

earthUI: earth(MARS) 패키지를 위한 대화형 인터페이스

한국어 번역본. 인공지능의 도움을 받아 번역하고 편집부가 검토했습니다. 영문판이 정본입니다.

William Bert Craytor

제 1 권 제 1 호 · 2026 년 여름호

발행일: 2026 년 7 월 1 일

DOI: [10.67330/7fj41b65](https://doi.org/10.67330/7fj41b65)

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

이 오픈 액세스 논문은 [크리에이티브 커먼즈 저작자표시 4.0 국제 라이선스 \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)의 조건에 따라 배포됩니다.

초록. earthUI 는 다변량 적응 회귀 스플라인 (MARS, multivariate adaptive regression splines) 을 구현한 R `earth` 패키지의 그래픽 사용자 인터페이스 (graphical user interface) 이다. 일반 예측 모형화, 부동산 감정평가, 시장지역 분석이라는 세 가지 용도 모드를 제공하며, 데이터 가져오기, 모형 구성, 적합, 그리고 `earth` 의 진단 정보와 그래픽 출력의 해석에 이르기까지 사용자를 안내한다. 이 논문은 earthUI 의 데이터 형식 요건, 모형화 작업 흐름, 출력 화면, 그리고 전체 기능 참조를 문서화한다.

키워드: MARS, 다변량 적응 회귀 스플라인, `earth`, R, 그래픽 사용자 인터페이스, 부동산 감정평가

발행인 주 (2026 년 7 월): 이 논문은 earthUI 0.10.0 을 다룬다. 인쇄 직전에 earthUI 0.11.0 이 자매 응용 프로그램인 `glmnetUI` 와 `mgcvUI`(이어지는 두 편의 논문에서 설명한다) 를 흡수하여 MARS, 엘라스틱 넷, GAM 작업 흐름을 하나의 응용 프로그램으로 통합하였다. 여기에 서 설명하는 인터페이스와 작업 흐름은 이번 통합의 영향을 받지 않는다.

목차

1 서론	3
1.1 earthUI 란 무엇인가?	3
1.2 하나의 패키지, 세 개의 루틴	4
1.3 세 가지 용도 모드	4
1.4 부동산 전용 기능	4
1.5 earthUI 와 vProlog 설치	5
1.6 시작하기	5
2 MLS 입력 데이터 요건	7
2.1 파일 형식과 구조	7
2.2 감정평가 모드에 필요한 열	7
2.3 특수 열 명명 규약	8
2.4 데이터 품질과 완전성	8
2.5 대상 부동산 행의 배치	8
3 General 용도 모드	9
3.1 개요	9
3.2 사이드바 작업 흐름	9
3.3 주 패널 탭	9
3.4 설정 지속성	10
3.5 다크 모드	10
3.6 로컬 및 지역 설정	10
3.6.1 지원 국가	11
3.6.2 용지 재정의	11
3.6.3 기본값 저장	11

4 Appraisal 모드	12
4.1 대상 부동산 행의 처리	12
4.2 기준일과 Sale Age	12
4.3 특수 열 지정	12
4.4 잔차 기여도 (회귀 특성 추정)	13
4.5 RCA 보정 개요	13
5 Market Area Analysis 모드	15
5.1 Appraisal 모드와의 차이	15
5.2 시장 모드를 쓸 때	15
6 변수 선택	16
6.1 목표 (반응) 변수	16
6.2 예측변수 표	16
6.3 자료형 감지와 재정의	16
6.4 Factor 및 Linear 플래그	17
6.5 특수 열 유형 참조	17
6.6 다중 목표변수	18
7 매개변수 선택	19
7.1 기본 매개변수	19
7.2 전진 단계	19
7.3 허용 상호작용	20
7.4 가지치기	20
7.5 교차검증	20
7.6 부분집합 필터링	20
7.7 권장 매개변수 값	21
7.7.1 전진 단계 매개변수	21
7.7.2 가지치기 매개변수	22
7.7.3 교차검증 매개변수	22
7.7.4 분산 모형	23
7.7.5 Earth 출력 파일	23
7.8 설정 기본값	23
8 Earth 모형 적합	24
8.1 Fit 버튼	24
8.2 적합 모달과 추적 출력	24
8.3 적합 로그	24
9 데이터 내려받기	25
9.1 출력 열	25
9.2 열 순서와 Excel 형식	25
9.3 CQA 점수	26
9.4 정렬과 대상 부동산 행	26
9.5 mgcvUI 자동 내보내기	26
9.6 glmnetUI 가져오기	26
9.6.1 주요 이점	27
9.6.2 중요: 가중치 열이 일치해야 한다	27
9.6.3 작업 흐름	27

10 RCA 계산과 내려받기	28
10.1 RCA 대화 상자 열기	28
10.2 CQA 점수 보간	28
10.3 출력 열	28
10.4 RCA Adjustments 탭	29
11 거래사례비교표	30
11.1 개요와 목적	30
11.2 비교사례 선택 모달	30
11.3 표 배치	30
11.4 매매가격, 양허, Net SP	31
11.5 묶음 행 (위치, 부지, 연령)	31
11.6 모형 변수 행	31
11.7 CQA Residual 및 잔차 특성 행	31
11.8 보정 매매가격 수식	32
11.9 시트 보호와 편집 가능 셀	32
11.10 Excel 에서 표 다루기	32
12 보고서 내려받기	33
12.1 Generate Quarto Report	33
12.2 Convert Quarto Report	33
12.3 보고서 내용	33
13 후처리 모듈: 엘라스틱 넷과 GAM	35
13.1 왜 보조 적합인가?	35
13.2 실행과 상태 공유	35
13.3 엘라스틱 넷 루틴	35
13.4 GAM 루틴	35
13.5 실무적 작업 흐름	35
14 데모 데이터셋: Appraisal_1.csv	37
14.1 설명	37
14.2 열	38
14.3 권장 빠른 시작	38
15 실험적 기능: Prolog 텍스트 추출과 도출 규칙	40
15.1 무엇을 하는가	40
15.2 요구 사항: SWI-Prolog 가 아니라 vProlog(Trealla)	40
15.3 활성화와 실행	41
15.4 도출 규칙	41
15.5 버전이 매겨진 활성 데이터 파일	41
15.6 프로그램 인터페이스와 함께 제공되는 파일	42
15.7 관련 실험: 기후 지역 사전정보	42
15.7.1 캘리포니아를 넘어: 관할권 팩과 시장지역 모듈	42

1 서론

1.1 earthUI 란 무엇인가?

earthUI 는 다변량 적응 회귀 스플라인 (MARS, multivariate adaptive regression splines) 을 구현한 R earth 패키지의 그래픽 사용자 인터페이스 (graphical user interface) 이다 (Friedman 1991).

모형화 엔진은 Stephen Milborrow 의 **earth** 패키지이며 (Milborrow 2024), Milborrow (n.d.[a]) 및 Milborrow (n.d.[b]) 에 상세히 문서화되어 있다. earthUI 는 로컬 Shiny 응용 프로그램으로 실행되며, 로그인도 서버도 계정도 없다. R 에서 실행한 뒤 데이터셋 (CSV 또는 Excel) 을 가져오고, 모형을 구성하여 대화형으로 적합시킨다.

이 응용 프로그램은 데이터 가져오기, 변수 구성, 모형 적합, 진단 그림, 변수 중요도, 모형 방정식, 그리고 HTML, Word, PDF 형식으로 내려받을 수 있는 보고서에 이르는 완전한 작업 흐름을 제공한다.

1.2 하나의 패키지, 세 개의 루틴

현재 릴리스 기준으로 earthUI 패키지는 두 개의 후처리 모듈 (Post Processing Module) 까지 포함하는데, 이전 릴리스에서는 glmnetUI 와 mgcvUI 라는 별도의 패키지였다. earthUI 를 설치하면 세 가지 루틴이 모두 설치되며, 각각은 자체 로컬 Shiny 응용 프로그램으로 실행된다:

- **earthUI::launch()** — 이 논문에서 설명하는 earth(MARS) 응용 프로그램으로, 포트 7878 을 사용한다. earth() 가 중심 감정평가 엔진이다.
- **earthUI::launch_glmnet()** — 엘라스틱 넷 후처리 모듈 (포트 7879) 로, earth 모형의 기저 함수 (basis function) 를 가져와 라쏘/엘라스틱 넷 정규화 아래에서 재추정한다 (아래 “glmnetUI Import” 참조).
- **earthUI::launch_mgcv()** — GAM 후처리 모듈 (포트 7880) 로, earth 모형의 매듭점 (knot) 위치를 mgcv 평활항의 출발점으로 사용한다 (아래 “mgcvUI Auto-Export” 참조).

세 루틴은 하나의 프로젝트 트리, 하나의 설정 데이터베이스, 그리고 지원 기반 구조의 사본 하나를 공유하므로, 한 응용 프로그램에서 연 프로젝트는 곧바로 다른 응용 프로그램에서도 사용할 수 있다. glmnet 과 mgcv 루틴은 서로 경쟁하는 작업 흐름이 아니라 earth 모형의 가치 결론을 뒷받침하는 증거이다. 의도된 방식은 먼저 earth 모형을 적합시키고 다듬은 다음, 보강이 필요할 때 같은 프로젝트에 대해 모듈을 실행하는 것이다.

1.3 세 가지 용도 모드

earthUI 는 작업을 프로젝트 단위로 조직한다. 프로젝트란 입력 파일, 출력물, 설정을 고정된 위치에 함께 보관하는 이름 붙은 개체이다. 모든 프로젝트는 생성 시 새 프로젝트 마법사에서 선택하는 **Purpose**(용도) 를 지닌다. 세 가지 모드가 있다:

- **General**(일반) — 모든 유형의 모집단이나 데이터셋에 대한 Earth 회귀분석. 기본 모드이다. 도메인 특화 기능 없이 전체 MARS 모형화 작업 흐름을 제공한다.
- **For Appraisal**(감정평가용) — 부동산 감정평가에 맞춘 Earth 회귀분석. 대상 부동산 처리, 특수 열 지정, 잔차 제약 접근법 (RCA, Residual Constraint Approach), 거래사례비교표 생성 등 단일 부동산 감정평가에 특화된 기능을 추가한다.
- **Market Area Analysis**(시장지역 분석) — 시장지역 연구에 맞춘 Earth 회귀분석. 단일 대상 부동산을 평가하는 대신 정의된 시장 내 부동산 집단을 분석하는 기능을 추가한다.

세 모드 모두에서 핵심 모형화 엔진은 동일하다. 사용자는 언제나 Earth(MARS) 모형을 적합시킨다. 용도 설정은 어떤 추가 도구와 인터페이스 요소를 사용할 수 있는지를 제어할 뿐이다.

용도는 프로젝트별로 고정된다. 프로젝트의 용도는 국가, 주, 카운티, 도시와 함께 프로젝트 생성 시점에 잠긴다. 다른 프로젝트로 전환하면 earthUI 는 깨끗한 기본 상태로 초기화되고 (가져온 데이터, 모형 결과, 탭, 변수 구성, earth 매개변수가 모두 지워진다), 이어서 해당 프로젝트의 용도에 저장된 구성을 복원한다.

1.4 부동산 전용 기능

For Appraisal 또는 **Market Area Analysis** 를 선택하면 earthUI 는 부동산 분석을 위해 설계된 여러 기능을 활성화한다:

- **특수 열 지정** — 각 예측변수에는 contract_date, dom, concessions, latitude, longitude, living_area, lot_size, actual_age, effective_age, area, site_dimensions,

`display_only` 같은 특수 역할을 부여할 수 있다. 이 지정은 적합, 출력, 거래사례비교표 생성 과정에서 해당 열이 어떻게 다루어지는지를 제어한다. 전체 참조는 제 6 장을 보라.

- **위도와 경도의 반올림** — 위도 또는 경도로 지정된 열은 과적합을 막기 위해 자동으로 소수점 셋째 자리까지 반올림된다.
- **Sale Age 열** — 어떤 열이 `contract_date` 로 지정되고 기준일 (Effective Date) 이 주어지면, `earthUI` 는 `sale_age` 열 (매매일과 기준일 사이의 일수) 을 계산하여 예측변수로 대체한다.
- **RCA 계산 (감정평가 모드 전용)** — 감정평가 모드에서는 모형을 적합시킨 뒤 `earthUI` 가 잔차 제약 접근법 (RCA) 출력을 계산할 수 있으며, 이는 비교사례별 보정치, 순보정/총보정 요약, 그리고 대상 부동산의 보정 매매가격을 산출한다.
- **거래사례비교표 (감정평가 모드 전용)** — 비교사례 보정, 잔차 특성 분해, 보정 매매가격에 대한 수식을 포함한 다중 시트 Excel 통합문서를 생성한다. 제 11 장을 보라.

팁: *General* 용도 모드는 금융, 과학, 공학, 부동산 등 어떤 데이터셋에도 적용된다. 감정평가 모드와 시장 모드는 부동산 전문가를 위한 편의 기능을 더할 뿐이다.

1.5 earthUI 와 vProlog 설치

`earthUI` 는 CRAN 에 공개되어 있으므로 표준 설치인 R 에서 한 줄이면 된다:

```
install.packages("earthUI")
```

이 논문은 현재 릴리스를 설명한다. 이 글을 읽는 시점에 CRAN 이 아직 이전 버전을 제공하고 있다면 개발 버전을 GitHub 에서 설치할 수 있다 (패키지가 저장소의 `pkg/` 폴더에 있으므로 `subdir` 에 유의하라):

```
remotes::install_github("wcraytor/earthUI", subdir = "pkg")
```

vProlog(선택 사항). 실험적인 비고 처리 기능 (이 논문 끝부분의 “Prolog 텍스트 추출” 참조) 은 `vProlog` 패키지, 즉 자체 완결형 내장 Prolog 엔진을 사용한다. 이는 의도적으로 필수 의존성이 아니며, `earthUI` 는 `vProlog` 없이도 설치되고 실행되며 Prolog 기능만 `vProlog` 가 있을 때까지 비활성 상태로 남는다. `vProlog` 는 저자의 `r-universe` 저장소를 통해 배포되며, Windows 와 macOS 용 사전 빌드 바이너리를 제공한다 (컴파일러 도구 모음이 필요 없다):

```
install.packages("vProlog",
  repos = c("https://wcraytor.r-universe.dev",
    getOption("repos")))
```

시스템에 Prolog(SWI-Prolog 등) 를 별도로 설치할 필요는 없다. Trealla 엔진이 패키지 자체에 컴파일되어 들어 있다. 이 논문의 나머지 내용은 모두 `vProlog` 없이도 동작한다.

1.6 시작하기

`earthUI` 를 사용하려면:

1. 위에서 설명한 대로 CRAN 에서 패키지를 설치한다.
2. **응용 프로그램을 실행한다:** R 에서 `earthUI::launch()` 를 실행하거나, 명령줄에서 `Rscript -e 'earthUI::launch()'` 를 실행한다. 응용 프로그램은 웹 브라우저의 포트 7878 에서 열린다. 브라우저에서 `http://localhost:7878` 로 직접 접속할 수도 있다. (후처리 모듈도 같은 방식으로 시작한다: `launch_glmnet()` 은 포트 7879, `launch_mgcv()` 는 포트 7880 을 사용한다.)
3. 사이드바의 1 절에서 프로젝트를 만들거나 연다. 새 프로젝트 마법사는 프로젝트 이름, Purpose(General, For Appraisal, Market Area Analysis 중 하나), 위치 (국가, 주, 카운티, 도시), 그리고 선택적으로 최초 데이터 파일을 받아 프로젝트의 `in/` 폴더로 복사한다.

4. 2 절에서 데이터를 가져온다. 2 절에는 활성 프로젝트의 in/ 폴더에 있는 CSV 및 Excel 파일이 나열된다. Finder 나 탐색기로 직접 추가한 파일은 Refresh 를 누르면 나타난다.
5. 변수를 구성한다 — 목표변수와 예측변수를 선택하고 자료형을 설정하며 특수 열 역할을 지정한다.
6. **Earth** 매개변수를 설정한다 — degree, nprune, 부분집합 필터를 비롯한 **earth()** 인수들.
7. 모형을 적합시킨다 — “Fit Earth Model” 을 클릭하고 주 패널에서 결과를 검토한다.
8. 내보낸다 — 예측값을 Excel 로 내려받거나, Quarto 보고서 묶음을 생성하거나, (감정평가 모드에서는) RCA 보정치와 거래사례비교표를 계산한다.

구성은 3 절과 4 절의 “Save current as default” 버튼을 통해 프로젝트와 용도를 키로 하여 프로젝트 데이터베이스에 서버 측으로 저장된다 (제 3 장 “설정 지속성” 참조).

2 MLS 입력 데이터 요건

부동산 감정평가 및 시장 분석 작업 흐름에서 입력 데이터는 대개 종합 부동산 정보 시스템 (MLS, Multiple Listing Service) 내보내기에서 나온다. 이 장에서는 기대되는 파일 구조와 earthUI 가 사용할 수 있는 열을 설명한다.

2.1 파일 형식과 구조

earthUI 는 CSV 와 Excel(.xlsx, .xls) 파일을 받는다. 가져오는 시점에 열 이름은 자동으로 snake_case 로 변환된다. 예를 들어 “Living SqFt” 는 living_sqft 로, “Contract Date” 는 contract_date 로, “Sale Price” 는 sale_price 로 바뀐다. 이 정규화 덕분에 작업 흐름 전체에서 열 참조가 일관되게 유지된다. 가져오기에 사용되는 CSV 구분자와 소수점 기호는 로컬 설정에 따라 결정된다 (제 3 장 “로컬 및 지역 설정” 참조).

데이터 파일은 부동산 한 건당 한 행, 속성 하나당 한 열로 이루어진 평면 표여야 한다. 파일의 첫 행에는 반드시 열 머리글이 있어야 한다.

2.2 감정평가 모드에 필요한 열

earthUI 는 어떤 열 구성에서도 동작하지만, 전체 감정평가 작업 흐름 (RCA 보정 + 거래사례비교표) 은 MLS 내보내기에 다음 열들이 있을 때 더 잘 작동한다:

열	특수 유형	용도
Sale Price	(목표변수)	모형의 반응변수
Contract Date	contract_date	sale_age(기준일로부터의 일수) 계산에 사용
Listing Date	listing_date	DOM 열이 없을 때 계약일과 함께 DOM 계산에 사용
Days on Market	dom	시장 노출 일수. 거래사례비교표에 표시
Concessions	concessions	매매 양허. 거래사례비교표에서 Net SP = Sale Price - Concessions
Living Area (SF)	living_area	제곱피트당 잔차 (residual_sf, cqa_sf) 사용 가능
Lot Size	lot_size	거래사례비교표의 “Site Size Dimensions” 행에 묶임
Site Dimensions	site_dimensions	거래사례비교표에서 부지 면적과 함께 묶임
Latitude	latitude	소수점 셋째 자리로 반올림. 근접성 계산과 Location 그룹에 사용
Longitude	longitude	소수점 셋째 자리로 반올림. 근접성 계산과 Location 그룹에 사용
Area ID	area	거래사례비교표의 “Loc: Long Lat Area” 행에 묶임
Actual Age	actual_age	거래사례비교표의 “Actual Age Effective Age” 행에 묶임
Effective Age	effective_age	거래사례비교표에서 실제 연령과 함께 묶임
Parcel Number	(없음)	거래사례비교표의 APN 행에 표시
Listing ID	(없음)	거래사례비교표의 APN 행에 표시
Address	display_only	거래사례비교표의 Address 행에 표시. 모형에서는 제외

스프레드시트의 열 이름은 외국어여도 무방하다. “특수” 이름이 영어인 것은 R 프로그램이 해당 열을 특별하게 처리할 수 있도록 하기 위함이다. 그 밖의 경우 주어진 열 이름이 회귀 모형, 그래프, 그리고 (감정평가를 수행한다면) 중간 거래사례비교표에 그대로 나타난다.

모든 열이 필수인 것은 아니다. earthUI 는 상황에 맞추어 동작한다. 어떤 열이 없으면 그에 대응하는 기능이 단순히 생략된다. 예를 들어 concessions 열이 지정되지 않았다면 거래사례비교표의 Net SP 행에는 양허 공제 없이 매매가격이 표시된다. 다만 부동산 가격 모형에서 받아들일 만한 적합도를 얻으려면 다음 열들이 매우 권장된다:

1. **Sale Age** — 계약 매매일과 감정평가 또는 분석의 기준일 사이의 일수. 여러 해에 걸친 매매 이력을 사용할 때, 특히 5 년을 넘는 기간에 대해서는 `sale_age` 가 매매가격 추정에서 중심적 역할을 하는 경우가 많다. 실제로 이 변수가 워낙 중요해서 그것이 없으면 earthUI 가 아예 모형을 만들어 내지 못하는 경우도 흔하다.
2. **Living Area** — “Living Sqft”, “GLA”(총 주거 면적) 등 여러 이름으로 불린다. 매매가격을 결정하는 또 하나의 주요 요인이다.
3. **Total Bath Count** — 전체, 4 분의 1, 절반, 4 분의 3 욕실을 모두 합한 수. 예를 들어 전체 욕실 둘과 반욕실 하나는 2.5 가 된다.
4. **Garage Bays** 또는 **Garage Area** — 차고 주차 칸 수 또는 차고 면적.
5. **Lot Size** — 부동산의 토지 면적. 보통 제곱피트나 에이커 단위이다.
6. **Longitude, Latitude**, 그리고 가능하다면 **Area ID**. 이들에 대한 보정은 거래사례비교 표에서 하나의 Location 보정으로 통합된다.

2.3 특수 열 명명 규약

earthUI 는 열 이름이 아니라 특수 유형 지정으로 열을 식별한다. MLS 내보내기에서 열 이름은 원하는 대로 붙여도 되며, 중요한 것은 변수 구성 표 (제 6 장) 에서 올바른 특수 유형을 부여하는 것이다.

예를 들어 MLS 가 주거 면적을 “GLA”, “Living SqFt”, “liv_area”, “gross_living_area” 중 어느 이름으로 내보내더라도, 가져오기 (그 과정에서 snake_case 가 된다) 이후 Special 드롭다운에서 `living_area` 로 지정하기만 하면 된다. 그러면 earthUI 는 원래 이름과 무관하게 그 열을 제곱피트당 잔차 계산과 거래사례비교표 묶음에 사용한다.

2.4 데이터 품질과 완전성

- **결측값 (NA)**: 예측변수나 목표변수 열에 NA 가 있는 행은 적합 전에 자동으로 제거된다. 내부적으로 `na.action` 은 항상 `na.fail` 로 설정되며, NA 제거는 `earth()` 호출 이전에 처리된다.
- **날짜 열**: 날짜는 R 이 해석할 수 있는 형식이어야 한다 (예: 2025-06-15, 06/15/2025, June 15, 2025). earthUI 는 값의 80% 이상이 성공적으로 해석되면 날짜 열로 자동 감지한다. 두 자리 연도 (예: 06/15/25) 는 가장 가까운 연도로 읽히므로 최근 MLS 내보내기는 올바르게 해석된다. 네 자리 연도는 항상 문자 그대로 받아들여진다.
- **숫자 열**: 매매가격, 주거 면적, 부지 면적, 양허 등의 필드는 숫자여야 한다. MLS 가 통화 기호나 천 단위 구분자와 함께 가격을 내보낸다면 (예: “\$350,000” 또는 “350.000,00”), 가져오기 전에 정리해야 할 수 있다.
- **요인 열**: 지역 ID, 양식, 상태 같은 범주형 변수는 관리 가능한 수의 고유값을 가져야 한다. earthUI 는 요인을 자동 감지하지만 감지 결과를 사용자가 재정의할 수 있다.

팁: 가져오기 후 예측변수 표의 T/F/NA 열을 확인하라. 각 열의 참/거짓/NA 개수를 보여 준다. 결측값이 30% 를 넘는 열은 빨간색으로 표시된다. NA 가 많은 열은 제외하거나 가져오기 전에 데이터를 정리하는 것을 고려하라.

2.5 대상 부동산 행의 배치

Appraisal 모드에서는 1 행이 반드시 대상 부동산이어야 한다. 나머지 행은 모두 비교 매매사례이다. 대상 부동산의 매매가격은 비워 두거나 (NA) 어떤 값을 넣어도 되며, earthUI 는 적합 시 이를 무조건 NA 로 취급한다.

Market Area Analysis 모드에서는 1 행에 대상 부동산을 두는 것이 선택 사항이며, 어떤 행도 자동으로 제외되지 않는다. 대상 행 (또는 다른 어떤 행) 을 적합에서 제외하려면 지정된 가중치 열에 0 값을 주거나 부분집합 필터를 사용하라.

General 모드에서는 특별한 행 처리가 없으며 모든 행이 동등하게 취급된다.

3 General 용도 모드

3.1 개요

General Purpose 모드는 earthUI 를 실행했을 때의 기본 모드이다. 부동산에 국한되지 않고 어떤 데이터셋에 대해서도 완전한 MARS 모형화 작업 흐름을 제공한다. 과학 데이터, 재무 분석, 공학 연구를 비롯하여 변수 사이의 비선형 관계와 상호작용을 탐색하려는 어떤 회귀 문제에도 earthUI 를 쓸 수 있다.

General 모드에서는 인터페이스가 부동산 전용 기능 (특수 열, sale age, 좌표 반올림, RCA) 을 생략한다. 사이드바는 변수 선택, 매개변수 구성, 모형 적합, 내보내기에 집중하도록 간소화된다.

3.2 사이드바 작업 흐름

사이드바는 번호가 매겨진 접이식 절들로 조직되어 있어 프로젝트 선택부터 내보내기까지 안내한다:

1. Project — 기존 프로젝트를 고르거나 새 프로젝트 마법사로 새로 만든다. 마법사는 용도, 국가, 행정 위치를 받으며 이는 생성 후 모두 잠긴다. 활성 프로젝트가 입력 파일을 읽어 오는 위치와 모든 출력이 기록되는 위치를 결정하며, 별도의 출력 폴더 입력란은 없다.

2. Import Data — 활성 프로젝트의 in/ 폴더에서 CSV 또는 Excel 파일을 고른다. Refresh 링크는 직접 추가한 파일을 찾기 위해 폴더를 다시 훑는다. 시트가 여럿인 Excel 파일에는 시트 선택기가 나타난다. 열 이름은 자동으로 snake_case 로 변환된다.

3. Variable Configuration — 목표변수 선택기 (다중 목표변수 지원), 기준일 (Effective Date), 그리고 Include, Factor, Linear, Latent 확인란이 있는 예측변수 표. 자세한 내용은 제 6 장을 보라.

4. Earth Call Parameters — earth() 함수의 모든 인수: degree, penalty, nk, 가지치기 방법, 교차검증 등. 전체 매개변수 참조는 제 7 장을 보라.

Execute Prolog Processing — 4 절과 5 절 사이에 있는 선택적이고 번호가 없는 절로, 설정에서 실험적 Prolog 기능을 켜지 않으면 비활성 상태이다. 15장을 보라.

5. Fit Earth Model — 난수 시드 입력란과 모형을 비동기적으로 실행하는 버튼. 적합에 관한 자세한 내용은 제 8 장을 보라.

6. Download Estimated Target Variable(s) & Residuals — (감정평가 및 시장 모드에서는 “Download Estimated Sale Prices & Residuals” 라는 제목이 된다) 예측값, 잔차, CQA 점수, g 함수별 기여도를 Excel 파일로 내보낸다. 제 9 장을 보라.

7-8. RCA Adjustments and Sales Grid — 감정평가 모드 전용이다. 제 10 장과 제 11 장을 보라.

9. Generate Quarto Report — 자체 완결형 Quarto 보고서 묶음 (.qmd 원본과 자산) 을 프로젝트의 출력 폴더에 기록한다. 감정평가 모드가 아닐 때는 번호가 7 이다. 제 12 장을 보라.

10. Convert Quarto Report — 임의의 .qmd 파일을 HTML, Word, PDF 로 렌더링한다. 제 12 장을 보라.

3.3 주 패널 탭

적합 후 주 패널에는 열 개의 탭이 제공된다:

탭	내용
Data	가져온 데이터의 미리보기. 감정평가 모드에서는 대상 부동산 표와 비교 매매사례 표로 나뉜다.
Equation	적합된 모형 방정식을 LaTeX/MathJax 로 표시하며, 각 기저 함수와 그 계수를 보여 준다.
Summary	주요 지표 (R^2 , GRSq, GCV, RSS, CV R^2) 와 계수 표.
Variable Importance	예측변수 중요도 점수의 막대그래프와 순위 표.
Contribution	g 함수 기여도 그림. 드롭다운에서 g 함수를 선택한다. 일변량 예측변수는 선 그림으로, 이변량 상호작용 ($\text{degree} \geq 2$) 은 3 차원 투시도와 등고선 그림으로 표시된다.
Correlation	예측변수 상관관계 히트맵.
Diagnostics	세 가지 그림: 잔차 대 적합값, 정규 Q-Q, 실제값 대 예측값.
RCA Adjustments	(감정평가 전용) 비교사례 전반의 Residual Adj %, Net Adj %, Gross Adj % 히스토그램. 7 단계 (Calculate RCA Adjustments) 가 완료된 뒤에 채워진다.
Residual Contribution	모형이 사용하지 않은 열들에 대해 모형 잔차를 결합 회귀시켜, 드물거나 제외된 특성의 시장 기여도를 추정한다 (subsection 4.4 참조).
ANOVA	각 기저 함수의 RSS 기여도를 보여 주는 ANOVA 분해 표.
Earth Output	가지치기 단계 세부 정보를 포함한 ‘earth::summary()’의 원시 텍스트 출력.

3.4 설정 지속성

earthUI 는 프로젝트와 용도를 키로 하여 프로젝트 데이터베이스 (regProj 루트의 projects.sqlite) 에 구성을 서버 측으로 저장한다. 저장은 버튼 방식이다. 3 절의 “Save current as default” 버튼은 목표변수, 기준일, 예측변수 구성을 저장하고, 4 절의 버튼은 earth 호출 매개변수와 Allowed Interactions 행렬을 저장한다. 각 버튼은 자기 몫만 갱신하므로 서로 덮어쓰지 않으며, **적합은 자동 저장하지 않으므로** 저장된 기본값을 덮어쓸 걱정 없이 자유롭게 실험할 수 있다. 한 프로젝트의 여러 입력 파일 (작은 시험 추출본, 전체 데이터셋) 은 하나의 구성을 공유하고, 설정이 프로젝트 트리 안에 있으므로 regProj 폴더를 동기화하거나 백업하거나 공유할 때 설정도 함께 따라간다.

3.5 다크 모드

오른쪽 위 모서리의 달/해 아이콘을 클릭하면 밝은 테마와 어두운 테마가 전환된다. 테마 설정은 로컬 저장소에 보관되어 세션이 바뀌어도 유지된다.

3.6 로캘 및 지역 설정

earthUI 는 국가 기반 로캘 체계를 통해 국제적인 숫자, 날짜, CSV 형식 관행을 지원한다. 상단 메뉴 막대의 **Settings**(톱니바퀴) 메뉴에 있는 **Country** 드롭다운으로 지원되는 31 개 국가의 사전 설정을 선택한다. 각 사전 설정은 다음을 구성한다:

- **CSV 구분자** — 미국/영국/일본은 쉼표 (,), 쉼표를 소수점 기호로 쓰는 유럽 대부분 지역은 세미콜론 (;).
- **소수점 기호** — 마침표 (.) 또는 쉼표 (,).
- **천 단위 구분자** — 쉼표 (미국/영국/일본), 마침표 (독일/이탈리아/스페인), 공백 (핀란드/프랑스/폴란드/발트 3 국/우크라이나/러시아), 또는 아포스트로피 (스위스).
- **날짜 형식** — MM/DD/YYYY(미국), DD/MM/YYYY(유럽 대부분), YYYY-MM-DD(스웨덴/리투아니아/일본/캐나다).
- **용지 크기** — Letter(미국/캐나다/멕시코) 또는 A4(그 외 모든 지역).

3.6.1 지원 국가

국가	CSV 구분자	소수점	천 단위	날짜	용지	
미국, 캐나다 (csv)	,	.	,	MDY/YMD	Letter	
영국, 아일랜드, 호주, 뉴질랜드	,	.	,	DMY	A4	
독일, 오스트리아, 벨기에, 네덜란드, 이탈리아, 스페인, 포르투갈, 덴마크, 브라질	;	,	.	DMY	A4	
프랑스, 폴란드, 체코, 노르웨이, 라트비아, 에스토니아	;	,	(공백)	DMY	A4	
핀란드, 스웨덴	;	,	(공백)	DMY/YMD	A4	
스위스	;	,	'	DMY	A4	
우크라이나, 러시아	;	,	(공백)	DMY	A4	
튀르키예	;	,	.	DMY	A4	
일본, 대한민국	,	.	,	YMD	A4	
멕시코	,	.	,	DMY	Letter	
리투아니아	;	,	(공백)	YMD	A4	

3.6.2 용지 재정의

국가 선택기 아래의 **Paper** 드롭다운 (Letter 또는 A4) 을 쓰면 국가를 바꾸지 않고도 보고서 페이지 크기를 재정의할 수 있다. CSV 구분자, 소수점 기호, 날짜 형식 순서는 선택한 국가의 사전 설정을 따르며, 국가를 바꾸면 용지 재정의는 그 국가의 기본값으로 초기화된다.

3.6.3 기본값 저장

Settings 메뉴에서 **Save** 를 클릭하면 로컬 설정이 전역으로 저장된다. 이 기본값은 어떤 프로젝트를 열든 이후 모든 세션에 적용된다. 프로젝트별 설정 (목표변수, 예측변수, 매개변수) 은 프로젝트 데이터베이스에 따로 저장되며, 로컬 기본값은 사용자 수준 SQLite 데이터베이스를 통해 모든 프로젝트에 걸쳐 유지된다. 이 두 층 구조는 여러 나라의 데이터를 다루는 감사 법인 같은 조직을 염두에 둔 것이다. 가장 자주 쓰는 국가를 기본값으로 정해 두고 필요할 때 조정하면 된다.

팁: 그림 축과 기울기 레이블의 숫자 형식은 로컬에 맞추어 자동으로 적응한다. 독일 로컬은 천 단위에 마침표를 쓰고 (200.000), 핀란드는 공백을 (200 000), 스위스는 아포스트로피를 (200'000) 쓴다. 통화 기호는 표시되지 않는다. earthUI 는 통화 중립적이다.

4 Appraisal 모드

Purpose 로 **For Appraisal** 을 선택하면 earthUI 는 단일 부동산 감정평가에 맞게 스스로를 구성한다. 제 3 장에서 설명한 모든 기능은 그대로 사용할 수 있으며, 이 장에서는 감정평가 전용 추가 사항만 다룬다.

4.1 대상 부동산 행의 처리

감정평가 모드에서는 데이터셋의 1 행이 대상 부동산이고 나머지 행은 모두 비교 매매사례이다. 입력 파일을 그에 맞게 구성해야 한다 (제 2 장 참조). 대상 부동산의 매매가격은 비워 두거나 어떤 값을 넣어도 되며, earthUI 는 적합 시 자동으로 NA 로 취급한다.

가져오기 후 Data 탭은 두 부분, 즉 **Subject Property**(1 행) 와 **Comparable Sales**(2 행 이후) 로 나뉜다. 1 행은 항상 모형 적합에서 제외되며, “Skipping row 1 (subject). Fitting on N rows.” 라는 알림이 이를 확인해 준다. 적합 후에도 모형은 대상 행에 대한 예측값을 생성하며, 출력에서 `est_<target>` 로 표시된다.

4.2 기준일과 Sale Age

Effective Date(기준일) 입력란은 모든 용도 모드의 변수 구성 절에 나타나며 (기본값은 오늘 날짜) — 감정평가뿐 아니라 일반 분석과 시장 분석에도 기준일이 있다 — 실제로 사용되는 것은 감정평가 모드와 시장 모드뿐이다. 이들 모드에서 Special 열 드롭다운으로 어떤 열을 `contract_date` 로 지정하면, earthUI 는 `sale_age` 열, 즉 각 매매의 계약일과 기준일 사이의 정수 일수를 계산한다. 이 열이 원래의 날짜 열을 예측변수로 대체한다.

계약일을 지정한 뒤 처음 Fit 을 클릭하면 earthUI 가 `sale_age` 열을 만들고, 이를 포함시키려면 Fit 을 한 번 더 클릭하라고 알려 준다.

4.3 특수 열 지정

감정평가 모드와 시장 모드에서는 변수 구성 표의 각 예측변수마다 **Special** 드롭다운이 나타난다. 특수 유형의 전체 목록과 그 효과는 다음과 같다:

특수 유형	효과
no	기본값 — 특별한 처리 없음
contract_date	기준일로부터 sale_age 를 자동 계산하도록 함
listing_date	dom 열이 없을 때 계약일과 함께 DOM 계산에 사용
dom	시장 노출 일수 열을 식별. 거래사례비교표에 표시
concessions	매매 양허를 식별. 거래사례비교표의 Net SP 수식 (Sale Price – Concessions) 에 사용
latitude	값을 소수점 셋째 자리로 반올림. 하버사인 근접성과 거래사례비교표의 Location 그룹에 사용
longitude	값을 소수점 셋째 자리로 반올림. 하버사인 근접성과 거래사례비교표의 Location 그룹에 사용
area	거래사례비교표의 "Loc: Long Lat Area" 행에 묶임
living_area	내려받기 출력에서 제곱피트당 잔차 (residual_sf, cqa_sf) 를 사용 가능하게 함
lot_size	거래사례비교표의 "Site Size Dimensions" 행에 묶임
site_dimensions	거래사례비교표에서 부지 면적과 함께 묶임
actual_age	거래사례비교표의 "Actual Age Effective Age" 행에 묶임
effective_age	거래사례비교표에서 실제 연령과 함께 묶임
display_only	내보내기에는 포함되지만 모형 적합에서는 제외되는 열 (여럿 지정 가능)
remarks	(실험적) 선택적 Prolog 단계가 특정 열로 구문 분석하는 자유 텍스트 비고 열 (여럿 지정 가능). 15장 참조
list	(실험적) 선택적 Prolog 단계가 원-핫 TRUE/FALSE 열로 분할하는 선평 구분 열 (여럿 지정 가능). 15장 참조

특수 유형당 열은 하나만 허용된다 (display_only, remarks, list 는 예외). 어떤 특수 유형을 두 번째 열에 부여하면 첫 번째 열에서 자동으로 해제된다. 변수 이름 옆에는 부여된 특수 유형을 보여 주는 작은 파란색 배지가 나타난다.

4.4 잔차 기여도 (회귀 특성 추정)

회귀 경로에 담기에는 너무 드문 특성 — 몇 건의 매매에만 나타나는 작업실 같은 것 — 은 RCA 잔차 층에 속한다. **Residual Contribution** 탭은 모형이 전혀 사용하지 않은 열들에 대해 적합한 모형의 잔차를 보통최소제곱으로 회귀시켜 그 시장 기여도를 추정한다. 여러 열을 선택하면 결합하여 추정하므로, 서로 상관된 편의 시설 (예: outbuilding_sqft 와 pr_has_workshop) 이 중복 계산되지 않고 기여를 나누어 갖는다. 이진 플래그는 총액 추정치를 내고 플래그가 붙은 각 매매의 잔차를 증거로 열거하며, 숫자 열은 단위당 기여도를 낸다. 보고되는 표준오차는 1 표준오차이다 (통상적인 유의성 기준으로는 $t \geq 2$). 관측치가 열 개 미만인 항목에는 생성되는 작업 파일 메모에 자동으로 CAUTION 이 붙는데, 이 메모는 방법, 추정치, 유의 사항을 한 문단으로 정리해 보고서에 쓸 수 있게 한다. 후보 목록에서 제외되는 것은 모형이 실제로 사용한 예측변수뿐이다. 적합에 제공되었으나 전진 단계에서 선택되지 않은 회귀 플래그는 여전히 후보로 남는데, 모형이 그 가치를 전혀 흡수하지 않았기 때문이다.

4.5 RCA 보정 개요

Calculate RCA Adjustments & Download 버튼 (사이드바 7 절, 감정평가 모드에서 적합 후에만 표시) 은 대상 부동산 대비 각 비교사례의 시장 도출 보정치를 계산한다. 전체 RCA 작업 흐름은 제 10 장에서 설명한다.

RCA 보정치를 계산한 뒤에는 **Generate Sales Grid & Download** 버튼 (사이드바 8 절) 을 사용할 수 있게 된다. 거래사례비교표 작업 흐름은 제 11 장에서 설명한다.

답: 대상 부동산에 부여하는 *CQA* 점수는 잔차 분포 가운데 얼마만큼을 대상 부동산에 귀속시킬지를 결정한다. 5.00 점은 대상 부동산을 비교사례의 중앙값에 놓는다.

5 Market Area Analysis 모드

Purpose 로 **Market Area Analysis** 를 선택하면 earthUI 는 감정평가 모드와 동일한 부동산 전용 기능 (특수 열, sale age, 좌표 반올림) 을 제공하되, 단일 대상 부동산을 평가하기보다 부동산 집단을 분석하는 쪽으로 방향을 맞춘다.

5.1 Appraisal 모드와의 차이

- 대상 부동산 자동 제외 없음 — 감정평가 모드와 달리 시장 모드는 1 행을 적합에서 자동으로 제외하지 않는다. 모든 행이 포함된다. 대상 행을 적합에서 빼려면 지정된 가중치 열에 0 값을 주거나 (데모 데이터셋처럼) 부분집합 필터를 사용하라.
- **RCA** 및 거래사례비교표 절 없음 — “Calculate RCA Adjustments & Download” 와 “Generate Sales Grid & Download” 단계는 제공되지 않는다. 시장 모드는 비교사례별 보정이 아니라 모형 적합과 출력에 집중한다.
- **사이드바 번호** — RCA 와 거래사례비교표 단계가 없으므로 Generate Quarto Report 절의 번호가 9 가 아니라 7 이 된다 (Convert Quarto Report 절은 그대로 10 이다).

5.2 시장 모드를 쓸 때

Market Area Analysis 모드는 다음과 같은 경우에 적합하다:

- 가치 결정 요인을 파악하기 위해 근린지역이나 시장지역의 회귀 모형을 구축할 때
- 면적, 연령, 부지 면적, 위치 같은 변수가 부동산 집단 전반의 매매가격에 어떤 영향을 주는지 분석할 때
- 시장 상황 분석이나 근린지역 회귀를 뒷받침할 자료를 준비할 때
- 1 행에 대상 부동산이 들어 있는 데이터셋을 다루면서 그것을 적합에 포함할지 제외할지 선택하고 싶을 때

팁: 시장 모드는 특수 열 기능 (sale age, 좌표 반올림) 은 쓰고 싶지만 RCA 보정 작업 흐름은 필요 없는 일반적인 부동산 회귀분석에도 유용하다.

6 변수 선택

사이드바 3 절 — **Variable Configuration** — 은 어떤 열이 모형에 참여하고 그 열들이 어떻게 다루어지는지를 정하는 곳이다.

6.1 목표 (반응) 변수

3 절 상단의 **Target (response) variable(s)** 드롭다운에는 데이터셋의 모든 열이 나열된다. 표준적인 단일 반응 모형에는 열 하나를, 다중 반응 모형에는 여러 열을 선택한다. 목표변수를 여럿 선택하면 모형은 `earth(cbind(y1, y2, ...) ~ .)` 를 사용하여 모든 반응변수를 동시에 적합시킨다.

목표변수로 선택된 열은 예측변수 목록에서 자동으로 제외된다.

6.2 예측변수 표

목표변수 선택기 아래에는 남은 모든 열을 다음 필드와 함께 나열하는 표가 있다:

열	설명
Variable	열 이름 (잘린 경우 전체 이름이 툴팁에 표시) 과 그 앞의 $\times/+$ 토글 (아래 참조). 감정평가/시장 모드에서는 파란색 배지가 부여된 특수 유형을 보여 준다.
Type	자료형 드롭다운: <code>numeric, integer, character, logical, factor, Date, POSIXct</code>
Include	확인란 — 이 열을 모형의 예측변수로 포함한다
Factor	확인란 — 이 열을 범주형 변수로 취급한다
Linear	확인란 — 선형 향으로만 들어가도록 강제한다 (힌지 함수 없음)
Latent	확인란 — 이 열을 모형에서 완전히 배제한다 (<code>earth</code> 와 <code>glmnetUI</code> / <code>mgcvUI</code> 로 내보내는 데이터 모두에서 제외), 잠재변수 작업을 위해 남겨 둔다
Special	드롭다운 (감정평가/시장 전용) — 아래 특수 열 유형 참조
T/F/NA	열마다 세 개의 개수: 논리형은 TRUE/FALSE/NA, 숫자형은 $\neq 0/=0/NA$. 결측값이 30% 를 넘으면 빨간색으로 표시된다

표 위의 안내 줄은 약어를 설명한다: “Type = column data type, Inc = include as predictor, Factor = treat as categorical, Linear = linear-only (no hinges), Special = column role (e.g. `contract_date`).”

삭제 대신 소프트 비활성화. 각 변수 이름 옆의 $\times$ 토글은 해당 행을 회색으로 만들고 그 열을 데이터에서 지우지 않은 채 모형에서 빼낸다. 토글은 초록색 $+$ 로 바뀌어 열을 다시 활성화한다. 열이 파괴되는 일은 결코 없다.

6.3 자료형 감지와 재정의

`earthUI` 는 가져오기 시점에 자료형을 자동으로 감지한다. 숫자, 정수, 논리, 요인, 날짜 열이 인식된다. 날짜처럼 보이는 문자 열 (값의 80% 이상이 일반 날짜 형식으로 해석되는 경우) 은 `Date` 로 분류된다.

Type 드롭다운을 바꾸어 어떤 감지 결과든 재정의할 수 있다. 열을 `character` 나 `factor` 로 바꾸면 Factor 확인란이 자동으로 선택된다. 자료형을 바꾸면 그 열이 `earth()` 함수에 전달되는 방식이 달라진다.

6.4 Factor 및 Linear 플러그

- **Factor** — 선택하면 해당 열이 범주형 변수로 취급된다. 연속적인 측정값이 아니라 이산적인 집단을 나타내는 **style**, **area_id**, **grade** 같은 열에 적절하다. 요인 열은 지시 (더미) 변수로 모형에 들어간다.
- **Linear** — 선택하면 해당 열이 선형으로만 모형에 들어가며, 힌지 (구간별 선형) 함수가만 들어가지 않는다. 어떤 변수가 목표변수와 엄밀히 선형 관계에 있음을 아는 경우에 유용하다.

6.5 특수 열 유형 참조

감정평가 모드와 시장 모드에서 **Special** 드롭다운은 다음 선택지를 제공한다. 각 유형은 최대 한 열에만 부여할 수 있다 (**display_only**, **remarks**, **list** 는 여럿 허용):

날짜 및 시간 유형:

- **contract_date** — **sale_age** 자동 계산을 촉발한다. 원래의 날짜 열은 매매일과 기준일 사이의 일수를 나타내는 정수 열로 대체된다.
- **listing_date** — 명시적인 **dom** 열이 지정되지 않은 경우 시장 노출 일수 (DOM = 계약일 - 등록일) 를 계산하는 대체 수단으로 사용된다.
- **dom** — 시장 노출 일수 열을 식별한다. 거래사례비교표의 APN 행과 Date of Sale 행에 표시된다.

금액 유형:

- **concessions** — 매매 양허 (매도인 크레딧, 매수인 유인책 등) 를 식별한다. 거래사례비교표에서 순매매가격을 계산하는 데 쓰인다: $\text{Net SP} = \text{Sale Price} - \text{Concessions}$.

규모 및 위치 유형:

- **latitude** — 과적합을 막기 위해 값이 자동으로 소수점 셋째 자리까지 반올림된다. 하버사인 근접성 계산 (대상 부동산에서 각 비교사례까지의 거리) 에 사용되며 거래사례비교표의 Location 행에 묶인다.
- **longitude** — 위도와 동일한 반올림 처리를 받는다. 근접성 계산과 Location 그룹에 사용된다.
- **area** — 대개 근린지역 또는 지역 식별자이다. 거래사례비교표의 “Loc: Long | Lat | Area” 행에서 위도, 경도와 함께 묶인다.
- **living_area** — 내려받기 출력에서 제곱피트당 잔차 계산 (**residual_sf**, **cqa_sf**) 을 가능하게 한다.
- **lot_size** — 거래사례비교표의 “Site Size | Dimensions” 행에 묶인다.
- **site_dimensions** — 거래사례비교표에서 부지 면적과 함께 묶인다 (예: “75x120”).

연령 유형:

- **actual_age** — 거래사례비교표의 “Actual Age | Effective Age” 행에 묶인다.
- **effective_age** — 거래사례비교표에서 실제 연령과 함께 묶인다.

표시 유형:

- **display_only** — 해당 열은 Excel 내보내기에는 포함되지만 모형 적합에서는 완전히 제외된다. 주소 필드, MLS 번호, 필지 ID 등 예측변수가 되어서는 안 되는 참조 데이터에 사용하라. 여러 열에 이 지정을 부여할 수 있다.

텍스트 추출 유형 (실험적):

- **remarks** — 선택적 Prolog 단계가 특성 열로 구문 분석할 자유 텍스트 열 (예: **public_remarks**). 해당 기능을 설정에서 켜기 전까지는 아무 일도 하지 않는다. 여러 열을 지정할 수 있다. 15장을 보라.
- **list** — 같은 단계가 원-핫 <col>_<value> TRUE/FALSE 열로 분할할 심표 구분 열 (예: **existing_structures**). 여러 열을 지정할 수 있다. 15장을 보라.

6.6 다중 목표변수

목표변수를 둘 이상 선택하면 다중 반응 Earth 모형이 적합된다. 여러 목표변수를 선택하면:

- 모형이 모든 반응변수를 동시에 예측하며 목표변수들 사이에서 기저 함수를 공유한다
- **wp (response weights)** 버튼이 활성화되어 각 목표변수에 숫자 가중치를 부여할 수 있다
- 분산 모형 (`varmod.method`) 은 비활성화된다 (다중 반응에서는 지원되지 않음)
- 결과 탭은 반응변수별 지표, 방정식, 진단 그림을 표시한다

팁: *display_only* 로 지정된 열은 데이터셋에 남아 *Excel* 내보내기에 나타나지만 예측변수 표에서는 완전히 제외된다. *ID* 열, 주소, 기타 참조 필드에 사용하라.

7 매개변수 선택

사이드바 4 절 — **Earth Call Parameters** — 은 `earth()` 함수가 받는 모든 인수에 접근할 수 있게 한다. 각 매개변수에는 툴팁 설명이 딸린 파란색 도움말 아이콘 (?) 이 있다. 매개변수는 접이식 하위 절로 조직되어 있다.

7.1 기본 매개변수

매개변수	기본값	설명
<code>subset</code>	(비어 있음)	행 필터 표현식. 아래 “부분집합 필터링” 참조.
<code>weights</code>	NULL	사례 (행) 가중치를 담은 열 선택기. 숫자 열만 나열된다.
<code>wp</code>	NULL	다중 목표 모형의 반응 가중치. 버튼을 누르면 목표변수마다 숫자 입력란이 있는 대화 상자가 열린다 (기본 값은 각각 1.0).
<code>keepxy</code>	꺼짐	x, y, subset, weights 를 모형 객체에 보존한다.
<code>trace</code>	0	적합 출력의 추적 수준 (0 에서 5 까지).
<code>glm</code>	없음	선택적 GLM 분포족: <code>gaussian</code> , <code>binomial</code> , <code>poisson</code> .
<code>degree</code>	1	최대 상호작용 차수. degree 를 ≥ 2 로 설정하면 교차검증이 자동으로 켜지고 Allowed Interactions 행렬이 나타난다.
<code>penalty</code>	2	매듭점당 GCV 벌점. 값이 클수록 더 단순한 모형이 만들어진다.

7.2 전진 단계

매개변수	기본값	설명
<code>nk</code>	자동	가지치기 전 최대 항 수. 적재된 데이터에 대한 권장 값으로 자동 입력된다 (아래 “권장 매개변수 값” 참조).
<code>thresh</code>	0.001	전진 단계 임계값. 값이 작을수록 더 많은 항이 허용된다.
<code>minspan</code>	자동	매듭점 사이의 최소 간격. 권장 값으로 자동 입력된다. 음수 값은 예측변수당 최대 매듭점 수를 정한다.
<code>endspan</code>	자동	끝 간격 — 매듭점에서 데이터 경계까지의 최소 거리. 권장 값으로 자동 입력된다.
<code>newvar.penalty</code>	0	새 변수를 도입하는 데 대한 벌점 (기존 예측변수의 재사용을 유도한다).

매개변수	기본값	설명
fast.k	20	빠른 MARS 알고리즘에서 고려할 부모 항의 수.
fast.beta	1	빠른 MARS 의 노화 계수를 제어한다.

7.3 허용 상호작용

degree 가 2 이상으로 설정되면 기본 매개변수 아래에 상호작용 행렬이 나타난다. 이는 예측변수 쌍마다 하나씩의 확인란으로 이루어진 상삼각 격자이다. 확인란이 선택되어 있으면 두 예측변수의 상호작용이 허용되고, 해제하면 그 특정 상호작용이 금지된다.

예측변수 이름 (행 또는 열 레이블) 을 클릭하면 그 변수의 모든 상호작용이 한꺼번에 전환된다. 상단의 **Allow All** 과 **Clear All** 확인란은 일괄 제어를 제공한다. 이 행렬은 고정 머리글을 사용하므로 스크롤해도 행과 열 레이블이 계속 보인다.

정보 알림은 다음을 상기시킨다: “Interaction terms increase the risk of overfitting. Cross-validation has been enabled (10-fold).”

7.4 가지치기

매개변수	기본값	설명
pmethod	backward	가지치기 방법: backward , none , exhaustive , forward , seqrep , cv .
nprune	NULL	가지치기 후 최대 항 수. 자동 선택을 원하면 비워 두라.

7.5 교차검증

매개변수	기본값	설명
nfold	10	교차검증 분할 (fold) 수. 0 으로 설정하면 교차검증이 꺼진다.
ncross	20	교차검증 반복 횟수.
stratify	켜짐	각 분할이 비슷한 반응 분포를 갖도록 교차검증 표본을 층화한다.

degree 가 ≥ 2 일 때는 상호작용으로 인한 과적합을 막는 데 도움이 되도록 교차검증이 자동으로 켜진다 (nfold 가 10 으로 설정된다).

7.6 부분집합 필터링

subset 텍스트 입력란은 모형 적합에 사용할 행을 걸러 내는 R 표현식을 받는다. 표현식을 직접 입력하거나 (예: `sale_age < 365 & area_id == 460`) **Build filter...** 버튼으로 시각적으로 구성할 수 있다.

Build Filter 대화 상자 — “Build filter...” 를 클릭하면 안내식 대화 상자가 열린다. 각 조건 행에는 열 드롭다운, 연산자 (<, >, <=, >=, ==, !=), 그리고 열 유형에 맞추어 바뀌는 값 입력란이 있다. 숫자에는 숫자 입력란, 날짜에는 날짜 선택기, 문자/요인 열에는 고유값 드롭다운이 제공된다. 조건은 AND(&) 또는 OR(|) 연결자로 결합된다. 하단의 미리보기가 표현식과 일치하는 행의 수를 보여 준다. **Apply** 를 클릭하면 표현식이 텍스트 입력란에 삽입된다.

날짜 열은 수동 표현식에서 `as.Date("...")` 또는 `as.POSIXct("...")` 래퍼를 사용해야 한다. Build Filter 대화 상자는 이를 자동으로 처리한다.

부분집합 필터링은 비파괴적이다. 제외된 행도 데이터셋에 남아 있으며 Excel 내보내기에서 예측값을 받는다.

팁: OR 조건을 AND 와 결합할 때는 괄호로 묶으라. 괄호가 없으면 $A \& B / C$ 는 $(A \& B) / C$ 로 평가되며, 이는 의도한 바가 아닐 수 있다.

7.7 권장 매개변수 값

earthUI 는 4 절의 주요 매개변수 아래에 권장 값을 표시한다. 이 권장값은 적합에 쓰이는 행의 수 (n) 와 선택된 예측변수의 수 (p) 에 따라 반응적으로 갱신된다. 공식은 Friedman 의 MARS 논문, earth 의 내부 알고리즘, 그리고 경험적 시험에서 도출되었다. **nk**, **minspan**, **endspan** 입력란은 데이터가 적재될 때 권장 값으로 미리 채워지며, 수동으로 편집한 값은 그대로 보존된다.

7.7.1 전진 단계 매개변수

nk(가지치기 전 최대 항 수):

$$nk = \min(100, \max(21, 2p + 1, \lfloor n/10 \rfloor))$$

n	$p=5$	$p=10$
30	21	21
50	21	21
100	21	21
200	21	21
500	50	50
1000	100	100
1500	100	100

earth 의 기본값인 $\min(200, \max(20, 2p)) + 1$ 은 데이터셋 크기를 고려하지 않아 큰 데이터셋에서는 지나치게 작을 수 있고, 전진 단계가 모든 예측변수를 충분히 탐색하기 전에 제약을 받게 된다. $\lfloor n/10 \rfloor$ 항은 관측치 10 개당 대략 한 항을 허용하는데, 이는 과다 모수화를 피하기 위한 표준적인 경험 법칙이다. 상한 100 은 수확 체감을 막는다.

minspan(매듭점 사이의 최소 관측치 수):

$$\text{minspan} = \min(16, \lfloor 5 + n/50 \rfloor)$$

endspan(데이터 경계로부터의 최소 관측치 수):

$$\text{endspan} = \min(16, \lfloor 5 + n/28 \rfloor)$$

n	minspan	endspan
30	5	6
50	6	6
100	7	8
200	9	12
300	11	15
500	15	16

n	minspan	endspan
1500	16	16

earth 가 자동 계산하는 minspan 은 Friedman 의 식 43 을 사용한다: $\lfloor (-\ln(-\ln 0.95) + \ln(p \cdot n)) / (2.5 \ln 2) \rfloor$. 이는 $\ln(np)$ 에 비례하여 커지므로 큰 데이터셋에서는 지나치게 느리게 증가한다. $n=1500$ 에서는 약 ≈ 7 밖에 되지 않아 연속 예측변수마다 대략 245 개의 후보 매듭점 위치를 허용하는데, 이는 너무 많아 전진 단계가 잡음을 적당하게 만든다. 권장 공식은 큰 n 에서 예측변수당 후보를 약 100 개로 맞추면서, 작은 n 에서는 (그 영역은 이미 충분히 검증되었으므로) earth 의 자동 계산을 따른다. endspan 은 minspan 보다 약간 크게 설정하여 경계 보호를 추가하는데, 이는 데이터의 꼬리가 얇을 때 (부동산에서는 흔한 일이다) 도움이 된다.

penalty(매듭점당 GCV 벌점):

$$\text{penalty} = \begin{cases} 3 & \text{if degree} > 1 \\ 2 & \text{if degree} = 1 \end{cases}$$

이는 earth 자체의 기본값이다. 가법 모형에는 2, 상호작용 모형에는 3(더 큰 탐색 공간을 상쇄하기 위해 벌점을 높인다).

newvar.penalty(새 예측변수 도입에 대한 벌점):

권장값: 예측변수들이 서로 상관되어 있을 때 (예: 주거 면적과 침실 수, 욕실 수, 부지 면적) 0.1. 이는 전진 단계가 상관된 대체 변수를 새로 들이기보다 이미 모형에 있는 예측변수를 재사용하도록 편향시킨다. 작동 방식은 이렇다. 새 예측변수의 RSS 개선분에 $1/(1 + \text{penalty})$ 가 곱해진다. **newvar.penalty** = 0.1 이면 새 변수는 기존 변수에 매듭점을 하나 더 두는 것보다 RSS 를 최소 10% 더 개선해야 한다. 그 결과 최종 계수 추정치에는 영향을 주지 않으면서 (선택 이후 벌점은 제거된다) 예측변수가 더 적은 단순한 모형이 만들어진다. 예측변수들이 상관되어 있지 않다면 0 으로 두라. earth 는 0 이 아닌 **newvar.penalty** 와 사례 가중치를 함께 지원하지 않으므로, 가중치 열이 지정되면 earthUI 가 이 입력을 비활성화한다 (0 으로 강제).

7.7.2 가지치기 매개변수

pmethod: 권장값은 **backward** 이다. 기본값인 GCV 기반 후진 가지치기는 결정론적이어서 같은 데이터가 언제나 같은 모형을 산출하며 난수 시드에 의존하지 않는다. 이는 재현성, 특히 감정 평가 업무에서 중요하다. **cv** 방법은 가지치기에 교차검증을 사용하므로 시드 의존성이 생긴다.

nprune: 권장값은 비워 두는 것 (NULL) 이다. GCV 가 최적 항 수를 선택하게 하라. **nprune** 을 설정하면 GCV 의 판단을 무시하는 엄격한 상한이 부과된다.

7.7.3 교차검증 매개변수

nfold(교차검증 분할 수): 표본 크기 구간별 권장값:

n (적합 행 수)	nfold
0-299	5
300-399	6
400-499	7
500-599	8
600-699	9
700+	10

분할이 많을수록 각 분할이 더 많은 데이터로 학습되어 (편향이 줄어든 추정치) 좋지만, 대신 검정용으로 남겨 두는 분할이 작아진다. 위 표는 표본이 커짐에 따라 표준적인 10 분할까지 단계적으로 올라가되 각 검정 분할이 안정적인 지표를 내기에 충분한 크기 (대략 50 행 이상) 를 유지하도록 하며, 작은 표본에서는 5 분할로 되돌아간다. 작은 표본에서는 CVR^2 추정치가 여전히 잡

음이 많으므로 `ncross` 를 높여 (예: 15–20) 분산을 평균해 내라. 상한을 10 으로 둔 것은 분할을 하나 더할 때마다 `earth` 적합이 통째로 한 번 더 필요한 데 비해 진단상의 이득은 무시할 만하기 때문이다. `pmethod = "backward"`에서는 교차검증이 모형 자체에는 영향을 주지 않고, 진단용 CVR^2 를 계산하고 분산 모형에 쓸 잔차를 제공할 뿐이다.

`ncross`(교차검증 반복 횟수): 표본 크기 구간별 권장값:

n (적합 행 수)	<code>ncross</code>
0–199	20
200–399	16
400–699	12
700–1199	10
1200–2999	7
3000+	5

반복은 분할 구성의 운을 평균해 내어 CVR^2 를 안정시킨다. 표본이 클수록 단 한 번의 교차검증만으로도 이미 더 안정적이므로 반복이 덜 필요하다. 반복된 교차검증의 잔차는 분산 모형에도 투입된다.

7.7.4 분산 모형

`varmod.method`: 권장값은 `lm` 으로, Milborrow 의 기본값이다. 예측구간을 위한 강건한 선형 분산 모형이다 (산포는 잡음이 많은 잔차로부터 추정되므로 단순한 모형이 더 잘 일반화된다). 데이터에서 `lm` 이 수렴하지 않으면 `rlm` 이나 `const`(등분산, 항상 수렴) 를 시도하거나 `earth` 로 되돌아가라 (비선형 산포도 다룬다). 이 기능에는 `nfold > 1` 과 `ncross > 1` 이 필요하다.

7.7.5 Earth 출력 파일

적합에 성공할 때마다 `earthUI` 는 출력 폴더에 `<filename>_earth_output_<timestamp>.txt` 를 기록한다. 이 파일에는 모형 항, 요약 통계량 (R^2 , $GRSq$, $CVR Sq$), 분산 모형 세부 정보, 추적 로그가 들어 있다. 적합마다 파일이 하나씩 만들어져 매개변수 구성을 비교할 수 있는 누적 기록이 된다.

7.8 설정 기본값

4 절 상단의 라디오 버튼이 어떤 기본값을 적재할지 제어한다:

- **Use last settings for input file** — 이 특정 파일에 대해 지난번에 사용한 설정을 복원한다
- **Use default settings** — 저장해 둔 사용자 지정 기본값을 적용한다
- **Earth defaults** — 모든 매개변수를 공장 초기값으로 되돌린다

4 절 하단의 **Save current as default** 를 클릭하면 현재 `earth` 매개변수와 Allowed Interactions 행렬이 저장된다. 3 절의 대응 버튼은 목표변수, 기준일, 예측변수 구성을 저장한다. 둘 다 프로젝트 데이터베이스에 프로젝트별, 용도별로 저장된다.

8 Earth 모형 적합

8.1 Fit 버튼

사이드바 5 절에는 **Fit Earth Model** 버튼과 **Random seed** 입력란이 있다. 시드는 교차검증 분할 배정을 재현 가능하게 만든다. 드롭다운에서 최근에 사용한 시드 열 개를 불러올 수 있고, 사용된 시드는 적합 후 Earth Output 탭에 기록된다. 적합 전에 earthUI 는 구성을 검증한다. 목표 변수가 선택되어 있어야 하고 예측변수가 최소 하나는 포함되어야 한다. 감정평가/시장 모드에서는 위도와 경도 열이 반올림되고, 기준일과 계약일 (지정된 경우) 로부터 sale age 가 계산된다.

8.2 적합 모달과 추적 출력

Fit 을 클릭하면 어두운 모달 오버레이가 나타나며 다음을 담고 있다:

- 머릿글 — “Fitting Earth Model” 과 함께 초 단위 경과 타이머, 그리고 배경 적합을 취소하는 **Abort** 버튼
- 추적 로그 — `earth()` 함수의 실시간 출력을 보여 주는 스크롤 가능한 고정폭 텍스트 영역: 데이터셋 정보, 전진 단계 진행 상황, 교차검증 분할, 완료 시각
- 색 구분 — 기본 줄은 초록색, 교차검증 줄은 노란색, 오류는 빨간색

earthUI 는 `callr::r_bg()` 를 사용하여 모형을 비동기적으로 적합시키며, 이는 계산을 별도의 R 프로세스에서 실행한다. 덕분에 오래 걸리는 적합 중에도 응용 프로그램이 반응성을 유지한다. 300 밀리초 주기의 폴링 관찰자가 배경 프로세스의 출력을 읽어 추적 로그로 흘려보낸다.

적합이 끝나면 “Done in X.Xs” 메시지가 나타나고 모달에 닫기 버튼 (X) 이 추가되며 배경막이 제거되어 결과 탭이 필요에 따라 그려질 수 있게 된다. 이후 보이는 출력이 모두 그려지면 모달은 저절로 사라지며, 그전에 직접 닫을 수도 있다.

성공하면 Fit 버튼에 초록색 체크 표시가 붙는다. 오류가 발생하면 오류 메시지가 추적 로그에 표시되고 알림이 나타난다.

8.3 적합 로그

모든 적합은 성공이든 실패든 출력 폴더에 로그 파일을 기록한다:

- 파일 이름: `<datafile>_earth_log_<YYYYMMDD_HHMMSS>.txt`
- 내용: 타임스탬프, 적합 과정의 모든 추적 출력, 그리고 오류 메시지
- 위치: 활성 프로젝트의 출력 폴더

팁: `callr` 가 설치되어 있지 않으면 `earthUI` 는 간단한 진행 표시줄과 함께 동기 적합으로 되돌아간다. 완전한 추적 출력을 보려면 `callr` 를 설치하라.

9 데이터 내려받기

적합 후 사이드바 6 절 — General 모드에서는 “Download Estimated Target Variable(s) & Residuals”, 감정평가 및 시장 모드에서는 “Download Estimated Sale Prices & Residuals” 라는 제목이다 — 은 예측값과 진단 정보가 담긴 Excel 파일을 내려받는다. 이 출력은 7 단계 (RCA) 에서 대상 부동산에 CQA(Condition-Quality-Appeal, 상태-품질-선호도) 등급을 부여하는 데 사용된다. 출력은 `residual_sf` 와 `cqa_sf` 로 정렬되어 대상 부동산이 순위에서 어디에 놓이는지 판단하는 데 도움을 준다. 모형의 품질이 좋다면 부동산들은 회귀분석에 들어가지 않은 잔차 특성에 근거하여 선호도가 가장 낮은 것부터 가장 높은 것까지 정렬되어야 한다. 가운데 값은 대략 0 이어야 하고, 아래쪽 절반은 음수, 위쪽 절반은 양수여야 한다. 품질이 가장 나쁜 주택, 즉 “수리 대상” 은 순위 아래쪽에, 가장 좋은 주택은 위쪽에 나타날 것이다. 압류, 숯세일, 상속 관련 매매, 직장 이동 등으로 인한 급매처럼 이례적인 경우에는 대개 예외가 생긴다. 이런 이상값을 조사해 보면 대개 가격 이상을 설명하는 타당한 이유가 발견된다.

버튼 레이블은 용도 모드에 따라 달라진다:

- General/Market: **Download Output (Excel)**
- Appraisal: **Download Intermediate Output (Excel)**

파일 이름 형식은 `<datafile>_modified_<YYYYMMDD_HHMMSS>.xlsx` 이다.

9.1 출력 열

목표변수마다 다음 열들이 덧붙는다:

열	설명
<code>est_<target></code>	모형 예측값. 예: <code>est_sale_price</code> (소수점 1 자리)
<code>residual</code>	실제값에서 예측값을 뺀 값 (소수점 1 자리)
<code>cqa</code>	상태-품질-선호도 점수 (소수점 2 자리, 범위 0-10)
<code>residual_sf</code>	잔차를 주거 면적으로 나눈 값 (지정된 경우, 소수점 1 자리)
<code>cqa_sf</code>	<code>residual_sf</code> 순위로부터 계산한 CQA(소수점 2 자리)
<code><variable>_contribution</code>	g 함수별 기여도 값 (소수점 1 자리, g 함수마다 한 열)
<code>basis</code>	절편 값의 기여분. 모든 부동산에서 동일하다 (소수점 1 자리)
<code>calc_residual</code>	실제값에서 (절편 + 모든 기여도) 를 뺀 값 — 검증용 열 (소수점 1 자리)

다중 목표 모형에서는 목표변수별 열을 구분하기 위해 `_<i>` 접미사가 붙는다.

9.2 열 순서와 Excel 형식

순위 열들은 출력 파일의 맨 왼쪽에 다음 순서로 놓인다: `residual_sf`, `cqa_sf`, `residual`, `cqa`. 덕분에 정렬된 목록을 훑어보며 대상 부동산이 CQA 분포에서 어디에 놓이는지 평가하기 쉽다.

이 열들에는 Excel 숫자 서식이 적용된다:

열	서식	예
<code>residual_sf</code>	숫자, 소수점 2 자리	12.34
<code>cqa_sf</code>	숫자, 소수점 2 자리	7.25
<code>residual</code>	숫자, 소수점 0 자리	15,200
<code>cqa</code>	숫자, 소수점 2 자리	6.80

9.3 CQA 점수

CQA(상태-품질-선호도) 점수는 각 행의 잔차를 다른 모든 잔차와 견주어 순위를 매긴다. 어떤 행의 CQA 는 그보다 작은 부호 있는 잔차를 갖는 행의 비율에 10 을 곱한 값이다. 그 결과 0-10 척도가 만들어지며 다음과 같이 해석한다:

- 높은 CQA($\approx 9-10$) — 그 부동산이 모형 예측보다 훨씬 비싸게 팔렸다 (큰 양의 잔차)
- 낮은 CQA($\approx 0-1$) — 그 부동산이 예측보다 훨씬 싸게 팔렸다 (큰 음의 잔차)
- CQA ≈ 5 — 잔차가 전체 잔차의 중앙값 근처에 있다

living_area 열이 지정되면 cqa_sf 가 제곱피트당 잔차에 근거한 동일한 순위를 제공한다.

9.4 정렬과 대상 부동산 행

감정평가 모드와 시장 모드에서는 비교사례 행이 residual_sf 내림차순으로 정렬된다 (주거 면적이 지정되지 않았다면 residual 기준). 대상 부동산 행 (1 행) 은 1 번 자리에 그대로 남고 정렬되지 않는다. General 모드에서는 행이 원래 순서대로 내보내진다.

이 정렬은 맨 왼쪽에 놓인 순위 열들과 함께, 감정평가사가 가장 과대 예측된 비교사례부터 가장 과소 예측된 비교사례까지 빠르게 훑어보고 대상 부동산에 부여된 CQA 점수가 그 분포의 어디에 놓이는지 평가할 수 있게 한다.

예측 데이터의 요인 수준은 predict() 를 호출하기 전에 학습 데이터와 정렬된다. 학습 때 보지 못한 요인 수준을 가진 행은 NA 예측값을 낸다.

팁: calc_residual 열은 (반올림 오차 범위 안에서) 언제나 residual 과 같아야 한다. 둘이 크게 다르다면 살펴볼 만한 계산 문제가 있다는 뜻이다.

9.5 mgcvUI 자동 내보내기

degree 가 ≤ 2 인 적합이 성공할 때마다 earthUI 는 전체 결과 객체를 .rds 파일로 프로젝트의 출력 폴더에 자동 저장한다. 파일 이름은 <datafile>_earthUI_result_<YYYYMMDD_HHMMSS>.rds 형식을 따른다.

이 파일은 earthUI 패키지에 포함되어 있고 earthUI::launch_mgcv() 로 시작하는 mgcv 후처리 모듈이 readRDS() 로 읽어 들일 수 있다. GAM 루틴은 earth 모형의 매듭점 위치와 기저 함수를 GAM 평활항의 출발점으로 사용하여 MARS 에서 GAM 모형화로 매끄럽게 넘어갈 수 있게 한다.

degree 가 2 를 넘는 모형은 mgcvUI 가 쌍별 상호작용만 지원하므로 건너뛴다. earthUI 를 Trilogy 모드 (Trilogy UI 에서) 로 실행하면 추가로 Lock Model Output 절이 생겨, 특정 적합의 .rds 를 그 예측변수 기저 및 공유 CQA 설정과 함께 glmnetUI 와 mgcvUI 가 가져올 정본 earth 모형으로 잠글 수 있다.

9.6 glmnetUI 가져오기

GAM 루틴을 위해 저장된 것과 같은 .rds 파일을 엘라스틱 넷 후처리 모듈로도 가져올 수 있다. 이 모듈 역시 earthUI 패키지에 포함되어 있으며 earthUI::launch_glmnet() 으로 시작한다. glmnet 응용 프로그램에서 Section 2: Import from earthUI 로 이동한 뒤 Browse 버튼으로 .rds 파일을 선택한다.

glmnetUI 는 표준 earth::model.matrix() 방식을 사용한다. earth 모형의 기저 함수 (힌지, 상호작용, 선형 항) 가 glmnet 설계 행렬의 열이 된다. 그러면 glmnet 은 이 기저 열들에 대해 정규화 회귀를 수행하여 라쏘/엘라스틱 넷으로 선택하고 축소한다. 이는 earth 의 적응적 기저 구성과 glmnet 의 정규화를 결합한다.

9.6.1 주요 이점

- **해석 가능한 상호작용:** earth 의 힌지 기반 상호작용 (3 원 항 포함) 은 명확한 기하학적 의미를 지니며, 이는 법정이나 감사 상황에서 설명하기 어려운 glmnet 고유의 교차곱 상호작용과 대조된다.
- **자동 비선형성:** earth 의 힌지 함수는 원 예측변수를 쓰는 표준 glmnet 모형이 놓칠 비선형 관계를 포착한다.
- **변수 선택:** glmnet 은 기여하지 않는 힌지 항을 걸러 내며 earth 기저를 한 번 더 가지치게 할 수 있다.

9.6.2 중요: 가중치 열이 일치해야 한다

경고

earthUI 가 가중치 열을 사용한다면 (예: 가중치 0 으로 대상 부동산 행을 제외하는 경우), glmnetUI 의 변수 구성에서도 동일한 가중치 열을 지정해야 한다. 가중치 설정이 다르면 행 수가 맞지 않아 earth 기저를 glmnet 의 모형 행렬에 덧붙일 수 없다.

9.6.3 작업 흐름

1. earthUI 에서 earth 모형을 적합시킨다 (degree 는 상관없다).
2. .rds 결과 파일이 프로젝트의 출력 폴더에 자동 저장된다.
3. glmnetUI 에서 **Section 2: Import from earthUI** 를 열고 .rds 파일을 찾는다.
4. 두 응용 프로그램에서 같은 데이터 파일과 같은 가중치 열을 쓰고 있는지 확인한다.
5. 5 절에서 glmnet 매개변수를 구성한다 (earthUI 가져오기가 활성화되면 상호작용 행렬은 자동으로 비활성화된다).
6. **Fit Model** 을 클릭한다. earth 기저 열들이 Equation, Coefficients, Contributions 탭에 나타난다.

10 RCA 계산과 내려받기

Calculate RCA Adjustments & Download 버튼은 사이드바 7 절에 나타나며, 감정평가 모드에서 모형을 적합시킨 뒤에만 보인다. RCA(잔차 제약 접근법, Residual Constraint Approach)는 대상 부동산 대비 비교사례별 시장 도출 보정치를 산출한다.

10.1 RCA 대화 상자 열기

버튼을 클릭하면 다음을 담은 작은 모달 대화 상자가 열린다:

- **Score type** — **CQA** 또는 **CQA per SF** 를 고르는 라디오 버튼 (SF 선택지는 **living_area** 열이 지정된 경우에만 사용할 수 있다). “CQA per SF” 를 고르면, 대상 부동산에 부여한 CQA 점수에 근거하여 그 잔차 점수는 주거 면적에 해당 CQA_SF 점수에 대응하는 **residual_sf** 를 곱한 값이 된다.
- **CQA value** — 대상 부동산의 CQA 점수 입력란 (0.00–10.00). 이 점수는 비교사례들의 품질 분포에서 대상 부동산이 어디에 놓인다고 보는지를 나타낸다.
- **Generate** 버튼 — RCA 출력을 계산하여 Excel 로 내려받는다. 유효한 점수를 입력하기 전까지는 버튼이 비활성 상태이다.

10.2 CQA 점수 보간

Generate 를 클릭하면 earthUI 는 비교사례의 CQA/잔차 쌍으로부터 대상 부동산의 잔차를 보간한다:

1. 비교사례의 CQA 점수와 잔차를 CQA 오름차순으로 정렬한다
2. 선형 보간 (`stats::approx()`) 으로 입력한 CQA 값을 잔차에 대응시킨다
3. CQA per SF 를 사용하는 경우, 제곱피트당 잔차에 대상 부동산의 주거 면적을 곱하여 총 잔차로 되돌린다
4. 대상 부동산의 추정 가치를 다음과 같이 계산한다: 모형 예측값 + 보간된 잔차

10.3 출력 열

RCA Excel 파일 (`<datafile>_adjusted_<YYYYMMDD_HHMMSS>.xlsx`) 에는 모든 중간 출력 열에 더해 다음이 포함된다:

열	설명
subject_value	모형 예측값 + 보간된 잔차 (1 행에만)
subject_cqa	입력한 CQA 점수 (1 행에만)
<variable>_adjustment	대상 부동산의 기여도에서 비교사례의 기여도를 뺀 값 (g 함수별)
residual_adjustment	대상 부동산의 잔차에서 비교사례의 잔차를 뺀 값
net_adjustments	모든 보정치의 합 (기여도 + 잔차)
gross_adjustments	모든 보정치 절댓값의 합
residual_pct	비교사례 매매가격 대비 잔차 보정의 비율
net_adj_pct	비교사례 매매가격 대비 순보정의 비율
gross_adj_pct	비교사례 매매가격 대비 총보정의 비율
adjusted_sale_price	비교사례 매매가격 + 순보정

보정 열들은 각 비교사례에 대해 모형이 그 차이를 각 변수에 얼마나 귀속시키는지를 알려 준다. **adjusted_sale_price** 는 모형에서 도출된 모든 보정을 적용한 뒤의 비교사례 매매가격으로, 이렇게 얻은 보정 매매가격들은 대상 부동산의 추정 가치 주변에 모여야 한다.

10.4 RCA Adjustments 탭

RCA 출력이 계산되면 주 패널의 **RCA Adjustments** 탭에 비교사례 전반의 Residual Adj %, Net Adj %, Gross Adj % 분포 히스토그램이 표시되어 (다중 목표 모형에서는 목표변수마다 한 별) 보정치가 얼마나 촘촘히 모여 있는지 시각적으로 빠르게 확인할 수 있다.

팁: CQA 5.00 점은 대상 부동산을 비교사례의 중앙값에 놓는다. 5 보다 높은 점수는 대상 부동산이 평균 이상의 품질임을, 5 보다 낮은 점수는 평균 이하임을 뜻한다.

11 거래사례비교표

11.1 개요와 목적

거래사례비교표 (Sales Comparison Grid) 는 RCA 보정치를 계산한 뒤 감정평가 모드 (사이드바 8 절) 에서 생성되는 Excel 통합문서이다. 대상 부동산과 선택된 비교사례를 구조화된 표 형식으로 나란히 제시하며, Excel 수식이 각 비교사례의 보정치와 보정 매매가격을 자동으로 계산한다.

이 표는 감정평가사의 작업 파일을 위해 설계되었다. Earth 모형에서 도출된 회귀 기반 보정치와, 감정평가사가 CQA 잔차를 특정 부동산 특성에 배분할 수 있는 편집 가능한 셀을 결합한다. 출력 파일 이름은 SalesGrid_<YYYYMMDD_HHMMSS>.xlsx 이다.

11.2 비교사례 선택 모달

“Generate Sales Grid & Download” 를 클릭하면 포함할 비교사례를 선택하는 모달 대화 상자가 열린다:

- 권장 비교사례가 미리 선택되어 있다. 이는 총보정 비율이 매매가격의 25% 미만인 비교사례로, sale age 오름차순 (가장 최근 매매부터) 으로 정렬되어 최대 30 건까지 표시된다.
- 추가 비교사례가 그 아래에 기본적으로 선택되지 않은 상태로 나열된다. 총보정이 25% 이상이지만 여전히 선택할 수 있는 것들이다 (최대 50 건 표시, 총보정 비율 오름차순 정렬).
- 최대 **30** 건의 비교사례를 선택할 수 있으며, 최대 **10** 개 시트가 만들어진다 (시트당 비교사례 3 건).

선택을 확정하면 선택된 비교사례는 총보정 비율 오름차순으로 정렬되므로 가장 유사한 비교사례가 시트 1 에 나타난다.

11.3 표 배치

각 시트는 대상 부동산과 비교사례 3 건을 수용하는 20 열 배치를 갖는다. 각 개체의 열 구성은 다음과 같다:

개체	열 (개체당 5 개)
대상 부동산	Label, Factual Value 1, Factual Value 2, Factual Value 3, Value Contribution (VC)
각 비교사례	Factual Value 1, Factual Value 2, Factual Value 3, Value Contribution (VC), Adjustment

행은 위에서 아래로 다음과 같다:

1. 제목 — 시트 이름 (예: “Intermediate Sales Comparable Grid — Sheet 1 of 3”)
2. 머리글 — “Subject” 와 주소가 포함된 비교사례 열 머리글
3. Address — 부동산의 전체 주소
4. APN | MLS# | DOM | Subj.Prox — 필지 번호, 등록 ID, 시장 노출 일수, 그리고 대상 부동산으로부터의 하버사인 거리 (마일)
5. Sales Price | Concess. | Net SP — 매매가격, 양허, 그리고 순매매가격 수식
6. Regression Features 머리글 행
7. BASE VALUE — 모형 절편
8. Date of Sale | OffMkt | OnMkt — 계약일, sale age, DOM
9. 묶음 행 (조건부) — Location, Site Size, Age 그룹
10. 모형 변수 행 — 예측변수마다 한 행 (묶음 변수 제외)
11. 빈 구분 행
12. Residual Features 머리글 행
13. CQA|Residual — CQA 점수와 남은 잔차 수식

14. 잔차 특성 행 — 이름이 붙은 행과 감정평가사 입력용 빈 행
15. **Total VC / Net Adjustment**
16. **Net Adjustment %**
17. **Gross Adjustment %**
18. **Adjusted Sale Price** — 주식 행
19. 저작권 꼬리말

11.4 매매가격, 양허, Net SP

5 행은 각 비교사례에 대해 세 가지 값을 보여 준다:

- **Sale Price** — 데이터에 있는 비교사례의 매매가격
- **Concessions** — concessions 로 지정된 열의 값 (지정되지 않았으면 0)
- **Net SP** — Excel 수식: Sale Price – Concessions

대상 부동산 열은 매매가격을 알 수 없으므로 “N/A” 를 표시하며, 양허 값이 있으면 그 값을 보여 준다.

11.5 묶음 행 (위치, 부지, 연령)

특정 특수 유형이 지정되어 있고 그 변수들이 모형에 나타나면, earthUI 는 관련 변수를 결합한 묶음 행을 만든다:

Loc: Long | Lat | Area — longitude, latitude, area 중 어느 하나라도 모형에 있으면 나타난다. 구성 변수 각각의 사실값, 결합된 가치 기여도 (개별 VC 의 합), 그리고 비교사례에 대해서는 보정치 (대상 부동산의 결합 VC – 비교사례의 결합 VC) 를 보여 준다. VC 셀은 연한 파란색 배경으로 꾸며진다.

Site Size | Dimensions — lot_size 나 site_dimensions 가 모형에 있으면 나타난다. 구조는 Location 그룹과 같다.

Actual Age | Effective Age — actual_age 나 effective_age 가 모형에 있으면 나타난다. 구조는 동일하다.

묶음 행에 쓰인 변수는 아래의 개별 모형 변수 행에서 제외되어 보정 합계에서 중복 계산되는 일을 막는다.

11.6 모형 변수 행

묶음 행 아래 (묶음 행이 없으면 BASE VALUE 바로 아래) 에는 남은 모형 예측변수마다 한 행씩 다음을 보여 준다:

- 대상 부동산과 각 비교사례의 사실값 (실제 데이터 값)
- 가치 기여도 (VC) — 해당 행의 예측값에 대한 g 함수의 기여분
- 보정치 (비교사례 전용) — 대상 부동산의 VC 에서 비교사례의 VC 를 뺀 값

11.7 CQA|Residual 및 잔차 특성 행

CQA|Residual 행은 각 부동산의 CQA 점수를 보여 주며 남은 잔차, 즉 아직 특정 특성에 배분되지 않은 잔차 부분을 구하는 수식을 담고 있다. 수식은 다음과 같다:

$$\text{Remaining Residual} = \text{Total Residual} - \sum (\text{Residual Feature VCs below})$$

그 아래에는 잔차 특성 행이 있다. 이름이 붙은 5 개 행 (Location, View/Appeal, Condition, Quality, Other) 과 빈 행 6 개이다. 이 셀들은 감정평가사가 모형이 포착하지 못한 특성의 가치 기여도를 입력하는 입력 셀이다. 감정평가사가 값을 채워 넣으면 남은 잔차 수식의 값이 자동으로 줄어든다.

각 비교사례의 보정 열에는 수식이 들어 있다: 대상 부동산의 특성 VC – 비교사례의 특성 VC.

11.8 보정 매매가격 수식

Adjusted Sale Price 행에는 Excel 수식이 들어 있다:

- **대상 부동산:** BASE VALUE 부터 마지막 잔차 특성 행까지의 모든 가치 기여도 셀의 합. 이는 대상 부동산에 대한 모형의 총 예측값에 감정평가사가 배분한 잔차 특성을 더한 값이다.
- **비교사례:** Net SP 에, 첫 묶음 행 또는 변수 행부터 마지막 잔차 특성 행까지의 모든 보정 셀의 합을 더한 값. 이는 회귀에서 도출된 보정과 감정평가사가 입력한 보정을 모두 적용한 뒤의 비교사례 매매가격이다.

11.9 시트 보호와 편집 가능 셀

각 시트는 수식과 데이터가 실수로 수정되는 것을 막기 위해 **보호**되어 있다. 보호는 다음을 잠근다:

- 모든 수식 셀 (Net SP, 남은 잔차, 보정치, 보정 매매가격 등)
- 모든 데이터 셀 (주소, 매매가격, 사실값, 모형에서 나온 가치 기여도)
- 레이블과 머리글

잠금이 풀린 (편집 가능한) 셀은 잔차 특성의 가치 기여도 입력란, 즉 CQA|Residual 행 아래에서 감정평가사가 내역을 입력하는 셀들뿐이다. 이 셀들은 편집 가능성을 나타내기 위해 연한 노란색 배경으로 꾸며져 있다.

11.10 Excel 에서 표 다루기

1. Excel 또는 호환되는 스프레드시트 프로그램에서 **파일을 연다**.
2. **회귀 보정치를 검토한다** — 모형에서 도출된 보정치는 미리 채워져 있고 잠겨 있다.
3. **잔차를 배분한다** — 잔차 특성 행 (노란색 셀) 에 남은 잔차 중 얼마만큼이 각 특성 (Location, View, Condition, Quality 등) 에 귀속되는지에 대한 판단을 입력한다. 배분함에 따라 남은 잔차 수식의 값이 줄어든다.
4. **보정 매매가격을 확인한다** — 잔차 특성 값을 입력하면 이 수식이 자동으로 갱신된다.
5. **여러 시트** — 비교사례를 3 건 넘게 선택했다면 시트 사이를 옮겨 다니라. 각 시트는 왼쪽 열에 같은 대상 부동산을 두고 서로 다른 비교사례 3 건을 보여 준다.

팁: 목표는 남은 잔차가 0 에 가까워질 때까지 CQA 잔차를 특성 행에 배분하는 것이다. 이는 각 비교사례의 회귀 추정치와 실제 매매가격의 차이를 설명할 수 있음을 보여 준다.

12 보고서 내려받기

모형을 적합시킨 뒤의 보고서 생성은 두 단계의 Quarto 작업 흐름이다. **Generate Quarto Report**(감정평가 모드에서는 사이드바 9 절, 그 외에는 7 절) 가 자체 완결형 보고서 묶음을 기록하고, **Convert Quarto Report**(10 절) 가 그것을 — 또는 다른 어떤 Quarto 파일이라도 — 최종 형식으로 렌더링한다.

12.1 Generate Quarto Report

Generate Quarto Report 를 클릭하면 프로젝트 출력 폴더 아래 `<base>_qmd/`에 자체 완결형 Quarto 묶음이 기록된다. 여기에는 값이 채워진 .qmd 원본, 미리 생성된 모든 그림 자산 (PNG 와 PDF), 보고서 데이터, 그리고 Word 참조 서식 파일이 들어 있다. 이 단계에서는 렌더링이 일어나지 않는다. 묶음은 직접 편집할 수도 있고, Quarto include 를 통해 다른 Quarto 원본과 결합할 수도 있으며, quarto preview 로 단독 미리보기를 할 수도 있다.

12.2 Convert Quarto Report

10 절은 파일 선택기 (기본값은 현재 프로젝트에서 가장 최근에 생성된 묶음이지만 어떤 .qmd 파일로도 찾아갈 수 있다) 와 형식 확인란을 제공한다:

- **HTML** — KaTeX 수식 렌더링, 내장 이미지, 목차를 갖춘 .html 파일. Flatly Bootstrap 테마와 Roboto Condensed 글꼴을 사용한다.
- **Word** — 편집과 배포에 적합한 .docx 파일. 목차와 쪽 번호를 포함하며, 일관된 서식을 위해 사용자 지정 참조 서식 파일을 사용한다.
- **PDF** — 전문가 수준의 출력을 위해 LuaLaTeX 으로 조판된다. 용지 크기는 로캘 설정 (Letter 또는 A4) 을 따른다. 큰 상호작용 행렬을 위한 가로 방향 페이지를 포함하며, 본문에는 Roboto Condensed, 수식에는 Latin Modern Math 를 사용한다. PDF 선택지는 LaTeX 배포판이 감지될 때에만 제공된다.

Convert 를 클릭하면 선택한 형식으로 렌더링된다. 렌더링은 배경에서 실행되며, 모달 대화 상자가 경과 시간과 Quarto 진행 상황을 보여 주는 동안에도 응용 프로그램은 반응성을 유지한다. 완료되면 알림이 파일 위치를 알려 준다.

모든 작업 (모형 적합, 데이터 내려받기, RCA 계산, 거래사례비교표 생성, 보고서 렌더링) 은 시작/종료 타임스탬프와 경과 시간과 함께 출력 폴더의 `<datafile>_earthui_log.txt` 에 기록되어 문제 해결과 성능 점검에 쓰인다.

12.3 보고서 내용

생성되는 모든 보고서에는 다음 절들이 들어 있다:

1. 데이터셋 설명 — 관측치 수, 목표변수, 사용된 예측변수, 범주형 예측변수
2. 모형 설정 — degree, 교차검증 여부, 항의 수, 모형에 들어간 예측변수
3. 허용 상호작용 (degree ≥ 2 인 경우에만) — 체크 표시가 있는 상호작용 행렬. 출력 형식에 맞게 서식이 갖추어진다
4. 결과 요약 — R^2 , GRSq, GCV, RSS, 그리고 (교차검증이 켜져 있었다면) $CV R^2$
5. 모형 방정식 — LaTeX 표기로 나타난 완전한 Earth 모형 방정식
6. 계수와 기저 함수 — 모든 항, 그 계수, 그리고 그것들을 정의하는 기저 함수의 표
7. 변수 중요도 — 예측변수 중요도 점수의 막대그래프와 순위 표
8. g 함수 기여도 — g 함수별 그림. 이변량 항은 선 그림, 이변량 상호작용은 3 차원 투시도와 등고선 그림
9. 상관 행렬 — 예측변수 상관관계 히트맵
10. 진단 — 잔차 대 적합값, 정규 Q-Q, 실제값 대 예측값 그림
11. ANOVA 분해 — 각 기저 함수의 RSS 기여도를 보여 주는 표

12. **Earth** 출력 — 가지치기 단계 세부 정보를 포함한 `earth::summary()` 의 원시 텍스트 출력

다중 반응 모형에서는 4-8 절과 10 절이 목표변수마다 반복된다.

팁: 보고서에는 *quarto R* 패키지와 *Quarto* 설치가 필요하다. *PDF* 출력에는 *LaTeX* 배포판도 설치되어 있어야 하며 (예: `tinytex::install_tinytex()`), 그렇지 않으면 *PDF* 확인란이 숨겨진다.

13 후처리 모듈: 엘라스틱 넷과 GAM

earthUI 패키지에는 두 개의 후처리 모듈이 함께 제공된다. 하나는 `glmnet` 기반의 엘라스틱 넷 루틴이고, 다른 하나는 `mgcv` 기반의 GAM 루틴이다. 각각은 이번 호의 별도 논문에서 깊이 다루므로, 이 절에서는 그것들이 earthUI 작업 흐름에서 무엇을 위한 것인지, 그리고 세 루틴이 어떻게 협력하는지를 설명한다.

13.1 왜 보조 적합인가?

가치 결론을 내는 것은 earth 모형이다. 후처리 모듈은 그것을 보강하기 위해 존재한다. 같은 프로젝트 데이터에 적합된, 방법론적으로 다른 엔진이 산출한 동일 대상 부동산에 대한 두 개의 추가 추정치인 것이다. earth 기저의 엘라스틱 넷 재추정과 독립적인 GAM 평활이 실질적으로 같은 가치에 도달한다면, 감정평가사의 결론은 더 이상 한 모형의 가정에만 기대지 않게 된다. 이는 심사와 소송 상황에서 의미 있는 이점이다. 반대로 결과가 어긋난다면 그 불일치 자체가 진단이 되어, 대개 데이터 부족, 지나치게 유연한 항, 또는 다시 살펴볼 만한 이상 비교사례를 가리킨다. 어느 모듈도 earth 모형보다 우위에 있지 않다. 이들은 보조 증거이며, 의도적으로 가중치 없이 제시된다.

13.2 실행과 상태 공유

각 루틴은 별도의 로컬 Shiny 응용 프로그램이다:

- `earthUI::launch_glmnet()` — 엘라스틱 넷 응용 프로그램, 포트 7879.
- `earthUI::launch_mgcv()` — GAM 응용 프로그램, 포트 7880.

세 응용 프로그램은 모두 같은 프로젝트 트리와 같은 설정 데이터베이스를 읽으므로, earthUI에서 열어 둔 프로젝트를 모듈에서도 곧바로 쓸 수 있다. earthUI가 그 프로젝트에 대해 저장하는 이월 필드, 즉 반응 (목표) 변수, 기준일, RCA 대상 부동산 CQA 유형과 값, 주거 면적 지정도 함께 넘어간다. 이 값들을 다시 입력할 필요는 없다. 모듈이 그것을 읽어 오므로 같은 값이 세 적합 모두에 흐른다.

13.3 엘라스틱 넷 루틴

`glmnet` 루틴은 earth 모형의 기저 함수 — 힌지, 상호작용, 선형 항 — 를 설계 행렬의 열로 가져와 라쏘/엘라스틱 넷 정규화 아래에서 재추정한다 (작동 방식과 가중치 열 관련 주의 사항은 위의 “`glmnetUI` 가져오기” 참조). 그 결과는 earth의 해석 가능하고 기하학적 의미가 분명한 항들을 유지하면서, 제 몫을 하지 못하는 항은 정규화가 축소하거나 제거하도록 한다. 자세한 내용은 이번 호의 `glmnetUI` 논문을 보라.

13.4 GAM 루틴

`mgcv` 루틴은 반대 방향으로 간다. earth의 기저를 재사용하는 대신, earth 모형의 매듭점 위치를 출발점으로 삼아 새로운 벌점 평활항을 적합시킨다 (위의 “`mgcvUI` 자동 내보내기” 참조). GAM은 다른 평활화 패러다임 아래에서 함수 형태를 다시 도출하므로, earth 모형과의 일치는 메아리가 아니라 진정으로 독립적인 보강이 된다. `degree`가 ≤ 2 인 earth 모형이 지원된다. 자세한 내용은 이번 호의 `mgcvUI` 논문을 보라.

13.5 실무적 작업 흐름

1. earthUI에서 만족스러울 때까지 **earth 모형**을 먼저 적합시키고 다듬는다. 성공한 적합 ($\text{degree} \leq 2$) 마다 모듈이 사용할 `.rds`가 자동으로 내보내진다.
2. 같은 프로젝트에 대해 **glmnet**과 **mgcv** 루틴을 열고 earth 결과를 가져와 적합시킨다. 한 번에 하나의 루틴만 다루라. 적합은 오래 걸리고 루틴들은 공유 프로젝트를 통해 협력하므로, 합리적인 리듬은 병렬이 아니라 순차이다.

3. 자유롭게 반복하라. 각 응용 프로그램은 전환하는 동안에도 결과를 유지하며, 도중에 보고서를 만들어야 할 이유는 전혀 없다.
4. 세 결과가 모두 만족스러울 때에만 각 루틴으로 돌아가 Quarto 보고서를 생성하라. 각 방법은 자신의 적합과 가치 결론을 프로젝트의 trilogy 기록에 등록하므로, 세 보고서가 하나의 비교 실행으로 추적된다.

팁: 모듈은 어느 수준에서도 선택 사항이다. 적합, RCA, 거래사례비교표, 보고서로 이어지는 *earthUI* 자체의 작업 흐름은 모듈 없이도 완결된다. 어려운 대상 부동산, 얇은 시장 부문, 또는 심사를 받게 될 가능성이 높은 업무처럼 결론에 보강이 필요할 때 모듈을 활용하라.

14 데모 데이터셋: Appraisal_1.csv

earthUI 에는 감정평가 작업 흐름을 살펴볼 수 있는 데모 MLS 데이터셋이 들어 있다. 프로그램 방식으로의 다음과 같이 불러온다:

```
demo_file <- system.file("extdata", "Appraisal_1.csv", package = "earthUI")
df <- import_data(demo_file)
```

또는 Shiny 응용 프로그램에서 이 파일을 프로젝트의 in/ 폴더로 복사하거나 (프로젝트를 만들 때 최초 데이터 파일로 첨부해도 된다) 2 절 (Import Data) 에서 선택한다.

14.1 설명

이 파일에는 모의 MLS 내보내기에서 나온 주거용 매매사례 1,501 건과 1 행의 대상 부동산 1 건 (총 1,502 행) 이 들어 있다. 데이터는 부동산 규모, 연령, 위치가 다양한 여러 지역으로 이루어진 시장의 단독주택 매매를 나타낸다.

이는 실제 데이터가 아니지만 북부 캘리포니아의 현실적인 근린지역에 기초하고 있다. 모든 식별 정보는 변경되거나 제거되었다.

14.2 열

열	자료형	특수 유형	설명
weight	numeric	—	관측치 가중치 (0 = 적합에서 제외)
id	numeric	display_only	내부 레코드 ID
property_id	numeric	display_only	MLS 부동산 식별자
listing_id	character	display_only	MLS 등록 번호
parcel_number	character	display_only	카운티 과세평가관 필지 번호 (APN)
street_address	character	display_only	부동산 주소
city_name	character	display_only	도시
postal_code	character	display_only	우편번호
county_name	character	display_only	카운티
contract_date	Date	contract_date	매매 계약일 (sale_age 계산)
sale_age	numeric	—	계약일부터 기준일까지의 일수 (미리 계산됨)
coe_date	Date	display_only	에스크로 종결일
listing_status	character	display_only	등록 상태 (예: "Sold")
sale_price	numeric	(목표변수)	매매가격 — 반응변수
rent	numeric	—	월 임대료 (다중 목표 모형용)
list_price	numeric	display_only	등록 가격
original_list_price	numeric	display_only	최초 등록 가격
living_sqft	numeric	living_area	총 주거 면적 (제곱피트)
beds_total	integer	—	침실 수
baths_total	numeric	—	총 욕실 수 (예: 2.5 = 전체 2 + 반 1)
lot_size	numeric	lot_size	부지 면적 (제곱피트)
area_id	integer	area	MLS 지역 식별자
area_text	character	display_only	지역명
age	numeric	actual_age	부동산 연령 (년)
year_built	integer	display_only	준공 연도
latitude	numeric	latitude	위도 (모형용으로 소수점 셋째 자리 반올림)
longitude	numeric	longitude	경도 (모형용으로 소수점 셋째 자리 반올림)
latitude6	numeric	display_only	전체 정밀도 위도
longitude6	numeric	display_only	전체 정밀도 경도
garage_spaces	integer	—	차고 주차 칸 수
fp_count	integer	—	벽난로 수
no_of_stories	numeric	—	층수
style	character	—	건축 양식
view	character	—	조망 유형 (예: "Neighborhood", "Hills")
days_on_market	integer	dom	시장 노출 일수
listing_date	Date	listing_date	등록일
sale_concessions	numeric	concessions	매도인 양허

14.3 권장 빠른 시작

1. earthUI 를 실행한다: `earthUI::launch()`
2. 새 프로젝트 마법사에서 Purpose 를 **For Appraisal** 로 하여 프로젝트를 만들고 **Appraisal_1.csv** 를 최초 데이터 파일로 첨부한다
3. 2 절 (Import Data) 에서 **Appraisal_1.csv** 를 선택한다
4. 목표변수로 **sale_price** 를 선택한다
5. 위 표에 나온 대로 특수 유형을 부여한다
6. 예측변수를 포함시킨다: **sale_age, living_sqft, baths_total, lot_size, area_id**(요인으로), **age, latitude, longitude, garage_spaces**
7. degree 를 1 로 설정하고 **Fit Earth Model** 을 클릭한다
8. 중간 출력을 내려받고 (6 단계) CQA 순위를 검토한다
9. CQA 점수 ~5.00 으로 RCA 보정치를 계산한다 (7 단계)
10. 거래사례비교표를 생성한다 (8 단계)

답: 데모 데이터셋은 자신의 *MLS* 데이터를 가져오기 전에 *earthUI* 의 모든 기능을 살펴보기에 좋은 방법이다. 위에 제시한 예측변수들은 대개 R^2 가 0.85 를 넘는 모형을 만들어 낸다.

15 실험적 기능: Prolog 텍스트 추출과 도출 규칙

0.10.0 버전은 MLS 내보내기의 서술형 텍스트 열 — 무엇보다도 “Public Remarks” 열, 그리고 중개인 비고와 방 특성이나 기존 구조물 같은 심포 구분 특성 목록 — 에서 모형에 바로 쓸 수 있는 특성을 추출하는 선택적이고 진정으로 실험적인 기능을 도입했다. 추출 엔진은 Prolog, 특히 한정절 문법 (DCG, Definite Clause Grammars) 이다. 비고 텍스트에서 “move in ready”, “3 car garage”, “owned solar” 같은 구절을 인식하는 규칙이 Prolog 문법 규칙으로 작성되어 있다.

이 기능은 아직 개발 중이다. 오늘도 추출에 부분적으로 쓸 수는 있다. 함께 제공되는 문법이 이미 유용한 특성 집합을 인식한다. 그러나 자신의 MLS 어휘에 맞게 조정하려면 Prolog, 그중에서도 DCG 를 읽고 쓸 수 있어야 한다. 그럴 형편이 아니라면 이 장은 통째로 건너뛰어도 된다. 이 기능은 기본적으로 꺼져 있으며, 그것 없이도 earthUI 의 나머지는 똑같이 동작한다.

경고

이것은 활발히 개발 중인 실험적 기능이다. 추출 문법은 완결되어 있지도, 실무 감정평가용으로 검증되어 있지도 않다. 추출된 열은 모형에서 신뢰하기 전에 원본 텍스트와 대조하여 검토해야 한다. 이 기능을 효과적으로 쓰려면 Prolog 와 DCG 에 대한 실무 지식이 필요하다. 이 기능이 하는 일은 모두 덧붙이기이고 되돌릴 수 있다. 문제가 생기면 설정 토글을 끄고 평소대로 계속하면 된다.

15.1 무엇을 하는가

MLS 등록 정보는 자유 텍스트와 구분자로 나뉜 목록 필드에 가치와 관련된 정보를 대단히 많이 담고 있다. 이 선택적 단계는 두 가지 특수 열 역할 (제 6 장) 을 통해 그 텍스트를 회귀분석에 쓸 수 있는 평범한 열로 바꾼다:

- **remarks** 열은 Prolog DCG 문법으로 구문 분석되어 추출된 속성마다 한 개의 특성 열이 만들어진다. 열 접두사는 원본 열 이름의 첫 글자들로 만들어진다. `public_remarks` 는 `pr_*` 열을, `agent_remark` 는 `ar_*` 열을 낳는 식이다.
- **list** 열 (예: `existing_structures = "Shed, Workshop"`) 은 원-핫 `<col>_<value>` TRUE/FALSE 열로 분할된다. 이 부분은 순수한 R 이며 Prolog 가 전혀 필요 없다.

함께 제공되는 문법 (`inst/prolog/mls_remarks.pl`) 은 현재 다음과 같은 것들을 추출한다: 1-5 범위의 `condition_score` 를 갖는 등급화된 상태 계층, 조망 플래그와 조망 유형, 주방 및 욕실 개조 플래그, 원목 및 기타 바닥재, 개선 목록과 개수, 차고 칸 수와 차고 면적, 수영장, 스파, 기타 옥외 시설, 벽난로 수, “40x50” 같은 치수에서 크기를 읽어 낸 부속 건물 (창고, 헛간, 작업실, 별채, 게스트하우스, 견사 등), RV 주차와 보관, 태양광의 소유 대 임대 여부와 배터리 저장, 부지 면적 (에이커-제곱피트 환산 포함), 대상 부동산과 결부된 매매 맥락상의 금액 (가격 인하, HOA 회비, 태양광 임대료 또는 상환액, Mello-Roos), 위치 단서, 그리고 $[-1, 1]$ 범위의 장점/우려 감성 점수.

15.2 요구 사항: SWI-Prolog 가 아니라 vProlog(Trealla)

비고 구문 분석은 자체 완결형 Trealla Prolog 엔진인 **vProlog** R 패키지 위에서 실행된다. 이 패키지는 CRAN 에 없다. 이 논문 앞부분의 “earthUI 와 vProlog 설치” 에 나온 대로 저자의 r-universe 저장소에서 설치하라 (Windows 와 macOS 용 사전 빌드 바이너리, 소스는 github.com/wcraytor/vProlog). 시스템에 Prolog(SWI-Prolog 등) 를 설치할 필요도 없고 사용하지도 않는다. vProlog 가 없으면 비고 구문 분석은 안내 메시지와 함께 건너뛰고 `list` 분할은 그대로 실행된다. 도우미 함수 `vProlog_available()` 은 엔진과 함께 제공되는 문법이 모두 있는지 알려 준다.

15.3 활성화와 실행

1. **Settings**(톱니바퀴) 를 열고 **Enable Prolog / list column processing** 을 선택한다. 이 토글은 기본적으로 꺼져 있고, “advanced feature — for users with Prolog expertise” 라는 안내가 붙어 있으며, 프로젝트별로 기억되고, 프로젝트 정보 대화 상자에 표시된다.
2. 예측변수 표에서 해당 열들의 **Special** 역할을 **remarks** 나 **list** 로 설정한다 (각 역할에 여러 열을 지정할 수 있다).
3. 사이드바의 4 절과 5 절 사이에 있는 **Execute Prolog Processing** 절을 사용한다. 두 단계를 제공한다:

1. **Expand lists & remarks** — 각 **remarks** 열에 DCG 문법을 실행하고 각 **list** 열을 원-핫으로 분할한다. 확장은 점증적이다. 마지막 실행 이후에 지정된 열만 처리되며, 지정된 것이 모두 확장되고 나면 버튼이 비활성화된다. 진행 표시줄이 구문 분석 진척을 보여 준다.

2. **Execute derivation rules** — 확장된 열들에 대해 사용자가 작성한 Prolog **derive/2** 규칙을 적용한다 (아래 참조). 규칙은 실행 전에 구문 검사와 예행 검증을 거친다.

15.4 도출 규칙

earthUI 는 데이터 행마다 그 행의 열들을 **cell(Name, Value)** 사실로 단언한 뒤 사용자의 **derive(Col, Val)** 규칙의 모든 해를 모은다. 서로 다른 Col 마다 새 열이 되며 열당 첫 번째로 찾은 값이 채택되므로, 뒤에 대체 절을 둘 수 있다. 도우미 술어 (**count_true/2**, **count_false/2**, **count_true_prefix/2**, **cell_or/3**, **is_na/1**) 가 집계와 NA 처리를 지원하며, **drop/1** 사실은 규칙 실행 후 열을 버린다. 예를 들면:

```
% Merge a collinear cluster of solar flags into one boolean.
derive(solar_any, B) :-
  ( cell(pr_has_solar, true)
  ; cell(pr_solar_owned, true)
  ; cell(pr_has_battery, true) ) -> B = true ; B = false.

% Count amenities instead of OR-ing them (numeric column).
derive(amenity_count, N) :- count_true_prefix(pr_has_, N).

% Threshold -> boolean, with a fallback clause.
derive(big_outbuilding, true) :- cell(pr_outbuilding_sqft, S), S >= 1000.
derive(big_outbuilding, false).

% Discard a raw granular column once consolidated.
drop(pr_price_solar_lease).
```

규칙은 프로젝트마다 하나씩 있는 **<project>_rules.pl** 파일에 들어 있고, 주석이 충실히 달린 서식 파일에서 출발하며, **Edit derivation rules...**를 통해 사이드바에 도킹할 수도 있는 떠 있는 이동 가능한 코드 편집기에서 편집한다. **Check** 버튼은 규칙을 실행하지 않고 검증한다 (구문 검사, 적재 검사, 그리고 데이터 1 행에 대한 예행 실행).

15.5 버전이 매겨진 활성 데이터 파일

두 단계 중 어느 쪽이든 실행하고 나면 증강된 데이터 프레임이 프로젝트의 입력 폴더에 **<original_name>_<YYMMDD_HHMMSS>.xlsx** 로 기록되어 프로젝트의 활성 데이터 파일이 된다. 프로젝트를 다시 열면 증강된 데이터가 적재되므로 확장을 다시 계산할 일이 없고, **glmnetUI** 와 **mgcvUI** 가 사용하는 것도 같은 파일이다. 원본 비교 열은 자동으로 모형에서 제외되며, 생성된 열은 예측변수로 자동 포함되지 않는다. 텍스트 구문 분석은 많은 열을 만들어 낼 수 있으므로, (아직) 회귀에 넣고 싶지 않은 열은 **Latent** 로 표시하거나 × 로 비활성화하라 (제 6 장).

15.6 프로그램 인터페이스와 함께 제공되는 파일

전체 파이프라인은 일괄 처리용으로 내보내져 있다. `execute_prolog_processing(df, remarks_cols, list_cols, rules_text)` 가 각 단계를 조율하며, 이는 `remarks_to_columns_()`, `split_list_column_()`, `apply_user_rules_()`, `validate_rules_()` 위에 세워져 있다. `inst/prolog/` 아래에 함께 제공되는 Prolog 원본은 `mls_remarks.pl`(DCG 문법과 특성 집계), `mls_analyze.pl`(일괄 구문 분석을 위한 파일 입출력 구동부), `mls_rules.pl`(도출 규칙에서 쓸 수 있는 도우미 술어), 그리고 `climate_regions_ca.pl`(아래 참조) 이다.

15.7 관련 실험: 기후 지역 사전정보

이와 짝을 이루는 실험적 기능은 데이터가 얇은 특성, 즉 비교사례 몇 건에만 나타나서 (이를 테면 증발식 냉방기) 회귀분석이 추정할 수 없는 특성을 다룬다. 세 개의 R 도우미 함수 — `climate_region_for()`, `market_area_profile()`, `climate_feature_priors()` — 가 프로젝트의 위치 (카운티 + 도시) 를 매수자가 인지하는 캘리포니아의 일곱 개 기후 지역 (캘리포니아 에너지위원회의 16 개 건축 기후대를 실용적으로 축약한 것) 가운데 하나로 분류하고, 지역별로 HVAC 및 냉방 특성 (난방, 에어컨, 증발식 냉방기, 전관 환기팬, 여과 장치) 에 대한 서열형 기여 가치 사전정보를 돌려준다. Prolog 모듈 `climate_regions_ca.pl` 은 도출 규칙 안에서 쓸 수 있도록 같은 분류를 그대로 반영한다. 양쪽 모두 하나의 정보 R 원본에서 생성되므로 서로 어긋날 수 없다. 이 사전정보는 특성 평가에 참고가 되고 추정 계수의 타당성 점검 역할을 하도록 의도된 것이다. 지역 배정 (특히 여러 지역에 걸친 카운티의 경우) 은 감정평가사가 검토해야 한다.

15.7.1 캘리포니아를 넘어: 관할권 팩과 시장지역 모듈

캘리포니아는 첫 번째 관할권이지 한계가 아니다. `earthUI` 는 국제적으로 사용되고 있으며, 기후 사전정보 지식은 관할권 팩으로 재구성되고 있다. 관할권 팩이란 프로젝트 시스템의 국가 및 행정 수준 코드에 결부된, 자체 완결형이며 데이터 형태를 띤 사실 표 (지대 정의, 행정 단위 배정, 특성 사전정보) 로서, 하나의 팩이 미국의 주, 독일의 란트, 또는 작은 나라 하나를 통째로 기술할 수 있다. 이와 짝을 이루는 시장지역 모듈 층은 이름이 붙은 시장지역의 지역 어휘와 정교화된 사전정보를 담는다.

두 가지 모두 감정평가공학과 감정평가사가 현장에서 자신의 언어로 작성하도록, 그리고 현실적으로는 LLM 의 도움을 받아 작성하도록 설계되었다. 세상에 Prolog 프로그래머는 많지 않으므로, 현장 산출물은 의도적으로 프로그램이 아니라 표로 되어 있다. 팩 형식은 완전한 실제 예시 (캘리포니아 자체) 를 갖춘 문서화된 스키마인데, 이는 Claude 같은 조력자가 시장에 대한 평이한 서술로부터 유효한 팩을 초안할 수 있게 하려는 의도가 명시적으로 담긴 것이다. `earthUI` 는 팩으로부터 바탕이 되는 Prolog 를 생성하며, 팩을 활성화하려면 반드시 차단형 검증기 (스키마, 적용 범위, 참조 무결성, 예행 실행 검사) 를 통과해야 한다. 현장에서 작성된 모든 팩은 의무적인 출처 정보 — 기록상의 작성자, 날짜, 그리고 인용된 근거 — 를 지니므로, 초안 작성 도구가 아니라 책임 있는 전문가가 그 지식을 뒷받침하며 보고서는 사용된 팩과 그 버전을 정확히 인용할 수 있다.

팁: 실무적인 방식은 이렇다. 비고를 확장하고, 잘게 나뉜 `pr_*` 플래그를 `derive/2` 규칙으로 몇 개의 파생 개수나 불리언으로 통합한 뒤, 통합에 쓰인 원 플래그는 `drop/1` 으로 버리고, 파생된 열만 후보 예측변수로 남기는 것이다. 이렇게 하면 예측변수 목록이 짧아지고 다중공선성도 관리할 수 있는 수준으로 유지된다.

참고문헌

- Friedman, Jerome H. (1991). “Multivariate Adaptive Regression Splines”. In: *The Annals of Statistics* 19.1, pp. 1–141. DOI: [10.1214/aos/1176347963](https://doi.org/10.1214/aos/1176347963).
- Milborrow, Stephen (Oct. 1, 2024). *earth: Multivariate Adaptive Regression Splines*. R package. URL: <https://CRAN.R-project.org/package=earth>.
- (n.d.[a]). *Notes on the earth Package*. URL: <http://www.milbo.org/doc/earth-notes.pdf>.

Milborrow, Stephen (n.d.[b]). *Variance Models in earth*. URL: <http://www.milbo.org/doc/earth-varmod.pdf>.