

Performance Comparison of Automated Valuation Models

**IAAO 71st Annual International Conference on Assessment Administration
Anchorage, Alaska**

J. Wayne Moore
Business Doctoral Program
Northcentral University

September 21, 2005
Session H
Room 3

1:30 – 3:00 pm

Abstract

Because it is necessary for assessors to prepare value estimates for huge numbers of properties, all at a specific point in time each year, a process called computer assisted mass appraisal (CAMA), which uses Automated Valuation Models (AVMs), has evolved during the past 35 years to handle the logistic challenge presented by the task. Several CAMA methodologies using different AVMs exist for determining the assessed value of residential properties. A large research study and controlled experiment that statistically tested the results of the several CAMA methodologies when applied under the same circumstances has not been reported until now. This research used all valid market sale transactions for the five years of 1999 through 2003 to estimate selling prices of an actual residential property population in an assessing jurisdiction by using four different CAMA methodologies. The model specification, calibration and value estimation work was done blindly by nine independent CAMA practitioners from across North America without knowledge of the source of the data, the actual 2004 sale prices, or even names of other participants. Six of the participants used a generic AVM market model specification approach and were free to select software tools. Three used a pre-specified transportable AVM. Actual sale prices for 2004 valid market transactions for properties from the same population were used to compare the predictive accuracy of each of the CAMA methods to determine if any statistical significance in predictive accuracy existed between the methods. This paper reports the results of the controlled experiment conducted under the supervision of an academic mentor.

Introduction

Six methodologies are used for performing computer assisted mass appraisal (CAMA) to estimate assessed value of residential properties for local property taxation. The first method is the direct sales comparison approach, which is widely used by fee appraisers to produce mortgage appraisals for home purchases. This method is used by property assessors, less frequently for the mass appraisal process, but is widely used to both challenge and defend individual property assessments. A second method, multiple regression analysis (MRA) using software such as SPSS, is a statistical extension of direct sales comparison, emerging in the past 30 years as the power of the computer became available to assessors. The third method is adaptive estimation procedure (AEP), also called feedback, which has its roots in numerical analysis and has also been available for about 30 years. The fourth and most commonly used method is the cost approach that relies upon local market analysis to provide an estimate of depreciation from all causes. The fifth is a hybrid approach referred to in this paper as the transportable cost-specified market (TCM) approach. These methods are used to varying degrees by local property assessors throughout the world. A sixth method exists, based upon artificial neural networks, but is not widely used.

The International Association of Assessing Officers (IAAO) is a professional organization that helps establish best practices and standards for performing tax assessment, including statistical measures for evaluating the quality of assessment work. Assessing professionals have presented numerous case study reports on AVM methodology, but a research study and controlled experiment has not been reported that statistically compares the results of the main AVM methodologies when applied to the same jurisdiction. A competition among vendors to select a computer-assisted mass appraisal system was conducted by the Board of County Commissioners of the County of Allegheny, Pennsylvania, in 1976-1977, has been reported and described (Carbone 1980). The Allegheny competition was patterned after one that had been conducted at Harvard and reported upon in 1977. Richard Ward and Lorraine Steiner presented a paper describing a comparison of feedback and nonlinear regression. At the time of their research, nonlinear regression was just beginning to appear and the stated purpose of the study was “to clarify for assessors some of the issues raised by these techniques with the hope that the comparison of these techniques will contribute to assessor education in the CAMA area”

(Ward and Steiner 1988, 43). Consistent with its educational purpose, the paper provided an overview of software available at the time and summarized statistical results from four different tests, but its main purpose was description of new CAMA techniques rather than performance comparison.

Charles Calhoun discussed the lack of independent testing of AVMs in his article in *Housing Finance International*, which reported upon property valuation methods and data in the United States, stating in footnote 38:

While there is increasing competition among various commercial models, independent evaluations are practically nonexistent given the proprietary nature of the data and models. Whether market forces will ultimately identify the most successful methodologies depends in part on the ability of consumers of these models to undertake their own validations (Calhoun 2001, 21)

The controlled experiment reported in this paper fills the research gap that Calhoun comments upon. The research used five years of sales data (1999-2003), property descriptive data and nine highly qualified CAMA practitioners from across the country to estimate the selling prices of an actual residential property population in an actual jurisdiction, with the practitioners using different CAMA methods as treatments. The model specification, calibration and value estimation work was done blindly by the nine independent CAMA practitioners without knowledge of the source of the data, the actual 2004 sale prices, or even names of other participants. Six of the participants used a generic AVM market model specification approach and were free to select software tools. Four of these six participants are listed as contributors or reviewers for the IAAO's Standard on Automated Valuation Models. The three other participants used a pre-specified transportable AVM. Articles cited in this paper that were published by the research participants in this experiment are noted with an asterisk (*). Standard IAAO statistical quality measures were applied to actual 2004 sales compared to calculated values of properties from the same population for each participant to determine if differences of any statistical significance existed between the methods.

Literature Review

The first attempts at using multiple regression analysis (MRA) to estimate property market value occurred around 1970 (Gouldemans* and Miller 1976). Prior to that time the most widely used

method was the traditional cost approach, primarily doing the work by hand with minimal market analysis. The formal description of the adaptive estimation procedure (AEP), also called feedback, first appeared in the late 1970's (Carbone 1980). Carbone's 1976 PhD dissertation provided a rigorous academic definition of the technique (Carbone 1976). The procedure tests and systematically adjusts model coefficients, converging upon the set of coefficients that minimize an error term (International Association of Assessing Officers [IAAO] 2003, 12). Schultz* (2001) makes a case for use of feedback in his 2001 Research & Technology Update article. There have been numerous conference papers, case studies and journal articles on the application of both MRA and AEP in the past 20 years. Typical of these is a 1995 paper describing the Denver County, Colorado, revaluation using multiple regression analysis presented by the jurisdiction's Chief Appraiser (White 1995). White's paper provides a very good description of the actual revaluation process as used throughout North America. The third method considered is the cost approach. The traditional cost approach is by definition not a market approach, even though in theory all three approaches to value (cost, market and income) should yield similar final values. The cost approach, with locally developed depreciation schedules, or with depreciation individually determined by appraisers, is widely used by assessors. The fundamentals of the cost approach have been well documented for more than 70 years in books such as *The Valuation of Property* (Bonbright 1937). Cost theoretically sets an upper limit on market value (assuming reasonable supply and time factors) and it is generally acknowledged that the main difficulty in using the cost approach is estimation of depreciation from all causes (physical, functional and economic) and the rapidly changing dynamics of the real estate market (Clapp 1977). Nevertheless, a number of states such as Alabama, Illinois, Indiana, Iowa, Nevada, and Michigan publish a state cost manual with a depreciation schedule and require or encourage its use by assessors in their respective states. Assessors were seeking ways to improve the fundamental assessment process as early as 1966. Franklin Graham, the Assessor of the City of Wisconsin Dells, published an article that proposed a new approach, beginning his paper by stating, "This method is a combination of the cost approach and the market data approach" (Graham 1966, 42). An article 14 years later, after the introduction of MRA into the assessment process, discussed a simplifying base home approach that was hinted at in Graham's article (Gloude-mans* 1981). In 1986 Eckert published a paper suggesting methods for calibrating the cost model to market that provided insight into what I have defined

as the fourth method considered in this research, the transportable cost-specified market (TCM) approach. “Much of the process of determining depreciation and fine tuning for location factors in the cost model can be done with the aid of linear and non-linear multiple regression, or feedback” (Eckert 1986, 14). In 1991 Ireland* presented a paper on transportability of a market calibrated cost model based upon the Illinois cost manual (Ireland* & Adams 1991). Ward provided a demonstration on use of feedback to calibrate cost models at the 1993 IAAO Annual Conference on Assessment Administration (Ward 1993). I presented a paper at the 1995 Annual Conference on a market correlated stratified cost approach that defined a hybrid, engineered cost model incorporating market factors (Moore 1995). This hybrid TCM model is now widely used and was employed by three of the participants in this experiment. At the 2005 Conference on Integrating CAMA and GIS in Savannah, Bob Gloudemans* presented a paper describing “how the District [of Columbia] used SPSS’s ‘Nonlinear’ MRA procedure to calibrate their cost structure using sales data in what can be called a fully ‘market calibrated cost model’” (Gloudemans* and Nelson 2005, 2 [Abstract]). All of these papers are presenting variations of the hybrid technique that I am calling the transportable cost-specified market (TCM) approach, the fourth method considered in this study.

A substantial body of literature exists on property tax assessment and the tax revolt that has occurred over the recent quarter century. In his dissertation submitted to the University of California, Berkeley, in fulfillment of the requirements for a Doctor of Philosophy in Sociology, Isaac William Martin does a marvelous job of tracing the historical and sociological issues that have culminated in the tax revolts experienced in the United States and elsewhere in the past 25 years. He argues that from a sociologist’s perspective the tax revolts were caused by the delayed modernization of the assessment system in the United States (Martin 2003). When the power of the computer was introduced to the assessment process in the late 1960’s and 1970’s, assessment procedures were rapidly modernized, and to a certain extent carelessly. At the same time inflation in home prices began escalating causing a corresponding rapid increase in property taxes. Local government officials failed to control the tax increases by reducing tax rates, and the tax revolts began. As the technology in using computer assisted mass appraisal matured, research continued and statistical standards were introduced to measure the quality of CAMA-produced values. An excellent example of these improvements is illustrated in Thomas Hamilton’s 1997 dissertation submitted at the University of Wisconsin. His work addresses the technical aspects

of how sales samples may not properly represent the population leading to value estimation problems. His paper presents his findings on how market value estimates can be improved by using a newly defined least squares estimation technique with distance metrics as weighting factors (Hamilton 1997). The dissertation confirms the advancements made since 1970 and the continuing research being done to improve the CAMA-based assessment process. The IAAO recently published a comprehensive standard on automated valuation models, which contains useful descriptive information about CAMA models and the automated appraisal process:

An automated valuation model (AVM) is a mathematically based computer software program that produces 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. The distinguishing feature of an AVM is that it is an estimate of market value produced through mathematical modeling. Credibility of an AVM is dependent on the data used and the skills of the modeler producing the AVM. ... The development of an AVM is an exercise in the application of mass appraisal principles and techniques, in which data are analyzed for a sample of properties to develop a model that can be applied to similar properties of the same type in the same market area. ... AVMs are characterized by the use and application of statistical and mathematical techniques. This distinguishes them from traditional appraisal methods in which an appraiser physically inspects properties and relies more on experience and judgment to analyze real estate data and develop an estimate of market value. Provided that the analysis is sound and consistent with accepted appraisal theory, an advantage to AVMs is the objectivity and efficiency of the resulting value estimates (IAAO 2003, 5-6).

An understanding of AVMs and their usage in tax assessment is important in grasping the significance of this research. A visit to the referenced IAAO URL and review of the full AVM standard is worthwhile. Even though a large body of literature exists about the subject of mass appraisal and the importance of accuracy in the application of CAMA AVM models, I have not been able to locate a single research project that compared the relative performance of the primary CAMA methodologies used throughout the world.

Method

The work reported in this paper was done for academic credit in the partial fulfillment of the author's degree requirements in a doctoral program. It was done independently and personally, without sponsorship by any organization, commercial or otherwise. It was undertaken for the single purpose of contributing to the body of available knowledge on CAMA techniques used by assessors throughout the world. The work was monitored by Dr. Robert Hausmann, research mentor at Northcentral University. This research and the controlled experiment have the primary purpose of comparing the performance of Automated Valuation Models (AVMs) used in computer assisted mass appraisal (CAMA). Its purpose is not intended to be educational in the use of the techniques themselves, as was the purpose of Ward and Steiner's 1988 research. Since property taxation depends upon having the underlying value assessments as accurate as possible, an important question to answer is whether any one of the methods produces statistically more accurate results than the others when applied under the same conditions. Professional appraisers must perform their work in adherence with the Uniform Standards of Professional Appraisal Practice (USPAP) and in particular, mass appraisal work must be performed according to Standard 6 (The Appraisal Foundation 2003). Most state oversight organizations, such as the Oregon Department of Revenue, have established standards for measuring assessment quality and performance (Oregon Department of Revenue 2004). The widely accepted measure of quality in the tax assessment field is the coefficient of dispersion (COD) about the median of assessment/sale ratios of a sales sample. As part of his consulting practice, Bob Gloudemans has done significant research into the COD statistic and published an important paper on confidence intervals for the coefficient of dispersion (Gloudemans 2001). The reason for wide acceptance of COD as the standard measure is that quality of assessment work is measured in terms of uniform treatment of every property to insure the highest degree of equity and fairness for individual property owners relative to one another. Hence, the practitioner wants the "scatter" of individual assessments (A) compared to their actual sale transaction amounts (S) when they later sell in the market (the A/S ratios) to have a normal distribution about the median of the A/S ratios for the entire sales sample and to be as small as possible, as measured by the COD. Therefore, the test statistic for AVM performance (the four mass appraisal methodologies that were tested) is the COD mean difference and the null hypothesis is stated as

$H_0: \mu COD_{MRA} = \mu COD_{AEP} = \mu COD_{TCM} = \mu COD_{COST}$, where H_0 : = the null hypothesis, and

μCOD_{MRA} = the population mean coefficient for the multiple regression analysis (MRA);
 μCOD_{AEP} = the population mean coefficient for the adaptive estimation procedure (AEP);
 μCOD_{TCM} = the population mean coefficient for transportable cost-specified market (TCM);
 μCOD_{COST} = the population mean coefficient for the cost approach (COST).

The null hypothesis is that the μCOD s will all be the same, not significantly influenced by the choice of method; the research hypothesis is that the selection of method will cause the μCOD s to not all be the same, with methods producing a significantly different COD mean at $p \leq 0.05$. The research hypothesis is stated as H_a : μCOD_{MRA} , μCOD_{AEP} , μCOD_{TCM} and μCOD_{COST} are not all equal; the research hypothesis states that when properly applied by knowledgeable appraisers, the four CAMA methods analyzed in this experiment yield value results with some COD mean differences that are statistically significant at $p \leq 0.05$.

The tests were conducted to measure the predictive accuracy of the four different “treatments” (automated valuation modeling methods). All tests were conducted using the same population and the same random sample drawn from that population. Some records that either had missing data or did not belong in a test of single family residences, such as duplexes and vacant properties, were eliminated prior to distribution to participants. This is different from human experiments because the exact same data can be put through the different treatments, effectively eliminating sampling error as a concern. The population was 22,785 existing single family residential properties and their descriptive characteristics from an actual Midwestern assessing jurisdiction, representing 52 distinct neighborhoods, which itself was a subset of randomly drawn neighborhoods from the entire jurisdiction. A “neighborhood” is a market area with homogeneous properties and similar economic influences. Neighborhood serves as a location variable for the jurisdiction. See the Oregon Sales Ratio Manual for more detail about sales sampling, sale validity and market areas (Oregon Department of Revenue 2004). The test sample was the 1,299 actual properties in the population that sold in 2004, after screening by the assessing staff throughout 2004 to verify the validity of each as an arms-length market transaction. This differs somewhat from generally accepted model testing methodology in that a portion of the model building sales sample (1999-2003 sales) was not set aside for testing and 2004 sales were used instead. In the Allegheny County test, 3,306 sale parcels were selected from the years 1974, 1975 and 1976 with 25% (779) placed in the “set aside” control group for testing, leaving 2,527 for the experimental model building group (Carbone 1980 164). Ward used

a total of 700 sale parcels from 1985 and 1986, with 500 parcels for model development and a control sample of 200 from the same years were used for model testing (Ward and Steiner 1988, 45). My justification for using the following year valid market sales as the control group is that it more closely resembles the reality faced by assessors each year. Also, it could possibly uncover instability in the models when attempting to predict future sale prices, rather than predicting the sale prices of a control group drawn from the model building sample. This decision was influenced in part by Hamilton's research and the desire to consider a "worst case" scenario in sales sample selection. In summary, the current experiment involved 6,845 sale transactions from a population of 22,785 parcels for the six years 1999 through 2004. A total of 5,546 jurisdiction-validated sales from the period 1999 – 2003, with characteristics as they were at the time of the sale, were available to be used for model development. Each modeler was free to use as many or as few of these as desired. Once their models were constructed they were used to blindly estimate the selling prices of the 1,299 jurisdiction-validated 2004 sales. None of the participants had information on current or prior assessed values for any of the parcels including the 5,546 available for model building, they did not know the jurisdiction from which the data had been extracted, and they did not even know who the other participants were. All 1,299 sales were used for testing all of the resultant value predictions, that is, no outliers were eliminated.

An observation was defined as the ratio produced by dividing the predicted sale price by the actual price for each of the 1,299 sold properties in the population. The test statistic was defined as the coefficient of dispersion (COD) obtained from the observations of one participant in the experiment, that is, the average percentage deviation about the median ratios of the observations for that participant. The randomness of the sample is insured by the random activity of the real estate market itself. The previously cited Hamilton dissertation shows that the sales sample created through random market activity may not be fully representative of the population for various reasons (Hamilton 1997), but that does not have an effect on the current research since it is assumed that any population representation errors would impact all the participants equally and not affect the relative difference of the CODs of the participants and the test outcome. The assessed values set by the jurisdiction on December 31, 2003 for the sold 2004 properties were also included as a TCM participant since they had been established prior to the actual sale dates of the 2004 test parcels. After reviewing this report, one of the participants in the experiment suggested that this may not be valid, so it is removed from one set of results.

Since the cost approach involves careful application of the costing procedure to the property characteristic data in a cookbook-like process without any modeling activity (the model and coefficients are pre-specified), I developed the cost estimates for the experiment by using two different AVMs based upon September 2003 Marshall & Swift cost data as the source. One AVM was based upon Section A of the September 2003 Marshall & Swift Residential Cost Handbook, implemented using a large Microsoft Excel spreadsheet. The other was based upon using the ProVal[®] software cost approach with a mass appraisal costing AVM that uses floor level calculations, created from Section B (Segregated Cost) and Section C (Unit-in-Place Cost) of the same September 2003 Marshall & Swift Residential Cost Handbook. Neither cost-based value prediction method used any market adjustments for location, house style, etc.

The first phase of the experiment consisted of recruiting highly qualified participants for the experiment. This was an area of uncertainty because of the potential differences in modeling skill between participants. The IAAO Standard on Automated Valuation Models states, concerning MRA model specification and calibration, “The availability of data will influence the specification of the model and may indicate the need for revisions in the specification and/or limit the usefulness of the resulting value estimates” (IAAO 2003, 8), and “No one software package is deemed superior to another, as success using MRA is a combination of modeling skills and software familiarity” (IAAO 2003, 12). Only qualified, experienced modelers were invited to participate in the experiment, including persons listed as contributors and reviewers of the AVM Standard (IAAO 2003, 2). Most of these persons are known to me, but obtaining the necessary time commitment was expected to be a problem, since the best practitioners are always extremely busy. In discussing time commitments with potential participants, we agreed that no participant should spend more than 24 hours on the research project. The participants were:

Fred Barker (Oregon)

Russ Beaudoin (Vermont)

Sue Cunningham (Virginia)

Bob Gloudemans* (Arizona)

Richard Horn (Iowa)

Michael Ireland* (Illinois)

Ron Schultz* (Florida)

Russ Thimgan (Arizona)

Michael Whitted and Char Cuthbertson as a team (Florida & Indiana)

The second phase involved extracting and organizing data files for distribution to participants. The six AEP and MRA model building participants were encouraged to use their favorite modeling software. The 40 data items listed in Table 1 for the 5,546 sales were extracted from the jurisdiction's SQL Server market database and placed into Microsoft Excel spreadsheets for the six modeling participants. A spreadsheet with the same layout but without sales information was provided for the 1,299 sale parcels in 2004 that comprised the control group to value for the test. All participants were provided the jurisdiction's established land values as of December 31, 2003 in Table 1, field 36 and were instructed to use them as a "given". No data was provided for computing new land values. Correct land values are a prerequisite for the cost approach, whereas land is not as important in the market approach since it looks at total property value. For the participants using the transportable cost-specified market (TCM) methodology, a backup of the SQL Server database used for the ProVal[®] software cost approach was supplied with all 2004 sales information removed, all assessment information removed and jurisdictional identity removed. Just as was the case for the AEP and MRA participants, the test was blind. However, the three TCM started from an existing model specification since they did have the cost approach AVM that was used to produce one set of cost-based predictions. Their task was to use the same sales information from 2003 and earlier that was available to the AEP and MRA modelers and add two market variables: the neighborhood number (Table 1, field 3) as a variable for location, the house type code (Table 1, field 17) as a variable for house type or style. They then were to use the standard analysis tools available within the software product to calibrate the cost approach values to the market using only these two additional variables. They did not use AEP or MRA tools for market calibration, but had a transportable version of these tools been available their results probably could have been improved.

To summarize, the six AEP and MRA participants had to build (specify) predictive models using their respective analytical tools and then calibrate (fit) them to the time trended sales sample from 2003 and earlier, using their own trending technique and judgment as to the age of the sales that should reasonably be used. They then applied their respective models to the 1,299 properties in the test group to estimate 2004 selling prices. The TCM participants had to use a cost-specified AVM as their starting point and then apply two additional market variables and use the standard analysis tools in the software, including its sales trending capability, to

estimate selling prices for the 1,299 properties in the 2004 test group. As described earlier, I provided the two sets of cost calculation results for the 1,299 properties in the test group based upon two different cost AVM model specifications using Marshall & Swift cost data from September, 2003. My notes and assumptions for the spreadsheet implementation of Section A of the Residential Cost Handbook are summarized in Appendix A. Finally, in order to have one other interesting perspective, the jurisdiction's statistics for the 2004 test group were included using their actual assessed values as of December 31, 2003. Based upon the jurisdiction's CAMA methodology, it would be considered a TCM participant (later removed from one result set at the suggestion of one participant). Thus, twelve distinct sets of 1,299 selling price predictions drawn from 15,588 individual observations were available for analysis. This process of estimating the 2004 selling prices of the test group, as performed by all participants, simulates the annual revaluation process that assessors must follow in order to establish assessed values as of January 1 (or other statutory tax lien date) each year for use in property taxation.

Phase 3 of the experiment involved processing each of the twelve distinct sets of 1,299 selling price predictions through exactly the same sales analysis process. Each set of values was extracted from its return source (Excel spreadsheet, text file or SQL Server database backup) and placed in a standard import format for sales analysis. Prior to the sales analysis processing, the 1,299 test group was carefully reviewed one last time to insure that no problems existed with the data. The only potential problem found was that six of the properties had sold twice in 2004. Since the jurisdiction had marked these as valid sales, I felt no need to make any changes, resulting in 1,305 actual ratios being calculated for each test group. The median A/S ratio, price related differential (PRD) and coefficient of dispersion (COD) for each of the 12 distinct AVM model sets were then computed (see Table 2).

The final step of the experiment was to enter the calculated CODs for each of the 12 AVM method groups into SPSS to produce descriptive statistics and perform a one-way analysis of variation (ANOVA) to test the strength of the null hypothesis about differences between the four COD group means. In performing the analysis I noticed a single potential outlier from among the six market model participants. Its COD fell more than two standard deviations from the mean of the market approach group. Therefore, I decided to present the results both with it included and with it excluded. Its inclusion or exclusion does not change the overall results.

Results

The four most commonly used mass appraisal automated valuation model types were tested in this experiment. A one-way analysis of variance (ANOVA) was conducted to evaluate the hypothesis that differences in market value estimating accuracy exist between the four major automated valuation model methods, that is, they are not all equal, and to evaluate the relationship between automated valuation model (AVM) type chosen and the resulting coefficient of dispersion (COD). A lower COD is an important indication of better quality assessments. The independent variable, AVM type, included four types: adaptive estimation procedure (AEP), multiple regression analysis including non-linear regression (MRA), the traditional cost approach (COST) and a hybrid transportable cost-specified market (TCM) method. The dependent variable was the COD that resulted from applying each AVM to predict the selling prices of the same set of 1,299 properties in the control test group. The analysis of variance was significant with or without the outlier, $F(3,7) = 22.28, p = .001$ with the outlying COD removed and $F(3,8) = 8.55, p = .007$ with it included. The strength of the relationship between the AVM type and the COD as assessed by η^2 was strong, with the AVM type accounting for 90% and 76% of the variance of the dependent variable, respectively. Post hoc tests were conducted to evaluate pair-wise differences among the means. Levene's test of equality of error variances was non-significant, $p = .470$ with the outlier eliminated and $p = .140$ with the outlier included. Considering the small sample size and differences between the two groups indicated by the tests of variance, the group with the outlier eliminated was assumed to have homogeneous variance and the results of the Tukey test were used to evaluate pair-wise differences among the means.

Table 2 contains statistics for each of the 12 sets of results including the outlier and the jurisdiction's results. At the request of one participant, the tests were re-run without the results from the jurisdiction. Its removal caused no material change in the results. Table 3 contains descriptive statistics for the result sets for (a) the outlier included, (b) the outlier excluded, and (c) the jurisdiction's results excluded. Table 4 shows the tests of between-subjects effects for the three data sets, (a), (b) and (c). Table 5 contains the results of the Tukey test evaluating pair-wise differences between the means for data set (b) with the outlier removed. Table 6 shows the same Tukey test with the jurisdiction's results removed. Figures 1 and 2 provide box plots with and without the outlier. Figure 3 shows box plots with the jurisdiction's results removed.

Discussion

Appendix B provides some additional information in non-technical terms to assist in understanding and interpretation of the statistical results. This experiment has shown that a statistically significant difference in results as measured by COD does exist between the major property valuation methodologies. It has provided clear statistical evidence to support what most CAMA practitioners believe to be the case and seems self-evident: A market-calibrated AVM will predict selling prices more accurately than a purely cost-based AVM. What may be surprising and not so evident is that the hybrid transportable cost-specified market (TCM) approach, as used in this experiment with only two market variables, and placed in the hands of skilled appraisers working in typical jurisdictions, appears to have performed as well as the other market AVMs as indicated by the Tukey test evaluating pair-wise differences between the means. This finding indicates a need for more research into TCM, which has evolved over the years without clear definition or documentation, but is widely used in various forms. The concept of what I am calling TCM was mentioned as early as 1966 by Graham, and has been discussed in papers through the years by Eckert, Ireland, Ward and others, with the most recent being a paper presented earlier this year by Gloudemans.

The research confirms a statement made in the introductory section of IAAO's Standard on AVMs that "Credibility of an AVM is dependent on the data used and the skills of the modeler producing the AVM" (IAAO 2003, 5). The data used for the experiment is from a jurisdiction where inadequate budget exists for proper field data collection and verification, so it was only possible to achieve CODs consistent with the quality and completeness of the available data. Skilled practitioners were recruited for the experiment since it was important that results not be influenced by varying skill levels. Participants were informed that the data was only adequate to achieve average results at best. Table 2 shows that five of the nine participants using market AVMs achieved CODs between 10.0 and 10.5 with three different software packages.

The research also provides a baseline for pursuit of several additional research questions. How would the results from mandated state appraisal manuals such as those published for use in Iowa, Illinois, Indiana and Michigan compare with the results reported in this paper under the same conditions? Would the addition of X-Y coordinates and use of response surface analysis improve results significantly? What additional market variables and methodology would make significant improvement in the performance of the TCM approach?

References

- The Appraisal Foundation. 2003. Uniform standards of professional appraisal practice, standard 6: Mass appraisal, development and reporting. Retrieved December 31, 2004 from: <http://www.appraisalfoundation.org/html/USPAP2003/standard6.htm>
- Bonbright, J. C. 1937. *The Valuation of property*. New York: McGraw-Hill.
- Calhoun, C. 2001, December. Property valuation methods and data in the United States. *Housing Finance International*, 16 (2), 12.
- Carbone, R. 1976. Design of an automated mass appraisal system using feedback. *Dissertation Abstracts International*, (UMI No. 7618072)
- Carbone, R., Ivory, E. L., & Longini, R. L. 1980. Competition used to select a computer-assisted mass appraisal system. [Electronic version]. *Assessors Journal*, 15(3), 163-167. Retrieved May 7, 2005, from <http://www.iaao.org/>
- Clapp, J. M. 1977. The cost approach to market value: Theory and evidence. *Assessors Journal*, 12(1), 43-46. Retrieved May 7, 2005, from <http://www.iaao.org/>
- Eckert, J. K. 1986. Calibrating the generic model using construction cost data. *Assessment Digest*, 8(1), 11-15. Retrieved May 7, 2005, from <http://www.iaao.org/>
- Gouldemans, R. J. & Miller, D. W. 1976, April. Multiple regression analysis applied to residential properties – a study of structural relationships over time. *Decision Sciences*, 7 (2), 294.
- Gouldemans, R. J. 1981. Simplifying MRA-based appraisal models: The base home approach. *Assessors Journal*, 16(4), 155-166. Retrieved May 7, 2005, from <http://www.iaao.org/>
- Gouldemans, R. J. 2001. Confidence intervals for the coefficient of dispersion. *Assessment Journal*, 8(6), 23-27. Retrieved May 7, 2005, from <http://www.iaao.org/>
- Gouldemans, R. J. 2005. Market calibration of cost models. [Paper]. *Proceedings of the integrating GIS & CAMA 2005 conference, February 15-18* [CD-ROM]. Savannah, GA.
- Graham, F. D. 1966. Comparative method for mass assessment of residential real estate. *Assessors Journal*, 1(3), 41-54. Retrieved May 7, 2005, from <http://www.iaao.org/>
- Hamilton, T. W. 1997. Sales sample data limitations and heteroskedasticity effects on property tax equity and incidence. *Dissertation Abstracts International*, 59 (02), 572A. (UMI No. 9803426)

- International Association of Assessing Officers. 2003. Standard on automated valuation models (AVMs). Retrieved December 24, 2004 from http://www.iaao.org/pdf/AVM_STANDARD.pdf
- Ireland, M. W. & Adams, L. 1991. Transportability of a general-purpose residential market-calibrated cost model. *Property Tax Journal*, 10(2), 203-224. Retrieved May 7, 2005, from <http://www.iaao.org/>
- Marshall & Swift. *The Residential Cost Handbook, September 2003 Data*. Los Angeles: Marshall & Swift
- Martin, I. W. 2003. The roots of retrenchment: Tax revolts and policy change in the United States and Denmark, 1945-1990. *Dissertation Abstracts International*, 65 (02), 719A. (UMI No. 3121598)
- Moore, J. W. 1995. The market correlated stratified cost approach. *Proceedings of the 61st annual conference of the International Association of Assessing Officers*, Chicago, 61, 223-236.
- Oregon Department of Revenue. 2004. Ratio manual. Retrieved January 2, 2005 from: http://egov.oregon.gov/DOR/PTD/ratio_manual.shtml
- Schultz, R. 2001. The other market model. *Assessment Journal*, 8(1), 48-50. Retrieved May 7, 2005, from <http://www.iaao.org/>
- Ward, R. D. & Steiner, L. C. 1988. Comparison of feedback and multivariate nonlinear regression analysis in computer-assisted mass appraisal. *Property Tax Journal*, 7(2), 43-67. Retrieved May 7, 2005, from <http://www.iaao.org/>
- Ward, R. D. 1993. Using feedback to calibrate a cost model. *Proceedings of the 59th annual conference of the International Association of Assessing Officers*, Chicago, 59, 253-256. Retrieved May 7, 2005, from <http://www.iaao.org/>
- White, B. 1995. 1995 Denver County revaluation using multiple regression analysis. *Proceedings of the 61st annual conference of the International Association of Assessing Officers*, Chicago, 61, 209-221.

Appendix A. Marshall & Swift Residential Cost Calculator Method Notes & Assumptions

1. The September 2003 Marshall & Swift Residential Cost Handbook was used as the cost source, following Form 1007 instructions on pages 12-16.
2. For the base rate, some interpolation was used between table entries where it was obvious that it would give more consistent results.
3. All houses were assumed to be stud frame since that is the building method commonly used in the jurisdiction and no data was available.
4. For the two quality groups higher than "Excellent" (originally derived from the High Quality Homes book), I added 20% & 33% respectively to the Excellent rates.
5. There was minimal roof cover data available, so I used the base for each quality class with no attempt to make adjustments.
6. "Base" was used for Energy Adjustment and Foundation Adjustment; no addition was made for seismic zones or hurricane wind (neither applies for the location).
7. "Base" was used for floor cover since no data was available.
8. The rate for "Warm and cooled air" was added by quality class where central air was indicated in the data.
9. For One and One Half story homes, they are assumed to be Cape Cod style with dormer linear feet estimated at 0.7% of TLA.
10. Dormer rate in Section A for Fair Quality appeared to be wrong (same as Avg Qual) so it was adjusted using page B-19 as the reference source.
11. Fireplaces are assumed to be direct-vent gas in fair & average homes, and the low end of the range of "single two-story" fireplaces for better homes.
12. Based upon the method of data collection in the jurisdiction, basement finish is assumed to be "partitioned" in the M&S basement cost tables.
13. Attic finish rate is from page B-24; M&S does not provide a rate for unfinished attic in Section A calculator method.
14. The base allowance was by quality class was added for built-in appliances.
15. Basements for Fair, Average & Good Quality are assumed to have 8" poured concrete walls; V-Good & Excellent have 12" poured concrete.
16. Porches are assumed to have roofs without ceilings (costs are added together for the porch rate); decks are without a roof.
17. Garage wall type is not in the data, so for Quality Fair - Good "Wood Shingle" cost was used, and for V-Good & Excellent, "Brick Veneer" cost.
18. The jurisdiction's "Extra cost amount" was added for pools and other significant yard items that they had found and entered in the data.
19. Marshall & Swift depreciation as described in Section E was applied.
20. Chronological (actual) age taken to a log base 1.25 produces effective age that closely approximates the effective age on page E-15 Life Cycle Chart.
21. Effective age was further adjusted by the jurisdiction's condition ratings using multipliers that caused the effective age to adjust within the high and low points on the chart on page E-15:
 0.33 for EX (a 12 year effective age becomes 4 years); 0.50 for VG; 0.75 for G; 1.00 for AV;
 1.40 for F; 2.00 for P.
22. The M&S Quarterly Multiplier for September 2003 was 1.02 and the local cost multiplier for frame construction was .98 for the jurisdiction, resulting in an overall factor of 1.00.

Appendix B. Understanding Experiments and the Significance of Statistical Results

A researcher conducts an experiment in order to test a theory or hypothesis, which is generally in the form of a question. The question for this research was, “Does the choice of AVM method have a significant impact on model performance?” The research hypothesis is that it does. The opposite hypothesis, called the null hypothesis, is that it does not. The researcher uses a research tool called hypothesis testing to answer the question by making a probabilistic statement about the true value of a test parameter. The researcher is looking for sufficient evidence from the sample to indicate that the true value of the parameter is not zero. For this experiment the random sample was created by having multiple AVM experts compute values in order to deal with the random error induced by the skill level of any single expert and be able to attach statistical significance to the results of the experiment. Statisticians use a t-test to make this probabilistic estimate. The probability of making an error when performing a t-test is called the level of significance of the t-test and is expressed as $p = .01$ or $.05$ or $.10$, meaning 1%, 5% or 10% respectively. For this experiment a level of significance of $p = .05$ was used for the hypothesis, meaning that if the probability of making an error was not greater than 5%, the results would be considered to be statistically significant. Another way to put this is that the level of confidence had to be 95% that the results were correct. SPSS was used to assess the statistical significance of the results of this experiment. The SPSS output shows a computed p -value that was used to compare to the experiment’s level of significance of $p = .05$, meaning the results would be considered significant if the p -level was shown by SPSS to be not greater than $.05$.

Dependent Variable: COD

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	27.918	3	9.306	22.277	.001	.905
Intercept	1357.323	1	1357.323	3249.221	.000	.998
AVM	27.918	3	9.306	22.277	.001	.905
Error	2.924	7	.418			
Total	1442.220	11				
Corrected Total	30.842	10				

R Squared = .905 (Adjusted R Squared = .865)

Above you see the SPSS output for a test of statistical significance. The column labeled “Sig.” is the p -value, and in this example $p = .001$, meaning the result has a high degree of statistical significance because there is only one chance in 1,000 that an error has been made. The column labeled “Partial Eta Squared” (Eta is the Greek letter η) is showing the coefficient of determination, also called “R Squared” or R^2 . It measures the fraction of the total variation in COD that is explained by choice of AVM, and has a value between $.00$ and 1.00 , so $R^2 = .905$ means nearly 91% of the variation in COD may be explained by choice of AVM. The statistical significance of the results of this research leaves little doubt about its accuracy.

Table 1*Parcel Variables*

Field & Name	Description
1 ParcelNo	Parcel identifier, numeric, ranging from 16 to 52100. (Note: Parcel Identifiers in the parcel population range from 3881 to 91011462 and do not have the same PINs as the historical sales data sample).
2 Class	Property class - all are residential, single family class 510
3 Neigh	Neighborhood number, 3-digit numeric, range 108 to 579 (52 total)
4 District	Tax district number, 6-digit numeric
5 SaleDate	Sale date in a single date field with the format 'mm/dd/yyyy' (total=5,546)
6 SaleAmt	Sale amount; range 17,400 - 1,823,000; median 139,900; mean 168,274
7 s1	Sale validity code for state reporting
8 s2	Sale validity code for arms-length market transaction, 'V' = valid
9 Acres	Parcel acreage where available
10 TLA_SF	Total finished living area square feet
11 FinSFB	Finished living area square feet - basement
12 FinSF1	Finished living area square feet - 1st floor
13 FinSF2	Finished living area square feet - full 2nd floor
14 FinSFUp	Finished living area square feet - partial upper floor such as half story
15 FinSFLL	Finished living area square feet - lower level of split or bi-level (split foyer)
16 Stories	Story height as a single numeric field; 100 = 1 story, 150 = 1 1/2 story, etc
17 H_Type	House type code, numeric, where 12 = old 1 or 1.5 story, 22 = older 2 story, 42 = newer 1 story, 52 = newer 1.5 story, 62 = newer 2 story, 71 = split foyer bi-level, 80 = split level
18 B_SF	Basement square feet (no basement = 0)
19 F_Baths	Number of full baths

20	H_Baths	Number of half baths
21	Tot_Fix	Number of total plumbing fixtures
22	AttGar_SF	Attached garage size in square feet (no attached garage = 0)
23	Gar_Cap	Attached garage car capacity (not always available)
24	DetG_SF	Detached garage size in square feet (no detached garage = 0)
25	C_Air	Central air conditioning (Y or N)
26	FP	Number of fireplaces
27	Year	Year constructed
28	EffYear	Effective year built - proxy for effective age
29	Cond	Condition - 94% = AV, 1% = EX, 1.5% = F, 2% = G, 1% = VG, 0.1% = P
30	Grade	Quality grade, numeric, ranging from 25 to 95 with 45 = avg, 25 = poor
31	Extra	Extra features flag, where 1 = yes
32	ExtraDesc	Free form description of extra features
33	ExtraAmt	Amount of value assigned to the extra features by the appraisal office
34	PorchSF	Total square feet of porch area
35	WdDkSF	Total square feet of wood deck area
36	Land_Cost	Estimated market land value placed on the lot by the appraisal office prior to time of sale
37	RoofMat	Roof cover material code
38	AtticSF	Total square feet of attic area
39	AtticFinSF	Finished living area square feet in the attic
40	Ext_Cov	Exterior cover material code

Table 2*Statistics for the 12 Sets of Results*

AVM Type	COD	Median Ratio	PRD
AEP	10.2	94	1.01
AEP	10.9	96	1.03
AEP	12.0	102	1.04
AEP	13.8	90	1.06
COST	14.4	98	.97
COST	14.9	94	.98
MRA	10.0	99	1.03
MRA	10.5	99	1.03
TCM	10.1	94	1.00
TCM	10.1	95	1.02
TCM ¹	10.2	89	1.01
TCM	11.3	96	1.01

¹ This is the jurisdiction's statistics

Table 3*Descriptive Statistics***(a) With outlier:**

Dependent Variable: COD

AVM Group	Mean	Std. Deviation	N
aep	11.725	1.5692	4
cost	14.650	.3536	2
mra	10.250	.3536	2
tcm	10.425	.5852	4
Total	11.533	1.8203	12

(b) Outlier removed:

Dependent Variable: COD

AVM Group	Mean	Std. Deviation	N
aep	11.033	.9074	3
cost	14.650	.3536	2
mra	10.250	.3536	2
tcm	10.425	.5852	4
Total	11.327	1.7562	11

(c) Jurisdiction results removed:

Dependent Variable: COD

AVM Group	Mean	Std. Deviation	N
aep	11.033	.9074	3
cost	14.650	.3536	2
mra	10.250	.3536	2
tcm	10.500	.6928	3
Total	11.440	1.8087	10

Table 4
Tests of Between-Subjects Effects

(a) With outlier:

Dependent Variable: COD

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	27.782	3	9.261	8.550	.007	.762
Intercept	1475.802	1	1475.802	1362.540	.000	.994
AVM	27.782	3	9.261	8.550	.007	.762
Error	8.665	8	1.083			
Total	1632.660	12				
Corrected Total	36.447	11				

R Squared = .762 (Adjusted R Squared = .673)

(b) Outlier removed:

Dependent Variable: COD

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	27.918	3	9.306	22.277	.001	.905
Intercept	1357.323	1	1357.323	3249.221	.000	.998
AVM	27.918	3	9.306	22.277	.001	.905
Error	2.924	7	.418			
Total	1442.220	11				
Corrected Total	30.842	10				

R Squared = .905 (Adjusted R Squared = .865)

(c) Jurisdiction results removed:

Dependent Variable: COD

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	26.587	3	8.862	18.614	.002	.903
Intercept	1293.633	1	1293.633	2717.081	.000	.998
AVM	26.587	3	8.862	18.614	.002	.903
Error	2.857	6	.476			
Total	1338.180	10				
Corrected Total	29.444	9				

R Squared = .903 (Adjusted R Squared = .854)

Table 5

Tukey test evaluating pair-wise differences between the means (outlier removed)

Multiple Comparisons

Dependent Variable: COD

	(I) AVM Group	(J) AVM Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	aep	cost	-3.617*	.5900	.002	-5.570	-1.664
		mra	.783	.5900	.576	-1.170	2.736
		tcm	.608	.4936	.628	-1.026	2.242
	cost	aep	3.617*	.5900	.002	1.664	5.570
		mra	4.400*	.6463	.001	2.261	6.539
		tcm	4.225*	.5597	.001	2.372	6.078
	mra	aep	-.783	.5900	.576	-2.736	1.170
		cost	-4.400*	.6463	.001	-6.539	-2.261
		tcm	-.175	.5597	.989	-2.028	1.678
	tcm	aep	-.608	.4936	.628	-2.242	1.026
		cost	-4.225*	.5597	.001	-6.078	-2.372
		mra	.175	.5597	.989	-1.678	2.028

Based on observed means.

* The mean difference is significant at the .05 level.

The Tukey test in SPSS evaluates pair-wise differences between the mean CODs (the dependent variable) and the AVM groups. The *p*-values of .001 and .002 when COST is evaluated against AEP, MRA and TCM indicate that there is only about one or two chances in 1,000 that the conclusion that the COD performance of a pure COST AVM is significantly less acceptable than the other three is wrong. The test also shows that no such conclusion may be drawn about the relative performance of AEP, MRA and TCM compared to one another. For example, the *p*-value of .989 when MRA is compared to TCM indicates that one would likely be wrong 989 times out of 1,000 claiming a statistically significant difference in the CODs produced by each.

Table 6

Tukey test evaluating pair-wise differences between the means (jurisdiction results removed)

Multiple Comparisons

Dependent Variable: COD

	(I) AVM Group	(J) AVM Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	aep	cost	-3.617*	.6299	.005	-5.797	-1.436
		mra	.783	.6299	.625	-1.397	2.964
		tcm	.533	.5634	.783	-1.417	2.484
	cost	aep	3.617*	.6299	.005	1.436	5.797
		mra	4.400*	.6900	.003	2.011	6.789
		tcm	4.150*	.6299	.002	1.970	6.330
	mra	aep	-.783	.6299	.625	-2.964	1.397
		cost	-4.400*	.6900	.003	-6.789	-2.011
		tcm	-.250	.6299	.977	-2.430	1.930
	tcm	aep	-.533	.5634	.783	-2.484	1.417
		cost	-4.150*	.6299	.002	-6.330	-1.970
		mra	.250	.6299	.977	-1.930	2.430

Based on observed means.

* The mean difference is significant at the .05 level.

This Tukey test evaluates pair-wise differences between the mean CODs when the jurisdiction’s TCM results are removed from the experiment. The *p*-values of .002 to .005 when COST is evaluated against AEP, MRA and TCM indicate that there is between two and five chances in 1,000 that the conclusion that the COD performance of a pure COST AVM is significantly less acceptable than the other three is wrong. On the other hand, the test also shows that no such conclusion may be drawn about the relative performance of AEP, MRA and TCM compared to one another. For example, the *p*-value of .977 when MRA is compared to TCM indicates that one would likely be wrong 977 times out of 1,000 claiming a statistically significant difference in the CODs produced by each.

Figure 1. Box plot of coefficient of dispersion (COD) means for AVM groups (with outlier)

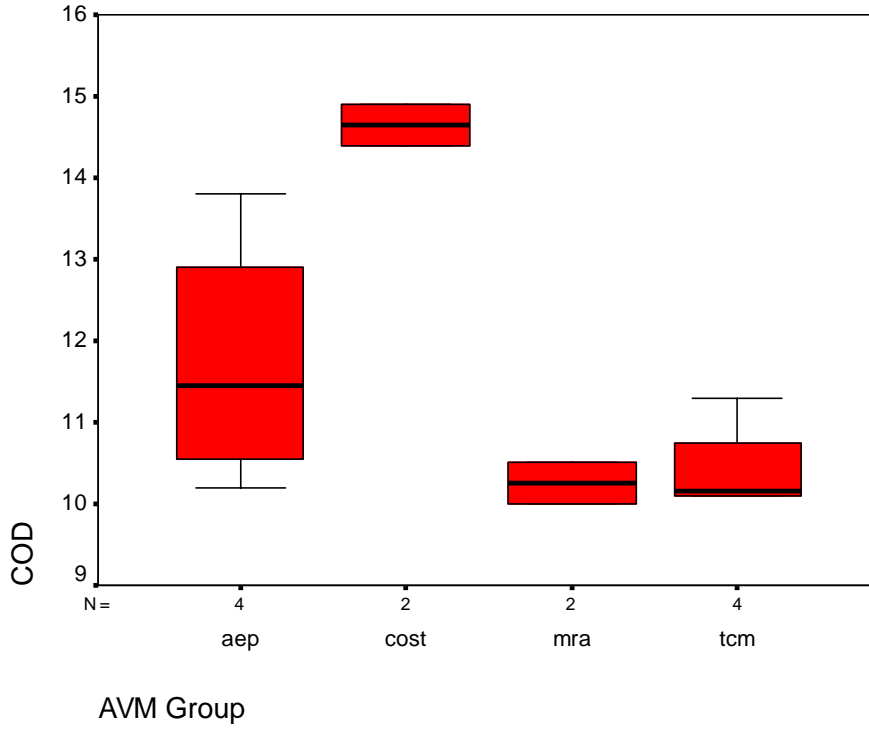


Figure 2. Box plot of coefficient of dispersion (COD) means for AVM groups (outlier removed)

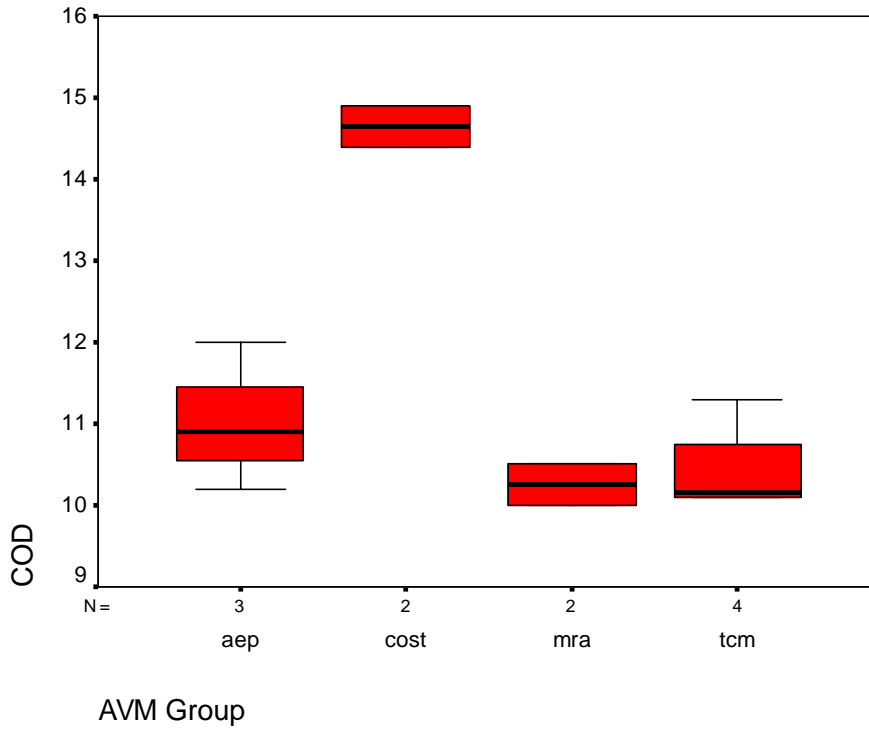
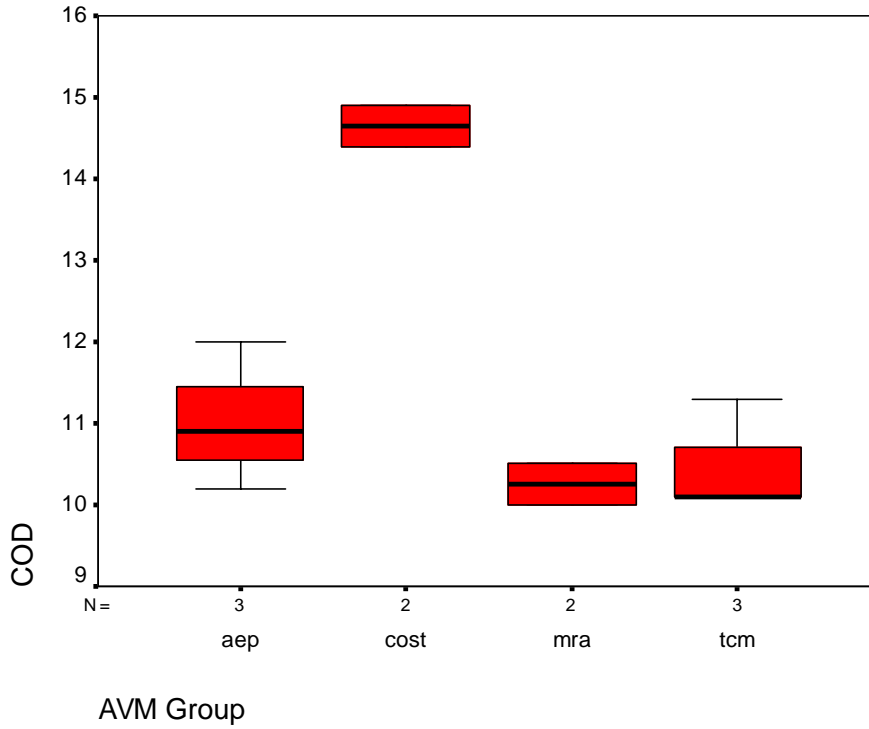


Figure 3. Box plot of coefficient of dispersion (COD) means for AVM groups (jurisdiction results removed)



About the Author

J. Wayne Moore has been a member of IAAO for 26 years. He has been involved directly or indirectly during the past 32 years in implementing AVM-based CAMA systems in more than 300 assessing jurisdictions in North America. He holds an undergraduate degree in Economics from the University of Delaware, a Master of Science degree in Systems Engineering from Southern Methodist University and is currently enrolled in a doctoral program at Northcentral University. He is the founder of ProVal Corporation and serves as Chief Application Architect at Manatron, Inc.



510 E. Milham Ave.
Portage, MI 49002
866.471.2900
www.manatron.com

