

Residual Constraint Approach (RCA)

RCA Framework & Protocol

William Bert Craytor

February 1, 2025

Contents

I.	Introduction	3
II.	Background	4
III.	Valuation Engineer	7
A.	Assumptions	8
IV.	RCA Workflow	10
A.	Types of MARS Analysis?	10
B.	Processing Workflow	10
C.	Data Workflow	11
1)	Data Requirements for MARS	11
2)	Protocols: Closed Sales Listings	12
D.	Project Folder Structure	14
E.	The Core Excel Workbook	14
F.	Backups	15
V.	Workbook Notation	15
A.	Descriptive Dimensions	15
1)	The Project Dimension ("a")	16
2)	Run Version Dimension ("r")	17
3)	Workflow Stages ("w")	17
4)	Excel Workbook Sheets ("s")	17
5)	Property Listings ("p") - MLSData rows	18
6)	MLS Variables ("v") MLSData Columns	18
7)	Final Notation Example	18
8)	Justification	19
9)	Matrices?	19
10)	First Simplification	19
11)	Second Simplification	19
12)	Variable Ontology	21
VI.	The RCA Sales Grid	23
1)	Second- and Third-Degree Terms	23
2)	Adjustment Variables	23
3)	Aggregation Variables	23

4)	Special Variable & Notation	24
5)	Breaking Down Residuals	24
A.	MARS Model	26
1)	Model Term Functions	26
B.	Core RCA Proofs	26
VII.	Data Characteristics	26
VIII.	MARS & R/earth	27
A.	R2 vs CVR2	27
1)	CVR2	27
2)	R2	27
IX.	Residuals	28
A.	Purpose of the Residual	28
B.	Data Errors	28
C.	What Is Accuracy? Perfect accuracy for a valuation requires:	29
D.	Ranking & Scoring	30
E.	CQA-Residual Curve	31
1)	The CQA Curve Characteristics	31
F.	Subject Residual	32
G.	RCA Value Conclusion	32
1)	General Considerations	32
2)	Review, Auditing, Complaints	33

Abstract

A significant limitation of conventional appraisal methods is their reliance on outdated manual techniques, such as matched pair analyses, which bypass property feature value contributions as if they didn't exist. In fact, value contributions provide the necessary basis for mathematical constraints that prevent over and undervaluation by traditional appraisers.

The Residual Constraint Approach (RCA) advances the traditional Sales Comparison Approach (SCA) by integrating a multi-phase valuation process that employs Multivariate Adaptive Regression Splines (MARS) alongside rigorous mathematical constraints on feature value contributions, in particular unmeasured features such as condition, quality, aesthetics and design whose values are typically subject to subjective judgment and bias by the traditional appraiser.

The mathematical constraints in RCA rely on the expert application of MARS regression to estimate the value contributions of measurable property features to the sale price. However, this estimate typically accounts for only about 80% of the actual sale price, excluding subjectively assessed features like condition, quality, design, and functional utility. In the San Francisco Bay Area, the residual—representing these subjective components—averages around 20% of the total property value. While MARS residuals are often treated as estimation errors in other contexts, in real estate valuation, they serve as an indirect measure of subjective feature value. Though this might initially seem impractical, proof demonstrates that the residual can be meaningfully decomposed into descriptive components, provided their contributions sum to the residual—without affecting the final valuation outcome.

This methodology emerges from two decades of empirical application by the author, developed through iterative MARS implementations in R and Python environments.

I. Introduction

This article aims to establish the mathematical and logical foundations of the Residual Constraint Approach (RCA), which offers markedly improved reliability and accuracy in estimating the market value of residential real estate. Although adaptable to various asset valuations, RCA is vital in high-value housing markets. For instance, homes commonly range between 2 and 4 million in communities like Burlingame, CA.

Accurate appraisals are crucial with the total residential real estate market in the US

valued at approximately \$45 trillion. Fannie Mae reports that typical appraisals are approximately 5% of the actual sale price. However, this should not be misconstrued as a measure of precision since appraisers often face pressure to align appraisals with sale prices. Moreover, as 60%-70% of appraisals serve refinancing purposes—where comparative sale prices are absent—the lack of actual sale price comparisons underscores the need for more precise valuation methods like RCA.

The Federal Reserve conducted an analysis in 2012 suggesting that in normal market conditions, about 15-20% of appraisals done prior to the 2008 financial crisis devi-

ated by more than $\pm 10\%$ from the eventual sale price. It is my opinion that the accuracy of standard home appraisals for refinancing purposes is still likely around $\pm 10\%$; lesser in areas of conforming houses such as newer subdivisions, and higher in areas of older homes that have gone through several updates. One of the principal sources of deviation is the adjustments for unmeasurable or difficult-to-measure variables such as condition, quality, view, functional utility, as well as a number of other variables.

Who cares about accuracy? On the one hand, if the LTV (Loan To Value) is low enough, the lenders and the GSEs are often willing to allow a waiver for an appraisal. For example, suppose a buyer with a strong credit history and income is making a down payment of \$500K on a 1M home. In that case, they assume that the risk of not recovering any outstanding debt in case of an eventual foreclosure is relatively low. However, in such cases, if the buyer foregoes any appraisal, he may later find out that he paid far more for the property than what it was worth. In fact, what happens is that in the frenzy of buying a home, significant defects in the property go unnoticed until it is too late to do anything about them. Of course, the buyer can be blamed for not being astute or knowledgeable, all the more reason that such issues lead to self-inflicted mental torture, perhaps living in that house for decades.

When it comes to the purchase of expensive homes, mistakes can be made, and the resulting consequences are often not pleasant. Why do people risk tens, hundreds of thousands, or possibly millions of dollars to save \$600 - \$1000 on an appraisal? The simple answer is that there is little confidence that spending more for an appraisal will make a difference. More thoughtful valuations employing advanced analytical methods take more time and resources than traditional residential home appraisals and are consequently

more expensive than traditional appraisals. However, as time progresses, experience is gained, and methods are improved, the time requirements should be reduced.

II. Background

The author's educational background is in mathematics, computer science, and statistics. By profession, he is a software engineer with secondary experience in appraisal and accounting. He has been doing appraisals, off and on, since 2002 and has applied MARS since about 2004 for determining property adjustments and price trends in residential real estate.

Appraising with MARS is "deep experience" in appraisal. MARS generates models that reflect in greater detail than any other statistical method both the variables that contribute to market value and how changes in those values impact sale price. Thus, in appraising with MARS, the appraiser is constantly investigating causal relationships between variable values and sale prices as generated by MARS models.

In the interest of transparency, the RCA methodology is best applied to residential areas with these characteristics:

1. Complex neighborhoods with properties that are quite different from one another
 - Older homes
 - Varied qualities of updating
 - Many different architectures and designs
 - Variation in property attributes like GLA, lot size, elevation, and view.
2. Complex market conditions
 - Quickly changing market conditions
 - Many different types of buyers

3. Good data sources: The RCA protocol is based on Data Mining and needs plentiful quality data. While it can be used for areas with sparse data, it might not be worth the time and cost required.
4. Accurate estimates of market value are desired. RCA should, in most cases, be able to provide $\pm 1\text{-}2\%$ accuracy. The valuation engineer report should comment on the conditions for accuracy in the final value conclusion. If data or information is lacking, or if consistent patterns in the market have not been uncovered through data mining, then the accuracy will suffer, and the client should be informed of the situation. Specifically, A high R^2 and CVR^2 , a meaningful MARS model, and a strong pattern in property CQA (condition-quality-appeal) features in the ranked residuals are the principal means for assessing accuracy and reliability in the value conclusion.

The development of these RCA protocols over the past 20+ years is based on the availability of MLS Listings Inc. (Sunnyvale, CA) and other MLS, which are currently accessible through MLS Listings, which comprises over 95% of the author's data analysis experience. The author's expertise stems from extensive appraisal work across 15 primary Northern California counties, with additional experience in 10 secondary counties.

The RCA methodology is most effective in markets that meet two key criteria:

1. Rich availability of high-quality appraisal data
2. Complex, heterogeneous residential areas characterized by:
 - Diverse architectural styles
 - Varying degrees of property updates
 - Multi-decade development patterns

California County Experience

The author's understanding is molded to a large extent by his experience as an appraiser in the following North California counties:

San Mateo	Santa Clara	Marin
Napa	Monterey	Santa Rosa
Alameda	San Francisco	Contra Costa
Sacramento	El Dorado	San Joaquin
Stanislaus	Placer	Solano
Mendocino	Lake	Colusa
Merced	Nevada	San Benito
Santa Cruz	Sutter	Yuba

The above areas essentially take in the Silicon Valley and its surrounding counties, extending outward into more rural area, including the Central Valley and Sierra foothills. It should be mentioned that the author has always had access to plentiful and *relatively* accurate data compared to what many appraisers in other less developed areas in the US have at their disposal.

It should be noted that the MLS data the author uses is RESO Certified, complying with the RCF format, which is derived from the old RETS format. Despite RCF certification, a data of usually secondary importance is often missing or incorrect for some properties, either through original tax assessor or subsequent sales agent MLS input errors. Often these errors follow a pattern, such as simply not making an input if the feature doesn't exist. For example, instead of entering 0 for "carport spaces", when there is no carport, nothing is entered. In areas where carports are rare, the analyst would be advised to substitute 0 for no data, in order to get a useful adjustment for carports, when occasionally there is a carport. MARS can handle missing data, it is up to the Valuation Engineer (VE) to decide on how to best handle it.

MLSListings data is also available from the MLS's of some other states, from Florida to Hawaii. This can be useful for testing how well protocols and code work in different environments.

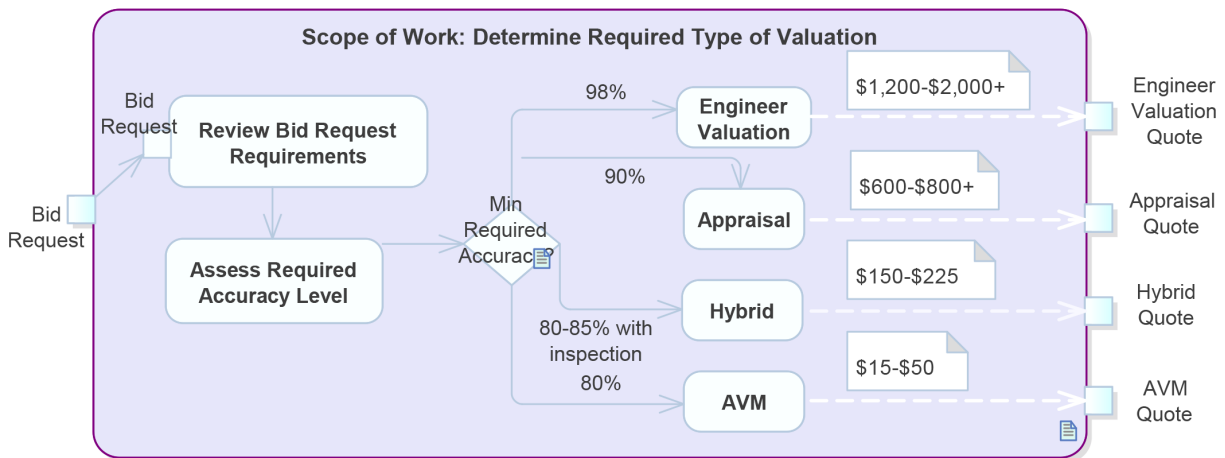
III. Valuation Engineer

Definition 1: Valuation Engineer

To prevent ambiguity, The term “Valuation Engineer” (or “VE” for short) is used to denote an individual possessing the expertise of a licensed appraiser, coupled with advanced knowledge, skills, and experience in utilizing data mining, statistics, programming, detailed protocols, and artificial intelligence to ascertain the market value of complex properties. It is also crucial to recognize that valuation is a task that necessitates some form of licensing due to the established potential for fraud.

Lenders typically offer several valuation options:

1. Automated Valuation Models (AVMs) - lowest cost option
 2. Hybrid appraisals - utilizing unlicensed inspectors with licensed appraiser oversight
 3. Standard appraisals - performed by licensed appraisers
- Absent from typical consideration for appraisal work are Engineer-level Appraisals. There simply aren't many appraisers around with the skill set to do these. There should be a need for them. They can provide significantly higher accuracy and nearly guarantee fairly good objectivity with the employment of RCA methodology (MARS and constraints).
- For conforming properties in newer subdivisions, standard appraisals generally suffice. However, engineer valuations are recommended for older homes in areas with significant renovation history spanning 30+ years. The choice between these methods should follow a structured decision process based on property characteristics and market conditions.



A. Assumptions

Assumption 1: Geographical Area & Property Type

The scope of this paper primarily focuses on single-family homes and condominiums in California’s metropolitan and urban areas. While the Sales Comparison Approach applies to other property types and geographical regions, income and cost approaches typically play a more dominant role when valuing agricultural land, multi-family properties, and commercial real estate. In such cases, the RCA approach may need modification or have limited applicability.

Assumption 2: Data Storage

CSV and Excel are heavily used by MLS systems and assumed to be the principal way to feed data to RCA programs. Although broker feeds can be used to feed data continually into SQL databases. Backup and intermediate storage often uses SQL databases and two that play big roles are the free and open source SQLite and PostgreSQL. Text or RTF files are also used for report snippets and PNG, JPG, TIFF files are used for graphics. Note that Microsoft’s SQL Server is also used by many, but the free edition “SQL Server Express” is very limited (e.g. only 1 CPU Core can be used) and other versions have hefty license fees for commercial use. Real Estate databases in California can be very large and joins can be very time consuming, so a 8+ core computer with 128G of RAM can be useful for large joins. SQL databases are also important when you need to find all of the properties in an arbitrarily bounded area. So, if for example, you used Cluster Analysis to define neighborhoods, you would be inclined to use some SQL database. However, different databases have different built-in functions for handling geographic information and thus are not compatible with one another. For this set of articles on RCA, PostgreSQL will be assumed for complex SQL tasks and SQLite for simple tasks.

Assumption 3: “R/earth” Is Used For MARS processing.

Until 2017, Salford Systems’ MARS software was an affordable solution at \$270 per year. However, after Minitab’s acquisition of Salford Systems, MARS was bundled into their SPM package at \$16,000 annually—a price point prohibitive for most small businesses. Fortunately, the R programming language’s “earth” package, maintained by Stephen Milborrow, has evolved to match or exceed the capabilities of Minitab’s SPM MARS. The earth package offers additional advantages: it’s freely available, integrates seamlessly with R’s comprehensive statistical ecosystem, and can be accessed through Python using Pandas, providing excellent flexibility for different workflows. Discussions related to MARS assume the specific use of R/earth.

Assumption 4: Notation

Standard Latin character notation is used to represent program external CSV or Excel spreadsheets, or the corresponding program internal R and Python data frames. But as the discussion turns towards variables which have numerous types and subsets, mathematical notation with Greek letters is employed to maintain clarity and *avoid* muddled thinking. However, for some, this may be a bit painstaking and tedious. In any case, assume if you see Greek characters, you are in the land of variables or lists or sets of variables. With only Latin characters, larger data structures are being referred to, such as data frames or sets of data frames or real estate property data (rows) of such data frames.

Assumption 5: Programming Languages

In this field R and Python w/Pandas are the dominant languages, with C or C++ being used when speed is important. R/earth is itself written in R and C.

Assumption 6: Indexing

In mathematics, indexing typically starts at 1, In C, C# and C++ programming languages and Python, indexing starts at 0. In R, indexing starts at 1. Very confusing since all of these have to be dealt with in practice - R and Python often call C code which is used in the RCA programs the author has written. To keep things simple, all list, matrix and data frame indexing is assumed in this paper to start at 0. And further, it is assumed that index 0, if it is used in some property list or array, represents the subject property. For matrices and data frames, row 0 is for the subject property and column 0 is a set of unique identifiers for the properties. A good practice is to first put the subject property in the first row and then rank all other properties under it from newest sale date to oldest and then sequentially assign an integer number to each of the properties starting with 0 for the subject property.

Also note that in implementing this papers math notation for R code, you will need to shift from 0 based to 1 based indexing by adding 1 to all indexes.

IV. RCA Workflow

Understanding the RCA workflow is essential for effective analysis. The process follows a structured sequence of analytical tasks and decision points

A. Types of MARS Analysis?

The workflow for MARS analysis depends on whether only a market analysis is needed or whether the valuation or appraisal of a given property is required. Note that a specific property appraisal typically requires an analysis of its market area.

- **Market Analysis:** focuses on broad areas such as counties, cities, or defined regions. This approach does not involve a specific subject property and thus requires no value conclusion. The workflow centers on model development, validation, and the generation of supporting analytics. The process concludes once the optimal model is identified and documented.
- **Single Property Analysis** encompasses all elements of market analysis plus the determination of a specific property's value. This differs fundamentally from traditional appraisal methods. In traditional approaches, value is typically derived by adjusting 3-9 comparable sales and computing a weighted average. In contrast, RCA determines value primarily through two components:

1. The MARS model's predicted value
2. The estimated property-specific residual (with the associated CQA score)

The Sales Grid in RCA serves primarily to explain value differences between the subject and comparables by illustrating variable contributions. A key advantage of

RCA is its stability: while clients may debate comparable selection, the value conclusion remains consistent regardless of which comparables are chosen for presentation, as all sales (whether 15 or 600) are adjusted using the same model. While older sales are automatically adjusted to current market conditions, recent sales are generally preferable. The model's effectiveness in capturing price factors diminishes with temporal distance, as unmeasured market characteristics become more significant over longer periods.

B. Processing Workflow

The overall processing workflow can be described by these steps:

1. Scope of work.
2. Setup initial project folders and files.
3. Download MLS data. Refer to Protocol 1-3 on the number of property records needed and setting limitations on date ranges.
4. Complete configurations sheets: Regression variables, Interactions, Aggregations, Calculations.
5. Run MARS (R/earth) regression to create price model.
6. Generate property price estimates & residuals.
7. Rank properties by residual/SF.
8. Create CQA scores for each comparable sale.
9. Assign CQA score to subject property.
10. Assign residual to property as residual/SF \cdot GLA.

11. **Add the MARS price estimate and assigned residual to Arrive at subject value conclusion.**
12. Break down the subject residual into descriptive components and value c.
13. Calculate comparable value contributions & adjustments, plus all subtotals and adjusted sale prices.
14. The adjusted sale prices will be equal to the subject final value conclusion.
15. Select the 6-12 most similar properties for the Sales Grid.
16. Break down the residuals for the comparables into component value contributions, setting the residual total to 0 as after the break down is completed.
17. Calculate the comparable residual component adjustments, setting the comparable residual total adjustment to 0.
18. Recalculate all subtotals and sales grid comparable adjusted sale price. Check values.
19. Upload Sales Grid to Report.

Note that with RCA, the final value conclusion does NOT directly derive from the comparable values in the Sales Grid. With the RCA method, the Sales Grid becomes a tool only for explaining why properties that appear similar to the subject, sell for more or less than the final value conclusion or the subject properties appraised value. This must be viewed a different way: The subject value conclusion depends on the MARS model price

estimate, and the residual estimated by the valuation engineer. But these two quantities are very dependent on all of the properties that went into the R/earth regression, which created the ranking of property residuals that determined the CQA score of the subject property thus determined its residual through placement in the ranking by the valuation engineer. The adjustments in the Sales Grid are really determined by the imposed mathematical constraints of RCA, although the residual breakdown is somewhat at the mercy of the VE, the break down value contribution totals must always add up to the associated residual. We know from Proof 2 that the while the actual values if the residual components may vary depending on the judgment of the VE, their totals for a given property are completely constrained and therefore will not change the adjusted sale price for the associated property, regardless of how the VE alters the residual breakdown.

C. Data Workflow

Note that in this paper, only a high-level workflow is presented for the RCA method. Other articles will delve into the various stages and substages of RCA analysis.

1) Data Requirements for MARS

We will need to specify a beginning date for sales transactions and an ending date based on the effective date of appraisal or valuation. This usually takes into consideration the estimated number of regression variables and interactions that will be needed for each degree of interaction over 1.

2) Protocols: Closed Sales Listings

Protocol 1: Determining the Number of MLS Closed Sales Listings Required

This protocol describes how to calculate the recommended minimum number of closed sales listings to input into MARS regression. You will need to make a rough estimate, to begin with, run MARS in first, second, and possibly third-degree runs to generate the initial model, and then review the number of rows according to the following rules to determine if more rows are needed. This protocol assumes the independent variables are 100% independent, which in real estate is rarely the case; for example, GLA and room counts are usually highly correlated, and GLA is also likely to show correlation with lot size. This gets into the calculation of "Degrees of Freedom (DOF)." Each truly independent variable gets one degree of freedom. But, when there are dependencies between the so-called independent variables, an accurate calculation of the DOF requires some analysis that is beyond the scope of this paper. Just keep in mind that due to likely co-linearity between variables, the estimates below are probably on the high side.

1. For first-order terms (no interactions), the ~ 15 rows per variable rule works well
2. For each allowed two-way interaction, you should add approximately 15-30 properties
3. For higher-order interactions (3+ variables), each should have 30+ properties.

However, this isn't a hard rule. The actual number needed can depend on:

- The signal-to-noise ratio in your data
- The complexity of relationships you're trying to model
- The variability in your response variable
- The quality of your data

Example:

So if you have:

- 10 main variables
- 5 allowed two-way interactions
- 2 allowed three-way interactions

You might calculate your minimum number of sales as:

$$\begin{aligned} N &= (10 \text{ variables} * 15 \text{ rows/variable}) + \\ &\quad (5 \text{ two-way-interactions} * 22 \text{ rows/two-way-interaction}) + \\ &\quad (2 \text{ three-way-interactions} * 30 \text{ rows/three-way-interaction}) \\ &= 320 \text{ rows} \end{aligned}$$

Alternatively, you may do the following: After running R/earth on a set of data, you can use the number of basis functions in the model, plus one as an estimate for the degrees of freedom and then take this number times 20-30 to get the number of input records needed.

Protocol 2: Sales Periods Should Encompass Minimal Change in Buyer Tastes

When collecting historical property data, there is no strict time limitation if changes in market preferences can be captured through appropriate variables. Consider these key points for handling architectural evolution in your market: Changes in dominant architectural styles (such as a shift from California Ranch to Mediterranean) can be modeled using:

- Subdivision identifiers
- MLS area codes
- Construction date ranges

However, if these style changes cannot be adequately captured through specific variables, you face an important constraint: while Date of Sale is typically used to adjust for market conditions, using it simultaneously to control for architectural preferences would confound these two distinct effects. In such cases, the preferred approach is to either:

1. Limit your data collection to the period after the introduction of the newer architectural style
2. Restrict your comparable selection to subdivisions sharing the same architectural style

This ensures your analysis maintains consistency in property characteristics while still allowing proper adjustment for market conditions over time. Alternatively, at least in this case, assuming you can extract the architectural style from the data, you could consider the interaction between these two variables. However, such extraction is not always possible.

Protocol 3: If In An Area Of Few Sales Then Do The Best You Can

When working with MARS modeling for sales data, you need approximately 15 transactions for meaningful analysis. If you have fewer sales (for example, 5 properties), you can duplicate each record twice to reach the 15-property threshold. Important considerations for this approach:

1. Use only first-degree regression to reduce over-fitting
2. Ignore or skip cross-validation since the data set is too limited
3. While MARS will likely outperform matched pairs or standard linear regression methods in this scenario, be aware that duplicating data affects the validity of uncertainty measurements and statistical significance

The best long-term solution is to expand your data-set with additional actual sales. However, if you must proceed with limited data, data duplication provides a workable temporary solution while acknowledging its statistical limitations.

D. Project Folder Structure

It would be a good idea to discuss project folder structure. It is a good idea to have a separate folder for Market Analyses and Single Property Appraisals. Both of these folders have a similar subfolder structure:

```

RCA_Projects/
├── MarketAnalysis/
│   ├── Burlingame_20240501/
│   ├── Pacifica_20240310/
│   ├── RedwoodCity_20240615/
│   └── ...
├── SingleFamily/
│   ├── Pacific_Rosewood_123_20240510/
│   │   ├── Code/
│   │   │   ├── R/
│   │   │   │   ├── ...
│   │   │   └── Python/
│   │   │       └── ...
│   │   ├── Data/
│   │   │   ├── MLS_Original.xlsx
│   │   │   ├── MLS/
│   │   │   │   ├── VER202402510153248/
│   │   │   │   │   ├── MLS.xlsx
│   │   │   │   │   ├── MLS_Stage1.xlsx
│   │   │   │   │   ├── MLS_Stage2.xlsx
│   │   │   │   │   ├── MLS_Stage3.xlsx
│   │   │   │   │   ├── MLS_Stage4.xlsx
│   │   │   │   │   └── Documents/
│   │   │   │   │       ├── LinearModels.pdf
│   │   │   │   │       ├── Cum_Dist.pdf
│   │   │   │   │       ├── Model.txt
│   │   │   │   │       ├── PartContrib.pdf
│   │   │   │   │       ├── Pairs.pdf
│   │   │   │   │       ├── Residual1.png
│   │   │   │   │       ├── Residual2.png
│   │   │   │   │       ├── ResidualSF.png
│   │   │   │   │       ├── Rpart.pdf
│   │   │   │   │       ├── VariablesImp.pdf
│   │   │   │   │       ├── VariableVsSP.pdf
│   │   │   │   │       └── ...
│   │   └── ...
│   └── ...
└── ...

```

Stage 1 processing will create the version

folder with its sub-folders. It will copy the `MLS_Original.xls`, even though this spreadsheet *should* never change, but reflect exactly what was downloaded from the MLS as of a certain date and time; yet there is the possibility on a lengthier project that a new MLS download will be needed. Next, it will make a copy of the expanded `MLS.xlsx` workbook with its configuration parameters, into the version folder. There can be many version folders each created for a complete run of MARS. Then the following stages will generate column additions, sorting and other changes to the workbook and write an updated version as `MLS.Stage(N)` to the project data folder when the appraisal is completed. To reiterate, the root `MLS.xlsx` folder is undergoing constant change, and so the version copy is an ongoing mutation of the original MLS workbook, with variable deletions and additions, plus other changes. Understand the processing R or Python code most likely will run with input from the `Data MLS.xlsx` workbook. Thus we always have a copy of all the data behind each version of the RCA run. We make as many versions as needed until we decide that we are not able to make any more improvements - and then go back and pick the best version. Each version folder contains a number of text, diagram and graph files that can be embedded into a report. The final `MLS.Stage4.xlsx` workbook contains as many sheets as necessary for the report. In particular it may have a spreadsheet for transfer to a Fannie Mae Sales Grid. It will also necessarily contain the computations or breakdown behind aggregations. This is potentially the most time-consuming part of the RCA approach - the struggle to find the perfect model in a difficult market area.

E. The Core Excel Workbook

Since most MLS services download property data to Excel CVS worksheets, Excel is

probably the most convenient way to store data and various configuration parameters. There is one core Excel Workbook for each project called `MLS.xlsx`, which has one sheet called "MLSData", to store the closed property sales listings that were downloaded from the MLS, as well as several configuration and parameter sheets:

```
Workbook:  MLS.xlsx/
├─ Sheet:  Project Data
├─ Sheet:  MLSData
├─ Sheet:  Regression Variables
├─ Sheet:  Calculations
├─ Sheet:  Allowed Interactions
├─ Sheet:  Aggregations
└─ Sheet:  (MARS run parameters)
```

The original data is stored in a workbook named something like `MLS.Original.xlsx`. A copy should be made with a different name, let's say "`MLS.xlsx`," which will load into the R or Python RCA program. Both of these workbooks are stored under the Data folder. Now, the `Data/MLS.xlsx` workbook will probably go through many iterations of runs, trying different combinations of variables, MARS parameters and changes to other settings. Each iteration creates a new input/output folder named with the date and time, with the format of `VER_YYYYMMDDHHMMSS`. Each folder will have a copy of the input `MLS.xlsx` workbook and contain the generated Excel Workbooks for each of the four stages of processing, as well as the output text documents, diagrams and graphs. Note that Stage 4, the final stage, may contain a number of other sheets for the report, depending on the type of client and SOW.

F. Backups

When writing new or modified Excel spreadsheets to the project version folders,

backups can also be made to SQLite or PostgreSQL databases in the project folder. The database files are less likely to become unintentionally corrupted over in subsequent periods, through review or reference use.

V. Workbook Notation

A. Descriptive Dimensions

There are roughly six descriptive dimensions for a workbook "M":

1. Appraisals "a"
2. Run Versions "r"
3. Workflow Stages "w"
4. Sheets "s"
5. Property Listing "p"
6. Variable Set "v"

So, an individual cell in a particular workbook M can be described as:

$$\begin{matrix} r_* \\ w_* \end{matrix} \begin{matrix} a_* \\ W \\ s_* \end{matrix} \begin{matrix} p_* \\ v_* \end{matrix} \quad (1)$$

(Note: "*" means some valid index)

or:

$$\begin{matrix} r_k \\ w_m \end{matrix} \begin{matrix} a_g \\ W \\ s_h \end{matrix} \begin{matrix} p_i \\ v_j \end{matrix} \quad (2)$$

$$\begin{aligned} i &= 1, \dots, n_p \\ j &= 1, \dots, n_v \\ k &= 1, \dots, n_r \\ m &= 1, \dots, n_m \\ g &= 1, \dots, n_g \\ h &= 1, \dots, n_h \end{aligned}$$

where:

- n_p = # of rows (depends on sheet)
- n_v = # of columns (depends on sheet)
- n_r = # of run versions
- n_m = # of workflow stages
- n_g = # of analysis projects
- n_h = # of workbook sheets

Understand that "W" references the entire Excel Workbook of five core data sheets

that each have their own table, which on running the associated R or Python program are loaded into separate internal data frames. Both R and Python have internal functions that can take a name and find which row or column of the data frame belongs to. For example, if in the subscript v_j , $v = (\text{"GLA"}, \text{"LotSize"}, \text{"BathRms"}, \text{"BedRms"}, \dots)$ and $j=0$, then $v_j = \text{"GLA"}$. Either R or Python will internally be able to resolve "GLA" to the column it is stored in, which for example, might be column 15 of the associated data frame. So, in this paper we index into our name arrays, which pass a name to the running program which re-indexes that name into the actual column or row index for the internal data frame. It is a bit complicated. But to be precise, this is necessary and most convenient. It is a "separation of responsibilities" between the analyst role and the developer role; although in reality the Valuation Engineer (VE) could be performing both roles. Note that an R "data frame" is called "dataframe" in Python.

Also note that the MARS run time parameters sheet is not a dimension of W, as W relates to the version of data being input into MARS, separate from the parameters that determine how MARS/earth handles that data.

For the sake of clarity, a complex multidimensional notation for the Workbook is presented which relates more directly to the internal R or Python data frames that the data workbook sheets are loaded into. It is "heavy" notation. But when we get into proofs and so on, we throw a lot of the baggage away by making assumptions about what we are working with. Nonetheless, this notation helps get a better overview of the workflow and data from the outset.

Let's start with notation for the set of workbooks we will be processing.

1) The Project Dimension ("a")

Once a Scope of Work (SOW) has been approved, a project folder needs to be setup to store input and output data, including configuration parameters, output graphs, diagrams, spreadsheets and report snippets. The VE should identify all projects with a simple sequential ID, such as an integer, so that if a report should go missing, it would be obvious. If projects are canceled, then there should be a consistent means to indicate the project has been cancelled within the given project folder, but it should never be deleted. We assume project IDs start at 0 and incremented by 1, although there are other possibilities. We further assume all valid project IDs are stored in an ordered set "a" created with the processing program.

The appraisal project data is in the project worksheet of the workbook, but could also be stored in some central SQL database,

Here are the following suggested project fields:

1. Local ID: A short integer ID that is sequentially generated for the company. This would make a great primary key for storage - and could be automatically generated by the database with each successive project.
2. Global GUID: If you want, you can create a GUID (Globally Universal ID) using

<https://www.uuidgenerator.net/>

3. Property Location (Market Analysis) or Address
4. Effective Date of Appraisal
5. Date of Order
6. Date of Inspection
7. Date of Original Report
8. Date of First Updated Report
9. Date of Second Updated Report
10. Date of Third Updated Report
11. Client Name
12. Client Address

13. Lender Name
14. Lender Address
15. Owner Name
16. Owner Address
17.

Example of list of project IDs:

```
a = ("1",
      "2",
      "...",
      "141",
      "142")
```

2) Run Version Dimension ("r")

Second in the workflow, we execute MARS regression on the input. This will likely involve several to many executions as we alter various parameters to improve the model. Each execution will create a new version folder for the output, with a name based on the date and time of execution. For example,

```
r = ("ver20241210_083124",
      "ver20241209_151130",
      "ver20241208_160944")
```

latest_version : $v_1 = \text{"ver20241210_083124"}$ stage2 : $w_2 = \text{"stage2"}$

Promising versions with competitive models and output, should be kept until a final version is decided upon, or perhaps longer if they have some utility. You may want to modify the version folder to indicate it is the final version, by placing "Final" at the end of the name.

3) Workflow Stages ("w")

The RCA Workflow is separated into usually five separate stages, as listed below. Each stage must successfully complete before starting the next stage. The VE should be

able to stop the processing at the end of a stage to review the results before proceeding. This is needed because market areas are fairly complex and often quite unique. Unusual problems can be encountered that take a new approach. The VE must be able to jump into R or Python code to make necessary changes and even try new approaches. And, in such cases, he may want to experiment and try different approaches requiring multiple runs of the same stage he is working on. Although, the VE should also be free to run all stages with one command, - with some caveats (e.g. 3rd Degree Aggregation must have already been determined through a previous run Stage 1). This means that when each stage completes, it should store all of its internal data in files or databases, so that the next stage can be started as if the previous stage were still in memory, by reloading the data from the previous stage into memory.

Here are the stages:

```
w = ("setup",    -Setup
      "stage1",   -MARS processing
      "stage2",   -Final Sales Grid
      "stage3",   -Report snippets
      "stage4")   -Review/Cleanup
```

4) Excel Workbook Sheets ("s")

In the RCA Core Excel Workbook, we have the following sheets that are each mapped into a internal data frame:

```
s = ("MLSData",
      "Regression Variables",
      "Calculations",
      "Interactions",
      "Aggregations 1st & 3rd Degree"),
      "Aggregations 2nd Degree")
```

The reason that 2nd degree aggregations are put on a separate sheet is that they can

be most easily defined with a symmetric table that shows all possible pairs of interactions, allowing the specification of an aggregation variable for each pair's associated value contributions and adjustments to be aggregated (added) to. We couldn't do this with 3rd degree aggregations because that would require a 3 dimensional symmetric table which is not possible in Excel. Further more 20x20x20 cells would 8,000 cells to deal with. It is easier to just list any third degree term names in the table with the 1st degree aggregations, for example:

Variable	Aggregate To
AboveGradeFinArea	GLA
BelowGradeFinArea	BGLA
UnfinArea	UBA
LotSize	LotSize
...	...
GLA__LotSize	GLA
GLA__DateOfSale	GLA
BathRms__DateOfSale	GLA
Lat__Long	Location
Lat__Long__GLA	Location
...	...

5) Property Listings ("p") - MLS-Data rows

Property listings form the rows of the MLSData sheet. They are not found in the other sheets, until Stage 4, where the final report snippets are created with the final selection of comparable properties. Here is a sample list of property IDs:

```
p = ("MLS823211",
      "MLS823456",
      "MLS823542",
      "MLS823721",
      "MLS823641",
      "...")
```

6) MLS Variables ("v") MLSData Columns

More will be said about handling variables later. Suffice it to state at this point that MLSData sheet typically has 30-40 variables of various types, but usually only 10-20 will be selected as independent variables for the MARS regression. These variable also appear in other sheets, although typically as a list of all possible variables that can be used for submission to MARS. An example of a list of variables is:

```
v = ("GLA",
      "SalePrice",
      "BedRms",
      "BathRms",
      "GarageSF",
      "...")
```

7) Final Notation Example

Lets pull the above notation into a final example. Using the above lists, let

$${}_{w_1}^{r_1} W_{s_0}^{a_{141}} {}_{v_0}^{p_2} \quad (3)$$

Note that we are using 0 indexing here, so that the first elements of lists is assumed to have an index of 0. Python also uses 0 indexing, by R uses 1 indexing.

So, the above means that we are talking about the cell value for:

- Variable v_0 or column "GLA"
- For property p_2 or ID "MLS823542"
- Of sheet s_0 or "MLSData"
- For workflow stage w_1 or "setup"
- Of version ersion r_1 or "ver20241209_151130"
- Of project a_{140} or "140"

The above notation can also then be written as:

$$\begin{array}{c} \text{"ver20241209_151130"} \\ \text{"setup"} \end{array} \begin{array}{c} \text{"140"} \\ W \\ \text{"MLSData"} \end{array} \begin{array}{c} \text{"MLS823542"} \\ \text{"GLA"} \end{array} \quad (4)$$

8) Justification

"Why don't we just use indexing or numbers in the notation?" We have two requirements for our notation:

1. Subsets, e.g., the analyst frequently selecting subsets of a list, such as subsets of variables, are used. For instance, there may be 35-40 variables in the MLSData sheet but only 10-20 of them. And, from one run to the next, The VE will very likely change the variables submitted to MARS/earth.
2. Iteration, e.g., $i = 1, 3 \dots, n_1$
Our algorithms, protocols, and proofs usually need to be able to iterate through all items in an ordered list, from the first to the last - or pick out some particular item.

So, we need a letter to represent the subset or list of items, and that must have another subscript that we can use to iterate over the items of the list to give is the strings that are used to index into the data frame rows or columns.

9) Matrices?

And what about the matrices of Linear Algebra? Matrices work only with numbers, either integer or real. Matrices as a data structure are supported in both R and Python. However, many of the variables we work with in MLS data are called "strings" or character data. Even if some variables are only numbers, they are often not real numbers but just names. For example, MLS Area Codes for a city may be

("660", "661", "662", "663", ..., "672"). Earth regression allows variables to be specified as "factor" variables, in which case the numbers are treated simply as names, without any sequence. In such cases, the model produced will merely provide a value contribution for each number, rather than a function that runs over the sequence. Other examples include "C1", ..., "C6" for Condition and "Q1", ..., "Q6" for quality.

10) First Simplification

The above notation for the core Excel project workbook is nice to tie things together, but too heavy to carry around for explaining algorithms or mathematical proofs. When we are doing the tedious work, we can assume it is within the given project, version, stage and sheet. We just say what we are working with, if it is not already implicit. All we really need from the above, once we indicate what we are working with is:

$$M_{s_h}^{p_i}_{v_j} \quad \begin{array}{l} i = 1, 3 \dots, n_p \\ j = 1, 3 \dots, n_v \\ h = 1, 3 \dots, n_s \end{array} \quad (5)$$

where:

$$\begin{array}{ll} n_p & = \# \text{ of rows (depends on sheet)} \\ n_v & = \# \text{ of columns (depends on sheet)} \\ n_s & = \# \text{ of sheets} \end{array}$$

11) Second Simplification

It is more convenient to rename the sheets to specific kinds of data frames they support rather than use indexes. For each sheet, the rows and columns have different functionalities. So, the W set of data frames is broken up into M, V, A, I, and C data frames as follows:

◇ **MLS Data**

$$M_{v_j}^{p_i} = W_{v_j}^{p_i} \quad \begin{array}{l} i = 0, 1, \dots, n_p \\ j = 0, 1, \dots, n_v \end{array} \quad (6)$$

$$\begin{array}{ll} \text{where: } n_p & = \# \text{ properties} \\ n_v & = \# \text{ input variables} \end{array}$$

Subject Is By Default Always First in Property Lists

In the "MLSData" Excel spreadsheet and associated data frame, the first row is always reserved for the subject property. It is never sorted with the other MLS properties below it. It is never entered into the MARS regression as independent data, it is entered rather as target data, or as the dependent variable. The same holds for property lists. In all property lists, the first element with index 0, is always assumed to be the subject property ID, unless stated otherwise.

Even if the task is to create a model for a market area without any subject property, the first row is nonetheless loaded with some dummy property data. This makes it far more convenient to write programs, specify algorithms, or do proofs. So, if we are using c to represent the list of properties, c_0 is the subject property and c_i , $i = 1, \dots, n_1$ are the comparable listings that will be input to MARS/earth regression.

◇ Regression Variable Specifications

$$V_{s_j}^{v_i} = W_1^{v_i} s_j \quad \begin{matrix} i = 1, 3 \dots, n_v \\ j = 1, 3 \dots, n_s \end{matrix} \quad (7)$$

where: $n_v = \#$ available variables
 $n_s = \#$ of specifications

Note, $n_s \geq n_v$, as we usually select a subset of available variables for input to regression.

◇ Calculations

$$C_{c_j}^{v_i} = W_2^{v_i} c_j \quad \begin{matrix} i = 1, 3 \dots, n_v \\ j = 1, 3 \dots, n_c \end{matrix} \quad (8)$$

where: $n_v = \#$ of variables
 $n_c = \#$ of calc specs

◇ Interactions

$$I_{t_j}^{f_i} = W_3^{f_i} t_j \quad \begin{matrix} i = 1, 3 \dots, n_f \\ j = 1, 3 \dots, n_t \end{matrix} \quad (9)$$

where: $n_f = \#$ from variables
 $n_t = \#$ to variables

Note that Data Frame I is a symmetric data frame, with variables alphabetically listed for the "from" rows and "to" columns. The cells are marked with a "1" to indicate an interaction between the associated pair of variables is allowed. Otherwise, "0" indicates that no interaction is permitted. For a second-degree interaction, this is relatively simple, with only one interaction possible. With a third-degree regression, we have three pair interactions between all three variables, where every pair of interactions must be allowed to generate the three variable interaction term (or variable).

◇ Aggregations

$$A_{v_j}^{v_i} = W_4^{v_i} v_j \quad \begin{matrix} i = 1, 3 \dots, n_v \\ j = 1, 3 \dots, n_v \end{matrix} \quad (10)$$

where: $n_v = \#$ of variables

12) Variable Ontology

The above notation will be refined even further, especially with respect to the v ordered subset of property variables. The handling of variables is the most complex subject in RCA. We can have different types of variables that fall into two basic ontologies: (1) Variable Origin and (2) Variable Application, i.e. Non-residual vs Residual

I. Variable Origin Ontology

1. First Degree Variables
 - 1.1 MLS Regression Variables
 - 1.2 External Regression Variables
 - 1.3 Synthetic Regression Variables
2. Factor Variables:

To handle factor variables, a function is needed that can identify them. Since they are specified in the Excel project configuration workbook, a function is created that returns true if a variable passed to it has been classified as a factor type.
3. Interaction Variables: These are created by MARS in generating the model, if the degree is set to 2 or higher. However, 3 interacting variables is the most that could be recommended due to the difficulty of interpretation, the probability of overfitting and the impact on Degrees of Freedom and the resultant increase in property records to ensure a robust model.
 - 3.1 Second Degree Variables: These are common and a good example would

be GLA:BathRms, GLA:LotSize, DateOfSale:BathRms, Latitude
Longitude.

- 3.2 Third Degree Variables: Less common interactions that suggest caution and possible over-fitting.

4. Residual Breakdown Variables

II. Variable Application Ontology

1. MARS Generated Value Contributions
 - 1.1 First Degree Variables
 - 1.1.1. MLS Regression Variables
 - 1.1.2. External Regression Variables
 - 1.1.3. Synthetic Regression Variables
 - 1.2 Interaction Variables
 - 1.2.1. Second Degree Variables
 - 1.2.2. Third Degree Variables
 - 1.3 Residual Breakdown Variables
 - 1.4 Aggregation Variables
2. Engineer Generated Variables & Values
 - 2.1 Residual Breakdown Variables
 - 2.1.1. Condition
 - 2.1.2. Quality
 - 2.1.3. Design/Aesthetics?
 - 2.1.4. Landscape?
 - 2.1.5. Functional Utility?
 - 2.1.6. View?
 - 2.1.7. ...

These variable groups will be given their notation and described in more detail in the next section

Protocol 4: Factor Variables

Factor variables, or simply factors, can occur in all terms, including interactions. They are important because they must be specially dealt with for MARS regression and graphing.

While variables with character values are necessarily factors, numerical variables must be flagged as factor variables for MARS to recognize them as such. For example, MLS areas are often identified by number. In particular, the neighborhood areas in the San Francisco Bay Area are often identified with unique integers. If these values are passed to MARS as factors, then MARS will attempt to discover if certain areas have an average contributing value to property sales within their boundaries without regard to any relationship with other areas. In contrast, with normal variables, it will assume there is some kind of continuity in value contributions as the variable value changes. This is an important distinction. Furthermore, these factor variables can appear in model terms with non-factor variables, which makes graphing difficult. Specifically, when graphing single variable terms with factors, we lump all such factors together and simply draw a bar graph showing their contributions, positive or negative.

If the factors appear with other variables in terms, they will always appear with one possible value, because factor variables can have only one value for a given property: If an area has the value 662, by definition it can't have any other value for that area. If you find this confusing, to make matters worse, in the MARS-generated model, there is not a single variable name for the factor; MARS will take the root name of the factor variable it was given as input, then create new columns with variable names formed by appending the possible values. So, a single column in the Excel spreadsheet, such as `MlsArea`, will have multiple data frame columns created for each possible value, such as `MlsArea660`, `MlsArea661`, etc. A variable such as `MlsArea661` will have a value of 1 for properties in area 661; otherwise, 0. Then, MARS will regress on those columns as if they represented Boolean variables. These modified names will be used in the model provided by MARS.

You will likely want to convert those model names back to their original name to make the graphs more understandable to clients who like it nice and simple (at least as simple as possible). Instead of seeing `MlsArea661=1`, we will see `MlsArea=661`. We then can graph terms by simply stating in the graph header something like "For MLS Area = 662". Thus, N degree terms with M factor variables are graphed as (N-M) degree terms, and the factor dimension disappears into the title.

Note that two or more factor variables could exist in a single term, so you would likely need a graph for each possible pair of values. If variable x with three distinct values and variable y with two distinct values appear in the same term with non-factor variable z, then you would need $2 \cdot 3 = 6$ graphs. An example would be "MlsArea" + "Architectural Style", e.g. Victorian homes in San Francisco might be worth more or less depending on the area of San Francisco they are in. Fortunately, in most areas, there are usually only 1-3 architectural styles important enough as value contributors to be used in the model (although there may be 6+ different architectural styles).

VI. The RCA Sales Grid

The RCA Sales Grid is an Excel spreadsheet labeled in this paper as "MLSData." It does not look like the Fannie Mae Sales Grid but is, roughly, a transpose of the Fannie Mae Sales Grid, with rows and columns switched: Each row contains data for some property, and each column contains data for some property variable. There can easily be hundreds or thousands of property rows. The first row is reserved for variable names, and the second for a subject property, even if we are only doing a market analysis. When the spreadsheet is read into a data frame, then the first row of the sheet becomes the names attribute, and the second row for the subject property becomes the first row of the data frame.

The columns in the RCA Sales Grid include identification columns, then primary, external, and synthetic variables combined and fed as the independent variables into MARS. MARS outputs from these variables those that it thinks have a significant impact on the Sale Price, also known as the dependent or target variable. In second-degree regression, we may get pairs of these variables that act as new variables. With third-degree regression, we may also get variable triplets that act as so-called independent variables, although they are, of course, highly correlated with the underlying variables.

1) Second- and Third-Degree Terms

With second-degree or third-degree regression, some of terms in the output model, may include two or respectively three variables. For example:

- $\text{BathRms_GLA} = -52 * \max(0, 1200 - \text{GLA}) * \max(0, \text{BathRms} - 1)$

Notice that we name these by alphabetically sorting the variable names and concatenating them with "Vs" (for vs. or versus). We do the same with third degree terms on the variable name triples produced in the model. For example, "BathRms_GLA_LotSz."

These variables get added to the primary input variables in Stage 2, when the RCA program calculates the contributions of all variables for each property, including the interaction variables.

2) Adjustment Variables

The RCA program will also calculate the adjustments for all variables and properties in Stage 2, after value contributions are calculated, by subtracting the corresponding comparable property value contribution for a feature from the subject contribution for that feature.

3) Aggregation Variables

Aggregation variables are defined to replace subsets of the variables, adding all their value contributions and adjustments into a common variable to be used in the report. For example, we may have split both finished and unfinished living area into finer detail variables:

1. Above Grade Legal Finished Area
2. Below Grade Legal Finished Area
3. Above Grade Illegal Finished Area
4. Below Graded Illegal Finished Area
5. Above Grade Unfinished Area
6. Below Grade Unfinished Area

Yes there are areas where this does make sense.

The MARS regression will regress on all of the variables and generate a model that will provide contributions and adjustments

for them. But for a form report, these different contributions will probably have to be combined or aggregated into fewer variables. The details of the aggregation should be supplied in an addenda. A more frequent example, would be aggregating interaction adjustments into one of the primary variable adjustments. For example, we might aggregate the adjustment for GLA_Lotsize to GLA or to LotSize, or even split them 50:50 between both, as determined by the Aggregation sheet.

4) Special Variable & Notation

Let:

$n_p = \#$ of comparable properties

$n_r = \#$ of regression variables

$n_c = \#$ of residual breakdown variables

π_i	Actual net sale price for property $i = 1, 2 \dots, n_p$
α_i	Adjusted net sale price for property $i = 1, 2 \dots, n_p$
ρ_i	MARS estimated net sale price for property $i = 0, 1 \dots, n_p$
β	Base value of the model
$\varphi_{i,j}$	Value contribution of regression variable $j = 1, 2 \dots, n_r$ for property $i = 1, 2 \dots, n_p$
$\delta_{i,j}$	Adjustment for regression variable $j = 1, 2 \dots, n_r$ for property $i = 1, 2 \dots, n_p$
ϵ_i	Total residual for property $i = 0, 1 \dots, n_p$

$\epsilon_{i,c}$	Residual component value contribution c , $c = 0, 1, \dots, n_c$ for property $i = 0, 1, \dots, n_p$
$\xi_{i,c}$	Adjustment for residual component c , $c = 0, 1, \dots, n_c$ for property $i = 0, 1, \dots, n_p$

5) Breaking Down Residuals

Once the MARS regression has created a model, it generates the estimated sale prices from the MARS-selected input variables. It subtracts that value from the corresponding Net Sale Price to get a residual value for each property. Now, the RCA protocol stipulates that once the VE has decided on the final property comparables, usually 6-12 property listings, he needs to go through each property listing and break down the calculated residual into the most likely features he thinks impact value, ignoring the possibility of random data errors. He might, for example, have these breakdown variables for a given property:

1. Condition
2. Quality
3. Design
4. Functional Utility
5. Over/Under Market Sale

Different properties will likely have some different residual breakdown variables, which need to be collected in one ordered set.

Notation: We use the notation in the above table to arrive at:

$$\epsilon_i = \epsilon_{i,1} + \epsilon_{i,2} + \dots + \epsilon_{i,n_c} \quad (11)$$

where n_c is the number of residual breakdown components.

The VE must use his judgment to break down the residual into the given components.

One might think this would allow bias to enter the final value conclusion. However, that is not the case because we have mathematical proof that if the breakdown adjustments total the residual, they will not impact the value conclusion. The purpose of the breakdown is not to establish value, but to explain why the residual is what it is in comparison to other property sale residuals. The only spot where the engineer's bias can impact the final value conclusion is the estimate of the residual for the subject since, of course, there is no actual sale price for the subject needed to calculate a residual for it.

Each property has its set of residual value contributions, but these need to be merged into one set C over all properties:

$$\begin{aligned} \text{Letting: } A_i &= \{\{\epsilon_{i,c}\}_{c=0}^{n_c}\} \\ C &= \{x \mid x \in \bigcup_{i=0}^{n_p} A_i\} \end{aligned}$$

where $[\epsilon_{i,c}]$ is the variable name for the value $\epsilon_{i,c}$ and A_i is the set of such variables and n_c is the number of values in each ϵ_i .

So, with the subject and comparable property data in separate columns, then under all the regression model variable value contributions for each property, we should see the set C of nc corresponding residual value contributions.

A. MARS Model

1) Model Term Functions

The MARS generated price model can be viewed as a sum of multiplicative terms of 0 to 3 variables, assuming that the max degree of regression is 3. So, suppose we have a 3rd degree regression model that is the sum of a base constant, n_1 first degree terms, n_2 second degree terms and n_3 third degree terms.

We can express it as:

Let:

$n_p = \#$ of comparable properties

$n_r = \#$ of regression variables

$n_c = \#$ of residual breakdown variables

$$\begin{aligned} \hat{P} = & f_{0,0} g_0^0 + \sum_{q=1}^{n_1} f_{1,q} g_q^1(x_*) \\ & + \sum_{r=1}^{n_2} f_{2,r} g_r^2(x_*, y_*) \\ & + \sum_{s=1}^{n_3} f_{3,s} g_s^3(x_*, y_*, z_*) \end{aligned} \quad (12)$$

B. Core RCA Proofs

Proof I: Each and every comparable adjusted sale price equals the estimated subject sale price.

For any comparable i , the adjusted sale price α_i is the sum of the net sale price π_i , the base value β , the value contributions $\varphi_{i,j}, j = 1, 2, \dots, n_r$ plus the comparable residual ϵ_i or:

$$\begin{aligned} \alpha_i = & \pi_i + (\beta - \beta) + (\varphi_{0,1} - \varphi_{i,1}) \\ & + (\varphi_{0,2} - \varphi_{i,2}) + \dots + \\ & (\varphi_{0,n_r} - \varphi_{i,n_r}) \\ & + (\epsilon_0 - \epsilon_i) \\ = & \pi_i + (\beta + \varphi_{0,1} + \varphi_{0,2} + \dots \\ & + \varphi_{0,n_r} + \epsilon_0) \\ & - (\beta + \varphi_{i,1} + \varphi_{i,2} + \dots + \\ & \varphi_{i,n_r} + \epsilon_i) \\ = & \pi_i + \pi_0 - \pi_i \\ = & \pi_0 \end{aligned} \quad (13)$$

Proof II: Altering the residual breakdown under the constraint that all residual component value contributions sum to the comparable residual, has no impact on the final value conclusion of the RCA protocol.

This is almost trivial. Yet without thinking about the math, it is easy to overlook this because traditional appraisers focus on the adjustments rather than the value contributions. It is the value contributions that cancel out here. - A decades long oversight. For some comparable k , $k > 0$, the total residual adjustment is equal to sum of the n_c breakdown residual components which in turn is equal to sum of the differences of value contributions, which is equal subject total residual minus the comparable total residual. And this is independent of the number of residual components created or what they represent. As long as the value of the residual components equals the total residual for both the subject and chosen comparable ξ_k will not change

$$\begin{aligned} \xi_k = & \xi_1 + \xi_2 + \dots + \xi_{n_c} \\ = & (\epsilon_{0,1} - \epsilon_{k,1}) + (\epsilon_{0,2} - \epsilon_{k,2}) + \dots + \\ & (\epsilon_{0,n_c} - \epsilon_{k,n_c}) \\ = & (\epsilon_{0,1} + \epsilon_{0,2} + \dots + \epsilon_{0,n_c}) - \\ & (\epsilon_{k,1} + \epsilon_{k,2} + \dots + \epsilon_{k,n_c}) \\ = & \epsilon_0 - \epsilon_k \\ = & \xi_k \end{aligned} \quad (14)$$

VII. Data Characteristics

The following section discusses MARS. MARS is entirely dependent on the data it receives, and most of this data will come from some MLS (Multiple Listing Service) or tax assessors. However, it may be routed through other data providers who maintain and improve data and then sell at some profit. In

some areas, it is fairly reliable; in others, it is not so reliable. In critical instances, the property inspector has to make his observations and measurements, and their accuracy can be questioned.

The valuation engineer needs to understand what the data represents and how good of a job it does representing what it is supposed to represent. This invites a whole list of issues:

- How do you classify a variable as being a measurement?
- If a variable is not a measurement, what else can it be?
- How can regression be useful if the data it receives is not accurate?

Most of the data we get from MLS listing services and tax assessor data for regression is numerical data that implies measurements or counts. Some of it is logical, True or False.

VIII. MARS & R/earth

A deeper dive into MARS and the R/earth package will be left for planned articles to follow.

If you are not already familiar with MARS then refer to Stephen Milborrow's earth package documentation [Milborrow \[a\]](#), as well as Jerome H. Friedman's foundation paper on Multivariate Adaptive Regression Splines [Friedman, Jerome H.](#) are good starting points for those who would like to learn more about Multivariate Adaptive Regression Splines and the popular open source program "earth" maintained by Stephen Milborrow. For more advanced resources, Recommended [Hastie et al.](#), [Kuhn and Johnson](#), [James et al.](#), [Zhang](#), [Milborrow \[a\]](#), [Milborrow \[b\]](#).

A. R2 vs CVR2

For this paper, which focuses on residuals, it is essential to discuss the importance of the R2 percentage in comparison to the lower CVR2 percentage: Is it correct to use the R2 value as the likely accuracy of the associated regression model when:

1. The CVR2 is usually 10% less.
2. T2. The subject property, which may not be listed but is being appraised for a refinance loan, has not been updated or modernized for decades like the sales listings used for regression.

1) CVR2

It is based on multiple runs of MARS regression on different randomly selected test representing 90% of the properties, with the remaining 10% used for testing. Data miners have found that this is the best way to achieve a robust model that works good at predicting as yet unseen properties.

However, the CVR2's primary value is not in reflecting the factual accuracy of the regression model for market area sales but in optimizing its ability to predict sale prices for hypothetical property sales as of a given effective date.

What, then, is the expected sale price for a property that hasn't been updated in the past 15 years, assuming it was listed last month? A model with cross-validation will likely provide the best estimate, but it assumes the property has been modernized like typical properties in the dataset. This assumption may not hold for every property.

2) R2

The R2 statistic is generally the best measure of how much price variance is accounted for by a regression model, assuming

the model is applied to properties that have undergone typical updating and modernization. The remaining variance can likely be attributed to factors not considered by the regression.

If appraising an older home without the updates common in recent sales, it might make sense to estimate the cost to bring it up to the norm and deduct that from the regression estimate. However, such adjustments are typically not allowed under GSE or lender guidelines.

IX. Residuals

As indicated by the name "Residual Constraint Approach," residuals are central to this methodology. Here, residual refers to the difference between a property's actual sale price and the sale price predicted by regression analysis.

A positive residual suggests that the property sold for more than its estimated price, based on typical measurable attributes like gross living area (GLA) and lot size, implying it might have more appeal than average. A negative residual indicates the property sold for less, potentially suggesting less appeal or other uncaptured negative attributes.

Given that larger homes likely have larger residuals, we might use residual per square foot based on GLA for a more normalized comparison. A more in-depth discussion of this issue is reserved for another paper.

A. Purpose of the Residual

The residual serves as an indirect gauge of the value added by variables not included in the regression model. By ordering properties based on their residual or residual per square foot, we achieve an objective ranking from properties with the least appeal to

those with the most. This ranking method is effective, provided a skilled analyst with a deep understanding of the local market crafts the MARS (Multivariate Adaptive Regression Splines) model. The following are some additional requirements:

- **Quality of Data and Model:** A well-constructed regression model and high-quality data are prerequisites for reliable residuals. In the San Francisco Bay Area, my goal is to develop a model with an R^2 of approximately 75-80% and a cross-validation R^2 (CV R^2) of 55-70%.
- **Avoid Over-fitting, Limit the R^2 :** The valuation engineer should aim for a maximum R^2 of about 80% in a most mature markets in higher priced areas like the SF Bay Area because much of a residential property's competitive value lies in its subjective appeal, which isn't quantifiable and thus can't be directly included in regression models. An R^2 significantly above 80% might indicate an over-fitted model, prompting further scrutiny. Yet a word of caution; There are market areas with R^2 values above 90% without over-fitting and there are difficult to appraise areas, where the best you might do is 50-60% or lower.
- **High R^2 in Specific Areas:** However, there are regions, like subdivisions or other uniform residential areas, where higher R^2 values are surprisingly common due to the homogeneity of the properties.

B. Data Errors

Based on social media comments, the California Multiple Listing Service (MLS) data is generally of higher quality than many other United States regions. Typically, MLS data should be expected to be of good qual-

ity in most mature metropolitan areas but poorer in rural areas.

While there are occasional errors in the data for various reasons, these errors should generally cancel each other out and have minimal impact on the averages used in price models if sufficient data exists.

However, systematic biases can occur in specific areas. For instance, consider a neighborhood where many homes were initially constructed with full second stories, but some buyers were given the option to remove second-floor rooms to create vaulted ceilings on one or more first-floor rooms, significantly reducing the gross living area (GLA). The changes were not subsequently updated in the assessor's records for one reason or another. Consequently, MLS might show 2,400 square feet when the measured value is only 2,000 square feet. In such scenarios, the MLS data should be corrected before performing MARS (Multivariate Adaptive Regression Splines) regression. It is important to note that obtaining this information for comparables may require some investigative effort. Discrepancies should be noted for future reference.

Given the foregoing, it is reasonable to assume that in high-quality MLS areas, the errors that spill over into the residual are minimal compared to the impact of subjective and unmeasured features such as condition, quality, and design. This is a crucial consideration when estimating the accuracy of the RCA method. If the R^2 value is 80%, then 20% of the residual is attributed to the CQA residual. A $\pm 10\%$ error in scoring the subject for the residual should result in a $\pm 10\% \cdot 20\% = \pm 2\%$ error. This is how you improve accuracy! In many cases, accurate value estimates well below $\pm 1\%$ are possible.

Therefore, treating residuals as indirect indicators of the collective value of unmodeled variables in MARS regression is reasonable, provided that the valuation engineer has sufficient local market experience to identify

where and why systemic measurement biases exist. This approach ensures that the analysis accurately reflects actual property values, acknowledging the limitations and nuances of real estate data.

C. What Is Accuracy? Perfect accuracy for a valuation requires:

1. **Conformity to Market Value:** The hypothetical sales transaction of the subject property adheres 100% to the given definition of Market Value
2. **Accurate and Robust Model Development:** A high-quality MARS model is developed with an R^2 of about 75-80% and a CVR2 of roughly 55-65% depending on location.
3. **Good Data Error Analysis:** All likely data errors for measured variables are understood and contained, with little spill over into residuals.
4. **Systemic Bias Accounting:** Any systemic bias in the data should be accounted and adjusted for.
5. **Residual Rank Analysis:** Ensure that residual rank properties follow a reasonable pattern of features from lowest to highest rank. Anomalies such as probate sales, shorts, and auction sales should be investigated, explained, and potentially removed.
6. **Subject Placement:** With a thorough understanding of residual ranking, the subject property should be accurately positioned. A property with a significantly higher appeal and another with a significantly lower appeal should be identified, and the subject placed between them and

scored. Then attempt to narrow the ranking between the higher and lower properties as much as possible.

Given the preceding list, the accuracy can be roughly calculated to be:

$$\begin{aligned} \text{CQA Error} &= \pm e\% \cdot b \\ \text{Expected Price Error} &= \epsilon \cdot (1 - R^2) \end{aligned} \quad (15)$$

So, if the regression R^2 is 80%, we expect about 20% of the subject value to be due to the residual. Now, this is only approximate. We know from the CQA-Residual curve that most of the distribution of residuals is at the lower and higher extremes of CQA. At extremely low or high CQA values, we expect lower accuracy and, in the middle, higher accuracy. But let's assume, for the sake of simplicity, that the distribution of residual from a CQA of 0% to 100% is linear. A $\pm 10\%$ difference in CQA value equates to a $\pm 10\%$ difference in the residual. That percentage difference will be applied to the residuals, which is expected to be 20% of the value. If your subject CQA is 4.2 and you are very sure that it can't possibly be more than 5.2 or less than 3.2, that is $\pm 10\%$ and $\pm 10\% \cdot 20\% = \pm 2\%$. Thus, if your value conclusion is \$1,000,000 then that would be $\pm \$20,000$. Of course, if your CQA is between 20% and 80% we assume a higher accuracy, if not less. In any case, quality's upper and lower ends are difficult to value. The RCA gives an objective valuation with constraints.

The only objective way to verify accuracy is to establish a strict protocol and verify its adherence. This is necessary because actual sale prices can deviate from market value due to imperfections in the world. Comparing appraisals and subsequent sales can provide some indication of accuracy. However, the complexities and instability of the real estate market necessitate the expertise of appraisers or highly qualified valuation engineers to

provide value estimates based on established standards. These standards should be stringent enough to ensure that competent valuation engineers will independently arrive at nearly the same value conclusion.

D. Ranking & Scoring

After creating the MARS (Multivariate Adaptive Regression Splines) model for measured variables, residuals are computed for each property. These residuals represent the difference between the observed net sale prices and the sale prices predicted by the model. An additional column should be created for the Residual/SF, that is, the residual divided by the living area of the residence, with the understanding that generally, the residual for specific properties is more significant with the size of the living area. Other possible statistics may work better in some neighborhoods, but they are a subject for another time

Assumption 7: "Ranking by Residual"

In this paper, references to "ranking by residual" should be taken to mean ranking by residual, by residual/SF, or ranking by some other function of residuals, as the valuation engineer deems most appropriate for the market area. The author's experience in SF Bay Area indicates residual/SF is generally the most reliable, especially when dealing with a subject property that has an unusually small or large GLA (Gross Living Area).

Next, the properties are sorted by their residuals from the most negative to the most positive. A Condition-Quality-Appeal (CQA) score is then generated for each property, ranging from 0.0 to 10.0. This score is

calculated as follows:

The CQA score corresponds to the percentile rank of a property’s residual among all properties. Specifically, the score is computed by determining what percentage of properties have residuals lower than that of the given property and dividing that percentage by 10.

For instance, if a property’s residual is lower than that of 50

Due to rounding, one property can receive a score of 10.0, representing the property with the highest residual. However, a perfect score of 10.0 is theoretically impossible because, logically, 100. On the other hand, the lowest possible score of 0.0 is possible since at least one property has no other properties with lower residuals. Note that more than one property might have a score of 0.0 due to rounding if many properties are being compared.

Therefore, the CQA score provides an intuitive measure of how well a property’s attributes align with the model’s expectations, with higher scores indicating better alignment with the model’s predictions.

It is essential to acknowledge that the CQA score can only be considered an indirect indicator of a comparable property’s appeal when the sale conditions are commensurate with its market value. The Valuation Engineer is responsible for ascertaining whether the residual value incorporates over- or under-market elements by providing a comprehensive explanation and thorough investigation of the sales transaction. While significant over- and under-market sales are not frequent, they can occur in circumstances such as forced sales, probate, incompetence, collusion, etc.

Table 1 presents a semi-realistic extract of MLS data, along with residuals and CQA scores, both in total and per square foot. The total number of properties analyzed in this study was approximately 600. It is noteworthy that several variables that were regressed

are not displayed. In other words, the residuals presented are after the impact of other variables, such as the date of sale, has been considered by MARS regression.

E. CQA-Residual Curve

A plot of the CQA vs. the Residual/SF for a neighborhood will invariably result in the CQA-Residual Characteristic Curve shown in Figure 2. Note that this plot is for Burlingame, CA, around 2022. You will get a similarly shaped curve if you plot the CQA scores against the Residual; I prefer to use the Residual/SF, which has its impact when you convert the CQA for the subject to a corresponding residual by multiplying the associated Residual/SF by the GLA of the subject. However, this also assumes the ranking of the properties is based on residual/SF rather than just residual. The valuation engineer must decide whether the magnitude of residuals is heavily influenced by the property GLA. For example, high end kitchen appliances and bathroom fixtures may have a significant impact on the residual. Still, in an area where homes typically have the same size kitchen and bathrooms, it may bear little relationship to GLA. On the other hand, expensive exterior features and materials, which improve the appeal and thus increase the residual, may be directly impacted by the size of the house. Consequently, it is recommended to plot both the CQA-Residual and CQA-Residual/SF foot curves and study the differences, which can indicate patterns in a market area. The valuation engineer can have his program compute both residuals and take some weighted average to calculate the residual to be assigned to the subject. It may or may not make much of a difference.

1) The CQA Curve Characteristics

Now, if we look at the CQA curve presented, we see that it has three components: on the left, as the CQA score approaches 0, we have exponential decay, on the right, as we approach 10, we have exponential growth, and in the middle, we have a nearly linear function with a steady increase in appeal as we move to higher end homes. It should not be surprising that this reflects reality in a mature market. As homes get old, without maintenance and updating, they do decay: The wood rots, colors become faded, metal gets rusty, and damage occurs. Interestingly, as the damage accrues, more damage can occur and does occur at a faster rate. We have a middle area that typically spans about 80-90% of the appeal range of homes. It reflects household wealth in society. At the upper end, we see the houses that the more wealthy tend to buy and maintain. There are typically only a small proportion of wealthy households in a market area. Only they have the income or wealth to buy the most appealing homes. As we move up in the scale of income and wealth, the percentage of households with a given income level decreases exponentially, and thus, it is not surprising that the perceived market value of these homes increases exponentially as well. Although it is to be noted that the exponential decay and growth functions are somewhat different from each other and different between different market areas, this pattern keeps emerging as we go from one market area to another.

The CQA-Residual curve is characteristic to each market area, and should not be overlooked when doing your analysis.

F. Subject Residual

Unlike the comparables, the subject property does not possess an objectively determined residual; it must be estimated by

the Valuation Expert (VE). This estimation introduces a degree of subjectivity into the Residual Constraint Approach (RCA) protocol, marking a critical juncture where personal judgment plays a pivotal role.

To estimate the Condition-Quality-Appeal (CQA) score, the VE meticulously analyzes the ranked properties, searching for discernible patterns within the rankings. This analysis often involves a detailed review of MLS photos associated with each property. An exhaustive review is impractical for a large dataset, such as 600 properties. Instead, the VE applies practical reasoning to approximate the subject's likely position within the overall ranking. Properties requiring extensive repairs, or "fixers," are generally positioned near the bottom, whereas those with superior updates and appeal tend to rank near the top.

The VE then adopts a more focused approach, 'jumping' within the ranking to identify comparables that closely match the subject's appeal. As the range narrows, the VE pays closer attention to similarities among the properties, particularly in aspects like quality, condition, design aesthetics (including elements like paint, woodwork, stonework, and windows), functional utility, and other relevant characteristics.

It is crucial to understand that the VE's judgments are fundamentally guided by the market-established rankings. This ensures that the subjective elements of the process are anchored by market realities, striving for a fair and accurate valuation of the subject property.

G. RCA Value Conclusion

1) General Considerations

Determining a CQA score is pivotal for establishing the estimated residual value of a property, as these elements are intrinsic

cally linked. While it is possible to assess a property without a CQA, this metric provides valuable context by illustrating where the property stands compared to others in the market area. For instance, a CQA score of 3.2 indicates that the property has greater appeal than 32% of comparable properties in the vicinity. If objections exist to the assigned score, disputants must consult the rankings and justify any proposed adjustments based on photographic evidence or other substantial data. If the ranking system is robust and the Valuation Expert (VE) has accurately positioned the property within this framework, it is unlikely that significant discrepancies will arise.

In the reporting phase, typically, 6 to 12 recent and relevant sales comparables are selected to support the valuation conclusion. The differences between these comparables and the subject property are analyzed, with each variable being adjusted accordingly. These adjustments are then applied to the net sale prices of the comparables to derive adjusted sale prices. As demonstrated mathematically, these adjusted prices should converge with the estimated sale price of the subject property.

2) Review, Auditing, Complaints

Once the estimated residual for the subject property is determined, residual adjustments can be calculated for all comparables. These adjustments serve as constraints for further manual breakdowns by the VE, tailored to specific CQA variables deemed pertinent. It is important to emphasize that these

detailed breakdowns are explanatory and do not influence the final valuation, as affirmed in Proof II.

Concerning complaints submitted by users not satisfied with the appraisal value conclusion, a traditional appraiser typically has to waste time explaining why this or that property with a higher or lower price does not have the user-expected impact on value. With RCA, all possible comparable sales (usually 100-600) are typically evaluated to precisely the same value as the price conclusion, so it takes little effort to show the adjustments for some arbitrary property. The entire RCA process is automated and objective, except for the scoring of the subject residual. And so another advantage of the RCA is that criticism for imperfections in the subjective judgment of the VE is mainly restricted to that one area - and it can be effectively and efficiently managed. Also, assuming an R^2 of about 80%, the impact of a $\pm 10\%$ error in estimating the subject residual amounts to a $\pm 2\%$ error in the final value estimate.

In many cases, the subject fits so perfectly in the residual ranking of properties, that the accuracy can be expected to be better $\pm 1\%$.

The sum of the estimated residual and the subject's MARS estimated sale price yields the final value conclusion known as the Residual Constraint Analysis (RCA). This process ensures a comprehensive and transparent approach to property valuation, fostering confidence in the accuracy and fairness of the assessment.

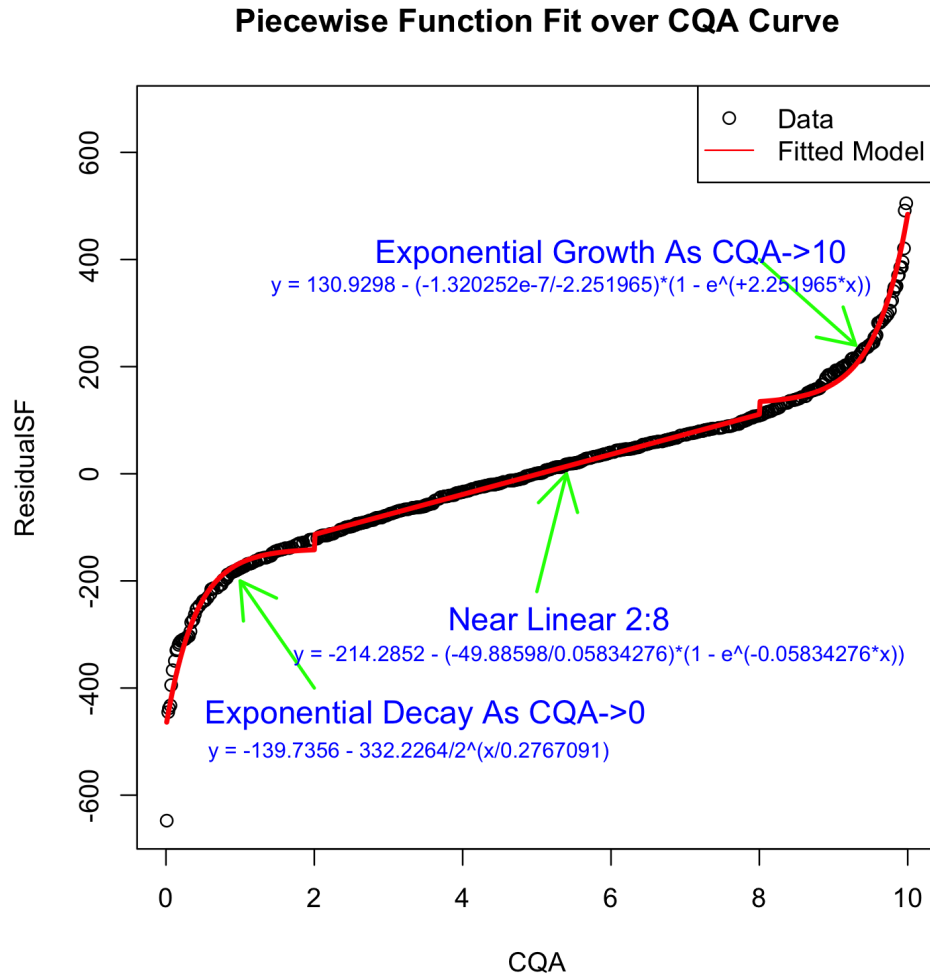
Table 1: Sales Comparables Sorted By Residual

SalePrice	Estimated SP	Residual	CQA	Residual/SF	CQA SF	GLA	SaleDate
\$3,450,000	\$2,318,865	\$1,131,135	10	\$420.50	9.98	2,690	2021-04
\$4,360,757	\$3,399,922	\$960,835	9.9	\$290.28	9.82	3,310	2021-09
\$4,000,000	\$3,134,444	\$865,556	9.8	\$234.57	9.60	3,690	2020-11
\$5,850,000	\$5,150,966	\$699,034	9.7	\$109.91	8.35	6,360	2021-12
\$3,600,000	\$3,023,366	\$576,634	9.5	\$269.46	9.79	2,140	2022-06
\$4,425,000	\$3,987,873	\$437,127	9.2	\$127.44	8.68	3,430	2021-08
\$2,530,000	\$2,164,564	\$365,436	8.9	\$180.02	9.28	2,030	2019-12
\$4,250,000	\$3,884,719	\$365,281	8.8	\$87.18	7.91	4,190	2021-08
\$3,350,000	\$3,034,613	\$315,387	8.5	\$130.87	8.75	2,410	2021-09
\$3,800,000	\$3,522,028	\$277,972	8.2	\$77.43	7.72	3,590	2020-09
\$4,200,000	\$3,963,748	\$236,252	7.8	\$56.12	6.89	4,210	2021-01
\$2,900,000	\$2,702,887	\$197,113	7.5	\$89.19	7.92	2,210	2021-07
\$4,250,000	\$4,103,885	\$146,115	6.9	\$39.38	6.29	3,710	2022-08
\$5,000,000	\$4,864,958	\$135,042	6.8	\$28.25	5.85	4,780	2021-07
\$4,320,000	\$4,199,672	\$120,328	6.5	\$33.90	6.04	3,550	2022-07
\$2,600,000	\$2,479,904	\$120,096	6.5	\$60.96	7.00	1,970	2021-10
\$3,400,000	\$3,301,146	\$98,854	6.2	\$30.32	5.90	3,260	2021-06
\$2,000,000	\$1,901,987	\$98,013	6.2	\$73.14	7.54	1,340	2019-11
\$3,000,000	\$2,919,718	\$80,282	5.9	\$26.50	5.79	3,030	2020-05
\$6,750,000	\$6,699,089	\$50,911	5.6	\$7.75	5.20	6,570	2020-08
\$5,652,000	\$5,601,570	\$50,430	5.5	\$12.83	5.37	3,930	2022-05
\$2,600,000	\$2,579,082	\$20,918	5.2	\$10.51	5.33	1,990	2022-03
\$1,925,000	\$1,905,441	\$19,559	5.1	\$11.51	5.36	1,700	2021-01
\$3,200,000	\$3,209,500	-\$9,500	4.8	-\$3.21	4.88	2,960	2020-11
\$4,150,000	\$4,187,560	-\$37,560	4.5	-\$10.49	4.68	3,580	2021-04
\$2,700,000	\$2,776,510	-\$76,510	3.9	-\$23.61	4.14	3,240	2021-07
\$3,450,000	\$3,567,668	-\$117,668	3.4	-\$31.05	3.90	3,790	2020-09
\$1,975,000	\$2,109,255	-\$134,255	3.3	-\$55.48	3.24	2,420	2020-02
\$1,950,000	\$2,085,431	-\$135,431	3.2	-\$68.40	2.86	1,980	2021-02
\$3,695,000	\$3,859,154	-\$164,154	3.1	-\$47.04	3.56	3,490	2021-05
\$2,288,000	\$2,455,747	-\$167,747	2.9	-\$56.86	3.19	2,950	2020-06
\$2,635,000	\$2,802,781	-\$167,781	2.9	-\$56.87	3.18	2,950	2020-09
\$2,050,000	\$2,217,794	-\$167,794	2.9	-\$82.66	2.46	2,030	2021-07
\$1,550,000	\$1,731,804	-\$181,804	2.7	-\$130.79	1.32	1,390	2021-05
\$2,800,000	\$3,028,411	-\$228,411	2.4	-\$81.58	2.51	2,800	2021-02
\$3,700,000	\$3,976,381	-\$276,381	1.9	-\$76.99	2.62	3,590	2021-08
\$2,517,500	\$2,834,974	-\$317,474	1.5	-\$120.71	1.48	2,630	2019-11
\$2,550,000	\$2,905,330	-\$355,330	1.1	-\$115.74	1.64	3,070	2019-12
\$1,815,000	\$2,263,944	-\$448,944	0.7	-\$273.75	0.15	1,640	2021-11
\$3,711,000	\$4,683,454	-\$972,454	0	-\$226.68	0.37	4,290	2021-12
\$2,225,000	\$3,332,960	-\$1,107,960	0	-\$370.56	0.03	2,990	2021-08
\$3,550,000	\$4,746,489	-\$1,196,489	0	-\$274.42	0.14	4,360	2022-06

Figure 1: Generalized Additive Model Components

Basis	$= g_0^0 = 4957150$
Age	$= g_0^1 = -3163.1809255 \cdot \text{Age}$
AreaNbr	$= g_1^1 = -623197.75104 \cdot I\{\text{AreaNbr} = 471\}$
Baths	$= g_1^1 = -111023.51496 \cdot \max(0, 2 - \text{Baths}) + 96735.52418 \cdot \max(0, \text{Baths} - 2)$
Beds	$= g_2^1 = -233489.66415 \cdot \max(0, \text{Beds} - 5)$
DaysOffMkt	$= g_3^1 = -2577.2188622 \cdot \max(0, 189 - \text{DaysOffMkt}) - 1282.8789534 \cdot \max(0, \text{DaysOffMkt} - 189)$
GLA	$= g_4^1 = -666.73587914 \cdot \max(0, 4027 - \text{GLA}) + 281.36312913 \cdot \max(0, \text{GLA} - 4027)$
Latitude	$= g_5^1 = -18141680.417 \cdot \max(0, \text{Latitude} - 37.573)$
LotSize	$= g_6^1 = -85.005683404 \cdot \max(0, 8000 - \text{LotSize}) - 23.562970183 \cdot \max(0, \text{LotSize} - 8000)$
Age_Baths	$= g_6^2 = -3916.3594872 \cdot \text{Age} \cdot \max(0, \text{Baths} - 3.5)$
Age_Garage	$= g_7^1 = -6503.1897339 \cdot \text{Age} \cdot \max(0, \text{Garage} - 2)$
AreaNbr_LotSize	$= g_2^2 = 78.54207825 \cdot I\{\text{AreaNbr} = 463\} \cdot \max(0, 8000 - \text{LotSize})$ $+ 87.263298012 \cdot I\{\text{AreaNbr} = 466\} \cdot \max(0, \text{LotSize} - 8000)$ $- 1070222.5995 \cdot I\{\text{AreaNbr} = 464\} \cdot \max(0, \text{Baths} - 2)$ $- 10759.444688 \cdot \text{Age} \cdot I\{\text{AreaNbr} = 460\} \cdot \max(0, \text{Baths} - 3.5)$
AreaNbr_Baths	$= g_3^2 = 17571072.548 \cdot I\{\text{AreaNbr} = 466\} \cdot \max(0, \text{Latitude} - 37.573)$
AreaNbr_Latitude	$= g_4^2 = -328.79173933 \cdot I\{\text{AreaNbr} = 471\} \cdot \max(0, \text{GLA} - 2550)$
AreaNbr_GLA	$= g_5^2 = -525.69503229 \cdot I\{\text{AreaNbr} = 471\} \cdot \max(0, 2550 - \text{GLA})$
DaysOffMkt_GLA	$= g_6^2 = 0.43839582286 \cdot \max(0, \text{DaysOffMkt} - 661) \cdot \max(0, 4027 - \text{GLA})$
FrPlcNbr_LotSize	$= g_7^2 = 22.059059591 \cdot \max(0, 2 - \text{FrPlcNbr}) \cdot \max(0, \text{LotSize} - 8000)$ $+ 39.084508159 \cdot \max(0, \text{FrPlcNbr} - 2) \cdot \max(0, \text{LotSize} - 8000)$
GLA_LotSize	$= g_8^2 = -0.45274160024 \cdot \max(0, \text{GLA} - 2994) \cdot \max(0, 8000 - \text{LotSize})$
ADU_Garage	$= g_0^3 = -5.4578633919 \cdot \max(0, \text{ADU} - 0) \cdot \text{Age} \cdot \max(0, 2 - \text{Garage})$
AreaNbr_Baths_LotSize	$= g_1^3 = 60.607126605 \cdot I\{\text{AreaNbr} = 460\} \cdot \max(0, \text{Baths} - 2) \cdot \max(0, \text{LotSize} - 2480)$
AreaNbr_Baths_FrPlcNbr	$= g_2^3 = 593779.41722 \cdot I\{\text{AreaNbr} = 464\} \cdot \max(0, \text{Baths} - 2) \cdot \max(0, \text{FrPlcNbr} - 0)$
AreaNbr_GLA_Latitude	$= g_3^3 = 54199.925713 \cdot I\{\text{AreaNbr} = 464\} \cdot \max(0, \text{GLA} - 2670) \cdot \max(0, \text{Latitude} - 37.573)$ $+ 58934.565747 \cdot I\{\text{AreaNbr} = 471\} \cdot \max(0, \text{GLA} - 2550) \cdot \max(0, \text{Latitude} - 37.568)$
AreaNbr_DaysOffMkt_GLA	$= g_5^3 = 0.37173306643 \cdot I\{\text{AreaNbr} = 470\} \cdot \max(0, 661 - \text{DaysOffMkt}) \cdot \max(0, 4027 - \text{GLA})$
Beds_GLA_LotSize	$= g_6^3 = -5.1607264026 \cdot \max(0, 4 - \text{Beds}) \cdot \max(0, \text{GLA} - 2994) \cdot \max(0, 8000 - \text{LotSize})$

Figure 2: CQA-Residual Characteristic Function



References

- Friedman, Jerome H. Multivariate Adaptive Regression Splines. 1991(1):1–141.
- Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, second edition edition. ISBN 978-0-387-84857-0 978-0-387-84858-7.
- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, editors. *An Introduction to Statistical Learning: With Applications in R*. Number 103 in Springer Texts in Statistics. Springer. ISBN 978-1-4614-7137-0.
- Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, corrected at 5th printing edition. ISBN 978-1-4614-6848-6 978-1-4614-6849-3.
- Stephen Milborrow. Notes on the earth package. a. URL <http://mtweb.cs.ucl.ac.uk/mus/lib/R/earth/doc/earth-notes.pdf>.
- Stephen Milborrow. Variance models in earth. b. URL <http://www.milbo.org/doc/earth-varmod.pdf>.
- Wengang Zhang. *MARS Applications in Geotechnical Engineering Systems: Multi-Dimension with Big Data*. Springer. ISBN 978-981-13-7421-0 978-7-03-061046-1.