Виды анализа и R Markdown

Баженов Владислав

25 10 2021

Разведочный анализ данных

Пример

Согласно определению, цель разведочного анализа данных – изучить данные и найти взаимосвязи, о которых ранее не было известно. Разведочный анализ данных:

- Изучает как могут быть связаны различные переменные.
- Полезно для обнаружения новых связей.
- Помогает формулировать гипотезы и управлять планированием будущих исследований и сбора данных.

Также одной из целей разведочного анализа данных является **обнаружение отклонений и аномалий** в данных.

Рассмотрим пример исследования (Doi: 10.21515/1990-4665-131-098), целью которого являлась разработка модели анализа стоимости недвижимости в условиях рынка недвижимости г. Краснодара. Задачи, необходимые для достижения поставленной цели, формулировались следующим образом:

— проведение разведочного анализа имеющихся данных на предмет выбросов и незначимых данных (при помощи построения линейных графиков и диаграмм рассеяния);
— проверка наличия возможных зависимостей между наблюдениями и между переменными (построение корреляционных матриц)

Иными словами, требовалось выявить возможные зависимости цены объекта недвижимости от определённых факторов.

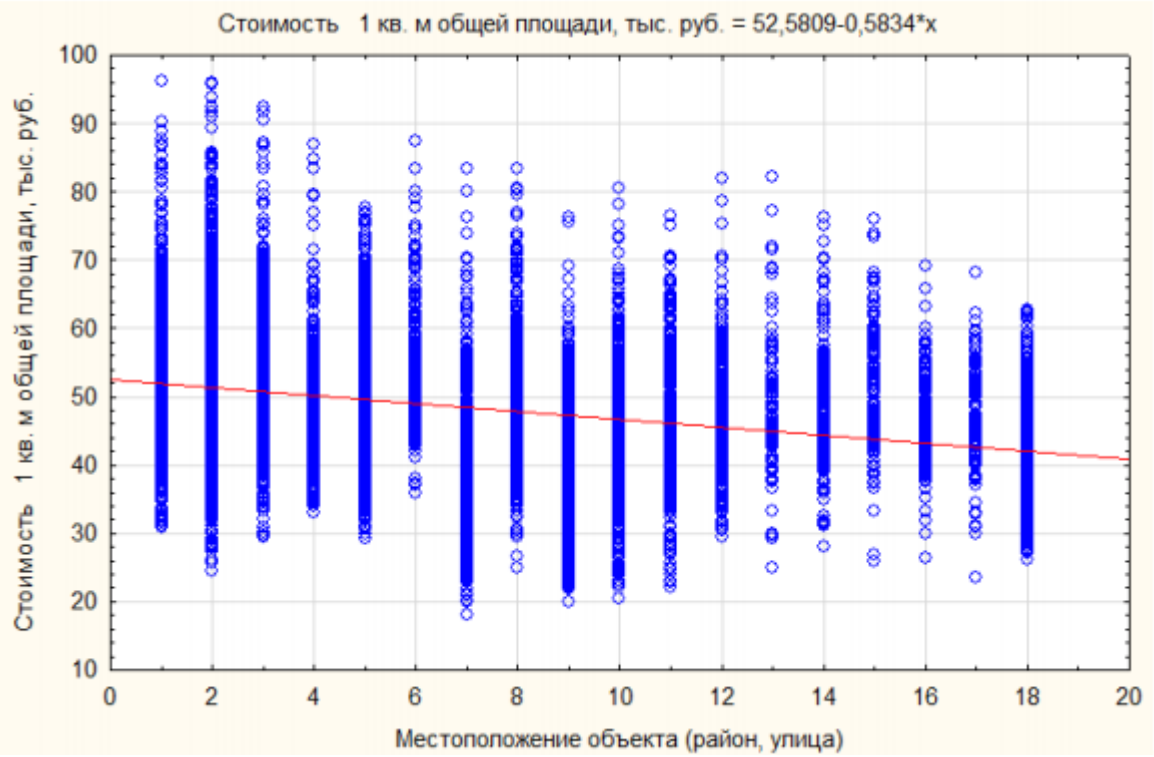
Исходные данные по квартирам были взяты с сайта Региональной энергетической комиссии – департамента цен и тарифов Краснодарского края. Исходный файл с данными содержал 8 переменных и 35653 наблюдения:

	Полные данные								
	1 Местопо-ложение объекта (район, улица)	2 Этажное располо-жение квартиры	3 Кол-во этажей в доме	4 Кол-во комнат	5 Общая площадь квартиры, кв.м.	6 Жилая площадь квартиры, кв.м.	7 Стоимость 1 кв. м общей площади, тыс. руб.	8 Средний номиналь-ный курс доллара США к рублю за период	9 Период
01.08.2011	1	5	10	1	60	20	48,3333333	28,77	01.08.2011
01.08.2011	1	3	5	1	30	18	73,3333333	28,77	01.08.2011
01.08.2011	1	1	4	1	56		37	28,77	01.08.2011
01.08.2011	1	3	4	1	38		43,4210526	28,77	01.08.2011
01.08.2011	1	1	9	1	35,5	17,5	54,9295775	28,77	01.08.2011
01.08.2011	1	5	16	1	40	20	47,5	28,77	01.08.2011
01.08.2011	1	1	3	1	38,1	18	44,6194226	28,77	01.08.2011
01.08.2011	1		7	1	31		37	28,77	01.08.2011
01.08.2011	1	5	5	1	41	20	51,2195122	28,77	01.08.2011
01.08.2011	1	1	4	1	36	19	38,8888889	28,77	01.08.2011
01.08.2011	1	7	9	1	32	18	56,25	28,77	01.08.2011
01.08.2011	1	6	9	1	42	25	52,3809524	28,77	01.08.2011
01.08.2011	1	3	7	1	56	22	53,5714286	28,77	01.08.2011
01.08.2011	1	15	16	1	40	20	45	28,77	01.08.2011
01.08.2011	1	6	7	1	55	24	43,6363636	28,77	01.08.2011
01.08.2011	1	1	3	1	33	18	51,5151515	28,77	01.08.2011

В названии наблюдений указан временной интервал: месяц и год. Список всех переменных представлен в таблице:

Номер	Переменная
1	Местоположение объекта (район, улица)
2	Этажное расположение квартиры
3	Количество этажей в доме
4	Количество комнат
5	Общая площадь квартиры, кв. м
6	Жилая площадь квартиры, кв. м
7	Стоимость 1 кв. м общей площади, тыс. руб.
8	Средний номинальный курс доллара США к рублю за период
9	Период

В рамках исследования была построена диаграмма рассеяния переменных «Стоимость 1 кв. м общей площади, тыс. руб.» и «Местоположения объекта», проведена нормализация данных путём избавления от явных выбросов:



В ходе исследования проведена перешифровка районов г. Краснодара от 1 до 18, где самым ценным районом (с шифром 1) будет Фестивальный микрорайон (ранее № 14), наименее ценным - ЛМР (ранее № 10):

1. ФМР;
2. Центр;
3. ЮМР;
4. ГМР;
5. ЧМР;
6. кинотеатр Аврора;
7. ЗИП;
8. КМР;
9. Российская;
10. 40 лет Победы;
11. ЖМР;
12. СМР;
13. Табачная фабрика;
14. Мосты;
15. завод Седина;
16. Авиагородок;
17. Школьная;
18. ЛМР.

На данном графике явно прослеживается линейная зависимость.

Выводы

Данное исследование использует методы **разведочного анализа данных**, чтобы обнаружить **выбросы и незначимые данные**, а также **взаимосвязи между переменными, о которых заранее неизвестно** (например, взаимосвязь между переменной, характеризующей местоположение объекта, и переменной, характеризующей стоимость единицы площади).

Механистический анализ данных

Пример

Согласно определению, цель механистического анализа данных – понять, какие именно изменения в переменных приводят к точным изменениям в других переменных. Механистический анализ данных применяется для:

- Применяется к простым ситуациям или ситуациям, которые хорошо моделируются детерминированными уравнениями.
- Обычно применяется к физическим или инженерным наукам.
- Например: биологические науки слишком «шумны» для использования механистического анализа.
- Часто единственный шум в данных – ошибки измерения.

Рассмотрим пример исследования (https://doi.org/10.3389/fevo.2015.00037), целью которого являлся анализ плотности населения в динамике популяции водорослей.

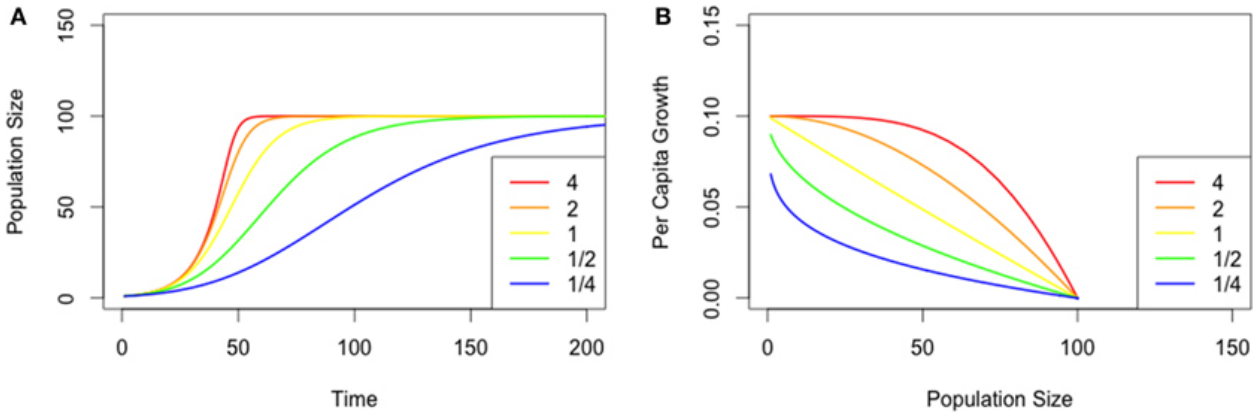
> Одним из фундаментальных принципов экологии является регулирование плотности населения, однако конкретный процесс, лежащий в основе моделей, описывающих, как темпы роста популяции замедляются по мере увеличения численности населения, ещё предстоит выяснить. Одна точка зрения утверждает, что модель динамики популяции одинакова для разных видов и не зависит от условий окружающей среды, тогда как другая заключается в том, что модель зависит от экзогенных и эндогенных процессов, действующих на популяцию.

Авторы статьи проводят исследование, изучающее динамику популяции водорослей «Chlamydomonas», которые выращивались в условиях градиента от низкой до высокой плотности питательных веществ.

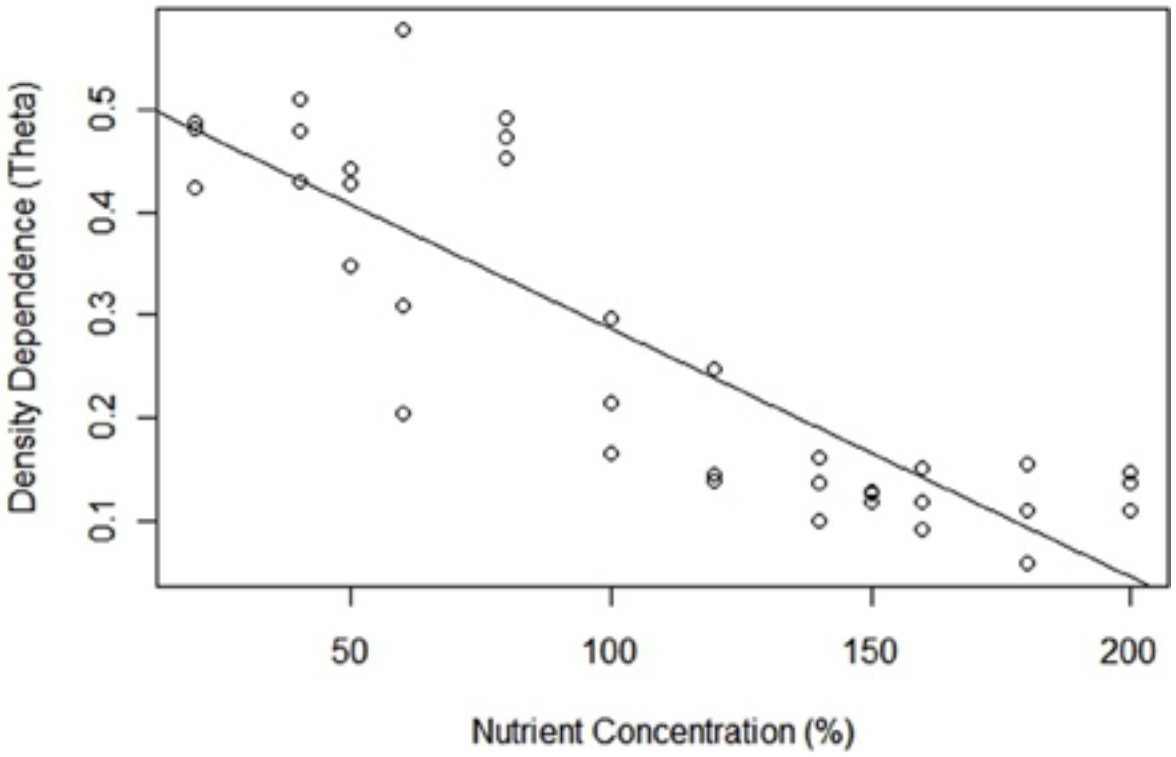
Динамика роста популяции исследовалась с помощью уравнения Рикера:

$$K_{t+1} = N_t e^{r(1-\frac{N_t}{K})^\theta}$$

где N_t - плотность населения в момент времени t, r - внутренняя скорость роста популяции, K - биологическая ёмкость среды. параметр θ отвечает за криволинейность в моделях роста численности населения в зависимости от численности популяции и характеризует интенсивность процесса регулирования плотности населения (рисунок «A», где показаны кривые с различными значениями θ):



В эксперименте исследователи выращивали одноклеточные зелёные водоросли Chlamydomonas reinhartti в средах с 12 различными уровнями концентрации питательных веществ. Среди прочих результатов было обнаружено, что экспериментально оценённый параметр θ из уравнения Рикера, с уровнем доверия в 95% имеет обратную зависимость от концентрации питательных веществ:



Выводы

Поскольку рассмотренное исследование имело дело с ситуацией, которая может быть смоделирована с помощью уравнения, а также обнаружило определённую связь между переменными с высоким уровнем доверия, можно считать, что исследование использовано механистический анализ данных.

Ссылка на репозиторий GitHub