

earthUI: Eine interaktive Benutzeroberfläche für das earth-(MARS-)Paket

Deutsche Übersetzung, mit KI-Unterstützung erstellt und redaktionell geprüft; die englische

Fassung ist maßgeblich.

William Bert Craytor

Jahrgang 1, Ausgabe 1 · Sommer 2026

Veröffentlicht am 1. Juli 2026

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

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

Dieser Open-Access-Artikel wird unter den Bedingungen der [Creative-Commons-Namensnennung-4.0-Lizenz](#) verbreitet.

Zusammenfassung. earthUI ist eine grafische Benutzeroberfläche für das R-Paket `earth`, das Multivariate Adaptive Regression Splines (MARS) implementiert. Es bietet drei Verwendungsmodi — allgemeine prädiktive Modellierung, Immobilien-Wertermittlung und Marktgebietsanalyse — und führt den Anwender durch Datenimport, Modellkonfiguration, Modellanpassung sowie die Interpretation der Diagnostik und grafischen Ausgaben von `earth`. Dieser Artikel dokumentiert die Datenformat-Anforderungen, den Modellierungs-Workflow, die Ausgabedarstellungen und die vollständige Funktionsreferenz von earthUI.

Schlüsselwörter: MARS, Multivariate Adaptive Regression Splines, `earth`, R, grafische Benutzeroberfläche, Immobilienbewertung

Anmerkung des Verlags (Juli 2026): Dieser Artikel beschreibt earthUI 0.10.0. Kurz vor Drucklegung hat earthUI 0.11.0 die Begleitanwendungen `glmnetUI` und `mgcvUI` (in den beiden folgenden Artikeln beschrieben) übernommen und damit die MARS-, Elastic-Net- und GAM-Workflows in einer einzigen Anwendung zusammengeführt. Die hier beschriebene Benutzeroberfläche und der Workflow bleiben von der Zusammenführung unberührt.

Inhaltsverzeichnis

1	Einleitung	4
1.1	Was ist earthUI?	4
1.2	Ein Paket, drei Routinen	4
1.3	Drei Purpose-Modi	4
1.4	Immobilien-spezifische Funktionen	5
1.5	Installation von earthUI und vProlog	5
1.6	Erste Schritte	6
2	Anforderungen an MLS-Eingabedaten	7
2.1	Dateiformat & Struktur	7
2.2	Erforderliche Spalten für den Appraisal-Modus	7
2.3	Besondere Konventionen für Spaltennamen	9
2.4	Datenqualität & Vollständigkeit	9
2.5	Platzierung der Zeile des Bewertungsobjekts	9

3	General Purpose Mode	11
3.1	Überblick	11
3.2	Der Arbeitsablauf in der Sidebar	11
3.3	Tabs des Hauptpanels	12
3.4	Persistenz der Einstellungen	12
3.5	Dark Mode	12
3.6	Locale & regionale Einstellungen	13
3.6.1	Unterstützte Länder	13
3.6.2	Übersteuerung des Papierformats	13
3.6.3	Speichern der Standardeinstellungen	14
4	Appraisal Mode (Gutachtenmodus)	15
4.1	Behandlung der Subject-Zeile	15
4.2	Effective Date & Sale Age	15
4.3	Special-Spalten-Ausweisungen	15
4.4	Residualer Wertbeitrag (Schätzung seltener Merkmale)	16
4.5	Überblick über die RCA-Anpassungen	17
5	Market Area Analysis Mode (Marktgebietsanalysemodus)	18
5.1	Unterschiede zum Appraisal Mode	18
5.2	Wann der Market Mode zu verwenden ist	18
6	Variablenauswahl	19
6.1	Zielvariable(n)	19
6.2	Die Prädiktortabelle	19
6.3	Datentyperkennung und Überschreibungen	20
6.4	Die Markierungen Factor und Linear	20
6.5	Referenz der speziellen Spaltentypen	20
6.6	Mehrere Zielvariablen	21
7	Parameterauswahl	22
7.1	Grundparameter	22
7.2	Vorwärtsdurchlauf (Forward Pass)	22
7.3	Allowed Interactions (zulässige Interaktionen)	23
7.4	Pruning	23
7.5	Kreuzvalidierung	23
7.6	Subset-Filterung	24
7.7	Empfohlene Parameterwerte	24
7.7.1	Parameter des Vorwärtsdurchlaufs	24
7.7.2	Pruning-Parameter	26
7.7.3	Kreuzvalidierungs-Parameter	26
7.7.4	Varianzmodell	27
7.7.5	Earth-Ausgabedatei	27
7.8	Voreinstellungen (Settings Defaults)	27
8	Anpassung des Earth-Modells	28
8.1	Die Schaltfläche „Fit“	28
8.2	Anpassungs-Modal & Trace-Ausgabe	28
8.3	Anpassungsprotokoll	28

9	Herunterladen der Daten	29
9.1	Ausgabespalten	29
9.2	Spaltenanordnung & Excel-Formatierung	29
9.3	CQA-Scores	30
9.4	Sortierung & Zeile des Bewertungsobjekts	30
9.5	Automatischer mgcvUI-Export	30
9.6	glmnetUI-Import	31
9.6.1	Wesentliche Vorteile	31
9.6.2	Wichtig: Die Gewichtungsspalte muss übereinstimmen	31
9.6.3	Arbeitsablauf	31
10	RCA-Berechnungen & Download	33
10.1	Öffnen des RCA-Dialogs	33
10.2	Interpolation des CQA-Scores	33
10.3	Ausgabespalten	33
10.4	Die Registerkarte RCA Adjustments	34
11	Sales Comparison Grid	35
11.1	Überblick & Zweck	35
11.2	Modaldialog zur Auswahl der Vergleichsobjekte	35
11.3	Rasterlayout	35
11.4	Sale Price, Concessions & Net SP	36
11.5	Gruppierte Zeilen (Location, Site, Age)	36
11.6	Modellvariablenzeilen	36
11.7	CQA Residual & Zeilen für Residualmerkmale	37
11.8	Formel für den Adjusted Sale Price	37
11.9	Blattschutz & editierbare Zellen	37
11.10	Arbeiten mit dem Raster in Excel	37
12	Herunterladen von Berichten	39
12.1	Generate Quarto Report	39
12.2	Convert Quarto Report	39
12.3	Berichtsinhalte	39
13	Die Post-Processing-Module: Elastic Net und GAM	41
13.1	Wozu unterstützende Anpassungen?	41
13.2	Start und gemeinsamer Zustand	41
13.3	Die Elastic-Net-Routine	41
13.4	Die GAM-Routine	41
13.5	Ein praktischer Arbeitsablauf	42
14	Demo-Datensatz: Appraisal_1.csv	43
14.1	Beschreibung	43
14.2	Spalten	44
14.3	Empfohlener Schnelleinstieg	44
15	Experimentell: Prolog-Textextraktion & Ableitungsregeln	46
15.1	Funktionsweise	46
15.2	Voraussetzungen: vProlog (Trealla), nicht SWI-Prolog	47
15.3	Aktivierung und Ausführung	47
15.4	Ableitungsregeln	47
15.5	Versionierte aktive Datendatei	48

15.6	Programmatische Schnittstelle & mitgelieferte Dateien	48
15.7	Verwandtes Experiment: Klimaregionen-Priors	48
15.7.1	Über Kalifornien hinaus: Jurisdiction Packs und Market-Area-Module . .	48

1 Einleitung

1.1 Was ist earthUI?

earthUI ist eine grafische Benutzeroberfläche für das R-Paket **earth**, das Multivariate Adaptive Regression Splines (MARS) implementiert (Friedman 1991). Die Modellierungsmaschine ist Stephen Milborrows Paket **earth** (Milborrow 2024), das von Milborrow (o.D.[a]) und Milborrow (o.D.[b]) ausführlich dokumentiert wird. Es läuft als lokale Shiny-Anwendung — es gibt keine Anmeldung, keinen Server und keine Benutzerkonten. Die Anwendung wird aus R heraus gestartet; anschließend wird ein Datensatz (CSV oder Excel) importiert, das Modell konfiguriert und interaktiv angepasst.

Die Anwendung stellt einen vollständigen Arbeitsablauf bereit: Datenimport, Variablenkonfiguration, Modellanpassung, Diagnosegrafiken, Variablenwichtigkeit, Modellgleichungen sowie herunterladbare Berichte im HTML-, Word- oder PDF-Format.

1.2 Ein Paket, drei Routinen

Seit der aktuellen Version *enthält* das Paket earthUI zudem seine beiden Post Processing Modules (Nachverarbeitungsmodule), die in früheren Versionen die separaten Pakete glmnetUI und mgcvUI waren. Mit der Installation von earthUI werden alle drei Routinen installiert; jede läuft als eigene lokale Shiny-Anwendung:

- **earthUI::launch()** — die in diesem Artikel beschriebene earth-(MARS-)Anwendung auf Port 7878. **earth()** ist die zentrale Bewertungsmaschine.
- **earthUI::launch_glmnet()** — das Elastic-Net Post Processing Module (Port 7879), das die Basisfunktionen des earth-Modells importiert und unter Lasso-/Elastic-Net-Regularisierung neu schätzt (siehe „glmnetUI Import“ weiter unten).
- **earthUI::launch_mgcv()** — das GAM Post Processing Module (Port 7880), das die Knotenpositionen des earth-Modells als Startpunkte für mgcv-Glättungsterme verwendet (siehe „mgcvUI Auto-Export“ weiter unten).

Die drei Routinen teilen sich einen Projektbaum, eine Einstellungsdatenbank und eine Kopie der unterstützenden Infrastruktur, sodass ein in einer Anwendung geöffnetes Projekt den anderen sofort zur Verfügung steht. Die glmnet- und mgcv-Routinen dienen als stützende Evidenz für die Wertschlussfolgerung des earth-Modells und stellen keine konkurrierenden Arbeitsabläufe dar: Das vorgesehene Vorgehen besteht darin, zunächst das earth-Modell anzupassen und zu verfeinern und anschließend — wenn eine Bestätigung gewünscht wird — die Module auf dasselbe Projekt anzuwenden.

1.3 Drei Purpose-Modi

earthUI organisiert die Arbeit in **Projekten** — benannten Einheiten, die Eingabedateien, Ausgaben und Einstellungen an einem festen Ort zusammenhalten. Jedes Projekt trägt einen **Purpose** (Verwendungszweck), der im New Project Wizard (Assistent zur Projektanlage) bei der Erstellung des Projekts gewählt wird. Es gibt drei Modi:

- **General** (Allgemein) — Earth-Regression für beliebige Populationen oder Datensätze. Dies ist der Standardmodus. Er stellt den vollständigen MARS-Modellierungsablauf ohne domänenspezifische Ergänzungen bereit.

- **For Appraisal** (für die Wertermittlung) — Earth-Regression, zugeschnitten auf die Wertermittlung von Immobilien. Ergänzt Funktionen speziell für die Bewertung einzelner Objekte, darunter die Behandlung des Bewertungsobjekts, spezielle Spaltenkennzeichnungen, den Residual Constraint Approach (RCA) (dt. etwa: Residualrestriktionsansatz) sowie die Erzeugung des Sales Comparison Grid (Vergleichsraster).
- **Market Area Analysis** (Marktgebietsanalyse) — Earth-Regression, zugeschnitten auf Marktgebietsstudien. Ergänzt Funktionen zur Analyse von Objektgruppen in einem definierten Markt anstelle der Bewertung eines einzelnen Bewertungsobjekts.

In allen drei Modi ist die zentrale Modellierungsmaschine identisch — es wird stets ein Earth-(MARS-)Modell angepasst. Die Purpose-Einstellung steuert, welche zusätzlichen Werkzeuge und Oberflächenelemente verfügbar sind.

Der Purpose ist je Projekt fest. Der Verwendungszweck eines Projekts — ebenso wie sein Land, Bundesstaat, County und seine Stadt — wird bei der Erstellung des Projekts festgeschrieben. Beim Wechsel zu einem anderen Projekt wird earthUI auf einen sauberen Ausgangszustand zurückgesetzt (importierte Daten, Modellergebnisse, Tabs, Variablenkonfiguration und earth-Parameter werden sämtlich gelöscht); anschließend wird die für den Purpose dieses Projekts gespeicherte Konfiguration wiederhergestellt.

1.4 Immobilienspezifische Funktionen

Wenn entweder **For Appraisal** oder **Market Area Analysis** gewählt ist, aktiviert earthUI mehrere Funktionen, die für die Immobilienanalyse konzipiert sind:

- **Spezielle Spaltenkennzeichnungen** — Jeder Prädiktor kann mit einer speziellen Rolle versehen werden, etwa `contract_date`, `dom`, `concessions`, `latitude`, `longitude`, `living_area`, `lot_size`, `actual_age`, `effective_age`, `area`, `site_dimensions` oder `display_only`. Diese Kennzeichnungen steuern, wie die Spalte bei der Modellanpassung, der Ausgabe und der Erzeugung des Sales Grid behandelt wird. Die vollständige Referenz findet sich in Kapitel 6.
- **Rundung von geografischer Breite und Länge** — Spalten, die als `latitude` oder `longitude` gekennzeichnet sind, werden automatisch auf 3 Dezimalstellen gerundet, um Überanpassung zu vermeiden.
- **Sale-Age-Spalte** — Wenn eine Spalte als `contract_date` gekennzeichnet ist und ein Effective Date (Wertermittlungstichtag) angegeben wird, berechnet earthUI eine Spalte `sale_age` (Tage zwischen Verkaufsdatum und Stichtag) und setzt sie als Prädiktor ein.
- **RCA-Berechnungen (nur Wertermittlung)** — Im Appraisal Mode kann earthUI nach der Modellanpassung Ausgaben des Residual Constraint Approach (RCA) berechnen; diese umfassen Anpassungen je Vergleichsobjekt, Zusammenfassungen der Netto-/Bruttoanpassungen sowie einen angepassten Verkaufspreis für das Bewertungsobjekt.
- **Sales Comparison Grid (nur Wertermittlung)** — Erzeugt eine Excel-Arbeitsmappe mit mehreren Blättern und Formeln für die Anpassungen der Vergleichsobjekte, Aufschlüsselungen der Residualmerkmale und angepasste Verkaufspreise. Siehe Kapitel 11.

***Hinweis:** Der Purpose-Modus General eignet sich für beliebige Datensätze — aus Finanzwesen, Wissenschaft, Ingenieurwesen oder Immobilienwirtschaft. Die Modi für Wertermittlung und Marktanalyse fügen lediglich Komfortfunktionen für Immobilienfachleute hinzu.*

1.5 Installation von earthUI und vProlog

earthUI ist auf CRAN veröffentlicht; die Standardinstallation besteht daher aus einer Zeile in R:

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

Dieser Artikel beschreibt die aktuelle Version. Sollte CRAN zum Zeitpunkt der Lektüre noch eine ältere Version führen, lässt sich die Entwicklungsversion von GitHub installieren (man beachte das `subdir` — das Paket liegt im Ordner `pkg/` des Repositorys):

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

vProlog (optional). Die experimentelle Funktion zur Verarbeitung von Anmerkungen (siehe „Prolog Text Extraction“ gegen Ende dieses Artikels) verwendet das Paket `vProlog` — eine eigenständige, eingebettete Prolog-Maschine. Es ist bewusst *keine* zwingende Abhängigkeit: `earthUI` lässt sich ohne `vProlog` installieren und ausführen, und die Prolog-Funktionen bleiben schlicht deaktiviert, bis das Paket vorhanden ist. `vProlog` wird über das `r-universe`-Repository des Autors verteilt, das vorkompilierte Binärdateien für Windows und macOS bereitstellt (keine Compiler-Werkzeugkette erforderlich):

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

Eine systemweite Prolog-Installation (SWI-Prolog oder anderes) ist nicht erforderlich — die Trealla-Maschine ist in das Paket selbst einkompiliert. Alles Übrige in diesem Artikel funktioniert ohne `vProlog`.

1.6 Erste Schritte

Zur Verwendung von `earthUI`:

1. **Installation des Pakets** von CRAN, wie oben beschrieben.
2. **Start der Anwendung:** In R wird `earthUI::launch()` ausgeführt, oder von der Kommandozeile aus `Rscript -e 'earthUI::launch()'`. Die Anwendung öffnet sich im Webbrowser auf Port 7878. Alternativ lässt sich die Anwendung direkt über `http://localhost:7878` im Browser aufrufen. (Die Post Processing Modules starten auf dieselbe Weise: `launch_glmnet()` auf Port 7879, `launch_mgcv()` auf Port 7880.)
3. **Erstellen oder Öffnen eines Projekts** in Section 1 der Seitenleiste. Der New Project Wizard erfasst den Projektnamen, den Purpose (General, For Appraisal oder Market Area Analysis), den Standort (Land, Bundesstaat, County, Stadt) und optional eine anfängliche Datendatei, die in den Ordner `in/` des Projekts kopiert wird.
4. **Import der Daten** in Section 2, wo die CSV- und Excel-Dateien im Ordner `in/` des aktiven Projekts aufgelistet werden. Dateien, die manuell über Finder oder Explorer hinzugefügt werden, erscheinen nach einem Klick auf Refresh (Aktualisieren).
5. **Konfiguration der Variablen** — Zielvariable(n) und Prädiktoren wählen, Datentypen festlegen und gegebenenfalls spezielle Spaltenrollen zuweisen.
6. **Festlegen der Earth-Parameter** — degree, nprune, Teilmengenfilter und weitere Argumente von `earth()`.
7. **Anpassen des Modells** — auf „Fit Earth Model“ klicken und die Ergebnisse im Hauptbereich prüfen.
8. **Export** — Prognosen als Excel herunterladen, Quarto-Berichtspakete erzeugen oder (im Appraisal Mode) RCA-Anpassungen und Sales Comparison Grids berechnen.

Die Konfiguration wird serverseitig in einer Projektdatenbank gespeichert, indiziert nach Projekt und Purpose, über die Schaltflächen „Save current as default“ in den Sections 3 und 4 (siehe Kapitel 3, „Settings Persistence“).

2 Anforderungen an MLS-Eingabedaten

Für Arbeitsabläufe der Immobilienwertermittlung und Marktanalyse stammen die Eingabedaten typischerweise aus einem Export eines Multiple Listing Service (MLS). Dieses Kapitel beschreibt die erwartete Dateistruktur sowie die Spalten, die earthUI verwenden kann.

2.1 Dateiformat & Struktur

earthUI akzeptiert **CSV**- und **Excel**-Dateien (**.xlsx**, **.xls**). Beim Import werden die Spaltennamen automatisch in **snake_case** umgewandelt — so wird beispielsweise aus „Living SqFt“ **living_sqft**, aus „Contract Date“ **contract_date** und aus „Sale Price“ **sale_price**. Diese Normalisierung stellt konsistente Spaltenreferenzen über den gesamten Arbeitsablauf hinweg sicher. Das beim Import verwendete CSV-Trennzeichen und das Dezimalzeichen werden durch die Locale-Einstellungen bestimmt (siehe Kapitel 3, „Locale & Regional Settings“).

Die Datendatei sollte eine flache Tabelle mit einer Zeile pro Objekt und einer Spalte pro Attribut sein. Die erste Zeile der Datei muss die Spaltenüberschriften enthalten.

2.2 Erforderliche Spalten für den Appraisal-Modus

Zwar arbeitet earthUI mit einem beliebigen Satz von Spalten, doch profitiert der vollständige Wertermittlungs-Arbeitsablauf (RCA-Anpassungen (Residual Constraint Approach, dt. etwa: Residualrestriktionsansatz) + Sales Comparison Grid) davon, wenn der MLS-Export die folgenden Spalten enthält:

Spalte	Spezialtyp	Zweck
Sale Price	(target)	Zielvariable für das Modell
Contract Date	contract_date	Zur Berechnung von sale_age (Tage ab dem Wertermittlungsstichtag)
Listing Date	listing_date	Zusammen mit dem Vertragsdatum zur Berechnung der DOM, falls keine DOM-Spalte vorhanden ist
Days on Market	dom	Vermarktungsdauer in Tagen; im Sales Grid (Vergleichsraster) angezeigt
Concessions	concessions	Verkaufszugeständnisse; Net SP = Sale Price – Concessions im Sales Grid
Living Area (SF)	living_area	Ermöglicht Residuen pro Quadratfuß (residual_sf , cqa_sf)
Lot Size	lot_size	In der Zeile SSite Size Dimensions des Sales Grid gruppiert
Site Dimensions	site_dimensions	Mit der Grundstücksgröße im Sales Grid gruppiert
Latitude	latitude	Auf 3 Dezimalstellen gerundet; für die Distanzberechnung und die Location-Gruppe verwendet
Longitude	longitude	Auf 3 Dezimalstellen gerundet; für die Distanzberechnung und die Location-Gruppe verwendet
Area ID	area	In der Zeile LLoc: Long Lat Area des Sales Grid gruppiert
Actual Age	actual_age	In der Zeile Äctual Age Effective Age des Sales Grid gruppiert
Effective Age	effective_age	Mit dem tatsächlichen Alter im Sales Grid gruppiert
Parcel Number	(none)	In der APN-Zeile des Sales Grid angezeigt
Listing ID	(none)	In der APN-Zeile des Sales Grid angezeigt
Address	display_only	In der Address-Zeile des Sales Grid angezeigt; vom Modell ausgeschlossen

Die Spaltennamen der Tabelle können in einer Fremdsprache vorliegen — die „speziellen“ Namen sind auf Englisch gehalten, damit das R-Programm sie gesondert behandeln kann. Ansonsten erscheinen die vergebenen Spaltennamen in den Regressionsmodellen, den Grafiken und (bei Gutachten) im Intermediate Sales Grid.

Nicht alle Spalten sind erforderlich. earthUI passt sich an — fehlt eine Spalte, wird das entsprechende Merkmal schlicht weggelassen. Wird beispielsweise keine **concessions**-Spalte zugewiesen, zeigt die Net-SP-Zeile im Sales Grid den Verkaufspreis ohne Abzug von Zugeständnissen. Für Preismodelle im Immobilienbereich werden jedoch bestimmte Spalten dringend empfohlen, um eine akzeptable Anpassungsgüte zu erreichen:

1. **Sale Age** (Verkaufsalter) — die Anzahl der Tage zwischen dem Vertragsdatum des Verkaufs und dem Stichtag der Wertermittlung oder Analyse. Wird eine mehrjährige Verkaufshistorie verwendet, insbesondere über Zeiträume von mehr als 5 Jahren, spielt **sale_age** häufig eine zentrale Rolle bei der Schätzung des Verkaufspreises. Tatsächlich ist diese Variable oft so bedeutsam, dass earthUI ohne sie überhaupt kein Modell liefern kann.
2. **Living Area** (Wohnfläche) — auch unter Bezeichnungen wie „Living Sqft“, „GLA“ (gross living area) und dergleichen bekannt. Dies ist eine weitere maßgebliche Bestimmungsgröße des Verkaufspreises.
3. **Total Bath Count** (Gesamtzahl der Bäder) — die Gesamtzahl der Voll-, Viertel-, Halb- und Dreiviertelbäder. Beispielsweise ergäben zwei Vollbäder und ein Halbbad den Wert 2.5.

4. **Garage Bays** (Garagenstellplätze) oder **Garage Area** (Garagenfläche) — die Anzahl der Garagenstellplätze oder die Garagenfläche in Quadratfuß.
5. **Lot Size** (Grundstücksgröße) — die Grundstücksfläche des Objekts, typischerweise in Quadratfuß oder Acres.
6. **Longitude**, **Latitude** und, sofern verfügbar, **Area ID**. Die Anpassungen für diese Variablen werden im Sales Grid zu einer einzigen Location-Anpassung (Lageanpassung) zusammengefasst.

2.3 Besondere Konventionen für Spaltennamen

earthUI identifiziert Spalten anhand ihrer **Spezialtyp-Zuweisung**, nicht anhand ihres Spaltennamens. Die Spalten im MLS-Export können beliebig benannt werden — entscheidend ist, dass in der Variable Configuration table (Variablenkonfigurationstabelle, Kapitel 6) der korrekte Spezialtyp zugewiesen wird.

Beispielsweise kann ein MLS die Wohnfläche als „GLA“, „Living SqFt“, „liv_area“ oder „gross_living_area“ exportieren. Nach dem Import (bei dem der Name in snake_case umgewandelt wird) wird die Spalte einfach im Special-Dropdown als `living_area` ausgewiesen. earthUI verwendet sie dann unabhängig von ihrem ursprünglichen Namen für die Berechnung der Residuen pro Quadratfuß und die Gruppierung im Sales Grid.

2.4 Datenqualität & Vollständigkeit

- **Fehlende Werte (NA)**: Zeilen mit NA-Werten in einer Prädiktor- oder Zielspalte werden vor der Modellanpassung automatisch entfernt. Intern ist `na.action` stets auf `na.fail` gesetzt; die Entfernung der NA-Werte erfolgt vor dem Aufruf von `earth()`.
- **Datumsspalten**: Datumsangaben sollten in einem Format vorliegen, das R interpretieren kann (z. B. 2025-06-15, 06/15/2025, June 15, 2025). earthUI erkennt Datumsspalten automatisch, wenn mindestens 80% der Werte erfolgreich interpretiert werden können. Zweistellige Jahresangaben (z. B. 06/15/25) werden als das nächstliegende passende Jahr gelesen, sodass aktuelle MLS-Exporte korrekt eingelesen werden; vierstellige Jahresangaben werden stets wörtlich übernommen.
- **Numerische Spalten**: Verkaufspreis, Wohnfläche, Grundstücksgröße, Zugeständnisse und ähnliche Felder müssen numerisch sein. Exportiert das MLS Preise mit Währungssymbolen oder Tausendertrennzeichen (z. B. „\$350,000“ oder „350.000,00“), müssen diese gegebenenfalls vor dem Import bereinigt werden.
- **Faktorspalten**: Kategoriale Variablen wie Area ID, Stil oder Zustand sollten eine überschaubare Anzahl eindeutiger Ausprägungen enthalten. earthUI erkennt Faktoren automatisch, die Erkennung kann jedoch übersteuert werden.

Tipp: Nach dem Import empfiehlt sich ein Blick auf die T/F/NA-Spalte in der Prädiktortabelle — sie zeigt die Anzahl der true-/false-/NA-Werte je Spalte. Spalten mit mehr als 30% fehlenden Werten werden rot markiert. Spalten mit hohem NA-Anteil sollten ausgeschlossen oder die Daten vor dem Import bereinigt werden.

2.5 Platzierung der Zeile des Bewertungsobjekts

Im **Appraisal-Modus** (Gutachtenmodus) **muss Zeile 1 das Bewertungsobjekt enthalten**. Alle übrigen Zeilen sind Vergleichsverkäufe. Der Verkaufspreis des Bewertungsobjekts kann leer gelassen (NA) oder auf einen beliebigen Wert gesetzt werden — earthUI behandelt ihn während der Modellanpassung in jedem Fall als NA.

Im **Market Area Analysis-Modus** (Marktgebietsanalyse) ist die Platzierung eines Bewertungsobjekts in Zeile 1 optional, und es wird keine Zeile automatisch ausgeschlossen. Um

die Zeile des Bewertungsobjekts (oder eine beliebige andere Zeile) aus der Modellanpassung herauszuhalten, ist ihr in einer dafür vorgesehenen Gewichtsspalte der Wert null zuzuweisen oder ein Subset-Filter zu verwenden.

Im **General**-Modus (allgemeiner Modus) gibt es keine besondere Zeilenbehandlung — alle Zeilen werden gleich behandelt.

3 General Purpose Mode

3.1 Überblick

Der General Purpose Mode (Allzweckmodus) ist der Standardmodus beim Start von earthUI. Er stellt den vollständigen MARS-Modellierungs-Arbeitsablauf für beliebige Datensätze bereit — nicht nur für Immobiliendaten. earthUI kann für wissenschaftliche Daten, Finanzanalysen, ingenieurwissenschaftliche Studien oder jedes Regressionsproblem eingesetzt werden, bei dem nichtlineare Zusammenhänge und Interaktionen zwischen Variablen untersucht werden sollen.

Im General Mode blendet die Oberfläche die immobilienpezifischen Funktionen aus (Spezialspalten, Verkaufsalter, Koordinatenrundung, RCA). Die Sidebar ist gestrafft und konzentriert sich auf Variablenauswahl, Parameterkonfiguration, Modellanpassung und Export.

3.2 Der Arbeitsablauf in der Sidebar

Die Sidebar ist in nummerierte, einklappbare Abschnitte gegliedert, die von der Projektauswahl bis zum Export führen:

1. Project — Auswahl eines bestehenden Projekts oder Anlage eines neuen Projekts über den New Project Wizard (Assistent für neue Projekte), der Zweck, Land und Verwaltungsstandort erfasst (nach der Anlage sämtlich gesperrt). Das aktive Projekt bestimmt, von wo Eingabedateien gelesen und wohin alle Ausgaben geschrieben werden — ein separates Feld für den Ausgabeordner existiert nicht.

2. Import Data — Auswahl einer CSV- oder Excel-Datei aus dem Ordner `in/` des aktiven Projekts; ein Refresh-Link durchsucht den Ordner erneut nach manuell hinzugefügten Dateien. Bei Excel-Dateien mit mehreren Blättern erscheint eine Blattauswahl. Spaltennamen werden automatisch in `snake_case` umgewandelt.

3. Variable Configuration — Auswahl der Zielvariablen (mehrere Zielvariablen werden unterstützt), Effective Date (Wertermittlungstichtag) sowie die Prädiktortabelle mit Kontrollkästchen für Include, Factor, Linear und Latent. Ausführliche Einzelheiten enthält Kapitel 6.

4. Earth Call Parameters — Alle Argumente der Funktion `earth()`: degree, penalty, nk, Pruning-Methode, Kreuzvalidierung und weitere. Die vollständige Parameterreferenz enthält Kapitel 7.

Execute Prolog Processing — Ein optionaler, unnummerierter Abschnitt zwischen den Abschnitten 4 und 5, der deaktiviert ist, sofern die experimentelle Prolog-Funktion nicht in den Settings aktiviert wurde. Siehe Kapitel 14.

5. Fit Earth Model — Ein Feld für den Zufallsstartwert (random seed) und die Schaltfläche, die die Modellanpassung asynchron ausführt. Einzelheiten zur Anpassung enthält Kapitel 8.

6. Download Estimated Target Variable(s) & Residuals — (in den Modi für Wertermittlung und Marktanalyse mit „Download Estimated Sale Prices & Residuals“ betitelt) Exportiert Vorhersagen, Residuen, CQA-Scores und die Wertbeiträge je g-Funktion als Excel-Datei. Siehe Kapitel 9.

7–8. RCA Adjustments und Sales Grid — (Vergleichsraster) Nur im Appraisal Mode. Siehe Kapitel 10 und 11.

9. Generate Quarto Report — Schreibt ein in sich geschlossenes Quarto-Berichtsbündel (`.qmd`-Quelldatei nebst Assets) in den Ausgabeordner des Projekts. Außerhalb des Appraisal Mode mit 7 nummeriert. Siehe Kapitel 12.

10. Convert Quarto Report — Rendert eine beliebige `.qmd`-Datei nach HTML, Word oder PDF. Siehe Kapitel 12.

3.3 Tabs des Hauptpanels

Nach der Modellanpassung stellt das Hauptpanel zehn Tabs bereit:

Tab	Inhalt
Data	Vorschau der importierten Daten. Im Appraisal Mode aufgeteilt in Tabellen für das Bewertungsobjekt (Subject Property) und die Vergleichsverkäufe (Comparable Sales).
Equation	Die angepasste Modellgleichung, dargestellt in LaTeX/MathJax, mit jeder Basisfunktion und ihrem Koeffizienten.
Summary	Zentrale Kennzahlen (R^2 , $GRSq$, GCV , RSS , $CV R^2$) sowie eine Koeffiziententabelle.
Variable Importance	Balkendiagramm und Rangtabelle der Wichtigkeitswerte der Prädiktoren.
Contribution	Diagramme der g-Funktions-Wertbeiträge. Auswahl einer g-Funktion über das Dropdown-Menü; univariate Prädiktoren werden als Liniendiagramme dargestellt, bivariate Interaktionen ($\text{degree} \geq 2$) als 3D-Perspektiv- und Konturdiagramme.
Correlation	Heatmap der Prädiktorkorrelationen.
Diagnostics	Drei Diagramme: Residuals vs Fitted, Normal Q-Q und Actual vs Predicted.
RCA Adjustments	(nur Wertermittlung) Histogramme von Residual Adj %, Net Adj % und Gross Adj % über die Vergleichsobjekte hinweg, befüllt nach Abschluss von Schritt 7 (Calculate RCA Adjustments).
Residual Contribution	Schätzt die Marktwertbeiträge seltener oder ausgeschlossener Merkmale, indem die Modellresiduen gemeinsam auf Spalten regressiert werden, die das Modell nie verwendet hat (siehe subsection 4.4).
ANOVA	ANOVA-Zerlegungstabelle, die den Beitrag jeder Basisfunktion zur RSS ausweist.
Earth Output	Unformatierte Textausgabe von 'earth::summary()', einschließlich der Details des Pruning-Durchlaufs.

3.4 Persistenz der Einstellungen

earthUI speichert die Konfiguration serverseitig in einer Projektdatenbank (`projects.sqlite` im `regProj`-Stammverzeichnis), indiziert nach Projekt und Zweck. Das Speichern erfolgt **schaltflächen-gesteuert**: Die Schaltfläche „Save current as default“ in Abschnitt 3 speichert Zielvariable, Stichtag und Prädiktorkonfiguration, während die Schaltfläche in Abschnitt 4 die Parameter des earth-Aufrufs sowie die Allowed-Interactions-Matrix speichert. Jede Schaltfläche aktualisiert nur ihre eigene Hälfte, sodass sie einander niemals überschreiben, und **die Modellanpassung speichert nicht automatisch** — es kann frei experimentiert werden, ohne die gespeicherten Standardwerte zu überschreiben. Die verschiedenen Eingabedateien eines Projekts (ein kleiner Testauszug, der vollständige Datensatz) teilen sich eine Konfiguration, und da die Einstellungen im Projektbaum liegen, wandern sie mit, wenn der `regProj`-Ordner synchronisiert, gesichert oder weitergegeben wird.

3.5 Dark Mode

Ein Klick auf das Mond-/Sonnensymbol in der oberen rechten Ecke wechselt zwischen hellem und dunklem Design. Die Design-Präferenz wird im lokalen Speicher (Local Storage) abgelegt und bleibt über Sitzungen hinweg erhalten.

3.6 Locale & regionale Einstellungen

earthUI unterstützt internationale Formatierungskonventionen für Zahlen, Datumsangaben und CSV-Dateien über ein länderbasiertes Locale-System. Das Dropdown-Menü **Country** im Menü **Settings** (Zahnradsymbol) der oberen Menüleiste wählt eine Voreinstellung für 31 unterstützte Länder aus. Jede Voreinstellung konfiguriert:

- **CSV-Trennzeichen** — Komma (,) für USA/UK/Japan oder Semikolon (;) für den Großteil Europas, wo das Komma als Dezimalzeichen dient.
- **Dezimalzeichen** — Punkt (.) oder Komma (,).
- **Tausendertrennzeichen** — Komma (USA/UK/Japan), Punkt (Deutschland/Italien/Spanien), Leerzeichen (Finnland/Frankreich/Polen/Baltikum/Ukraine/Russland) oder Apostroph (Schweiz).
- **Datumsformat** — MM/DD/YYYY (USA), DD/MM/YYYY (Großteil Europas) oder YYYY-MM-DD (Schweden/Litauen/Japan/Kanada).
- **Papierformat** — Letter (USA/Kanada/Mexiko) oder A4 (überall sonst).

3.6.1 Unterstützte Länder

Land	CSV-Trennz.	Dezimal	Tausender	Datum	Papier
USA, Kanada (csv)	,	.	,	MDY/YMD	Letter
UK, Irland, AU, NZ	,	.	,	DMY	A4
Deutschland, Österreich, Belgien, Niederlande, Italien, Spanien, Portugal, Dänemark, Brasilien	;	,	.	DMY	A4
Frankreich, Polen, Tschechien, Norwegen, Lettland, Estland	;	,	(Leerzeichen)	DMY	A4
Finnland, Schweden	;	,	(Leerzeichen)	DMY/YMD	A4
Schweiz	;	,	,	DMY	A4
Ukraine, Russland	;	,	(Leerzeichen)	DMY	A4
Türkei	;	,	.	DMY	A4
Japan, Südkorea	,	.	,	YMD	A4
Mexiko	,	.	,	DMY	Letter
Litauen	;	,	(Leerzeichen)	YMD	A4

3.6.2 Übersteuerung des Papierformats

Unterhalb der Länderauswahl ermöglicht ein Dropdown-Menü **Paper** (Letter oder A4), das Seitenformat des Berichts zu übersteuern, ohne das Land zu wechseln. CSV-Trennzeichen, Dezimalzeichen und die Reihenfolge des Datumsformats folgen der Voreinstellung des gewählten Landes; beim Wechsel des Landes wird die Papierformat-Übersteuerung auf den Standardwert dieses Landes zurückgesetzt.

3.6.3 Speichern der Standardeinstellungen

Ein Klick auf **Save** im Settings-Menü speichert die Locale-Präferenzen global. Diese Standardwerte gelten für alle künftigen Sitzungen, unabhängig davon, welches Projekt geöffnet wird. Projektspezifische Einstellungen (Zielvariable, Prädiktoren, Parameter) werden separat in der Projektdatenbank gespeichert; die Locale-Standardwerte bleiben über alle Projekte hinweg in einer benutzerbezogenen SQLite-Datenbank erhalten. Dieser zweistufige Ansatz ist für Organisationen wie Wirtschaftsprüfungsgesellschaften konzipiert, die mit Daten aus mehreren Ländern arbeiten — das am häufigsten benötigte Land wird als Standard festgelegt und bei Bedarf angepasst.

Hinweis: Die Zahlenformatierung auf Diagrammachsen und Steigungsbeschriftungen passt sich automatisch an das eingestellte Locale an. Das deutsche Locale verwendet Punkte als Tausendertrennzeichen (200.000), das finnische Leerzeichen (200 000), das schweizerische Apostrophe (200'000). Währungssymbole werden nicht angezeigt — earthUI ist währungsneutral.

4 Appraisal Mode (Gutachtenmodus)

Wird als Purpose (Zweck) die Option **For Appraisal** gewählt, konfiguriert sich earthUI für die Bewertung eines einzelnen Objekts. Alle in Kapitel 3 beschriebenen Funktionen bleiben verfügbar; das vorliegende Kapitel behandelt ausschließlich die gutachtenspezifischen Ergänzungen.

4.1 Behandlung der Subject-Zeile

Im Appraisal Mode ist **Zeile 1 des Datensatzes das Bewertungsobjekt**, und alle übrigen Zeilen sind Vergleichsverkäufe. Die Eingabedatei muss entsprechend organisiert sein (siehe Kapitel 2). Der Verkaufspreis des Bewertungsobjekts kann leer gelassen oder auf einen beliebigen Wert gesetzt werden — earthUI behandelt ihn während der Anpassung automatisch als NA.

Nach dem Import teilt sich der Data tab in zwei Abschnitte: **Subject Property** (Zeile 1) und **Comparable Sales** (Zeilen 2+). Zeile 1 wird stets von der Modellanpassung ausgeschlossen — die Benachrichtigung „Skipping row 1 (subject). Fitting on N rows.“ bestätigt dies. Nach der Anpassung erzeugt das Modell dennoch Prognosen für die Subject-Zeile, die in der Ausgabe als `est_<target>` ausgewiesen werden.

4.2 Effective Date & Sale Age

Ein Feld **Effective Date** (Wertermittlungstichtag) erscheint im Abschnitt Variable Configuration in jedem Zweckmodus (mit dem heutigen Datum als Voreinstellung) — auch allgemeine Analysen und Marktanalysen verfügen über einen Stichtag, nicht nur Gutachten —, *verwendet* wird es jedoch nur im Appraisal Mode und im Market Mode. Wird dort eine Spalte im Dropdown-Menü Special als `contract_date` ausgewiesen, berechnet earthUI eine Spalte `sale_age` — die Anzahl ganzer Tage zwischen dem Vertragsdatum des jeweiligen Verkaufs und dem Stichtag. Diese Spalte ersetzt die ursprüngliche Datumsspalte als Prädiktor.

Beim ersten Klick auf Fit nach der Ausweisung eines Vertragsdatums erstellt earthUI die Spalte `sale_age` und weist darauf hin, dass erneut auf Fit zu klicken ist, um sie einzubeziehen.

4.3 Special-Spalten-Ausweisungen

Im Appraisal Mode und im Market Mode erscheint für jeden Prädiktor in der Tabelle Variable Configuration ein Dropdown-Menü **Special**. Die vollständige Liste der Special-Typen und ihrer Wirkungen:

Special-Typ	Wirkung
no	Voreinstellung — keine besondere Behandlung
contract_date	Löst die automatische Berechnung von <code>sale_age</code> aus dem Effective Date aus
listing_date	Wird zusammen mit dem Vertragsdatum zur Berechnung von DOM verwendet, wenn keine Spalte <code>dom</code> existiert
dom	Kennzeichnet die Spalte „Days on Market“; wird im Sales Grid angezeigt
concessions	Kennzeichnet Verkaufszugeständnisse; wird in der Net-SP-Formel (Verkaufspreis – Zugeständnisse) im Sales Grid verwendet
latitude	Werte werden auf 3 Dezimalstellen gerundet; verwendet für die Haversine-Näheberechnung und die Gruppe Location im Sales Grid
longitude	Werte werden auf 3 Dezimalstellen gerundet; verwendet für die Haversine-Näheberechnung und die Gruppe Location im Sales Grid
area	Gruppiert in der Zeile LLoc: Long Lat Area im Sales Grid
living_area	Aktiviert Residuen je Quadratfuß (<code>residual_sf</code> , <code>cqa_sf</code>) in der Download-Ausgabe
lot_size	Gruppiert in der Zeile SSite Size Dimensions im Sales Grid
site_dimensions	Gruppiert mit der Grundstücksgröße im Sales Grid
actual_age	Gruppiert in der Zeile Actual Age Effective Age im Sales Grid
effective_age	Gruppiert mit dem tatsächlichen Alter im Sales Grid
display_only	Spalte wird in Exporte einbezogen, jedoch von der Modellanpassung ausgeschlossen (mehrfach zulässig)
remarks	(Experimentell) Freitext-Bemerkungsspalte, die durch den optionalen Prolog-Schritt in Merkmalsspalten zerlegt wird (mehrfach zulässig); siehe Kapitel 14
list	(Experimentell) Kommagetrennte Spalte, die durch den optionalen Prolog-Schritt in One-Hot-Spalten mit TRUE/FALSE aufgeteilt wird (mehrfach zulässig); siehe Kapitel 14

Pro Special-Typ ist nur eine Spalte zulässig (mit Ausnahme von `display_only`, `remarks` und `list`). Wird ein Special-Typ einer zweiten Spalte zugewiesen, wird er automatisch von der ersten entfernt. Neben dem Variablennamen erscheint eine kleine blaue Kennzeichnung, die den zugewiesenen Special-Typ anzeigt.

4.4 Residualer Wertbeitrag (Schätzung seltener Merkmale)

Merkmale, die für den Regressionskanal zu selten sind — etwa eine Werkstatt, die nur in einer Handvoll Verkäufe vorkommt —, gehören zur Residualschicht des RCA. Der Tab **Residual Contribution** (residualer Wertbeitrag) schätzt ihren Marktbeitrag, indem die Residuen des angepassten Modells mittels gewöhnlicher Kleinst-Quadrat-Schätzung auf Spalten regressiert werden, die das Modell nie verwendet hat. Die Auswahl mehrerer Spalten schätzt diese *gemeinsam*, sodass korrelierte Ausstattungsmerkmale (z. B. `outbuilding_sqft` zusammen mit `pr_has_workshop`) den Wertbeitrag untereinander aufteilen, statt ihn doppelt zu zählen. Binäre Indikatoren liefern Pauschalschätzungen und führen das Residuum jedes gekennzeichneten Verkaufs als Beleg auf; numerische Spalten liefern Wertbeiträge je Einheit; der ausgewiesene Standardfehler entspricht einem SE ($t \geq 2$ als üblicher Signifikanzmaßstab). Terme, die auf weniger als zehn Beobachtungen beruhen, erhalten automatisch den Vermerk CAUTION in der erzeugten Arbeitsdatei-Notiz, die Methode, Schätzwerte und Vorbehalte in einem Absatz für das Gutachten dokumentiert. Von der Kandidatenliste ausgeschlossen werden nur Prädiktoren, die das Modell tatsächlich *verwendet* hat — ein seltener Indikator, der der Anpassung angeboten, vom Vorwärtsdurchlauf jedoch nie ausgewählt wurde, bleibt wählbar, da das Modell keinen Teil

seines Wertes absorbiert hat.

4.5 Überblick über die RCA-Anpassungen

Die Schaltfläche **Calculate RCA Adjustments & Download** (Seitenleistenabschnitt 7, nur im Appraisal Mode nach der Anpassung sichtbar) berechnet marktabgeleitete Anpassungen für jedes Vergleichsobjekt relativ zum Bewertungsobjekt. Der vollständige RCA-Arbeitsablauf ist in Kapitel 10 beschrieben.

Nach der Berechnung der RCA-Anpassungen wird die Schaltfläche **Generate Sales Grid & Download** (Seitenleistenabschnitt 8) verfügbar. Der Sales-Grid-Arbeitsablauf ist in Kapitel 11 beschrieben.

***Hinweis:** Der dem Bewertungsobjekt zugewiesene CQA-Wert steuert, welcher Anteil der Residualverteilung dem Bewertungsobjekt zugerechnet wird. Ein Wert von 5.00 platziert das Bewertungsobjekt am Median der Vergleichsobjekte.*

5 Market Area Analysis Mode (Marktgebietsanalysemodus)

Wird als Purpose die Option **Market Area Analysis** gewählt, stellt earthUI dieselben immobilien-spezifischen Funktionen wie der Appraisal Mode bereit (Special-Spalten, Verkaufsalter, Rundung von Koordinaten), ist jedoch auf die Analyse einer Gruppe von Objekten ausgerichtet und nicht auf die Bewertung eines einzelnen Bewertungsobjekts.

5.1 Unterschiede zum Appraisal Mode

- **Kein automatischer Ausschluss des Bewertungsobjekts** — anders als der Appraisal Mode schließt der Market Mode Zeile 1 nicht automatisch von der Anpassung aus. Alle Zeilen werden einbezogen; um eine Subject-Zeile aus der Anpassung herauszuhalten, ist ihr in einer ausgewiesenen Gewichtsspalte der Wert null zuzuweisen (wie im Demonstrationsdatensatz) oder ein Teilmengenfilter zu verwenden.
- **Keine RCA- oder Sales-Grid-Abschnitte** — die Schritte „Calculate RCA Adjustments & Download“ und „Generate Sales Grid & Download“ sind nicht verfügbar. Der Market Mode konzentriert sich auf Modellanpassung und Ausgabe, nicht auf Anpassungen je Vergleichsobjekt.
- **Nummerierung der Seitenleiste** — ohne die RCA- und Sales-Grid-Schritte trägt der Abschnitt Generate Quarto Report die Nummer 7 statt 9 (der Abschnitt Convert Quarto Report behält die Nummer 10).

5.2 Wann der Market Mode zu verwenden ist

Der Market Area Analysis Mode ist angemessen, wenn

- ein Regressionsmodell für ein Wohnviertel oder Marktgebiet erstellt werden soll, um Werttreiber zu verstehen,
- analysiert werden soll, wie Variablen wie Wohnfläche, Alter, Grundstücksgröße und Lage die Verkaufspreise über eine Gruppe von Objekten hinweg beeinflussen,
- Belegmaterial für eine Marktbedingungsanalyse oder eine Abgrenzung von Wohnvierteln vorbereitet werden soll,
- mit einem Datensatz gearbeitet wird, der ein Bewertungsobjekt in Zeile 1 enthält, dessen Einbeziehung in die oder Ausschluss aus der Anpassung jedoch wahlweise möglich bleiben soll.

Hinweis: Der Market Mode eignet sich auch für allgemeine Immobilienregressionen, bei denen die Special-Spalten-Funktionen (Verkaufsalter, Rundung von Koordinaten) benötigt werden, der RCA-Anpassungsablauf jedoch nicht.

6 Variablenauswahl

In Abschnitt 3 der Seitenleiste — **Variable Configuration** (Variablenkonfiguration) — wird festgelegt, welche Spalten in das Modell eingehen und wie sie behandelt werden.

6.1 Zielvariable(n)

Das Dropdown-Menü **Target (response) variable(s)** (Ziel- bzw. Antwortvariable(n)) am Anfang von Abschnitt 3 führt jede Spalte des Datensatzes auf. Für ein gewöhnliches Modell mit einer einzelnen Antwortvariablen wird eine Spalte ausgewählt, für ein Mehrfachantwort-Modell mehrere Spalten. Werden mehrere Zielvariablen ausgewählt, passt das Modell alle Antwortvariablen simultan mittels `earth(cbind(y1, y2, ...) ~ .)` an.

Als Zielvariablen ausgewählte Spalten werden automatisch aus der Liste der Prädiktoren ausgeschlossen.

6.2 Die Prädiktortabelle

Unterhalb der Zielvariablenauswahl führt eine Tabelle jede verbleibende Spalte mit den folgenden Feldern auf:

Spalte	Beschreibung
Variable	Spaltenname (bei Abschneidung wird der vollständige Name im Tooltip angezeigt), dem ein $\times$ / $+$ - Umschalter vorangestellt ist (siehe unten). In den Modi Wertermittlung/Markt zeigt ein blaues Abzeichen den zugewiesenen Spezialtyp an.
Type	Dropdown-Menü für den Datentyp: <code>numeric</code> , <code>integer</code> , <code>character</code> , <code>logical</code> , <code>factor</code> , <code>Date</code> , <code>POSIXct</code>
Include	Kontrollkästchen — diese Spalte als Prädiktor in das Modell aufnehmen
Factor	Kontrollkästchen — diese Spalte als kategoriale Variable behandeln
Linear	Kontrollkästchen — ausschließlich linearen Eingang erzwingen (keine Hinge-Funktionen)
Latent	Kontrollkästchen — diese Spalte vollständig aus dem Modell heraushalten (ausgeschlossen aus <code>earth</code> und aus den nach <code>glmnetUI</code> / <code>mgcvUI</code> exportierten Daten) und für Arbeiten mit latenten Variablen vorbehalten
Special	Dropdown-Menü (nur Wertermittlung/Markt) — siehe die nachstehende Referenz der speziellen Spaltentypen
T/F/NA	Drei Zählwerte je Spalte: bei logischen Spalten TRUE/FALSE/NA; bei numerischen Spalten $\neq 0$ / $= 0$ /NA. Rot dargestellt, wenn mehr als 30 % der Werte fehlen

Eine Hinweiszeile oberhalb der Tabelle erläutert die Abkürzungen: „Type = column data type, Inc = include as predictor, Factor = treat as categorical, Linear = linear-only (no hinges), Special = column role (e.g. `contract_date`).“

Weiches Deaktivieren statt Löschen. Der $\times$ -Umschalter neben jedem Variablennamen graut die Zeile aus und hält die Spalte aus dem Modell heraus, ohne sie aus den Daten zu entfernen; er wechselt zu einem grünen $+$, das die Spalte wieder aktiviert. Spalten werden niemals zerstört.

6.3 Datentyperkennung und Überschreibungen

earthUI erkennt Datentypen beim Import automatisch. Numerische, ganzzahlige, logische, faktorielle und Datumsspalten werden erkannt. Zeichenspalten, die wie Datumsangaben aussehen (mindestens 80 % der Werte lassen sich gegen gängige Datumsformate parsen), werden als **Date** klassifiziert.

Jede Erkennung kann durch Änderung des Dropdown-Menüs **Type** überschrieben werden. Wird eine Spalte auf **character** oder **factor** geändert, wird das Kontrollkästchen Factor automatisch aktiviert. Typänderungen wirken sich darauf aus, wie die Spalte an die Funktion `earth()` übergeben wird.

6.4 Die Markierungen Factor und Linear

- **Factor** — Ist das Kontrollkästchen aktiviert, wird die Spalte als kategoriale Variable behandelt. Dies ist angemessen für Spalten wie **style**, **area_id** oder **grade**, die diskrete Gruppen und keine stetigen Messgrößen darstellen. Faktorspalten gehen als Indikatorvariablen (Dummy-Variablen) in das Modell ein.
- **Linear** — Ist das Kontrollkästchen aktiviert, wird der Spalte ein streng linearer Eingang in das Modell auferlegt — es werden für sie keine Hinge-Funktionen (stückweise lineare Funktionen) erzeugt. Dies ist nützlich, wenn bekannt ist, dass eine Variable in einem streng linearen Zusammenhang mit der Zielvariablen steht.

6.5 Referenz der speziellen Spaltentypen

In den Modi Wertermittlung und Markt stellt das Dropdown-Menü **Special** die folgenden Optionen bereit. Jeder Typ kann höchstens einer Spalte zugewiesen werden (mit Ausnahme von **display_only**, **remarks** und **list**, die mehrfach zulässig sind):

Datums- und Zeittypen:

- **contract_date** — Löst die automatische Berechnung von **sale_age** aus. Die ursprüngliche Datumsspalte wird durch eine ganzzahlige Spalte ersetzt, welche die Anzahl der Tage zwischen dem Verkaufsdatum und dem Wertermittlungstichtag (Effective Date) misst.
- **listing_date** — Dient als Rückfalloption zur Berechnung der Vermarktungsdauer (Days on Market, DOM = Vertragsdatum – Listing-Datum), wenn keine explizite **dom**-Spalte bestimmt ist.
- **dom** — Kennzeichnet die Spalte der Vermarktungsdauer (Days on Market). Wird in der APN-Zeile und der Verkaufsdatumszeile des Sales Grid (Vergleichsraster) angezeigt.

Monetäre Typen:

- **concessions** — Kennzeichnet Verkaufszugeständnisse (Verkäufergutschriften, Käuferanreize usw.). Wird im Sales Grid zur Berechnung des Nettoverkaufspreises verwendet: $\text{Net SP} = \text{Verkaufspreis} - \text{Zugeständnisse}$.

Größen- und Lagetypen:

- **latitude** — Die Werte werden automatisch auf 3 Dezimalstellen gerundet, um Überanpassung zu vermeiden. Wird für Haversine-Näherungsberechnungen (Entfernung vom Bewertungsobjekt zu jedem Vergleichsobjekt) verwendet und in der Lagezeile des Sales Grid gruppiert.
- **longitude** — Gleiche Rundungsbehandlung wie die geographische Breite. Wird für die Entfernungsberechnung und die Lagegruppe verwendet.
- **area** — Typischerweise eine Nachbarschafts- oder Gebietskennung. Wird gemeinsam mit geographischer Breite und Länge in der Zeile „Loc: Long | Lat | Area“ des Sales Grid gruppiert.

- **living_area** — Ermöglicht Residualberechnungen je Quadratfuß (**residual_sf** und **cqa_sf**) in der Download-Ausgabe.
- **lot_size** — Wird in der Zeile „Site Size | Dimensions“ des Sales Grid gruppiert.
- **site_dimensions** — Wird gemeinsam mit der Grundstücksgröße im Sales Grid gruppiert (z. B. „75x120“).

Alterstypen:

- **actual_age** — Wird in der Zeile „Actual Age | Effective Age“ des Sales Grid gruppiert.
- **effective_age** — Wird gemeinsam mit dem tatsächlichen Alter im Sales Grid gruppiert.

Anzeigetypen:

- **display_only** — Die Spalte wird in Excel-Exporte aufgenommen, jedoch vollständig von der Modellanpassung ausgeschlossen. Diese Kennzeichnung eignet sich für Adressfelder, MLS-Nummern, Flurstückskennungen oder sonstige Referenzdaten, die nicht als Prädiktor dienen sollen. Mehrere Spalten können diese Kennzeichnung tragen.

Typen zur Textextraktion (experimentell):

- **remarks** — Eine Freitextspalte (z. B. **public_remarks**), die durch den optionalen Prolog-Schritt in Merkmalspalten zerlegt wird. Ohne Aktivierung dieser Funktion in den Settings bleibt sie wirkungslos. Mehrere Spalten zulässig. Siehe Kapitel 14.
- **list** — Eine kommasetrennte Spalte (z. B. **existing_structures**), die durch denselben Schritt in One-Hot-Spalten der Form `<col>_<value>` mit TRUE/FALSE-Werten aufgespalten wird. Mehrere Spalten zulässig. Siehe Kapitel 14.

6.6 Mehrere Zielvariablen

Die Auswahl mehrerer Zielvariablen führt zur Anpassung eines Earth-Modells mit mehreren Antwortvariablen. Bei Auswahl mehrerer Zielvariablen gilt:

- Das Modell prognostiziert alle Antwortvariablen simultan, wobei die Basisfunktionen über die Zielvariablen hinweg geteilt werden
- Die Schaltfläche **wp (response weights)** (Antwortgewichte) wird aktiv und erlaubt die Zuweisung eines numerischen Gewichts an jede Zielvariable
- Varianzmodelle (**varmod.method**) sind deaktiviert (für Modelle mit mehreren Antwortvariablen nicht unterstützt)
- Die Ergebnisreiter zeigen Kennzahlen, Gleichungen und Diagnosedigramme je Antwortvariable an

Hinweis: Als **display_only** gekennzeichnete Spalten verbleiben im Datensatz und erscheinen im Excel-Export, werden jedoch vollständig aus der Prädiktortabelle ausgeschlossen. Diese Kennzeichnung eignet sich für ID-Spalten, Adressen oder sonstige Referenzfelder.

7 Parameterauswahl

Abschnitt 4 der Seitenleiste — **Earth Call Parameters** (Parameter des Earth-Aufrufs) — bietet Zugriff auf alle Argumente, die von der Funktion `earth()` akzeptiert werden. Jeder Parameter verfügt über ein blaues Hilfesymbol (?) mit einer Tooltip-Erläuterung. Die Parameter sind in einklappbare Unterabschnitte gegliedert.

7.1 Grundparameter

Parameter	Standardwert	Beschreibung
subset	(empty)	Ausdruck zur Zeilenfilterung. Siehe „Subset-Filterung“ weiter unten.
weights	NULL	Spaltenauswahl für Fall-(Zeilen-)Gewichte. Es werden nur numerische Spalten aufgeführt.
wp	NULL	Antwortgewichte für Modelle mit mehreren Zielvariablen. Die Schaltfläche öffnet einen Dialog mit einem numerischen Eingabefeld je Zielvariable (Standard jeweils 1.0).
keepxy	off	Behält x, y, subset und weights im Modellobjekt.
trace	0	Trace-Stufe für die Ausgabe während der Anpassung (0 bis 5).
glm	none	Optionale GLM-Familie: gaussian , binomial oder poisson .
degree	1	Maximale Interaktionsordnung. Ein Wert $\text{degree} \geq 2$ aktiviert automatisch die Kreuzvalidierung und blendet die Matrix Allowed Interactions (zulässige Interaktionen) ein.
penalty	2	GCV-Strafterm je Knoten. Höhere Werte führen zu einfacheren Modellen.

7.2 Vorwärtsthroughlauf (Forward Pass)

Parameter	Standardwert	Beschreibung
nk	auto	Maximale Anzahl an Termen vor dem Pruning. Wird automatisch mit dem für die geladenen Daten empfohlenen Wert vorbelegt (siehe „Empfohlene Parameterwerte“ weiter unten).
thresh	0.001	Schwellenwert des Vorwärtsschritts. Kleinere Werte lassen mehr Terme zu.

Parameter	Standardwert	Beschreibung
minspan	auto	Minimale Spannweite zwischen Knoten. Wird automatisch mit dem empfohlenen Wert vorbelegt. Negative Werte legen die maximale Knotenzahl je Prädiktor fest.
endspan	auto	Randspannweite — minimaler Abstand eines Knotens vom Rand der Daten. Wird automatisch mit dem empfohlenen Wert vorbelegt.
newvar.penalty	0	Strafterm für die Einführung einer neuen Variablen (begünstigt die Wiederverwendung bereits vorhandener Prädiktoren).
fast.k	20	Anzahl der im Fast-MARS-Algorithmus berücksichtigten Elternterme.
fast.beta	1	Steuert den Alterungsfaktor des Fast-MARS-Algorithmus.

7.3 Allowed Interactions (zulässige Interaktionen)

Wird **degree** auf 2 oder höher gesetzt, erscheint unterhalb der Grundparameter eine Interaktionsmatrix. Dabei handelt es sich um ein oberes Dreiecksraster aus Kontrollkästchen, eines für jedes Prädiktorpaar. Ein gesetztes Häkchen bedeutet, dass die beiden Prädiktoren interagieren dürfen; wird es entfernt, ist diese spezifische Interaktion untersagt.

Ein Klick auf einen Prädiktornamen (Zeilen- oder Spaltenbeschriftung) schaltet sämtliche Interaktionen dieser Variablen um. Die Kontrollkästchen **Allow All** und **Clear All** am oberen Rand ermöglichen eine Sammelsteuerung. Die Matrix verwendet fixierte Kopfzeilen (Sticky Headers), sodass Spalten- und Zeilenbeschriftungen beim Scrollen sichtbar bleiben.

Ein Hinweissfeld erinnert daran: „Interaction terms increase the risk of overfitting. Cross-validation has been enabled (10-fold).“ (Interaktionsterme erhöhen das Risiko der Überanpassung; die Kreuzvalidierung wurde aktiviert, 10-fach.)

7.4 Pruning

Parameter	Standardwert	Beschreibung
pmethod	backward	Pruning-Methode: backward , none , exhaustive , forward , seqrep oder cv .
nprune	NULL	Maximale Anzahl an Termen nach dem Pruning. Für die automatische Auswahl leer lassen.

7.5 Kreuzvalidierung

Parameter	Standardwert	Beschreibung
nfold	10	Anzahl der Folds (Teilstichproben) der Kreuzvalidierung. Ein Wert von 0 deaktiviert die CV.
ncross	20	Anzahl der Wiederholungen der Kreuzvalidierung.
stratify	on	Stratifiziert die CV-Stichproben, sodass jeder Fold eine ähnliche Verteilung der Zielvariablen aufweist.

Bei $\text{degree} \geq 2$ wird die Kreuzvalidierung automatisch aktiviert (nfold auf 10 gesetzt), um einer Überanpassung durch Interaktionen vorzubeugen.

7.6 Subset-Filterung

Das Texteingabefeld **subset** akzeptiert einen R-Ausdruck, der festlegt, welche Zeilen für die Modellanpassung verwendet werden. Der Ausdruck kann direkt eingegeben werden (z. B. `sale_age < 365 & area_id == 460`) oder mithilfe der Schaltfläche **Build filter...** visuell erstellt werden.

Build Filter Dialog — Ein Klick auf „Build filter...“ öffnet einen geführten Dialog. Jede Bedingungszeile besitzt eine Spaltenauswahlliste, einen Operator (`<`, `>`, `<=`, `>=`, `==`, `!=`) sowie ein Werteingabefeld, das sich dem Spaltentyp anpasst: numerische Eingabe für Zahlen, Datumsauswahl für Datumsangaben und eine Auswahlliste der eindeutigen Werte für Zeichen- bzw. Faktorspalten. Die Bedingungen werden mit AND- (&) oder OR-Verknüpfungen (|) kombiniert. Eine Vorschau am unteren Rand zeigt den Ausdruck sowie die Anzahl der übereinstimmenden Zeilen. Ein Klick auf **Apply** fügt den Ausdruck in das Texteingabefeld ein.

Datumsspalten müssen in manuell erstellten Ausdrücken mit `as.Date("...")` oder `as.POSIXct("...")` umschlossen werden. Der Build-Filter-Dialog erledigt dies automatisch.

Die Subset-Filterung ist nicht-destruktiv — ausgeschlossene Zeilen verbleiben im Datensatz und erhalten im Excel-Export Prognosewerte.

***Tipp:** Beim Kombinieren mit AND sollten OR-Bedingungen durch Klammern gruppiert werden. Ohne Klammern wird $A \& B \mid C$ als $(A \& B) \mid C$ ausgewertet, was unter Umständen nicht dem beabsichtigten Ergebnis entspricht.*

7.7 Empfohlene Parameterwerte

earthUI zeigt unterhalb zentraler Parameter in Abschnitt 4 einen empfohlenen Wert an. Diese Empfehlungen werden reaktiv auf Grundlage der Anzahl der für die Anpassung verwendeten Zeilen (n) und der ausgewählten Prädiktoren (p) aktualisiert. Die Formeln beruhen auf Friedmans MARS-Aufsatz, den internen Algorithmen von earth sowie empirischen Tests. Die Felder **nk**, **minspan** und **endspan** werden beim Laden der Daten mit ihren empfohlenen Werten vorbelegt; eine manuelle Änderung bleibt erhalten.

7.7.1 Parameter des Vorwärtsthroughlaufs

nk (maximale Anzahl an Termen vor dem Pruning):

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

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

Der Standardwert von `earth`, $\min(200, \max(20, 2p)) + 1$, berücksichtigt die Größe des Datensatzes nicht und kann für große Datensätze zu klein ausfallen, wodurch der Vorwärtsdurchlauf eingeschränkt wird, bevor er alle Prädiktoren hinreichend untersucht hat. Der Term $\lfloor n/10 \rfloor$ erlaubt näherungsweise einen Term je 10 Beobachtungen — eine gängige Faustregel zur Vermeidung einer Überparametrisierung. Die Obergrenze von 100 beugt abnehmenden Erträgen vor.

minspan (minimale Anzahl an Beobachtungen zwischen Knoten):

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

endspan (minimale Anzahl an Beobachtungen zu den Datenrändern):

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

n	minspan	endspan
30	5	6
50	6	6
100	7	8
200	9	12
300	11	15
500	15	16
1500	16	16

Der von `earth` automatisch berechnete `minspan`-Wert verwendet Friedmans Gleichung 43: $\lfloor (-\ln(-\ln 0.95) + \ln(p \cdot n)) / (2.5 \ln 2) \rfloor$, die mit $\ln(np)$ skaliert und für große Datensätze zu langsam wächst. Bei $n=1500$ ergibt sie lediglich ≈ 7 , was rund 245 Kandidaten-Knotenpositionen je stetigem Prädiktor entspricht — zu viele, sodass der Vorwärtsdurchlauf Rauschen anpassen kann. Die empfohlene Formel zielt für große n auf etwa 100 Kandidaten je Prädiktor ab, während sie sich für kleine n (wo diese gut erprobt ist) an der automatischen Berechnung von `earth` orientiert. `Endspan` wird geringfügig größer als `minspan` gewählt, um zusätzlichen Schutz an den Datenrändern zu bieten; dies ist hilfreich, wenn die Daten dünn besetzte Ränder aufweisen (im Immobilienbereich häufig).

penalty (GCV-Strafterm je Knoten):

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

Dies ist der eigene Standardwert von `earth` — 2 für additive Modelle, 3 für Interaktionsmodelle (der höhere Strafterm kompensiert den größeren Suchraum).

newvar.penalty (Strafterm für die Einführung eines neuen Prädiktors):

Empfohlen: **0.1**, wenn die Prädiktoren korreliert sind (z. B. Wohnfläche mit Schlafzimmerzahl, Badezimmerzahl, Grundstücksgröße). Dies lenkt den Vorwärtsthroughlauf dahin, bereits im Modell enthaltene Prädiktoren wiederzuverwenden, statt korrelierte Alternativen einzuführen. Der Mechanismus: Die RSS-Verbesserung jedes neuen Prädiktors wird mit $1/(1+\text{penalty})$ multipliziert. Bei `newvar.penalty = 0.1` muss eine neue Variable die RSS um mindestens 10% stärker verbessern als ein weiterer Knoten auf einer bereits vorhandenen Variablen. Dies führt zu einfacheren Modellen mit weniger Prädiktoren, ohne die endgültigen Koeffizientenschätzungen zu beeinflussen (der Strafterm wird nach der Auswahl entfernt). Sind die Prädiktoren nicht korreliert, sollte der Wert bei 0 belassen werden. Zu beachten ist, dass `earth` einen von null verschiedenen Wert für `newvar.penalty` nicht in Kombination mit Fallgewichten unterstützt — `earthUI` deaktiviert das Eingabefeld (und erzwingt 0), sobald eine Gewichtsspalte festgelegt ist.

7.7.2 Pruning-Parameter

pmethod: Empfohlen: **backward**. Das standardmäßige GCV-basierte Rückwärts-Pruning ist deterministisch — dieselben Daten erzeugen stets dasselbe Modell, ohne Abhängigkeit von Zufallsstartwerten. Dies ist für die Reproduzierbarkeit wichtig, insbesondere in der Wertermittlung. Die Methode `cv` verwendet für das Pruning eine Kreuzvalidierung, wodurch eine Abhängigkeit vom Startwert (Seed) entsteht.

nprune: Empfohlen: **leer lassen** (NULL). Die optimale Anzahl an Termen sollte durch GCV bestimmt werden. Das Setzen von `nprune` erzwingt eine feste Obergrenze, die das GCV-Kriterium außer Kraft setzt.

7.7.3 Kreuzvalidierungs-Parameter

nfold (CV-Folds): Empfehlung nach Bereich der Stichprobengröße:

n (Anpassungszeilen)	nfold
0–299	5
300–399	6
400–499	7
500–599	8
600–699	9
700+	10

Mehr Folds trainieren jeden Fold auf mehr Daten (eine weniger verzerrte Schätzung), verkleinern jedoch den zurückgehaltenen Test-Fold. Die Tabelle steigert die Anzahl mit wachsender Stichprobe schrittweise bis zur üblichen 10-fachen Kreuzvalidierung und hält dabei jeden Test-Fold groß genug (rund 50+ Zeilen) für eine stabile Kennzahl je Fold; für kleine Stichproben fällt sie auf die 5-fache Kreuzvalidierung zurück — dort ist die CVR^2 -Schätzung weiterhin veräuscht, weshalb `ncross` erhöht werden sollte (z. B. 15–20), um die Varianz herauszumitteln. Die Obergrenze von 10 trägt dem Umstand Rechnung, dass jeder zusätzliche Fold eine vollständige `earth`-Anpassung bei vernachlässigbarem diagnostischem Zugewinn bedeutet. Mit `pmethod = "backward"` beeinflusst die Kreuzvalidierung das Modell nicht — sie berechnet lediglich das diagnostische CVR^2 und liefert Residuen für das Varianzmodell.

ncross (CV-Wiederholungen): Empfehlung nach Bereich der Stichprobengröße:

n (Anpassungszeilen)	ncross
0–199	20
200–399	16
400–699	12

n (Anpassungszeilen)	ncross
700–1199	10
1200–2999	7
3000+	5

Die Wiederholungen mitteln den Zufall der Fold-Aufteilung heraus und stabilisieren so CVR²; größere Stichproben benötigen weniger Wiederholungen, da eine einzelne Kreuzvalidierung bereits stabiler ist. Die Residuen der wiederholten CV fließen zudem in das Varianzmodell ein.

7.7.4 Varianzmodell

varmod.method: Empfohlen: **lm** — Milborrows Standard: ein robustes lineares Varianzmodell für Prognoseintervalle (die Streuung wird aus verrauschten Residuen geschätzt, weshalb ein einfaches Modell besser generalisiert). Konvergiert **lm** auf den vorliegenden Daten nicht, empfiehlt sich **rlm** oder **const** (homoskedastisch, konvergiert stets), ersatzweise **earth** (erfasst auch eine nichtlineare Streuung). Erforderlich sind **nfold** > 1 und **ncross** > 1.

7.7.5 Earth-Ausgabedatei

Nach jeder erfolgreichen Anpassung schreibt earthUI die Datei `<filename>_earth_output_<timestamp>.txt` in den Ausgabeordner. Diese Datei enthält die Modellterme, zusammenfassende Statistiken (R^2 , GRSq, CVRSq), Einzelheiten zum Varianzmodell sowie das Trace-Protokoll. Je Anpassung wird eine Datei erzeugt, sodass eine fortlaufende Aufzeichnung zum Vergleich von Parameterkonfigurationen entsteht.

7.8 Voreinstellungen (Settings Defaults)

Ein Optionsfeld am oberen Rand von Abschnitt 4 steuert, welche Voreinstellungen geladen werden:

- **Use last settings for input file** — stellt die Einstellungen wieder her, die zuletzt mit genau dieser Datei verwendet wurden
- **Use default settings** — übernimmt die gespeicherten benutzerdefinierten Voreinstellungen
- **Earth defaults** — setzt alle Parameter auf die Werksvorgaben zurück

Über **Save current as default** am unteren Rand von Abschnitt 4 werden die aktuellen earth-Parameter sowie die Matrix Allowed Interactions gespeichert; die entsprechende Schaltfläche in Abschnitt 3 speichert die Konfiguration von Zielvariable, Stichtag und Prädiktoren. Beides wird je Projekt und Verwendungszweck in der Projektdatenbank abgelegt.

8 Anpassung des Earth-Modells

8.1 Die Schaltfläche „Fit“

Abschnitt 5 der Seitenleiste enthält die Schaltfläche **Fit Earth Model** (Earth-Modell anpassen) zusammen mit einem Feld **Random seed** (Zufalls-Seed). Der Seed macht die Zuordnung der Kreuzvalidierungs-Folds reproduzierbar — ein Dropdown-Menü ruft die letzten 10 verwendeten Seeds ab, und der verwendete Seed wird nach der Anpassung im Reiter **Earth Output** protokolliert. Vor der Anpassung validiert earthUI die Konfiguration — eine Zielvariable muss ausgewählt und mindestens ein Prädiktor muss einbezogen sein. In den Modi Appraisal/Market werden die Spalten für geografische Breite und Länge gerundet, und das Verkaufsalter wird aus dem Stichtag und dem Vertragsdatum berechnet (sofern zugewiesen).

8.2 Anpassungs-Modal & Trace-Ausgabe

Beim Klick auf **Fit** erscheint eine dunkle modale Überlagerung mit:

- **Kopfzeile** — „Fitting Earth Model“ mit einem laufenden Zeitzähler in Sekunden sowie einer Schaltfläche **Abort** (Abbrechen), die die im Hintergrund laufende Anpassung abbricht
- **Trace-Log** — ein scrollbarer Textbereich in nichtproportionaler Schrift, der die Echtzeitausgabe der Funktion `earth()` anzeigt: Informationen zum Datensatz, Fortschritt des Forward Pass, Kreuzvalidierungs-Folds und Abschlusszeit
- **Farbcodierung** — Standardzeilen in Grün, Kreuzvalidierungszeilen in Gelb, Fehler in Rot

earthUI passt Modelle asynchron mittels `callr::r_bg()` an, wodurch die Berechnung in einem separaten R-Prozess ausgeführt wird. Dadurch bleibt die Anwendung während lang laufender Anpassungen reaktionsfähig. Ein Polling-Observer mit 300 ms Intervall liest die Ausgabe des Hintergrundprozesses und leitet sie fortlaufend an das Trace-Log weiter.

Nach Abschluss der Anpassung erscheint eine Meldung „Done in X.Xs“, dem Modal wird eine Schließen-Schaltfläche (X) hinzugefügt, und der Hintergrund wird entfernt, sodass die Ergebnis-Reiter bei Bedarf gerendert werden können. Das Modal schließt sich anschließend automatisch, sobald die sichtbaren Ausgaben fertig gerendert sind — alternativ kann es auch früher geschlossen werden.

Bei Erfolg wird der Schaltfläche **Fit** ein grünes Häkchen hinzugefügt. Tritt ein Fehler auf, wird die Fehlermeldung im Trace-Log angezeigt und eine Benachrichtigung erscheint.

8.3 Anpassungsprotokoll

Jede Anpassung (Erfolg oder Fehlschlag) schreibt eine Protokolldatei in den Ausgabeordner:

- **Dateiname:** `<datafile>_earth_log_<YYYYMMDD_HHMMSS>.txt`
- **Inhalt:** Zeitstempel, die gesamte Trace-Ausgabe des Anpassungsprozesses sowie etwaige Fehlermeldungen
- **Speicherort:** der Ausgabeordner des aktiven Projekts

Tipp: Ist `callr` nicht installiert, weicht earthUI auf eine synchrone Anpassung mit einem einfachen Fortschrittsbalken aus. Für die vollständige Trace-Ausgabe sollte `callr` installiert werden.

9 Herunterladen der Daten

Nach der Anpassung lädt Abschnitt 6 der Seitenleiste — im allgemeinen Modus mit „Download Estimated Target Variable(s) & Residuals“ und in den Modi Appraisal und Market mit „Download Estimated Sale Prices & Residuals“ betitelt — eine Excel-Datei mit Prognosen und Diagnostiken herunter. Diese Ausgabe wird in Schritt 7 (RCA) verwendet, um dem Bewertungsobjekt eine CQA-Bewertung (Condition-Quality-Appeal, d. h. Zustand-Qualität-Attraktivität) zuzuweisen. Die Ausgabe ist nach `residual_sf` und `cqa_sf` sortiert, um die Einordnung des Bewertungsobjekts im Ranking zu erleichtern. Bei guter Modellqualität sollten die Objekte auf Grundlage der residualen Merkmale, die nicht in die Regression eingegangen sind, von der geringsten bis zur höchsten Attraktivität geordnet sein. Der mittlere Wert sollte annähernd 0 betragen, die untere Hälfte negative und die obere Hälfte positive Werte aufweisen. Die Objekte schlechtester Qualität, sogenannte „Fixers“ (renovierungsbedürftige Objekte), sollten sich am unteren Ende des Rankings finden und die attraktivsten Objekte am oberen Ende. Ausnahmen ergeben sich üblicherweise bei Anomalien wie Zwangsversteigerungen, Short Sales (Verkäufen unter Restschuld), Nachlassverkäufen (erbschaftsbedingt) sowie Schnellverkäufen aufgrund eines Arbeitsplatzwechsels oder aus anderen Gründen. Die Untersuchung von Anomalien fördert in der Regel einen einschlägigen Grund für die Preisanomalie zutage.

Die Beschriftung der Schaltfläche passt sich dem Zweckmodus an:

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

Das Dateinamensformat lautet `<datafile>_modified_<YYYYMMDD_HHMMSS>.xlsx`.

9.1 Ausgabespalten

Für jede Zielvariable werden die folgenden Spalten angefügt:

Spalte	Beschreibung
<code>est_<target></code>	Modellprognose, z. B. <code>est_sale_price</code> (1 Dezimalstelle)
<code>residual</code>	Ist-Wert minus Prognosewert (1 Dezimalstelle)
<code>cqa</code>	Condition-Quality-Appeal-Score (2 Dezimalstellen, Wertebereich 0–10)
<code>residual_sf</code>	Residuum dividiert durch die Wohnfläche (sofern zugewiesen, 1 Dezimalstelle)
<code>cqa_sf</code>	CQA, berechnet aus dem Ranking über <code>residual_sf</code> (2 Dezimalstellen)
<code><variable>_contribution</code>	Wertbeitrag je g-Funktion (1 Dezimalstelle, eine Spalte je g-Funktion)
<code>basis</code>	Wertbeitrag des Achsenabschnitts, für alle Objekte identisch (1 Dezimalstelle)
<code>calc_residual</code>	Ist-Wert minus (basis + alle Wertbeiträge) — Verifikationsspalte (1 Dezimalstelle)

Bei Modellen mit mehreren Zielvariablen wird ein Suffix `_<i>` angefügt, um die Spalten der einzelnen Zielvariablen zu unterscheiden.

9.2 Spaltenanordnung & Excel-Formatierung

Die Ranking-Spalten werden in der Ausgabedatei an der **äußersten linken Position** platziert, und zwar in dieser Reihenfolge: `residual_sf`, `cqa_sf`, `residual`, `cqa`. Dies erleichtert das

Durchsehen der sortierten Liste und die Beurteilung, wo das Bewertungsobjekt in der CQA-Verteilung liegt.

Auf diese Spalten wird eine Excel-Zahlenformatierung angewendet:

Spalte	Format	Beispiel
<code>residual_sf</code>	Numerisch, 2 Dezimalstellen	12.34
<code>cqa_sf</code>	Numerisch, 2 Dezimalstellen	7.25
<code>residual</code>	Numerisch, 0 Dezimalstellen	15,200
<code>cqa</code>	Numerisch, 2 Dezimalstellen	6.80

9.3 CQA-Scores

Der CQA-Score (Condition-Quality-Appeal) stuft das Residuum jeder Zeile gegenüber allen anderen Residuen ein. Für eine gegebene Zeile ist der CQA der Prozentsatz der Zeilen mit einem kleineren vorzeichenbehafteten Residuum, multipliziert mit 10. Daraus ergibt sich eine Skala von 0 bis 10, wobei gilt:

- **Hoher CQA** ($\approx 9-10$) — das Objekt wurde für deutlich mehr verkauft, als das Modell prognostiziert hat (großes positives Residuum)
- **Niedriger CQA** ($\approx 0-1$) — das Objekt wurde für deutlich weniger als prognostiziert verkauft (großes negatives Residuum)
- **CQA ≈ 5** — das Residuum liegt nahe am Median aller Residuen

Ist eine Spalte `living_area` zugewiesen, liefert `cqa_sf` dasselbe Ranking auf Grundlage der Residuen je Quadratfuß.

9.4 Sortierung & Zeile des Bewertungsobjekts

In den Modi Appraisal und Market werden die Zeilen der Vergleichsobjekte absteigend nach `residual_sf` sortiert (bzw. nach `residual`, wenn keine Wohnfläche zugewiesen ist). Die Zeile des Bewertungsobjekts (Zeile 1) verbleibt an Position 1 und wird nicht mitsortiert. Im allgemeinen Modus werden die Zeilen in ihrer ursprünglichen Reihenfolge exportiert.

Diese Sortierung ermöglicht es dem Gutachter in Verbindung mit den ganz links platzierten Ranking-Spalten, die Vergleichsobjekte rasch von der stärksten Überprognose bis zur stärksten Unterprognose durchzusehen und zu beurteilen, wo der dem Bewertungsobjekt zugewiesene CQA-Score in dieser Verteilung liegt.

Die Faktorstufen in den Prognosedaten werden vor dem Aufruf von `predict()` an die Trainingsdaten angeglichen. Zeilen mit unbekannten Faktorstufen erzeugen NA-Prognosen.

Tipp: Die Spalte `calc_residual` sollte stets (im Rahmen der Rundung) gleich `residual` sein. Weichen die Werte erheblich voneinander ab, deutet dies auf ein untersuchenswertes Berechnungsproblem hin.

9.5 Automatischer mgcvUI-Export

Bei jeder erfolgreichen Anpassung mit `degree` ≤ 2 speichert earthUI das vollständige Ergebnisobjekt automatisch als `.rds`-Datei im Ausgabeordner des Projekts. Der Dateiname folgt dem Muster `<datafile>_earthUI_result_<YYYYMMDD_HHMMSS>.rds`.

Diese Datei kann vom mgcv Post Processing Module (mgcv-Nachbearbeitungsmodul) — das im earthUI-Paket enthalten ist und mit `earthUI::launch_mgcv()` gestartet wird — mittels `readRDS()` geladen werden. Die GAM-Routine verwendet die Knotenpositionen und Basisfunktionen des Earth-Modells als Ausgangspunkte für die GAM-Glättungsterme und ermöglicht so einen nahtlosen Übergang von der MARS- zur GAM-Modellierung.

Modelle mit `degree > 2` werden übersprungen, da `mgcvUI` ausschließlich paarweise Interaktionen unterstützt. Wird `earthUI` im Trilogy-Modus (aus der Trilogy UI heraus) gestartet, erlaubt ein zusätzlicher Abschnitt **Lock Model Output** (Modellausgabe fixieren), die `.rds`-Datei einer bestimmten Anpassung — zusammen mit ihrer Prädiktorbasis und den gemeinsamen CQA-Einstellungen — als kanonisches Earth-Modell zu fixieren, das `glmnetUI` und `mgcvUI` importieren.

9.6 glmnetUI-Import

Dieselbe für die GAM-Routine gespeicherte `.rds`-Datei kann auch in das **elastic net Post Processing Module** (Elastic-Net-Nachbearbeitungsmodul) importiert werden — ebenfalls im `earthUI`-Paket enthalten und mit `earthUI::launch_glmnet()` zu starten. In der `glmnet`-Anwendung ist zu **Section 2: Import from earthUI** zu navigieren und die `.rds`-Datei über die Schaltfläche **Browse** auszuwählen.

`glmnetUI` verwendet den Standardansatz über `earth::model.matrix()`: Die Basisfunktionen des Earth-Modells (Hinges, Interaktionen und lineare Terme) werden zu den Spalten der Designmatrix von `glmnet`. `glmnet` führt anschließend eine regularisierte Regression auf diesen Basisspalten durch und selektiert und schrumpft sie mittels Lasso/Elastic Net. Dies verbindet die adaptive Basiskonstruktion von `earth` mit der Regularisierung von `glmnet`.

9.6.1 Wesentliche Vorteile

- **Interpretierbare Interaktionen:** Die Hinge-basierten Interaktionen von `earth` (einschließlich Dreifachtermen) besitzen eine klare geometrische Bedeutung, im Gegensatz zu den nativen Kreuzprodukt-Interaktionen von `glmnet`, die vor Gericht oder in Prüfungskontexten schwer zu erläutern sind.
- **Automatische Nichtlinearität:** Die Hinge-Funktionen von `earth` erfassen nichtlineare Zusammenhänge, die ein Standard-`glmnet`-Modell mit unbearbeiteten Prädiktoren übersehen würde.
- **Variablenselektion:** `glmnet` kann die Earth-Basis weiter ausdünnen und Hinge-Terme entfernen, die keinen Beitrag leisten.

9.6.2 Wichtig: Die Gewichtungsspalte muss übereinstimmen

Warnung

Verwendet `earthUI` eine Gewichtungsspalte (z. B. um die Zeile des Bewertungsobjekts mit `weight=0` auszuschließen), muss dieselbe Gewichtungsspalte in der Variablenkonfiguration von `glmnetUI` zugewiesen werden. Weichen die Gewichtungseinstellungen voneinander ab, stimmen die Zeilenzahlen nicht überein, und die Earth-Basis kann nicht an die Modellmatrix von `glmnet` angefügt werden.

9.6.3 Arbeitsablauf

1. Anpassung eines Earth-Modells in `earthUI` (beliebiger Grad).
2. Die `.rds`-Ergebnisdatei wird automatisch im Ausgabeordner des Projekts gespeichert.
3. In `glmnetUI` **Section 2: Import from earthUI** öffnen und zur `.rds`-Datei navigieren.
4. Überprüfen, dass in beiden Anwendungen **dieselbe Datendatei** und **dieselbe Gewichtungsspalte** verwendet werden.
5. Konfiguration der `glmnet`-Parameter in Abschnitt 5 (die Interaction Matrix wird bei aktivem `earthUI`-Import automatisch deaktiviert).

6. Auf **Fit Model** klicken. Die Earth-Basispalten erscheinen in den Reitern Equation, Coefficients und Contributions.

10 RCA-Berechnungen & Download

Die Schaltfläche **Calculate RCA Adjustments & Download** (RCA-Anpassungen berechnen und herunterladen) erscheint in Abschnitt 7 der Seitenleiste und ist nur im Appraisal Mode sichtbar, nachdem ein Modell angepasst wurde. RCA (Residual Constraint Approach) erzeugt marktabgeleitete Anpassungen je Vergleichsobjekt relativ zum Bewertungsobjekt.

10.1 Öffnen des RCA-Dialogs

Ein Klick auf die Schaltfläche öffnet einen kleinen modalen Dialog mit:

- **Score type** (Score-Typ) — Optionsfelder zur Auswahl von **CQA** oder **CQA per SF** (CQA je Quadratfuß; die SF-Option ist nur verfügbar, wenn eine Spalte `living_area` festgelegt wurde). Wird „CQA per SF“ gewählt, so ergibt sich der Residual-Score des Bewertungsobjekts auf Grundlage des ihm zugewiesenen CQA-Scores als dessen Wohnfläche multipliziert mit dem `residual_sf`-Wert, der dem gegebenen CQA_SF-Score entspricht.
- **CQA value** (CQA-Wert) — Eingabefeld für den CQA-Score des Bewertungsobjekts (0.00–10.00). Dieser Score gibt an, wo das Bewertungsobjekt nach Einschätzung des Anwenders in der Qualitätsverteilung der Vergleichsobjekte liegt.
- Schaltfläche **Generate** (Erzeugen) — berechnet die RCA-Ausgabe und lädt sie als Excel-Datei herunter. Die Schaltfläche bleibt deaktiviert, bis ein gültiger Score eingegeben wurde.

10.2 Interpolation des CQA-Scores

Beim Klick auf **Generate** interpoliert earthUI das Residuum des Bewertungsobjekts aus den CQA-/Residuen-Paaren der Vergleichsobjekte:

1. Die CQA-Scores und Residuen der Vergleichsobjekte werden nach CQA aufsteigend sortiert
2. Eine lineare Interpolation (`stats::approx()`) bildet den eingegebenen CQA-Wert auf ein Residuum ab
3. Bei Verwendung von CQA per SF wird das Residuum je Quadratfuß durch Multiplikation mit der Wohnfläche des Bewertungsobjekts in ein Gesamtresiduum zurückgerechnet
4. Der geschätzte Wert des Bewertungsobjekts wird berechnet als: Modellprognose + interpoliertes Residuum

10.3 Ausgabespalten

Die RCA-Excel-Datei (`<datafile>_adjusted_<YYYYMMDD_HHMMSS>.xlsx`) enthält alle Zwischenausgabespalten sowie zusätzlich:

Spalte	Beschreibung
<code>subject_value</code>	Modellprognose + interpoliertes Residuum (nur Zeile 1)
<code>subject_cqa</code>	Der eingegebene CQA-Score (nur Zeile 1)
<code><variable>_adjustment</code>	Wertbeitrag des Bewertungsobjekts abzüglich des Wertbeitrags des Vergleichsobjekts (je g-Funktion)
<code>residual_adjustment</code>	Residuum des Bewertungsobjekts abzüglich des Residuums des Vergleichsobjekts
<code>net_adjustments</code>	Summe aller Anpassungen (Wertbeitrag + Residuum)
<code>gross_adjustments</code>	Summe der Absolutbeträge aller Anpassungen
<code>residual_pct</code>	Residualanpassung als Anteil am Verkaufspreis des Vergleichsobjekts

Spalte	Beschreibung
<code>net_adj_pct</code>	Nettoanpassungen als Anteil am Verkaufspreis des Vergleichsobjekts
<code>gross_adj_pct</code>	Bruttoanpassungen als Anteil am Verkaufspreis des Vergleichsobjekts
<code>adjusted_sale_price</code>	Verkaufspreis des Vergleichsobjekts + Nettoanpassungen

Die Anpassungsspalten geben für jedes Vergleichsobjekt an, in welchem Umfang das Modell die Differenz den einzelnen Variablen zuschreibt. `adjusted_sale_price` ist der Verkaufspreis des Vergleichsobjekts nach Anwendung aller modellabgeleiteten Anpassungen — eine Menge angepasster Verkaufspreise, die sich eng um den geschätzten Wert des Bewertungsobjekts gruppieren sollten.

10.4 Die Registerkarte RCA Adjustments

Sobald die RCA-Ausgabe berechnet wurde, zeigt die Registerkarte **RCA Adjustments** (RCA-Anpassungen) im Hauptpanel Histogramme der Verteilungen von Residual Adj %, Net Adj % und Gross Adj % über die Vergleichsobjekte an (bei Modellen mit mehreren Zielgrößen ein Satz je Zielgröße) und ermöglicht so eine schnelle visuelle Kontrolle, wie eng die Anpassungen gruppiert sind.

Tipp: Ein CQA-Score von 5.00 platziert das Bewertungsobjekt am Median der Vergleichsobjekte. Scores über 5 zeigen an, dass das Bewertungsobjekt eine überdurchschnittliche Qualität aufweist; Werte unter 5 deuten auf eine unterdurchschnittliche Qualität hin.

11 Sales Comparison Grid

11.1 Überblick & Zweck

Das **Sales Comparison Grid** (Vergleichsraster) ist eine Excel-Arbeitsmappe, die im Appraisal Mode (Sidebar-Abschnitt 8) nach der Berechnung der RCA-Anpassungen erzeugt wird. Es stellt das Bewertungsobjekt zusammen mit ausgewählten Vergleichsobjekten in einem strukturierten Rasterformat dar, wobei Excel-Formeln die Anpassungen sowie einen angepassten Verkaufspreis für jedes Vergleichsobjekt automatisch berechnen.

Das Raster ist für die Arbeitsakte des Gutachters konzipiert — es kombiniert die aus der Regression abgeleiteten Anpassungen des Earth-Modells mit editierbaren Zellen, in denen der Gutachter das CQA-Residuum bestimmten Objektmerkmalen zuordnen kann. Der Name der Ausgabedatei lautet **SalesGrid_<YYYYMMDD_HHMMSS>.xlsx**.

11.2 Modaldialog zur Auswahl der Vergleichsobjekte

Ein Klick auf „Generate Sales Grid & Download“ öffnet einen modalen Dialog, in dem ausgewählt wird, welche Vergleichsobjekte einbezogen werden sollen:

- **Recommended comps** (empfohlene Vergleichsobjekte) sind vorab angehakt. Dabei handelt es sich um Vergleichsobjekte mit einem Brutto-Anpassungsprozentsatz von unter 25% des Verkaufspreises, aufsteigend nach Verkaufsalter sortiert (jüngste Verkäufe zuerst) und auf 30 begrenzt.
- **Additional comps** (weitere Vergleichsobjekte) werden darunter aufgeführt und sind standardmäßig nicht ausgewählt. Diese weisen Brutto-Anpassungen von 25% oder mehr auf, stehen jedoch weiterhin zur Auswahl (angezeigt werden bis zu 50, aufsteigend nach Brutto-Anpassungsprozentsatz sortiert).
- Es können maximal **30 Vergleichsobjekte** ausgewählt werden, woraus sich bis zu **10 Arbeitsblätter** ergeben (3 Vergleichsobjekte pro Blatt).

Nach Bestätigung der Auswahl werden die ausgewählten Vergleichsobjekte aufsteigend nach Brutto-Anpassungsprozentsatz sortiert, sodass die ähnlichsten Vergleichsobjekte auf Blatt 1 erscheinen.

11.3 Rasterlayout

Jedes Arbeitsblatt besitzt ein Layout mit 20 Spalten, das das Bewertungsobjekt sowie 3 Vergleichsobjekte aufnimmt. Die Spalten je Entität sind:

Entität	Spalten (5 je Entität)
Bewertungsobjekt	Label, Factual Value 1, Factual Value 2, Factual Value 3, Value Contribution (VC)
Jedes Vergleichsobjekt	Factual Value 1, Factual Value 2, Factual Value 3, Value Contribution (VC), Adjustment

Die Zeilen von oben nach unten sind:

1. **Titel** — Blattname (z. B. „Intermediate Sales Comparable Grid — Sheet 1 of 3“)
2. **Kopfzeilen** — „Subject“ sowie die Spaltenüberschriften der Vergleichsobjekte mit Adresse
3. **Adresse** — vollständige Objektadresse
4. **APN | MLS# | DOM | Subj.Prox** — Flurstücksnummer, Listing-ID, Vermarktungsdauer (days on market) und Haversine-Distanz (in Meilen) zum Bewertungsobjekt

5. **Sales Price | Concess. | Net SP** — Verkaufspreis, Preiszugeständnisse und Formel für den Nettoverkaufspreis
6. Kopfzeile **Regression Features**
7. **BASE VALUE** — der Achsenabschnitt (Intercept) des Modells
8. **Date of Sale | OffMkt | OnMkt** — Vertragsdatum, Verkaufsalter und DOM
9. **Gruppierte Zeilen** (bedingt) — Gruppen für Location, Site Size und/oder Age
10. **Modellvariablenzeilen** — eine Zeile je Prädiktor (ohne gruppierte Variablen)
11. **Leere Trennzeile**
12. Kopfzeile **Residual Features**
13. **CQA|Residual** — CQA-Score + Formel für das verbleibende Residuum
14. **Zeilen für Residualmerkmale** — benannte + leere Zeilen für Eintragungen des Gutachters
15. **Total VC / Net Adjustment** (Gesamt-Wertbeitrag / Netto-Anpassung)
16. **Net Adjustment %** (Netto-Anpassung in Prozent)
17. **Gross Adjustment %** (Brutto-Anpassung in Prozent)
18. **Adjusted Sale Price** (angepasster Verkaufspreis) — Formelzeile
19. **Copyright-Fußzeile**

11.4 Sale Price, Concessions & Net SP

Zeile 5 zeigt für jedes Vergleichsobjekt drei Werte:

- **Sale Price** — der Verkaufspreis des Vergleichsobjekts aus den Daten
- **Concessions** — der Wert aus der als **concessions** ausgewiesenen Spalte (0, falls keine Spalte ausgewiesen ist)
- **Net SP** — eine Excel-Formel: Verkaufspreis – Preiszugeständnisse

Die Spalte des Bewertungsobjekts zeigt „N/A“ für den Verkaufspreis (da dieser unbekannt ist) sowie gegebenenfalls einen vorhandenen Wert für Preiszugeständnisse.

11.5 Gruppierte Zeilen (Location, Site, Age)

Wenn bestimmte Spezialtypen ausgewiesen sind und die entsprechenden Variablen im Modell erscheinen, erzeugt earthUI gruppierte Zeilen, die verwandte Variablen zusammenfassen:

Loc: Long | Lat | Area — Erscheint, wenn **longitude**, **latitude** oder **area** im Modell enthalten sind. Zeigt die faktischen Werte jeder beteiligten Variablen, einen kombinierten Wertbeitrag (Summe der einzelnen Wertbeiträge) sowie für Vergleichsobjekte eine Anpassung (kombinierter VC des Bewertungsobjekts – kombinierter VC des Vergleichsobjekts). Die VC-Zellen sind mit hellblauem Hintergrund formatiert.

Site Size | Dimensions — Erscheint, wenn **lot_size** oder **site_dimensions** im Modell enthalten sind. Gleiche Struktur wie die Location-Gruppe.

Actual Age | Effective Age — Erscheint, wenn **actual_age** oder **effective_age** im Modell enthalten sind. Gleiche Struktur.

Von gruppierten Zeilen erfasste Variablen werden **aus den darunterliegenden einzelnen Modellvariablenzeilen ausgeschlossen**, um Doppelzählungen in den Anpassungssummen zu vermeiden.

11.6 Modellvariablenzeilen

Unterhalb der gruppierten Zeilen (bzw. direkt unterhalb von BASE VALUE, falls keine gruppierten Zeilen vorhanden sind) zeigt je eine Zeile pro verbleibendem Modellprädiktor:

- **Faktische Werte** für das Bewertungsobjekt und jedes Vergleichsobjekt (die tatsächlichen Datenwerte)

- **Wertbeitrag (Value Contribution, VC)** — der Beitrag der g-Funktion zum prognostizierten Wert in dieser Zeile
- **Anpassung** (nur Vergleichsobjekte) — VC des Bewertungsobjekts minus VC des Vergleichsobjekts

11.7 CQA|Residual & Zeilen für Residualmerkmale

Die Zeile **CQA|Residual** zeigt den CQA-Score jedes Objekts und enthält eine Formel für das **verbleibende Residuum** — den Teil des Residuums, der noch keinen konkreten Merkmalen zugeordnet wurde. Die Formel lautet:

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

Darunter befinden sich **Zeilen für Residualmerkmale** — 5 benannte Zeilen (Location, View/Appeal, Condition, Quality, Other) sowie 6 leere Zeilen. Dies sind Eingabezellen, in die der Gutachter Wertbeiträge für Merkmale einträgt, die das Modell nicht erfasst. Mit jedem eingetragenen Wert verringert sich das verbleibende Residuum gemäß der Formel automatisch.

Für jedes Vergleichsobjekt enthält die Anpassungsspalte eine Formel: Merkmal-VC des Bewertungsobjekts – Merkmal-VC des Vergleichsobjekts.

11.8 Formel für den Adjusted Sale Price

Die Zeile **Adjusted Sale Price** enthält Excel-Formeln:

- **Bewertungsobjekt:** Summe aller Wertbeitragszellen von BASE VALUE bis zur letzten Residualmerkmalzeile. Dies entspricht der Gesamtprognose des Modells für das Bewertungsobjekt zuzüglich der vom Gutachter zugeordneten Residualmerkmale.
- **Vergleichsobjekte:** Net SP + Summe aller Anpassungszellen von der ersten gruppierten Zeile bzw. Variablenzeile bis zur letzten Residualmerkmalzeile. Dies ergibt den Verkaufspreis des Vergleichsobjekts nach allen aus der Regression abgeleiteten und vom Gutachter eingetragenen Anpassungen.

11.9 Blattschutz & editierbare Zellen

Jedes Arbeitsblatt ist **geschützt**, um versehentliche Änderungen an Formeln und Daten zu verhindern. Der Schutz sperrt:

- Alle Formelzellen (Net SP, verbleibendes Residuum, Anpassungen, angepasster Verkaufspreis usw.)
- Alle Datenzellen (Adressen, Verkaufspreise, faktische Werte, Wertbeiträge aus dem Modell)
- Beschriftungen und Kopfzeilen

Die einzigen **entsperrten** (editierbaren) Zellen sind die Eingabezellen für die Wertbeiträge der Residualmerkmale — die Zellen unterhalb der Zeile CQA|Residual, in die der Gutachter die Aufschlüsselung einträgt. Diese sind mit hellgelbem Hintergrund formatiert, um ihre Editierbarkeit anzuzeigen.

11.10 Arbeiten mit dem Raster in Excel

1. **Öffnen der Datei** in Excel oder einer kompatiblen Tabellenkalkulationsanwendung.
2. **Prüfung der Regressionsanpassungen** — die aus dem Modell abgeleiteten Anpassungen sind vorausgefüllt und gesperrt.
3. **Zuordnung des Residuums** — in den Zeilen für Residualmerkmale (gelbe Zellen) wird die Einschätzung eingetragen, welcher Anteil des verbleibenden Residuums dem jeweiligen Merkmal (Location, View, Condition, Quality usw.) zuzurechnen ist. Das verbleibende Residuum verringert sich gemäß der Formel mit jeder Zuordnung.

4. **Kontrolle des Adjusted Sale Price** — diese Formel aktualisiert sich automatisch, sobald Werte für Residualmerkmale eingetragen werden.
5. **Mehrere Arbeitsblätter** — wurden mehr als 3 Vergleichsobjekte ausgewählt, kann zwischen den Blättern navigiert werden. Jedes Blatt enthält dasselbe Bewertungsobjekt in der linken Spalte mit 3 unterschiedlichen Vergleichsobjekten.

Tipp: Ziel ist es, das CQA-Residuum so auf die Merkmalzeilen zu verteilen, dass das verbleibende Residuum nahe null liegt. Damit wird gezeigt, dass die Differenz zwischen dem Regressionsschätzwert und dem tatsächlichen Verkaufspreis jedes Vergleichsobjekts erklärt werden kann.

12 Herunterladen von Berichten

Nach der Anpassung eines Modells erfolgt die Berichtserstellung als zweistufiger Quarto-Arbeitsablauf: **Generate Quarto Report** (Seitenleistenabschnitt 9 im Appraisal Mode, andernfalls 7) schreibt ein in sich geschlossenes Berichtspaket, und **Convert Quarto Report** (Abschnitt 10) rendert dieses — oder jede andere Quarto-Datei — in die endgültigen Formate.

12.1 Generate Quarto Report

Ein Klick auf **Generate Quarto Report** schreibt ein in sich geschlossenes Quarto-Paket in den Ausgabeordner des Projekts unter `<base>_qmd/`, das die befüllte `.qmd`-Quelldatei, alle vorab erzeugten Diagramm-Assets (PNG und PDF), die Berichtsdaten sowie eine Word-Referenzvorlage enthält. In diesem Schritt findet kein Rendern statt — das Paket kann von Hand bearbeitet, über Quarto-Includes mit anderen Quarto-Quellen kombiniert oder eigenständig mit `quarto preview` in der Vorschau betrachtet werden.

12.2 Convert Quarto Report

Abschnitt 10 stellt eine Dateiauswahl bereit (voreingestellt auf das zuletzt erzeugte Paket des aktuellen Projekts — es kann jedoch zu jeder beliebigen `.qmd`-Datei navigiert werden) sowie Kontrollkästchen für die Formate:

- **HTML** — eine `.html`-Datei mit KaTeX-Mathematikdarstellung, eingebetteten Bildern und einem Inhaltsverzeichnis. Verwendet das Flatly-Bootstrap-Thema und die Schriftart Roboto Condensed.
- **Word** — eine `.docx`-Datei, die sich zur Bearbeitung und Verteilung eignet. Enthält ein Inhaltsverzeichnis sowie Seitenzahlen und verwendet eine angepasste Referenzvorlage für eine einheitliche Formatierung.
- **PDF** — mit LuaLaTeX gesetzt für eine Ausgabe in professioneller Qualität. Das Papierformat richtet sich nach der Gebietsschema-Einstellung (Letter oder A4). Enthält Querformatseiten für große Interaktionsmatrizen und verwendet Roboto Condensed mit Latin Modern Math für Gleichungen. Die PDF-Option wird nur angeboten, wenn eine LaTeX-Distribution erkannt wird.

Ein Klick auf **Convert** rendert die ausgewählten Formate. Das Rendern läuft im Hintergrund — ein modaler Dialog zeigt die verstrichene Zeit und den Quarto-Fortschritt an, während die Anwendung weiterhin reaktionsfähig bleibt. Nach Abschluss bestätigt eine Benachrichtigung den Speicherort der Datei.

Alle Operationen (Modellanpassung, Daten-Downloads, RCA-Berechnungen, Sales-Grid-Erzeugung und Berichts-Rendering) werden mit Start-/End-Zeitstempeln und verstrichenen Zeiten in `<datafile>_earthui_log.txt` im Ausgabeordner protokolliert, zur Fehlersuche und Leistungsüberwachung.

12.3 Berichtsinhalte

Jeder erzeugte Bericht enthält die folgenden Abschnitte:

1. **Datensatzbeschreibung** — Anzahl der Beobachtungen, Zielvariable(n), verwendete Prädiktoren, kategoriale Prädiktoren
2. **Modellspezifikation** — Grad, Kreuzvalidierungsstatus, Anzahl der Terme, im Modell enthaltene Prädiktoren
3. **Zugelassene Interaktionen** (nur bei $\text{Grad} \geq 2$) — die Interaktionsmatrix mit Häkchen, formatiert für den jeweiligen Ausgabebetyp

4. **Ergebniszusammenfassung** — R^2 , GRSq, GCV, RSS und CV R^2 (sofern die Kreuzvalidierung aktiviert war)
5. **Modellgleichung** — die vollständige Gleichung des Earth-Modells in LaTeX-Notation
6. **Koeffizienten & Basisfunktionen** — Tabelle aller Terme, ihrer Koeffizienten und der sie definierenden Basisfunktionen
7. **Variablenwichtigkeit** — Balkendiagramm und Rangtabelle der Wichtigkeitswerte der Prädiktoren
8. **g-Funktions-Beiträge** — Diagramme für jede g-Funktion: Liniendiagramme für univariate Terme, perspektivische 3D- und Konturdiagramme für bivariate Interaktionen
9. **Korrelationsmatrix** — Heatmap der Prädiktorkorrelationen
10. **Diagnostik** — Diagramme „Residuals vs Fitted“, „Normal Q-Q“ und „Actual vs Predicted“
11. **ANOVA-Zerlegung** — Tabelle mit dem Beitrag jeder Basisfunktion zur RSS
12. **Earth-Ausgabe** — unformatierte Textausgabe von `earth::summary()`, einschließlich der Details des Pruning-Durchlaufs

Bei Modellen mit mehreren Zielvariablen werden die Abschnitte 4-8 und 10 für jede Zielvariable wiederholt.

Hinweis: Berichte erfordern das R-Paket *quarto* sowie eine Quarto-Installation. Für die PDF-Ausgabe muss zusätzlich eine LaTeX-Distribution installiert sein (z.B. über `tinytex::install_tinytex()`) — andernfalls wird das PDF-Kontrollkästchen ausgeblendet.

13 Die Post-Processing-Module: Elastic Net und GAM

Das earthUI-Paket wird mit zwei Post-Processing-Modulen ausgeliefert — einer Elastic-Net-Routine auf Basis von `glmnet` und einer GAM-Routine auf Basis von `mgcv`. Beide werden jeweils in einem eigenen Beitrag dieser Ausgabe eingehend behandelt; dieser Abschnitt erläutert, *wozu* sie im earthUI-Arbeitsablauf dienen und wie die drei Routinen zusammenwirken.

13.1 Wozu unterstützende Anpassungen?

Das Earth-Modell liefert das Wertschlussfolgerung. Die Post-Processing-Module dienen dazu, dieses zu **bestätigen**: zwei zusätzliche Schätzungen desselben Bewertungsobjekts, erzeugt durch methodisch unterschiedliche Verfahren und angepasst an dieselben Projektdaten. Wenn eine Elastic-Net-Neuschätzung der Earth-Basis und eine unabhängige GAM-Glättung zu einem im Wesentlichen gleichen Wert gelangen, beruht die Schlussfolgerung des Gutachters nicht mehr auf den Annahmen eines einzelnen Modells — ein bedeutsamer Vorteil in Prüfungs- und Gerichtskontexten. Wenn sie hingegen *voneinander abweichen*, ist diese Abweichung selbst diagnostisch aussagekräftig und deutet in der Regel auf dünne Daten, einen zu flexiblen Term oder ein Ausreißer-Vergleichsobjekt hin, das einer erneuten Prüfung wert ist. Keines der Module steht über dem Earth-Modell; sie stellen unterstützende Evidenz dar, die bewusst ohne Gewichtung präsentiert wird.

13.2 Start und gemeinsamer Zustand

Jede Routine ist eine eigenständige lokale Shiny-Anwendung:

- `earthUI::launch_glmnet()` — die Elastic-Net-Anwendung, auf Port 7879.
- `earthUI::launch_mgcv()` — die GAM-Anwendung, auf Port 7880.

Alle drei Anwendungen lesen denselben Projektbaum und dieselbe Einstellungsdatenbank, sodass das in earthUI geöffnete Projekt in den Modulen unmittelbar verfügbar ist — zusammen mit den von earthUI dafür gespeicherten Übernahmeefeldern: der Response (Zielvariable), dem Effective Date (Bewertungstichtag), dem RCA-CQA-Typ und -Wert des Bewertungsobjekts sowie der Wohnflächen-Zuweisung. Diese Angaben müssen nicht erneut eingegeben werden; die Module lesen sie ein, sodass dieselben Werte in alle drei Anpassungen einfließen.

13.3 Die Elastic-Net-Routine

Die `glmnet`-Routine importiert die Basisfunktionen des Earth-Modells — seine Hinge-Funktionen, Interaktionen und linearen Terme — als Spalten einer Designmatrix und schätzt sie unter Lasso-/Elastic-Net-Regularisierung neu (siehe „`glmnetUI` Import“ oben zur Funktionsweise und zum Vorbehalt bezüglich der Gewichtsspalte). Das Ergebnis behält die interpretierbaren, geometrisch aussagekräftigen Terme von earth bei, während der Regularisierer diejenigen Terme schrumpfen oder entfernen kann, die ihren Beitrag nicht rechtfertigen. Eine ausführliche Darstellung findet sich im `glmnetUI`-Beitrag dieser Ausgabe.

13.4 Die GAM-Routine

Die `mgcv`-Routine geht den umgekehrten Weg: Statt die Basis von earth wiederzuverwenden, passt sie neue penalisierte Glättungsterme an und nutzt dabei die Knotenpositionen des Earth-Modells als Ausgangspunkte (siehe „`mgcvUI` Auto-Export“ oben). Da das GAM die funktionalen Formen unter einem anderen Glättungsparadigma neu herleitet, stellt seine Übereinstimmung mit dem Earth-Modell eine tatsächlich unabhängige Bestätigung dar und nicht bloß ein Echo. Unterstützt werden Earth-Modelle mit $\text{degree} \leq 2$. Eine ausführliche Darstellung findet sich im `mgcvUI`-Beitrag dieser Ausgabe.

13.5 Ein praktischer Arbeitsablauf

1. Zunächst wird das **Earth-Modell** in earthUI angepasst und verfeinert, bis es den Anforderungen genügt. Jede erfolgreiche Anpassung ($\text{Grad} \leq 2$) exportiert automatisch die **.rds**-Datei, die von den Modulen eingelesen wird.
2. Anschließend werden die **glmnet**- und die **mgcv**-Routine für dasselbe Projekt geöffnet, das Earth-Ergebnis importiert und die Anpassung durchgeführt. Es empfiehlt sich, jeweils nur in einer Routine zu arbeiten — die Anpassungen sind langlaufend und die Routinen koordinieren sich über das gemeinsame Projekt, sodass ein sequentielles und kein paralleles Vorgehen sinnvoll ist.
3. Es kann frei iteriert werden. Jede Anwendung behält ihre Ergebnisse beim Wechsel zwischen den Anwendungen bei, und es besteht keine Notwendigkeit, unterwegs Berichte zu erzeugen.
4. Erst wenn alle drei Ergebnisse zufriedenstellend sind, wird zu jeder Routine zurückgekehrt und deren Quarto-Bericht erzeugt. Jede Methode registriert ihre Anpassung und ihr Wertschlussfolgerung im Trilogie-Datensatz des Projekts, sodass die drei Berichte auf einen gemeinsamen Vergleichslauf rückführbar bleiben.

Hinweis: Die Module sind auf jeder Ebene optional: Der eigene Arbeitsablauf von earthUI — Anpassung, RCA, Sales Grids, Berichte — ist auch ohne sie vollständig. Sie empfehlen sich, wenn eine Schlussfolgerung einer Bestätigung bedarf: bei einem schwierigen Bewertungsobjekt, einem dünnen Marktsegment oder einem Auftrag, der voraussichtlich einer Prüfung unterzogen wird.

14 Demo-Datensatz: Appraisal_1.csv

earthUI enthält einen MLS-Demo-Datensatz zur Erkundung des Arbeitsablaufs der Wertermittlung. Er lässt sich programmatisch laden mit:

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

Alternativ kann die Datei in der Shiny-App in den Ordner `in/` eines Projekts kopiert (oder beim Anlegen des Projekts als initiale Datendatei angehängt) und in Section 2 (Import Data) ausgewählt werden.

14.1 Beschreibung

Die Datei enthält 1,501 Wohnimmobilienverkäufe sowie 1 Bewertungsobjekt in Zeile 1 (insgesamt 1,502 Zeilen) aus einem simulierten MLS-Export. Die Daten repräsentieren Verkäufe von Einfamilienhäusern in einem Markt mit mehreren Teilgebieten und einer Bandbreite an Objektgrößen, Gebäudealtern und Lagen.

Es handelt sich nicht um reale Daten; sie basieren jedoch auf einer realistischen Nachbarschaft in Nordkalifornien. Sämtliche Identifikationsangaben wurden verändert oder entfernt.

14.2 Spalten

Spalte	Typ	Spezialtyp	Beschreibung
weight	numeric	—	Beobachtungsgewicht (0 = von der Anpassung ausgeschlossen)
id	numeric	display_only	Interne Datensatz-ID
property_id	numeric	display_only	MLS-Objektkennung
listing_id	character	display_only	MLS-Listing-Nummer
parcel_number	character	display_only	Flurstücksnummer des County Assessor (APN)
street_address	character	display_only	Objektadresse
city_name	character	display_only	Stadt
postal_code	character	display_only	Postleitzahl (ZIP)
county_name	character	display_only	County
contract_date	Date	contract_date	Kaufvertragsdatum (dient der Berechnung von sale_age)
sale_age	numeric	—	Tage vom Vertragsdatum bis zum Wertermittlungstichtag (vorberechnet)
coe_date	Date	display_only	Datum des Escrow-Abschlusses
listing_status	character	display_only	Listing-Status (z. B. SSold)
sale_price	numeric	(target)	Verkaufspreis — Zielvariable
rent	numeric	—	Monatsmiete (für Modelle mit mehreren Zielgrößen)
list_price	numeric	display_only	Angebotspreis
original_list_price	numeric	display_only	Ursprünglicher Angebotspreis
living_sqft	numeric	living_area	Bruttowohnfläche in Quadratfuß
beds_total	integer	—	Anzahl der Schlafzimmer
baths_total	numeric	—	Gesamtzahl der Bäder (z. B. 2.5 = 2 volle + 1 halbes)
lot_size	numeric	lot_size	Grundstücksgröße in Quadratfuß
area_id	integer	area	MLS-Gebietskennung
area_text	character	display_only	Gebietsname
age	numeric	actual_age	Gebäudealter in Jahren
year_built	integer	display_only	Baujahr
latitude	numeric	latitude	Geographische Breite (für das Modell auf 3 Nachkommastellen gerundet)
longitude	numeric	longitude	Geographische Länge (für das Modell auf 3 Nachkommastellen gerundet)
latitude6	numeric	display_only	Geographische Breite in voller Präzision
longitude6	numeric	display_only	Geographische Länge in voller Präzision
garage_spaces	integer	—	Anzahl der Garagenstellplätze
fp_count	integer	—	Anzahl der Kamine
no_of_stories	numeric	—	Anzahl der Stockwerke
style	character	—	Architekturstil
view	character	—	Aussichtstyp (z. B. NNeighborhood", "Hills")
days_on_market	integer	dom	Angebotstage am Markt
listing_date	Date	listing_date	Listing-Datum
sale_concessions	numeric	concessions	Verkäuferzugeständnisse

14.3 Empfohlener Schnelleinstieg

1. earthUI starten: `earthUI::launch()`
2. Im New-Project-Assistenten (New Project wizard) ein Projekt mit dem Purpose **For Appraisal** anlegen und dabei `Appraisal_1.csv` als initiale Datendatei anhängen
3. `Appraisal_1.csv` in Section 2 (Import Data) auswählen
4. `sale_price` als Zielvariable auswählen
5. Die Spezialtypen gemäß der obigen Tabelle zuweisen
6. Als Prädiktoren einbeziehen: `sale_age`, `living_sqft`, `baths_total`, `lot_size`, `area_id` (als Faktor), `age`, `latitude`, `longitude`, `garage_spaces`
7. Den Grad (degree) auf 1 setzen und auf **Fit Earth Model** klicken
8. Die Zwischenausgabe herunterladen (Schritt 6) und das CQA-Ranking prüfen
9. Die RCA-Anpassungen (Schritt 7) mit einem CQA-Score von ~5.00 berechnen

10. Das Sales Comparison Grid (Vergleichsraster) generieren (Schritt 8)

Tipp: Der Demo-Datensatz eignet sich gut, um sämtliche Funktionen von earthUI zu erkunden, bevor eigene MLS-Daten importiert werden. Die oben vorgeschlagenen Prädiktoren liefern typischerweise ein Modell mit einem R^2 (Bestimmtheitsmaß) über 0.85.

15 Experimentell: Prolog-Textextraktion & Ableitungsregeln

Version 0.10.0 führt eine **optionale und ausdrücklich experimentelle** Funktionalität zur Extraktion modellfertiger Merkmale aus den beschreibenden Textspalten eines MLS-Exports ein — allen voran der Spalte „Public Remarks“, aber auch Maklerbemerkungen und kommasetrennte Merkmalslisten wie Raumausstattungen oder bestehende Baulichkeiten. Die Extraktions-Engine ist Prolog, insbesondere **Definite Clause Grammars (DCGs)**: Die Regeln, die Wendungen wie „move in ready“, „3 car garage“ oder „owned solar“ im Bemerkungstext erkennen, sind als Prolog-Grammatikregeln formuliert.

Diese Funktion befindet sich noch in der Entwicklung. Sie kann bereits heute *teilweise* zur Extraktion eingesetzt werden — die mitgelieferte Grammatik erkennt bereits einen nützlichen Satz von Merkmalen —, doch ihre Anpassung an das Vokabular des eigenen MLS setzt voraus, Prolog und insbesondere DCGs lesen und schreiben zu können. Wer diese Voraussetzung nicht mitbringt, kann dieses Kapitel vollständig überspringen: Die Funktion ist standardmäßig deaktiviert, und der Rest von earthUI arbeitet ohne sie exakt gleich.

Warnung

Dies ist eine experimentelle Funktion in aktiver Entwicklung. Ihre Extraktionsgrammatik ist weder vollständig noch für den produktiven Einsatz in der Wertermittlung validiert — extrahierte Spalten müssen vor ihrer Verwendung in einem Modell gegen den Quelltext geprüft werden. Ein wirksamer Einsatz der Funktion erfordert praktische Kenntnisse in Prolog und DCGs. Alles, was sie tut, ist additiv und reversibel; bei Fehlverhalten genügt es, den Schalter in den Settings zu deaktivieren und normal fortzufahren.

15.1 Funktionsweise

MLS-Listings enthalten in Freitext- und in listenartigen Feldern eine Fülle wertrelevanter Informationen. Dieser optionale Schritt überführt diesen Text mittels zweier spezieller Spaltenrollen (Kapitel 6) in gewöhnliche Spalten, die von der Regression genutzt werden können:

- **remarks-Spalten** werden von der Prolog-DCG-Grammatik in je eine Merkmalspalte pro extrahiertem Attribut überführt. Das Spaltenpräfix wird aus den Anfangsbuchstaben des Namens der Quellspalte gebildet: `public_remarks` liefert `pr_*`-Spalten, `agent_remark` liefert `ar_*` und so fort.
- **list-Spalten** (z. B. `existing_structures = SSshed, Workshop`) werden in One-Hot-Spalten `<col>_<value>` mit TRUE/FALSE-Werten aufgespalten. Dieser Teil ist reines R und benötigt überhaupt kein Prolog.

Die mitgelieferte Grammatik (`inst/prolog/mls_remarks.pl`) extrahiert derzeit unter anderem: eine abgestufte Zustandsklasse mit einem `condition_score` von 1–5; Aussichts-Flags und Aussichtstypen; Flags für Küchen- und Badmodernisierung; Hartholz- und andere Bodenbeläge; eine Liste und Anzahl von Aufwertungen; Garagenstellplätze und Garagenfläche; Pool, Spa und weitere Außenanlagen; die Anzahl der Kamine; Nebengebäude (Schuppen, Scheune, Werkstatt, Casita, Gästehaus, Zwinger usw.) mit Größen, die aus Maßangaben wie „40x50“ geparkt werden; Stellplätze und Unterstellmöglichkeiten für Wohnmobile (RV); Solareigentum versus Solar-Leasing, Batteriespeicher; Grundstücksgröße (einschließlich der Umrechnung von Acres in Quadratfuß); auf ein Subjekt bezogene Preisangaben aus dem Verkaufskontext (Preisreduktion, HOA-Beiträge, Solar-Leasingrate oder -Ablöse, Mello-Roos); Lagehinweise; sowie einen Sentiment-Score für Vorzüge und Bedenken in $[-1, 1]$.

15.2 Voraussetzungen: vProlog (Trealla), nicht SWI-Prolog

Das Parsen der Remarks läuft über das R-Paket **vProlog** — eine in sich geschlossene Trealla-Prolog-Engine. Es ist nicht auf CRAN verfügbar; die Installation erfolgt aus dem r-universe-Repository des Autors, wie in „Installing earthUI and vProlog“ zu Beginn dieses Artikels gezeigt (vorgefertigte Binärdateien für Windows und macOS, Quelltext unter github.com/wcraytor/vProlog). Eine systemweite Prolog-Installation (SWI-Prolog oder anderes) ist weder erforderlich noch wird sie verwendet. Fehlt vProlog, wird das Parsen der Remarks mit einem Hinweis übersprungen, während die `list`-Aufspaltung weiterhin ausgeführt wird. Die Hilfsfunktion `vProlog_available()` meldet, ob Engine und mitgelieferte Grammatik beide vorhanden sind.

15.3 Aktivierung und Ausführung

1. **Settings** (Zahnrad) öffnen und **Enable Prolog / list column processing** aktivieren. Der Schalter ist standardmäßig deaktiviert, trägt den Hinweis „advanced feature — for users with Prolog expertise“, wird pro Projekt gespeichert und im Project-Information-Dialog angezeigt.
2. In der Prädiktortabelle die Rolle **Special** der betreffenden Spalten auf **remarks** oder **list** setzen (mehrere Spalten können jede der beiden Rollen tragen).
3. Den Abschnitt **Execute Prolog Processing** verwenden, der zwischen den Sections 4 und 5 der Sidebar liegt. Er bietet zwei Schritte:

1. **Expand lists & remarks** — führt die DCG-Grammatik über jede **remarks**-Spalte aus und spaltet jede **list**-Spalte one-hot auf. Die Expansion erfolgt *inkrementell*: Es werden nur die seit dem letzten Lauf neu designierten Spalten verarbeitet, und die Schaltfläche wird deaktiviert, sobald alles Designierte bereits expandiert wurde. Ein Fortschrittsbalken begleitet den Parsevorgang.

2. **Execute derivation rules** — wendet die eigenen Prolog-Regeln `derive/2` auf die expandierten Spalten an (siehe unten). Die Regeln werden vor der Ausführung syntaktisch geprüft und in einem Probelauf validiert.

15.4 Ableitungsregeln

Für jede Datenzeile assertiert earthUI die Spalten der Zeile als Fakten `cell(Name, Value)` und sammelt jede Lösung der eigenen `derive(Col, Val)`-Regeln — jedes distinkte `Col` wird zu einer neuen Spalte, und pro Spalte gewinnt der *erste* gefundene Wert, sodass eine Fallback-Klausel folgen kann. Hilfsprädikate (`count_true/2`, `count_false/2`, `count_true_prefix/2`, `cell_or/3`, `is_na/1`) unterstützen Aggregation und NA-Behandlung, und `drop/1`-Fakten verwerfen Spalten nach dem Regellauf. Beispiel:

```
% 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).
```

Die Regeln liegen in einer projektspezifischen Datei `<project>_rules.pl`, die aus einer vollständig kommentierten Vorlage initialisiert wird, und werden über **Edit derivation rules...** in einem frei schwebenden, verschiebbaren Code-Editor bearbeitet, der sich auch in die Sidebar andocken lässt. Eine Schaltfläche **Check** validiert die Regeln (Syntaxprüfung, Ladeprüfung und ein Probelauf gegen Zeile 1 der Daten), ohne sie auszuführen.

15.5 Versionierte aktive Datendatei

Nach jedem der beiden Schritte wird der erweiterte Data Frame als `<original_name>_<YYMMDD_HHMMSS>.xlsx` in den Eingabeordner des Projekts geschrieben und zur **aktiven** Datendatei des Projekts — beim erneuten Öffnen des Projekts werden die erweiterten Daten geladen, sodass Expansionen nie neu berechnet werden, und dieselbe Datei wird auch von `glmnetUI` und `mgcvUI` konsumiert. Die rohen Remarks-Quellspalten werden automatisch aus dem Modell herausgehalten; die erzeugten Spalten werden *nicht* automatisch als Prädiktoren aufgenommen. Da die Textextraktion viele Spalten erzeugen kann, sollten diejenigen, die (noch) nicht in die Regression eingehen sollen, als **Latent** markiert oder per $\times$ deaktiviert werden (Kapitel 6).

15.6 Programmatische Schnittstelle & mitgelieferte Dateien

Die gesamte Pipeline ist für den Batch-Einsatz exportiert: `execute_prolog_processing(df, remarks_cols, list_cols, rules_text)` orchestriert die Durchläufe und baut auf `remarks_to_columns()`, `split_list_column()`, `apply_user_rules()` und `validate_rules()` auf. Die mitgelieferten Prolog-Quellen unter `inst/prolog/` sind `mls_remarks.pl` (die DCG-Grammatik und die Merkmalsaggregation), `mls_analyze.pl` (ein Datei-E/A-Treiber für das Batch-Parsing), `mls_rules.pl` (die den Ableitungsregeln zur Verfügung stehenden Hilfsprädikate) und `climate_regions_ca.pl` (siehe unten).

15.7 Verwandtes Experiment: Klimaregionen-Priors

Eine begleitende experimentelle Funktionalität adressiert Merkmale mit *dünnere Datenlage* — solche, die die Regression nicht schätzen kann, weil sie nur in einer Handvoll Vergleichsobjekte vorkommen (etwa ein Verdunstungskühler, engl. „swamp cooler“). Drei R-Hilfsfunktionen — `climate_region_for()`, `market_area_profile()` und `climate_feature_priors()` — ordnen den Standort eines Projekts (County + Stadt) einer von sieben käuferseitig wahrgenommenen kalifornischen Klimaregionen zu (eine praxisorientierte Reduktion der 16 Gebäudeklimazonen der California Energy Commission) und liefern je Region ordinale Priors für den Wertbeitrag von HLK- und Kühlmerkmalen (Heizung, Klimaanlage, Verdunstungskühler, Whole-House-Fan, Filtration). Das Prolog-Modul `climate_regions_ca.pl` spiegelt dieselbe Klassifikation für die Verwendung innerhalb von Ableitungsregeln; beide Seiten werden aus einer einzigen kanonischen R-Quelle generiert, sodass sie nicht auseinanderlaufen können. Die Priors sollen die Merkmalsbewertung *informieren* und als Plausibilitätsprüfung geschätzter Koeffizienten dienen — Regionszuordnungen (insbesondere bei Counties, die sich über mehrere Regionen erstrecken) sollten vom Gutachter überprüft werden.

15.7.1 Über Kalifornien hinaus: Jurisdiction Packs und Market-Area-Module

Kalifornien ist die erste Jurisdiktion, nicht die Grenze. `earthUI` wird international eingesetzt, und das Wissen über Klima-Priors wird derzeit als **Jurisdiction Packs** restrukturiert: in sich geschlossene, datenförmige Faktentabellen (Zonendefinitionen, Zuordnungen von Verwaltungseinheiten, Merkmals-Priors), die nach den Länder- und Verwaltungsebenen-Codes des Projektsystems

geschlüsselt sind, sodass ein Pack einen US-Bundesstaat, ein deutsches Land oder ein kleines Land insgesamt beschreiben kann. Eine begleitende Schicht von **Market-Area-Modulen** erfasst lokales Vokabular und verfeinerte Priors für ein benanntes Marktgebiet.

Beide sind darauf ausgelegt, *im Feld* von Valuation Engineers und Gutachtern in deren eigener Sprache verfasst zu werden — und, realistisch betrachtet, mit LLM-Unterstützung: Die Welt verfügt nicht über viele Prolog-Programmierer, weshalb Feldartefakte bewusst Tabellen und keine Programme sind. Das Pack-Format ist ein dokumentiertes Schema mit einem vollständig ausgearbeiteten Beispiel (Kalifornien selbst), ausdrücklich damit ein Assistent wie Claude aus einer umgangssprachlichen Beschreibung eines Marktes ein gültiges Pack entwerfen kann; earthUI generiert das zugrunde liegende Prolog aus dem Pack, und ein blockierender Validator (Schema-, Abdeckungs- und referenzielle Integritätsprüfungen sowie ein Probelauf) muss bestanden werden, bevor ein Pack aktiviert werden kann. Jedes im Feld verfasste Pack trägt verpflichtende Provenienzangaben — den verantwortlichen Autor, das Datum und die zitierte Autorität —, sodass der verantwortliche Fachmann und nicht das Entwurfswerkzeug für das Wissen entsteht und Berichte das exakt verwendete Pack samt Version zitieren können.

Tipp: Ein bewährtes Muster: die Remarks expandieren, die granularen `pr_*`-Flags mit `derive/2`-Regeln zu einer Handvoll abgeleiteter Zähler oder Boolescher Werte konsolidieren, die konsolidierten Roh-Flags per `drop/1` verwerfen und nur die abgeleiteten Spalten als Kandidaten-Prädiktoren behalten. So bleibt die Prädiktorliste kurz und die Kollinearität beherrschbar.

Literatur

- Friedman, Jerome H. (1991). „Multivariate Adaptive Regression Splines“. In: *The Annals of Statistics* 19.1, S. 1–141. DOI: [10.1214/aos/1176347963](https://doi.org/10.1214/aos/1176347963).
- Milborrow, Stephen (1. Okt. 2024). *earth: Multivariate Adaptive Regression Splines*. R package. URL: <https://CRAN.R-project.org/package=earth>.
- (o. D.[a]). *Notes on the earth Package*. URL: <http://www.milbo.org/doc/earth-notes.pdf>.
- (o. D.[b]). *Variance Models in earth*. URL: <http://www.milbo.org/doc/earth-varmod.pdf>.