

Performance Comparison of Automated Valuation Models

BY J. WAYNE MOORE

This article is based on a presentation given September 21, 2005, at the IAAO 71st Annual International Conference on Assessment Administration in Anchorage, Alaska.

Author's Note: The work reported in this paper was completed for academic credit in partial fulfillment of 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 sole purpose of contributing to the body of available knowledge on CAMA techniques used by assessors throughout the world.

Because it is necessary for assessors to prepare value estimates for huge numbers of properties, all by 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 this task. Six CAMA methodologies presently exist for determining the 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 employed less frequently by assessors for the mass appraisal process, but it 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. This method has emerged in the past 30 years as the power of the computer has become 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 research as the transportable cost-specified market (TCM) approach. These five methods are used to varying degrees by local property assessors throughout the world. A sixth method exists, based upon artificial neural networks, but it is not widely used.

J. Wayne Moore, in his 32-year career, has been involved directly or indirectly 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 and a Master of Science degree in Systems Engineering from Southern Methodist University. He is currently enrolled in a doctoral program at Northcentral University. Mr. Moore was the founder of ProVal Corporation and serves as Application Architect at Manatron, Inc.

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 which was conducted by the Board of County Commissioners of the County of Allegheny, Pennsylvania, in 1976–1977 has been reported and described (Carbone, Ivory, and Longini 1980). In 1988, 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* in which he reported on property valuation methods and data in the United States.

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, n. 38)

The controlled experiment reported here fills the research gap that Calhoun describes. The research used valid mar-

ket sales transactions, including property descriptive data, for the five years from 1999 through 2003 to estimate the 2004 selling prices of an existing residential property population in an actual jurisdiction using different CAMA methods as treatments. The model specification, calibration, and value estimation work was done blindly by nine independent CAMA practitioners without knowledge of the source of the data, the actual 2004 sale prices, or even the names of other participants. Six of the participants used a generic AVM market model specification approach with the software tools of their choice. The three other participants used a pre-specified transportable AVM. For comparison, valuations using the cost approach were prepared by the author. Standard IAAO statistical quality measures (IAAO 1999, 41–44) were applied to actual 2004 sales and 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. Neither direct sales comparison nor artificial neural networks were considered in this research.

Literature Review

The first attempts at using multiple regression analysis (MRA) to estimate property market value occurred around 1970 (Gloudemans and Miller 1976). Prior to that time, the most widely used method was the traditional cost approach, done primarily by hand with minimal market analysis.

The formal description of the adaptive estimation procedure (AEP), also called feedback, first appeared in the late 1970s (Carbone, Ivory, and Longini 1980). Carbone’s PhD dissertation provided a rigorous academic definition of the technique (Carbone 1976). This procedure tests and systematically adjusts model coefficients, converging upon the set of coefficients that minimize an error term (IAAO 2003, 12). Schultz makes a case for the use of feedback in his 2001

Assessment Journal 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, Ben White (1995). White's paper provides an informative discussion of the revaluation process used throughout North America.

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). 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.

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.

Variations of the hybrid technique referred to in this research as the transportable cost-specified market (TCM) approach have been the subject of numerous papers by assessment professionals. As early as 1966, Franklin Graham, Assessor of the City of Wisconsin Dells, Wisconsin, 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 the 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 and Adams 1991). Ward provided a demonstration on the use of feedback to calibrate cost models at the 1993 IAAO Annual Conference on Assessment Administration (Ward 1993). This author presented a paper at the 1995 IAAO 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. At the Integrating GIS and CAMA 2005 Conference, Gloude-mans and Nelson presented a paper describing "how the District [of Columbia] used SPSS's 'Non-linear' MRA procedure to calibrate their cost structure using sales data in what can be called a fully 'market calibrated cost model.'" (Gloude-mans and Nelson 2005, 2 [Abstract])

As the technology for using computer-assisted mass appraisal matured, statistical standards were introduced to measure the quality of CAMA-produced values. An excellent example of these improvements is described 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 property 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.

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)

Even though a large body of literature exists on the subject of mass appraisal and the importance of accuracy in the

application of CAMA AVM models, there was not a single paper that reported on the proposed topic of this research—the evaluation of the relative performance of the primary CAMA methodologies used throughout the world.

Method

The primary purpose of this controlled experiment was to compare the performance of the automated valuation models used in computer-assisted mass appraisal. It was not intended to be educational in the use of the techniques themselves, as was Ward and Steiner's 1988 research. Since equitable property taxation depends upon having underlying value assessments that are 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 conformance with the *Uniform Standards of Professional Appraisal Practice* (USPAP). In particular, mass appraisal work must be conducted according to Standard 6 (Appraisal Foundation 2003). The quality of assessment work is measured in terms of uniform treatment of every property to ensure the highest degree of equity and fairness for individual property owners. 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. Gloudemans has done extensive research into the COD statistic and his 2001 paper provides a useful discussion of confidence intervals for the coefficient of dispersion (Gloudemans 2001).

To have assessments that exhibit uniformity, the practitioner wants the “scatter” of individual assessments (A) compared to their actual sale transaction

amounts (S) when they subsequently sell in the market (the A/S ratios) to approximate 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 used for the four mass appraisal methodologies applied in this research is the COD mean difference.

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) approach; μCOD_{COST} = the population mean coefficient for the cost approach (COST).

The null hypothesis is that the μ CODs will all be the same, that is, not significantly influenced by the choice of method. The research hypothesis is that the selection of method will cause the μ CODs 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 $H_a: \mu COD_{MRA}, \mu COD_{AEP}, \mu COD_{TCM},$ and μCOD_{COST} are not all equal. The research hypothesis further 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 $p \leq 0.05$.

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. The population, obtained

from a Midwestern assessing jurisdiction, included 22,785 existing single family residential properties with their descriptive characteristics, representing 52 distinct neighborhoods, which 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* [Oregon Department of Revenue 2004] for a more detailed description of sales sampling, sale validity, and market areas.)

The test sample consisted of the 1,299 properties in the population that sold in 2004. These sales had been screened by the assessing staff to verify that they were arm’s-length market transactions. 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 but 2004 sales were used instead. For example, 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, Ivory, and Longini 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 for model testing (Ward and Steiner 1988, 45).

The justification for using the following year’s valid market sales as the control group was that it more closely resembled 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, from a population of 22,785

parcels from the period 1999–2003, a total of 5,546 jurisdiction-validated sales, with characteristics as they were at the time of the sale, were available for use in model development. Each modeler was free to use as many or as few of these historical sales as desired. Once their models were constructed, they were used to blindly estimate the selling prices of the 1,299 jurisdiction-validated 2004 sales. All 1,299 sales were used for testing the resultant value predictions, that is, no outliers were eliminated. 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 know who the other participants were.

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 was ensured by the random activity characteristic of the real estate market. Although Hamilton (1997) states that a sales sample created through random market activity may not be fully representative of the population for various reasons, this was not considered a factor in the current research because it was 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 included as a TCM participant since they had been established prior to the actual sale dates of the 2004 test parcels. After reviewing the initial research report, one participant suggested that this may not be valid, so

the assessed values were 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), the cost estimates for the experiment were calculated by the author using two different AVMs based upon Marshall & Swift cost data (2003). One AVM was based upon Section A of the September 2003 *Marshall & Swift Residential Cost Handbook*, implemented using a large Microsoft® Excel spreadsheet. The notes and assumptions used for this spreadsheet implementation are summarized in Appendix A. The other was developed 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, or other such factors.

The Experiment

The first phase consisted of recruiting highly qualified participants for the experiment. Potential differences in modeling skill among participants represented an area of uncertainty. As the IAAO *Standard on Automated Valuation Models* states in its discussion of 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) Therefore, only qualified, experienced modelers were invited to participate. Among them were contributors and reviewers of the IAAO

AVM standard (2003, 2). The practitioners who participated in the research were as follows. Their home states are provided in parentheses.

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/Char Cuthbertson
as a team (Florida/Indiana)

In discussing potential time commitments, it was agreed that no participant should spend more than 24 hours on the research project.

The second phase of the experiment involved extracting and organizing data files for distribution to participants. The six AEP and MRA model-building participants were provided with the 40 data items listed in table 1 for the 5,546 sales. These were extracted from the jurisdiction's SQL Server market database and placed into Microsoft® Excel spreadsheets. A spreadsheet with the same layout but without sales information was provided for the 1,299 sale parcels from 2004 that comprised the control group to be valued for the test. Each AEP and MRA model builder was encouraged to use their preferred modeling software.

All participants were supplied with the jurisdiction's established land values as of December 31, 2003 (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 is based on 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 calculation was supplied with all 2004 sales information removed,

all assessment information removed, and jurisdictional identity removed. Although the test was blind for all participants, the three who used 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, and 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 three TCM participants had to use a cost-specified AVM as their starting point and then apply two additional market variables before using 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. Two sets of cost calculation results for the 1,299 properties in the test group were furnished by the author based upon two different cost AVM model specifications using Marshall & Swift cost data from September 2003. Finally, in order to have one other interesting perspective, the jurisdiction's statistics for the 2004 test

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 arm's-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	FinSFL	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 = 1½ 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

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. (The jurisdiction's figures were later removed from one result set at the suggestion of a participant).

Thus, 12 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 for use in property taxation as of January 1 (or other statutory tax lien date) each year.

Phase 3 of the experiment involved processing each of the 12 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 ensure 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, it was determined that both sales should be included, 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 phase 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. During this analysis, a single potential outlier surfaced within the results of the six market model participants.

Its COD fell more than two standard deviations from the mean of the market approach group. Therefore, it was decided to present the results both with the outlier included and with it excluded. Its inclusion or exclusion does not change the overall results.

Statistical Results

The four automated valuation model types most commonly used in mass appraisal 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 these major AVM methods and to analyze the relationship between AVM type chosen and the resulting coefficient of dispersion (COD). A lower COD is an important indicator 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

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

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 overall results. Table 3 contains descriptive statistics for the result sets with (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. Appendix B provides some additional information in non-technical terms to assist in the understanding and interpretation of the statistical results.

Discussion

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 true: a market-calibrated AVM will predict selling prices more accurately than a purely cost-based

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

AVM. What may be surprising is that the hybrid transportable cost-specified market (TCM) approach, using only two market variables, 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 nonetheless widely used in various forms.

The research further served to confirm 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) Skilled practitioners were recruited for the experiment so that results would not be influenced by varying skill levels. However, the data used was from a jurisdiction where there is an inadequate budget for proper field data collection and verification. As a result, it was anticipated that the quality and completeness of the available data would only be adequate to achieve average results at best. As table 2 shows, five of the nine

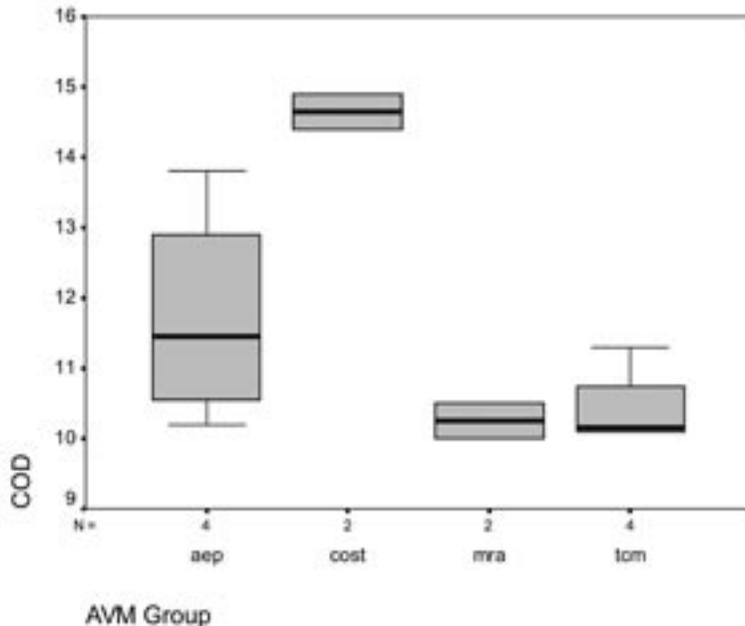
participants using market AVMs achieved CODs between 10.0 and 10.5 with three different software packages.

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 the 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

- Appraisal Foundation. 2003. Standard 6: Mass appraisal, development and reporting. In *Uniform standards of professional appraisal practice*. <http://www.appraisalfoundation.org/html/USPAP2003/standard6.htm> (accessed December 31, 2004).
- Bonbright, J.C. 1937. *The valuation of property*. New York: McGraw-Hill.

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



Calhoun, C. 2001. 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. PhD diss. Abstract in *Dissertation Abstracts International*. UMI No. 7618072.

Carbone, R., E.L. Ivory, and R.L. Longini. 1980. Competition used to select a computer-assisted mass appraisal system. [Electronic version]. *Assessors Journal* 15 (3): 163–167. <http://www.iaao.org/> (accessed May 7, 2005).

Clapp, J.M. 1977. The cost approach to mar-

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

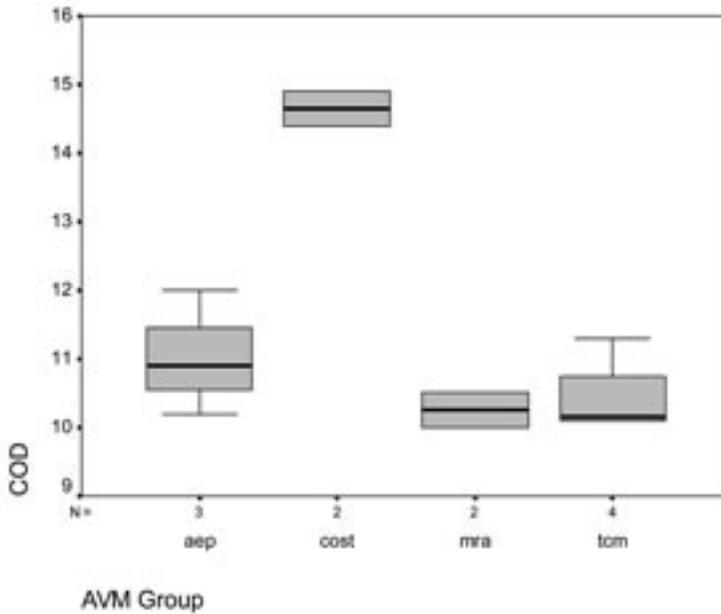
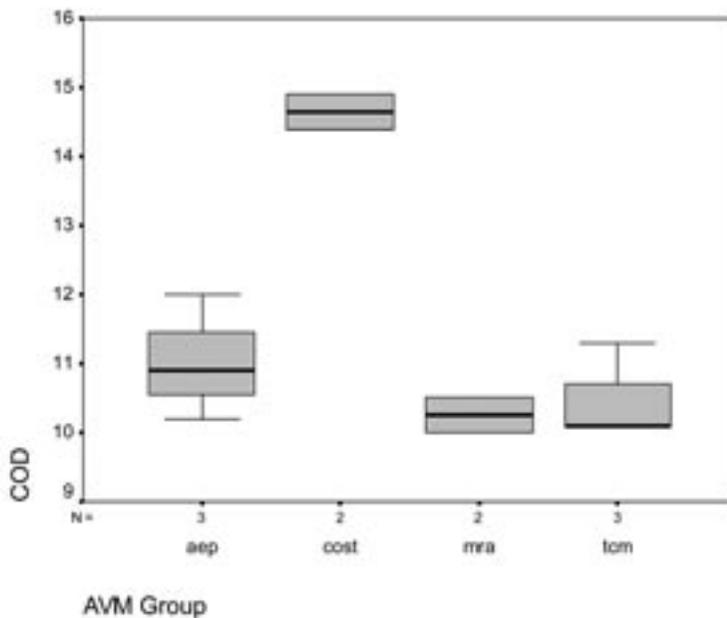


Figure 3. Box plot of coefficient of dispersion (COD) means for AVM groups (jurisdiction results removed)



ket value: Theory and evidence. *Assessors Journal* 12 (1): 43–46. <http://www.iaao.org/> (accessed May 7, 2005).

Eckert, J.K. 1986. Calibrating the generic model using construction cost data. *Assessment Digest* 8 (1): 11–15. <http://www.iaao.org/> (accessed May 7, 2005).

Gloude-mans, R.J. 1981. Simplifying MRA-based appraisal models: The base

home approach. *Assessors Journal* 16 (4): 155–166. <http://www.iaao.org/> (accessed May 7, 2005).

Gloude-mans, R.J. 2001. Confidence intervals for the coefficient of dispersion. *Assessment Journal* 8 (6): 23–27. <http://www.iaao.org/> (accessed May 7, 2005).

Gloude-mans, R.J., and D.W. Miller. 1976. Multiple regression analysis applied to

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)

residential properties—A study of structural relationships over time. *Decision Sciences* 7 (2): 294.

Gloude-mans, R.J., and W.R. Nelson, Jr. 2005. Market calibration of cost models. [Paper]. *Proceedings of the Integrating GIS & CAMA 2005 Conference* [CD-ROM] February 15-18, Savannah, GA.

Graham, F.D. 1966. Comparative method for mass assessment of residential real estate. *Assessors Journal* 1 (3): 41–54. <http://www.iaao.org/> (accessed May 7, 2005).

Hamilton, T.W. 1997. Sales sample data limitations and heteroskedasticity effects on property tax equity and incidence. Abstract in *Dissertation Abstracts International* 59 (02), 572A. (UMI No. 9803426).

IAAO. 1999. Standard on ratio studies. *As-*

essment Journal 6 (5): 23–65. <http://www.iaao.org/> (accessed February 27, 2005).

IAAO. 2003. *Standard on automated valuation models (AVMs)*. http://www.iaao.org/pdf/AVM_STANDARD.pdf (accessed December 24, 2004).

Ireland, M.W., and L. Adams. 1991. Transportability of a general-purpose residential market-calibrated cost model. *Property Tax Journal* 10 (2): 203–224. <http://www.iaao.org/> (accessed May 7, 2005).

Marshall & Swift. 2003. *The residential cost handbook, September 2003 data*. Los Angeles: Marshall & Swift.

Moore, J.W. 1995. The market correlated stratified cost approach. In *Proceedings of the 61st annual conference*

Table 5. Tukey test evaluating pair-wise differences between the means (outlier removed)

Multiple Comparisons
Dependent Variable: COD

			Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
	(I) AVM Group	(J) AVM Group				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.

of the International Association of Assessing Officers. Chicago: IAAO, 223–236.

Oregon Department of Revenue. 2004. *Ratio manual*. http://egov.oregon.gov/DOR/PTD/ratio_manual.shtml (accessed January 2, 2005).

Schultz, R. 2001. The other market model. *Assessment Journal* 8 (1): 48–50. <http://www.iaao.org/> (accessed May 7, 2005).

Ward, R.D. 1993. Using feedback to calibrate a cost model. In *Proceedings of the 59th annual conference of the International Association of Assessing Officers*. Chicago: IAAO, 209–221.

IAAO, 253–256. <http://www.iaao.org/> (accessed May 7, 2005).

Ward, R.D., and L.C. Steiner. 1988. Comparison of feedback and multivariate nonlinear regression analysis in computer-assisted mass appraisal. *Property Tax Journal* 7 (2): 43–67. <http://www.iaao.org/> (accessed May 7, 2005).

White, B. 1995. 1995 Denver County revaluation using multiple regression analysis. In *Proceedings of the 61st annual conference of the International Association of Assessing Officers*. Chicago: IAAO, 209–221.

Table 6. Tukey test evaluating pair-wise differences between the means (jurisdiction results removed)

Multiple Comparisons
Dependent Variable: COD

			Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
	(I) AVM Group	(J) AVM Group				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.

Appendix A.

Marshall & Swift Residential Cost Calculator Method Notes and 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 most commonly used building method 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), 20% and 33% respectively were added to the Excellent rates.
5. There was minimal roof cover data available, so the base for each quality class was used 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 of which 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. One-and-One-Half-Story homes, were 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 Average Quality) so it was adjusted using page B-19 as the reference source.
11. Fireplaces were assumed to be direct-vent gas in Fair and Average homes, and in 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 was assumed to be “partitioned” in the M&S basement cost tables.
13. Attic finish rate was from page B-24; M&S does not provide a rate for unfinished attic in Section A calculator method.
14. The base allowance by quality class was added for built-in appliances.
15. Basements for Fair, Average, and Good Quality are assumed to have 8-inch poured concrete walls; those for V-Good and Excellent have 12-inch poured concrete.
16. Porches were assumed to have roofs without ceilings (costs are added together for the porch rate); decks are without a roof.
17. Garage wall type was not in the data, so for the Quality Fair through Good, “Wood Shingle” cost was used. For V-Good and Excellent, “Brick Veneer” cost was used.
18. The jurisdiction’s “Extra cost amount” was added for pools and other significant yard items that had been 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 the Significance of the Statistical Results

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. A research tool called hypothesis testing is used 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 avoid any random error that might be induced by the skill level of any single expert and to 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. The SPSS output would indicate that the results were considered significant if the p -level was less than $.05$.

The SPSS output for a test of statistical significance is detailed in table B.1. The column labeled “Sig.” is the p -value, and in this example, $p = .001$. This means that 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. It 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 B.1. Example SPSS output for a test of statistical significance

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)

