AVM Model Building Study: Skinning the Cat

Robert J. Gloudemans and Erin Montgomery

Robert J. Gloudemans is a mass appraisal consultant and partner in Almy, Gloudemans, Jacobs & Denne. Erin Montgomery, PhD, works for the Northern Ireland Statistics and Research Agency and at the time of this paper is on special assignment with the Valuation Lands Agency of Northern Ireland.

Abstract

This paper was prepared as part of the competition on development of Automated Valuation Models (AVM) presented at the URISA/IAAO *2007 Conference on Integrating GIS and CAMA*. Professor John Clapp of the University of Connecticut provided participants with data from Fairfax County, Virginia divided into a model group and test group[1]. Participants were asked to develop a best possible model on the model group and apply the model to the test group, for which sales prices were not provided. This paper describes the data provided, our approach to the problem, final model, and estimated performance for the holdout sales. Although the data were less complete than typical of an assessor's office, the project was a fascinating challenge with particular emphasis on the important issues of time trend analysis and location adjustments.

Examination of Data

The model data set contains 51,190 sales over the almost 25-year period, January 1967 through June 1991 (except for a small number of July 1991 sales). The test or holdout data set contains 7,177 sales from January 1972 through December 1991. Notice that the holdout data set begins five years later than the model data set and runs through the end of 1991, whereas the sales effectively end in June 1991. In fact, 30.3% of the holdout sales are from the second half of 1991.

Since the test data set began in 1972, we decided to drop sales from 1967 to 1970 in model development (leaving 50,040 sales in the model group). Sales from 1971 were retained to better anchor the start point and bolster sample size at the beginning of the period where sales were far less plentiful than later.

The data sets contain land area, number of rooms, number of bedrooms, number of baths and half baths, number of fireplaces, year built, census track, x-y coordinates, and the 15-closest neighbors to each sale or holdout and distance thereto. Although some additional census data was also provided, it was constant across each census tract. The databases did not contain the following data items commonly used to develop CAMA models:

[1] For a description of work by Professor Clapp and colleagues on the same data, see Bradford Case, John Clapp, Robin Dubin, and Mauricio Rodriguez, Modeling Spatial and Temporal House Price Trends: a Comparison of Four Models, *The Journal of Real Estate Finance and Economics*, vol. 29, no. 2 (September 2004).

- Market area and neighborhood
- Living area
- Construction quality/grade
- Building style
- Effective year built or condition
- Secondary size variables and amenities such as basement areas, garages, porches, balconies, pools, etc.
- Land or location attributes such as water, golf course, greenbelt, view, and traffic influences

These significant omissions challenged participants to utilize available data in more imaginative ways. Clearly, census tracts and the nearest neighbor data (or x-y coordinates on which they were based) will have to proxy for more traditional neighborhood and location variables. Determining accurate time trends over (and somewhat beyond) so long a time span was a second daunting challenge.

Model Specification and Calibration

The model data set was robust in terms of having over 50,000 sales to work with. Aside from some apparent vacant land sales, the data appeared rationale and consistent, due in part to that fact that the project organizers had pre-filtered it to remove missing and out-of-range data. For consistency with the test data set, we removed a small number of sales from the model data set that exceeded the range of values for the holdout data (e.g., sales with more than 16 rooms). We also excluded 81 sales with extreme prices relative to number of rooms or extreme prices for their census tract. Some of which we suspect to be vacant land sales. Of course, the obvious challenge in any filtering of this sort is to remove likely invalid sales while preserving the ability to predict the lowest and highest price ends of the holdout data set.

In the absence of living area information, the available data on number of rooms and bathrooms took on added importance. After some early exploration, number of bedrooms was dropped because of its very high collinearity with rooms and baths. Baths and half baths were converted into a single variable (computed as full baths plus ½ of half baths) and, alongside rooms and number of fireplaces, were incorporated into the model as binary variables (the base or reference property being a house with 8 rooms, 2.5 baths, and 1 fireplace).

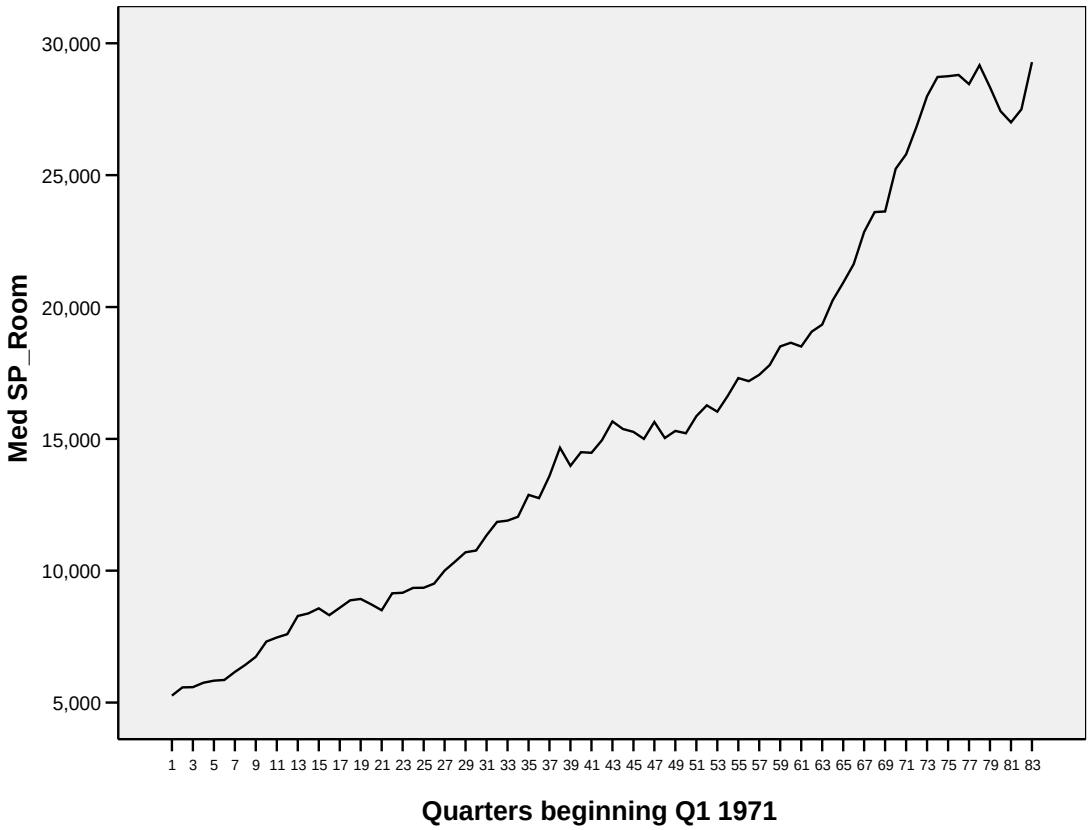
Land area was very effectively incorporated into the model by dividing it by number of rooms to yield lot size per room. In addition a percent good variable was included in the model based on a transformation of effective age.

Partly for convenience and because market areas were not provided, we chose to develop a single multiplicative model with census tracts serving as proxies for neighborhood variables (several of the smallest census tracts were combined with other ones). The large range of sales prices over so broad an area dictated use of a multiplicative model over a somewhat simpler additive or linear model.

Time Adjustment

We developed an initial time variable based on sale quarter (1971Q1 = 1, 1970Q2 = 2, … , 1991Q2 = 82). Figure 1 shows the trend in sale price per room over the 82-quarter period. As the figure shows, the Fairfax real estate market trended generally upward during this time, but stagnated in the second quarter of 1989 and then fell briefly before recovering. Accordingly we capped our time variable at 74 (1989Q2) and found that it was best represented in the model by raising it to the 1.15 power, reflecting a gradual acceleration in prices until 1989.

Figure 1 – Graph of Median Price Per Room with Time



In our final model we reversed the time variable (1989Q2 = 1, 1989Q1 = 2, …, 1971Q1=74), so that the model would be anchored at the critical end of the period, where most of the holdout sales occurred, and the model coefficients would reflect value at that point in time (rather than at the beginning of the period). With the time variable reversed in this manner, the optimal exponent, not unexpectedly, was 0.85, indicating that prices trended downward but at a decelerating rate as one went farther and farther back (reflecting the reverse of the accelerating trend going forward in time). In addition, we included sale year binaries to capture departures from the overall trend and pseudo time binary variables (one time variable coded 1-74 for each census tract other than the chosen base area) to allow for the fact that some areas may have appreciated faster or slower than the overall rate. Notice that since the model is anchored at the end of the period and includes time of sale variable, it will predict value as of the sale date, as desired in this case.

Location

In addition to census tract binaries, we utilized the nearest neighbor (NN) variables to capture location effects as described below.

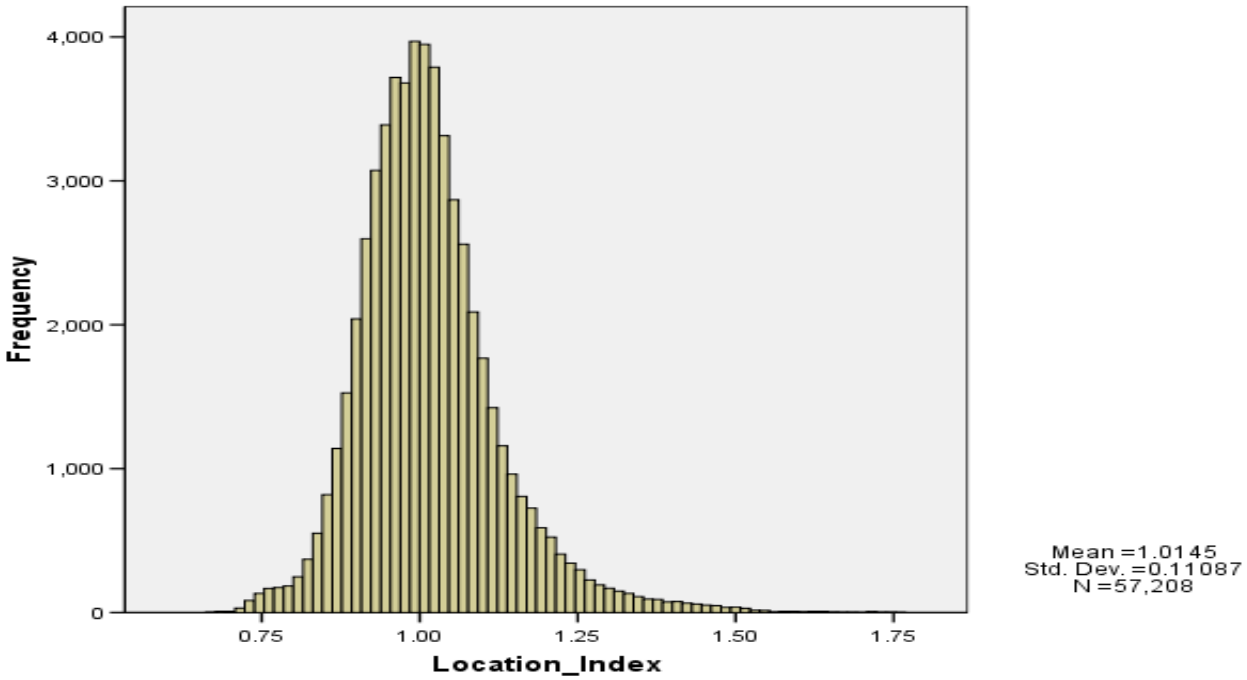
We developed a model using the variables described above except for NN variables, identified and removed a modest number of outliers, reran the model, and saved the predicted values. We then divided sales prices by predicted values to form sale/assessment ratios (SARs) and identified the 15 NNs and corresponding SARs for each sale. Hence it was possible to create a location variable whose value took account firstly of these SAR's and secondly of the physical distance of each sale to its 15 NNs.

Initially the units in which the distances for each NN were given proved somewhat problematic in that the distances were represented in units of "squared decimal degrees" (which for technical reasons could not be accurately converted into feet, miles, or some other standard measure.) The algorithm employed however overcame this difficulty in that it utilized only the relative distances, i.e. the unit of measure did not matter.

The question before us was how to weight each of the 15 NNs. Naturally it was found that some sales had a small number, perhaps just one or two, NNs that were in close proximity while the remainders were relatively far away. Conversely some sales had NNs that were all very close except for a few that were relatively far away. After some deliberation and partly for simplicity, we decided to weight each NN based half on distance and half on rank. As it turned out, the location index was somewhat normally distributed and centered on a mean of approximately 1, as demonstrated in figure 2.

This measure thus provided a "smooth" surface layer that could compensate for missing land attributes: a high index represented a desirable location, while a low index represented an undesirable one. With the location index successfully created the improvement on the MRA model was quite dramatic. (A limitation of this approach, of course, is that all distances are assumed equal, regardless of whether the nearest neighbors border a lake or heavily traveled street).

Figure 2 – Location Index



Model Results

By the conclusion of the MRA modeling the sales records used were as follows:

Sales Includes	49,764
Sales Excluded	267
Holdouts	7,177

Of 50,040 sales in the model data set from 1971 onward, only 267 (0.5%) were excluded for one reason or another (atypical price for neighborhood, ratio outlier, etc.). The adjusted R-square is 0.940. Figure 3 shows the final model in abbreviated form (only three neighborhood binary, three neighborhood time variables, and three sale year binary variables are shown). All variables have reasonable coefficients and show the expected progression (recall that the base property has 8 rooms, 2,5 baths, and one fireplace). Sales ratio statistics, both with outliers included and excluded, are show below.

Outliers Included

Sales	50,031
Median	1.000
Weighted Mean	0.988
Minimum	0.205
Maximum	8.311
PRD	1.032
COD	0.112

Outliers Excluded

Sales	49,764
Median	1.000
Weighted Mean	0.986
Minimum	0.390
Maximum	2.226
PRD	1.024
COD	0.102

Importantly, the COD is 11.2 with all sales in the model set included (other than the 1967-1970 sales as previously explained) and 10.2 with the worst data anomalies and outliers excluded. It should be noted that inclusion of the location index had a hugely beneficial effect on model performance (reducing the COD by approximately 1.5). Although not shown, sales ratio plots demonstrate excellent equity across all available property attributes and date of sale.

Figure 3 – Final Model

Model: 67

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	10.276	.009		1095.322	.000
rooms__4	-.262	.007	-.044	-36.999	.000
rooms__5	-.185	.004	-.073	-51.079	.000
rooms__6	-.121	.002	-.077	-51.983	.000
rooms__7	-.057	.002	-.041	-29.284	.000
rooms__9	.070	.002	.049	36.932	.000
rooms__10	.122	.003	.058	46.029	.000
rooms__11	.153	.004	.048	38.934	.000
rooms__12__13	.186	.006	.039	31.606	.000
rooms__14__15__16	.269	.018	.017	15.030	.000
totbaths__1	-.151	.004	-.056	-38.280	.000
totbaths__1.5	-.104	.003	-.044	-32.698	.000
totbaths__2	-.084	.003	-.044	-33.524	.000
totbaths__3	.018	.002	.012	9.457	.000
totbaths__3.5	.069	.002	.041	31.711	.000
totbaths__4	.122	.007	.021	18.024	.000
totbaths__4.5	.206	.006	.042	33.717	.000
totbaths__5	.256	.016	.018	15.559	.000
totbaths__5.5__6	.285	.016	.021	17.955	.000
totbaths__6.5__7__7.5	.419	.056	.008	7.443	.000
fire__0	-.068	.002	-.048	-34.536	.000
fire__2	.097	.002	.061	49.149	.000
fire__3	.215	.004	.064	52.129	.000
fire__4	.272	.008	.038	33.016	.000
ln_pctgood	1.269	.010	.181	121.163	.000
ln_land_punit	.164	.001	.298	182.784	.000
Q.85	-.044	.000	-.799	-279.529	.000
Location_Index	.940	.006	.187	160.475	.000
NBHD__1	.221	.015	.027	14.590	.000
NBHD__2	.310	.012	.045	26.670	.000
NBHD__3	.200	.009	.033	21.246	.000
Q_N1	-.005	.001	-.010	-5.594	.000
Q_N2	-.009	.001	-.019	-11.711	.000
Q_N3	-.002	.001	-.004	-2.915	.004
syear1988	.018	.003	.010	6.930	.000
syear1989	.010	.003	.006	3.627	.000
syear1991	-.036	.004	-.013	-9.972	.000

Generation of Values for Holdout Data Set

As a final step, we applied the model to the holdout data set, allowing our time variables to capture time of sale influences (prices had on average approximately quintupled over the 20-year period from 1971 to 1991). Since the holdout data set has not been purged of data problems or extreme prices, we of course must estimate performance based on the full model set with outliers included. Further, some additional deterioration, both in the level and uniformity of values, may occur due to the fact that over 30% of the holdouts are from beyond the timeframe for which sales prices were provided.

Given these considerations, we estimate a median assessment level of 0.98 to 1.02 and a COD of 11.3 to 11.5 for the 7,177 sales in the test data set.

Conclusions

Multiplicative MRA with the variables described proved to be an effective choice for the problem at hand. Naturally the model's predictive ability suffered due to the lack of more traditional property attribute data. Additional obstacles included outliers and holdouts beyond the sale period.

Still, the model was able to accurately capture the underlying time trends over a 20-year time span in which prices had risen enormously. Location too was reasonably accounted for by drilling into the NN information for each sale. The chosen model is simple in structure and easily explained. The coefficients too are explainable and intuitive or, in short, make good appraisal sense.

Lastly, given the available data, the final model exhibits good performance statistics, particularly so when obvious outliers are removed. While the data clearly pales in comparison to that found in the modern assessor's office, the results demonstrate the ability of MRA to skin the cat even when the database falls short of what might normally be available.