Preface

Recently, the AVM Performance Testing Subcommittee of the Collateral Risk Management Consortium (CRC) performed an extensive review of Automated Valuation Model (AVM) performance testing practices. The Consortium initiated this process out of concern for the general lack of guidance available with respect to this issue for residential mortgage market participants, as well as the absence of general uniform standards and methodologies. This conclusion was clear to the subcommittee as many Consortium member institutions had engaged in AVM performance testing themselves and recognized the need for both uniformity and consistent standards.

The hope in publishing this report is to foster adoption of uniform standards and methods for AVM performance testing. By presenting some likely standards and methods, the Consortium hopes to invite further debate among the mortgage community at large. Such standards and methodologies could be used to test the likely performance of the available AVMs.

Because mortgage market participants use AVMs differently according to need, any final set of standards and methodologies needs to accommodate variations in approach. In addition, standards must fulfill the need for users to gather objective support for the development and deployment of credit policy and other processes that might be based on AVM results.

The subcommittee feels that it has provided a reasonably comprehensive discussion of the AVM performance testing process and associated issues with this document. It cautions, however, that this is a preliminary document. Much more discussion is in order to accommodate the disparate issues related to different aspects of the mortgage market. The Consortium and this subcommittee invite those organizations to share their experiences to aid in further refining the information presented here.
**Testing Process Overview**

The testing process itself is intended to obtain AVM performance results from the different commercially available (or homegrown) AVM products and services based on a sample of residential properties. Results are then analyzed and compared to ascertain which AVM or AVMs are most appropriate for use.

There are four aspects of the process:

1. Identification of and Relationship with Vendors
2. Sample Selection
3. Interpretation of Results
4. Application of Results
**Glossary of Terms**

To ensure a certain amount of clarity with respect to the process, the subcommittee has included this glossary of terms. This glossary is not intended to cover all of the terms that arise in the context of AVM performance testing. Additional terms are defined directly in the text. The subcommittee invites suggestions on other terms for inclusion in this glossary.

**Automated Valuation Model (AVM).** For our purposes, an automated valuation model (AVM) is the generic term for any electronic analytic algorithm, process or model that is intended to estimate the value of an individual property, without human assistance (other than the initial entry of the data). Specific to this report, the term applies to models designed to value single-family residential properties (which typically includes single family residential, condominiums, planned unit developments (PUD) and sometimes cooperatives). While not considered in this report, the subcommittee recognizes that there are models that are intended to value commercial-industrial properties.

**Methodology.** Methodology is the generic term for the variety of methods that a statistical model practitioner may use in the development of an AVM. Occasionally the word “technology” is used in this circumstance where methodology is intended.

Examples of familiar methodologies applied to AVMs include hedonic methods and repeat sales analysis.

**Hedonic Model.** The Bureau of Labor Statistics informally describes the hedonic model this way: “A hedonic model decomposes the price of a consumer product into implicit prices for each of its important features and components, thereby providing an estimate of the value for each price-influencing feature and component.”

Applied to property valuation the Hedonic model uses specific physical and sales characteristics of each property in assigning a value estimate by taking into account the contribution that each physical characteristic may bring to value. For example, Hedonic models will typically look, first individually and then in combination, at various significant characteristics such as:

- gross living area (square footage) of the house
- number of bedrooms
- number of bathrooms
- size of lot
- year built
- other characteristics

By analyzing how these characteristics contribute to value in different locations, the model assigns individual weights to characteristics. When applied to the property in question, those weights form a complete value estimate.

Generally, hedonic models, because of this ability to account for major characteristics that appeal to buyers, are more accurate than other techniques in estimating value. By its very nature, however, the Hedonic model is very dependent upon property data to use in its analysis.

The inherent weaknesses of this dependency appear in two different forms. First, there is the difficulty in assessing the subjective aspects of some characteristics. For example, strong positive influences on value, such as views or proximity to significant features—notably water—are very difficult to measure statistically. In like fashion, strong negative influences, especially those that affect only a few properties in a neighborhood, such as poor access, also prove difficult to measure. Second, many jurisdictions in the United States simply do not have adequate data available to support hedonic models.
Repeat Sales Analysis Model. By aggregating changes in value for properties sold more than once during a specified period of time in a given geographic area, statistical means estimate market-level housing price changes. If an individual property has not been substantially changed since its last sale, this analysis matches each pair of sales transactions (thus the name “repeat sales”). The amount of appreciation (or depreciation) is calculated from the time of the first sale to the second and so on, providing an estimate of the overall appreciation of that local housing market during that time period.

As can be imagined, the larger the number of available sales pairs, the more statistically reliable the estimate of overall housing price trends will be. Because this analysis is based on identifying properties where more than one sale has occurred, the challenge is to identify enough observations to provide a meaningful index of housing values, while keeping to as small a geographic area as possible.

A repeat sales index may also overestimate market appreciation if the data contains pairs of sales in which the second sales price reflects substantial improvements (or other alterations) made to the property after the first sale.

As this method looks at movements in housing prices within a geographic area, it is inherently less precise than a hedonic model that can differentiate between individual properties.

On the other hand, repeat sales indices can and do provide very useful valuation estimates in jurisdictions where the data is insufficient to support hedonic models. In addition, they may prove more accurate in tracking housing values for the houses that a hedonic model may struggle with (especially those subject to extreme positive or negative influences) when a prior sale is known on the property.

Appraisal Emulation Model. The appraisal emulation model follows the steps that an appraiser might follow in forming a value estimate (although not with the same insight or flexibility that a qualified appraiser brings to the assignment). The model selects “comparable sales” using some standard criteria. It then rates those comparable sales by suitability based on the physical and sales characteristics of each comparable sales. By adjusting the varying elements (much as is done on an appraisal form), the model then forms an opinion of value.

This method combines some elements of the hedonic model (in that it will typically carry some sense of the weights needed to adjust between significant physical characteristics) and the index models (in that it might need to adjust the sales information forward for older sales in order to gather enough sales information).

Hybrid Model. A hybrid model is one where more than one modeling technique is used in deriving the estimate of value. Typically the technique involves the running of a hedonic model and a repeat sales index. The results of each are compared and evaluated. Based on each result, the hybrid model reports a final estimate of value. In addition to the hedonic model and repeat sales index, many hybrid models also include the results of a tax assessed value model.

By using more than one approach in estimating value, modelers can overcome potential weaknesses of one methodology by using the strength of another.

As a matter of course, many AVM developers have incorporated elements of more than one approach into their models, regardless of how that model may be characterized.

Tax Assessed Value Model. Tax-assessed value models derive an estimate of value by examining market values attributed to properties by the local taxing authorities. Taxing authorities are typically county officials applying policies at the county level. Sometimes, authorities work at the township level within the county.

As a matter of local law and custom, the values reported by the taxing authorities often (but not always) vary from the current market value in some reasonably predictable manner. For example, some jurisdictions require the taxing authority to report the value at 25% of likely market value. In others, values are re-assessed only on an infrequent basis. Some jurisdictions report multiple values : assessed , appraised and market values. By examining local laws and customs with respect to how that value is derived, it is often possible to provide a general adjustment to values reported by taxing authorities to better approximate current market value.
Given that the adjustment is at the county (or township) level, the model cannot take into account individual variations within the county. (Note however, that some repeat sales index models are only effective at the county level as there is not enough data available to report a credible index for a smaller geographic segment.)

**Test Sample.** The test sample is the selection of properties that the testing institution wishes the AVM providers to value and return. It is the basis for AVM performance testing.

The test sample is intended to be sufficiently broad to assure a degree of statistical validity. While it is possible to select a “generic” test sample, test samples typically reflect the market preferences of the testing institution as to geography, type of property and so on.

**Confidence Score.** The confidence score (some prefer the term reliability score) is the term applied to the rating given by the individual AVM to the level of confidence in the relative accuracy of the value estimate.

Each AVM provider uses its own scale. There is no apparent correlation between the different confidence score scales used.

Because the score reflects “confidence” in the result, in some ways it may be seen as measure of the factors that assure confidence in the value estimate, especially the quality of relevant data, activity in the market and the relative homogeneity of the properties.

**Response Rate (“Hit Rate”).** The response rate (often called the hit rate) is a measure of the number of valuation estimates received as a percentage of the total in the test sample. That is, how many valuation requests were “hits.”

Some institutions also measure effective response rate or useful response rate. This is a measure of response rate, where a confidence score that meets or exceeds some minimum level accompanies the valuation estimate. That minimum level is typically related to the institution’s own policy regarding what is considered an acceptable AVM value.

**Accuracy.** In the narrow sense, as used in this report, accuracy (as in accuracy of an AVM) is the measure of the degree to which the estimate of value by the AVM mirrors the reported sales price of the property. In cases where sales price is not available, non purchase mortgages for example, accuracy is measured against the appraised value of a property.

The actual discussion of accuracy is much more complex, of course, since the likely sales price of a property occurs within a range of values, and each of the interested parties has a different view of what accuracy means to them.

**Standardized Address.** A standardized address is one where the elements of the address are reviewed and corrected to conform to consistent standards. The most important address standard is the Coding Accuracy Support System (CASS) certification from the United States Postal Service. CASS certification tests the results of commercially available address correction and standardization software for accuracy. CASS certification, however, is primarily intended to ensure the accuracy of the addresses for mail delivery purposes which may be at a different physical location than the property's site address – a common occurrence in the rural areas.

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 (a very common occurrence), 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 2 character state code: FL (without a period at the end).
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 a computer.

Even with address standardization, it is possible that not all addresses will be matched, as different address standardization software programs do not always approach every issue the same (for example, some are more forgiving of slight variations than others). Knowing which address standardization technique is used by the AVM provider helps both the lender and the AVM provider ensure the best possible performance.

**Error.** Error is the term applied to the difference between the “true value” and the value estimate reported by the AVM. (The “true value” is defined as the stated sales price where a purchase transaction or the institution’s appraised value (or stated value) if a refinance transaction.)

Error is the fundamental measurement used to evaluate the likely performance of any AVM. It is typically reported as an absolute percentage difference.
Process Outline

Identification of Vendors

National v. regional AVM providers. Should the test include both national and regional providers?

The decision whether to include national and regional providers (or for that matter a large number of individual providers) depends on several factors that the testing institution will need to take into account.

- Assumption is that including regional providers will improve the overall performance of AVMs for the institution. There is an assumption implicit in the question of whether to include both national and regional providers. That assumption is that by including regional providers the institution will obtain “better” results because different providers perform differently at the local level. By picking and choosing individual providers at the local level, the institution’s response rate and accuracy will improve.
  - On the other hand, some institutions make effective use of an AVM even if they make the decision based only on a few providers (national or regional) that suit their lending market areas.
  - Moreover, the administration of AVM procurement that includes a larger number of regional providers is more difficult and requires more attention than one with a small number of national providers.
  - Thus, the extent to which the testing institution expects to make decisions about usage on a national rather than a regional scale is the first factor.

- Testing both national and regional providers requires more time and people.

Including both national and regional providers in a test requires more time and people to develop the sample and analyze the results.

  - The size of the sample needed to test performance needs to be larger in order to evaluate both national and regional providers. It is possible to draw inferences about a national provider with a smaller sample than necessary to evaluate regional providers. In order to select an appropriate sample size for regional providers, the sample size itself needs to be increased to ensure that every geographic area is tested.
  - As both sample size and the number of providers increase, the amount of work necessary to analyze the results and form an opinion about the performance of the different models also increases. This factor often affects the depth and latitude of the tests that institutions perform.

Is the underlying methodology used by the AVM provider important to the testing of AVMs?

There are several different recognized methodologies used by AVM providers, including hedonic models, repeat sales index models, appraisal emulation models, hybrid models, tax assessed value models. Those in turn may use various different mathematical techniques, including neural network technology, probabilistic methods and so on.

In general, the purpose of AVM performance testing is to assess the relative performance of the different models with respect to two measures: response rate (“hit rate”) and accuracy. Regardless of the underlying methodology the AVM provider uses, the analysis of test results will use the same criteria. Thus, the underlying methodology is not significant to the process.

In like manner, the underlying methodology used by a particular vendor should have little consequence to whether that vendor should be included or not included in the group of providers tested.
Adoption of a particular model, however, may well depend on the user's understanding of the methodology behind the model (See Advisory Opinion 18, Appraisal Foundation, 2000).

To what extent is the distribution method used by the AVM provider important to its inclusion in the testing group?

The method by which the AVM provider distributes its results (or receives its requests) is both a question of testing convenience and of operational convenience.

Sending a large test sample to the AVM provider is a different proposition than the ordering of AVMs and receiving results within the institution’s day-to-day operations. Thus, a small inconvenience in distribution method in the testing process is not the same as an inconvenient ordering and receipt distribution method when used in the mortgage process itself. An ill designed process in operation may be an important element in deciding whether to use the AVM after the analysis of results is completed.

As it happens, most AVM providers are much more automated today than they have been even in the near past. At least as to testing convenience, nearly all of them can easily accommodate the batch submissions of large test samples.

What kind of advance agreement should the institution have with each AVM provider?

In general, the testing of AVMs is covered by a confidentiality/non-disclosure agreement between institution and AVM provider. This protects the confidentiality of the information the institution provides to the AVM provider and protects in like fashion the results the provider supplies back to the institution.

Testing institutions report that AVM providers typically supply the governing confidentiality/non-disclosure agreement. While this seems unusual, it is consistent with the need for the AVM provider to protect its technology and results reporting. (It should be noted that while there are different forms to confidentiality agreements, they typically cover the same items in the same way.)

There are two items that may need a little more attention than is usually covered in the typical agreement.

- **Gramm-Leach-Bliley Act.** The first is the question of the privacy of nonpublic personal information under Gramm-Leach-Bliley (G-L-B). Typically the information provided to the AVM provider does not include nonpublic personal information. However, as a matter of caution, (whether the institution feels it may inadvertently reveal certain information or not), the institution should include G-L-B language in their agreement with the AVM provider.

- **Sharing results with the AVM provider.** The second is the question of how the AVM provider can use the results derived by the institution. It is not unusual for institutions to share summary, anonymous results of their AVM analysis with the AVM providers. Those providers have a real interest in using that information to further their marketing efforts. The institution needs to settle on its own policy with respect to this issue and make sure that it is addressed very clearly in the agreement.

What are the other details regarding how the institution will work with the AVM provider?

There are several other details that institutions and AVM providers need to address.

How much time will the AVM provider have to respond to the test?

The consensus is that the AVM provider should have no more than 24 hours to process and return the test sample to the institution. Some institutions allow up to 48 hours. In either case, this keeps the test closer to loan production conditions.

It is important that the results mirror the sort of results that the institution would likely see in production. By restricting the amount of time the AVM providers have, there is less chance for the AVM providers to examine the results and make revisions to the records to assure better results. This is not to say that this is inappropriate, but results where the AVM provider has examined results (and perhaps make corrections in the address) would not reflect production results and thus would be less instructive to the institution.
By requiring vendors to return the test results within a short period of time, institutions can ensure that recently recorded sale prices do not unduly affect the AVM results.

**How are the records handed to each vendor?**

The question here is whether the files given to each AVM provider are identical or is there any accommodation for differences in format or other items that an individual vendor would prefer. This has become less of an issue than it has been in the past. The records included in the test are fairly short. The sample is generally delivered as a comma delimited value (CSV) file or a spreadsheet format. The AVM providers appear to be able to interpret these readily.

Some AVM providers in the past have asked that certain property types be excluded from the test. As a matter of course, the consensus is that all test samples should be identical in content and form.

**Are there any costs involved that need to be borne by either party?**

Not surprisingly, it is the view of the AVM user community that the cost of the AVM testing results should be borne by the AVM providers. However, testing institutions need to be careful not to overuse this privilege and only ask for testing on a reasonable basis.

Institutions should bear their own costs with respect to the selection, management and analysis of the test samples.

**Sample Selection**

The next step is to select the sample of properties to be included in the test. The process of selecting samples not only involves ensuring the statistical validity of the test but also the likely practical application of the results to the institution’s production use of AVMs.

**Who selects the sample? Does it matter?**

Traditionally the selection of the test sample has centered within the operational group that uses AVMs. Typically this has been the Appraisal Department or similar operational department within the institution.

This selection method gives the sample a clearly operational slant. The kinds of loans and location of loans that the operational people are seeing are reflected in the sample.

This approach has proven reasonably successful in the minds of the institutions, but as AVMs gain greater and greater acceptance, other parts of the institution may want to participate in the selection process. Likely participants may be credit risk managers and other who are looking at not only current market areas, but also expanded or reexamined market areas.

Traditionally, AVM providers have not participated in the sample selection process.

**Size of Sample**

The simple answer is the sample should be large enough to provide useful results without being too large to analyze easily. This leaves much latitude.

The size of the sample will vary depending on whether the sample is intended to reach an overall conclusion regarding an AVM provider (especially a national provider) or whether the sample needs to permit useful conclusions at a specific geographic level (such as MSA or county).

**Is there a statistical answer to this question?**

There is, but it is not very helpful. In general, a sample can be statistically sound at relatively small numbers. For purposes of AVM performance testing, however, most institutions felt bare statistical validity alone was insufficient to demonstrate suitability of this valuation method. As a result, they have chosen to use a greater number than would be required statistically.
For “national” samples, where the results are intended to identify relative performance on a single, global level (not individual county or MSA level), sample sizes have ranged from 5,000 to 10,000 records.

- Testers have looked to include at least 100 records in each county/MSA included in the test
- Testers tend to look at 50 – 100 MSA’s (by population) as a sample base
- This leads to a sample size of approximately 5,000 to 10,000, which seems to be a manageable number for analysis.

For “regional” samples, where the results are intended to identify relative performances on an individual MSA or county level, sample sizes must increase.

- Testers have looked to include at least 200-500 records in each county/MSA included in the test (as a matter of practice institutions tend to use larger sample sizes in areas where they have no prior experience)
- Where the sample is intended to identify the likely performance at the county/MSA level, this size test sample can more precisely reflect the lending areas of greatest concern by the testing institution.

Geographic Distribution

The geographic distribution of the sample properties can be selected based on several different approaches.

- Most testing institutions select their sample from within their own lending area. That is, they select the sample properties in sufficient quantities to represent the location of their typical loan production.
- Others use the top 50 or 100 MSAs as their sample area
- County is the most significant geographic subdivision. Regardless of what approach the testing institution takes, the smallest geographic area of significance is the county, even if we tend to think in terms of MSA’s (preconditioned by HMDA). It is primarily county (or township in New England) data that fuels AVMs.
  - County practices with respect to assessment information (where descriptions of properties may be located) dictate what information is available about individual properties, how current that information may be and how available that information may be (especially in electronic form).
  - County practices also dictate how current information about land recordings may be, how well properties are identified, and often how complete the information may be.
  - Because of the importance of county, the sample needs to contain enough properties in the county to ensure a thorough examination of likely results. As discussed above, the size of that sample will be dictated by whether the sample is intended to test the entire sample (such as a national sample) or the individual county level.
  - It is more difficult to assure coverage within the county. Limitations in the sample selection process and the wide variations in the size of counties have prevented the even distribution of samples across the county itself. The general view is that any sample selection process will be sufficiently “random” enough to account for the likely results across a county.

Note that since the sample is typically taken from very current sales transactions, the sample will tend to represent current market activity in the county. That may not reflect general county activity, or likely future market activity.

What is the nature of the underlying transaction used as a sample case?

Should the sample include purchase transactions only or should it include both purchase and refinance transactions?

The stronger view is that sample properties should only represent purchase transactions. By using purchase transactions only, the analyst will have the reported sales price to which to compare the results of the AVM. In refinance transactions, the analyst can only compare the results to an appraisal (if one is available).
The counter-argument is that the normal run of production includes both purchase and refinance transactions. The individual institution will need to decide whether it believes it can derive useful results by including refinance transactions.

To the extent that refinance transactions are included, the number of refinance transactions should be overshadowed by purchase transactions. In addition, the number of refinances should not reduce the number of purchase transactions used in order to assure an adequate sample.

**Should the underlying transaction include only Single Family Detached Residences or should it include Condominiums/Cooperatives?**

There is no consensus on this issue. Testing institutions approach this differently. Everyone wants to know the likely results with respect to single family residential detached properties. Many are also very active condominium lenders. Some testing institutions sample the two property types separately or by increasing the sample size based on representation of property type.

Some use the “run of production” mix. That is, they use the same property types as they see in their current run of production. The risk in this approach is that the institution’s policies or marketing may change in a way that the institution does not have the best sample for future lending activities.

*Note that there are circumstances where the analyst may not be able to properly identify the property type because not all property records identify the property type.*

**Should the underlying transaction amount and/or reported sales price be taken into account in the selection of the sample?**

Typically the transaction amount and/or reported sales price are not taken into account in the selection of the sample. The sample tends to reflect the institution’s preferred loans, especially conforming v. jumbo transactions.

Institutions have not expressed any specific reservations about this practice. In fact, some believe that reflecting the institution’s preferred loan offerings may give them a better understanding of how the AVM might perform in production. The counter argument is that institutions do change their market preferences and any skewing of the sample may turn out to work against the performance of the AVMs in production.

Some analysts have reported interesting variations in performance between different transaction sizes. This may or may not be a factor in sample selection, depending on whether it would have the tendency to increase the sample size.

**To evoke an impartial result, the sales transactions reflected in the sample need to be very current. How current should they be?**

Testing institutions place great importance on the use of recent sales. There is substantial concern that the model results will be “steered” by allowing the AVM to “see” the most recent sale.

As a result, testing institutions recommend reaching back in time only as far as necessary to ensure that a sufficient sample is available. Typically this means sales transactions that have occurred in the last 2 – 4 weeks.

In any case, most testing institutions use different strategies in the analysis to attempt to isolate those transactions that may have been heavily influenced by a current sales transaction.

One method used by testing institutions is to exclude any test sample where the sales price and the AVM value are the same. The view is that predicting the value exactly is highly unlikely, leading to an inference that the model was steered.

A more direct, and recommended, approach is to ask the AVM provider to not only include value estimates with each property valued, but to add the last recorded sales transaction it shows in its files (sales price and date) for the property. This allows the testing institution to have more direct control over the
way it handles transactions that might be steered. Each institution can then set its own policy regarding
the exclusion of certain records (e.g., if the AVM provider has a reported sales transaction within 60, 90,
120 or some other number of days).

As an aside, some desire to use AVMs for retrospective valuations. In this case, it might select sales from
past periods to assess how well the AVM will respond. The risk still remains that the model may be
steered by any current transactions.

To what extent should the testing institution attempt to ensure “randomness” in the sample?

The current approach used by testers is to assume that the sample, by its nature is sufficiently random to
test the AVMs. That is, transactions that occurred in the past few weeks are fairly random.

However, there is a small concern that current transactions reflect only that part of the market where there
is activity. That is, certain houses in certain markets at certain price ranges may be selling right now, but
that market experience may change. As a result, not all of the market may be sampled.

Part of this problem may be obviated by increasing the size of the sample beyond the number required.
This will allow the testing institution the latitude to make the sample “more random.”

What is the minimum amount of information the testing institution needs to know about each of the
sample properties?

The institution should know at least:

- Property address (especially street address, city, state and county)
- Property type (if the institution considers that significant)
- Purchase price and date (if testing sales transactions)
- Appraised value and date (if the underlying transaction is a re-finance or the institution uses
  AVMs as an appraisal review tool)
- Stated value and date (if no other value appears available)

What is the minimum amount of information the testing institution needs to supply about each of the
sample properties to the AVM vendors?

The testing institution typically provides the AVM vendor with the following information:

- Property address (especially street address, city, state and county)
- Unique record identifier (unrelated to loan number or other nonpublic personal information—
  remember Gramm-Leach-Bliley)

The institution does not generally provide any value estimates or property type information.

Are the addresses provided to the AVM vendor “standardized” or not?

The majority view is that all addresses submitted for testing should be standardized. The act of address
standardization increases the “hit rate” and reduces the “false positive” responses.

This allows the institution to test the performance of the model itself, without introducing issues of
address matching.

The minority view is that addresses should be submitted as is. This strategy tends to reveal the most likely
response rate under actual production conditions.

That aside, all institutions recommend that addresses should be standardized in their production
system to improve their response rate (in order to obtain the greatest benefit of using AVMs).

What information should the vendor supply in response?

Testers believe that the vendor should supply at least the following:

- Valuation result and date
• Valuation result high and low (sometimes called the value range)
• “Confidence score” (or other indication of the vendor’s own rating of the reliability of the result)
• Last known sales transaction and date for the property valued

How often should testing be performed?

The consensus is that testing should occur at least every 6 months (provided the testing institution is prepared to perform the analysis that frequently).

This view arises from several different factors. First, as noted above, the transactions an institution may find in its market (or its future markets) are likely to change over time. Rather than trying to predict that, more frequent testing tends to pick up these differences.

In addition, the general view is that most AVMs are changing more often now than before. Rather than trying to re-test a model every time it changes (which may prove inconvenient and burdensome) the model changes can be detected at least every 6 months (which appears to be often enough).
Analyzing the Results

Once each tested AVM provider returns its results, the results must be analyzed. It is this analysis that provides the uniform means of comparison between the models. The tester should record all assumptions he or she used along with any and all methods used for making decisions.

What are the core comparisons?

There are several core measurements in any analysis.

1. Ascertain the Response Rate.

   Ascertain the response rate or “hit rate”; that is, for what percentage of records were results (AVM values) received?

   • “Raw” response rate
     This is the total response rate. That is, how many sample properties returned a value?

   • “Useful” response rate
     This is the total response rate, minus those values where the reliability (confidence) score falls below some pre-determined minimum level.

     By implication, the “useful” response rate may vary by testing institution based on its own internal risk decisions.

   • By relevant geographic area
     As many institutions wish to arrive at preferred AVM providers by geographic area, they need to gather response rate data by the geographic area by which they will subdivide their preferred AVM providers.

2. Compare AVM value to “true value”

   The “true value” is defined as the stated sales price where a purchase transaction or the institution’s appraised value (or stated value) if a refinance transaction

3. Measure the variation between the AVM value and the true value as a percentage error

   Query, is the percentage error calculated by comparing to the true value or the AVM value? Is the True Value or the AVM Value the baseline number from which to compare, that is, is it (True Value – AVM Value) or (AVM Value – True Value)? There is a split of opinion on this issue. As a result, we have reported both equations.

   \[
   \frac{\text{True Value} - \text{AVM Value}}{\text{AVM Value}} = \text{Percentage Error}
   \]

   OR

   \[
   \frac{\text{AVM Value} - \text{True Value}}{\text{True Value}} = \text{Percentage Error}
   \]

   If the “true value” is the first number, a positive (+) Percentage Error indicates that the AVM value is less than (or undervalued) the property in relationship to the True value.
If the “true value” is the first number, a negative (-) Percentage Error indicates that the AVM value is greater than (or overvalued) the property in relationship to the True value.

A lower Percentage Error means that the AVM result is more “accurate”. That is, Percentage Error is a rough measure of accuracy.

4. Inventory all of the percentage errors

- Is the model “biased”? That is, does the model appear to “undervalue” or “overvalue” properties as a matter of course?
  - A truly unbiased model should produce an average percentage error of 0.00% (positive percentage errors offset the negative percentage errors).
  - The further away from a 0.00% error, the more biased the model is. Regardless of whether the testing institution calculates its percentage error where the negative percentage error means undervalued or overvalued, it is very important to know whether a model is biased or not and in which direction. This will need to be taken into account when the institution sets its credit policies.

- What is the distribution or profile of errors?
  - The profile of errors is typically done two different ways. First it is done with the absolute value of the percentage error (regardless of whether the percentage error is positive or negative). This allows the analyst to assess in an “absolute” sense, whether a model reached some threshold level of accuracy (regardless of whether over or under the “true value”).
  - The profile is then done using actual percentage error (both positive and negative). This profile is most helpful because it allows the analyst to depict the percentage errors in graphic form. This makes it easier to illustrate both “bias” and the consistency of the AVM results. It also allows for a graphic comparison of the different results obtained by the different AVM providers.

- What percentage of errors are < 5%, < 10%, <20% and so on

- What percentage of error is considered to be “outliers”? (how do we know something is an outlier?)
  - There are differing opinions on what is considered an outlier. Some apply a statistical test, where they exclude (or at least identify) records that exceed a certain percentage error, such as records where the percentage error exceeds 100%. Others apply a more empirical test depending on individual test results, looking at records that appear to depart in substantial ways from the general sample. In reality, the statistical tests initially came from using the empirical view of the data, where the results started to reveal at what level records started to become outliers.

  The notion of outliers is important when it comes to analyzing the results because measures such as average percentage error can be significantly affected by a few large (outlier) errors.

- What is the “average” error (that is “magnitude”)?
  - Expressed as a “mean”
  - Expressed as a “median” (most prefer the median). The overwhelming majority of the tests use mean and not the median.

- What is the profile of percentage error expressed in standard deviations? Is there a statistical standard by which the variations in error expressed as standard deviations can be measured?
  - The consensus was that the use of standard deviations in this analysis has limitations. The use of standard deviations usually involves assumptions regarding “normal distributions” which may not occur in this analysis. As a result, the use of standard deviations has not proven especially helpful for reporting purposes.

  However, the use of standard deviations has proven helpful to understanding each AVM Vendor’s “confidence scores”. It has also proven helpful when comparing results between AVM providers. A larger standard deviation indicates less consistency in the valuations provided by that model.
The distribution of errors can also be profiled by
- Purchase transactions v. refinance transactions (and/or equity transactions)
- Property type (SFR, condominium etc)

5. Are some records excluded from the analysis?

The short answer is yes. In the process, a certain number of records are normally excluded from the analysis as these records are expected to distort the final results. Records normally excluded include:
- “Outliers”
  An “outlier” is any record where the disparity between the true value and AVM value is very large. This could occur for various reasons, including a “false positive”. That is, the AVM may have identified a different property than the intended to be valued.
  Testers use different rules of thumb for exclusion, such as where the percentage error exceeds 100% or some other relatively disproportionate amount (see discussion above).
- Records where the result may be “tainted” by the current sales price.
  Testers universally exclude any records where they believe the results have been tainted by the sales price. Some testers exclude all records where the True Value is the same as the AVM Value. Others ask the AVM Vendors to report the most recent sales transactions the Vendor has available (the recommended approach). Where the current sales price occurs, that record is excluded (although some might examine the currency of that sale before excluding it).
- Are excluded records accounted for in a different way?
  Yes. While they do not influence the statistical results, they are typically included in the response rate calculations.

6. Compare the percentage error to the vendor’s reported “confidence score”

- To what extent does the vendor reported confidence score correlate to the percentage error?
- Is there a reasonable basis for describing an AVM valuation result as accurate or inaccurate based on the confidence score (that is, for a confidence score of “X”, the value will be more or less accurate than with a confidence score of “Y”—with the caveat that we are using percentage error as a measure of accuracy)
- To what extent can the confidence score from one vendor be associated with the confidence score of another?
  - Is there any detected uniformity?
  - Does this allow more direct comparison between vendors, or is it more reliable to compare percentage error?

The issue here has to do with the setting of policy. That is, there is a strong desire to set minimum standards in loan policy. Thus, a policy would like to identify a level of accuracy by reference to an explicit guide such as the confidence score.

In practice, because there is not correlation between different confidence scores, most institutions set a minimum confidence score rating on an individual model basis.

7. Rank the different valuation models by performance

- Rank the different valuation models by response rate from highest to lowest
- Rank the different valuation models by adjusted or effective response rate from highest to lowest
- Rank the different valuation models by percentage error profile from lowest to highest
- Rank the different models by performance within different geographic areas (e.g. MSA or County)
- Response (raw and effective) rate
- Percentage error

Query whether there should be any differentiation between models based on the underlying technology. That is, would it be useful in the setting of policy to take into account the different results that might be reached using different technologies?
Application of Results

The ultimate purpose of AVM performance testing is to make AVMs useful within the mortgage process. To do so, it is important for institutions to draw certain conclusions from their tests, to convert those conclusions into credit policy and apply them to the loan process.

- Based on the results of the test, is the institution comfortable using AVMs in certain circumstances?
- Is response rate more important than percentage error?
- Are the test results such that the institution may not feel comfortable using any AVM in some jurisdictions?
- Are the test results such that the institution can easily give a rank order to the use of the different AVMs?
  - That is, use this AVM in this county as long as the results carry a confidence score of a certain level
  - If the first AVM does not return a result, use the following that have also been found acceptable in this geographic area
- Is it appropriate for the institution to shop from AVM to AVM until it receives a result that supports the proposed transaction?

This issue raised significant concerns within the group. Some institutions take the view that they should request a valuation from their top rated valuation source for the county in which the subject property is located. If that AVM provider provides a value estimate, but one that does not support the transaction, they will not request a different AVM. Their belief is that “shopping” for a value may cause them to overlook issues involving value that carry unanticipated risks.

[It should also be noted that AVM values as a general rule lag the market because of the delay in receiving and managing sales transaction data. Thus, in a rising market, an AVM will often underestimate the value of a property, where an appraisal, taking into account the most current transactions and other market factors will more accurately reflect the current value. Of course, in a declining market an AVM will tend to overestimate.]

The actual statistical analysis to assure accuracy when using more than one model if the first one does not report a high enough value is quite complex. That is, it is no longer possible to evaluate each AVM separately. The analysis focuses more to estimating the probabilities that AVM model B is accurate at the transaction value where AVM model A reported a value less than the transaction value. This involves a very large sample of data and is an analysis that an institution is unlikely to attempt.

Each institution decides its own level of risk, yet there are certain common standards that apply, especially those imposed by mortgage investors.

- Are there any standards available either from the Agencies or the other mortgage investors that an institution can use to help set its own risk policies? What do the rating agencies say?
- How does an institution report AVM results it obtains in a transaction to an investor in that transaction (mortgage)? Are there any representations or warranties that are likely required as part of the process?

There are no real answers to the questions above. They are presented to help each institution in forming policy around AVMs and how they will use them. Suffice it to say that many institutions currently use AVMs in many different ways. Some use AVMs exclusively (but very successfully) in their home equity lending. Others
use them for funding purchase money conforming first mortgages sold in the secondary market (by agreement). Others use them for quality assurance or pre-qualifying only. These approaches represent the different practices of the individual institutions, and their appetite for risk. Others have started with a modest use of AVMs then expanded their application as they become more comfortable with how AVMs can be used. In any case, AVMs are here to stay.

We wish you good luck and success with your AVM performance testing and, more important, AVM use in your mortgage activities.
For More Information

For more information about this document or any other aspect of the Collateral Risk Management Consortium please feel free to contact:

Vicky Cassens  
Director, Strategic Data Alliances  
FNC, Inc  
T (619) 281-3043  
C (619) 889-1672  
v cassens@fncinc.com

Neil Olson  
Chief Strategic Officer  
FNC, Inc  
3100 Bristol Street, Suite 200  
Costa Mesa, CA 92626  
T (714) 866-1099 x475  
F (714) 866-1141  
C (949) 632-6410  
nolson@fncinc.com