



Mercury Network

AVM Validation

Evaluating AVM performance

The responsible use of Automated Valuation Models in any application begins with a thorough understanding of the models' performance in absolute and relative terms. This document discusses the methodology used by Mercury Network to evaluate AVM performance.

Table of contents

Introduction	3
Process	4
01 Benchmark data	5
02 Data gathering	6
03 Data combination	6
04 Data analysis	7
05 Model scoring and cascade development	9
Scoring and model preference table	10
Cascade testing	11
Contact information	11

Introduction

The purpose of this document is to detail the process employed by Mercury Network to test Automated Valuation Model (AVM) performance and develop an AVM cascade(s) from the performance data on the AVMs.

The use of AVMs in the mortgage industry is widespread and, in many cases, is related to a lending decision. Regardless of the use, it is a critical component of any risk management program that AVM performance be understood so that unnecessary risk is avoided.

AVMs are mathematical or statistical models which use home characteristic data to estimate the current value of an individual property. AVMs are comprised of two basic components: the data used to fuel them and the analytics which turn the data into an estimate of value.

In order to understand AVM performance, the models must be tested and they must be tested using the right data, the right analysis, and by personnel who are familiar with AVM technology and how to assess the models' performance. This document will describe the process used by Mercury Network to evaluate AVM performance.

Process

Mercury Network tests 21 AVMs on a continuous basis to evaluate how well they perform absolutely and relative to one another. Nine of these AVMs are commercially available for use in a cascade. There are several steps to the process for AVM evaluation including identifying good benchmark data, creating test files, combining the AVM estimates with the benchmarks, and analyzing the data.

The table below lists the vendors and their respective models which are included in the AVM test results.

AVM Vendor	Model(s)
Equifax	AVM Insight
Freddie Mac	HVE
HouseCanary	HouseCanary Value Report
Black Knight Financial Services	SiteX, RVM, ValueSure
CoreLogic	PASS, VP4
RELAR	RELAR
Veros	VeroValue

01 Benchmark data

The data used as a benchmark against which the AVMs' estimated value will be compared is perhaps the most crucial aspect of AVM validation. Typically, it is a recent sale price against which the AVM estimate of value is compared. Some lenders are large enough that they generate sufficient data internally. But, for those organizations which do not have sufficient internal data, purchasing benchmark data from a third party may be necessary.

Without good, clean benchmark data it is almost impossible to definitively evaluate AVM performance. There are two key data issues which exist in AVM validation today. First is the use of closed loan transactions which have been recorded and, subsequently, aggregated by one of the national data aggregators. One of the easiest ways to get benchmark data if you cannot generate it internally is to purchase it from a data aggregator. Although this provides a large number of transactions to use in testing, many of the benchmark (sales prices) values will be known to the AVMs being tested. This happens because all AVMs purchase recorded sales transactions to fuel their models. Once the AVM provider has the same data that you are testing, they know the sale price of the property which is being tested. The model's ability to estimate the value of the property in this case is not very impressive as they know the answer to the test. The second issue with benchmark data which is not as common as the first but is increasing is the use of Multiple Listing Service (MLS) data. As the use of MLS data by AVM providers increases, it becomes more difficult to create a purchase transaction benchmark which is unfamiliar to the AVM vendors. Even if the sale of the test property has not been recorded, it has most likely been listed and that listing data is available to the AVMs providing some insight to the subject property's value via a sale price to list price ratio.

Mercury Network uses proprietary benchmark data which is collected internally through its flagship product RealView. Through RealView, we collect 15,000 to 16,000 benchmark properties each week. For each property we have a contract price, 1004 appraisal value or both. There are two key characteristics to these benchmarks: 1) the transactions are purchases or refinances which are still in process and have not closed or been recorded. As such, this data is unknown to the AVM vendors and the data aggregators, and 2) the use of 1004 appraised values as one type of benchmark value mitigates the impact of MLS data available to the AVM models.

The benchmark data is collected on Monday each week and is extracted from the previous 7 days of transactions insuring that the data is current.

Any geographic bias introduced by the use of data derived from RealView transactions is moot. Because the cascade will only be created for counties where there are enough transactions to score and rank the AVMs, any counties for which data does not exist will simply not have a cascade.

02 Data gathering

Once the benchmark data is collected, it is filtered and formatted into a test file to be sent to the AVM vendors. The filtering process eliminates any transactions where the contract price or appraised value is extremely low or extremely high. For example, transactions with benchmark values of less than \$20,000 are automatically removed. If the client wishes to further filter the range of the benchmark values, this can be accomplished easily and quickly.

The benchmark file sent to the AVM vendors contains the following data fields: Reference ID, address, city, state, and zip code. This data is all that is required for the AVMs to identify the property and render a value estimate.

The vendors are given 72 hours to process the test file and return their appended results. Appended AVM data includes the AVM estimate, AVM low value, AVM high value, confidence score, Forecast Standard Deviation (FSD), last known sale price, last known sale date, and any error codes which define why the AVM failed to render a value estimate.

Not all AVMs will return all data fields. As each week's return values are received, they are aggregated in a master database containing both the benchmarks as well as the AVMs' estimated values, the AVM estimate, AVM low value, AVM high value, confidence score, Forecast Standard Deviation (FSD), last known sale price, last known sale date, and any error codes which define why the AVM failed to render a value estimate. Not all AVMs will return all data fields. As each week's return values are received, they are aggregated in a master database containing both the benchmarks as well as the AVMs' estimated values.

03 Data combination

As mentioned earlier, each property record has a unique Ref ID which is used as the identifier for the benchmark. This Reference ID is used to link the benchmark data to the AVM return data in preparation for analysis.


04 Data analysis

All analysis of the AVM data is performed using Tableau, which is an advanced data analytics and visualization application. The default parameter for analysis is the county. This is an industry standard and stems from the fact that the primary source of data fueling the AVMs comes from county recordation data and that the depth and quality of the data can vary by county. In addition to county-level, state and national level analytics are performed as well. National level metrics are never used for any model evaluation, but are often requested by end users. State level metrics are frequently used when an individual county does not have sufficient data to perform a county level analysis and the state level analytics are used as a proxy for those counties. In addition to geography-based analytics, examining AVM performance by price band is also available. The use of this analysis is largely dependent upon the end user and their specific requirements.

For each county and state, a host of metrics are calculated as part of the performance evaluation. The exploration of AVM performance centers on several key concepts – two of which are centrality and dispersion.

These concepts are applied when examining the distribution of errors for each model. The error refers to the % difference between the AVM value and the benchmark value. When these errors are charted, centrality refers to how close the peak of the error curve is to zero. Dispersion refers to how spread out the errors are over the range of possible values.

This base calculation of percentage error is extended to various forms such as absolute percentage error, mean and median percentage error, mean and median absolute percentage error, and standard deviation of percentage error. Additionally, the PPE metric is calculated at various levels. PPE refers to the percentage of error observations which fall within a certain percentage range. For example, PPE10 defines the percentage of observations where the error was between -10% and +10%. For the purposes of complete analysis, we calculate the PPE at the 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50 levels. This approach provides a thorough understanding of the distribution of errors and when combined in a score, will convey both centrality and dispersion in a concise manner. Metrics such as Mean % Error and Median % Error are simply point estimates which can be greatly influenced by the distribution of the data and may not be indicative of the AVM's performance.



Beyond the examination of percentage error and the various metrics describing its behavior is the analysis of the AVMs' confidence score or FSD and its relationship to AVM accuracy. Every AVM provides some indication of how accurate the estimated value should be based on factors such as currency and quality of data.

This indication is commonly referred to as a confidence score and is typically a proprietary number but may also take the form of a Forecast Standard Deviation (FSD) which has become somewhat of an industry standard. In either case, the value of a confidence score is that it can provide the end user with some indication as to how accurate the AVM value estimate is assuming that the AVM knows when the value estimate is good or not. While some AVMs calculate an FSD natively, many of the FSDs provided by other models are calculated by subtracting the proprietary confidence score from 100, for example. Because of this lack of consistency in how the FSD is calculated and its definition, the proprietary confidence scores are used for the analysis with the exception of HVE which has FSD as its proprietary confidence score.

To establish whether the confidence score/FSD is a reliable indicator of the AVM's accuracy, an analysis is performed to examine this relationship. The analysis ranges from simple visual review of how the accuracy of the AVM changes as a function of the confidence score/FSD, but may evolve to include more advanced statistical analysis using Bayesian principles or logistic regression. In any case, the objective is to determine if a valid relationship exists between confidence score/FSD and the AVM's accuracy and if there are inflection points in that relationship which can be incorporated into the policies which govern AVM use.

05 Model scoring and cascade development

It is widely known and accepted that no single AVM performs optimally across all geographies and, as such, one of the primary outcomes of an AVM evaluation is to create a Model Preference Table (MPT) or cascade of AVMs. The purpose of an AVM cascade is to establish a predefined sequence of AVMs, on a county-by-county basis, which identifies the “best” AVM, second-best AVM, etc. But, in order to develop a cascade, there must be a method for rank ordering the AVMs and to do that, they must be scored.

Scoring AVMs consists of assigning different weights to the various performance metrics so that each AVM's performance can be expressed numerically in a single number. AVM scores are typically calculated using measures such as how frequently the AVM values a property (hit rate), the accuracy (ie. Median % error, PPE10), and volatility (standard deviation). There is no single, definitive approach to scoring and Mercury Network engages with clients to jointly define the appropriate metrics and method for scoring. Mercury Network has developed a default scoring approach which is presented to the end user but can be adjusted upon request.

Once the metrics and method have been defined by Mercury Network and the client, the AVMs are scored within each county and the cascade is developed. The cascade will now define which AVM will be run first for any property in a given county and, if that first AVM fails, which AVM will be run second, and so on.

Cascades, by default, are created at the county level; however, in some cases there is insufficient data to assess AVM performance for a given county. In this case, some end users will elect to employ a state-level cascade to be used when a county-level cascade is not available.

In any case, the industry standard is to employ a cascade which is not more than 3 AVMs in depth. This is a prudent practice as it is unlikely that if a value cannot be generated by one of the top three AVMs, then an AVM is not the appropriate valuation for the subject property.

Scoring and model preference table

Presented below is the full Model Preference Table which was created by rank ordering the AVMs by county. Scoring methodologies and formulas are subjective and easily modified to reflect specific risk parameters. Generally, there are two primary components to the score: Hit Rate and Accuracy. The hit rate is defined as the number of properties for which the AVM rendered a value divided by the total number of properties submitted. There are a host of accuracy metrics which can be combined to score the AVMs. Point estimates like mean % error and median absolute % error provide no information about the distribution of errors and can be swayed by outliers or an unusual concentration of values.

However, a combination of PPE metrics is ideal for gauging accuracy as it captures both centrality and dispersion of the error distribution. Specifically, using PPE5, PPE10,...,PPE50 provides a very clear picture of the AVM's performance.

As the PPE_x metric represents the % of errors which fall within +/- x% and given that PPE50 contains PPE45 which contains PPE40, etc., it is a self-weighting formula. This accounts for the dispersion in the distribution. If the distribution of errors is skewed, then the PPE5 and PPE10 values will be smaller and the overall value will be lower. In this way, it also accounts for centrality.

A perfect score of 10 is theoretically possible. In order for this to occur, the Hit Rate would need to be 100% and the PPE5 metric would need to be 100% also. PPE5 equal to 100% means that all errors were between -5% and +5%. If this was the case, then the PPE10 would be 100% as would PPE15, etc. So, the sum of the PPEs would be 1000% divided by 10 equals 100%.

The number of ways to score the AVMs for the purpose of rank ordering them within a given county is endless and modifications to the scoring formula to better align with the end user's risk profile and management practices is available.

A key characteristic of this approach to scoring is that it is symmetrical around the benchmark valuing placing equal weight on AVM values that are greater than the benchmark as those values which are less than the benchmark. When using AVM values to support a loan decision, this characteristic may need to be modified so as to penalize the AVMs for overvaluation or, at least, place a greater positive weight on AVM values which are close to, but less than, the benchmark and a lesser weight on AVM values which are greater than the benchmark.

Cascade testing

An AVM cascade is not a model that is derived from any methodologies such as regression. There is no need for a hold-out sample or to confirm the accuracy of the model given that the model does not render an output as one would expect from an EPD model or pre-payment model. End users can, and should, track the use of any AVM cascade to ensure that the rate of a returned value (a “hit”) is consistent with general expectations and that the production environment yields outcomes which are aligned with the policies and procedures defining AVM use.

Contact information

Craig Zielazny

Senior Professional, Product Solutions

crzielazny@corelogic.com
800 434 7260 x307