Introduction

The International Association of Assessing Officers (IAAO) defines the market approach: "*In its broadest use, it might denote any valuation procedure intended to produce an estimate of* ***market*** *value, or any valuation procedure that incorporates* ***market****-derived data, such as the stock and debt technique, gross rent multiplier method, and allocation by ratio. In its narrowest use, it might denote the sales comparison approach*."

Under this definition, all three approaches to value could be considered a "market approach", however for the purpose of this document, the term market approach is used interchangeably with the term sales comparison approach. IAAO defines this approach as: *One of three approaches to value, estimates a property's value by reference to comparable sales.* Also referred to as the "comparable sales approach".

The market approach has two components, the model, which consists of coefficients used to adjust individual comparable sales to a subject and the comparable selection criteria and weights which will determine how comparability is measured and how properties will be selected as comps.

The goal of the appraiser calibrating a market model is to fit the model to a specific set of data. The data is the sales set, which represents the "market" for that location at a specific point in time. A perfectly fit model has an R square of 1 which means the variables selected in the model result in a perfect estimate of the dependent variable (sale price) for every sale. While we may not get a perfectly fit model, by monitoring the market, collecting pertinent data when possible, and carefully selecting variables, we should get close.

Some of the analyses for this stage in the process include:

- Initial sales ratio to determine the current overall level of value.
- Number of sales vacant and improved, by neighborhood.
- Investigate real estate listings and note the amenities and locations that are considered desirable; e.g. if many listings mention the school system, then consider adding school district as a variable or as part of the comparable selection process.
- Analyze neighborhoods with comparable selection in mind. Are there enough properties in each neighborhood to allow for ample comps? Are there properties assigned to the same neighborhood that are geographically very far apart? Should they be separated for comparable selection purposes?

Once neighborhood lines are drawn and finalized, each neighborhood is assigned to a group. Neighborhood groups are primarily used for comparable selection, so this should be foremost in the appraiser's mind. Neighborhood group may also be used as a variable. A group could be

made up of a single neighborhood, or several similar neighborhoods. The goal is to ensure that there are an adequate number of sales in each group for selection as comps for the majority of properties.

Groups are then assigned to clusters or market areas. Each cluster is assigned to a model. Depending on the size of the jurisdiction, there may be just one model or there may be several. The determining factor will be based on whether the same variables will explain sales price in each area. It is common for there to be separate models for different types of properties, such as dwellings versus condominiums. Also, consideration may be given to rural versus urban areas or any other influence(s) that cannot be captured in a single variable. Keep in mind that locational differences will be captured in the land variable so this alone does not warrant a separate model. The more data available to calibrate the model, the better the results will be, so fewer models is preferable.

Variable Selection (Model Specification)

A variable is a data element that is used, or could potentially be used, in a regression model. Some variables are a means of assigning value to a character field (such as CDU) and will never be used in the model, but are used by other variables that would be included. Having over 100 variables would not be uncommon.

A coefficient, on the other hand, represents the adjustment used in a final market model to adjust comps to the subject. Only a small percentage of variables will end up as coefficients in the final market model.

As mentioned previously, the goal of the appraiser calibrating a market model is to fit the model to a specific set of data. Selecting the variables that best explain why properties sell for what they do is an iterative process, and often requires trial and error to perfect.

The regression process will attempt to allocate the sale price into buckets – the buckets represent the variables used in the model. It can only allocate sale price to the buckets that are present, so if one of the buckets is missing, the value associated with it will be allocated into one or several of the other buckets. When regression is run, it is common for several of the buckets to be removed because they contribute little towards explaining sales price. This does not necessarily mean that the variable should ultimately be eliminated.

It's important for the appraiser to consider not only what variables the market indicates drive value, but also any data item that needs to result in a change of value. For example, you may find that the market doesn't recognize any value for fireplaces; however, the appraiser, or client, may want to recognize a difference in value between a house with and one without a fireplace. In this case, the variable will be forced into the model (constrained).

Some guidelines for variable selection:

- The number of variables will depend on the number of sales in the sales set. The more sales, the more variables that can be used. A rule of thumb is 5 sales for every variable so a sales set with 100 sales could produce a model with no more than 20 variables.

- Due to the limitation of the number of variables, the appraiser may need to combine items such as decks and porches, rather than using individual variables for each.

- The contribution of land and outbuildings, with the exception of pools or garages, perhaps, will be measured by using the cost value for these items. The resulting coefficient will represent the percentage of the value the model determines is appropriate. A coefficient of 1 would indicate that the cost model and market model indicate the same value.

- There must be sales representing every variable. If you include a variable to capture the contribution of multi-families, then there would need to be several sales of multi-families in the sales set.

- Variables can be straight data elements (SFLA) or may be transformed (square root of SFLA).

- Binary variables (yes or no) will result in a flat rate adjustment.

- Multiplying a binary variable by square footage converts it from a flat rate adjustment to an adjustment per square foot of living area.

- It's important to decide whether the SFLA being used in the manner above should include finished basement or not. If it is included, the variable to capture the value of finished basement will most likely be negative, as it is typically worth less than above-grade living area.

- Subjective data, such as condition or CDU, must be assigned numerical values in order to be used in the model. These values must be carefully considered and tested.

- Data elements may be splined, or separated into groups. Age is a variable that is often introduced as a spline variable, which would indicate that properties do not depreciate in a straight-line manner, but rather the amount of adjustment required would differ for each grouping.

There will be very few variables that reflect data exactly as it exists in the database. Most variables will be transformed from their original state. Some examples of transformations:

- Binary – converts to yes/no (1/0). Examples include:

- Pool Y/N – this will capture the **flat value** of a pool (in ground) regardless of size.

- Style variables – this will capture the **flat value** of a specific style, such as raised ranches. Multiplying this variable (0 or 1) by the SFLA will create a variable that captures the **value per square foot living area**.

- Mathematical – adding, subtracting, multiplying or dividing variables. Examples include:

  - Living unit – 1 – this will capture the value associated with each additional unit in a multi-family. A single family has a living unit = 1 so once you subtract 1 – you get 0.

  - Basement – 4 – Basement is a code. A full basement is a code 4, while no basement is a code 1. Subtracting 4 solves for properties that have less than a full basement. The adjustment will be a negative, as 1 – 4 = -3 and the adjustment will be greater for no basement (-3) than for part basement (code 2, or -2). Full basement would have no adjustment as 4-4 = 0.

  - SFLA*Grade factor * CDU factor – By combining these variables, you are accounting for the amount of living area, the quality of construction and the relative condition of the property in a single variable, much like buyers in the market do.

- Exponential – are used to capture non-linear relationships. An exponent > 1 expands the differences, < 1 contracts the differences and < 0 reverses the direction of the numbers. Examples include:

  - SQRT of fireplaces – square root = exponent of .5. This variable assumes that the value of each fireplace decreases with each increment. For example, if the adjustment for SQRT of fireplace is +5,000, 1 fireplace would net an adjustment of 5,000 while 2 would net an adjustment of +7,071 (SQRT of 2 = 1.414 * 5,000)

- Reciprocal – 1/X

- Logarithmic – A logarithm is an exponent to which a given base must be raised to obtain a specified number. The common log (or base 10 log) is the power to which 10 must be raised to obtain a given number. The natural log (ln) is the power to which the number 2.71828 (base .e.) must be raised to obtain a given number. Logarithmic adjustments are useful for size variables, such as SFLA, where economy of scale is a factor.

- Scalar – are used to assign a factor to categorical data, such as CDU. Factors can be centered around 1 or 0, where the average or typical is set at 1 or 0.

- CDU – since CDU (or condition) is a character, a value must be assigned. Age will almost certainly be a variable in the model, so this variable is strictly condition. This is a tricky variable to deal with since on the cost size, the difference between EX and AV for a new house is negligible, however for an older house, the difference can be quite significant. This can be handled through data collection (being careful not to assign EX to new houses), or by solving it in the model itself.

Model Calibration

Once the candidate variables have been selected based, in part, by the composition of the sales set, regression modeling can begin. Sales should be reflective of the entire population and unique properties, such as multiple dwellings on a single parcel, mobile homes, or incomplete construction, should be eliminated from the sales set (although not invalidated as they are still valid sales). After the sales are extracted, the output should be carefully reviewed to ensure there are no data quality issues. Below is an example of the output from IAS.

VARIABLE STATISTICS (NON-ZERO ONLY) MODEL 0001

VNUM		SUM	SUM OF SQUARES	MEAN	STD DEVIATION	LO VALUE	HI VALUE	NON ZERO
0	SALPRICE	1.248838E+08	6.285511E+13	333023.3500	238454.8200	83000.000	2310000.000	375
0	DEP VAR	1.248838E+08	6.285511E+13	333023.3500	238454.8200	83000.000	2310000.000	375
0	SALE YEAR	7.537590E+05	1.515194E+09	2010.1040	0.6436	2009.000	2011.000	375
0	SALE MON	-1.746700E+04	8.298770E+05	46.5786	6.5991	- 60.000	- 36.000	375
0	COST VAL	1.176078E+08	5.475998E+13	313620.8000	218623.1000	85800.000	2121000.000	375
1	TOTVAL	1.176078E+08	5.475998E+13	313620.7500	218623.2000	85800.000	2121000.000	375
2	SP	1.248838E+08	6.285512E+13	333023.3500	238454.9000	83000.000	2310000.000	375
3	LUC	3.788000E+04	3.834604E+06	101.0133	4.6878	13.000	111.000	375
5	GRADEFACT	4.295007E+02	5.293501E+02	1.1514	0.3058	0.650	3.250	373
6	DOS	3.815231E+07	3.896289E+12	101739.5000	6267.6290	91002.000	110930.000	375
9	LAND	5.040620E+07	1.317919E+13	134775.9400	130842.2300	47700.000	1342000.000	374
13	LIVUNIT	4.030000E+02	4.770000E+02	1.0775	0.3385	1.000	4.000	374
15	TRUEGFA	3.641500E+05	3.924019E+08	976.2734	314.9155	280.000	2338.000	373
16	TRUETLA	6.321340E+05	1.267464E+09	1694.7292	726.1779	552.000	5583.000	373
17	STYLE	1.776000E+03	1.346200E+04	4.7613	3.6682	1.000	17.000	373
18	TLA	6.559370E+05	1.341528E+09	1758.5442	710.9630	552.000	5583.000	373
19	STORYHT	5.311001E+02	8.320701E+02	1.4238	0.4515	1.000	2.500	373
22	YRBLT	7.311820E+05	1.433987E+09	1960.2734	42.4518	1717.000	2010.000	373
30	STORYHT	5.311001E+02	8.320701E+02	1.4238	0.4515	1.000	2.500	373
91	CDUFACT	4.086005E+02	4.518297E+02	1.0896	0.1330	0.150	1.500	375
92	AGE	2.294300E+04	9.717621E+06	61.1813	149.0965	1.000	2011.000	375
117	LU-1	4.030000E+02	4.770000E+02	1.0775	0.3385	1.000	4.000	374

Upon review, there are 375 sales in the sales set, yet only 374 of them have a living unit assigned. Additionally, one of the sales does not have a land value and 2 parcels have no living area, style or year built. Prior to continuing, these data issues should be resolved. Keep in mind that there may not be data for every variable, such as porches, as not every property will have a porch. And since the sales should represent the entire parcel inventory, any edits done on the sales, should also be done on the population, as well.

There are several statistics that will be used to measure the quality of the model's value predictions. There are two categories of statistics; measures of goodness-of-fit and measure of variable importance.

- Goodness-of-fit measures

- R square (coefficient of determination) – R square measures the percentage of the variation in sales prices explained by the model. An R square of 100% would mean the model explains every variation in sales price (not likely); however a value of greater than 85% is acceptable.
- Adjusted R square – R square adjusted for degrees of freedom (number of observations in the set minus 1).
- Standard error of the estimate (SEE) – the standard deviation of the regression errors. The regression error is the difference between the model value estimate and the dependent variable (sale price or adjusted sale price).
- Coefficient of variation (COV) – the standard error divided by the average sale price.

- Variable importance measures

  - T-value – the ratio of a regression coefficient to its standard error. **The higher the ratio, the more significant the variable.**

  - F-value – the square of the t-value. The F Limit is used in the regression procedures to control the inclusion and deletion of variables from the model based upon the F-value. As a rule of thumb, an F-value of 4 indicates significance at the 95% level, 8 indicates significance at the 98% level, although most regression models are set much lower to include a greater number of key variables.

  - Coefficient of correlation – measures the linear correlation between two variables, ranging from -1 to 1. The closer to 1 (or -1), the more highly correlated the variables, meaning they are measuring the same item. Below is a correlation matrix output from IAS.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	LANDVAL	OBYVAL	BSMT	FIXTOT	FBLA	RECRM	SFLA*	AIRSF	CDUSF	DOSSF	T_ATGARA	T_CPAREA	AGE*SF	AGE10SF	AGE25SF	AGE40SF
01	1.00000	-.02217	.67606	.27390	.21747	.17018	.39781	-.10804	.27152	-.06602	.34383	.08649	.24647	.12516	-.23928	-.16107
02		1.00000	-.05021	-.02310	-.02287	.01275	-.05772	-.03800	-.06341	-.07277	-.21708	-.09548	-.05918	.17797	-.06403	-.07395
03			1.00000	.05749	.07409	.18938	.16448	-.28918	-.02807	-.01359	.08364	.08266	.19902	.08049	-.36672	-.21465
04				1.00000	.07586	.27304	.83764	.07053	.33554	-.11510	.46699	-.07696	.42043	.24756	.13036	.10117
05					1.00000	-.08279	.15321	.39369	.26281	.13248	.05792	.03252	-.04712	-.04345	-.10277	.20947
06						1.00000	.15418	-.09758	.09576	-.03161	.25531	-.01428	-.03451	-.03803	.06540	.09009
07							1.00000	.10306	.37343	-.11016	.42886	-.03068	.64712	.16811	.04291	.05126
08								1.00000	.35673	.04898	.11127	-.02189	-.07089	-.08468	.36837	.36967
09									1.00000	-.05931	.47772	-.12199	-.10707	.02551	.24623	.32523
10										1.00000	.00905	.01180	-.03385	-.10005	-.08954	.13069
11											1.00000	-.06272	.00054	.28054	.19772	.22890
12												1.00000	.08904	-.04848	-.11466	-.14085
13													1.00000	-.19017	-.37039	-.39143
14														1.00000	.18882	.07339
15															1.00000	.56138
16																1.00000

Line 01 compares variable 1 (LANDVAL) to each variable as they are displayed across the top. Line 02 compares variable 2 (OBYVAL). The highest correlation is with variable SFLA* (variable 7) and FIXTOT (variable 4). They are highly correlated at .83764, which seems logical as the bigger the house is, the more likely it is to have additional bathrooms. The analyst has to consider whether there is a need for the model to adjust for bathrooms, considering the model has

told us that much of the value associated with that item is already being explained in SFLA.

As mentioned previously, multiple regression analysis is an iterative process and the analyst should expect to process the regression model a number of times. In addition to the statistical indicators of a successful model, the analyst must consider whether the resulting coefficients make sense. **This cannot be overstated.** Some examples:

- The rec room coefficient is greater than the finished basement coefficient
- Fireplace is a negative coefficient
- Age is a positive coefficient
- The coefficient results in an adjustment that is not logical, such as a deck being worth $40/sf.

The market model that results from the reappraisal effort will be used to maintain values for years to come. It's important for the values calculated when data changes are made to the property, such as the addition of a deck or porch, make sense and are not significantly different from the values generated using the cost method.

If you recall the bucket metaphor, the model is going to allocate to sale price to the buckets available and since there is most likely data driving market value that is simply not captured in the number of variables available or not even in the CAMA database itself, it can result in anomalies that must be addressed.

To identify additional variables to include in a model to improve the results of goodness-of-fit statistics, the analyst should focus on parcels with the greatest standard error, meaning those where the regression estimate differ the greatest from the dependent variable (sale price). Attempt to find a pattern:

- Are they in the same neighborhood or group? If so, add a variable to the model to calculate either a flat or per sf adjustment for that neighborhood or group.

- Are they the same style? If so, introduce a style variable, preferably as a sf variable.

- Are they all the same condition? If so, review the factor assigned to the condition, or consider a variable that adjusts for just that condition (usually fair or poor).

Constraining variables

Constraining refers to forcing a variable into a model at a specified value. It can be used to force a variable into the model that was not significant or assigning a different rate than what the model predicted. It's a good idea to perfect the model prior to constraining any variables. Constraining will diminish the statistical performance of the model. Reasons to constrain variables include:

- To force a variable into the model that was eliminated as insignificant based on its f-value. Example: central ac is constrained at the value at which the model indicated prior to its being eliminated (stepwise regression).
- To force a variable with no sales on which to model. Example: unfinished area constrained to the same $/ sf value as the cost model.
- To assign a more logical value to a variable that came in as significant. Example: constrain the fireplace to a positive value when the model indicates a negative value.

Comparable Sales Model

There are two primary comparable sales models, one based on coefficients from regression or other statistical means, the other with adjustments based on cost estimates. The second type is sometimes referred to as "cosmetic comps" as the adjustments are lump sums based on the total cost value or land and building values separately.

The first step in the development of the comparable sales model (either type) should be the selection of variables used to determine the comparability "distance". This distance (not to be confused with physical distance) is then used to mathematically select those sales with the lowest distance, which should indicate that they are most similar to the subject.

The same thought process that went into the variable selection for the regression model applies here. What makes a property comparable to another in this market?

- Location (neighborhood and group)
- Style
- Amount of living area
- Age
- Grade
- Condition
- Living Units

In order for the model to calculate a market value, there will need to be ample comps so setting the criteria too high in the comparable selection will reduce the number of quality comps, while setting it too low may result in a value based on less than the best set of comps. Like the regression model, several passes may be required to ensure the model is performing as expected.

Once the variables are selected, weights are assigned. There are two types of weights:

- Constant: The weight is to be applied as a lump sum or constant amount, whenever the value of the variable differs between the sale and subject property, e.g., a weight of 100 may be entered for NBHD (neighborhood). If the sale is in a different NBHD than the subject, a weight of 100 will contribute toward the comparability distance calculation.

- Variable: The weight is to be applied to difference between the value of this variable (characteristic) for the sale and subject property, e.g. a weight of 0.1 might be applied as a "variable" weight to the difference in SFLA (sqft of living area). If there is a difference of 500 square feet between the sale and subject property a weighted difference of 50 will be contributed toward the comparability distance calculation.

As with the number of variables, the weighting has to be carefully applied to ensure that the number of useable comps is balanced by the quality of the comps selected. For example, if the maximum distance for a comp to be selected is 500 and the weight for neighborhood is set to 250 and neighborhood group at 300, a comp that is identical to the subject in every way except neighborhood group (and therefore neighborhood) would never be selected.

The Models

In North Salem, there was one residential model.

Model 1

VNAME	DESCRIPTION	COEFF
A/C	If has AC = ADJSFLA, else 0	45.55854073
ADJGRASF	(Grade Fact-1)*AdjSFLA	103.9175433
AGEMAX	Effage maxed at 120 yrs	-5722.342271
ATTGAR	Attached garage area	35
CONSTANT		219369.6667
DECK	Deck square footage	5
DETGARVAL	Detached garage value	1
FBLATOT	Total finished basement area	20
LANDVALC	Land value for cost value	1.39751684
NG3ADJSF	If Nbhd group = 3, then ADJSFLA, else 0	28.29445909
PORCH	Pool RCNLD	50
REC-AR	Rec Room Area	15
TOTOBY	Total value of other obys	1.75695003
Adjsfla =	SFLA - finished basement area	

Model Results

MODEL	SALESUSED	RSQUARED
1	115	.9572

The following were used to determine comparability:

Component	Component
Age	Condition
Finished basement area	Quality grade
Land size	Building style
Neighborhood (location)	Story height
# bathrooms	Living units