AVM Testing Guide & Glossary

Contents

AVM Testing Methods

Over the past few years, rapid progress made in the fields of machine learning and data science has spurred a renaissance of automated valuation models (AVMs), and in particular, the development of the “lending-grade” or “professional-grade” AVM. The growing movement toward using these professional-grade AVMs in higher-stakes situations has created a new responsibility for AVM providers: to approach AVM testing with the same rigor and accuracy they’ve built into their model.

AVMs are becoming less of a “gut check” tool, and more of a pillar in certain situations, like in home equity lending or secondary/securitization workflows. AVM testing methods vary in their approaches and benchmarking data, and lenders should be aware that because of the lack of standardization, there are loopholes AVM providers can leverage to “game” these tests. Using flawed test results may lead lenders to false comfort in the accuracy or performance of an AVM in normal, day-to-day use.

To test an AVM, you need to compare the AVM’s value estimate against a real property value (benchmark) to see how accurate the model is in aggregate. This can be done nationally, or by region (state, county, ZIP code, etc.).

The chosen benchmark has implications on the performance stats, and how easy it is to tailor AVM results to reflect the benchmark and provide an inflated accuracy measurement. Currently, most AVMs are compared to three benchmarks: sale price, contract price in purchase appraisal, and refinance appraisal valuation. Below is an overview of the different types of benchmarks and the nuances of each one.

Sale price

The market price a property sold for, documented by the county recorder, or multiple listing service (MLS) sources.

Pros

- Reflects the real market value of a property.
- There are plenty of benchmarks available (large sample size).

Cons

- Most AVMs have access to these prices, which allows for gaming of the test. For example, if the AVM provider can look up the test’s addresses and see the last sale value, it is easy to tailor their AVM results to better fit the benchmark.
- AVMs use this data to train their model. In statistics, using your benchmark data in the training data leads to biased (more favorable) results.

AVM Testing Methods (cont.)

Concerns with testing against sale price benchmarks

Since the benchmark data being used consists of sale prices from public records and MLS data, any AVM provider can easily look up the same sales data they have and send back an "AVM" that is very close to the benchmark. The industry should be aware of these potential testing loopholes because there is money to be made with overinflated testing results, and it is tempting for an AVM provider to take advantage of the situation.

A note on preventing the gaming of third-party tests: AVM testers are not blind to the fact that some tests can be gamed to make an AVM look better than it actually is. They highly discourage this practice and will take steps to prevent it from happening; however, it is not always easy to detect.

Contract price in purchase appraisal

The contract price agreed to by a buyer and seller, denoted on an appraisal report for a property currently pending sale (typically).

Pros

- More difficult for an AVM to game, since most AVM providers do not have access to these benchmarks, and they are not yet reflected in the MLS (usually).
- Allows for benchmarking/testing in non-disclosure states where sale prices are difficult to find.

Cons

- Most AVMs have access to the listing and pending price on these properties since they are typically listed for sale on the MLS. The listing price is easily used by AVMs to predict what the property is under contract for (they are highly correlated, but not perfect).
- There are not as many purchase appraisal benchmarks as sale prices to form a large testing sample.

Concerns with testing against purchase appraisal benchmarks

A test against purchase appraisals has room to be gamed since AVM providers can usually see what the listing price or pending price is for these properties in MLS data. This is tempting for an AVM provider to tailor their results. If the AVM provider also has access to view all appraisals that happen each week, they could tailor the results that way.

A note on comparing AVMs to appraisals: While there are national, third-party aggregators of appraisals that provide this service, they do not disclose the actual appraised values publicly. Even the AVM provider is blind to the actual data in the appraisal reports from these services. The valuations are only used in aggregate to answer the question that everyone has: "Are AVMs as good as appraisals?" (The answer in 2020: They are not. There are too many factors that appraisers take into account that a machine cannot.)

AVM Testing Methods (cont.)

Refinance appraisal valuation

The appraiser's expert opinion of value on a property that is undergoing the refinance process on a loan.

Pros

- Not easily gamed. These properties are not typically for sale, so the AVM does not know what they are listed for, what they will sell for, or any MLS-based characteristic information on them.
- Most AVM providers do not have access to these appraised values, so their model is forced to be blind to the benchmark.
- The appraiser's research and valuation is the most informed and accurate valuation methodology for this scenario.
- Allows for benchmarking/testing in non-disclosure states where sale prices are difficult to find.

Cons

- Not as many benchmarks available when refis are slow.
- The benchmark is an opinion of value, not the actual market price the property sold for.

Choosing the right AVM provider up front is critical. Properly understanding how an AVM will perform in day-to-day operation by validating the performance metrics methodology will save lenders from having to repeat the selection process. An AVM provider should make this information easy and discoverable, enabling a hassle-free switch to a strong-performing AVM.

AVMs can be a useful tool in situations where they've traditionally had difficulty being consistently accurate (e.g. home equity lending). Since most tests do not shed light on how AVMs perform in scenarios where the benchmark is not known, it is important to ensure the AVM can stand up in those situations. Testing an AVM purely against the same set of refinance appraisals gives the best indication of what AVM will perform the best in lending situations.

Most AVM testers produce a few common measurements to gauge accuracy. They don't all tell the same story, and are typically used in conjunction with each other. The best AVMs find a balance between hit rate and accuracy measures, so it can be tailored to the use case (sometimes really strong accuracy is preferred over hit rate, other times the opposite is desired). A good AVM provider can adjust their model to suit both needs.

How to conduct an AVM test

So, how can we conduct a rigorous AVM test before use in lending or securitization workflows? In other words, how can we measure accuracy so the AVM can't "game" the test, or make itself look better than it actually is?

The best way to properly test an AVM is to control your benchmarks and understand what information an AVM might have access to that could skew the results. For example, a set of benchmark property addresses that include recent sale prices will most likely be available to any modern AVM that constantly ingests sale price information and uses it to train the valuation model. In such a case, the tester may not get a feel for how an AVM will behave on properties that have not recently sold. Taking this further, even if the AVM does not yet know the sale price, it's possible the AVM is using listing prices sourced from MLS (multiple listing service) data to gain insight into how much the property could sell for.

Since AVMs are commonly used in situations where the property has not recently sold or is not currently listed on the market, there must be a way to test real-world, day-to-day use. Rather than relying on the AVM provider to perform an unbiased, out-of-sample test and trust that it was conducted in a completely honest fashion, we suggest conducting your own test.

At Clear Capital, we test ClearAVM™ — our proprietary AVM that consistently outperforms many other major AVMs — in several ways, and have found that the refinance appraisal valuation is the best measure for the quality of an AVM because it proves if the AVM can handle tricky property scenarios. When multiple AVMs are measured against the same set of benchmarks, it's easier to get a feel for the more professional-grade AVMs versus the more consumer-grade AVMs.

Without further ado, here are the seven steps we suggest to conduct a bulletproof AVM test.

1. Select the AVM vendors you plan to test ahead of gathering any AVM values.

This step ensures all the AVMs will use the same set of benchmarks and deliver results to you in the same time period.

2. Prepare a large set of benchmarks (sample size matters).

Again, the best benchmark for an AVM is a recently completed appraisal on a refinance transaction. Ideally there are more than 1,000 addresses in this set — tests on tens of thousands of properties are common — and they all have effective dates within 90 days. These property addresses should properly represent where you conduct business. If you conduct business nationally, the addresses should be dispersed, but have a proportionate amount in urban and suburban areas (versus highly rural areas).

How to conduct an AVM test (cont.)

3. Ask the vendor to provide AVMs on the benchmark.

This can be done in a couple ways. Ideally, you ask them to provide test access to their instant AVM API, so you can gather AVMs on your terms, and limit the possibility for an AVM vendor to tailor or influence the results. Knowing most valuation providers are honest, opting for a bulk spreadsheet match and append is not a bad alternative. Keep the benchmark values (the appraised values) to yourself, and only send the property addresses and a tracking ID for each row. Request that the results — at minimum the AVM values and an FSD — be returned as soon as possible. It's reasonable to ask for a same-day turn around with most modern AVM vendors.

4. Calculate the error rates on the results.

Once you have the AVMs returned, pull them in next to your benchmark values and run some stats. We like to look at AVM accuracy from several views, but most commonly we measure hit rate, mean absolute error (MAE), and PPE10 (see the glossary below for definitions). For each benchmark, tabulate how often the AVM produced a "hit" — this percentage is your hit rate. For all the hits, calculate the percent variance between the AVM and the benchmark. Take the absolute value of this, average it across the whole set, and you have the MAE. You can take the number of AVMs that were within plus or minus 10 percent of the benchmark on the whole set to produce the PPE10.

5. Optional: Compare FSD/confidence score to the accuracy of the AVMs.

A good AVM has a confidence score highly correlated to the valuations variance to the benchmark value. This can be used to determine if you trust the AVM's confidence in day-to-day use.

6. Compare the AVMs.

Now that you have error rates on all the AVM vendors, you can compare them. Plot the MAE versus hit rate for all AVM vendors on the same chart. In most cases, you should be looking for the AVM that has a good balance of high hit rate and low MAE (or high PPE10).

> **Note:** If you value higher accuracy but less hits (or vice versa), bring it up to your vendor. They should be able to accomodate results that fit your needs.

7. Choose your candidate vendors.

By this point, you can narrow down the results to a few leaders. This is when it is important to look at other factors, including product offering fit, model governance, and unbiased error rates against sale prices.

How to conduct an AVM test (cont.)

These steps represent a perfect scenario of data availability, and we realize not all consumers of AVMs have the benchmarks or capacity to conduct these tests. Variations of these tests can be conducted as long as it's understood what the possible pitfalls may be and how the results could be skewed.

Fortunately, there are third-party AVM testers that conduct similar tests. While not all third-party AVM tests are perfect, they do provide a good representation of AVM accuracy when interpreted correctly.

AVM Glossary

Below is a short glossary that breaks down some of the common terms one might expect to come across in the world of AVM testing and validation. As AVMs become more prevalent in lending or securitization workflows, this is the jargon we're commonly asked to define.

Accuracy

With regard to individual AVMs, accuracy is the measure of the variance between the estimated value and the true value of the property (which can be the sale price or appraised value). Here are our thoughts on choosing the right benchmark. Accuracy is more commonly defined as how much error the AVM produces as compared the benchmarks, in aggregate on a large testing sample.

> **Note:** The actual discussion of accuracy is much more complex. As I mentioned in my previous post, most AVM testers use a few common benchmarks to gauge accuracy and measure the error in different ways. They don't all tell the same story, and are typically used in conjunction with each other. The best AVMs find a balance between hit rate and accuracy measures, so they can be tailored to the use case (sometimes really strong accuracy is preferred over hit rate, other times the opposite is desired). A good AVM provider can adjust their model to suit both needs depending on the use case.

Absolute error

The absolute value of the percent variance between the model valuation and true value or benchmark. (See also: MAE, MdAE, and true value)

MAE (mean absolute error)

Determined by calculating the percent variance between each AVM and benchmark, taking the absolute value of each, then averaging them over the whole test/sample set of benchmarks. A strong measure if using consistent benchmarks since it accounts for the times the AVM was very wrong and produces a large error. In other words, it accounts for the outlier predictions which are important when judging an AVM for day-to-day use.

MdAE (median absolute error)

Similar to MAE, but since it uses the median (instead of the mean), it hides the times the AVM was really far off; i.e. the outliers. This is the most common measure for consumer-grade (not professional/lending-grade) AVMs since it generally makes an AVM appear more accurate than it is.

P10 (or PPE10)

The percentage of time the AVM is within 10 percent of the benchmark. A standard measure for the AVM industry, but does not capture or consider the AVM predictions that are very far from the benchmark (the spread).

AVM Glossary (cont.)

Outlier
In terms of data, this is an observation that has unusual values as compared to the norm, markedly differing from a measure of central tendency. In terms of AVMs, an outlier is a value prediction that is significantly off from the benchmark, or having an unusually large error.

Hit rate
The number of benchmark addresses the AVM was able to predict on. The hit rate is based on the AVM's ability to locate the property address, or how much confidence the AVM has on its prediction of value for the property. Often, AVMs do not (and should not) produce results when they don't have strong confidence in the valuation.

Confidence score
The AVM confidence score is typically based on the standard deviation of the valuation prediction (see also: FSD) that indicates the level to which each of multiple models "agree" with each other for a given property.

Coverage
Typically defined on the national level, the ratio between the total number of valuations returned by the model and the total number of valuations requested. This metric can be further "decomposed" in coverage ratios for various reasons for which valuations cannot be produced. A typical example is to decipher if the AVM "cannot find the property" or if it "cannot produce a valuation that meets a desired confidence level." (See also: hit rate)

FSD (forecast standard deviation)
Most AVM confidence scores are based on the calculated FSD produced along with the value prediction. FSD is a statistical measure that scores the likeliness that the valuation is accurate. The FSD can be used to determine a highly probable value range around the property's value. The FSD is typically based on a measure of the spread or deviation in possible estimates that the model found while attempting to conclude a final value estimate. With this metric on each valuation, a modeling team can then measure how correlated these standard deviations are with actual AVM errors and build a model to predict any future AVM error. The result of this model produces the FSD (forecast standard deviation), which can be communicated in various ways. The most common way is as a decimal value (e.g. 0.07) or as a percentage (e.g. 7 percent). The lower the FSD, the better quality/accuracy of the AVM. This percentage is sometimes turned into a confidence score, by subtracting it from 100 percent, (e.g. an FSD of 0.07 or 7 percent is a confidence score of 93 percent).

> **Note:** Since every AVM vendor typically builds a proprietary model to produce the FSD statistic, the communication of this measure can vary between AVMs. There is no standardized way to produce an FSD in the valuation or financial services industry. This means a 0.07 FSD from one AVM can have a very different measure of accuracy or confidence from another vendor. This makes it difficult to use multiple AVMs together, or to apply a framework on how to use AVMs in general. Our goal is to educate the lending and financial services industry on this nuance and to provide guidance on ways to solve that problem. Clear Capital has taken the science of predicting our AVM errors to the next level. Since we have found that prediction errors are not normally distributed, we forced ourselves to come up with a better way to predict our errors and provide an FSD metric along with every AVM that accurately reflects the validity of the valuation.

AVM Glossary (cont.)

Error
Error is the term applied to the measured difference between the true value benchmark and the value estimate reported by the AVM. Error is the fundamental measurement used to evaluate the likely performance of any AVM. It is typically reported as an absolute percentage difference.

Methodology
With regard to AVMs, methodology is the generic term for the variety of methods that a provider may use in the development of an AVM. Occasionally the word “technology” is used in this circumstance where methodology is intended. Some common approaches are index-based, hedonic, appraisal emulation, or tree-based.

Test sample
The test sample is the selection of properties/addresses that the AVM tester asks the AVM providers to return with an estimated predicted value. It is the basis for AVM performance testing. The test sample is intended to be sufficiently broad to assure a degree of statistical validity to measure error nationwide. While it is possible to select a “generic” test sample, test samples typically reflect the market preferences of the AVM tester as to geography, type of property, and so on. A good sample of properties to be used as an AVM benchmark should be large, diverse, and recent to prevent bias or gaming.

True value
Or, “the benchmark.” The sale price or market value of a purchase transaction, or the appraised value on a refinance transaction.

Address standardization
A key part of obtaining an AVM on any given property is to make sure we know the right property address, so we can locate the property. Address standardization is where the different components of the address are parsed and checked to conform with the typical format of U.S. addresses. Clear Capital uses the Coding Accuracy Support System (CASS) by the United States Postal Service, along with third-party services to parse and locate property addresses. There are several different components to address standardization: parsing, error correction, and standardization.

Address parsing is the process by which individual components of the address are broken up and stored separately to allow better management and quality control. Typically parsing looks at separating these components: street number, pre-direction (e.g., N, S, E, W), street name, post-direction (e.g. NW, SE), street suffix (e.g., street, drive, avenue), and so on.

Error correction fixes problems with misspellings, incorrect city names, incorrect zip codes and so on.

Standardization is the process that ensures that the same name is reported the same way. For example, Florida may be represented by Florida, Fla., FL., and so on. Once standardized, it will always be represented by the “standard” which is the two-character state code: FL (without a period at the end).

AVM Glossary (cont.)

With regard to AVMs, why is address standardization so important? A substantial number of “misses” by AVMs are not the result of failures in the model itself, but rather simply a failure to properly identify the address within the AVM’s database. Addresses can be represented in many different ways, and even the slightest variations are confusing to an automated system. Knowing which address standardization technique is used by the AVM provider helps both the lender and the AVM provider ensure the best possible performance. Since that isn’t always the case, Clear Capital invests heavily to be able to find any property address entered into our system.

Learn more at

clearcapital.com/avmguide