잔차 제약 접근법 (RCA): 방법론적 틀과 프로토콜 (한국어판)

2025 년 *3* 월 *5* 일 *Zenodo* 에 최초 발표된 논문의 개정판에 대한 한국어 번역본. 인공지능의 도움을 받아 번역하고 편집부가 검토했습니다. 영문판이 정본입니다.

William Bert Craytor, SRA

제 1 권 제 1 호 · 2026 년 여름
2026 년 7 월 1 일 발행
DOI: 10.67330/kgebxf46

© 2026 William Bert Craytor. Published by Pacific Vista Net.
이 오픈 액세스 논문은 크리에이티브 커먼즈 저작자표시 4.0 국제 라이선스 (CC BY 4.0)의 조건에 따라 배포됩니다.

초록. 전통적 감정평가 (appraisal) 방법의 중대한 한계는 짝지은 사례 비교분석 (matched pair analysis) 과 같은 낡은 수작업 기법에 의존한다는 점에 있다. 이러한 기법은 부동산 특성의 가치기여 (value contribution) 를 마치 존재하지 않는 것처럼 건너뛴다. 전통적 거래사례비교표 (sales grid) 에는 가치기여를 적을 난이 없고 보정 (adjustment) 난만 있을 뿐이다. 그러나 실제로 가치기여야말로 전통적 감정평가사 (appraiser) 의 과대평가와 과소평가를 막아 주는 수학적 제약 (mathematical constraint) 의 필수적 토대를 제공한다.

잔차 제약 접근법 (Residual Constraint Approach, RCA) 은 다변량 적응 회귀 스플라인 (Multivariate Adaptive Regression Splines, MARS) 과 특성 가치기여에 대한 엄밀한 수학적 제약을 함께 사용하는 다단계 감정평가 절차를 통합함으로써 전통적 거래사례비교법 (Sales Comparison Approach, SCA) 을 발전시킨다. 이때 특히 문제가 되는 것은 측정되지 않은 특성, 즉 미관, 디자인, 상태, 품질과 같은 잠재변수 (latent variable) 이며, 그 가치는 전통적 감정평가사의 주관적 판단과 편의 (bias) 에 좌우되는 것이 보통이다.

RCA 의 수학적 제약은 측정 가능한 부동산 특성이 거래가격에 기여하는 가치를 추정하기 위해 MARS 회귀분석 (regression) 을 전문적으로 적용하는 데 의존한다. 그러나 이 추정치는 통상 실제 거래가격의 약 80% 만을 설명하며, 디자인, 기능적 효용, 상태, 품질과 같이 주관적으로 평가되는 특성은 여기에서 제외된다. 샌프란시스코 베이 에어리어 (San Francisco Bay Area) 에서는 이러한 주관적 구성요소를 나타내는 잔차 (residual) 가 부동산 총가치의 약 20% 를 차지한다. 이 80/20 이라는 수치는 전형적인 데이터 여건 아래 베이 에어리어에서 저자가 겪은 경험을 반영한 것이며, 시장을 불문하고 변하지 않는 상수로 읽어서는 안 된다. 달성 가능한 R^2 가 더 낮은 시장에서 이 틀이 제시하는 대응은, 그러한 한계를 그대로 받아들이고 더 넓은 오차범위를 보고하는 것이 아니라 측정되지 않은 가치결정요인을 찾아내어 측정하는 것이다. 해안 지역에서의 바다까지의 거리, 구릉지 주거지역에서의 표고, 그리고 MLS 데이터가 제공하지 않을 때 외부 자료원에서 수집하는 유사한 변수들이 그 예이다. 다른 분야에서는 MARS 의 잔차를 흔히 추정오차로 취급하지만, 부동산 감정평가에서 잔차는 잠재변수 가치의 간접적 척도 역할을 한다. 처음에는 실용적이지 않아 보일 수 있으나, 잔차는 각 구성요소의 기여액 합계가 잔차와 같기만 하다면 최종 감정평가 결론에 영향을 주지 않으면서 서술적 구성요소들로 의미 있게 분해될 수 있음이 증명으로 밝혀진다.

이 방법론은 저자가 20 년에 걸쳐 실증적으로 적용해 온 경험에서 비롯되었으며, R 및 Python 환경에서 MARS 를 반복적으로 구현하는 과정을 통해 발전해 왔다.

주제어: 부동산 감정평가, 주거용 감정평가, 잔차 제약 접근법, MARS, 다변량 적응 회귀 스플라인, 거래사례비교법, 잔차분석, 잠재변수, 자동 감정평가

목차

1 서론 **4**

2 배경 **6**

3 감정평가공학자 **8**

4 가정 **9**

5 RCA 작업 흐름 **11**

5.1 MARS 분석의 유형 11

5.2 처리 작업 흐름 11

5.3 데이터 작업 흐름 12

5.3.1 MARS 를 위한 데이터 요건 12

5.3.2 프로토콜: 종결 거래 리스팅 13

5.4 프로젝트 폴더 구조 15

5.4.1 프로토콜: 다중 MARS(earth) 실행 16

5.5 핵심 Excel 워크북 17

5.6 백업 17

6 워크북 표기법 **18**

6.1 서술적 차원 18

6.1.1 프로젝트 차원 (“a”) 19

6.1.2 실행 버전 차원 (“r”) 20

6.1.3 작업 흐름 단계 (“w”) 20

6.1.4 Excel 워크북 시트 (“s”) 20

6.1.5 부동산 리스팅 (“p”) — MLSData 행 21

6.1.6 MLS 변수 (“v”) — MLSData 열 21

6.1.7 최종 표기 예시 21

6.1.8 정당화 22

6.1.9 행렬은? 22

6.1.10 첫 번째 단순화 22

6.1.11 두 번째 단순화 23

7 RCA 거래사례비교표 **25**

8 변수 **28**
8.1 변수 기원 온톨로지 . . . 28
8.2 변수 용도 온톨로지 . . . 29
8.3 2 차 및 3 차 항 . . . 30
8.3.1 보정 변수 . . . 31
8.3.2 집계 변수 . . . 31
8.3.3 특수 변수 및 표기법 . . . 32
8.3.4 잔차의 분해 . . . 32

9 RCA 의 핵심 증명 **33**

10 MARS 모형 **35**
10.1 모형 변수 및 상호작용 "g" 함수 . . . 35
10.2 g 함수의 색인 . . . 35
10.3 그래프의 차원 . . . 36
10.4 최종 모형 식 . . . 36
10.5 모형을 그래프로 나타내는 데 필요한 차트의 수 . . . 37

11 데이터의 특성 **39**
11.1 잔차의 목적 . . . 39
11.2 근사 회귀변수로서의 상태 및 품질 등급 . . . 39
11.3 데이터 오류 . . . 40
11.4 정확성이란 무엇인가 . . . 41
11.5 이상치로서의 대상 부동산 . . . 43
11.6 예상 밖 데이터를 다루는 데서 CVR2 의 중요성 . . . 43

12 잔차 **44**
12.1 순위와 점수 . . . 44
12.2 CQA-잔차 곡선 . . . 46
12.3 CQA 곡선의 특성 . . . 46
12.3.1 CQA 곡선 구성요소의 분석 . . . 46
12.3.2 시장 동학과 사회적 함의 . . . 47
12.4 대상 부동산의 잔차 . . . 48
12.5 RCA 가치 결론 . . . 48
12.5.1 일반적 고려사항 . . . 48
12.5.2 검토, 감사, 이의 . . . 48

13 결론 **50**

1 서론

이 논문은 주거용 부동산을 비롯한 각종 부동산의 시장가치 (market value) 를 정확하게 추정하기 위한 현대적 접근법의 논리적, 수학적 토대를 제시한다. 이 방법은 다변량 적응 회귀 스플라인 (Multivariate Adaptive Regression Splines, MARS) 소프트웨어를 북부 캘리포니아의 주택 — 특히 샌프란시스코 베이 에어리어 일대 지역사회의 주택 — 에 적용해 온 20 년의 경험에서 개발된 것으로, 감정평가 정밀도에서 상당한 진전을 제공한다. MARS 는 부동산 특성과 가치 사이의 비선형 관계를 능란하게 포착하는 한편, 영향력이 가장 큰 변수를 효율적으로 식별하고 그 우선순위를 정함으로써 감정평가의 질을 높인다. 이러한 능력 덕분에 감정평가사 (appraiser) 는 새로운 시장 지역의 가치 동학을 전통적 방법이 허용하는 것보다 훨씬 빠르게 밝혀낼 수 있다. MARS 는 여기서 제안하는 잔차 제약 접근법 (Residual Constraint Approach, RCA) 의 중추를 이룬다. RCA 라는 이름은 다음과 같은 그 엄밀한 틀에서 비롯된다.

- RCA 는 측정된 특성과 측정되지 않은 특성 모두의 가치기여 (value contribution) 에 엄격한 제약을 부과한다.
  1. 잔차 (residual) 를 포함하여 MARS 에서 도출된 모든 가치기여의 합은 각 비교사례 (comparable) 부동산의 순 거래가격과 같아야 한다.
  2. 잔차는 하위 구성요소로 분해하여 각각에 값을 배정할 수 있으나, 그 합계는 해당 비교사례의 잔차와 같아야 한다.
- RCA 는 미적 매력, 상태, 품질과 같은 주관적 특성의 가치 총액을 잔차에 묶어 둠으로써 그 추정에서 발생하는 편의 (bias) 를 최소화한다. 유능한 감정평가공학자 (valuation engineer) 가 MARS 모형을 구축한 경우 이 잔차는 정확하게 계량된다.

주관적 특성이 시장가치에 큰 영향을 미치는 주거용 부동산을 일차 대상으로 설계되기는 했으나, RCA 는 더 넓은 범위의 자산 감정평가에도 적용될 잠재력을 지닌다. 미국 주거용 부동산 시장의 규모가 약 45 조 달러로 추정된다는 점을 고려하면 정확한 감정평가는 매우 중요하다. 패니메이 (Fannie Mae) 는 통상적인 감정평가가 실제 거래가격에서 약 5% 벗어난다고 밝히고 있으나, 감정평가사가 거래가격에 맞추라는 압력을 받는 경우가 많으므로 이 수치는 편의를 가린다. 더욱이 감정평가의 60%–70% 가 — 비교 가능한 거래가격이 존재하지 않는 — 재융자 (refinancing) 를 뒷받침하는 것이므로, RCA 가 제공하는 정밀성의 필요는 자명해진다. 연방준비제도 (Federal Reserve) 의 2012 년 분석은 2008 년 위기 이전에 이루어진 감정평가의 15%–20% 가 거래가격에서 ±10% 넘게 벗어났다고 추정했는데, 이 범위는 부동산 유형과 시장에 따라 다르기는 하나 오늘날의 재융자 감정평가에도 여전히 유효할 가능성이 높다. 정확성은 왜 중요한가? 담보인정비율 (loan-to-value, LTV) 이 낮은 거래에서는 — 예컨대 신용도가 우수한 매수인이 100 만 달러 주택에 50 만 달러를 계약금으로 넣는 경우 — 대출기관이 경매 (foreclosure) 위험이 미미하다고 보아 감정평가를 면제해 주기도 한다. 그러나 감정평가를 건너뛴 매수인은 과다 지급의 위험을 지며, 매수 후에야 비로소 하자를 발견하게 된다. 이는 RCA 와 같은 견고한 감정평가 방법의 가치를 부각한다. 감정평가의 정확성을 계량하는 일은 여전히 어렵다. 연구자들에게는 "시장가치" 로부터의 편차를 직접 재는 척도가 없어 간접적인 대리지표에 의존한다. 주관적 판단과 들쭉날쭉한 감정평가사 역량에 뿌리를 둔 전통적 감정평가 방법은 분석을 한층 복잡하게 만든다. 사용되는 전략에는 다음과 같은 것들이 있다.

1. Tzioumis (2016) 는 11 개 주의 감정평가사 4,782 명과 감정평가 259,436 건을 연구하여, 3 분의 1 은 계약가격의 1% 이내, 절반은 계약가격보다 1%–3% 높은 구간, 그리고 소수는 평균 5% 높은 수준이라는 사실을 발견했다. 이는 상향 편의를 시사한다. 이 연구는 감정평가사를 두 유형 — 계약가격에 근접하는 "시장가치형" 과 일관되게 더 높게 평가하는 "시장가치 초과형" — 으로 구분했다.

2. Krivorotov and LaCour-Little (2020) 는 매매 목적 감정평가의 90% 가 계약가격 이상이었음을 발견하고, 특히 재융자 맥락에서 희망 가치에 맞추려는 경향인 "정박 편의 (anchor bias)" 를 부각했다.

이러한 연구들은 감정평가 실무에 편의가 존재함을, 그리고 그것이 전통적 도구의 한계로 증폭됨을 드러낸다. 만약 모든 감정평가사가 한결같이 유능하고 편향되지 않았다면 그들의 평가액은 수렴했을 것이다. 일부 집단에서 과대평가가 지속된다는 사실은 구조적 문제를 알리는 신호다.

위에서 인용한 편의 연구들과 그것이 가리키는 구조적 양상은 전통적 거래사례비교 실무의 더 깊은 특징을 반영하며, RCA 는 바로 이 특징을 다루기 위해 설계되었다. 전통적 방법론은 인식론적 지위가 크게 다른 두 개의 경로로 작동한다. 측정 가능한 특성 경로 — GLA(주거면적), 방 수, 대지면적, 연식, 위치 — 에서 감정평가사는 짝지은 사례 비교분석 (matched-pair analysis) 이나 비교거래사례에 대한 국지적 선형 회귀분석을 통해 보정치를 객관적으로 도출할 것이 요구된다. 유능한 감정평가사는 이 작업을 수행하며, 그 결과로 나온 보정치는 적어도 데이터에 비추어 대체로 방어 가능하다. 그러나 여기서조차 전통적 비교표는 보정치만 기록할 뿐, 가치기여를 적을 난은 아예 갖고 있지 않다. 또한 전통적 보정치는 가치기여 사이의 차액으로 계산되지도 않는다. 그것은 밑에 깔린 기여액을 한 번도 계산하지 않은 채 짝지은 사례나 유사한 절차로부터 직접 추정된다. 이와 달리 RCA 에서는 모든 보정치가 해당 특성에 대한 대상 부동산과 비교사례의 가치기여 사이의 차액이며, 증명 II 가 그 귀결을 정밀하게 밝히는 구조가 바로 이것이다. 주관적 특성 경로 — 상태, 시공 품질, 디자인, 기능적 효용, 미관 — 에서는 동원할 수 있는 뒷받침 절차가 빈약한 것에서 실낱같은 것에 이른다. 주관적 특성의 보정치는 직접적인 비교사례 집합 밖에 있는 부동산에 대한 짝지은 사례 비교분석으로, 인정된 동료 실무 관행에 대한 원용으로, 또는 감정평가사가 해당 시장에서 문서로 남긴 경험에 대한 참조로 뒷받침될 수 있다. 이 세 가지 모두 전문직 기준에는 형식상 부합한다. 그러나 그 어느 것도 측정 가능한 경로가 일상적으로 달성하는 통계적 엄밀성에는 미치지 못한다.

그 결과는 기준이 요구하는 바와 방법론이 제공할 수 있는 바 사이의 절차적 비대칭이다. 기준은 측정 가능하든 주관적이든 모든 보정치에 대한 뒷받침을 요구한다. 방법론은 한 경로에서는 견고한 뒷받침을 제공하지만, 다른 경로에서는 표본 크기가 하나뿐인 일화적 짝지은 사례에서부터 다른 감정평가사들이 관례적으로 쓰는 값의 인용에 이르기까지 여러 층위의 더 약한 절차들을 내놓는다. 조망 보정치에 대해 짝지은 사례 작업을 성실히 수행하는 정직하고 유능한 감정평가사는 근거 없이 동료 관행을 인용하는 감정평가사보다 의미 있게 나은 뒷받침을 산출하지만, 표준 양식으로 두 감정평가서를 비교하는 심사자는 이 둘을 쉽게 구별할 수 없다. 양식이 기록하는 것은 보정치의 값이지 그 도출의 엄밀성이 아니기 때문이다. 주관적 특성에 대해 더 강력한 절차가 실무 감정평가사들에게 일반적으로 제공된 적이 없었기에, 방법론은 이 비대칭을 용인해 왔다.

RCA 는 두 경로 모두를 개선하지만, 그 방식은 범주적으로 다르다. 측정 가능한 경로에서 MARS 회귀분석은 짝지은 사례 비교분석이나 통상적인 선형 회귀분석이 표현할 수 없는 비선형 관계, 변수 간 상호작용, 기저함수 (basis function) 구조를 포착한다. 전통적 비교표의 3 9 개 비교사례가 아니라 해당 시장 지역의 적격 거래사례 전체 — 통상 수백 건 — 에 대해 적합시킨 결과, MARS 에서 도출된 가치기여는 객관적이면서 동시에 그것이 대체하는 방법론보다 실질적으로 더 정확하다. 주관적 경로에서 RCA 는 전통적 뒷받침 절차 어느 것과도 종류가 다른 일을 한다. RCA 는 개별 주관적 특성의 보정치를 추정하려는 시도 자체를 하지 않는다. 모든 주관적 특성 — 잠재변수 (latent variable) — 의 기여는 집합적으로 잔차로 나타난다. 즉 각 비교사례의 실제 순 거래가격과 MARS 가 예측한 가격의 차액이며, 이는 감정평가사의 개입 없이 데이터만으로 산출된다. 그 잔차를 디자인, 기능적 효용, 조망, 상태, 품질 등 명명된 구성요소로 분해할 때는 그 합계가 잔차 총액과 같아야 한다는 제약이 걸리며, 증명 II(9절) 에 따라 그 제약을 만족하는 어떤 분해도 동일한 가치 결론을 낳는다. 작업 흐름에 남는 판단은 비교사례들의 순위화된 잔차 분포 안에서 대상 부동산의 위치를 정하는 일이며, 이는 검토 가능한 순위표에 비추어 이루어지고 감정평가서에 명시적으로 문서화된다. 부실한 RCA 감정평가도 여전히 가능하다 — 경험이 부족한 VE 는 과적합된 MARS 모형을 만들거나, 대상 부동산을 개연성 없는 자리에 놓거나, 데이터의 이상치를 놓칠 수 있다 — 그러나 그 부실함은 표준 양식 심사로는 드러낼 수 없는 방식의 부실함이 아니라, 심사자가 모형과 순위표를 검토함으로써 짚어낼 수 있는 방식의 부실함이다.

재융자 문제를 넘어, 감정평가사는 소유자에게 정확한 평가액을 제공할 의무를 지며, 오차가 수만 내지 수십만 달러의 손실로 이어질 수 있는 고액 매수에서는 더욱 그러하다. 그럼에도 많은 이들이 감정평가에 600 1,000 달러를 쓰는 것을 주저하며 그 부가가치를 의심한다. RCA 와 같은 진보된 방법은 초기에는 자원 집약적이지만 더 높은 정밀도를 약속한다. 개선과 경험이 쌓이면 그 효

율성 — 그리고 채택률 — 은 높아질 것이다.

2 배경

저자의 학문적 배경은 수학, 컴퓨터과학, 통계학이다. 직업상으로는 소프트웨어 엔지니어이며 감정평가와 회계 분야의 부수적 경력을 갖고 있다. 2002 년부터 간헐적으로 감정평가를 수행해 왔고, 2004 년경부터는 주거용 부동산의 보정치와 가격 추세를 산정하는 데 MARS 를 적용해 왔다.

MARS 로 감정평가를 한다는 것은 감정평가에서 "깊은 경험" 을 쌓는 일이다. MARS 는 시장가치에 기여하는 변수와 그 값의 변화가 거래가격에 미치는 영향을 다른 어떤 통계적 방법보다도 상세하게 반영하는 모형을 생성한다. 따라서 MARS 로 감정평가를 하는 감정평가사는 MARS 모형이 산출하는 변수값과 거래가격 사이의 인과관계를 끊임없이 탐구하게 된다.

투명성을 위해 밝히자면, RCA 방법론은 다음과 같은 특징을 가진 주거지역에 가장 잘 적용된다.

1. 부동산들이 서로 상당히 다른 복합적 근린지역
    - 노후 주택
    - 개보수 품질의 편차
    - 다양한 건축양식과 디자인
    - GLA(주거면적), 대지면적, 표고, 조망 등 부동산 속성의 편차.
2. 복합적 시장 여건
    - 빠르게 변하는 시장 여건
    - 다양한 유형의 매수인
3. 양질의 데이터 원천: RCA 프로토콜은 데이터 마이닝에 기초하므로 풍부한 양질의 데이터를 필요로 한다. 데이터가 희박한 지역에도 쓸 수는 있으나, 소요되는 시간과 비용에 비해 그만한 가치가 없을 수 있다.
4. 시장가치의 정확한 추정이 요구되는 경우. RCA 는 대부분의 경우 ± 1–2% 의 정확도를 제공할 수 있어야 한다. 감정평가공학자의 보고서는 최종 가치 결론의 정확성에 관한 조건을 언급해야 한다. 데이터나 정보가 부족하거나, 데이터 마이닝을 통해 시장의 일관된 양상이 드러나지 않았다면 정확도는 떨어질 것이며, 이 사정을 의뢰인에게 알려야 한다. 구체적으로, 높은 R2 와 CVR2(교차검증 R2. MLS 데이터 집합을 무작위 훈련 분할과 검정 분할로 반복 분할하여 계산한다), 의미 있는 MARS 모형, 그리고 순위화된 잔차에서 나타나는 부동산 CQA(상태-품질-매력) 특성의 뚜렷한 양상이 가치 결론의 정확성과 신뢰성을 판정하는 주된 수단이다.

지난 20 여 년에 걸친 이들 RCA 프로토콜의 개발은 MLSListings Inc.(Sunnyvale, CA) 및 현재 MLSListings 를 통해 접근 가능한 다른 MLS 의 이용 가능성에 기반한 것으로, 저자의 데이터 분석 경험의 95% 이상이 여기서 나왔다. 저자의 전문성은 북부 캘리포니아의 주요 15 개 카운티에 걸친 방대한 감정평가 업무와 부차적인 10 개 카운티에서의 추가 경험에서 비롯되었다.

RCA 방법론은 다음 두 가지 핵심 요건을 갖춘 시장에서 가장 효과적이다.

1. 고품질 감정평가 데이터의 풍부한 이용 가능성
2. 다음 특징을 갖는 복합적, 이질적 주거지역:
    - 다양한 건축양식
    - 개보수 정도의 편차
    - 수십 년에 걸친 개발 양상

캘리포니아 카운티 경험

이 논문에서 제시하는 서술은 다음 북부 캘리포니아 카운티들에서 부동산을 평가한 경험에 기초한다.

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

위 지역들은 실질적으로 실리콘밸리와 그 주변 카운티를 포괄하며, 센트럴 밸리와 시에라 산기슭을 포함한 보다 농촌적인 지역까지 뻗어 있다. 저자는 미국의 덜 발달된 다른 지역의 많은 감정평가사가 쓸 수 있는 것에 비해 언제나 풍부하고 상대적으로 정확한 데이터에 접근할 수 있었다는 점을 밝혀 둔다.

사용된 MLS 데이터는 RESO 인증을 받았으며 RCF(RESO Common Format) 를 준수한다. RCF 인증에도 불구하고 일부 부동산에서는 통상 부차적 중요성을 갖는 데이터가 누락되거나 부정확한 경우가 있는데, 이는 최초 세무 평가관의 오류이거나 이후 중개인의 MLS 입력 오류에서 비롯된다. 이러한 오류에는 특성이 존재하지 않을 때 아예 입력을 하지 않는 것과 같은 일정한 양상이 있는 경우가 많다. 예컨대 카포트가 없을 때 “carport spaces”(카포트 주차대수) 에 0 을 입력하는 대신 아무것도 입력하지 않는 식이다. 카포트가 드문 지역에서는 분석자가 카포트에 대한 쓸모 있는 보정치를 얻기 위해 결측 데이터를 0 으로 대체할 수 있다. R/earth 를 비롯한 MARS 구현체들이 결측 데이터를 처리할 수 있기는 하지만, 이를 어떻게 다룰지는 감정평가공학자 (Valuation Engineer, VE) 가 결정할 몫이다.

MLSListings 데이터는 플로리다에서 하와이에 이르는 다른 몇몇 주의 MLS 에서도 얻을 수 있다. 이는 프로토콜과 코드가 여러 주에서 얼마나 잘 작동하는지 시험하는 데 유용하다.

3 감정평가공학자

정의 1: 감정평가공학자

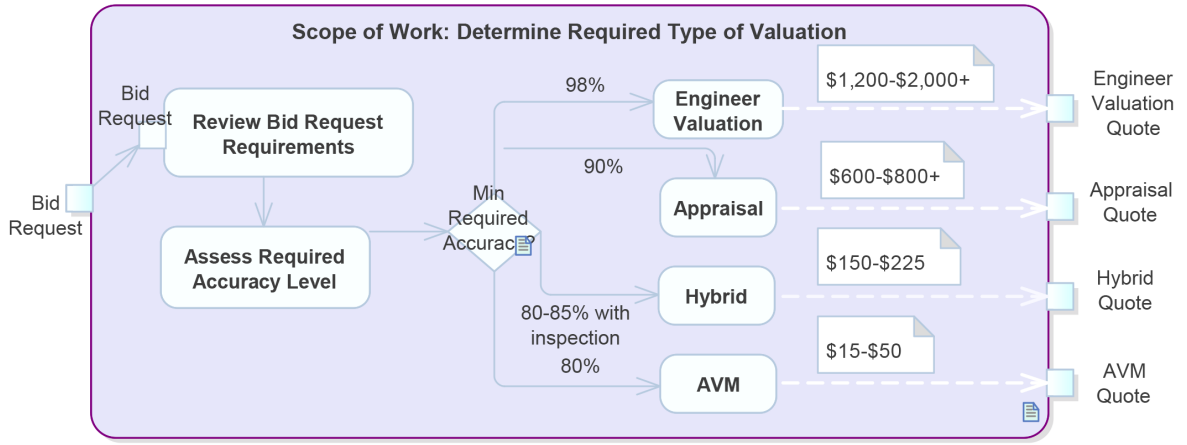
모호함을 피하기 위해, “감정평가공학자”(Valuation Engineer, 줄여서 “VE”) 라는 용어는 면허를 가진 감정평가사의 전문성에 더해 데이터 마이닝, 통계학, 프로그래밍, 상세한 프로토콜, 인공지능을 활용하여 복합적 부동산의 시장가치를 판정하는 데 관한 고급 지식과 기술, 경험을 겸비한 사람을 가리킨다. 아울러 감정평가는 사기 가능성이 이미 확인된 영역이므로 어떤 형태로든 면허를 요하는 업무라는 점을 인식하는 것도 중요하다.

대출기관은 통상 다음과 같은 몇 가지 감정평가 선택지를 제시한다.

1. 자동가치산정모형 (Automated Valuation Models, AVM) — 비용이 가장 낮은 선택지
2. 하이브리드 감정평가 — 무면허 조사자를 활용하되 면허 감정평가사가 감독
3. 표준 감정평가 — 면허 감정평가사가 수행

감정평가 업무에서 통상 고려 대상에 들지 않는 것이 공학자 수준의 감정평가 (Engineer-level Appraisal) 다. 이를 수행할 기술 역량을 갖춘 감정평가사가 애초에 많지 않다. 그러나 그 수요는 마땅히 있어야 한다. 이러한 감정평가는 RCA 방법론 (MARS 와 제약) 을 활용함으로써 현저히 높은 정확도를 제공하고 상당히 양호한 객관성을 거의 보장할 수 있다.

신규 택지개발지구의 규격화된 부동산에는 표준 감정평가로 대체로 충분하다. 그러나 30 년 이상에 걸친 상당한 개보수 이력이 있는 지역의 노후 주택에는 공학자 감정평가가 권장된다. 두 방법 사이의 선택은 부동산의 특성과 시장 여건에 기초한 체계적 의사결정 절차를 따라야 한다.



4 가정

가정 1: 지리적 범위 및 부동산 유형

이 논문의 범위는 주로 캘리포니아의 대도시권 및 도시 지역에 있는 단독주택과 콘도미니엄에 초점을 맞춘다. 거래사례비교법은 다른 부동산 유형과 지역에도 적용되지만, 농지, 다세대 부동산, 상업용 부동산을 평가할 때는 통상 수익방식과 원가방식이 더 지배적인 역할을 한다. 그러한 경우 RCA 접근법은 수정이 필요하거나 적용 범위가 제한될 수 있다.

가정 2: 데이터 저장

CSV 와 Excel 은 MLS 시스템에서 널리 사용되며 RCA 프로그램에 데이터를 공급하는 주된 방식으로 상정한다. 다만 브로커 피드를 이용해 SQL 데이터베이스로 데이터를 지속적으로 공급할 수도 있다. 백업과 중간 저장에는 SQL 데이터베이스가 자주 쓰이는데, 그중 큰 역할을 하는 것이 무료 오픈소스인 SQLite 와 PostgreSQL 이다. 보고서 조각에는 텍스트나 RTF 파일이, 그래픽에는 PNG, JPG, TIFF 파일이 사용된다. Microsoft 의 SQL Server 를 쓰는 이들도 많지만, 무료판인 "SQL Server Express" 는 제약이 매우 크고 (예컨대 CPU 코어를 하나만 쓸 수 있다) 다른 판본은 상업적 이용 시 상당한 라이선스 비용이 든다. 캘리포니아의 부동산 데이터베이스는 매우 클 수 있고 조인 (join) 에 시간이 오래 걸릴 수 있으므로, 대규모 조인에는 8 코어 이상에 128G RAM 을 갖춘 컴퓨터가 유용하다. SQL 데이터베이스는 임의로 경계 지어진 구역 안의 모든 부동산을 찾아야 할 때도 중요하다. 따라서 예컨대 군집분석 (Cluster Analysis) 으로 근린지역을 정의했다면 SQL 데이터베이스를 쓰게 될 것이다. 그러나 데이터베이스마다 지리정보를 다루는 내장 함수가 달라 서로 호환되지 않는다. RCA 에 관한 이 연재에서는 복잡한 SQL 작업에는 PostgreSQL 을, 단순한 작업에는 SQLite 를 전제한다.

가정 3: MARS 처리에는 "R/earth" 를 사용한다.

2017 년까지 Salford Systems 의 MARS 소프트웨어는 연 270 달러로 감당할 만한 해결책이었다. 그러나 Minitab 이 Salford Systems 를 인수한 뒤 MARS 는 연 16,000 달러의 SPM 패키지에 묶였고, 이는 대부분의 소규모 사업자에게는 감당하기 어려운 가격이다. 다행히 Stephen Milborrow 가 관리하는 R 프로그래밍 언어의 "earth" 패키지 (Milborrow 2024a) 가 Minitab SPM MARS 의 기능에 필적하거나 그것을 능가하는 수준으로 발전했다. earth 패키지에는 추가적인 이점도 있다. 무료로 이용할 수 있고, R 의 포괄적인 통계 생태계와 매끄럽게 통합되며, Pandas 를 통해 Python 에서도 접근할 수 있어 다양한 작업 흐름에 뛰어난 유연성을 제공한다. MARS 에 관한 논의는 R/earth 의 사용을 전제한다.

가정 4: 표기법

CSV 파일이나 Excel 스프레드시트와 같은 외부 원천에서 온 원자료는 표준 라틴 문자로 표기한다. 여기에는 MLS(Multiple Listing Service) 부동산 데이터, R/earth 패키지를 이용한 MARS(다변량 적응 회귀 스플라인) 설정값, 그리고 프로젝트별 데이터가 포함된다. 이 표기법은 R 이나 Python 에서 그에 대응하는 데이터 프레임 표현에도 적용된다.
부동산 특성과 그 여러 유형 및 부분집합을 논할 때는 정밀성과 분석적 명료성을 높이기 위해 그리스 문자로 표기를 바꾼다. 이러한 수학적 표기가 더 엄격해 보일 수 있으나, 여기에는 추상적인 부동산 특성을 그 원자료 표현과 구별한다는 중요한 목적이 있다. 관행상 그리스 문자는 일관되게 부동산 특성을, 라틴 문자는 스프레드시트 데이터 구조를 가리킨다. 예컨대 기호 M 은 행에 부동산을, 열에 그에 대응하는 특성을 배치한 MLS 데이터 행렬을 나타낸다.

가정 5: 프로그래밍 언어

이 분야에서는 R 과 Python(w/Pandas) 이 지배적인 언어이며, 속도가 중요할 때는 C 나 C++ 가 쓰인다. R/earth 자체도 R 과 C 로 작성되어 있다.

가정 6: 색인

색인 방식은 프로그래밍 언어와 수학에서 서로 달라 혼동을 일으킬 수 있다.

- 수학과 R 은 색인이 1 에서 시작한다.
- C, C#, $C++$, Python 은 0 에서 시작한다.

이 논문에서는 일관성과 단순성을 위해 모든 색인을 (Python 이나 C 처럼) 0 에서 시작한다. 구체적인 방식은 다음과 같다.

- 리스트, 행렬, 데이터 프레임의 경우:
  - 색인 0 은 "대상 부동산"(연구 대상이 되는 주된 부동산) 에 예약된다.
  - 예: 행이 부동산인 표에서 행 0 은 대상 부동산이고, 열 0 은 모든 부동산의 고유 ID 를 담는다.
  - 실무 요령: 대상 부동산을 맨 앞 (행 0) 에 두고, 나머지는 거래일이 최근인 것부터 오래된 것 순으로 나열하며 순차적으로 (0, 1, 2 등) 번호를 매긴다.
- 대상 부동산이 없는 경우:
  - 데이터 구조에 부동산이 포함되지 않는다면 (예컨대 변수나 함수만 있는 경우) 색인 0 은 건너뛰고 1 에서 시작한다.
  - R 사용자를 위한 참고: R 은 1 기반 색인을 쓰므로, 이 논문의 수식을 R 로 코딩할 때는 0 기반에서 1 기반으로 맞추기 위해 모든 색인에 1 을 더해야 한다.

5 RCA 작업 흐름

RCA 작업 흐름을 이해하는 것은 효과적인 분석에 필수적이다. 이 절차는 분석 과업과 의사결정 지점의 구조화된 순서를 따른다

5.1 MARS 분석의 유형

MARS 분석의 작업 흐름은 시장분석만 필요한지, 아니면 특정 부동산의 평가나 감정평가가 요구되는지에 따라 달라진다. 특정 부동산의 감정평가에는 통상 그 시장 지역에 대한 분석이 필요하다는 점에 유의하라.

- 시장분석 (Market Analysis): 카운티, 시, 또는 획정된 권역과 같은 넓은 지역에 초점을 맞춘다. 이 접근에는 특정 대상 부동산이 없으므로 가치 결론도 필요하지 않다. 작업 흐름은 모형 개발, 검증, 그리고 뒷받침 분석자료의 생성을 중심으로 한다. 최적 모형이 식별되고 문서화되면 절차는 종료된다.
- 단일 부동산 분석 (Single Property Analysis) 은 시장분석의 모든 요소에 더해 특정 부동산의 가치 판정을 포함한다. 이는 전통적 감정평가 방법과 근본적으로 다르다. 전통적 접근에서 가치는 통상 3 9 개의 비교 거래사례를 보정하고 가중평균을 계산하여 도출한다. 이와 달리 RCA 는 주로 다음 두 요소를 통해 가치를 판정한다.

  1. MARS 모형의 예측가치
  2. 추정된 부동산별 잔차 (그리고 그에 결부된 CQA 점수)

  RCA 에서 거래사례비교표 (Sales Grid) 는 주로 변수별 기여를 보여줌으로써 대상 부동산과 비교사례 사이의 가치 차이를 설명하는 역할을 한다. RCA 의 핵심 장점은 안정성이다. 의뢰인이 비교사례의 선정을 두고 이견을 제기할 수는 있으나, 제시를 위해 어떤 비교사례를 고르든 가치 결론은 일관되게 유지된다. 모든 거래사례가 (그것이 15 건이든 600 건이든) 동일한 모형으로 보정되기 때문이다. 오래된 거래도 현재의 시장 여건에 맞추어 자동으로 보정되지만, 일반적으로는 최근 거래가 바람직하다. 기간이 길어질수록 측정되지 않은 시장 특성의 비중이 커지므로, 모형이 가격 요인을 포착하는 효력은 시간적 거리가 멀어질수록 약해진다.

5.2 처리 작업 흐름

전체 처리 작업 흐름은 다음 단계들로 서술할 수 있다.

1. 업무 범위 (Scope of work).
2. 초기 프로젝트 폴더와 파일 설정.
3. MLS 데이터 다운로드. 필요한 부동산 레코드 수와 기간 범위 제한 설정에 관해서는 프로토콜 1.1–1.3 을 참조하라.
4. 설정 시트 완성: 회귀 변수, 상호작용, 집계, 계산.
5. MARS(R/earth) 회귀분석을 실행하여 가격 모형 생성.
6. MARS 모형으로부터 부동산 가치기여 및 잔차 (순 잔차와 잔차/SF) 생성.
7. 잔차/SF 로 부동산 순위 매기기.
8. 각 비교 거래사례의 CQA 점수 산출.

9. 부동산의 전반적 매력을 참되게 대표하는 품질 순위표가 산출되었다는 전제 아래, 순위표의 전반적 양상과 추세를 검토하여 대상 부동산의 최적 위치를 찾고 CQA 점수를 배정한다.
10. 그에 대응하는 잔차/SF 를 잔차/SF · (대상 부동산) GLA 의 형태로 대상 부동산에 배정한다.
11. **MARS 가격 추정치와 배정된 잔차를 더하여 대상 부동산의 가치 결론에 이른다**.
12. 대상 부동산의 잔차를 서술적 구성요소와 가치기여로 분해한다.
13. 비교사례에 대해 모형이 산출한 가치기여와 보정치, 그리고 모든 소계와 보정 후 거래가격을 계산한다.
14. 보정 후 거래가격은 대상 부동산의 최종 가치 결론과 같아진다.
15. 거래사례비교표에 넣을 가장 유사한 부동산 6 12 건을 선정한다.
16. 비교사례의 잔차 총액을 잔차 구성요소별 가치기여로 분해하고, 분해가 끝나면 잔차 총액 항목의 값을 0 으로 설정한다.
17. 모든 소계와 거래사례비교표의 비교사례 보정 후 거래가격을 재계산한다. 값을 확인한다.
18. 거래사례비교표를 보고서에 올린다.

RCA 에서 최종 가치 결론은 거래사례비교표의 비교사례 값들로부터 직접 도출되는 것이 **아니**라는 점에 유의하라. RCA 방법에서 거래사례비교표는 대상 부동산과 비슷해 보이는 부동산들이 왜 최종 가치 결론 또는 대상 부동산의 평가액보다 비싸게 또는 싸게 팔리는지를 설명하는 도구일 뿐이다. 이는 달리 보아야 한다. 대상 부동산의 가치 결론은 MARS 모형의 가격 추정치와 감정평가공학자가 추정한 잔차에 달려 있다. 그런데 이 두 양은 R/earth 회귀분석에 투입된 모든 부동산에 크게 의존한다. 그 회귀분석이 부동산 잔차의 순위표를 만들어 냈고, 그 순위표가 대상 부동산의 CQA 점수를 결정했으며, 감정평가공학자가 순위표 안에서 위치를 정함으로써 대상 부동산의 잔차가 확정되었기 때문이다. 거래사례비교표의 보정치는 실은 RCA 가 부과한 수학적 제약에 의해 결정된다. 잔차의 분해는 다소간 VE 의 재량에 맡겨져 있지만, 분해된 가치기여의 합계는 언제나 그에 결부된 잔차와 같아야 한다. 증명 2 로부터 우리는 잔차 구성요소의 실제 값이 VE 의 판단에 따라 달라질 수 있더라도 특정 부동산에 대한 그 합계는 완전히 제약되어 있으며, 따라서 VE 가 잔차 분해를 어떻게 바꾸든 해당 부동산의 보정 후 거래가격은 달라지지 않는다는 것을 안다.

5.3 데이터 작업 흐름

이 논문에서는 RCA 방법의 상위 수준 작업 흐름만을 제시한다는 점에 유의하라. RCA 분석의 여러 단계와 하위 단계는 다른 논문에서 다룬다.

5.3.1 MARS 를 위한 데이터 요건

거래 개시일과, 감정평가 또는 가치산정의 기준일에 근거한 종료일을 지정해야 한다. 이때 보통 필요한 회귀 변수의 추정 개수와 1 을 넘는 각 상호작용 차수별로 필요한 상호작용의 개수를 고려한다.

5.3.2 프로토콜: 종결 거래 리스팅

프로토콜 1.1: 필요한 MLS 종결 거래 리스팅 건수의 결정

이 프로토콜은 MARS 회귀분석에 투입할 종결 거래 리스팅의 권장 최소 건수를 계산하는 방법을 설명한다. 우선 대략적인 추정을 한 다음, 1 차, 2 차, 경우에 따라 3 차 차수로 MARS 를 실행하여 초기 모형을 생성하고, 아래 규칙에 따라 행 수를 검토하여 더 많은 행이 필요한지 판단해야 한다. 이 프로토콜은 독립변수가 100% 독립임을 전제하는데, 부동산에서는 그런 경우가 드물다. 예컨대 GLA 와 방 수는 대체로 상관관계가 높고, GLA 는 대지면적과도 상관을 보일 가능성이 크다. 여기서 "자유도 (Degrees of Freedom, DOF)" 의 계산 문제가 나온다. 참으로 독립적인 변수 하나가 자유도 하나를 갖는다. 그러나 이른바 독립변수들 사이에 종속관계가 있으면 정확한 DOF 계산에는 이 논문의 범위를 넘어서는 분석이 필요하다. 변수 간 공선성이 있을 가능성이 크므로 아래 추정치는 아마도 다소 높은 편이라는 점만 유념하라.

1. 1 차 항 (상호작용 없음) 에 대해서는 변수당 ~ 15 행 규칙이 잘 통한다
2. 허용되는 2 원 상호작용 하나마다 부동산 약 15 30 건을 추가해야 한다
3. 고차 상호작용 (변수 3 개 이상) 에는 각각 30 건 이상의 부동산이 필요하다.

다만 이것이 절대적인 규칙은 아니다. 실제로 필요한 건수는 다음에 좌우될 수 있다.

- 데이터의 신호 대 잡음 비율
- 모형화하려는 관계의 복잡성
- 반응변수의 변동성
- 데이터의 품질

예:

다음과 같은 경우라면:

- 주요 변수 10 개
- 허용되는 2 원 상호작용 5 개
- 허용되는 3 원 상호작용 2 개

최소 거래 건수는 다음과 같이 계산할 수 있다.

N = (변수 10 개 * 변수당 15 행) +
(2 원 상호작용 5 개 * 2 원 상호작용당 22 행) +
(3 원 상호작용 2 개 * 3 원 상호작용당 30 행)
= 320 행

또는 다음과 같이 할 수도 있다. 어떤 데이터 집합에 R/earth 를 실행한 뒤, 모형의 기저함수 개수에 1 을 더한 값을 자유도의 추정치로 삼고, 여기에 20 30 을 곱하여 필요한 입력 레코드 수를 구할 수 있다.

프로토콜 1.2: 거래 기간은 매수인 선호의 변화가 최소인 구간을 포괄해야 한다

과거 부동산 데이터를 수집할 때, 시장 선호의 변화를 적절한 변수로 포착할 수 있다면 엄격한 기간 제한은 없다. 해당 시장의 건축양식 변천을 다룰 때는 다음의 핵심 사항을 고려하라. 지배적 건축양식의 변화 (예컨대 California Ranch 에서 Mediterranean 으로의 이행)는 다음을 이용해 모형화할 수 있다.

- 택지개발지구 식별자
- MLS 지역 코드
- 건축연도 구간

그러나 이러한 양식 변화를 특정 변수로 충분히 포착할 수 없다면 중요한 제약에 부딪힌다. 거래일 (Date of Sale) 은 통상 시장 여건을 보정하는 데 쓰이는데, 이를 동시에 건축 선호를 통제하는 데까지 쓰면 이 두 개의 별개 효과가 뒤섞이게 된다. 그러한 경우 바람직한 접근은 다음 중 하나다.

1. 데이터 수집을 새로운 건축양식이 도입된 이후 기간으로 한정한다
2. 비교사례 선정을 동일한 건축양식을 공유하는 택지개발지구로 제한한다

이렇게 하면 시간에 따른 시장 여건의 적절한 보정은 그대로 허용하면서도 부동산 특성의 일관성이 유지된다. 아니면 적어도 이 경우에는, 데이터에서 건축양식을 추출할 수 있다는 전제 아래 이 두 변수 사이의 상호작용을 고려해 볼 수도 있다. 그러나 그러한 추출이 언제나 가능한 것은 아니다.

프로토콜 1.3: 거래가 적은 지역이라면 할 수 있는 최선을 다하라

거래 데이터로 MARS 모형화를 할 때 의미 있는 분석을 위해서는 대략 15 건의 거래가 필요하다. 거래가 그보다 적다면 (예컨대 부동산 5 건) 각 레코드를 두 번씩 복제하여 15 건 기준선에 이르게 할 수 있다. 이 접근에서 중요하게 고려할 사항은 다음과 같다.

1. 과적합을 줄이기 위해 1 차 회귀만 사용한다
2. 데이터 집합이 지나치게 제한적이므로 교차검증은 무시하거나 건너뛴다
3. 이러한 상황에서도 MARS 가 짝지은 사례 비교나 표준 선형 회귀분석보다 나은 성능을 보일 가능성이 크지만, 데이터 복제는 불확실성 측정치와 통계적 유의성의 타당성에 영향을 준다는 점에 유의하라

장기적으로 가장 좋은 해법은 실제 거래를 추가하여 데이터 집합을 늘리는 것이다. 그러나 제한된 데이터로 진행할 수밖에 없다면, 통계적 한계를 인정하는 전제 아래 데이터 복제가 쓸 만한 임시 해법이 된다.

5.4 프로젝트 폴더 구조

프로젝트 폴더 구조를 논해 두는 것이 좋겠다. 시장분석과 단일 부동산 감정평가는 별도의 폴더를 두는 것이 좋다. 두 폴더는 유사한 하위 폴더 구조를 갖는다.

```
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
                        └── ...
```

1 단계 처리는 버전 폴더와 그 하위 폴더들을 생성한다. 이때 MLS_Original.xlsx 를 복사하는데, 이 스프레드시트는 특정 일시에 MLS 에서 내려받은 내용을 그대로 반영하며 결코 변경되어서는 안 되는 것이지만, 기간이 긴 프로젝트에서는 새로운 MLS 다운로드가 필요해질 가능성이 있기 때문이다. 다음으로 이 단계는 설정 파라미터가 포함된 확장 MLS.xlsx 워크북의 사본을 버전 폴더에 만든다. MARS 를 완전히 한 번 실행할 때마다 하나씩, 버전 폴더는 여러 개가 있을 수 있다.

이어지는 단계들은 워크북에 열을 추가하고 정렬하는 등의 변경을 가하며, 감정평가가 완료되면 갱신된 판본을 MLS_Stage(N) 으로 프로젝트 데이터 폴더에 기록한다. 다시 말하면, 루트의 MLS.xlsx 폴더는 끊임없이 변화하며, 버전 사본은 변수의 삭제와 추가를 비롯한 여러 변경을 거치는, 원본 MLS 워크북의 진행 중인 변형이다. 처리용 R 또는 Python 코드는 대개 Data 의

MLS.xlsx 워크북을 입력으로 실행된다. 따라서 RCA 실행의 각 버전 뒤에 있는 모든 데이터의 사본이 언제나 보존된다.

더는 개선이 불가능해질 때까지 필요한 만큼 여러 버전을 만들고, 그 시점에서 최선의 버전을 최종본으로 선정한다. 디렉터리를 훑어볼 때 바로 알아볼 수 있도록 폴더 이름 끝에 "_Final" 을 붙일 수도 있다. 흥미롭거나 경합하는 모형을 담은 옛 버전 일부는 대체로 보존해 두는 것이 좋다. 각 버전 폴더에는 보고서에 삽입할 수 있는 여러 텍스트, 도표, 그래프 파일이 들어 있다. 최종 MLS_Stage5.xlsx 워크북에는 보고서에 필요한 만큼의 시트가 들어 있다. 특히 Fannie Mae 거래사례비교표로 옮길 스프레드시트가 있을 수 있다. 또한 집계의 계산 과정이나 그 분해도 반드시 포함된다. 어려운 시장 지역에서 완벽한 모형을 찾아내려는 이 씨름은 RCA 접근법에서 가장 시간이 많이 드는 부분일 수 있다.

거래사례비교표를 재현하는 데 필요한 모든 것을 버전 폴더에 복사해 두어야 하며, 코드가 자주 바뀌는 경우에는 사용된 R 또는 Python 코드도 포함된다.

5.4.1 프로토콜: 다중 MARS(earth) 실행

프로토콜 2.1: 시장 지역의 가격 요인을 정밀하게 파악하기 위한 과적합 실행

초기 earth 실행 단계에서는 밑에 깔린 가격 요인을 발견하기 위해 과적합을 전략적으로 활용해야 한다. 예컨대 거래 800 건으로 대표되는 시장 지역에 작업장 (workshop) 이 딸린 부동산이 단 3 건뿐일 수 있는데, 그 소수 사례에서는 작업장이 최종 거래가격에 큰 영향을 줄 수 있다. 전진 단계만 실행하면 이러한 작업장의 대략적인 가치기여를 보여주는 과적합 모형이 산출된다. 그러나 800 건 중 그러한 거래가 3 건뿐이므로 작업장 항은 후진 단계에서, 그리고 교차검증에서는 확실히 제거될 것이다. 전진 단계만의 실행이 작업장에 부여한 값은 기록해 두었다가 잠재변수 분석으로 옮겨, 거래사례비교표에 등장하는 작업장 보유 부동산에 사용할 수 있다. 이러한 과적합 모형은 후진 단계의 가지치기나 교차검증 실행에서 제거될 다른 여러 희소 데이터 특성에 대해서도 상세한 추정치를 제공한다.
이러한 탐색적 실행은 각각 별도의 버전 폴더로 저장하라. 기록된 가치기여는 나중에 잔차를 분해할 때 VE 가 인용하고자 할 증거이다. 물론 VE 는 회귀분석이 한 번도 보지 못한 부동산 — 이후에 등장하는 대상 부동산 같은 — 에 대해서도 모형이 견고하도록 궁극적으로는 가지치기와 교차검증에 의존해야 한다는 점을 이해해야 한다. 과적합 실행은 연구를 위한 도구일 뿐 결코 최종 모형이 아니다.

프로토콜 2.2: 서로 다른 earth() 파라미터 아래에서의 다중 교차검증 실행

시장 지역은 구조와 규모가 서로 다르므로, VE 는 대개 일련의 실행을 통해 실험해야 한다. 어떤 변수가 가격에 가장 큰 영향을 주는지 이해하기 위해서, 기존의 측정 변수 집합이 가치를 추정하는 데 충분한지 아니면 새로운 데이터 (수변까지의 거리, 표고, 상업시설까지의 거리 등) 를 추가해야 하는지 판단하기 위해서, 그리고 비효율적이거나 오도하는 상호작용은 배제하면서 새로운 종류의 상호작용을 발견하기 위해서다. 최종적으로 VE 가 원하는 것은 훈련 데이터에 대한 양호한 R2 와 검정 데이터에 대한 최대한 좋은 CVR2 이다. 그에 이르려면 새로운 데이터를 수집하거나, 기존 데이터의 정확성을 높이거나, 시장 지역의 규모와 성격에 맞게 (예컨대 가용 자유도에 맞추어 — 프로토콜 1.1 참조) earth() 파라미터를 조정해야 할 수 있다.
위의 프로젝트 구조가 다수의 실행 버전 폴더를 지원하는 이유가 바로 이것이다. 각 후보 모형이 그 파라미터 및 R2/CVR2 와 함께 보존되므로 여러 실행에 걸친 궤적을 비교할 수 있다. R2 는 오르는데 CVR2 는 정체하거나 떨어지는 버전은 시장 신호가 아니라 잡음에 적합한 것이며, 그 사실을 드러내 주는 것이 바로 버전 이력이다.

프로토콜 2.3: Remarks 및 서술적 특성 열의 Prolog 확장

아직 다소 실험적인 단계에 있기는 하지만, earthUI 에는 이제 내장 Prolog 엔진 (vProlog) 이 탑재되어 있으며 확정절 문법 (Definite Clause Grammars, DCG) 을 이용해 MLS 의 자유 서술 열 — 무엇보다 Public Remarks, 그 밖에 중개인 remarks 와 방 특성이나 기존 구조물 같은 쉼표 구분 특성 목록 — 을 파싱한다. 이 확장은 “owned solar”, “3 car garage”, “40x50 workshop” 과 같은 표현을 회귀분석에 투입할 수 있는 통상의 특성 열로 바꾸어 주며, 이렇게 복원된 희소 데이터 특성은 잠재변수 분석에 곧바로 반영된다 (뒤에 나오는 근사 회귀변수로서의 상태 및 품질 등급에 관한 소절을 보라. 이 기능 자체는 이번 호의 earthUI 관련 논문에 서술되어 있다).

텍스트 확장은 후보 예측변수 집합을 크게 늘리며, 그에 따라 검증을 위한 다수의 earth/RCA 실행이 필요해진다. 확장된 열은 모형에 넣고 빼기를 시험해 보아야 하고, 공선적인 경우에는 파생 계수나 불리언으로 통합해야 하며, 잡음이 아니라 시장 신호를 담고 있는지 CVR2 로 확인해야 한다. 그러한 실험 각각을 프로토콜 2.1 및 2.2 에서와 똑같이 별도의 실행 버전 폴더로 다루어라. 또한 파싱 문법 자체가 개발 중이므로, VE 는 추출된 열을 모형에서 신뢰하기에 앞서 원문과 대조하여 검토해야 한다.

5.5 핵심 Excel 워크북

대부분의 MLS 서비스가 부동산 데이터를 Excel CSV 워크시트로 내려받게 하므로, 데이터와 각종 설정 파라미터를 저장하는 데는 Excel 이 아마도 가장 편리한 수단일 것이다. 프로젝트마다 MLS.xlsx 라는 핵심 Excel 워크북이 하나씩 있으며, 여기에는 MLS 에서 내려받은 종결 부동산 거래 리스팅을 저장하는 “MLSData” 시트 하나와 여러 개의 설정 및 파라미터 시트가 들어 있다.

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

원자료는 MLS_Original.xlsx 같은 이름의 워크북에 저장한다. 여기서 다른 이름, 이를테면 “MLS.xlsx” 로 사본을 만들어 R 또는 Python RCA 프로그램에 적재한다. 이 두 워크북은 모두 Data 폴더 아래에 둔다. Data/MLS.xlsx 워크북은 변수의 여러 조합, MARS 파라미터, 그 밖의 설정 변경을 시험하며 아마도 여러 차례의 실행을 거치게 된다. 각 회차는 VER_YYYYMMDDHHMMSS 형식으로 날짜와 시각을 이름에 담은 새 입출력 폴더를 만든다. 각 폴더에는 입력 MLS.xlsx 워크북의 사본과 네 개의 처리 단계 (2–5 단계) 각각에 대해 생성된 Excel 워크북, 그리고 출력 텍스트 문서, 도표, 그래프가 들어간다. 최종 단계인 5 단계에는 의뢰인의 유형과 SOW(업무 범위) 에 따라 보고서용 시트가 여럿 더 포함될 수 있다.

5.6 백업

새로 작성했거나 수정한 Excel 스프레드시트를 프로젝트 버전 폴더에 기록할 때, 프로젝트 폴더 안의 SQLite 나 PostgreSQL 데이터베이스에도 백업을 만들 수 있다. 데이터베이스 파일은 이후 검토나 참조 과정에서 의도치 않게 손상될 가능성이 더 낮다.

6 워크북 표기법

6.1 서술적 차원

워크북 “M” 에는 대략 여섯 개의 서술적 차원이 있다.

1. 감정평가 프로젝트 “a”
2. 실행 버전 “r”
3. 작업 흐름 단계 “w”
4. 시트 “s”
5. 부동산 리스팅 “p”
6. 변수 집합 “v”

따라서 특정 워크북 M 안의 개별 셀은 다음과 같이 서술할 수 있다.

$$\begin{matrix}\mathrm{r}_* \\ \mathrm{w}_*\end{matrix} \overset{\mathrm{a}_*}{\underset{\mathrm{s}_*}{W}} \begin{matrix}\mathrm{p}_* \\ \mathrm{v}_*\end{matrix} \tag{1}$$

(참고: “*” 는 유효한 임의의 색인을 뜻한다)

또는:

$$\begin{matrix}\mathrm{r}_k \\ \mathrm{w}_m\end{matrix} \overset{\mathrm{a}_g}{\underset{\mathrm{s}_h}{W}} \begin{matrix}\mathrm{p}_i \\ \mathrm{v}_j\end{matrix} \qquad \begin{matrix} i = 1, \ldots, n_p \\ j = 1, \ldots, n_v \\ k = 1, \ldots, n_r \\ m = 1, \ldots, n_m \\ g = 1, \ldots, n_g \\ h = 1, \ldots, n_h \end{matrix} \tag{2}$$

여기서:

n_p = 행의 수 (시트에 따라 다름)
n_v = 열의 수 (시트에 따라 다름)
n_r = 실행 버전의 수
n_m = 작업 흐름 단계의 수
n_g = 분석 프로젝트의 수
n_h = 워크북 시트의 수

“W” 는 각자 자기 표를 갖는 여섯 개의 핵심 데이터 시트로 이루어진 Excel 워크북 전체를 가리키며, 연관된 R 또는 Python 프로그램을 실행하면 이들이 각각 별개의 내부 데이터 프레임으로 적재된다는 점을 이해하라. R 과 Python 모두 이름을 받아 그것이 데이터 프레임의 어느 행 또는 열에 해당하는지 찾아 주는 내부 함수를 갖고 있다. 예컨대 아래 첨자 v_j 에서 v = (“GLA”, “LotSize”, “BathRms”, “BedRms”, …) 이고 j=0 이면 v_j = “GLA” 이다. R 이든 Python 이든 내부적으로 “GLA” 를 그것이 저장된 열 — 예컨대 해당 데이터 프레임의 15 번째 열 — 로 해석해 낼 수 있다. 따라서 이 논문에서는 이름 배열에 색인을 걸고, 그 이름이 실행 중인 프로그램에 전달되면 프로그램이 그 이름을 내부 데이터 프레임의 실제 열 또는 행 색인으로 다시 변환한다. 다소 복잡하기는 하다. 그러나 정밀을 기하자면 이것이 필요하고 또 가장 편리하다. 이는 분석자 역할과 개발자 역할 사이의 “책임 분리” 인데, 실제로는 감정평가공학자 (VE) 가 두 역할을 모두 수행할 수도 있다. R 의 “data frame” 은 Python 에서 “dataframe” 이라 불린다는 점에 유의하라.

또한 MARS 실행 파라미터 시트는 W 의 차원이 아니라는 점에 유의하라. W 는 MARS 에 투입되는 데이터의 버전에 관한 것이고, MARS/earth 가 그 데이터를 어떻게 다룰지를 정하는 파라미터는 그와 별개이기 때문이다.

명료성을 위해, 데이터 워크북 시트가 적재되는 R 또는 Python 내부 데이터 프레임과 보다 직접적으로 대응하는 복합적 다차원 표기법을 워크북에 대해 제시한다. “무거운” 표기법이다. 그러나

증명 등에 들어가면 우리가 다루는 대상에 관한 가정을 둠으로써 이 짐의 상당 부분을 덜어낸다. 그럼에도 이 표기법은 처음부터 작업 흐름과 데이터의 전체 그림을 파악하는 데 도움이 된다.

우리가 처리할 워크북 집합에 대한 표기법에서 시작하자.

6.1.1 프로젝트 차원 (“a”)

업무 범위 (SOW) 가 승인되고 나면, 설정 파라미터, 출력 그래프, 도표, 스프레드시트, 보고서 조각을 포함한 입출력 데이터를 저장할 프로젝트 폴더를 마련해야 한다. VE 는 보고서가 분실되었을 때 그것이 곧바로 드러나도록 모든 프로젝트를 정수와 같은 단순한 순차 ID 로 식별해야 한다. 프로젝트가 취소되는 경우에는 해당 프로젝트 폴더 안에서 취소되었음을 일관되게 표시할 수단이 있어야 하되, 폴더를 삭제해서는 결코 안 된다. 다른 방식도 가능하지만, 여기서는 프로젝트 ID 가 0 에서 시작하여 1 씩 증가한다고 가정한다. 또한 유효한 모든 프로젝트 ID 가 처리 프로그램으로 생성된 순서집합 “a” 에 저장되어 있다고 가정한다.

감정평가 프로젝트 데이터는 워크북의 프로젝트 워크시트에 있으나, 중앙 SQL 데이터베이스에 저장할 수도 있다.

프로젝트 필드로는 다음을 제안한다.

1. Local ID: 회사 단위로 순차 생성되는 짧은 정수 ID. 저장용 기본키로 삼기에 좋으며, 프로젝트가 하나씩 추가될 때마다 데이터베이스가 자동으로 생성할 수 있다.
2. Global GUID: 원한다면 다음을 이용해 GUID(전역 고유 ID) 를 생성할 수 있다.

   https://www.uuidgenerator.net/

3. 부동산 소재지 (시장분석) 또는 주소
4. 감정평가 기준일
5. 의뢰일
6. 임장활동일
7. 최초 보고서 작성일
8. 1 차 수정 보고서 작성일
9. 2 차 수정 보고서 작성일
10. 3 차 수정 보고서 작성일
11. 의뢰인명
12. 의뢰인 주소
13. 대출기관명
14. 대출기관 주소
15. 소유자명
16. 소유자 주소
17. ....

프로젝트 ID 목록의 예:

```
a =     (“1”,
         “2”,
         “...,”
         “141”,
         “142”)
```

6.1.2 실행 버전 차원 (“r”)

작업 흐름의 두 번째로 입력 데이터에 MARS 회귀분석을 수행한다. 모형을 개선하기 위해 여러 파라미터를 바꾸어 가며 몇 차례에서 여러 차례에 이르는 실행을 하게 될 것이다. 각 실행은 실행 일시에 기반한 이름으로 출력용 새 버전 폴더를 생성한다. 예를 들면,

r = (“ver20241210_083124”,
“ver20241209_151130”,
“ver20241208_160944”)

latest_version : r_1 = “ver20241210_083124”

경쟁력 있는 모형과 출력을 가진 유망한 버전은 최종본이 결정될 때까지, 또는 쓸모가 있다면 그보다 더 오래 보존해야 한다. 최종 버전임을 표시하기 위해 폴더 이름 끝에 “Final” 을 붙여 수정할 수도 있다.

6.1.3 작업 흐름 단계 (“w”)

RCA 작업 흐름은 아래에 열거한 것처럼 통상 다섯 개의 개별 단계로 나뉜다. 각 단계는 다음 단계를 시작하기 전에 성공적으로 완료되어야 한다. VE 는 한 단계가 끝난 지점에서 처리를 멈추고 결과를 검토한 뒤 진행할 수 있어야 한다. 시장 지역은 상당히 복잡하고 종종 매우 특이하기 때문에 이것이 필요하다. 새로운 접근을 요하는 이례적 문제에 부딪힐 수 있다. VE 는 R 이나 Python 코드에 뛰어들어 필요한 수정을 가하고 나아가 새로운 접근을 시도할 수 있어야 한다. 그러한 경우 작업 중인 동일 단계를 여러 번 실행하면서 다양한 접근을 실험해 보고 싶을 수 있다. 다만 VE 는 — 몇 가지 단서 (예컨대 3 차 집계는 2 단계의 이전 실행에서 이미 결정되어 있어야 한다) 와 함께 — 명령 하나로 모든 단계를 실행할 자유도 가져야 한다. 이는 각 단계가 완료될 때 그 모든 내부 데이터를 파일이나 데이터베이스에 저장하여, 다음 단계가 이전 단계의 데이터를 메모리로 다시 적재함으로써 마치 이전 단계가 여전히 메모리에 있는 것처럼 시작될 수 있어야 함을 뜻한다.

단계는 다음과 같다.

w = (“stage1”, -설정
“stage2”, -MARS 처리
“stage3”, -최종 거래사례비교표
“stage4”, -보고서 조각
“stage5”) -검토/정리

stage2 : w_2 = “$stage2''$

6.1.4 Excel 워크북 시트 (“s”)

RCA 핵심 Excel 워크북에는 각각 내부 데이터 프레임으로 대응되는 다음 시트들이 있다.

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

2 차 집계를 별도의 시트에 두는 이유는, 가능한 모든 상호작용 쌍을 보여주는 대칭 표로 정의하는 것이 가장 손쉬워서 각 쌍에 결부된 가치기여와 보정치를 어느 집계 변수로 모을 (더할) 것인지 지정할 수 있기 때문이다. 3 차 집계는 이렇게 할 수 없는데, 3 차원 대칭 표가 필요하지만 Excel 에서는 불가능하기 때문이다. 게다가 20x20x20 표라면 다루어야 할 셀이 8,000 개가 된다. 3 차 항의 이름은 그냥 1 차 집계 표에 함께 나열하는 편이 더 쉽다. 예를 들면 다음과 같다.

변수	**집계 대상**
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 부동산 리스팅 (“p”) — MLSData 행

부동산 리스팅은 MLSData 시트의 행을 이룬다. 다른 시트에는 나타나지 않다가, 최종 비교사례 선정으로 보고서 조각을 만드는 4 단계에 이르러 등장한다. 부동산 ID 목록의 예는 다음과 같다.

$$\mathrm{p} = \quad (\text{“MLS823211”, “MLS823456”, “MLS823542”, “MLS823721”, “...”})$$

6.1.6 MLS 변수 (“v”) — MLSData 열

변수의 처리에 관해서는 뒤에서 더 다룬다. 여기서는 MLSData 시트에 통상 다양한 유형의 변수 30 40 개가 있으나 MARS 회귀분석의 독립변수로는 대체로 10 20 개만 선정된다는 점만 밝혀 두면 충분하다. 이 변수들은 다른 시트에도 나타나는데, 통상 MARS 에 제출할 수 있는 모든 가능한 변수의 목록 형태이다. 변수 목록의 예는 다음과 같다.

$$\mathrm{v} = \quad (\text{“GLA”,“SalePrice”,“BedRms”,“BathRms”,“GarageSF”,“...”})$$

6.1.7 최종 표기 예시

위 표기법을 하나의 최종 예시로 모아 보자. 위 목록들을 사용하여 다음과 같이 두자.

$$\overset{\mathrm{r}_1}{\underset{\mathrm{w}_1}{}}\overset{\mathrm{a}_{141}}{\underset{\mathrm{s}_0}{W}}\overset{\mathrm{p}_2}{\underset{\mathrm{v}_0}{}} \tag{3}$$

여기서는 0 기반 색인을 쓰고 있으므로 목록의 첫 원소는 색인 0 을 갖는 것으로 본다는 점에 유의하라. Python 도 0 기반 색인을 쓰지만 R 은 1 기반 색인을 쓴다.

따라서 위 표기는 다음에 해당하는 셀 값을 가리킨다.

- 변수 v_0, 즉 열 “GLA”

- 부동산 p_2, 즉 ID “MLS823542”
- 시트 s_0, 즉 “MLSData”
- 작업 흐름 단계 w_1, 즉 “stage2”
- 버전 r_1, 즉 “ver20241209_151130”
- 프로젝트 a_{141}, 즉 “142”

위 표기는 다음과 같이 쓸 수도 있다.

$$\begin{matrix}\text{“ver20241209\_151130”}\\ \text{“stage2”}\end{matrix}\ \overset{\text{“142”}}{\underset{\text{“MLSData”}}{W}}\ \begin{matrix}\text{“MLS823542”}\\ \text{“GLA”}\end{matrix} \tag{4}$$

6.1.8 정당화

“표기법에 그냥 색인이나 숫자를 쓰면 안 되는가?” 우리의 표기법에는 두 가지 요건이 있다.

1. 부분집합. 예컨대 분석자는 변수의 부분집합처럼 목록의 부분집합을 자주 선택한다. MLSData 시트에는 변수가 35 40 개 있을 수 있지만 특정 실행에는 그중 10 20 개만 사용된다. VE 는 실행할 때마다 MARS/earth 에 제출하는 변수를 자주 바꾼다.
2. 반복. 예컨대 $i = 1, 2, \ldots, n_1$
   우리의 알고리즘, 프로토콜, 증명은 대개 순서 목록의 모든 항목을 처음부터 끝까지 순회하거나 특정 항목을 골라낼 수 있어야 한다.

따라서 항목의 부분집합이나 목록을 나타내는 문자가 필요하고, 그 문자에는 목록의 항목들을 순회하여 데이터 프레임의 행이나 열에 색인을 거는 데 쓰이는 문자열을 얻어낼 또 하나의 아래 첨자가 있어야 한다.

6.1.9 행렬은?

선형대수의 행렬은 어떤가? 행렬은 정수든 실수든 숫자만 다룬다. 자료구조로서의 행렬은 R 과 Python 모두에서 지원된다. 그러나 MLS 데이터에서 우리가 다루는 변수 중 상당수는 “문자열” 또는 문자 데이터라 불린다. 어떤 변수가 숫자로만 되어 있더라도 그것은 실수가 아니라 그저 이름인 경우가 많다. 예컨대 어느 도시의 MLS 지역 코드는 (“660”, “661”, “662”, “663”, ..., “672”) 일 수 있다. earth 회귀분석은 변수를 “factor” 변수로 지정할 수 있게 하며, 이 경우 숫자는 순서 없이 단순한 이름으로 취급된다. 그러한 경우 산출되는 모형은 순서를 따라 흐르는 함수가 아니라 각 숫자에 대한 가치기여만을 제공한다. 다른 예로는 상태에 대한 “C1”,...,“C6” 과 품질에 대한 “Q1”,...,“Q6” 이 있다.

6.1.10 첫 번째 단순화

핵심 Excel 프로젝트 워크북에 대한 위 표기법은 여러 요소를 한데 묶기에는 좋지만, 알고리즘이나 수학적 증명을 설명하며 늘 지고 다니기에는 너무 무겁다. 지루한 작업을 할 때는 주어진 프로젝트, 버전, 단계, 시트 안에 있다고 가정할 수 있다. 이미 함축되어 있지 않다면 무엇을 다루고 있는지만 밝히면 된다. 무엇을 다루는지 일단 지시하고 나면 위에서 실제로 필요한 것은 다음뿐이다.

$$M_{\mathrm{s}_h}^{\ \mathrm{p}_i\ \mathrm{v}_j} \qquad \begin{matrix} i = 1, 2, \ldots, n_p \\ j = 1, 2, \ldots, n_v \\ h = 1, 2, \ldots, n_s \end{matrix} \tag{5}$$

여기서: n_p = 부동산 수
n_v = 입력 변수 수
n_s = 입력 시트 수

6.1.11 두 번째 단순화

색인을 쓰기보다 시트를 그것이 뒷받침하는 데이터 프레임의 종류에 따라 이름 붙이는 편이 더 편리하다. 시트마다 행과 열의 기능이 다르다. 그리하여 W 라는 데이터 프레임 집합을 다음과 같이 M, V, A, I, C 데이터 프레임으로 나눈다.

◇ **MLS 데이터**

$$M_{v_j}^{\ p_i} = W_{0\ v_j}^{\ p_i} \qquad \begin{matrix} i = 0, 1, \ldots, n_p \\ j = 0, 1, \ldots, n_v \end{matrix} \tag{6}$$

여기서: n_p = 부동산 수
n_v = 입력 변수 수

참고: 대상 부동산은 기본적으로 언제나 부동산 목록의 첫머리에 온다

"MLSData" Excel 스프레드시트와 그에 대응하는 데이터 프레임에서 첫 행은 언제나 대상 부동산을 위해 예약된다. 그 아래의 다른 MLS 부동산들과 함께 정렬되는 일이 결코 없다. MARS 회귀분석에 독립 데이터로 투입되지 않으며, 목표 데이터 즉 종속변수로 투입된다. 부동산 목록에서도 마찬가지다. 모든 부동산 목록에서 색인 0 의 첫 원소는 달리 밝히지 않는 한 언제나 대상 부동산 ID 로 본다.

대상 부동산 없이 시장 지역의 모형을 만드는 과업이더라도 첫 행에는 어떤 가상의 부동산 데이터를 적재한다. 이렇게 하면 프로그램을 작성하거나 알고리즘을 규정하거나 증명을 하기가 훨씬 편해진다. 따라서 c 로 부동산 목록을 나타낸다면, c_0 는 대상 부동산이고 c_i, $i = 1, \ldots, n_p$ 는 MARS/earth 회귀분석에 입력될 비교사례 리스팅이다.

◇ **회귀 변수 지정**

$$V_{s_j}^{v_i} = W_{1\ s_j}^{\ v_i} \qquad \begin{matrix} i = 1, 2, \ldots, n_v \\ j = 1, 2, \ldots, n_s \end{matrix} \tag{7}$$

여기서: n_v =사용 가능한 변수 수
n_s=지정의 수

통상 회귀분석 입력으로 사용 가능한 변수의 부분집합을 선택하므로 $n_s \leq n_v$ 이다.

◇ **계산**

$$C_{c_j}^{v_i} = W_{2\ c_j}^{\ v_i} \qquad \begin{matrix} i = 1, 2, \ldots, n_v \\ j = 1, 2, \ldots, n_c \end{matrix} \tag{8}$$

여기서: n_v = 변수 수
n_c = 계산 지정 수

◇ **상호작용**

$$I_{t_j}^{f_i} = W_{3\ t_j}^{\ f_i} \qquad \begin{matrix} i = 1, 2, \ldots, n_f \\ j = 1, 2, \ldots, n_t \end{matrix} \tag{9}$$

여기서: n_f = from 변수 수
n_t = to 변수 수

데이터 프레임 I 는 대칭 데이터 프레임으로, “from” 행과 “to” 열에 변수가 알파벳순으로 나열된다. 셀에 “1” 을 표시하면 해당 변수 쌍 사이의 상호작용이 허용됨을 뜻한다. 그렇지 않고 “0” 이면 상호작용이 허용되지 않음을 뜻한다. 2 차 상호작용에서는 가능한 상호작용이 하나뿐이므로 비교적 단순하다. 3 차 회귀에서는 세 변수 사이에 세 개의 쌍 상호작용이 있으며, 세 변수 상호작용 항 (또는 변수) 을 생성하려면 그 모든 쌍의 상호작용이 허용되어야 한다.

◇ **집계**

$$A_{v_j}^{v_i} = \underset{4}{W}{}_{v_j}^{v_i} \qquad \begin{matrix} i = 1, 2, \ldots, n_v \\ j = 1, 2, \ldots, n_v \end{matrix} \tag{10}$$

여기서: n_v = 변수 수

7 RCA 거래사례비교표

RCA 프로토콜에는 R/earth 에 접근하여 모든 세부 단계를 수행할 수 있는 R 또는 Python 프로그램이 필요하다는 점을 유념하라. 데이터 집약적이어서 그 작업을 수작업으로 수행하기란 도저히 불가능하다. 동시에 프로그램에 버그가 없다는 전제 아래, 필요한 작업이 사람의 실수 없이 정확하게 수행되도록 보장하려면 소프트웨어 프로그램이 있어야 한다.

또한 이 일을 하는 VE 는 특히 예상하지 못한 새로운 시장 지역, 근린지역, 부동산이나 그 밖의 새로운 요구사항, 문제에 부딪혔을 때 그러한 프로그램을 필요에 따라 수정할 수 있어야 한다.

RCA 처리에는 세 가지 유형의 거래사례비교표가 있다.

1. RCA 처리용 거래사례비교표. 이 논문에서 “MLSData” 로 표기하는 것으로, 매우 큰 데이터프레임과 그에 대응하는 스프레드시트이며, 집계 과정을 거쳐 아래에서 말하는 RCA 거래사례비교표에 이르게 될 여러 중간 세부 항목을 담고 있을 수 있다. 예시를 여기에 싣기에는 너무 크다. 감정평가공학자들은 대개 이 스프레드시트를 자신의 취향에 맞게 다듬을 것이다. 그러한 표의 개발에 관해서는 앞으로 별도의 논문을 쓸 생각이다.

2. RCA 최종 거래사례비교표. 비 GSE 보고서에 쓰이며 표 2과 비슷한 형태다. 이는 독자에게 어느 정도 익숙함을 주기 위해 예컨대 Fannie Mae 양식을 어느 정도 본뜬 것이다. 다만 비교사례 하나당 열의 수가 Fannie Mae 표보다 많다는 점에 유의하라. 각 비교사례에는 “Measure”(측정값) 와 “Adjustment”(보정) 열 옆에 “Value Contribution”(가치기여) 열이 있고 대상 부동산에도 하나가 있는데, 이는 Fannie Mae 양식에 없는 열이다. 더 단순한 형태는 표 1에 있다.

3. GSE 업로드용 거래사례비교표. Alamode 같은 소프트웨어에 거래사례비교표를 올릴 때 쓴다. 다만 모든 감정평가 양식 소프트웨어가 스프레드시트에서 거래사례비교표 데이터를 올리는 기능을 제공하는 것은 아니다. 부득이하게 그러한 표에는 가치기여를 위한 별도의 열을 둘 수 없는데, Fannie Mae 같은 GSE 는 가치기여라는 것을 전혀 이해하지 못하기 때문이다. 전략은 그들이 요구하는 양식을 주되, RCA 거래사례비교표와 변수 집계의 세부와 같은 그 밖의 뒷받침 문서로 이를 보강하는 것이다.

RCA 처리용 거래사례비교표는 Fannie Mae 거래사례비교표처럼 생기지 않았고, 대략 그것의 전치 (transpose) 에 해당하여 행과 열이 뒤바뀌어 있다. 각 행은 어느 부동산의 데이터를, 각 열은 어느 부동산 변수의 데이터를 담는다. 부동산 행은 수백 내지 수천 개, 열은 200 개 이상이 되기 쉽다. 첫 행은 변수명에, 둘째 행은 대상 부동산에 예약되는데, 시장분석만 하는 경우에도 그러하다. 스프레드시트를 데이터 프레임으로 읽어들이면 시트의 첫 행은 names 속성이 되고, 대상 부동산이 있는 둘째 행이 데이터 프레임의 첫 행이 된다.

RCA 거래사례비교표의 열에는 식별용 열이 있고, 그다음에 1 차 변수, 외부 변수, 합성 변수가 결합되어 독립변수로 MARS 에 투입된다. MARS 는 이 변수들 가운데 거래가격 (종속변수 또는 목표변수라고도 한다) 에 유의미한 영향을 준다고 판단되는 것들을 출력한다. 2 차 회귀에서는 새로운 변수 역할을 하는 변수 쌍이 나올 수 있다. 3 차 회귀에서는 이른바 독립변수 역할을 하는 변수 삼중항도 얻을 수 있는데, 물론 이는 밑에 깔린 변수들과 높은 상관을 갖는다.

표 1: RCA 스프레드시트 예시: 세로 배치

대상 부동산			비교사례 1			비교사례 2		
	측정값	가치기여	측정값	가치기여	보정	측정값	가치기여	보정
거래가격			$1,150,000			$1,500,000		
기초가치		$500,000		$500,000	$0.00		$500,000	$0
측정된 특성			측정된 특성			측정된 특성		
주거면적	1,530	$463,500	1,350	$382,500	$81,000	1,700	$540,000	-$76,500
대지면적	6,000	$200,000	6,500	$220,000	-$20,000	7,000	$240,000	-$40,000
연식	53	$63,600	57	$68,400	-$4,800	51	$61,200	$2,400
욕실 수	2	$15,000	3	$30,000	-$15,000	3	$30,000	-$15,000
침실 수	3	$6,000	4	$9,000	-$3,000	4	$9,000	-$3,000
측정 합계		$1,248,100.00		$1,209,900.00	$38,200		$1,380,200.00	-$132,100
잔차 특성			잔차 특성			잔차 특성		
잔차	CQA 5.8	$30,000	CQA 2.5	-$59,900	$89,900	CQA 8.2	$119,800	-$89,800
잔차 분해			잔차 분해			잔차 분해		
디자인		7,000		-$15,000	$22,000		$35,000	-$28,000
상태		5,000		-$8,000	$13,000		$20,000	-$15,000
품질		8,000		-$10,000	$18,000		$30,000	-$22,000
기능적 효용		5,000		-$12,000	$17,000		$15,000	-$10,000
부대시설		5,000		-$14,900	$19,900		$19,800	-$14,800
소계		30,000		-59,900	$89,900		119,800	-$89,800
합계			합계			합계		
가치기여 총계		$1,278,100.00		$1,150,000.00			$1,500,000.00	
보정 후 거래가격					$1,278,100.00			$1,278,100.00

표 2: 사망일 기준 감정평가에 사용된 실제 RCA 거래사례비교표 사본 (부동산 식별 정보 삭제)

Sales Comparable Grid: Comps 1-3																				
Feature	**Subject**			**Comparable 1**					**Comparable 2**					**Comparable 3**						
Street, City, State Zip	<subject address>, Pacifica,CA 94044			<comp 1 address>, Pacifica,CA 94044					<comp 2 address>, Pacifica,CA 94044					<comp 3 address>, Pacifica,CA 94044						
APN | MLS# | Subj.Proximity	N/A		0 miles	009-xxx-xxx	MLxxxxxxxx		0.37 miles		009-xxx-xxx	MLxxxxxxxx		0.43 miles		009-xxx-xxx	CCxxxxxxxx		0.11 miles			
Sales Price|Concess.|Net SP		009-xxx-xxx	0 miles		$1,250,000		0	$1,250,000		$1,200,000		0	$1,200,000			$1,458,000		2000	$1,456,000	
Regression Features	**Factual Value**		**Value Contrib.**	**Factual Value**			**Value Contrib.**	**Adjustment**	**Factual Value**			**Value Contrib.**	**Adjustment**	**Factual Value**			**Value Contrib.**		**Adjustment**	
BASE VALUE			$1,361,209				$1,361,209	$0				$1,361,209	$0					$1,361,209	$0	
Sale Date | Close Date | Off Mkt	11/30/22	11/30/22	0 -$351,389	08/24/22	09/23/22	98	-$260,288	-$91,101	05/26/22	06/27/22	188	-$176,624	-$174,765	05/18/22	06/07/22	196	-$169,187		-$182,202	
Location: Longitude | Latitude	-122.483	37.658	-$66,040	-122.48	37.663		-$76,256	$10,216	-122.477	37.654		-$88,170	$22,130	-122.483	37.657		$0		-$66,040	
Site Size | Dimensions	4312 sf	N/A	0	4000 sf	N/A		0	$0	4700 sf	N/A		$0	$0	4510 sf	N/A		$0		$0	
Actual Age | Effective Age	58	N/A	$1,880	59	N/A		($14,556)	$16,436	58	N/A		$1,880	$0	56	N/A		$34,750		-$32,870	
Beds | Baths	3	1	-$118,379	4	2		-$39,460	-$78,919	3	2		-$78,919	-$39,460	4	3		$0		-$118,379	
Legal Living Area Above Gnd	1050 sf		$182,452	1270 sf			$220,680	-$38,228	1240 sf			$215,467	-$33,015	1210 sf			$210,254		-$27,802	
Legal Living Area Below Gnd	0 sf		$0	0 sf			0	$0	0 sf			$0	$0	0 sf			$0		$0	
Non-legal Living Area	0 sf		-$46,406	670 sf			0	-$46,406	920 sf			$0	-$46,406	729 sf			$0		-$46,406	
Secondary/ADU	sf		0	0 sf			0	$0	0 sf			$0	$0	0 sf			$0		$0	
Garage SF	800 sf		$74,406	600 sf			$46,328	$28,078	250 sf			$0	$74,406	351 sf			$11,371		$63,035	
Fireplaces	1		0	0			$0	$0	1			$0	$0	1			$0		$0	
Stories			0	2			$0	$0	2			$0	$0	2			$0		$0	
			$0				$0	$0				$0	$0				$0		$0	
			$0				$0	$0				$0	$0				$0		$0	
			$0				$0	$0				$0	$0				$0		$0	
			$0				$0	$0				$0	$0				$0		$0	
Residual Features	**CQA/Description**		**Value Contrib.**																	
CQA | Residual | Remaining	3.5	-24,150	$0	5.79	12,343		$0	$0	3.05	-34,842		$0	$0	5.6	7,603		$0		$0	
View	Average		$0	Average			$0	$0	Average			$0	$0	Average			$0		$0	
Design	Average		$0	Average			$0	$0	Average			$0	$0	Average			$0		$0	
Quality of Construction	Average		$0	Upgrades			$4,000	-$4,000	Average			$0	$0	Upgrades			$4,000		-$4,000	
Condition	Needs Updating		-$14,150	Weill maintained			$7,000	-$21,150	Average			$0	-$14,150	Condition			$3,000		-$17,150	
Functional Utility	Small Bathroom		-$5,000	Average			$0	-$5,000	Small 1 car garage			-$34,842	$29,842	Average			$0		-$5,000	
Heating & A/c	Average		$0	Average			$0	$0	Average			$0	$0	Average			$0		$0	
Porch/Patio	Needs improvement		-$5,000	Nice yard			$1,343	-$6,343	Average			$0	-$5,000	Average			$0		-$5,000	
Other			$0				$0	$0				$0	$0	Nice Shed			$603		-$603	
			$0				$0	$0				$0	$0				$0		$0	
			$0				$0	$0				$0	$0				$0		$0	
			$0				$0	$0				$0	$0				$0		$0	
			$0				$0	$0				$0	$0				$0		$0	
			$0				$0	$0				$0					$0			
Total VC / Net Adjustment			$1,013,583				$1,250,000	-$236,417				$1,200,001	-$186,418				$1,456,000		-$442,417	
Adjusted Sale Price							**Net Adj: -18.9%**					**Net Adj: -15.5%**					**Net Adj: -30.4%**			
of Comparables							**Grs Adj: 24.8%**	$1,013,583				**Grs Adj: 32.5%**	$1,013,582				**Grs Adj: 36.9%**		$1,013,583	
Copyright 2022-3, Pacific Vista Net ©																				

8 변수

MARS 를 실행하는 입력 측에는 다음과 같은 단순 변수만 있다.

1. MLS 스프레드시트에서 적재된 부동산 특성.
2. MLS 변수를 다듬거나 기존 변수의 변환을 추가하기 위해 VE 가 만든 파생 변수. 예컨대 VE 는 거래일 (Date of Sale) 을 감정평가 기준일에서 거래일을 뺀 값인 “DaysOffMarket” 으로 변환하기를 선호할 수 있다.
3. VE 는 주택담보대출 금리와 같은 제 3 자 자료원의 데이터를 들여올 수 있다. 또는 MLS 데이터에 위도, 경도, 표고 데이터가 없다면 다른 자료원에서 가져올 수 있다. 이러한 변수는 “외부 변수” 라 부를 수 있다.

MARS 의 출력 측에서 우리가 다루는 것은 다음 중 하나에 해당하는 항들의 조합으로 이루어진 가격 모형이다.

1. 입력 변수 하나만을 포함하며 그 변수의 가치기여에 대한 식을 제공하는 항. 이를 1 차 항 또는 (출력) 변수라 부른다. 항 자체가 가치기여 값으로 취급되며 이름이 부여된다.
2. 입력 변수 두 개를 포함하며 그 두 변수의 상호작용에 대한 가치기여의 식인 항. 이를 2 차 항 또는 (출력) 변수라 부를 수 있다.
3. 입력 변수 세 개를 포함하며 그 세 변수 사이 상호작용에 대한 가치기여의 식인 항. 이 논문에서는 이를 3 차 항 또는 (출력) 변수라 부를 수 있다.

그리하여 MARS 의 출력 변수는 잔차 가치기여를 포함한 가치기여들이다. 이들은 다시 보정치 계산이나 그래프 작성과 같은 거래사례비교표 계산에서 변수로 취급된다.

변수에는 두 가지 기본 온톨로지로 나뉘는 여러 유형이 있다. (1) 변수의 기원, (2) 변수의 용도, 즉 비잔차 대 잔차.

8.1 변수 기원 온톨로지

1. 1 차 변수

   1.1 MLS 회귀 변수
   1.2 외부 회귀 변수
   1.3 파생 회귀 변수

2. 팩터 변수 (factor variable):

   팩터 변수를 다루려면 그것을 식별할 수 있는 함수가 필요하다. 팩터 변수는 Excel 프로젝트 설정 워크북에 지정되므로, 전달받은 변수가 팩터 유형으로 분류되어 있으면 참을 반환하는 함수를 만든다.

3. 상호작용 변수: 차수가 2 이상으로 설정된 경우 MARS 가 모형을 생성하면서 만들어 낸다. 다만 해석의 어려움, 과적합의 가능성, 자유도에 미치는 영향과 그로 인해 견고한 모형을 위해 필요한 부동산 레코드의 증가를 고려하면 권장할 수 있는 상호작용 변수는 최대 3 개다.

   3.1 2 차 변수: 흔히 나타나며, GLA:BathRms, GLA:LotSize, DateOfSale:BathRms, Latitude_Longitude 가 좋은 예다.
   3.2 3 차 변수: 덜 흔한 상호작용으로, 신중을 요하며 과적합의 소지가 있다.

4. 잔차 분해 변수

8.2 변수 용도 온톨로지

1. MARS 가 생성한 가치기여

   1.1 1 차 변수

      1.1.1. MLS 회귀 변수
      1.1.2. 외부 회귀 변수
      1.1.3. 파생 회귀 변수

   1.2 상호작용 변수

      1.2.1. 2 차 변수
      1.2.2. 3 차 변수

   1.3 잔차 분해 변수

   1.4 집계 변수

2. 공학자가 생성한 변수 및 값

   2.1 잔차 분해 변수

      2.1.1. 디자인/미관
      2.1.2. 조경
      2.1.3. 기능적 효용
      2.1.4. 조망
      2.1.5. 상태
      2.1.6. 품질
      2.1.7. ...

이 변수 집단들에 대한 표기법과 상세한 설명은 다음 절에서 다룬다.

프로토콜 3.1: 팩터 변수

팩터 변수, 줄여서 팩터는 상호작용을 포함한 모든 항에 나타날 수 있다. MARS 회귀분석과 그래프 작성에서 특별한 처리를 요하므로 중요하다.
문자값을 갖는 변수는 필연적으로 팩터이지만, 수치 변수는 MARS 가 팩터로 인식하도록 팩터 변수로 표시해 주어야 한다. 예컨대 MLS 지역은 흔히 번호로 식별된다. 특히 샌프란시스코 베이 에어리어의 근린 구역은 흔히 고유한 정수로 식별된다. 이 값들을 팩터로 MARS 에 넘기면, MARS 는 다른 구역과의 관계와 무관하게 특정 구역이 그 경계 안의 부동산 거래에 평균적으로 기여하는 값이 있는지를 찾아내려 한다. 이와 달리 일반 변수의 경우 MARS 는 변수값이 변함에 따라 가치기여에도 어떤 연속성이 있다고 가정한다. 이는 중요한 구별이다. 더욱이 이러한 팩터 변수는 모형의 항에서 팩터가 아닌 변수와 함께 나타날 수 있어 그래프 작성을 어렵게 한다. 구체적으로, 팩터가 포함된 단일 변수 항을 그래프로 그릴 때는 그러한 팩터를 모두 한데 모아 그 기여를 양이든 음이든 막대그래프로 단순히 그린다.
팩터가 항에서 다른 변수와 함께 나타나는 경우, 팩터 변수는 특정 부동산에 대해 값을 하나만 가질 수 있으므로 언제나 하나의 가능한 값으로 나타난다. 어떤 구역의 값이 662 라면 정의상 그 구역이 다른 값을 가질 수 없다. 이것이 혼란스럽다면 사정을 더 복잡하게 만드는 점도 있다. MARS 가 생성한 모형에는 팩터에 대한 단일 변수명이 없다. MARS 는 입력으로 받은 팩터 변수의 어근 이름을 취한 다음, 가능한 값들을 덧붙여 만든 변수명으로 새 열들을 만든다. 그래서 MlsArea 같은 Excel 스프레드시트의 단일 열에 대해 MlsArea660, MlsArea661 등 가능한 값마다 데이터 프레임 열이 여러 개 생성된다. MlsArea661 같은 변수는 661 구역의 부동산에 대해 값 1 을, 그렇지 않으면 0 을 갖는다. 그런 다음 MARS 는 그 열들을 불리언 변수처럼 취급하여 회귀분석한다. 이렇게 변형된 이름이 MARS 가 제공하는 모형에 사용된다.
깔끔하고 단순한 것을 좋아하는 의뢰인에게 그래프를 (가능한 한) 더 이해하기 쉽게 만들려면 그 모형 이름들을 원래 이름으로 되돌리고 싶을 것이다. MlsArea661=1 대신 MlsArea=662 와 같이 보게 되는 것이다. 그러면 그래프 머리글에 “MLS Area = 662 에 대하여” 와 같은 문구를 적어 항을 그래프로 그릴 수 있다. 이렇게 하여 팩터 변수 M 개를 가진 N 차 항은 (N-M) 차 항으로 그려지고, 팩터 차원은 제목 속으로 사라진다.
한 항에 팩터 변수가 둘 이상 있을 수 있으므로, 값의 조합마다 그래프가 하나씩 필요할 수 있다는 점에 유의하라. 서로 다른 값 3 개를 갖는 변수 x 와 값 2 개를 갖는 변수 y 가 팩터가 아닌 변수 z 와 같은 항에 나타난다면 $2 \cdot 3 = 6$ 개의 그래프가 필요하다. 예를 들면 “MlsArea” + “Architectural Style” 이 있는데, 샌프란시스코의 빅토리아식 주택은 어느 구역에 있느냐에 따라 가치가 더하거나 덜할 수 있다. 다행히 대부분의 지역에서는 (건축양식이 6 가지 이상 있을 수 있더라도) 모형에 쓸 만큼 가치기여자로서 중요한 건축양식은 통상 1 3 가지뿐이다.

8.3 2 차 및 3 차 항

2 차 또는 3 차 회귀에서는 출력 모형의 일부 항에 변수가 각각 둘 또는 셋 포함될 수 있다. 예를 들면 다음과 같다.

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

이러한 항의 이름은 변수명을 알파벳순으로 정렬하여, 대개 밑줄을 넣어 이어 붙여 짓는다는 점에 유의하라. 3 차 항에 대해서도 모형이 산출한 변수명 삼중항에 같은 방식을 적용한다. 예컨대 “BathRms___GLA___LotSz” 이다.

이 변수들은 RCA 프로그램이 상호작용 변수를 포함한 모든 변수의 부동산별 기여를 계산하는 3 단계에서 1 차 입력 변수에 추가된다.

8.3.1 보정 변수

RCA 프로그램은 가치기여가 계산된 뒤 3 단계에서, 어떤 특성에 대한 대상 부동산의 기여에서 그에 대응하는 비교사례의 가치기여를 빼는 방식으로 모든 변수와 부동산에 대한 보정치도 계산한다.

8.3.2 집계 변수

집계 변수는 변수들의 부분집합을 대체하기 위해 정의되며, 그 가치기여와 보정치를 모두 더하여 보고서에 쓸 공통 변수로 모은다. 예컨대 마감된 주거면적과 마감되지 않은 주거면적을 다음과 같이 더 세분한 변수로 나누어 두었을 수 있다.

1. 지상 합법 마감면적
2. 지하 합법 마감면적
3. 지상 불법 마감면적
4. 지하 불법 마감면적
5. 지상 미마감면적
6. 지하 미마감면적

실제로 이런 구분이 의미 있는 지역이 있다.

MARS 회귀분석은 이 모든 변수에 대해 회귀를 수행하여 그 기여와 보정치를 제공하는 모형을 생성한다. 그러나 양식 보고서를 위해서는 이렇게 서로 다른 기여들을 더 적은 수의 변수로 결합하거나 집계해야 할 것이다. 집계의 세부는 부록에 제시해야 한다. 더 자주 있는 예는 상호작용 보정치를 1 차 변수 보정치 중 하나로 집계하는 것이다. 예컨대 GLA_Lotsize 에 대한 보정치를 GLA 로 또는 LotSize 로 집계하거나, 집계 시트에서 정한 대로 둘 사이에 50:50 으로 나눌 수도 있다.

8.3.3 특수 변수 및 표기법

π_i	부동산 $i = 1, 2, \ldots, n_p$ 의 실제 순 거래가격. 대상 부동산의 경우 π_0 는 가치 결론을 나타낸다 (증명 I 참조)
α_i	부동산 $i = 1, 2 \ldots, n_p$ 의 보정 후 순 거래가격
ρ_i	부동산 $i = 0, 1 \ldots, n_p$ 에 대한 MARS 추정 순 거래가격
β	모형의 기초가치
$\varphi_{i,j}$	부동산 $i = 0, 1, \ldots, n_p$ 에 대한 회귀 변수 $j = 1, 2, \ldots, n_r$ 의 가치기여
$\delta_{i,j}$	부동산 $i = 1, 2 \ldots, n_p$ 에 대한 회귀 변수 $j = 1, 2 \ldots, n_r$ 의 보정치
ϵ_i	부동산 $i = 0, 1 \ldots, n_p$ 의 총 잔차
$\epsilon_{i,c}$	부동산 $i = 0, 1, \ldots, n_p$ 에 대한 잔차 구성요소 c, $c = 1, 2, \ldots, n_c$ 의 가치기여
$\xi_{i,c}$	부동산 $i = 0, 1, \ldots, n_p$ 에 대한 잔차 구성요소 c, $c = 1, 2, \ldots, n_c$ 의 보정치

여기서

$$n_p = \text{비교사례 부동산 수}$$
$$n_r = \text{회귀 변수 수}$$
$$n_c = \text{잔차 분해 변수 수}$$

8.3.4 잔차의 분해

MARS 회귀분석이 모형을 만들고 나면, MARS 가 선택한 입력 변수들로부터 추정 거래가격을 생성한다. 그리고 그 값을 대응하는 순 거래가격에서 빼서 각 부동산의 잔차값을 얻는다. 이제 RCA 프로토콜은 VE 가 최종 비교사례 부동산 — 통상 6 12 개의 부동산 리스팅 — 을 정하고 나면, 각 부동산 리스팅을 하나씩 살펴 계산된 잔차를 가치에 영향을 준다고 판단되는 가장 유력한 특성 (잠재변수) 들로 분해할 것을 요구한다. 이때 무작위적 데이터 오류의 가능성은 고려하지 않는다. 예컨대 어떤 부동산에 대해 다음과 같은 분해 변수를 둘 수 있다.

1. 디자인
2. 기능적 효용
3. 시장가 초과/미달 거래
4. 상태

5. 품질

부동산마다 잔차 분해 변수가 얼마간 다를 것이므로, 이들을 하나의 순서집합으로 모아야 한다.

희소 특성의 기여 추정. 가치를 지닌 특성 가운데 회귀 경로에 태우기에는 지나치게 드문 것들이 있다. 780 건의 거래 중 3 건에만 있는 작업장은 방어 가능한 계수를 얻을 수 없고, 전진 단계는 그러한 변수를 아예 건너뛰는 것이 보통이다. 이들은 잔차 특성이며, 그 기여는 잔차 공간에서 추정할 수 있다. 즉 해당 특성 없이 모형을 적합시킨 다음, 모형의 잔차를 제외된 열 (들) 에 대해 통상최소제곱으로 회귀시키는 것이다. 이진 플래그의 경우 계수는 일시금 형태의 기여 (달리 말해 플래그가 붙은 거래들의 평균 잔차 편차) 이고, 수치 열의 경우에는 단위당 기여다. 상관된 여러 부대시설을 동시에 추정하는 경우 (예컨대 부속 건물 면적을 작업장 플래그와 함께), 회귀분석은 같은 금액을 두 개의 분해 행에 이중으로 계상하지 않고 그 몫을 둘 사이에 나눈다. 관측치가 약 열 개에 미치지 못하는 추정치는 참고용일 뿐이므로 감가상각 원가나 넓힌 표본 (예컨대 public remarks 에서 복원한 플래그) 으로 뒷받침해야 한다. 잔차 제약 자체가 최종적으로 배정되는 값의 상한을 지운다. earthUI(v0.11 이상) 는 이 방법을 *Residual Contribution*(잔차 기여) 탭으로 직접 구현하고 있으며, 동시 추정, 거래별 잔차 증거, 작업 파일용 문서화 노트를 함께 제공한다.

표기: 위 표의 표기법을 사용하면 다음을 얻는다.

$$\epsilon_i = \epsilon_{i,1} + \epsilon_{i,2} + \ldots + \quad \epsilon_{i,n_c} \tag{11}$$

여기서 n_c 는 잔차 분해 구성요소의 개수다.

VE 는 자신의 판단으로 잔차를 주어진 구성요소들로 분해해야 한다. 이 때문에 최종 가치 결론에 편의가 끼어들 수 있다고 생각할 수도 있다. 그러나 그렇지 않다. 분해된 보정치의 합계가 잔차와 같다면 가치 결론에 영향을 주지 않는다는 수학적 증명이 있기 때문이다. 분해의 목적은 가치를 확정하는 데 있지 않고, 다른 부동산 거래의 잔차와 견주어 그 잔차가 왜 그러한 값인지를 설명하는 데 있다. 공학자의 편의가 최종 가치 결론에 영향을 줄 수 있는 유일한 지점은 대상 부동산의 잔차 추정이다. 물론 대상 부동산에는 잔차를 계산하는 데 필요한 실제 거래가격이 없기 때문이다.

부동산마다 잔차 가치기여의 집합이 있으나, 이를 모든 부동산에 걸친 하나의 집합 C 로 병합해야 한다.

다음과 같이 두면: $A_i \quad = \{\{\epsilon_{i,c}\}_{c=1}^{n_c}\}$

$$\mathrm{C} \qquad = \{x \mid x \in \bigcup_{i=0}^{n_p} A_i\}$$

여기서 $[\epsilon_{i,c}]$ 는 값 $\epsilon_{i,c}$ 에 대한 변수명이고, A_i 는 그러한 변수들의 집합이며, n_c 는 각 ϵ_i 에 들어 있는 값의 개수다.

따라서 대상 부동산과 비교사례의 데이터를 별개의 열에 두면, 각 부동산에 대한 모든 회귀 모형 변수의 가치기여 아래에 그에 대응하는 n_c 개의 잔차 가치기여 집합 C 가 보여야 한다.

9 RCA 의 핵심 증명

다음과 같이 두자.

$$n_p = \text{부동산 수}$$
$$n_r = \text{회귀 변수 수}$$
$$n_c = \text{잔차 구성요소 변수 수}$$

증명 I: 모든 비교사례의 보정 후 거래가격은 대상 부동산의 추정 거래가격과 같다.

임의의 비교사례 i 에 대하여, 보정 후 거래가격 α_i 는 순 거래가격 π_i 에 보정치들을 더한 값이다. 보정치란 기초가치, 회귀 변수 $j = 1, 2, \ldots, n_r$, 그리고 잔차 각각에 대한 대상 부동산과 비교사례의 가치기여 차액을 말한다. 실제 거래가 없는 대상 부동산의 경우 순 거래가격 π_0 는 가치 결론 $\beta + \varphi_{0,1} + \varphi_{0,2} + \ldots + \varphi_{0,n_r} + \epsilon_0$ 로 정의한다. 그러면 다음과 같다.

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

증명 II: 모든 잔차 구성요소의 가치기여 합계가 비교사례의 잔차와 같다는 제약 아래에서 잔차 분해를 바꾸더라도 RCA 프로토콜의 최종 가치 결론에는 아무런 영향이 없다.

이는 거의 자명하다. 그럼에도 수학을 생각해 보지 않으면 놓치기 쉬운데, 전통적 감정평가사는 가치기여가 아니라 보정치에 주목하기 때문이다. 여기서 상쇄되는 것은 바로 가치기여이며, 이는 수십 년간 간과되어 온 사실이다. 어떤 비교사례 k($k > 0$) 에 대하여, 총 잔차 보정치는 n_c 개의 분해된 잔차 구성요소의 합과 같고, 이는 다시 가치기여 차액들의 합과 같으며, 이는 곧 대상 부동산의 총 잔차에서 비교사례의 총 잔차를 뺀 값과 같다. 그리고 이는 만들어진 잔차 구성요소의 개수나 그것이 무엇을 나타내는지와 무관하다. 대상 부동산과 선택된 비교사례 모두에서 잔차 구성요소들의 값이 총 잔차와 같기만 하다면 ξ_k 는 달라지지 않는다.

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

10 MARS 모형

10.1 모형 변수 및 상호작용 "g" 함수

RCA 방법의 핵심 원리 하나는 "설명적 추론" 이 무엇보다 중요하다는 것이다. 안타깝게도 특히 상호작용과 팩터 변수가 쓰이는 경우, 여러 가치기여 식에 대해 의미 있는 그래프를 제공하기란 쉽지 않을 수 있다. 그래서 이러한 그래프를 만드는 명확하고 엄격한 프로토콜을 제시하는 것이 긴요하다고 본다. 이를 여기서 제시하되, 이 프로토콜은 그래프를 생성하는 R 또는 Python 코드를 작성하는 개발자를 겨냥한 것이며 코딩에 관여하지 않는 이들에게는 표면적인 관심사에 그칠 수 있음을 이해하기 바란다.

MARS 가 생성한 가격 모형은 기초 상수에 각각 변수 1 3 개를 담은 여러 항을 더한 것이다. 이를 분석하기 위해 먼저 모형을 개별 항으로 나눈다. 그런 다음 동일한 변수 집합을 공유하며 하나의 상호작용을 나타내는 여러 항을 하나의 상호작용 식으로 결합한다. 회귀의 최대 차수가 3 이라고 하자. 기초 상수, n_1 개의 1 차 식, n_2 개의 2 차 상호작용 식, n_3 개의 3 차 상호작용 식으로 이루어진 3 차 회귀 모형을 생각해 보자. 이렇게 하면 단일 변수이거나 다변수 상호작용인 네 개의 구별되는 식 집단이 생기며, 각각은 두 가지 요인에 따라 다르게 그려진다. (1) 팩터 변수 (범주형 변수) 를 포함하는지 여부, 그리고 (2) 그 안에 들어 있는 서로 다른 변수의 개수. 구체적으로는 다음과 같다.

- 1 차 식 (변수 1 개) 은 2D 로 그린다.
- 2 차 식 (변수 2 개) 은 3D 로 그린다.
- 3 차 식 (변수 3 개) 은 3D 그래프의 집합으로 그린다.
- 식에 팩터 변수가 포함되면 서로 다른 팩터 하나마다 그래프의 차원이 하나씩 줄어든다 (예컨대 한 변수가 팩터이면 3D 그래프가 2D 가 된다).

각 함수 g 의 색인은 그것을 어떻게 그려야 하는지를 판정하는 데 필요한 정보를 제공하며, 서로 다른 변수의 개수와 팩터 변수의 존재 여부를 반영한다.

10.2 g 함수의 색인

각 함수 g 는 ${}^{f_{j,k}}g_k^j$ 로 색인되며, 여기서:

- $f_{j,k}$ 는 첫 번째 색인이다 (g 앞의 위 첨자).
  - j 는 0 에서 3 까지의 값을 가지며 집단을 식별한다.
    - $j = 0$: 기초 상수 (통상적 색인 규칙의 예외).
    - $j = 1$: 1 차 식.
    - $j = 2$: 2 차 식.
    - $j = 3$: 3 차 식.
  - k 는 집단 안에서 특정 함수의 색인이다 (예컨대 집단 j 의 첫 함수는 $k = 1$).
  - $f_{j,k}$ 의 값은 그 함수 안에 있는 서로 다른 팩터 변수의 개수를 나타낸다.
    - $f_{j,k} = 0$: 팩터 변수 없음.
    - $f_{j,k} > 0$: 팩터 변수의 개수 (예컨대 1, 2, 3).
- g 에 뒤따르는 색인:
  - g 뒤의 위 첨자 j(예: g^j) 는 명료성을 위해 집단 번호를 반복한다.

  - g 뒤의 아래 첨자 k(예: g_k) 는 $f_{j,k}$ 의 k 와 일치하여 집단 안에서 함수를 식별한다.

f 와 g 의 색인은 수치상 일치하지만 그 역할은 다르다. $f_{j,k}$ 는 그래프 작성 판단에 결정적인 팩터 변수의 개수를 세는 반면, g 의 j 와 k 는 그 집단과 위치를 규정한다.

10.3 그래프의 차원

함수 ${}^{f_{j,k}}g_k^j$ 에 대한 그래프의 차원 d 는 다음과 같이 계산한다.

$$d = j - f_{j,k} \tag{14}$$

여기서:

- j 는 서로 다른 변수의 개수 (상수의 경우를 제외하면 집단 번호) 다.
- $f_{j,k}$ 는 팩터 변수의 개수다.
- d 는 그래프로 그릴 연속 (비팩터) 변수의 개수이며 최대 3 이다 (최대 차수가 3 이므로). 연속 변수가 d 개인 함수는 $d+1$ 차원으로 그린다. $d = 1$ 이면 2D 그래프, $d = 2$ 면 3D 그래프 (또는 2D 열지도), $d = 3$ 이면 3D 그래프의 집합이다.

예를 들면:

- $j = 3$(변수 3 개) 이고 $f_{j,k} = 0$(팩터 변수 없음) 이면 $d = 3 - 0 = 3$ 이므로, 변수 하나씩을 고정한 3D 그래프의 집합이 필요하다.
- $j = 3$ 이고 $f_{j,k} = 1$(팩터 변수 1 개) 이면 $d = 3 - 1 = 2$ 이므로, 모형에 나타나는 팩터 변수의 값마다 3D 그래프 (또는 2D 열지도) 가 하나씩 필요하다.
- $j = 0$(상수) 이면 값이 하나뿐이므로 통상 그래프가 필요하지 않다.

$d = 3$ 인 경우 (예컨대 팩터 변수가 없는 3 차 식) 함수 전체를 시각화하려면 여러 개의 3D 그래프가 필요할 수 있으며, 각각은 변수 하나를 서로 다른 값에 고정하여 함수 거동의 한 부분을 보여준다.

10.4 최종 모형 식

MARS 모형은 네 개의 식 집단 (상수, 1 차, 2 차, 3 차) 으로 나뉜다. 색인 ${}^{f_{j,k}}g_k^j$ 는 다음을 알려준다.

- j: 집단과 변수의 개수.
- $f_{j,k}$: 팩터 변수의 개수.
- $d = j - f_{j,k}$: 그래프의 차원.

이 구조는 각 g 함수를 어떻게 시각화할지 이끌어 주며, 그래프 작성 방법이 식의 복잡성과 변수 유형에 부합하도록 보장한다.

MARS 모형이 제공하는 거래가격 추정치 $\hat{P}$ 는 다음과 같이 표현할 수 있다.

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

$n_1 = 1$ 차 항의 수
$n_2 = 2$ 차 항의 수
$n_3 = 3$ 차 항의 수

10.5 모형을 그래프로 나타내는 데 필요한 차트의 수

모형 안의 모든 함수를 시각화하려면 그래프가 몇 개나 필요한가? 답은 변수, 그 상호작용, 그리고 그것이 팩터 변수인지 여부에 달려 있다.

VE 는 R/earth 를 위한 Excel 설정 파일에서 변수와 상호작용을 제한할 수 있으며, 이는 어떤 상호작용이 허용되는지에 영향을 준다. 실용적인 접근은 R/earth 로 하여금 가능한 모든 상호작용을 평가하게 하여 거래가격 추정에 유의미한 기여를 하는 것들을 고르게 하되, 과적합이나 공선성 같은 문제를 함께 점검하는 것이다. 일부 상호작용은 배제될 수 있고, 이는 더 높은 차수의 상호작용에도 영향을 준다 (예컨대 3 차 상호작용에서 한 쌍이 제외되면 그 조합 전체가 제외된다).

팩터 변수의 경우 (예컨대 값이 12 개인 MLS Area 가 GLA 와 상호작용하는 경우) 그래프 작성이 더 복잡해진다. 구역마다 하나씩 12 개의 그래프가 필요할 수 있다. 여기에 값이 6 개인 Architectural Style 같은 또 다른 팩터가 MLS Area 및 GLA 와 상호작용한다면 최대 $12 \cdot 6 = 60$ 개의 그래프가 필요할 수 있다. 그러나 R/earth 가 그 모든 조합을 유의미하다고 보는 일은 드물다. 통상은 모형이 짚어 주는 몇 개의 핵심 상호작용 (예컨대 특정 구역과 양식) 만이 그래프로 그릴 가치가 있다.

필요한 그래프 수에 대한 간략한 지침은 다음과 같다.

1. **단일 변수 함수 (비팩터):** 그래프 1 개.
2. **단일 변수 함수 (팩터):** 모든 팩터 값을 합친 막대그래프 1 개.
3. **기초 함수:** 그래프 불필요.
4. **2 차 상호작용 (팩터 없음):** 그래프 1 개 (예컨대 거래가격을 색으로 나타내는 2D 열지도).
5. **2 차 상호작용 (팩터 1 개):** 모형에 나타나는 팩터 값마다 그래프 1 개 (예컨대 값이 2 개면 그래프 2 개).
6. **2 차 상호작용 (팩터 2 개):** 팩터 조합의 행렬을 담은 그래프 1 개 (예컨대 구역 3 개 × 양식 2 개 = 셀 6 개).
7. **3 차 상호작용 (팩터 없음):** 여러 개의 그래프 (예컨대 욕실 수 = 1, 2, 3, 4 에 대해 4 개).
8. **3 차 상호작용 (팩터 1 개):** 모형에 나타나는 팩터 값마다 그래프 1 개.
9. **3 차 상호작용 (팩터 2 개):** 모형에 나타나는 팩터 값 조합마다 그래프 1 개.
10. **3 차 상호작용 (팩터 3 개):** 서로 다른 값의 수가 가장 적은 팩터의 값마다 그래프 1 개를 그리고, 나머지 두 팩터는 색 행렬로 나타낸다.

더 복잡한 상황에서는 고급 그래프 기법이 필요할 수 있다.

더 정교한 그래프를 요하는 다른 가능성들도 있다.

MARS 가 생성한 완전한 모형의 예는 그림 1에 있다.

그림 1: MARS 모형의 구성요소

$$
\begin{aligned}
\text{Basis} &= {}^{0}g_{1}^{0} = 4957150 \\
\text{Age} &= {}^{0}g_{1}^{1} = -3163.1809255 \cdot \text{Age} \\
\text{AreaNbr} &= {}^{1}g_{2}^{1} = -623197.75104 \cdot I\{\text{AreaNbr} = 471\} \\
\text{Baths} &= {}^{0}g_{3}^{1} = -111023.51496 \cdot \max(0, 2 - \text{Baths}) + 96735.52418 \cdot \max(0, \text{Baths} - 2) \\
\text{Beds} &= {}^{0}g_{4}^{1} = -233489.66415 \cdot \max(0, \text{Beds} - 5) \\
\text{DaysOffMkt} &= {}^{0}g_{5}^{1} = -2577.2188622 \cdot \max(0, 189 - \text{DaysOffMkt}) - 1282.8789534 \cdot \max(0, \text{DaysOffMkt} - 189) \\
\text{GLA} &= {}^{0}g_{6}^{1} = -666.73587914 \cdot \max(0, 4027 - \text{GLA}) + 281.36312913 \cdot \max(0, \text{GLA} - 4027) \\
\text{Latitude} &= {}^{0}g_{7}^{1} = -18141680.417 \cdot \max(0, \text{Latitude} - 37.573) \\
\text{LotSize} &= {}^{0}g_{8}^{1} = -85.005683404 \cdot \max(0, 8000 - \text{LotSize}) - 23.562970183 \cdot \max(0, \text{LotSize} - 8000) \\
\text{Age\_Baths} &= {}^{0}g_{1}^{2} = -3916.3594872 \cdot \text{Age} \cdot \max(0, \text{Baths} - 3.5) \\
\text{Age\_Garage} &= {}^{0}g_{2}^{2} = -6503.1897339 \cdot \text{Age} \cdot \max(0, \text{Garage} - 2) \\
\text{AreaNbr\_LotSize} &= {}^{1}g_{3}^{2} = 78.54207825 \cdot I\{\text{AreaNbr} = 463\} \cdot \max(0, 8000 - \text{LotSize}) \\
&\qquad + 87.263298012 \cdot I\{\text{AreaNbr} = 466\} \cdot \max(0, \text{LotSize} - 8000) \\
\text{AreaNbr\_Baths} &= {}^{1}g_{4}^{2} = -1070222.5995 \cdot I\{\text{AreaNbr} = 464\} \cdot \max(0, \text{Baths} - 2) \\
\text{AreaNbr\_Latitude} &= {}^{1}g_{5}^{2} = 17571072.548 \cdot I\{\text{AreaNbr} = 466\} \cdot \max(0, \text{Latitude} - 37.573) \\
\text{AreaNbr\_GLA} &= {}^{1}g_{6}^{2} = -328.79173933 \cdot I\{\text{AreaNbr} = 471\} \cdot \max(0, \text{GLA} - 2550) \\
&\qquad + 525.69503229 \cdot I\{\text{AreaNbr} = 471\} \cdot \max(0, 2550 - \text{GLA}) \\
\text{DaysOffMkt\_GLA} &= {}^{0}g_{7}^{2} = 0.43839582286 \cdot \max(0, \text{DaysOffMkt} - 661) \cdot \max(0, 4027 - \text{GLA}) \\
\text{FrPlcNbr\_LotSize} &= {}^{0}g_{8}^{2} = 22.059059591 \cdot \max(0, 2 - \text{FrPlcNbr}) \cdot \max(0, \text{LotSize} - 8000) \\
&\qquad + 39.084508159 \cdot \max(0, \text{FrPlcNbr} - 2) \cdot \max(0, \text{LotSize} - 8000) \\
\text{GLA\_LotSize} &= {}^{0}g_{9}^{2} = -0.45274160024 \cdot \max(0, \text{GLA} - 2994) \cdot \max(0, 8000 - \text{LotSize}) \\
\text{ADU\_Age\_Garage} &= {}^{0}g_{1}^{3} = -5.4578633919 \cdot \max(0, \text{ADU} - 0) \cdot \text{Age} \cdot \max(0, 2 - \text{Garage}) \\
\text{AreaNbr\_Baths\_LotSize} &= {}^{1}g_{2}^{3} = 60.607126605 \cdot I\{\text{AreaNbr} = 460\} \cdot \max(0, \text{Baths} - 2) \cdot \max(0, \text{LotSize} - 2480) \\
\text{AreaNbr\_Baths\_FrPlcNbr} &= {}^{1}g_{3}^{3} = 593779.41722 \cdot I\{\text{AreaNbr} = 464\} \cdot \max(0, \text{Baths} - 2) \cdot \max(0, \text{FrPlcNbr} - 0) \\
\text{AreaNbr\_GLA\_Latitude} &= {}^{1}g_{4}^{3} = 54199.925713 \cdot I\{\text{AreaNbr} = 464\} \cdot \max(0, \text{GLA} - 2670) \cdot \max(0, \text{Latitude} - 37.573) \\
&\qquad + 58934.565747 \cdot I\{\text{AreaNbr} = 471\} \cdot \max(0, \text{GLA} - 2550) \cdot \max(0, \text{Latitude} - 37.568) \\
\text{AreaNbr\_DaysOffMkt\_GLA} &= {}^{1}g_{5}^{3} = 0.37173306643 \cdot I\{\text{AreaNbr} = 470\} \cdot \max(0, 661 - \text{DaysOffMkt}) \cdot \max(0, 4027 - \text{GLA}) \\
\text{Beds\_GLA\_LotSize} &= {}^{0}g_{6}^{3} = -5.1607264026 \cdot \max(0, 4 - \text{Beds}) \cdot \max(0, \text{GLA} - 2994) \cdot \max(0, 8000 - \text{LotSize}) \\
\text{Age\_AreaNbr\_Baths} &= {}^{1}g_{7}^{3} = -10759.444688 \cdot \text{Age} \cdot I\{\text{AreaNbr} = 460\} \cdot \max(0, \text{Baths} - 3.5)
\end{aligned}
$$

11 데이터의 특성

이 절에서는 MARS 를 다룬다. MARS 는 전적으로 자신이 받는 데이터에 의존하며, 그 데이터의 대부분은 어떤 MLS(Multiple Listing Service) 나 세무 평가관에게서 나온다. 다만 데이터를 관리하고 개선하여 이윤을 붙여 파는 다른 데이터 공급자를 거칠 수도 있다. 어떤 지역에서는 상당히 신뢰할 만하고, 다른 지역에서는 그렇지 않다. 결정적인 사안에서는 부동산 조사자가 직접 관찰하고 측정해야 하는데, 그 정확성 또한 의문의 대상이 될 수 있다.

감정평가공학자는 데이터가 무엇을 나타내는지, 그리고 그것이 나타내야 할 바를 얼마나 잘 나타내는지 이해해야 한다. 여기서 다음과 같은 여러 쟁점이 따라 나온다.

- 어떤 변수를 측정값으로 분류하는 기준은 무엇인가?
- 변수가 측정값이 아니라면 달리 무엇일 수 있는가?
- 받은 데이터가 정확하지 않다면 회귀분석이 어떻게 쓸모 있을 수 있는가?

MLS 리스팅 서비스와 세무 평가관 데이터에서 회귀분석용으로 얻는 데이터의 대부분은 측정값이나 개수를 함의하는 수치 데이터다. 일부는 참/거짓의 논리값이다.

11.1 잔차의 목적

잔차는 회귀 모형에 포함되지 않은 변수들이 더한 가치의 간접적 척도 역할을 한다. 헤도닉(hedonic) 문헌의 용어로 하자면 이러한 측정되지 않은 특성은 잠재변수다. 즉 거래가격에 영향을 주지만 직접 측정되지 않는 변수이며, 대개는 그것을 객관적으로 측정할 실용적 방법이 없기 때문이다. 그 값은 대신 추론되어야 하는데, 실무에서 잠재변수의 값은 흔히 주관적 판단으로 추정된다. RCA 에서는 그 판단이 잔차 순위표에 닻을 내린다. 부동산을 잔차 또는 제곱피트당 잔차로 정렬함으로써 매력이 가장 낮은 부동산에서 가장 높은 부동산에 이르는 객관적 순위를 얻는다. 지역 시장을 깊이 이해하는 숙련된 분석자가 MARS(다변량 적응 회귀 스플라인) 모형을 만든다는 전제 아래 이 순위 매기기 방법은 효과적이다. 추가로 요구되는 사항은 다음과 같다.

- 데이터와 모형의 품질: 잘 구축된 회귀 모형과 양질의 데이터는 믿을 만한 잔차의 전제조건이다. 샌프란시스코 베이 에어리어에서 나의 목표는 R2 약 75 80%, 교차검증 R2(CV R2) 55 70% 의 모형을 개발하는 것이다.
- 과적합을 피하고 R2 를 제한하라: SF 베이 에어리어처럼 가격대가 높고 성숙한 대부분의 시장에서 감정평가공학자는 R2 상한을 약 80% 로 잡아야 한다. 주거용 부동산의 경쟁력 있는 가치 가운데 상당 부분이 계량 불가능하여 회귀 모형에 직접 넣을 수 없는 주관적 매력에 있기 때문이다. R2 가 80% 를 크게 웃돈다면 과적합된 모형일 수 있으므로 더 면밀히 살펴야 한다. 다만 한 가지 유의할 점이 있다. 과적합 없이도 R2 가 90% 를 넘는 시장 지역이 있고, 감정평가가 어려워 잘해야 50 60% 또는 그 이하에 머무는 지역도 있다.
- 특정 지역에서의 높은 R2: 다만 택지개발지구나 그 밖의 균질한 주거지역처럼 부동산의 동질성 덕분에 놀라울 만큼 흔히 높은 R2 값이 나오는 지역도 있다.

11.2 근사 회귀변수로서의 상태 및 품질 등급

Fannie Mae 의 상태 및 품질 등급 척도 (C1–C6, Q1–Q6) 에는 특별한 언급이 필요하다. 이 척도들은 참된 측정 특성으로 취급하기에는 너무 거칠고, 감정평가사가 부여하는 것이므로 실무상 주관적 등급 — 그 자체로 잠재변수 — 이다. 두 층위를 명시적으로 구별해 두는 것이 유용하다. 하나는 회귀변수로 쓰이는 감정평가사의 거친 등급, 즉 상태-조립도 (condition-1) 와 품질-조립도 (quality-1) 이고, 다른 하나는 잔차에서 복원되는 세밀한 잠재변수, 즉 상태-정밀도 (condition-2)

와 품질-정밀도 (quality-2) 다. VE 는 감정평가사의 조립도 등급을 근사치로 받아들여 그에 대해 회귀분석을 수행할 수 있으며, 그렇게 얻은 근사적 가치기여는 세밀한 잠재변수의 초기 값 배정에 유용하다. 그렇게 하는 경우의 프로토콜은, 이 두 조립도 변수의 회귀 추정 가치기여를 모형 추정치에서 빼내어 — 거래사례비교표의 회귀 구역 아래 측정된 속성에서 옮겨내어 — 잔차에 다시 더하는 것이다. 그런 다음 이를 잔차에서 제거하여 그에 대응하는 세밀한 잠재변수 res_condition1 과 res_quality1 에 배정한다. 이어서 MLS 의 public remarks(public_remarks) 와 관련 문자열 목록에서 추출한 추가 정보로 이들 잠재변수 값을 미세하게 다듬는다. 최종 결과는 세밀한 가치기여가 조립도 가치기여를 대체하는 것이다. 거친 등급은 근사 모형을 위한 발판 역할만 할 뿐이며, 상태와 품질에 대해 최종적으로 보고되는 값은 잔차에서 도출된 세밀한 값이다.

11.3 데이터 오류

소셜 미디어의 언급들에 비추어 볼 때, 캘리포니아 MLS 데이터는 미국의 다른 여러 지역에 비해 대체로 품질이 높다. 통상 MLS 데이터는 성숙한 대도시권에서는 양질일 것으로, 농촌 지역에서는 그보다 못할 것으로 예상해야 한다.

여러 이유로 데이터에 간간이 오류가 있기는 하나, 데이터가 충분하다면 이러한 오류는 대체로 서로 상쇄되어 가격 모형에 쓰이는 평균에는 미미한 영향만 미친다.

그러나 특정 지역에서는 체계적 편의가 발생할 수 있다. 예컨대 많은 주택이 애초에 2 층을 온전히 갖추어 지어졌으나, 일부 매수인에게 2 층 방을 없애 1 층의 한두 방에 높은 천장을 만드는 선택권이 주어져 총 주거면적 (GLA) 이 크게 줄어든 근린지역을 생각해 보라. 그러한 변경이 어떤 이유로든 이후 평가관 기록에 반영되지 않았다고 하자. 그 결과 실측값이 2,000 제곱피트인데도 MLS 에는 2,400 제곱피트로 나타날 수 있다. 그러한 경우 MARS 회귀분석을 수행하기 전에 MLS 데이터를 바로잡아야 한다. 비교사례에 대해 이 정보를 얻으려면 얼마간의 조사 노력이 필요할 수 있다는 점에 유의해야 한다. 불일치 사항은 후일의 참조를 위해 기록해 두어야 한다.

이상을 고려하면, 양질의 MLS 지역에서는 잔차로 흘러드는 오류가 디자인, 상태, 품질과 같은 주관적, 미측정 특성의 영향에 비해 미미하다고 보는 것이 합당하다. 이는 RCA 방법의 정확도를 추정할 때 중요한 고려사항이다. 예컨대 어느 시장 지역의 R^2 값이 80% 라면 가격의 약 20% 가 CQA 잔차에 귀속된다. 따라서 대상 부동산을 잔차 순위표에 놓는 데서 생기는 $\pm 10\%$ 의 오차는 최종 가치 결론에서 $\pm 10\% \cdot 20\% = \pm 2\%$ 의 오차로 전파된다. 정확도는 이렇게 높이는 것이다! 많은 경우 $\pm$ 1% 를 훨씬 밑도는 정확한 가치 추정이 가능하다.

그러므로 잔차를 MARS 회귀분석에서 모형화되지 않은 변수들의 집합적 가치에 대한 간접 지표로 취급하는 것은 합당하다. 다만 감정평가공학자가 체계적 측정 편의가 어디에 왜 존재하는지를 식별할 만큼 충분한 지역 시장 경험을 갖추고 있어야 한다. 이 접근은 부동산 데이터의 한계와 미묘함을 인정하면서도 분석이 실제 부동산 가치를 정확히 반영하도록 보장한다.

실례: Sea Ranch 에서 측정되지 않은 가치결정요인 측정하기

어느 시장 지역에서 달성 가능한 R^2 가 그 시장의 외견상 동질성에서 기대할 법한 수준보다 낮을 때, 이 틀이 제시하는 대응은 그 한계를 받아들이는 것이 아니라 빠진 변수를 찾아내어 측정하는 것이다. 소노마 카운티의 Sea Ranch 지역사회가 이를 직접적으로 보여준다.

측정하라, 받아들이지 말라는 대응의 구체적 사례다. 소노마 카운티의 Sea Ranch 지역사회는 바다까지의 거리가 주된 가치결정요인인 해안 개발지다. 다른 특성이 같다면 절벽에서 수백 피트 떨어진 주택은 내륙으로 4 분의 1 마일 떨어진 주택과 상당히 다른 가격을 받는다. 이 지역사회의 MLS 데이터에는 바다까지의 거리가 변수로 기록되어 있지 않다. 사용 가능한 MLS 변수들로 적합시킨 초기 MARS 모형은 그 시장의 외견상 동질성이 뒷받침했어야 할 수준보다 낮은 설명력을 냈고, 잔차 구조는 데이터에서 보이는 어떤 특성으로도 부동산을 충분히 정렬하지 못했다. 이는 시장이 진정으로 잡음이 많다는 것이 아니라 중요한 결정요인이 모형에서 빠져 있다는 신호였다.

해법은 직접 측정이었다. 저자는 해안선을 기준으로 부동산 배치를 보여주는 지역사회 지도를 구

해 수백 채의 주택에 대해 바다까지의 거리를 측정했다. 측정된 거리를 MLS 데이터 집합에 변수로 추가하고 MARS 회귀분석을 다시 실행했다. 새 변수는 모형에서 유의미한 것으로 선택되었고, 설명력은 크게 높아졌으며, 그에 따라 잔차 순위표는 이전까지 그것을 오염시켰던 잠재적 바다-거리 신호가 아니라 남은 주관적 특성 — 조망, 상태, 품질 — 과 부합하는 양상으로 부동산을 정렬하게 되었다. MLS 변수가 없거나 신뢰할 수 없는 구릉지 근린지역에서는 표고를 같은 방식으로 다루어 왔다.

이 방법론적 원리는 일반적이다. 달성 가능한 R^2 가 그 시장의 외견상 동질성이 뒷받침해야 할 수준에 못 미치면, 첫 번째 가설은 중요한 변수가 모형에서 빠져 있다는 것이며, 적절한 대응은 그것이 무엇인지 찾아내어 직접 측정치를 수집하는 것이다. 자료원으로는 지역사회 지도, GIS 데이터, 세무 평가관 기록, 건축 당국 기록, 위성 영상, 현장 조사 등이 있다. 그러한 측정 작업의 비용은 실재하지만 한정되어 있다. 특정 시장 지역에 대해 변수를 한 번 수집해 두면 그 지역의 이후 업무에 그대로 이어 쓸 수 있다. RCA 는 새로운 변수를 깔끔하게 흡수하는데, 이 틀은 어떤 변수가 MARS 회귀분석에 들어가는지에 무관하기 때문이다. 새 변수가 실재하는 신호를 담고 있는지는 잔차 구조가 VE 에게 알려 주며, 가치기여는 그에 따라 조정된다.

대안적 대응 — 낮은 R^2 를 받아들이고 더 넓은 오차범위를 보고하는 것 — 은 빠진 결정요인이 진정으로 측정 불가능할 때 (매수인 개인의 선호, 일시적 시장 심리, 특이한 거래 사정) 에는 방어 가능하지만, 단지 수집되지 않았을 뿐일 때는 그렇지 않다. 모형이 약할 때 VE 가 먼저 할 일은 빠진 신호가 후자인지를 묻는 것이다.

11.4 정확성이란 무엇인가

감정평가에서 완전한 정확성을 갖추려면 다음이 필요하다.

1. 시장가치에의 부합: 대상 부동산의 가상 거래가 주어진 시장가치 정의에 100% 부합할 것
2. 정확하고 견고한 모형 개발: 지역에 따라 R2 약 75 80%, CVR2 대략 55 70% 의 고품질 MARS 모형이 개발될 것
3. 충실한 데이터 오류 분석: 측정 변수의 있음직한 데이터 오류가 모두 파악되어 통제되고 잔차로 흘러드는 부분이 거의 없을 것
4. 체계적 편의의 반영: 데이터의 체계적 편의가 모두 반영되고 보정될 것
5. 잔차 순위 분석: 잔차 순위상 부동산들이 최하위에서 최상위까지 특성의 합리적 양상을 따르는지 확인할 것. 상속재산 처분 거래, 숏세일, 경매 거래와 같은 이상치는 조사하고 설명하여 필요하면 제외할 것
6. 대상 부동산의 위치 결정: 잔차 순위를 충분히 이해한 상태에서 대상 부동산을 정확히 자리매김할 것. 매력이 뚜렷이 높은 부동산과 뚜렷이 낮은 부동산을 각각 찾아 그 사이에 대상 부동산을 놓고 점수를 매긴다. 그런 다음 위아래 부동산 사이의 간격을 가능한 한 좁혀 나간다.
7. 설명적 추론: 정확성을 위해 설명적 추론을 결코 희생하지 말 것. 감정평가사는 적어도 통상의 지력을 갖춘 사람에게 대상 부동산과 비교사례 사이의 가치 차이를 설명할 수 있어야 한다. 예컨대 R2 와 CVR2 가 최적인 좋은 MARS 모형은 측정된 특성의 가치기여를 설명하는 데 쓸 수 있고, 감정평가공학자의 잔차 분해는 그 밖의 특성의 가치기여를 설명하는 데 쓸 수 있다.
8. 과적합: 단독으로든 조합으로든 특정 부동산 한둘을 식별할 수 있는 특성을 사용하지 않음으로써 모형의 과적합을 피할 것. 특히 그 값이 소수의 부동산을 특정할 수 있는 팩터 변수의 사용을 피하라. 규모 있는 데이터 집합에서 근린지역을 식별하는 팩터 변수는 적절하지만, 수영장이 거의 없는 지역에서 수영장 보유 부동산을 식별하는 변수는 신중하

> 게 써야 한다. 정확한 GIS 좌표, 필지번호, 주소처럼 개별 부동산을 특정하는 특성은 절대 허용되지 않는다.

회귀 R2 가 80% 라면 대상 부동산 가치의 약 20% 가 잔차에서 비롯되리라 예상한다. 다만 이는 근사일 뿐이다. CQA-잔차 곡선으로부터 우리는 잔차 분포의 대부분이 CQA 의 낮은 쪽과 높은 쪽 극단에 있음을 안다. CQA-잔차 곡선과 낮은 CQA 값 및 높은 CQA 값에서의 가파른 기울기에 비추어 보면, 이 구간에서는 잔차의 추정이 덜 정확하리라 예상할 수 있다. 그러나 이는 이 구간에 부동산이 상대적으로 드물다는 점과 인접한 부동산 사이의 매력 차이가 더 크다는 점으로 상쇄된다. 그리하여 중간 구간에서는 ±5% 범위의 영향을 말할 수 있는 반면, 극단에서는 다소 좁은 ±2 ∼ ±3% 범위를 더 유념하게 된다.

RCA 의 정확도를 계산하는 일은 CQA-잔차 곡선의 비선형성 때문에 일반적으로 다소 어렵다. 그러나 합리적인 잔차 순위표가 있다면 사안별로는 가능하다. CQA 점수의 범위는 0.0 에서 10.0 이다. 좋은 MARS 모형이 있다면 잔차 순위표는 매력이 가장 낮은 것에서 가장 높은 것까지 뚜렷한 양상을 보여야 하며, 전체 범위의 ±5% 이내, 즉 CQA 점수로 ±0.5 이내로 평정하기가 어렵지 않을 것이다. 회귀분석에 부동산 100 건이 들어갔다면 이는 대상 부동산을 10 건의 무리 안에서 정확히 자리매김하는 것에 해당한다. 그러니 표 3의 Burlingame 거래 데이터를 참조하여, VE 가 추정한 CQA 가 7.5, 5.1, 1.9 이고 대상 부동산에 대한 그의 가치 결론이 $3,000,000 일 때의 근사적 정확도 (즉 오차) 를 계산해 보자. 표 3에는 사용된 수백 건 가운데 표본만 실려 있으므로 필요한 곳에서는 (선형성을 가정하여) 잔차를 보간해야 한다. 분명히 해 두자면, 여기서는 VE 가 각 경우에 100 건뿐인 잔차 순위표에서 대상 부동산의 순위를 ±5 건 범위 안으로 매우 확신한다고 가정한다.

CQA= 7.5 ±0.5					
CQA	잔차	절대차	예상 오차	가상 거래가격	오차율
8.00	$257,112				
7.50	$197,113	$59,999			
7.00	$154,615	$42,498	±$51, 249	$3,000,000	±1.71%

CQA=5.1 ±0.5					
CQA	잔차	절대차	예상 오차	가상 거래가격	오차율
5.60	$50,911				
5.10	$19,559	$31,352			
4.60	-$28,207	$47,766	±$39, 559	$3,000,000	±1.32%

CQA= 1.9 ±0.5					
CQA	잔차	절대차	예상 오차	가상 거래가격	오차율
2.40	-$228,411				
1.90	-$276,381	$47,970			
1.40	-$326,938	$50,557	±$49, 264	$3,000,000	±1.64%

끝으로, 정확성을 검증하는 유일하게 객관적인 방법은 엄격한 프로토콜 (이 논문은 그 큰 줄기만 부분적으로 시사할 뿐이다) 을 세우고 그 준수 여부를 검증하는 것이다. 실제 거래가격은 시장의 불완전성 때문에 시장가치에서 벗어날 수 있으므로 이것이 필요하다. 감정평가와 이후의 거래를 비교하면 정확성에 관한 어느 정도의 시사를 얻을 수 있다. 그러나 부동산 시장의 복잡성과 불안정성 때문에, 확립된 기준과 프로토콜에 기초하여 가치 추정치를 제공하려면 감정평가사나 고도로 자격을 갖춘 감정평가공학자의 전문성이 필요하다. 그 기준과 프로토콜은 유능한 감정평가공학자들이 독립적으로 거의 동일한 가치 결론에 이르도록 보장할 만큼 엄격해야 한다. 좋은 데이터가 있는 대부분의 경우 이는 달성 가능한 목표일 것이다.

11.5 이상치로서의 대상 부동산

부동산의 시장가치를 추정하려면 최근 거래사례의 분석이 필요하므로, 현대화와 수선이 이루어지지 않은 노후한 대상 부동산의 여러 특성을 최근 거래사례가 잘 대표하는지 물어보는 것은 마땅하다. 최근 거래사례가 통상 매도가격을 높이려고 개보수를 거친 주택인 샌프란시스코 베이 에어리어에서는 특히 그러하다. 그리하여 비교 거래사례의 평균적 품질과 상태는 꽤 오랫동안 시장에 나오지 않았던 노후 주택보다 대체로 우월하다. 이러한 가치 차이는 대상 부동산의 잔차 추정, 달리 말해 그 CQA 점수에서 포착되어야 한다. 이는 대상 부동산을 평가할 때 데이터 품질과 정확성과 관련하여 중요한 고려사항이다.

11.6 예상 밖 데이터를 다루는 데서 CVR2 의 중요성

MARS 회귀분석은 무작위로 선택된 부동산 부분집합에 R/earth 를 여러 차례 반복 적용하는 데 기초하는데, 이 부분집합은 통상 MLS 데이터 집합의 80% 에서 90% 에 해당하며 훈련 집합이라 불린다. 나머지 10% 20% 의 부동산이 검정 집합이 된다. 이 단계에서 훈련 데이터 집합으로 MARS 모형을 만들고, 그 모형으로 검정 데이터 집합에 속한 부동산의 거래가격을 예측한 뒤 실제 거래가격과 비교한다. 이 방법론은 아직 관측되지 않은 부동산의 거래가격을 예측할 수 있는 견고한 모형을 개발하는 가장 효과적인 전략으로 데이터 마이닝 분야에서 인정받고 있다.

그러나 CVR2 지표의 일차적 효용은 실제 시장 지역 거래에 관한 회귀 모형의 사실적 정확성을 반영하는 데 있다기보다, 특정 기준일 현재의 가상적 부동산 거래에 관한 모형의 예측력을 최적화하는 데 있다. 그러한 부동산은 훈련 데이터 집합에 포함되어 있지 않으며, 훈련 데이터와 검정 데이터 어디에도 없는 새롭거나 예상치 못한 특성 조합을 보일 수 있다.

이러한 맥락에서, 예컨대 감정평가 기준일 한 달 전에 매물로 나왔고 지난 15 년간 개보수가 없었던 부동산의 거래가격은 어떻게 추정할 수 있을까? 반복적 교차검증을 거치지 않은 모형보다는 강한 CVR2 를 보이는 모형이 가장 정확한 추정치를 제공할 가능성이 높다. 또한 데이터 집합에 개보수, 수선, 현대화의 정도가 다양한 비교사례가 충분히 포함되어 있는 것이 긴요하다.

궁극적으로, 비교사례 데이터 집합에서 여러 부동산 특성의 가치기여를 정확히 예측하는 견고한 모형을 갖고 있다면, 잔차로 정렬된 그 부동산들의 순위표 안에 새로운 대상 부동산을 효과적으로 자리매김할 수 있다. 이렇게 위치를 정하면 해당 부동산에 적절한 CQA 점수와 잔차를 배정할 수 있다.

12 잔차

"잔차 제약 접근법" 이라는 이름이 말해 주듯, 이 방법론에서 잔차는 중심적이다. 여기서 잔차란 어느 부동산의 순 거래가격과 회귀분석이 예측한 거래가격의 차액을 말한다.

양의 잔차는 총 주거면적 (GLA) 이나 대지면적과 같은 통상의 측정 가능한 속성에 근거한 추정가격보다 그 부동산이 비싸게 팔렸음을 시사하며, 이는 평균보다 매력이 클 수 있음을 함의한다. 음의 잔차는 더 싸게 팔렸음을 뜻하며, 매력이 덜하거나 포착되지 않은 다른 부정적 속성이 있음을 시사할 수 있다.

큰 주택일수록 잔차도 클 가능성이 높으므로, 보다 정규화된 비교를 위해 GLA 기준 제곱피트당 잔차를 쓸 수 있다. 이 문제에 대한 더 깊은 논의는 다른 논문으로 미룬다.

12.1 순위와 점수

측정 변수에 대한 MARS(다변량 적응 회귀 스플라인) 모형을 만든 뒤 각 부동산에 대해 잔차를 계산한다. 이 잔차는 관측된 순 거래가격과 모형이 예측한 거래가격의 차액이다. 잔차/SF, 즉 잔차를 주택의 주거면적으로 나눈 값을 담을 열을 하나 더 만들어야 한다. 대체로 특정 부동산의 잔차는 주거면적이 클수록 크다는 이해에서다. 근린지역에 따라서는 다른 통계량이 더 잘 맞을 수도 있으나, 그것은 다음 기회의 주제다.

가정 7: "잔차에 의한 순위"

이 논문에서 "잔차에 의한 순위" 라는 표현은 잔차에 의한 순위, 잔차/SF 에 의한 순위, 또는 감정평가공학자가 해당 시장 지역에 가장 적합하다고 판단하는 잔차의 다른 함수에 의한 순위를 뜻하는 것으로 읽어야 한다. SF 베이 에어리어에서의 저자의 경험에 따르면, 특히 GLA(총 주거면적) 가 유난히 작거나 큰 대상 부동산을 다룰 때는 잔차/SF 가 대체로 가장 믿을 만하다.

다음으로 부동산을 잔차가 가장 작은 (음의 값이 가장 큰) 것에서 가장 큰 것 순으로 정렬한다. 그런 다음 각 부동산에 0.0 에서 10.0 범위의 상태-품질-매력 (CQA) 점수를 부여한다. 이 점수는 다음과 같이 계산한다.

CQA 점수는 전체 부동산 가운데 그 부동산 잔차의 백분위 순위에 대응한다. 구체적으로는 해당 부동산보다 잔차가 낮은 부동산의 비율을 구한 뒤 그 비율을 10 으로 나누어 계산한다. 예컨대 어떤 부동산의 잔차가 전체의 50% 보다 높다면 그 점수는 5.0 이 된다.

반올림 때문에 잔차가 가장 큰 부동산 하나가 10.0 을 받을 수 있다. 그러나 논리적으로 자기 자신을 포함한 100% 의 부동산보다 잔차가 높을 수는 없으므로 만점 10.0 은 이론상 불가능하다. 반면 최저점 0.0 은 가능한데, 적어도 한 부동산은 자기보다 잔차가 낮은 부동산이 없기 때문이다. 비교 대상 부동산이 많으면 반올림 때문에 둘 이상이 0.0 을 받을 수도 있다는 점에 유의하라.

따라서 CQA 점수는 어느 부동산의 속성이 모형의 기대에 얼마나 부합하는지를 직관적으로 보여주는 척도이며, 점수가 높을수록 모형의 예측에 더 잘 부합함을 뜻한다.

CQA 점수가 비교사례의 매력에 대한 간접 지표로 여겨질 수 있는 것은 거래 조건이 그 시장가치에 상응할 때뿐임을 인식하는 것이 긴요하다. 감정평가공학자는 거래에 대한 포괄적 설명과 철저한 조사를 통해 잔차값에 시장가 초과 또는 미달 요소가 포함되어 있는지를 확인할 책임이 있다. 시장가를 크게 넘거나 밑도는 거래가 잦은 것은 아니지만, 강제 매각, 상속재산 처분, 무능력, 담합 등의 사정에서 일어날 수 있다.

표 3는 MLS 데이터의 준현실적 발췌를 잔차 및 CQA 점수와 함께, 총액 기준과 제곱피트당 기준으로 제시한다. 이 연구에서 분석된 부동산의 총수는 약 600 건이었다. 회귀에 투입된 여러 변수가 표에 표시되지 않았다는 점은 유의할 만하다. 달리 말해 여기 제시된 잔차는 거래일과 같은 다른 변수의 영향을 MARS 회귀분석이 이미 고려한 뒤의 값이다.

표 3: 잔차로 정렬한 비교 거래사례

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 CQA-잔차 곡선

어느 근린지역에 대해 CQA 대 잔차/SF 를 그리면 예외 없이 그림 2에 보이는 CQA-잔차 특성 곡선이 나온다. 이 그림은 2022 년경 캘리포니아 Burlingame 의 데이터를 나타낸다. CQA 점수를 잔차값에 대해 그릴 때도 대체로 비슷한 곡선이 관찰되지만, 나는 대상 부동산의 총 주거면적 (GLA) 과 관련하여 CQA 를 맥락화하는 데 유용하다는 점에서 잔차/SF 를 선호한다. 잔차/SF 에 GLA 를 곱하면 어느 부동산의 CQA 를 그에 대응하는 잔차값으로 환산할 수 있다.

감정평가공학자는 부동산의 규모, 구체적으로는 총 주거면적 (GLA) 이 잔차의 크기에 유의미한 영향을 주는지를 신중히 판정해야 한다. 이 판정은 부동산 가치를 계산하는 가장 정확한 방법을 정할 수 있으므로 결정적이다. 예컨대 고급 주방기기와 호화로운 욕실 설비는 주택의 잔차값을 크게 높일 수 있다. 그러나 주방과 욕실의 크기와 배치가 대체로 균일한 근린지역에서는 이러한 업그레이드가 GLA 와 강하게 상관하지 않을 수 있다. 그것이 잔차값에 미치는 영향은 차지하는 공간보다는 그것이 나타내는 품질에 더 관계될 수 있다.

반대로 부동산의 외관 매력을 눈에 띄게 높이는 값비싼 외부 요소와 자재는 부동산의 규모와 더 직접적인 관계를 가질 수 있다. 큰 주택일수록 광범위한 조경, 우수한 지붕 자재, 넓은 파티오 같은 더 정교한 외부 구성을 수용할 수 있고, 이는 자연스럽게 부동산의 전반적 매력과 그에 따른 잔차값을 높인다.

그러므로 감정평가공학자가 잔차를 분석할 때 이러한 동학을 고려하는 것은 긴요하다. 고품질 설비와 특성이 더하는 내재적 가치를 판정하는 데 그치지 않고, 그러한 개선이 부동산의 물리적 규모와 어떻게 상호작용하는지도 이해해야 한다. 이러한 포괄적 접근은 감정평가가 부동산의 질적 속성과 양적 속성을 모두 정확히 반영하도록 보장한다.

따라서 CQA-잔차 곡선과 CQA-잔차/SF 곡선을 모두 그려 그 차이를 살펴볼 것을 권한다. 그 차이가 시장 지역의 양상을 시사할 수 있다. 감정평가공학자는 프로그램으로 두 잔차를 모두 계산한 뒤 가중평균을 취하여 대상 부동산에 배정할 잔차를 산출할 수 있다. 차이가 클 수도, 크지 않을 수도 있다.

12.3 CQA 곡선의 특성

상태-품질-매력 (CQA) 곡선은 특정 지역 안에서 부동산 가치가 매력, 상태, 시장 동학에 따라 어떻게 변동하는지를 이해하는 정교한 모형을 제공한다. 이 모형은 부동산 시장에 영향을 주는 경제적, 물리적 요인의 상호작용을 포착하며, 시장 부문별 감정평가의 미묘한 실상을 반영한다.

12.3.1 CQA 곡선 구성요소의 분석

1. 곡선 좌측의 지수적 감쇠:

   CQA 점수가 2 부근에서 0 으로 낮아짐에 따라 부동산은 급속한 노후화의 특징을 보인다. 곡선의 이 부분은 지수적 감쇠 함수로 잘 표현되며, 적절한 유지관리와 개보수 없이 부동산이 얼마나 빠르게 매력과 가치를 잃는지를 부각한다. 목재의 부식, 색의 바램, 금속의 녹과 같은 물리적 열화가 이러한 하락을 가속하며, 정기적으로 관리되지 않는 주택에서 특히 그러하다.

2. 중간 구간의 선형적 증가:

   시장의 대부분이 통상 이 범주에 들며, 점수 2 부근에서 8 까지 매력이 거의 선형으로 증가한다. 이 구간은 대체로 표준적으로 관리되고 있으나 반드시 특출한 품질이나 입지를 갖추지는 않은 부동산을 반영한다. 이는 주택의 광범위한 중간층을 대표하며, 표준적 특성과 전반적 상태를 근거로 경쟁적으로 가격이 매겨지는 주택시장의 대다수를 아우른다.

3. 곡선 우측의 지수적 증가:

   여기서는 점수 8 에서 10 까지 곡선이 지수적으로 상승하며, 시장의 매력이 높은 쪽 끝을 반영한다. 이 구간의 부동산은 흔히 더 부유한 이들의 것으로, 그들은 뛰어난 유지관리뿐 아니라

부동산의 매력과 가치를 크게 높이는 개선에도 투자한다. 지수적 증가는 그러한 고품질 부동산의 희소성과 부유한 매수인들 사이의 높은 수요를 반영한다. 그리고 가격이 낮은 주택이 매력 면에서는 매우 높게 평정될 수 있고, 크고 비싼 주택이 낮게 평정될 수도 있다는 점을 이해하라. 이러한 동학은 MARS 를 전문적으로 활용할 때 드러나며, 대부분의 감정평가사는 대개 이를 알아채지 못한다.

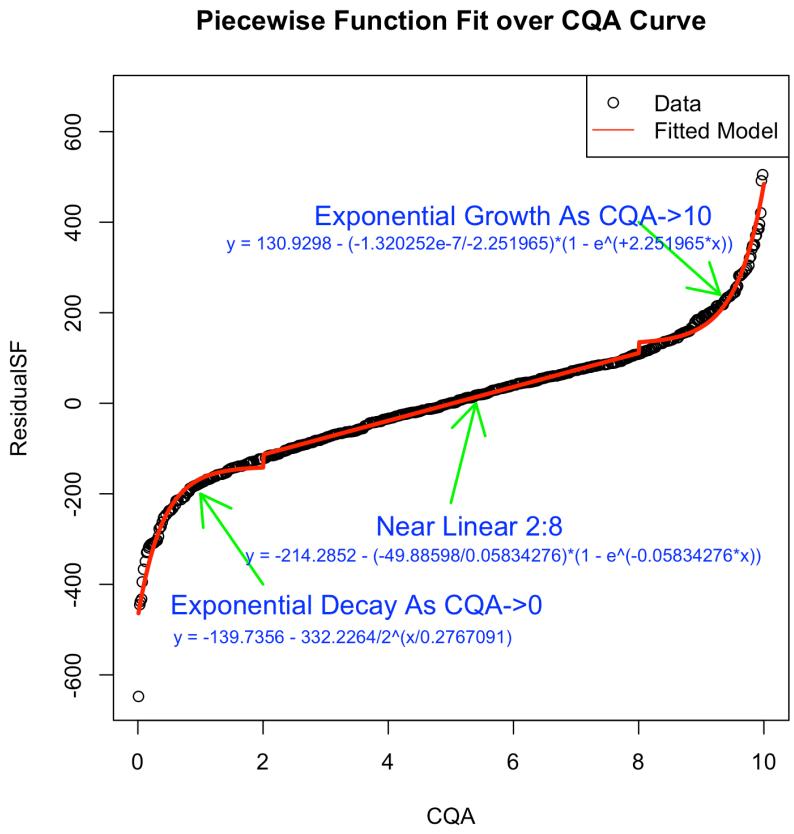
12.3.2 시장 동학과 사회적 함의

CQA 곡선은 지역마다 다른 상대적 부와 경제적 격차의 이야기를 은연중에 서술한다. San Mateo 카운티 같은 곳에서 Burlingame 과 Hillsborough, Woodside, Atherton 같은 더 부유한 근린지역 사이의 부동산 가치 편차는 상당한 사회경제적 분단을 보여준다. 이 격차는 부동산 유지관리에서 시장 평가에 이르는 모든 것에 영향을 주며, 궁극적으로 매수인 구성과 부동산 시장의 추세에도 영향을 미친다.

나아가, 비교적 부유한 사람이 스스로 편안하게 여기고 여러 이유로 살고 싶어 하는 중간 등급 근린지역에 고품질 주택을 마련한다는 일화는 부동산 경제의 또 다른 결정적 측면 — “과잉건축(overbuilt)” 부동산이라는 개념 — 을 부각한다. 그러한 부동산은 매우 매력적이지만, 더 높은 가격대의 시장 지역에서라면 받았을 값을 치를 수 있는 현지 매수인을 찾지 못할 수 있다.

CQA-잔차 곡선은 시장 지역마다 고유하므로, 분석을 할 때 간과해서는 안 된다.

그림 2: CQA-잔차 특성함수



12.4 대상 부동산의 잔차

비교사례와 달리 대상 부동산에는 객관적으로 확정된 잔차가 없으므로, 감정평가공학자 (VE) 가 이를 추정해야 한다. 이 추정은 잔차 제약 접근법 (RCA) 프로토콜에 어느 정도의 주관성을 들여오며, 개인의 판단이 결정적 역할을 하는 중대한 분기점이 된다.

상태-품질-매력 (CQA) 점수를 추정하기 위해 VE 는 순위화된 부동산들을 면밀히 분석하여 순위 안에서 식별 가능한 양상을 찾는다. 이 분석에는 각 부동산에 딸린 MLS 사진의 상세한 검토가 포함되는 경우가 많다. 600 건 같은 큰 데이터 집합에서 이를 남김없이 검토하기란 현실적이지 않다. 그 대신 VE 는 실무적 추론을 적용하여 전체 순위 안에서 대상 부동산의 있음직한 위치를 어림한다. 대대적인 수선이 필요한 부동산, 이른바 "픽서 (fixer)" 는 대체로 하위에 놓이고, 개보수와 매력이 뛰어난 부동산은 상위에 오르는 경향이 있다.

그런 다음 VE 는 더 좁혀 들어가는 접근을 취하여, 순위 안에서 "건너뛰며" 대상 부동산의 매력에 근접하는 비교사례를 찾는다. 범위가 좁혀질수록 VE 는 품질, 상태, 디자인 미감 (도장, 목공, 석공, 창호 등의 요소를 포함), 기능적 효용, 그 밖의 관련 특성 등에서 부동산들 사이의 유사성에 더욱 주의를 기울인다.

VE 의 판단이 근본적으로 시장이 형성한 순위에 이끌린다는 점을 이해하는 것이 결정적이다. 이로써 절차의 주관적 요소가 시장의 실상에 닻을 내리게 되며, 대상 부동산에 대한 공정하고 정확한 평가를 지향하게 된다.

12.5 RCA 가치 결론

12.5.1 일반적 고려사항

CQA 점수의 판정은 부동산의 추정 잔차값을 확정하는 데 결정적인데, 이 둘이 본질적으로 연결되어 있기 때문이다. CQA 없이도 부동산을 평가할 수는 있으나, 이 지표는 해당 부동산이 시장 지역의 다른 부동산과 견주어 어디에 서 있는지를 보여줌으로써 유용한 맥락을 제공한다. 예컨대 CQA 점수 3.2 는 그 부동산이 인근 비교 부동산의 32% 보다 매력이 크다는 뜻이다. 부여된 점수에 이의가 있다면, 이의 제기자는 순위표를 참조하여 사진 증거나 그 밖의 실질적 데이터에 근거해 제안하는 조정을 정당화해야 한다. 순위 체계가 견고하고 감정평가공학자 (VE) 가 그 틀 안에서 부동산을 정확히 자리매김했다면 유의미한 불일치가 생길 가능성은 낮다.

보고 단계에서는 통상 가치 결론을 뒷받침할 최근의 적절한 비교 거래사례 6 12 건을 선정한다. 이 비교사례와 대상 부동산의 차이를 분석하고 변수마다 그에 맞게 보정한다. 그런 다음 이 보정치를 비교사례의 순 거래가격에 적용하여 보정 후 거래가격을 도출한다. 수학적으로 밝힌 바와 같이, 이 보정 후 가격들은 대상 부동산의 추정 거래가격으로 수렴해야 한다.

12.5.2 검토, 감사, 이의

대상 부동산의 추정 잔차가 확정되면 모든 비교사례에 대한 잔차 보정치를 계산할 수 있다. 이 보정치는 VE 가 적절하다고 판단한 특정 CQA 변수에 맞추어 이후 수작업으로 분해할 때의 제약으로 작용한다. 이러한 상세 분해는 설명을 위한 것일 뿐 최종 평가액에는 영향을 주지 않는다는 점을 강조해 둘 필요가 있으며, 이는 증명 II 가 확인한 바다.

감정평가 가치 결론에 만족하지 못한 이용자의 이의와 관련하여, 전통적 감정평가사는 가격이 더 높거나 낮은 이런저런 부동산이 왜 이용자가 기대한 만큼 가치에 영향을 주지 않는지 설명하느라 시간을 허비하는 것이 보통이다. RCA 에서는 가능한 모든 비교 거래사례 (통상 100 600 건) 가 대체로 가격 결론과 정확히 같은 값으로 평가되므로, 임의의 부동산에 대한 보정치를 보여주는 데 품이 거의 들지 않는다. RCA 절차 전체는 대상 부동산의 잔차 점수 매기기를 제외하면 자동화되어 있고 객관적이다. 그리하여 RCA 의 또 다른 이점은 VE 의 주관적 판단이 지닌 불완전성에 대한 비판이 주로 그 한 영역에 국한된다는 것이며, 이는 효과적이고 효율적으로 관리할 수 있다. 또

한 R2 가 약 80% 라고 할 때, 대상 부동산 잔차 추정에서의 ± 10% 오차는 최종 가치 추정치의 ± 2% 오차에 해당한다.

많은 경우 대상 부동산이 잔차 순위표에 매우 잘 들어맞아 정확도가 ± 1% 이내일 것으로 기대할 수 있다.

추정 잔차와 대상 부동산의 MARS 추정 거래가격의 합이 잔차 제약 접근법 (RCA) 가치 결론이라 불리는 최종 가치 결론을 이룬다. 이 절차는 부동산 감정평가에 대한 포괄적이고 투명한 접근을 보장하여 평가의 정확성과 공정성에 대한 신뢰를 키운다.

13 결론

잔차 제약 접근법은 세 가지 실질적 기여를 갖는 틀이다. 첫째, 소수의 비교사례 집합이 아니라 시장 전체에 대해 모형을 적합시키는 데서 도출되는 구조적 반정박 (anti-anchoring) 성질 (증명 I). 둘째, 잔차 분해가 가치 결론을 흐트러뜨리지 않으면서 설명의 층위 역할을 할 수 있게 하는 불변성 결과 (증명 II 와 그에 이어지는 논평). 셋째, 지수-선형-지수의 특징적 형태로 여러 시장 지역에 걸쳐 되풀이하여 나타나는 것으로 보이는 경험적 규칙성 — CQA-잔차 특성곡선. 이 틀의 한계 또한 그에 못지않게 실질적이다. 잔차 분해는 불변이지만 식별적이지는 않고, CQA 위치 결정 단계는 아직 정식 평정자 간 신뢰도 연구로 검증되지 않은 VE 의 판단에 의존하며, 정확성에 관한 주장은 대안적 방법과의 벤치마킹을 기다리고 있다.

이 논문이 제시하는 것은 하나의 틀과 프로토콜이다. 아직 제시하지 못했으나 현재 진행 중인 더 큰 연구 프로그램이 제시하는 것은, 방어 가능한 기본값과 함께 그 프로토콜을 소프트웨어로 구현한 실행 가능하고 감사 추적이 가능한 파이프라인이다. 그 파이프라인 — MARS 기저함수 단계를 위한 CRAN 의 earthUI 패키지, 그 기저 위에서 정칙화된 구조 모형을 다루는 glmnetUI 패키지, 공간 평활을 위한 mgcvUI 패키지 — 은 이번 호의 관련 논문들이 다루는 주제이며, 방법으로서의 RCA 에서 산출물로서의 RCA 로 건너가는 다리다. 감정평가 규제에서의 교리-실무 간극 (Craytor, 2026, 이번 호) 은 그 다리가 왜 중요한지를 알려 준다. 감정평가의 산출물은 경제적으로 미래 지향적 위험 투입값으로 소비되며, 현재가치 의견을 감사 가능하고 재현 가능하게 만드는 틀은 그러한 미래 지향적 용도와 정직하게 대면하기 위한 선결조건이다. RCA 는 그 더 큰 프로젝트의 한 수 (手) 일 뿐이다. 프로젝트 전체가 아니며, 이 논문도 그렇게 주장하지 않는다.

참고문헌

Ambrose, Brent W. et al. (2025). “Do Appraiser and Borrower Race Affect Mortgage Collateral Valuation?” In: *Review of Finance*. DOI: 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, pp. 1–141.

Hastie, Trevor, Robert Tibshirani, and 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 and 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.

Kuhn, Max and Kjell Johnson (2016). *Applied Predictive Modeling*. Corrected at 5th printing. New York: Springer. ISBN: 978-1-4614-6848-6.

Milborrow, Stephen (2024a). *Earth: Multivariate Adaptive Regression Splines*. R Package. URL: https://CRAN.R-project.org/package=earth.

— (2024b). *Notes on the Earth Package*. Tech. rep. URL: http://www.milbo.org/doc/earth-notes.pdf.

— (2024c). *Variance Models in Earth*. Tech. rep. URL: http://www.milbo.org/doc/earth-varmod.pdf.

Perry, Andre M., Jonathan Rothwell, and David Harshbarger (2018). *The Devaluation of Assets in Black Neighborhoods*. Tech. rep. 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*. Tech. rep. Brookings Institution.

Property Appraisal and Valuation Equity Task Force (2022). *Action Plan to Advance Property Appraisal and Valuation Equity*. Tech. rep. U.S. Department of Housing and Urban Development.

Tzioumis, Konstantinos (2016). "Appraisers and Valuation Bias: An Empirical Analysis". In: *Real Estate Economics* 45.3, pp. 679–712. DOI: 10.1111/1540-6229.12133.

Williamson, Jake and Mark Palim (2022). *Appraising the Appraisal: A Closer Look at Divergent Appraisal Values for Black and White Borrowers Refinancing Their Home*. Tech. rep. Fannie Mae.

Zhang, Wengang (2020). *MARS Applications in Geotechnical Engineering Systems: Multi-Dimension with Big Data*. Singapore: Springer. ISBN: 978-981-13-7421-0.