sloping trend actually represents the same underlying reality: the models improved throughout the year. Their MAE metrics, unaided by anchoring on listing prices, ranged from ~11.5% to ~13% in this National test. The high performing model(s) posted MAE scores less than 10%, declining to ~8%, which is still very respectable.

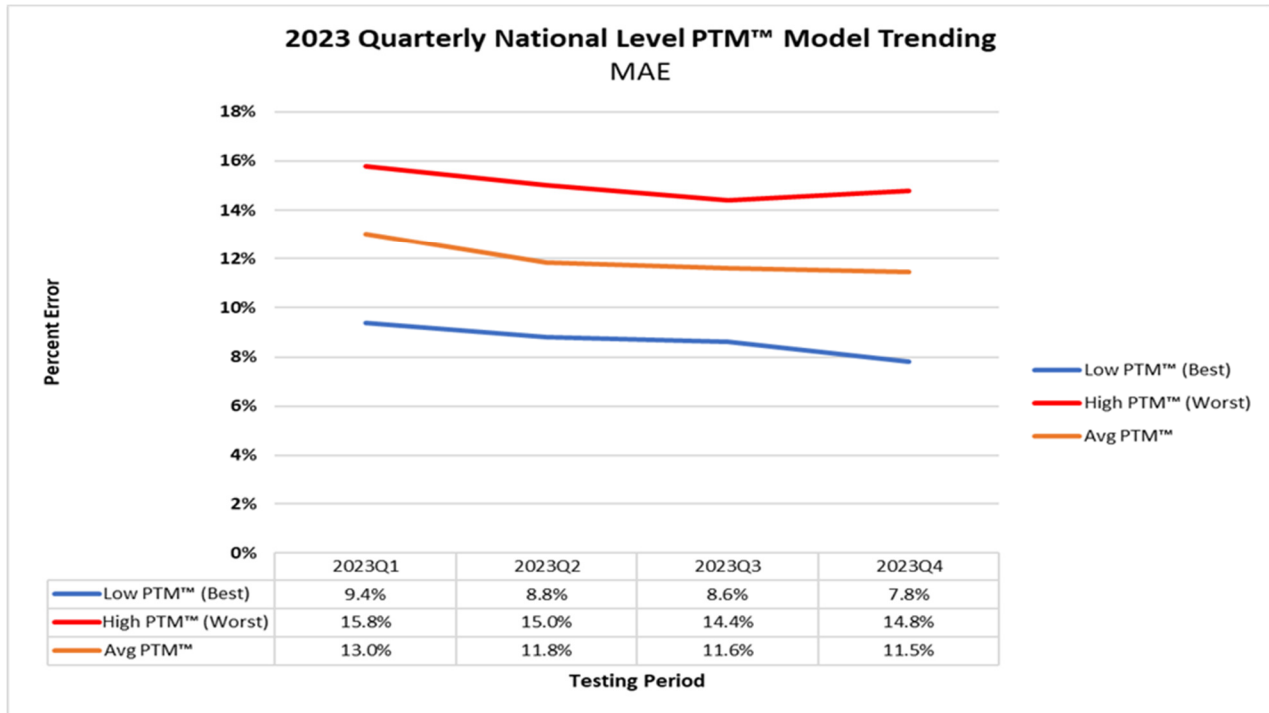


Figure 8 - 2023 overall model performance trend using MAE.

For Q4 of 2023, the nationwide results vary significantly between the models. The PTM™ analysis, using value estimates without listing price anchoring, averaged an MAE of 11.6%. With listing prices available to them, the same models averaged 7.0% MAE.

It is interesting to note that some models were able to perform better without listing prices than other models performed with them. For example, models H, K and O performed better without listing prices than models P, L and A performed with the listing prices.

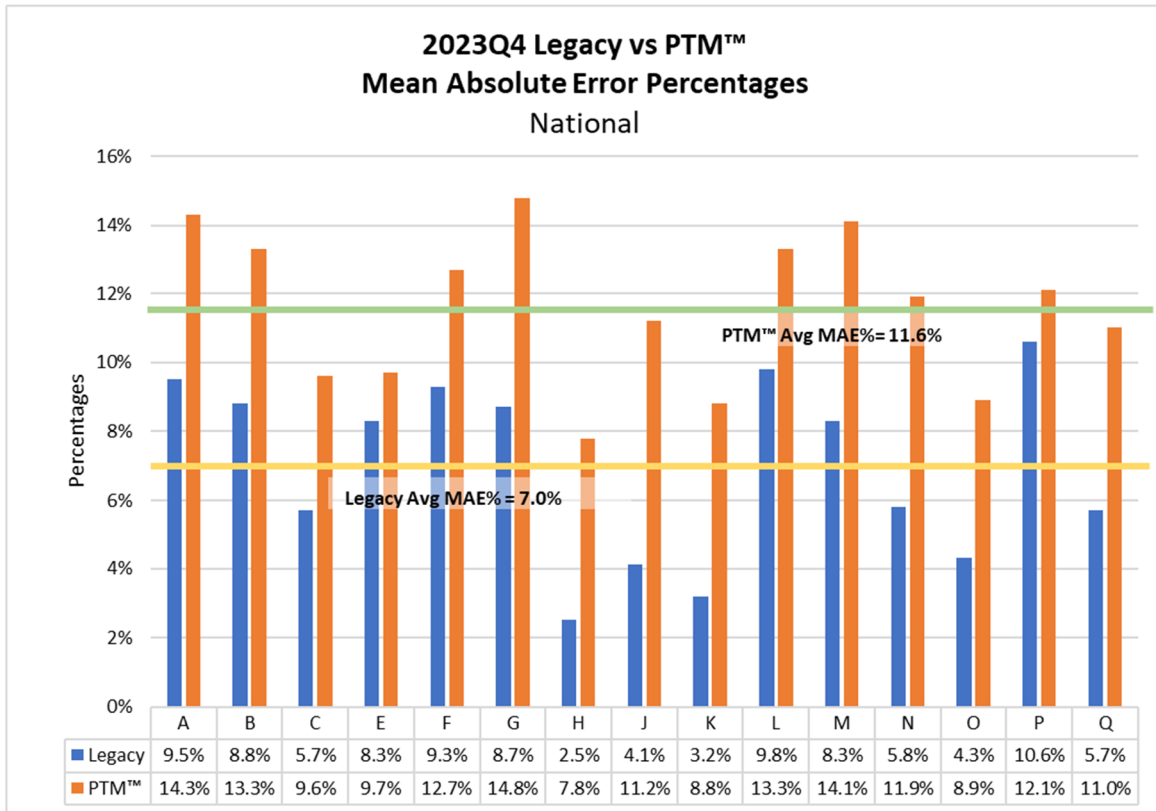


Figure 9 - Q4 Performance of each model, as measured by MAE, with and without anchoring, across all 50 states.

Figures 10 & 11 reflect the same message: Every model performed better with listing prices than without when the sample is limited to only California or only Los Angeles California.

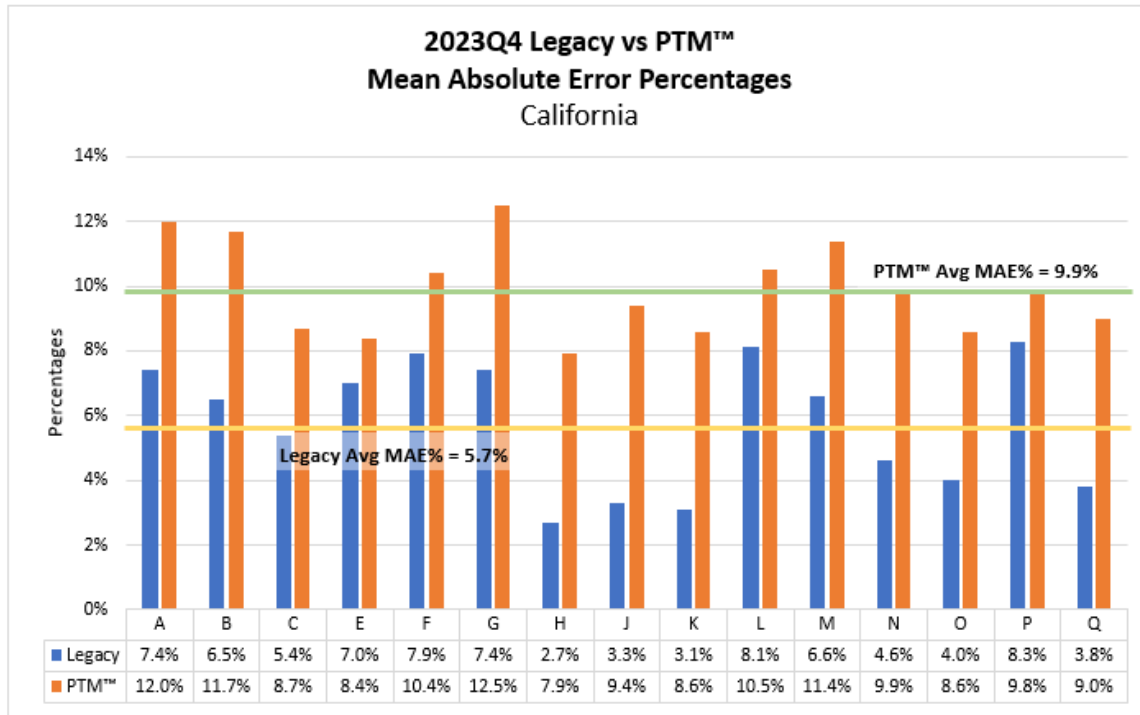


Figure 10 - Q4 Performance of each model, as measured by MAE, with and without anchoring, in California only.

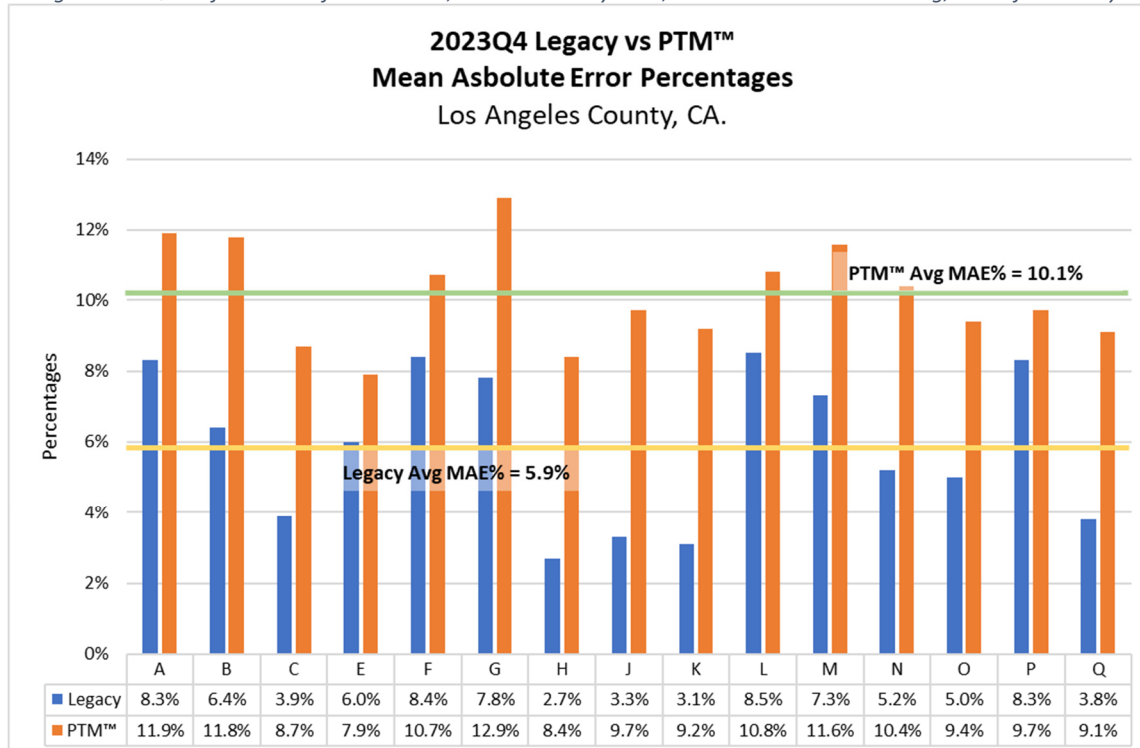


Figure 11 - Q4 Performance of each model, as measured by MAE, with and without anchoring, in Los Angeles County only.

The national level trend analysis average of the greater than +20% (overvaluation) error rate (unanchored) dropped from a high of 12.7% in Q1 to 8.4% in Q4 for the models as they and the market stabilized in 2023.

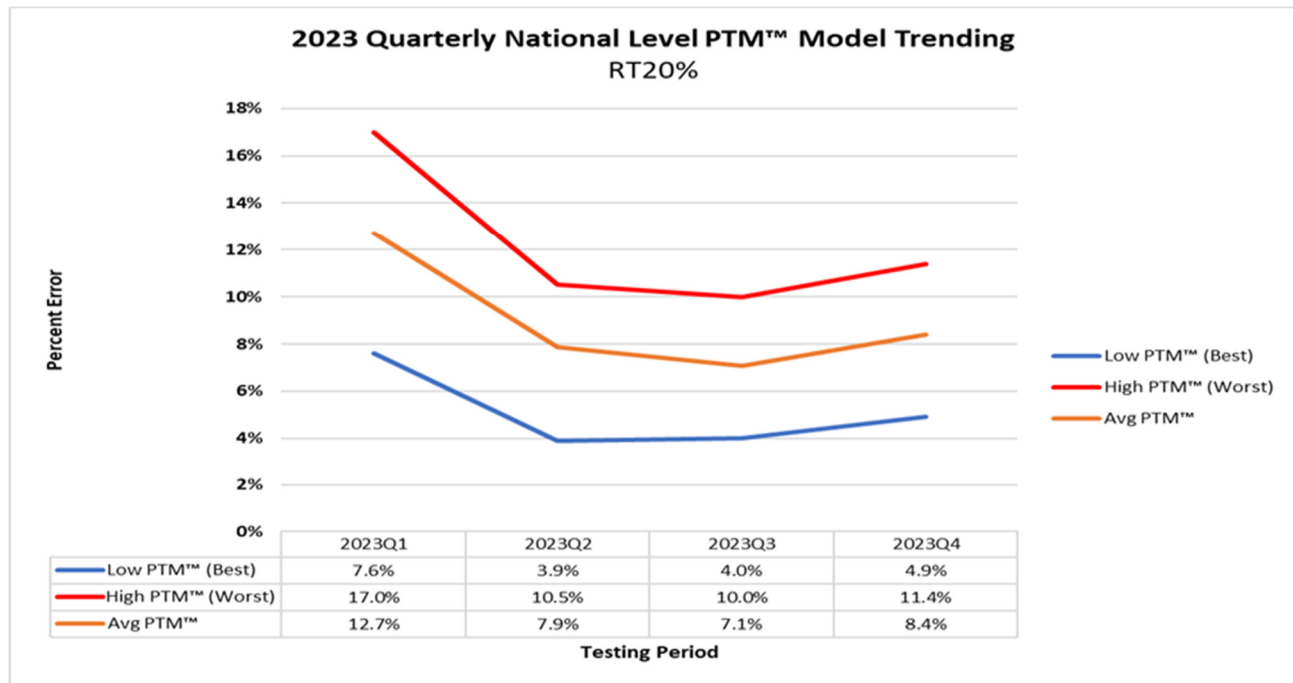


Figure 12 - 2023 overall model performance trend using RT20%.

For Q4 of 2023, as with the other metrics, the nationwide RT20% results vary significantly between the models. The PTM™ analysis, without listing price anchoring, produced an average RT20% of 8.4%. With listing prices available to them, the same models averaged 5.3% (see Fig. 13). That 3.1% improvement exemplifies how substantial the anchoring effect is.

Again, some models were able to perform better without listing prices than other models performed with them. Once again, models H, K and O performed better without listing prices than six other models performed with the listing prices.

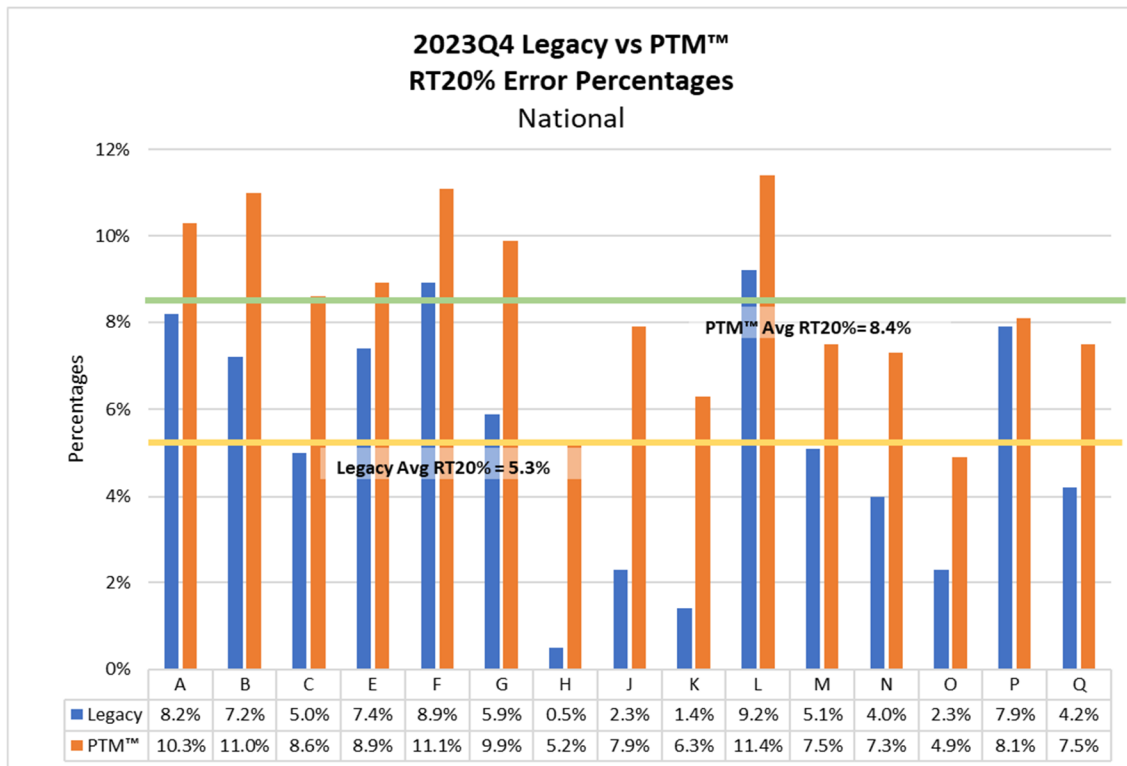


Figure 13- Q4 Performance of each model, as measured by RT20%, with and without anchoring, across all 50 states.

Figures 14 & 15 again reflect those same conclusions using data from only California and only Los Angeles County.

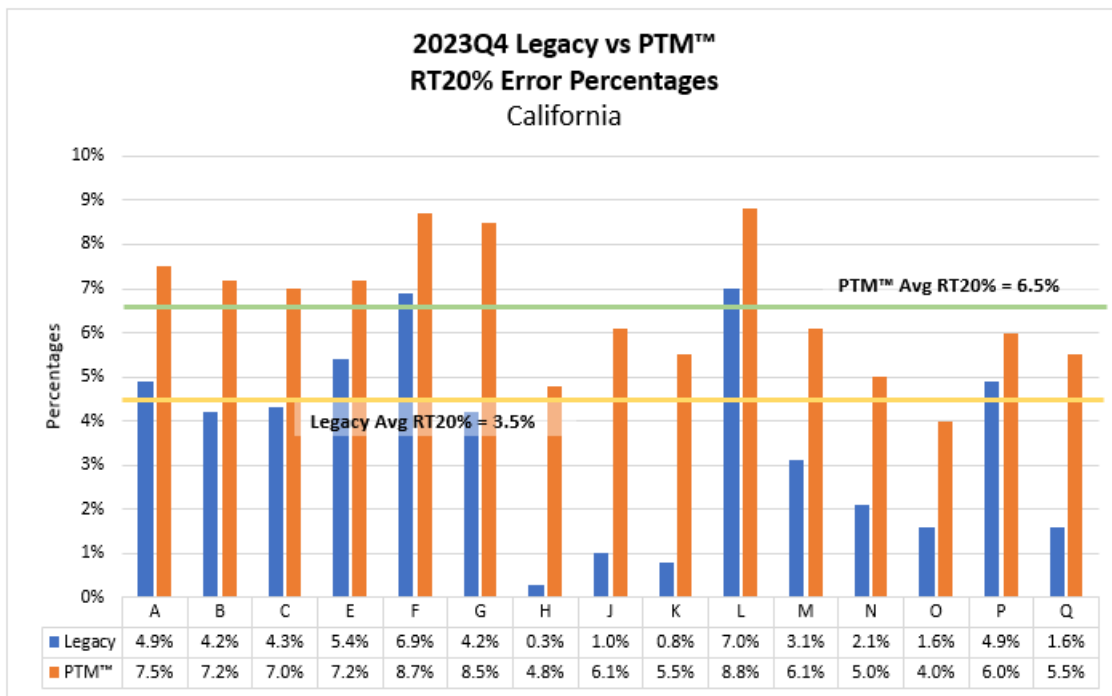


Figure 14- Q4 Performance of each model, as measured by RT20%, with and without anchoring, in California only.

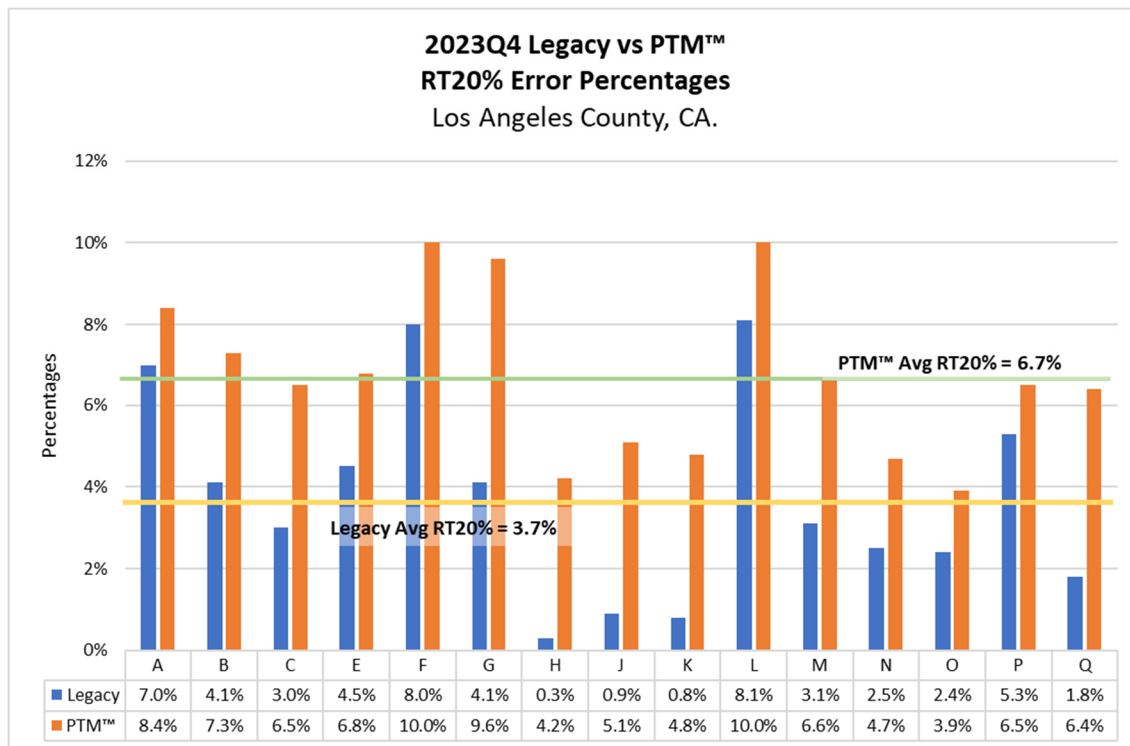


Figure 15 - Q4 Performance of each model, as measured by RT20%, with and without anchoring, in Los Angeles County only.

## Consistency of Results Across Geographies

The preceding analysis focused on California and Los Angeles County to provide the reader with the ability to observe different metrics being applied to the same geographic data set. The same analysis could be completed across every different county in the United States, and we have performed it, but the result would be an overwhelming number of charts if viewed in a similar fashion. So, we have summarized those results in the following table. It shows the difference between the Legacy PPE10% and PTM™ PPE10% (PPE10% being the most commonly referenced overall metric) for the top 25 counties, which account for 20% of the total records in our analysis.

Figure 16 below details how participating models in every county performed better when informed of listing prices (i.e., using the Legacy testing method). Note that every single percentage in the chart is positive. This demonstrates the universality of the conclusion, across all geographies and for all models: knowing the listing price doesn't just tilt the results in some places and times; it is always and everywhere a tremendous influence.

## AVMETRICS' PREDICTIVE TESTING METHODOLOGY

|             |                |         |            | Difference between Legacy and PTM average values for PPE10<br>(per Model) |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
|-------------|----------------|---------|------------|---------------------------------------------------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| State       | County         | Count   | % of Total | A                                                                         | B     | C     | E     | F     | G     | H     | J     | K     | L     | M     | N     | O     | P     | Q     |
| AZ          | Maricopa       | 14,236  | 1.88%      | 19.8%                                                                     | 18.4% | 20.0% | 10.5% | 13.1% | 22.0% | 23.7% | 30.2% | 27.9% | 13.1% | 25.7% | 28.0% | 23.1% | 7.4%  | 23.3% |
| IL          | Cook           | 13,432  | 1.78%      | 25.7%                                                                     | 31.3% | 24.8% | 12.8% | 25.4% | 34.2% | 28.4% | 35.9% | 31.7% | 25.8% | 30.5% | 38.7% | 31.6% | 11.3% | 22.1% |
| CA          | Los Angeles    | 12,617  | 1.67%      | 19.4%                                                                     | 26.4% | 22.7% | 11.1% | 13.1% | 23.2% | 28.3% | 31.8% | 30.6% | 13.1% | 23.3% | 28.8% | 25.2% | 6.9%  | 21.6% |
| NV          | Clark          | 8,367   | 1.11%      | 18.0%                                                                     | 17.5% | 14.1% | 7.1%  | 9.4%  | 21.8% | 19.0% | 20.7% | 19.9% | 9.5%  | 23.2% | 18.2% | 16.2% | 7.4%  | 16.5% |
| FL          | Broward        | 7,934   | 1.05%      | 22.0%                                                                     | 22.2% | 17.8% | 8.7%  | 21.9% | 30.9% | 26.2% | 30.1% | 27.7% | 22.1% | 34.0% | 28.6% | 24.6% | 7.7%  | 20.0% |
| FL          | Palm Beach     | 6,638   | 0.88%      | 31.3%                                                                     | 25.3% | 18.8% | 8.1%  | 24.2% | 34.4% | 24.9% | 29.4% | 27.0% | 24.4% | 34.4% | 25.3% | 21.9% | 9.2%  | 24.1% |
| FL          | Miami-Dade     | 6,559   | 0.87%      | 32.5%                                                                     | 25.9% | 15.1% | 6.2%  | 32.3% | 34.3% | 28.6% | 30.6% | 29.6% | 32.2% | 33.6% | 30.2% | 26.1% | 7.4%  | 20.4% |
| WA          | King           | 6,126   | 0.81%      | 24.8%                                                                     | 22.7% | 5.8%  | 12.9% | 16.4% | 26.8% | 25.3% | 26.8% | 26.4% | 16.4% | 24.7% | 24.9% | 22.6% | 5.7%  | 20.1% |
| CA          | San Diego      | 5,776   | 0.76%      | 32.8%                                                                     | 30.1% | 22.0% | 12.6% | 13.2% | 26.5% | 24.8% | 27.4% | 26.6% | 13.2% | 33.6% | 29.9% | 27.0% | 9.9%  | 20.8% |
| CA          | Riverside      | 5,438   | 0.72%      | 26.3%                                                                     | 24.3% | 16.9% | 9.2%  | 13.0% | 23.8% | 22.8% | 24.6% | 23.7% | 13.0% | 26.5% | 22.1% | 19.8% | 7.8%  | 16.7% |
| CA          | Orange         | 5,319   | 0.70%      | 36.6%                                                                     | 27.3% | 20.5% | 10.7% | 15.2% | 35.4% | 27.3% | 29.1% | 28.3% | 15.3% | 37.4% | 27.2% | 25.2% | 6.3%  | 20.6% |
| OH          | Cuyahoga       | 4,967   | 0.66%      | 33.2%                                                                     | 27.7% | 32.3% | 17.4% | 31.2% | 39.9% | 36.7% | 47.0% | 40.7% | 31.3% | 37.4% | 47.5% | 35.4% | 20.6% | 35.6% |
| FL          | Hillsborough   | 4,812   | 0.64%      | 19.1%                                                                     | 22.4% | 17.3% | 9.8%  | 19.2% | 25.6% | 20.5% | 25.3% | 23.3% | 19.1% | 28.5% | 18.3% | 18.0% | 7.6%  | 18.4% |
| FL          | Pinellas       | 4,659   | 0.62%      | 27.5%                                                                     | 27.8% | 24.4% | 13.7% | 27.4% | 37.9% | 27.2% | 37.2% | 29.9% | 27.5% | 36.8% | 35.2% | 29.6% | 9.8%  | 24.9% |
| FL          | Lee            | 4,514   | 0.60%      | 19.5%                                                                     | 23.4% | 19.8% | 1.5%  | 19.3% | 31.5% | 25.2% | 26.8% | 26.2% | 19.5% | 28.7% | 22.1% | 20.3% | 7.6%  | 20.9% |
| OH          | Franklin       | 4,500   | 0.59%      | 28.1%                                                                     | 22.1% | 23.8% | 14.0% | 18.8% | 32.8% | 23.8% | 32.9% | 27.1% | 19.2% | 33.3% | 37.4% | 25.3% | 15.7% | 30.3% |
| MN          | Hennepin       | 4,464   | 0.59%      | 24.4%                                                                     | 26.8% | 25.2% | 14.4% | 21.9% | 29.7% | 27.1% | 28.5% | 28.0% | 21.8% | 24.9% | 27.9% | 25.4% | 5.1%  | 14.0% |
| MI          | Wayne          | 4,447   | 0.59%      | 32.3%                                                                     | 29.7% | 26.3% | 13.8% | 26.3% | 35.6% | 27.4% | 39.5% | 30.8% | 26.7% | 35.3% | 43.0% | 29.4% | 7.2%  | 30.7% |
| MI          | Oakland        | 4,078   | 0.54%      | 30.1%                                                                     | 28.8% | 27.7% | 15.9% | 20.4% | 36.8% | 29.9% | 36.1% | 33.2% | 20.9% | 34.9% | 37.4% | 33.4% | 8.7%  | 25.5% |
| NC          | Mecklenburg    | 4,063   | 0.54%      | 12.1%                                                                     | 22.0% | 20.0% | 9.9%  | 11.5% | 28.5% | 21.2% | 25.8% | 23.9% | 12.1% | 29.5% | 22.8% | 18.3% | 8.6%  | 18.0% |
| FL          | Duval          | 3,663   | 0.48%      | 30.6%                                                                     | 22.3% | 19.5% | 10.6% | 17.0% | 30.1% | 22.2% | 31.0% | 25.6% | 17.5% | 32.6% | 29.3% | 20.9% | 11.5% | 25.3% |
| CA          | Sacramento     | 3,660   | 0.48%      | 22.6%                                                                     | 22.2% | 3.5%  | 0.1%  | 10.8% | 20.5% | 24.1% | 26.3% | 25.2% | 10.9% | 23.0% | 28.4% | 24.4% | 5.8%  | 19.3% |
| FL          | Orange         | 3,659   | 0.48%      | 28.8%                                                                     | 22.9% | 16.2% | 7.6%  | 19.4% | 33.4% | 19.1% | 21.8% | 20.4% | 19.5% | 32.7% | 12.0% | 16.6% | 6.4%  | 17.3% |
| CA          | San Bernardino | 3,640   | 0.48%      | 25.7%                                                                     | 22.2% | 16.5% | 9.6%  | 12.2% | 23.1% | 20.5% | 25.5% | 23.5% | 12.3% | 26.5% | 22.3% | 19.9% | 11.1% | 20.1% |
| MA          | Middlesex      | 3,467   | 0.46%      | 21.5%                                                                     | 23.6% | 21.2% | 9.7%  | 17.1% | 29.5% | 26.0% | 28.3% | 27.4% | 17.1% | 27.5% | 30.2% | 26.2% | 8.7%  | 30.6% |
|             |                | 151,035 | 20.0%      |                                                                           |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| Total Count |                | 2,877   | 756,505    |                                                                           |       |       |       |       |       |       |       |       |       |       |       |       |       |       |

Figure 16 Difference between Legacy and PTM™ PPE10% performance for every model in the top 25 counties in the U.S. in Q4

## The Conclusions

Current testing practices utilize available best practices to build and maintain a qualified set of transactions that satisfy benchmarking standards, including back-end filtering methodologies. And yet, the practices still allow models to be informed and influenced by listing prices, thereby skewing testing metrics and making models appear more accurate than their unanchored predictions. If AVM testing prohibited the use of MLS listing prices (as in our PTM™), the predictive accuracy of the models would be much more realistic.

Furthermore, the data shows different models are more willing and able to leverage the MLS pricing data, creating an uneven playing field.

As a result, not only are old testing methods unable to measure how accurate AVMs are in real world use cases, they are not able to effectively compare models and rank them appropriately.

[In our previous analysis](#) we restricted the testing to just California with a much smaller test data set and fewer models. As we expanded our new testing methodology with many more models and a much larger data set to a national scope, we expected to see similar results with variations based upon each model's access to MLS by geography.

This expanded assessment of the new test methodology clearly shows that it met our expectations and that PTM™ can remove the influence of MLS list price information being used by the models on a

localized and national level, thus, answering the question: “Is this methodology applicable in all geographies?” The short answer is, “Yes,” as demonstrated in Figure 16.

At AVMetrics, our mission is to support the AVM industry through effective, truly independent<sup>e</sup> model testing and validation, and we believe that this innovation in testing methodology advances that mission in two important ways. First, it levels the playing field for AVM builders, some of whom have more access to or reliance on listing prices. Second, this new testing methodology supports a higher standard of care and best practice for valuation tool(s) assessment and therefore meets/exceeds many of the objectives of the recently announced AVM Quality Control Standards.

---

<sup>a</sup> Most MLS data can be valuable to models in many ways that do not corrupt the testing process, for example, by providing updated information on property characteristics.

<sup>b</sup> Filtering was used to align the Legacy model value returns with the PTM™ value returns so that both were valuing the same properties for each model.

<sup>c</sup> Q4 results are shown, but all four quarters' analysis reveals similar relationships. Q4 has the advantage of being the freshest data. We typically rely on a quarter's worth of data, because the sample is large enough to support reliable analysis in narrow subsets, but a quarter is brief enough that it captures a meaningful snapshot of a model's performance. Averaging over the course of a year would obscure both market changes and model updates.

<sup>d</sup> AEI recently [published a study](#) with these same conclusions using a data set limited to only five models in 34 counties but extending over 6 years.

<sup>e</sup> Independent Testing/Validation is a model validation process designed and executed by parties that are separate (independent) from those that build, sell, resell or are using or running the AVM system being validated.