

Residual Constraint Approach (RCA): estrutura metodológica e protocolo

Tradução para o português (Brasil) da versão atualizada do artigo originalmente publicado no Zenodo em 5 de março de 2025. Tradução elaborada com assistência de IA e revisada pela redação; a versão em inglês é a autorizada.

William Bert Craytor, SRA

Volume 1, Número 1 · Verão de 2026

Publicado em 1º de julho de 2026

DOI: [10.67330/kgebx46](https://doi.org/10.67330/kgebx46)

© 2026 William Bert Craytor. Published by Pacific Vista Net.

Este artigo de acesso aberto é distribuído sob os termos da [licença Creative Commons Atribuição 4.0](#).

Resumo. Uma limitação importante dos métodos convencionais de avaliação está na sua dependência de técnicas manuais ultrapassadas, como a análise de pares comparativos (*matched pairs*), que ignoram as contribuições de valor das características do imóvel como se elas não existissem: o *Sales Grid* (quadro de homogeneização) tradicional não possui coluna para contribuições de valor, apenas para ajustes. Na verdade, as contribuições de valor fornecem a base necessária para as restrições matemáticas que impedem a supervalorização e a subvalorização pelos avaliadores tradicionais.

O Residual Constraint Approach (RCA, pt.: abordagem por restrição residual) faz avançar o método comparativo direto de dados de mercado (*Sales Comparison Approach*, SCA) tradicional ao integrar um processo de avaliação multifásico que emprega Multivariate Adaptive Regression Splines (MARS) juntamente com restrições matemáticas rigorosas sobre as contribuições de valor das características — em particular as características não medidas: variáveis latentes como estética, projeto, estado de conservação e padrão construtivo, cujos valores costumam ficar sujeitos ao juízo subjetivo e ao viés do avaliador tradicional.

As restrições matemáticas do RCA apoiam-se na aplicação especializada da regressão MARS para estimar as contribuições de valor das características mensuráveis do imóvel ao preço de venda. Contudo, essa estimativa normalmente responde por apenas cerca de 80% do preço de venda efetivo, excluindo características avaliadas subjetivamente, como projeto, utilidade funcional, estado de conservação e padrão construtivo. Na San Francisco Bay Area, o resíduo — que representa esses componentes subjetivos — gira em torno de 20% do valor total do imóvel. A cifra de 80/20 reflete a experiência do autor na San Francisco Bay Area sob condições típicas de dados e não deve ser lida como uma constante invariante entre mercados. Em mercados nos quais o R^2 alcançável é menor, a resposta da estrutura metodológica consiste em identificar e medir os determinantes de valor não medidos — a distância ao oceano em comunidades litorâneas, a altimetria em bairros de encosta e variáveis semelhantes coletadas de fontes externas quando os dados do MLS não as fornecem — em vez de aceitar a limitação e reportar faixas de erro mais amplas. Embora os resíduos do MARS sejam frequentemente tratados como erros de estimação em outros contextos, na avaliação de imóveis eles funcionam como medida indireta do valor das variáveis latentes. Ainda que isso possa parecer pouco prático à primeira vista, uma demonstração estabelece que o resíduo pode ser decomposto de maneira significativa em componentes descritivos, desde que suas contribuições somem o resíduo — sem afetar o resultado final da avaliação.

Esta metodologia resulta de duas décadas de aplicação empírica por parte do autor, desenvolvida por meio de implementações iterativas do MARS em ambientes R e Python.

Palavras-chave: avaliação de imóveis, avaliação residencial, Residual Constraint Approach, MARS, Multivariate Adaptive Regression Splines, método comparativo direto de dados de mercado, análise de resíduos, variáveis latentes, avaliação automatizada

Sumário

1	Introdução	4
2	Antecedentes	6
3	Engenheiro de Avaliações	9
4	Premissas	10
5	Fluxo de Trabalho do RCA	13
5.1	Tipos de Análise MARS	13
5.2	Fluxo de Processamento	13
5.3	Fluxo de Dados	14
5.3.1	Requisitos de Dados para o MARS	15
5.3.2	Protocolos: Listagens de Vendas Fechadas	16
5.4	Estrutura de Pastas do Projeto	18
5.4.1	Protocolos: Múltiplas Execuções do MARS (earth)	19
5.5	A Pasta de Trabalho Excel Central	21
5.6	Backups	21
6	Notação da Pasta de Trabalho	22
6.1	Dimensões Descritivas	22
6.1.1	A Dimensão do Projeto (“a”)	23
6.1.2	Dimensão da Versão de Execução (“r”)	24
6.1.3	Etapas do Fluxo de Trabalho (“w”)	24
6.1.4	Planilhas da Pasta de Trabalho Excel (“s”)	25
6.1.5	Listagens de Imóveis (“p”) — linhas do MLSData	25
6.1.6	Variáveis do MLS (“v”) — colunas do MLSData	25
6.1.7	Exemplo Final de Notação	26
6.1.8	Justificativa	26
6.1.9	Matrizes?	27
6.1.10	Primeira Simplificação	27
6.1.11	Segunda Simplificação	27
7	O Sales Grid do RCA	29

8	Variáveis	33
8.1	Ontologia da Origem das Variáveis	33
8.2	Ontologia da Aplicação das Variáveis	34
8.3	Termos de Segundo e Terceiro Graus	36
8.3.1	Variáveis de Ajuste	36
8.3.2	Variáveis de Agregação	36
8.3.3	Variáveis e Notação Especiais	37
8.3.4	Decompondo os Resíduos	37
9	Demonstrações Fundamentais do RCA	39
10	Modelo MARS	41
10.1	Funções “ g ” de Variáveis e Interações do Modelo	41
10.2	Indexação das Funções g	41
10.3	Dimensões dos Gráficos	42
10.4	A Equação Final do Modelo	42
10.5	Número de Gráficos Necessários Para Representar um Modelo	43
11	Características dos Dados	46
11.1	Finalidade do Resíduo	46
11.2	Notas de Estado de Conservação e Padrão Construtivo Como Regressores Aproximados	47
11.3	Erros nos Dados	47
11.4	O Que É Exatidão?	49
11.5	O Imóvel Avaliando Como Anomalia	50
11.6	A Importância do CVR2 no Trato de Dados Inesperados	51
12	Resíduos	52
12.1	Ordenação e Pontuação	52
12.2	Curva CQA-Resíduo	55
12.3	As Características da Curva CQA	55
12.3.1	Análise dos Componentes da Curva CQA	55
12.3.2	Dinâmica de Mercado e Implicações Sociais	56
12.4	Resíduo do Avaliando	56
12.5	Conclusão de Valor do RCA	58
12.5.1	Considerações Gerais	58
12.5.2	Revisão, Auditoria, Reclamações	58
13	Conclusão	60

1 Introdução

Este artigo apresenta o fundamento lógico e matemático de uma abordagem moderna para estimar com exatidão o valor de mercado de imóveis residenciais e de outros tipos de bens imóveis. O método, desenvolvido ao longo de 20 anos de experiência na aplicação do software Multivariate Adaptive Regression Splines (MARS) a residências do norte da Califórnia — em particular em comunidades da San Francisco Bay Area —, representa um avanço significativo na precisão da avaliação. O MARS eleva a qualidade da avaliação ao capturar com habilidade as relações não lineares entre as características do imóvel e o valor, ao mesmo tempo em que identifica e prioriza de forma eficiente as variáveis de maior impacto. Essa capacidade permite que os avaliadores descubram a dinâmica do valor em novas regiões de mercado muito mais rapidamente do que os métodos tradicionais permitem. O MARS constitui a espinha dorsal do Residual Constraint Approach (RCA) (pt.: abordagem por restrição residual) aqui proposto. O método RCA faz jus ao seu nome graças ao seu arcabouço rigoroso:

- Impõe restrições estritas sobre as contribuições de valor das características do imóvel, tanto medidas quanto não medidas:
 1. Todas as contribuições de valor derivadas do MARS, inclusive o resíduo, devem somar o preço de venda líquido de cada elemento comparativo.
 2. O resíduo pode ser decomposto em subcomponentes com valores atribuídos, desde que seu total seja igual ao resíduo do elemento comparativo.
- O RCA minimiza o viés na estimativa das características subjetivas — tais como o apelo estético, o estado de conservação e o padrão construtivo — ao vincular seu valor total ao resíduo, o qual é quantificado com exatidão quando um engenheiro de avaliações competente constrói o modelo MARS.

Embora concebido principalmente para imóveis residenciais, nos quais as características subjetivas influenciam de maneira significativa o valor de mercado, o RCA tem potencial para aplicações mais amplas na avaliação de ativos. Avaliações exatas são críticas diante do valor estimado de \$45 trilhões do mercado residencial de imóveis dos EUA. A Fannie Mae observa que as avaliações típicas se desviam cerca de 5% dos preços de venda efetivos, embora essa cifra mascare o viés, já que os avaliadores frequentemente enfrentam pressão para se alinhar aos preços de venda. Além disso, como 60%–70% das avaliações dão suporte a refinanciamentos — nos quais não há preços de venda comparáveis —, a necessidade de precisão, tal como a oferecida pelo RCA, torna-se evidente. A análise do Federal Reserve de 2012 estimou que 15%–20% das avaliações anteriores à crise de 2008 se desviavam em mais de $\pm 10\%$ dos preços de venda, faixa provavelmente ainda pertinente hoje para as avaliações de refinanciamento, com variações conforme o tipo de imóvel e o mercado. Por que a exatidão importa? Em operações com baixa relação empréstimo-valor (loan-to-value, LTV) — por exemplo, um comprador com bom histórico de crédito dando \$500,000 de entrada em uma residência de \$1 milhão — os credores podem dispensar a avaliação, supondo risco mínimo de execução hipotecária. Contudo, os compradores que dispensam a avaliação correm o risco de pagar demais e de descobrir defeitos somente após a compra. Isso ressalta o valor de métodos robustos de avaliação como o RCA. Quantificar a exatidão das avaliações continua sendo um desafio. Os pesquisadores carecem de métricas diretas dos desvios em relação ao “valor de mercado” e recorrem, em seu lugar, a indicadores indiretos. Os métodos tradicionais de avaliação, enraizados no juízo subjetivo e na competência variável dos avaliadores, complicam a análise. Entre as estratégias empregadas estão:

1. Tzioumis (2016) estudou 4,782 avaliadores em 259,436 avaliações em 11 estados e constatou que um terço ficava dentro de 1% do preço de contrato, metade entre 1%–3% acima e uma

minoria em média 5% acima — o que sugere viés de alta. O estudo identificou dois tipos de avaliador: os “ao valor de mercado” (próximos ao preço de contrato) e os “acima do valor de mercado” (consistentemente mais altos).

2. Krivorotov e LaCour-Little (2020) constatou que 90% das avaliações de compra igualavam ou superavam os preços de contrato, o que evidencia o “viés de ancoragem” — a tendência a se alinhar a um valor desejado —, especialmente em contextos de refinanciamento.

Esses estudos revelam viés nas práticas de avaliação, amplificado pelas limitações das ferramentas tradicionais. Se os avaliadores fossem uniformemente competentes e imparciais, suas avaliações convergiriam; a supervalorização persistente por parte de certos grupos sinaliza problemas sistêmicos.

Os estudos sobre viés citados acima, e o padrão sistêmico que eles apontam, refletem um traço mais profundo da prática tradicional de comparação de vendas que o RCA foi concebido para enfrentar. A metodologia tradicional opera em dois canais com estatuto epistêmico bastante distinto. No canal das características mensuráveis — GLA (área bruta habitável), número de cômodos, área do terreno, idade, localização — espera-se que o avaliador derive os valores de ajuste de maneira objetiva, por meio da análise de pares comparativos ou de regressão linear local sobre dados de mercado. Avaliadores competentes realizam esse trabalho, e os ajustes resultantes são, no mínimo, aproximadamente defensáveis diante dos dados. Mesmo aqui, contudo, o quadro de homogeneização tradicional registra somente ajustes; ele sequer possui uma coluna para contribuições de valor. Tampouco os ajustes tradicionais são calculados como diferenças entre contribuições de valor — eles são estimados diretamente, a partir de pares comparativos ou de procedimentos semelhantes, sem que as contribuições subjacentes cheguem a ser calculadas. No RCA, ao contrário, cada ajuste é a diferença entre as contribuições de valor do imóvel avaliando e as do elemento comparativo para a característica considerada — a estrutura cujas consequências a Demonstração II torna precisas. No canal das características subjetivas — estado de conservação, padrão construtivo, projeto, utilidade funcional, estética — os procedimentos de fundamentação disponíveis vão do tênue ao precário. Os ajustes de características subjetivas podem ser fundamentados por análise de pares comparativos de imóveis fora do conjunto imediato de comparáveis, por apelo à prática reconhecida dos pares ou por referência à experiência documentada do avaliador no mercado. Todos os três são formalmente conformes às normas profissionais. Nenhum deles atinge o rigor estatístico que o canal mensurável alcança rotineiramente.

O resultado é uma assimetria procedimental entre o que as normas exigem e o que a metodologia consegue entregar. As normas exigem fundamentação para cada ajuste, mensurável ou subjetivo. A metodologia entrega fundamentação robusta em um canal e um espectro de procedimentos mais fracos no outro, indo de um par comparativo anedótico de tamanho amostral um até a citação daquilo que outros avaliadores costumam utilizar. O avaliador honesto e competente que faz um trabalho cuidadoso de pares comparativos para um ajuste de vista produz fundamentação sensivelmente melhor do que aquele que cita a prática dos pares sem embasamento, mas um revisor que compare as duas avaliações no formulário padrão não consegue distingui-las facilmente, porque o formulário registra o valor do ajuste, e não o rigor de sua derivação. A metodologia tolera a assimetria porque nenhum procedimento mais forte para características subjetivas esteve, de modo geral, à disposição dos avaliadores em atividade.

O RCA aprimora ambos os canais, mas de maneiras categoricamente distintas. No canal mensurável, a regressão MARS captura relações não lineares, interações entre variáveis e uma estrutura de funções de base que a análise de pares comparativos e a regressão linear ordinária não conseguem representar. Ajustado contra o conjunto completo de vendas qualificadas da região de mercado — tipicamente várias centenas — em vez de contra os três a nove elementos comparativos do quadro tradicional, as contribuições de valor derivadas do MARS são ao mesmo

tempo objetivas e substancialmente mais exatas do que a metodologia que substituem. No canal subjetivo, o RCA faz algo diferente em espécie de qualquer dos procedimentos tradicionais de fundamentação. Ele não tenta, de modo algum, estimar valores de ajuste para características subjetivas individuais. As contribuições de todas as características subjetivas — variáveis latentes — aparecem coletivamente como o resíduo: a diferença entre o preço de venda líquido efetivo de cada elemento comparativo e o preço previsto pelo MARS, calculável a partir dos dados sem intervenção do avaliador. A decomposição desse resíduo em componentes nomeados — Projeto, Utilidade Funcional, Vista, Estado de Conservação, Padrão Construtivo e assim por diante — é restringida a somar o total do resíduo, e, pela Demonstração II (Seção 9), qualquer decomposição que satisfaça essa restrição produz a mesma conclusão de valor. O juízo que permanece no fluxo de trabalho é o posicionamento do imóvel avaliando na distribuição ordenada dos resíduos dos elementos comparativos, feito contra uma classificação inspecionável e explicitamente documentada no laudo. Uma avaliação RCA fraca continua sendo possível — um VE inexperiente pode produzir um modelo MARS sobreajustado, posicionar o avaliando de forma inverossímil ou deixar passar anomalias nos dados —, mas ela é fraca de maneiras que um revisor pode identificar ao inspecionar o modelo e a classificação, e não de maneiras que uma revisão por formulário padrão não consiga trazer à tona.

Para além do refinanciamento, os avaliadores devem aos proprietários avaliações exatas, especialmente em compras de alto valor, nas quais os erros podem custar dezenas ou centenas de milhares. Ainda assim, muitos hesitam em gastar \$600–\$1,000 em avaliações, duvidando de seu valor agregado. Métodos avançados como o RCA, embora inicialmente intensivos em recursos, prometem maior precisão. Com refinamento e experiência, sua eficiência — e sua adoção — devem aumentar.

2 Antecedentes

A formação acadêmica do autor é em matemática, ciência da computação e estatística. Por profissão, é engenheiro de software, com experiência secundária em avaliação e contabilidade. Realiza avaliações, de forma intermitente, desde 2002 e aplica o MARS desde cerca de 2004 para determinar ajustes de imóveis e tendências de preço em imóveis residenciais.

Avaliar com MARS é “experiência profunda” em avaliação. O MARS gera modelos que refletem, com mais detalhe do que qualquer outro método estatístico, tanto as variáveis que contribuem para o valor de mercado quanto o modo como mudanças nesses valores impactam o preço de venda. Assim, ao avaliar com MARS, o avaliador investiga constantemente as relações causais entre valores de variáveis e preços de venda tal como geradas pelos modelos MARS.

No interesse da transparência, a metodologia RCA é mais bem aplicada a áreas residenciais com estas características:

1. Bairros complexos, com imóveis bastante diferentes entre si
 - Residências mais antigas
 - Qualidades variadas de reforma
 - Muitas arquiteturas e projetos distintos
 - Variação em atributos do imóvel como GLA, área do terreno, altimetria e vista.
2. Condições de mercado complexas
 - Condições de mercado em rápida mudança
 - Muitos tipos distintos de compradores

3. Boas fontes de dados: o protocolo RCA baseia-se em mineração de dados e necessita de dados abundantes e de qualidade. Embora possa ser usado em áreas com dados escassos, talvez não compense o tempo e o custo exigidos.
4. Deseja-se estimativas exatas do valor de mercado. O RCA deve, na maioria dos casos, conseguir fornecer exatidão de $\pm 1-2\%$. O laudo do engenheiro de avaliações deve comentar as condições de exatidão na conclusão final de valor. Se faltarem dados ou informações, ou se padrões consistentes do mercado não tiverem sido descobertos por meio da mineração de dados, a exatidão será prejudicada e o cliente deverá ser informado da situação. Especificamente, um R2 e um CVR2 elevados (R2 validado cruzadamente, calculado pelo particionamento repetido do conjunto de dados MLS em partições aleatórias de treino e teste), um modelo MARS significativo e um padrão forte nas características CQA (condition-quality-appeal: estado de conservação, padrão construtivo e apelo) do imóvel nos resíduos ordenados são os principais meios de aferir exatidão e confiabilidade na conclusão de valor.

O desenvolvimento destes protocolos RCA ao longo dos últimos 20 e poucos anos apoia-se na disponibilidade dos dados da MLSListings Inc. (Sunnyvale, CA) e de outros MLS, atualmente acessíveis por meio da MLSListings, que representam mais de 95% da experiência de análise de dados do autor. A expertise do autor decorre de extenso trabalho de avaliação em 15 condados primários do norte da Califórnia, com experiência adicional em 10 condados secundários.

A metodologia RCA é mais eficaz em mercados que atendem a dois critérios essenciais:

1. Ampla disponibilidade de dados de avaliação de alta qualidade
2. Áreas residenciais complexas e heterogêneas, caracterizadas por:
 - Estilos arquitetônicos diversos
 - Graus variados de reforma dos imóveis
 - Padrões de urbanização ao longo de várias décadas

Experiência nos Condados da Califórnia

As afirmações feitas neste artigo baseiam-se na experiência de avaliar imóveis nos seguintes condados do norte da Califórnia:

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

As áreas acima abrangem essencialmente o Vale do Silício e os condados que o cercam, estendendo-se para fora em direção a regiões mais rurais, incluindo o Vale Central e o sopé da Sierra. Cabe mencionar que o autor sempre teve acesso a dados abundantes e *relativamente* exatos em comparação com aquilo de que muitos avaliadores dispõem em outras áreas menos desenvolvidas dos EUA.

Os dados de MLS utilizados são certificados RESO, em conformidade com o RCF (RESO Common Format). Apesar da certificação RCF, dados de importância geralmente secundária às vezes estão ausentes ou incorretos para alguns imóveis, seja por erros de lançamento do órgão fiscal de origem, seja por erros posteriores de digitação do corretor no MLS. Com frequência esses

erros seguem um padrão, como simplesmente não fazer um lançamento quando a característica não existe. Por exemplo, em vez de digitar 0 em “carport spaces” (vagas cobertas sem fechamento) quando não há carport, nada é digitado. Em áreas onde carports são raros, o analista pode substituir a ausência de dado por 0, a fim de obter um ajuste útil para carports. Embora o R/earth e outras implementações de MARS consigam lidar com dados faltantes, cabe ao Valuation Engineer (VE, engenheiro de avaliações) decidir como tratá-los.

Os dados da MLSListings também estão disponíveis a partir dos MLS de alguns outros estados, da Flórida ao Havaí. Isso pode ser útil para testar o quão bem os protocolos e o código funcionam em estados distintos.

3 Engenheiro de Avaliações

Definição 1: Engenheiro de Avaliações

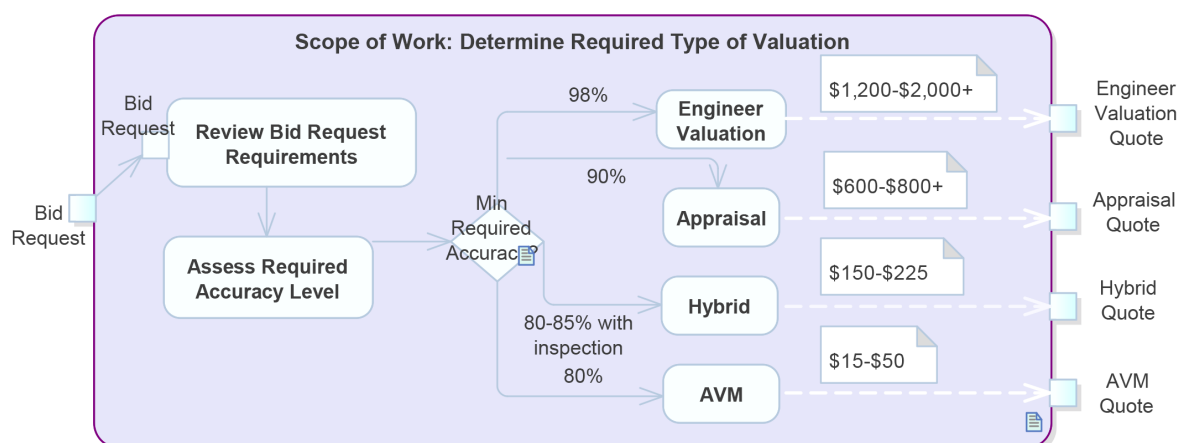
Para evitar ambiguidade, o termo “Valuation Engineer” (ou “VE”, em português engenheiro de avaliações) é usado para designar um indivíduo que reúne a expertise de um avaliador habilitado com conhecimentos, habilidades e experiência avançados no uso de mineração de dados, estatística, programação, protocolos detalhados e inteligência artificial para apurar o valor de mercado de imóveis complexos. Também é importante reconhecer que a avaliação é uma tarefa que exige alguma forma de habilitação, dado o potencial já estabelecido de fraude.

Os credores tipicamente oferecem várias opções de avaliação:

1. Automated Valuation Models (AVMs, modelos automatizados de avaliação) — a opção de menor custo
2. Avaliações híbridas — que utilizam inspetores não habilitados sob supervisão de avaliador habilitado
3. Avaliações padrão — realizadas por avaliadores habilitados

Ausentes da consideração típica para trabalhos de avaliação estão as avaliações em nível de engenharia. Simplesmente não há muitos avaliadores por aí com o conjunto de competências necessário para realizá-las. Deveria haver demanda por elas. Elas podem proporcionar exatidão significativamente maior e assegurar quase por completo uma objetividade razoavelmente boa com o emprego da metodologia RCA (MARS e restrições).

Para imóveis padronizados em loteamentos mais novos, as avaliações padrão em geral bastam. Contudo, avaliações de engenharia são recomendadas para residências mais antigas em áreas com histórico significativo de reformas ao longo de mais de 30 anos. A escolha entre esses métodos deve seguir um processo decisório estruturado, baseado nas características do imóvel e nas condições de mercado.



4 Premissas

Premissa 1: Área Geográfica e Tipo de Imóvel

O escopo deste artigo concentra-se primordialmente em residências unifamiliares e condomínios nas áreas metropolitanas e urbanas da Califórnia. Embora o método comparativo direto de dados de mercado se aplique a outros tipos de imóveis e regiões geográficas, os métodos da renda e do custo costumam desempenhar papel mais dominante na avaliação de terras agrícolas, imóveis multifamiliares e imóveis comerciais. Nesses casos, a abordagem RCA pode exigir modificação ou ter aplicabilidade limitada.

Premissa 2: Armazenamento de Dados

CSV e Excel são amplamente usados pelos sistemas MLS e assumidos como a principal forma de alimentar os programas RCA com dados. Ainda que feeds de corretoras possam ser usados para alimentar continuamente bancos de dados SQL. Backup e armazenamento intermediário frequentemente usam bancos de dados SQL, e dois que desempenham papéis importantes são os gratuitos e de código aberto SQLite e PostgreSQL. Arquivos de texto ou RTF também são usados para trechos de laudo, e arquivos PNG, JPG e TIFF são usados para gráficos. Note que o SQL Server da Microsoft também é usado por muitos, mas a edição gratuita “SQL Server Express” é muito limitada (por exemplo, só é possível usar 1 núcleo de CPU) e outras versões têm taxas de licença pesadas para uso comercial. Bancos de dados imobiliários na Califórnia podem ser muito grandes e as junções podem consumir muito tempo, de modo que um computador com 8 ou mais núcleos e 128G de RAM pode ser útil para junções grandes. Bancos de dados SQL também são importantes quando é preciso encontrar todos os imóveis em uma área delimitada arbitrariamente. Assim, se, por exemplo, você usasse análise de agrupamentos para definir bairros, tenderia a usar algum banco de dados SQL. Contudo, bancos de dados diferentes possuem funções internas distintas para o tratamento de informação geográfica e, portanto, não são compatíveis entre si. Para este conjunto de artigos sobre RCA, presume-se o PostgreSQL para tarefas SQL complexas e o SQLite para tarefas simples.

Premissa 3: “R/earth” É Usado Para o Processamento MARS.

Até 2017, o software MARS da Salford Systems era uma solução acessível, a \$270 por ano. Contudo, após a aquisição da Salford Systems pela Minitab, o MARS passou a ser vendido no pacote SPM a \$16,000 anuais — um patamar de preço proibitivo para a maioria das pequenas empresas. Felizmente, o pacote “earth” da linguagem de programação R (Milborrow 2024a), mantido por Stephen Milborrow, evoluiu a ponto de igualar ou superar as capacidades do MARS do SPM da Minitab. O pacote earth oferece vantagens adicionais: é gratuito, integra-se perfeitamente ao abrangente ecossistema estatístico do R e pode ser acessado a partir do Python usando Pandas, proporcionando excelente flexibilidade para diferentes fluxos de trabalho. As discussões relativas ao MARS pressupõem o uso específico do R/earth.

Premissa 4: Notação

Dados brutos de fontes externas, como arquivos CSV e planilhas Excel, são representados com caracteres latinos padrão. Isso inclui os dados de imóveis do Multiple Listing Service (MLS), as configurações do MARS (Multivariate Adaptive Regression Splines) usando o pacote R/earth e dados específicos do projeto. Essa notação estende-se às respectivas representações em data frame no R ou no Python.

Ao tratar de *características do imóvel* e de seus vários tipos e subconjuntos, a notação passa a usar letras gregas, para maior precisão e clareza analítica. Embora essa notação matemática possa parecer mais rigorosa, ela cumpre um propósito crucial: distinguir as características abstratas do imóvel de suas representações em dados brutos. Por convenção, os caracteres gregos denotam consistentemente características do imóvel, ao passo que os caracteres latinos indicam estruturas de dados de planilha. Por exemplo, o símbolo M representa a matriz de dados do MLS, organizada com os imóveis nas linhas e as respectivas características nas colunas.

Premissa 5: Linguagens de Programação

Nesta área, R e Python com Pandas são as linguagens dominantes, com C ou C++ sendo usados quando a velocidade é importante. O próprio R/earth é escrito em R e C.

Premissa 6: Indexação

A indexação pode variar entre linguagens de programação e matemática, o que pode gerar confusão:

- A matemática e o R iniciam a indexação em 1.
- C, C#, C++ e Python iniciam em 0.

Neste artigo, para manter tudo consistente e simples, toda a indexação começa em 0 (como em Python ou C). Veja como funciona:

- Para listas, matrizes ou data frames:
 - O índice 0 é reservado ao “imóvel avaliando” (o imóvel principal em estudo).
 - Exemplo: em uma tabela cujas linhas são imóveis, a linha 0 é o imóvel avaliando, e a coluna 0 contém identificadores únicos de todos os imóveis.
 - Uma dica prática: organize os imóveis com o avaliando primeiro (linha 0) e, em seguida, liste os demais da data de venda mais recente para a mais antiga, numerando-os sequencialmente (0, 1, 2 etc.).
- Quando não há imóvel avaliando:
 - Se uma estrutura de dados não inclui imóveis (por exemplo, apenas variáveis ou funções), o índice 0 é pulado e a indexação começa em 1.
 - Nota especial para usuários de R: como o R usa indexação baseada em 1, ao codificar em R a matemática deste artigo, some 1 a todos os índices para ajustar de base 0 para base 1.

5 Fluxo de Trabalho do RCA

Compreender o fluxo de trabalho do RCA é essencial para uma análise eficaz. O processo segue uma sequência estruturada de tarefas analíticas e pontos de decisão

5.1 Tipos de Análise MARS

O fluxo de trabalho da análise MARS depende de se é necessária apenas uma análise de mercado ou se é exigida a avaliação de um determinado imóvel. Note que a avaliação de um imóvel específico normalmente requer a análise de sua região de mercado.

- Análise de mercado: concentra-se em áreas amplas, como condados, cidades ou regiões delimitadas. Essa abordagem não envolve um imóvel avaliando específico e, portanto, não exige conclusão de valor. O fluxo de trabalho gira em torno do desenvolvimento do modelo, da validação e da geração de análises de apoio. O processo se encerra assim que o modelo ótimo é identificado e documentado.
- A análise de imóvel único abrange todos os elementos da análise de mercado, acrescidos da determinação do valor de um imóvel específico. Isso difere fundamentalmente dos métodos tradicionais de avaliação. Nas abordagens tradicionais, o valor é tipicamente obtido pelo ajuste de 3 a 9 dados de mercado e pelo cálculo de uma média ponderada. Em contraste, o RCA determina o valor primordialmente por meio de dois componentes:
 1. O valor previsto pelo modelo MARS
 2. O resíduo estimado, específico do imóvel (com o respectivo escore CQA)

O Sales Grid (quadro de homogeneização) no RCA serve primordialmente para explicar as diferenças de valor entre o avaliando e os elementos comparativos, ilustrando as contribuições das variáveis. Uma vantagem central do RCA é sua estabilidade: embora os clientes possam discutir a seleção dos comparáveis, a conclusão de valor permanece consistente independentemente de quais comparáveis sejam escolhidos para apresentação, pois todas as vendas (sejam 15 ou 600) são ajustadas com o mesmo modelo. Ainda que vendas mais antigas sejam automaticamente ajustadas às condições atuais de mercado, vendas recentes são em geral preferíveis. A eficácia do modelo em capturar os fatores de preço diminui com a distância temporal, na medida em que características de mercado não medidas se tornam mais significativas em períodos mais longos.

5.2 Fluxo de Processamento

O fluxo geral de processamento pode ser descrito por estas etapas:

1. Escopo do trabalho.
2. Configurar as pastas e os arquivos iniciais do projeto.
3. Baixar os dados do MLS. Consulte os Protocolos 1.1–1.3 quanto ao número de registros de imóveis necessários e à definição de limites para os intervalos de datas.
4. Completar as planilhas de configuração: variáveis de regressão, interações, agregações, cálculos.
5. Executar a regressão MARS (R/earth) para criar o modelo de preço.

6. Gerar as contribuições de valor e os resíduos dos imóveis (resíduo puro e resíduo/SF) a partir do modelo MARS.
7. Ordenar os imóveis por resíduo/SF.
8. Criar escores CQA para cada venda comparável.
9. Supondo que tenha sido produzida uma ordenação de qualidade dos imóveis que represente fielmente o apelo geral dos imóveis, encontrar a melhor posição do avaliando na ordenação e atribuir-lhe um escore CQA, estudando as tendências gerais do padrão na ordenação.
10. Atribuir o resíduo/SF correspondente ao imóvel avaliando como resíduo/SF · GLA (do avaliando).
11. **Somar a estimativa de preço do MARS e o resíduo atribuído para chegar à conclusão de valor do avaliando.**
12. Decompor o resíduo do avaliando em componentes descritivos e contribuições de valor.
13. Calcular as contribuições de valor e os ajustes dos comparáveis produzidos pelo modelo, além de todos os subtotais e dos preços de venda ajustados.
14. Os preços de venda ajustados serão iguais à conclusão final de valor do avaliando.
15. Selecionar os 6 a 12 imóveis mais semelhantes para o Sales Grid.
16. Decompor os montantes totais de resíduo dos comparáveis em contribuições de valor de componentes residuais, zerando o valor total do resíduo após concluída a decomposição.
17. Recalcular todos os subtotais e o preço de venda ajustado dos comparáveis do quadro. Conferir os valores.
18. Carregar o Sales Grid no laudo.

Observe que, com o RCA, a conclusão final de valor NÃO deriva diretamente dos valores dos comparáveis no Sales Grid. Com o método RCA, o Sales Grid torna-se apenas uma ferramenta para explicar por que imóveis que parecem semelhantes ao avaliando são vendidos por mais ou por menos do que a conclusão final de valor ou do que o valor apurado do imóvel avaliando. Isso precisa ser visto de outra maneira: a conclusão de valor do avaliando depende da estimativa de preço do modelo MARS e do resíduo estimado pelo engenheiro de avaliações. Mas essas duas quantidades dependem fortemente de todos os imóveis que entraram na regressão do R/earth, a qual criou a ordenação dos resíduos dos imóveis que determinou o escore CQA do imóvel avaliando e, assim, determinou seu resíduo por meio do posicionamento na ordenação feito pelo engenheiro de avaliações. Os ajustes no Sales Grid são, na verdade, determinados pelas restrições matemáticas impostas pelo RCA; embora a decomposição do resíduo fique de certo modo à mercê do VE, os totais das contribuições de valor da decomposição devem sempre somar o resíduo associado. Sabemos pela Demonstração 2 que, ainda que os valores efetivos dos componentes do resíduo possam variar conforme o juízo do VE, seus totais para um dado imóvel são completamente restringidos e, portanto, não alterarão o preço de venda ajustado do imóvel associado, independentemente de como o VE modifique a decomposição do resíduo.

5.3 Fluxo de Dados

Note que, neste artigo, apenas um fluxo de trabalho de alto nível é apresentado para o método RCA. Outros artigos aprofundarão as diversas etapas e subetapas da análise RCA.

5.3.1 Requisitos de Dados para o MARS

Será preciso especificar uma data inicial para as transações de venda e uma data final baseada na data-base da avaliação. Isso normalmente leva em conta o número estimado de variáveis de regressão e de interações que serão necessárias para cada grau de interação acima de 1.

5.3.2 Protocolos: Listagens de Vendas Fechadas

Protocolo 1.1: Determinação do Número de Listagens de Vendas Fechadas do MLS Necessárias

Este protocolo descreve como calcular o número mínimo recomendado de listagens de vendas fechadas a serem inseridas na regressão MARS. Você precisará fazer uma estimativa grosseira para começar, executar o MARS em rodadas de primeiro, segundo e possivelmente terceiro grau para gerar o modelo inicial e, em seguida, revisar o número de linhas conforme as regras a seguir para determinar se são necessárias mais linhas. Este protocolo pressupõe que as variáveis independentes sejam 100% independentes, o que raramente é o caso em imóveis; por exemplo, GLA e número de cômodos costumam ser altamente correlacionados, e é provável que a GLA também apresente correlação com a área do terreno. Isso nos leva ao cálculo dos “graus de liberdade (DOF)”. Cada variável verdadeiramente independente recebe um grau de liberdade. Mas, quando há dependências entre as chamadas variáveis independentes, o cálculo exato dos DOF exige uma análise que foge ao escopo deste artigo. Tenha apenas em mente que, em razão da provável colinearidade entre variáveis, as estimativas abaixo estão provavelmente superestimadas.

1. Para termos de primeira ordem (sem interações), a regra de ~ 15 linhas por variável funciona bem
2. Para cada interação de duas vias permitida, convém acrescentar aproximadamente 15-30 imóveis
3. Para interações de ordem superior (3 ou mais variáveis), cada uma deve ter mais de 30 imóveis.

Contudo, essa não é uma regra rígida. O número efetivamente necessário pode depender de:

- A relação sinal-ruído dos seus dados
- A complexidade das relações que você está tentando modelar
- A variabilidade da sua variável resposta
- A qualidade dos seus dados

Exemplo:

Assim, se você tiver:

- 10 variáveis principais
- 5 interações de duas vias permitidas
- 2 interações de três vias permitidas

Você poderia calcular seu número mínimo de vendas como:

$$\begin{aligned}
 N &= (10 \text{ variáveis} * 15 \text{ linhas/variável}) + \\
 &\quad (5 \text{ interações-de-duas-vias} * 22 \text{ linhas/interação-de-duas-vias}) + \\
 &\quad (2 \text{ interações-de-três-vias} * 30 \text{ linhas/interação-de-três-vias}) \\
 &= 320 \text{ linhas}
 \end{aligned}$$

Alternativamente, você pode fazer o seguinte: depois de executar o R/earth sobre um conjunto de dados, é possível usar o número de funções de base do modelo, mais um, como estimativa dos graus de liberdade e, então, multiplicar esse número por 20-30 para obter o número de registros de entrada necessários.

Protocolo 1.2: Os Períodos de Venda Devem Abranger Mudança Mínima no Gosto dos Compradores

Ao coletar dados históricos de imóveis, não há limitação temporal rígida se as mudanças nas preferências de mercado puderem ser capturadas por variáveis apropriadas. Considere estes pontos-chave para lidar com a evolução arquitetônica no seu mercado: mudanças nos estilos arquitetônicos dominantes (como a transição do California Ranch para o mediterrâneo) podem ser modeladas usando:

- Identificadores de loteamento
- Códigos de área do MLS
- Intervalos de data de construção

Contudo, se essas mudanças de estilo não puderem ser adequadamente capturadas por variáveis específicas, você enfrenta uma restrição importante: enquanto a Data da Venda é tipicamente usada para ajustar as condições de mercado, usá-la simultaneamente para controlar as preferências arquitetônicas confundiria esses dois efeitos distintos. Nesses casos, a abordagem preferível é:

1. Limitar a coleta de dados ao período posterior à introdução do estilo arquitetônico mais recente
2. Restringir a seleção de comparáveis a loteamentos que compartilhem o mesmo estilo arquitetônico

Isso assegura que sua análise mantenha consistência nas características dos imóveis, ao mesmo tempo em que permite o ajuste adequado às condições de mercado ao longo do tempo. Alternativamente, ao menos neste caso, supondo que você consiga extrair dos dados o estilo arquitetônico, seria possível considerar a interação entre essas duas variáveis. Contudo, tal extração nem sempre é possível.

Protocolo 1.3: Se Estiver Numa Área Com Poucas Vendas, Faça o Melhor Que Puder

Ao trabalhar com modelagem MARS para dados de vendas, são necessárias aproximadamente 15 transações para uma análise significativa. Se você tiver menos vendas (por exemplo, 5 imóveis), pode duplicar cada registro duas vezes para atingir o limiar de 15 imóveis. Considerações importantes sobre essa abordagem:

1. Use apenas regressão de primeiro grau, para reduzir o sobreajuste
2. Ignore ou dispense a validação cruzada, uma vez que o conjunto de dados é demasiadamente limitado
3. Embora o MARS provavelmente supere os pares comparativos ou os métodos de regressão linear padrão nesse cenário, esteja ciente de que duplicar dados afeta a validade das medidas de incerteza e da significância estatística

A melhor solução de longo prazo é ampliar seu conjunto de dados com vendas reais adicionais. Contudo, se você tiver de prosseguir com dados limitados, a duplicação de dados oferece uma solução temporária viável, reconhecidas as suas limitações estatísticas.

5.4 Estrutura de Pastas do Projeto

Convém discutir a estrutura de pastas do projeto. É uma boa ideia manter pastas separadas para Análises de Mercado e Avaliações de Imóvel Único. Ambas as pastas têm estrutura de subpastas semelhante:

```

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_Stage2.xlsx
│   │       │       ├── MLS_Stage3.xlsx
│   │       │       ├── MLS_Stage4.xlsx
│   │       │       ├── MLS_Stage5.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
│   │       │           └── ...

```

O processamento da Etapa 1 criará a pasta da versão com suas subpastas. Ele copiará o `MLS_Original.xlsx`, ainda que essa planilha *devesse* nunca mudar e refletir exatamente o que foi baixado do MLS em determinada data e hora; contudo, existe a possibilidade, em um projeto mais longo, de que seja necessário um novo download do MLS. Em seguida, esta etapa fará uma cópia da pasta de trabalho `MLS.xlsx` expandida, com seus parâmetros de configuração, para a pasta da versão. Pode haver muitas pastas de versão, cada uma criada para uma execução completa do MARS.

Depois, as etapas seguintes gerarão acréscimos de colunas, ordenações e outras alterações na

pasta de trabalho e gravarão uma versão atualizada como `MLS_Stage(N)` na pasta de dados do projeto, quando a avaliação estiver concluída. Reiterando, a pasta raiz `MLS.xlsx` está em constante mudança, e assim a cópia da versão é uma mutação contínua da pasta de trabalho `MLS` original, com exclusões e inclusões de variáveis, além de outras alterações. O código R ou Python de processamento muito provavelmente será executado com entrada a partir da pasta de trabalho `Data MLS.xlsx`. Assim, sempre temos uma cópia de todos os dados por trás de cada versão da execução do RCA.

São criadas tantas versões quantas forem necessárias até que não seja mais possível obter melhorias, momento em que a melhor versão é selecionada como versão final. Pode-se alterar o nome da pasta acrescentando “_Final” ao seu fim, de modo que fique evidente ao percorrer o diretório. Algumas versões mais antigas, com modelos interessantes ou concorrentes, devem geralmente ser preservadas. Cada pasta de versão contém uma série de arquivos de texto, diagramas e gráficos que podem ser incorporados a um laudo. A pasta de trabalho final `MLS_Stage5.xlsx` contém tantas planilhas quantas forem necessárias ao laudo. Em particular, ela pode ter uma planilha para transferência a um Sales Grid da Fannie Mae. Também conterá necessariamente os cálculos ou a decomposição por trás das agregações. Esta é potencialmente a parte mais demorada da abordagem RCA — o esforço para encontrar o modelo perfeito em uma região de mercado difícil.

Tudo o que é necessário para recriar os quadros de homogeneização precisa ser copiado para a pasta da versão, inclusive o código R ou Python utilizado — caso o código sofra alterações frequentes.

5.4.1 Protocolos: Múltiplas Execuções do MARS (earth)

Protocolo 2.1: Execuções Sobreajustadas Para Estudar Cirurgicamente os Fatores de Preço da Região de Mercado

Durante as execuções iniciais do earth, o sobreajuste deve ser usado estrategicamente para descobrir fatores de preço subjacentes. Por exemplo, uma região de mercado representada por 800 vendas pode conter apenas 3 imóveis com oficina, e nesses casos isolados a oficina pode ter grande impacto no preço final de venda. Uma execução apenas com passo direto (forward pass) produzirá um modelo sobreajustado que mostra a contribuição de valor aproximada dessas oficinas. Contudo, como há apenas 3 vendas desse tipo em 800, os termos de oficina serão removidos no passo reverso e, certamente, sob validação cruzada. O valor que o passo apenas direto atribui às oficinas pode ser anotado e transferido para a análise de variáveis latentes, a ser usado para qualquer imóvel com oficina que apareça no Sales Grid. Esses modelos sobreajustados também fornecem estimativas detalhadas para muitas outras características com dados escassos, que seriam eliminadas pela poda do passo reverso ou por execuções validadas cruzadamente.

Salve cada uma dessas execuções exploratórias em sua própria pasta de versão: as contribuições de valor registradas são evidências que o VE vai querer citar mais tarde, ao decompor os resíduos. Compreenda, evidentemente, que o VE deve, em última instância, apoiar-se na poda e na validação cruzada para tornar o modelo final robusto diante de imóveis que a regressão nunca viu — como um imóvel avaliando que apareça posteriormente. Execuções sobreajustadas são instrumentos de estudo; jamais são o modelo final.

Protocolo 2.2: Múltiplas Execuções Validadas Cruzadamente Sob Diferentes Parâmetros de `earth()`

Regiões de mercado diferem em estrutura e tamanho, e o VE normalmente precisa experimentar uma série de execuções: para desenvolver a compreensão de quais variáveis mais impactam o preço; para julgar se o conjunto existente de variáveis medidas cumpre adequadamente a tarefa de estimar valor, ou se novos dados precisam ser acrescentados (distância à água, altimetria, distância a centros comerciais e assim por diante); e para descobrir novos tipos de interação, ao mesmo tempo em que se desautorizam outros por ineficientes ou enganosos. Ao final, o VE quer um bom R^2 nos dados de treino e o melhor CVR2 possível contra os dados de teste. Chegar lá pode exigir coletar novos dados, melhorar a exatidão dos dados existentes ou ajustar os parâmetros de `earth()` ao tamanho e ao caráter da região de mercado (por exemplo, aos graus de liberdade disponíveis — ver o Protocolo 1.1).

É precisamente por isso que a estrutura de projeto acima comporta muitas pastas de versão de execução: cada modelo candidato é preservado com seus parâmetros e seu R^2 /CVR2, de modo que a trajetória ao longo das execuções possa ser comparada. Uma versão cujo R^2 sobe enquanto o CVR2 estagna ou cai está ajustando ruído, e não sinal de mercado, e é o histórico de versões que torna isso visível.

Protocolo 2.3: Expansão em Prolog de Observações e Colunas de Características Descritivas

Embora esse recurso ainda esteja em fase um tanto experimental, o `earthUI` agora vem com um motor Prolog embutido (`vProlog`) e usa Definite Clause Grammars (DCGs, gramáticas de cláusulas definidas) para analisar as colunas de texto livre do MLS — Public Remarks (observações públicas) acima de tudo, mas também as observações do corretor e listas de características separadas por vírgula, como características dos cômodos ou benfeitorias existentes. A expansão converte expressões como “owned solar” (solar próprio), “3 car garage” (garagem para 3 carros) ou “40x50 workshop” (oficina de 40x50) em colunas de características comuns, que podem ser oferecidas à regressão, e as características com dados escassos assim recuperadas alimentam diretamente a análise de variáveis latentes (ver a subseção adiante sobre notas de estado de conservação e padrão construtivo como regressores aproximados; o recurso em si está documentado no artigo companheiro sobre o `earthUI` nesta edição).

A expansão de texto multiplica o conjunto de preditores candidatos, e isso torna necessárias múltiplas execuções `earth/RCA` para teste: as colunas expandidas devem ser testadas dentro e fora do modelo, consolidadas em contagens derivadas ou booleanos quando forem colineares, e conferidas contra o CVR2 para confirmar que carregam sinal de mercado, e não ruído. Trate cada experimento desses como sua própria pasta de versão de execução, exatamente como nos Protocolos 2.1 e 2.2. E, uma vez que a própria gramática de análise está em desenvolvimento, o VE deve revisar as colunas extraídas em confronto com o texto de origem antes de confiar nelas em um modelo.

5.5 A Pasta de Trabalho Excel Central

Como a maioria dos serviços de MLS baixa os dados de imóveis em planilhas CSV do Excel, o Excel é provavelmente a forma mais conveniente de armazenar dados e diversos parâmetros de configuração. Há uma pasta de trabalho Excel central para cada projeto, chamada `MLS.xlsx`, que possui uma planilha chamada “`MLSData`”, para armazenar as listagens de vendas fechadas de imóveis baixadas do MLS, além de várias planilhas de configuração e parâmetros:

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

Os dados originais são armazenados em uma pasta de trabalho com nome do tipo `MLS_Original.xlsx`. Deve-se fazer uma cópia com nome diferente, digamos “`MLS.xlsx`”, que será carregada no programa RCA em R ou Python. Ambas as pastas de trabalho ficam armazenadas na pasta `Data`. Agora, a pasta de trabalho `Data/MLS.xlsx` provavelmente passará por muitas iterações de execuções, testando diferentes combinações de variáveis, parâmetros do MARS e alterações em outras configurações. Cada iteração cria uma nova pasta de entrada/saída nomeada com a data e a hora, no formato `VER_YYYYMMDDHHMMSS`. Cada pasta terá uma cópia da pasta de trabalho de entrada `MLS.xlsx` e conterá as pastas de trabalho Excel geradas para cada uma das quatro etapas de processamento (Etapas 2–5), bem como os documentos de texto, diagramas e gráficos de saída. Note que a Etapa 5, a etapa final, pode conter diversas outras planilhas para o laudo, dependendo do tipo de cliente e do escopo do trabalho.

5.6 Backups

Ao gravar planilhas Excel novas ou modificadas nas pastas de versão do projeto, também podem ser feitos backups em bancos de dados SQLite ou PostgreSQL na pasta do projeto. Os arquivos de banco de dados têm menor probabilidade de se corromper involuntariamente em períodos subsequentes, por uso em revisão ou consulta.

6 Notação da Pasta de Trabalho

6.1 Dimensões Descritivas

Há aproximadamente seis dimensões descritivas para uma pasta de trabalho “M”:

1. Avaliações “a”
2. Versões de Execução “r”
3. Etapas do Fluxo de Trabalho “w”
4. Planilhas “s”
5. Listagem de Imóveis “p”
6. Conjunto de Variáveis “v”

Assim, uma célula individual de uma determinada pasta de trabalho M pode ser descrita como:

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

(Nota: “*” significa algum índice válido)

ou:

$$\begin{matrix} r_k \\ w_m \end{matrix} \begin{matrix} a_g \\ W \\ s_h \end{matrix} \begin{matrix} p_i \\ v_j \end{matrix} \quad \begin{matrix} i = 1, \dots, n_p \\ j = 1, \dots, n_v \\ k = 1, \dots, n_r \\ m = 1, \dots, n_m \\ g = 1, \dots, n_g \\ h = 1, \dots, n_h \end{matrix} \quad (2)$$

onde:

- n_p = n° de linhas (depende da planilha)
- n_v = n° de colunas (depende da planilha)
- n_r = n° de versões de execução
- n_m = n° de etapas do fluxo de trabalho
- n_g = n° de projetos de análise
- n_h = n° de planilhas da pasta de trabalho

Entenda que “W” faz referência à pasta de trabalho Excel inteira, com seis planilhas de dados centrais, cada uma com sua própria tabela, as quais, ao se executar o programa R ou Python associado, são carregadas em data frames internos separados. Tanto o R quanto o Python possuem funções internas capazes de tomar um nome e descobrir a que linha ou coluna do data frame ele pertence. Por exemplo, se no subscrito v_j tivermos $v = (\text{“GLA”}, \text{“LotSize”}, \text{“BathRms”}, \text{“BedRms”}, \dots)$ e $j=0$, então $v_j = \text{“GLA”}$. Tanto o R quanto o Python conseguirão internamente resolver “GLA” para a coluna em que está armazenada, que, por exemplo, pode ser a coluna 15 do data frame associado. Assim, neste artigo indexamos nossos vetores de nomes, os quais passam um nome ao programa em execução, que reindexa esse nome no índice real de coluna ou linha do data frame interno. É um pouco complicado. Mas, para ser preciso, isso é necessário e bastante conveniente. Trata-se de uma “separação de responsabilidades” entre o papel do analista e o papel do desenvolvedor; embora, na prática, o Valuation Engineer (VE, engenheiro de avaliações) possa desempenhar ambos os papéis. Note que um “data frame” do R é chamado de “dataframe” em Python.

Note também que a planilha de parâmetros de execução do MARS não é uma dimensão de W, já que W diz respeito à versão dos dados que entram no MARS, separadamente dos parâmetros que determinam como o MARS/earth trata esses dados.

Em prol da clareza, apresenta-se uma notação multidimensional complexa para a pasta de trabalho, que se relaciona mais diretamente aos data frames internos do R ou do Python nos quais as planilhas da pasta de dados são carregadas. É uma notação “pesada”. Mas, quando chegarmos às demonstrações e assim por diante, descartaremos boa parte dessa bagagem ao fazer suposições sobre aquilo com que estamos lidando. Ainda assim, essa notação ajuda a obter, desde o início, uma visão geral melhor do fluxo de trabalho e dos dados.

Começemos pela notação do conjunto de pastas de trabalho que iremos processar.

6.1.1 A Dimensão do Projeto (“a”)

Uma vez aprovado o Escopo do Trabalho (SOW), é preciso montar uma pasta de projeto para armazenar dados de entrada e de saída, incluindo parâmetros de configuração, gráficos, diagramas, planilhas e trechos de laudo. O VE deve identificar todos os projetos com um ID sequencial simples, como um inteiro, de modo que, se um laudo vier a sumir, isso fique evidente. Se projetos forem cancelados, deve haver um meio consistente de indicar que o projeto foi cancelado dentro da respectiva pasta de projeto, mas ela jamais deve ser excluída. Assumimos que os IDs de projeto começam em 0 e são incrementados de 1, embora haja outras possibilidades. Assumimos ainda que todos os IDs de projeto válidos são armazenados em um conjunto ordenado “a” criado com o programa de processamento.

Os dados do projeto de avaliação ficam na planilha de projeto da pasta de trabalho, mas também poderiam ser armazenados em algum banco de dados SQL central,

Eis os seguintes campos de projeto sugeridos:

1. ID local: um ID inteiro curto, gerado sequencialmente para a empresa. Faria uma excelente chave primária para armazenamento — e poderia ser gerado automaticamente pelo banco de dados a cada projeto sucessivo.
2. GUID global: se quiser, você pode criar um GUID (Globally Universal ID) usando

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

3. Localização do imóvel (Análise de Mercado) ou endereço
4. Data-base da avaliação
5. Data do pedido
6. Data da vistoria
7. Data do laudo original
8. Data do primeiro laudo atualizado
9. Data do segundo laudo atualizado
10. Data do terceiro laudo atualizado
11. Nome do cliente
12. Endereço do cliente
13. Nome do credor
14. Endereço do credor
15. Nome do proprietário
16. Endereço do proprietário
17.

Exemplo de lista de IDs de projeto:

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

6.1.2 Dimensão da Versão de Execução (“r”)

Em segundo lugar no fluxo de trabalho, executamos a regressão MARS sobre a entrada. Isso provavelmente envolverá de várias a muitas execuções, à medida que alteramos diversos parâmetros para aprimorar o modelo. Cada execução criará uma nova pasta de versão para a saída, com nome baseado na data e na hora da execução. Por exemplo,

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

```
latest_version : r1 = "ver20241210_083124"
```

Versões promissoras, com modelos e saídas competitivos, devem ser mantidas até que se decida por uma versão final, ou talvez por mais tempo se tiverem alguma utilidade. Talvez você queira modificar a pasta da versão para indicar que é a versão final, colocando “Final” ao fim do nome.

6.1.3 Etapas do Fluxo de Trabalho (“w”)

O fluxo de trabalho do RCA é separado em geralmente cinco etapas distintas, listadas a seguir. Cada etapa deve ser concluída com sucesso antes do início da etapa seguinte. O VE deve poder interromper o processamento ao fim de uma etapa para revisar os resultados antes de prosseguir. Isso é necessário porque as regiões de mercado são bastante complexas e frequentemente bem singulares. Podem surgir problemas incomuns que exijam uma nova abordagem. O VE precisa poder entrar no código R ou Python para fazer as alterações necessárias e até mesmo tentar novas abordagens. E, nesses casos, ele pode querer experimentar e tentar abordagens diferentes, exigindo múltiplas execuções da mesma etapa em que está trabalhando. Ainda assim, o VE também deve ter liberdade para executar todas as etapas com um único comando — com algumas ressalvas (por exemplo, a Agregação de 3º Grau já deve ter sido determinada em uma execução anterior da Etapa 2). Isso significa que, ao concluir cada etapa, ela deve armazenar todos os seus dados internos em arquivos ou bancos de dados, de modo que a etapa seguinte possa ser iniciada como se a etapa anterior ainda estivesse em memória, recarregando os dados da etapa anterior para a memória.

Eis as etapas:

```
w = ("stage1", -Configuração
     "stage2", -Processamento MARS
     "stage3", -Sales Grid final
     "stage4", -Trechos de laudo
     "stage5") -Revisão/Limpeza
```

```
stage2 : w2 = "stage2"
```

6.1.4 Planilhas da Pasta de Trabalho Excel (“s”)

Na Pasta de Trabalho Excel Central do RCA, temos as seguintes planilhas, cada uma mapeada em um data frame interno:

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

A razão de as agregações de 2º grau ficarem em planilha separada é que elas podem ser definidas mais facilmente com uma tabela simétrica que mostre todos os pares possíveis de interações, permitindo especificar uma variável de agregação à qual as contribuições de valor e os ajustes associados a cada par serão agregados (somados). Não poderíamos fazer isso com agregações de 3º grau, porque isso exigiria uma tabela simétrica tridimensional, o que não é possível no Excel. Além disso, uma tabela 20x20x20 teria 8,000 células a manejar. É mais fácil simplesmente listar quaisquer nomes de termos de terceiro grau na tabela junto com as agregações de 1º grau, por exemplo:

Variável	Agregar Para
AboveGradeFinArea	GLA
BelowGradeFinArea	BGLA
UnfinArea	UBA
LotSize	LotSize
...	...
GLA__LotSize	GLA
GLA__DateOfSale	GLA
BathRms__DateOfSale	GLA
Lat__Long	Location
Lat__Long__GLA	Location
...	...

6.1.5 Listagens de Imóveis (“p”) — linhas do MLSData

As listagens de imóveis formam as linhas da planilha MLSData. Elas não aparecem nas demais planilhas até a Etapa 4, na qual os trechos finais do laudo são criados com a seleção final de imóveis comparáveis. Eis uma lista exemplificativa de IDs de imóveis:

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

6.1.6 Variáveis do MLS (“v”) — colunas do MLSData

Mais será dito adiante sobre o tratamento das variáveis. Basta afirmar, neste ponto, que a planilha MLSData tipicamente tem de 30 a 40 variáveis de vários tipos, mas normalmente apenas de 10 a 20 serão selecionadas como variáveis independentes para a regressão MARS. Essas variáveis também aparecem em outras planilhas, ainda que tipicamente como uma lista de todas

as variáveis possíveis que podem ser submetidas ao MARS. Um exemplo de lista de variáveis é:

$$v = (\text{"GLA"}, \text{"SalePrice"}, \text{"BedRms"}, \text{"BathRms"}, \text{"GarageSF"}, \text{"..."})$$

6.1.7 Exemplo Final de Notação

Vamos reunir a notação acima em um exemplo final. Usando as listas acima, seja

$$\begin{matrix} r_1 & & a_{141} \\ w_1 & \overline{W} & p_2 \\ & s_0 & v_0 \end{matrix} \quad (3)$$

Note que aqui estamos usando indexação em 0, de modo que se assume que o primeiro elemento das listas tem índice 0. O Python também usa indexação em 0, mas o R usa indexação em 1.

Assim, o acima significa que estamos falando do valor da célula para:

- A variável v_0 , ou coluna “GLA”
- Do imóvel p_2 , ou ID “MLS823542”
- Da planilha s_0 , ou “MLSDData”
- Para a etapa de fluxo de trabalho w_1 , ou “stage2”
- Da versão r_1 , ou “ver20241209_151130”
- Do projeto a_{141} , ou “142”

A notação acima também pode então ser escrita como:

$$\begin{matrix} & & \text{"142"} \\ \text{"ver20241209_151130"} & \overline{W} & \text{"MLS823542"} \\ \text{"stage2"} & \text{"MLSDData"} & \text{"GLA"} \end{matrix} \quad (4)$$

6.1.8 Justificativa

“Por que não usamos simplesmente índices ou números na notação?” Temos dois requisitos para nossa notação:

1. Subconjuntos: por exemplo, o analista frequentemente seleciona subconjuntos de uma lista, como subconjuntos de variáveis. Pode haver de 35 a 40 variáveis na planilha MLSDData, mas apenas de 10 a 20 delas são usadas em uma dada execução. De uma execução para outra, o VE frequentemente muda as variáveis submetidas ao MARS/earth.
2. Iteração: por exemplo, $i = 1, 2, \dots, n_1$
Nossos algoritmos, protocolos e demonstrações normalmente precisam poder iterar por todos os itens de uma lista ordenada, do primeiro ao último — ou selecionar algum item em particular.

Assim, precisamos de uma letra para representar o subconjunto ou a lista de itens, e essa letra deve ter outro subscrito que possamos usar para iterar sobre os itens da lista, fornecendo-nos as cadeias de caracteres usadas para indexar as linhas ou colunas do data frame.

6.1.9 Matrizes?

E quanto às matrizes da álgebra linear? Matrizes só funcionam com números, inteiros ou reais. Matrizes como estrutura de dados são suportadas tanto em R quanto em Python. Contudo, muitas das variáveis com que trabalhamos nos dados de MLS são chamadas de “strings”, ou dados de caracteres. Mesmo que algumas variáveis sejam apenas números, muitas vezes não são números reais, e sim apenas nomes. Por exemplo, os códigos de área do MLS para uma cidade podem ser (“660”, “661”, “662”, “663”, ..., “672”). A regressão do earth permite que variáveis sejam especificadas como variáveis “factor” (fator), caso em que os números são tratados simplesmente como nomes, sem qualquer sequência. Nesses casos, o modelo produzido apenas fornecerá uma contribuição de valor para cada número, em vez de uma função que percorra a sequência. Outros exemplos incluem “C1”, ..., “C6” para Condition (estado de conservação) e “Q1”, ..., “Q6” para quality (padrão construtivo).

6.1.10 Primeira Simplificação

A notação acima para a pasta de trabalho Excel central do projeto é boa para amarrar as coisas, mas pesada demais para carregar ao explicar algoritmos ou demonstrações matemáticas. Quando estivermos no trabalho metódico, podemos assumir que ele se dá dentro de um dado projeto, versão, etapa e planilha. Apenas dizemos com o que estamos trabalhando, se isso já não estiver implícito. Tudo de que realmente precisamos do acima, uma vez indicado com o que estamos trabalhando, é:

$$M_{s_h}^{p_i v_j} \quad \begin{matrix} i = 1, 2, \dots, n_p \\ j = 1, 2, \dots, n_v \\ h = 1, 2, \dots, n_s \end{matrix} \quad (5)$$

onde: n_p = n° de imóveis
 n_v = n° de variáveis de entrada
 n_s = n° de planilhas de entrada

6.1.11 Segunda Simplificação

É mais conveniente renomear as planilhas segundo os tipos específicos de data frames que elas suportam, em vez de usar índices. Para cada planilha, as linhas e as colunas têm funcionalidades diferentes. Assim, o conjunto W de data frames é decomposto nos data frames M, V, A, I e C, da seguinte forma:

$$\diamond \text{ Dados do MLS} \quad M_{v_j}^{p_i} = W_0^{p_i v_j} \quad \begin{matrix} i = 0, 1, \dots, n_p \\ j = 0, 1, \dots, n_v \end{matrix} \quad (6)$$

onde: n_p = n° de imóveis
 n_v = n° de variáveis de entrada

Nota: o avaliando vem, por padrão, sempre em primeiro lugar nas listas de imóveis

Na planilha Excel “MLSData” e no data frame associado, a primeira linha está sempre reservada ao imóvel avaliando. Ela nunca é ordenada junto com os demais imóveis do MLS abaixo dela. Ela nunca entra na regressão MARS como dado independente; entra, isto sim, como dado-alvo, ou seja, como variável dependente. O mesmo vale para as listas de imóveis.

Em todas as listas de imóveis, o primeiro elemento, de índice 0, é sempre presumido como o ID do imóvel avaliando, salvo indicação em contrário.

Mesmo que a tarefa seja criar um modelo para uma região de mercado sem qualquer imóvel avaliando, a primeira linha é, ainda assim, carregada com alguns dados fictícios de imóvel. Isso torna muito mais conveniente escrever programas, especificar algoritmos ou fazer demonstrações. Assim, se estivermos usando c para representar a lista de imóveis, c_0 é o imóvel avaliando e c_i , $i = 1, \dots, n_p$, são as listagens comparáveis que serão entrada da regressão MARS/earth.

◇ **Especificações das Variáveis de Regressão**

$$V_{s_j}^{v_i} = W_1^{v_i}_{s_j} \quad \begin{matrix} i = 1, 2, \dots, n_v \\ j = 1, 2, \dots, n_s \end{matrix} \quad (7)$$

onde: $n_v = n^\circ$ de variáveis disponíveis
 $n_s = n^\circ$ de especificações

Note que $n_s \leq n_v$, pois normalmente selecionamos um subconjunto das variáveis disponíveis para entrada na regressão.

◇ **Cálculos**

$$C_{c_j}^{v_i} = W_2^{v_i}_{c_j} \quad \begin{matrix} i = 1, 2, \dots, n_v \\ j = 1, 2, \dots, n_c \end{matrix} \quad (8)$$

onde: $n_v = n^\circ$ de variáveis
 $n_c = n^\circ$ de especificações de cálculo

◇ **Interações**

$$I_{t_j}^{f_i} = W_3^{f_i}_{t_j} \quad \begin{matrix} i = 1, 2, \dots, n_f \\ j = 1, 2, \dots, n_t \end{matrix} \quad (9)$$

onde: $n_f = n^\circ$ de variáveis “de”
 $n_t = n^\circ$ de variáveis “para”

Note que o data frame I é um data frame simétrico, com as variáveis listadas em ordem alfabética nas linhas “from” (de) e nas colunas “to” (para). As células são marcadas com “1” para indicar que é permitida uma interação entre o par de variáveis associado. Caso contrário, “0” indica que nenhuma interação é permitida. Para uma interação de segundo grau, isso é relativamente simples, com apenas uma interação possível. Com uma regressão de terceiro grau, temos três interações de pares entre as três variáveis, e cada par de interações precisa ser permitido para que se gere o termo (ou variável) de interação de três variáveis.

◇ **Agregações**

$$A_{v_j}^{v_i} = W_4^{v_i}_{v_j} \quad \begin{matrix} i = 1, 2, \dots, n_v \\ j = 1, 2, \dots, n_v \end{matrix} \quad (10)$$

onde: $n_v = n^\circ$ de variáveis

7 O Sales Grid do RCA

Tenha em mente que o protocolo RCA exige um programa em R ou Python capaz de acessar o R/earth e executar todas as etapas detalhadas. É intensivo em dados, e o trabalho não poderia, de modo algum, ser feito manualmente. Ao mesmo tempo, supondo que o programa esteja livre de defeitos, é preciso um programa de software para assegurar que o trabalho exigido seja feito com precisão e sem erro humano.

É também necessário que os VEs que realizam esse trabalho consigam modificar tais programas conforme necessário, sobretudo quando estão lidando com novas regiões de mercado, bairros, imóveis imprevistos ou outras exigências e problemas inéditos.

No processamento do RCA temos três tipos de Sales Grid:

1. O Sales Grid de Processamento do RCA, aqui designado como “MLSDData”, que é um data frame muito grande, com a respectiva planilha, e que pode conter muitos detalhes intermediários que acabam passando por um processo de agregação até se chegar ao Sales Grid do RCA mencionado adiante. Um exemplo seria grande demais para exibir aqui. Muito provavelmente os engenheiros de avaliações ajustarão essa planilha às suas preferências. Pretendo escrever um artigo separado sobre o desenvolvimento de um quadro desses em alguma data futura.
2. O Sales Grid Final do RCA, usado em laudos não destinados às GSE, que se parece com a [Table 2](#). Ele imita, por exemplo, o formulário da Fannie Mae, o suficiente para oferecer ao leitor alguma base de familiaridade. Contudo, observe que ele traz mais colunas por venda comparável do que o quadro da Fannie Mae: cada comparável tem uma coluna de “Contribuição de valor” ao lado de suas colunas de “Medida” e “Ajuste”, e o avaliando também tem uma — colunas de que o formulário da Fannie Mae carece. Um formato mais simples é também o da [Table 1](#).
3. O Sales Grid de Upload para as GSE, usado para carregar o quadro em softwares como o Alamode. Note, contudo, que nem todos os pacotes de software de formulários de avaliação oferecem a possibilidade de carregar dados de quadro de homogeneização a partir de planilha. Por necessidade, tais quadros não podem incluir coluna separada para contribuições de valor, das quais as GSE, como a Fannie Mae, não têm compreensão alguma. A estratégia é lhes dar o formulário exigido, mas respaldá-lo com um Sales Grid do RCA, bem como com outra documentação de apoio, tal como detalhes sobre eventuais agregações de variáveis.

Os Sales Grids de Processamento do RCA não se parecem com os Sales Grids da Fannie Mae, mas são, grosso modo, uma transposta do Sales Grid da Fannie Mae, com linhas e colunas trocadas: cada linha contém dados de algum imóvel, e cada coluna contém dados de alguma variável do imóvel. Pode haver facilmente centenas ou milhares de linhas de imóveis e 200 ou mais colunas. A primeira linha é reservada aos nomes das variáveis, e a segunda a um imóvel avaliando, mesmo que estejamos fazendo apenas uma análise de mercado. Quando a planilha é lida para dentro de um data frame, a primeira linha da planilha torna-se o atributo de nomes, e a segunda linha, do imóvel avaliando, torna-se a primeira linha do data frame.

As colunas do Sales Grid do RCA incluem colunas de identificação e, em seguida, variáveis primárias, externas e sintéticas, combinadas e alimentadas como variáveis independentes ao MARS. Dessas variáveis, o MARS devolve aquelas que julga ter impacto significativo sobre o Preço de Venda, também conhecido como variável dependente ou alvo. Na regressão de segundo grau, podemos obter pares dessas variáveis atuando como novas variáveis. Com a regressão de terceiro grau, podemos também obter triplas de variáveis atuando como as

chamadas variáveis independentes, embora sejam, evidentemente, altamente correlacionadas com as variáveis subjacentes.

Tabela 1: Planilha RCA de exemplo: disposição vertical

Avaliando			Comp 1			Comp 2		
Medida	Contribuição de valor		Medida	Contribuição de valor	Ajuste	Medida	Contribuição de valor	Ajuste
Preço de venda			\$1,150,000			\$1,500,000		
Valor-base	\$500,000			\$500,000	\$0.00		\$500,000	\$0
Características medidas								
Área habitável	1,530	\$463,500	1,350	\$382,500	\$81,000	1,700	\$540,000	-\$76,500
Área do terreno	6,000	\$200,000	6,500	\$220,000	-\$20,000	7,000	\$240,000	-\$40,000
Idade	53	\$63,600	57	\$68,400	-\$4,800	51	\$61,200	\$2,400
Banheiros	2	\$15,000	3	\$30,000	-\$15,000	3	\$30,000	-\$15,000
Dormitórios	3	\$6,000	4	\$9,000	-\$3,000	4	\$9,000	-\$3,000
Total medido		\$1,248,100.00		\$1,209,900.00	\$38,200		\$1,380,200.00	-\$132,100
Características residuais								
Resíduo	CQA 5.8	\$30,000	CQA 2.5	-\$59,900	\$89,900	CQA 8.2	\$119,800	-\$89,800
Decomposição do resíduo			Decomposição do resíduo			Decomposição do resíduo		
Projeto		7,000		-\$15,000	\$22,000		\$35,000	-\$28,000
Estado de conservação		5,000		-\$8,000	\$13,000		\$20,000	-\$15,000
Padrão construtivo		8,000		-\$10,000	\$18,000		\$30,000	-\$22,000
Utilidade funcional		5,000		-\$12,000	\$17,000		\$15,000	-\$10,000
Comodidades		5,000		-\$14,900	\$19,900		\$19,800	-\$14,800
Subtotal		30,000		-\$59,900	\$89,900		119,800	-\$89,800
Total			Total			Total		
Contribuição de valor total		\$1,278,100.00		\$1,150,000.00			\$1,500,000.00	
Preço de venda ajustado					\$1,278,100.00			\$1,278,100.00

8 Variáveis

Do lado da entrada da execução do MARS, temos apenas variáveis simples, que são:

1. Características do imóvel carregadas da planilha do MLS.
2. Variáveis derivadas, criadas pelo VE para refinar as variáveis do MLS ou acrescentar transformações de variáveis existentes. Por exemplo, o VE pode preferir transformar a Data da Venda em “DaysOffMarket”, que equivale à data-base da avaliação menos a Data da Venda.
3. O VE pode importar dados de fontes de terceiros, como taxas de juros hipotecários. Ou, se dados de latitude, longitude ou altimetria estiverem faltando nos dados do MLS, ele pode importá-los de outras fontes. Essas variáveis podem ser chamadas de “variáveis externas”.

Do lado da saída do MARS, trabalhamos com um modelo de preço que é uma combinação de termos que:

1. Contêm apenas uma das variáveis de entrada e fornecem a equação da contribuição de valor dessa variável. São chamados de termos (ou variáveis de saída) de primeiro grau. O próprio termo é tratado como um valor de contribuição de valor e recebe um nome.
2. Contêm duas das variáveis de entrada e constituem a equação da contribuição de valor da interação entre essas duas variáveis. Esses termos podem ser chamados de termos (ou variáveis de saída) de segundo grau.
3. Contêm três das variáveis de entrada e constituem a equação da contribuição de valor da interação entre as três variáveis. Neste artigo, esses termos podem ser chamados de termos (ou variáveis de saída) de terceiro grau.

As variáveis de saída do MARS são, então, contribuições de valor, inclusive a contribuição de valor residual. Elas, por sua vez, são tratadas como variáveis nos cálculos do Sales Grid, tais como o cálculo de ajustes ou a criação de gráficos.

Podemos ter diferentes tipos de variáveis, que se enquadram em duas ontologias básicas: (1) Origem da Variável e (2) Aplicação da Variável, isto é, não residual versus residual

8.1 Ontologia da Origem das Variáveis

1. Variáveis de primeiro grau
 - 1.1 Variáveis de regressão do MLS
 - 1.2 Variáveis de regressão externas
 - 1.3 Variáveis de regressão derivadas

2. Variáveis fator (factor):

Para lidar com variáveis fator, é preciso uma função capaz de identificá-las. Uma vez que elas são especificadas na pasta de trabalho Excel de configuração do projeto, cria-se uma função que retorna verdadeiro se a variável passada a ela tiver sido classificada como do tipo fator.

3. Variáveis de interação: são criadas pelo MARS ao gerar o modelo, se o grau estiver definido como 2 ou mais. Contudo, 3 variáveis interagentes é o máximo que se poderia recomendar, em razão da dificuldade de interpretação, da probabilidade de sobreajuste e do impacto sobre os graus de liberdade, com o consequente aumento do número de registros de imóveis necessários para assegurar um modelo robusto.
 - 3.1 Variáveis de segundo grau: são comuns, e bons exemplos seriam GLA:BathRms, GLA:LotSize, DateOfSale:BathRms, Latitude _Longitude.
 - 3.2 Variáveis de terceiro grau: interações menos comuns, que sugerem cautela e possível sobreajuste.
4. Variáveis de decomposição do resíduo

8.2 Ontologia da Aplicação das Variáveis

1. Contribuições de valor geradas pelo MARS
 - 1.1 Variáveis de primeiro grau
 - 1.1.1. Variáveis de regressão do MLS
 - 1.1.2. Variáveis de regressão externas
 - 1.1.3. Variáveis de regressão derivadas
 - 1.2 Variáveis de interação
 - 1.2.1. Variáveis de segundo grau
 - 1.2.2. Variáveis de terceiro grau
 - 1.3 Variáveis de decomposição do resíduo
 - 1.4 Variáveis de agregação
2. Variáveis e valores gerados pelo engenheiro
 - 2.1 Variáveis de decomposição do resíduo
 - 2.1.1. Projeto/Estética
 - 2.1.2. Paisagismo
 - 2.1.3. Utilidade funcional
 - 2.1.4. Vista
 - 2.1.5. Estado de conservação
 - 2.1.6. Padrão construtivo
 - 2.1.7. ...

Esses grupos de variáveis receberão sua notação e serão descritos com mais detalhe na próxima seção

Protocolo 3.1: Variáveis Fator

Variáveis fator, ou simplesmente fatores, podem ocorrer em todos os termos, inclusive nas interações. Elas são importantes porque precisam de tratamento especial para a regressão MARS e para a construção de gráficos.

Ainda que variáveis com valores de caractere sejam necessariamente fatores, variáveis numéricas precisam ser sinalizadas como variáveis fator para que o MARS as reconheça como tal. Por exemplo, as áreas do MLS são muitas vezes identificadas por número. Em particular, as áreas de bairro na San Francisco Bay Area costumam ser identificadas por inteiros únicos. Se esses valores forem passados ao MARS como fatores, o MARS tentará descobrir se determinadas áreas têm um valor médio de contribuição às vendas de imóveis dentro de seus limites, sem levar em conta qualquer relação com outras áreas. Em contraste, com variáveis normais, ele suporá que há algum tipo de continuidade nas contribuições de valor conforme o valor da variável muda. Essa é uma distinção importante. Além disso, essas variáveis fator podem aparecer em termos do modelo junto com variáveis que não são fatores, o que dificulta a construção de gráficos. Especificamente, ao construir gráficos de termos de variável única com fatores, agrupamos todos esses fatores e simplesmente desenhamos um gráfico de barras mostrando suas contribuições, positivas ou negativas.

Se os fatores aparecerem em termos junto com outras variáveis, eles sempre aparecerão com um único valor possível, porque variáveis fator só podem ter um valor para um dado imóvel: se uma área tem o valor 662, por definição ela não pode ter nenhum outro valor. Se você achar isso confuso, para piorar, no modelo gerado pelo MARS não há um único nome de variável para o fator; o MARS toma o nome-raiz da variável fator que lhe foi dada como entrada e então cria novas colunas, com nomes de variável formados pela anexação dos valores possíveis. Assim, uma única coluna da planilha Excel, como `MlsArea`, terá múltiplas colunas de data frame criadas para cada valor possível, tais como `MlsArea660`, `MlsArea661` etc. Uma variável como `MlsArea661` terá valor 1 para imóveis da área 661; caso contrário, 0. Em seguida, o MARS fará a regressão sobre essas colunas como se representassem variáveis booleanas. Esses nomes modificados serão usados no modelo fornecido pelo MARS.

Você provavelmente vai querer converter esses nomes do modelo de volta ao nome original, para tornar os gráficos mais compreensíveis a clientes que gostam das coisas simples (ao menos tão simples quanto possível). Em vez de ver `MlsArea661=1`, veremos `MlsArea=661`. Podemos então construir gráficos dos termos simplesmente declarando no cabeçalho do gráfico algo como “Para MLS Area = 662”. Assim, termos de grau N com M variáveis fator são graficados como termos de grau (N-M), e a dimensão do fator desaparece no título.

Note que duas ou mais variáveis fator podem existir em um único termo, de modo que você provavelmente precisaria de um gráfico para cada par possível de valores. Se a variável x, com três valores distintos, e a variável y, com dois valores distintos, aparecerem no mesmo termo junto com a variável não fator z, então você precisaria de $2 \cdot 3 = 6$ gráficos. Um exemplo seria “`MlsArea`” + “`Architectural Style`” (estilo arquitetônico); por exemplo, casas vitorianas em San Francisco podem valer mais ou menos dependendo da área de San Francisco em que estejam. Felizmente, na maioria das áreas há em geral apenas de 1 a 3 estilos arquitetônicos importantes o bastante como contribuidores de valor para serem usados no modelo (embora possa haver 6 ou mais estilos arquitetônicos diferentes).

8.3 Termos de Segundo e Terceiro Grau

Com a regressão de segundo ou de terceiro grau, alguns dos termos do modelo de saída podem incluir duas ou, respectivamente, três variáveis. Por exemplo:

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

Note que os nomeamos ordenando alfabeticamente os nomes das variáveis e concatenando-os, talvez com um underscore. Fazemos o mesmo com os termos de terceiro grau, sobre as triplas de nomes de variáveis produzidas no modelo. Por exemplo, “BathRms__GLA__LotSz”.

Essas variáveis são acrescentadas às variáveis primárias de entrada na Etapa 3, quando o programa RCA calcula as contribuições de todas as variáveis para cada imóvel, inclusive as variáveis de interação.

8.3.1 Variáveis de Ajuste

O programa RCA também calculará os ajustes de todas as variáveis e imóveis na Etapa 3, depois de calculadas as contribuições de valor, subtraindo da contribuição do avaliando para uma característica a correspondente contribuição de valor do imóvel comparável para essa característica.

8.3.2 Variáveis de Agregação

Variáveis de agregação são definidas para substituir subconjuntos de variáveis, somando todas as suas contribuições de valor e ajustes em uma variável comum a ser usada no laudo. Por exemplo, podemos ter dividido tanto a área habitável acabada quanto a não acabada em variáveis de detalhe mais fino:

1. Área acabada legal acima do nível do solo
2. Área acabada legal abaixo do nível do solo
3. Área acabada irregular acima do nível do solo
4. Área acabada irregular abaixo do nível do solo
5. Área não acabada acima do nível do solo
6. Área não acabada abaixo do nível do solo

Sim, há áreas em que isso faz sentido.

A regressão MARS fará a regressão sobre todas as variáveis e gerará um modelo que fornecerá contribuições e ajustes para elas. Mas, para um laudo em formulário, essas diferentes contribuições provavelmente terão de ser combinadas ou agregadas em menos variáveis. Os detalhes da agregação devem ser fornecidos em anexo. Um exemplo mais frequente seria agregar os ajustes de interação a um dos ajustes das variáveis primárias. Por exemplo, poderíamos agregar o ajuste de GLA_LotSize a GLA ou a LotSize, ou mesmo dividi-lo 50:50 entre ambos, conforme determinado pela planilha Aggregation.

8.3.3 Variáveis e Notação Especiais

π_i	Preço de venda líquido efetivo do imóvel $i = 1, 2, \dots, n_p$; para o avaliando, π_0 denota a conclusão de valor (ver a Demonstração I)
α_i	Preço de venda líquido ajustado do imóvel $i = 1, 2, \dots, n_p$
ρ_i	Preço de venda líquido estimado pelo MARS para o imóvel $i = 0, 1, \dots, n_p$
β	Valor-base do modelo
$\varphi_{i,j}$	Contribuição de valor da variável de regressão $j = 1, 2, \dots, n_r$ para o imóvel $i = 0, 1, \dots, n_p$
$\delta_{i,j}$	Ajuste da variável de regressão $j = 1, 2, \dots, n_r$ para o imóvel $i = 1, 2, \dots, n_p$
ϵ_i	Resíduo total do imóvel $i = 0, 1, \dots, n_p$
$\epsilon_{i,c}$	Contribuição de valor do componente residual c , $c = 1, 2, \dots, n_c$, para o imóvel $i = 0, 1, \dots, n_p$
$\xi_{i,c}$	Ajuste do componente residual c , $c = 1, 2, \dots, n_c$, para o imóvel $i = 0, 1, \dots, n_p$

onde

n_p = nº de imóveis comparáveis

n_r = nº de variáveis de regressão

n_c = nº de variáveis de decomposição do resíduo

8.3.4 Decompondo os Resíduos

Depois que a regressão MARS cria um modelo, ela gera os preços de venda estimados a partir das variáveis de entrada selecionadas pelo MARS. Subtrai esse valor do correspondente Preço de Venda Líquido para obter um valor residual de cada imóvel. Ora, o protocolo RCA estipula que, uma vez que o VE tenha decidido quais serão os comparáveis finais, normalmente de 6 a 12 listagens de imóveis, ele precisa percorrer cada listagem e decompor o resíduo calculado nas características (variáveis latentes) que julga mais provavelmente impactarem o valor, ignorando a possibilidade de erros aleatórios nos dados. Ele poderia, por exemplo, ter estas variáveis de

decomposição para um dado imóvel:

1. Projeto
2. Utilidade funcional
3. Venda acima/abaixo do mercado
4. Estado de conservação
5. Padrão construtivo

Imóveis diferentes provavelmente terão algumas variáveis de decomposição do resíduo distintas, que precisam ser reunidas em um único conjunto ordenado.

Estimando contribuições de características raras. Algumas características portadoras de valor são raras demais para o canal da regressão: uma oficina presente em 3 de 780 vendas não consegue receber um coeficiente defensável, e o passo direto normalmente ignorará por completo uma variável dessas. Trata-se de características residuais, e suas contribuições podem ser estimadas *no espaço residual*: ajusta-se o modelo sem a característica e, em seguida, regride-se os resíduos do modelo sobre a(s) coluna(s) excluída(s) por mínimos quadrados ordinários. Para um indicador binário, o coeficiente é uma contribuição de montante fixo (equivalentemente, o deslocamento médio do resíduo das vendas sinalizadas); para uma coluna numérica, é uma contribuição por unidade. Quando várias comodidades correlacionadas são estimadas *conjuntamente* (por exemplo, a área construída de edificações acessórias junto com um indicador de oficina), a regressão reparte o crédito entre elas, em vez de contar duas vezes os mesmos valores em duas linhas da decomposição. Estimativas apoiadas em menos de cerca de dez observações são apenas indicativas e devem ser corroboradas pelo custo depreciado ou por uma amostra ampliada (por exemplo, indicadores recuperados das observações públicas); a própria restrição residual limita qualquer valor que venha a ser finalmente atribuído. O earthUI (v0.11+) implementa esse método diretamente na aba *Residual Contribution* (contribuição residual), incluindo a estimativa conjunta, a evidência residual por venda e uma nota de documentação para o arquivo de trabalho.

Notação: usamos a notação da tabela acima para chegar a:

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

onde n_c é o número de componentes da decomposição do resíduo.

O VE precisa usar seu juízo para decompor o resíduo nos componentes dados. Poder-se-ia pensar que isso permitiria a entrada de viés na conclusão final de valor. Contudo, não é o caso, pois temos demonstração matemática de que, se os ajustes da decomposição somarem o resíduo, eles não impactarão a conclusão de valor. O propósito da decomposição não é estabelecer valor, mas explicar por que o resíduo é o que é em comparação com os resíduos das vendas de outros imóveis. O único ponto em que o viés do engenheiro pode impactar a conclusão final de valor é a estimativa do resíduo do avaliando, já que, evidentemente, não há preço de venda efetivo do avaliando que permita calcular um resíduo para ele.

Cada imóvel tem seu conjunto de contribuições de valor residuais, mas estas precisam ser fundidas em um único conjunto C sobre todos os imóveis:

$$\begin{aligned} \text{Fazendo: } A_i &= \{\{\epsilon_{i,c}\}_{c=1}^{n_c}\} \\ C &= \{x \mid x \in \bigcup_{i=0}^{n_p} A_i\} \end{aligned}$$

onde $[\epsilon_{i,c}]$ é o nome de variável para o valor $\epsilon_{i,c}$, A_i é o conjunto de tais variáveis e n_c é o número de valores em cada ϵ_i .

Assim, com os dados do avaliando e dos imóveis comparáveis em colunas separadas, então, abaixo de todas as contribuições de valor das variáveis do modelo de regressão de cada imóvel, deveríamos ver o conjunto C das n_c correspondentes contribuições de valor residuais.

9 Demonstrações Fundamentais do RCA

Seja:

$$\begin{aligned} n_p &= \text{n}^\circ \text{ de imóveis} \\ n_r &= \text{n}^\circ \text{ de variáveis de regressão} \\ n_c &= \text{n}^\circ \text{ de variáveis de componentes residuais} \end{aligned}$$

Demonstração I: Todos e cada um dos preços de venda ajustados dos comparáveis são iguais ao preço de venda estimado do avaliando.

Para qualquer comparável i , o preço de venda ajustado α_i é a soma do preço de venda líquido π_i mais os ajustes: as diferenças entre as contribuições de valor do avaliando e do comparável para o valor-base, para as variáveis de regressão $j = 1, 2, \dots, n_r$ e para o resíduo. Para o imóvel avaliando, que não tem venda efetiva, o preço de venda líquido π_0 é definido como a conclusão de valor $\beta + \varphi_{0,1} + \varphi_{0,2} + \dots + \varphi_{0,n_r} + \epsilon_0$. Então:

$$\begin{aligned} \alpha_i &= \pi_i + (\beta - \beta) + (\varphi_{0,1} - \varphi_{i,1}) \\ &\quad + (\varphi_{0,2} - \varphi_{i,2}) + \dots + \\ &\quad (\varphi_{0,n_r} - \varphi_{i,n_r}) \\ &\quad + (\epsilon_0 - \epsilon_i) \\ &= \pi_i + (\beta + \varphi_{0,1} + \varphi_{0,2} + \dots \\ &\quad + \varphi_{0,n_r} + \epsilon_0) \\ &\quad - (\beta + \varphi_{i,1} + \varphi_{i,2} + \dots + \\ &\quad \varphi_{i,n_r} + \epsilon_i) \\ &= \pi_i + \pi_0 - \pi_i \\ &= \pi_0 \end{aligned} \tag{12}$$

Demonstração II: Alterar a decomposição do resíduo, sob a restrição de que todas as contribuições de valor dos componentes residuais somem o resíduo do comparável, não tem impacto algum sobre a conclusão final de valor do protocolo RCA.

Isto é quase trivial. Ainda assim, sem pensar na matemática, é fácil deixar passar, porque os avaliadores tradicionais focam os ajustes, e não as contribuições de valor. São as contribuições de valor que se cancelam aqui — um descuido de décadas. Para algum comparável k , $k > 0$, o ajuste residual total é igual à soma dos n_c componentes residuais da decomposição, que, por sua vez, é igual à soma das diferenças de contribuições de valor, o que é igual ao resíduo total do avaliando menos o resíduo total do comparável. E isso independe do número de componentes residuais criados ou do que eles representem. Desde que o valor dos componentes residuais seja igual ao resíduo total tanto para o avaliando quanto para o comparável escolhido, ξ_k não mudará.

$$\begin{aligned}
\xi_k &= \xi_{k,1} + \xi_{k,2} + \dots + \xi_{k,n_c} \\
&= (\epsilon_{0,1} - \epsilon_{k,1}) + (\epsilon_{0,2} - \epsilon_{k,2}) + \dots + \\
&\quad (\epsilon_{0,n_c} - \epsilon_{k,n_c}) \\
&= (\epsilon_{0,1} + \epsilon_{0,2} + \dots + \epsilon_{0,n_c}) - \\
&\quad (\epsilon_{k,1} + \epsilon_{k,2} + \dots + \epsilon_{k,n_c}) \\
&= \epsilon_0 - \epsilon_k
\end{aligned} \tag{13}$$

10 Modelo MARS

10.1 Funções “ g ” de Variáveis e Interações do Modelo

Um princípio central do método RCA é que o “raciocínio explicativo” é da máxima importância. Infelizmente, fornecer gráficos significativos para as várias equações de contribuição de valor, especialmente quando se usam interações e variáveis fator, pode ser desafiador. Considera-se, portanto, imperativo fornecer um protocolo claro e estrito para a criação desses gráficos. É o que se faz aqui, mas entenda que o protocolo se dirige ao desenvolvedor que escreve o código R ou Python para gerar os gráficos, e talvez interesse apenas superficialmente a quem não está envolvido com a codificação.

O modelo de preço gerado pelo MARS é a soma de uma constante-base mais termos adicionais, cada um envolvendo de 1 a 3 variáveis. Para analisá-lo, primeiro decompomos o modelo em termos individuais. Em seguida, múltiplos termos que compartilham o mesmo conjunto de variáveis, representando uma única interação, são combinados em expressões de interação únicas. Suponha que o grau máximo da regressão seja 3. Considere um modelo de regressão de terceiro grau composto por uma constante-base, n_1 expressões de primeiro grau, n_2 expressões de interação de segundo grau e n_3 expressões de interação de terceiro grau. Isso nos dá quatro grupos distintos de expressões, que são interações de variável única ou de múltiplas variáveis, cada qual graficada de modo diverso conforme dois fatores: (1) se incluem uma variável fator (uma variável categórica) e (2) o número de variáveis distintas que contêm. Especificamente:

- Expressões de primeiro grau (1 variável) são graficadas em 2D.
- Expressões de segundo grau (2 variáveis) são graficadas em 3D.
- Expressões de terceiro grau (3 variáveis) são graficadas como um *conjunto* de gráficos 3D.
- Se uma expressão contém uma variável fator, cada fator distinto reduz em um a dimensionalidade do gráfico (por exemplo, um gráfico 3D torna-se 2D se uma variável for fator).

A indexação de cada função g fornece a informação necessária para determinar como ela deve ser graficada, refletindo o número de variáveis distintas e a presença de variáveis fator.

10.2 Indexação das Funções g

Cada função g é indexada como $f_{j,k} g_k^j$, onde:

- $f_{j,k}$ é o primeiro índice (um sobrescrito antes de g):
 - j varia de 0 a 3 e identifica o grupo:
 - * $j = 0$: a constante-base (uma exceção à indexação usual).
 - * $j = 1$: expressões de primeiro grau.
 - * $j = 2$: expressões de segundo grau.
 - * $j = 3$: expressões de terceiro grau.
 - k é o índice de uma função específica dentro de seu grupo (por exemplo, $k = 1$ para a primeira função do grupo j).
 - O valor de $f_{j,k}$ indica o número de variáveis fator distintas na função:
 - * $f_{j,k} = 0$: sem variáveis fator.
 - * $f_{j,k} > 0$: o número de variáveis fator (por exemplo, 1, 2 ou 3).

- Índices posteriores em g :
 - O sobrescrito j após g (por exemplo, g^j) repete o número do grupo, por clareza.
 - O subscrito k após g (por exemplo, g_k) corresponde ao k de $f_{j,k}$, identificando a função dentro de seu grupo.

Embora os índices de f e g coincidam numericamente, seus papéis diferem: $f_{j,k}$ conta variáveis fator, o que é crítico para as decisões de construção de gráficos, ao passo que j e k em g definem seu grupo e sua posição.

10.3 Dimensões dos Gráficos

A dimensão d do gráfico de uma função $f_{j,k} g_k^j$ é calculada como:

$$d = j - f_{j,k} \quad (14)$$

onde:

- j é o número de variáveis distintas (ou o número do grupo, desconsiderado o caso da constante).
- $f_{j,k}$ é o número de variáveis fator.
- d é o número de variáveis contínuas (não fator) a serem graficadas, com máximo de 3 (já que o grau máximo é 3). Uma função com d variáveis contínuas é desenhada em $d + 1$ dimensões: $d = 1$ dá um gráfico 2D, $d = 2$ um gráfico 3D (ou um mapa de calor 2D) e $d = 3$ um conjunto de gráficos 3D.

Por exemplo:

- Se $j = 3$ (3 variáveis) e $f_{j,k} = 0$ (sem variáveis fator), então $d = 3 - 0 = 3$, exigindo um conjunto de gráficos 3D, cada um fixando uma das variáveis.
- Se $j = 3$ e $f_{j,k} = 1$ (uma variável fator), então $d = 3 - 1 = 2$, exigindo um gráfico 3D (ou mapa de calor 2D) para cada valor da variável fator que aparece no modelo.
- Se $j = 0$ (a constante), normalmente não é preciso gráfico, pois se trata de um único valor.

Quando $d = 3$ (por exemplo, uma expressão de terceiro grau sem variáveis fator), visualizar a função completa pode exigir múltiplos gráficos 3D, cada um fixando uma variável em valores diferentes para mostrar um subconjunto do comportamento da função.

10.4 A Equação Final do Modelo

O modelo MARS é decomposto em quatro grupos de expressões (constante, primeiro, segundo e terceiro graus). A indexação $f_{j,k} g_k^j$ nos informa:

- j : o grupo e o número de variáveis.
- $f_{j,k}$: o número de variáveis fator.
- $d = j - f_{j,k}$: a dimensão do gráfico.

Essa estrutura orienta como cada função g é visualizada, assegurando que o método gráfico corresponda à complexidade da expressão e aos tipos de variáveis.

A estimativa do preço de venda $\hat{P}$ fornecida pelo modelo MARS pode ser expressa como:

$$\begin{aligned} \hat{P} = & f_{0,1}g_1^0 + \sum_{q=1}^{n_1} f_{1,q}g_q^1(x_*) \\ & + \sum_{r=1}^{n_2} f_{2,r}g_r^2(x_*, y_*) \\ & + \sum_{s=1}^{n_3} f_{3,s}g_s^3(x_*, y_*, z_*) \end{aligned} \quad (15)$$

$n_1 = \text{n}^\circ \text{ de termos de primeiro grau}$
 $n_2 = \text{n}^\circ \text{ de termos de segundo grau}$
 $n_3 = \text{n}^\circ \text{ de termos de terceiro grau}$

10.5 Número de Gráficos Necessários Para Representar um Modelo

Quantos gráficos são necessários para visualizar todas as funções de um modelo? A resposta depende das variáveis, de suas interações e de serem elas variáveis fator ou não.

O VE pode limitar variáveis e interações no arquivo de configuração Excel destinado ao R/earth, influenciando quais interações são permitidas. Uma abordagem prática é deixar o R/earth avaliar todas as interações possíveis e selecionar aquelas com contribuições significativas para a estimativa do preço de venda, verificando ao mesmo tempo problemas como sobreajuste ou colinearidade. Algumas interações podem ser desautorizadas, afetando interações de grau superior (por exemplo, se um par é excluído em uma interação de terceiro grau, a combinação inteira fica excluída).

Para variáveis fator (por exemplo, MLS Area com 12 valores interagindo com GLA), a construção de gráficos torna-se mais complexa. Isso poderia exigir 12 gráficos — um por área. Se outro fator, como Architectural Style (6 valores), interagir com MLS Area e GLA, poderiam ser necessários até $12 \cdot 6 = 60$ gráficos. Contudo, o R/earth raramente considera todas as combinações significativas; tipicamente, apenas algumas poucas interações-chave (por exemplo, áreas e estilos específicos) merecem gráfico, conforme identificado pelo modelo.

Eis um guia simplificado do número de gráficos necessários:

1. **Funções de variável única (não fator):** 1 gráfico.
2. **Funções de variável única (fator):** 1 gráfico de barras combinando todos os valores do fator.
3. **Função-base:** nenhum gráfico necessário.
4. **Interações de 2º grau (sem fatores):** 1 gráfico (por exemplo, mapa de calor 2D com cores para o preço de venda).
5. **Interações de 2º grau (1 fator):** 1 gráfico por valor do fator presente no modelo (por exemplo, 2 valores = 2 gráficos).
6. **Interações de 2º grau (2 fatores):** 1 gráfico com uma matriz das combinações dos fatores (por exemplo, 3 áreas $\times$ 2 estilos = 6 células).
7. **Interações de 3º grau (sem fatores):** múltiplos gráficos (por exemplo, 4 gráficos para número de banheiros = 1, 2, 3, 4).
8. **Interações de 3º grau (1 fator):** 1 gráfico por valor do fator presente no modelo.

9. **Interações de 3º grau (2 fatores):** 1 gráfico por combinação de valores dos fatores presente no modelo.
10. **Interações de 3º grau (3 fatores):** 1 gráfico por valor do fator com o menor número de valores distintos, usando uma matriz de cores para os outros dois.

Cenários mais complexos podem exigir técnicas gráficas avançadas.

Há outras possibilidades, com gráficos provavelmente mais sofisticados.

Um exemplo de modelo completo gerado pelo MARS está na [Figure 1](#).

Figura 1: Componentes do Modelo MARS

Basis	$= {}^0g_1^0 = 4957150$
Age	$= {}^0g_1^1 = -3163.1809255 \cdot \text{Age}$
AreaNbr	$= {}^1g_2^1 = -623197.75104 \cdot I\{\text{AreaNbr} = 471\}$
Baths	$= {}^0g_3^1 = -111023.51496 \cdot \max(0, 2 - \text{Baths}) + 96735.52418 \cdot \max(0, \text{Baths} - 2)$
Beds	$= {}^0g_4^1 = -233489.66415 \cdot \max(0, \text{Beds} - 5)$
DaysOffMkt	$= {}^0g_5^1 = -2577.2188622 \cdot \max(0, 189 - \text{DaysOffMkt}) - 1282.8789534 \cdot \max(0, \text{DaysOffMkt} - 189)$
GLA	$= {}^0g_6^1 = -666.73587914 \cdot \max(0, 4027 - \text{GLA}) + 281.36312913 \cdot \max(0, \text{GLA} - 4027)$
Latitude	$= {}^0g_7^1 = -18141680.417 \cdot \max(0, \text{Latitude} - 37.573)$
LotSize	$= {}^0g_8^1 = -85.005683404 \cdot \max(0, 8000 - \text{LotSize}) - 23.562970183 \cdot \max(0, \text{LotSize} - 8000)$
Age_Baths	$= {}^0g_1^2 = -3916.3594872 \cdot \text{Age} \cdot \max(0, \text{Baths} - 3.5)$
Age_Garage	$= {}^0g_2^2 = -6503.1897339 \cdot \text{Age} \cdot \max(0, \text{Garage} - 2)$
AreaNbr_LotSize	$= {}^1g_3^2 = 78.54207825 \cdot I\{\text{AreaNbr} = 463\} \cdot \max(0, 8000 - \text{LotSize})$ $+ 87.263298012 \cdot I\{\text{AreaNbr} = 466\} \cdot \max(0, \text{LotSize} - 8000)$
AreaNbr_Baths	$= {}^1g_4^2 = -1070222.5995 \cdot I\{\text{AreaNbr} = 464\} \cdot \max(0, \text{Baths} - 2)$
AreaNbr_Latitude	$= {}^1g_5^2 = 17571072.548 \cdot I\{\text{AreaNbr} = 466\} \cdot \max(0, \text{Latitude} - 37.573)$
AreaNbr_GLA	$= {}^1g_6^2 = -328.79173933 \cdot I\{\text{AreaNbr} = 471\} \cdot \max(0, \text{GLA} - 2550)$ $+ 525.69503229 \cdot I\{\text{AreaNbr} = 471\} \cdot \max(0, 2550 - \text{GLA})$
DaysOffMkt_GLA	$= {}^0g_7^2 = 0.43839582286 \cdot \max(0, \text{DaysOffMkt} - 661) \cdot \max(0, 4027 - \text{GLA})$
FrPlcNbr_LotSize	$= {}^0g_8^2 = 22.059059591 \cdot \max(0, 2 - \text{FrPlcNbr}) \cdot \max(0, \text{LotSize} - 8000)$ $+ 39.084508159 \cdot \max(0, \text{FrPlcNbr} - 2) \cdot \max(0, \text{LotSize} - 8000)$
GLA_LotSize	$= {}^0g_9^2 = -0.45274160024 \cdot \max(0, \text{GLA} - 2994) \cdot \max(0, 8000 - \text{LotSize})$
ADU_Age_Garage	$= {}^0g_1^3 = -5.4578633919 \cdot \max(0, \text{ADU} - 0) \cdot \text{Age} \cdot \max(0, 2 - \text{Garage})$
AreaNbr_Baths_LotSize	$= {}^1g_2^3 = 60.607126605 \cdot I\{\text{AreaNbr} = 460\} \cdot \max(0, \text{Baths} - 2) \cdot \max(0, \text{LotSize} - 2480)$
AreaNbr_Baths_FrPlcNbr	$= {}^1g_3^3 = 593779.41722 \cdot I\{\text{AreaNbr} = 464\} \cdot \max(0, \text{Baths} - 2) \cdot \max(0, \text{FrPlcNbr} - 0)$
AreaNbr_GLA_Latitude	$= {}^1g_4^3 = 54199.925713 \cdot I\{\text{AreaNbr} = 464\} \cdot \max(0, \text{GLA} - 2670) \cdot \max(0, \text{Latitude} - 37.573)$ $+ 58934.565747 \cdot I\{\text{AreaNbr} = 471\} \cdot \max(0, \text{GLA} - 2550) \cdot \max(0, \text{Latitude} - 37.568)$
AreaNbr_DaysOffMkt_GLA	$= {}^1g_5^3 = 0.37173306643 \cdot I\{\text{AreaNbr} = 470\} \cdot \max(0, 661 - \text{DaysOffMkt}) \cdot \max(0, 4027 - \text{GLA})$
Beds_GLA_LotSize	$= {}^0g_6^3 = -5.1607264026 \cdot \max(0, 4 - \text{Beds}) \cdot \max(0, \text{GLA} - 2994) \cdot \max(0, 8000 - \text{LotSize})$
Age_AreaNbr_Baths	$= {}^1g_7^3 = -10759.444688 \cdot \text{Age} \cdot I\{\text{AreaNbr} = 460\} \cdot \max(0, \text{Baths} - 3.5)$

11 Características dos Dados

A seção seguinte trata do MARS. O MARS depende inteiramente dos dados que recebe, e a maior parte desses dados virá de algum MLS (Multiple Listing Service) ou de órgãos fiscais. Contudo, eles podem passar por outros provedores de dados que os mantêm e aprimoram para depois vendê-los com algum lucro. Em algumas áreas, são razoavelmente confiáveis; em outras, nem tanto. Em situações críticas, o vistoriador do imóvel precisa fazer suas próprias observações e medições, cuja exatidão pode ser questionada.

O engenheiro de avaliações precisa compreender o que os dados representam e o quão bem cumprem o papel de representar aquilo que deveriam representar. Isso suscita toda uma lista de questões:

- Como se classifica uma variável como sendo uma medição?
- Se uma variável não é uma medição, o que mais ela pode ser?
- Como a regressão pode ser útil se os dados que recebe não são exatos?

A maior parte dos dados que obtemos dos serviços de listagem do MLS e dos dados dos órgãos fiscais para regressão são dados numéricos que implicam medições ou contagens. Alguns são lógicos, verdadeiro ou falso.

11.1 Finalidade do Resíduo

O resíduo serve como medida indireta do valor agregado por variáveis não incluídas no modelo de regressão. Na linguagem da literatura hedônica, essas características não medidas são *variáveis latentes*: variáveis que influenciam o preço de venda mas não são medidas diretamente, na maioria dos casos porque não há maneira prática de medi-las objetivamente. Seus valores precisam, em vez disso, ser inferidos e, na prática, o valor de uma variável latente é muitas vezes estimado por juízo subjetivo — no RCA, um juízo ancorado na ordenação dos resíduos. Ao ordenar os imóveis com base em seu resíduo ou resíduo por pé quadrado, obtemos uma ordenação objetiva, dos imóveis de menor apelo aos de maior apelo. Esse método de ordenação é eficaz, desde que um analista habilidoso, com profundo conhecimento do mercado local, construa o modelo MARS (Multivariate Adaptive Regression Splines). Seguem alguns requisitos adicionais:

- Qualidade dos dados e do modelo: um modelo de regressão bem construído e dados de alta qualidade são pré-requisitos para resíduos confiáveis. Na San Francisco Bay Area, minha meta é desenvolver um modelo com R^2 de aproximadamente 75-80% e R^2 de validação cruzada (CV R^2) de 55-70%.
- Evitar o sobreajuste, limitar o R^2 : o engenheiro de avaliações deve buscar um R^2 máximo em torno de 80% na maioria dos mercados maduros de áreas de preço mais alto, como a SF Bay Area, porque boa parte do valor competitivo de um imóvel residencial reside em seu apelo subjetivo, que não é quantificável e, portanto, não pode ser incluído diretamente em modelos de regressão. Um R^2 significativamente acima de 80% pode indicar um modelo sobreajustado, exigindo exame adicional. Ainda assim, uma palavra de cautela: existem regiões de mercado com valores de R^2 acima de 90% sem sobreajuste, e existem áreas difíceis de avaliar, nas quais o melhor que se consegue talvez seja 50-60% ou menos.
- R^2 elevado em áreas específicas: contudo, há regiões, como loteamentos ou outras áreas residenciais uniformes, nas quais valores de R^2 mais altos são surpreendentemente comuns, em razão da homogeneidade dos imóveis.

11.2 Notas de Estado de Conservação e Padrão Construtivo Como Regressores Aproximados

As escalas de estado de conservação e padrão construtivo da Fannie Mae (C1–C6 e Q1–Q6) merecem comentário especial. Essas escalas são grosseiras demais para serem tratadas como verdadeiras características medidas e, por serem atribuídas pelo avaliador, são na prática notas subjetivas — variáveis latentes por direito próprio. É útil distinguir explicitamente dois níveis: a nota grosseira do avaliador usada como regressor, *condition-coarse-grain* (condition-1, estado de conservação de granulação grossa) e *quality-coarse-grain* (quality-1, padrão construtivo de granulação grossa), versus as variáveis latentes de granulação fina recuperadas do resíduo, *condition-fine-grain* (condition-2) e *quality-fine-grain* (quality-2). O VE pode aceitar as notas de granulação grossa do avaliador como aproximações e regredir sobre elas, obtendo contribuições de valor aproximadas que são úteis como atribuições iniciais de valor para as variáveis latentes de granulação fina. Onde isso é feito, o protocolo consiste em subtrair da estimativa do modelo as contribuições de valor dessas duas variáveis de granulação grossa estimadas pela regressão — retirando-as dos atributos medidos, sob a seção de regressão do quadro de homogeneização — e devolvê-las ao resíduo. Dali, elas são subsequentemente removidas do resíduo e atribuídas às correspondentes variáveis latentes de granulação fina, *res_condition1* e *res_quality1*. Informações adicionais extraídas das observações públicas do MLS (*public_remarks*) e de listas de texto relacionadas são então usadas para fazer microrrefinamentos nesses valores de variáveis latentes. O resultado final é que as contribuições de valor de granulação fina substituem as de granulação grossa: as notas grosseiras servem apenas de andaime para o modelo aproximado, e os valores reportados ao final para estado de conservação e padrão construtivo são os de granulação fina, derivados do resíduo.

11.3 Erros nos Dados

Com base em comentários de redes sociais, os dados do Multiple Listing Service (MLS) da Califórnia são em geral de qualidade superior aos de muitas outras regiões dos Estados Unidos. Tipicamente, deve-se esperar que os dados de MLS sejam de boa qualidade na maioria das áreas metropolitanas maduras, porém piores em áreas rurais.

Embora haja erros ocasionais nos dados por diversos motivos, esses erros devem em geral cancelar-se mutuamente e ter impacto mínimo sobre as médias usadas nos modelos de preço, caso exista dado suficiente.

Contudo, vieses sistemáticos podem ocorrer em áreas específicas. Por exemplo, considere um bairro em que muitas casas foram inicialmente construídas com segundo pavimento completo, mas alguns compradores tiveram a opção de suprimir cômodos do segundo pavimento para criar pés-direitos duplos em um ou mais cômodos do térreo, reduzindo significativamente a área bruta habitável (GLA). As alterações não foram posteriormente atualizadas nos registros do órgão fiscal, por uma razão ou outra. Em consequência, o MLS pode mostrar 2,400 pés quadrados quando o valor medido é de apenas 2,000 pés quadrados. Nesses cenários, os dados do MLS devem ser corrigidos antes de se executar a regressão MARS (Multivariate Adaptive Regression Splines). É importante notar que obter essa informação para os comparáveis pode exigir algum esforço investigativo. Discrepâncias devem ser registradas para referência futura.

Diante do exposto, é razoável supor que, em áreas de MLS de alta qualidade, os erros que transbordam para o resíduo sejam mínimos em comparação com o impacto de características subjetivas e não medidas, como projeto, estado de conservação e padrão construtivo. Essa é uma consideração importante ao estimar a exatidão do método RCA. Por exemplo, se o valor de R^2 de uma dada região de mercado é 80%, então aproximadamente 20% do preço é atribuído ao resíduo CQA. Um erro de $\pm 10\%$ no posicionamento do avaliando dentro da ordenação de

resíduos propaga-se, portanto, para um erro de $\pm 10\% \cdot 20\% = \pm 2\%$ na conclusão final de valor. É assim que se melhora a exatidão! Em muitos casos, estimativas de valor exatas bem abaixo de $\pm 1\%$ são possíveis.

Portanto, é razoável tratar os resíduos como indicadores indiretos do valor coletivo das variáveis não modeladas na regressão MARS, desde que o engenheiro de avaliações tenha experiência local de mercado suficiente para identificar onde e por que existem vieses sistêmicos de medição. Essa abordagem assegura que a análise reflita com exatidão os valores efetivos dos imóveis, reconhecendo as limitações e as nuances dos dados imobiliários.

Exemplo Prático: Medindo Determinantes Não Medidos em Sea Ranch

Quando o R^2 alcançável em uma região de mercado é menor do que a aparente homogeneidade do mercado levaria a esperar, a resposta da estrutura metodológica é identificar e medir a variável faltante, em vez de aceitar a limitação. A comunidade de Sea Ranch, no condado de Sonoma, ilustra isso diretamente.

Uma instância concreta da resposta “medir, não aceitar”. A comunidade de Sea Ranch, no condado de Sonoma, é um empreendimento litorâneo em que a distância ao oceano é um determinante primário de valor: uma casa a algumas centenas de pés da falésia alcança preço substancialmente diferente do de outra a um quarto de milha para o interior, mantidas constantes as demais características. Os dados do MLS da comunidade não registram a distância ao oceano como variável. Um modelo MARS inicial, ajustado sobre as variáveis disponíveis no MLS, produziu poder explicativo inferior ao que a aparente homogeneidade do mercado deveria sustentar, com uma estrutura de resíduos que não ordenava suficientemente os imóveis por nenhuma característica visível nos dados — sinal de que um determinante importante faltava ao modelo, e não de que o mercado fosse genuinamente ruidoso.

O remédio foi a medição direta. O autor obteve um mapa da comunidade mostrando a disposição dos lotes em relação à linha da costa e mediu a distância ao oceano de várias centenas de casas. As distâncias medidas foram acrescentadas como variável ao conjunto de dados do MLS e a regressão MARS foi refeita. A nova variável foi selecionada pelo modelo como significativa, o poder explicativo aumentou substancialmente e a ordenação dos resíduos passou a ordenar os imóveis por padrões consistentes com as características subjetivas remanescentes — vista, estado de conservação, padrão construtivo — em vez de pelo sinal latente da distância ao oceano, que antes a contaminava. A altimetria tem sido tratada de modo semelhante em bairros de encosta em que a variável do MLS está ausente ou não é confiável.

O princípio metodológico é geral. Quando o R^2 alcançável fica abaixo do que a aparente homogeneidade do mercado deveria sustentar, a primeira hipótese é a de que falta ao modelo uma variável importante, e a resposta apropriada é identificar qual é e coletar as medições diretamente. As fontes incluem mapas comunitários, dados de SIG, registros de órgãos fiscais, registros da prefeitura, imagens de satélite e vistoria de campo. O custo desse trabalho de medição é real, mas limitado; uma vez coletada a variável para uma dada região de mercado, ela pode ser aproveitada em trabalhos futuros naquela área. O RCA absorve novas variáveis de forma limpa, porque a estrutura metodológica é indiferente a quais variáveis entram na regressão MARS — a estrutura dos resíduos diz ao VE se a nova variável captura sinal real, e as contribuições de valor se ajustam em conformidade.

A resposta alternativa — aceitar um R^2 baixo e reportar faixas de erro mais amplas — é defensável quando os determinantes faltantes são genuinamente imensuráveis (preferências específicas do comprador, sentimento transitório de mercado, circunstâncias idiossincráticas da transação), mas não quando eles são apenas não coletados. O primeiro movimento do VE diante de um modelo fraco deve ser perguntar se o sinal faltante é deste último tipo.

11.4 O Que É Exatidão?

A exatidão perfeita de uma avaliação exige:

1. Conformidade com o valor de mercado: a transação hipotética de venda do imóvel avaliando adere 100% à definição dada de valor de mercado
2. Desenvolvimento de modelo exato e robusto: desenvolve-se um modelo MARS de alta qualidade, com R2 em torno de 75-80% e CVR2 de aproximadamente 55-70%, conforme a localização.
3. Boa análise de erros nos dados: todos os prováveis erros de dados das variáveis medidas são compreendidos e contidos, com pouco transbordamento para os resíduos.
4. Consideração de vieses sistêmicos: qualquer viés sistêmico nos dados deve ser contabilizado e ajustado.
5. Análise da ordenação dos resíduos: assegurar que os imóveis, na ordenação por resíduo, sigam um padrão razoável de características, do menor ao maior posto. Anomalias como vendas de inventário, vendas em short sale e leilões devem ser investigadas, explicadas e possivelmente removidas.
6. Posicionamento do avaliando: com uma compreensão completa da ordenação dos resíduos, o imóvel avaliando deve ser posicionado com exatidão. Devem ser identificados um imóvel de apelo significativamente maior e outro de apelo significativamente menor, posicionando-se o avaliando entre eles e pontuando-o. Depois, procure estreitar ao máximo a faixa da ordenação entre os imóveis superior e inferior.
7. Raciocínio explicativo: nunca sacrifique o raciocínio explicativo em nome da exatidão. O avaliador precisa ser capaz de explicar as diferenças de valor entre o avaliando e os imóveis comparáveis, ao menos a alguém de inteligência razoável. Por exemplo, um bom modelo MARS com R2 e CVR2 ótimos pode ser usado para explicar as contribuições de valor das características medidas, ao passo que a decomposição do resíduo pelo engenheiro de avaliações pode ser usada para explicar as contribuições de valor das demais características.
8. Sobreajuste: evite modelos sobreajustados, não usando características do imóvel capazes de identificar, isoladamente ou em combinação, um ou vários imóveis. Em particular, evite usar variáveis fator cujos valores possam especificar um número pequeno de imóveis. Uma variável fator que identifique bairros em conjuntos de dados de tamanho considerável é apropriada, mas uma que identifique imóveis com piscina em uma área em que existam pouquíssimas piscinas precisa ser usada com cautela. Uma característica que identifique imóveis únicos, como coordenadas GIS exatas, números de matrícula ou endereços, é absolutamente proibida.

Se o R2 da regressão é 80%, esperamos que cerca de 20% do valor do avaliando se deva ao resíduo. Ora, isso é apenas aproximado. Sabemos, pela curva CQA-Resíduo, que a maior parte da distribuição dos resíduos está nos extremos inferior e superior do CQA. Com base na curva CQA-Resíduo e na inclinação extrema nos valores baixos e altos de CQA, poderíamos esperar uma estimativa menos exata do resíduo nessas faixas. Contudo, isso é contrabalançado pela relativa escassez de imóveis nessas faixas e pela maior diferença de apelo que se observa entre imóveis adjacentes. Assim, embora possamos falar do impacto de uma faixa de $\pm 5\%$ na região central, nos extremos estaríamos mais preocupados com uma faixa um pouco mais estreita, de $\pm 2 - \pm 3\%$

Calcular a exatidão do RCA é, em geral, algo difícil, em razão da não linearidade da curva CQA-Resíduo. Mas é possível caso a caso, desde que se disponha de uma ordenação de resíduos razoável. A amplitude dos escores CQA é de 0,0 a 10,0. Com um bom modelo MARS, sua ordenação de resíduos deve mostrar um padrão claro do menor ao maior apelo, e você não deve ter dificuldade em pontuar dentro de $\pm 5\%$ da amplitude total, ou seja, ± 0.5 no escore CQA. Se houver 100 imóveis na regressão, isso equivaleria a classificar o imóvel avaliando com exatidão dentro de um grupo de 10 imóveis. Assim, com referência aos dados de vendas de Burlingame da Tabela 3, calculemos a exatidão (ou o erro) aproximada para um CQA estimado pelo VE de 7,5, de 5,1 e de 1,9, sendo sua conclusão de valor para o avaliando de \$3,000,000. Note que, como a Tabela 3 contém apenas uma amostra das várias centenas de imóveis utilizados, será necessário interpolar o resíduo quando cabível (supondo linearidade). Para ser claro, estamos supondo que o VE esteja muito seguro, em cada caso, de que sua classificação do avaliando se encontra dentro de uma faixa de ± 5 imóveis na ordenação de resíduos de apenas 100 imóveis.

CQA= 7.5 ± 0.5					
CQA	Resíduo	Dif. abs.	Erro esp.	Preço hipot. de venda	Erro pct.
8.00	\$257,112				
7.50	\$197,113	\$59,999			
7.00	\$154,615	\$42,498	$\pm \$51,249$	\$3,000,000	$\pm 1.71\%$

CQA=5.1 ± 0.5					
CQA	Resíduo	Dif. abs.	Erro esp.	Preço hipot. de venda	Erro pct.
5.60	\$50,911				
5.10	\$19,559	\$31,352			
4.60	-\$28,207	\$47,766	$\pm \$39,559$	\$3,000,000	$\pm 1.32\%$

CQA= 1.9 ± 0.5					
CQA	Resíduo	Dif. abs.	Erro esp.	Preço hipot. de venda	Erro pct.
2.40	-\$228,411				
1.90	-\$276,381	\$47,970			
1.40	-\$326,938	\$50,557	$\pm \$49,264$	\$3,000,000	$\pm 1.64\%$

Por fim, o único modo objetivo de verificar a exatidão é estabelecer um protocolo estrito (que este artigo apenas esboça em linhas gerais) e verificar sua observância. Isso é necessário porque os preços de venda efetivos podem desviar-se do valor de mercado em razão de imperfeições do mercado. Comparar avaliações e vendas subsequentes pode fornecer alguma indicação de exatidão. Contudo, as complexidades e a instabilidade do mercado imobiliário exigem a expertise de avaliadores ou de engenheiros de avaliações altamente qualificados para fornecer estimativas de valor baseadas em normas e protocolos estabelecidos. Estes devem ser rigorosos o bastante para assegurar que engenheiros de avaliações competentes cheguem independentemente a conclusões de valor quase idênticas. Essa deve ser uma meta alcançável na maioria dos casos em que haja bons dados disponíveis.

11.5 O Imóvel Avaliando Como Anomalia

Uma vez que estimar o valor de mercado de um imóvel exige a análise de transações de venda recentes, é justo perguntar se as transações recentes representam bem as diversas características de um imóvel avaliando mais antigo, carente de modernização e reparos. Isso é particularmente verdadeiro na San Francisco Bay Area, onde as transações de venda recentes tipicamente são de

casas que passaram por reformas para melhorar seu preço de venda. Assim, o padrão construtivo e o estado de conservação médios dos dados de mercado são em geral superiores aos de casas mais antigas que não estão no mercado há bastante tempo. Essa diferença de valor deve ser captada pela estimativa do resíduo do imóvel avaliando, ou seja, por seu escore CQA. Trata-se de uma consideração importante quanto à qualidade dos dados e à exatidão na avaliação do imóvel avaliando.

11.6 A Importância do CVR2 no Trato de Dados Inesperados

A regressão MARS baseia-se em múltiplas iterações do R/earth aplicadas a um subconjunto de imóveis selecionado aleatoriamente, que tipicamente compreende de 80% a 90% do conjunto de dados do MLS, denominado conjunto de treino. Os 10%-20% restantes de imóveis servem como conjunto de teste. Nessa fase, o modelo MARS é criado a partir do conjunto de treino e depois usado para prever os preços de venda dos imóveis do conjunto de teste, os quais são então comparados aos preços de venda efetivos. Essa metodologia tem sido reconhecida pelos profissionais de mineração de dados como a estratégia mais eficaz para desenvolver um modelo robusto, capaz de prever os preços de venda de imóveis ainda não observados.

Contudo, a principal utilidade da métrica CVR2 não está em refletir a exatidão factual do modelo de regressão quanto às vendas efetivas da região de mercado, mas sim em otimizar as capacidades preditivas do modelo quanto a vendas hipotéticas de imóveis em uma dada data-base. Esses imóveis não estão incluídos no conjunto de treino e podem apresentar combinações novas ou inesperadas de características, ausentes tanto nos dados do conjunto de treino quanto nos do conjunto de teste.

Nesse contexto, como podemos estimar o preço de venda de um imóvel que não passa por reforma há 15 anos, supondo que tenha sido anunciado, por exemplo, um mês antes da data-base da avaliação? Um modelo que demonstre CVR2 forte tem maior probabilidade de fornecer a estimativa mais exata do que um modelo que não tenha passado por essa validação cruzada repetitiva. É também vital que o conjunto de dados inclua um número suficiente de imóveis comparáveis com graus variados de reforma, reparo e modernização.

Em última análise, se dispusermos de um modelo robusto que preveja com exatidão as contribuições de valor de diversas características dos imóveis em nosso conjunto de dados de comparáveis, poderemos posicionar efetivamente o novo imóvel avaliando dentro da ordenação desses imóveis por resíduo produzida pelo modelo. Esse posicionamento nos permite atribuir ao imóvel um escore CQA e um resíduo apropriados.

12 Resíduos

Como indica o nome “Residual Constraint Approach”, os resíduos são centrais para esta metodologia. Aqui, resíduo refere-se à diferença entre o preço de venda líquido de um imóvel e o preço de venda previsto pela análise de regressão.

Um resíduo positivo sugere que o imóvel foi vendido por mais do que seu preço estimado, com base em atributos mensuráveis típicos como a área bruta habitável (GLA) e a área do terreno, o que implica que ele talvez tenha apelo acima da média. Um resíduo negativo indica que o imóvel foi vendido por menos, sugerindo potencialmente menor apelo ou outros atributos negativos não captados.

Dado que casas maiores provavelmente têm resíduos maiores, podemos usar o resíduo por pé quadrado com base na GLA para uma comparação mais normalizada. Uma discussão mais aprofundada dessa questão fica reservada a outro artigo.

12.1 Ordenação e Pontuação

Depois de criado o modelo MARS (Multivariate Adaptive Regression Splines) para as variáveis medidas, calculam-se os resíduos de cada imóvel. Esses resíduos representam a diferença entre os preços de venda líquidos observados e os preços de venda previstos pelo modelo. Deve-se criar uma coluna adicional para o Resíduo/SF, isto é, o resíduo dividido pela área habitável da residência, entendendo-se que, em geral, o resíduo de determinados imóveis é mais significativo conforme o tamanho da área habitável. Outras estatísticas possíveis podem funcionar melhor em alguns bairros, mas são assunto para outra ocasião.

Premissa 7: “Ordenação por Resíduo”

Neste artigo, as referências a “ordenação por resíduo” devem ser entendidas como ordenação por resíduo, por resíduo/SF, ou ordenação por alguma outra função dos resíduos, conforme o engenheiro de avaliações julgue mais apropriado à região de mercado. A experiência do autor na SF Bay Area indica que o resíduo/SF é em geral o mais confiável, especialmente quando se trata de um imóvel avaliando com GLA (Gross Living Area) incomumente pequena ou grande.

Em seguida, os imóveis são ordenados por seus resíduos, do mais negativo ao mais positivo. Gera-se então um escore de estado de conservação, padrão construtivo e apelo (Condition-Quality-Appeal, CQA) para cada imóvel, variando de 0,0 a 10,0. Esse escore é calculado da seguinte forma:

O escore CQA corresponde ao posto percentílico do resíduo de um imóvel entre todos os imóveis. Especificamente, o escore é calculado determinando-se o percentual de imóveis com resíduos inferiores ao do imóvel em questão e dividindo-se esse percentual por 10. Por exemplo, se o resíduo de um imóvel é superior aos resíduos de 50% dos imóveis, seu escore seria 5,0.

Em razão do arredondamento, um imóvel pode receber escore 10,0, representando o imóvel de maior resíduo. Contudo, um escore perfeito de 10,0 é teoricamente impossível porque, logicamente, nenhum imóvel pode ter resíduo superior ao de 100% dos imóveis, ele próprio incluído. Por outro lado, o escore mínimo possível de 0,0 é viável, pois ao menos um imóvel não tem nenhum outro imóvel com resíduo inferior. Note que mais de um imóvel pode ter escore 0,0 por arredondamento, se muitos imóveis estiverem sendo comparados.

Portanto, o escore CQA fornece uma medida intuitiva de quão bem os atributos de um imóvel se alinham às expectativas do modelo, com escores mais altos indicando melhor alinhamento

com as previsões do modelo.

É essencial reconhecer que o escore CQA só pode ser considerado indicador indireto do apelo de um imóvel comparável quando as condições da venda são compatíveis com seu valor de mercado. Cabe ao engenheiro de avaliações apurar se o valor residual incorpora elementos acima ou abaixo do mercado, fornecendo explicação abrangente e investigação minuciosa da transação de venda. Embora vendas significativamente acima ou abaixo do mercado não sejam frequentes, elas podem ocorrer em circunstâncias como vendas forçadas, inventário, incompetência, conluio etc.

A Tabela 3 apresenta um extrato semirrealista de dados do MLS, junto com os resíduos e os escores CQA, tanto totais quanto por pé quadrado. O número total de imóveis analisados neste estudo foi de aproximadamente 600. É digno de nota que várias variáveis submetidas à regressão não são exibidas. Em outras palavras, os resíduos apresentados são posteriores à consideração, pela regressão MARS, do impacto de outras variáveis, como a data da venda.

Tabela 3: Dados de mercado ordenados por resíduo

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

12.2 Curva CQA-Resíduo

Um gráfico do CQA versus o Resíduo/SF de um bairro resultará invariavelmente na curva característica CQA-Resíduo mostrada na [Figure 2](#). Este gráfico em particular representa dados de Burlingame, CA, por volta de 2022. Curvas semelhantes são em geral observadas ao se plotar escores CQA contra valores de resíduo; contudo, prefiro usar o Resíduo/SF por sua pertinência em contextualizar o CQA em relação à área bruta habitável (GLA) do imóvel avaliando. Multiplicando-se o Resíduo/SF pela GLA, pode-se converter o CQA de um imóvel em um valor residual correspondente.

O engenheiro de avaliações precisa avaliar cuidadosamente se o tamanho do imóvel, especificamente a área bruta habitável (GLA), influencia de modo significativo a magnitude dos resíduos. Essa aferição é crítica porque pode determinar o método mais exato de cálculo dos valores dos imóveis. Por exemplo, eletrodomésticos de cozinha de alto padrão e louças e metais de banheiro luxuosos podem elevar substancialmente o valor residual de uma casa. Contudo, em bairros nos quais o tamanho e o leiaute de cozinhas e banheiros são geralmente uniformes, essas melhorias podem não se correlacionar fortemente com a GLA. Seu impacto sobre o valor residual pode ter menos a ver com o espaço que ocupam e mais com a qualidade que representam.

Inversamente, características e materiais externos caros, que melhoram notavelmente o apelo visual do imóvel, podem ter relação mais direta com o tamanho do imóvel. Casas maiores comportam exteriores mais elaborados, como paisagismo extenso, materiais de cobertura superiores ou terraços amplos, que naturalmente ampliam o apelo geral do imóvel e, por consequência, seu valor residual.

Portanto, é essencial que os engenheiros de avaliações considerem essas dinâmicas ao analisar resíduos. Eles devem não apenas aferir o valor intrínseco agregado por acabamentos e características de alta qualidade, mas também compreender como essas melhorias interagem com as dimensões físicas do imóvel. Essa abordagem abrangente assegura que as avaliações reflitam com exatidão tanto os atributos qualitativos quanto os quantitativos de um imóvel.

Em consequência, recomenda-se plotar tanto a curva CQA-Resíduo quanto a curva CQA-Resíduo/SF e estudar as diferenças, que podem indicar padrões de uma região de mercado. O engenheiro de avaliações pode fazer seu programa calcular ambos os resíduos e tomar alguma média ponderada para calcular o resíduo a ser atribuído ao avaliando. Pode ou não fazer muita diferença.

12.3 As Características da Curva CQA

A curva de estado de conservação, padrão construtivo e apelo (CQA) oferece um modelo sofisticado para compreender como os valores dos imóveis flutuam em função do apelo, do estado de conservação e da dinâmica de mercado dentro de uma dada área. Esse modelo capta a interação de fatores econômicos e físicos que influenciam o mercado imobiliário, refletindo realidades matizadas da avaliação de imóveis nos diferentes segmentos de mercado.

12.3.1 Análise dos Componentes da Curva CQA

1. Decaimento exponencial no lado esquerdo da curva:

À medida que o escore CQA decresce de cerca de 2 para 0, os imóveis exibem características de rápida deterioração. Essa parte da curva é bem representada por uma função de decaimento exponencial, evidenciando com que rapidez um imóvel pode perder apelo e valor sem manutenção e reformas adequadas. A degradação física, como apodrecimento

da madeira, desbotamento de cores e enferrujamento de metais, acelera esse declínio, particularmente em casas que não recebem manutenção regular.

2. Crescimento linear na faixa intermediária:

O grosso do mercado tipicamente se enquadra nesta categoria, na qual o apelo cresce quase linearmente de escores em torno de 2 a 8. Esta seção reflete imóveis geralmente mantidos dentro do padrão, mas que não apresentam necessariamente qualidades ou localizações excepcionais. Representa a ampla classe média habitacional, abrangendo a maior parte do mercado de moradia, no qual as casas são precificadas competitivamente com base em suas características padrão e em seu estado geral.

3. Crescimento exponencial no lado direito da curva:

Aqui, a curva sobe exponencialmente dos escores 8 a 10, espelhando a extremidade de maior apelo do mercado. Os imóveis desse segmento com frequência pertencem a pessoas mais abastadas, que investem não apenas em manutenção superior, mas também em melhorias que ampliam significativamente o apelo e o valor do imóvel. O crescimento exponencial reflete a escassez de imóveis de tão alta qualidade e a forte demanda entre compradores de alta renda. E entenda: uma casa de preço mais baixo pode receber nota muito alta em termos de apelo, ao passo que uma casa grande e cara pode receber nota baixa. Essas dinâmicas são reveladas pelo uso especializado do MARS e geralmente passam despercebidas pela maioria dos avaliadores.

12.3.2 Dinâmica de Mercado e Implicações Sociais

A curva CQA narra implicitamente a história da riqueza relativa e da disparidade econômica em diferentes regiões. Em lugares como o condado de San Mateo, a variação dos valores dos imóveis entre áreas como Burlingame e bairros mais ricos, como Hillsborough, Woodside ou Atherton, ilustra divisões socioeconômicas significativas. Essa disparidade afeta tudo, da manutenção dos imóveis às avaliações de mercado, e em última instância influencia o perfil demográfico dos compradores e as tendências do mercado imobiliário.

Além disso, o caso de uma pessoa relativamente abastada que investe em uma casa de alta qualidade dentro de um bairro de faixa intermediária com o qual se sente à vontade e no qual, por diversas razões, prefere morar, resalta outro aspecto crítico da economia imobiliária — o conceito de imóveis “supervalorizados por sobreconstrução” (*overbuilt*). Tais imóveis, embora muito atraentes, podem não encontrar compradores locais capazes de pagar o preço que alcançariam em uma região de mercado de preços mais altos.

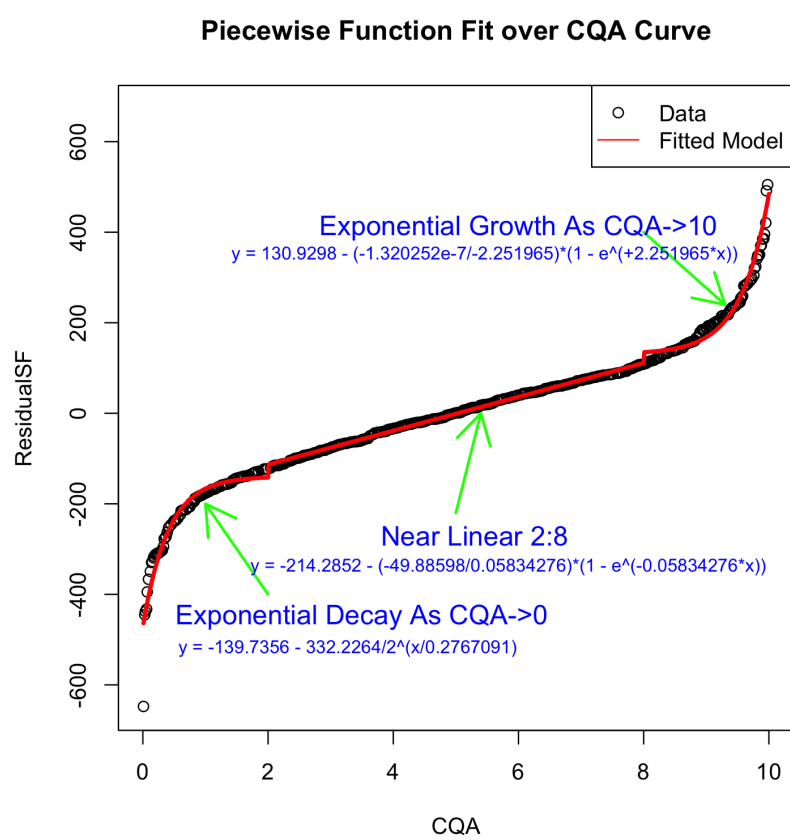
A curva CQA-Resíduo é característica de cada região de mercado e não deve ser negligenciada na sua análise.

12.4 Resíduo do Avaliando

Diferentemente dos comparáveis, o imóvel avaliando não possui resíduo determinado objetivamente; ele precisa ser estimado pelo engenheiro de avaliações (VE). Essa estimativa introduz um grau de subjetividade no protocolo do Residual Constraint Approach (RCA), marcando um ponto crítico em que o juízo pessoal desempenha papel decisivo.

Para estimar o escore de estado de conservação, padrão construtivo e apelo (CQA), o VE analisa meticulosamente os imóveis ordenados, buscando padrões discerníveis dentro da classificação. Essa análise frequentemente envolve exame detalhado das fotos do MLS associadas a cada imóvel. Um exame exaustivo é impraticável para um conjunto de dados grande, como o de 600 imóveis.

Figura 2: Função Característica CQA-Resíduo



Em vez disso, o VE aplica raciocínio prático para aproximar a provável posição do avaliando na classificação geral. Imóveis que exigem reparos extensos, ou “fixers” (imóveis para reforma), são geralmente posicionados perto do fim, ao passo que aqueles com melhores reformas e apelo tendem a se classificar perto do topo.

O VE então adota uma abordagem mais focada, “saltando” dentro da classificação para identificar comparáveis que correspondam de perto ao apelo do avaliando. À medida que a faixa se estreita, o VE presta atenção maior às semelhanças entre os imóveis, particularmente em aspectos como padrão construtivo, estado de conservação, estética do projeto (incluindo elementos como pintura, marcenaria, cantaria e esquadrias), utilidade funcional e outras características pertinentes

É crucial entender que os juízos do VE são fundamentalmente guiados pelas classificações estabelecidas pelo mercado. Isso assegura que os elementos subjetivos do processo fiquem ancorados nas realidades de mercado, buscando uma avaliação justa e exata do imóvel avaliando.

12.5 Conclusão de Valor do RCA

12.5.1 Considerações Gerais

Determinar um escore CQA é decisivo para estabelecer o valor residual estimado de um imóvel, pois esses elementos estão intrinsecamente ligados. Embora seja possível avaliar um imóvel sem CQA, essa métrica fornece contexto valioso ao ilustrar onde o imóvel se situa em comparação com os demais da região de mercado. Por exemplo, um escore CQA de 3,2 indica que o imóvel tem apelo maior do que 32% dos imóveis comparáveis das imediações. Se houver objeções ao escore atribuído, os contestantes precisam consultar as classificações e justificar quaisquer ajustes propostos com base em evidência fotográfica ou em outros dados substanciais. Se o sistema de ordenação for robusto e o engenheiro de avaliações (VE) tiver posicionado o imóvel com exatidão dentro desse arcabouço, é improvável que surjam discrepâncias significativas.

Na fase de elaboração do laudo, tipicamente são selecionados de 6 a 12 dados de mercado recentes e pertinentes para embasar a conclusão de valor. As diferenças entre esses comparáveis e o imóvel avaliando são analisadas, com cada variável sendo ajustada em conformidade. Esses ajustes são então aplicados aos preços de venda líquidos dos comparáveis para se obterem os preços de venda ajustados. Como demonstrado matematicamente, esses preços ajustados devem convergir para o preço de venda estimado do imóvel avaliando.

12.5.2 Revisão, Auditoria, Reclamações

Uma vez determinado o resíduo estimado do imóvel avaliando, os ajustes residuais podem ser calculados para todos os comparáveis. Esses ajustes servem como restrições para decomposições manuais adicionais pelo VE, adaptadas às variáveis CQA específicas consideradas pertinentes. É importante enfatizar que essas decomposições detalhadas são explicativas e não influenciam a avaliação final, como afirmado na Demonstração II.

Quanto a reclamações apresentadas por usuários insatisfeitos com a conclusão de valor da avaliação, um avaliador tradicional tipicamente precisa perder tempo explicando por que este ou aquele imóvel de preço mais alto ou mais baixo não tem o impacto sobre o valor que o usuário esperava. Com o RCA, todas as vendas comparáveis possíveis (normalmente de 100 a 600) são tipicamente avaliadas em precisamente o mesmo valor da conclusão de preço, de modo que é pouco custoso mostrar os ajustes de um imóvel qualquer. Todo o processo RCA é automatizado e objetivo, exceto pela pontuação do resíduo do avaliando. E assim outra vantagem do RCA é que a crítica a imperfeições no juízo subjetivo do VE fica restrita principalmente a essa única área — e pode ser gerida de forma eficaz e eficiente. Além disso, supondo um R2 em torno de

80%, o impacto de um erro de $\pm 10\%$ na estimativa do resíduo do avaliando corresponde a um erro de $\pm 2\%$ na estimativa final de valor.

Em muitos casos, o avaliando se encaixa tão perfeitamente na ordenação dos resíduos dos imóveis que se pode esperar exatidão dentro de $\pm 1\%$.

A soma do resíduo estimado com o preço de venda estimado pelo MARS para o avaliando produz a conclusão final de valor, conhecida como Conclusão de Valor do Residual Constraint Approach (RCA). Esse processo assegura uma abordagem abrangente e transparente da avaliação de imóveis, promovendo confiança na exatidão e na justeza da apuração.

13 Conclusão

O Residual Constraint Approach é uma estrutura metodológica com três contribuições substantivas: uma propriedade estrutural antiancoragem, derivada do ajuste do modelo sobre todo o mercado, e não sobre um pequeno conjunto de comparáveis (Demonstração I); um resultado de invariância que permite à decomposição do resíduo servir como camada explicativa sem perturbar a conclusão de valor (Demonstração II e a observação que a segue); e uma regularidade empírica — a curva característica CQA-Resíduo — que parece recorrer em diferentes regiões de mercado sob a forma característica exponencial-linear-exponencial. As limitações da estrutura são igualmente substantivas: a decomposição do resíduo é invariante, mas não identificadora; a etapa de posicionamento CQA apoia-se em juízo do VE que ainda não foi validado por estudo formal de concordância entre avaliadores; e as alegações de exatidão aguardam comparação com alternativas.

O que este artigo apresenta é uma estrutura metodológica e um protocolo. O que ele ainda não apresenta, mas o programa de pesquisa mais amplo agora em curso apresenta, é um pipeline executável e auditável que operacionaliza o protocolo em software com padrões defensáveis. Esse pipeline — o pacote earthUI no CRAN para a etapa das funções de base do MARS, o pacote glmnetUI para o modelo estrutural regularizado sobre essas bases e o pacote mgcvUI para a suavização espacial — é objeto de artigos companheiros nesta edição e constitui a ponte do RCA-como-método ao RCA-como-entregável. A lacuna entre doutrina e prática na regulação da atividade de avaliação (Craytor, 2026, nesta edição) fornece a razão pela qual essa ponte importa: os produtos da avaliação são economicamente consumidos como insumos prospectivos de risco, e uma estrutura que torne a opinião de valor presente auditável e reproduzível é pré-condição para qualquer engajamento honesto com esse uso prospectivo. O RCA é um movimento nesse projeto maior. Não é o projeto inteiro, e este artigo não pretende que seja.

Referências

- Ambrose, Brent W. et al. (2025). “Do Appraiser and Borrower Race Affect Mortgage Collateral Valuation?” Em: *Review of Finance*. DOI: [10.1093/rof/rfaf046](https://doi.org/10.1093/rof/rfaf046).
- Craytor, William Bert (2026). “The Doctrine–Practice Gap in Real Estate Appraisal: A Structured Account of Functions, Boundaries, and Tensions”. Em: *Valuation Engineer Journal* 1.1.
- Friedman, Jerome H. (1991). “Multivariate Adaptive Regression Splines”. Em: *The Annals of Statistics* 19.1, pp. 1–141.
- Hastie, Trevor, Robert Tibshirani e Jerome H. Friedman (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-0-387-84857-0.
- James, Gareth et al. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics 103. New York: Springer. ISBN: 978-1-4614-7137-0.
- Krivorotov, George e Michael LaCour-Little (2020). “AVM versus Appraisal-Based Underwriting in Refinance Mortgages: The Trade-off Between Noise and Bias”. Em: *SSRN Electronic Journal*. DOI: [10.2139/ssrn.3603327](https://doi.org/10.2139/ssrn.3603327).
- Kuhn, Max e Kjell Johnson (2016). *Applied Predictive Modeling*. Corrected at 5th printing. New York: Springer. ISBN: 978-1-4614-6848-6.
- Milborrow, Stephen (2024a). *Earth: Multivariate Adaptive Regression Splines*. R Package. URL: <https://CRAN.R-project.org/package=earth>.
- (2024b). *Notes on the Earth Package*. Rel. técn. URL: <http://www.milbo.org/doc/earth-notes.pdf>.

- Milborrow, Stephen (2024c). *Variance Models in Earth*. Rel. técn. URL: <http://www.milbo.org/doc/earth-varmod.pdf>.
- Perry, Andre M., Jonathan Rothwell e David Harshbarger (2018). *The Devaluation of Assets in Black Neighborhoods*. Rel. técn. Brookings Institution.
- Perry, Andre M., Jonathan Rothwell, J. Williamson et al. (2024). *How Racial Bias in Appraisals Affects the Devaluation of Homes in Majority-Black Neighborhoods*. Rel. técn. Brookings Institution.
- Property Appraisal and Valuation Equity Task Force (2022). *Action Plan to Advance Property Appraisal and Valuation Equity*. Rel. técn. U.S. Department of Housing e Urban Development.
- Tzioumis, Konstantinos (2016). “Appraisers and Valuation Bias: An Empirical Analysis”. Em: *Real Estate Economics* 45.3, pp. 679–712. DOI: [10.1111/1540-6229.12133](https://doi.org/10.1111/1540-6229.12133).
- Williamson, Jake e Mark Palim (2022). *Appraising the Appraisal: A Closer Look at Divergent Appraisal Values for Black and White Borrowers Refinancing Their Home*. Rel. técn. Fannie Mae.
- Zhang, Wengang (2020). *MARS Applications in Geotechnical Engineering Systems: Multi-Dimension with Big Data*. Singapore: Springer. ISBN: 978-981-13-7421-0.