

earthUI: интерактивный интерфейс для пакета earth (MARS)

Русский перевод, подготовленный с использованием искусственного интеллекта и прошедший редакционную проверку; английская версия является основной.

William Bert Craytor

Том 1, Выпуск 1 · Лето 2026

Опубликовано: 1 июля 2026

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

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

Эта статья открытого доступа распространяется на условиях [лицензии Creative Commons Attribution 4.0](#)

Аннотация. earthUI — графический интерфейс пользователя для пакета R earth, реализующего многомерные адаптивные регрессионные сплайны (MARS). Он предлагает три режима работы — общее прогнозное моделирование, оценку недвижимости и анализ рыночной зоны — и проводит пользователя через импорт данных, настройку модели, подгонку, графики g-функций и загружаемые отчёты. В статье описаны требования к формату данных, рабочий процесс моделирования, форматы вывода и полный справочник возможностей earthUI.

Ключевые слова: MARS, многомерные адаптивные регрессионные сплайны, earth, R, графический интерфейс, оценка недвижимости

Publisher's Note (July 2026): This article describes earthUI 0.10.0. Shortly before press, earthUI 0.11.0 absorbed the companion applications glmnetUI and mgcvUI (described in the following two articles), consolidating the MARS, elastic-net, and GAM workflows in a single application. The interface and workflow described here are unchanged by the consolidation.

Содержание

1	Введение	4
1.1	Что такое earthUI?	4
1.2	Один набор, три процедуры	4
1.3	Три режима работы	4
1.4	Недвижимость — специальные материалы	5
1.5	Установка earthUI и vProlog	5
1.6	Начало работы	6
2	Требования к данным для ввода MLS	7
2.1	Формат файла & Структура	7
2.2	Обязательные столбцы для оценки Режим	7
2.3	Особые правила именования столбец	9
2.4	Качество данных& Полнота данных	9
2.5	Размещение строк объекта оценки	10
3	Режим общего назначения	11
3.1	Обзор	11
3.2	Рабочий процесс с боковой панелью	11
3.3	Вкладки главной панели	12
3.4	Сохранение настроек	12
3.5	Тёмный режим	12
3.6	Язык и регион & Региональные настройки	13

3.6.1	Поддерживаемые страны	13
3.6.2	Paper Override	13
3.6.3	Сохранение настроек по умолчанию	14
4	Режим оценки	15
4.1	Обработка строк объекта оценки	15
4.2	Дата оценки & Возраст, с которого разрешена продажа	15
4.3	Специальный столбец Обозначения	15
4.4	Остаточный вклад (оценка редких признаков)	16
4.5	Обзор корректировок RCA	17
5	Режим анализа рыночных зон	18
5.1	Отличия от режима оценки Режим «	18
5.2	Когда следует использовать режим «Рынок»	18
6	Отбор переменных	19
6.1	Целевые переменные	19
6.2	Таблица предикторов	19
6.3	Определение типов данных& Переопределения	20
6.4	Флаги факторов и линейных моделей	20
6.5	Специальные типы столбцов См.	20
6.6	Несколько целей	21
7	Выбор параметров	22
7.1	Основные параметры	22
7.2	«Forward Pass»	23
7.3	Допустимые взаимодействия	24
7.4	Обрезка	24
7.5	Перекрёстная проверка	24
7.6	Фильтрация подмножеств	24
7.7	Рекомендуемые значения параметров	25
7.7.1	Параметры передачи вперед	25
7.7.2	Параметры обрезки	26
7.7.3	Перекрёстная проверка Параметры	27
7.7.4	Модель дисперсии	27
7.7.5	Файл вывода Earth	28
7.8	Настройки по умолчанию	28
8	Подгонка модели Земли	29
8.1	Кнопка «Fit»	29
8.2	Подгонка модальной кривой & Trase Вывод	29
8.3	Журнал тренировок	29
9	Загрузка данных	30
9.1	Столбцы вывода	30
9.2	Упорядочивание столбцов & Excel Формат	31
9.3	Результаты CQA: балл	31
9.4	Сортировка & Строка объекта оценки	31
9.5	mgcvUI Auto-Export	31
9.6	glmnetUI: импорт	32
9.6.1	Основные преимущества	32
9.6.2	Важно: столбец «Вес» должен соответствовать	32

9.6.3	Рабочий процесс	32
10	Расчеты RCA& Скачивание	34
10.1	Открытие диалогового окна RCA	34
10.2	Интерполяция баллов CQA	34
10.3	Столбцы вывода	34
10.4	Вкладка «Корректировки RCA»	35
11	Таблица сравнения продаж	36
11.1	Обзор & Цель	36
11.2	Модальное окно выбора аналогов	36
11.3	Макет таблицы	36
11.4	Цена продажи, уступки & Net Цена продажи	37
11.5	Сгруппированные строки (Местоположение, Объект, Возраст)	37
11.6	Строки с переменными модели	37
11.7	CQA Остаточная величина & Остаточная характеристика Строки	38
11.8	Скорректированная цена продажи Формула	38
11.9	Защита листа & Редактируемые ячейки	38
11.10	Работа с таблицей в Excel	38
12	Скачивание отчетов	40
12.1	Создать отчет Quarto	40
12.2	Конвертировать отчет Quarto	40
12.3	Содержание отчёта	40
13	Модули постобработки: эластичная сеть и GAM	42
13.1	Почему стоит поддерживать Fits?	42
13.2	Запуск и общее состояние	42
13.3	Процедура «эластичная сеть»	42
13.4	Процедура GAM	42
13.5	Практический алгоритм действий	43
14	Демо-набор данных: Оценка _ 1.csv	44
14.1	Описание	44
14.2	Столбцы	45
14.3	Рекомендуемое руководство по быстрому запуску	46
15	Экспериментальная часть: Извлечение текста из Prolog & Вывод Правила	47
15.1	Описание функции	47
15.2	Требования: vProlog (Trealla), не SWI-Prolog	48
15.3	Включение и запуск	48
15.4	Правила вывода	48
15.5	Файл активных данных с версиями	49
15.6	Программный интерфейс & Набор файлов	49
15.7	Связанный эксперимент: «Климатический регион» Исходные распределения:	49
15.7.1	За пределами Калифорнии: пакеты юрисдикций и модули рыночных зон	50

1 Введение

1.1 Что такое earthUI?

earthUI — это графический интерфейс пользователя для пакета R «earth», реализующего метод многомерных адаптивных регрессионных сплайнов (MARS) (Friedman 1991). В качестве ядра моделирования используется пакет earthСтивена Милбороу (Stephen Milborrow) (Milborrow 2024), подробно описанный на сайтах Milborrow ([6.r.\[a\]](#)) и Milborrow ([6.r.\[b\]](#)). Он работает как локальное приложение Shiny — без входа в систему, без сервера и без учетных записей. Вы запускаете его из R, импортируете набор данных (CSV или Excel), настраиваете модель и выполняете её подгонку в интерактивном режиме.

Приложение обеспечивает полный рабочий процесс: импорт данных, настройку переменных, подгонку модели, графики диагностики, оценку значимости переменных, уравнения модели, а также возможность скачивания отчетов в форматах HTML, Word или PDF.

1.2 Один набор, три процедуры

Начиная с текущей версии, пакет earthUI также содержит два модуля постобработки, которые в предыдущих версиях представляли собой отдельные пакеты glmnetUI и mgcvUI. При установке earthUI устанавливаются все три процедуры; каждая из них запускается как отдельное локальное приложение Shiny:

- `earthUI::launch()` — приложение earth (MARS), описанное в данной статье, на порту 7878. `earth()` — это центральный механизм оценки.
- `earthUI::launch_glmnet()` — модуль постобработки эластичной сети (порт 7879), который импортирует базисные функции модели earth и переоценивает их с использованием регуляризации Lasso/эластичной сети (см. раздел «Импорт glmnetUI» ниже).
- `earthUI::launch_mgcv()` — модуль постобработки GAM (порт 7880), который использует положения узлов модели Earth в качестве начальных точек для процесса сглаживания слагаемых mgcv (см. раздел «mgcvUI Auto-Export» ниже).

Эти три процедуры используют одно дерево проекта, одну базу данных настроек и одну копию вспомогательной инфраструктуры, поэтому проект, открытый в одном приложении, сразу становится доступным для остальных. Процедуры glmnet и mgcv служат подтверждением выводов о значении модели Земли, а не конкурирующими рабочими процессами: предполагаемый алгоритм заключается в том, чтобы сначала провести подгонку и уточнение модели Земли, а затем запустить модули на том же проекте, когда требуется подтверждение результатов.

1.3 Три режима работы

earthUI организует работу в рамках проектов — именованных объектов, которые объединяют в одном фиксированном месте вводные файлы, выводы и настройки. Каждый проект имеет назначение, которое выбирается в мастере создания нового проекта при его создании. Существует три режима:

- Общий — Регрессия по методу Earth для любого типа совокупности или набора данных. Это режим по умолчанию. Он обеспечивает полный рабочий процесс моделирования MARS без каких-либо дополнений, специфичных для конкретной области.
- Для оценки — Регрессия Earth, адаптированная для оценки недвижимости. Добавляет функции, специфичные для оценки отдельных объектов недвижимости, включая работу с объектом оценки, специальные обозначения столбцов, Метод остаточных ограничений (RCA) и .

- Анализ рыночной зоны — регрессия Earth, адаптированная для исследований рыночных зон. Добавляет функции для анализа групп объектов недвижимости на определённом рынке вместо оценки отдельного объекта оценки.

Во всех трёх режимах основной движок моделирования одинаков — вы всегда проводите подгонку модели Земли (MARS). Настройка «Цель» определяет, какие дополнительные инструменты и элементы интерфейса будут доступны.

Цель задается для каждого проекта отдельно. Цель проекта — наряду с указанием страны, штата, округа и города — фиксируется при создании проекта. Переход к другому проекту сбрасывает earthUI в чистое состояние по умолчанию (импортированные данные, результаты моделирования, вкладки, настройки переменных и параметры earth — всё это очищается), а затем восстанавливает сохранённую конфигурацию данного проекта для его цели.

1.4 Недвижимость — специальные материалы

При выборе пункта «Для оценки» или «Анализ рыночной зоны» earthUI активирует ряд функций, предназначенных для анализа рынка недвижимости:

- Обозначения специальных столбцов — Каждому предиктору можно присвоить специальную роль, такую как « contract_date », « dom », « concessions », « latitude », « longitude », « living_area », « lot_size », « actual_age », « effective_age », « area », « site_dimensions » или « display_only ». Эти обозначения определяют, как столбец будет обрабатываться во время подгонки модели, вывода результатов и формирования таблицы продаж. Полную справочную информацию см. в главе 6.
- Округление широты и долготы — Столбцы, обозначенные как широта или долгота, автоматически округляются до 3 знаков после запятой, чтобы предотвратить переобучение.
- Столбец «Срок продажи» — Когда столбец обозначен как contract_date и указана дата оценки, earthUI вычисляет столбец « sale_age » (количество дней между датой продажи и датой оценки) и использует его в качестве предиктора.
- Расчеты RCA (только при оценке) — В режиме оценки, после подгонки модели earthUI может вычислить результаты метода Residual Constraint (RCA), который выводит корректировки для каждого аналога, сводки чистых/валовых корректировок и скорректированную цену продажи объекта оценки.
- Таблица сравнения продаж (только для оценки) — Генерирует многолистовую книгу Excel с формулами для корректировок аналогов, разбивкой остаточных характеристик и скорректированными ценами продаж. См. главу 11.

Совет: Режим «Общее назначение» подходит для любых наборов данных — финансовых, научных, инженерных или связанных с недвижимостью. Режимы «Оценка» и «Рынок» просто предоставляют дополнительные удобные функции для специалистов в сфере недвижимости.

1.5 Установка earthUI и vProlog

Пакет earthUI опубликован на CRAN, поэтому для его стандартной установки в R достаточно одной строки:

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

В этой статье описывается текущий выпуск. Если на момент чтения этой статьи на CRAN по-прежнему доступна более старая версия, версию для разработчиков можно установить с GitHub (обратите внимание на subdir — пакет находится в папке pkg/ репозитория):

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

vProlog (необязательно). Экспериментальный механизм обработки комментариев (см. раздел «Извлечение текста из Prolog» в конце данной статьи) использует пакет vProlog — автономный встроенный движок Prolog. Он намеренно не является обязательной зависимостью: earthUI устанавливается и работает без него, а функции Prolog просто остаются отключенными до тех пор, пока он не будет установлен. vProlog распространяется через репозиторий r-universe автора, в котором доступны готовые бинарные файлы для Windows и macOS (компиляторный инструментарий не требуется):

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

Установка системы Prolog (SWI-Prolog или любой другой) не требуется — движок Trealla встроен непосредственно в сам пакет. Все остальное, описанное в этой статье, работает и без vProlog.

1.6 Начало работы

Чтобы использовать earthUI:

- 1 . Установите пакет из CRAN, как описано выше.
- 2 . Запустите приложение : выполните команду `earthUI::launch()` в R, или из командной строки выполните `Rscript -e 'earthUI::launch()'` . Приложение откроется в вашем веб-браузере на порту 7878. Вы также можете получить доступ к приложению напрямую, перейдя по адресу `http://localhost:7878` в своем браузере. (Модули постобработки запускаются таким же образом: `launch_glmnet()` на порту 7879, `launch_mgcv()` на порте 7880.)
- 3 . Создайте или откройте проект в разделе 1 боковой панели. Мастер создания нового проекта запрашивает название проекта, цель (Общее, Для оценки или «Анализ рыночной зоны»), местоположение (страна, штат, округ, город), а также, по желанию, исходный файл данных, который копируется в папку проекта «in/».
- 4 . Импортируйте данные в разделе 2, где перечислены файлы CSV и Excel, находящиеся в папке «in/» активного проекта. Файлы, добавленные вручную через Finder или Проводник, появятся после нажатия кнопки «Обновить».
- 5 . Настройка переменных — выберите целевые переменные и предикторы, укажите типы данных и назначьте специальные роли столбцам.
- 6 . Настройка параметров Земли — degree, prune, фильтры подмножеств и другие аргументы `earth()` .
- 7 . Подгонка модели — нажмите «Fit Earth Model» и просмотрите результаты на главной панели.
- 8 . Экспорт — загрузите прогнозы в формате Excel, сгенерируйте наборы отчетов Quarto или (в режиме оценки) рассчитайте корректировки RCA и таблицы сравнения продаж.

Настройки сохраняются на стороне сервера в базе данных проекта с указанием проекта и назначения с помощью кнопок «Сохранить текущие настройки как стандартные» в разделах 3 и 4 (см. главу 3 «Сохранение настроек»).

2 Требования к данным для ввода MLS

В рамках рабочих процессов по оценке недвижимости и анализу рынка исходные данные обычно получают из экспорта из системы Multiple Listing Service (MLS). В этой главе описывается ожидаемая структура файла и столбцы, которые может использовать earthUI.

2.1 Формат файла & Структура

earthUI поддерживает файлы в форматах CSV и Excel (.xlsx, .xls). При импорте названия столбцов автоматически преобразуются в формат `snake_case` — например, «Living SqFt» превращается в «`living_sqft`», «Contract Date» — в «`contract_date`», а «Цена продажи» — в «`sale_price`». Такая нормализация обеспечивает единообразные ссылки на столбцы на протяжении всего рабочего процесса. Разделитель CSV и знак десятичной запятой, используемые при импорте, определяются настройками локали (см. главу 3, «& региональных настроек»).

Ваш файл с данными должен представлять собой плоскую таблицу, в которой каждая строка соответствует одному объекту недвижимости, а каждый столбец — одной характеристике. Первая строка файла должна содержать заголовки столбцов.

2.2 Обязательные столбцы для оценки Режим

Хотя earthUI поддерживает работу с любым набором столбцов, для полноценного рабочего процесса оценки (корректировки по методу RCA + таблица сравнительных продаж) рекомендуется включить следующие столбцы в экспорт из MLS:

Колонка «	» Специальный тип	Назначение
Цена продажи	(целевая)	Переменная отклика для модели
Дата заключения договора	contract_date	Используется для расчета sale_age (количество дней с даты оценки)
Дата размещения объявления	listing_date	Используется вместе с датой заключения договора для расчета DOM, если колонка DOM отсутствует
Количество дней на рынке	dom	Количество дней на рынке; отображается в таблице продаж
Уступки	concessions	Уступки при продаже; чистая цена продажи (Net SP) = цена продажи – Уступки в таблице продаж
Жилая площадь (SF)	living_area	Позволяет рассчитывать остатки на квадратный фут (residual_sf , cqa_sf)
Площадь участка	lot_size	Сгруппировано в строке «Площадь участка Размеры» в таблице продаж
Размеры участка	site_dimensions	Сгруппировано с площадью участка в таблице продаж
Широта	latitude	Округлено до 3 знаков после запятой; используется для расчета близости и группы «Местоположение»
Долгота	longitude	Округлено до 3 знаков после запятой; используется для расчета близости и группы «Местоположение»
Идентификатор района	area	Сгруппировано в строке «Местоположение: Долгота Широта Район» в таблице продаж
Фактический возраст	actual_age	Сгруппировано в строке «Фактический возраст Эффективный возраст» в таблице продаж
Эффективный возраст	effective_age	Сгруппировано с фактическим возрастом в таблице продаж
Номер участка	(нет)	Отображается в строке APN таблицы продаж
Идентификатор объявления	(нет)	Отображается в строке APN таблицы продаж
Адрес	display_only	Отображается в строке «Адрес» таблицы продаж; исключено из модели

Названия столбцов в таблице могут быть на иностранном языке — «специальные» названия указаны на английском языке, чтобы программа R могла обрабатывать их особым образом. В противном случае указанные названия столбцов будут отображаться в регрессионных моделях, графиках и (при проведении оценок) в промежуточной таблице продаж.

Не все столбцы являются обязательными. earthUI адаптируется — если какой-либо столбец отсутствует, соответствующий показатель просто опускается. Например, если столбец «concessions» не указан, строка «Net SP» в таблице «Sales» Таблица отображает цену продажи без вычета скидок. Однако для моделей ценообразования в сфере недвижимости настоятельно рекомендуется использовать определённые столбцы для достижения приемлемой подгонки:

- 1 . Срок с момента продажи (Sale Age) — количество дней между датой заключения договора купли-продажи и датой оценки или анализа. Если используется многолетняя история продаж, особенно за периоды свыше 5 лет, показатель « sale_age » часто играет ключевую роль в оценке цены продажи. Фактически этот показатель зачастую настолько важен, что без него earthUI не может предложить никакой модели вообще.
- 2 . Жилая площадь — также известна под такими названиями, как «Living Sqft», «GLA» (общая жилая площадь) и т. д. Это ещё один основной фактор, определяющий цену продажи.

- 3 . Общее количество ванных комнат — общее количество полных, четвертных, половинных и 3/4 ванных комнат. Например, две полные ванные комнаты и одна половинная будут иметь значение 2,5.
- 4 . Гаражные места или площадь гаража — количество гаражных мест или площадь гаража в квадратных футах.
- 5 . Площадь участка — площадь земельного участка, обычно в квадратных футах или акрах.
- 6 . Долгота, широта и , если доступно, идентификатор района . Корректировки по этим параметрам будут объединены в одну корректировку по местоположению в таблице продаж.

2.3 Особые правила именования столбцов

earthUI идентифицирует столбцы по их специальным обозначениям типов, а не по названиям столбцов. При экспорте в формат MLS вы можете называть столбцы как угодно — главное, чтобы в таблице «Настройка переменных» (глава 6) был назначен правильный специальный тип.

Например, ваша система MLS может экспортировать данные о жилой площади как «GLA», «Living SqFt», «liv _ area» или «gross _ living _ area». После импорта (когда название преобразуется в формат snake_case) вам достаточно просто указать его как «living_area» в раскрывающемся списке «Special». После этого earthUI будет использовать его для расчёта остаточной стоимости на квадратный фут и группировки в таблице продаж, независимо от его исходного названия.

2.4 Качество данных& Полнота данных

- Отсутствующие значения (NA) : Строки, содержащие значения NA в любом столбце предикторов или целевом столбце, автоматически удаляются перед подгонкой модели. Параметр « na.action » внутренне всегда устанавливается в значение « na.fail », при этом удаление значений NA происходит до вызова функции « earth() ».
- Столбцы с датами : Даты должны быть представлены в формате, который R может обработать (например, 2025-06-15 , 06/15/2025 , June 15, 2025). earthUI автоматически определяет столбцы с датами, если успешно анализируется не менее 80% значений. . Двухзначные годы (например, 06/15/25) интерпретируются как ближайший соответствующий год, поэтому недавние экспорты из MLS обрабатываются правильно; четырёхзначные годы всегда принимаются буквально.
- Числовые столбцы : цена продажи, жилая площадь, площадь участка, уступки и подобные поля должны быть числовыми. Если ваши экспортные данные из MLS содержат цены с символами валют или разделителями тысяч (например, «\$ 350,000» или «350.000,00»), вам, возможно, потребуется очистить их перед импортом.
- Столбцы факторов : категориальные переменные, такие как идентификатор района, стиль или состояние, должны содержать управляемое количество уникальных значений. earthUI автоматически определяет факторы, но вы можете переопределить это определение.

Совет: после импорта проверьте столбец «T/F/NA» в таблице предикторов — в нём указано количество значений «true», «false» и «NA» для каждого столбца. Столбцы, в которых более 30 % пропущенных значений, отмечены красным цветом. Рекомендуется исключить столбцы с большим количеством пропущенных значений или очистить данные перед импортом.

2.5 Размещение строк объекта оценки

В режиме оценки в первой строке должна быть указана оценка объекта. Все остальные строки — это аналоги сделок купли-продажи. Цена продажи объекта оценки может быть оставлена пустой (NA) или установлена на любое значение — earthUI в любом случае будет рассматривать её как NA во время подгонки.

В режиме «Анализ рыночной зоны» размещение объекта в строке 1 является необязательным, и ни одна строка не исключается автоматически. Чтобы исключить строку с объектом (или любую другую строку) из подбора, присвойте ей нулевое значение в соответствующем столбце весов или воспользуйтесь фильтром подмножества.

В режиме «Общий» специальной обработки строк не предусмотрено — все строки обрабатываются одинаково.

3 Режим общего назначения

3.1 Обзор

Режим «Общего назначения» является режимом по умолчанию при запуске earthUI. Он обеспечивает полный рабочий процесс моделирования MARS для любых наборов данных — не только для недвижимости. Вы можете использовать earthUI для работы с научными данными, финансового анализа, инженерных исследований или решения любых регрессионных задач, в которых необходимо изучить нелинейные зависимости и взаимодействия между переменными.

В общем режиме интерфейс не отображает функции, специфичные для рынка недвижимости (специальные столбцы, срок продажи, округление координат, RCA). Боковая панель упрощена, чтобы сосредоточить внимание на отборе переменных, настройке параметров, подгонке модели и экспорте.

3.2 Рабочий процесс с боковой панелью

Боковая панель состоит из пронумерованных сворачиваемых разделов, которые помогут вам пройти весь путь от выбора проекта до экспорта:

1. Проект — выберите существующий проект или создайте новый с помощью мастера «Новый проект», в котором указываются цель, страна и административное местоположение (все эти параметры блокируются после создания). Активный проект определяет, откуда считываются файлы ввода и куда записываются все результаты — отдельного поля для указания папки вывода нет.

2. Импорт данных — выберите файл CSV или Excel из папки in/ текущего проекта; ссылка «Обновить» запускает повторное сканирование папки на наличие файлов, добавленных вручную. Для файлов Excel с несколькими листами появляется окно выбора листа. Названия столбцов автоматически преобразуются в регистр «snake» _.

3. Настройка переменных — выбор целевой переменной (поддерживается несколько целевых переменных), дата оценки и таблица предикторов с флажками «Включить», «Фактор», «Линейный» и «Скрытый». Подробную информацию см. в главе 6.

4. Параметры Earth Call — все аргументы функции 'earth()': degree, штраф, nk, метод прореживания, перекрёстная проверка и др. Полный справочник по параметрам см. в главе 7.

Выполнение обработки Prolog — необязательный раздел без номера, расположенный между разделами 4 и 5; он отключен, если экспериментальная функция Prolog не включена в настройках. См. главу «15».

5. Подгонка модели Земли — поле для ввода случайного начального значения и кнопка, запускающая модель в асинхронном режиме. Подробности о подгонке см. в главе 8.

6. Скачать остатки по оцениваемой целевой переменной (переменным) & — (в режимах оценки и рыночном режиме функция называется «Скачать остатки по оцениваемым ценам продажи & ») Экспортирует прогнозы, остатки, оценки CQA и вклады по g-функциям в файл Excel. См. главу 9.

- 7–8. Корректировки RCA и таблица цен — только в режиме оценки. См. главы 10 и 11.

9. Сгенерировать отчёт Quarto — записывает автономный набор отчёта Quarto (исходный код .qmd и ресурсы) в папку вывода проекта. Имеет номер 7 вне режима оценки. См. главу 12.

10. Конвертация отчёта Quarto — преобразует любой файл формата «.qmd» в HTML, Word или PDF. См. главу 12.

3.3 Вкладки главной панели

После подгонки на главной панели отображается десять вкладок:

Вкладка «	» Содержание
Данные	Предварительный просмотр импортированных данных. В режиме оценки данные разбиты на таблицы «Объект оценки» и «Аналоги».
Уравнение	Уравнение построенной модели, отображаемое в формате LaTeX/MathJax, с указанием каждой базисной функции и её коэффициента.
Сводка	Ключевые показатели (R^2 , $GRSq$, GCV , RSS , $CV R^2$) и таблица коэффициентов.
Важность переменных	Гистограмма и таблица с ранжированными баллами важности предикторов.
Вклад	Графики вклада g-функции. Выберите g-функцию из раскрывающегося списка; для одномерных предикторов отображаются линейные графики, для двумерных взаимодействий (степень ≥ 2) — 3D-перспективные и контурные графики.
Корреляция	Тепловая карта корреляций предикторов.
Диагностика	Три графика: «Остатки против прогнозируемых значений», «Нормальный Q-Q» и «Фактические против прогнозируемых значений».
Корректировки RCA	(только для оценки) Гистограммы показателей «Корректировка по остаткам»%, «Чистая корректировка»% и «Валовая корректировка»% по аналогам, заполняемые после завершения шага 7 («Расчет корректировок RCA»).
Остаточный вклад	Оценивает вклад редких или исключенных признаков в рынок путем совместной регрессии остатков модели по столбцам, которые модель никогда не использовала (см. subsection 4.4).
ANOVA	Таблица декомпозиции ANOVA, показывающая вклад каждой базисной функции в RSS .
Вывод Earth	Необработанный текстовый вывод из <code>'earth::summary()'</code> , включая детали прохода обрезки.

3.4 Сохранение настроек

earthUI сохраняет вашу конфигурацию на стороне сервера в базе данных проекта (projects.sqlite в корневом каталоге regProj) с ключом, определяемым проектом и целью. Сохранение осуществляется с помощью кнопок: кнопка «Сохранить текущее как по умолчанию» в разделе 3 сохраняет цель, дату оценки и конфигурацию предиктора, а кнопка в разделе 4 — параметры вызова earth и матрицу «Допустимые взаимодействия». Каждая кнопка обновляет только свою половину, поэтому они никогда не перезаписывают друг друга, а результаты подгонки не сохраняются автоматически — вы можете свободно экспериментировать, не перезаписывая сохраненные настройки по умолчанию. Несколько файлов ввода проекта (небольшой тестовый фрагмент, полный набор данных) используют одну конфигурацию, и поскольку настройки хранятся в дереве проекта, они переносятся вместе с ним при синхронизации, создании резервной копии или совместном использовании папки regProj.

3.5 Темный режим

Нажмите на значок луны/солнца в правом верхнем углу, чтобы переключаться между светлой и темной темами. Настройки темы сохраняются в локальном хранилище и сохраняются

между сеансами.

3.6 Язык и регион & Региональные настройки

earthUI поддерживает международные стандарты форматов чисел, дат и файлов CSV посредством системы локалей, основанной на странах. В раскрывающемся списке «Страна» в меню «Настройки» (значок шестеренки) на верхней панели меню можно выбрать одну из 31 предустановки для поддерживаемых стран. Каждая предустановка настраивает:

- Разделитель CSV — запятая (,) для США/Великобритании/Японии или точка с запятой (;) для большинства стран Европы, где запятая используется в качестве десятичного знака.
- Десятичный знак — точка (.) или запятая (,).
- Разделитель тысяч — запятая (США/Великобритания/Япония), точка (Германия/Италия/Испания), пробел (Финляндия/Франция/Польша/Прибалтика/Украина/Россия) или апостроф (Швейцария).
- Формат даты — ММ/ДД/ГГГГ (США), ДД/ММ/ГГГГ (большая часть Европы), или ГГГГ-ММ-ДД (Швеция/Литва/Япония/Канада).
- Формат бумаги — Letter (США/Канада/Мексика) или A4 (везде else).

3.6.1 Поддерживаемые страны

Страна	CSV Сентябрь	Десятичный формат	Тысячи	Дата	Бумага
США, Канада (CSV)	,	.	,	MDY/YMD	Формат «Letter»
Великобритания, Ирландия, Австралия, Новая Зеландия	,	.	,	DMY	Формат A4
Германия, Австрия, Бельгия, Нидерланды, Италия, Испания, Португалия, Дания, Бразилия	;	,	.	DMY	Формат A4
Франция, Польша, Чехия, Норвегия, Латвия, Эстония	;	,	(пробел)	DMY	A4
Финляндия, Швеция	;	,	(пробел)	DMY/YMD	A4
Швейцария	;	,	,	DMY	A4
Украина, Россия	;	,	(пробел)	DMY	A4
Турция	;	,	.	DMY	A4
Япония, Южная Корея	,	.	,	YMD	A4
Мексика	,	.	,	DMY	Letter
Литва	;	,	(пробел)	YMD	A4

3.6.2 Paper Override

Под полем выбора страны находится раскрывающийся список «Формат бумаги» (Letter или A4), позволяющий изменить размер страницы отчёта без смены страны. Разделитель

CSV, знак десятичной запятой и порядок форматирования даты соответствуют настройкам выбранной страны; при смене страны настройка формата бумаги сбрасывается на значения по умолчанию для этой страны.

3.6.3 Сохранение настроек по умолчанию

Нажмите «Сохранить» в меню «Настройки», чтобы сохранить настройки региональных параметров глобально. Эти настройки по умолчанию будут применяться ко всем будущим сеансам, независимо от того, какой проект вы откроете. Настройки для отдельных проектов (целевая переменная, предикторы, параметры) сохраняются отдельно в базе данных проекта; настройки региональных параметров по умолчанию сохраняются для всех проектов в базе данных SQLite на уровне пользователя. Этот двухуровневый подход разработан для таких организаций, как аудиторские фирмы, которые работают с данными из разных стран — установите наиболее часто используемую страну в качестве значения по умолчанию, а затем при необходимости вносите корректировки.

Совет: Формат чисел на осях графика и в подписях к наклонам автоматически адаптируется к вашей языковой настройке. В немецкой языковой настройке для разделителя тысяч используются точки (200.000), в финской — пробелы (200 000), а в швейцарской — апострофы (200'000). Символы валют не отображаются — earthUI не зависит от валюты.

4 Режим оценки

При выборе цели «Для оценки» earthUI автоматически настраивается на оценку одного объекта недвижимости. Все функции, описанные в главе 3, остаются доступными; в данной главе рассматриваются только дополнительные возможности, специфичные для оценки.

4.1 Обработка строк объекта оценки

В режиме оценки первая строка вашего набора данных представляет собой объект оценки, а все остальные строки — аналогичные сделки. Ваш вводный файл должен быть организован соответствующим образом (см. главу 2). Цена продажи объекта может быть оставлена пустой или установлена на любое значение — earthUI автоматически рассматривает ее как NA во время подгонки модели.

После импорта вкладка «Данные» разделяется на два раздела: «Объект оценки недвижимости» (строка 1) и «Аналоги» (строки 2 и далее). Строка 1 всегда исключается из подгонки модели — об этом свидетельствует уведомление «Пропуск строки 1 (объект оценки). Подгонка модели на N строках». После подгонки модель по-прежнему генерирует прогнозы для строки объекта оценки, которые отображаются в выводе как «`est_<target>`».

4.2 Дата оценки & Возраст, с которого разрешена продажа

Поле «Дата оценки» отображается в разделе «Настройка переменных» во всех режимах работы (по умолчанию установлена сегодняшняя дата) — не только в режимах оценки, но и в режимах общего и рыночного анализа — однако оно используется только в режимах оценки и рыночного анализа. Если в раскрывающемся списке «Special column» («Специальный столбец») вы укажете столбец «`contract_date`» (Дата вступления в силу), earthUI вычислит столбец «`sale_age`» (Количество дней) — количество целых дней между датой заключения договора каждой сделки и датой оценки. Этот столбец заменяет исходный столбец с датами в качестве предикторов.

При первом нажатии кнопки «Подгонка» после указания даты заключения договора earthUI создает столбец «`sale_age`» и выдает уведомление о необходимости нажать кнопку «Подгонка» еще раз, чтобы включить его.

4.3 Специальный столбец Обозначения

В режимах оценки и рынка для каждого предиктора в таблице «Настройка переменных» появляется специальный раскрывающийся список. Полный список специальных типов и их эффектов:

Специальный тип	Эффект
no	По умолчанию — без специальной обработки
contract_date	Запускает автоматический sale_age расчет с датой оценки
listing_date	Используется вместе с датой заключения договора для расчета DOM, если столбец dom отсутствует
dom	Обозначает столбец «Дни на рынке»; отображается в таблице продаж
concessions	Обозначает скидки при продаже; используется в формуле чистой продажной цены (цена продажи — скидки) в таблице продаж
latitude	Значения округляются до 3 знаков после запятой; используются для расчёта близости по формуле Хаверсина и определения группы местоположения в таблице продаж
longitude	Значения округлены до 3 знаков после запятой; используются для расчёта близости по формуле Хаверсина и определения группы местоположения в таблице продаж
area	Сгруппированы в строке «Местоположение: Долг. Шир. Площадь» в таблице продаж
living_area	Включает остатки на квадратный фут (residual_sf, cqa_sf) в выводе
lot_size	Сгруппированы в строке «Размер участка Размеры» в таблице продаж
site_dimensions	Сгруппировано по площади участка в таблице продаж
actual_age	Сгруппировано в строке «Фактический возраст Эффективный возраст» в таблице продаж
effective_age	Сгруппировано по фактическому возрасту в таблице продаж
display_only	Столбец включен в экспортные данные, но исключен из подгонки модели (допускается несколько)
remarks	(Экспериментально) Колонка с комментариями в виде свободного текста, разбитая на столбцы признаков с помощью опционального шага Prolog (допускается несколько); см. главу 15
list	(Экспериментально) Колонка, разделенная запятыми, разбитая на столбцы «one-hot» TRUE/FALSE с помощью опционального шага Prolog (допускается несколько); см. главу 15

Для каждого специального типа допускается только один столбец (за исключением display_only, remarks и list). Присвоение специального типа второму столбцу автоматически удаляет его из первого. Рядом с именем переменной появляется небольшой синий значок, указывающий присвоенный ей специальный тип.

4.4 Остаточный вклад (оценка редких признаков)

Характеристики, слишком редкие для регрессионного канала — мастерская, представленная лишь в нескольких сделках, — относятся к слою остаточных величин RCA. На вкладке «Вкладка остаточных величин» оценивается их вклад в рынок путем регрессии остатков подгоняемой модели по столбцам, которые модель никогда не использовала, с помощью метода обычных малейших квадратов. Выбор нескольких столбцов позволяет оценить их совместно, благодаря чему коррелирующие характеристики (например, outbuilding_sqft плюс pr_has_workshop) распределяют вклад между собой, а не учитываются дважды. Двоичные флаги дают суммарные оценки и перечисляют остатки по каждой отмеченной сделке в качестве доказательства; числовые столбцы дают вклад на единицу; в отчёте указана стандартная ошибка в размере одной SE ($t \geq 2$ в качестве обычного критерия значимости). Термины, подкреплённые менее чем десятью наблюдениями, автоматически получают отметку «caution» в примечании сгенерированного рабочего файла, где в одном

абзаце для отчёта документируются метод, оценки и предостережения. Из списка кандидатов исключаются только те предикторы, которые фактически использовались моделью — редкий флаг, предложенный при подгонке, но так и не выбранный при прямом проходе, остаётся приемлемым, поскольку модель не учла ни одной из его значений.

4.5 Обзор корректировок RCA

Кнопка «Рассчитать корректировки RCA» & Download (боковая панель, раздел 7, отображается только в режиме оценки после подгонки аналогов) позволяет рассчитать корректировки, основанные на рыночных данных, для каждого аналога по отношению к объекту оценки. Полный рабочий процесс RCA описан в главе 10.

После расчёта корректировок RCA становится доступной кнопка «Сгенерировать таблицу продаж» & (раздел 8 боковой панели). Порядок действий с таблицей продаж описан в главе 11.

Совет: Балл CQA, присваиваемый объекту оценки, определяет, какая доля остаточного распределения относится именно к этому объекту. Балл 5,00 означает, что объект оценки находится на медиане среди аналогов.

5 Режим анализа рыночных зон

Если в качестве цели выбрать «Анализ рыночной зоны», earthUI предоставляет те же функции, характерные для рынка недвижимости, что и в режиме оценки (специальные столбцы, срок с момента продажи, округление координат), но ориентирован на анализ группы объектов недвижимости, а не на оценку отдельного объекта.

5.1 Отличия от режима оценки Режим «

»

- Отсутствие автоматического исключения объектов оценки — в отличие от режима оценки, в режиме рынка строка 1 не исключается из подгонки автоматически. Все строки включаются в подгонку; чтобы исключить строку объекта оценки из подгонки, присвойте ей нулевое значение в соответствующем столбце весов (как в демонстрационном наборе данных) или используйте фильтр подмножества.
- Отсутствие разделов RCA и «Таблица продаж» — шаги «Рассчитать корректировки RCA & Скачать» и «Сгенерировать таблицу продаж& Скачать» недоступны. Режим «Market» ориентирован на подгонку модели и вывод результатов, а не на корректировки по каждому аналогу.
- Нумерация боковых панелей — без шагов RCA и «Sales Grid» раздел «Generate Quarto Report» имеет номер 7 вместо 9 (раздел «Convert Quarto Report» по-прежнему имеет номер 10).

5.2 Когда следует использовать режим «Рынок»

Режим «Анализ рыночной зоны» подходит в следующих случаях:

- Построение регрессионной модели для микрорайона или рыночной зоны с целью понимания факторов, влияющих на стоимость
- Анализ того, как такие переменные, как площадь, возраст, площадь участка и расположение, влияют на цены продажи в группе объектов недвижимости
- Подготовка материалов для анализа рыночной конъюнктуры или определения границ района
- Работа с набором данных, в котором рассматриваемый объект оценки находится в первой строке, но вы хотите иметь возможность включать или исключать его из модели

Совет: Режим «Market» также полезен для общей регрессионного анализа на рынке недвижимости, когда требуются специальные характеристики столбцов (срок продажи, округление координат), но не нужен рабочий процесс корректировки RCA.

6 Отбор переменных

В разделе 3 боковой панели — «Настройка переменных» — вы выбираете, какие столбцы будут включены в модель и как с ними будет проводиться работа.

6.1 Целевые переменные

В раскрывающемся списке «Целевая переменная (ответ)» в верхней части раздела 3 перечислены все столбцы вашего набора данных. Выберите один столбец для стандартной модели с одним ответом или несколько столбцов для модели с несколькими ответами. При выборе нескольких целевых переменных модель подгоняется для всех ответов одновременно с использованием метода «`earth(cbind(y1, y2, ...) ~ .)`».

Столбцы, выбранные в качестве целевых, автоматически исключаются из списка предикторов.

6.2 Таблица предикторов

Ниже селектора целей в таблице перечислены все оставшиеся столбцы со следующими полями:

Столбец	Описание
Переменная	Название столбца (полное название, отображаемое в всплывающей подсказке, если оно сокращено), перед которым стоит переключатель $\times / +$ (см. ниже). В режимах оценки/рынка синий значок указывает на назначенный специальный тип.
Тип	Раскрывающийся список типов данных: <code>numeric</code> , <code>integer</code> , <code>character</code> , <code>logical</code> , <code>factor</code> , <code>Date</code> , <code>POSIXct</code>
Включить	Флажок — включить этот столбец в качестве предиктора в модель
Фактор	Флажок — рассматривать этот столбец как категориальную переменную
Линейный	Флажок — принудительно использовать только линейный ввод (без функций типа « <code>hinge</code> »)
Латентная	Флажок — полностью исключить этот столбец из модели (исключить из <code>earth</code> и из данных, экспортируемых в <code>glmnetUI</code> / <code>mgcvUI</code>), зарезервировав его для работы с латентными переменными
Специальный	Раскрывающийся список (только для оценки/рынка) — см. справочник по типам столбцов «Специальные»
T/F/NA	Справочник по специальным типам столбцов ниже Три значения на столбец: для логических значений — <code>TRUE/FALSE/NA</code> ; для числовых значений — $\neq 0 / = 0 / NA$. Отображается красным, если отсутствует более 30% значений

В подсказке над таблицей дано объяснение сокращений: «`Type` = тип данных столбца, `Inc` = включить в качестве предиктора, `Factor` = рассматривать как категориальную переменную, `Linear` = только линейная модель (без переходных членов), `Special` = роль столбца (например, дата заключения договора `_`)».

«Мягкое отключение» вместо удаления. Переключатель «X» рядом с каждым именем переменной делает строку серым цветом и исключает столбец из модели, не удаляя его из данных; при переключении на зеленый значок «+» столбец снова включается. Столбцы никогда не удаляются.

6.3 Определение типов данных & Переопределения

earthUI автоматически определяет типы данных при импорте. Распознаются столбцы с числовыми, целочисленными, логическими, факторными и датными значениями. Символьные столбцы, которые выглядят как даты (не менее 80 % значений соответствуют распространенным форматам дат), классифицируются как Date.

Вы можете переопределить любое распознавание, изменив значение в раскрывающемся списке «Тип». При изменении типа столбца на «character» или «factor» флажок «Factor» автоматически устанавливается. Изменение типа влияет на то, как столбец передается в функцию «earth()».

6.4 Флаги факторов и линейных моделей

- «Фактор» — Если этот флажок установлен, столбец рассматривается как категориальная переменная. Это подходит для таких столбцов, как style, area_id или grade, которые представляют дискретные группы, а не непрерывные измерения. Столбцы-факторы включаются в модель в качестве индикаторных (фиктивных) переменных.
- Линейный — Если этот флажок установлен, столбец принудительно включается в модель линейно — для него не создаются функции типа «hinge» (кусочно-линейные). Это полезно, когда известно, что переменная имеет строго линейную зависимость от целевой переменной.

6.5 Специальные типы столбцов См.

В режимах «Оценка» и «Рынок» в раскрывающемся списке «Special» доступны следующие параметры. Каждый тип можно назначить не более чем одному столбцу (за исключением «display_only», «remarks» и «list», которые допускают назначение нескольким столбцам):

Дата & . Типы времени:

- contract_date — Запускает автоматический расчет показателя «sale_age». Исходный столбец даты заменяется столбцом с целыми числами, отражающим количество дней между датой продажи и датой оценки.
- listing_date — Используется в качестве резервного варианта для расчета показателя «Days on Market» (DOM = дата заключения договора — дата размещения объявления), если явно не указан столбец «dom».
- dom — Идентифицирует столбец «Дни на рынке». Отображается в строках «APN» и «Дата продажи» таблицы продаж.

Типы денежных единиц:

- concessions — Обозначает скидки при продаже (кредиты продавца, поощрения для покупателя и т. д.). Используется в таблице продаж для расчета чистой цены продажи: Чистая цена продажи = Цена продажи — Скидки.

Размер & . Типы мест:

- latitude — Значения автоматически округляются до 3 знаков после запятой для предотвращения переобучения. Используются для расчета близости по формуле Хаверсина (расстояние от объекта оценки до каждого аналога) и сгруппированы в строке «Местоположение» таблицы продаж.

- `longitude` — Округление такое же, как для широты. Используется для расчета близости и в группе «Местоположение».
- `area` — Обычно служит идентификатором района или территории. Группируется вместе с широтой и долготой в строке «Местоположение: Долгота | Широта | Площадь» таблицы продаж.
- `living_area` — Позволяет производить расчёты остатков на квадратный фут (`residual_sf` и `sqa_sf`) в выводе данных.
- `lot_size` — Сгруппировано в строке «Размеры участка» таблицы продаж.
- `site_dimensions` — Сгруппировано с площадью участка в таблице продаж (например, «75x120»).

Возрастные группы:

- `actual_age` — Сгруппировано в строке «Фактический возраст | Эффективный возраст» таблицы продаж.
- `effective_age` — Сгруппировано по фактическому возрасту в таблице продаж.

Типы дисплеев:

- `display_only` — Этот столбец включается в экспортные файлы Excel, но полностью исключается из процесса подгонки модели. Используйте эту настройку для полей с адресами, номерами MLS, идентификаторами земельных участков или другими справочными данными, которые не должны выступать в качестве предикторов. Эту настройку можно применить к нескольким столбцам.

Типы извлечения текста (экспериментальные):

- `remarks` — Столбец с произвольным текстом (например, `public_remarks`), который будет разбит на столбцы с характеристиками с помощью дополнительного шага Prolog. Не выполняет никаких действий, пока данная функция не будет включена в настройках. Допускается указание нескольких столбцов. См. главу 15.
- `list` — Столбец с разделителем запятой (например, `existing_structures`), который будет разбит на однозначные `<col>_<value>` столбцы TRUE/FALSE в рамках того же этапа. Допускается использование нескольких столбцов. См. главу 15.

6.6 Несколько целей

Выбор более одной целевой переменной позволяет провести подгонку модели Земли с несколькими откликами. При выборе нескольких целевых переменных:

- Модель прогнозирует все ответы одновременно, используя общие базисные функции для всех целей
- Кнопка «wp» (веса ответов) становится активной, что позволяет присвоить числовой вес каждому ответу
- Модели дисперсии (`varmod.method`) отключены (не поддерживаются для многовариантного прогнозирования)
- На вкладках «Результаты» отображаются показатели по каждому ответу, уравнения и графики диагностики

Совет: Столбцы, помеченные как «`display_only`» (не используемые для прогнозирования), остаются в наборе данных и отображаются при экспорте в Excel, но полностью исключаются из таблицы предикторов. Используйте эту функцию для столбцов с идентификаторами, адресами или другими справочными полями.

7 Выбор параметров

Раздел 4 боковой панели — «Параметры Earth Call» — предоставляет доступ ко всем аргументам, поддерживаемым функцией «earth()». У каждого параметра есть синий значок справки (?) с всплывающей подсказкой. Параметры сгруппированы в сворачиваемые подразделы.

7.1 Основные параметры

Параметр	По умолчанию	Описание
подмножество	(пустое)	Выражение фильтрации строк. См. раздел «Фильтрация подмножеств» ниже.
веса	NULL xml-ph-0028@de. NULL xml-ph-0028@de. веса NULL xml-ph-0028@de. веса NULL xml-ph-0028@de. веса NULL xml-ph-0028@de. веса NULL xml-ph-0028@de. веса NULL xml-ph-0028@de. веса NULL xml-ph-0028@de. веса NULL xml-ph-0028@de. веса NULL xml-ph-0028@de. веса NULL xml-ph-0028@de. веса NULL xml-ph-0028@de. Фильтрация» ниже. веса NULL	Селектор столбцов для весов случаев (строк). В списке указаны только числовые столбцы.
wp	NULL	Веса отклика для многоцелевых моделей. Кнопка открывает диалоговое окно с одним числовым полем ввода для каждой цели (по умолчанию по 1,0 для каждой).
keepxy	off	Сохранить x, y, подмножество и веса в объекте модели.
trace	0	Уровень трассировки для вывода результатов подгонки (от 0 до 5).
glm	none	Дополнительное семейство GLM: gaussian, binomial или poisson.

Параметр	По умолчанию	Описание
degree	1	Максимальный порядок взаимодействия. Установка значения $\text{degree} \geq 2$ автоматически включает перекрёстную проверку и отображает матрицу «Допустимые взаимодействия»
штраф	2	Штраф GCV на узел. Более высокие значения дают более простые модели.

7.2 «Forward Pass»

Параметр	По умолчанию	Описание
nk	auto	Максимальное количество членов до обрезки. Автоматически заполняется рекомендуемым значением для загруженных данных (см. раздел «Рекомендуемые значения параметров» ниже).
thresh	0.001	Пороговое значение для прямого шага. Меньшие значения допускают большее количество членов.
minspan	auto	Минимальный интервал между узлами. Заполняется автоматически рекомендуемым значением. Отрицательные значения устанавливают максимальное количество узлов на одного предиктора.
endspan	auto	Конечный интервал — минимальное расстояние от узла до края данных. Заполняется автоматически рекомендуемым значением.
newvar.штраф	0	Штраф за введение новой переменной (поощряет повторное использование существующих предикторов).
fast.k	20	Количество родительских членов, которые следует учитывать в алгоритме fast MARS.
fast.beta	1	Регулирует коэффициент старения в алгоритме fast MARS.

7.3 Допустимые взаимодействия

Если значение параметра «degree» установлено на 2 или выше, под основными параметрами появляется таблица взаимодействий. Это верхнетреугольная таблица из флажков, по одному для каждой пары предикторов. Установленный флажок означает, что между двумя предикторами допускается взаимодействие; снятие флажка запрещает данное взаимодействие.

При нажатии на название предиктора (заголовок строки или столбца) включаются или отключаются все взаимодействия для этой переменной. Флажки «Разрешить все» и «Снять все» в верхней части окна позволяют управлять настройками для всех элементов сразу. В матрице используются фиксированные заголовки, поэтому заголовки столбцов и строк остаются видимыми при прокрутке.

В информационном уведомлении содержится следующее напоминание: «Термины взаимодействия повышают риск переобучения. Включена перекрёстная проверка (10 блоков).»

7.4 Обрезка

Параметр	По умолчанию	Описание
pmethod	backward	Метод обрезки: backward , none , exhaustive , forward , seqrep , или cv .
nprune	NULL	Максимальное количество слагаемых после обрезки. Оставьте пустым для автоматического выбора.

7.5 Перекрёстная проверка

Параметр	По умолчанию	Описание
nfold	10	Количество блоков перекрёстной проверки. Установите значение 0, чтобы отключить перекрёстную проверку.
ncross	20	Количество повторений перекрёстной проверки.
stratify	on	Стратифицировать выборки перекрёстной проверки, чтобы каждая складка имела схожее распределение ответов.

Когда степень $\geq$ равна 2, автоматически включается перекрёстная проверка (nfold установлено на 10) для предотвращения переобучения из-за взаимодействий.

7.6 Фильтрация подмножеств

В поле ввода «Подмножество» можно указать выражение на языке R, которое определяет, какие строки будут использоваться для подгонки модели. Выражение можно ввести непосредственно (например, `sale_age < 365 & area_id == 460`) или создать его визуально с помощью кнопки «Построить фильтр» (...).

Диалоговое окно «Создать фильтр» — Нажмите «Создать фильтр ...», чтобы открыть пошаговое диалоговое окно. В каждой строке условия есть раскрывающийся список столбцов, оператор (<, >, <=, >=, ==, !=) и поле ввода значения, которое адаптируется к типу столбца: числовое поле для чисел, выбор даты для дат и раскрывающийся список уникальных значений для столбцов с символами или факторами. Условия объединяются с помощью операторов AND (&) или OR (|). В нижней части окна в режиме предварительного просмотра отображается выражение и количество соответствующих строк. Нажмите «Применить», чтобы вставить выражение в поле ввода текста.

В ручных выражениях для столбцов даты необходимо использовать обертки `as.Date("...")` или `as.POSIXct("...")`. Диалоговое окно «Создать фильтр» обрабатывает это автоматически.

Фильтрация по подмножествам является неразрушающей — исключенные строки остаются в наборе данных и получают прогнозы при экспорте в Excel.

Совет: Используйте скобки для группировки условий «ИЛИ» при их объединении с условиями «И». Без скобок выражение «A & B | C» будет интерпретироваться как «(A & B) | C», что может не соответствовать вашим намерениям.

7.7 Рекомендуемые значения параметров

В разделе 4 earthUI отображает рекомендуемые значения под ключевыми параметрами. Эти рекомендации динамически обновляются в зависимости от количества строк, используемых для подгонки (n), и выбранных предикторов (p). Формулы получены на основе статьи Фридмана о методе MARS, внутренних алгоритмов earth и эмпирических испытаний. Поля `nk`, `minspan` и `endspan` заполняются рекомендуемыми значениями при загрузке данных; изменения, внесенные вручную, сохраняются.

7.7.1 Параметры передачи вперед

`nk` (максимальное количество термов до обрезки):

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

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

Значение по умолчанию для параметра Earth, $\min(200, \max(20, 2p)) + 1$, не учитывает размер набора данных и может оказаться слишком малым для больших наборов данных, что ограничивает прямой проход, не позволяя ему адекватно проанализировать все предикторы. Параметр $\lfloor n/10 \rfloor$ допускает примерно один член на каждые 10 наблюдений — это стандартное эмпирическое правило, позволяющее избежать чрезмерной параметризации. Ограничение в 100 предотвращает снижение эффективности.

`minspan` (минимальное количество наблюдений между узлами):

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

endspan (минимальное количество наблюдений от границ данных):

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

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

При автоматическом расчете minspan в Earth используется уравнение 43 Фридмана: $\lfloor (-\ln(-\ln 0.95) + \ln(p \cdot n)) / (2.5 \ln 2) \rfloor$, которое масштабируется как $\ln(np)$ и растёт слишком медленно для больших наборов данных. По адресу $n=1500$ оно даёт лишь ≈ 7 , что соответствует примерно 245 возможным местам расположения узлов на каждый непрерывный предиктор — это слишком много, что позволяет прямому проходу проводить подгонку под шум. Рекомендуемая формула нацелена на примерно 100 кандидатов на каждый предиктор для больших значений n , в то время как для малых значений n (где она хорошо протестирована) используется автоматический расчёт earth. Endspan устанавливается немного больше, чем minspan, чтобы обеспечить дополнительную защиту границ, что помогает, когда данные имеют тонкие хвосты (что часто встречается в сфере недвижимости).

штраф (штраф GCV за каждый узел):

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

Это стандартное значение в Earth — 2 для аддитивных моделей, 3 для моделей взаимодействия (более высокий штраф компенсирует большее пространство поиска).

newvar.penalty (штраф за добавление нового предиктора):

Рекомендуемое значение: 0,1, если предикторы коррелируют между собой (например, жилая площадь с количеством спален, количеством ванных комнат, площадью участка). Это приводит к тому, что при пошаговом прохождении модель будет отдавать предпочтение повторному использованию предикторов, уже присутствующих в модели, а не введению коррелирующих альтернативных вариантов. Механизм: улучшение RSS для каждого нового предиктора умножается на $1/(1 + \text{penalty})$. При значении newvar.penalty = 0.1 новая переменная должна улучшать RSS как минимум на 10 % больше, чем очередной узел существующей переменной. Это позволяет получать более простые модели с меньшим количеством предикторов без ущерба для окончательных оценок коэффициентов (штраф снимается после отбора). Оставьте значение 0, если предикторы не коррелируют. Обратите внимание, что earth не поддерживает ненулевое значение newvar.penalty вместе с весами по случаям — earthUI отключает ввод (принудительно устанавливая значение 0) при указании столбца с весами.

7.7.2 Параметры обрезки

pmethod: Рекомендуется: backward. Основанная на GCV обрезка по методу «backward» по умолчанию является детерминированной — одни и те же данные всегда дают одну и ту же

модель, независимо от случайных начальных значений. Это важно для воспроизводимости результатов, особенно при проведении оценок. Метод «cv» использует перекрёстную проверку для обрезки, что влечет за собой зависимость от начальных значений.

`prune`: Рекомендуется: оставить пустым (NULL). Позвольте GCV выбрать оптимальное количество термов. Установка значения `prune` накладывает жесткое ограничение, которое перебивает решение GCV.

7.7.3 Перекрёстная проверка Параметры

`nfold` (блоки перекрестных проверок): рекомендуемое значение в зависимости от диапазона размера выборки:

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

Увеличение количества блоков позволяет обучать каждый блок на большем объеме данных (что дает менее смещенную оценку), но приводит к уменьшению размера выделенного тестового блока. В таблице показано, как с ростом выборки количество блоков постепенно увеличивается до стандартного 10-кратного, при этом каждый тестовый блок остается достаточно большим (примерно 50+ строк) для получения стабильной метрики по каждому блоку, а при небольших выборках количество блоков уменьшается до 5-блоков для небольших выборок — где оценка CVR^2 по-прежнему содержит шум, поэтому следует увеличить `ncross` (например, до 15–20), чтобы усреднить дисперсию. Ограничение в 10 отражает то, что каждый дополнительный блок представляет собой полную подгонку по данным, приносящую пренебрежимо малую выгоду для диагностики. При использовании `pmethod = "backward"` перекрёстная проверка не влияет на модель — она лишь вычисляет диагностический CVR^2 и предоставляет остатки для модели дисперсии.

`ncross` (количество повторений CV): рекомендуемое значение в зависимости от диапазона размера выборки:

n (подгонка строк)	ncross
0–199	20
200–399	16
400–699	12
700–1199	10
1200–2999	7
3000+	5

Повторы выравнивают случайные отклонения, связанные с разбиением на блоки, что способствует стабилизации CVR^2 ; для больших выборок требуется меньше повторов, поскольку однократная перекрёстная проверка уже обеспечивает более стабильные результаты. Остатки от повторной перекрёстной проверки также используются в модели дисперсии.

7.7.4 Модель дисперсии

`varmod.method`: Рекомендуется: `lm` — модель Милборроу по умолчанию: робастная линейная модель дисперсии для интервалов предсказания (разброс оценивается по шумным остаткам,

поэтому простая модель обеспечивает лучшую обобщаемость). Если lm не сходится на ваших данных, попробуйте rlm или $const$ (гомоскедастичность, всегда сходится), либо воспользуйтесь $earth$ (также обрабатывает нелинейный разброс). Для этого требуются $nfold > 1$ и $ncross > 1$.

7.7.5 Файл вывода Earth

После каждой успешной подгонки `earthUI` записывает файл «`<filename>_earth_output_<timestamp>.txt`» в папку вывода. Этот файл содержит термины модели, сводные статистические показатели (R^2 GRSq, CVRSq), сведения о модели дисперсии и журнал трассировки. Для каждой подгонки создается отдельный файл, что обеспечивает накопленную историю данных для сравнения конфигураций параметров.

7.8 Настройки по умолчанию

Радио-кнопка в верхней части раздела 4 определяет, какие настройки по умолчанию будут загружены:

- Использовать последние настройки для файла ввода — восстанавливает настройки, которые вы использовали в последний раз с этим конкретным файлом
- Использовать настройки по умолчанию — применяет ваши сохраненные пользовательские настройки по умолчанию
- Настройки по умолчанию Earth — сбрасывает все параметры к заводским значениям

Нажмите кнопку «Сохранить текущие значения в качестве значений по умолчанию» внизу раздела 4, чтобы сохранить текущие параметры Земли и матрицу «Допустимые взаимодействия»; кнопка «Сопоставить» в разделе 3 позволяет сохранить цель, дату оценки и конфигурацию предикторов. Обе настройки сохраняются для каждого проекта и цели в базе данных проекта.

8 Подгонка модели Земли

8.1 Кнопка «Fit»

В разделе 5 боковой панели расположена кнопка «Fit Earth Model» вместе с полем «Random seed». Это начальное значение обеспечивает воспроизводимость распределения блоков при перекрёстной проверке — в раскрывающемся списке отображаются последние 10 использованных начальных значений, а использованное значение фиксируется на вкладке «Earth Output» после подгонки. Перед подгонкой earthUI проверяет вашу конфигурацию — необходимо выбрать целевую переменную и включить как минимум один предиктор. В режимах «Оценка» и «Рынок» столбцы «Широта» и «Долгота» округляются, а срок реализации рассчитывается на основе даты оценки и даты заключения договора (если указана).

8.2 Подгонка модальной кривой & Trace Вывод

При нажатии кнопки «Подгонка» появляется тёмное модальное окно, в котором отображается:

- Заголовок — «Подгонка модели Земли» с таймером, отсчитывающим секунды, и кнопкой «Прервать», которая отменяет фоновую подгонку
- Журнал трассировки — прокручиваемая область текста с моноширинным шрифтом, отображающая вывод в режиме реального времени от функции `earth()` : информацию о наборе данных, ход прямого прогона, блоки перекрёстной проверки и время завершения
- Цветовая кодировка — линии по умолчанию — зелёные, блоки перекрёстной проверки — жёлтые, ошибки — красные

earthUI выполняет подгонку моделей асинхронно с помощью `callr::r_bg()`, который запускает вычисления в отдельном процессе R. Это позволяет приложению оставаться отзывчивым во время длительных процессов подгонки. Наблюдатель с периодичностью опроса 300 мс считывает вывод из фонового процесса и передает его в журнал трассировки.

По завершении подгонки появляется сообщение «Готово за X,X секунд», в модальном окне появляется кнопка закрытия (X), а фоновый слой удаляется, чтобы вкладки с результатами могли отображаться по требованию. Затем модальное окно автоматически закрывается, как только завершится отображение видимых выводов — либо вы можете закрыть его раньше.

В случае успешного выполнения на кнопке «Подгонка» появляется зелёная галочка. Если происходит ошибка, сообщение об ошибке отображается в журнале трассировки и появляется уведомление.

8.3 Журнал тренировок

Каждый запуск (успешный или неудачный) записывает файл журнала в вашу папку вывода:

- Имя файла : `<datafile>_earth_log_<YYYYMMDD_HHMMSS>.txt`
- Содержимое : временная метка, весь вывод трассировки процесса подгонки и любые сообщения об ошибках
- Расположение : папка вывода активного проекта

Совет: Если `callr` не установлен, earthUI переключается на синхронную подгонку с отображением простой полосы прогресса. Установите `callr`, чтобы получить полный набор возможностей по выводу трассировок.

9 Загрузка данных

После подгонки раздел боковой панели № 6 — под названием «Скачать оценочные целевые переменные & и остатки» в общем режиме и «Скачать оценочные цены продажи & и остатки» в режимах оценки и рынка — загружает файл Excel с прогнозами и диагностикой. Этот вывод используется на этапе 7 (RCA) для присвоения рейтинга CQA (Condition-Quality-Appeal — состояние-качество-привлекательность) объекту оценки. Вывод отсортирован по residual_sf и cqa_sf, чтобы помочь вам оценить, какое место занимает объект оценки в рейтинге. Если модель хорошего качества, то объекты недвижимости должны быть ранжированы от наименее привлекательных до наиболее привлекательных на основе остаточных признаков, которые не вошли в регрессионную модель. Среднее значение должно быть приблизительно равно 0, нижняя половина — отрицательные значения, а верхняя половина — положительные значения. Вы должны найти дома самого низкого качества, или «объекты для ремонта», в нижней части рейтинга, а самые привлекательные дома — в верхней. Обычно бывают исключения в виде аномалий, таких как обращение взыскания, продажи с убытком, продажи в рамках наследственного процесса и срочные продажи, необходимые в связи со сменой работы или по другим причинам. Изучение таких аномалий обычно позволяет выявить вескую причину ценовой аномалии.

Надпись на кнопке меняется в зависимости от режима работы:

- Общие сведения/Рынок: Скачать вывод (Excel)
- Оценка: Скачать промежуточный вывод (Excel)

Формат имени файла: <datafile>_modified_<YYYYMMDD_NHMMSS>.xlsx.

9.1 Столбцы вывода

Для каждой целевой переменной добавляются следующие столбцы:

Столбец	Описание
est_<target>	Прогноз по модели, например, est_sale_price (1 десятичное место)
residual	Разница между фактическим и прогнозным значением (1 десятичное место)
cqa	Оценка состояния, качества и привлекательности (2 десятичных места, диапазон 0–10)
residual_sf	Остаток, деленный на жилую площадь (если указано, 1 знак десятичной)
cqa_sf	Показатель CQA, рассчитанный на основе ранжирования по остатку_sf (2 знака десятичной)
<variable>_contribution	Значение вклада по каждой g-функции (1 dp, по одному столбцу на g-функцию)
basis	Вклад значения пересечения, одинаковый для всех объектов недвижимости (1 dp)
calc_residual	Фактическое значение минус (базис + все вклады) — столбец верификации (1 dp)

В моделях с несколькими целями добавляется суффикс «_<i>», чтобы различать столбец для каждой цели.

9.2 Упорядочивание столбцов & Excel Формат

Столбцы ранжирования размещаются в самом левом положении в выводе в следующем порядке: `residual_sf`, `cqa_sf`, `residual`, `cqa`. Это упрощает просмотр отсортированного списка и определение места, которое занимает объект оценки в распределении CQA.

К этим столбцам применен формат чисел в Excel:

Столбец	Формат	Пример
<code>residual_sf</code>	Числовой, 2 знака после запятой	12,34
<code>cqa_sf</code>	Числовой, 2 знака после запятой	7,25
<code>residual</code>	Числовой, 0 знаков после запятой	15 200
<code>cqa</code>	Числовой, 2 знака после запятой	6,80

9.3 Результаты CQA: балл

балл

Балл CQA (Condition-Quality-Appeal) ранжирует остаток каждой строки по отношению ко всем остальным остаткам. Для данной строки балл CQA представляет собой процент строк с остатком меньшего значения, умноженный на 10. В результате получается шкала от 0 до 10, где:

- Высокий показатель CQA ($\approx 9-10$) — объект недвижимости был продан значительно дороже, чем прогнозировала модель (большой положительный остаток)
- Низкий показатель CQA ($\approx 0-1$) — недвижимость была продана значительно дешевле, чем прогнозировалось (большой отрицательный остаток)
- $CQA \approx 5$ — остаток близок к медиане всех остатков

Если обозначен столбец « `living_area` », сайт `cqa_sf` предоставляет аналогичный рейтинг, основанный на остатках стоимости на квадратный фут.

9.4 Сортировка & Строка объекта оценки

В режимах «Оценка» и «Рынок» аналогичные строки сортируются по показателю « `residual_sf` » в порядке убывания (или по показателю « `residual` », если жилая площадь не указана). Строка объекта оценки (строка 1) остается на первой позиции и не подвергается сортировке. В общем режиме строки экспортируются в исходном порядке.

Такая сортировка в сочетании с крайними левыми столбцами рейтинга позволяет оценить быстро просматривать аналоги от тех, стоимость которых была завышена в наибольшей степени, до тех, стоимость которых была занижена в наибольшей степени, и определить, где находится присвоенный объекту оценки балл CQA в рамках данного распределения.

Перед вызовом функции « `predict()` » уровни факторов в данных прогнозирования приводятся в соответствие с данными обучения. Строки с неизвестными уровнями факторов будут давать прогнозы с значением «NA».

Совет: Столбец « `calc_residual` » всегда должен равняться значению « `residual` » (с учетом округления). Если эти значения значительно различаются, это указывает на проблему в вычислениях, которую стоит изучить.

9.5 mgcvUI Auto-Export

При каждой успешной подгонке с помощью $\text{degree} \leq 2$, `earthUI` автоматически сохраняет полный объект результата в виде файла `.rds` в папку вывода проекта. Имя файла соответствует шаблону `<datafile>_earthUI_result_<YYYYMMDD_HHMMSS>.rds`.

Этот файл можно загрузить с помощью модуля постобработки `mgcv` — входящего в пакет `earthUI` и запускаемого командой «`earthUI::launch_mgcv()`» — с помощью команды «`readRDS()`». Процедура GAM использует координаты узлов и базисные функции модели Земли в качестве исходных точек для сглаживающих слагаемых GAM, что обеспечивает плавный переход от моделирования по методу MARS к моделированию по методу GAM.

Модели с `degree > 2` пропускаются, поскольку `mgcvUI` поддерживает только парные взаимодействия. При запуске `earthUI` в режиме Trilogy (из пользовательского интерфейса Trilogy) появляется дополнительный раздел «Lock Model» «Вывод» позволяет зафиксировать конкретную подгонку `.rds` — вместе с её базисом предикторов и общими настройками CQA — в качестве канонической модели `earth`, которую импортируют `glmnetUI` и `mgcvUI`.

9.6 glmnetUI: импорт

Тот же файл `.rds`, сохраненный для процедуры GAM, также можно импортировать в модуль постобработки эластичной сети — он также входит в пакет `earthUI` и запускается с помощью команды `earthUI::launch_glmnet()`. В приложении `glmnet` перейдите к разделу 2: «Импорт из `earthUI`» и с помощью кнопки «Обзор» выберите файл `.rds`.

`glmnetUI` использует стандартный подход «`earth::model.matrix()`»: базисные функции модели `earth` (шарниры, взаимодействия и линейные члены) становятся столбцами матрицы дизайна `glmnet`. Затем `glmnet` выполняет регуляризованную регрессию по этим базисным столбцам, выбирая и сокращая их с помощью `lasso`/эластичной сети. Таким образом, создается адаптивное построение базиса в `earth` с регуляризацией `glmnet`.

9.6.1 Основные преимущества

- Интерпретируемые взаимодействия : взаимодействия на основе «шарниров» в модели `Earth` (включая трёхсторонние члены) имеют чёткое геометрическое значение, в отличие от встроенных взаимодействий типа «векторное произведение» в `glmnet`, которые трудно объяснить в суде или при проведении аудита.
- Автоматическая нелинейность : функции шарниров `earth` позволяют учитывать нелинейные зависимости, которые стандартная модель `glmnet` с исходными предикторами не смогла бы учесть.
- Отбор переменных : `glmnet` может дополнительно прореживать базис `earth`, исключая термины шарниров, которые не вносят вклада.

9.6.2 Важно: столбец «Вес» должен соответствовать

Warning

Если `earthUI` использует столбец весов (например, для исключения строки с весом = 0), тот же самый столбец весов необходимо указать в настройках переменных `glmnetUI`. Если настройки весов различаются, количество строк не совпадёт, и базис `earth` не сможет быть добавлен к матрице модели `glmnet`.

9.6.3 Рабочий процесс

- 1 . Построить модель Земли в `earthUI` (любая степень).
- 2 . Результирующий файл «`.rds`» автоматически сохраняется в папке «вывод» проекта.
- 3 . В `glmnetUI` откройте раздел 2: «Import from `earthUI`» и выберите файл «`.rds`».
- 4 . Убедитесь, что в обоих приложениях используются один и тот же файл данных и один и тот же столбец весов используются в обоих приложениях.

- 5 . Настройте параметры glmnet в разделе 5 (матрица взаимодействий автоматически отключается при активном импорте из earthUI).
- 6 . Нажмите «Fit Model» (Подгонка модели). Столбцы базиса earth появятся на вкладках «Equation» (Уравнение), «Coefficients» (Коэффициенты) и «Contributions» (Вклады).

10 Расчеты RCA& Скачивание

Кнопка «Рассчитать корректировки по методу RCA» (&) появляется в секции 7 боковой панели и отображается только в режиме оценки после того, как модель была подгонена. Метод остаточных ограничений позволяет получить основанные на рыночных данных корректировки для каждого аналога относительно объекта оценки.

10.1 Открытие диалогового окна RCA

При нажатии на кнопку открывается небольшое модальное диалоговое окно со следующим содержанием:

- Тип балла — переключатели для выбора «CQA» или «CQA по SF» (опция «SF» доступна только в том случае, если указан столбец «living_area»). Если вы выберете «CQA на SF», то на основе оценки CQA, присвоенной объекту оценки, его остаточный балл будет равен жилой площади, умноженной на остаточную площадь, соответствующую заданной оценке CQA_SF.
- Значение CQA — ввод оценки CQA объекта оценки (0,00–10,00). Этот балл отражает ваше мнение о том, где объект оценки находится в распределении качества аналогов.
- Кнопка «Generate» — вычисляет вывод RCA и сохраняет его в формате Excel. Кнопка остается неактивной до тех пор, пока не будет введено допустимое значение балла.

10.2 Интерполяция баллов CQA

При нажатии кнопки «Generate» earthUI интерполирует остаточную величину объекта оценки на основе аналогичных пар CQA/остаточных величин:

1. Баллы CQA и остатки аналогов отсортированы по CQA в порядке возрастания
2. Линейная интерполяция (stats::approx()) преобразует введенное вами значение CQA значение на остаток
3. При использовании показателя CQA на квадратный фут остаток на квадратный фут преобразуется обратно в общий остаток путем умножения на жилую площадь объекта оценки
4. Расчетная оценка объекта оценки вычисляется следующим образом: прогноз модели + интерполированный остаток

10.3 Столбцы вывода

Файл RCA в формате Excel (<datafile>_adjusted_<YYYYMMDD_NHMMSS>.xlsx) содержит все промежуточные столбцы вывода, а также:

Колонка	Описание
subject_value	Прогноз модели + интерполированное отклонение (только строка 1)
subject_cqa	Введенный вами балл CQA (только строка 1)
<variable>_adjustment	Вклад испытуемого за вычетом вклада контрольной группы (по g-функции)
residual_adjustment	Отклонение испытуемого за вычетом отклонения контрольной группы
net_adjustments	Сумма всех поправок (вклад + отклонение)
gross_adjustments	Сумма абсолютных значений всех корректировок
residual_pct	Корректировка остатка стоимости как доля от цены продажи

Колонка	Описание
net_adj_pct	Чистые корректировки как доля от цены продажи
gross_adj_pct	Валовые корректировки как доля от цены продажи
adjusted_sale_price	Цена продажи + чистые корректировки

В столбцах «Корректировки» для каждого аналога указано, в какой степени модель характеризует разницу с каждой переменной. «adjusted_sale_price» — это цена продажи аналога после применения всех корректировок, вычисленных по модели, — набор скорректированных цен продажи, которые должны группироваться вокруг оценочной стоимости объекта оценки.

10.4 Вкладка «Корректировки RCA»

После вычисления значения RCA на вкладке «RCA Adjustments» на главной панели отображаются гистограммы распределений показателей Residual Adj %, Чистая корректировка % и Валовая корректировка % по аналогам (по одному набору на каждый целевой объект для моделей с несколькими целевыми объектами), что позволяет быстро визуально оценить, насколько плотно сгруппированы корректировки.

Совет: Балл CQA равный 5,00 означает, что объект оценки находится на медиане среди аналогов. Баллы выше 5 указывают на то, что качество объекта выше среднего; ниже 5 — ниже среднего.

11 Таблица сравнения продаж

11.1 Обзор & Цель

Таблица сравнения продаж представляет собой рабочую книгу Excel, формируемую в режиме оценки (раздел 8 боковой панели) после расчета корректировок RCA. В ней объект оценки представлен наряду с выбранными аналогами в формате структурированной таблицы, с использованием формул Excel, которые автоматически рассчитывают корректировки и скорректированную цену продажи для каждого аналога.

Эта таблица предназначена для рабочего файла оценщика — она объединяет корректировки, полученные на основе регрессионного анализа из модели «Земля», с редактируемыми ячейками, в которых оценщик может распределить остаток CQA по конкретным характеристикам объектов недвижимости. Имя вывода — SalesGrid_<YYYYMMDD_HHMMSS>.xlsx.

11.2 Модальное окно выбора аналогов

При нажатии кнопки «Сгенерировать таблицу продаж & Скачать» открывается модальное диалоговое окно, в котором вы выбираете, какие аналоги включить:

- Рекомендуемые аналоги отмечены по умолчанию. Это аналоги с валовой корректировкой ниже 25% от цены продажи, отсортированные по сроку продажи в порядке возрастания (сначала самые последние продажи) и с ограничением в 30.
- Дополнительные аналоги перечислены ниже; по умолчанию они не отмечены. Их процент валовой корректировки составляет 25% или более, но они по-прежнему доступны для выбора (отображается до 50 объектов, отсортированных по проценту валовой корректировки в порядке возрастания).
- Можно выбрать не более 30 аналогов, что даст до 10 листов (по 3 аналога на лист).

После подтверждения выбора выбранные аналоги сортируются по проценту валовой корректировки в порядке возрастания, благодаря чему наиболее похожие аналоги отображаются на листе 1.

11.3 Макет таблицы

Каждый лист имеет макет из 20 столбцов, в которых размещаются данные по объекту оценки и 3 аналога. Столбцы для каждого объекта:

Сущность	Столбцы (5 на каждую сущность)
Объект оценки	Метка, Фактическое значение 1, Фактическое значение 2, Фактическое значение 3, Вклад в стоимость (VC)
Каждый аналог	Фактическое значение 1, Фактическое значение 2, Фактическое значение 3, Вклад в стоимость (VC), Корректировка

Строки сверху вниз:

1. Название — название листа (например, «Промежуточные аналоги продаж Таблица — Лист 1 из 3»)
2. Заголовки — «Объект оценки» и заголовки столбцов аналогов с адресом
3. Адрес — полный адрес объекта недвижимости
4. APN | MLS # | DOM | Subj.Prox — номер участка, ID объявления, количество дней на рынке и расстояние по формуле Хаверсина (миль) от объекта оценки
5. Цена продажи | Уступки | Чистая цена продажи — формула расчета, уступки и формула цены продажи

- 6 . Заголовок строк переменных регрессии
- 7 . БАЗОВОЕ ЗНАЧЕНИЕ — пересечение с осью абсцисс модели
- 8 . Дата продажи | Вне рынка | На рынке — дата заключения договора, срок продажи и количество дней на рынке
- 9 . Сгруппированные строки (условные) — Местоположение, Площадь участка и/или Возрастные группы
- 10 . Строки с переменными модели — по одной строке на каждый предиктор (за исключением сгруппированных переменных)
- 11 . Пустая разделительная строка
- 12 . Строка заголовка остаточных характеристик
- 13 . CQA |Остаток — балл CQA + оставшаяся формула остатка
- 14 . Строки остаточных характеристик — именованные + пустые строки для записи оценщика ввод
- 15 . Общая стоимость VC / Чистая корректировка
- 16 . Чистая корректировка %
- 17 . Валовая корректировка %
- 18 . Скорректированная цена продажи — строка формулы
- 19 . Нижний колонтитул с указанием авторских прав

11.4 Цена продажи, уступки & Net Цена продажи

В строке 5 для каждого аналога указаны три значения:

- Цена продажи — цена продажи аналога из данных
- Льготы — значение из столбца, обозначенного как concessions (0, если не обозначено)
- Чистая цена продажи — формула Excel: Цена продажи– Льготы

В столбце «Объект оценки» для цены продажи указано «N/A» (поскольку она неизвестна) и, при наличии, сумма льгот.

11.5 Сгруппированные строки (Местоположение, Объект, Возраст)

Когда заданы определенные специальные типы и эти переменные появляются в модели, earthUI создает сгруппированные строки, объединяющие связанные переменные:

Долг.| Шир.| Площадь — Отображается , когда в модели присутствует любой из следующих элементов: longitude, latitude или area. Показывает фактические значения для каждой составляющей переменной, совокупный вклад в значение (сумма отдельных значений VC), а также для аналогов — корректировку (— совокупного значения VC аналога по сравнению с совокупным значением VC аналога). Оформлено с светло-голубым фоном в ячейках VC.

Размеры объекта «|» — Отображаются, когда в модели присутствуют элементы « lot_size » или « site_dimensions ». Имеют ту же структуру, что и группа «Location».

Фактический возраст | Эффективный возраст — Отображается, когда в модели присутствуют actual_age или effective_age. Имеет ту же структуру.

Переменные, используемые в сгруппированных строках, исключаются из строк с переменными отдельных моделей, расположенных ниже, что позволяет избежать двойного учета в итоговых значениях корректировок.

11.6 Строки с переменными модели

Ниже сгруппированных строк (или непосредственно под строкой «BASE VALUE», если сгруппированных строк нет) отображается по одной строке для каждого оставшегося предиктора модели:

- Фактические значения для объекта оценки и каждого аналога (фактические значения данных)

- Вклад значения (VC) — вклад g-функции в прогнозируемое значение для данной строки
- Корректировка (только для аналогов) — VC объекта оценки минус VC аналога

11.7 CQA | Остаточная величина & Остаточная характеристика Строки

В строке «Остаточная величина CQA» (|) отображается балл CQA для каждого объекта недвижимости, а также содержится формула для расчета оставшейся части остатка — той части, которая ещё не была отнесена к конкретным характеристикам. Формула выглядит следующим образом:

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

Ниже расположены строки с остаточными характеристиками — 5 именованных строк («Местоположение», «Вид/привлекательность», «Состояние», «Качество», «Прочее») плюс 6 пустых строк. Это ячейки ввода, в которые оценщик вводит вклады в стоимость для характеристик, не учтенных моделью. По мере того как оценщик вводит значения, значение формулы «Остаточная величина» автоматически уменьшается.

Для каждого аналога в столбце «Корректировка» содержится формула: характеристика объекта оценки VC — характеристика аналога VC.

11.8 Скорректированная цена продажи Формула

Строка «Скорректированная цена продажи» содержит формулы Excel:

- Объект оценки : Сумма всех ячеек «Вклад в стоимость» от «БАЗОВОЙ СТОИМОСТИ» до последней строки с остаточными характеристиками. Это отражает общий прогноз модели для данного объекта оценки плюс любые остаточные характеристики, заданные оценщиком.
- Сравнимые объекты : Чистая цена продажи (Net SP) + сумма всех ячеек «Корректировка» (Adjustment) от первой группированной или переменной строки до последней строки с остаточными характеристиками. Это даёт цену продажи аналога после всех корректировок, полученных в результате регрессионного анализа и введенных оценщиком.

11.9 Защита листа & Редактируемые ячейки

Каждый лист защищен от случайного изменения формул и данных. Защита блокирует:

- Все ячейки с формулами (чистая прибыль, оставшаяся сумма остатка, корректировки, скорректированная цена продажи и т. д.)
- Все ячейки с данными (адреса, цены продажи, фактические значения, вклады в стоимость по модели)
- Метки и заголовки

Единственными разблокированными (доступными для редактирования) ячейками являются поля ввода вклада остаточных признаков — ячейки в строке «Остаточные значения» (|) раздела «CQA», в которые оценщик вводит данные по разбивке. Они выделены светло-желтым фоном, что указывает на возможность их редактирования.

11.10 Работа с таблицей в Excel

1. Откройте файл в Excel или другом совместимом приложении для работы с электронными таблицами.

- 2 . Просмотрите корректировки регрессии — корректировки, полученные на основе модели, уже введены и заблокированы.
- 3 . Распределите остаточную величину — в строках с остаточными характеристиками (желтые ячейки) введите свою оценку того, какая часть оставшейся остаточной величины приходится на каждую характеристику (местоположение, вид, состояние, качество и т. д.). Формула «Оставшаяся остаточная величина» будет уменьшаться по мере распределения.
- 4 . Проверьте «Скорректированную цену продажи» — эта формула автоматически обновляется по мере ввода значений остаточных характеристик.
- 5 . Несколько листов — если вы выбрали более 3 аналогов, переключайтесь между листами. В левом столбце каждого листа указан один и тот же объект оценки с 3 различными аналогами.

Совет: Цель состоит в том, чтобы распределить остаток СQA по строкам признаков до тех пор, пока показатель «Остаточный остаток» не станет близок к нулю. Это демонстрирует, что можно учесть разницу между регрессионной оценкой каждого аналога и его фактической ценой продажи.

12 Скачивание отчетов

После подгонки модели формирование отчёта представляет собой двухэтапный рабочий процесс в Quarto: «Генерация отчёта Quarto» (раздел 9 в боковой панели в режиме оценки, 7 — в остальных случаях) создаёт автономный набор отчётов, а «Конвертация отчёта Quarto» (раздел 10) преобразует его — или любой другой файл Quarto — в конечные форматы.

12.1 Создать отчёт Quarto

Нажатие кнопки «Generate Quarto Report» приводит к созданию автономного набора Quarto в папке вывода проекта по адресу `<base>_qmd/`, содержащего заполненный исходный файл `.qmd`, все предварительно сгенерированные графические ресурсы (PNG и PDF), данные отчёта и шаблон справочного документа в формате Word. На этом этапе рендеринг не происходит — набор можно редактировать вручную, объединять с другими исходными файлами Quarto с помощью операторов `include` или просматривать в автономном режиме с помощью программы `quarto preview`.

12.2 Конвертировать отчёт Quarto

В разделе 10 представлен выбор файлов (по умолчанию — последний сгенерированный набор текущего проекта, но вы можете выбрать любой файл `.qmd`) и флажки для выбора формата:

- HTML — файл `.html` с отображением математических формул с помощью KaTeX, встроенными изображениями и оглавлением. Использует тему Flatly Bootstrap и шрифт Roboto Condensed.
- Word — файл `.docx`, подходящий для редактирования и распространения. Включает оглавление, нумерацию страниц и использует пользовательский шаблон ссылок для единообразного стилизового оформления.
- PDF — набран с помощью LuaLaTeX для получения вывода профессионального качества. Размер бумаги соответствует настройкам локали (Letter или A4). Включает страницы в альбомной ориентации для больших матриц взаимодействия и использует шрифт Roboto Condensed и Latin Modern Math для уравнений. Вариант в формате PDF предлагается только при обнаружении дистрибутива LaTeX.

Нажмите «Конвертировать», чтобы выполнить рендеринг выбранных форматов. Рендеринг выполняется в фоновом режиме — в модальном диалоговом окне отображается прошедшее время и ход выполнения Quarto, при этом приложение продолжает реагировать на действия пользователя. По завершении процесса появляется уведомление с указанием местоположения файла.

Все операции (подгонка модели, загрузка данных, расчеты RCA, создание таблицы продаж и формирование отчетов) регистрируются с указанием меток времени начала и окончания, а также продолжительности выполнения в файле `<datafile>_earthui_log.txt`, расположенном в папке вывода, для целей устранения неполадок и мониторинга производительности.

12.3 Содержание отчёта

Каждый сгенерированный отчёт содержит следующие разделы:

1. Описание набора данных — количество наблюдений, целевая переменная (переменные), используемые предикторы, категориальные предикторы
2. Спецификация модели — степень, статус перекрёстной проверки, количество членов, предикторы в модели

- 3 . Допустимые взаимодействия (только для степени ≥ 2) — матрица взаимодействий с галочками, отформатированная для данного формата вывода
- 4 . Сводка результатов — R^2 , GRSq, GCV, RSS и CV R^2 (если была включена перекрёстная проверка)
- 5 . Уравнение модели — полное уравнение модели Земли в нотации LaTeX нотация
- 6 . Коэффициенты & Базисные функции — таблица всех слагаемых, их коэффициентов и базисных функций, которые их определяют
- 7 . Важность переменных — гистограмма и таблица с ранжированием оценок важности предикторов
- 8 . Вклады g-функций — графики для каждой g-функции: линейные для одномерных членов, 3D-перспективные и контурные графики для двумерных взаимодействий
- 9 . Матрица корреляций — тепловая карта корреляций предикторов
- 10 . Диагностика — графики «Остатки против подгоночных значений», «Нормальный Q-Q» и «Фактические против прогнозируемых
- 11 . Разложение ANOVA — таблица, показывающая вклад каждой базисной функции в RSS
- 12 . Вывод Earth — исходный текстовый вывод из `earth::summary()` , включая детали прохода обрезки

В случае моделей с несколькими ответами разделы 4–8 и 10 повторяются для каждой целевой переменной.

Совет: Для создания отчетов требуется пакет R «quarto» и установленная среда Quarto. Для вывода в формате PDF также необходимо установить дистрибутив LaTeX (например, с сайта `tinytex::install_tinytex()`) — в противном случае флажок «PDF» будет скрыт.

13 Модули постобработки: эластичная сеть и GAM

Пакет earthUI содержит два модуля постобработки — процедуру «эластичная сеть», основанную на glmnet, и процедуру GAM, основанную на mgcv. Каждая из них подробно рассматривается в отдельной статье в этом выпуске; в данном разделе объясняется, какое место они занимают в рабочем процессе earthUI и как эти три процедуры взаимодействуют друг с другом.

13.1 Почему стоит поддержать Fits?

Модель «Earth» выдает итоговое значение. Модули постобработки предназначены для его подтверждения: две дополнительные оценки одного и того же объекта оценки, полученные с помощью методологически разных движков и подогнанные к одинаковым проектным данным. Когда переоценка базы «Earth» с помощью метода «эластичная сеть» и независимое сглаживание оценки GAM приводят к практически одинаковым значениям, вывод оценщика больше не опирается на допущения одной единственной модели — что является значимым преимуществом в контексте экспертизы и судебных разбирательств. Когда они расходятся, само это расхождение носит характер диагностики и, как правило, указывает на недостаток данных, чрезмерно гибкий термин или выпадающий аналог, который стоит пересмотреть. Ни один из модулей не имеет приоритета над моделью Earth; они являются вспомогательными доказательствами, намеренно представленными без присвоения весов.

13.2 Запуск и общее состояние

Каждая процедура представляет собой отдельное локальное приложение Shiny:

- earthUI::launch_glmnet() — приложение эластичной сети, на порту 7879.
- earthUI::launch_mgcv() — приложение GAM, на порту 7880.

Все три приложения обращаются к одному и тому же дереву проекта и одной и той же базе данных настроек, поэтому проект, открытый вами в earthUI, немедленно становится доступным в модулях — вместе с переносимыми полями, которые earthUI сохраняет для него: «Ответ» (цель), «Дата оценки», тип и значение RCA «subject-CQA», а также обозначение жилой площади. Вам не нужно вводить их заново; модули считывают их, поэтому одни и те же значения передаются во все три подгонки.

13.3 Процедура «эластичная сеть»

Процедура glmnet импортирует базисные функции модели «земля» — её шарниры, взаимодействия и линейные слагаемые — в качестве столбцов матрицы плана эксперимента и заново оценивает их с применением регуляризации Lasso/эластичная сеть (см. раздел «glmnetUI Import» выше для ознакомления с механизмом работы и предостережением относительно столбца весов). В результате сохраняются интерпретируемые и геометрически значимые члены модели earth, в то время как регуляризатор уменьшает или исключает те, которые не оправдывают себя. См. статью о glmnetUI в этом выпуске для полного описания.

13.4 Процедура GAM

Процедура mgcv действует в обратном направлении: вместо повторного использования базиса модели «Earth» она проводит подгонку новых сглаженных членов с штрафными коэффициентами, используя координаты узлов модели «Earth» в качестве начальных точек (см. раздел «mgcvUI Auto-Export» выше). Поскольку GAM заново выводит функциональные формы в рамках иной парадигмы сглаживания, её согласие с моделью Земли является

подлинным независимым подтверждением, а не просто повторением. Поддерживаются модели Земли с $\text{degree} \leq 2$. Полное описание см. в статье о `mgcvUI` в этом выпуске.

13.5 Практический алгоритм действий

- 1 . Сначала подгоните и уточните модель «earth» в `earthUI`, пока не будете ею удовлетворены. Каждая успешная подгонка ($\text{degree} \leq 2$) автоматически экспортирует файл `.rds`, который используют модули.
- 2 . Запустите процедуры `glmnet` и `mgcv` для того же проекта, импортируйте результат `earth` и выполните подгонку. Работайте с одной процедурой за раз — подбор занимает много времени, а процедуры взаимодействуют через общий проект, поэтому разумно работать последовательно, а не параллельно.
- 3 . Свободно повторяйте итерации. Каждое приложение сохраняет свои результаты при переключении между ними, и ничто не заставляет вас создавать отчёты по ходу работы.
- 4 . Только когда все три результата вас устроят, вернитесь к каждой процедуре и сгенерируйте для неё отчёт Quarto. Каждый метод регистрирует свою подгонку и вывод о значении в записи «трилогии» проекта, благодаря чему все три отчёта можно отследить как относящиеся к одному сравнительному запуску.

Совет: Модули не являются обязательными на любом уровне: собственный рабочий процесс `earthUI` — подгонка персонала, анализ причин и следствий (RCA), таблицы продаж, отчёты — является полноценным и без них. Прибегайте к ним, когда выводы требуют подтверждения: в случае сложной темы, узкого сегмента рынка или задания, которое, вероятно, будет подвергнуто проверке.

14 Демо-набор данных: Оценка _ 1.csv

earthUI содержит демонстрационный набор данных MLS для ознакомления с рабочим процессом оценки. Загрузите его программным способом с помощью:

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

Либо в приложении Shiny скопируйте его в папку « in/» проекта (или прикрепите в качестве исходного файла данных при создании проекта) и выберите его в разделе 2 («Импорт данных»).

14.1 Описание

Файл содержит данные о 1 501 сделке по продаже жилой недвижимости, а также об одном объекте оценки в первой строке (всего 1 502 строки), полученные в результате моделирования экспорта данных из системы MLS. Данные отражают продажи односемейных домов на рынке, охватывающем несколько регионов, с различными размерами, возрастом и местоположением объектов оценки.

Это не реальные данные, но они основаны на реальном районе в Северной Калифорнии. Вся идентифицирующая информация была изменена или удалена.

14.2 Столбцы

Столбец	Тип	Специальный тип	Описание
вес	числовой	—	Вес наблюдения (0 = исключить из подгонки)
id	числовой	отображение_ только	Внутренний идентификатор записи
объект недвижимости_ id	числовой	отображение_ только	Идентификатор объекта недвижимости в системе MLS
объявление_ id	символьный	отображение_ только	Номер объявления в системе MLS
участок_ номер	символьный	отображение_ только	Номер участка, присвоенный оценщиком округа (APN)
улица_ адрес	символьный	отображение_ только	Адрес объекта недвижимости
город_ название (рассчитывает_)			
—	числовой	—	Количество дней от даты заключения договора до даты оценки (рассчитано заранее)
coe_	Дата	отображение_ только	Дата закрытия эскроу
listing_ статус	символ	отображение_ только	Статус объявления (например, «Продано»)
цена_ аренда	числовой	(целевая)	Цена продажи — переменная отклика
аренда	числовой	—	Ежемесячная арендная плата (для моделей с несколькими целями)
список_ цена	числовой	отображение_ только	Цена объявления
исходная_ список_ цена	числовой	отображение_ только	Исходная цена объявления
жилая_ кв. футов	числовой	жилая_ площадь	Общая жилая площадь в квадратных футах
спальни_ всего	целое число	—	Количество спален
ванные_ всего	числовой	—	Общее количество ванных комнат (например, 2,5 = 2 полные + 1 половина)
участок_ размер	числовой	участок_ размер	Площадь участка в квадратных футах
площадь_ id	целое число	площадь	Идентификатор площади в системе MLS
площадь_ текст	символ	отображение_ только	Название площади
возраст	числовой	фактический_ возраст	Возраст объекта недвижимости в годах
год_ постройки	целое число	отображение_ только	Год постройки
широта	числовой	широта	Широта (округленная до 3 знаков после запятой для модели)
долгота	числовой	долгота	Долгота (округлённая до 3 знаков после запятой для модели)
широтаб	числовой	отображение_ только	Широта с полной точностью
долготаб	числовой	отображение_ только	Долгота с полной точностью
гараж_ парковочных мест	целое число	—	Количество парковочных мест в гараже
с плавающей запятой_ количество	целое число	—	Количество каминов
количество_ этажей_	числовой	—	Количество этажей
style	character	—	Архитектурный стиль
view	character	—	Тип вида (например, «Район», «Холмы»)
days_ on_ рынок	integer	dom	Количество дней на рынке
объявление_ date	Дата	объявление_ date	Дата размещения объявления

14.3 Рекомендуемое руководство по быстрому запуску

1. Запустите earthUI: `earthUI::launch()`
2. Создайте проект с целью «Для оценки» в мастере «New Project», указав файл `Appraisal_1.csv` в качестве исходного файла данных
3. В разделе 2 («Import Data») выберите « `Appraisal_1.csv` »
4. В качестве целевого файла выберите `sale_price`
5. Назначьте специальные типы, как показано в таблице выше
6. Включите предикторы: `sale_age` , `living_sqft` , `baths_total` , `lot_size` , `area_id` (в качестве фактора), `age` , `latitude` , `longitude` , `garage_spaces`
7. Установите степень равной 1, нажмите «Fit Earth Model» (Подгонка модели Земли)
8. Загрузите промежуточный вывод (Шаг 6), просмотрите рейтинг CQA
9. Вычислите корректировки RCA (Шаг 7) с баллом CQA~ 5,00
10. Сгенерируйте таблицу сравнения продаж (Шаг 8)

Совет: Демонстрационный набор данных — отличный способ ознакомиться со всеми возможностями earthUI перед импортом ваших собственных данных MLS. Указанные выше рекомендуемые предикторы, как правило, позволяют построить модель с коэффициентом R^2 выше 0,85.

15 Экспериментальная часть: Извлечение текста из Prolog & Вывод Правила

В версии 0.10.0 введена дополнительная и сугубо экспериментальная функция извлечения готовых к использованию в модели характеристик из описательных текстовых столбцов экспорта MLS — прежде всего из столбца «Публичные примечания» в столбец помещаются данные о характеристиках комнат или существующих конструкциях, а также замечания агента и разделенные запятыми списки характеристик, такие как характеристики комнат или существующие конструкции. В качестве механизма извлечения используется Prolog, а в частности — грамматики определённых предложений (DCG): правила, распознающие фразы типа «готово к заселению», «гараж на 3 машины» или «собственная солнечная батарея» в тексте примечаний, написаны в виде грамматических правил Prolog.

Эта функция пока находится в стадии разработки. Сегодня её можно частично использовать для извлечения информации — входящий в набор грамматика уже распознаёт полезный набор признаков, — однако для её адаптации к лексике вашей собственной системы MLS требуется умение читать и писать на языке Prolog, в частности, с использованием DCG. Если это не ваш случай, пропустите эту главу полностью: функция по умолчанию отключена, и остальные части earthUI работают точно так же и без неё.

Warning

Это экспериментальная функция, находящаяся в стадии активной разработки. Её грамматика извлечения не является ни полной, ни проверенной для оценки в производственных условиях — извлечённые столбцы необходимо сверять с исходным текстом, прежде чем использовать их в модели. Для эффективного использования этой функции требуются практические знания языка Prolog и DCG. Все её действия носят аддитивный и обратимый характер; если что-то работает некорректно, отключите переключатель в настройках и продолжайте работу в обычном режиме.

15.1 Описание функции

Объявления в системе MLS содержат большой объем информации, влияющей на стоимость, в виде свободного текста и в полях списка с разделителями. Этот необязательный шаг преобразует этот текст в обычные столбцы, которые можно использовать в регрессионном анализе, с помощью двух специальных ролей столбцов (глава 6):

- `remarks` Столбцы анализируются грамматикой Prolog DCG с формированием одного столбца-особенности на каждый извлеченный атрибут. Префикс столбца создается из первых букв имени исходного столбца: из выражения « `public_remarks` » получаются столбцы « `pr_*` », из выражения « `agent_remark` » получаются столбцы « `ag_*` » и т. д.
- Столбцы « `list` » (например, « `existing_structures = "Shed, Workshop"` ») разбиваются на столбцы « `one-hot` » `<col>_<value>` с значениями TRUE/FALSE. Эта часть реализована исключительно на языке R и не требует использования Prolog.

Набор грамматики (`inst/prolog/mls_remarks.pl`) в настоящее время извлекает, помимо прочего: уровень состояния с шкалой от 1 до 5 `condition_score`; флаги и типы видов из окон; флаги ремонта кухни и ванной; паркет и другие виды напольных покрытий; список и количество улучшений; места в гараже и площадь гаража; бассейн, спа и другие элементы внешней инфраструктуры; количество каминов; хозяйственные постройки (сарай, амбар, мастерская, касита, гостевой домик, питомник и т. д.) с указанием размеров, полученных из таких данных, как «40x50»; стоянки и хранилища для автофургонов; владение солнечными батареями по сравнению с арендой, аккумуляторные батареи; площадь участка (включая

пересчёт акров в квадратные футы); цены в контексте продажи, связанные с конкретным объектом оценки (снижение цены, взносы в ТСЖ, аренда или выкуп солнечных батарей, налог Мелло-Рус); характеристики местоположения; а также балл настроения по ключевым моментам и проблемам в формате « $[-1, 1]$ ».

15.2 Требования: vProlog (Trealla), не SWI-Prolog

Анализ комментариев осуществляется с помощью пакета vProlog R — автономного движка Trealla Prolog. Его нет на CRAN; установите его из репозитория r-universe автора, как показано в разделе «Установка earthUI и vProlog» в начале этой статьи (готовые бинарные файлы для Windows и macOS, исходный код по адресу github.com/wcraytor/vProlog). Системная установка Prolog (SWI-Prolog или иная) не требуется и не используется. Если vProlog отсутствует, разбор комментариев пропускается с выводом соответствующего уведомления, в то время как разделение на фрагменты list все равно выполняется. Вспомогательная функция vProlog_available() предоставляет отчёт о том, имеются ли как движок, так и входящий в набор набор грамматики.

15.3 Включение и запуск

1. Откройте «Настройки» (значок шестеренки) и установите флажок «Включить обработку столбцов Prolog / списка». По умолчанию этот переключатель выключен, сопровождается примечанием «расширенная функция — для пользователей, владеющих Prolog», настройка сохраняется для каждого проекта и отображается в диалоговом окне «Информация о проекте».
2. В таблице предикторов установите для соответствующих столбцов специальную роль «remarks» или «list» (каждая роль может присваиваться нескольким столбцам).
3. Воспользуйтесь разделом «Выполнить обработку Prolog», расположенным между разделами 4 и 5 на боковой панели. Он предлагает два шага:

1. Развернуть списки «&» — запускает грамматику DCG для каждого столбца «remarks» и выполняет «one-hot»-разбиение каждого столбца «list». Развертывание происходит инкрементально: обрабатываются только те столбцы, которые были отмечены после последнего запуска, а кнопка отключается, как только все отмеченные столбцы уже развернуты. Индикатор выполнения отслеживает ход разбора.

2. Выполнить правила деривации — применяются ваши собственные правила Prolog derive/2 к развернутым столбцам (см. ниже). Перед выполнением правила проходят синтаксическую проверку и проверку в режиме «сухого запуска».

15.4 Правила вывода

Для каждой строки данных earthUI проверяет столбцы этой строки на соответствие фактам cell(Name, Value) и собирает все решения ваших правил derive(Col, Val) — каждый отдельный Col становится новым столбцом, и для каждого столбца принимается первое найденное значение, поэтому за ним может следовать резервное условие. Вспомогательные предикаты (count_true/2, count_false/2, count_true_prefix/2, cell_or/3, is_na/1) поддерживают агрегацию и обработку значений NA, а факты drop/1 удаляют столбцы после выполнения правил. Например:

```
% 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).
```

Правила хранятся в файле `<project>_rules.pl` для каждого проекта, с начальным значением подразумеваемого шаблона, и редактируются с помощью функции «Редактировать правила деривации» ... в плавающем редакторе кода, который можно перетаскивать и который также можно закрепить на боковой панели. Кнопка «Проверить» проверяет правила (проверка синтаксиса, проверка загрузки и пробный запуск на первой строке ваших данных) без их выполнения.

15.5 Файл активных данных с версиями

После выполнения любого из этих шагов расширенная таблица данных записывается в папку ввода проекта как `<original_name>_<YYMMDD_NHMMSS>.xlsx` и становится активным файлом данных проекта — при повторном открытии проекта загружаются расширенные данные, поэтому расширения никогда не пересчитываются, и именно этот файл используется программами `glmnetUI` и `mgcvUI`. Исходные столбцы с комментариями автоматически исключаются из модели; сгенерированные столбцы не включаются автоматически в качестве предикторов. В результате разбора текста может быть сгенерировано много столбцов, поэтому те из них, которые вы (пока) не хотите использовать в регрессии, пометьте как «Latent» или отключите с помощью флага «x» (глава 6).

15.6 Программный интерфейс & Набор файлов

Весь конвейер экспортирован для пакетного использования: `execute_prolog_processing(df, remarks_cols, list_cols, rules_text)` координирует проходы, построенные на основе `remarks_to_columns()`, `split_list_column()`, `apply_user_rules()` и `validate_rules()`. Исходные коды Prolog, включённые в набор по адресу `inst/prolog/`, представляют собой: `mls_remarks.pl` (грамматика DCG и агрегация признаков), `mls_analyze.pl` (драйвер ввода-вывода файлов для пакетного синтаксического анализа), `mls_rules.pl` (вспомогательные предикаты, доступные для правил деривации) и `climate_regions_ca.pl` (ниже).

15.7 Связанный эксперимент: «Климатический регион» Исходные распределения:

Дополнительная экспериментальная установка позволяет изучить характеристики с небольшим объемом данных — те, которые регрессионная модель не может оценить, поскольку они встречаются лишь в нескольких аналогах (например, болотный охладитель). Три вспомогательные программы на языке R — `climate_region_for()`, `market_area_profile()` и `climate_feature_priors()` — классифицируют местоположение объекта (округ + город) по одному из семи признанных покупателями климатических регионов Калифорнии (практическое упрощение 16 климатических зон Калифорнийской энергетической комиссии 16 климатических зон для зданий) и возвращают порядковые априорные значения вклада для систем ОВКВ и средств охлаждения (отопление, кондиционер, болотный охладитель, вентилятор для всего дома, фильтрация) по регионам. Модуль Prolog `climate_regions_ca.pl`

отражает ту же классификацию для использования в правилах вывода; обе стороны генерируются из одного канонического исходного кода на языке R, поэтому они не могут расходиться. Априорные распределения предназначены для обоснования оценки характеристик и служат в качестве проверки правдоподобия оцениваемых коэффициентов — отнесение к регионам (особенно в случае округов, охватывающих несколько регионов) должно проверяться оценщиком.

15.7.1 За пределами Калифорнии: пакеты юрисдикций и модули рыночных зон

Калифорния — это первая юрисдикция, а не граница. earthUI используется во всем мире, и исходные данные о климате в настоящее время реорганизуются в виде пакетов по юрисдикциям: автономных таблиц фактов, сформированных на основе данных (определения зон, сопоставления административных единиц, априорные данные об объектах), привязанные к кодам стран и административных уровней проектной системы, благодаря чему один пакет может описывать штат США, федеральную землю Германии или небольшую страну в целом. Сопутствующий слой модулей рыночных зон фиксирует местную терминологию и уточнённые исходные данные для конкретной рыночной зоны.

Оба предназначены для создания в полевых условиях инженерами по оценке и оценщиками на их родном языке — и, если смотреть реалистично, с помощью моделей LLM: в мире не так много программистов на Prolog, поэтому полевые артефакты намеренно представляют собой таблицы, а не программы. Формат `rack` представляет собой документированную схему с полным примером (сам штат Калифорния), специально для того, чтобы такой помощник, как Claude, мог составить корректный файл `rack` на основе описания рынка простым языком; earthUI генерирует базовый код на Prolog из файла `rack`, и перед тем, как файл `rack` можно будет активировать, он должен пройти (проверки схемы, охвата, референциальной целостности и пробного запуска) должны пройти успешно, прежде чем пакет можно будет активировать. Каждый `rack`, созданный специалистами на местах, содержит обязательную информацию о происхождении — имя автора, дату и цитируемый авторитетный источник — так что за знаниями стоит ответственный специалист, а не инструмент составления, и в отчётах можно указывать точный `rack` и версию, использованные при составлении.

Совет: Практический подход: разверните примечания, объедините детализированные флаги `pr_*` в несколько производных значений или булевых переменных с помощью правил `derive/2`, `drop/1` исходные флаги, которые вы объединили, и оставьте в качестве потенциальных предикторов только производные столбцы. Это позволяет сократить список предикторов и упростить управление коллинеарностью.

References

- 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>.