

Residual Constraint Approach (RCA) : cadre méthodologique et protocole

Traduction française de la version actualisée de l'article initialement publié sur Zenodo le 5 mars 2025. Traduction préparée avec l'assistance d'une IA et revue par la rédaction ; la version anglaise fait foi.

William Bert Craytor, SRA

Volume 1, Numéro 1 · Été 2026

Publié le 1^{er} juillet 2026

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

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

Cet article en libre accès est diffusé selon les termes de la [licence Creative Commons Attribution 4.0](#).

Résumé. Une limite importante des méthodes d'évaluation conventionnelles réside dans leur dépendance à des techniques manuelles dépassées, telles que l'analyse par paires appariées (*matched pairs*), qui ignorent les contributions à la valeur des caractéristiques du bien comme si elles n'existaient pas : la grille de comparaison traditionnelle (*Sales Grid*) ne comporte aucune colonne pour les contributions à la valeur, seulement des ajustements. Or, les contributions à la valeur fournissent la base nécessaire aux contraintes mathématiques qui préviennent la surévaluation et la sous-évaluation par les évaluateurs traditionnels.

Le Residual Constraint Approach (RCA, fr. : approche par contraintes résiduelles) fait progresser le Sales Comparison Approach (SCA, comparable à la méthode par comparaison française) traditionnel en intégrant un processus d'évaluation en plusieurs phases qui associe les Multivariate Adaptive Regression Splines (MARS) à des contraintes mathématiques rigoureuses sur les contributions à la valeur des caractéristiques — en particulier les caractéristiques non mesurées : des variables latentes telles que l'esthétique, la conception, l'état et la qualité, dont les valeurs sont généralement soumises au jugement subjectif et aux biais de l'évaluateur traditionnel.

Les contraintes mathématiques du RCA reposent sur l'application experte de la régression MARS pour estimer les contributions à la valeur des caractéristiques mesurables du bien au prix de vente. Cette estimation ne représente toutefois, en règle générale, qu'environ 80% du prix de vente réel, à l'exclusion des caractéristiques appréciées subjectivement comme la conception, l'utilité fonctionnelle, l'état et la qualité. Dans la San Francisco Bay Area, le résidu — qui représente ces composantes subjectives — s'élève en moyenne à environ 20% de la valeur totale du bien. Le ratio 80/20 reflète l'expérience de l'auteur dans la San Francisco Bay Area dans des conditions de données typiques et ne doit pas être interprété comme une constante invariante d'un marché à l'autre. Dans les marchés où le R^2 atteignable est plus faible, la réponse du cadre méthodologique consiste à identifier et à mesurer les déterminants de la valeur non mesurés — la distance à l'océan dans les communes littorales, l'altitude dans les quartiers à flanc de colline et d'autres variables similaires collectées auprès de sources externes lorsque les données MLS ne les fournissent pas — plutôt que d'accepter cette limite et de publier des bandes d'erreur plus larges. Alors que les résidus MARS sont souvent traités comme des erreurs d'estimation dans d'autres contextes, ils servent, en évaluation immobilière, de mesure indirecte de la valeur des variables latentes. Bien que cela puisse d'abord sembler peu praticable, une démonstration établit que le résidu peut être décomposé de manière significative en composantes descriptives, pourvu que leurs contributions se somment au résidu — sans affecter le résultat final de l'évaluation.

Cette méthodologie est le fruit de deux décennies d'application empirique par l'auteur et a été développée au fil d'implémentations itératives de MARS dans les environnements R et Python.

Mots-clés : évaluation immobilière, évaluation résidentielle, Residual Constraint Approach, MARS, Multivariate Adaptive Regression Splines, Sales Comparison Approach, analyse des résidus, variables latentes, évaluation automatisée

Table des matières

1	Introduction	4
2	Contexte	6
3	Valuation Engineer	9
4	Hypothèses	10
5	Workflow RCA	12
5.1	Types d'analyse MARS	12
5.2	Workflow de traitement	12
5.3	Workflow des données	13
5.3.1	Exigences de données pour MARS	13
5.3.2	Protocoles : listings de ventes conclues	14
5.4	Structure des dossiers de projet	16
5.4.1	Protocoles : exécutions MARS (earth) multiples	17
5.5	Le classeur Excel central	19
5.6	Sauvegardes	19
6	Notation du classeur	20
6.1	Dimensions descriptives	20
6.1.1	La dimension projet (« a »)	21
6.1.2	La dimension des versions d'exécution (« r »)	22
6.1.3	Étapes du flux de travail (« w »)	22
6.1.4	Feuilles du classeur Excel (« s »)	23
6.1.5	Listings de biens (« p ») - lignes de MLSData	23
6.1.6	Variables MLS (« v ») colonnes de MLSData	24
6.1.7	Exemple final de notation	24
6.1.8	Justification	24
6.1.9	Matrices ?	25
6.1.10	Première simplification	25
6.1.11	Seconde simplification	25
7	La Sales Grid RCA	27

8 Variables	31
8.1 Ontologie de l'origine des variables	31
8.2 Ontologie de l'application des variables	32
8.3 Termes du deuxième et du troisième degré	34
8.3.1 Variables d'ajustement	34
8.3.2 Variables d'agrégation	34
8.3.3 Variables spéciales & notation	35
8.3.4 Décomposition des résidus	35
9 Démonstrations fondamentales de la RCA	37
10 Modèle MARS	39
10.1 Fonctions « g » de variables du modèle et d'interaction	39
10.2 Indexation des fonctions g	39
10.3 Dimensions de représentation graphique	40
10.4 L'équation finale du modèle	40
10.5 Nombre de graphiques nécessaires pour représenter un modèle	41
11 Caractéristiques des données	44
11.1 Rôle du résidu	44
11.2 Les notations d'état et de qualité comme régresseurs approximatifs	45
11.3 Erreurs de données	45
11.4 Qu'est-ce que l'exactitude ?	47
11.5 Le bien à évaluer comme anomalie	48
11.6 L'importance du CVR2 face aux données inattendues	49
12 Résidus	50
12.1 Ranking & Scoring (classement et notation)	50
12.2 Courbe CQA-résidu	53
12.3 Caractéristiques de la courbe CQA	53
12.3.1 Analyse des composantes de la courbe CQA	53
12.3.2 Dynamique du marché et implications sociales	54
12.4 Résidu du bien à évaluer	54
12.5 Conclusion de valeur RCA	56
12.5.1 Considérations générales	56
12.5.2 Révision, audit et réclamations	56
13 Conclusion	58

1 Introduction

Cet article présente les fondements logiques et mathématiques d'une approche moderne visant à estimer avec exactitude la valeur de marché (valeur vénale) des biens immobiliers résidentiels et d'autres types de biens. La méthode, élaborée au cours de 20 années d'expérience dans l'application du logiciel Multivariate Adaptive Regression Splines (MARS) à des maisons du nord de la Californie — en particulier dans les communes de la San Francisco Bay Area — constitue un progrès significatif en matière de précision des évaluations. MARS améliore la qualité des évaluations en captant habilement les relations non linéaires entre les caractéristiques du bien et la valeur, tout en identifiant et en hiérarchisant efficacement les variables les plus influentes. Cette capacité permet aux évaluateurs de mettre au jour la dynamique de la valeur dans de nouvelles zones de marché bien plus rapidement que ne le permettent les méthodes traditionnelles. MARS constitue l'ossature de la Residual Constraint Approach (RCA) (fr. : approche par contraintes résiduelles) proposée ici. La méthode RCA doit son nom à la rigueur de son cadre :

- Elle impose des contraintes strictes sur les contributions à la valeur des caractéristiques du bien, tant mesurées que non mesurées :
 1. Toutes les contributions à la valeur dérivées de MARS, y compris le résidu, doivent avoir pour somme le prix de vente net de chaque bien comparable.
 2. Le résidu peut être décomposé en sous-composantes dotées de valeurs assignées, à condition que leur total soit égal au résidu du bien comparable.
- La RCA minimise le biais dans l'estimation des caractéristiques subjectives — telles que l'attrait esthétique, l'état et la qualité — en liant leur valeur totale au résidu, lequel est quantifié avec exactitude lorsqu'un Valuation Engineer (VE) compétent construit le modèle MARS.

Bien que principalement conçue pour l'immobilier résidentiel, où les caractéristiques subjectives influencent sensiblement la valeur de marché, la RCA présente un potentiel d'application à l'évaluation d'autres catégories d'actifs. Des rapports d'évaluation exacts revêtent une importance critique au regard de la valeur estimée du marché immobilier résidentiel américain, soit environ 45 billions de dollars (\$45 trillion). Fannie Mae relève que les évaluations typiques s'écartent d'environ 5% du prix de vente effectif, encore que ce chiffre masque un biais, les évaluateurs subissant souvent des pressions pour s'aligner sur les prix de vente. En outre, dès lors que 60%–70% des rapports d'évaluation servent à des refinancements — où l'on ne dispose pas de prix de vente comparables — la nécessité d'une précision telle que celle offerte par la RCA devient manifeste. L'analyse conduite par la Réserve fédérale en 2012 estimait que 15%–20% des évaluations antérieures à la crise de 2008 s'écartaient de plus de $\pm 10\%$ des prix de vente, une fourchette vraisemblablement encore pertinente aujourd'hui pour les évaluations de refinancement, variable selon le type de bien et le marché. Pourquoi l'exactitude importe-t-elle ? Pour les transactions à faible ratio prêt-valeur (loan-to-value, LTV) — par exemple, un acheteur au dossier de crédit solide versant un apport de \$500,000 sur une maison de \$1 million — les prêteurs peuvent renoncer au rapport d'évaluation, en supposant un risque de saisie minimal. Toutefois, les acheteurs qui se passent d'évaluation risquent de payer trop cher et de ne découvrir les défauts qu'après l'achat. Cela souligne l'intérêt de méthodes d'évaluation robustes telles que la RCA. Quantifier l'exactitude des évaluations demeure difficile. Les chercheurs ne disposent pas de mesures directes des écarts par rapport à la « valeur de marché » et s'appuient sur des indicateurs indirects. Les méthodes d'évaluation traditionnelles, fondées sur le jugement subjectif et sur une compétence variable des évaluateurs, compliquent l'analyse. Parmi les stratégies employées :

1. TZIOUMIS (2016) a étudié 4,782 évaluateurs et 259,436 rapports d'évaluation dans 11 États, constatant qu'un tiers se situait à moins de 1% du prix contractuel, que la moitié se situait

entre 1% et 3% au-dessus, et qu'une minorité se situait en moyenne 5% au-dessus — ce qui suggère un biais à la hausse. L'étude a identifié deux types d'évaluateurs : ceux « à la valeur de marché » (proches du prix contractuel) et ceux « au-dessus de la valeur de marché » (systématiquement plus élevés).

2. KRIVOROTOV et LACOUR-LITTLE (2020) a constaté que 90% des évaluations d'achat atteignaient ou dépassaient le prix contractuel, mettant en évidence le « biais d'ancrage » — la tendance à s'aligner sur une valeur souhaitée — en particulier dans les contextes de refinancement.

Ces études révèlent un biais dans les pratiques d'évaluation, amplifié par les limites des outils traditionnels. Si les évaluateurs étaient uniformément compétents et exempts de biais, leurs évaluations convergeraient ; la surévaluation persistante de la part de certains groupes signale des problèmes systémiques.

Les études sur le biais citées ci-dessus, et le schéma systémique qu'elles mettent en évidence, reflètent un trait plus profond de la pratique traditionnelle de comparaison des ventes que la RCA est conçue pour corriger. La méthodologie traditionnelle opère sur deux canaux au statut épistémique très différent. Sur le canal des caractéristiques mesurables — GLA (surface habitable brute), nombre de pièces, superficie du terrain, âge, localisation — l'évaluateur est censé dériver les valeurs d'ajustement de manière objective, par analyse de paires appariées ou par régression linéaire locale sur les ventes comparables. Les évaluateurs compétents effectuent ce travail, et les ajustements qui en résultent sont au moins approximativement défendables au regard des données. Même là, cependant, la grille traditionnelle ne consigne que des ajustements ; elle ne comporte même pas de colonne pour les contributions à la valeur. Les ajustements traditionnels ne sont pas non plus calculés comme des différences entre contributions à la valeur — ils sont estimés directement, à partir de paires appariées ou de procédures analogues, sans que les contributions sous-jacentes ne soient jamais calculées. Dans la RCA, au contraire, chaque ajustement est la différence entre les contributions à la valeur du bien à évaluer et du bien comparable pour la caractéristique considérée — la structure dont la Démonstration II précise les conséquences. Sur le canal des caractéristiques subjectives — état, qualité de construction, conception, utilité fonctionnelle, esthétique — les procédures de justification disponibles vont du ténu à l'inconsistant. Les ajustements relatifs aux caractéristiques subjectives peuvent être étayés par une analyse de paires appariées portant sur des biens extérieurs à l'ensemble immédiat des comparables, par un renvoi à la pratique reconnue des pairs, ou par référence à l'expérience documentée de l'évaluateur sur le marché considéré. Ces trois procédés sont formellement conformes aux normes professionnelles. Aucun n'atteint la rigueur statistique que le canal mesurable atteint de manière routinière.

Il en résulte une asymétrie procédurale entre ce que les normes exigent et ce que la méthodologie peut fournir. Les normes exigent une justification pour chaque ajustement, qu'il soit mesurable ou subjectif. La méthodologie fournit une justification robuste sur un canal et un éventail de procédures plus faibles sur l'autre, allant d'une paire appariée anecdotique de taille d'échantillon un à la citation de ce que d'autres évaluateurs utilisent habituellement. L'évaluateur honnête et compétent qui effectue un travail soigné de paires appariées pour un ajustement de vue produit une justification sensiblement meilleure que celui qui invoque la pratique des pairs sans fondement, mais un réviseur comparant les deux rapports d'évaluation sur le formulaire standard ne peut aisément les distinguer, car le formulaire consigne la valeur de l'ajustement, et non la rigueur de sa dérivation. La méthodologie tolère cette asymétrie parce qu'aucune procédure plus solide pour les caractéristiques subjectives n'a été, de manière générale, à la disposition des évaluateurs en exercice.

La RCA améliore les deux canaux, mais de manières catégoriquement différentes. Sur le canal mesurable, la régression MARS capte des relations non linéaires, des interactions entre variables et une structure de fonctions de base que l'analyse de paires appariées et la régression linéaire

ordinaire ne peuvent représenter. Ajustées sur l'ensemble complet des ventes qualifiées de la zone de marché — typiquement plusieurs centaines — plutôt que sur les trois à neuf comparables de la grille traditionnelle, les contributions à la valeur dérivées de MARS sont à la fois objectives et sensiblement plus exactes que la méthodologie qu'elles remplacent. Sur le canal subjectif, la RCA fait quelque chose de différent en nature de toutes les procédures de justification traditionnelles. Elle ne tente aucunement d'estimer des valeurs d'ajustement pour les caractéristiques subjectives individuelles. Les contributions de toutes les caractéristiques subjectives — variables latentes — apparaissent collectivement sous la forme du résidu : la différence entre le prix de vente net effectif de chaque bien comparable et son prix prédit par MARS, calculable à partir des données sans intervention de l'évaluateur. La décomposition de ce résidu en composantes nommées — Conception, Utilité fonctionnelle, Vue, État, Qualité, etc. — est contrainte à avoir pour somme le résidu total, et, en vertu de la Démonstration II (Section 9), toute décomposition satisfaisant cette contrainte conduit à la même conclusion de valeur. Le jugement qui subsiste dans le flux de travail est le placement du bien à évaluer dans la distribution ordonnée des résidus des comparables, effectué au regard d'un classement inspectable et explicitement documenté dans le rapport. Un rapport d'évaluation RCA faible demeure possible — un VE inexpérimenté peut produire un modèle MARS surajusté, placer le bien à évaluer de manière peu plausible, ou manquer des anomalies dans les données — mais il est alors faible de manières qu'un réviseur peut identifier en inspectant le modèle et le classement, et non de manières qu'une revue sur formulaire standard ne saurait mettre au jour.

Au-delà du refinancement, les évaluateurs doivent aux propriétaires des évaluations exactes, en particulier pour les achats de grande valeur où les erreurs peuvent coûter des dizaines, voire des centaines de milliers de dollars. Pourtant, beaucoup rechignent à dépenser \$600–\$1,000 pour un rapport d'évaluation, doutant de sa valeur ajoutée. Les méthodes avancées telles que la RCA, quoique initialement exigeantes en ressources, promettent une précision accrue. Avec le perfectionnement des méthodes et l'expérience acquise, leur efficience — et leur adoption — devraient s'accroître.

2 Contexte

La formation universitaire de l'auteur relève des mathématiques, de l'informatique et de la statistique. De profession, il est ingénieur logiciel, avec une expérience secondaire en évaluation immobilière et en comptabilité. Il pratique l'évaluation immobilière, de manière intermittente, depuis 2002 et applique MARS depuis environ 2004 pour déterminer les ajustements de biens et les tendances de prix dans l'immobilier résidentiel.

Évaluer avec MARS constitue une « expérience approfondie » de l'évaluation immobilière. MARS génère des modèles qui reflètent, plus finement que toute autre méthode statistique, à la fois les variables qui contribuent à la valeur de marché (valeur vénale) et la manière dont les variations de ces valeurs influent sur le prix de vente. Ainsi, en évaluant avec MARS, l'évaluateur examine en permanence les relations causales entre les valeurs des variables et les prix de vente telles que générées par les modèles MARS.

Par souci de transparence, la méthodologie RCA (fr. : approche par contraintes résiduelles) s'applique de manière optimale aux zones résidentielles présentant les caractéristiques suivantes :

1. Des quartiers complexes, dont les biens diffèrent sensiblement les uns des autres
 - Maisons anciennes
 - Qualités de rénovation variées
 - Grande diversité d'architectures et de conceptions

- Variation des attributs des biens, tels que la GLA (surface habitable brute), la superficie du terrain, l'altitude et la vue.
- 2. Des conditions de marché complexes
 - Conditions de marché évoluant rapidement
 - Nombreux types d'acheteurs différents
- 3. De bonnes sources de données : le protocole RCA repose sur l'exploration de données (*data mining*) et requiert des données abondantes et de qualité. Bien qu'il puisse être utilisé dans des zones où les données sont rares, le temps et le coût nécessaires pourraient ne pas en valoir la peine.
- 4. Des estimations exactes de la valeur de marché sont recherchées. La RCA devrait, dans la plupart des cas, pouvoir fournir une exactitude de $\pm 1-2\%$. Le rapport du Valuation Engineer devrait commenter les conditions d'exactitude de la conclusion de valeur finale. Si les données ou les informations font défaut, ou si aucune régularité cohérente du marché n'a été mise au jour par l'exploration de données, l'exactitude en pâtira et le client devra être informé de la situation. Plus précisément, un R^2 (coefficient de détermination) élevé et un CVR2 élevé (R^2 validé de manière croisée, calculé en partitionnant de façon répétée le jeu de données MLS en partitions aléatoires d'entraînement et de test), un modèle MARS pertinent et une structure marquée des caractéristiques CQA (état, qualité, attrait) des biens dans les résidus classés constituent les principaux moyens d'apprécier l'exactitude et la fiabilité de la conclusion de valeur.

Le développement de ces protocoles RCA au cours des 20 dernières années et plus repose sur la disponibilité de MLSListings Inc. (Sunnyvale, CA) et d'autres MLS, actuellement accessibles via MLSListings, ce qui représente plus de 95% de l'expérience de l'auteur en matière d'analyse de données. L'expertise de l'auteur découle d'un travail d'évaluation approfondi mené dans 15 comtés principaux de Californie du Nord, complété par une expérience dans 10 comtés secondaires.

La méthodologie RCA est la plus efficace sur les marchés répondant à deux critères clés :

1. Une riche disponibilité de données d'évaluation de haute qualité
2. Des zones résidentielles complexes et hétérogènes, caractérisées par :
 - Des styles architecturaux variés
 - Des degrés divers de rénovation des biens
 - Des schémas de développement s'étalant sur plusieurs décennies

Expérience des comtés californiens

Les affirmations formulées dans le présent article reposent sur l'expérience de l'évaluation de biens dans les comtés suivants de Californie du Nord :

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

Les zones ci-dessus recouvrent pour l'essentiel la Silicon Valley et les comtés environnants, en s'étendant vers des zones plus rurales, notamment la Central Valley et les contreforts de la Sierra. Il convient de mentionner que l'auteur a toujours eu accès à des données abondantes et

relativement exactes par rapport à celles dont disposent de nombreux évaluateurs dans d'autres régions moins développées des États-Unis.

Les données MLS utilisées sont certifiées RESO et conformes au RCF (RESO Common Format). Malgré la certification RCF, des données d'importance généralement secondaire sont parfois manquantes ou erronées pour certains biens, en raison d'erreurs de saisie initiales du Tax Assessor (administration d'évaluation fiscale) ou d'erreurs de saisie MLS ultérieures des agents immobiliers. Ces erreurs suivent souvent un schéma, tel que l'absence pure et simple de saisie lorsque la caractéristique n'existe pas. Par exemple, au lieu de saisir 0 pour « carport spaces » lorsqu'il n'y a pas de carport, rien n'est saisi. Dans les zones où les carports sont rares, l'analyste peut substituer 0 à l'absence de donnée afin d'obtenir un ajustement utile pour les carports. Bien que R/earth et d'autres implémentations de MARS puissent gérer les données manquantes, il appartient au Valuation Engineer (VE) de décider de la manière de les traiter.

Les données de MLSListings sont également disponibles auprès des MLS de certains autres États, de la Floride à Hawaï. Cela peut être utile pour tester le bon fonctionnement des protocoles et du code dans différents États.

3 Valuation Engineer

Définition 1 : Valuation Engineer

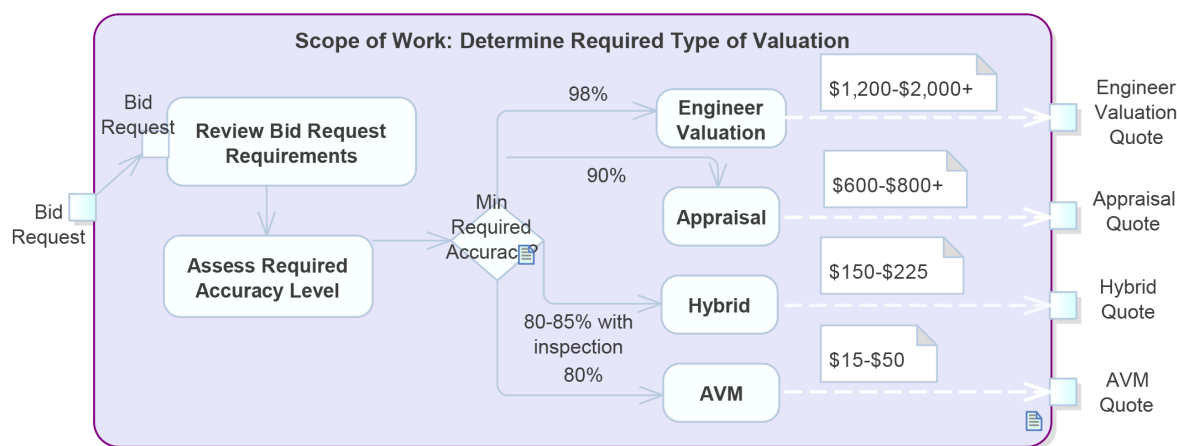
Afin d'éviter toute ambiguïté, le terme « Valuation Engineer » (ou « VE » en abrégé) désigne une personne possédant l'expertise d'un évaluateur agréé, associée à des connaissances, des compétences et une expérience avancées dans l'utilisation de l'exploration de données, de la statistique, de la programmation, de protocoles détaillés et de l'intelligence artificielle en vue de déterminer la valeur de marché de biens complexes. Il importe également de reconnaître que l'évaluation est une tâche qui nécessite une forme d'agrément en raison du potentiel avéré de fraude.

Les prêteurs proposent généralement plusieurs options d'évaluation :

1. Les modèles d'évaluation automatisée (*Automated Valuation Models*, AVM) - l'option la moins coûteuse
2. Les évaluations hybrides - recourant à des inspecteurs non agréés sous la supervision d'un évaluateur agréé
3. Les évaluations standard - réalisées par des évaluateurs agréés

Les évaluations de niveau ingénieur sont absentes des considérations habituelles en matière de travaux d'évaluation. Il existe tout simplement peu d'évaluateurs disposant des compétences nécessaires pour les réaliser. Le besoin devrait pourtant exister. Elles peuvent offrir une exactitude nettement supérieure et garantir presque une objectivité assez bonne grâce à l'emploi de la méthodologie RCA (MARS et contraintes).

Pour les biens conformes situés dans des lotissements récents, les évaluations standard suffisent généralement. En revanche, les évaluations d'ingénieur sont recommandées pour les maisons anciennes situées dans des zones présentant un historique de rénovation important s'étendant sur plus de 30 ans. Le choix entre ces méthodes devrait suivre un processus décisionnel structuré fondé sur les caractéristiques du bien et les conditions de marché.



4 Hypothèses

Hypothèse 1 : Zone géographique & type de bien

Le champ du présent article porte principalement sur les maisons individuelles et les copropriétés (condominiums) des zones métropolitaines et urbaines de Californie. Bien que la Sales Comparison Approach (comparable à la méthode par comparaison française) s'applique à d'autres types de biens et à d'autres régions géographiques, les approches par le revenu et par le coût jouent généralement un rôle plus dominant lors de l'évaluation de terres agricoles, d'immeubles collectifs et d'immobilier commercial. Dans de tels cas, l'approche RCA peut nécessiter des modifications ou n'avoir qu'une applicabilité limitée.

Hypothèse 2 : Stockage des données

Les formats CSV et Excel sont massivement utilisés par les systèmes MLS et sont supposés constituer le principal moyen d'alimenter en données les programmes RCA, bien que des flux de courtiers puissent être utilisés pour alimenter en continu des bases de données SQL. La sauvegarde et le stockage intermédiaire recourent souvent à des bases de données SQL, dont deux jouent un rôle majeur : SQLite et PostgreSQL, libres et open source. Des fichiers texte ou RTF sont également utilisés pour des fragments de rapport, et des fichiers PNG, JPG et TIFF pour les graphiques. Notons que SQL Server de Microsoft est également utilisé par beaucoup, mais l'édition gratuite « SQL Server Express » est très limitée (p. ex. un seul cœur de CPU peut être utilisé) et les autres versions comportent des frais de licence élevés pour un usage commercial. Les bases de données immobilières en Californie peuvent être très volumineuses et les jointures très chronophages ; un ordinateur de 8 cœurs ou plus doté de 128G de RAM peut donc être utile pour les grandes jointures. Les bases de données SQL sont également importantes lorsqu'il faut retrouver tous les biens situés dans une zone délimitée de manière arbitraire. Ainsi, si l'on a par exemple utilisé la classification automatique (« Cluster Analysis ») pour définir des quartiers, on sera enclin à utiliser une base de données SQL. Toutefois, les différentes bases de données possèdent des fonctions intégrées différentes pour le traitement de l'information géographique et ne sont donc pas compatibles entre elles. Pour la présente série d'articles sur la RCA, PostgreSQL sera supposé pour les tâches SQL complexes et SQLite pour les tâches simples.

Hypothèse 3 : « R/earth » est utilisé pour le traitement MARS.

Jusqu'en 2017, le logiciel MARS de Salford Systems constituait une solution abordable à \$270 par an. Cependant, après l'acquisition de Salford Systems par Minitab, MARS a été intégré à leur suite SPM au prix de \$16,000 par an — un tarif prohibitif pour la plupart des petites entreprises. Heureusement, le package « earth » du langage de programmation R (MILBORROW 2024a), maintenu par Stephen Milborrow, a évolué au point d'égaliser, voire de dépasser, les capacités du MARS de SPM de Minitab. Le package earth offre des avantages supplémentaires : il est disponible gratuitement, s'intègre parfaitement à l'écosystème statistique complet de R et peut être utilisé depuis Python au moyen de Pandas, offrant une excellente flexibilité pour différents flux de travail. Les discussions relatives à MARS supposent l'utilisation spécifique de R/earth.

Hypothèse 4 : Notation

Les données brutes issues de sources externes telles que les fichiers CSV et les feuilles de calcul Excel sont représentées à l'aide de caractères latins standard. Cela inclut les données de biens du Multiple Listing Service (MLS), les paramètres de configuration de MARS (Multivariate Adaptive Regression Splines) utilisant le package R/earth, ainsi que les données propres au projet. Cette notation s'étend à leurs représentations correspondantes sous forme de data frames en R ou en Python.

Lorsqu'il est question des *caractéristiques des biens* et de leurs divers types et sous-ensembles, la notation passe aux lettres grecques pour une précision et une clarté analytique accrues. Si cette notation mathématique peut paraître plus rigoureuse, elle sert un objectif crucial : distinguer les caractéristiques abstraites des biens de leurs représentations sous forme de données brutes. Par convention, les caractères grecs désignent systématiquement les caractéristiques des biens, tandis que les caractères latins indiquent les structures de données de type feuille de calcul. Par exemple, le symbole M représente la matrice de données MLS, organisée avec les biens en lignes et leurs caractéristiques correspondantes en colonnes.

Hypothèse 5 : Langages de programmation

Dans ce domaine, R et Python avec Pandas sont les langages dominants, C ou C++ étant utilisés lorsque la vitesse est importante. R/earth est lui-même écrit en R et en C.

Hypothèse 6 : Indexation

L'indexation peut varier selon les langages de programmation et les mathématiques, ce qui peut prêter à confusion :

- Les mathématiques et R commencent l'indexation à 1.
- C, C#, C++ et Python commencent à 0.

Dans le présent article, par souci de cohérence et de simplicité, toute indexation commence à 0 (comme en Python ou en C). Voici comment cela fonctionne :

- Pour les listes, matrices ou data frames :
 - L'indice 0 est réservé au « bien à évaluer » (le bien principal étudié).
 - Exemple : dans un tableau dont les lignes sont des biens, la ligne 0 est le bien à évaluer et la colonne 0 contient les identifiants uniques de tous les biens.
 - Conseil pratique : disposer les biens en plaçant le bien à évaluer en premier (ligne 0), puis lister les autres de la date de vente la plus récente à la plus ancienne, en les numérotant séquentiellement (0, 1, 2, etc.).
- Lorsqu'il n'y a pas de bien à évaluer :
 - Si une structure de données ne contient pas de biens (p. ex. seulement des variables ou des fonctions), l'indice 0 est omis et l'indexation commence à 1.
 - Note particulière pour les utilisateurs de R : R utilisant une indexation à base 1, lors du codage en R des mathématiques du présent article, il convient d'ajouter 1 à tous les indices pour passer de la base 0 à la base 1.

5 Workflow RCA

Comprendre le workflow RCA est essentiel pour une analyse efficace. Le processus suit une séquence structurée de tâches analytiques et de points de décision

5.1 Types d'analyse MARS

Le workflow d'une analyse MARS dépend du point de savoir si seule une analyse de marché est requise, ou si la valorisation ou l'évaluation immobilière d'un bien donné est demandée. On notera qu'un rapport d'évaluation portant sur un bien particulier exige en général une analyse de sa zone de marché.

- Analyse de marché : elle porte sur des zones étendues telles que des comtés, des villes ou des régions définies. Cette approche ne fait intervenir aucun bien à évaluer particulier et n'exige donc aucune conclusion de valeur. Le workflow est centré sur le développement du modèle, sa validation et la production d'analyses à l'appui. Le processus s'achève une fois le modèle optimal identifié et documenté.
- L'analyse d'un bien individuel englobe tous les éléments de l'analyse de marché, auxquels s'ajoute la détermination de la valeur d'un bien particulier. Elle diffère fondamentalement des méthodes d'évaluation traditionnelles. Dans les approches traditionnelles, la valeur est habituellement obtenue en ajustant 3-9 ventes comparables et en calculant une moyenne pondérée. En revanche, la RCA détermine la valeur principalement à partir de deux composantes :
 1. La valeur prédite par le modèle MARS
 2. Le résidu propre au bien estimé (avec le score CQA associé)

Dans la RCA, le Sales Grid sert avant tout à expliquer les différences de valeur entre le bien à évaluer et les comparables en illustrant les contributions des variables. Un avantage clé de la RCA est sa stabilité : si les clients peuvent débattre de la sélection des comparables, la conclusion de valeur demeure constante quels que soient les comparables retenus pour la présentation, puisque toutes les ventes (qu'il y en ait 15 ou 600) sont ajustées au moyen du même modèle. Bien que les ventes plus anciennes soient automatiquement ajustées aux conditions actuelles du marché, les ventes récentes sont généralement préférables. L'efficacité du modèle à capturer les facteurs de prix diminue avec l'éloignement temporel, les caractéristiques non mesurées du marché prenant davantage d'importance sur de longues périodes.

5.2 Workflow de traitement

Le workflow de traitement global peut être décrit par les étapes suivantes :

1. Périmètre de mission.
2. Mise en place des dossiers et fichiers initiaux du projet.
3. Téléchargement des données MLS. Se reporter aux Protocoles 1.1–1.3 concernant le nombre d'enregistrements de biens nécessaires et la fixation de limites sur les plages de dates.
4. Renseignement des feuilles de configuration : Regression Variables, Interactions, Aggregations, Calculations.
5. Exécution de la régression MARS (R/earth) pour créer le modèle de prix.
6. Génération des contributions à la valeur des biens & des résidus (résidu pur et résidu/SF) à partir du modèle MARS.
7. Classement des biens selon le résidu/SF.
8. Création des scores CQA pour chaque vente comparable.

9. En supposant qu'un classement des biens par qualité a été produit et qu'il représente fidèlement l'attrait global des biens, trouver la meilleure position du bien à évaluer dans ce classement et lui attribuer un score CQA, en étudiant les tendances d'ensemble du classement.
10. Attribuer au bien à évaluer le résidu/SF correspondant, soit $\text{résidu/SF} \cdot \text{GLA}$ (surface habitable brute) du bien à évaluer.
11. **Additionner l'estimation de prix MARS et le résidu attribué pour aboutir à la conclusion de valeur du bien à évaluer.**
12. Décomposer le résidu du bien à évaluer en composantes descriptives et en contributions à la valeur.
13. Calculer les contributions à la valeur produites par le modèle pour les comparables & les ajustements, ainsi que tous les sous-totaux et les prix de vente ajustés.
14. Les prix de vente ajustés seront égaux à la conclusion de valeur finale du bien à évaluer.
15. Sélectionner les 6-12 biens les plus similaires pour le Sales Grid.
16. Décomposer les montants de résidu total des comparables en contributions à la valeur des composantes résiduelles, en fixant la valeur totale du résidu à 0 une fois la décomposition achevée.
17. Recalculer tous les sous-totaux et les prix de vente ajustés des comparables du Sales Grid. Vérifier les valeurs.
18. Charger le Sales Grid dans le rapport.

On notera qu'avec la RCA, la conclusion de valeur finale ne découle PAS directement des valeurs des comparables figurant dans le Sales Grid. Avec la méthode RCA, le Sales Grid devient un outil servant uniquement à expliquer pourquoi des biens qui paraissent similaires au bien à évaluer se vendent plus ou moins cher que la conclusion de valeur finale ou que la valeur d'expertise du bien à évaluer. Il faut envisager les choses autrement : la conclusion de valeur du bien à évaluer dépend de l'estimation de prix du modèle MARS et du résidu estimé par le Valuation Engineer. Mais ces deux quantités dépendent très fortement de l'ensemble des biens entrés dans la régression R/earth , laquelle a produit le classement des résidus des biens ayant déterminé le score CQA du bien à évaluer et, partant, son résidu, par le placement dans le classement effectué par le Valuation Engineer. Les ajustements du Sales Grid sont en réalité déterminés par les contraintes mathématiques imposées par la RCA ; bien que la décomposition du résidu dépende dans une certaine mesure du jugement du VE, les totaux des contributions à la valeur issues de la décomposition doivent toujours correspondre au résidu associé. Nous savons par la Démonstration 2 que, si les valeurs effectives des composantes résiduelles peuvent varier selon le jugement du VE, leurs totaux pour un bien donné sont entièrement contraints et ne modifieront donc pas le prix de vente ajusté du bien associé, quelle que soit la manière dont le VE modifie la décomposition du résidu.

5.3 Workflow des données

On notera que le présent article ne présente qu'un workflow de haut niveau pour la méthode RCA. D'autres articles approfondiront les différentes étapes et sous-étapes de l'analyse RCA.

5.3.1 Exigences de données pour MARS

Il conviendra de spécifier une date de début pour les transactions de vente et une date de fin fondée sur la date d'effet du rapport d'évaluation ou de l'évaluation. Cela prend habituellement en considération le nombre estimé de variables de régression et d'interactions qui seront nécessaires pour chaque degré d'interaction supérieur à 1.

5.3.2 Protocoles : listings de ventes conclues

Protocole 1.1 : Détermination du nombre requis de listings MLS de ventes conclues

Ce protocole décrit comment calculer le nombre minimal recommandé de listings de ventes conclues à fournir en entrée de la régression MARS. Il faudra commencer par une estimation approximative, exécuter MARS en passes de premier, de deuxième et éventuellement de troisième degré afin de générer le modèle initial, puis examiner le nombre de lignes selon les règles suivantes pour déterminer si des lignes supplémentaires sont nécessaires. Ce protocole suppose que les variables indépendantes sont indépendantes à 100%, ce qui est rarement le cas en immobilier ; par exemple, la GLA et les nombres de pièces sont généralement fortement corrélés, et la GLA est également susceptible de présenter une corrélation avec la superficie du terrain. On touche ici au calcul des « degrés de liberté (DOF) ». Chaque variable véritablement indépendante reçoit un degré de liberté. Mais lorsqu'il existe des dépendances entre les variables dites indépendantes, un calcul exact des DOF exige une analyse qui dépasse le cadre du présent article. Il faut simplement garder à l'esprit qu'en raison d'une colinéarité probable entre les variables, les estimations ci-dessous sont sans doute plutôt élevées.

1. Pour les termes du premier ordre (sans interactions), la règle des ~ 15 lignes par variable donne de bons résultats
2. Pour chaque interaction à deux variables autorisée, il convient d'ajouter environ 15-30 biens
3. Pour les interactions d'ordre supérieur (3 variables ou plus), chacune devrait disposer de 30 biens ou plus.

Il ne s'agit toutefois pas d'une règle absolue. Le nombre réellement nécessaire peut dépendre :

- Du rapport signal/bruit des données
- De la complexité des relations que l'on cherche à modéliser
- De la variabilité de la variable réponse
- De la qualité des données

Exemple :

Ainsi, si l'on dispose de :

- 10 variables principales
- 5 interactions à deux variables autorisées
- 2 interactions à trois variables autorisées

Le nombre minimal de ventes peut alors se calculer ainsi :

$$\begin{aligned}
 N &= (10 \text{ variables} * 15 \text{ lignes/variable}) + \\
 &\quad (5 \text{ interactions-à-deux-variables} * 22 \text{ lignes/interaction-à-deux-variables}) + \\
 &\quad (2 \text{ interactions-à-trois-variables} * 30 \text{ lignes/interaction-à-trois-variables}) \\
 &= 320 \text{ lignes}
 \end{aligned}$$

On peut également procéder comme suit : après avoir exécuté R/earth sur un jeu de données, on peut utiliser le nombre de fonctions de base du modèle, plus un, comme estimation des degrés de liberté, puis multiplier ce nombre par 20-30 pour obtenir le nombre d'enregistrements d'entrée nécessaires.

Protocole 1.2 : Les périodes de vente doivent couvrir une évolution minimale des goûts des acheteurs

Lors de la collecte de données historiques sur les biens, il n'existe pas de limite temporelle stricte dès lors que les évolutions des préférences du marché peuvent être capturées par des variables appropriées. Les points clés suivants sont à considérer pour traiter l'évolution architecturale du marché étudié : les changements de styles architecturaux dominants (comme le passage du style California Ranch au style méditerranéen) peuvent être modélisés à l'aide :

- D'identifiants de lotissement
- De codes de zone MLS
- De plages de dates de construction

Toutefois, si ces évolutions de style ne peuvent être capturées de manière adéquate par des variables spécifiques, on se heurte à une contrainte importante : la date de vente étant habituellement utilisée pour l'ajustement aux conditions de marché, l'utiliser simultanément pour contrôler les préférences architecturales confondrait ces deux effets distincts. Dans de tels cas, l'approche à privilégier consiste :

1. Soit à limiter la collecte de données à la période postérieure à l'introduction du style architectural le plus récent
2. Soit à restreindre la sélection des comparables aux lotissements partageant le même style architectural

On garantit ainsi la cohérence des caractéristiques des biens dans l'analyse, tout en permettant un ajustement correct aux conditions de marché au fil du temps. Alternativement, du moins dans ce cas, à supposer que le style architectural puisse être extrait des données, on pourrait considérer l'interaction entre ces deux variables. Une telle extraction n'est cependant pas toujours possible.

Protocole 1.3 : En zone de ventes rares, faire au mieux avec les données disponibles

Lorsque l'on travaille avec la modélisation MARS sur des données de ventes, environ 15 transactions sont nécessaires pour une analyse significative. Si l'on dispose de moins de ventes (par exemple 5 biens), on peut dupliquer chaque enregistrement deux fois afin d'atteindre le seuil de 15 biens. Considérations importantes pour cette approche :

1. N'utiliser qu'une régression au premier degré afin de réduire le surajustement
2. Ignorer ou omettre la validation croisée, le jeu de données étant trop limité
3. Bien que MARS surpasse vraisemblablement les méthodes de paires appariées ou de régression linéaire standard dans ce scénario, il faut être conscient que la duplication des données affecte la validité des mesures d'incertitude et de la significativité statistique

La meilleure solution à long terme consiste à étendre le jeu de données avec des ventes réelles supplémentaires. Toutefois, s'il faut procéder avec des données limitées, la duplication des données constitue une solution temporaire praticable, dont il convient de reconnaître les limites statistiques.

5.4 Structure des dossiers de projet

Il n'est pas inutile d'aborder la structure des dossiers de projet. Il est judicieux de disposer d'un dossier distinct pour les analyses de marché et pour les rapports d'évaluation de biens individuels. Ces deux dossiers présentent une structure de sous-dossiers similaire :

```

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
│   │       │           └── ...

```

Le traitement de l'étape 1 crée le dossier de version avec ses sous-dossiers. Il copie le fichier `MLS_Original.xlsx`, bien que ce classeur ne *devrait* jamais changer, mais refléter exactement ce qui a été téléchargé depuis le MLS à une date et une heure données ; il reste toutefois possible, sur un projet de plus longue durée, qu'un nouveau téléchargement MLS soit nécessaire. Ensuite, cette étape copie dans le dossier de version le classeur `MLS.xlsx` étendu, avec ses paramètres de configuration. Il peut exister de nombreux dossiers de version, chacun créé pour une exécution complète de MARS.

Puis, les étapes suivantes génèrent des ajouts de colonnes, des tris et d'autres modifications

du classeur, et écrivent une version mise à jour sous le nom `MLS_Stage(N)` dans le dossier de données du projet lorsque le rapport d'évaluation est achevé. Répétons-le : le classeur racine `MLS.xlsx` subit des changements constants, de sorte que la copie de version est une mutation continue du classeur `MLS` d'origine, avec des suppressions et des ajouts de variables, ainsi que d'autres modifications. Le code de traitement R ou Python s'exécutera très probablement avec, en entrée, le classeur `MLS.xlsx` du dossier `Data`. Nous disposons ainsi toujours d'une copie de toutes les données sous-jacentes à chaque version de l'exécution RCA.

Autant de versions sont créées que nécessaire, jusqu'à ce qu'aucune amélioration ne soit plus possible ; la meilleure version est alors retenue comme version finale. On peut modifier le nom du dossier en lui ajoutant le suffixe « `_Final` », afin que cela apparaisse clairement lors du parcours du répertoire. Certaines versions plus anciennes, présentant des modèles intéressants ou concurrents, devraient généralement être conservées. Chaque dossier de version contient un certain nombre de fichiers de texte, de diagrammes et de graphiques pouvant être intégrés dans un rapport. Le classeur final `MLS_Stage5.xlsx` contient autant de feuilles que nécessaire pour le rapport. Il peut notamment comporter une feuille de calcul destinée au transfert vers un Sales Grid Fannie Mae. Il contiendra aussi nécessairement les calculs ou la décomposition sous-jacents aux agrégations. C'est potentiellement la partie la plus chronophage de l'approche RCA - la quête du modèle parfait dans une zone de marché difficile.

Tout ce qui est nécessaire pour recréer les Sales Grids doit être copié dans le dossier de version, y compris le code R ou Python utilisé - si ce code fait l'objet de modifications fréquentes.

5.4.1 Protocoles : exécutions MARS (earth) multiples

Protocole 2.1 : Exécutions en surajustement pour étudier finement les facteurs de prix d'une zone de marché

Lors des premières exécutions de `earth`, le surajustement doit être employé de manière stratégique pour découvrir les facteurs de prix sous-jacents. Par exemple, une zone de marché représentée par 800 ventes peut ne contenir que 3 biens dotés d'un atelier, et dans ces cas isolés l'atelier peut avoir un impact majeur sur le prix de vente final. Une exécution limitée à la passe avant produira un modèle surajusté qui montre la contribution à la valeur approximative de ces ateliers. Cependant, comme il n'y a que 3 ventes de ce type sur 800, les termes relatifs à l'atelier seront éliminés lors de la passe arrière, et à plus forte raison en validation croisée. La valeur que la passe avant seule attribue aux ateliers peut être consignée et reportée dans l'analyse des variables latentes, pour être utilisée pour tout bien doté d'un atelier figurant dans le Sales Grid. De tels modèles surajustés fournissent également des estimations détaillées pour de nombreuses autres caractéristiques à données rares qui seraient éliminées par l'élagage de la passe arrière ou par les exécutions en validation croisée.

Chacune de ces exécutions exploratoires doit être sauvegardée dans son propre dossier de version : les contributions à la valeur consignées constituent des éléments que le VE voudra citer plus tard lors de la décomposition des résidus. Il faut bien entendu comprendre que le VE doit en dernier ressort s'appuyer sur l'élagage et la validation croisée pour rendre le modèle final robuste face à des biens que la régression n'a jamais vus — tel un bien à évaluer qui apparaît ultérieurement. Les exécutions surajustées sont des instruments d'étude ; elles ne constituent jamais le modèle final.

Protocole 2.2 : Exécutions multiples en validation croisée sous différents paramètres de earth()

Les zones de marché diffèrent par leur structure et leur taille, et le VE doit généralement expérimenter une série d'exécutions : pour comprendre quelles variables influent le plus sur le prix ; pour juger si l'ensemble existant de variables mesurées estime la valeur de manière adéquate, ou si de nouvelles données doivent être ajoutées (distance à l'eau, altitude, distance aux centres commerciaux, etc.) ; et pour découvrir de nouveaux types d'interactions tout en écartant d'autres comme inefficaces ou trompeuses. Au final, le VE recherche un bon R^2 sur les données d'entraînement et le meilleur CVR2 possible sur les données de test. Y parvenir peut exiger de collecter de nouvelles données, d'améliorer l'exactitude des données existantes, ou d'ajuster les paramètres de earth() à la taille et au caractère de la zone de marché (p. ex. au regard des degrés de liberté disponibles — voir le Protocole 1.1).

C'est précisément pourquoi la structure de projet présentée ci-dessus prend en charge de nombreux dossiers de versions d'exécution : chaque modèle candidat est conservé avec ses paramètres et son R^2 /CVR2, de sorte que la trajectoire d'une exécution à l'autre peut être comparée. Une version dont le R^2 grimpe tandis que son CVR2 stagne ou chute ajuste du bruit plutôt que le signal du marché, et c'est l'historique des versions qui rend ce phénomène visible.

Protocole 2.3 : Expansion Prolog des remarques et des colonnes descriptives de caractéristiques

Bien que cette fonctionnalité soit encore dans une phase quelque peu expérimentale, earthUI est désormais livré avec un moteur Prolog intégré (vProlog) et utilise des grammaires à clauses définies (Definite Clause Grammars, DCG) pour analyser les colonnes de texte libre du MLS — Public Remarks avant tout, mais aussi les remarques d'agents et les listes de caractéristiques délimitées par des virgules, telles que les caractéristiques des pièces ou les structures existantes. Cette expansion transforme des expressions comme « owned solar », « 3 car garage » ou « 40x50 workshop » en colonnes de caractéristiques ordinaires pouvant être proposées à la régression, et les caractéristiques à données rares ainsi récupérées alimentent directement l'analyse des variables latentes (voir la sous-section ultérieure sur les notations d'état et de qualité comme régresseurs approchés ; la fonctionnalité elle-même est documentée dans l'article compagnon consacré à earthUI dans ce numéro). L'expansion textuelle multiplie l'ensemble des prédicteurs candidats, ce qui impose de multiples exécutions earth/RCA à des fins de test : les colonnes issues de l'expansion doivent être essayées dans le modèle et hors du modèle, consolidées en comptages ou en booléens dérivés lorsqu'elles sont colinéaires, et vérifiées au regard du CVR2 afin de confirmer qu'elles portent un signal de marché plutôt que du bruit. Chaque expérience de ce type doit être traitée comme son propre dossier de version d'exécution, exactement comme dans les Protocoles 2.1 et 2.2. Et puisque la grammaire d'analyse est elle-même en cours de développement, le VE devrait vérifier les colonnes extraites au regard du texte source avant de s'y fier dans un modèle.

5.5 Le classeur Excel central

Puisque la plupart des services MLS téléchargent les données des biens vers des feuilles de calcul Excel CVS, Excel est probablement le moyen le plus commode de stocker les données et les divers paramètres de configuration. Il existe un classeur Excel central pour chaque projet, nommé `MLS.xlsx`, qui comporte une feuille appelée « `MLSData` » pour stocker les listings de ventes conclues téléchargés depuis le MLS, ainsi que plusieurs feuilles de configuration et de paramètres :

```
Classeur : MLS.xlsx/  
├─ Feuille : Project Data  
├─ Feuille : MLSData  
├─ Feuille : Regression Variables  
├─ Feuille : Calculations  
├─ Feuille : Allowed Interactions  
├─ Feuille : Aggregations  
└─ Feuille : (paramètres d'exécution MARS)
```

Les données d'origine sont stockées dans un classeur nommé par exemple `MLS_Original.xlsx`. Une copie doit en être faite sous un autre nom, disons « `MLS.xlsx` », qui sera chargée dans le programme RCA en R ou en Python. Ces deux classeurs sont stockés dans le dossier `Data`. Or, le classeur `Data/MLS.xlsx` passera vraisemblablement par de nombreuses itérations d'exécutions, au fil d'essais de différentes combinaisons de variables, de paramètres MARS et de modifications d'autres réglages. Chaque itération crée un nouveau dossier d'entrée/sortie nommé avec la date et l'heure, au format `VER_YYYYMMDDHHMMSS`. Chaque dossier contiendra une copie du classeur `MLS.xlsx` d'entrée ainsi que les classeurs Excel générés pour chacune des quatre étapes de traitement (étapes 2–5), de même que les documents texte, diagrammes et graphiques produits en sortie. On notera que l'étape 5, l'étape finale, peut contenir un certain nombre d'autres feuilles pour le rapport, selon le type de client et le SOW (périmètre de mission).

5.6 Sauvegardes

Lors de l'écriture de feuilles de calcul Excel nouvelles ou modifiées dans les dossiers de version du projet, des sauvegardes peuvent également être effectuées vers des bases de données SQLite ou PostgreSQL dans le dossier du projet. Les fichiers de base de données risquent moins d'être corrompus involontairement au cours des périodes ultérieures, lors de consultations ou d'utilisations de référence.

6 Notation du classeur

6.1 Dimensions descriptives

Un classeur « M » comporte environ six dimensions descriptives :

1. Rapports d'évaluation « a »
2. Versions d'exécution « r »
3. Étapes du flux de travail « w »
4. Feuilles « s »
5. Listing de bien « p »
6. Ensemble de variables « v »

Ainsi, une cellule individuelle d'un classeur M donné peut être décrite comme suit :

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

(Remarque : « * » désigne un indice valide quelconque)

ou :

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

où :

- n_p = # de lignes (selon la feuille)
- n_v = # de colonnes (selon la feuille)
- n_r = # de versions d'exécution
- n_m = # d'étapes du flux de travail
- n_g = # de projets d'analyse
- n_h = # de feuilles du classeur

Il convient de comprendre que « W » désigne l'intégralité du classeur Excel, composé de six feuilles de données principales possédant chacune leur propre table, lesquelles, lors de l'exécution du programme R ou Python associé, sont chargées dans des data frames internes distincts. R comme Python disposent de fonctions internes capables, à partir d'un nom, de déterminer à quelle ligne ou colonne du data frame celui-ci appartient. Par exemple, si dans l'indice v_j , $v =$ (« GLA », « LotSize », « BathRms », « BedRms », ...) et $j=0$, alors $v_j =$ "GLA". R comme Python sauront résoudre en interne « GLA » vers la colonne dans laquelle cette variable est stockée, qui pourrait être, par exemple, la colonne 15 du data frame associé. Ainsi, dans le présent article, nous indexons nos tableaux de noms, lesquels transmettent un nom au programme en cours d'exécution, qui ré-indexe ce nom en l'indice réel de ligne ou de colonne du data frame interne. C'est un peu compliqué. Mais, pour être précis, cela est nécessaire et des plus commodes. Il s'agit d'une « séparation des responsabilités » entre le rôle d'analyste et le rôle de développeur, bien qu'en réalité le Valuation Engineer (VE) puisse assumer les deux rôles. Notons qu'un « data frame » de R s'appelle « dataframe » en Python.

Notons également que la feuille des paramètres d'exécution de MARS n'est pas une dimension de W, car W se rapporte à la version des données fournies en entrée à MARS, indépendamment des paramètres qui déterminent la manière dont MARS/earth traite ces données.

Par souci de clarté, une notation multidimensionnelle complexe du classeur est présentée, qui correspond plus directement aux data frames internes de R ou Python dans lesquels les feuilles de données du classeur sont chargées. Il s'agit d'une notation « lourde ». Mais lorsque nous aborderons les démonstrations, etc., nous nous délesterons d'une grande partie de ce bagage en formulant des hypothèses sur les objets manipulés. Néanmoins, cette notation aide à obtenir d'emblée une meilleure vue d'ensemble du flux de travail et des données.

Commençons par la notation de l'ensemble des classeurs que nous allons traiter.

6.1.1 La dimension projet (« a »)

Une fois qu'un Scope of Work (SOW, « périmètre de mission ») a été approuvé, un dossier de projet doit être créé pour stocker les données d'entrée et de sortie, y compris les paramètres de configuration, les graphiques de sortie, les diagrammes, les feuilles de calcul et les extraits de rapport. Le VE devrait identifier tous les projets par un simple ID séquentiel, tel qu'un entier, de sorte que la disparition éventuelle d'un rapport soit manifeste. Si des projets sont annulés, il devrait exister un moyen cohérent d'indiquer l'annulation du projet au sein du dossier de projet concerné, mais celui-ci ne devrait jamais être supprimé. Nous supposons que les ID de projet commencent à 0 et sont incrémentés de 1, bien que d'autres possibilités existent. Nous supposons en outre que tous les ID de projet valides sont stockés dans un ensemble ordonné « a » créé avec le programme de traitement.

Les données du projet d'évaluation figurent dans la feuille de calcul projet du classeur, mais pourraient également être stockées dans une base de données SQL centrale.

Voici les champs de projet suggérés :

1. ID local : un identifiant entier court généré séquentiellement pour l'entreprise. Il constituerait une excellente clé primaire pour le stockage - et pourrait être généré automatiquement par la base de données à chaque nouveau projet.
2. GUID global : si vous le souhaitez, vous pouvez créer un GUID (Globally Universal ID) à l'aide de

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

3. Localisation du bien (analyse de marché) ou adresse
4. Date d'effet de l'évaluation
5. Date de la commande
6. Date de l'inspection
7. Date du rapport original
8. Date du premier rapport mis à jour
9. Date du deuxième rapport mis à jour
10. Date du troisième rapport mis à jour
11. Nom du client
12. Adresse du client
13. Nom du prêteur
14. Adresse du prêteur
15. Nom du propriétaire
16. Adresse du propriétaire
17.

Exemple de liste d'ID de projet :

a = (« 1 »,
 « 2 »,
 « ... »,
 « 141 »,
 « 142 »)

6.1.2 La dimension des versions d'exécution (« r »)

En deuxième lieu dans le flux de travail, nous exécutons la régression MARS sur les données d'entrée. Cela impliquera vraisemblablement de plusieurs à de nombreuses exécutions, à mesure que nous modifions divers paramètres pour améliorer le modèle. Chaque exécution créera un nouveau dossier de version pour la sortie, dont le nom repose sur la date et l'heure d'exécution. Par exemple,

r = (« ver20241210_083124 »,
 « ver20241209_151130 »,
 « ver20241208_160944 »)

latest_version : r_1 = "ver20241210_083124"

Les versions prometteuses, dotées de modèles et de sorties compétitifs, devraient être conservées jusqu'à ce qu'une version finale soit arrêtée, voire plus longtemps si elles présentent une certaine utilité. Vous pouvez modifier le dossier de version pour indiquer qu'il s'agit de la version finale, en plaçant « Final » à la fin du nom.

6.1.3 Étapes du flux de travail (« w »)

Le flux de travail RCA est habituellement séparé en cinq étapes distinctes, énumérées ci-dessous. Chaque étape doit s'achever avec succès avant le démarrage de la suivante. Le VE devrait pouvoir arrêter le traitement à la fin d'une étape afin d'examiner les résultats avant de poursuivre. Cela est nécessaire car les zones de marché sont relativement complexes et souvent tout à fait singulières. Des problèmes inhabituels exigeant une nouvelle approche peuvent survenir. Le VE doit pouvoir intervenir dans le code R ou Python pour effectuer les modifications nécessaires, voire essayer de nouvelles approches. Et, dans de tels cas, il peut vouloir expérimenter et essayer différentes approches nécessitant plusieurs exécutions de la même étape sur laquelle il travaille. Cela étant, le VE devrait aussi être libre d'exécuter toutes les étapes en une seule commande, - sous certaines réserves (p. ex. l'agrégation du 3e degré doit déjà avoir été déterminée lors d'une exécution antérieure de l'étape 2). Cela signifie qu'à l'achèvement de chaque étape, celle-ci devrait stocker toutes ses données internes dans des fichiers ou des bases de données, de sorte que l'étape suivante puisse démarrer comme si l'étape précédente était encore en mémoire, en rechargeant en mémoire les données de l'étape précédente.

Voici les étapes :

w = (« stage1 », -Configuration
 « stage2 », -Traitement MARS
 « stage3 », -Sales Grid finale
 « stage4 », -Extraits de rapport
 « stage5 », -Revue/Nettoyage

stage2 : $w_2 = \text{"stage2"}$

6.1.4 Feuilles du classeur Excel (« s »)

Dans le classeur Excel RCA Core, nous disposons des feuilles suivantes, chacune étant chargée dans un data frame interne :

```
s = (« MLSDData »,
      « Regression Variables »,
      « Calculations »,
      « Interactions »,
      « Aggregations 1st & 3rd Degree »,
      « Aggregations 2nd Degree »)
```

La raison pour laquelle les agrégations du 2e degré sont placées sur une feuille séparée est qu'elles se définissent le plus aisément au moyen d'une table symétrique présentant toutes les paires d'interactions possibles, permettant de spécifier, pour chaque paire, une variable d'agrégation à laquelle les contributions à la valeur et les ajustements associés seront agrégés (additionnés). Cela n'était pas possible pour les agrégations du 3e degré, car il faudrait une table symétrique tridimensionnelle, ce qui est impossible dans Excel. En outre, une table de 20x20x20 représenterait 8,000 cellules à gérer. Il est plus simple de lister les noms des éventuels termes du 3e degré dans la table des agrégations du 1er degré, par exemple :

Variable	Agréger vers
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 Listings de biens (« p ») - lignes de MLSDData

Les listings de biens forment les lignes de la feuille MLSDData. Ils n'apparaissent pas dans les autres feuilles avant l'étape 4, où les extraits finaux de rapport sont créés avec la sélection finale de biens comparables. Voici un exemple de liste d'ID de biens :

```
p = (« MLS823211 », « MLS823456 », « MLS823542 », « MLS823721 », « ... »)
```

6.1.6 Variables MLS (« v ») colonnes de MLSData

Nous reviendrons plus loin sur le traitement des variables. Qu'il suffise d'indiquer à ce stade que la feuille MLSData comporte typiquement 30-40 variables de types divers, mais que seules 10-20 d'entre elles seront généralement retenues comme variables indépendantes pour la régression MARS. Ces variables apparaissent également dans d'autres feuilles, généralement toutefois sous la forme d'une liste de toutes les variables susceptibles d'être soumises à MARS. Voici un exemple de liste de variables :

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

6.1.7 Exemple final de notation

Rassemblons la notation ci-dessus dans un exemple final. À l'aide des listes précédentes, posons

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

Notons que nous utilisons ici une indexation à partir de 0, de sorte que le premier élément d'une liste est supposé avoir l'indice 0. Python utilise également l'indexation à partir de 0, tandis que R utilise l'indexation à partir de 1.

Ainsi, ce qui précède signifie que nous parlons de la valeur de la cellule correspondant à :

- La variable v_0 , soit la colonne « GLA »
- Pour le bien p_2 , soit l'ID « MLS823542 »
- De la feuille s_0 , soit « MLSData »
- Pour l'étape de flux de travail w_1 , soit « stage2 »
- De la version r_1 , soit
« ver20241209_151130 »
- Du projet a_{141} , soit « 142 »

La notation ci-dessus peut alors également s'écrire :

$$\begin{matrix} & & \text{«142»} \\ \text{«ver20241209_151130»} & W & \text{«MLS823542»} \\ \text{«stage2»} & \text{«MLSData»} & \text{«GLA»} \end{matrix} \quad (4)$$

6.1.8 Justification

« Pourquoi ne pas simplement utiliser des index ou des numéros dans la notation ? » Notre notation doit satisfaire deux exigences :

1. Les sous-ensembles : l'analyste sélectionne fréquemment des sous-ensembles d'une liste, par exemple des sous-ensembles de variables. Ainsi, la feuille MLSData peut contenir 35-40 variables, dont seulement 10-20 sont utilisées pour une exécution donnée. D'une exécution à l'autre, le VE change fréquemment les variables soumises à MARS/earth.
2. L'itération, p. ex. $i = 1, 2, \dots, n_1$
Nos algorithmes, protocoles et démonstrations doivent généralement pouvoir itérer sur tous les éléments d'une liste ordonnée, du premier au dernier – ou en extraire un élément particulier.

Il nous faut donc une lettre représentant le sous-ensemble ou la liste d'éléments, laquelle doit être munie d'un indice supplémentaire permettant d'itérer sur les éléments de la liste afin d'obtenir les chaînes de caractères servant à indexer les lignes ou les colonnes du data frame.

6.1.9 Matrices ?

Et qu'en est-il des matrices de l'algèbre linéaire ? Les matrices ne fonctionnent qu'avec des nombres, entiers ou réels. Les matrices en tant que structure de données sont prises en charge tant en R qu'en Python. Toutefois, nombre des variables que nous manipulons dans les données MLS sont des « chaînes de caractères » (strings), c'est-à-dire des données de type caractère. Même lorsque certaines variables ne contiennent que des nombres, il s'agit souvent non pas de nombres réels, mais de simples noms. Par exemple, les codes de zone MLS d'une ville peuvent être (« 660 », « 661 », « 662 », « 663 », ..., « 672 »). La régression earth permet de spécifier des variables comme variables « factor », auquel cas les nombres sont traités simplement comme des noms, sans aucun ordre. Dans ce cas, le modèle produit fournira simplement une contribution à la valeur pour chaque nombre, plutôt qu'une fonction définie sur la séquence. D'autres exemples incluent « C1 », ..., « C6 » pour l'état et « Q1 », ..., « Q6 » pour la qualité.

6.1.10 Première simplification

La notation ci-dessus pour le classeur Excel central du projet est utile pour relier les éléments entre eux, mais trop lourde à manipuler lorsqu'il s'agit d'expliquer des algorithmes ou des démonstrations mathématiques. Lorsque nous effectuons le travail de détail, nous pouvons supposer que nous nous trouvons dans le projet, la version, l'étape et la feuille considérés. Nous indiquons simplement l'objet sur lequel nous travaillons, s'il n'est pas déjà implicite. Tout ce dont nous avons réellement besoin dans la notation ci-dessus, une fois précisé l'objet du travail, est :

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

où : n_p = # biens
 n_v = # variables d'entrée
 n_s = # feuilles d'entrée

6.1.11 Seconde simplification

Il est plus commode de renommer les feuilles selon les types particuliers de data frames qu'elles prennent en charge plutôt que d'utiliser des index. Pour chaque feuille, les lignes et les colonnes ont des fonctionnalités différentes. Ainsi, l'ensemble W de data frames est décomposé en data frames M, V, A, I et C comme suit :

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

où : n_p = # biens
 n_v = # variables d'entrée

Note : le bien à évaluer figure par défaut toujours en tête des listes de biens

Dans la feuille de calcul Excel « MLSData » et le data frame associé, la première ligne est toujours réservée au bien à évaluer. Elle n'est jamais triée avec les autres biens MLS situés en dessous. Elle n'est jamais introduite dans la régression MARS comme donnée indépendante ; elle y entre plutôt comme donnée cible, c'est-à-dire comme variable dépendante. Il en va de même pour les listes de biens. Dans toute liste de biens, le premier élément, d'index 0, est toujours supposé être l'identifiant du bien à évaluer, sauf indication contraire.

Même lorsque la tâche consiste à créer un modèle pour une zone de marché sans aucun bien à évaluer, la première ligne est néanmoins remplie avec des données de bien fictives. Cela rend beaucoup plus commode l'écriture de programmes, la spécification d'algorithmes ou la rédaction de démonstrations. Ainsi, si nous utilisons c pour représenter la liste des biens, c_0 est le bien à évaluer et les c_i , $i = 1, \dots, n_p$, sont les listings comparables qui seront fournis en entrée à la régression MARS/earth.

◇ **Spécifications des variables de régression**

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

où : $n_v = \#$ variables disponibles
 $n_s = \#$ de spécifications

On notera que $n_s \leq n_v$, car nous sélectionnons habituellement un sous-ensemble des variables disponibles comme entrée de la régression.

◇ **Calculs**

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

où : $n_v = \#$ de variables
 $n_c = \#$ de spécifications de calcul

◇ **Interactions**

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

où : $n_f = \#$ variables origine (from)
 $n_t = \#$ variables destination (to)

On notera que le data frame I est un data frame symétrique, dont les variables sont énumérées par ordre alphabétique pour les lignes « from » (origine) et les colonnes « to » (destination). Les cellules sont marquées d'un « 1 » pour indiquer qu'une interaction entre la paire de variables correspondante est autorisée. À l'inverse, un « 0 » indique qu'aucune interaction n'est permise. Pour une interaction du second degré, la situation est relativement simple, une seule interaction étant possible. Avec une régression du troisième degré, il existe trois interactions par paires entre les trois variables, chacune de ces paires devant être autorisée pour engendrer le terme (ou la variable) d'interaction à trois variables.

◇ **Agrégations**

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

où : $n_v = \#$ de variables

7 La Sales Grid RCA

Il convient de garder à l'esprit que le protocole RCA exige un programme R ou Python capable d'accéder à R/earth et d'exécuter l'ensemble des étapes détaillées. Le traitement est intensif en données et le travail ne pourrait en aucun cas être effectué manuellement. En même temps, à supposer que le programme soit exempt d'erreurs, un logiciel est nécessaire pour garantir que le travail requis soit accompli avec précision et sans erreur humaine.

Il est également indispensable que les VE qui effectuent ce travail soient en mesure de modifier de tels programmes selon les besoins, en particulier lorsqu'ils sont confrontés à de nouvelles zones de marché imprévues, à de nouveaux quartiers, à de nouveaux biens ou à d'autres exigences ou problèmes inédits.

Au sein du traitement RCA, on distingue trois types de Sales Grids :

1. La RCA Processing Sales Grid, désignée dans le présent article par « MLSDData », qui est un très grand data frame accompagné d'une feuille de calcul associée, pouvant contenir de nombreux détails intermédiaires qui passent finalement par un processus d'agrégation afin d'aboutir à la Sales Grid RCA mentionnée ci-dessous. Un exemple serait trop volumineux pour être présenté ici. Selon toute vraisemblance, les Valuation Engineers adapteront cette feuille de calcul à leurs préférences. J'ai l'intention de consacrer ultérieurement un article distinct à l'élaboration d'une telle grille.
2. La RCA Final Sales Grid, utilisée dans les rapports non destinés aux GSE, qui ressemble à [Table 2](#). Elle imite, par exemple, le formulaire de Fannie Mae, suffisamment pour offrir au lecteur un certain repère de familiarité. On remarquera toutefois qu'elle comporte davantage de colonnes par bien comparable que la grille de Fannie Mae : chaque comparable dispose d'une colonne « Contribution à la valeur » aux côtés de ses colonnes « Mesure » et « Ajustement », et le bien à évaluer en possède également une — des colonnes qui font défaut au formulaire de Fannie Mae. Une forme plus simple est également présentée dans la [Table 1](#).
3. La GSE Upload Sales Grid, qui sert à charger la Sales Grid dans des logiciels tels qu'Alamode. On notera toutefois que tous les logiciels de formulaires d'évaluation ne permettent pas de charger les données d'une Sales Grid à partir d'une feuille de calcul. Par nécessité, de telles grilles ne peuvent inclure de colonne distincte pour les contributions à la valeur, notion dont les GSE, telles que Fannie Mae, n'ont aucune compréhension. La stratégie consiste à leur fournir le formulaire qu'elles exigent, tout en l'étayant par une Sales Grid RCA ainsi que par d'autres pièces justificatives, telles que des précisions sur les éventuelles agrégations de variables.

Les RCA Processing Sales Grids ne ressemblent pas aux Sales Grids de Fannie Mae mais en constituent, grosso modo, une transposée, lignes et colonnes étant interverties : chaque ligne contient les données d'un bien, et chaque colonne les données d'une variable du bien. On peut facilement compter des centaines, voire des milliers de lignes de biens et 200 colonnes ou plus. La première ligne est réservée aux noms des variables, et la deuxième à un bien à évaluer, même lorsque l'on ne réalise qu'une analyse de marché. Lorsque la feuille de calcul est lue dans un data frame, la première ligne de la feuille devient l'attribut names, et la deuxième ligne, celle du bien à évaluer, devient la première ligne du data frame.

Les colonnes de la Sales Grid RCA comprennent des colonnes d'identification, puis les variables primaires, externes et synthétiques, combinées et fournies à MARS en tant que variables indépendantes. MARS sélectionne parmi ces variables celles qu'il estime avoir une incidence significative sur le prix de vente, également appelé variable dépendante ou variable cible. En régression de degré deux, on peut obtenir des paires de ces variables agissant comme de nouvelles variables. En régression de degré trois, on peut également obtenir des triplets de variables agissant

comme de prétendues variables indépendantes, bien qu'elles soient, évidemment, fortement corrélées avec les variables sous-jacentes.

TABLE 1 – Exemple de feuille de calcul RCA : disposition verticale

Bien à évaluer			Comparable 1			Comparable 2		
Mesure	Contribution à la valeur		Mesure	Contribution à la valeur	Ajustement	Mesure	Contribution à la valeur	Ajustement
Prix de vente			\$1,150,000			\$1,500,000		
Valeur de base		\$500,000		\$500,000	\$0.00		\$500,000	\$0
Caractéristiques mesurées			Caractéristiques mesurées			Caractéristiques mesurées		
Surface habitable	1,530	\$463,500	1,350	\$382,500	\$81,000	1,700	\$540,000	-\$76,500
Superficie du terrain	6,000	\$200,000	6,500	\$220,000	-\$20,000	7,000	\$240,000	-\$40,000
Âge	53	\$63,600	57	\$68,400	-\$4,800	51	\$61,200	\$2,400
Salles de bains	2	\$15,000	3	\$30,000	-\$15,000	3	\$30,000	-\$15,000
Chambres	3	\$6,000	4	\$9,000	-\$3,000	4	\$9,000	-\$3,000
Total mesuré		\$1,248,100.00		\$1,209,900.00	\$38,200		\$1,380,200.00	-\$132,100
Caractéristiques résiduelles			Caractéristiques résiduelles			Caractéristiques résiduelles		
Résidu	CQA 5.8	\$30,000	CQA 2.5	-\$59,900	\$89,900	CQA 8.2	\$119,800	-\$89,800
Conception	Décomposition du résidu		Décomposition du résidu		Décomposition du résidu		Décomposition du résidu	
État		7,000		-\$15,000	\$22,000		\$35,000	-\$28,000
Qualité		5,000		-\$8,000	\$13,000		\$20,000	-\$15,000
Utilité fonct.		8,000		-\$10,000	\$18,000		\$30,000	-\$22,000
Équipements		5,000		-\$12,000	\$17,000		\$15,000	-\$10,000
		5,000		-\$14,900	\$19,900		\$19,800	-\$14,800
Sous-total		30,000		-\$59,900	\$89,900		119,800	-\$89,800
Total			Total			Total		
Contribution totale à la valeur		\$1,278,100.00		\$1,150,000.00			\$1,500,000.00	
Prix de vente ajusté					\$1,278,100.00			\$1,278,100.00

8 Variables

Du côté des entrées de l'exécution de MARS, on ne dispose que de variables simples, à savoir :

1. Les caractéristiques du bien chargées depuis la feuille de calcul MLS.
2. Les variables dérivées créées par le VE afin d'affiner les variables MLS ou d'ajouter des transformations de variables existantes. Par exemple, le VE peut préférer transformer la date de vente en « DaysOffMarket », qui est égale à la date d'effet de l'évaluation moins la date de vente.
3. Le VE peut importer des données de sources tierces, telles que les taux d'intérêt hypothécaires. De même, si les données de latitude, de longitude ou d'altitude sont absentes des données MLS, il peut les importer d'autres sources. Ces variables pourraient être appelées « variables externes ».

Du côté des sorties de MARS, on travaille avec un modèle de prix qui est une combinaison de termes qui, soit :

1. Contiennent une seule des variables d'entrée et fournissent l'équation de la contribution à la valeur de cette variable. On les appelle termes du premier degré ou variables (de sortie). Le terme lui-même est traité comme une valeur de contribution à la valeur et reçoit un nom.
2. Contiennent deux des variables d'entrée et constituent l'équation de la contribution à la valeur de l'interaction des deux variables considérées. Ces termes peuvent être appelés termes du deuxième degré ou variables (de sortie).
3. Contiennent trois des variables d'entrée et constituent l'équation de la contribution à la valeur de l'interaction entre les trois variables. Ces termes peuvent être appelés, dans le présent article, termes du troisième degré ou variables (de sortie).

Les variables de sortie de MARS sont donc des contributions à la valeur, y compris la contribution à la valeur résiduelle. Elles sont à leur tour traitées comme des variables dans les calculs de la Sales Grid, tels que le calcul des ajustements ou la création de graphiques.

On peut avoir différents types de variables qui relèvent de deux ontologies fondamentales : (1) l'origine des variables et (2) l'application des variables, c'est-à-dire non résiduelles vs résiduelles.

8.1 Ontologie de l'origine des variables

1. Variables du premier degré
 - 1.1 Variables de régression MLS
 - 1.2 Variables de régression externes
 - 1.3 Variables de régression dérivées

2. Variables facteur :

Pour traiter les variables facteur, une fonction capable de les identifier est nécessaire. Puisqu'elles sont spécifiées dans le classeur Excel de configuration du projet, une fonction est créée qui renvoie vrai si la variable qui lui est passée a été classée comme étant de type facteur.

3. Variables d'interaction : elles sont créées par MARS lors de la génération du modèle, si le degré est fixé à 2 ou plus. Toutefois, trois variables en interaction constituent le maximum qui puisse être recommandé, en raison de la difficulté d'interprétation, de la probabilité de surajustement et de l'impact sur les degrés de liberté, ainsi que de l'augmentation résultante du nombre d'enregistrements de biens nécessaires pour garantir un modèle robuste.

- 3.1 Variables du deuxième degré : elles sont courantes, et de bons exemples seraient GLA :BathRms, GLA :LotSize, DateOfSale :BathRms, Latitude __Longitude.
- 3.2 Variables du troisième degré : interactions moins courantes qui appellent à la prudence et suggèrent un possible surajustement.
- 4. Variables de décomposition du résidu

8.2 Ontologie de l'application des variables

- 1. Contributions à la valeur générées par MARS
 - 1.1 Variables du premier degré
 - 1.1.1. Variables de régression MLS
 - 1.1.2. Variables de régression externes
 - 1.1.3. Variables de régression dérivées
 - 1.2 Variables d'interaction
 - 1.2.1. Variables du deuxième degré
 - 1.2.2. Variables du troisième degré
 - 1.3 Variables de décomposition du résidu
 - 1.4 Variables d'agrégation
- 2. Variables & valeurs générées par l'ingénieur
 - 2.1 Variables de décomposition du résidu
 - 2.1.1. Conception/Esthétique
 - 2.1.2. Aménagement paysager
 - 2.1.3. Utilité fonctionnelle
 - 2.1.4. Vue
 - 2.1.5. État
 - 2.1.6. Qualité
 - 2.1.7. ...

Ces groupes de variables recevront leur notation et seront décrits plus en détail dans la section suivante.

Protocole 3.1 : Variables facteur

Les variables facteur, ou simplement les facteurs, peuvent apparaître dans tous les termes, y compris les interactions. Elles sont importantes parce qu'elles doivent faire l'objet d'un traitement particulier pour la régression MARS et la représentation graphique.

Si les variables à valeurs de type caractère sont nécessairement des facteurs, les variables numériques doivent être signalées comme variables facteur pour que MARS les reconnaisse comme telles. Par exemple, les zones MLS sont souvent identifiées par un numéro. En particulier, les quartiers de la San Francisco Bay Area sont souvent identifiés par des entiers uniques. Si ces valeurs sont transmises à MARS en tant que facteurs, MARS tentera alors de découvrir si certaines zones présentent une valeur contributive moyenne pour les ventes de biens situés à l'intérieur de leurs limites, sans égard à une quelconque relation avec d'autres zones. En revanche, avec des variables ordinaires, il supposera qu'il existe une certaine forme de continuité dans les contributions à la valeur à mesure que la valeur de la variable change. C'est une distinction importante. En outre, ces variables facteur peuvent apparaître dans des termes du modèle avec des variables non-facteur, ce qui rend la représentation graphique difficile. Plus précisément, lors du tracé des termes à variable unique comportant des facteurs, on regroupe tous ces facteurs et on trace simplement un diagramme en barres montrant leurs contributions, positives ou négatives. Si les facteurs apparaissent avec d'autres variables dans des termes, ils apparaîtront toujours avec une seule valeur possible, car les variables facteur ne peuvent avoir qu'une seule valeur pour un bien donné : si une zone a la valeur 662, par définition elle ne peut avoir aucune autre valeur pour cette zone. Si cela vous semble déroutant, pour ne rien arranger, dans le modèle généré par MARS, il n'existe pas un nom de variable unique pour le facteur ; MARS prendra le nom racine de la variable facteur qui lui a été fournie en entrée, puis créera de nouvelles colonnes dont les noms de variables sont formés en y accolant les valeurs possibles. Ainsi, une seule colonne de la feuille de calcul Excel, telle que `MlsArea`, donnera lieu à la création de plusieurs colonnes de data frame pour chaque valeur possible, telles que `MlsArea660`, `MlsArea661`, etc. Une variable telle que `MlsArea661` aura la valeur 1 pour les biens situés dans la zone 661, et 0 sinon. MARS effectuera alors la régression sur ces colonnes comme si elles représentaient des variables booléennes. Ces noms modifiés seront utilisés dans le modèle fourni par MARS.

On voudra vraisemblablement reconvertir ces noms du modèle vers leur nom d'origine afin de rendre les graphiques plus compréhensibles pour des clients qui apprécient la simplicité (du moins autant que possible). Au lieu de voir `MlsArea661=1`, on verra `MlsArea=661`. On peut alors représenter graphiquement les termes en indiquant simplement dans l'en-tête du graphique une mention du type « Pour MLS Area = 662 ». Ainsi, les termes de degré N comportant M variables facteur sont tracés comme des termes de degré (N-M), et la dimension du facteur disparaît dans le titre.

Il convient de noter que deux variables facteur ou plus pourraient exister dans un même terme, de sorte qu'un graphique serait vraisemblablement nécessaire pour chaque paire de valeurs possible. Si la variable x, avec trois valeurs distinctes, et la variable y, avec deux valeurs distinctes, apparaissent dans le même terme avec la variable non-facteur z, il faudrait alors $2 \cdot 3 = 6$ graphiques. Un exemple serait « `MlsArea` » + « `Architectural Style` » : par exemple, les maisons victoriennes de San Francisco peuvent valoir plus ou moins selon le quartier de San Francisco où elles se trouvent. Heureusement, dans la plupart des zones, il n'y a généralement que 1-3 styles architecturaux suffisamment importants comme contributeurs de valeur pour être utilisés dans le modèle (bien qu'il puisse exister 6+ styles architecturaux différents).

8.3 Termes du deuxième et du troisième degré

Avec une régression du deuxième ou du troisième degré, certains termes du modèle de sortie peuvent inclure respectivement deux ou trois variables. Par exemple :

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

On remarquera que l'on nomme ces termes en triant les noms de variables par ordre alphabétique puis en les concaténant, éventuellement avec un tiret bas. On procède de même pour les termes du troisième degré, sur les triplets de noms de variables produits dans le modèle. Par exemple, « BathRms_GLA_LotSz ».

Ces variables sont ajoutées aux variables d'entrée primaires à l'étape 3, lorsque le programme RCA calcule les contributions de toutes les variables pour chaque bien, y compris les variables d'interaction.

8.3.1 Variables d'ajustement

Le programme RCA calculera également, à l'étape 3, les ajustements pour l'ensemble des variables et des biens, une fois les contributions à la valeur calculées, en soustrayant la contribution à la valeur du bien comparable correspondant, pour une caractéristique donnée, de la contribution du bien à évaluer pour cette même caractéristique.

8.3.2 Variables d'agrégation

Les variables d'agrégation sont définies pour remplacer des sous-ensembles de variables, en additionnant toutes leurs contributions à la valeur et leurs ajustements dans une variable commune destinée à être utilisée dans le rapport. Par exemple, on peut avoir scindé la surface habitable, achevée comme inachevée, en variables de détail plus fines :

1. Surface achevée légale hors sol
2. Surface achevée légale en sous-sol
3. Surface achevée illégale hors sol
4. Surface achevée illégale en sous-sol
5. Surface inachevée hors sol
6. Surface inachevée en sous-sol

Oui, il existe des zones où cela a effectivement du sens.

La régression MARS portera sur toutes ces variables et générera un modèle qui fournira des contributions et des ajustements pour chacune d'elles. Mais pour un rapport sur formulaire, ces différentes contributions devront probablement être combinées ou agrégées en un nombre plus restreint de variables. Les détails de l'agrégation devront être fournis dans une annexe. Un exemple plus fréquent consisterait à agréger les ajustements d'interaction dans l'un des ajustements des variables primaires. Par exemple, on pourrait agréger l'ajustement de GLA_LotSize dans GLA ou dans LotSize, ou même le répartir 50 :50 entre les deux, selon ce que détermine la feuille Aggregation.

8.3.3 Variables spéciales & notation

π_i	Prix de vente net effectif du bien $i = 1, 2, \dots, n_p$; pour le bien à évaluer, π_0 désigne la conclusion de valeur (voir Démonstration I)
α_i	Prix de vente net ajusté du bien $i = 1, 2, \dots, n_p$
ρ_i	Prix de vente net estimé par MARS pour le bien $i = 0, 1, \dots, n_p$
β	Valeur de base du modèle
$\varphi_{i,j}$	Contribution à la valeur de la variable de régression $j = 1, 2, \dots, n_r$ pour le bien $i = 0, 1, \dots, n_p$
$\delta_{i,j}$	Ajustement de la variable de régression $j = 1, 2, \dots, n_r$ pour le bien $i = 1, 2, \dots, n_p$
ϵ_i	Résidu total du bien $i = 0, 1, \dots, n_p$
$\epsilon_{i,c}$	Contribution à la valeur de la composante résiduelle c , $c = 1, 2, \dots, n_c$, pour le bien $i = 0, 1, \dots, n_p$
$\xi_{i,c}$	Ajustement de la composante résiduelle c , $c = 1, 2, \dots, n_c$, pour le bien $i = 0, 1, \dots, n_p$

où

$n_p = \#$ of comparable properties

$n_r = \#$ of regression variables

$n_c = \#$ of residual breakdown variables

8.3.4 Décomposition des résidus

Une fois que la régression MARS a créé un modèle, elle génère les prix de vente estimés à partir des variables d'entrée sélectionnées par MARS. Elle soustrait cette valeur du prix de vente net correspondant afin d'obtenir une valeur résiduelle pour chaque bien. Le protocole RCA stipule alors qu'une fois que le VE a arrêté son choix définitif de biens comparables, généralement 6-12 listings, il doit passer en revue chaque listing et décomposer le résidu calculé selon les caractéristiques (variables latentes) les plus vraisemblables qui, selon lui, influencent la valeur, en ignorant la possibilité d'erreurs de données aléatoires. Il pourrait, par exemple, retenir les variables de décomposition suivantes pour un bien donné :

1. Conception
2. Utilité fonctionnelle
3. Vente au-dessus/au-dessous du marché
4. État
5. Qualité

Des biens différents auront vraisemblablement des variables de décomposition du résidu en partie différentes, qu'il convient de rassembler dans un seul ensemble ordonné.

Estimation des contributions des caractéristiques rares. Certaines caractéristiques porteuses de valeur sont trop rares pour le canal de la régression : un atelier présent dans 3 ventes sur 780 ne peut se voir attribuer un coefficient défendable, et la passe avant ignorera généralement une telle variable entièrement. Ce sont des caractéristiques résiduelles, et leurs contributions peuvent être estimées *dans l'espace des résidus* : on ajuste le modèle sans la caractéristique, puis on régresse les résidus du modèle sur la ou les colonnes exclues par les moindres carrés ordinaires. Pour un indicateur binaire, le coefficient est une contribution forfaitaire (de manière équivalente, le décalage résiduel moyen des ventes signalées) ; pour une colonne numérique, il s'agit d'une contribution unitaire. Lorsque plusieurs aménités corrélées sont estimées *conjointement* (p. ex. la surface des dépendances conjointement avec un indicateur d'atelier), la régression répartit le crédit entre elles plutôt que de compter deux fois les mêmes dollars dans deux lignes de décomposition. Les estimations fondées sur moins d'une dizaine d'observations sont seulement indicatives et devraient être corroborées par le coût déprécié ou par un échantillon élargi (p. ex. des indicateurs récupérés à partir des remarques publiques) ; la contrainte résiduelle elle-même plafonne quelle que soit la valeur finalement attribuée. earthUI (v0.11+) met en œuvre cette méthode directement sous la forme de l'onglet *Residual Contribution*, y compris l'estimation conjointe, les éléments de preuve résiduels par vente et une note documentaire pour le dossier de travail.

Notation : on utilise la notation du tableau ci-dessus pour arriver à :

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

où n_c est le nombre de composantes de la décomposition du résidu.

Le VE doit exercer son jugement pour décomposer le résidu selon les composantes données. On pourrait penser que cela permettrait à un biais de s'introduire dans la conclusion de valeur finale. Or, il n'en est rien, car on dispose d'une démonstration mathématique établissant que si les ajustements de la décomposition totalisent le résidu, ils n'auront aucune incidence sur la conclusion de valeur. L'objet de la décomposition n'est pas d'établir la valeur, mais d'expliquer pourquoi le résidu est ce qu'il est en comparaison des résidus des ventes des autres biens. Le seul point où le biais de l'ingénieur peut influencer sur la conclusion de valeur finale est l'estimation du résidu du bien à évaluer puisque, bien entendu, il n'existe pas de prix de vente effectif pour le bien à évaluer permettant de calculer un résidu pour celui-ci.

Chaque bien possède son propre ensemble de contributions à la valeur résiduelles, mais ceux-ci doivent être fusionnés en un seul ensemble C couvrant tous les biens :

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

où $[\epsilon_{i,c}]$ est le nom de variable associé à la valeur $\epsilon_{i,c}$, A_i est l'ensemble de ces variables et n_c est

le nombre de valeurs dans chaque ϵ_i .

Ainsi, les données du bien à évaluer et des biens comparables figurant dans des colonnes séparées, on devrait voir, sous l'ensemble des contributions à la valeur des variables du modèle de régression pour chaque bien, l'ensemble C des n_c contributions à la valeur résiduelles correspondantes.

9 Démonstrations fondamentales de la RCA

Posons :

$$\begin{aligned} n_p &= \# \text{ of properties} \\ n_r &= \# \text{ of regression variables} \\ n_c &= \# \text{ of residual component variables} \end{aligned}$$

Démonstration I : Chaque prix de vente ajusté de chaque bien comparable est égal au prix de vente estimé du bien à évaluer.

Pour tout bien comparable i , le prix de vente ajusté α_i est la somme du prix de vente net π_i et des ajustements : les différences entre les contributions à la valeur du bien à évaluer et celles du bien comparable pour la valeur de base, les variables de régression $j = 1, 2, \dots, n_r$, et le résidu. Pour le bien à évaluer, qui ne fait l'objet d'aucune vente effective, le prix de vente net π_0 est défini comme la conclusion de valeur $\beta + \varphi_{0,1} + \varphi_{0,2} + \dots + \varphi_{0,n_r} + \epsilon_0$. Alors :

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

Démonstration II : Modifier la décomposition du résidu, sous la contrainte que toutes les contributions à la valeur des composantes résiduelles se somment au résidu du bien comparable, n'a aucune incidence sur la conclusion de valeur finale du protocole RCA.

C'est presque trivial. Pourtant, sans réfléchir aux mathématiques, il est facile de passer à côté, car les évaluateurs traditionnels se concentrent sur les ajustements plutôt que sur les contributions à la valeur. Ce sont les contributions à la valeur qui s'annulent ici — un oubli vieux de plusieurs décennies. Pour un bien comparable k , $k > 0$, l'ajustement du résidu total est égal à la somme des n_c composantes résiduelles de la décomposition, laquelle est à son tour égale à la somme des différences de contributions à la valeur, elle-même égale au résidu total du bien à évaluer moins le résidu total du bien comparable. Et cela est indépendant du nombre de composantes résiduelles créées ou de ce qu'elles représentent. Tant que la valeur des composantes résiduelles

est égale au résidu total, tant pour le bien à évaluer que pour le bien comparable choisi, ξ_k ne changera pas.

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

10 Modèle MARS

10.1 Fonctions « g » de variables du modèle et d'interaction

Un principe fondamental de la méthode RCA est que le « raisonnement explicatif » revêt une importance capitale. Malheureusement, produire des graphiques parlants pour les diverses équations de contribution à la valeur, en particulier lorsque des interactions et des variables facteurs sont utilisées, peut s'avérer difficile. Il est donc jugé impératif de fournir un protocole clair et strict pour la création de ces graphiques. C'est ce qui est fait ici ; il convient toutefois de comprendre que ce protocole s'adresse au développeur qui écrit le code R ou Python générant les graphiques, et qu'il ne présente peut-être qu'un intérêt superficiel pour les personnes non impliquées dans la programmation.

Le modèle de prix généré par MARS est la somme d'une constante de base et de termes additionnels, chacun faisant intervenir 1 à 3 variables. Pour l'analyser, on décompose d'abord le modèle en termes individuels. Ensuite, les termes partageant le même ensemble de variables, représentant une seule interaction, sont regroupés en expressions d'interaction uniques. Supposons que le degré maximal de la régression soit 3. Considérons un modèle de régression de troisième degré composé d'une constante de base, de n_1 expressions de premier degré, de n_2 expressions d'interaction de deuxième degré et de n_3 expressions d'interaction de troisième degré. On obtient ainsi quatre groupes distincts d'expressions, qui sont soit des expressions à variable unique, soit des interactions à plusieurs variables, chacune représentée graphiquement de manière différente selon deux facteurs : (1) la présence ou non d'une variable facteur (une variable catégorielle), et (2) le nombre de variables uniques qu'elles contiennent. Plus précisément :

- Les expressions de premier degré (1 variable) sont représentées en 2D.
- Les expressions de deuxième degré (2 variables) sont représentées en 3D.
- Les expressions de troisième degré (3 variables) sont représentées par un *ensemble* de graphiques 3D.
- Si une expression contient une variable facteur, chaque facteur unique réduit la dimensionnalité du graphique d'une unité (p. ex. un graphique 3D devient 2D si l'une des variables est un facteur).

L'indexation de chaque fonction g fournit les informations nécessaires pour déterminer comment elle doit être représentée graphiquement, en reflétant le nombre de variables uniques et la présence de variables facteurs.

10.2 Indexation des fonctions g

Chaque fonction g est indexée sous la forme $f_{j,k} g_k^j$, où :

- $f_{j,k}$ est le premier indice (un exposant placé avant g) :
 - j varie de 0 à 3 et identifie le groupe :
 - $j = 0$: la constante de base (une exception à l'indexation habituelle).
 - $j = 1$: expressions de premier degré.
 - $j = 2$: expressions de deuxième degré.
 - $j = 3$: expressions de troisième degré.
- k est l'indice d'une fonction particulière au sein de son groupe (p. ex. $k = 1$ pour la première fonction du groupe j).
- La valeur de $f_{j,k}$ indique le nombre de variables facteurs uniques dans la fonction :
 - $f_{j,k} = 0$: aucune variable facteur.
 - $f_{j,k} > 0$: le nombre de variables facteurs (p. ex. 1, 2 ou 3).

- Indices postposés de g :
 - L'exposant j après g (p. ex. g^j) répète le numéro du groupe par souci de clarté.
 - L'indice inférieur k après g (p. ex. g_k) correspond au k de $f_{j,k}$ et identifie la fonction au sein de son groupe.

Bien que les indices de f et de g coïncident numériquement, leurs rôles diffèrent : $f_{j,k}$ compte les variables facteurs, ce qui est déterminant pour les choix de représentation graphique, tandis que j et k sur g définissent son groupe et sa position.

10.3 Dimensions de représentation graphique

La dimension d du graphique d'une fonction $f_{j,k} g_k^j$ se calcule comme suit :

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

où :

- j est le nombre de variables uniques (ou le numéro du groupe, en ignorant le cas de la constante).
- $f_{j,k}$ est le nombre de variables facteurs.
- d est le nombre de variables continues (non facteurs) à représenter, avec un maximum de 3 (puisque le degré maximal est 3). Une fonction à d variables continues se dessine en $d + 1$ dimensions : $d = 1$ donne un graphique 2D, $d = 2$ un graphique 3D (ou une carte thermique 2D), et $d = 3$ un *ensemble* de graphiques 3D.

Par exemple :

- Si $j = 3$ (3 variables) et $f_{j,k} = 0$ (aucune variable facteur), alors $d = 3 - 0 = 3$, ce qui exige un ensemble de graphiques 3D, chacun fixant l'une des variables.
- Si $j = 3$ et $f_{j,k} = 1$ (une variable facteur), alors $d = 3 - 1 = 2$, ce qui exige un graphique 3D (ou une carte thermique 2D) pour chaque valeur de la variable facteur apparaissant dans le modèle.
- Si $j = 0$ (la constante), une représentation graphique n'est généralement pas nécessaire, puisqu'il s'agit d'une valeur unique.

Lorsque $d = 3$ (p. ex. une expression de troisième degré sans variable facteur), la visualisation de la fonction complète peut nécessiter plusieurs graphiques 3D, chacun fixant une variable à différentes valeurs afin de montrer un sous-ensemble du comportement de la fonction.

10.4 L'équation finale du modèle

Le modèle MARS est décomposé en quatre groupes d'expressions (constante, premier, deuxième et troisième degré). L'indexation $f_{j,k} g_k^j$ nous indique :

- j : le groupe et le nombre de variables.
- $f_{j,k}$: le nombre de variables facteurs.
- $d = j - f_{j,k}$: la dimension du graphique.

Cette structure guide la visualisation de chaque fonction g , en garantissant que la méthode de représentation graphique corresponde à la complexité de l'expression et aux types de variables.

L'estimation du prix de vente $\hat{P}$ fournie par le modèle MARS peut s'exprimer ainsi :

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

$n_1 = \#$ of first degree terms
 $n_2 = \#$ of second degree terms
 $n_3 = \#$ of third degree terms

10.5 Nombre de graphiques nécessaires pour représenter un modèle

Combien de graphiques faut-il pour visualiser toutes les fonctions d'un modèle ? La réponse dépend des variables, de leurs interactions et de leur nature de variable facteur ou non.

Le VE peut restreindre les variables et les interactions dans le fichier de configuration Excel destiné à R/earth, ce qui influe sur les interactions autorisées. Une approche pratique consiste à laisser R/earth évaluer toutes les interactions possibles et sélectionner celles qui apportent une contribution significative à l'estimation du prix de vente, tout en vérifiant l'absence de problèmes tels que le surajustement ou la colinéarité. Certaines interactions peuvent être interdites, ce qui se répercute sur les interactions de degré supérieur (p. ex. si une paire est exclue dans une interaction de troisième degré, la combinaison entière est exclue).

Pour les variables facteurs (p. ex. MLS Area avec 12 valeurs en interaction avec GLA), la représentation graphique devient plus complexe. Cela pourrait nécessiter 12 graphiques—un par zone. Si un autre facteur comme Architectural Style (6 valeurs) interagit avec MLS Area et GLA, jusqu'à $12 \cdot 6 = 60$ graphiques pourraient être nécessaires. Toutefois, R/earth ne juge que rarement toutes les combinaisons significatives ; en général, seules quelques interactions clés (p. ex. certaines zones et certains styles) méritent d'être représentées, ainsi que le modèle l'identifie.

Voici un guide simplifié du nombre de graphiques nécessaires :

1. **Fonctions à variable unique (non facteur) :** 1 graphique.
2. **Fonctions à variable unique (facteur) :** 1 graphique en barres regroupant toutes les valeurs du facteur.
3. **Fonction de base :** aucun graphique nécessaire.
4. **Interactions de 2^e degré (sans facteur) :** 1 graphique (p. ex. carte thermique 2D avec des couleurs pour le prix de vente).
5. **Interactions de 2^e degré (1 facteur) :** 1 graphique par valeur du facteur présente dans le modèle (p. ex. 2 valeurs = 2 graphiques).
6. **Interactions de 2^e degré (2 facteurs) :** 1 graphique avec une matrice des combinaisons de facteurs (p. ex. 3 zones $\times$ 2 styles = 6 cellules).
7. **Interactions de 3^e degré (sans facteur) :** plusieurs graphiques (p. ex. 4 graphiques pour un nombre de salles de bains = 1, 2, 3, 4).
8. **Interactions de 3^e degré (1 facteur) :** 1 graphique par valeur du facteur présente dans le modèle.
9. **Interactions de 3^e degré (2 facteurs) :** 1 graphique par combinaison de valeurs des facteurs présente dans le modèle.
10. **Interactions de 3^e degré (3 facteurs) :** 1 graphique par valeur du facteur ayant le moins de valeurs distinctes, avec une matrice de couleurs pour les deux autres.

Des scénarios plus complexes peuvent exiger des techniques de représentation graphique plus avancées.

D'autres possibilités existent, avec des graphiques vraisemblablement plus sophistiqués.

Un exemple de modèle complet généré par MARS figure dans [Figure 1](#).

11 Caractéristiques des données

La section suivante traite de MARS. MARS dépend entièrement des données qu'il reçoit, et la plupart de ces données proviendront d'un MLS (Multiple Listing Service) ou des Tax Assessors (administration d'évaluation fiscale). Toutefois, elles peuvent transiter par d'autres fournisseurs de données qui entretiennent et améliorent les données, puis les revendent avec profit. Dans certaines zones, elles sont assez fiables ; dans d'autres, elles le sont moins. Dans les cas critiques, l'inspecteur du bien doit effectuer lui-même ses observations et ses mesures, dont l'exactitude peut être mise en question.

Le Valuation Engineer doit comprendre ce que les données représentent et dans quelle mesure elles remplissent bien leur rôle de représentation. Cela soulève toute une série de questions :

- Comment classer une variable comme étant une mesure ?
- Si une variable n'est pas une mesure, que peut-elle être d'autre ?
- Comment la régression peut-elle être utile si les données qu'elle reçoit ne sont pas exactes ?

La plupart des données que nous obtenons des services de listings MLS et des Tax Assessors aux fins de la régression sont des données numériques qui impliquent des mesures ou des dénombrements. Certaines sont logiques, Vrai ou Faux.

11.1 Rôle du résidu

Le résidu sert d'indicateur indirect de la valeur ajoutée par les variables non incluses dans le modèle de régression. Dans le langage de la littérature hédonique, ces caractéristiques non mesurées sont des *variables latentes* : des variables qui influencent le prix de vente mais ne sont pas mesurées directement, le plus souvent parce qu'il n'existe aucun moyen pratique de les mesurer objectivement. Leurs valeurs doivent au contraire être inférées et, en pratique, la valeur d'une variable latente est souvent estimée par un jugement subjectif — dans la RCA, un jugement ancré dans le classement des résidus. En ordonnant les biens selon leur résidu ou leur résidu par pied carré, on obtient un classement objectif allant des biens présentant le moins d'attrait à ceux qui en présentent le plus. Cette méthode de classement est efficace, à condition qu'un analyste qualifié doté d'une connaissance approfondie du marché local élabore le modèle MARS (Multivariate Adaptive Regression Splines). Voici quelques exigences supplémentaires :

- Qualité des données et du modèle : un modèle de régression bien construit et des données de haute qualité sont des préalables à des résidus fiables. Dans la San Francisco Bay Area, mon objectif est de développer un modèle présentant un R^2 d'environ 75-80% et un R^2 de validation croisée (CV R^2) de 55-70%.
- Éviter le surajustement, limiter le R^2 : le Valuation Engineer devrait viser un R^2 maximal d'environ 80% dans la plupart des marchés matures des zones à prix élevés comme la SF Bay Area, car une grande partie de la valeur concurrentielle d'un bien résidentiel réside dans son attrait subjectif, qui n'est pas quantifiable et ne peut donc pas être intégré directement dans les modèles de régression. Un R^2 nettement supérieur à 80% peut indiquer un modèle surajusté et appeler un examen plus approfondi. Une mise en garde s'impose toutefois : il existe des zones de marché présentant des valeurs de R^2 supérieures à 90% sans surajustement, et il existe des zones difficiles à évaluer où le mieux que l'on puisse obtenir est de 50-60% ou moins.
- R^2 élevé dans certaines zones : il existe cependant des régions, comme les lotissements ou d'autres zones résidentielles uniformes, où des valeurs de R^2 plus élevées sont étonnamment courantes en raison de l'homogénéité des biens.

11.2 Les notations d'état et de qualité comme régresseurs approximatifs

Les échelles de notation d'état et de qualité de Fannie Mae (C1–C6 et Q1–Q6) méritent un commentaire particulier. Ces échelles sont trop grossières pour être traitées comme de véritables caractéristiques mesurées et, parce qu'elles sont attribuées par l'évaluateur, elles constituent en pratique des notations subjectives — des variables latentes à part entière. Il est utile de distinguer explicitement deux niveaux : la notation grossière de l'évaluateur utilisée comme régresseur, *état à gros grain* (condition-1) et *qualité à gros grain* (quality-1), par opposition aux variables latentes à grain fin extraites du résidu, *état à grain fin* (condition-2) et *qualité à grain fin* (quality-2). Le VE peut accepter les notations à gros grain de l'évaluateur comme approximations et effectuer la régression sur celles-ci, obtenant ainsi des contributions à la valeur approximatives qui sont utiles comme valeurs initiales pour les variables latentes à grain fin. Lorsque l'on procède ainsi, le protocole consiste à soustraire de l'estimation du modèle les contributions à la valeur estimées par la régression pour ces deux variables à gros grain — en les retirant des attributs mesurés figurant dans la section régression de la Sales Grid — et à les rajouter au résidu. De là, elles sont ensuite retirées du résidu et affectées aux variables latentes à grain fin correspondantes, `res_condition1` et `res_quality1`. Des informations supplémentaires extraites des remarques publiques du MLS (`public_remarks`) et des listes de chaînes de caractères associées sont ensuite utilisées pour apporter des micro-ajustements aux valeurs de ces variables latentes. Le résultat final est que les contributions à la valeur à grain fin remplacent les contributions à la valeur à gros grain : les notations grossières ne servent que d'échafaudage pour le modèle approximatif, et les valeurs finalement rapportées pour l'état et la qualité sont celles à grain fin, dérivées du résidu.

11.3 Erreurs de données

D'après des commentaires sur les réseaux sociaux, les données du Multiple Listing Service (MLS) de Californie sont généralement de meilleure qualité que celles de nombreuses autres régions des États-Unis. En règle générale, on peut s'attendre à ce que les données MLS soient de bonne qualité dans la plupart des zones métropolitaines matures, mais de moindre qualité dans les zones rurales.

Bien que des erreurs occasionnelles surviennent dans les données pour diverses raisons, ces erreurs devraient généralement se compenser mutuellement et n'avoir qu'un impact minime sur les moyennes utilisées dans les modèles de prix, pour autant que les données soient suffisantes.

Cependant, des biais systématiques peuvent survenir dans certaines zones. Considérons par exemple un quartier où de nombreuses maisons ont été initialement construites avec un étage complet, mais où certains acheteurs se sont vu offrir la possibilité de supprimer des pièces de l'étage afin de créer des plafonds cathédrale dans une ou plusieurs pièces du rez-de-chaussée, réduisant considérablement la surface habitable brute (GLA). Ces modifications n'ont pas été reportées par la suite dans les registres du Tax Assessor, pour une raison ou une autre. Par conséquent, le MLS peut indiquer 2,400 pieds carrés alors que la valeur mesurée n'est que de 2,000 pieds carrés. Dans de tels scénarios, les données MLS devraient être corrigées avant d'effectuer la régression MARS (Multivariate Adaptive Regression Splines). Il est important de noter que l'obtention de cette information pour les biens comparables peut exiger un certain effort d'investigation. Les divergences devraient être consignées pour référence future.

Compte tenu de ce qui précède, il est raisonnable de supposer que, dans les zones où les données MLS sont de haute qualité, les erreurs qui se répercutent sur le résidu sont minimales par rapport à l'impact des caractéristiques subjectives et non mesurées telles que la conception, l'état et la qualité. C'est une considération importante lors de l'estimation de l'exactitude de la méthode RCA. Par exemple, si la valeur du R^2 pour une zone de marché donnée est de 80%, alors environ 20% du prix est attribué au résidu CQA. Une erreur de $\pm 10\%$ dans le positionnement du bien à

évaluer au sein du classement des résidus se propage donc en une erreur de $\pm 10\% \cdot 20\% = \pm 2\%$ dans la conclusion de valeur finale. C'est ainsi que l'on améliore l'exactitude ! Dans de nombreux cas, des estimations de valeur exactes bien en deçà de $\pm 1\%$ sont possibles.

Par conséquent, traiter les résidus comme des indicateurs indirects de la valeur collective des variables non modélisées dans la régression MARS est raisonnable, à condition que le Valuation Engineer dispose d'une expérience suffisante du marché local pour identifier où et pourquoi existent des biais de mesure systémiques. Cette approche garantit que l'analyse reflète fidèlement les valeurs réelles des biens, tout en reconnaissant les limites et les nuances des données immobilières.

Exemple détaillé : mesurer les déterminants de la valeur non mesurés à Sea Ranch

Lorsque le R^2 atteignable dans une zone de marché est inférieur à ce que l'homogénéité apparente du marché laisserait attendre, la réponse du cadre méthodologique consiste à identifier et à mesurer la variable manquante plutôt qu'à accepter cette limitation. La communauté de Sea Ranch, dans le comté de Sonoma, l'illustre directement.

Voici un cas concret de la réponse « mesurer plutôt qu'accepter ». La communauté de Sea Ranch, dans le comté de Sonoma, est un développement côtier où la distance à l'océan constitue un déterminant de la valeur de premier ordre : une maison située à quelques centaines de pieds de la falaise se vend à un prix sensiblement différent de celui d'une maison située à un quart de mile à l'intérieur des terres, toutes autres caractéristiques égales par ailleurs. Les données MLS de la communauté n'enregistrent pas la distance à l'océan comme variable. Un modèle MARS initial ajusté sur les variables MLS disponibles a produit un pouvoir explicatif inférieur à ce que l'homogénéité apparente du marché aurait dû permettre, avec une structure des résidus qui n'ordonnait pas suffisamment les biens selon une caractéristique visible dans les données — signe qu'un déterminant important manquait au modèle, plutôt que d'un marché véritablement bruité.

Le remède fut la mesure directe. L'auteur s'est procuré une carte de la communauté montrant la disposition des propriétés par rapport au littoral et a mesuré la distance à l'océan pour plusieurs centaines de maisons. Les distances mesurées ont été ajoutées comme variable au jeu de données MLS et la régression MARS a été relancée. La nouvelle variable a été retenue par le modèle comme significative, le pouvoir explicatif a augmenté substantiellement, et le classement des résidus a ensuite ordonné les biens selon des motifs cohérents avec les caractéristiques subjectives restantes — vue, état, qualité — plutôt que selon le signal latent de distance à l'océan qui l'avait auparavant contaminé. L'altitude a été traitée de manière similaire dans des quartiers à flanc de colline où la variable MLS est manquante ou peu fiable.

Le principe méthodologique est général. Lorsque le R^2 atteignable tombe en dessous de ce que l'homogénéité apparente du marché devrait permettre, la première hypothèse est qu'une variable importante manque au modèle, et la réponse appropriée consiste à identifier de quoi il s'agit et à recueillir les mesures directement. Les sources comprennent les cartes communautaires, les données SIG, les registres du Tax Assessor, les registres du service des permis de construire, l'imagerie satellitaire et l'inspection sur le terrain. Le coût d'un tel travail de mesure est réel mais borné ; une fois la variable recueillie pour une zone de marché donnée, elle peut être reportée dans les missions futures dans cette zone. La RCA absorbe proprement les nouvelles variables parce que le cadre méthodologique est indifférent aux variables qui entrent dans la régression MARS — la structure des résidus indique au VE si la nouvelle variable capte un signal réel, et les contributions à la valeur s'ajustent en conséquence.

La réponse alternative — accepter un R^2 faible et rapporter des bandes d'erreur plus larges — est défendable lorsque les déterminants manquants sont véritablement non mesurables (préférences propres à l'acheteur, sentiment de marché transitoire, circonstances de transaction

idiosyncratiques), mais pas lorsqu'ils sont simplement non collectés. Face à un modèle faible, le premier réflexe du VE devrait être de se demander si le signal manquant relève de ce second cas.

11.4 Qu'est-ce que l'exactitude ?

Une exactitude parfaite pour une évaluation exige :

1. Conformité à la valeur de marché : la transaction de vente hypothétique du bien à évaluer adhère à 100% à la définition donnée de la valeur de marché
2. Développement d'un modèle exact et robuste : un modèle MARS de haute qualité est développé, avec un R^2 d'environ 75-80% et un CVR2 d'environ 55-70% selon la localisation.
3. Bonne analyse des erreurs de données : toutes les erreurs de données probables pour les variables mesurées sont comprises et contenues, avec peu de répercussions sur les résidus.
4. Prise en compte des biais systémiques : tout biais systémique dans les données doit être pris en compte et faire l'objet d'un ajustement.
5. Analyse du classement des résidus : s'assurer que les biens classés par résidu suivent un schéma raisonnable de caractéristiques, du rang le plus bas au rang le plus élevé. Les anomalies telles que les ventes successorales, les ventes en « short sale » et les ventes aux enchères doivent être examinées, expliquées et éventuellement retirées.
6. Positionnement du bien à évaluer : avec une compréhension approfondie du classement des résidus, le bien à évaluer doit être positionné avec exactitude. Il convient d'identifier un bien présentant un attrait nettement supérieur et un autre présentant un attrait nettement inférieur, puis de placer le bien à évaluer entre les deux et de le noter. On s'efforce ensuite de resserrer autant que possible le classement entre le bien supérieur et le bien inférieur.
7. Raisonnement explicatif : ne jamais sacrifier le raisonnement explicatif à l'exactitude. L'évaluateur doit être capable d'expliquer les différences de valeur entre le bien à évaluer et les biens comparables, du moins à une personne d'intelligence raisonnable. Par exemple, un bon modèle MARS présentant un R^2 et un CVR2 optimaux peut être utilisé pour expliquer les contributions à la valeur des caractéristiques mesurées, tandis que la décomposition du résidu effectuée par le Valuation Engineer peut être utilisée pour expliquer les contributions à la valeur des autres caractéristiques.
8. Surajustement : éviter le surajustement des modèles en n'utilisant pas de caractéristiques du bien susceptibles d'identifier, individuellement ou en combinaison, un ou plusieurs biens. En particulier, éviter d'utiliser des variables facteurs dont les valeurs peuvent désigner un petit nombre de biens. Une variable facteur qui identifie des quartiers dans des jeux de données de taille conséquente est appropriée, mais une variable qui identifie les biens dotés d'une piscine dans une zone où il en existe très peu doit être utilisée avec prudence. Une caractéristique qui identifie des biens uniques, telle que des coordonnées SIG exactes, des numéros de parcelle ou des adresses, est absolument proscrite.

Si le R^2 de la régression est de 80%, on s'attend à ce qu'environ 20% de la valeur du bien à évaluer provienne du résidu. Or, cela n'est qu'approximatif. Nous savons, d'après la courbe CQA-Résidu, que la majeure partie de la distribution des résidus se situe aux extrémités basses et hautes du CQA. Sur la base de la courbe CQA-Résidu et de la pente extrême pour les valeurs basses et hautes du CQA, on pourrait s'attendre à une estimation moins exacte du résidu dans ces zones. Toutefois, cela est contrebalancé par la rareté relative des biens dans ces zones et par la plus grande différence d'attrait observée entre biens adjacents. Ainsi, tandis que l'on pourrait parler de l'impact d'une fourchette de $\pm 5\%$ dans la plage médiane, aux extrêmes on se préoccuperait plutôt d'une fourchette un peu plus étroite de $\pm 2 - \pm 3\%$

Le calcul de l'exactitude de la RCA est, en général, assez difficile en raison de la non-linéarité de la courbe CQA-Résidu. Mais il est possible au cas par cas, à condition de disposer d'un classement des résidus raisonnable. La plage des scores CQA va de 0.0 à 10.0. Avec un bon modèle MARS, votre classement des résidus devrait présenter un schéma clair allant du moindre au plus grand attrait, et il ne devrait pas être difficile de noter à $\pm 5\%$ près de la plage totale, soit ± 0.5 pour le score CQA. Si la régression comprend 100 biens, cela revient à classer le bien à évaluer avec exactitude au sein d'un groupe de 10 biens. Ainsi, en nous référant aux données de vente de Burlingame du tableau 3, calculons l'exactitude (ou l'erreur) approximative pour des CQA estimés par le VE de 7.5, 5.1 et 1.9, sa conclusion de valeur pour le bien à évaluer étant de \$3,000,000. Notons que, le tableau 3 ne contenant qu'un échantillon des plusieurs centaines de biens utilisés, il faudra interpoler le résidu si nécessaire (en supposant la linéarité). Pour être clair, nous supposons que le VE est, dans chaque cas, très sûr que son classement du bien à évaluer se situe dans une fourchette de ± 5 biens au sein du classement des résidus portant sur 100 biens seulement.

CQA= 7.5 ± 0.5					
CQA	Résidu	Diff. abs.	Erreur att.	Prix de vente hyp.	Erreur en %
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	Résidu	Diff. abs.	Erreur att.	Prix de vente hyp.	Erreur en %
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	Résidu	Diff. abs.	Erreur att.	Prix de vente hyp.	Erreur en %
2.40	-\$228,411				
1.90	-\$276,381	\$47,970			
1.40	-\$326,938	\$50,557	$\pm \$49,264$	\$3,000,000	$\pm 1.64\%$

Enfin, la seule manière objective de vérifier l'exactitude est d'établir un protocole strict (que le présent article ne fait qu'esquisser à grands traits) et d'en vérifier le respect. Cela est nécessaire parce que les prix de vente réels peuvent s'écarter de la valeur de marché en raison des imperfections du marché. La comparaison des rapports d'évaluation et des ventes ultérieures peut fournir une certaine indication de l'exactitude. Cependant, les complexités et l'instabilité du marché immobilier exigent l'expertise d'évaluateurs ou de Valuation Engineers hautement qualifiés pour fournir des estimations de valeur fondées sur des normes et des protocoles établis. Ceux-ci devraient être suffisamment rigoureux pour garantir que des Valuation Engineers compétents parviennent indépendamment à une conclusion de valeur presque identique. Cela devrait constituer un objectif atteignable dans la plupart des cas où de bonnes données sont disponibles.

11.5 Le bien à évaluer comme anomalie

Puisque l'estimation de la valeur de marché d'un bien exige l'analyse de transactions de vente récentes, il est légitime de se demander si les transactions récentes représentent bien les diverses

caractéristiques d'un bien à évaluer plus ancien, dépourvu de modernisation et de réparations. Cela est particulièrement vrai dans la San Francisco Bay Area, où les transactions récentes concernent généralement des maisons ayant fait l'objet d'une remise au goût du jour destinée à en améliorer le prix de vente. Ainsi, l'état et la qualité moyens des biens comparables vendus sont généralement supérieurs à ceux de maisons plus anciennes absentes du marché depuis longtemps. Cette différence de valeur devrait être captée par l'estimation du résidu du bien à évaluer, autrement dit par son score CQA. C'est une considération importante en matière de qualité des données et d'exactitude dans l'évaluation du bien à évaluer.

11.6 L'importance du CVR2 face aux données inattendues

La régression MARS repose sur de multiples itérations de R/earth appliquées à un sous-ensemble de biens sélectionné aléatoirement, qui comprend typiquement 80% à 90% du jeu de données MLS et que l'on appelle l'ensemble d'entraînement. Les 10% - 20% de biens restants constituent l'ensemble de test. Dans cette phase, le modèle MARS est créé à partir des données d'entraînement, puis utilisé pour prédire les prix de vente des biens de l'ensemble de test, lesquels sont ensuite comparés aux prix de vente réels. Cette méthodologie est reconnue par les spécialistes de la fouille de données comme la stratégie la plus efficace pour développer un modèle robuste capable de prédire les prix de vente de biens non encore observés.

Cependant, l'utilité première de la métrique CVR2 ne réside pas dans le reflet qu'elle donne de l'exactitude factuelle du modèle de régression au regard des ventes réelles de la zone de marché, mais plutôt dans l'optimisation des capacités prédictives du modèle concernant des ventes hypothétiques de biens à une date d'effet donnée. Ces biens ne figurent pas dans les données d'entraînement et peuvent présenter des combinaisons de caractéristiques nouvelles ou inattendues, absentes à la fois des données d'entraînement et des données de test.

Dans ce contexte, comment estimer le prix de vente d'un bien qui n'a pas été rénové au cours des 15 dernières années, en supposant qu'il ait été mis en vente, par exemple, un mois avant la date d'effet de l'évaluation ? Un modèle présentant un CVR2 solide est plus susceptible de fournir l'estimation la plus exacte qu'un modèle n'ayant pas subi cette validation croisée répétée. Il est également essentiel que le jeu de données comprenne un nombre suffisant de biens comparables présentant divers degrés de rénovation, de réparation et de modernisation.

En définitive, si nous disposons d'un modèle robuste qui prédit avec exactitude les contributions à la valeur des diverses caractéristiques des biens de notre jeu de données de biens comparables, nous pouvons placer efficacement le nouveau bien à évaluer dans le classement par résidu que le modèle établit pour ces biens. Ce placement nous permet d'attribuer au bien un score CQA et un résidu appropriés.

12 Résidus

Comme l'indique le nom « Residual Constraint Approach », les résidus occupent une place centrale dans cette méthodologie. Le terme résidu désigne ici la différence entre le prix de vente net d'un bien et le prix de vente prédit par l'analyse de régression.

Un résidu positif suggère que le bien s'est vendu plus cher que son prix estimé sur la base des attributs mesurables usuels, tels que la surface habitable brute (GLA) et la superficie du terrain, ce qui laisse penser qu'il présente un attrait supérieur à la moyenne. Un résidu négatif indique que le bien s'est vendu moins cher, ce qui peut suggérer un attrait moindre ou d'autres attributs négatifs non captés.

Étant donné que les maisons plus grandes présentent vraisemblablement des résidus plus importants, on peut utiliser le résidu par pied carré, calculé à partir de la GLA, afin d'obtenir une comparaison plus normalisée. Une discussion plus approfondie de cette question est réservée à un autre article.

12.1 Ranking & Scoring (classement et notation)

Après la création du modèle MARS (Multivariate Adaptive Regression Splines) pour les variables mesurées, les résidus sont calculés pour chaque bien. Ces résidus représentent la différence entre les prix de vente nets observés et les prix de vente prédits par le modèle. Une colonne supplémentaire doit être créée pour le Residual/SF, c'est-à-dire le résidu divisé par la surface habitable de la résidence, étant entendu qu'en général, le résidu d'un bien donné croît avec la taille de la surface habitable. D'autres statistiques peuvent donner de meilleurs résultats dans certains quartiers, mais elles feront l'objet d'un autre travail.

Hypothèse 7 : « Classement par résidu »

Dans cet article, les références au « classement par résidu » doivent s'entendre comme un classement par résidu, par résidu/SF, ou par toute autre fonction des résidus que le Valuation Engineer juge la plus appropriée pour la zone de marché. L'expérience de l'auteur dans la SF Bay Area indique que le résidu/SF est généralement le plus fiable, en particulier lorsque le bien à évaluer présente une GLA (Gross Living Area) inhabituellement petite ou grande.

Ensuite, les biens sont triés selon leurs résidus, du plus négatif au plus positif. Un score de Condition-Quality-Appeal (CQA — état, qualité, attrait) est alors généré pour chaque bien, sur une échelle de 0.0 à 10.0. Ce score est calculé comme suit :

Le score CQA correspond au rang centile du résidu d'un bien parmi l'ensemble des biens. Plus précisément, le score est obtenu en déterminant le pourcentage de biens dont les résidus sont inférieurs à celui du bien considéré, puis en divisant ce pourcentage par 10. Par exemple, si le résidu d'un bien est supérieur aux résidus de 50% des biens, son score sera de 5.0.

En raison des arrondis, un bien peut recevoir un score de 10.0 : celui qui présente le résidu le plus élevé. Toutefois, un score parfait de 10.0 est théoriquement impossible car, en toute logique, aucun bien ne peut avoir un résidu supérieur à celui de 100% des biens, lui-même compris. En revanche, le score minimal de 0.0 est possible, puisqu'au moins un bien n'a aucun autre bien présentant un résidu inférieur au sien. Notons que plusieurs biens peuvent recevoir un score de 0.0 en raison des arrondis lorsque de nombreux biens sont comparés.

Le score CQA fournit donc une mesure intuitive du degré d'adéquation entre les attributs d'un bien et les attentes du modèle, les scores les plus élevés indiquant une meilleure concordance

avec les prédictions du modèle.

Il est essentiel de reconnaître que le score CQA ne peut être considéré comme un indicateur indirect de l'attrait d'un bien comparable que lorsque les conditions de la vente sont en adéquation avec sa valeur de marché. Il incombe au Valuation Engineer de déterminer si la valeur résiduelle incorpore des éléments de sur- ou de sous-évaluation par rapport au marché, en fournissant une explication complète et une investigation approfondie de la transaction de vente. Bien que les ventes significativement au-dessus ou au-dessous du marché ne soient pas fréquentes, elles peuvent survenir dans des circonstances telles que les ventes forcées, les successions, l'incompétence, la collusion, etc.

Le tableau 3 présente un extrait semi-réaliste de données MLS, accompagné des résidus et des scores CQA, en valeur totale et par pied carré. Le nombre total de biens analysés dans cette étude s'élevait à environ 600. Il convient de noter que plusieurs variables ayant fait l'objet de la régression ne sont pas affichées. Autrement dit, les résidus présentés s'entendent après prise en compte, par la régression MARS, de l'impact d'autres variables, telles que la date de vente.

TABLE 3 – Biens comparables triés par résidu

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 Courbe CQA-résidu

Le tracé du CQA en fonction du Residual/SF pour un quartier produit invariablement la courbe caractéristique CQA-résidu présentée à la [Figure 2](#). Ce graphique particulier représente des données de Burlingame, CA, vers 2022. Des courbes similaires sont généralement observées lorsque l'on trace les scores CQA en fonction des valeurs de résidu ; je préfère toutefois utiliser le Residual/SF pour sa pertinence à contextualiser le CQA par rapport à la Gross Living Area (GLA) du bien à évaluer. En multipliant le Residual/SF par la GLA, on peut convertir le CQA d'un bien en une valeur résiduelle correspondante.

Le Valuation Engineer doit évaluer avec soin si la taille du bien, en particulier la Gross Living Area (GLA), influence de manière significative l'ampleur des résidus. Cette appréciation est critique, car elle peut déterminer la méthode la plus exacte pour calculer les valeurs des biens. Par exemple, des équipements de cuisine haut de gamme et des installations de salle de bains luxueuses peuvent accroître substantiellement la valeur résiduelle d'une maison. Cependant, dans les quartiers où la taille et l'agencement des cuisines et des salles de bains sont généralement uniformes, ces améliorations peuvent ne pas être fortement corrélées avec la GLA. Leur impact sur la valeur résiduelle peut tenir moins à l'espace qu'elles occupent qu'à la qualité qu'elles représentent.

À l'inverse, des éléments et matériaux extérieurs coûteux, qui améliorent notablement l'attrait extérieur d'un bien, peuvent entretenir une relation plus directe avec la taille de celui-ci. Les maisons plus grandes peuvent accueillir des extérieurs plus élaborés, tels que des aménagements paysagers étendus, des matériaux de toiture supérieurs ou de vastes patios, qui rehaussent naturellement l'attrait global du bien et, par conséquent, sa valeur résiduelle.

Il est donc essentiel que les Valuation Engineers prennent en compte ces dynamiques lors de l'analyse des résidus. Ils doivent non seulement apprécier la valeur intrinsèque ajoutée par des installations et des caractéristiques de haute qualité, mais aussi comprendre comment ces améliorations interagissent avec les dimensions physiques du bien. Cette approche globale garantit que les évaluations reflètent avec exactitude les attributs tant qualitatifs que quantitatifs d'un bien.

Par conséquent, il est recommandé de tracer à la fois les courbes CQA-résidu et CQA-résidu/SF et d'en étudier les différences, lesquelles peuvent révéler des tendances au sein d'une zone de marché. Le Valuation Engineer peut faire calculer les deux résidus par son programme et en prendre une moyenne pondérée afin de déterminer le résidu à attribuer au bien à évaluer. Cela peut faire une différence notable, ou non.

12.3 Caractéristiques de la courbe CQA

La courbe Condition-Quality-Appeal (CQA) offre un modèle sophistiqué pour comprendre comment les valeurs des biens fluctuent en fonction de l'attrait, de l'état et de la dynamique du marché au sein d'une zone donnée. Ce modèle capte l'interaction des facteurs économiques et physiques qui influencent le marché immobilier, reflétant les réalités nuancées de l'évaluation immobilière à travers différents segments de marché.

12.3.1 Analyse des composantes de la courbe CQA

1. Décroissance exponentielle sur la partie gauche de la courbe :

À mesure que le score CQA diminue d'environ 2 vers 0, les biens présentent des caractéristiques de détérioration rapide. Cette partie de la courbe est bien représentée par une

fonction de décroissance exponentielle, ce qui souligne la rapidité avec laquelle un bien peut perdre son attrait et sa valeur en l'absence d'entretien et de remises à niveau appropriés. La dégradation physique, telle que le bois qui pourrit, les couleurs qui passent et le métal qui rouille, accélère ce déclin, en particulier dans les maisons qui ne sont pas entretenues régulièrement.

2. Croissance linéaire dans la plage médiane :

L'essentiel du marché s'inscrit typiquement dans cette catégorie, où l'attrait augmente de façon quasi linéaire entre des scores d'environ 2 et 8. Cette section reflète des biens généralement entretenus selon les standards, mais qui ne présentent pas nécessairement de qualités ou d'emplacements exceptionnels. Elle représente la vaste classe moyenne du logement, englobant la majorité du marché immobilier, où les maisons sont proposées à des prix compétitifs sur la base de leurs caractéristiques standards et de leur état général.

3. Croissance exponentielle sur la partie droite de la courbe :

Ici, la courbe s'élève de façon exponentielle entre les scores 8 et 10, reflétant le segment du marché au plus fort attrait. Les biens de ce segment appartiennent souvent à des personnes plus fortunées, qui investissent non seulement dans un entretien supérieur, mais aussi dans des améliorations qui rehaussent considérablement l'attrait et la valeur du bien. La croissance exponentielle reflète la rareté de ces biens de grande qualité et la forte demande émanant d'acheteurs aisés. Et il faut bien comprendre qu'une maison de prix modeste peut obtenir une note très élevée en termes d'attrait, tandis qu'une maison grande et coûteuse peut obtenir une note faible. Ces dynamiques sont mises au jour par un usage expert de MARS et passent généralement inaperçues auprès de la plupart des évaluateurs.

12.3.2 Dynamique du marché et implications sociales

La courbe CQA raconte implicitement l'histoire de la richesse relative et des disparités économiques au sein de différentes régions. Dans des endroits comme le comté de San Mateo, la variation des valeurs immobilières entre des secteurs comme Burlingame et des quartiers plus fortunés tels que Hillsborough, Woodside ou Atherton illustre des clivages socio-économiques importants. Cette disparité affecte tout, de l'entretien des biens aux évaluations de marché, et influence en dernier ressort la démographie des acheteurs et les tendances du marché immobilier.

En outre, l'anecdote de la personne relativement fortunée qui investit dans une maison de grande qualité au sein d'un quartier de gamme intermédiaire où elle se sent à l'aise et où, pour diverses raisons, elle préfère vivre, met en lumière un autre aspect critique de l'économie immobilière — le concept de biens « surconstruits » (*overbuilt*). De tels biens, bien que très attrayants, peuvent ne pas trouver d'acheteurs locaux capables de payer le prix qu'ils auraient pu atteindre dans une zone de marché plus chère.

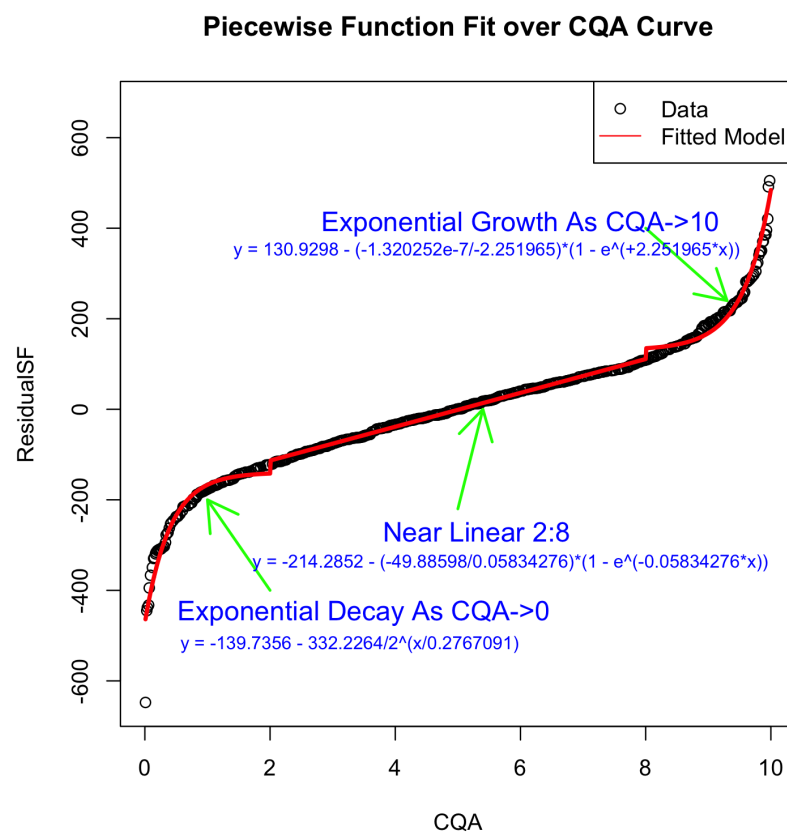
La courbe CQA-résidu est caractéristique de chaque zone de marché et ne doit pas être négligée lors de l'analyse.

12.4 Résidu du bien à évaluer

Contrairement aux biens comparables, le bien à évaluer ne possède pas de résidu déterminé de manière objective ; celui-ci doit être estimé par le Valuation Engineer (VE). Cette estimation introduit un degré de subjectivité dans le protocole du Residual Constraint Approach (RCA), marquant un point critique où le jugement personnel joue un rôle déterminant.

Pour estimer le score de Condition-Quality-Appeal (CQA), le VE analyse méticuleusement les biens classés, à la recherche de motifs discernables au sein du classement. Cette analyse implique

FIGURE 2 – Fonction caractéristique CQA-résidu



souvent un examen détaillé des photos MLS associées à chaque bien. Un examen exhaustif est impraticable pour un vaste jeu de données, tel que 600 biens. Le VE applique plutôt un raisonnement pratique pour approcher la position probable du bien à évaluer dans le classement global. Les biens nécessitant des réparations importantes, ou « fixers », se situent généralement près du bas du classement, tandis que ceux qui présentent des rénovations et un attrait supérieurs tendent à se classer près du sommet.

Le VE adopte ensuite une approche plus ciblée, en « sautant » au sein du classement pour identifier les biens comparables dont l'attrait correspond étroitement à celui du bien à évaluer. À mesure que la plage se resserre, le VE prête une attention croissante aux similitudes entre les biens, en particulier sur des aspects tels que la qualité, l'état, l'esthétique de la conception (y compris des éléments comme la peinture, les boiseries, la maçonnerie de pierre et les fenêtres), l'utilité fonctionnelle et d'autres caractéristiques pertinentes.

Il est crucial de comprendre que les jugements du VE sont fondamentalement guidés par les classements établis par le marché. Cela garantit que les éléments subjectifs du processus restent ancrés dans les réalités du marché, en visant une évaluation juste et exacte du bien à évaluer.

12.5 Conclusion de valeur RCA

12.5.1 Considérations générales

La détermination d'un score CQA est essentielle pour établir la valeur résiduelle estimée d'un bien, ces éléments étant intrinsèquement liés. S'il est possible d'évaluer un bien sans CQA, cette métrique fournit un contexte précieux en illustrant la position du bien par rapport aux autres dans la zone de marché. Par exemple, un score CQA de 3.2 indique que le bien présente un attrait supérieur à celui de 32% des biens comparables du voisinage. Si le score attribué fait l'objet d'objections, les contestataires doivent consulter les classements et justifier tout ajustement proposé sur la base de preuves photographiques ou d'autres données substantielles. Si le système de classement est robuste et que le Valuation Engineer (VE) a positionné le bien avec exactitude dans ce cadre, il est peu probable que des divergences significatives apparaissent.

Lors de la phase de rapport, on sélectionne typiquement 6 à 12 ventes comparables récentes et pertinentes pour étayer la conclusion de valeur. Les différences entre ces biens comparables et le bien à évaluer sont analysées, chaque variable faisant l'objet d'un ajustement correspondant. Ces ajustements sont ensuite appliqués aux prix de vente nets des comparables afin d'obtenir des prix de vente ajustés. Comme démontré mathématiquement, ces prix ajustés doivent converger vers le prix de vente estimé du bien à évaluer.

12.5.2 Révision, audit et réclamations

Une fois le résidu estimé du bien à évaluer déterminé, des ajustements résiduels peuvent être calculés pour tous les biens comparables. Ces ajustements servent de contraintes pour les décompositions manuelles supplémentaires effectuées par le VE, adaptées aux variables CQA spécifiques jugées pertinentes. Il importe de souligner que ces décompositions détaillées sont explicatives et n'influencent pas l'évaluation finale, comme l'affirme la Démonstration II.

Concernant les réclamations soumises par des utilisateurs insatisfaits de la conclusion de valeur de l'évaluation, un évaluateur traditionnel doit typiquement perdre du temps à expliquer pourquoi tel ou tel bien, dont le prix est plus élevé ou plus bas, n'a pas l'impact sur la valeur attendu par l'utilisateur. Avec le RCA, toutes les ventes comparables possibles (habituellement 100-600) sont typiquement évaluées précisément à la même valeur que la conclusion de prix, de sorte qu'il faut peu d'efforts pour présenter les ajustements relatifs à un bien arbitraire. L'ensemble du processus

RCA est automatisé et objectif, à l'exception de la notation du résidu du bien à évaluer. Ainsi, un autre avantage du RCA est que la critique des imperfections du jugement subjectif du VE se limite pour l'essentiel à ce seul point — et elle peut être gérée de manière efficace et efficiente. De plus, en supposant un R^2 d'environ 80%, l'impact d'une erreur de $\pm 10\%$ dans l'estimation du résidu du bien à évaluer se traduit par une erreur de $\pm 2\%$ dans l'estimation finale de la valeur.

Dans de nombreux cas, le bien à évaluer s'insère si parfaitement dans le classement des biens par résidu que l'on peut s'attendre à une exactitude de l'ordre de $\pm 1\%$.

La somme du résidu estimé et du prix de vente du bien à évaluer estimé par MARS donne la conclusion de valeur finale, appelée conclusion de valeur du Residual Constraint Approach (RCA). Ce processus garantit une approche complète et transparente de l'évaluation immobilière, renforçant la confiance dans l'exactitude et l'équité de l'appréciation.

13 Conclusion

La Residual Constraint Approach est un cadre porteur de trois contributions substantielles : une propriété structurelle d'anti-ancrage découlant de l'ajustement du modèle sur l'ensemble du marché plutôt que sur un petit ensemble de biens comparables (Démonstration I) ; un résultat d'invariance qui permet à la décomposition du résidu de servir de couche explicative sans perturber la conclusion de valeur (Démonstration II et la remarque qui la suit) ; et une régularité empirique — la courbe caractéristique CQA-Résidu — qui semble se reproduire d'une zone de marché à l'autre sous une forme caractéristique exponentielle-linéaire-exponentielle. Les limites du cadre sont tout aussi substantielles : la décomposition du résidu est invariante mais non identifiante, l'étape de positionnement CQA repose sur le jugement du VE, lequel n'a pas encore été validé par une étude formelle de concordance inter-évaluateurs, et les affirmations relatives à l'exactitude demeurent en attente d'une évaluation comparative face aux approches alternatives.

Ce que le présent article propose, c'est un cadre et un protocole. Ce qu'il ne propose pas encore, mais que le programme de recherche plus vaste actuellement en cours propose, c'est une chaîne de traitement exécutable et traçable à des fins d'audit, qui opérationnalise le protocole dans un logiciel doté de paramètres par défaut défendables. Cette chaîne de traitement — le package earthUI sur CRAN pour l'étape des fonctions de base MARS, le package glmnetUI pour le modèle structurel régularisé sur ces bases, et le package mgcvUI pour le lissage spatial — fait l'objet d'articles compagnons dans le présent numéro et constitue le pont entre la RCA en tant que méthode et la RCA en tant que livrable. L'écart entre doctrine et pratique dans la réglementation de l'évaluation immobilière (CRAYTOR, 2026, ce numéro) fournit la raison pour laquelle ce pont importe : les résultats d'évaluation sont économiquement consommés comme des intrants de risque prospectifs, et un cadre qui rend l'opinion de valeur actuelle auditable et reproductible est une condition préalable à tout engagement honnête avec cet usage prospectif. La RCA n'est qu'une étape de ce projet plus vaste. Elle n'en constitue pas la totalité, et le présent article ne prétend pas le contraire.

Références

- AMBROSE, Brent W. et al. (2025). « Do Appraiser and Borrower Race Affect Mortgage Collateral Valuation? » In : *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 ». In : *Valuation Engineer Journal* 1.1.
- FRIEDMAN, Jerome H. (1991). « Multivariate Adaptive Regression Splines ». In : *The Annals of Statistics* 19.1, p. 1-141.
- HASTIE, Trevor, Robert TIBSHIRANI et 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 et Michael LACOUR-LITTLE (2020). « AVM versus Appraisal-Based Underwriting in Refinance Mortgages: The Trade-off Between Noise and Bias ». In : *SSRN Electronic Journal*. DOI : [10.2139/ssrn.3603327](https://doi.org/10.2139/ssrn.3603327).
- KUHN, Max et 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>.

- MILBORROW, Stephen (2024b). *Notes on the Earth Package*. Rapp. tech. URL : <http://www.milbo.org/doc/earth-notes.pdf>.
- (2024c). *Variance Models in Earth*. Rapp. tech. URL : <http://www.milbo.org/doc/earth-varmod.pdf>.
- PERRY, Andre M., Jonathan ROTHWELL et David HARSHBARGER (2018). *The Devaluation of Assets in Black Neighborhoods*. Rapp. tech. 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*. Rapp. tech. Brookings Institution.
- PROPERTY APPRAISAL AND VALUATION EQUITY TASK FORCE (2022). *Action Plan to Advance Property Appraisal and Valuation Equity*. Rapp. tech. U.S. Department of Housing et Urban Development.
- TZIOUMIS, Konstantinos (2016). « Appraisers and Valuation Bias: An Empirical Analysis ». In : *Real Estate Economics* 45.3, p. 679-712. DOI : [10.1111/1540-6229.12133](https://doi.org/10.1111/1540-6229.12133).
- WILLIAMSON, Jake et Mark PALIM (2022). *Appraising the Appraisal: A Closer Look at Divergent Appraisal Values for Black and White Borrowers Refinancing Their Home*. Rapp. tech. Fannie Mae.
- ZHANG, Wengang (2020). *MARS Applications in Geotechnical Engineering Systems: Multi-Dimension with Big Data*. Singapore : Springer. ISBN : 978-981-13-7421-0.