

earthUI: earth (MARS) 软件包的交互式界面

简体中文译本，由人工智能辅助翻译并经编辑部审校；英文版为权威版本。

William Bert Craytor

第 1 卷，第 1 期 · 2026 年夏季

发表日期: 2026 年 7 月 1 日

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

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

本开放获取文章依据 [知识共享署名 4.0 国际许可协议 \(CC BY 4.0\)](#) 的条款发布。

摘要。 earthUI 是 R 语言 earth 软件包的图形用户界面，该软件包实现了多元自适应回归样条 (MARS, multivariate adaptive regression splines)。它提供三种用途模式——通用预测建模、房地产估价和市场区域分析——并引导用户完成数据导入、模型配置、模型拟合，以及对 earth 诊断信息和图形输出的解读。本文记录了 earthUI 的数据格式要求、建模工作流程、输出显示界面以及完整的功能参考。

关键词： MARS, 多元自适应回归样条, earth, R, 图形用户界面, 房地产估价

出版者按 (2026 年 7 月): 本文描述的是 earthUI 0.10.0。付印前不久, earthUI 0.11.0 吸收了配套应用 glmnetUI 和 mgcvUI (见随后两篇文章), 将 MARS、弹性网络和 GAM 工作流程整合到单一应用程序中。本文所述的界面与工作流程不受此次整合的影响。

目录

1	引言	3
1.1	什么是 earthUI?	3
1.2	一个包, 三个例程	4
1.3	三种用途模式	4
1.4	房地产专用功能	4
1.5	安装 earthUI 与 vProlog	5
1.6	快速入门	5
2	MLS 输入数据要求	7
2.1	文件格式与结构	7
2.2	估价模式所需的列	7
2.3	特殊列命名约定	8
2.4	数据质量与完整性	8
2.5	估价对象行的位置	8
3	通用模式	9
3.1	概述	9
3.2	侧边栏 workflow	9
3.3	主面板选项卡	9
3.4	设置持久化	10
3.5	深色模式	10
3.6	语言环境与区域设置	10
3.6.1	支持的国家	11
3.6.2	纸张覆盖设置	11
3.6.3	保存默认设置	11

4	估价模式	12
4.1	估价对象行的处理	12
4.2	生效日期与销售日龄	12
4.3	特殊列指定	12
4.4	残差贡献（稀有特征估计）	13
4.5	RCA 调整概述	13
5	市场区域分析模式	14
5.1	与估价模式的区别	14
5.2	何时使用市场模式	14
6	变量选择	15
6.1	目标变量	15
6.2	预测变量表	15
6.3	数据类型检测与覆盖	15
6.4	Factor 与 Linear 标志	15
6.5	特殊列类型参考	16
6.6	多个目标变量	16
7	参数选择	17
7.1	基本参数	17
7.2	前向过程	17
7.3	允许的交互	18
7.4	剪枝	18
7.5	交叉验证	18
7.6	子集过滤	18
7.7	推荐参数值	19
7.7.1	前向过程参数	19
7.7.2	剪枝参数	20
7.7.3	交叉验证参数	20
7.7.4	方差模型	21
7.7.5	Earth 输出文件	21
7.8	设置默认值	21
8	拟合 Earth 模型	22
8.1	Fit 按钮	22
8.2	拟合模态框与跟踪输出	22
8.3	拟合日志	22
9	下载数据	23
9.1	输出列	23
9.2	列排序与 Excel 格式	23
9.3	CQA 评分	24
9.4	排序与估价对象行	24
9.5	mgcvUI 自动导出	24
9.6	glmnetUI 导入	24
9.6.1	主要优势	24
9.6.2	重要提示：权重列必须一致	25
9.6.3	工作流程	25

10 RCA 计算与下载	26
10.1 打开 RCA 对话框	26
10.2 CQA 评分插值	26
10.3 输出列	26
10.4 RCA Adjustments 选项卡	27
11 销售比较网格	28
11.1 概述与用途	28
11.2 可比案例选择模态框	28
11.3 网格布局	28
11.4 售价、让价优惠与 Net SP	29
11.5 分组行（位置、地块、房龄）	29
11.6 模型变量行	29
11.7 CQA Residual 与残差特征行	29
11.8 调整后售价公式	29
11.9 工作表保护与可编辑单元格	30
11.10 在 Excel 中使用网格	30
12 下载报告	31
12.1 生成 Quarto 报告	31
12.2 转换 Quarto 报告	31
12.3 报告内容	31
13 后处理模块：弹性网络与 GAM	33
13.1 为什么需要辅助拟合?	33
13.2 启动与共享状态	33
13.3 弹性网络例程	33
13.4 GAM 例程	33
13.5 一个实用的工作流	33
14 演示数据集：Appraisal_1.csv	35
14.1 描述	35
14.2 数据列	35
14.3 建议的快速上手步骤	36
15 实验性功能：Prolog 文本提取与推导规则	37
15.1 它的作用	37
15.2 环境要求：vProlog (Trealla)，而非 SWI-Prolog	37
15.3 启用与运行	37
15.4 推导规则	38
15.5 带版本号的活动数据文件	38
15.6 编程接口与随附文件	38
15.7 相关实验：气候区先验	39
15.7.1 超越加利福尼亚：辖区包与市场区域模块	39

1 引言

1.1 什么是 earthUI?

earthUI 是 R 语言 `earth` 包的图形用户界面，该包实现了多元自适应回归样条 (MARS, multivariate adaptive regression splines) (Friedman 1991)。建模引擎是 Stephen Milborrow 的

earth 包 (Milborrow 2024), Milborrow (n.d.[a]) 与 Milborrow (n.d.[b]) 对其有详细的文档说明。它以本地 Shiny 应用程序的形式运行——无需登录、无需服务器、也没有账户。您从 R 中启动它, 导入数据集 (CSV 或 Excel), 配置模型, 然后交互式地进行拟合。

该应用程序提供了完整的工作流程: 数据导入、变量配置、模型拟合、诊断图、变量重要性、模型方程, 以及可下载的 HTML、Word 或 PDF 格式报告。

1.2 一个包, 三个例程

从当前版本起, earthUI 包还包含了它的两个后处理模块, 这两个模块在早期版本中是独立的包 glmnetUI 和 mgcvUI。安装 earthUI 即安装了全部三个例程; 每个例程都作为独立的本地 Shiny 应用程序运行:

- `earthUI::launch()` ——本文所描述的 earth (MARS) 应用程序, 运行在端口 7878 上。`earth()` 是核心估值引擎。
- `earthUI::launch_glmnet()` ——弹性网后处理模块 (端口 7879), 它导入 earth 模型的基函数, 并在 lasso / 弹性网正则化下重新估计它们 (参见下文“glmnetUI Import”)。
- `earthUI::launch_mgcv()` ——GAM 后处理模块 (端口 7880), 它以 earth 模型的节点位置作为 mgcv 平滑项的起点 (参见下文“mgcvUI Auto-Export”)。

这三个例程共享同一个项目树、同一个设置数据库和同一套支撑基础设施, 因此在一个应用程序中打开的项目可以立即供其他应用程序使用。glmnet 和 mgcv 例程是对 earth 模型价值结论的佐证, 而不是相互竞争的工作流程: 预期的使用模式是先拟合并完善 earth 模型, 然后在需要佐证时针对同一项目运行这些模块。

1.3 三种用途模式

earthUI 以项目的形式组织工作——项目是有名称的实体, 将您的输入文件、输出和设置集中保存在一个固定位置。每个项目都带有一个 **Purpose** (用途), 在创建项目时于 New Project wizard (新建项目向导) 中选择。共有三种模式:

- **General** (通用) ——适用于任何类型的总体或数据集的 Earth 回归。这是默认模式。它提供完整的 MARS 建模工作流程, 不含任何特定领域的附加功能。
- **For Appraisal** (估价用途) ——为房地产估价量身定制的 Earth 回归。增加了针对单一房产估值的功能, 包括估价对象 (subject) 处理、特殊列指定、残差约束法 (RCA), 以及销售比较网格 (Sales Comparison Grid) 的生成。
- **Market Area Analysis** (市场区域分析) ——为市场区域研究量身定制的 Earth 回归。增加了在既定市场中分析成组房产的功能, 而不是对单一估价对象进行估值。

在这三种模式下, 核心建模引擎完全相同——您始终是在拟合一个 Earth (MARS) 模型。用途设置控制的是哪些附加工具和界面元素可用。

每个项目的用途是固定的。项目的用途——连同其国家、州、县和城市——在项目创建时即被锁定。切换到另一个项目会将 earthUI 重置为干净的默认状态 (已导入的数据、模型结果、标签页、变量配置和 earth 参数全部清空), 然后恢复该项目针对其用途保存的配置。

1.4 房地产专用功能

当选择 **For Appraisal** 或 **Market Area Analysis** 时, earthUI 会激活若干为房地产分析设计的功能:

- **特殊列指定**——每个预测变量都可以被标记为某种特殊角色, 例如 `contract_date`、`dom`、`concessions`、`latitude`、`longitude`、`living_area`、`lot_size`、`actual_age`、`effective_age`、`area`、`site_dimensions` 或 `display_only`。这些指定控制该列在拟合、输出和销售网格生成过程中的处理方式。完整参考请参见第 6 章。
- **经纬度取整**——被指定为 `latitude` 或 `longitude` 的列会自动取整到小数点后 3 位, 以防止过拟合。

- **Sale Age 列**——当某列被指定为 `contract_date` 且提供了 Effective Date（生效日期）时，earthUI 会计算一个 `sale_age` 列（销售日期与生效日期之间的天数），并将其作为预测变量代入。
- **RCA 计算（仅限估价模式）**——在估价模式下，模型拟合完成后，earthUI 可以计算残差约束法（RCA）输出，生成逐个可比销售案例的调整、净 / 总调整汇总，以及估价对象的调整后销售价格。
- **销售比较网格（仅限估价模式）**——生成一个多工作表的 Excel 工作簿，其中包含可比案例调整、残差特征分解和调整后的销售价格的公式。参见第 11 章。

提示：General 用途模式适用于任何数据集——金融、科学、工程或房地产。估价模式和市场模式只是为房地产专业人士增加了便利功能。

1.5 安装 earthUI 与 vProlog

earthUI 已发布在 CRAN 上，因此标准安装只需在 R 中执行一行命令：

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

本文描述的是当前版本。如果您阅读本文时 CRAN 上仍是较旧的版本，可以从 GitHub 安装开发版（注意 `subdir` 参数——该包位于仓库的 `pkg/` 文件夹中）：

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

vProlog（可选）。实验性的备注处理功能（参见本文末尾附近的“Prolog Text Extraction”）使用 vProlog 包——一个自包含的嵌入式 Prolog 引擎。它被有意设计为非强制依赖：earthUI 在没有它的情况下也能安装并运行，Prolog 相关功能只是在其安装之前保持禁用状态。vProlog 通过作者的 `r-universe` 仓库分发，该仓库为 Windows 和 macOS 提供预编译的二进制文件（无需编译器工具链）：

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

不需要系统级的 Prolog 安装（无论是 SWI-Prolog 还是其他）——Trealla 引擎已编译进包本身。本文中的其他所有内容在没有 vProlog 的情况下均可正常使用。

1.6 快速入门

使用 earthUI 的步骤：

1. **安装包**——按上文所述从 CRAN 安装。
2. **启动应用程序**：在 R 中运行 `earthUI::launch()`，或从命令行运行 `Rscript -e 'earthUI::launch()'`。应用程序将在您的网页浏览器中打开，端口为 7878。您也可以直接在浏览器中访问 `http://localhost:7878` 来打开应用程序。（后处理模块以同样的方式启动：`launch_glmnet()` 使用端口 7879，`launch_mgcv()` 使用端口 7880。）
3. **创建或打开项目**——在侧边栏的第 1 节中进行。New Project wizard 会记录项目名称、Purpose（General、For Appraisal 或 Market Area Analysis）、位置（国家、州、县、城市），以及可选的初始数据文件，该文件会被复制到项目的 `in/` 文件夹中。
4. **导入数据**——在第 2 节中进行，该节列出活动项目 `in/` 文件夹中的 CSV 和 Excel 文件。您通过 Finder 或资源管理器手动添加的文件，在点击 Refresh（刷新）后会显示出来。
5. **配置变量**——选择目标变量和预测变量，设置数据类型，并指定任何特殊列角色。
6. **设置 Earth 参数**——`degree`、`nprune`、子集过滤器以及其他 `earth()` 参数。
7. **拟合模型**——点击“Fit Earth Model”（拟合 Earth 模型），并在主面板中查看结果。

8. 导出——将预测结果下载为 Excel，生成 Quarto 报告包，或（在估价模式下）计算 RCA 调整和销售比较网格。

配置通过第 3 节和第 4 节中的“Save current as default”（将当前设置保存为默认）按钮保存在服务器端的项目数据库中，以项目和用途为键（参见第 3 章“Settings Persistence”）。

2 MLS 输入数据要求

对于房地产估价和市场分析工作流程，您的输入数据通常来自多重挂牌服务（MLS，Multiple Listing Service）的导出文件。本章描述预期的文件结构以及 earthUI 可以使用的列。

2.1 文件格式与结构

earthUI 接受 CSV 和 Excel (.xlsx、.xls) 文件。导入时，列名会被自动转换为 snake_case——例如，“Living SqFt”变为 living_sqft，“Contract Date”变为 contract_date，“Sale Price”变为 sale_price。这种规范化确保了整个工作流程中列引用的一致性。导入时使用的 CSV 分隔符和小数点符号由区域设置决定（参见第 3 章“Locale & Regional Settings”）。

您的数据文件应为平面表格，每个房产占一行，每个属性占一列。文件的第一行必须包含列标题。

2.2 估价模式所需的列

虽然 earthUI 可以处理任何列集合，但完整的估价工作流程（RCA 调整 + 销售比较网格）在您的 MLS 导出文件中包含以下列时效果最佳：

列	特殊类型	用途
Sale Price	(目标变量)	模型的响应变量
Contract Date	contract_date	用于计算 sale_age（距生效日期的天数）
Listing Date	listing_date	在没有 DOM 列时，与合同日期一起用于计算 DOM
Days on Market	dom	在市天数；显示在销售网格中
Concessions	concessions	销售让步；销售网格中 Net SP = Sale Price – Concessions
Living Area (SF)	living_area	启用每平方英尺残差 (residual_sf、cqa_sf)
Lot Size	lot_size	在销售网格中归入”Site Size Dimensions”行
Site Dimensions	site_dimensions	在销售网格中与地块面积归为一组
Latitude	latitude	取整到小数点后 3 位；用于邻近度计算和 Location 组
Longitude	longitude	取整到小数点后 3 位；用于邻近度计算和 Location 组
Area ID	area	在销售网格中归入”Loc: Long Lat Area”行
Actual Age	actual_age	在销售网格中归入”Actual Age Effective Age”行
Effective Age	effective_age	在销售网格中与实际房龄归为一组
Parcel Number	(无)	显示在销售网格的 APN 行中
Listing ID	(无)	显示在销售网格的 APN 行中
Address	display_only	显示在销售网格的 Address 行中；不参与模型

电子表格的列名可以使用外语——“特殊”名称使用英语，以便 R 程序能对它们进行特殊处理。除此之外，给定的列名会出现在回归模型、图表，以及（在进行估价时）中间销售网格 (Intermediate Sales Grid) 中。

并非所有列都是必需的。earthUI 会自行适应——如果某列缺失，相应的功能会被直接省略。例如，如果没有指定 concessions 列，销售网格中的 Net SP 行将显示未扣除让步的销售价格。不过，对于房地产定价模型，强烈建议提供某些列以获得可接受的拟合效果：

1. **Sale Age**——合同销售日期与估价或分析的生效日期之间的天数。如果使用多年的销售历史数据，尤其是超过 5 年的时段，sale_age 往往在销售价格估计中扮演核心角色。事实上，它常常重要到没有它 earthUI 根本无法给出任何模型。
2. **Living Area**——也称作“Living Sqft”、“GLA” (gross living area, 总居住面积) 等。这是销售价格的另一个主要决定因素。

3. **Total Bath Count**——全卫、四分之一卫、半卫和四分之三卫浴的总数。例如，两个全卫加一个半卫的值为 2.5。
4. **Garage Bays** 或 **Garage Area**——车库车位数量或车库面积（平方英尺）。
5. **Lot Size**——房产的土地面积，通常以平方英尺或英亩为单位。
6. **Longitude、Latitude**，如果有的话再加上 **Area ID**。这些变量的调整将在销售网格中合并为单一的 Location 调整项。

2.3 特殊列命名约定

earthUI 通过列的特殊类型指定而非列名来识别列。您可以在 MLS 导出文件中为列取任何名称——关键在于您在变量配置表（第 6 章）中指定了正确的特殊类型。

例如，您的 MLS 可能将居住面积导出为“GLA”、“Living SqFt”、“liv_area”或“gross_living_area”。导入之后（此时它已变为 snake_case），您只需在 Special dropdown（特殊类型下拉框）中将其指定为 living_area。此后，无论其原始名称如何，earthUI 都会将其用于每平方英尺残差计算和销售网格分组。

2.4 数据质量与完整性

- **缺失值 (NA)**：任何预测变量或目标变量列中含有 NA 值的行会在拟合前自动删除。na.action 在内部始终设置为 na.fail，NA 的删除在调用 earth() 之前完成。
- **日期列**：日期应采用 R 可解析的格式（例如 2025-06-15、06/15/2025、June 15, 2025）。当至少 80% 的值能够成功解析时，earthUI 会自动检测日期列。两位数年份（例如 06/15/25）会被读取为最接近的匹配年份，因此近期的 MLS 导出文件可以正确解析；四位数年份始终按字面值处理。
- **数值列**：销售价格、居住面积、地块面积、让步及类似字段必须是数值型。如果您的 MLS 导出的价格带有货币符号或千位分隔符（例如“\$350,000”或“350.000,00”），您可能需要在导入前清理这些数据。
- **因子列**：区域 ID、风格或状况等分类变量应包含数量适中的唯一值。earthUI 会自动检测因子，但您可以覆盖该检测结果。

提示：导入后请查看预测变量表中的 T/F/NA 列——它显示每列的 true/false/NA 计数。缺失值超过 30% 的列会以红色标出。请考虑排除高缺失率的列，或在导入前清理数据。

2.5 估价对象行的位置

在 **Appraisal**（估价）模式下，第 1 行必须是估价对象（subject）。其余所有行均为可比销售案例。估价对象的销售价格可以留空（NA）或设为任意值——无论如何，earthUI 在拟合时都将其视为 NA。

在 **Market Area Analysis** 模式下，将估价对象放在第 1 行是可选的，且不会自动排除任何行。若要将估价对象所在行（或任何其他行）排除在拟合之外，可在指定的权重列中将其值设为零，或使用子集过滤器。

在 **General** 模式下，没有特殊的行处理——所有行一视同仁。

3 通用模式

3.1 概述

通用模式 (General Purpose mode) 是启动 earthUI 时的默认模式。它为任何数据集——而不仅仅是房地产数据——提供完整的多元自适应回归样条 (MARS, multivariate adaptive regression splines) 建模工作流。您可以将 earthUI 用于科学数据、金融分析、工程研究, 或任何希望探索变量间非线性关系与交互作用的回归问题。

在通用模式下, 界面会省略房地产专用功能 (特殊列、销售日龄、坐标舍入、RCA)。侧边栏经过精简, 聚焦于变量选择、参数配置、模型拟合与导出。

3.2 侧边栏工作流

侧边栏按编号组织为可折叠的分区, 引导您从项目选择一路走到导出:

1. Project (项目)——选择一个现有项目, 或通过 New Project (新建项目) 向导创建新项目; 该向导会记录用途、国家和行政区位置 (创建后全部锁定)。当前活动项目决定了输入文件的读取位置以及所有输出的写入位置——不存在单独的输出文件夹字段。

2. Import Data (导入数据)——从活动项目的 in/ 文件夹中选择 CSV 或 Excel 文件; Refresh (刷新) 链接会重新扫描该文件夹, 以发现手动添加的文件。对于包含多个工作表的 Excel 文件, 会出现工作表选择器。列名会被自动转换为 snake_case。

3. Variable Configuration (变量配置)——目标变量选择器 (支持多个目标变量)、Effective Date (生效日期), 以及带有 Include、Factor、Linear 和 Latent 复选框的预测变量表。详见第 6 章。

4. Earth Call Parameters (Earth 调用参数)——earth() 函数的全部参数: degree、penalty、nk、剪枝方法、交叉验证等。完整的参数参考见第 7 章。

Execute Prolog Processing (执行 Prolog 处理)——位于第 4 与第 5 分区之间的一个可选、无编号分区, 除非在 Settings (设置) 中启用了实验性的 Prolog 功能, 否则处于禁用状态。见第 14 章。

5. Fit Earth Model (拟合 Earth 模型)——一个随机种子字段和异步运行模型的按钮。拟合的详细说明见第 8 章。

6. Download Estimated Target Variable(s) & Residuals (下载估计的目标变量与残差)——(在估价模式和市场模式下标题为“Download Estimated Sale Prices & Residuals”, 即下载估计售价与残差) 将预测值、残差、CQA 评分以及各 g 函数的贡献导出为一个 Excel 文件。见第 9 章。

7–8. RCA Adjustments and Sales Grid (RCA 调整与销售网格)——仅限估价模式。见第 10 章和第 11 章。

9. Generate Quarto Report (生成 Quarto 报告)——将一个自包含的 Quarto 报告包 (.qmd 源文件及资源文件) 写入项目的输出文件夹。在非估价模式下编号为 7。见第 12 章。

10. Convert Quarto Report (转换 Quarto 报告)——将任意 .qmd 文件渲染为 HTML、Word 或 PDF。见第 12 章。

3.3 主面板选项卡

拟合完成后, 主面板提供十个选项卡:

选项卡	内容
Data (数据)	导入数据的预览。在估价模式下，拆分为“Subject Property”（估价对象）和“Comparable Sales”（可比销售案例）两个表格。
Equation (方程)	以 LaTeX/MathJax 显示的拟合模型方程，展示每个基函数及其系数。
Summary (摘要)	关键指标 (R^2 、GRSq、GCV、RSS、 $CV R^2$) 和系数表。
Variable Importance (变量重要性)	预测变量重要性得分的条形图和排名表。
Contribution (贡献)	g 函数贡献图。从下拉菜单中选择一个 g 函数；单变量预测变量显示折线图，双变量交互项 ($\text{degree} \geq 2$) 显示 3D 透视图和等高线图。
Correlation (相关性)	预测变量相关性的热力图。
Diagnostics (诊断)	三幅图：残差 vs 拟合值、正态 Q-Q 图，以及实际值 vs 预测值。
RCA Adjustments (RCA 调整)	（仅估价模式）各可比销售案例的残差调整百分比 (Residual Adj %)、净调整百分比 (Net Adj %) 和总调整百分比 (Gross Adj %) 的直方图，在第 7 步 (Calculate RCA Adjustments) 完成后填充。
Residual Contribution (残差贡献)	通过将模型残差对模型从未使用过的列进行联合回归，估计稀有特征或被排除特征的市场贡献（见 subsection 4.4 ）。
ANOVA	ANOVA 分解表，显示每个基函数对 RSS 的贡献。
Earth Output (Earth 输出)	来自 ‘earth::summary()’ 的原始文本输出，包括剪枝过程的详细信息。

3.4 设置持久化

earthUI 将您的配置保存在服务器端的项目数据库中（位于 regProj 根目录下的 `projects.sqlite`），以项目和用途为键。保存是按钮驱动的：第 3 分区中的“Save current as default”（将当前设置保存为默认）按钮保存目标变量、生效日期和预测变量配置，而第 4 分区中的按钮保存 earth 调用参数和 Allowed Interactions（允许的交互项）矩阵。每个按钮只更新自己负责的那一半，因此它们绝不会相互覆盖，并且拟合不会自动保存——您可以自由试验而不必担心覆盖已保存的默认设置。一个项目的多个输入文件（一个小型测试抽样、完整数据集）共享同一份配置；而且由于设置存放在项目目录树中，当您同步、备份或共享 regProj 文件夹时，设置会随之迁移。

3.5 深色模式

点击右上角的月亮/太阳图标可在浅色与深色主题之间切换。主题偏好保存在本地存储中，并在会话之间持续生效。

3.6 语言环境与区域设置

earthUI 通过基于国家的语言环境系统支持国际化的数字、日期和 CSV 格式约定。顶部菜单栏 **Settings**（设置，齿轮图标）菜单中的 **Country**（国家）下拉菜单可为 31 个受支持的国家选择预设。每个预设配置以下内容：

- **CSV 分隔符**——美国/英国/日本使用逗号（,），欧洲大部分国家使用分号（;），因为那里逗号被用作小数点。
- **小数点符号**——句点（.）或逗号（,）。
- **千位分隔符**——逗号（美国/英国/日本）、句点（德国/意大利/西班牙）、空格（芬兰/法国/波兰/波罗的海国家/乌克兰/俄罗斯）或撇号（瑞士）。

- 日期格式——MM/DD/YYYY（美国）、DD/MM/YYYY（欧洲大部分国家）或 YYYY-MM-DD（瑞典/立陶宛/日本/加拿大）。
- 纸张尺寸——Letter（美国/加拿大/墨西哥）或 A4（其他所有国家）。

3.6.1 支持的国家

国家	CSV 分隔符	小数点	千位分隔符	日期	纸张	
美国、加拿大 (csv)	,	.	,	MDY/YMD	Letter	
英国、爱尔兰、澳大利亚、新西兰	,	.	,	DMY	A4	
德国、奥地利、比利时、荷兰、意大利、西班牙、葡萄牙、丹麦、巴西	;	,	.	DMY	A4	
法国、波兰、捷克、挪威、拉脱维亚、爱沙尼亚	;	,	(空格)	DMY	A4	
芬兰、瑞典	;	,	(空格)	DMY/YMD	A4	
瑞士	;	,	'	DMY	A4	
乌克兰、俄罗斯	;	,	(空格)	DMY	A4	
土耳其	;	,	.	DMY	A4	
日本、韩国	,	.	,	YMD	A4	
墨西哥	,	.	,	DMY	Letter	
立陶宛	;	,	(空格)	YMD	A4	

3.6.2 纸张覆盖设置

在国家选择器下方，**Paper**（纸张）下拉菜单（Letter 或 A4）允许您在不切换国家的情况下覆盖报告页面尺寸。CSV 分隔符、小数点符号和日期格式顺序仍遵循所选国家的预设；当您更改国家时，纸张覆盖设置会重置为该国家的默认值。

3.6.3 保存默认设置

点击 Settings 菜单中的 **Save**（保存）可全局存储您的语言环境偏好。这些默认设置适用于今后所有会话，无论您打开哪个项目。项目级设置（目标变量、预测变量、参数）单独保存在项目数据库中；语言环境默认设置则通过用户级 SQLite 数据库在所有项目间持续生效。这种两级设计面向审计事务所等需要处理来自多个国家数据的机构——将您最常用的国家设为默认，需要时再进行调整。

提示：图表坐标轴和斜率标签上的数字格式会自动适配您的语言环境。德国语言环境使用句点作为千位分隔符（200.000），芬兰使用空格（200 000），瑞士使用撇号（200'000）。不显示任何货币符号——*earthUI* 与货币无关。

4 估价模式

当您选择 **For Appraisal**（用于估价）作为用途（Purpose）时，earthUI 会将自身配置为面向单一物业估值。第 3 章描述的所有功能仍然可用；本章仅涵盖估价模式特有的附加功能。

4.1 估价对象行的处理

在估价模式下，数据集的第 1 行是估价对象（**subject**）物业，其余所有行均为可比销售案例。您的输入文件必须按此方式组织（见第 2 章）。估价对象的售价可以留空或设为任意值——earthUI 在拟合时会自动将其视为 NA。

导入后，Data 选项卡会拆分为两个部分：**Subject Property**（估价对象物业，第 1 行）和 **Comparable Sales**（可比销售案例，第 2 行及之后）。第 1 行始终被排除在模型拟合之外——通知消息“Skipping row 1 (subject). Fitting on N rows.”（跳过第 1 行（估价对象）。在 N 行上拟合。）会对此加以确认。拟合后，模型仍会为估价对象行生成预测值，在输出中显示为 `est_<target>`。

4.2 生效日期与销售日龄

在每种用途模式下，Variable Configuration 分区中都会出现一个 **Effective Date**（生效日期）字段（默认为当天日期）——通用分析和市场分析同样有生效日期，而不仅仅是估价——但它只在估价模式和市场模式下才会被使用。在这两种模式下，如果您在 Special（特殊列）下拉菜单中将某列指定为 `contract_date`，earthUI 会计算一个 `sale_age` 列——即每笔销售的合同日期与生效日期之间的整数天数。该列会取代原始日期列作为预测变量。

在指定合同日期后第一次点击 Fit 时，earthUI 会创建 `sale_age` 列，并通知您再次点击 Fit 以将其纳入模型。

4.3 特殊列指定

在估价模式和市场模式下，Variable Configuration 表中的每个预测变量旁都会出现一个 **Special**（特殊）下拉菜单。特殊类型及其作用的完整列表如下：

特殊类型	作用
no	默认——无特殊处理
contract_date	触发基于生效日期的 sale_age 自动计算
listing_date	在不存在 dom 列时，与合同日期一起用于计算 DOM
dom	标识“在售天数”(Days on Market) 列；显示在销售网格 (Sales Grid) 中
concessions	标识销售让步 (concessions)；用于销售网格中的净售价公式 (售价 - 让步)
latitude	数值舍入到小数点后 3 位；用于 Haversine 邻近度计算及销售网格中的 Location 分组
longitude	数值舍入到小数点后 3 位；用于 Haversine 邻近度计算及销售网格中的 Location 分组
area	在销售网格中归入“Loc: Long Lat Area”行
living_area	在下载输出中启用按平方英尺计算的残差 (residual_sf、cqa_sf)
lot_size	在销售网格中归入“Site Size Dimensions”行
site_dimensions	在销售网格中与地块面积归为一组
actual_age	在销售网格中归入“Actual Age Effective Age”行
effective_age	在销售网格中与实际房龄归为一组
display_only	该列包含在导出中，但排除在模型拟合之外（允许多列）
remarks	（实验性）自由文本备注列，由可选的 Prolog 步骤解析为特征列（允许多列）；见第 14 章
list	（实验性）逗号分隔的列，由可选的 Prolog 步骤拆分为独热 (one-hot) TRUE/FALSE 列（允许多列）；见第 14 章

每种特殊类型只允许指定一列 (display_only、remarks 和 list 除外)。将某个特殊类型指定给第二列时，会自动从第一列上清除该指定。变量名旁边会出现一个蓝色小徽章，显示其被指定的特殊类型。

4.4 残差贡献（稀有特征估计）

对于因过于稀有而无法进入回归通道的特征——例如仅在少数几笔销售中出现的车间 (workshop)——它们属于 RCA 残差层。**Residual Contribution** (残差贡献) 选项卡通过普通最小二乘法，将拟合模型的残差对模型从未使用过的列进行回归，以估计这些特征的市场贡献。选择多列时会对它们进行联合估计，因此相关的配套设施 (例如 outbuilding_sqft 加上 pr_has_workshop) 会分摊贡献，而不会重复计算。二元标志列产生一次性总额估计，并列出每笔带该标志的销售的残差作为证据；数值列产生每单位贡献；报告的标准误为一个 SE (通常以 $t \geq 2$ 作为显著性判断标准)。支持观测数少于十个的项会在生成的工作底稿注记中自动获得 CAUTION (警示) 标记；该注记以一个段落记录方法、估计值和注意事项，供报告使用。只有模型实际使用过的预测变量才会被排除在候选列表之外——一个曾提供给拟合但从未被前向过程选中的稀有标志列仍然符合条件，因为模型没有吸收它的任何价值。

4.5 RCA 调整概述

Calculate RCA Adjustments & Download (计算 RCA 调整并下载) 按钮 (侧边栏第 7 分区，仅在估价模式下拟合完成后可见) 计算每个可比销售案例相对于估价对象的市场推导调整值。完整的 RCA 工作流在第 10 章中描述。

计算完 RCA 调整后，**Generate Sales Grid & Download** (生成销售网格并下载) 按钮 (侧边栏第 8 分区) 变为可用。销售比较网格 (Sales Comparison Grid) 工作流在第 11 章中描述。

提示：您为估价对象指定的 CQA 评分决定了残差分布中有多大比例归属于估价对象。评分 5.00 会将估价对象置于可比销售案例的中位数位置。

5 市场区域分析模式

当您选择 **Market Area Analysis**（市场区域分析）作为用途时，earthUI 提供与估价模式相同的房地产专用功能（特殊列、销售日龄、坐标舍入），但其定位是分析一组物业，而不是对单一估价对象进行估值。

5.1 与估价模式的区别

- **不自动排除估价对象**——与估价模式不同，市场模式不会自动将第 1 行排除在拟合之外。所有行都会被纳入；若要将估价对象行排除在拟合之外，可在指定的权重列中为其赋零值（如演示数据集所示），或使用子集过滤器。
- **没有 RCA 或销售网格分区**——“Calculate RCA Adjustments & Download”和“Generate Sales Grid & Download”步骤不可用。市场模式专注于模型拟合与输出，而非逐个可比案例的调整。
- **侧边栏编号**——由于没有 RCA 和销售网格步骤，Generate Quarto Report 分区的编号为 7 而非 9（Convert Quarto Report 分区仍为 10）。

5.2 何时使用市场模式

以下情形适合使用市场区域分析模式：

- 为某个社区或市场区域构建回归模型，以理解价值驱动因素
- 分析建筑面积、房龄、地块面积和区位等变量如何影响一组物业的售价
- 为市场状况分析或社区界定准备支持材料
- 使用的数据集第 1 行包含一个估价对象物业，但您希望可以自由选择将其纳入或排除在拟合之外

提示：市场模式也适用于一般的房地产回归场景——您需要特殊列功能（销售日龄、坐标舍入），但不需要 RCA 调整 workflow。

6 变量选择

侧边栏第 3 节——**Variable Configuration**（变量配置）——用于选择哪些列参与模型，以及它们如何被处理。

6.1 目标变量

第 3 节顶部的 **Target (response) variable(s)**（目标（响应）变量）下拉框列出数据集中的每一列。选择一列可拟合标准的单响应模型；选择多列则拟合多响应模型。选择多个目标变量时，模型使用 `earth(cbind(y1, y2, ...) ~ .)` 同时拟合所有响应。

被选为目标变量的列会自动从预测变量列表中排除。

6.2 预测变量表

在目标变量选择器下方，一个表格列出其余每一列，包含以下字段：

列	说明
Variable （变量）	列名（若被截断，完整名称显示在工具提示中），前面带有一个 ×/+ 切换按钮（见下文）。在估价/市场模式下，蓝色徽章显示所分配的特殊类型。
Type （类型）	数据类型下拉框：numeric、integer、character、logical、factor、Date、POSIXct
Include （纳入）	复选框——将此列作为预测变量纳入模型
Factor （因子）	复选框——将此列视为分类变量
Linear （线性）	复选框——强制仅以线性方式进入模型（不生成铰链函数）
Latent （潜变量）	复选框——将此列完全排除在模型之外（既不进入 earth，也不包含在导出到 glmnetUI / mgcvUI 的数据中），保留用于潜变量分析
Special （特殊类型）	下拉框（仅估价/市场模式）——见下文“特殊列类型参考”
T/F/NA	每列三个计数：对逻辑列为 TRUE/FALSE/NA；对数值列为 ≠ 0/= 0/NA。当缺失值超过 30% 时以红色显示

表格上方的提示行解释了这些缩写：“Type = 列数据类型，Inc = 作为预测变量纳入，Factor = 视为分类变量，Linear = 仅线性（无铰链），Special = 列角色（例如 contract_date）。”

软禁用而非删除。每个变量名旁的 × 切换按钮会将该行变灰，并使该列不参与模型，但不会将其从数据中删除；按钮随之变为绿色的 +，可重新启用该列。列永远不会被销毁。

6.3 数据类型检测与覆盖

earthUI 在导入时自动检测数据类型。数值、整数、逻辑、因子和日期列都会被识别。看起来像日期的字符列（至少 80% 的值能按常见日期格式解析）会被归类为 Date。

你可以通过更改 **Type** 下拉框来覆盖任何检测结果。当你将某列改为 character 或 factor 时，Factor 复选框会自动勾选。更改类型会影响该列传递给 earth() 函数的方式。

6.4 Factor 与 Linear 标志

- **Factor**——勾选后，该列被视为分类变量。这适用于 style、area_id 或 grade 等表示离散组别而非连续测量值的列。因子列以指示（哑）变量的形式进入模型。
- **Linear**——勾选后，该列被强制以线性方式进入模型——不会为其创建铰链（分段线性）函数。当你确定某个变量与目标变量之间存在严格的线性关系时，此选项很有用。

6.5 特殊列类型参考

在估价和市场模式下，**Special** 下拉框提供以下选项。每种类型最多只能分配给一列（**display_only**、**remarks** 和 **list** 除外，它们允许多列）：

日期与时间类型：

- **contract_date**——触发自动计算 **sale_age**。原始日期列被替换为一个整数列，度量销售日期与 Effective Date（评估基准日）之间的天数。
- **listing_date**——当未指定明确的 **dom** 列时，用作计算在市天数（DOM = 合同日期 - 挂牌日期）的后备来源。
- **dom**——标识在市天数（Days on Market）列。显示在 Sales Grid（销售比较网格）的 APN 行和 Date of Sale 行中。

货币类型：

- **concessions**——标识销售让步（卖方信贷、买方激励等）。在 Sales Grid 中用于计算净售价：Net SP = 售价 - 让步。

面积与位置类型：

- **latitude**——值自动四舍五入到小数点后 3 位以防止过拟合。用于 Haversine 邻近度计算（估价对象（subject）到每个可比销售案例的距离），并归入 Sales Grid 的 Location 行。
- **longitude**——与 latitude 相同的舍入处理。用于邻近度计算和 Location 组。
- **area**——通常为社区或区域标识符。与 latitude、longitude 一起归入 Sales Grid 的 “Loc: Long | Lat | Area” 行。
- **living_area**——启用下载输出中的每平方英尺残差计算（**residual_sf** 和 **cqa_sf**）。
- **lot_size**——归入 Sales Grid 的 “Site Size | Dimensions” 行。
- **site_dimensions**——与 lot size 一起归入 Sales Grid（例如 “75x120”）。

房龄类型：

- **actual_age**——归入 Sales Grid 的 “Actual Age | Effective Age” 行。
- **effective_age**——与 actual age 一起归入 Sales Grid。

显示类型：

- **display_only**——该列包含在 Excel 导出中，但完全不参与模型拟合。适用于地址字段、MLS 编号、宗地编号或其他不应作为预测变量的参考数据。允许多列使用此指定。

文本提取类型（实验性）：

- **remarks**——一个自由文本列（例如 **public_remarks**），由可选的 Prolog 步骤解析为特征列。在 Settings（设置）中启用该功能之前不起任何作用。允许多列。参见第 14 章。
- **list**——一个逗号分隔的列（例如 **existing_structures**），由同一步骤拆分为独热编码的 **<col>_<value>** TRUE/FALSE 列。允许多列。参见第 14 章。

6.6 多个目标变量

选择多个目标变量会拟合一个多响应 Earth 模型。选择多个目标变量时：

- 模型同时预测所有响应，各目标变量共享基函数
- **wp (response weights)**（响应权重）按钮被激活，允许为每个目标变量分配数值权重
- 方差模型（**varmod.method**）被禁用（多响应模型不支持）
- 结果标签页会显示每个响应的指标、方程和诊断图

提示：指定为 **display_only** 的列保留在数据集中并出现在 Excel 导出中，但完全不出现在预测变量表中。适用于 ID 列、地址或其他参考字段。

7 参数选择

侧边栏第 4 节——**Earth Call Parameters** (Earth 调用参数)——提供对 `earth()` 函数所接受的全部参数的访问。每个参数都带有一个蓝色帮助图标 (?) 及工具提示说明。参数按可折叠的小节组织。

7.1 基本参数

参数	默认值	说明
subset	(空)	行过滤表达式。见下文“子集过滤”。
weights	NULL	用于个案 (行) 权重的列选择器。仅列出数值列。
wp	NULL	多目标模型的响应权重。按钮会打开一个对话框，为每个目标变量提供一个数值输入 (默认各为 1.0)。
keepxy	关闭	在模型对象中保留 x、y、subset 和 weights。
trace	0	拟合输出的跟踪级别 (0 到 5)。
glm	无	可选的 GLM 族: gaussian、binomial 或 poisson。
degree	1	最大交互阶数。将 degree 设为 ≥ 2 会自动启用交叉验证，并显示 Allowed Interactions (允许的交互) 矩阵。
penalty	2	每个节点的 GCV 惩罚。值越高，模型越简单。

7.2 前向过程

参数	默认值	说明
nk	自动	剪枝前的最大模型项数。自动填充为所加载数据的推荐值 (见下文“推荐参数值”)。
thresh	0.001	前向步骤阈值。值越小，允许的模型项越多。
minspan	自动	节点之间的最小间隔。自动填充为推荐值。负值设置每个预测变量的最大节点数。
endspan	自动	端部间隔——节点到数据边缘的最小距离。自动填充为推荐值。
newvar.penalty	0	引入新变量的惩罚 (鼓励重用已有的预测变量)。
fast.k	20	快速 MARS 算法中考虑的父项数量。
fast.beta	1	控制快速 MARS 的老化因子。

7.3 允许的交互

当 **degree** 设为 2 或更高时，基本参数下方会出现一个交互矩阵。它是一个上三角形的复选框网格，每个预测变量对对应一个复选框。勾选表示允许这两个预测变量交互；取消勾选则禁止该特定交互。

点击预测变量名称（行或列标签）可切换该变量的所有交互。顶部的 **Allow All**（全部允许）和 **Clear All**（全部清除）复选框提供批量控制。矩阵使用固定表头，滚动时行、列标签保持可见。

一条信息提示会提醒你：“交互项会增加过拟合的风险。交叉验证已启用（10 折）。”

7.4 剪枝

参数	默认值	说明
pmethod	backward	剪枝方法：backward、none、exhaustive、forward、seqrep 或 cv。
nprune	NULL	剪枝后的最大模型项数。留空则自动选择。

7.5 交叉验证

参数	默认值	说明
nfold	10	交叉验证折数。设为 0 可禁用交叉验证。
ncross	20	交叉验证重复次数。
stratify	开启	对交叉验证样本进行分层，使每一折具有相似的响应分布。

当 $\text{degree} \geq 2$ 时，交叉验证会自动启用（**nfold** 设为 10），以帮助防范交互项带来的过拟合。

7.6 子集过滤

subset 文本输入框接受一个 R 表达式，用于筛选哪些行参与模型拟合。你可以直接输入表达式（例如 `sale_age < 365 & area_id == 460`），也可以使用 **Build filter...**（构建过滤器…）按钮以可视化方式构建。

Build Filter Dialog（构建过滤器对话框）——点击“Build filter...”打开一个引导式对话框。每个条件行包含一个列下拉框、运算符（`<`、`>`、`<=`、`>=`、`==`、`!=`），以及一个根据列类型自适应的值输入框：数值列为数字输入，日期列为日期选择器，字符/因子列为唯一值下拉框。各条件通过 **AND**（`&`）或 **OR**（`|`）连接符相连。底部的预览会显示表达式以及匹配的行数。点击 **Apply**（应用）将表达式插入文本输入框。

日期列在手动表达式中必须使用 `as.Date("...")` 或 `as.POSIXct("...")` 包装。Build Filter 对话框会自动处理这一点。

子集过滤是非破坏性的——被排除的行仍保留在数据集中，并在 Excel 导出中获得预测值。

提示：将 **OR** 条件与 **AND** 组合时，请使用括号进行分组。若不加括号，`A & B | C` 会被解释为 `(A & B) | C`，这可能不符合你的本意。

7.7 推荐参数值

earthUI 会在第 4 节的关键参数下方显示推荐值。这些推荐值会根据拟合行数 (n) 和所选预测变量数 (p) 动态更新。这些公式来源于 Friedman 的 MARS 论文、earth 的内部算法以及经验测试。**nk**、**minspan** 和 **endspan** 字段会在数据加载时预填充其推荐值；手动编辑的值会被保留。

7.7.1 前向过程参数

nk (剪枝前的最大模型项数):

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

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

Earth 的默认值 $\min(200, \max(20, 2p)) + 1$ 未考虑数据集规模，对大数据集而言可能过小，会在前向过程充分探索所有预测变量之前就对其加以限制。 $\lfloor n/10 \rfloor$ 一项允许大约每 10 个观测值对应一个模型项——这是避免过度参数化的标准经验法则。100 的上限可避免收益递减。

minspan (节点之间的最小观测数):

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

endspan (距数据边界的最小观测数):

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

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

Earth 自动计算的 minspan 使用 Friedman 的公式 43: $\lfloor (-\ln(-\ln 0.95) + \ln(p \cdot n)) / (2.5 \ln 2) \rfloor$, 它按 $\ln(np)$ 缩放，对大数据集增长过慢。在 $n=1500$ 时它仅得出 ≈ 7 ，使每个连续预测变量约有 245 个候选节点位置——太多了，会让前向过程去拟合噪声。推荐公式的目标是在大 n 时每个预测变量约有 100 个候选位置，而在小 n 时（此时 earth 的自动计算经过了充分检验）沿用其自动计算结果。endspan 设置得略大于 minspan，以提供额外的边界保护，这在数据尾部稀薄时（房地产数据中很常见）很有帮助。

penalty (每个节点的 GCV 惩罚):

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

这正是 earth 自身的默认值——加性模型为 2，交互模型为 3（更高的惩罚补偿了更大的搜索空间）。

newvar.penalty（引入新预测变量的惩罚）：

推荐值：当预测变量之间存在相关性时（例如居住面积与卧室数、浴室数、地块面积相关）取 **0.1**。这会使前向过程倾向于重用模型中已有的预测变量，而不是引入相关的替代变量。其机制是：每个新预测变量的 RSS 改进量会乘以 $1/(1 + \text{penalty})$ 。当 **newvar.penalty** = 0.1 时，新变量的 RSS 改进必须比在已有变量上再加一个节点至少高出 10%。这会产生预测变量更少、更简单的模型，且不影响最终的系数估计（惩罚在选择完成后即被移除）。若预测变量之间不相关，则保持为 0。注意，earth 不支持在使用个案权重的同时设置非零的 **newvar.penalty**——一旦指定了权重列，earthUI 就会禁用该输入（强制为 0）。

7.7.2 剪枝参数

pmethod：推荐值：**backward**。默认的基于 GCV 的后向剪枝是确定性的——相同的数据总是产生相同的模型，不依赖任何随机种子。这对可复现性很重要，在估价工作中尤其如此。cv 方法使用交叉验证进行剪枝，会引入对种子的依赖。

nprune：推荐值：留空 (NULL)。让 GCV 选择最优的模型项数。设置 **nprune** 会施加一个硬性上限，凌驾于 GCV 的判断之上。

7.7.3 交叉验证参数

nfold（交叉验证折数）：按样本量区间推荐：

<i>n</i> （拟合行数）	nfold
0–299	5
300–399	6
400–499	7
500–599	8
600–699	9
700+	10

折数越多，每一折的训练数据越多（估计偏差更小），但留出的测试折会随之缩小。该表随着样本增大逐级升至标准的 10 折，同时保持每个测试折足够大（约 50 行以上）以获得稳定的单折指标；对小样本则回退到 5 折——此时 CVR^2 估计仍有噪声，因此应提高 **ncross**（例如 15–20）以平均掉方差。上限 10 反映的是：每增加一折都是一次完整的 earth 拟合，而诊断收益微乎其微。当 **pmethod** = "backward" 时，交叉验证不影响模型——它只计算诊断性的 CVR^2 ，并为方差模型提供残差。

ncross（交叉验证重复次数）：按样本量区间推荐：

<i>n</i> （拟合行数）	ncross
0–199	20
200–399	16
400–699	12
700–1199	10
1200–2999	7
3000+	5

重复可以平均掉折划分中的运气成分，从而稳定 CVR^2 ；样本越大所需重复越少，因为单次交叉验证本身已经更稳定。重复交叉验证的残差还会输入方差模型。

7.7.4 方差模型

varmod.method: 推荐值: **lm**——Milborrow 的默认值：用于预测区间的稳健线性方差模型（离散度是从带噪声的残差中估计的，因此简单模型的泛化能力更好）。如果 **lm** 在你的数据上无法收敛，可尝试 **rlm** 或 **const**（同方差，总能收敛），或退回到 **earth**（也能处理非线性的离散度）。它要求 $nfold > 1$ 且 $ncross > 1$ 。

7.7.5 Earth 输出文件

每次成功拟合后，earthUI 会将 `<filename>_earth_output_<timestamp>.txt` 写入输出文件夹。该文件包含模型项、汇总统计量（ R^2 、 $GRSq$ 、 $CVRSq$ ）、方差模型详情和跟踪日志。每次拟合创建一个文件，为比较不同参数配置提供累积记录。

7.8 设置默认值

第 4 节顶部的单选按钮控制加载哪一组默认值：

- **Use last settings for input file**（使用该输入文件的上次设置）——恢复你上次针对这个特定文件使用的设置
- **Use default settings**（使用默认设置）——应用你保存的自定义默认值
- **Earth defaults**（Earth 默认值）——将所有参数重置为出厂值

点击第 4 节底部的 **Save current as default**（将当前设置保存为默认）以存储当前的 **earth** 参数和 Allowed Interactions 矩阵；第 3 节中对应的按钮存储目标变量、评估基准日和预测变量配置。两者都按项目和用途保存在项目数据库中。

8 拟合 Earth 模型

8.1 Fit 按钮

侧边栏第 5 节包含 **Fit Earth Model** (拟合 Earth 模型) 按钮, 以及一个 **Random seed** (随机种子) 字段。该种子使交叉验证折的分配具有可重现性——下拉列表可调出最近使用过的 10 个种子, 拟合完成后所用的种子会记录在 Earth Output (Earth 输出) 选项卡中。在拟合之前, earthUI 会验证您的配置——必须选定一个目标变量, 并且至少包含一个预测变量。在估价/市场模式下, 纬度和经度列会被四舍五入, 销售时距则根据估价基准日和合同日期 (若已指定) 计算得出。

8.2 拟合模态框与跟踪输出

当您点击 Fit 时, 会出现一个深色模态覆盖层, 其中包含:

- 标题栏——显示“Fitting Earth Model”, 附带以秒计数的已用时间计时器, 以及一个可取消后台拟合的 **Abort** (中止) 按钮
- 跟踪日志——一个可滚动的等宽文本区域, 实时显示 `earth()` 函数的输出: 数据集信息、前向遍历进度、交叉验证折以及完成时间
- 颜色编码——默认行为绿色, 交叉验证行为黄色, 错误为红色

earthUI 使用 `callr::r_bg()` 异步拟合模型, 将计算放在单独的 R 进程中运行。这使应用程序在长时间拟合期间保持响应。一个 300 毫秒的轮询观察器从后台进程读取输出并将其流式写入跟踪日志。

拟合完成后, 会出现“Done in X.Xs”消息, 模态框上会添加一个关闭按钮 (X), 并移除背景遮罩, 以便结果选项卡可以按需渲染。待可见输出渲染完毕后, 模态框会自动关闭——您也可以提前关闭它。

拟合成功时, Fit 按钮上会附加一个绿色对勾。如果发生错误, 错误消息会显示在跟踪日志中, 并弹出一条通知。

8.3 拟合日志

每次拟合 (无论成功或失败) 都会在您的输出文件夹中写入一个日志文件:

- 文件名: `<datafile>_earth_log_<YYYYMMDD_HHMMSS>.txt`
- 内容: 时间戳、拟合过程中的全部跟踪输出, 以及所有错误消息
- 位置: 当前活动项目的输出文件夹

提示: 如果未安装 `callr`, `earthUI` 会回退到带简单进度条的同步拟合。安装 `callr` 可获得完整的跟踪输出体验。

9 下载数据

拟合完成后，侧边栏第 6 节——在通用模式下标题为“Download Estimated Target Variable(s) & Residuals”（下载估计的目标变量与残差），在估价和市场模式下标题为“Download Estimated Sale Prices & Residuals”（下载估计的销售价格与残差）——可下载一个包含预测值和诊断信息的 Excel 文件。该输出用于第 7 步（RCA），为估价对象（subject）指定 CQA（状况—质量—吸引力）评分。输出按 `residual_sf` 和 `cqa_sf` 排序，以帮助您评估估价对象在排名中的位置。如果模型质量良好，那么各房产应按照未纳入回归的残差特征，从吸引力最低到吸引力最高进行排序。中间值应约为 0，下半部分为负值，上半部分为正值。您应当在排名底部附近发现质量最差的住宅，即“待修房”（fixers），而最好的住宅位于顶部。通常会有一些异常情形的例外，例如法拍、亏本急售（short sale）、遗嘱认证（与继承相关的）销售，以及因工作变动或其他原因需要的快速出售。对异常案例的调查通常会揭示出价格异常的相关原因。

按钮标签会根据用途模式而变化：

- 通用/市场：**Download Output (Excel)**（下载输出（Excel））
- 估价：**Download Intermediate Output (Excel)**（下载中间输出（Excel））

文件名格式为 `<datafile>_modified_<YYYYMMDD_HHMMSS>.xlsx`。

9.1 输出列

对于每个目标变量，将追加以下各列：

列名	说明
<code>est_<target></code>	模型预测值，例如 <code>est_sale_price</code> （保留 1 位小数）
<code>residual</code>	实际值减去预测值（1 位小数）
<code>cqa</code>	状况—质量—吸引力评分（2 位小数，范围 0–10）
<code>residual_sf</code>	残差除以居住面积（若已指定，1 位小数）
<code>cqa_sf</code>	通过 <code>residual_sf</code> 排名计算得出的 CQA（2 位小数）
<code><variable>_contribution</code>	每个 g 函数的贡献值（1 位小数，每个 g 函数一列）
<code>basis</code>	截距贡献值，对所有房产相同（1 位小数）
<code>calc_residual</code>	实际值减去（ <code>basis</code> + 所有贡献值）——验证列（1 位小数）

对于多目标模型，会添加 `_<i>` 后缀以区分各目标对应的列。

9.2 列排序与 Excel 格式

排名列被放置在输出文件的最左侧位置，顺序为：`residual_sf`、`cqa_sf`、`residual`、`cqa`。这样便于浏览排序后的列表，并评估估价对象在 CQA 分布中的位置。

这些列应用了以下 Excel 数字格式：

列名	格式	示例
<code>residual_sf</code>	数值，2 位小数	12.34
<code>cqa_sf</code>	数值，2 位小数	7.25
<code>residual</code>	数值，0 位小数	15,200
<code>cqa</code>	数值，2 位小数	6.80

9.3 CQA 评分

CQA（状况—质量—吸引力）评分将每一行的残差与所有其他残差进行排名比较。对于给定的行，CQA 是带符号残差小于该行的行所占的百分比乘以 10。由此得到一个 0–10 的评分尺度，其中：

- 高 CQA ($\approx 9-10$) ——该房产的成交价远高于模型预测（较大的正残差）
- 低 CQA ($\approx 0-1$) ——该房产的成交价远低于预测（较大的负残差）
- CQA ≈ 5 ——残差接近所有残差的中位数

当指定了 `living_area` 列时，`cqa_sf` 提供基于每平方英尺残差的同类排名。

9.4 排序与估价对象行

在估价和市场模式下，可比销售案例行按 `residual_sf` 降序排序（若未指定居住面积则按 `residual` 排序）。估价对象所在行（第 1 行）保持在第 1 位，不参与排序。在通用模式下，各行按其原始顺序导出。

这种排序方式与最左侧的排名列相结合，使估价师能够快速浏览从预测偏高最多到预测偏低最多的可比销售案例，并评估估价对象被赋予的 CQA 评分在该分布中的位置。

在调用 `predict()` 之前，预测数据中的因子水平会与训练数据对齐。含有未见过的因子水平的行将产生 NA 预测值。

提示：`calc_residual` 列应始终等于 `residual`（在舍入误差范围内）。如果两者差异显著，则表明存在值得调查的计算问题。

9.5 mgcvUI 自动导出

每次成功拟合且 $\text{degree} \leq 2$ 时，`earthUI` 会自动将完整的结果对象保存为 `.rds` 文件，存入项目的输出文件夹。文件名遵循如下模式：`<datafile>_earthUI_result_<YYYYMMDD_HHMMSS>.rds`。

该文件可由 `mgcv` 后处理模块加载——该模块包含在 `earthUI` 包中，通过 `earthUI::launch_mgcv()` 启动——使用 `readRDS()` 读取。GAM 例程使用 `earth` 模型的节点位置和基函数作为 GAM 平滑项的起点，从而实现从 MARS 到 GAM 建模的无缝过渡。

$\text{degree} > 2$ 的模型会被跳过，因为 `mgcvUI` 仅支持两两交互。当 `earthUI` 以 Trilogy 模式（从 Trilogy UI）启动时，额外的 **Lock Model Output**（锁定模型输出）区块可让您将某次拟合的 `.rds`——连同其预测变量基和共享的 CQA 设置——锁定为供 `glmnetUI` 和 `mgcvUI` 导入的规范 `earth` 模型。

9.6 glmnetUI 导入

为 GAM 例程保存的同一 `.rds` 文件也可导入弹性网络后处理模块——同样包含在 `earthUI` 包中，通过 `earthUI::launch_glmnet()` 启动。在 `glmnet` 应用中，导航至 **Section 2: Import from earthUI**（第 2 节：从 `earthUI` 导入），使用 Browse（浏览）按钮选择该 `.rds` 文件。

`glmnetUI` 采用标准的 `earth::model.matrix()` 方法：`earth` 模型的基函数（铰链、交互项和线性项）成为 `glmnet` 设计矩阵的列。`glmnet` 随后对这些基列执行正则化回归，通过 lasso/弹性网络对其进行选择和收缩。这将 `earth` 的自适应基构造与 `glmnet` 的正则化相结合。

9.6.1 主要优势

- 可解释的交互项：`earth` 基于铰链的交互项（包括三阶项）具有清晰的几何含义，而 `glmnet` 原生的叉积交互项在法庭或审计场合难以解释。
- 自动非线性：`earth` 的铰链函数可以捕捉使用原始预测变量的标准 `glmnet` 模型会遗漏的非线性关系。
- 变量选择：`glmnet` 可以进一步修剪 `earth` 基，剔除没有贡献的铰链项。

9.6.2 重要提示：权重列必须一致

警告

如果 earthUI 使用了权重列（例如，用 `weight=0` 排除估价对象所在行），则必须在 glmnetUI 的变量配置中指定相同的权重列。如果权重设置不同，行数将无法对齐，earth 基也就无法附加到 glmnet 的模型矩阵中。

9.6.3 工作流程

1. 在 earthUI 中拟合一个 earth 模型（任意阶数）。
2. `.rds` 结果文件会自动保存到项目的输出文件夹。
3. 在 glmnetUI 中，打开 **Section 2: Import from earthUI**，浏览至该 `.rds` 文件。
4. 确认两个应用中使用的是同一数据文件和同一权重列。
5. 在第 5 节中配置 glmnet 参数（当 earthUI 导入处于活动状态时，交互矩阵会自动禁用）。
6. 点击 **Fit Model**（拟合模型）。earth 基列将出现在 Equation（方程）、Coefficients（系数）和 Contributions（贡献）选项卡中。

10 RCA 计算与下载

Calculate RCA Adjustments & Download (计算 RCA 调整并下载) 按钮位于侧边栏第 7 节, 仅在估价模式下且模型拟合完成后可见。RCA (Residual Constraint Approach, 残差约束法) 生成相对于估价对象 (subject) 的、源自市场的逐个可比案例调整。

10.1 打开 RCA 对话框

点击该按钮会打开一个小型模态对话框, 其中包含:

- **Score type** (评分类型) —— 单选按钮, 用于选择 **CQA** 或 **CQA per SF** (仅当已指定 `living_area` 列时, SF 选项才可用)。如果选择“CQA per SF”, 则根据你为估价对象指定的 CQA 评分, 其残差评分将是其居住面积乘以与给定 CQA_SF 评分匹配的 `residual_sf`。
- **CQA value** (CQA 值) —— 用于输入估价对象的 CQA 评分 (0.00–10.00)。该评分表示你认为估价对象在可比案例质量分布中所处的位置。
- **Generate** (生成) 按钮——计算 RCA 输出并以 Excel 格式下载。在输入有效评分之前, 该按钮保持禁用状态。

10.2 CQA 评分插值

当你点击 Generate 时, earthUI 会根据可比案例的 CQA/残差对来插值估价对象的残差:

1. 将可比案例的 CQA 评分和残差按 CQA 升序排序
2. 通过线性插值 (`stats::approx()`) 将你输入的 CQA 值映射为一个残差
3. 如果使用 CQA per SF, 则将每平方英尺残差乘以估价对象的居住面积, 换算回总残差
4. 估价对象的估计价值计算为: 模型预测值 + 插值残差

10.3 输出列

RCA Excel 文件 (`<datafile>_adjusted_<YYYYMMDD_HHMMSS>.xlsx`) 包含所有中间输出列, 另加:

列	说明
<code>subject_value</code>	模型预测值 + 插值残差 (仅第 1 行)
<code>subject_cqa</code>	你输入的 CQA 评分 (仅第 1 行)
<code><variable>_adjustment</code>	估价对象贡献减去可比案例贡献 (按 g 函数逐项计算)
<code>residual_adjustment</code>	估价对象残差减去可比案例残差
<code>net_adjustments</code>	所有调整之和 (贡献 + 残差)
<code>gross_adjustments</code>	所有调整绝对值之和
<code>residual_pct</code>	残差调整占可比案例售价的比例
<code>net_adj_pct</code>	净调整占可比案例售价的比例
<code>gross_adj_pct</code>	总调整占可比案例售价的比例
<code>adjusted_sale_price</code>	可比案例售价 + 净调整

各调整列告诉你, 对于每个可比案例, 模型将差异归因于各变量的金额。`adjusted_sale_price` 是应用所有模型推导调整后的可比案例售价——这组调整后售价应聚集在估价对象的估计价值附近。

10.4 RCA Adjustments 选项卡

RCA 输出计算完成后，主面板中的 **RCA Adjustments** (RCA 调整) 选项卡会显示各可比案例的 Residual Adj %、Net Adj % 和 Gross Adj % 分布的直方图 (多目标模型则每个目标一组)，从而可以快速直观地检查调整的聚集程度。

提示：CQA 评分 5.00 表示估价对象处于可比案例的中位数位置。评分高于 5 表示估价对象质量高于平均水平；低于 5 则表示低于平均水平。

11 销售比较网格

11.1 概述与用途

销售比较网格 (Sales Comparison Grid) 是在估价模式下 (侧边栏第 8 节)、计算 RCA 调整之后生成的 Excel 工作簿。它以结构化的网格形式将估价对象与选定的可比案例并排呈现, 并附带自动计算调整和每个可比案例调整后售价的 Excel 公式。

该网格专为估价师的工作底稿而设计——它将 Earth 模型推导出的回归调整与可编辑单元格相结合, 估价师可在其中将 CQA 残差分配到具体的房产特征上。输出文件名为 SalesGrid_<YYYYMMDD_HHMMSS>.xlsx。

11.2 可比案例选择模态框

点击“Generate Sales Grid & Download”(生成销售网格并下载) 会打开一个模态对话框, 供你选择要包含的可比案例:

- 推荐可比案例已预先勾选。这些可比案例的总调整百分比低于售价的 25%, 按销售时间由近及远排序 (最近的销售在前), 上限为 30 个。
- 其他可比案例列在下方, 默认不勾选。它们的总调整达到或超过 25%, 但仍可供选择 (最多显示 50 个, 按总调整百分比升序排序)。
- 最多可选择 30 个可比案例, 最多生成 10 张工作表 (每张工作表 3 个可比案例)。

确认选择后, 所选可比案例按总调整百分比升序排序, 因此最相似的可比案例会出现在 Sheet 1 上。

11.3 网格布局

每张工作表采用 20 列布局, 容纳估价对象加 3 个可比案例。各实体的列为:

实体	列 (每个实体 5 列)
估价对象	标签、事实值 1、事实值 2、事实值 3、价值贡献 (VC)
每个可比案例	事实值 1、事实值 2、事实值 3、价值贡献 (VC)、调整

各行从上到下依次为:

1. 标题——工作表名称 (例如“Intermediate Sales Comparable Grid — Sheet 1 of 3”)
2. 表头——“Subject”与各可比案例的列标题及地址
3. 地址——完整的房产地址
4. APN | MLS# | DOM | Subj.Prox (宗地号 | 挂牌编号 | 在售天数 | 与估价对象的距离)——宗地号、挂牌编号、在售天数, 以及与估价对象的 Haversine 距离 (英里)
5. Sales Price | Concess. | Net SP (售价 | 让价优惠 | 净售价)——售价、让价优惠, 以及 Net Sale Price (净售价) 公式
6. Regression Features (回归特征) 表头行
7. BASE VALUE (基准值)——模型截距
8. Date of Sale | OffMkt | OnMkt (成交日期 | 下市 | 上市)——合同日期、销售距今时间和 DOM
9. 分组行 (视情况而定)——Location (位置)、Site Size (地块面积) 和/或 Age (房龄) 组
10. 模型变量行——每个预测变量一行 (不含已分组的变量)
11. 空白分隔行
12. Residual Features (残差特征) 表头行
13. CQA|Residual (CQA | 残差)——CQA 评分 + 剩余残差公式
14. 残差特征行——命名行加空白行, 供估价师填写

15. **Total VC / Net Adjustment** (总 VC / 净调整)
16. **Net Adjustment %** (净调整百分比)
17. **Gross Adjustment %** (总调整百分比)
18. **Adjusted Sale Price** (调整后售价) ——公式行
19. 版权页脚

11.4 售价、让价优惠与 Net SP

第 5 行为每个可比案例显示三个值:

- **Sale Price**——来自数据的可比案例售价
- **Concessions**——指定为 **concessions** 的列中的值 (未指定时为 0)
- **Net SP**——一个 Excel 公式: $\text{Sale Price} - \text{Concessions}$

估价对象列的售价显示为“N/A”(因为未知),若有让价优惠值则一并显示。

11.5 分组行 (位置、地块、房龄)

当指定了某些特殊类型且这些变量出现在模型中时, earthUI 会创建将相关变量合并的分组行:

Loc: Long | Lat | Area (位置: 经度 | 纬度 | 区域) ——当 **longitude**、**latitude** 或 **area** 中的任一变量出现在模型中时出现。显示每个组成变量的事实值、合并的价值贡献 (各 VC 之和), 以及针对可比案例的调整 (估价对象合并 VC - 可比案例合并 VC)。VC 单元格以浅蓝色背景标示。

Site Size | Dimensions (地块面积 | 尺寸) ——当 **lot_size** 或 **site_dimensions** 出现在模型中时出现。结构与 **Location** 组相同。

Actual Age | Effective Age (实际房龄 | 有效房龄) ——当 **actual_age** 或 **effective_age** 出现在模型中时出现。结构相同。

被分组行占用的变量将从下方的单个模型变量行中排除, 以防止调整合计中的重复计算。

11.6 模型变量行

在分组行下方 (若无分组行, 则直接在 **BASE VALUE** 下方), 每个剩余的模型预测变量占一行, 显示:

- **事实值**——估价对象和每个可比案例的实际数据值
- **价值贡献 (VC)** ——该行 **g** 函数对预测值的贡献
- **调整 (仅可比案例)** ——估价对象 VC 减去可比案例 VC

11.7 CQA|Residual 与残差特征行

CQA|Residual 行显示每个房产的 CQA 评分, 并包含剩余残差公式——即尚未分配到具体特征的残差部分。公式为:

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

其下是残差特征行——5 个命名行 (**Location**、**View/Appeal**、**Condition**、**Quality**、**Other**, 即位置、景观/吸引力、状况、品质、其他) 加 6 个空白行。这些是输入单元格, 供估价师为模型未捕捉到的特征填写价值贡献。随着估价师填入数值, **Remaining Residual** 公式的值会自动减少。

对于每个可比案例, 调整列包含一个公式: 估价对象特征 VC - 可比案例特征 VC。

11.8 调整后售价公式

Adjusted Sale Price 行包含 Excel 公式:

- 估价对象：从 BASE VALUE 到最后一个残差特征行的所有价值贡献单元格之和。这代表模型对估价对象的总预测值，加上估价师分配的残差特征。
- 可比案例：Net SP + 从第一个分组行或变量行到最后一个残差特征行的所有调整单元格之和。这给出应用所有回归推导调整和估价师填写调整后的可比案例售价。

11.9 工作表保护与可编辑单元格

每张工作表均已受保护，以防止公式和数据被意外修改。保护锁定了：

- 所有公式单元格（Net SP、剩余残差、调整、调整后售价等）
- 所有数据单元格（地址、售价、事实值、来自模型的价值贡献）
- 标签和表头

唯一未锁定（可编辑）的单元格是残差特征价值贡献输入格——即 CQA|Residual 行下方供估价师填写分解值的单元格。这些单元格以浅黄色背景标示，表示可编辑。

11.10 在 Excel 中使用网格

1. 打开文件——使用 Excel 或兼容的电子表格应用程序。
2. 审核回归调整——模型推导的调整已预先填入并锁定。
3. 分配残差——在残差特征行（黄色单元格）中，填写你对剩余残差中归因于各特征（位置、景观、状况、品质等）的金额的判断。随着分配的进行，Remaining Residual 公式的值会相应减少。
4. 检查 **Adjusted Sale Price**——该公式会随你填入残差特征值而自动更新。
5. 多张工作表——如果选择了 3 个以上的可比案例，请在工作表之间切换。每张工作表的左列均为同一估价对象，并配以 3 个不同的可比案例。

提示：目标是将 CQA 残差分配到各特征行，直至 *Remaining Residual* 接近于零。这表明你能够解释每个可比案例的回归估计值与其实际售价之间的差异。

12 下载报告

模型拟合完成后，报告生成是一个两步式的 Quarto 工作流：**Generate Quarto Report**（生成 Quarto 报告，估价模式下为侧边栏第 9 节，其他模式下为第 7 节）写出一个自包含的报告包，而 **Convert Quarto Report**（转换 Quarto 报告，第 10 节）将其——或任何其他 Quarto 文件——渲染为最终格式。

12.1 生成 Quarto 报告

点击 **Generate Quarto Report** 会在项目输出文件夹下的 `<base>_qmd/` 中写出一个自包含的 Quarto 包，其中包含已填充内容的 `.qmd` 源文件、所有预生成的图表资源（PNG 和 PDF）、报告数据，以及一个 Word 参考模板。此步骤不进行任何渲染——该报告包可以手工编辑、通过 Quarto includes 与其他 Quarto 源文件合并，或使用 `quarto preview` 单独预览。

12.2 转换 Quarto 报告

第 10 节提供一个文件选择器（默认指向当前项目最近生成的报告包——但你可以浏览到任意 `.qmd` 文件）以及格式复选框：

- **HTML**——一个 `.html` 文件，带有 KaTeX 数学公式渲染、内嵌图片和目录。使用 Flatly Bootstrap 主题和 Roboto Condensed 字体。
- **Word**——一个适合编辑和分发的 `.docx` 文件。包含目录、页码，并使用自定义参考模板以保持样式一致。
- **PDF**——使用 LuaLaTeX 排版，输出达到专业质量。纸张尺寸遵循区域设置（Letter 或 A4）。包含用于大型交互矩阵的横向页面，并使用 Roboto Condensed 搭配 Latin Modern Math 排版公式。仅当检测到 LaTeX 发行版时才会提供 PDF 选项。

点击 **Convert**（转换）以渲染所选格式。渲染在后台运行——一个模态对话框显示已用时间和 Quarto 进度，应用保持可响应状态。完成后会有通知确认文件位置。

所有操作（模型拟合、数据下载、RCA 计算、销售网格生成和报告渲染）都会连同起止时间戳和耗时记录到输出文件夹中的 `<datafile>_earthui_log.txt`，以便排查问题和监控性能。

12.3 报告内容

每份生成的报告都包含以下各节：

1. 数据集描述——观测数量、目标变量、所用预测变量、分类预测变量
2. 模型规格——阶数（degree）、交叉验证状态、项数、模型中的预测变量
3. 允许的交互项（仅当 $\text{degree} \geq 2$ 时）——带勾选标记的交互矩阵，按输出类型格式化
4. 结果摘要—— R^2 、GRSq、GCV、RSS，以及 $\text{CV } R^2$ （若启用了交叉验证）
5. 模型方程——以 LaTeX 记号表示的完整 Earth 模型方程
6. 系数与基函数——所有项、其系数以及定义它们的基函数的表格
7. 变量重要性——预测变量重要性得分的条形图和排名表
8. g 函数贡献——每个 g 函数的图：单变量项为折线图，双变量交互项为 3D 透视图和等高线图
9. 相关矩阵——预测变量相关性的热图
10. 诊断——残差对拟合值（Residuals vs Fitted）、正态 Q-Q 图，以及实际值对预测值（Actual vs Predicted）图
11. ANOVA 分解——显示每个基函数对 RSS 贡献的表格
12. Earth 输出——来自 `earth::summary()` 的原始文本输出，包括剪枝过程的细节

对于多响应模型，第 4–8 节和第 10 节会针对每个目标变量重复一次。

提示：生成报告需要 `quarto R` 包和 `Quarto` 安装。若要输出 *PDF*，还必须安装 *LaTeX* 发行版（例如通过 `tinytex::install_tinytex()`）——否则 *PDF* 复选框会被隐藏。

13 后处理模块：弹性网络与 GAM

earthUI 包附带两个后处理模块——一个基于 `glmnet` 的弹性网络 (elastic net) 例程和一个基于 `mgcv` 的广义可加模型 (GAM) 例程。本期各有一篇文章对它们分别做了深入论述；本节说明它们在 earthUI 工作流中的用途，以及三个例程如何协作。

13.1 为什么需要辅助拟合？

earth 模型给出价值结论。后处理模块的存在是为了印证它：由方法学上不同的引擎、针对同一估价对象 (subject)、基于相同项目数据产生的两个额外估计。当对 earth 基函数进行弹性网络重估计与独立的 GAM 光滑拟合得出实质上相同的价值时，估价师的结论便不再仅仅依赖于单一模型的假设——这在复核和诉讼场景中是一项实实在在的优势。当它们不一致时，这种分歧本身就是诊断信息，通常指向数据稀薄、某个过度灵活的项，或某个值得重新审视的离群可比销售案例。两个模块都不凌驾于 earth 模型之上；它们是佐证证据，并有意不做加权呈现。

13.2 启动与共享状态

每个例程都是一个独立的本地 Shiny 应用：

- `earthUI::launch_glmnet()`——弹性网络应用，端口 7879。
- `earthUI::launch_mgcv()`——GAM 应用，端口 7880。

三个应用读取同一项目树和同一设置数据库，因此你在 earthUI 中打开的项目在模块中立即可用——连同 earthUI 为其保存的沿用字段：Response (响应/目标变量)、Effective Date (估价基准日)、RCA 估价对象的 CQA 类型和取值，以及居住面积 (living-area) 指定。这些无需重新输入；模块会读取它们，使同样的取值贯穿全部三次拟合。

13.3 弹性网络例程

`glmnet` 例程将 earth 模型的基函数——其铰链函数、交互项和线性项——作为设计矩阵的列导入，并在套索回归 (lasso) /弹性网络正则化下对其重新估计（其机制及权重列的注意事项参见上文“`glmnetUI Import`”）。其结果保留了 earth 可解释、具有几何意义的项，同时让正则化器收缩或删除那些不能证明自身价值的项。完整论述参见本期的 `glmnetUI` 文章。

13.4 GAM 例程

`mgcv` 例程走的是另一个方向：它不复用 earth 的基函数，而是重新拟合带惩罚的光滑项，以 earth 模型的节点位置作为起点（参见上文“`mgcvUI Auto-Export`”）。由于 GAM 在不同的光滑范式下重新推导函数形式，它与 earth 模型的一致性真正独立的印证，而非回声。支持 $\text{degree} \leq 2$ 的 earth 模型。完整论述参见本期的 `mgcvUI` 文章。

13.5 一个实用的工作流

1. 先在 earthUI 中拟合并打磨 earth 模型，直到你对其满意为止。每次成功的拟合 ($\text{degree} \leq 2$) 都会自动导出供模块使用的 `.rds` 文件。
2. 针对同一项目打开 `glmnet` 和 `mgcv` 例程，导入 earth 结果并进行拟合。一次只在一个例程中工作——拟合耗时较长，且各例程通过共享项目进行协调，因此合理的节奏是顺序进行，而非并行。
3. 自由迭代。在你于各应用之间切换时，每个应用都会保留其结果，没有任何环节强制你在过程中生成报告。
4. 只有当三个结果都令你满意时，再回到各个例程生成其 Quarto 报告。每种方法都会将其拟合和价值结论登记到项目的三部曲 (trilogy) 记录中，使三份报告可追溯到同一次比较运行。

提示：这些模块在任何层面上都是可选的：*earthUI* 自身的工作流——拟合、*RCA*、销售网格、报告——没有它们也是完整的。当某个结论需要印证时再使用它们：棘手的估价对象、数据稀薄的市场细分，或很可能面临复核的委托任务。

14 演示数据集：Appraisal_1.csv

earthUI 包含一个演示 MLS 数据集，用于探索估价 workflow。可通过编程方式加载：

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

或者，在 Shiny 应用中，将其复制到项目的 in/ 文件夹（或在创建项目时将其作为初始数据文件附加），然后在第 2 节（Import Data，导入数据）中选择它。

14.1 描述

该文件包含来自一份模拟 MLS 导出的 1,501 条住宅销售记录，外加第 1 行的 1 个估价对象 (subject)（共 1,502 行）。数据代表一个多区域市场中的独栋住宅销售，涵盖多种房屋面积、房龄和位置。

这不是真实数据，但基于北加州一个真实感很强的社区。所有识别信息均已被更改或删除。

14.2 数据列

列名	类型	特殊类型	描述
weight	numeric	—	观测权重 (0 = 不参与拟合)
id	numeric	display_only	内部记录 ID
property_id	numeric	display_only	MLS 房产标识符
listing_id	character	display_only	MLS 挂牌编号
parcel_number	character	display_only	县税务评估员宗地号 (APN)
street_address	character	display_only	房产地址
city_name	character	display_only	城市
postal_code	character	display_only	邮政编码 (ZIP)
county_name	character	display_only	县
contract_date	Date	contract_date	销售合同日期 (用于计算 sale_age)
sale_age	numeric	—	从合同日期到估价基准日的天数 (预先计算)
coe_date	Date	display_only	过户完成 (close of escrow) 日期
listing_status	character	display_only	挂牌状态 (例如 "Sold")
sale_price	numeric	(target)	销售价格——响应变量
rent	numeric	—	月租金 (用于多目标模型)
list_price	numeric	display_only	挂牌价格
original_list_price	numeric	display_only	原始挂牌价格
living_sqft	numeric	living_area	总居住面积 (平方英尺)
beds_total	integer	—	卧室数量
baths_total	numeric	—	浴室总数 (例如 2.5 = 2 个全卫 + 1 个半卫)
lot_size	numeric	lot_size	地块面积 (平方英尺)
area_id	integer	area	MLS 区域标识符
area_text	character	display_only	区域名称
age	numeric	actual_age	房龄 (年)
year_built	integer	display_only	建成年份
latitude	numeric	latitude	纬度 (建模时四舍五入到小数点后 3 位)
longitude	numeric	longitude	经度 (建模时四舍五入到小数点后 3 位)
latitude6	numeric	display_only	全精度纬度
longitude6	numeric	display_only	全精度经度
garage_spaces	integer	—	车库车位数
fp_count	integer	—	壁炉数量
no_of_stories	numeric	—	楼层数
style	character	—	建筑风格
view	character	—	景观类型 (例如 "Neighborhood"、"Hills")
days_on_market	integer	dom	在市天数
listing_date	Date	listing_date	挂牌日期
sale_concessions	numeric	concessions	卖方让步 (concessions) 金额

14.3 建议的快速上手步骤

1. 启动 earthUI: `earthUI::launch()`
2. 在 New Project (新建项目) 向导中创建一个 Purpose (用途) 为 **For Appraisal** (用于估价) 的项目, 并附加 `Appraisal_1.csv` 作为初始数据文件
3. 在第 2 节 (Import Data) 中选择 `Appraisal_1.csv`
4. 选择 `sale_price` 作为目标变量
5. 按上表所示指定特殊类型
6. 纳入预测变量: `sale_age`、`living_sqft`、`baths_total`、`lot_size`、`area_id` (作为因子)、`age`、`latitude`、`longitude`、`garage_spaces`
7. 将 `degree` 设为 1, 点击 **Fit Earth Model** (拟合 Earth 模型)
8. 下载中间输出 (第 6 步), 审阅 CQA 排名
9. 计算 RCA 调整 (第 7 步), CQA 得分约为 ~5.00
10. 生成销售比较网格 (第 8 步)

提示: 在导入你自己的 *MLS* 数据之前, 演示数据集是探索 *earthUI* 全部功能的好方法。上面建议的预测变量通常能得到 R^2 高于 0.85 的模型。

15 实验性功能：Prolog 文本提取与推导规则

0.10.0 版本引入了一项可选且确属实验性的功能，用于从 MLS 导出数据的描述性文本列中提取可供模型使用的特征——首要的是“Public Remarks”（公开备注）列，此外还包括经纪人备注以及以逗号分隔的特征列表，例如房间特征或既有建筑物。提取引擎是 Prolog，具体而言是确定子句文法（**Definite Clause Grammars, DCG**）：在备注（remarks）文本中识别诸如“move in ready”“3 car garage”或“owned solar”等短语的规则，均以 Prolog 语法规则的形式编写。

该功能仍在开发之中。目前它可以部分用于提取——随附的文法已能识别一组实用的特征——但要将其适配到您所在 MLS 的用语习惯，您需要能够阅读并编写 Prolog，尤其是 DCG。如果这不适合您，可以完全跳过本章：该功能默认关闭，关闭时 earthUI 的其余部分工作方式完全不变。

Warning

这是一项处于积极开发中的实验性功能。其提取文法既不完整，也未经过面向实际估价工作的验证——在将提取出的列用于模型之前，必须将其与源文本核对。要有效使用该功能，需要具备 Prolog 和 DCG 的实际工作知识。它所做的一切都是附加且可逆的；如果出现任何异常，只需关闭 Settings 中的开关并照常继续即可。

15.1 它的作用

MLS 房源在自由文本和分隔列表字段中承载着大量与价值相关的信息。这一可选步骤通过两种特殊的列角色（见第 6 章），将这些文本转化为回归可以使用的普通列：

- **remarks** 列由 Prolog DCG 文法解析，每个提取出的属性生成一个特征列。列前缀由源列名称各单词的首字母构成：public_remarks 生成 pr_* 列，agent_remark 生成 ar_* 列，依此类推。
- **list** 列（例如 existing_structures = "Shed, Workshop"）被拆分为独热编码的 <col>_<value> TRUE/FALSE 列。这一部分是纯 R 实现，完全不需要 Prolog。

随附的文法（inst/prolog/mls_remarks.pl）目前可提取的内容包括：带有 1-5 级 condition_score 的分级状况等级；景观标志与景观类型；厨房与卫浴翻新标志；硬木及其他地板材质；升级项目列表及数量；车库车位数与车库面积；泳池、水疗池及其他室外设施；壁炉数量；附属建筑（棚屋、谷仓、工作间、casita、客房、犬舍等）及从“40x50”这类尺寸中解析出的面积；房车停放与储物空间；太阳能自有与租赁之分、电池储能；地块面积（含英亩到平方英尺的换算）；与主体相关联的交易背景价格金额（降价、HOA 会费、太阳能租赁或买断、Mello-Roos）；区位线索；以及取值范围为 [-1, 1] 的亮点 / 顾虑情感评分。

15.2 环境要求：vProlog (Trealla)，而非 SWI-Prolog

备注解析运行于 vProlog R 包之上——一个自包含的 Trealla Prolog 引擎。它不在 CRAN 上；请按照本文开头“Installing earthUI and vProlog”（安装 earthUI 与 vProlog）一节所示，从作者的 r-universe 仓库安装（提供 Windows 和 macOS 的预编译二进制包，源代码位于 github.com/wcraytor/vProlog）。无需也不会使用任何系统级 Prolog 安装（无论是 SWI-Prolog 还是其他实现）。如果缺少 vProlog，备注解析会被跳过并给出提示，而 list 列拆分仍照常运行。辅助函数 vProlog_available() 可报告引擎与随附文法是否均已就绪。

15.3 启用与运行

1. 打开 Settings（设置，齿轮图标），勾选 **Enable Prolog / list column processing**（启用 Prolog / 列表列处理）。该开关默认关闭，附有“高级功能——面向具备 Prolog 专业知识用户”的说明，按项目记忆设置，并显示在 Project Information（项目信息）对话框中。

2. 在预测变量表中，将相关列的 **Special**（特殊）角色设为 `remarks` 或 `list`（每种角色都可以指定给多个列）。
3. 使用位于侧边栏第 4 节与第 5 节之间的 **Execute Prolog Processing**（执行 Prolog 处理）区域。它提供两个步骤：

1. **Expand lists & remarks**（展开列表与备注）——对每个 `remarks` 列运行 DCG 文法，并对每个 `list` 列进行独热拆分。展开是增量式的：只处理自上次运行以来新指定的列，一旦所有已指定的列都已展开，按钮即被禁用。解析过程由进度条跟踪。

2. **Execute derivation rules**（执行推导规则）——在展开后的列上应用您自己的 Prolog `derive/2` 规则（见下文）。规则在运行前会经过语法检查和试运行验证。

15.4 推导规则

对于每一数据行，`earthUI` 将该行的各列断言为 `cell(Name, Value)` 事实，并收集您的 `derive(Col, Val)` 规则的所有解——每个不同的 `Col` 都成为一个新列，且每列以找到的第一个值为准，因此可以在后面附加一条兜底子句。辅助谓词 (`count_true/2`、`count_false/2`、`count_true_prefix/2`、`cell_or/3`、`is_na/1`) 支持聚合与 NA 处理，而 `drop/1` 事实会在规则运行后丢弃相应列。例如：

```
% 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).
```

规则存放在每个项目独立的 `<project>_rules.pl` 文件中，该文件由一份带完整注释的模板初始化，并通过 **Edit derivation rules...**（编辑推导规则……）在一个可浮动、可拖动、也可停靠到侧边栏的代码编辑器中编辑。**Check**（检查）按钮可在不执行规则的情况下对其进行验证（语法扫描、加载检查，以及针对数据第 1 行的试运行）。

15.5 带版本号的的活动数据文件

任一步骤完成后，增补后的数据框都会以 `<original_name>_<YYMMDD_HHMMSS>.xlsx` 的形式写入项目的输入文件夹，并成为该项目的活动数据文件——重新打开项目时加载的即是增补后的数据，因此展开结果永远不会被重复计算，而 `glmnetUI` 和 `mgcvUI` 使用的也正是同一个文件。原始的备注源列会被自动排除在模型之外；生成的列则不会被自动纳入预测变量。文本解析可能产生大量列，因此请将您（暂时）不希望进入回归的列标记为 **Latent**（潜在），或用 `×` 将其禁用（见第 6 章）。

15.6 编程接口与随附文件

整个流水线均已导出，可供批处理使用：`execute_prolog_processing(df, remarks_cols, list_cols, rules_text)` 负责编排各个处理阶段，其底层基于 `remarks_to_columns()`、`split_list_column()`、`apply_user_rules()` 和 `validate_rules()`。位于 `inst/prolog/`

下随附的 Prolog 源文件包括: `mls_remarks.pl` (DCG 文法与特征聚合)、`mls_analyze.pl` (用于批量解析的文件 IO 驱动)、`mls_rules.pl` (可供推导规则使用的辅助谓词), 以及 `climate_regions_ca.pl` (见下文)。

15.7 相关实验: 气候区先验

一项配套的实验性功能针对数据稀薄的特征——即那些因仅出现在极少数可比销售案例中而无法由回归估计的特征 (例如蒸发式冷风机)。三个 R 辅助函数——`climate_region_for()`、`market_area_profile()` 和 `climate_feature_priors()`——将项目所在位置 (县+城市) 归入七个购房者认知层面的加利福尼亚气候区之一 (这是对加州能源委员会 16 个建筑气候区的实用性归并), 并按区域返回 HVAC 与制冷特征 (供暖、空调、蒸发式冷风机、全屋换气扇、空气过滤) 的序数化贡献价值先验。Prolog 模块 `climate_regions_ca.pl` 镜像了同一套分类, 可在推导规则内部使用; 两侧均由同一个规范的 R 源生成, 因此不会产生偏差。这些先验旨在为特征估值提供参考, 并作为对估计系数的合理性核查——区域归属 (尤其是横跨多个区域的县) 应由估价师复核。

15.7.1 超越加利福尼亚: 辖区包与市场区域模块

加利福尼亚是第一个辖区, 而非边界。earthUI 在国际范围内被使用, 气候先验知识正被重构为辖区包 (**jurisdiction packs**): 自包含、数据化的事实表 (分区定义、行政单元归属、特征先验), 以项目系统的国家代码和行政层级代码为键, 因此一个包可以描述一个美国的州、一个德国的联邦州 (Land), 乃至直接描述一个小国。与之配套的市场区域模块 (**market-area modules**) 层则捕捉某一具名市场区域的本地用语与更精细的先验。

二者的设计目标都是由 Valuation Engineer 和估价师在一线现场、用他们自己的语言编写——并且, 从现实出发, 可借助 LLM 完成: 世界上并没有多少 Prolog 程序员, 因此现场产出的成果被有意设计为表格而非程序。包格式是一个有文档的模式 (schema), 并附有完整的成品示例 (即加利福尼亚本身), 其明确目的就是让 Claude 这样的助手能够根据对某一市场的自然语言描述起草出一个有效的包; earthUI 会从包生成底层的 Prolog, 并且一个具有阻断性的验证器 (模式、覆盖度、引用完整性和试运行检查) 必须通过后, 包才能被启用。每个现场编写的包都必须携带出处信息——署名作者、日期以及所引用的权威依据——因此为这些知识背书的是负有责任的专业人员而非起草工具, 报告也能准确引用所使用的包及其版本。

提示: 一个实用的模式: 先展开备注, 再用 *derive/2* 规则将细粒度的 *pr_** 标志整合为少数几个派生的计数或布尔列, 用 *drop/1* 丢弃已整合的原始标志, 只保留派生列作为候选预测变量。这样可以保持预测变量列表精简, 并使共线性可控。

参考文献

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