An Exposition of AVM Performance Metrics

Mark D. Ecker

Mathematics Department
University of Northern Iowa

Hans R. Isakson

Economics Department
University of Northern Iowa

Lee Kennedy

Managing Director
AVMetrics, LLC

December 10, 2019

Abstract

This case presents a thorough exposition of the state-of-the-art in the calculation and interpretation of Automated Valuation Model ("AVM") Performance Metrics, including the Forecast Standard Deviation ("FSD"), confidence scores, vertical and horizontal equity and error buckets. It also discusses the Failure Rate, Failure Magnitude, and Failure MAPE metrics, which focus on tails of an AVM's distribution of errors. In addition, this case demonstrates the calculations and relationships between the AVM Performance Metrics using a regression model and property sales from a medium-sized, midwestern, college city.

Keywords: AVM, AVM Performance Metrics, Confidence Score, Cross-Validation, Error Buckets, Failure Magnitude, Failure MAPE, Failure Rate, FSD, Horizontal Inequity, PRESS, Sales Tier Analysis, Unbiasedness, Vertical Inequity

1. Introduction

Automated Valuation Models[1] ("AVMs") are becoming an increasingly important tool when estimating the market values of residential properties, due in part, to the Federal Deposit Insurance Company ("FDIC"), the Board of Governors of the Federal Reserve System, and the Office of the Comptroller of the Currency jointly issued notice to increase 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, 2018). As a result, lenders will be allowed to make more residential mortgages secured by properties that are valued using an AVM, rather than a traditional appraisal. Although there is no trade association of AVM vendors nor any government agency that collects data regarding the AVM industry, CoreLogic, the largest AVM vendor, claims to have 4.5 billion property records from nearly all counties in the United States spanning fifty years.[2] Other AVM vendors include The Federal Home Loan Mortgage Corporation ("Freddie Mac"), VeroValue, Clear Capital, and Equifax.

There is no universal definition of an AVM. CoreLogic (2016) defines an AVM as "a computerized system that analyzes data to provide an estimate of market value for a property at a given point in time." The IAAO (2018, p. 4) states that an AVM is

> (a) mathematically based computer software program that market analysts use to produce an estimate of market value based on market analysis of location, market conditions, and real estate characteristics from information that was previously and separately collected.

For purposes of this work, an AVM is defined as a computer software program that produces an estimate of market value, called the *AVM valuation*, along with statistics that assess the accuracy and precision of the AVM (called *AVM Performance Metrics*), for a single target property, given the address of the target property and property sales and property characteristics data. This definition distinguishes AVMs from computer assisted mass assessment ("CAMA") systems, which simultaneously produce market value estimates for multiple target properties. As a result, Moore (2006) and Matysiak (2017) consider AVMs to be a subset of CAMAs.

The International Association of Assessment Officers ("IAAO") (2018, p. 6) states, "the purpose of an AVM is to efficiently provide an accurate, uniform, equitable estimate of fair market value" of a target property.[3] Compared to the valuation obtained from an appraisal with a physical

inspection, an AVM can quickly and inexpensively produce an its valuation, for a target property at a given point in time, called the *AVM valuation date* (either contemporaneously or retrospectively).

Even though the internal workings of an AVM are a closely guarded proprietary secret, in general, nearly all AVMs contain four primary ingredients: (i) a large database of recent property sales, which includes locations and characteristics of these sold properties (together with their selling prices and dates of sale); (ii) a dataset of all properties, regardless of whether the property has recently sold or not, that also contains the characteristics of these properties; (iii) a theoretical property valuation model that defines, mathematically, the relationship between the value of a target property and some or all of the characteristics of that target property; and (iv) an algorithmic mechanism or statistical procedure that fits the theoretical valuation model. Once the valuation model is fit (using i, iii, and iv), the AVM looks up the characteristics of a target property using (ii) and applies the valuation model to the target property to obtain the AVM valuation.

There are many formulations of theoretical property valuation models that can estimate a property's market value from a set of housing sales. The Mortgage Bankers Association ("MBA") (2019, p. 11-12) identifies hedonic regression, appraisal emulation, repeat-sales index, hybrid or blended, and cascade models as valuation approaches used by AVMs. Matysiak (2017) observes that hedonic regression analysis is the traditional method of choice for residential valuation models, while Grover (2016) states that regression is the orthodox approach in mass appraisal models, including AVMs. Other types of valuation models include tax assessed value models (CRC, 2003), artificial neural networks and expert systems (Epley, 2017), nonparametric regressions (Filho and Bin, 2005), k-nearest neighbors (Isakson, 1986), regression trees (Fan, *et al*., 2006) and fuzzy logic models (Theriault, *et al*., 2005). However, details regarding the specific type of valuation model contained within the AVM is a copyright protected, proprietary secret, vigorously guarded by the AVM vendor.[4] Yet, an AVM provider might disclose snippets of its model as part of its marketing efforts.[5] As a consequence, academicians are not able to fully examine the algorithmic mechanism or statistical procedure used within an AVM. Instead, the performance of an AVM is typically evaluated by comparing AVM valuations to selling prices.[6]

Generally stated, the market value of a target property is estimated by the AVM as some function (i.e., the valuation model) of a set of recent property sales. These property sales, called the *training data,* are used to fit (calibrate or estimate the coefficients of) the theoretical valuation model and may consist of three to five comparable sales[7] in an appraisal emulation AVM, or thousands of sales in a regression AVM. The training dataset is employed by the valuation model to estimate the numerical relationship(s) between the selected property characteristics and price. These gleaned relationships, for properties in the training dataset, are then applied to the characteristics of the target property to provide the AVM valuation. If the AVM returns a valuation for the target property, then the AVM has successfully provided a 'hit'[8] (CoreLogic, 2016). The Hit Rate is a common AVM Performance Metric that measures the percentage of target properties, for which the AVM returned a valuation (MBA, 2019, p. 10).

Output from an AVM (called an *AVM report*)[9] for a target property typically includes an AVM valuation, along with high and low ranges of value, together with AVM Performance Metrics, such as a Forecast Standard Deviation ("FSD") and/or a Confidence Score. A more detailed AVM report may also include the target property's recent transaction history, together with market area metrics, such as neighborhood median house price, maps containing nearby land-uses, a list of property sales used by the AVM, and Multiple Listing Service ("MLS") information on the target property.

AVM Performance Metrics typically quantify how well an AVM predicts selling prices.[10] Unfortunately, several AVM Performance Metrics are not universally defined, nor consistently calculated. Consequentially, the AVM user is often not able to compare the performance of competing AVMs, even when valuing the same target property. Worse yet, AVMs deliver measures of precision that often do not meet widely-accepted scientific standards.

Lastly, qualifying how well the AVM is performing, via AVM Performance Metrics, is generally achieved by comparing the values of the individual AVM Performance Metrics to pre-set thresholds. This work examines the available literature that provides thresholds for AVM Performance Metrics. Unfortunately, only a few peer-reviewed articles (Gloudemans (2001), Benmamqun (2006), Moore (2006) and SanPeitro *et. al.* (2019)) contain specific threshold values (or confidence intervals/hypothesis tests) for AVM Performance Metrics. In contrast, most AVM Performance Metric suggested performance thresholds have been presented in unpublished

manuscripts (Rossini and Kershaw, 2008; AVMetrics, 2018), self-published books (Kirchmeyer, 2004; Kirchmeyer and Staas, 2008), or recent trade publications (Gloudemans, 2011; IAAO, 2013; Veros, 2017; IAAO, 2018; Freddie Mac, 2019a; and MBA, 2019). Exhibit 1 provides an overview of existing AVM Performance Metric thresholds that will be explained in this work.

2. Accuracy and Precision

Most generally, accuracy measures the level of conformity of an estimate to a known benchmark. For property sales, the accuracy of an estimate from any valuation model, including an AVM, is established by comparing a property's AVM valuation to its selling price,[11] via its sales error:[12]

$$Sales\ Error = AVM\ Valuation - Selling\ Price.$$

A sales error is a measure of accuracy calculated in dollars. But, because a $30,000 sales error for a $300,000 property (10% error) is not equivalent to a $30,000 sales error for a $3,000,000 property (1% error), a relative measure of accuracy can be obtained by the percentage sales error, which is:

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

As a result, a positive percentage sales error means that the AVM has overvalued the target property, while a negative percent indicates undervaluation.

EXHIBIT 1 About Here

A set of recent housing sales is typically needed by the AVM to produce its valuation for a target property. But not all properties will have recently sold before the valuation date. Therefore, an individual AVM valuation is the observed value of a random variable (property value). Had a different set of properties sold before the valuation date, then the AVM would have produced a different valuation. In addition, a valuation model can never fully explain 100 percent of the observed variation in selling prices contained in the training dataset. Therefore, the AVM valuation has a built-in margin of error or set of *within-AVM estimates*. In statistical terms, an AVM valuation for a single target property has its own sampling distribution.

The sampling distribution or set of within-AVM estimates can be theoretically derived or empirically calculated. For example, the final estimate of (predicted) value from a regression

AVM theoretically follows a t-distribution (Neter *et. al.*, 1996, Section 6.6) provided the regression sales errors follow a normal distribution. If normality is not a reasonable assumption, then the AVM valuation's sampling distribution can be empirically computed, using, for example, a bootstrap resampling procedure (Montgomery *et. al.*, 2001, Section 14.4).

AVM Performance Metrics often summarize the distribution of the percentage sales errors. In particular, the mean percentage sales error measures the center of the distribution (accuracy). The FSD conveys its spread (precision), while error buckets, which count the number of sales having their AVM valuations accurate to within a certain percentage (ex, +/- 10%) of selling prices, quantify AVM accuracy at a given level of precision. See Morris and Langari (2016, Section 2.3) for a general discussion regarding accuracy and precision.

3. Accuracy of the AVM Valuations - Unbiasedness

Typically, AVM performance is assessed by examining the aggregate level of concordance between a set of AVM valuations and their corresponding selling prices for many target properties, in a *holdout dataset*,[13] in contrast to the assessment of the AVM, described above, for valuing only one target property. If all percentage sales errors are zero then the AVM provides 100 percent (aggregate) accuracy with perfect precision.

If the mean (percentage) sales error for a set of target properties in the holdout set is zero, then the AVM is producing *unbiased* estimates of market value.[14] A high-quality AVM will be unbiased, and also have its median (percentage) sales error be zero (*unbiased at the median*). Should either or both the mean and median (percentage) sales errors be statistically significantly different from zero, then, the AVM is systematically overvaluing or undervaluing properties. As a result, the AVM would be producing biased valuations.

Additionally, CoreLogic (2011, p. 6) states that AVMs need to be evaluated to "ensure unbiased test results" using "actual sales data in a specific trade area or market prior to the information being available to the model." As a consequence, AVM providers would benefit from more transparency, by providing a detailed description of the 'actual sales data' (holdout datasets) used to test the AVM's mean (and median) percentage sales errors for unbiasedness.

4. Precision of AVM Valuations – MAPE

Generally, precision measures the consistency of a set of observations, typically calculated with respect to the center.[15] Two precision metrics frequently reported by AVMs are the mean and median absolute (percentage) sales error. In particular, the Median Absolute Percentage Error ("MAPE") provides "a useful comparison across different models and across different data sets and locations" Rossini and Kershaw, 2008, p. 4).

Kirchmeyer and Staas (2008, p. 91) state [see Exhibit 1] that MAPE values less than 10 percent "are indicative of a strong AVM," while those between 10 and 15 percent might "be acceptable for some lending programs." Kirchmeyer and Staas (2008, p. 91) further add that AVMs "that exhibit an average or median margin of error [MAPE] in excess of 20 percent are generally not appropriate for use in riskier [*e.g.*, lower credit scores] applications."

Rossini and Kershaw (2008) present AVM performance thresholds called "reasonable" and "absolute minimum/maximum" threshold levels for several AVM Performance Metrics [see Exhibit 1], including the MAPE in their study of house prices in Australia. Rossini and Kershaw (2008) consider an AVM that produces a MAPE value of 13 or higher to be "of no real value to its users" and that this "AVM provides no better accuracy than basic submarket averages" (Rossini and Kershaw, 2008, p. 8).

5. Precision of AVM Valuations – FSD

For each individual target property being valued, AVM vendors may also report (CoreLogic, 2014) the FSD, which was coined by Freddie Mac for use with its Home Value Explorer® AVM in the late 1990s to early 2000s. Today, reporting of the FSD by AVM vendors is ubiquitous, however, its definition 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 an FSD is that it is the standard deviation of the percentage sales errors (Gayler *et. al.,* 2015, p. 5).[16] In addition, the method of calculating the FSD, for the same target property valued by different AVMs, is not consistent, meaning that it is not clear how an AVM vendor is using the sampling distribution and/or parsing a holdout dataset to provide a unique FSD value for a particular target property. Note that the FSD is different from the standard deviation of the sales errors, because the percentage sales error is the ratio of sales error divided by selling price.

The usual industry interpretation of the FSD is that one can be 68.26% confident that the true market value of a target property lies within +/- one FSD of the AVM valuation (CoreLogic, 2017). If the target property has an AVM valuation of $300,000 and an FSD of 19, then one has 68.26% confidence that the true market value of this target property lies between $243,000 and $357,000. Because this interpretation of an FSD assumes that the sales errors are normally distributed, AVM providers should verify this normality assumption using a hypothesis test, such as the Shapiro–Wilk test (Shapiro-Wilk, 1965). When faced with a non-normal distribution of errors, the AVM vendor should investigate why these errors are not well-behaved. Isakson, Ecker and Kennedy (2019) propose five AVM-related principles that if followed, will greatly improve the chances of having well-behaved errors.

Freddie Mac (2019a) qualifies the value of the FSD generated from its Home Value Explorer® (HVE®) AVM as having *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 estimates with an FSD greater than 20 [see Exhibit 1].[17] Freddie Mac (2019a) reports that "(o)ver 70% of our HVE estimates are High Confidence estimates. About 25% of HVE estimates are Medium Confidence estimates, and less than 5% of HVE estimates are Low." For the hypothetical property with a $300,000 AVM valuation together with an FSD of 19, the Freddie Mac criterion would ascribe 'Medium Confidence' to the $300,000 valuation, despite having a +/- factor of $57,000. This 'Medium Confidence' label is confusing, given that one would have only 68.26% confidence that the market value lies within this 'Medium Confidence' +/- $57,000 range.

According to the IAAO (2018, p. 14), "confidence intervals are the most commonly used measure of reliability" of AVM valuations. An AVM report usually includes a High/Low range of likely

market values for the target property, often created from: AVM valuation +/- 1×FSD (CoreLogic, 2017) and, as a result, offers only 68.26% confidence. While the 68.26% confidence level attached to the plus and minus one FSD confidence interval appears high, it falls far below the scientific standard of 95%, and makes the AVM appear to be more precise than it really is.[18] Also, the IAAO (2018, p. 14) advises, "it is important to form conclusions about AVM quality assurance measures through statistical hypothesis testing because the AVM will ultimately be applied to the population of properties from which the [training] sample was drawn." The most widely used scientific standard for testing hypotheses is the five percent significance level (Cowles and Davis, 1982; Kaye and Freedman, 2011, p. 380), which translates to a 95% confidence level.[19]

Consequentially, if the AVM vendor is not providing a confidence interval for its High/Low range, then it should follow the advice of the IAAO (2018) and do so. Transparency in reporting the confidence level will allow the AVM client to match a particular choice of confidence level, for example 68.26%, 90%, 95% or 99%, to their level of acceptable risk, given their intended use[20] of the AVM (MBA, 2019, p. 10). Preferably, the AVM vendor should allow the client to pre-specify their desired confidence level, or at least not assume the default should be 68.26%. To illustrate, assuming that the FSD is the only precision metric reported by the AVM,[21] a 95% FSD-based confidence interval can be calculated by: AVM valuation +/- 1.96×FSD.[22]

Lastly, if the AVM can produce a confidence interval for each property in the holdout dataset (for example, from a regression AVM), then the AVM's precision can be measured by the average or median 95 (or 90 or 99) percent confidence interval width. The wider the confidence interval, the less precise is its corresponding AVM valuation. Then, the number of properties in the training dataset that have a confidence interval width larger than a specified dollar or percentage amount,[23] provides valuable information about the precision of the AVM valuation.

6. Confidence Scores

AVM providers may also produce a Confidence Score, "which is often interpreted as meaning that the AVM valuation is within plus or minus 10 percent of the 'true' market value of the property with a high degree of confidence" (Follain and Follain, 2007). Confidence Scores can be related to FSDs (Gordon, 2005) but, unfortunately, like the FSD, their definition and use are not consistent across AVM vendors. Each vendor has created its own scale and "there is no apparent correlation

between the different confidence score scales used" (CoreLogic, 2011, p. 6). For example, Veros® (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 an RVM Confidence Score of zero to five stars without any explanation of what each star represents. CoreLogic's (2017) PASS® produces a Confidence Score between 60 and 100 that measures how well "sales data, property information, and comparable sales support the property valuation process." Gordon (2005) states 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." As a result, the confidence scores reported by one AVM vendor may not be comparable to those of another.

7. AVM Error Buckets

Calculated for housing sales in the holdout dataset, AVM error buckets assess accuracy at a preset level of precision, by counting the number of sales that are deemed accurate for each individual bucket. Cumulative error buckets are symmetric intervals around the target of zero sales error. Common cumulative error buckets, also called percent (predicted) errors ("PEs"), include +/- 10%, +/- 15% and +/- 20% (CoreLogic, 2011). Incremental error buckets are typically not symmetric around the target of zero sales error. An example of an incremental error bucket is the AVM's percentage sales errors that only fall between 5 and 10 percent. Only the number of property sales in the dataset will limit how many error buckets can be created or how fine the distribution of the percentage sales errors can be discretized, as each error bucket calculation needs to be computed using "a large pool of estimates" (Rossini and Kershaw, 2008, p. 8).

Kirchmeyer (2004, p. 5) suggests AVM performance thresholds for cumulative PE buckets: at least 50 percent of an AVM's valuations should fall within +/- 10 percent of selling prices, and at least 70 percent should fall within +/- 15% [see Exhibit 1]. In other words, PE10 should be at least 50 percent and PE15 should be at least 70 percent. For this work, a PE performance threshold is denoted as PE/%, where PE represents the percentage error bucket (i.e. +/- 15%) and % represents the percentage of the valuations that must fall within +/- PE of the selling price. For example, the Kirchmeyer (2004) suggested error bucket thresholds would be written as PE10/50 and PE15/70. Rossini and Kershaw (2008) advocate that the PE10/65 threshold be an 'absolute minimum' level of AVM performance, together with PE15/80 being a 'reasonable' level of AVM performance.

More recently, AVMetrics (2018), (a third-party independent tester of AVM performance) observes how well the AVMs that it evaluates are performing and suggests AVM performance thresholds of PE10/68-70 and PE15/80. In addition, the Mortgage Bankers Association (2019, p. 28) indicates that (i) "[a]lmost all counties in the United States experience [PE10] rates north of 70 percent", suggesting PE10/70, and (ii) a majority of AVM valuations analyzed by county have realized PE15 rates "north of 80-90 percent." Lastly, Veros (2017) states that "[t]op-tier AVMs can estimate the value of a home (in a blind purchase transaction) within 10% about 80% to 90% of the time", suggesting PE10/80-90 for high quality AVMs [see Exhibit 1].

A potential weakness of any PE/% performance threshold, for example Kirchmeyer's (2004) PE15/70, is that this threshold ignores the magnitude of the errors in the complementary set of housing sales, where the AVM fails to predict selling prices accurately. For example, using Kirchmeyer's (2004) PE15/70 threshold, an AVM can be in error by any amount, for the (up to) 30 percent of valuations that fail and yet, the AVM and *all* of its valuations would be deemed acceptable, if the PE/% threshold were used in a pass/fail fashion. Thus, the complement of the PE calculation is defined as the ***Failure Rate*** of the AVM, to focus on the properties where the AVM fails to accurately predict selling prices.[24] In Exhibit 1, the AVM performance thresholds suggested in the literature for PE/%s have been translated to Failure Rate thresholds.

The concept of Failure Rate has been customarily associated with tests of systems and components. For example, Failure Rate is a term common in engineering, where it is defined as the frequency with which a component fails a particular test (Finkelstein, 2008). Other Failure Rate examples include 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), corporate bankruptcies (Platt and Platt, 1990), and commercial bank insolvencies (Ashcraft, 2005). The Failure Rate seamlessly extends to AVMs, because the presence of a large number of extreme percentage sales errors, given accurate sales data, suggests that the AVM is failing to accurately predict selling prices.

In addition, the ***Failure Magnitude*** and ***Failure MAPE*** are defined as the mean and median absolute percentage sales error, respectively, but only for properties that count towards the Failure Rate; only those with a percentage sales error larger, in absolute value, than the PE bucket

percentage. To illustrate for a PE15 bucket, the Failure Magnitude and Failure MAPE must each be at least 15 percent, provided that at least one sale has an AVM valuation that is more than 15 percent different than its corresponding selling price. How much larger the Failure Magnitude is, compared to the PE bucket percentage and Failure MAPE, will allow the AVM user to assess how the AVM performs in the tails of its distribution of percentage sales errors. Together, these three AVM Performance Metrics, the Failure Rate, Failure Magnitude and Failure MAPE, quantify the errors of the AVM within its tails, by counting the number (percentage) of housing sales for which the AVM is poorly predicting, together with the average (mean and median) absolute percentage error for those poorly predicted sales.

AVMetrics (2018, p. 25) also suggests a criterion for a particular incremental error bucket: no more than 10 percent of the AVM's valuations should be more than 20 percent larger than their corresponding selling prices (Right Tail 20%, in Exhibit 1). From a lender's perspective, potential overvaluation is a key risk component because overvaluations (especially by more than 20 percent) can expose the lender to more collateral risk than factored into the original loan-to-value ratio.

Percentage sales errors can be examined across the spectrum of sales prices, in a Sales Tier Analysis, where selling prices of the sales in the holdout dataset are first stratified into price tiers or buckets. The individual dollar values chosen for the price tiers should reflect relevant housing values in a particular geographic region. For example, at a national level, AVMetrics (2018, p. 24) advocates using the following price tiers: Under $100,000, $100,000 to 300,000, $300,000 to $500,000, $500,000 to $700,000, $700,000 to $900,000, $900,000 to $1,100,000 and Over $1,100,000. Alternatively, the selling prices could be divided into groups with (essentially) equal numbers of sales, such as by selling price quartiles or deciles. Regardless of how the distribution of selling prices is discretized, average percentage sales errors are calculated for all properties within each tier or stratum. In the Sales Tier Analysis, the resulting percentage sales error in each price tier can be compared to zero, and to each other across tiers, to investigate any potential patterns of over or undervaluation, by price level, produced by the AVM. Moreover, multiple sets of sales tiers, in which percentage sales errors are calculated, would ensure that any uncovered AVM shortcomings are invariant to the stratification (i.e., quartiles, deciles, or the particular dollar choices for each tier).

8. Horizontal and Vertical Equity

The IAAO (2018, p. 13) states that, "[s]ales-based ratio [AVM valuation to selling price] studies are among the most objective methods for testing the performance and quality of any valuation system." While most of the IAAO recommended metrics have already been discussed, two topics, typically found in the study of property taxes and assessed values, deserve further attention: horizontal and vertical equity. Horizontal equity is the economic principle that similar properties should have similar tax assessed values, while vertical equity suggests that higher valued properties should bear higher tax burdens.

The IAAO (2018, p. 13) provides horizontal and vertical equity metrics and performance thresholds that measure and evaluate "the overall quality of the AVM value estimates" by examining "the degree of variability (uniformity) in the results of any AVM model." Lack of uniformity indicates that properties are not valued by the AVM consistently; the AVM is producing flawed valuations. In particular, one can measure horizontal inequity by the variability within the distribution of property sales using the center (mean or median), while vertical inequity can be measured by comparing properties values along the entire spectrum of sales, but with extra scrutiny typically paid to sales in the tails of the distribution. Horizontal inequity AVM Performance Metrics detailed by the IAAO (2018) include the Coefficient of Variation and Coefficient of Dispersion, while the Price Related Difference and Price Related Bias measure and test for vertical inequity.

The Coefficient of Variation ("COV") is a unitless AVM Performance Metric, defined as the standard deviation divided by the mean. The COV permits comparison of datasets with different scales and allows for an assessment of the spread in terms of the mean. As seen in Exhibit 1, Rossini and Kershaw (2008) indicate that a reasonably performing AVM will have its COV below 13, with an absolute maximum COV value of 17. Unfortunately, the COV is a biased estimator, as its estimates are systematically too low. Adidi (2010, p. 3) provides an unbiased COV statistic, whereby the COV is adjusted by the factor $(1 + \frac{1}{4n})$, where n is the number of sales.

The Coefficient of Dispersion ("COD") [25] measures dispersion about the median and is recommended by the IAAO (2018, p. 13) as the "variability statistic of choice", when compared to the COV, due to the influence of outliers on the mean and standard deviation. For housing sales,

the COD measures the average percentage deviation of the AVM valuation-to-sales price ratios from the median AVM valuation-to-sales price ratios (IAAO, 2013, pp. 13-14). In other words, the COD is a measure of the (horizontal) dispersion of these ratios about the median. A formal confidence interval for the COD can be computed by assuming normality of the population of AVM valuation-to-sales price ratios, from which the sample of properties (ratios) is drawn (Gloudemans, 2001), or employing the bootstrap statistical resampling procedure (Benmamqun, 2010). Moreover, the IAAO (2018, Table 2) states that the COD should fall between 5 and 20 for residential properties (as seen in Exhibit 1), with a COD between 5 and 10 for newer, more homogeneous properties.

The Price-Related Differential ("PRD") is the mean (valuation-to-selling price) ratio divided by the weighted (by selling prices) mean ratio (IAAO, 2013, p. 14). The PRD assesses the level of uniformity in AVM valuation-to-sales price ratios between low and high valued properties. The IAAO (2013, p. 14) calls the PRD an "index statistic measurement" of vertical equity, whereby, a PRD statistic of 1.00 indicates that there are no systematic differences in AVM valuation-to-price ratio for low and high valued properties (no vertical inequity). However, because the PRD statistic can have a value of one, while the AVM is systematically undervaluing *all* properties by a consistent, say 5%, a Sales Tier Analysis is advocated to be reported together with the PRD statistic. A PRD statistic different from one indicates that the AVM valuations exhibit vertical inequity. In other words, the AVM has a tendency to value high-value properties differently than low-value properties in relation to their selling prices.

The IAAO (2013, p. 14) indicates that a PRD statistic should lie between 0.98 and 1.03 [see Exhibit 1]. A PRD value above 1.03 indicates regressivity; the AVM systematically undervalues high-valued properties more so (at a lower ratio) than low-valued properties. In other words, lower priced properties are overvalued *more than* higher priced properties are undervalued. A PRD value below 0.98 indicates progressivity or the systematic undervaluation of low-valued properties (at a lower ratio) compared to high-valued properties. Thus, lower price properties are undervalued at a higher rate than high priced properties are overvalued.

The IAAO (2013, p. 15) cautions not to actively create two sets of sales that consist of high- and low-priced properties and then calculate the PRD for each, because the choice of the cuts that

determine these two sets of extreme sales can affect the resulting PRD statistics. Furthermore, because the PRD statistic is a ratio of two random variables (where each of the random variables is already a ratio; the PRD is the mean of a ratio divided by the ratio of means), it is biased upwards towards regressivity (Gloudemans, 2011, p. 5). Moreover, the PRD is highly affected by outliers and heteroscedasticity[26] (Denne, 2011) and may be less meaningful with small samples (IAAO, 2018, p. 29; SanPietro*, et. al,* 2019, p. 14). As a result, the PRD should be used as a "first-line indicator" or one component when measuring vertical inequity (Gloudemans, 2011, p. 8).

The Price-Related Bias ("PRB") coefficient measures and tests whether the AVM valuation-to-sales price ratios "tend to be systematically lower, higher, or steady as market value increases" (Denne, 2011, p. 4). In particular, the PRB coefficient allows for an evaluation of the strength of any vertical inequality by assessing the elasticity of a percentage change in AVM valuation-to-sales price ratios relative to the percent change in property values (where the property 'value' is an average of the AVM valuation, weighted by the median AVM valuation, and selling price[27]). Negative values of the PRB coefficient indicates regressivity, while progressivity results in a positive PRB coefficient. To illustrate, the interpretation of a PRB coefficient of -0.02 is that for a one-unit change in value, (which corresponds to doubling value in the construction of the PRB coefficient), the AVM valuation-to-sales price ratio would fall by 2 percent. As seen in Exhibit 1, a statistically significantly, different-from-zero PRB coefficient indicates vertical inequity (Gloudemans, 2011). Moreover, the IAAO (2018, p. 29) states that a PRB coefficient should generally fall between +/- 0.05, while a PRB coefficient larger than 0.10 in absolute value, with a statistically significant p-value, indicates significant vertical inequity.

While the use of the PRB mitigates the effect of outliers and eliminates the bias towards regressivity associated with the PRD, the PRB coefficient can still suffer from heteroscedasticity (Gloudemans, 2011, p. 7). Moreover, the PRB's use of a combination of selling price and the AVM valuation as the independent variable in a regression violates a fundamental regression assumption (that the independent variable is not a random variable; X is a fixed-effect, measured without error, see Neter *et. al.*, 1996, p. 44). The randomness in a regression analysis is in the dependent variable, additively modeled as a non-random mean structure coupled with a normally distributed error term.[28] Properly modeling an independent variable that contains a random component, in a regression, results in a random effects model (Neter *et. al.,* 1996, Chapter 24).

When the random component for an independent variable is disregarded, the regression suffers a "measurement error" or "error in variables" ("EIV") bias (SanPietro, *et. al.,* 2019), as the measurement error component becomes part of the overall error variance. Consequentially, the resulting PRB coefficient becomes weaker in assessing departures from vertical equity; it is biased downwards towards zero (towards vertical equity) if its estimate is positive (progressivity), while being biased upwards towards zero (towards vertical equity) if its estimate is negative (regressivity) (Carter, 2016, p. 6; Pischke, 2006; Raviv, 2016; Also see Neter *et. al.*, 1996, Section 2.11). The size of the bias is determined by amount of measurement error found in the independent variable, although Gloudemans (2011, p. 8) indicates that the use of their particular form of the proxy value variable "addresses the bias problem." Overall, a limitation in using the PRB coefficient is that it is potentially biased, due to the EIV problem, together with its p-value needed for assessing the strength of the vertical inequity requires normality and may be additionally biased, due to heteroscedasticity.

SanPietro *et. al.* (2019) evaluated several vertical inequity metrics and tests, including the PRD and PRB, and found that both the PRD and PRB detected regressivity in their example, however the PRD outperformed the PRB using the Kullback-Leibler divergence. As a result, following the advice of Carter (2016, p. 6) (who advocates computing multiple measures of vertical inequity), the calculation of the PRD statistic and the PRB coefficient, coupled with its p-value, for the entire dataset of housing sales, together with a Sales Tier Analysis using multiple tiers, for example quartiles, deciles, etc., are valuable tools that can shed light on any vertical inequities generated by an AVM.

9. An Empirical Example

In this section, details regarding a research AVM (not a commercial AVM), called the Test Valuation Model (the "TEST-VM") are provided. The TEST-VM was created by the authors to illustrate the calculations of, and relationship between, the various AVM Performance Metrics for an arbitrary house in the midwestern city of Cedar Falls, Iowa. Details regarding the TEST-VM, needed to support the subsequent discussion of the AVM Performance Metrics, are provided below.[29] The TEST-VM would require further testing, if it were used to value properties in a city other than Cedar Falls, Iowa or across many cities.

The TEST-VM is a regression model (Linne, Kane and Dell, 2000, Chapters 7-8), in which (log-scaled) selling price is regressed on a total of fifteen independent variables, including living area, number of bathrooms, date of the sale during the year, locational binary variables, distances to important externalities in the submarket and neighborhood characteristics, such as median age. These variables were chosen to align with the factors that buyers and sellers find important in this particular housing submarket (IAAO, 2018, Section 6.5). Specifically, each of these fifteen variables in the TEST-VM was statistically significant in explaining property prices, and was culled, via a backwards elimination model building strategy[30] (Neter *et. al.*, 1996, p. 435), from an initial set of fifty independent variables that included, hedonic housing characteristics, date of sale, location, distances to important features, aggregate census block level neighborhood variables, school test scores and macro-economic indicators, including the current mortgage rate at the time of sale and the quarterly revenue of the largest employer in the area (John Deere). More specific details regarding these variables can be found in Rosberg, *et. al*. (2017).

The training dataset employed consists of 53 housing sales from 2012 located in a specific submarket in Cedar Falls, Iowa. Exhibit 2 shows the relative locations of these 53 sales in the central part of Cedar Falls, denoted with circles ("○"). These 53 housing sales were identified to be in the same school district and were among more than 500 properties that were sold city-wide in 2012. An even larger dataset of over 2,000 housing sales, from 2009 to 2013 in Cedar Falls, was analyzed in Rosberg, *et. al*. (2017).

The first property that sold in 2013, in the same submarket as these 53 properties, was arbitrarily chosen to be the target property. It was built in 1951 on almost a third of an acre of land and is indicated by the filled-in box ("■"), in Exhibit 2. This property has 1,181 square feet of living area with four bedrooms and one bathroom. The TEST-VM produced a valuation of this target property as of January 1, 2013, just before the sale of the property. The TEST-VM selected the 53 housing sales, which are in the same school district and sold within one year of the valuation date for the target property (January 1, 2013), as the training dataset. Summary statistics for several of the TEST-VM's variables are seen in Exhibit 3.

EXHIBITS 2 and 3 About Here

As seen in the TEST-VM Report in Exhibit 4, the TEST-VM valuation for the target property is $159,427 on January 1, 2013. This property sold for $152,000 a few weeks later. As a result, the TEST-VM overvalued the target property by $7,427, resulting in a +4.89% sales error. The 95% regression-based confidence interval ranged from $133,033 to $191,056, a width of $58,023 or 38.2% of its selling price.[31] This 95% confidence interval is valid because the regression residuals are tested to be normally distributed (Shapiro-Wilk test p-value is 0.2701, which fails to reject the null hypothesis of normality). To demonstrate the level of transparency advocated at the end of the Introduction section, the methodology used to calculate the TEST-VM's Performance Metrics seen in Exhibits 4, 5 and 6 is disclosed below.

EXHIBIT 4 About Here

The TEST-VM's Performance Metrics for to this target property, including the FSD of 19.6 in Exhibit 5, are calculated reusing the training dataset of 53 housing sales as the holdout dataset, via the Predicted Residual Sum of Squares ("PRESS") cross-validation technique. In particular, a TEST-VM PRESS predicted value for each of the 53 housing sales is obtained by fitting the original regression model, but with the *i*th observation withheld (Neter *et. al.*, 1996, p. 345). Specifically, for the Cedar Falls data, the TEST-VM regression is re-run 53 times,[32] each time omitting a different observation. Then, 53 predict values are generated, one for each of the successively omitted observations. The PRESS leave-one-out, cross-validation technique allows for model assessment using "an analysis independent from the data used in developing the model" (IAAO, 2018, p. 15). Furthermore, the PRESS technique avoids sacrificing valid observations needed to fit the model, by reusing the training dataset as the holdout set. In other words, it avoids the need to partition the 53 housing sales into mutually exclusive training and holdout datasets.

The FSD of 19.6, which is the standard deviation of the percentage sales errors for the 53 Cedar Falls housing sales, is reported in Exhibit 4. Additional AVM Performance Metrics associated with the target property are reported in Exhibit 5, while the Sales Tier Analysis is presented in Exhibit 6. The mean sales and percentage sales errors of $ 414 and 2.18%, respectively, in Exhibit 5, are not statistically significantly different from zero (t-test p-values of 0.9140 and 0.4224, respectively; see Ott and Longnecker, 2016, Section 5.7). Thus, the TEST-VM is providing unbiased estimates of market value. The median sales and percentage sales errors of $-2,936 and

-2.04%, respectively, are also not statistically significantly different from zero (p-value for the sign test is 0.4101; see Ott and Longnecker, 2016, Section 5.9).

EXHIBITS 5 and 6 About Here

Comparing to the AVM performance thresholds seen in Exhibit 1, the MAPE of 8.5 for the TEST-VM reflects 'Strong' AVM performance, according to Kirchmeyer and Staas (2008), while the FSD of 19.6 results in a 'Medium Confidence' label from Freddie Mac (2019a).

Comparing the TEST-VM's PEs in Exhibit 5 to each of the Kirchmeyer error buckets criteria (PE10/50 and PE15/70), the TEST-VM exceeds the PE10/50 (54.7%) while it falls short of the PE15/70 criterion (62.2%). The corresponding Failure Rates for the TEST-VM at 10 and 15 percent are 45.3% and 37.8%. Further study of the individual errors reveals that the TEST-VM performs poorly for the four least expensive houses (37.7% sales error) and for four of the six most expensive houses (29.1% sales error, for these four), which contributes to its high Failure Magnitudes, for example, 25.6% at +/- 10 percent. In other words, for the 24 Cedar Falls housing sales (out of 53, for its 45.3% Failure Rate at +/- 10 percent) where the TEST-VM valuation differs from its corresponding selling price by more than 10 percent, the mean (absolute percentage sales) errors is 25.6 percent. Also, the Failure MAPE at +/- 10 percent of 22.2% is close to the Failure Magnitude (25.6%), and both of which are much larger than the 10% error bucket, reinforcing that the TEST-VM fails to predict selling prices accurately for a substantial number of sales.

In addition, seven of the 53 (13.2%) housing sales have TEST-VM valuations that are more than 20 percent larger than their respective selling prices, which fails AVMetrics' Right Tail 20% criterion, while the TEST-VM exceeds the +/- one FSD 68.26% interpretation criterion (73.6%). Lastly, the Sales Tier Analysis in Exhibit 6 confirms that these large sales errors occur for the highest and lowest priced houses, while the TEST-VM performs reasonably well for houses in the $115,000 to $180,000 price range.

The TEST-VM meets the IAAO (2018) standard for horizontal equity, as its COD value of 14.5 falls within the range of 5 and 20, but fails Rossini and Kershaw's (2008) threshold for the COV, due to its high Failure Magnitudes affecting the standard deviation. In addition, the PRD value of 1.0188 in Exhibit 5 falls within the IAAO's recommended range of 0.98 and 1.03, while the PRB

coefficient of -0.047 is not statistically significant (p-value = 0.5886). Both the PRD value and PRB coefficient do suggest a lean towards regressivity, which can be seen in the Sales Tier Analysis in Exhibit 6. The TEST-VM is slightly overvaluing inexpensive properties more so than it undervalues expensive properties, illustrated by the 14.47% overvaluation of the lowest selling price quartile of properties, in comparison to the TEST-VM's undervaluation, by 7.74%, of the highest price quartile. The other sales tiers (halves, thirds and quintiles) in Exhibit 6 reinforce that the TEST-VM's slight lean towards regressivity is not an artifact owning to the subjective choice of tier. The Sales Tier Analysis also shows that the TEST-VM performs reasonably well for properties close to the median and mean selling prices of $130,000 and $143,767, respectively, but its performance, as with many statistical models, declines the more a property's selling price differs from the mean and median.[33]

10. Conclusions and Recommendations

A thorough exposition of the calculation, interpretation and evaluation AVM Performance Metrics, including FSDs, Confidence Scores, horizontal and vertical inequity, and error buckets is presented. In addition, the Failure Rate, Failure Magnitude and Failure MAPE, which evaluate how the AVM performs in the tails of its predictions, are examined. In particular, the Failure Rate counts (the frequency of) properties for which the AVM fails to predict selling prices accurately, to within +/- 5, 10, 15, and/or 20 percent, while the Failure Magnitude and Failure MAPE are the mean and median absolute percentage sales error for those poorly predicted sales. The calculation and relationships between these AVM Performance Metrics using the TEST-VM, which is a research valuation model created to estimate the market value of an arbitrarily selected house in a medium-sized, midwestern college city, is demonstrated.

In addition, the authors advocate for more transparency and uniformity in the AVM Performance Metrics typically reported by commercial AVMs. AVM providers should report the mean and median percentage sales errors, along with a detailed description of the actual sales data, i.e. the properties used to produce the mean and median percentage sales errors. Moreover, the authors urge AVM providers to adopt common definitions for AVM Performance Metrics, especially the FSD and Confidence Score. Lastly, the AVM vendor should allow the client to pre-specify their desired confidence level, when reporting High/Low values, or at least be transparent and inform the client that the default range, based upon the FSD's industry interpretation, is 68.26%.

11. Glossary

Absolute Percentage Sales Error – The absolute value of the quantity: (AVM value minus selling price)/selling price, for a target property.

Accuracy – Accuracy measures the level of conformity of an estimate to a known benchmark. The mean and median (percentage) sales errors are AVM Performance Metrics that measure the accuracy of the AVM.

Automated Valuation Model (AVM) – A computer software program that produces an AVM valuation, along with AVM Performance Metrics, for a single target property, given the address of the target property and property sales and characteristics data.

AVM report – The document provided to a client of an AVM vendor containing the AVM valuation. It typically also includes the high and low ranges of value, together with AVM Performance Metrics.

AVM valuation – An estimate of market value of a target property produced by an AVM.

AVM valuation date – The date, either contemporaneously or retrospectively, for which the AVM valuation is produced.

AVM Performance Metrics – The set of calculations, typically involving the (percentage) sales errors, that measure the accuracy and/or precision of the AVM valuations.

Coefficient of Determination (R-squared) – The percent of variability explained by a statistical (regression) model.

Coefficient of Dispersion (COD) – An AVM Performance Metric that measures precision. The COD measures the average percentage deviation of the AVM valuation-to-sales price ratios from the median AVM valuation-to-sales price ratio.

Coefficient of Variation (COV) – An AVM Performance Metric that measures a combination of accuracy and precision. The COV is the standard deviation divided by the mean AVM valuation-to-price ratio.

Confidence Interval – A range of values in which the target parameter is likely enclosed, where likelihood is measured by the confidence level.

Confidence Score – An AVM Performance Metric that varies by AVM vendor. Confidence Scores should not be used to compare the performance of one AVM vendor to another, for the same target property.

Cross-Validation –a statistical procedure to assess how well a model, fit using a training dataset, predicts outcomes in a holdout dataset, drawn from the same population as the training dataset.

Failure Magnitude – An AVM Performance Metric that measures a combination of accuracy and precision. The mean absolute error percentage sales error for properties for which the AVM fails to predict selling prices to within +/- (a given percentage, for example, +/- 10%).

Failure MAPE – An AVM Performance Metric that measures a combination of accuracy and precision. The median absolute error percentage sales error for properties for which the AVM fails to predict selling prices to within +/- (a given percentage, for example, +/- 10%).

Failure Rate – An AVM Performance Metric that measures a combination of accuracy and precision. The percentage of properties for which the AVM fails to predict selling prices to within +/- (a given percentage, for example, +/- 10%).

Forecast Standard Deviation (FSD) – An AVM Performance Metric that measures precision. Specifically, the FSD is the standard deviation of the percentage sales errors.

Hit Rate – The percentage of properties that have been valued by the AVM.

Holdout Dataset – A dataset, drawn from the same population as the training dataset, that is not used to calculate the AVM valuations. Cross-validation techniques allow sales in the training dataset to be reused as the holdout set.

Horizontal Equity – The principle that people in the same circumstances should be treated the same or that similar properties should have similar tax assessed values.

Mean Absolute Percentage Sales Error – An AVM Performance Metric that measures precision; specifically, the mean or average of a set of absolute percentage sales errors.

Median Absolute Percentage (Sales) Error (MAPE) – An AVM Performance Metric that measures precision. The MAPE is the median of a set of absolute percentage sales errors.

Mean (Percentage) Sales Error – An AVM Performance Metric that measures accuracy. The mean (percentage) sales error is the mean or average of a set of (percentage) sales errors.

Median (Percentage) Sales Error – An AVM Performance Metric that measures accuracy. The median (percentage) sales error is the median of a set of (percentage) sales errors.

Percent (Predicted) Error (PE) Bucket – An AVM Performance Metric that measures a combination of accuracy and precision. The PE is the percentage of properties for which the AVM predicts selling prices to within +/- a given percentage. The complement of the Failure Rate.

Percentage Sales Error – An AVM Performance Metric that measures accuracy; specifically, the AVM value minus selling price for a particular target property, which is then divided by the selling price.

PRESS Predicted Value –The Predicted Residual Sum of Squares (PRESS) predicted value is the (regression) model's predicted value of the *i*th observation in the training dataset, obtained by fitting the original regression model with the *i*th observation withheld.

Precision – Precision measures the consistency of a set of observations, typically calculated with respect to the center. As an illustration, the standard deviation is a measure of precision using the center (mean), while the range is a precision metric that does not use the center.

Price Related Bias (PRB) Coefficient – An AVM Performance Metric that measures precision. The PRB is the slope in the regression of the AVM valuation-to-sales price ratios (adjusted by the median ratio) on a proxy variable that measures the value of the house. The proxy value variable is the natural log of the average of selling price and AVM valuation (all divided by 0.693).

Price Related Differential (PRD) – An AVM Performance Metric that measures precision. The PRD is the mean (valuation-to-selling price) ratio divided by the weighted (by selling prices) mean ratio

Progressivity – Lower price properties are undervalued at a higher rate than high priced properties are overvalued.

Regression – A statistical method that measures the relationship, if any, between a dependent variable and a set of independent variables that are thought to influence (correlate with) the dependent variable. For housing data, (natural log of) selling prices are often regressed on a set of variables, including house characteristics (i.e., size, age, condition), location, neighborhood variables, and any market conditions that are expected to influence its value (such as the closure of a school, the state of the economy, interest rates, etc.).

Regressivity –Lower priced properties are overvalued more than higher priced properties are undervalued.

RMSE – The square root of the mean squared error.

Sales Error – The AVM value minus selling price for a target property.

Sales Tier Analysis – A comparison of the percent sales errors (or another metric) after creating price bins or tiers. These stratified price bins may be chosen using specific monetary values or they can be chosen to provide (roughly) equal numbers of sales in each bin.

Training Dataset – The dataset used to fit (estimate the parameters of) a model (AVM).

Unbiasedness – A statistic is unbiased if its expected value equals the target parameter. For housing sales data, an AVM valuation will be unbiased if its mean sales error equals zero.

Vertical Equity – The principle that people with higher income (wealth) should pay higher taxes or that higher valued properties should bear proportionally higher assessed values.

12. References

Abdi, H. (2010). Coefficient of Variation. In Neil Salkind (Ed.), *Encyclopedia of Research*

Design. Thousand Oaks, CA: Sage. 5 pages. Found at www.utdallas.edu/~herve/abdi-cv2010-pretty.pdf .

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.

Benmamqun, M. (2006). Bootstrap confidence intervals and Gloudemans' COD tolerance test using SPSS and Stata. *Journal of Property Tax Assessment & Administration.* **3**(4), pp 51-58.

Bennedsen, J, and M. E. Caspersen. (2007). Failure rates in introductory programming. *SIGCSE Bull*, **39**(2). pp 32-36.

Boyle, M. A. and K. A. Kiel. (2001). A survey of house prices hedonic studies of the impact of environmental externalities. *Journal of Real Estate Literature*. **9**(2), pp 117-144.

Carter, M.J. (2016). Methods for Determining Vertical Inequity in Mass Appraisals. *Fair & Equitable.* **14(**6), pp 3-8.

Collateral Assessment & Technologies Committee. (2009). Best Practices in Automated Valuation Model (AVM) Validation. Found at: www.federalreserve.gov/SECRS/2009/February/20090202/OP-1338/OP-1338_012009_179_286926910293_1.pdf .

Collateral Risk Management Consortium. (2003). The CRC Guide to Automated Valuation Model (AVM) Performance. Found at professional.sauder.ubc.ca/re_creditprogram/course_resources/courses/content/344/CRC_AVM.pdf .

CoreLogic. (2010). MLS Data in AVMs: The Case for – and Against. Found at: www.corelogic.com/research/avm-industry/mls-data-in-avms.pdf .

CoreLogic. (2011). Automated Valuation Model Testing. Whitepaper. Found at www.corelogic.com/downloadable-docs/automated-valuation-model-testing.pdf .

CoreLogic. (2014). AVM FAQs. Found at www.corelogic.com/downloadable-docs/avm-faqs.pdf .

CoreLogic. (2016). Automated Valuation Solutions. Found at www.corelogic.com/downloadable-docs/1-avmcval-0610-03.pdf .

CoreLogic. (2017). Forecast Standard Deviation & AVM Confidence Scores Found at: www.corelogic.com/downloadable-docs/fsd-and-avm-confidence.pdf .

CoreLogic. (2018). CoreLogic Launches new Total Home Value Suite of Automated Valuation Models. Found at www.corelogic.com/news/corelogic-launches-new-total-home-value-suite-of-automated-valuation-models.aspx .

Cowles, M. and C. Davis. (1982). On the Origins of the .05 Level of Statistical Significance. *American Psychologist*. **37**(5), pp 553-558.

Denne, R.C. (2011). The PRB and Other Potential Successors to the Flawed PRD as a Measure of Vertical Inequity. *Fair & Equitable.* **9**(11), pp 3-11.

DeSimone, B. (2015). What are Comps? Understanding a Key Real Estate Tool. Found at www.zillow.com/blog/what-are-comps-179631/ .

Epley, D. (2017). The Need to Reference Automatic Valuation Models to The Valuation Process. *Journal of Real Estate Literature*. **25**(1), pp 237-251.

FDIC. (2018). Appraisal Threshold for Residential Real Estate Loans. Found at www.fdic.gov/news/news/financial/2018/fil18076.html .

Filho, C.M., Z.S. Ong and H.C. Koh. (2005). Determinants of property price: A decision tree approach. *Urban Studies*. **43**(12), pp 2301-2315.

Filho, C.M. and O. Bin (2005). Estimation of hedonic price functions via additive nonparametric regression. *Empirical Economics*. **30**(1), pp 93-114.

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 www.ficonsulting.com/avms-have-feelings-too/ .

Freddie Mac. (2019a). Confidence Levels. Found at www.freddiemac.com/hve/confidence_scores.html .

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

Gloudemans, R.J. (2001). Confidence Intervals for the Coefficient of Dispersion. *Assessment Journal*. **8**(6), pp 23-27.

Gloudemans, R.J. (2011). The Coefficient of Price-Related-Bias: A Measure of Vertical Inequity. *Fair & Equitable.* **9**(8), pp 3-8.

Gordon, D. (2005). Metrics Matter. Found at www.freddiemac.com/hve/pdf/dougwhitepaper_metricsmatter.pdf .

Grover, R. (2016). Mass valuations. *Journal of Property Valuation & Investment*. **34**(2), pp 191-204.

IAAO. (2010). *Standard on Verification and Adjustment of Sales*. Kansas City, MO. 18 pages. Found at www.iaao.org/media/standards/Verification_Adjustment_of_Sales.pdf .

IAAO. (2013). *Standard on Ratio Studies*, Kansas City, MO. 63 pages. Found at www.iaao.org/media/standards/Standard_on_Ratio_Studies.pdf .

IAAO. (2018). *Standard on Automated Valuation Models (AVMs)*, Kansas City, MO. 35 pages. Found at 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. (1986). The Nearest Neighbors Appraisal Technique: An Alternative to the Adjustment Grid Methods. *American Real Estate and Urban Economics Association Journal*. **14**(2), pp 274-286.

Isakson, H. R., Ecker, M. D. and L. Kennedy. (2019). Principles for Calculating AVM Performance Metrics. Technical Report. Mathematics Department, University of Northern Iowa. Found at www.math.uni.edu/~ecker/research .

Kaye, D. and D. Freedman. (2011). Reference Guide on Statistics. Published in *Reference Manual on Scientific Evidence*, 3rd edition, National Academies Press.

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.

Lane, B. (2018). Amrock ordered to pay $706 million for stealing trade secrets from HouseCanary. Found at www.housingwire.com/articles/42762-amrock-ordered-to-pay-706-million-for-stealing-trade-secrets-from-propertycanary .

Linne, M.R., Kane, M.S. and G. Dell. (2000). A Guide to Appraisal Valuation Modeling. Appraisal Institute. 95 pages.

Matysiak, G. (2017). The Accuracy of Automated Valuation Models (AVMs). Report for TEGoVA. 23 pages.

Metzner, S. and A. Kindt. (2018). Determination of the parameters of automated valuation models for the hedonic property valuation of residential properties. *International Journal of Housing Markets and Analysis*. **11**(1). pp 73-100.

Montgomery, D.C., Peck, E.A. and G.G. Vining. (2001). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Moore, J. W. (2006). Performance comparison of automated valuation models. *Journal of Property Tax Assessment & Administration*. **3**(1), pp 43-59.

Morris, A. S. and R. Langari. (2016). *Measurement and Instrumentation: Theory and Application.* Second Edition. Elsevier, Inc.

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 .

Neter, J., Kutner, M. H., Nachtsheim, C. J. and W. Wasserman. (1996). *Applied Linear Statistical Models*. Irwin. 1,408 pages.

Ott, R. L. and M. Longnecker. (2016). *An Introduction to Statistical Methods and Data Analysis, 7th edition*. Brooks/Cole. 1,174 pages.

Platt, H. D. and M. B. Platt. (1990). Development of a Class of Stable Predictive Variables: The Case of Bankruptcy Prediction. *Journal of Business Finance and Accounting*. **17**(1), pp 31-51.

Pischke, S. (2006). Lecture notes on Measurement Error. Found at econ.lse.ac.uk/staff/spischke/ec524/Merr_new.pdf .

Raviv, E. (2016). Measurement Error Bias. Found at eranraviv.com/measurement-error-bias/ .

Rosberg, A., Isakson, H., Ecker, M. and T. Strauss. (2017). Beyond Standardized Test Scores: The Impact of a Public School Closure on Property Prices. *Journal of Housing Research*. **26**(2). pp 119-135.

RPR. (2018). What is an AVM or RVM®. Confidence Score. Found at support.narrpr.com/hc/en-us/articles/204964670-What-is-an-AVM-or-RVM-confidence-score- .

Rossini, P. and P. Kershaw. (2008). Automated Valuation Model Accuracy: Some Empirical Testing. 14th Pacific Rim Real Estate Society Conference. 12 pages.

SanPietro, F.J., Sunderman, M.S., Radetskiy, E. and B. Janz. (2019). Using Information Theory to Evaluate Measures of Vertical Inequity in Property Taxation. *Journal of Property Tax Assessment and Administration*. **16**(1), pp 5-23.

Shapiro, S. S. and M. B. Wilk. (1965). An analysis of variance test for normality (complete samples). *Biometrika*. **52**(3 and 4). pp 591-611.

Sirmans, S., Macpherson, D., and E. Zietz. (2005). The composition of hedonic pricing models, *Journal of Real Estate Literature*. **13**(1), pp 3-43.

Theriault, M., F. Des Rosers and F. Joerin. (2005). Modelling accessibility to urban services using fuzzy logic. A Comparative analysis of two methods. *Journal of Property Investment and Finance*. **23**(1), pp 22-54.

Veros. (2017). When do AVMs make the most sense? Found at www.housingwire.com/articles/40935-when-do-avms-make-the-most-sense?page=2 .

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

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.

Exhibit 1: AVM Performance Metric thresholds suggested in the literature.

AVM Performance Metric	Reference	AVM Performance Metric Threshold (with comment)
MAPE*	Kirchmeyer and Staas (2008) Rossini and Kershaw (2008)	10 (Strong) and 15 (Acceptable) 10 (Reasonable) and 13 (Abs Min)
FSD*	Freddie Mac (2019a)	<13 (High); 13-20 (Medium); >20 (Low)
Error Buckets**		
+/- 10%	Rossini and Kershaw (2008) Kirchmeyer (2004) AVMetrics (2018) MBA (2019) Veros (2017)	50% (Abs Min) and 65% (Reasonable) 50% 68-70% 70% 80-90% (Top-Tier AVM)
+/- 15%	Rossini and Kershaw (2008) Kirchmeyer (2004) AVMetrics (2018)	65% (Abs Min) and 80% (Reasonable) 70% 80%
+/- 20%	Rossini and Kershaw (2008)	80% (Abs Min) and 90% (Reasonable)
Right Tail 20%	AVMetrics (2018)	10%
Failure Rates**		
+/- 10%	Rossini and Kershaw (2008) Kirchmeyer (2004) AVMetrics (2018) MBA (2019) Veros (2017)	35% (Reasonable) and 50% (Abs Max) 50% 30-32% 30% 10-20% (Top Tier AVM)
+/- 15%	Rossini and Kershaw (2008) Kirchmeyer (2004) AVMetrics (2018)	20% (Reasonable) and 35% (Abs Max) 30% 20%
+/- 20%	Rossini and Kershaw (2008)	10% (Reasonable) and 20% (Abs Max)
IAAO Metrics		
COV**	Rossini and Kershaw (2008)	13 (Reasonable) and 17 (Abs Max)
COD*	IAAO (2018) Rossini and Kershaw (2008)	5 to 20 for residential properties 10 (Reasonable) and 13 (Abs Max)
PRD*	IAAO (2018)	0.98 to 1.03
PRB*	Gloudemans (2011) IAAO (2018)	Statistically significantly different from zero +/- 0.05

* AVM Performance Metric measuring Precision

** AVM Performance Metric measuring both Accuracy and Precision

Exhibit 2: Relative Locations of the Property Sales in Cedar Falls, Iowa

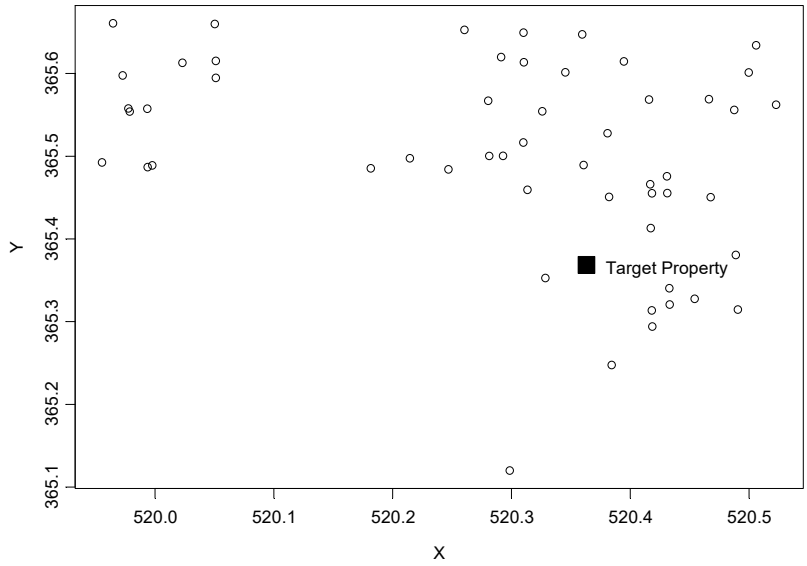

Exhibit 3: Summary Statistics for 53 housing sales in the Cedar Falls housing dataset. Areas are reported in square feet and distances in miles.

Variable	Mean	Median	Standard Deviation	Low	High
Selling Price	$ 143,767	$ 130,000	$ 34,585	$ 72,000	$ 223,900
Living Area	934.4	900	268.5	480	1,682
Non-Main Living Area	407.8	428	421.4	0	1,665
Number of Bathrooms	1.13	1	0.34	1	2
Distance to University	0.660	0.658	0.14	0.378	0.991
Distance to High School	0.681	0.616	0.24	0.252	1.274

Exhibit 4: TEST-VM Report for the Target Property. The exact address, location of the property on a map and owner information are omitted for confidentiality reasons. A more detailed description of the sales data used to produce the valuation is provided by Exhibits 2 and 3 and in the text.

Metric	Value
County	Black Hawk
Land Use	Single Family Residential
TEST-VM Valuation	$ 159,427
TEST-VM Valuation Date	January 1, 2013
Selling Price	$ 152,000
Sales Error	$ 7,427
Percent Sales Error	4.89%
FSD	19.6
Lower 95% Confidence Interval	$ 133,033
Upper 95% Confidence Interval	$ 191,056
95% Confidence Interval Width	$ 58,023
Number of Housing Sales Used	53
Mean Selling Price of the 53 Housing Sales	$ 143,767
Median Selling Price of the 53 Housing Sales	$ 130,000

Exhibit 5: TEST-VM Report - AVM Performance Metrics

AVM Performance Metric	Value
Mean Sales Error	$ 416
Mean Percentage Sales Error	2.18%
Median Sales Error	$ -2,939
Median Percentage Sales Error	-2.04%
Mean Absolute Sales Error	$ 20,516
Median Absolute Sales Error	$ 14,256
Mean Absolute Percentage (Sales) Error	14.4%
Median Absolute Percentage (Sales) Error (MAPE)	8.5%
FSD	19.6
Average Confidence Interval Width	$ 48,927
Median Confidence Interval Width	$ 42,841
Mean (CI Width/Selling Price)	34.2%
Percent of CIs with a Width of $50,000 or less	34/53 for 64.2%
Percent of CIs with a Width of $100,000 or less	52/53 for 98.1%
Percent of Estimates within +/- 5% (PE5)	14/53 for 26.4%
Failure Rate at +/- 5%	39/53 for 73.6%
Failure Magnitude at +/- 5%	18.7%
Failure MAPE at +/- 5%	16.0%
Percent of Estimates within +/- 10% (PE10)	29/53 for 54.7%
Failure Rate at +/- 10%	24/53 for 45.3%
Failure Magnitude at +/- 10%	25.6%
Failure MAPE at +/- 10%	22.2%
Percent of Estimates within +/- 15% (PE15)	33/53 for 62.3%
Failure Rate at +/- 15%	20/53 for 37.7%
Failure Magnitude at +/- 15%	28.2%
Failure MAPE at +/- 15%	24.0%
Percent of Estimates within +/- 1 FSD (+/- 19.6%)	39/53 for 73.6%
Percent of Estimates within +/- 20% (PE20)	39/53 for 73.6%
Failure Rate at +/- 20%	14/53 for 26.4%
Failure Magnitude at +/- 20%	32.7%
Failure MAPE at +/- 20%	30.3%
Right Tail 20%	7/53 for 13.2%
Coefficient of Variation (COV) of TEST-VM/Sale Price	0.19199 or 19.2
Unbiased COV estimator	0.19289 or 19.3
Coefficient of Dispersion (COD) of TEST-VM/Sale Price	14.5
PRD of TEST-VM/Sale Price	1.0188
PRB Coefficient	-0.047
P-value for PRB Coefficient	0.5886
Regression R-Squared (Coefficient of Determination)	0.7272
Adjusted R-Squared	0.6165
RMSE	0.1538

Exhibit 6: TEST-VM Sales Tier Analysis

Selling Prices

Tier	Number of Properties	Lower Bound	Upper Bound	Mean Percent Sales Error
All Comps	53	$ 72,000	$ 223,900	2.18%
Halves: 1	26	$ 72,000	130,000	8.17
2	27	130,000	223,900	-2.48
Thirds: 1	18	$ 72,000	125,000	10.27
2	17	127,500	158,750	0.66
3	18	163,000	223,900	-5.34
Quartiles: 1	13	$ 72,000	119,000	14.47
2	13	122,000	130,000	1.84
3	13	130,000	171,420	-0.80
4	14	174,500	223,900	-7.74
Quintiles: 1	10	$ 72,000	115,000	17.50
2	11	116,000	128,000	-0.43
3	11	128,000	150,000	1.00
4	11	156,000	181,500	-2.55
5	10	182,900	223,900	-6.12

13. Endnotes

[1] Throughout this work, the term "AVM" will be used to refer to commercial or professional grade AVMs. 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 (2019, p. 9-10).

[2] See https://www.corelogic.com/about-us/our-company.aspx .

[3] The IAAO (2018, p. 6) defines fair market value as "the amount in terms of money that a well-informed buyer is justified in paying and a well-informed seller is justified in accepting for a property in an open and competitive market, assuming the parties are acting without undue compulsion."

[4] For example, recently Amroc was ordered to pay HouseCanary $706 million for the theft of trade secrets from HouseCanary, an AVM vendor (Lane, 2018).

[5] *See* www.platdata.com/pub/download/pdf/products/AVM%20Product%20Descriptions.pdf for a description of CoreLogic's individual AVMs taken from their marketing materials.

[6] For the purposes of this work, we assume that the selling price is known and that it meets all of the requirements to represent the best indicator of the market value of the property. See IAAO (2013, Appendix A – Sales Validation Guidelines). Also, CoreLogic (2010, p. 2) states that "a purchase price is really the best (and only) benchmark for a property's true value." The Collateral Assessment & Technologies Committee (2009, p. 11-12) of the Real Estate Information Professionals Association states, "[a]rms-length purchase money transactions provide the best indicator of value since there is a willing buyer and a willing seller in the open market (a.k.a. arms-length transaction)."

[7] Comparable sales "are located in the same area and very similar in size, condition and features" (DeSimone, 2015) as the target property.

[8] An AVM "hit" generally means that the AVM has found the target property in its database of all properties in the market area. It may or may not mean that the AVM was able to return a meaningful valuation of the property.

[9] For sample AVM reports, see www.avantus.com/samples/AvantusAVM.pdf or www.corelogic.com/downloadable-docs/11-geoavm-va-0814-00_geoavmcore_sample.pdf or www.eamc1.com/files/HomeJunction.pdf .

[10] The use of selling price as market value is so ubiquitous in the academic literature that authors almost never explicitly state that the selling price of a house represents its market value. Instead, academicians simply *use* the selling price in an arms-length transaction as the market value of a property. For example, Sirmans, Macpherson, and Zietz (2005), cite over 100 articles published in real estate academic journals that use selling price as the value of residential property in hedonic pricing models. Other peer-reviewed research, that reviews academic papers treating selling prices as market values, include Metzner and Kindt (2018) and Boyle and Kiel (2001).

[11] The comparison of an AVM valuation to a "benchmark value, typically a recently observed purchase price, is one way to assess model accuracy" MBA (2019, p. 15). As noted by a reviewer, the benchmark value is a choice made by the researcher or AVM provider. In this work, the selling price for an arm's length transaction in a competitive environment is selected as the benchmark.

[12] Note that the sales error is of the form Predicted – Actual, to facilitate an overvalued/undervalued interpretation of a (percentage) sales error, in contrast to a standard statistical residual which takes the form Actual – Predicted.

[13] A holdout dataset is a dataset that is drawn from the same population as the training dataset, but the holdout dataset is not directly used to calculate the AVM valuation. Cross-validation techniques, such as the Predicted Residual Sum of Squares ("PRESS") technique can allow the sales in the training dataset to be properly reused as the holdout set.

[14] Unbiased has a formal statistical definition. See Montgomery *et. al.,* (2001). p. 20. A statistic is unbiased if its expected value equals the target parameter.

[15] As an illustration, the standard deviation is a measure of precision using the center (mean), while the range is a precision metric that does not use the center.

[16] The CoreLogic white paper, Gayler *et. al.* (2015), defines the FSD as the standard deviation of the percentage sales errors. This Gaylor definition is used for any FSD calculation perform by the authors in this work.

[17] Rossini and Kershaw (2008) also provide FSD performance thresholds, but they are not discussed in this work, because their FSD definition (p. 3) does not involve selling prices.

[18] The IAAO (2018, p. 14) cautions that choosing a lower confidence level, resulting in a narrower confidence interval, may appear to provide more reliable results, but at the expense of added uncertainty.

[19] If the intended use of the AVM is to statistically decide if the AVM is undervaluing (or overvaluing) properties, then a one-tailed hypothesis test may be warranted, in contrast to a (two-tailed) confidence interval.

[20] Common intended uses of AVMs include performing appraisal quality control, mortgage underwriting (including home equity loans and refinancing decisions), loss mitigation and credit risk management activities for financial institutions, valuing by assessment jurisdictions and providing estimates to the public (IAAO, 2018, section 6.2)

[21] Regression AVMs can routinely provide a High/Low range of values, at any desired level of confidence, from the theoretical sampling distribution of the predicted value (Neter *et. al.*, 1996, Section 4.2), which does not involve the FSD.

[22] For 90% confidence, use +/- 1.645×FSD. For 99% confidence, use +/- 2.576×FSD.

[23] The subsequent choice of $50,000 and $100,000 for the Cedar Falls sales may not be applicable for houses in all geographic regions. For example, because the median house price in New York City is over a million dollars, then a $250,000 or $400,000 confidence interval width might be more appropriate.

[24] Note that an AVM cannot be properly vetted using sales that fail to represent market values.

[25] Note that the COD should not be confused with the Coefficient of Determination or regression R-squared, which measures the percent of variability in a dependent variable explained by a regression model. The regression AVM's R^2 computes the squared correlation between selling prices and AVM valuations, assuming the regression contains an intercept term. See Neter, *et. al.* (1996, Section 6.5).

[26] Heteroscedasticity, or lack of constant variance, is a common ailment in statistical (regression) models, whereby the variability of one variable is correlated with another. For example, older houses tend to have more variability in price than newer houses, assuming the houses examined are otherwise relatively comparable. Heteroscedasticity means that the variability of house prices is not constant with respect to age. Formal statistical hypothesis tests for heteroscedasticity, for example, the Breusch-Pagan test or Modified Levene test (Neter *et. al.*, 1996, Section 3.6), require the user to specify the exact form of the anticipated heteroscedasticity, for example, linear with X (fan-shaped) or the divide the dataset into the groups that are expected to display different

levels of variability. Heteroscedasticity in the PRD/PRB statistic means that higher valued houses tend to have more variability in their assessment ratios than lower valued houses.

[27] SanPietro *et. al.* (2019, p. 9) state that the proxy *value* variable, the average of selling price and AVM valuation, "more accurately reflects the *unobservable* market value and thus addresses the EIV [Error in Variables] problem."

[28] Normality is not required to fit the regression model (provide point estimates). For example, normality is not required to estimate the slope(s) and intercept, together with provide predicted values. However, normality is required for any formal statistical inference, including calculating p-values for the slope(s), performing a hypotheses test for the intercept, and/or providing confidence or prediction intervals for predicted values. See Neter *et. al.* (1996), Section 1.8.

[29] These TEST-VM SAS code can be found at: github.com/markecker/Exposition-of-AVM-Performance-Metrics .

[30] The backwards elimination regression starts with all available independent variables and first eliminates the variable with the largest insignificant p-value (in this work, larger than α=0.1). Then the regression is refit, after omitting the most insignificant variable. The process is then iterated, removing insignificant variables, one at a time, until all remaining variables are statistically significant (or all variables have been removed).

[31] The confidence interval of $133,033 to $191,056 reflects the average selling price of a house with the target property's characteristics. Conceptually, it is a confidence interval for the mean selling price. In contrast, the TEST-VM's 95% prediction interval for the target property ranges from $111,179 to $228,612, producing a $117,433 prediction interval width. Confidence intervals for the mean are narrower than prediction intervals for a new observation. The end user typically makes the philosophical choice in whether to use a confidence interval for the non-random mean or a prediction interval for a random new observation. See Neter *et. al.* (1996), Sections 4.2 and 4.3.

[32] Technically, re-running the TEST-VM regression $n=53$ times is not required, as the PRESS predicted value can be calculated using the original regression that valued the target property. See Montgomery *et. al.* (2001). pp. 598-600, together with the SAS file TEST-VM.txt.

[33] "AVMs perform best in the case of houses that are of average quality and size with typical attributes in an area with many sales." MBA (2019, p. 10).