

Principles for Calculating AVM Performance Metrics

Hans R. Isakson

Economics Department
University of Northern Iowa

Mark D. Ecker

Mathematics Department
University of Northern Iowa

Lee Kennedy

Managing Director
AVMetrics, LLC

December 10, 2019

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 on this site are presented for informational purposes only and do not constitute product approval or endorsement by the authors.

Abstract

An analysis of 5.3 million housing sales suggests that there are fundamental shortcomings with how AVM vendors currently calculate their AVM Performance Metrics, in particular the FSD. We propose consistent methodologies, i.e., five best practice principles and a modified leave-one-out, cross-validation technique, to calculate AVM Performance Metrics that comply with well-established appraiser principles and allow a consistent evaluation and comparison of AVM performance. A research AVM, developed to estimate a specific target property's market value from a dataset of comparable housing sales in Cedar Falls, Iowa, empirically illustrates that not following these principles yields overly optimistic AVM Performance Metric values.

Keywords: AVMs; AVM Performance Metrics; Cross-Validation; Error Buckets; Failure Rate; FSD; Leave-One-Out; PRESS Statistic

1. Introduction

An Automated Valuation Model (“AVM”)¹ is a computer software program that produces an estimate of the market value, called the *AVM valuation*, of a target or subject property given (i) the address of the target property and (ii) property sales and characteristics data. AVM vendors blend many property transactions, acquired from public sources or data aggregators, with one or more valuation models, acquired from academic and professional publications or developed by their own analysts, into a product called an AVM, the details of which are a closely guarded trade secret. An AVM produces a valuation along with certain statistics, called *AVM Performance Metrics*, that assess the validity, accuracy and precision of the AVM valuation. The focus of this paper is on these AVM Performance Metrics.

Assessing overall AVM performance is becoming increasingly important, due, in part, to two recent events. First, the Interagency Appraisal and Evaluation Guidelines (2010) require, among other things, that lending institutions independently assess the reliability of the AVMs they use. Second, the Federal Deposit Insurance Company (“FDIC”), the Board of Governors of the Federal Reserve System, and the Office of the Comptroller of the Currency, have jointly increased the *de minimis* threshold, from \$250,000 to \$400,000, for residential real estate transactions that do not require an appraisal with a physical inspection of the property and neighborhood (FDIC, 2019). As a result, lenders will be allowed to make more residential mortgages secured by properties that are valued utilizing an AVM as part of the valuation process, rather than a traditional appraisal.

Due to the proprietary, intellectual property contained within an AVM, assessing AVM credibility, *i.e.* its validity, accuracy and precision, is accomplished through an examination of its AVM Performance Metrics.² Typically, users of AVMs are dependent upon AVM vendors to provide reliable Performance Metrics, including, for example, the Forecast Standard Deviation (“FSD”).³

¹ Throughout this work, the term “AVM” will be used to refer to commercial or professional grade AVMs that value residential properties. That is, AVMs whose output is sold by AVM vendors to clients, in contrast to consumer facing AVMs that typically provide output free of charge. See Mortgage Bankers Association (“MBA”) (2019, p. 9-10).

² The terms “AVM Performance Metric(s)” and “Performance Metric(s)” are used interchangeably.

³ See Exhibit 1 for a list of common Performance Metrics.

However, as Kane, *et al.* (2004, p. 171) state, “[t]hird party verification is critical.” These third parties, including credit rating agencies (Fitch, Standard & Poor’s, and Moody’s) and independent AVM testing firms (such as AVMetrics, LLC), assess AVM reliability using Performance Metrics.

The purpose of this study is to first demonstrate that the calculation of Performance Metrics are not standardized across the AVM industry (AVM vendors) which often results in AVM vendors underreporting their FSDs. As a result, we suggest five best-practice principles for AVMs, and a supporting statistical procedure to implement these principles, that, if followed would bring AVMs in better alignment with current appraisal practices. Moreover, we contend that if these principles are respected, then the values of the Performance Metrics associated with one model are directly comparable to those of another. We also demonstrate that not following these principles can result in an overly optimistic assessment of an AVM’s performance. As a result, these principles should be adopted by AVM vendors and requested by the users of AVM.

2. Review of the Literature

Most of the literature regarding AVM Performance Metrics appears in unpublished manuscripts (Rossini and Kershaw, 2008; AVMetrics, 2018), self-published books (Kirchmeyer, 2004; Kirchmeyer and Staas, 2008), industry websites (Veros, 2019; Freddie Mac, 2019) or recent trade publications (IAAO, 2018; MBA, 2019). Exhibit 1 contains a list of Performance Metrics, along with associated abbreviations and definitions. For example, Gayler, *et. al.* (2015) recognize mean percentage sales error, mean absolute percentage sales error, FSD and hit rate as important metrics for the evaluation of the performance of an AVM. The Collateral Risk Management Consortium (2003) suggests using percentage sales errors, mean percentage sales error and error buckets to assess AVMs. CoreLogic (2011) recommends evaluating AVMs using the mean percentage sales error, median percentage sales error, FSD, and error buckets. AVMetrics (2018, p. 25) advocates that no more than 10 percent of AVM valuations should be more than 20% higher than their corresponding selling prices, suggesting a right tail 20% Performance Metric. Kirchmeyer and Staas (2008) state that median absolute percentage (sales) errors (“MAPE”) of less than 10 percent “are indicative of a strong AVM, while those ranging from 11 percent to 15 percent might also be acceptable for some lending programs.”

Error buckets, also called Percent (Predicted) Error (“PE”) buckets, count the number of sales that are deemed accurate (*i.e.* the *success* rate of the AVM), at a given level of precision, typically +/- 5, 10, 15 and 20% (Kirchmeyer, 2004; Slump & Arora, 2019). We use a notation PE_{xx} to refer to a specific error bucket, at a given (xx) +/- percentage. For example, PE10 represents the +/- 10% Error Bucket. Kirchmeyer (2004) originally suggested a success rate that at least 50 percent of AVM valuations should be within +/- 10% of selling prices. That is, the (percentage) success rate of an AVM at PE10 should be at least 50 percent. More recently, the Mortgage Bankers Association (2019, p. 28) reports that “[a]lmost all counties in the United States experience [PE10] rates north of 70 percent,” suggesting a success rate at PE10 of 70 percent or more.

To focus on the sales where the AVM fails to accurately predict selling prices, we define the **Failure Rate** of an AVM in a particular Error Bucket (*e.g.* PE10) as the (percentage) frequency with which an AVM *fails* to predict the value of a target property within the tolerance given by the Error Bucket (*e.g.* +/- 10%).⁴ An AVM’s Failure Rate in a given Error Bucket is the complement of the AVM’s success rate within that Error Bucket. The Failure Rate is a concept common in engineering, where it is defined as the frequency with which a component fails (Finkelstein, 2008). The Failure Rate concept is also found in other fields where the process fails to perform well, such as the percent of small business failures (Watson and Everett, 1996), the percent of students failing a computer programming course (Bennedsen and Caspersen, 2007), hotel failures (Ingram and Baum, 1997), and commercial banks insolvencies (Ashcraft, 2005).

In addition, AVM vendors typically provide a Confidence Score, “which is often interpreted as meaning that the AVM estimate is within plus or minus 10% of the ‘true’ market value of the property with a high degree of confidence” (Follain and Follain, 2007). However, the definition and use of a Confidence Score is not standardized across AVM vendors.⁵ For example, Veros®

⁴ The Failure Rate of an AVM is not exclusively limited to the +/- 10% error bucket. It can be used at any error bucket, for example, the Failure Rate +/- 5, 10, 15, or 20%. Also see Ecker *et. al.* (2019).

⁵ The Confidence Score of an AVM should not be confused with the confidence interval (level) of a statistical estimate. A Confidence Score most generally informs the level of the “AVM provider’s confidence in the estimated values” (CoreLogic, 2014), which may or may not involve a confidence interval.

(2019) describes its Confidence Score as a measure of accuracy between zero and 100 for which each decile generally corresponds to a 5 percent variance. Realtors Property Resources®, LLC (RPR®) (2018) uses a RVM Confidence Score of zero to five stars. CoreLogic's PASS® (2017) produces a Confidence Score between 60 and 100 that measure how well "sales data, property information, and comparable sales support the property valuation process." Gordon (2005) states that a Confidence Score may or may not be related to the FSD and that "[s]uch a confusion of [confidence] scores and lack of connection to statistical performance in actual use forces lenders to guess at their risk management."

For each individual target property being valued, AVM vendors may also report (CoreLogic, 2014) its FSD, which was originally coined by the Federal Home Loan Mortgage Corporation ("Freddie Mac") for use with its Home Value Explorer® AVM in the late 1990s to early 2000s. Gayler, *et al.* (2015, p. 5) define an FSD as "the standard deviation of the percentage error, where the percentage error describes the relative difference between [AVM] valuation and price." Freddie Mac (2019) qualifies the value of the FSD generated from its Home Value Explorer® (HVE®) AVM as High, Medium or Low Confidence. 'High Confidence' requires an FSD of 13 or less. 'Medium Confidence' arises from an FSD between 13 and 20, while 'Low Confidence' occurs for valuations with an FSD greater than 20.

Reporting of the FSD by AVM providers is ubiquitous, however, its description is not standardized across the industry. CoreLogic (2017, p. 1) states that "[t]he FSD is a statistic that measures the likely range or dispersion an AVM estimate will fall within, based on the consistency of the information available to the AVM at the time of estimation." Matysiak (2017, p. 7) writes that the FSD is an "estimate of the amount of variation that can occur between the actual sales price and the forecast (the most probable market value) made by the AVM." Another definition for the FSD (Gordon 2005, p. 1) is "an AVM value's expected (forecasted) proportional standard deviation around actual subsequent sales price for the given property value estimate."

The mathematically clearest definition of the FSD is that it is the standard deviation of the percentage sales errors for a collection of valuations (Gayler *et. al.*, 2015, p. 5).⁶ However, the method of calculating an FSD for an *individual target property* is not consistent, meaning that it is not clear how an AVM provider is using the sampling distribution and/or parsing a dataset to provide a unique FSD value for any one particular target property.

An AVM report typically contains a high/low range of value based upon a +/- 1×FSD confidence interval around the AVM valuation (CoreLogic, 2017). This 1×FSD interval is often interpreted by assuming that the underlying sales errors are normally distributed. Under normality, an AVM vendor has 68.26% confidence that the true market value of the target property lies within +/- one FSD of the AVM valuation, or 95% confidence that the market value of the target property falls within +/- 1.96×FSD of the AVM valuation. The assumption of normality allows the client to use the FSD-based confidence interval to test hypotheses regarding the market value of the target property. Therefore, it behooves the AVM vendor to test the distribution of percentage sales errors for normality before rendering any FSD-based inference, including a high/low value range.

Additional studies using Performance Metrics include Clapp and O’Connor (2008), who report the mean absolute percentage sales error and its 25th, 50th, and 75th percentiles to evaluate seven different valuation models. Pokryshevskaya and Antipov (2011) use mean average percentage sales error and the coefficient of dispersion, (“COD”), the average percentage deviation of the median sales error (IAAO, 2013). Rossini and Kershaw (2008) describe several Performance Metrics for which they also suggest performance thresholds. They report the mean absolute percentage sales error, FSD, and COD, in addition to three error buckets: PE10, PE15 and PE20. Kane *et. al.*, (2004) suggest using the COD, together with the coefficient of variation (“COV”), both of which assess horizontal equity.⁷ Lastly, the IAAO (2018) advocates an additional metric,

⁶ The FSD definition by Gayler *et. al.* (2015), the standard deviation of the percentage sales errors, is used for any FSD calculation perform by the authors in this work.

⁷ Horizontal equity is the notion that people in the same circumstances should be treated the same or that similar properties should have similar tax assessed values. Following IAAO (2018), both the COD and COV assess horizontal equity as they measure spread of AVM valuation to selling price ratios about the

namely, the price related difference (“PRD”), which measures the vertical equity of the property tax system.⁸

3. An AVM Valuation Example

To illustrate the calculations of these Performance Metrics, we construct a research AVM, labeled the Test Valuation Model (“TVM”). The TVM is a regression model (Kane, *et. al.*, 2004, Chapter 8) containing fifteen housing characteristics employed as independent variables. See Ecker, *et. al.* (2019) for additional details about the model. The purpose of this analysis is to illustrate that Performance Metrics are sensitive to their calculation methodology and to additionally show how these metrics change when applying several best-practice principles.

We demonstrate below how different statistical methodologies, using the same valuation model and the same dataset, result in different Performance Metric values. We start with a base case of Performance Metrics calculated using *internally* fitted (regression) valuations. The dataset employed in this empirical demonstration consists of 53 housing sales, denoted with circles (“○”) in Exhibit 2, from 2012 located in a submarket of Cedar Falls, Iowa. We arbitrarily chose the first house to sell in 2013 as our target property, which is indicated by the filled-in box (“■”), in Exhibit 2. The TVM uses the 53 sales to produce a valuation of \$159,427 as of January 1, 2013 for the target property. The Performance Metrics associated with the target property, reported in Exhibit 3, are calculated from the 53 comparable property sales based on the internally generated predicted values from the regression.

EXHIBITS 2 and 3 About Here

The TVM Performance Metrics reported in Exhibit 3 are indicative of an acceptable AVM. The TVM performs well with regard to the Rossini and Kershaw’s (2008) thresholds, the Kirchmeyer (2004) Error Buckets and the TVM’s Failure Rate (at +/- 10%) is 39.6 percent. With an FSD of

center (mean ratio for COV and median ratio for COD).

⁸ Vertical equity means that higher valued properties have higher assessed values than lower valued properties. For this analysis, following IAAO (2018), vertical equity is assessed through the PRD statistic by comparing AVM value to selling price ratios for the most and least expensive houses.

13.4, the TVM earns a ‘Medium Confidence’ score based on Freddie Mac’s (2019) thresholds (and only 0.5 away from attaining ‘High Confidence’). Only four of the 53 (7.5 percent) comparable sales had their TVM values more than 20% larger than their respective selling prices, which is within AVMetrics’ (2018) right tail 10 percent suggested threshold. Lastly, the TVM has a PRD value of 1.0156, which suggests that it is slightly overvaluing inexpensive houses, more so than it undervalues expensive houses, but the model performs reasonably well for houses close to the median and mean selling prices.

The TVM regression-based predicted values for each of the 53 comparable sales, which produce the Performance Metrics seen in Exhibit 3, provide internal measures of model performance. That is, all 53 housing sales, that produced the target property’s valuation, are re-used to determine each comparable sales’ valuation. As such, the resulting metrics in Exhibit 3 tend to be overly optimistic, compared to the prediction of a new, external-from-the-model observation.⁹ Although the internal calculations yield favorable Performance Metric values, an AVM should use an *external*, cross-validation methodology to judge how well the model predicts market values for housing sales that were *not* used to construct the model. We recommend the leave-one-out (“LOO”) cross-validation procedure, which removes each sale in the original dataset, one at a time, and generates a valuation for that left-out property from the remaining $n-1$ sales.¹⁰ This process is repeated until each property in the original dataset has been valued. Consequentially, under a LOO validation methodology, each house will never be used in the model to (indirectly) value itself.

4. Lack of Standardization of AVM Performance Metrics

In this section, we empirically illustrate that current methodologies used by AVM vendors to (self-) report Performance Metrics, in particular, the FSD, for an individual target property are not

⁹ “A result of this model development process is that the error mean square MSE will tend to understate the inherent variability in making future predictions from the selected model.” See Neter, *et. al.* (1996, p. 435).

¹⁰ Another common validation technique is a k-Fold analysis, which is a mass appraisal technique that cycles through valuing $100(k/n)$ percent of the data. The leave-one-out (“LOO”) process is a k-Fold analysis that values one observation at a time.

consistent, meaning that it is not clear how an AVM parses a dataset of housing sales to provide a unique value of, for example, the FSD, for an individual target property.

Third party companies, such as AVMetrics, LLC, independently test commercial AVMs bi-weekly by providing participating AVM vendors with sales of hundreds of thousands of test properties. Vendors are allowed 48 hours from receipt of the test properties to return AVM valuations, along with an FSD, for each test property (AVMetrics, 2018, p. 3). Each test property provided to the vendor by AVMetrics has passed a number of quality checks to ensure that its sales price is representative of market value (an arms-length transaction – see IAAO, 2013 Appendix A – Sales Validation Guidelines) and is unknown to the vendor.¹¹ AVMetrics then compares each returned AVM valuation to its corresponding selling prices to calculate the AVM’s Performance Metrics, including the observed FSD, which is calculated using all housing sales with a common vendor-reported FSD.

An analysis that looks at the accuracy and precision of AVMs by aggregating housing sales to a common, vendor-reported FSD is called an AVM-by-FSD analysis. This analysis allows AVM testers, such as AVMetrics, to corroborate a vendor-reported FSD with an observed FSD. Exhibit 4 shows the performance of fourteen AVMs, whose identity is blinded for confidentiality, across a total of 683,802 properties, where the FSD for each of these properties had a vendor-reported value of eight (the horizontal line in Exhibit 4). In Exhibit 4, Models 5 and 6 have their observed FSDs (8.1 and 7.7, respectively) closest to their self-reported value of eight.

Alarmedly, as seen in Exhibit 4, the majority of the fourteen AVMs are underreporting their FSDs, with eight models having their observed FSDs above nine (when it should be eight). That is, the vendor-reported FSD of eight appears to be overly optimistic in assessing the AVM’s precision for the majority of AVMs in Exhibit 4. In particular, the three AVMs (Models 8, 9 and 10) with

¹¹ AVMetrics takes considerable proprietary effort to uncover whether or not the property’s selling price is truly unknown to the AVM that is valuing it, including, but not limited to, having the vendor report the last known selling price (and date) for each test property being valued.

the highest observed FSDs (10.7, 13.9 and 13.8, respectively) are off by more than 60 percent (from the vendor-reported FSD value of eight).

EXHIBIT 4 About Here

The underreporting of the FSD is not unique to only those AVMetrics test properties with a vendor-reported FSD of eight. Exhibit 5 shows the performance for most of the 327 total AVM/FSD combinations, each with at least 100 hits, which results in over 5.3 million housing sales. In Exhibit 5, 138,118 properties, each with a vendor-reported FSD of twelve, had a calculated or observed FSD of 20.4 (averaged over fifteen AVMs). That is, the vendor-reported FSD of twelve is underreported by 8.4 or by 70.3% on average.

Inspecting Exhibit 5, AVMs are consistently underreporting their FSDs, with the best performing having vendor-reported FSDs of seven and eight. Overall, for the 327 total AVM/FSD combinations, 77.1% (252 of the 327) have underreported their FSDs. Moreover, the average FSD difference (observed FSD minus vendor-reported FSD) for all 327 is 4.3 or, on average, a 54.9% underreporting percentage computed using all 5.3 million sales. In addition, only 25 (7.6% of the 327) AVM/FSD combinations have their observed FSD within +/- 10% of the vendor-reported FSD, which suggests a lack of consistency regarding the FSD's definition and/or how AVM providers calculate an FSD for each target property. These findings are not a trivial matter, as an overly optimistic vendor-reported FSD, typically underreported by 4.3 (on average, for 5.3 million sales) makes AVMs appear substantially more reliable than they actually are.

EXHIBIT 5 About Here

Lastly, to help mitigate the above findings, AVM providers should adopt and employ uniform methodologies when calculating Performance Metrics, in particular the FSD and percentage sales errors. Moreover, lenders who use FSDs should demand that vendors align their statistical calculation methodologies to comply with well-established appraiser principles, such as those set forth in the next section.

5. Valuation Principles for AVMs

In this section, we propose four valuation principles, that appraisers already observe when performing a traditional appraisal, for AVMs. Because both an appraiser and an AVM have the common purpose of valuing exactly one target property, we advocate that AVM vendors adopt these principles. Moreover, these principles, if implemented, would provide greater uniformity in assessing the reliability of AVM's when i) building a valuation model that uses comparable sales to estimate the target property's market value and ii) constructing a dataset used to compute Performance Metrics, which may be different from the original dataset of comparable sales. We recalculate the Performance Metrics, found in Exhibit 3, following these principles to illustrate their application.

The first principle in the AVM building and validation process is that *the comparable properties should be very close substitutes for the target property*. Comparable sales “are located in the same area and are very similar in size, condition and features” (DeSimone, 2015) as the target property. Appraisers understand the concept of comparable properties, but it is difficult to build an AVM that can select a set of comparable properties as well as a well-trained appraiser. For all practical purposes, comparable properties are essentially equivalent to each other (and the target property), as these comparable properties and the target property compete for the same set of buyers. Much appraisal and academic literature addresses the issue of substitutability among properties using submarkets. For example, Palm (1978), Schnare and Struyk (1976), and Watkins (2001) use predefined submarkets, such as ZIP codes, census block groups, school districts, etc. from which comparable properties are selected. Bourassa *et. al.* (1999), Goodman and Thibodeau (2003), Tu (2003) and Isakson and Ecker (2018) use statistical analyses to allow the housing sales data, itself, to identify (possibly non-contiguous) submarkets containing similar properties.

The second principle is that *a property should never be used to value itself*. Obviously, the sale of a target property should never be included in the original dataset that values itself. Again, appraisers would almost never violate this principle. However, if the target property has been recently flipped (bought and resold within a few months), then it is possible for the sale of the target property to be included in the original dataset used by an AVM. The more common violation of this principle occurs indirectly, when the valuation model uses the coefficients that it calculated from all properties, including the property being valued, to estimate that property's market value.

Comparable sales should i) never include the target property, even indirectly, ii) be selected from the same submarket as the target property and iii) be sold as close in time to the valuation date as possible, but never after it. Therefore, the third principle is that ***the dataset should contain no sales that post-date the valuation date of the target property.*** That is, sales that occur after the valuation date should not be used to value the target property.¹² Doing so produces what Thanos, *et al.* (2016) call “arrow of time” violations.

To avoid arrow of time violations, neither the entire post-dated sale itself, nor any individual variable that is post-dated should be included in the dataset. Specifically, mixing and matching pre- and post-dated variables for an individual house is especially egregious, because housing characteristics can change over time, by, for example, the owner’s remodeling. A common violation of the arrow of time principle occurs when the assessed value, in a Tax Assessed Value AVM, does not comport to the date of sale for a comparable house and/or the valuation date for the target property.

The fourth, another time-related principle, is that ***sales of comparable properties should span similar market conditions.*** That is, comparable sales should be chosen that sold during market conditions similar to those that existed as of the target property’s valuation date. Of course, all comparable sales must occur on or before the valuation date, but the question of how old the comparable sales can be, requires knowledge of the local market conditions. See Stevenson (2008), Novy-Marx (2009), Vander Hoff (1996) and Dale-Johnson and Hamilton (1998) for discussions of comparative market analyses.

As a general rule, it is easier to control for changing market conditions by selecting comparable properties closer in time to the valuation date, than to build time-related trends into a valuation model. In other words, it is difficult, at best, to decide the appropriate statistical methodology to account for changing market conditions. It is simpler to choose comparable sales that occur closer

¹² The calibration (and validation) of retrospective AVMs can easily include sales that post-date the valuation date. In addition, using the internal residuals from a regression analysis as sales errors, as we demonstrate in Exhibit 3, would also include some predicted values based upon post-dated sales.

in time to the target property's valuation date (e.g. hold market conditions constant), than it is to decide, mathematically, how the model should reflect these changing market conditions, when including sales that span changing market conditions. If the valuation model within the AVM does not account for changing market conditions or fails to use comparable sales that reflect similar, market conditions, then the AVM will produce non-credible (biased and/or imprecise) valuations.

6. Improved AVM Performance Methodology

Although the internal metrics presented in Exhibit 3 indicate that the TVM is an acceptable valuation model, the calculation of these metrics violate the second principle above, namely that a property should never be used to value itself. Specifically, the metrics in Exhibit 3 are an internal measure of model performance because the TVM predicted values for each of the 53 houses were created using the regression coefficients already derived from these 53 houses. As a result, we advocate using the LOO methodology to provide the TVM's Performance Metrics, whereby, the original dataset of comparable sales would 'do double duty'; it can be used to value the target property and also provide, using a LOO strategy, the values of the Performance Metrics.

In a regression, the Predicted Residual Sum of Squares ("PRESS") statistic has traditionally been used to identify unique individual observations (outliers) and/or to determine a set of independent variables that statistically significantly contribute to explaining the dependent variable (Montgomery *et. al.*, 2001). The PRESS statistic implements a LOO methodology that systematically pulls each of the n sales, one at a time, from the original dataset and uses the remaining $n-1$ sales to estimate the market value of that removed property. This process is repeated by cycling through each of the n sales, one-at-a-time, employing the same AVM that was originally used to value the target property.¹³ A PRESS predicted value can be used to calculate (with the property's selling price) the PRESS sales error. In fact, any Performance Metric seen in Exhibit 3 can be calculated using the PRESS predicted value. For example, the PRESS-based FSD, for the target property evaluated using the TVM, is 19.6, an increase from the internal-prediction-based FSD of 13.4. Such a large increase in the FSD indicates that the TVM does not predict new

¹³ Technically, re-running of the AVM regression n times is not required, as the PRESS residual can be calculated using the original regression that valued the target property. See Montgomery *et. al.* (2001). pp. 598-600.

observations nearly as well as it explains the house prices for the existing 53 comparable sales and, as a result, the initial FSD of 13.4 is providing an overly optimistic assessment of model's predictive performance.

Although the PRESS methodology abides by our second principle (a property should never be used to value itself), it violates principle three (the dataset should contain no sales that post-date the valuation date of the target property). In fact, all but the most recent comparable sale will have at least one sale post-dating it. In particular, to value the oldest of the 53 sales in the Cedar Falls dataset, the PRESS procedure uses the remaining 52 sales, that each post-date this oldest sale. As a result, we do not present any of the PRESS-based metrics, other than the FSD of 19.6, discussed above. Instead, we advocate modifying the PRESS methodology to abide by our four principles, discussed in the previous section. We call this modified LOO procedure a Generalized PRESS ("GenPRESS") methodology, because, as long as the AVM uses a set of comparable sales to produce its valuation, regardless of the valuation model being a regression or not, then the GenPRESS procedure can provide the values of the Performance Metrics.

To enforce the no-post-dated-sales principle of the GenPRESS methodology, additional housing sales are needed to value the oldest comparable properties. For example, if an AVM uses n comparable sales to value the target property, then the oldest of these n sales cannot be valued using the remaining $n - 1$ post-dated sales; new comparable sales would need to be selected. These newly chosen comparable properties must be selected in concordance with the time and submarket criteria established when picking the original comparable sales used to value the target property. Then, the AVM is run, using these newly gathered sales, exactly as it was when valuing the target property, to value each of the original comparable sales.

To calculate the TVM's GenPRESS regression-based predicted value for oldest of the 53 properties in the Cedar Falls dataset, property sales in 2011 were gathered. As seen in Exhibit 2, 40 property sales, denoted with plus ("+") symbols, occurred in the same submarket as the target property in 2011, within one year of the selling date for the oldest sale from 2012 (January 1, 2012) in the original dataset. A market analysis (House Price Index) in Rosburg *et. al.* (2018, p. 124, Exhibit 3) indicates that relatively stable market conditions existed in 2011 and 2012 in Cedar Falls. As a result, the TVM was run with these 40 property sales in 2011 to value this oldest

comparable property. When using a twelve-month moving window, the number of property sales valuing each of the 53 original comparable sales in the original dataset ranged from 40 to 57.

A GenPRESS predicted value was calculated for each of the original 53 comparable properties and was used to (re-) calculate the values of the Performance Metrics reported in Exhibit 6. Any substantial difference, in the respective statistics between these Exhibits 3 and 6, provides an assessment of “the applicability of the model to data beyond those on which the model is based” (Neter *et. al.*, 1996, p. 435).

EXHIBIT 6 About Here

All of the GenPRESS-based Performance Metrics in Exhibit 6 show a poorly performing AVM, compared to the corresponding metrics in Exhibit 3. In particular, the FSD has risen from 13.4 to 24.7, while the Failure Rate (at +/- 10%) has increased from 39.6 to 66.0 percent. The difference in the respective values of the Performance Metrics in these two tables is entirely attributed to using the GenPRESS methodology in which no post-dated sales were used to value each comparable sale. Exhibit 6 reveals that the results seen in Exhibit 3 make the AVM appear to be more accurate, precise, and reliable. As a result, the methodology used to calculate the values of Performance Metrics meaningfully impacts an AVM’s credibility.

Therefore, we introduce a fifth principle that is applicable when resampling sales in a cross-validation analysis to calculate Performance Metrics. Specifically, ***valuations produced by an AVM in a cross-validation analysis should use the same methodology originally used to value the target property.*** That is, the Performance Metrics associated with a target property (for example, the FSD) should be calculated using the dataset that was originally used to value the target property using a leave-one-out methodology.¹⁴ Our recommendation to use the GenPRESS

¹⁴ Other philosophical choices exist for cross-validation, such as creating a holdout dataset, for example, by setting aside n of the 53 comparable sales in the Cedar Falls dataset. The remaining $53 - n$ sales are then used to value the target property, while the n withheld sales are used to calculate the values of the Performance Metrics. See Kane *et. al.* (2004, p. 171). We advocate including all comparable property sales in the original dataset to value the target property and then using the LOO GenPRESS methodology to calculate the values of the Performance Metrics, primarily because the GenPRESS avoids sacrificing any comparable sales needed to value the target property.

methodology is more than just a method to compute the AVM's Performance Metrics following our fifth principle. It also provides i) guidance on how to resample a sales dataset to calculate Performance Metrics, and ii) a straightforward and consistent method to calculate a unique FSD value for each target property.

The GenPRESS methodology also assesses the AVM's accuracy for a dataset that was not used to create the model, especially when including additional comparable properties when valuing the earliest sales in the original dataset. The GenPRESS-based Performance Metrics evaluate the quality of the AVM's prediction of new or external properties because each target property is left out of the training dataset, in contrast to the usual internally-based predicted value.

7. Discussion and Conclusions

We advocate that AVM vendors adopt the five best practice principles recommended in this paper, coupled with the GenPRESS methodology, to calculate Performance Metrics to allow a more realistic assessment of AVM performance. In addition, we have established, for the Cedar Falls dataset, that applying these five principles, together with the GenPRESS methodology, produces a degradation in the values of Performance Metrics, when comparing the results seen in Exhibit 3 to that in Exhibit 6. In other words, not following these principles provides an overly optimistic evaluation of the AVM's performance.

Using a research AVM, we demonstrate that the values of Performance Metrics highly depend upon their calculation methodologies. For example, we show that when using three different statistical methodologies, the target property's FSD changes from 13.4, using regular regression predicted values, to 19.6, using traditional PRESS predicted values (that allow the use of post-dated sales), and finally to 24.7, using GenPRESS predicted values (that do *not* allow the use of post-dated sales). In particular, the latter two FSD values (19.6 and 24.7) are each calculated using a LOO methodology. The increase from 19.6 to 24.7 reveals the substantial effect from eliminating the use of post-dated sales. Moreover, the values of nearly all Performance Metrics substantially deteriorate when enforcing the two principles that no property should be used in the model to value itself, even indirectly, and no sales should post-date the property being valued.

We have also shown that the vendor-reported FSDs for commercial AVMs are not nearly as credible as currently being reported. Our analysis of 5.3 million housing sales in Section 4,

indicates that 85.0 percent of AVMs with a vendor-reported FSD of 15 or below are overly optimistic in their reported precision (by 5.3 or 83.3%, on average). Overall, the observation that AVM vendors are inconsistent in their calculation of Confidence Scores, along with their underreporting of FSDs, should be a concern for the currently under-regulated AVM industry. Standardization of the calculation of Performance Metrics should be employed by AVM vendors, as this step could lessen the impact of future regulatory mandates of the Consumer Finance Protection Board (“CFPB”), by adopting quality control standards.

Unfortunately, exactly how AVM vendors calculate Performance Metrics for any one target property is part of the AVM’s proprietary intellectual property. The analysis of 327 AVM/FSD combinations using 5.3 million sales in Section 4 suggests that AVM vendors are underreporting their FSDs. Although not resulting from a direct testing of our five principles, these overly optimistic vendor-reported FSDs suggest that our five principles, together with the GenPRESS methodology, can potentially serve as a means to comparatively evaluate AVM performance properly and consistently. That is, we would fully expect to see similar results for commercial AVMs, as presented for the TVM (for the Cedar Falls dataset), because Neter, *et. al.*, (1996, p. 435)¹⁵ have already established that internally-based metrics produce overly optimistic results.

What remains unclear is by how much the values of the Performance Metrics would change, especially the FSD, if our five principles (together with the GenPRESS methodology) were adopted by AVM vendors. However, only through the scrutiny of the internal workings of AVMs, by the AVM vendors themselves, can the efficacy of any set of best practice principles be determined. In sum, we encourage AVM vendors to adopt a set of principles, such as those detailed in this work, that comply with well-established appraisal practices and allow AVM clients to trust the credibility and comparability of AVMs, measured through their Performance Metrics.

¹⁵ There may be AVM vendors currently abiding by a set of valuation principles like those for which we advocate, and, as a result, little to no improvement in AVM credibility would be seen.

8. References

- Ashcraft, A. B. (2005). Are Banks Really Special? New Evidence from the FDIC-Induced Failure of Healthy Banks, *American Economic Review*. **95**(5), pp. 1712-1730.
- AVMetrics. (2018). Automated Valuation Model (AVM) Tests. 58 pages.
- Bennedsen, J, and M. E. Caspersen. (2007). Failure rates in introductory programming. *SIGCSE Bull*, **39**(2). pp. 32-36.
- Bourassa, S. C., Hamelink, F., Hoesli, M., and B. D MacGregor. (1999). Defining housing submarkets. *Journal of Housing Economics*, **8**, pp. 160-183.
- Clapp J. M. and P. M. O'Connor. (2008). Automated Valuation Models of Time and Space: Best Practice. *Journal of Property Tax Assessment & Administration*. **5**(2), pp. 57-67.
- Collateral Risk Management Consortium. (2003). The CRC Guide to Automated Valuation Model (AVM) Performance. Found at https://professional.sauder.ubc.ca/re_creditprogram/course_resources/courses/content/344/CRC_AVM.pdf .
- 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> .
- CoreLogic. (2017). Forecast Standard Deviation & AVM Confidence Scores Found at: <https://www.corelogic.com/downloadable-docs/fsd-and-avm-confidence.pdf> .
- Dale-Johnson, D. and S.W. Hamilton. (1998). Housing Market Conditions, Listing Choice and MLS Market Share, *Real Estate Economics*. **26**(2), pp. 275-307.
- DeSimone, B. (2015). What are Comps? Understanding a Key Real Estate Tool. Found at: <https://www.zillow.com/blog/what-are-comps-179631/> .
- Ecker, M. D., Isakson, H. R. and L. Kennedy. (2019). An Exposition of AVM Performance Metrics. Technical Report, University of Northern Iowa. Found at: <http://www.math.uni.edu/~ecker/research> .
- Falessi, D., Narayana, L., Thai, J.F., and B. Turhan. (2018). Preserving Order of Data When Validating Defect Prediction Model. Found at: <https://arxiv.org/abs/1809.01510> .
- FDIC. (2019). New Appraisal Threshold for Residential Real Estate Loans. Found at <https://www.fdic.gov/news/news/financial/2019/fil19053.html> .
- Finkelstein, M. (2008). Introduction. *Failure Rate Modelling for Reliability and Risk*.

- Springer Series in Reliability Engineering. pp. 1–84.
- Follain, J.R. and B.A. Follain. (2007). AVMs Have Feelings, too. Found at: <https://www.ficonsulting.com/avms-have-feelings-too/> .
- Freddie Mac. (2019). Confidence Levels. Found at http://www.freddiemac.com/hve/confidence_scores.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 https://www.corelogic.com.au/sites/default/files/2018-03/20151028-CL-RP_AVM.pdf .
- Goodman A. C. and T. G. Thibodeau. (2003). Housing Market Segmentation and Hedonic Prediction Accuracy, *Journal of Housing Economics*. **12**, pp. 181–201.
- Gordon, D. (2005). Metrics Matter. Found at http://www.freddiemac.com/hve/pdf/dougwhitepaper_metricsmatter.pdf .
- IAAO. (2013). *Standard on Ratio Studies*, Kansas City, MO. 63 pages. Found at https://www.iaao.org/media/standards/Standard_on_Ratio_Studies.pdf .
- IAAO. (2018). *Standard on Automated Valuation Models (AVMs)*. Found at https://www.iaao.org/media/standards/AVM_STANDARD_2018.pdf .
- Ingram, P. and J. A. C. Baum. (1997). Chain Affiliation and the Failure of Manhattan Hotels, 1898-1980, *Administrative Science Quarterly*. **42**(1), pp. 68-102.
- Interagency Appraisal and Evaluation Guidelines. (2010). Found at www.federalregister.gov/documents/2010/12/10/2010-30913/interagency-appraisal-and-evaluation-guidelines .
- Isakson, H. R. and M. D. Ecker. (2018). The Influence of Leaking Underground Storage Tanks on Nearby House Prices. *Journal of Economic Insight*. **44**(1), pp 45-67.
- Kane, M.S., M.R. Linné, J.A. Johnson. (2004). *Practical Applications in Appraisal Valuation Modeling*, Appraisal Institute: Chicago, 200 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.
- Montgomery, D.C., Peck, E.A. and G.G. Vining. (2001). *Introduction to Linear Regression Analysis*. John Wiley & Sons.
- 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_(0).pdf).
- Neter, J., Kutner, M. H., Nachtsheim, C. J. and W. Wasserman. (1996). *Applied Linear Statistical Models*. Irwin. 1,408 pages.
- Novy-Marx, R. (2009). Hot and Cold Markets, *Real Estate Economics*. **37**(1), pp 1-22.
- Palm R. (1978). Spatial Segmentation of the Urban Housing Market, *Economic Geography*. **54**, pp. 210-221.
- Pokryshevskaya, E. B. and E. A. Antipov. (2011). Applying a CART-based approach for the diagnostics of mass appraisal models. *Economics Bulletin*. **31**(3), pp 1-8.
- Rosberg, A., Isakson, H., Ecker, M. and T. Strauss. (2017). Beyond Standardized Test Scores: The Impact of a Public School Closure on House Prices. *Journal of Housing Research*. **26**(2), pp. 119-135.
- Rossini, P. and P. Kershaw. (2008). Automated Valuation Model Accuracy: Some Empirical Testing. 14th Pacific Rim Real Estate Society Conference. Found at <https://pdfs.semanticscholar.org/43ff/e969edd70eb86e56638f1b2ea353e13b7682.pdf>.
- RPR. (2018). What is an AVM or RVM®. Confidence Score. Found at: <https://support.narrpr.com/hc/en-us/articles/204964670-What-is-an-AVM-or-RVM-confidence-score->.
- Schnare A. and R. Struyk. (1976). Segmentation in Urban Housing Markets, *Journal of Urban Economics*. **3**, pp. 146 – 166.
- Slump, R. and A. Arora. (2019). Property Valuations as a Part of the Rating Process, Fitch Ratings, CRN Presentation, Las Vegas, NV.
- Stevenson, S. (2008). Modeling Housing Market Fundamentals: Empirical Evidence of Extreme Market Conditions, *Real Estate Economics*. **36**(1), pp. 1-29.
- Thanos, S., J. Dube, and D. Legros. (2016). Putting Time into Space: The Temporal Coherence of Spatial Applications in the Housing Market, *Regional Science and Urban Economics*. **58**, pp. 78-88.
- Tu, Y. (2003). in O’Sullivan, T and Gibb, K (eds). *Housing Economics and Public Policy*, Blackwell Science, Oxford, UK.
- Vander Hoff, J. (1996). Adjustable and Fixed Rate Mortgage Termination, Option Values and Local Market Conditions: An Empirical Analysis. *Real Estate Economics*. **24**(3), pp. 379-406.

Veros. (2019). Veros Confidence Scores: VCS. Found at:
http://www.veros.com/files/2715/1665/6640/Veros_Confidence_Score_Insert.pdf .

Watkins, C. A. (2001). The Definition and Identification of Housing Submarkets. *Environment and Planning*. **33**, pp. 2235-2253.

Watson, J. and J. E. Everett. (1996). Do Small Businesses Have High Failure Rates? Evidence from Australian Retailers, *Journal of Small Business Management*. **34**(4), pp 45-62.

9. Additional Materials:

Chagani, E. (2018). The Potential of Machine Learning Real Estate Valuation Models. Blog. Found at <https://blog.realestate.cornell.edu/2018/03/28/machine-learning/> .

CoreLogic. (2017). About Automated Valuation Models (AVMs). Found at <https://www.corelogic.com/downloadable-docs/about-automated-valuation-models.pdf> .

CoreLogic. (2019). So – What is an AVM?
Found at; www.corelogic.com/landing-pages/realavm.aspx .

Various Federal Agencies. Interagency Appraisal and Evaluation Guidelines. Found at <https://www.fdic.gov/news/news/financial/2010/fil10082a.pdf> .

Exhibit 1: Glossary of common AVM Performance Metrics.

AVM Performance Metric	Abbreviation	Definition (Source)
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 (Pokryshevskaya and Antipov, 2011)
Coefficient of Variation	COV	The standard deviation divided by the mean AVM valuation-to-price ratio (Kane <i>et. al.</i> , 2004)
Confidence Score	None	A value that indicates the AVM vendor's confidence in its AVM valuation (CoreLogic, 2014)
Failure Rate	None	The complement of the PPE bucket (Ecker <i>et. al.</i> , 2019)
Forecast Standard Deviation	FSD	The standard deviation of a set of percentage sales errors (Gayler, <i>et. al.</i> , 2015)
Hit Rate	None	The percent of properties for which an AVM returns a value (MBA, 2019)
Mean Percentage Sales Error	MPE	The mean of a set of percentage sales errors (CoreLogic, 2011)
Median Absolute Percentage Sales Error	MAPE	The median of a set of absolute percentage sales errors (Kirchmeyer and Staas, 2008)
Median Percentage Sales Error	None	The median of a set of percentage sales errors (CoreLogic, 2011)
Percentage Sales Error	None	The AVM valuation minus its selling price, for a target property, which is then divided by the selling price (CRC, 2003)
Percent Predicted Error bucket	PE%	The percent of AVM valuations within a specified +/- percentage of selling prices (Kirchmeyer, 2004; CoreLogic, 2011)
Price Related Difference	PRD	The mean valuation-to-selling price ratio divided by the weighted (by selling prices) mean ratio (IAAO, 2018)
Right Tail 20%	None	The percentage of AVM valuations more than 20% higher than their corresponding selling prices (AVMetrics, 2018)

Exhibit 2: Location of house sales in Cedar Falls, Iowa. The X-Y coordinates are State Plane Coordinates, Iowa North, NAD 1983, where each unit represents 10,000 feet.

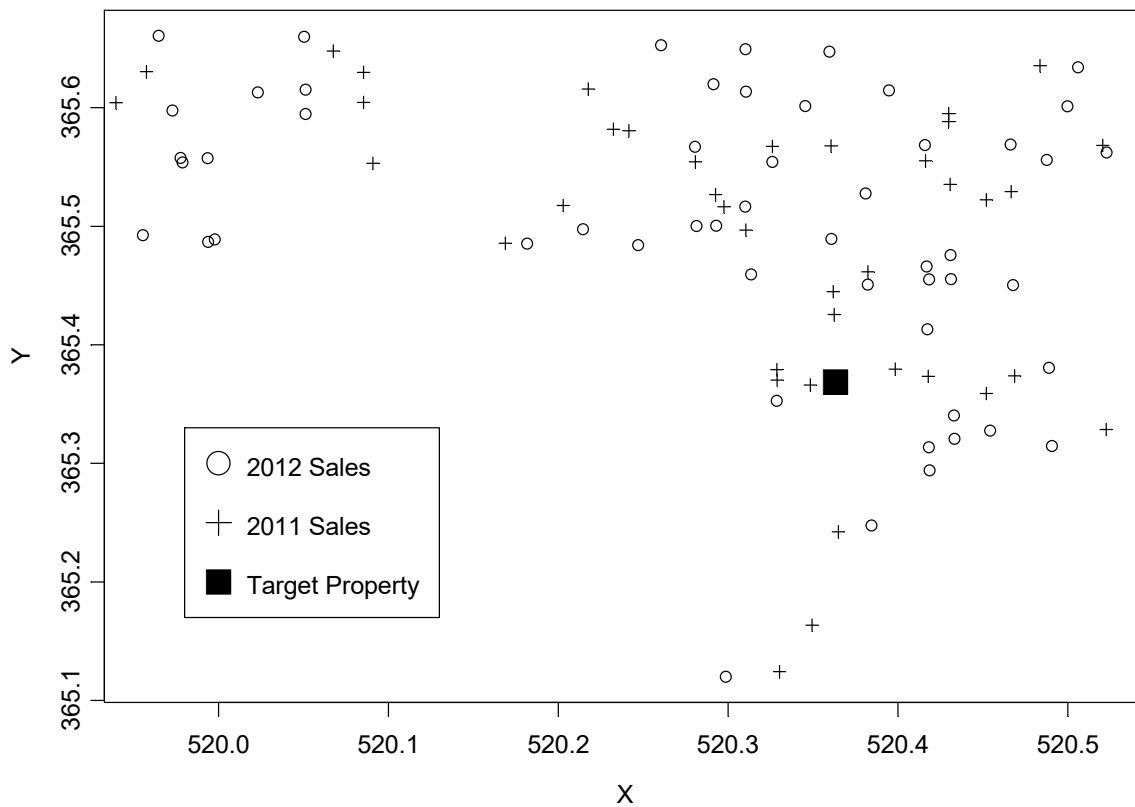


Exhibit 3. TVM Performance Metrics

AVM Metric	Value
Mean Sales Error	\$ 1,014
Mean Percentage Sales Error (MPE)	0.84%
Median Sales Error	\$ -1,832
Median Percentage Sales Error	-1.51%
Mean Absolute Sales Error	\$ 13,788
Median Absolute Sales Error	\$ 10,611
Mean Absolute Percentage (Sales) Error	9.8%
Median Absolute Percentage (Sales) Error (MAPE)	6.5%
FSD	13.4
Percent of Estimates within +/- 10% (PE10)	32/53 for 60.4%
Failure Rate at +/- 10%	21/53 for 39.6%
Percent of Estimates within +/- 15% (PE15)	40/53 for 75.5%
Percent of Estimates within +/- 20% (PE20)	47/53 for 88.7%
Percent of Estimates more than 20% (Right Tail 20%)	4/53 for 7.5%
Coefficient of Variation (COV) of TVM/Sale Price	0.13317 or 13.3
Coefficient of Dispersion (COD) of TVM/Sale Price	9.86
Regression R-Squared (Coefficient of Determination)	0.7272
Adjusted R-Squared	0.6165
PRD of TVM/Sale Price	1.0156
Mean Selling Price of the 53 Sales	\$ 143,767
Median Selling Price of the 53 Sales	\$ 130,000
Mean TVM valuation for the 53 Sales	\$ 143,062
Median TVM valuation for the 53 Sales	\$ 138,650

Exhibit 4. AVM Performance for 683,802 target properties valued by fourteen AVMs, each with a self-reported FSD of 8. Two AVMs produce nearly identical results for their mean percentage sales errors and the observed FSDs (Models 11 and 13; Models 7 and 14) and, as a result, only twelve AVMs are presented in this exhibit.

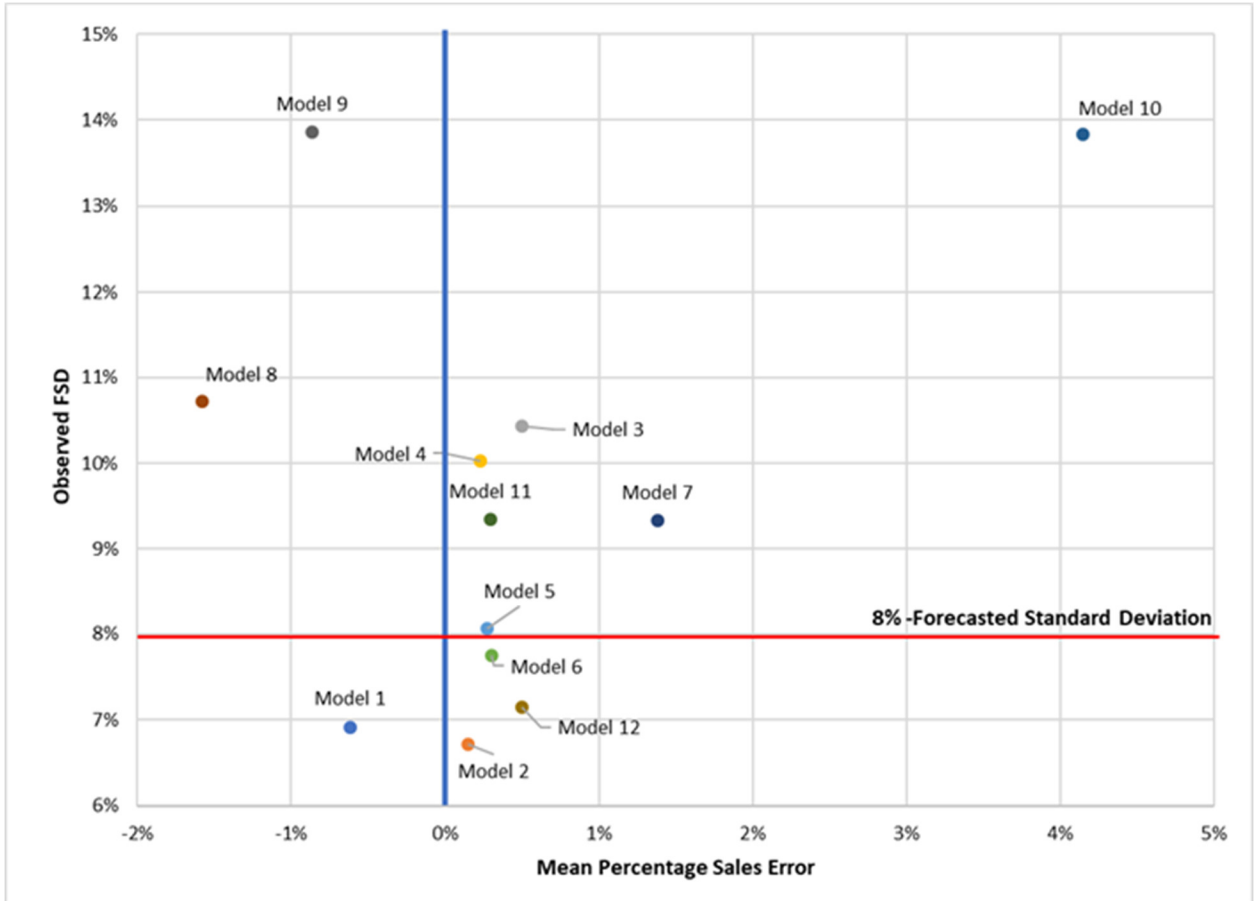


Exhibit 5. AVM performance by vendor-reported FSDs for 5.3 million target properties valued by as many as fifteen AVMs.

Vendor Reported FSD	Number of AVMs	Hits	Observed FSD	FSD Difference: Observed FSD minus Vendor Reported FSD	Percent FSD Difference
1	5	29,076	9.0	8.0	802.0
2	7	69,671	7.1	5.1	375.0
3	9	406,350	7.6	4.6	152.6
4	12	620,462	7.4	3.4	83.8
5	14	870,006	7.1	2.1	42.7
6	15	485,109	8.0	2.0	33.4
7	14	715,546	8.3	1.3	19.0
8	14	683,802	9.5	1.5	18.5
9	14	305,621	12.5	3.5	38.9
10	15	230,958	15.5	5.5	54.7
11	15	155,956	19.7	8.7	78.8
12	15	138,118	20.4	8.4	70.3
13	15	120,975	22.1	9.1	70.0
14	15	84,143	21.8	7.8	55.4
15	15	74,297	23.7	8.7	58.1
20	12	79,238	27.1	7.1	35.3
25	5	7,475	28.3	3.3	13.4
30	4	3,980	32.0	2.0	6.8
40	1	530	24.9	-15.1	-37.8
Total	327	5,344,833	19.6	4.3	54.9

Exhibit 6. GenPRESS Performance Metrics for the TVM.

AVM Metric	Value
Mean Sales Error	\$ 806
Mean Percentage Sales Error (MPE)	3.5%
Median Sales Error	\$ -5,245
Median Percentage Sales Error	-3.1%
Mean Absolute Sales Error	\$ 24,374
Median Absolute Sales Error	\$ 17,764
Mean Absolute Percentage (Sales) Error	18.2%
Median Absolute Percentage (Sales) Error (MAPE)	13.9%
FSD	24.7
Percent of Estimates within +/- 10% (PE10)	18/53 for 34.0%
Failure Rate at +/- 10%	35/53 for 66.0%
Percent of Estimates within +/- 15% (PE15)	32/53 for 60.4%
Percent of Estimates within +/- 20% (PE20)	36/53 for 83.7%
Percent of Estimates more than 20% (Right Tail 20%)	10/53 for 18.9%
Coefficient of Variation (COV) of TVM/Price	0.23909 or 23.9
Coefficient of Dispersion (COD) of TVM/Price	18.51
PRD of TVM/Price	1.0290
Average Regression R-Squared (Coefficient of Determination) for the 53 Regressions	0.7524
Average Adjusted R-Squared for the 53 Regressions	0.6463
Average Number of Sales for the 53 Regressions	50
Mean Selling Price of the 53 Sales	\$ 143,767
Median Selling Price of the 53 Sales	\$ 130,000
Mean TVM GenPRESS valuation for the 53 Sales	\$ 144,573
Median TVM GenPRESS valuation for the 53 Sales	\$ 139,216