AVM Testing and Evaluation using AVM Performance Metrics

**Lee Kennedy**

Managing Director

AVMetrics, LLC

https://orcid.org/0000-0002-0124-4109

Lee.Kennedy@avmetrics.net

**Mark D. Ecker**

Professor of Statistics

Mathematics Department

University of Northern Iowa

http://orcid.org/0000-0003-3602-7558

mark.ecker@uni.edu

 **Hans R. Isakson**

 Professor Emeritus

 Economics Department

University of Northern Iowa

http://orcid.org/0000-0002-6914-2432

hans.isakson@uni.edu

May 14, 2021

NOTE: The authors have been retained as expert witnesses to assess AVMs in dozens of high-profile litigations. Any opinions or points of view expressed in this study represent a consensus of the authors and do not necessarily represent the official position or policies of the University of Northern Iowa or AVMetrics, LLC. Any products and manufacturers discussed in this study are presented for informational purposes only and do not constitute product approval or endorsement by the authors.

**Abstract**

Currently-used industry procedures independently test the credibility of AVM valuations, via their AVM Performance Metrics, calibrated using actual sales data. In particular, the vendor-reported Forecast Standard Deviation (“FSD”), self-generated by an AVM provider, together with the valuation for a target property, is empirically corroborated using 455,563 housing sales, each valued by as many as 15 AVMs in an AVM-by-FSD analysis. The AVM-by-FSD analysis reveals that AVM providers are empirically underreporting their vendor-reported FSDs for 267 (72.8%) of the 367 AVM/FSD combinations.

Furthermore, a composite index of any collection of AVM Performance Metrics, called the AVM Error Score (“AVM-ES”), is introduced to compare, rank and grade competing AVMs, at any desired geography and sampling unit. The AVM-ES is illustrated using two sets of housing sales data. The first expands the AVM-by-FSD analysis by using eleven AVM Performance Metrics (not just the FSD) to rank each of the 367 AVM/FSD combinations. The second analysis compares the performance of twelve AVMs for a common set of 1,193 sales in Clark County, Nevada using ten (mostly different) AVM Performance Metrics.  Lastly, given a set of AVM performance thresholds, the AVM-ES can assign a collective ordinal grade (e.g. ‘Strong’, ‘Reasonable’, ‘Acceptable but Weak’ or ‘Poorly’ performing) to classify overall AVM performance.

**Keywords:** Automated Valuation Models, AVM-by-FSD Analysis, AVM Performance Metrics, Failure Magnitude, Failure MAPE, Failure Rate, FSD, Percentage Sales Error, Principal Component Analysis

**1. Introduction**

Mortgage originators and others looking to quickly and inexpensively value a target property[[1]](#footnote-1) are increasingly turning to an Automated Valuation Model (“AVM”),[[2]](#footnote-2) instead of a traditional appraisal, to estimate the property’s market value (Ekram and Fricke, 2019). The growth in AVM use will most likely accelerate, given the Federal Deposit Insurance Corporation (“FDIC”), the Board of Governors of the Federal Reserve System, and the Office of the Comptroller of the Currency (“OCC”) joint notice, on November 21, 2018 to increase the threshold for residential real estate transactions, requiring a traditional appraisal, from $250,000 to $400,000 (FDIC, 2018).

Regulatory agencies currently require an assessment of AVM compliance with applicable federal regulations. In particular, the Consumer Finance Protection Board (“CFPB”) develops minimum standards for appraisal management companies (together with the Board of Governors of the Federal Reserve System, the OCC, the FDIC, the National Credit Union Administration (“NCUA”), and the Federal Housing Finance Agency (“FHFA”)) in support of the Dodd-Frank Wall Street Reform and Consumer Protection Act to the Financial Institutions Reform, Recovery, and Enforcement Act (“FIRREA”). See CFPB (2014). Moreover, Government Sponsored Enterprises (”GSE”s), such as the Federal Home loan Mortgage Corporation (“Freddie Mac”) and the Federal National Mortgage Association (“Fannie Mae”) have been given the mandate by the FHFA to update the lending process, including the modernization of the appraisal process (U.S. Department of the Treasury, 2018, pp. 103-106). At present, FHFA is offering waivers for about 10 percent of loans being made, where no appraisal is required, and reports that that number could increase to as much as 40 percent.[[3]](#footnote-3)

The Interagency Appraisal and Evaluation Guidelines (2010, Appendix B) require that institutions using AVMs in lieu of standard appraisals provide clarity regarding, among other things, AVM vendor selection, model development, together with *independent AVM testing* and monitoring (Long, 2010). Institutions that fail to comply with these guidelines “will be cited in supervisory letters or examination reports and may be criticized for unsafe and unsound banking practices” (Interagency Appraisal and Evaluation Guidelines, 2010, Section III). Consequentially, the need for more independent AVM testing and better statistical tools to evaluate AVM performance has never been greater. This study helps to fill this gap.

**2. Testing and Evaluating AVM Performance**

Current industry practices to test the accuracy and reliability of AVM valuations use AVM Performance Metrics, where, given a set of housing sales, these metrics assess how well a set of AVM valuations compare to their respective selling prices.[[4]](#footnote-4) AVMs typically provide the client with a self-generated Forecast Standard Deviation (“FSD”), as a measure of AVM performance (henceforth written as the “vendor-reported FSD”), with each AVM valuation. However, CoreLogic (2011, p. 4) states that the AVM user “should not rely solely on validation representations provided by the AVM vendor.” The Interagency Appraisal and Evaluation Guidelines (2010, Section III) suggest that institutions should evaluate whether vendor-reported AVM Performance Metrics (e.g. a vendor-reported FSD), “provide an appropriate indicator of model reliability by property types and geographic locations.” As a result, the first objective of this study is to detail the currently-used, industry testingof AVM valuations, by independent third-parties, to corroborate vendor-reported FSDs.

CoreLogic (2011, p. 12) states that “[e]ach vendor should be able to comment on their model’s [AVM’s] performance relative to their competitors”. Regardless of the business use of any AVM (e.g., portfolio management, portfolio marketing, mortgage underwriting, appraisal review), the client may ultimately have a collection of AVM reports, each with its own set of AVM Performance Metrics and be faced with making a comparative assessment of AVMs, to decide which AVM valuation best reflects the target property’s market value. However, Gaylor, *et al*. (2015, p. 3) state that “it is not an easy matter to reliably measure relative performance between AVMs.”[[5]](#footnote-5) As a result, the second objective of this study is to introduce statistical tools to compare, rank and grade the performance of multiple competing AVMs, via AVM Performance Metrics.

 **2a. Independent Third-Party AVM Testing**

Unfortunately, the user of an AVM cannot discern the accuracy of any AVM valuation without additional information regarding the AVM’s performance. Although AVM providers offer such information to their clients in the form of AVM Performance Metrics (e.g. a vendor-reported FSD), this information may or may not be accurate. As a result, third parties (apart from AVM providers and clients), such as AVMetrics, regularly perform *independent* *AVM testing* to evaluate the accuracy of the AVM’s valuations, against known selling prices, for hundreds of thousands of target properties in multiple markets.

AVM testing requires the choice of a sampling unit to perform model validation for any desired market segment. Most generally, one may partition housing sales to allow model testing and evaluation between or amongst different: i) types of properties/structures (e.g., residential single-family houses, planned unit developments (“PUDs”), attached condominiums); ii) housing price tiers (e.g., inexpensive, moderately priced, expensive houses); iii) geography (e.g., zip code, city, Federal Information Processing Standards (“FIPS”), Core Based Statistical Area (“CBSA”), Metropolitan Statistical Area (“MSA”), county, state, national); iv) vendor-reported FSD values and v) model demographics, such as by AVM vendor or by type of AVM (e.g., hedonic regression, appraisal emulation). For example, one might compare, rank and grade competing AVMs when i) valuing expensive custom-built houses in Maricopa County, Arizona or ii) using all appraisal emulation models to value detached single-family residential properties in California or iii) valuing all properties, nationwide, with a common vendor-reported FSD.

To draw meaningful comparisons from AVM testing, a continuous stream of high-quality housing sales data is needed (Allen, 2009). AVM testing first evaluates the quality of the data needed for the validation process, “including the data sources and types, frequency of updates, quality control performed on the data, and the sources of the data in states where public real estate sales data are not disclosed.”[[6]](#footnote-6) CoreLogic reports that they internally value 90 million residential properties on a rolling basis using each of their AVMs resulting in over 1 billion property valuations per month.[[7]](#footnote-7) However, CoreLogic (2011, p. 4.) also states that “[i]nstitutions must gather information on the performance of AVMs from someone other than the vendor”, i.e. independent, third parties. These third-parties (e.g. AVMetrics) provide AVM vendors with a set of properties whose locations (but not their selling prices) are disclosed to the vendors. The AVM then values each property and returns its AVM valuation, together with a vendor-reported FSD, to the third-party.

Comprehensive AVM testing by an independent third-party employs hundreds of thousands of housing transactions, to regularly blind test the performance (accuracy and geographic coverage) of almost every commercially available AVM.[[8]](#footnote-8) To illustrate, AVMetrics delivers housing sales to AVM providers bi-weekly to test the accuracy of the vendors’ AVMs. These AVM testing sales consist of target properties that have passed the arms-length transaction criteria[[9]](#footnote-9) and, as a result, qualify as “benchmark” properties. AVM vendors are provided only the addresses for hundreds of thousands of benchmark properties and asked to return their AVM valuation and vendor-reported FSD for each benchmark property.

A fundamental problem in AVM testing occurs when the AVM vendor already has the sales price of a benchmark property in its database. CRC (2003, p. 12) state that “There is substantial concern that the model results will be ‘steered’ [or informed] by allowing the AVM to ‘see’ the most recent sale.”[[10]](#footnote-10) Independent third-parties (e.g., AVMetrics) take numerous precautions to ensure that the AVM vendors cannot simply “look-up” the selling price from the vendor’s own sales data and then simply report that selling price as the valuation. For example, AVMetrics requires that AVM providers return the last known sales price in their database for each benchmark property, together with its date of sale.[[11]](#footnote-11) AVMetrics also requires AVM providers return an AVM valuation, together with a vendor-reported FSD for each benchmark property, within 48 hours from receipt of these test data.[[12]](#footnote-12)

 **2b. AVM Performance Metrics**

An AVM is often described as a “black box” due to the hidden and proprietary nature of its internal algorithmic mechanism or statistical procedure used to produce an AVM valuation for a target property. As a result, the internal mechanics of an AVM are difficult to evaluate. Therefore, the credibility of an AVM is typically assessed by evaluating the accuracy and precision of its valuations, using AVM Performance Metrics.

The basic and most fundamental AVM Performance Metric, for any target property, is the percentage sales error calculated as:

$Percentage Sales Error= \frac{AVM Valuation-Selling Price}{Selling Price}\*100\%$.

Most AVM Performance Metrics are derived from a set of percentage sales errors, calculated using multiple target properties. Several metrics are shown in Exhibit 1.[[13]](#footnote-13) For example, the mean and median percentage sales errors measure the accuracy of an AVM. The mean and median absolute percentage sales errors, together with the FSD, quantify an AVM’s precision, while many of the remaining metrics seen in Exhibit 1 assess a combination of an AVM’s accuracy and precision.

Exhibit 1 About Here

Reporting an AVM valuation, together with its corresponding vendor-reported FSD, to the client is ubiquitous, however, the definition and/or calculation of the FSD is not standardized across the industry.[[14]](#footnote-14) The mathematically clearest definition of an FSD is that it is the standard deviation of a set of percentage sales errors.[[15]](#footnote-15) However, standards for selecting the housing sales required to calculate an FSD do not currently exist.

The first goal in this study is to demonstrate how the vendor-reported, self-generated FSD is currently independently tested, by third parties in the AVM industry, using actual sales data. In particular, the accuracy of the vendor-reported FSD is corroborated using 455,563 properties across the United States in the third quarter (July through September) of 2018. Each of these 455,563 sales is valued by as many as 15 commercial AVMs (whose identities are blinded in this study for confidentiality reasons), resulting in a total of 5.36 million AVM valuations.

AVM Performance Metrics seen in Exhibit 1 include the Failure Rate of an AVM,[[16]](#footnote-16) which is the percentage of properties for which the AVM fails to predict selling prices accurately, to within a specified +/- factor of precision (e.g. 5, 10, 15 and 20 percent). The Failure Rate is the complement of the Percent (Predicted) Error (“PE”) bucket, a popular metric also reported in Exhibit 1. Also, for each Failure Rate (at a given +/- percent), the Failure Magnitude and Failure MAPE are the mean and median percentage sales error for all sales that were in error by more than the given +/- percentage, respectively. The Failure Rate counts (the percentage of) properties for which the AVM fails to predict selling prices accurately, to within, for example, +/- 10 percent, while the Failure Magnitude and Failure MAPE measure the mean and median absolute percentage sales error, respectively for those poorly predicted sales. These three AVM Performance Metrics assess the impact of the AVM’s extreme prediction errors or outliers.

As seen in Exhibit 1, the Right Tail 20% (“RT20”) metric measures the percentage of properties whose AVM valuation exceeds its corresponding selling price by more than 20 percent. RT20 indicates the potential overvaluation that can expose a lender to more risk than factored into an original loan-to-value ratio (typically 80 percent). An additional metric is the +/- 1 FSD criterion, which is the percent of AVM valuations within one FSD of selling prices. A frequent industry interpretation of the FSD is that one can be 68.26% confident that the true market value of a target property lies within +/- 1 FSD of the AVM valuation. However, an assumption of normality is required to make this interpretation.

AVM vendors typically provide Confidence Scores, along with FSD, and as seen in Exhibit 1, a Confidence Score also indicates the model’s confidence in its AVM valuation.  Confidence Scores can be related to FSDs (Gordon, 2005) but, unfortunately, like the FSD, their definition and use are not consistent across AVM vendors (Ecker et al. 2021). AVM users, however, have started to view the FSD as the next generation confidence scoring method in efforts to come up with a more universal self-scoring metric.

Four International Association of Assessing Officers (“IAAO”) metrics are also seen in Exhibit 1. The Coefficient of Variation (“COV”) and Coefficient of Dispersion (“COD”) measure horizonal inequity, while the Price Related Differential (“PRD”) and Price Related Bias (“PRB”) quantifies vertical inequity (IAAO, 2018, Section 7.3). Horizontal inequity measures variability within the distribution of housing sales, while vertical inequity emphasizes the comparison of properties valued in the two tails of the distribution. For a set of housing sales, the IAAO (2018) advocates calculating the COV, COD, PRB and PRD statistics using the ratio of AVM value to sales price.

These AVM Performance Metrics, especially the FSD, are the tools that assess AVM accuracy and reliability, in particular, by third parties who perform independent testing of the AVM vendor’s self-reported metrics (e.g. the vendor-reported FSD) using actual sales data. Moreover, multiple AVM Performance Metrics (more than just the FSD) can be used to compare, rank, and grade the performance of competing AVMs, as discussed in the next subsection.

#  2c. A Composite Index using AVM Performance Metrics

The second goal for this study is to introduce a composite index of AVM Performance Metrics, called the AVM Error Score (“AVM-ES”), to compare, rank, and grade the performance of competing AVMs. For a set of housing sales, distinguishing between AVMs using any single AVM Performance Metric is straightforward, for example, by identifying the AVM having its mean sales error closest to zero or choosing the AVM with the lowest Observed Standard Deviation.[[17]](#footnote-17) However, the AVM-ES evaluates AVMs using a multi-dimensional, composite index derived from any chosen collection of metrics to identify top-performing AVMs. In addition to ranking competing AVMs, the AVM-ES can grade overall AVM performance, using a set of AVM Performance Metrics, by providing an ordinal grade (e.g., ‘Strong’, ‘Reasonable’, Acceptable but Weak’ or ‘Poorly’ performing), analogous to what Freddie Mac (2019) currently offers for its FSD.[[18]](#footnote-18)

Given a set of AVM Performance Metrics, the AVM-ES quantifies the overall performance of AVM valuations and is calculated for all properties aggregated to a chosen sampling unit, using a linear combination of any collection of AVM Performance Metrics, such as those seen in Exhibit 1. The coefficients for the variables in the AVM-ES linear combination are determined through a statistical Principal Component Analysis (“PCA”).[[19]](#footnote-19) Note that the PCA, outlined in this section, is not used to compute any AVM valuation; it is used to aggregate a collection of AVM Performance Metrics to compare, rank, and grade competing AVMs.

Given a set of *p* AVM Performance Metrics, the AVM-ES, for the *k*th sampling unit (e.g. an AVM/FSD combination, where *k* = 1, 2, … 367 in the first analysis in Section 5a, or an AVM, where *k* = 1, 2, … 12 for the second in Section 5a), is the first principal component, defined as:

$$AVM­ES\_{k}=\sum\_{i=1}^{p}c\_{i}M\_{i}$$

where $M\_{i}$ is the *i*th AVM Performance Metric error value [[20]](#footnote-20) and $c\_{i}$ is the weight assigned to the *i*th metric. The AVM-ES is a linear combination of *p* AVM Performance Metrics, where the individual coefficients, called component loadings, reflect the variability observed in each variable, while simultaneously accounting for the correlation between variables. If all variables used in the PCA are positively correlated, then each coefficient, $c\_{i}$, in the AVM-ES will be positive, producing a weighted average of the individual AVM Performance Metrics (as demonstrated in the first analysis in Section 5a). Note that other data mining or machine learning techniques, such as decision trees, cluster analysis, random forests, etc., can alternatively be employed to extract patterns that differentiate between AVMs using AVM Performance Metrics. However, these machine leaning models offer little explanation of any potential cause-effect relationships between variables (Valier, 2020). In contrast, the internal coefficients in a PCA (the set of $c\_{i}$’s), offer meaningful interpretations (as demonstrated empirically in Section 5a).

The PCA coefficients, the set of $c\_{i}'s$, represent the components of the eigenvector oriented in the direction of maximum variability for the AVM Performance Metrics, and whose corresponding eigenvalue allows calculation of the percentage of total variability explained by the first principal component. If each of the *p* AVM Performance Metrics used in the AVM-ES is based upon the percentage sales errors, then the common scale allows use of the variance-covariance matrix for the PCA (as illustrated in the first analysis in Section 5a), however, if not, (as illustrated in the second analysis in Section 5a), then a PCA based upon the correlation matrix is warranted.

The AVM-ES is an inverse measure of AVM performance where the lower the AVM-ES value, the better performing the AVM. An AVM that perfectly predicts all target properties would have no error, and thus, would have the lowest AVM-ES value,[[21]](#footnote-21) while large values of the AVM-ES indicate a poorly performing AVM. A desirable feature of the AVM-ES is that its internal coefficients are derived from the dataset itself, rather than using pre-set weights. Another strength of the AVM-ES is that it allows one to compare, rank, and grade AVMs, at any chosen sampling unit and geography, using any collection of AVM Performance Metrics.

**3. Datasets for Testing and Evaluating AVMs**

We now return to the first goal in this study, to empirically test the accuracy and reliability of AVM valuations, through an ***AVM-by-FSD analysis***. This analysis tests the accuracy of the FSD self-generated by each AVM vendor, by independently corroborating the vendor-reported FSD to an Observed Standard Deviation calculated from actual sales data. An accurate FSD “enables a lender to develop an integrated credit and collateral policy”[[22]](#footnote-22). Moreover, an AVM-by-FSD analysis supports the Interagency Appraisal and Evaluation Guidelines (2010, Appendix B) by providing one element in assessing “the necessary level of due diligence on AVM vendors and their models” by identifying the AVM vendors (via AVM/FSD combinations) that most accurately report FSDs.

The testing dataset in this study consists of 455,563 nationwide sales (associated with single-family, condominium/townhouses and PUD properties), which are valued by as many as 15 commercial AVMs (whose identities are blinded in this study for confidentiality reasons) across the United States in the third quarter (July through September) of 2018. The median selling price for these 455,563 benchmark properties is $267,900 (mean = $341,641), while the maximum was capped at $2 million. The most sales occurred, as seen in Exhibit 2, in California and Florida.

Exhibit 2 About Here

The second objective of this study is to move beyond the FSD by employing multiple AVM Performance Metrics to compare, rank and grade competing AVMs using the AVM-ES. Evaluating AVM performance, via the AVM-ES, is illustrated using two empirical datasets. The first employs the original 455,563 nationwide sales in an ***Extended AVM-by-FSD analysis***, extended to include any set of AVM Performance Metrics (beyond just the Observed Standard Deviation). The purpose of the Extended AVM-by-FSD analysis is to compare, rank, and grade the performance of the 15 AVMs across the various vendor-reported FSDs. The second empirical example uses a commonly-valued subset of 1,193 benchmark properties in Clark County, Nevada, which were culled from the 455,563 nationwide sales. Twelve of the 15 AVMs were able to value all 1,193 sales. The median selling price for these 1,193 housing sales is $286,999 (mean = $311,374), while the maximum selling price was $1.215 million. The purpose of the second empirical analysis is to compare and rank the performance of the 12 AVMs to determine the top-performing AVM in this market segment.

**4. AVM Testing – An AVM-by-FSD Analysis**

An AVM-by-FSD analysis allows lenders, federal regulators, and others to provide oversight by having third parties independently test which AVMs are most accurately reporting their FSDs. One component of independent AVM testing, as performed by third-parties such as AVMetrics, is to corroborate the vendor-reported FSD’s accuracy by calculating an Observed Standard Deviation for a set of benchmark properties, each having the same vendor-reported FSD. That is, the set of benchmark properties is divided, by AVM, into groups of sales having the same vendor-reported FSD. Then, the Observed Standard Deviation, calculated for each set of benchmark properties having a common vendor-reported FSD, is compared to that common vendor-reported FSD. In other words, the sampling unit for an AVM-by-FSD analysis is the set of benchmark properties, valued by each AVM, having the same vendor-reported FSD.

In an AVM-by-FSD analysis, each AVM is not required to return a value for each benchmark property. Evaluation of AVMs with different hit rates (the percentage of sales valued by an AVM) is not uncommon in the AVM industry,[[23]](#footnote-23) because one would not expect an AVM with a vendor-reported FSD of 20 to outperform another AVM (from the same or from a different vendor) with a vendor-reported FSD of 5, regardless of whether the same set of properties are valued by each AVM. In addition, valuable information regarding AVM performance would be lost if a common set of valued benchmark properties, for each AVM, were required to conduct an AVM-by-FSD analysis.[[24]](#footnote-24)

Exhibit 3 About Here

Exhibit 4 shows an AVM-by-FSD analysis for a particular AVM (Model 1), where Model 1 valued 339,346 of the 455,563 nationwide sales for a 75.5% hit rate. As seen in Exhibit 3, Model 1 produces 19 different vendor-reported FSDs, ranging from 2 to 20 for the 339,346 benchmark properties (that Model 1 valued). For example, 40,278 of Model 1’s valuations had a vendor-reported FSD of 6 (written as “Model 1, FSD 6”). That is, the vendor owning Model 1 proffers that the FSD for each of these 40,278 benchmark properties is 6, despite the AVM provider not knowing their selling prices.[[25]](#footnote-25) Since AVMetrics knows the selling price for each of the 40,278 benchmark properties for Model 1, FSD 6, the percentage sales error for each can be calculated. The Observed Standard Deviation for these 40,278 benchmark properties for Model 1, FSD 6 is computed as the standard deviation of these 40,278 percent sales errors.[[26]](#footnote-26) The Observed Standard Deviation for Model 1, FSD 6 is 7.6. Consequentially, the AVM provider that owns Model 1 is underreporting its FSD for these 40,278 benchmark properties by 1.6 or 26.7 percent. That is, the vendor-reported FSD for Model 1 is statistically significantly inaccurate (p-value < 0.0001).[[27]](#footnote-27) The AVM-by-FSD analysis in Exhibit 3 indicates that no vendor-reported FSD from Model 1 is corroborated as accurate. Model 1 is statistically significantly underreporting its FSD, by at least 20 percent, for each of its 19 different vendor-reported FSDs (p-value for each < 0.001).

Exhibit 4 About Here

Overall, Model 1 is underreporting its vendor-reported FSDs by 57.7%, on average, and the aggregate (weighted average) Observed Standard Deviation for Model 1 is higher by almost 3, when compared to its corresponding vendor-reported FSD. The AVM-by-FSD analysis reveals that Model 1 is most accurate[[28]](#footnote-28) at vendor-reported FSDs of 6 and 7 (1.6 underreporting error for each) and least accurate at vendor-reported FSDs of 2 (245.0% underreporting error) and 19 (underreporting by 8.9).

In the complete AVM-by-FSD analysis, each of the 15 AVMs has a similar set of results as that seen for Model 1 in Exhibit 3. Specifically, the vendor associated with each of these 15 AVMs has returned valuations with as many 25 different vendor-reported FSD values, resulting in a total of 367 AVM/FSD combinations. The complete AVM-by-FSD analysis independently tests the accuracy of the vendor-reported FSD for each of the 15 AVMs across all 367 AVM/FSD combinations. The AVM-by-FSD analysis reveals that the vendor-reported FSD is corroborated as accurate for only 27 (7.4%) of the 367 AVM/FSD combinations. Moreover, AVM providers have statistically significantly (α = 0.05) underreported their FSDs for 267 of the 367 (72.8%) AVM/FSD combinations.[[29]](#footnote-29) As seen in Exhibit 4, after aggregating the results over the vendor-reported FSDs for each AVM, 13 of the 15 AVMs have underreported their FSDs by as much as 6.7, while the average FSD underreporting error is 2.2.

Exhibit 4 About Here

# 5. AVM Evaluation – Comparing, Ranking, and Grading

The AVM-by-FSD analysis in Section 4 tests AVMs for accuracy of their vendor-reported FSDs against actual sales data. The Observed Standard Deviation measures only one component (precision) of the distribution of percentage sales errors, while the AVM Performance Metrics presented in Exhibit 1 assess accuracy, precision, or their combination (accuracy at a given level of precision). A second objective of this study is to introduce and demonstrate statistical methodology to determine which AVM performs best, by comparing and ranking AVMs using any set of AVM Performance Metrics detailed in Exhibit 1 (more than just an FSD).

 **5a. AVM Ranking**

The first empirical analysis in this section employs the AVM-ES, a composite index of AVM Performance Metrics, to compare, rank, and grade the performance of the 15 AVMs (over the 367 AVM/FSD combinations) in an ***Expanded AVM-by-FSD analysis***. In particular, the Expanded AVM-by-FSD analysis uses eleven AVM Performance Metrics (not just the FSD) to rank each of the AVM/FSD combinations from best performing to worst. The second empirical analysis uses the AVM-ES to evaluate twelve commercial AVMs using a commonly-valued set of 1,193 housing sales in Clark County, Nevada. These 1,193 benchmark properties, which were culled from the 455,536 nationwide sales in the first analysis, employ ten AVM Performance Metrics to identify the top-performing AVM in this market segment.[[30]](#footnote-30)

 **5a i. Expanded AVM-by-FSD Analysis**

The Expanded AVM-by-FSD analysis employs eleven metrics, as seen in Exhibit 5, to rank competing AVMs. Specifically, the AVM-ES is employed to rank the 367 AVM/FSD combinations from 1 (best performance) to 367 (worst performance). The Expanded AVM-by-FSD analysis can also identify the vendor-reported FSD value at which each AVM performs best and illuminates which AVM performs best for all property sales with a given vendor-reported FSD.

Because all eleven AVM Performance Metrics in the Expanded AVM-by-FSD analysis are based upon the percentage sales error, the common scale allows the AVM-ES to use the variance-covariance matrix, instead of the correlation matrix, in the PCA. In other words, because a one-unit change for any of these eleven metrics has the exact same numeric interpretation, then the metrics having the most variability allow the AVM-ES to compare, rank and grade AVMs. In general, the less variability, the lower the weight (coefficient) that the marginal metric receives in the AVM-ES. For the Expanded AVM-by-FSD analysis in this section, the AVM-ES explains 83.5 percent of the total variability in the ten AVM Performance Metrics.

Exhibit 5 About Here

Exhibit 5 also shows the standard deviations for the eleven metrics used in the Expanded AVM-by-FSD analysis,[[31]](#footnote-31) together with the normalized weight assigned to each metric by the AVM-ES. The four Failure Rates (FR5, FR10, FR15 and FR20, respectively) have the highest variability and, as a result, receive the most weight in the AVM-ES. The median percentage sales error receives the smallest normalized weight (2.8%). Note that each weight is not directly proportional to the standard deviation, because the coefficient in the AVM-ES is a function of the correlations between the eleven metrics (in addition to the variability of each metric). To illustrate, the median percentage sales error and the MAPE have similar standard deviations (0.08 and 0.09, respectively), however the MAPE receives a 5.3 percent normalized AVM-ES weight, which is nearly twice as large as the weight of 2.8 percent assigned by the AVM-ES to the median percentage sales error. Lastly, the AVM-ES weights in Exhibit 5 are data dependent, meaning that any change to the dataset (the 455,563 sales), the sampling unit (AVM-by-FSDs) or the set of AVM Performance Metrics, would produce a different set of weights.

Exhibit 6 shows the values of the AVM Performance Metrics for the five top-performing AVM/FSD combinations in the Expanded AVM-by-FSD analysis of the 367 AVM/FSD combinations. All five come from the same AVM: Model 6. The best performing AVM/FSD combination, according to the AVM-ES, is Model 6, FSD 4; its AVM-ES value is 0.1247. The worst performing is Model 13, FSD 36; its AVM-ES value is 2.651 (primarily because it poorly valued only two properties). It is noteworthy that Exhibit 6 shows that Model 6, FSD 3 is top-performing for eight of the eleven marginal metrics (colored in green in Exhibit 6), yet it finishes fourth best overall, due to its performing relatively poorly (compared to the other top AVM/FSD combinations) for two metrics (Mean Percentage Sales Error and the +/- 1 FSD metrics).

Exhibit 6 About Here

A strength of the Expanded AVM-by-FSD analysis is that it can evaluate the credibility of the valuations from any one AVM.  For example, Exhibit 7 shows the AVM-ES values for the 18 different vendor-reported FSDs for only Model 15. Model 15 performs best at a vendor-reported FSD of 7 (AVM-ES = 0.2326 for Model 15, FSD 7), while, not surprising, Model 15, FSD 20 performs worst (AVM-ES = 1.2411). The AVM-ES results for Model 15 demonstrate relative consistency of the FSDs, whereby the lower vendor-reported FSDs tend to perform better than higher ones, however, vendor-reported FSDs of 7, 8 and 9 from Model 15 outperform its vendor-reported FSDs of 3, 4, 5 and 6.

Exhibit 7 About Here

Lastly, the AVM-ES can illuminate which AVM performs best for all property sales with a given vendor-reported FSD. To illustrate, Model 1 is the top-performing of all 15 AVMs having a vendor-reported FSD of 6, as it has the lowest AVM-ES value (0.1290 for Model 1, FSD 6, see Exhibit 6). Model 12 is worst performing (AVM-ES = 0.6582 for Model 12, FSD 6) of the 15 AVMs valuing benchmark properties with a vendor-reported FSD of 6.

 **5a ii. Clark County Analysis**

The second empirical analysis directly compares and ranks twelve AVMs, using ten AVM Performance Metrics for a commonly-valued set of 1,193 sales from Clark County, Nevada. That is, each AVM valued each of the 1,193 benchmark properties, which were culled from the 455,563 sales used in the AVM-by-FSD analysis. The ten AVM Performance Metrics used in this analysis can be seen in Exhibit 8, while Exhibit 9 shows the individual values of these metrics.

Exhibits 8 and 9 About Here

The objective of this second analysis is to identify the top-performing AVMs in this market. Marginally, Models 2 and 7 are the most accurate, having their mean percentage errors closest to zero. However, Model 7 has the third highest Failure Rate and highest values of the Failure Magnitude and Failure MAPE, respectively, suggesting that Model 7 is performing well on average, but apparently, at the cost of many extreme predictions or outliers. Model 2 has the lowest Failure Rate, but the second highest Failure Magnitude, which indicates that Model 2 also provides several poor predictions, but far fewer than Model 7

As seen in Exhibit 8, the standard deviations for three of the IAAO metrics (COV, PRD, PRB) used in the Clark County analysis are orders of magnitude smaller than the remaining seven metrics. That is, as seen in Exhibit 9, the PRD value for the twelve AVMs only ranges from 0.990 to 1.002. As a result, the AVM-ES requires the use of the correlation matrix in the Clark County analysis when using these ten AVM Performance Metrics. For the Clark County analysis, the AVM-ES explains 61.3 percent of the total variability in the ten metrics.

Looking at the coefficients in the AVM-ES for the Clark County analysis in Exhibit 8, seven of the ten metrics each receive similar weights, between 0.316 and 0.397, with the COD and Failure Rate being the most important. In contrast, the Failure Magnitude, Failure MAPE and the PRD are the least important metrics, each receiving the lowest weights, with the Failure Magnitude and PRD being assigned negative weights. These negative weights act as a counterbalance to the remaining AVM Performance Metrics. In particular, the negative Failure Magnitude coefficient arises because the models with the five lowest Failure Rates have four of the five highest Failure Magnitudes. The combined contribution[[32]](#footnote-32) to the AVM-ES for these three metrics (Failure Rate, Failure Magnitude, Failure MAPE), each measuring a different facet of outlier performance of an AVM, shows (in the Outlier Contribution column in Exhibit 10) that Models 2 and 11 each substantially outperform the remaining AVMs. In terms of outliers, however, Model 2 slightly edges Model 11. The AVM-ES facilitates a direct comparison of two similarly performing models. Model 2 has produced fewer outliers (lower Failure Rate), but they are slightly more severe (higher Failure Magnitude and Failure MAPE) than Model 11.

Exhibit 10 About Here

As seen in Exhibit 8, the PRD receives a negative weight in the AVM-ES, due to it being inversely correlated with all other metrics, except for the PRB coefficient. The contribution to the AVM-ES from the PRD, as seen in Exhibit 10, while also negative, only differs in the third decimal place for the twelve AVMs. That is, the PRD’s lack of variability provides little comparative power to differentiate between AVMs, and together with the PRD values being very close to one, indicates that these twelve AVMs do not suffer from vertical inequity (see Ecker et al., 2021). The twelve AVMs demonstrate little error and little variability, in terms of the PRD.

The overall goal of this Clark County analysis is to provide a ranking, using the AVM-ES, of the twelve AVMs to identify the top-performing AVM when valuing the 1,193 sales. As seen in Exhibit 10, Model 11 is the top-performing AVM because it received the lowest AVM-ES value (3.88) and slightly outperforms Model 2 (AVM-ES for Model 2 is 4.22). As seen in Exhibit 9, Model 11 is more precise than Model 2 (Model 11 has a lower Observed Standard Deviation and MAPE), but Model 2 is more accurate (Model 2 has its mean percentage sales error closer to zero). Overall, Model 11 outperforms Model 2 on eight of the ten individual AVM Performance Metrics, while Model 2 outperforms Model 11 in terms of outliers (the combined contribution of the Failure Rate, Failure Magnitude and Failure MAPE, seen in Exhibit 10 and discussed above). However, the AVM-ES incorporates the magnitude of each marginal metric, when comparing Models 2 and 11. Lastly, Models 8 and 9 perform worst, as they have the two highest values of the AVM-ES and the highest values for each of the following marginal metrics: the mean percentage sales error (in absolute value), MAPE, FSD, Failure Rate, COV and COD.

Lastly, we show how the Observed Standard Deviation, together with additional AVM Performance Metrics, are used to corroborate the *collective* vendor-reported FSDs in a particular housing market. Exhibit 11 shows how AVM performance is assessed for 12 AVMs, by vendor-reported FSD, using the 1,193 sold houses in Clark County, Nevada. In total, 14,316 AVM valuations (12 times 1,193), each have a vendor-reported FSD ranging from 1 to 36 (depicted in Exhibit 11 as 1-21+). The 14,316 valuations are aggregated by vendor-reported FSD, and an Observed Standard Deviation[[33]](#footnote-33) is calculated for all sales having a common vendor-reported FSD.

Exhibit 11 About Here

As reported in Exhibit 11, the vendor-reported FSD values of 8, 9, and 11 are statistically indistinguishable from their respective Observed Standard Deviations. In other words, using a high degree of statistical confidence (α = 0.01), only the vendor-reported FSDs of 8, 9 and 11 are deemed to be statistically accurate. In addition, the lower vendor-reported FSDs, ranging from 1-7, are statistically significantly understated, when compared to their respective Observed Standard Deviations. Note that the majority (10,320 or 72.1%) of the 14,316 valuations in Clark County have vendor-reported FSDs ranging between 1 and 7. In addition, the PE10 metric, which measures the percentage of AVM valuations within 10 percent of selling prices, ranges from 85 to 100 percent for all AVM valuations with vendor-reported FSD between 1 and 7. Also, the RT20, which measures the percentage of AVM valuations exceeding selling prices by 20 percent or more, ranges from 0.55 to 1.18, for these same vendor-reported FSDs.

In contrast, the larger vendor-reported FSDs, ranging from 8-21+, have higher Observed Standard Deviations, and are, for the most part, statistically significantly overstated as reported in Exhibit 11. Also, the lower vendor-reported FSDs have lower PE10% percentages than that for higher vendor-reported FSDs. Lastly, the RT20 metric ranged from 0.88 to 3.24 percent for the higher vendor-reported FSDs, which tend to be higher than that for the lower vendor-reported FSDs.

Exhibit 11 indicates that a majority of AVM valuations (72.1%) for the Clark County data have a corresponding vendor reported FSD (ranging from 1 to 7) that is statistically significantly understated. The lower vendor-reported FSDs, nevertheless, outperform the vendor-reported FSDs that are higher than 7. The valuations having vendor-reported FSDs ranging from 1 to 7 are just not as accurate as the AVM vendor had originally advertised.

 **5b. AVM Grading**

The Interagency Appraisal and Evaluation Guidelines (2010, Appendix B) require selection of “acceptable minimum AVM performance criteria prior to and independent of the validation process” as one of its six tenets for model selection. Moreover, CoreLogic “recommends that [AVM] accuracy thresholds be set prior to testing”.[[34]](#footnote-34) As discussed in Ecker, Isakson, and Kennedy (2021), AVM performance criteria or accuracy thresholds suggested in the literature are mostly found in self-published books, industry websites, unpublished manuscripts, or recent trade publications.

Exhibit 12 proffers AVM performance criteria or accuracy thresholds for AVM Performance Metrics, where many of which were taken from the above sources.[[35]](#footnote-35) However, the end user should adopt a set of AVM performance thresholds, such as those seen in Exhibit 12, that best reflect their intended use of the valuations from AVMs, as suggested by the Interagency Appraisal and Evaluation Guidelines (2010, Appendix B).

Exhibit 12 About Here

The Expanded AVM-by-FSD analysis in Section 5a ranks the 367 AVM/FSD combinations from top performing to worst using multiple AVM Performance Metrics. In contrast, one can collectively grade how well any AVM (at a given vendor-reported FSD, for this example) is meeting a set of individual performance thresholds, such as those seen in Exhibit 12. Moreover, the AVM can be assigned one overall grade, instead of, for example, providing a checklist of marginal grades. That is, the 40,278 housing sales from Model 1, FSD 6 can be assigned a checklist of grades for each individual AVM Performance Metric, measured marginally against the respective AVM performance thresholds shown in Exhibit 12. Model 1, FSD 6 receives five ‘Strong’ performance grades, together with five ‘Reasonable’ grades. It does not receive any ‘Acceptable but Weak’ nor ‘Poor’ grades. However, one might be interested in one overall, collective grade for all 40,278 sales from Model 1, FSD 6.

Given a set of AVM performance thresholds, the AVM-ES can merge a checklist of marginal grades to assign a collective ordinal grade (e.g. ‘Strong’, ‘Reasonable’, ‘Acceptable but Weak’ or ‘Poorly’ performing) to qualify AVM performance. Following a golf analogy, par can be established for the AVM-ES using specific values for each set (column) of individual AVM performance thresholds observed in Exhibit 12. For ‘Strong Performance’, the AVM-ES par score is 0.2752. For ‘Reasonable Performance’, the AVM-ES par is 0.6411, while for ‘Acceptable but Weak Performance’ the AVM-ES par is 0.9444. The interpretation of these par scores is that a set of properties with an associated AVM-ES value below 0.2752 indicates a strongly performing AVM, while an AVM-ES value between 0.2752 and 0.6411 reflects a reasonably performing AVM. If the AVM-ES value is between 0.6411 and 0.9444, then the AVM is acceptable, but weakly performing and an AVM-ES value above 0.9444 suggests poor AVM performance. As a result, each of the 367 AVM/FSD combinations in the Expanded AVM-by-FSD analysis can receive an overall performance grade, according to the pre-set AVM performance thresholds provided in Exhibit 12, in contrast to each of the 367 AVM/FSD combinations receiving a checklist of eleven marginal grades. To illustrate, the 40,278 sales from Model 1, FSD 6 has an AVM-ES value of 0.3606, indicating a collective grade of ‘Reasonable Performance’.

Returning to Model 15, the color coding seen in Exhibit 7 (Green, Light Green, Yellow and Light Red) reflects ‘Strong Performance’, ‘Reasonable Performance’, ‘Acceptable but Weak Performance’, and ‘Poor Performance’ grading, respectively, for its 394,952 valuations (housing sales). Colored in green, valuations from Model 15 with FSD’s 7 and 8 have an AVM-ES below 0.2752, indicating strong AVM performance. Model 15 has 51.1 percent of its valuations having an FSD labeled as Strongly Performing. Light green colored valuations meet the reasonable AVM performance criteria, respectably, but not the strong performance criteria. Model 15 has 33.9 percent of its valuations with an FSD labeled as Reasonable. FSDs colored in yellow suggest acceptable but weak AVM performance. Model 15 has 8.7 percent of its valuations associated with and FSD that is labeled as Acceptable but Weak. Lastly, FSDs colored in light red do not meet any set of AVM performance criteria seen in Exhibit 6. Model 15, having an FSD of 15 or higher, has 6.3 percent of its valuations with an FSD labeled as Poorly Performing. Consequentially, an example of the utility offered by the AVM-ES, is a recommendation for Model 15 that it performs best when providing the client with a vendor-reported FSD of 7 or 8, and it performs poorly when the vendor-reported FSD is 15 or higher.

**6. Discussion and Conclusions**

This study shows how AVMs are calibrated, by detailing the currently-used, industry testingof AVM valuations by independent third-parties. An AVM-by-FSD analysis is performed to test the accuracy of the FSD, reported by the AVM vendor along with the value estimate for a target property. An AVM-by-FSD analysis allows federal regulators to assess the credibility of an AVM using actual sales data, as recommended by the Interagency Appraisal and Evaluation Guidelines (2010, Appendix B). The AVM-by-FSD analysis of 455,563 housing sales across the United States in the third quarter of 2018 reveals that the vendor-reported FSD is corroborated as accurate for only 27 (7.4%) of the 367 AVM/FSD combinations. In other words, 340 of the 367 (92.6%) AVM/FSD combinations have their vendor-reported FSDs being statistically significantly different from the Observed Standard Deviation. The observation that most AVM vendor-reported FSDs are inaccurate when empirically tested, only reinforces the lack of uniformity, in both the FSDs definition and its calculation standards, across the AVM industry. Moreover, of the 340 AVM/FSD combinations that are statistically significantly different, 267 (78.5%) are underreported, which gives the appearance that AVM valuations are more precise than they really are.

Because the FSD only measures one component (precision) of a set of sales errors, additional metrics should be employed to assess overall AVM performance. As such, a composite index, the AVM Error Score, called the AVM-ES, is constructed to compare, rank and grade AVMs. The AVM-ES is a data-driven, linear combination of any collection of AVM Performance Metrics, allowing the user to choose the set of metrics that make the most sense for their individual application.

The use of the AVM-ES is first empirically demonstrated in the Expanded AVM-by-FSD analysis, which ranks 367 different AVM/FSD combinations using 455,563 properties and eleven AVM Performance Metrics. The Expanded AVM-by-FSD analysis also identifies the vendor-reported FSD value at which each AVM performs best and illuminates which AVM performs best for all property sales with a given FSD. A second empirical example, illustrating the AVM-ES, directly compares and ranks twelve AVMs by valuing 1,193 sales in Clark County, Nevada using ten metrics. These two analyses demonstrate how the AVM-ES can combine any desired set of AVM Performance Metrics into a single composite index to rank the performance of competing AVMs, when applied to housing sales at any chosen geography and/or property type. Such ranking allows clients employing AVMs (e.g., lenders, real-estate brokers, investment bankers and consumers) to better support appraisal quality control, mortgage underwriting, including home equity loans and refinancing decisions, together with loss mitigation and credit risk management activities for financial institutions.

Lastly, given a set of performance thresholds for AVM Performance Metrics, a collective ordinal AVM classification or grade, by, for example, applying labels: ‘Strong’, ‘Reasonable’, ‘Acceptable but Weak’ or ‘Poorly’ performing to each AVM/FSD combination, is demonstrated using the AVM-ES in the expanded AVM-by-FSD analysis. Such assessment allows the user to qualify or grade the overall performance of the AVM, given a set of performance thresholds, to better manage each client’s specific risk tolerances and be in concordance with the policies and procedures recommended by the Interagency Appraisal and Evaluation Guidelines (2010).

**7. References**

Allen, S. (2009). Innovation in AVM Testing. Whitepaper. Found at

[www.corelogic.com/downloadable-docs/freddie-mac-home-value-suite/asset\_upload\_file698\_27248.pdf](http://www.corelogic.com/downloadable-docs/freddie-mac-home-value-suite/asset_upload_file698_27248.pdf) .

AVMetrics. (2018). Automated Valuation Model (AVM) Tests. 58 pages.

AVMetrics. (2021). AVMetrics Homepage. Found at [www.avmetrics.net/](http://www.avmetrics.net/) .

Consumer Finance Protection Board (CFPB). (2014). CFR Citation: 12 CFR 34; 12 CFR 1026.

Found at: [reginfo.gov/public/do/eAgendaViewRule?pubId=201410&RIN=3170-AA44](https://reginfo.gov/public/do/eAgendaViewRule?pubId=201410&RIN=3170-AA44) .

CoreLogic. (2011). Automated Valuation Model Testing. Whitepaper. Found at

<https://www.corelogic.com/downloadable-docs/automated-valuation-model-testing.pdf> .

CoreLogic. (2014). AVM FAQs. Found at:

 <https://www.corelogic.com/downloadable-docs/avm-faqs.pdf> .

CRC. (2003). The CRC Guide to Automated Valuation Model (AVM) Performance Testing.

Available at [https://docplayer.net/8833631-The-crc-guide-to-automated-valuation-model-avm-performance-TESTing-preliminary-report-of-the-avm-performance-TESTing-subcommittee.html](https://docplayer.net/8833631-The-crc-guide-to-automated-valuation-model-avm-performance-testing-preliminary-report-of-the-avm-performance-testing-subcommittee.html) .

Ecker, M. D., Isakson, H. R. and L. Kennedy. (2021). An Exposition of AVM Performance

Metrics. *Journal of Real Estate Practice and Education*. **22** (1), pp 22-39.

DOI: <https://doi.org/10.1080/15214842.2020.1757352> .

Ekram, L. and W. Fricke. (2019). Moody’s – Automated valuation models for home

appraisals a mixed bag for RMBS credit quality. Found at: [www.moodys.com/research/Moodys-Automated-valuation-models-for-home-appraisals-a-mixed-bag--PBS\_1182443?WT.mc\_id=AM%7eRmluYW56ZW4ubmV0X1JTQl9SYXRpbmdzX05ld3NfTm9fVHJhbnNsYXRpb25z%7e20190624\_PBS\_1182443](http://www.moodys.com/research/Moodys-Automated-valuation-models-for-home-appraisals-a-mixed-bag--PBS_1182443?WT.mc_id=AM%7eRmluYW56ZW4ubmV0X1JTQl9SYXRpbmdzX05ld3NfTm9fVHJhbnNsYXRpb25z%7e20190624_PBS_1182443) .

## Federal Deposit Insurance Corporation (FDIC). (2018). Appraisal Threshold for Residential Real Estate Loans. Found at <https://www.fdic.gov/news/news/financial/2018/fil18076.html> .

Federal Housing Finance Agency (FHFA) Office of Inspector General. (2018). An Overview of Enterprise Appraisal Waivers. Found at <https://www.fhfaoig.gov/sites/default/files/WPR-2018-006.pdf> .

Freddie Mac. (2019). Single-Family Business, Forecast Standard Deviation. Found at

[www.freddiemac.com/hve/fsd.html](http://www.freddiemac.com/hve/fsd.html) .

Gayler, R., Sanyal, D., Pugh, R. and S. King. (2015). Best Practice Validation and Comparison

for Automated Valuation Models (AVMs). 22 pages. Found at [www.corelogic.com.au/sites/default/files/2018-03/20151028-CL-RP\_AVM.pdf](http://www.corelogic.com.au/sites/default/files/2018-03/20151028-CL-RP_AVM.pdf) .

Gordon, D. (2005). Metrics Matter. Found at

[www.freddiemac.com/hve/pdf/dougwhitepaper\_metricsmatter.pdf](http://www.freddiemac.com/hve/pdf/dougwhitepaper_metricsmatter.pdf) .

International Association of Assessing Officers (IAAO). (2013). *Standard on Ratio Studies*, Kansas City, MO. 63 pages. Found at

 [www.iaao.org/media/standards/Standard\_on\_Ratio\_Studies.pdf](http://www.iaao.org/media/standards/Standard_on_Ratio_Studies.pdf) .

International Association of Assessing Officers (IAAO). (2018). *Standard on Automated*

 *Valuation Models (AVMs)*, Kansas City, MO. 35 pages. [www.iaao.org/media/standards/AVM\_STANDARD\_2018.pdf](http://www.iaao.org/media/standards/AVM_STANDARD_2018.pdf) .

International Association of Assessing Officers (IAAO). (2020). *Standard on Verification and Adjustment of Sales*. Kansas City, MO. 59 pages. Found at [www.iaao.org/media/standards/Verification\_Adjustment\_of\_Sales.pdf](http://www.iaao.org/media/standards/Verification_Adjustment_of_Sales.pdf) .

Interagency Appraisal and Evaluation Guidelines*.* (2010). OCC-2010-0012. Found at:

[www.federalregister.gov/documents/2010/12/10/2010-30913/interagency-appraisal-and-evaluation-guidelines](http://www.federalregister.gov/documents/2010/12/10/2010-30913/interagency-appraisal-and-evaluation-guidelines) .

Isakson, H. R., Ecker, M. D. and L. Kennedy. (2020). Principles for Calculating AVM

Performance Metrics. *The Appraisal Journal*.  **88** (1), pp. 14–29.

Johnson R.A. and D. W. Wichern. (2007). Applied Multivariate Statistical Analysis, 6th edition.

Pearson Prentice Hall. 773 pages.

Kirchmeyer, J. (2004). A Guide to Automated Valuation Models AVMs: 101. Real Info,

Kirchmeyer & Assoc. 38 pages.

Kirchmeyer, J. and P. Staas. (2008). AVMs 201: A Practical Guide to the Implementation of Automated Valuation Models. 273 pages.

Long, T. (2010). Interagency Appraisal and Evaluation Guidelines. OCC 2010-42. Available at:

 <https://www.occ.gov/news-issuances/bulletins/2010/bulletin-2010-42.html> .

Mortgage Bankers Association (MBA). (2019). The State of Automated Valuation Models

in the Age of Big Data. Found at [www.mba.org/Documents/MBA\_Real\_Estate\_Appraisals\_(0).pdf](http://www.mba.org/Documents/MBA_Real_Estate_Appraisals_%280%29.pdf) .

Ott, R. L. and M. Longnecker. (2016). An Introduction to Statistical Methods & Data Analysis.

 Seventh Edition. Cengage Learning. 1,173 pages.

Steurer, M. and R. Hill. (2020). Metrics for Measuring the Performance of Machine Learning

 Prediction Models: An Application to the Housing Market. Found at: [www.researchgate.net/publication/338992074\_Metrics\_for\_Measuring\_the\_Performance \_of\_Machine\_Learning\_Prediction\_Models\_An\_Application\_to\_the\_Housing\_Market](http://www.researchgate.net/publication/338992074_Metrics_for_Measuring_the_Performance%09_of_Machine_Learning_Prediction_Models_An_Application_to_the_Housing_Market) .

United States Department of the Treasury. (2018). A Financial System that Creates Economic Opportunities Nonbank Financials, Fintech, and Innovation. Found at:

 [https://home.treasury.gov/sites/default/files/2018-08/A-Financial-System-that-Creates- Economic-Opportunities---Nonbank-Financials-Fintech-and-Innovation.pdf](https://home.treasury.gov/sites/default/files/2018-08/A-Financial-System-that-Creates-%09Economic-Opportunities---Nonbank-Financials-Fintech-and-Innovation.pdf) .

Valier, A. (2020). Who performs better? AVMs vs hedonic models. *Journal of Property*

 *Investment & Finance*. **38** (3). pp. 213-225.

Vyas, S and L, Kumaranayake. (2006). Constructing socio-economic status indices: How to use principal components analysis. *Health Policy and Planning*. **21** (6). pp. 459-468.

**Exhibit 1:** Glossary of Common AVM Performance Metrics

|  |  |
| --- | --- |
| **AVM Performance Metric (Abbreviation)** | **Definition** **(Reference)** |
| Hit Rate | The percent of properties for which an AVM returns a value (MBA, 2019) |
| Confidence Score | A value that indicates the AVM vendor’s confidence in its AVM valuation (CoreLogic, 2014) |
| **Accuracy Metrics** |  |
| Mean Percentage Sales Error | The mean of a set of percentage sales errors(CoreLogic, 2011) |
| Median Percentage Sales Error | The median of a set of percentage sales errors(Steurer and Hill, 2020) |
| **Precision Metrics** |  |
| Mean Absolute Percentage Error  | The mean of a set of absolute percentage sales errors (Steurer and Hill, 2020) |
| Median Absolute Percentage Error (MAPE) | The median of a set of absolute percentage sales errors (Kirchmeyer and Staas, 2008) |
| Forecast Standard Deviation (FSD) | The standard deviation of a set of percentage sales errors (Gayler, et. al., 2015) |
| **Measures of Accuracy at a Given Level of Precision** |  |
| Percent (Predicted) Error (PE) bucket | The percent of AVM valuations within a specified +/- percentage of selling prices (MBA 2019;Kirchmeyer, 2004; CoreLogic, 2011) |
| Failure Rate (FR) | The complement of the PE bucket(Ecker *et. al.,* 2021) |
| Failure Magnitude (FMag)  | The mean absolute percentage sales error for poorly predicted sales (Ecker *et. al.,* 2021) |
| Failure MAPE (FMAPE) | The median absolute percentage sales error for poorly predicted sales (Ecker *et. al.,* 2021) |
| Right Tail 20% (RT20) | The percentage of AVM valuations more than 20% higher than their corresponding selling prices (AVMetrics, 2018) |
| +/- 1×FSD | The percent of AVM valuations within one FSD of selling prices (AVMetrics, 2018) |
| **IAAO Ratio Metrics** |  |
| Coefficient of Variation (COV) | The standard deviation divided by the mean AVM valuation-to-price ratio (IAAO, 2018) |
| Coefficient of Dispersion (COD) | The average percentage deviation of the AVM’s valuation-to-sales price ratios from the median AVM valuation-to-sales price ratio (IAAO, 2018) |
| Price Related Difference (PRD) | The mean valuation-to-selling price ratio divided by the weighted (by selling prices) mean ratio (IAAO, 2018) |
| Price Related Bias (PRB) | The slope found by regressing percentage differences from the median AVM’s valuation-to-sales price ratio on a proxy variable that measures value (IAAO, 2018) |

**Exhibit 2:** Choropleth state map of the 455,563 benchmark properties in Q3 of 2018, evaluating the performance of 15 different AVMs.



**Exhibit 3:** AVM-by-FSD analysis for AVM Model 1. Note that the vendor-reported FSD is statistically significantly different from the Observed Standard Deviation for each of the 19 different vendor-reported FSDs for Model 1 (p-value < 0.001, for each).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| AVM | Hits | Vendor-Reported FSD | Observed St Dev | Difference | PercentDifference |
| Model 1 | 10,421 | 2 | 6.9 | -4.9 | 245.0% |
| Model 1 | 51,162 | 3 | 7.0 | -4.0 | 133.3% |
| Model 1 | 58,399 | 4 | 6.2 | -2.2 | 55.0% |
| Model 1 | 68,828 | 5 | 6.7 | -1.7 | 34.0% |
| Model 1 | 40,278 | 6 | 7.6 | -1.6 | 26.7% |
| Model 1 | 27,098 | 7 | 8.6 | -1.6 | 22.9% |
| Model 1 | 12,694 | 8 | 10.4 | -2.4 | 30.0% |
| Model 1 | 8,506 | 9 | 10.8 | -1.8 | 20.0% |
| Model 1 | 12,778 | 10 | 12.5 | -2.5 | 25.0% |
| Model 1 | 2,451 | 11 | 16.8 | -5.8 | 52.7% |
| Model 1 | 7,044 | 12 | 16.5 | -4.5 | 37.5% |
| Model 1 | 4,928 | 13 | 19.2 | -6.2 | 47.7% |
| Model 1 | 2,951 | 14 | 20.1 | -6.1 | 43.6% |
| Model 1 | 4,915 | 15 | 21.4 | -6.4 | 42.7% |
| Model 1 | 6,458 | 16 | 22.3 | -6.3 | 39.4% |
| Model 1 | 4,863 | 17 | 25.1 | -8.1 | 47.6% |
| Model 1 | 6,436 | 18 | 24.7 | -6.7 | 37.2% |
| Model 1 | 2,841 | 19 | 27.9 | -8.9 | 46.8% |
| Model 1 | 6,295 | 20 | 28.0 | -8.0 | 40.0% |
|  |  |  |  |  |  |
| Total | 339,346 | 6.66 | 9.65 | -2.98 | 57.7% |

**Exhibit 4:** Summary of the AVM-by-FSD analysis, where the results are averaged over all 15 AVMs using 455,563 sales across the United States in Q3 of 2018. Note that the vendor-reported FSD is statistically significantly different from the Observed Standard Deviation for each of the 15 different vendor-reported FSDs for the 15 AVMs (p-value < 0.001, for each).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AVM | Hits | Average Vendor-Reported FSD | Average Observed St Dev | Difference |
| Model 1 | 339,346 | 6.66 | 9.65 | -3.0 |
| Model 2 | 360,821 | 7.15 | 10.58 | -3.4 |
| Model 3 | 143,924 | 5.05 | 11.71 | -6.7 |
| Model 4 | 414,194 | 9.02 | 11.43 | -2.4 |
| Model 5 | 386,081 | 11.15 | 8.37 | 2.8 |
| Model 6 | 377,779 | 9.95 | 7.46 | 2.5 |
| Model 7 | 406,019 | 6.78 | 10.07 | -3.3 |
| Model 8 | 417,242 | 7.13 | 10.82 | -3.7 |
| Model 9 | 285,740 | 11.58 | 15.81 | -4.2 |
| Model 10 | 338,408 | 6.83 | 12.37 | -5.5 |
| Model 11 | 306,157 | 6.06 | 9.33 | -3.3 |
| Model 12 | 383,724 | 4.21 | 6.11 | -1.9 |
| Model 13 | 386,591 | 4.31 | 6.42 | -2.1 |
| Model 14 | 416,178 | 9.40 | 10.43 | -1.0 |
| Model 15 | 394,900 | 8.87 | 10.48 | -1.6 |
|  |  |  |  |  |
| Total | 5,357,104 | 7.70 | 9.90 | -2.2 |

**Exhibit 5:** Expanded AVM-by-FSD analysis results, including standard deviations for individual AVM Performance Metrics across the 367 AVM/FSD combinations, together with the normalized AVM-ES weights (percentages) for 455,563 property sales.

|  |  |  |
| --- | --- | --- |
| **AVM Performance Metric** | **Standard** **Deviation** | **Normalized Percentage for AVM-ES** |
| Mean Percentage Sales Error | 0.081 | 3.7% |
| Median Percentage Sales Error | 0.080 | 2.8% |
| Mean Absolute Percentage Error | 0.103 | 6.5% |
| Median Absolute Percent Error (MAPE) | 0.090 | 5.3% |
| Observed Standard Deviation | 0.121 | 6.5% |
| +/- 1 FSD | 0.173 | 7.6% |
| Failure Rate at +/- 5% (FR5) | 0.253 | 15.8% |
| Failure Rate at +/- 10% (FR10) | 0.248 | 16.3% |
| Failure Rate at +/- 15% (FR15) | 0.221 | 14.5% |
| Failure Rate at +/- 20% (FR20) | 0.196 | 12.7% |
| Right Tail (RT20) | 0.143 | 8.3% |

**Exhibit 6:** Expanded AVM-by-FSD analysis results. Five best performing AVM/FSD combinations, out of the 367 total.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rank | 1 | 2 | 3 | 4 | 5 |
| Model | 6 | 6 | 6 | 6 | 6 |
| Vendor-Reported FSD | 4 | 5 | 6 | 3 | 7 |
| AVM-ES | 0.1247 | 0.1275 | 0.1290 | 0.1327 | 0.1347 |
| Number of Properties | 28,075 | 31,536 | 31,782 | 21,069 | 30,233 |
| Mean Percentage Sales Error | -0.1 | 0.0 | 0.1 | -0.5 | 0.1 |
| Median Percentage Sales Error | 0.2 | 0.3 | 0.4 | -0.2 | 0.4 |
| Mean Absolute Percentage Error | 2.6 | 2.7 | 2.8 | 2.4 | 2.9 |
| Median Absolute Percent Error (MAPE) | 1.4 | 1.4% | 1.4 | 1.4 | 1.3 |
| Observed Standard Deviation | 5.5 | 5.7 | 5.8 | 5.0 | 6.1 |
|  +/- 1 FSD | 85.3 | 89.7 | 92.3 | 76.9 | 93.0 |
| Failure Rate at +/- 5% (FR5) | 8.6 | 10.3 | 11.0 | 7.3 | 11.7 |
| Failure Rate at +/- 10% (FR10) | 3.5 | 4.0 | 4.4 | 3.1 | 4.8 |
| Failure Rate at +/- 15% (FR15) | 2.4 | 2.6 | 2.8 | 2.1 | 3.0 |
| Failure Rate at +/- 20% (FR20) | 1.3 | 1.4 | 1.6 | 1.2 | 1.7 |
| Right Tail (RT20) | 0.7 | 0.9 | 0.9 | 0.6 | 1.0 |

|  |
| --- |
| Best (among the top five) |
| Worst (among the top five) |

**Exhibit 7:** Expanded AVM-by-FSD analysis results: Ranking and grading of the vendor-reported FSDs from Model 15. The last column reflects the relative performance of the Model 15 vendor-reported FSDs compared to all 367 AVM/FSD combinations.



|  |  |
| --- | --- |
| **AVM Performance** | **Model 15 Results** |
| Strong | 51.1% of hits |
| Reasonable  | 33.9% of hits |
| Acceptable but Weak | 8.7% of hits |
| Poor | 6.3% of hits |

**Exhibit 8:** Clark County Analysis: Standard deviations for individual AVM Performance Metrics across the twelve AVMs using a common set of 1,193 sales from Clark County, Nevada, together with the coefficients in the AVM-ES.

|  |  |  |
| --- | --- | --- |
| **AVM Performance Metric** | **Standard** **Deviation** |  **Coefficient in AVM-ES** |
| Mean Percentage Sales Error | 1.937  |  0.354 |
| Median Absolute Percent Error (MAPE) |  1.780 | 0.382 |
| Observed Standard Deviation |  1.331 |  0.361 |
| Failure Rate at +/- 10% (FR10) |  9.820 |  0.390 |
| Failure Magnitude at +/- 10% (FMag10) | 1.065 | -0.166 |
| Failure MAPE at +/- 10% (FMAPE10) | 0.815 | 0.099 |
| Coefficient of Variation (COV) | 0.015  | 0.376  |
| Coefficient of Dispersion (COD) |  1.225 | 0.397  |
| Price Related Difference (PRD) |  0.003 | -0.102 |
| Price Related Bias (PRB) |  0.009 |  0.316 |

**Exhibit 9:** Clark County Analysis: Values of ten AVM Performance Metrics for twelve AVMs using a common set of 1,193 sales in Clark County, Nevada.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **AVM** | **Mean****Percentage** **Error** | **MAPE** | **Obs****St Dev** | **Failure****Rate 10%** | **Failure****Magnitude 10%** | **Failure****MAPE** **10%** | **COV** | **COD** | **PRD** | **PRB** |
| 1 | -2.67 | 3.60 | 5.50 | 5.11 | 15.56 | 11.82 | 0.056 | 3.397 | 0.994 | 0.010 |
| 2 | -0.06 | 2.36 | 5.56 | 3.94 | 18.17 | 14.21 | 0.056 | 3.363 | 0.995 | 0.010 |
| 3 | 1.01 | 2.22 | 5.96 | 6.20 | 17.16 | 13.84 | 0.059 | 3.418 | 0.995 | 0.012 |
| 4 | 1.03 | 2.20 | 6.04 | 6.20 | 17.32 | 13.41 | 0.060 | 3.398 | 0.995 | 0.013 |
| 5 | -2.83 | 4.41 | 6.48 | 8.72 | 16.09 | 13.03 | 0.067 | 4.390 | 0.988 | 0.023 |
| 6 | -1.55 | 3.52 | 6.53 | 8.47 | 16.15 | 13.17 | 0.066 | 4.432 | 0.994 | 0.024 |
| 7 | 0.39 | 2.17 | 8.06 | 11.06 | 18.77 | 14.92 | 0.080 | 4.477 | 1.002 | 0.004 |
| 8 | -7.02 | 7.91 | 8.86 | 37.89 | 15.57 | 13.85 | 0.095 | 6.834 | 0.994 | 0.035 |
| 9 | -3.30 | 5.22 | 8.94 | 22.05 | 16.56 | 14.69 | 0.093 | 6.415 | 1.000 | 0.023 |
| 10 | -0.71 | 3.26 | 6.87 | 8.13 | 16.67 | 13.92 | 0.069 | 4.429 | 0.992 | 0.020 |
| 11 | 0.84 | 1.62 | 5.02 | 4.02 | 17.56 | 13.81 | 0.050 | 2.770 | 0.993 | 0.006 |
| 12 | 0.54 | 2.15 | 5.51 | 6.79 | 15.48 | 13.32 | 0.055 | 3.592 | 0.990 | 0.022 |

**Exhibit 10**: Clark County Analysis: Overall ranking of twelve AVMs that valued a common set of 1,193 sales in Clark County, Nevada. Outlier contribution to the AVM-ES is the combined effect of the Failure Rate, Failure Magnitude and Failure MAPE.

| Rank | Model Number | Outlier Contribution to AVM-ES | PRBContribution to AVM-ES | AVM-ES |
| --- | --- | --- | --- | --- |
| 1 | Model 11 | 0.02 | -0.101 | 3.88 |
| 2 | Model 2 | -0.07 | -0.102 | 4.22 |
| 3 | Model 4 | 0.87 | -0.101 | 5.64 |
| 4 | Model 3 | 0.94 | -0.101 | 5.69 |
| 5 | Model 12 | 1.40 | -0.101 | 5.86 |
| 6 | Model 1 | 0.58 | -0.101 | 6.26 |
| 7 | Model 10 | 1.78 | -0.101 | 7.55 |
| 8 | Model 6 | 1.92 | -0.101 | 7.97 |
| 9 | Model 7 | 2.68 | -0.102 | 8.37 |
| 10 | Model 5 | 2.02 | -0.101 | 8.83 |
| 11 | Model 9 | 7.30 | -0.102 | 16.29 |
| 12 | Model 8 | 13.56 | -0.101 | 25.04 |

**Exhibit 11:** Twelve Models combined by FSD showing the difference between Vendor Forecasted Standard Deviations and the Observed Standard Deviations with other standard AVM Performance Metrics used to evaluate the performance of AVM’s.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Vendor-Reported FSD** | **Observed Standard Deviation** | **Deviation Difference** | **Count of AVM Values** | **Test Stat** | **p-value** | **Mean Absolute Error** | **PE10%** | **RT20%** |
| 1 | 3.07% | **(2.07)** | 19 | 169.6 | 0.0000 | 2.95% | 100.0% |   |
| 2 | 3.59% | **(1.59)** | 103 | 328.6 | 0.0000 | 3.15% | 96.1% |   |
| 3 | 4.65% | **(1.65)** | 1697 | 4074.6 | 0.0000 | 2.91% | 95.7% | 0.65% |
| 4 | 5.94% | **(1.94)** | 2661 | 5865.9 | 0.0000 | 4.05% | 93.3% | 0.64% |
| 5 | 5.97% | **(0.97)** | 3069 | 4373.9 | 0.0000 | 4.14% | 93.8% | 0.59% |
| 6 | 7.94% | **(1.94)** | 911 | 1593.6 | 0.0000 | 5.59% | 84.9% | 0.55% |
| 7 | 7.93% | **(0.93)** | 1860 | 2385.8 | 0.0000 | 5.11% | 87.0% | 1.18% |
| 8 | 7.72% | **0.28**  | 1636 | 1522.6 | 0.0434 | 4.72% | 88.3% | 1.04% |
| 9 | 8.58% | **0.42**  | 660 | 598.9 | 0.0859 | 6.04% | 80.9% | 1.21% |
| 10 | 7.81% | **2.19**  | 453 | 275.7 | 0.0000 | 5.88% | 81.9% | 0.88% |
| 11 | 10.89% | **0.11**  | 340 | 332.3 | 0.7843 | 7.29% | 77.9% | 3.24% |
| 12 | 10.58% | **1.42**  | 262 | 202.9 | 0.0055 | 7.01% | 77.9% | 2.67% |
| 13 | 10.56% | **2.44**  | 171 | 112.2 | 0.0003 | 7.92% | 71.4% | 2.92% |
| 14 | 7.89% | **6.11**  | 109 | 34.3 | 0.0000 | 7.01% | 72.5% |   |
| 15 | 8.78% | **6.22**  | 69 | 23.3 | 0.0000 | 6.68% | 76.8% |   |
| 16 | 6.45% | **9.55**  | 67 | 10.7 | 0.0000 | 5.26% | 85.1% |   |
| 17 | 11.92% | **5.08**  | 36 | 17.2 | 0.0068 | 8.88% | 63.9% |   |
| 18 | 7.81% | **10.19**  | 49 | 9.0 | 0.0000 | 5.61% | 81.6% |   |
| 19 | 10.06% | **8.94**  | 26 | 7.0 | 0.0002 | 8.66% | 65.4% |   |
| 20 | 7.71% | **12.29**  | 75 | 11.0 | 0.0000 | 5.55% | 85.3% |   |
| 21+ | 12.21% | **8.79**  | 43 | 13.5 | 0.0000 | 8.36% | 76.7% |   |

|  |  |
| --- | --- |
|   | Statistically Significantly Underreported Vendor-Reported FSDs |
|  | Statistically Accurate Vendor-Reported FSDs (α = 0.01) |
|  | Statistically Significantly Overreported Vendor-Reported FSDs. |

**Exhibit 12:** Suggested thresholds for AVM Performance Metrics for AVM Grading. An additional AVM Performance Metric is the +/- 1 FSD criterion, which has a 68.26% threshold. An AVM receives a ‘Pass’ grade above 68.26% and a ‘Fail’ grade below.

| AVM Performance Metric | Acceptable but WeakThresholds | ReasonableThresholds | Strongly PerformingThresholds |
| --- | --- | --- | --- |
| Mean Percentage Sales Error\*Median Percentage Sales Error\*Mean Absolute Percentage ErrorMedian Absolute Percentage Error (MAPE) Forecast Standard Deviation (FSD)Failure Rate at +/- 5% (FR5)Failure Rate at +/- 10% (FR10)Failure Rate at +/- 15% (FR15)Failure Rate at +/- 20% (FR20)Right Tail 20% (RT20) |  2 to 52 to 510 to 1310 to 1315 to 19 40 to 6035 to 5020 to 3510 to 2010 to 15 |  1 to 21 to 25 to 105 to 105 to 1520 to 4010 to 32 5 to 202 to 102 to 10 |  < 1< 1< 5< 5< 5< 20< 10< 5< 2< 2 |

\* Thresholds presented in absolute value

1. The target property is the same as the subject property in an appraisal. [↑](#footnote-ref-1)
2. An AVM is a computer software program that produces an estimate of market value for a single target

property. In this study, the term “AVM” will be used to refer to a commercial or professional grade AVM.

That is, an AVM whose output is sold by AVM vendors to clients, in contrast to a consumer facing

AVM that typically provides output free of charge. See Mortgage Bankers Association (2019, p. 9-10). [↑](#footnote-ref-2)
3. Recent data have illustrated that approximately 90 percent of residential mortgage originations are

eligible for appraisal exceptions established since the enactment of FIRREA by the designated federal

regulatory agencies. An updated appraisal statute should account for the development of automated and

hybrid appraisal practices and sanction their use, where the characteristics of the transaction and market

conditions indicate it is prudent to do so. (FHFA Office of Inspector General, 2018, p. 9) [↑](#footnote-ref-3)
4. AVM testing requires that selling prices be known and meet the requirements to represent the best

indicator of the market value of each property. See IAAO (2013, Appendix A – Sales Validation

Guidelines) and IAAO (2020). [↑](#footnote-ref-4)
5. [↑](#footnote-ref-5)
6. Interagency Appraisal and Evaluation Guidelines (2010), Appendix B. [↑](#footnote-ref-6)
7. Allen (2009), p. 1 [↑](#footnote-ref-7)
8. AVMetrics, (2021). [↑](#footnote-ref-8)
9. See IAAO (2013), Appendix A - Sales Validation Guidelines or IAAO (2020). [↑](#footnote-ref-9)
10. CRC (2003), p. 12. [↑](#footnote-ref-10)
11. AVMetrics makes considerable proprietary effort to uncover whether the property’s selling price is truly unknown to the AVM that is valuing it, in addition to having the vendor report the last known selling price (and date) for each benchmark property being valued. [↑](#footnote-ref-11)
12. AVMetrics (2018), p. 3. [↑](#footnote-ref-12)
13. See Steurer and Hill (2020) and Ecker, et al. (2021) for a comprehensive list and additional discussion

of AVM Performance Metrics. [↑](#footnote-ref-13)
14. Ecker, et al. (2021). [↑](#footnote-ref-14)
15. Gayler *et al. (*2015), p. 5. [↑](#footnote-ref-15)
16. Isakson et al. (2020). [↑](#footnote-ref-16)
17. An Observed Standard Deviation is the standard deviation of a set of percentage sales errors, calculated from actual sales data. The Observed Standard Deviation is then compared to the vendor-reported FSD by

third-parties, such as AVMetrics. [↑](#footnote-ref-17)
18. Freddie Mac (2019) assigns a qualitative grade of ‘High’, ‘Medium’ or ‘Low Confidence’ for the value

of the FSD generated by its Home Value Explorer® (HVE®) AVM. [↑](#footnote-ref-18)
19. Johnson and Wichern (2007), Chapter 8 and Vyas and Kumaranayake (2006). [↑](#footnote-ref-19)
20. An AVM Performance Metric may need to be converted to an “error”, where zero reflects the value that

the AVM Performance Metric would attain if the AVM perfectly valued all target properties. For

example, a PE bucket needs to be converted to a Failure Rate, where a zero Failure Rate value indicates

the model has successfully valued all properties (to within, for example, 5, 10, 15 or 20 percent). [↑](#footnote-ref-20)
21. If all variables used in the AVM-ES are positively correlated, then the AVM-ES cannot be negative. [↑](#footnote-ref-21)
22. Gordon (2005). [↑](#footnote-ref-22)
23. Allen (2009). [↑](#footnote-ref-23)
24. Only 57,290 of the 455,563 (12.6%) benchmark properties were valued by each of the 15 AVMs. As a result, 87.4 percent of benchmark properties would be eliminated, if a common set of valued sales were required. [↑](#footnote-ref-24)
25. See footnote 4. [↑](#footnote-ref-25)
26. Gaylor, *et al.* (2015), p. 3. [↑](#footnote-ref-26)
27. Ott and Longnecker (2016), Section 7.2. [↑](#footnote-ref-27)
28. Note that Model 1, FSD 4 has the lowest Observed Standard Deviation of 6.2, however, both Model 1,

FSD 6 and Model 1, FSD 7 have Observed Standard Deviations that most closely match their respective

vendor-reported FSDs. [↑](#footnote-ref-28)
29. In contrast, only 73 of the 367 (19.9%) AVM/FSD combinations have AVM venders statistically

significantly overreporting their FSDs. [↑](#footnote-ref-29)
30. The AVMs have been randomized between the two empirical examples, meaning that Model 1, in the first analysis, is not the same Model 1 in the second. Such precautions are taken to discourage drawing

any general “change of scale” conclusions for AVMs, and to avoid any possibility of model identification. [↑](#footnote-ref-30)
31. The AVM Performance Metrics shown in Exhibit 5 (and Exhibit 8) are illustrative and are not an

endorsement by the authors, regarding a *best* or *preferred* set of AVM Performance Metrics. The authors

advocate that each client should choose a set of AVM Performance Metrics that are in concordance with

the business need for using the AVM. To further emphasize this point, the authors choose a (mostly)

different set of AVM Performance Metrics for each of the two empirical analyses in this study. [↑](#footnote-ref-31)
32. The contribution to the AVM-ES for any one metric is the value of the metric times its corresponding

AVM-ES coefficient. [↑](#footnote-ref-32)
33. See footnote 17. [↑](#footnote-ref-33)
34. CoreLogic (2011), p. 2. [↑](#footnote-ref-34)
35. The suggested thresholds in Exhibit 12 are only presented to illustrate the grading of AVMs, using the

statistical technique developed in this study (the AVM-ES). They do not reflect an endorsement, by the

authors, of any one, or any set of AVM performance thresholds. [↑](#footnote-ref-35)