

Работа с гео-данными и временными данными

Цифровая гуманитаристика 2024/2025

Г. А. Мороз

Международная лаборатория языковой конвергенции, НИУ ВШЭ

24.04.2025

Пространственные данные

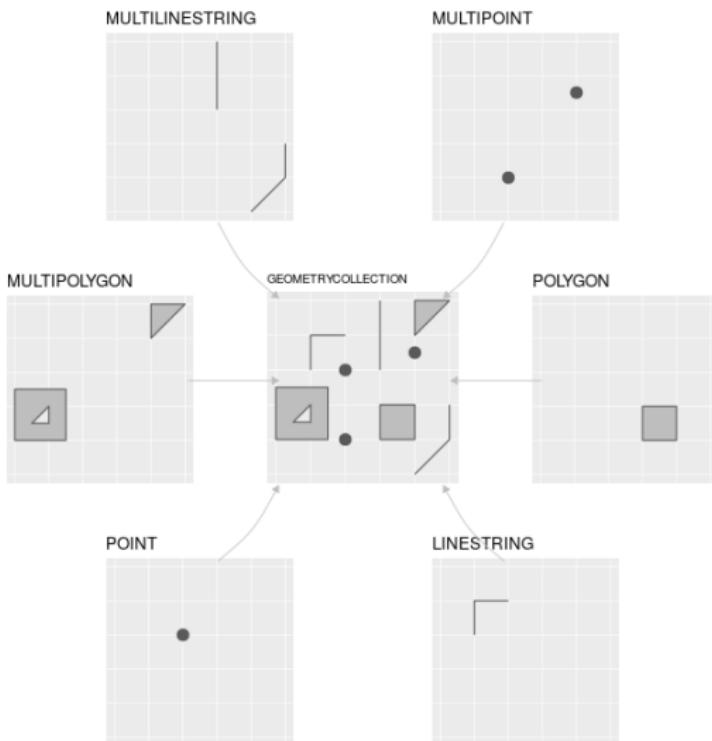
Анализ данных может включать

- сбор данных
- очистку данных и их предобработку
- визуализацию данных
- моделирование данных
- дескриптивный анализ
- предиктивный анализ
- машинное обучение
- ...

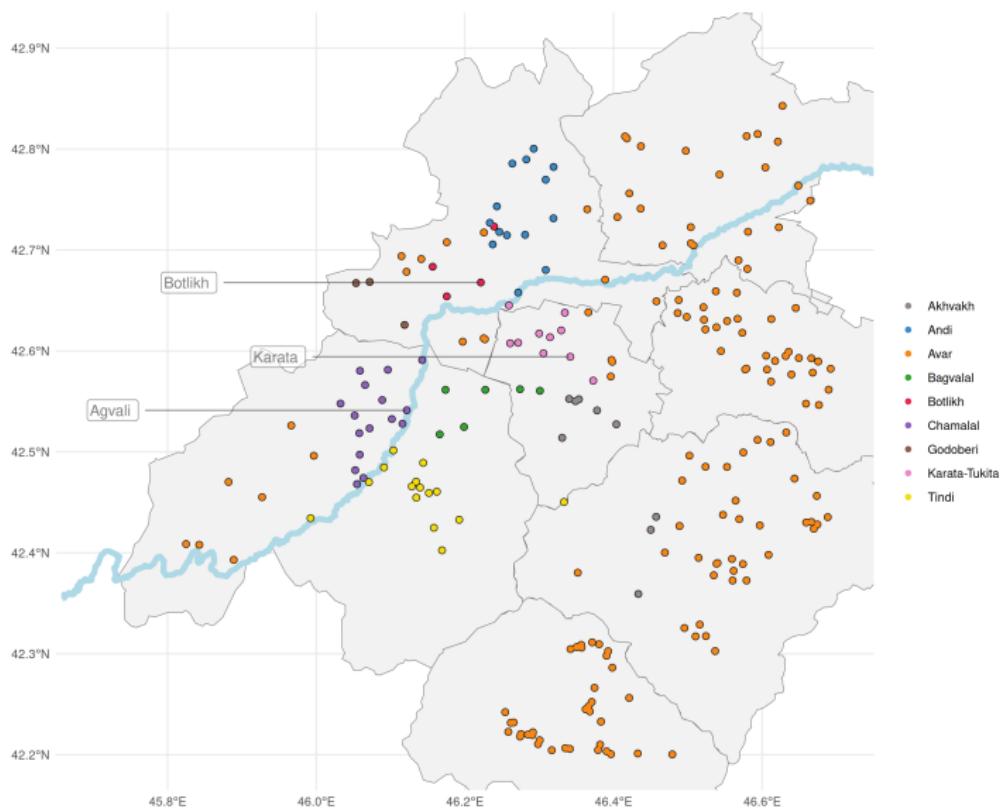
Анализ пространственных данных — это

анализ данных, который основывается на понятиях места, расстояний и пространственного взаимодействия как ключевых признаков данных и использует особые инструменты и методы для хранения, визуализации и исследования такого типа данных.

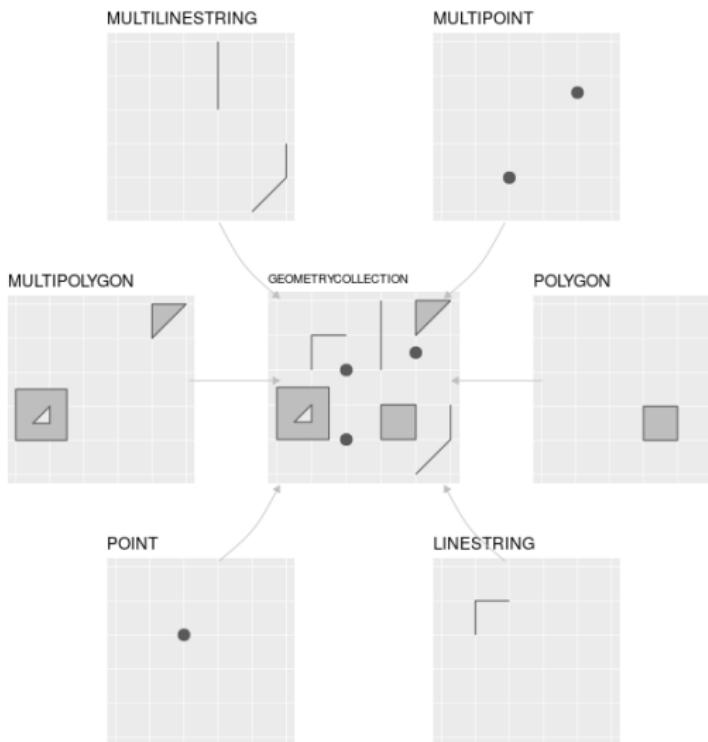
Пространственные примитивы



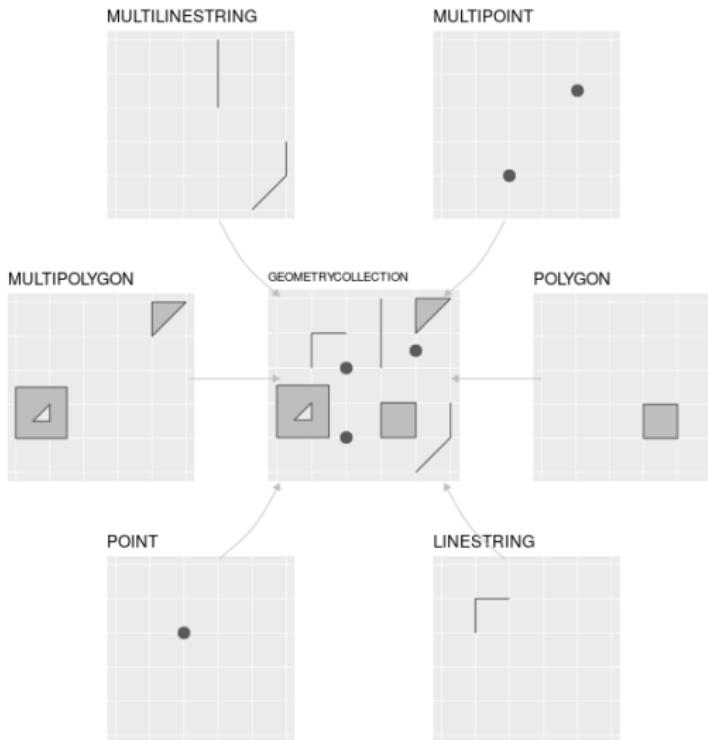
Какие пространственные примитивы можно здесь найти?



Чего, как Вам кажется, здесь не хватает?



Чего, как Вам кажется, здесь не хватает?



Мне не хватает объема (т. е. учета высотности).

Растровые данные

Иногда географические данные не представляют собой набор пространственных примитивов.

- сетка некоторой частоты, с некоторым приписанным значением каждой ячейке

Растровые данные

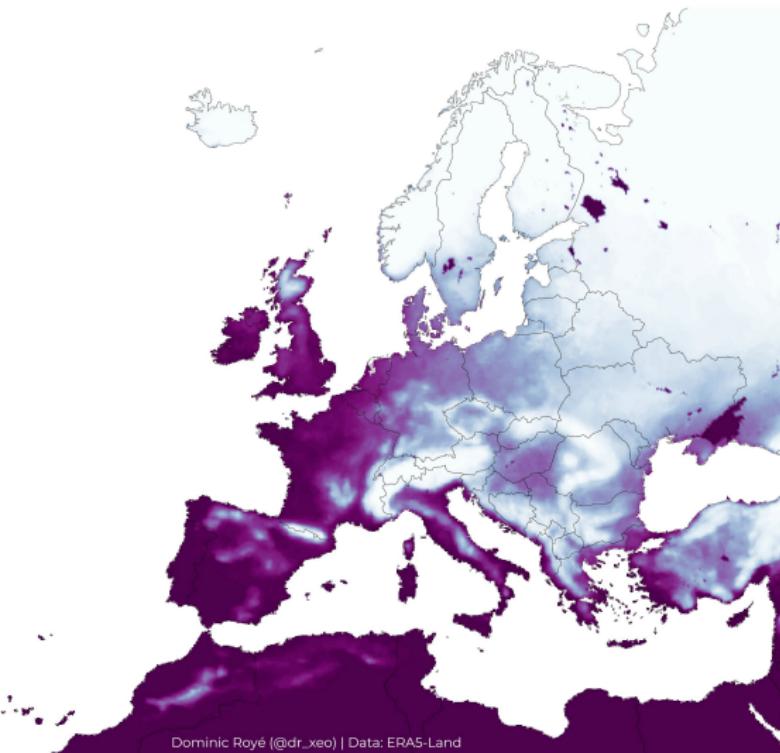
Иногда географические данные не представляют собой набор пространственных примитивов.

- сетка некоторой частоты, с некоторым приписанным значением каждой ячейке
- растровый объект, например, карта XVI века, которая даже не имеет привязки к современной системе координат

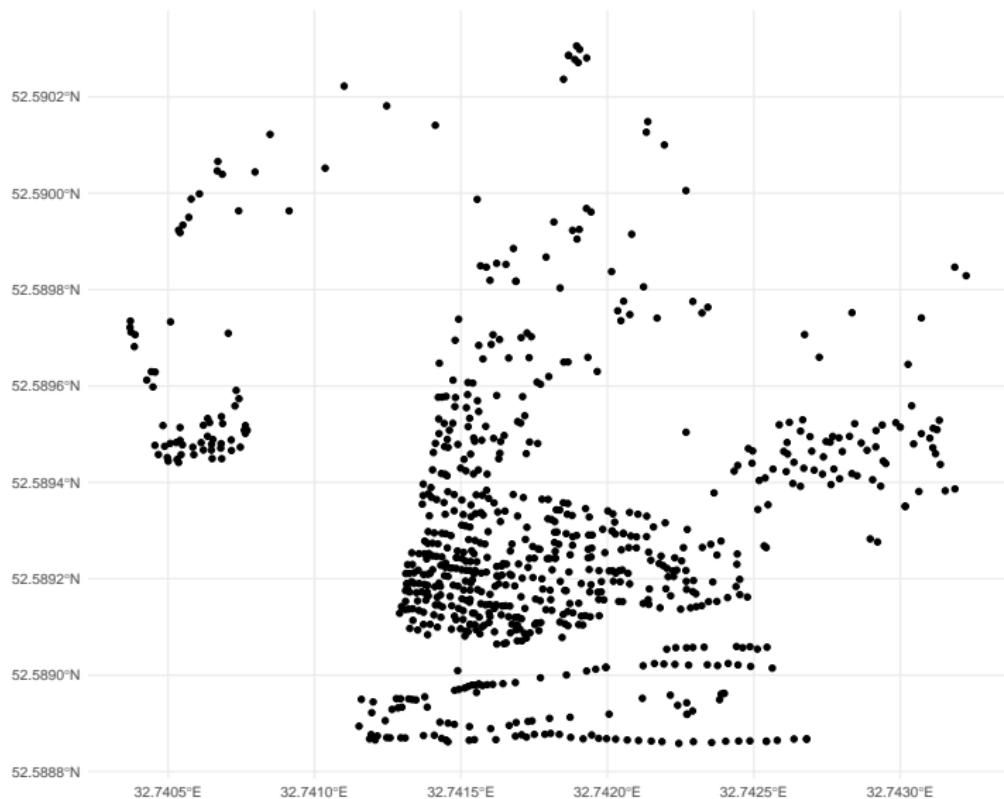
Растровые данные

Probability of snow on Christmas Eve in Europe

Snow depth > 1 cm. Reference period 1950-2020



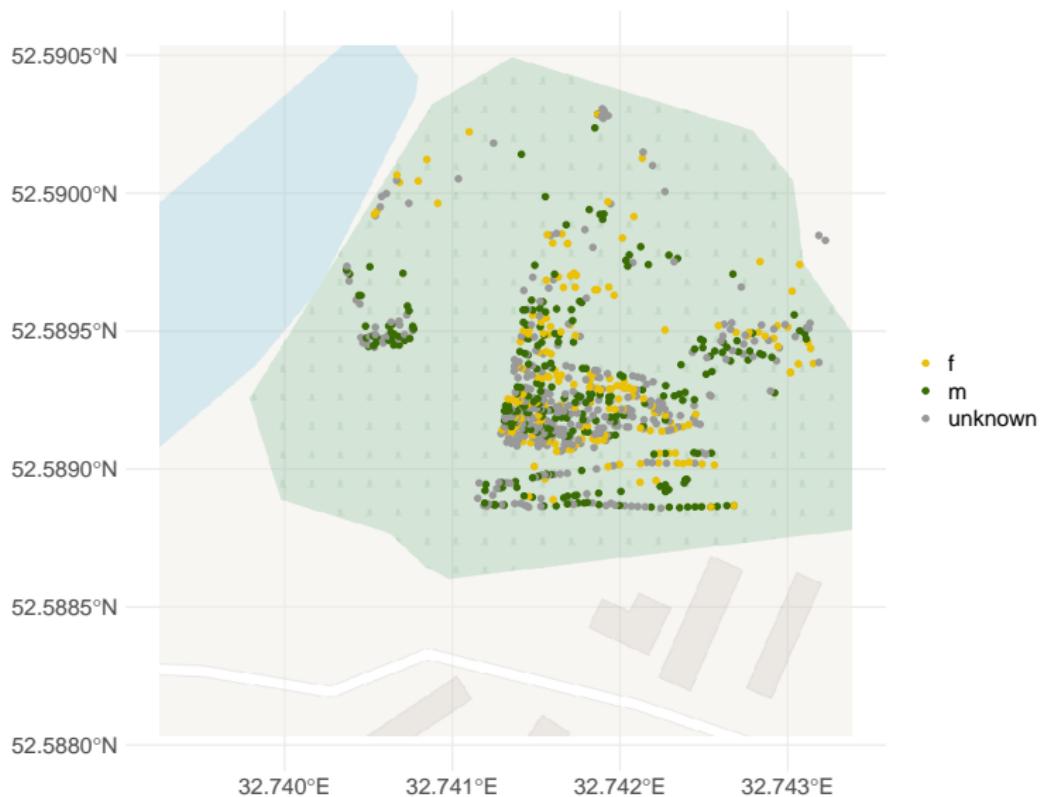
Кладбище Стародуб (данные полевого архива SFIRA)



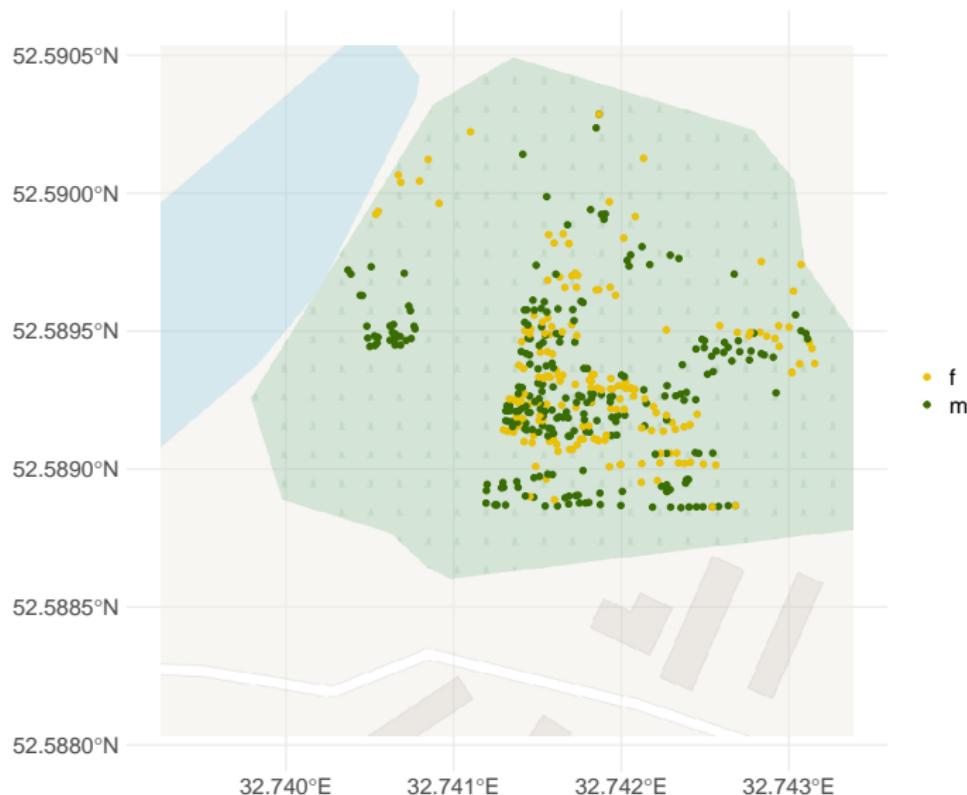
Кладбище Стародуб (данные полевого архива SFIRA)



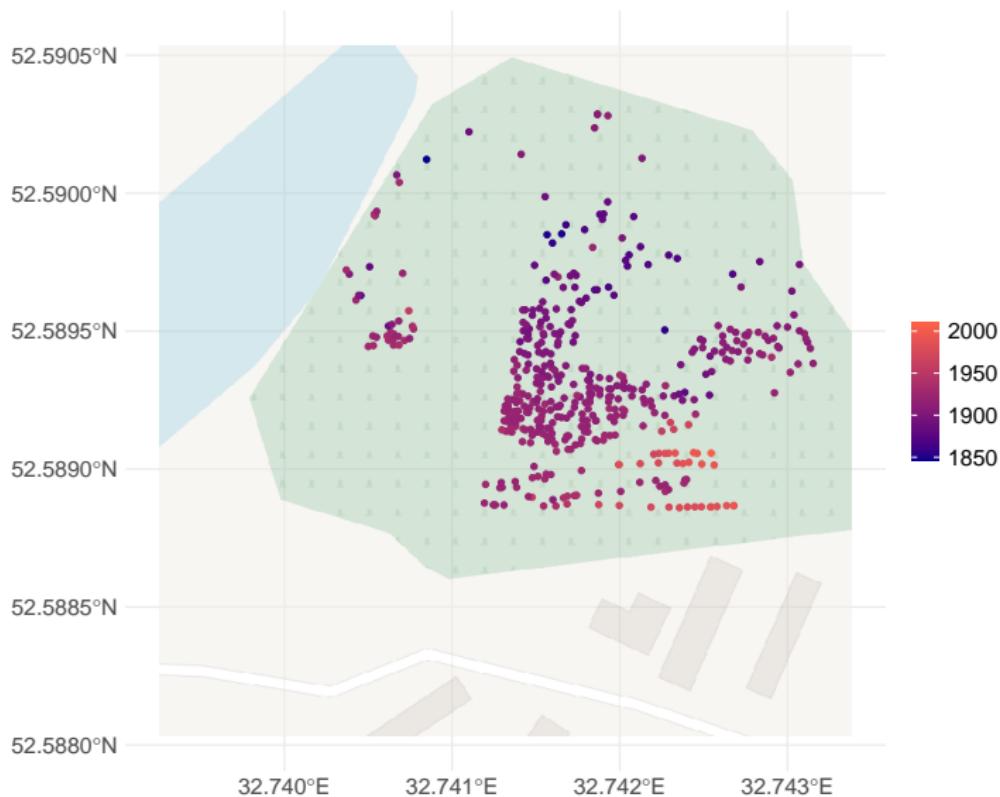
Кладбище Стародуб (данные полевого архива SFIRA)



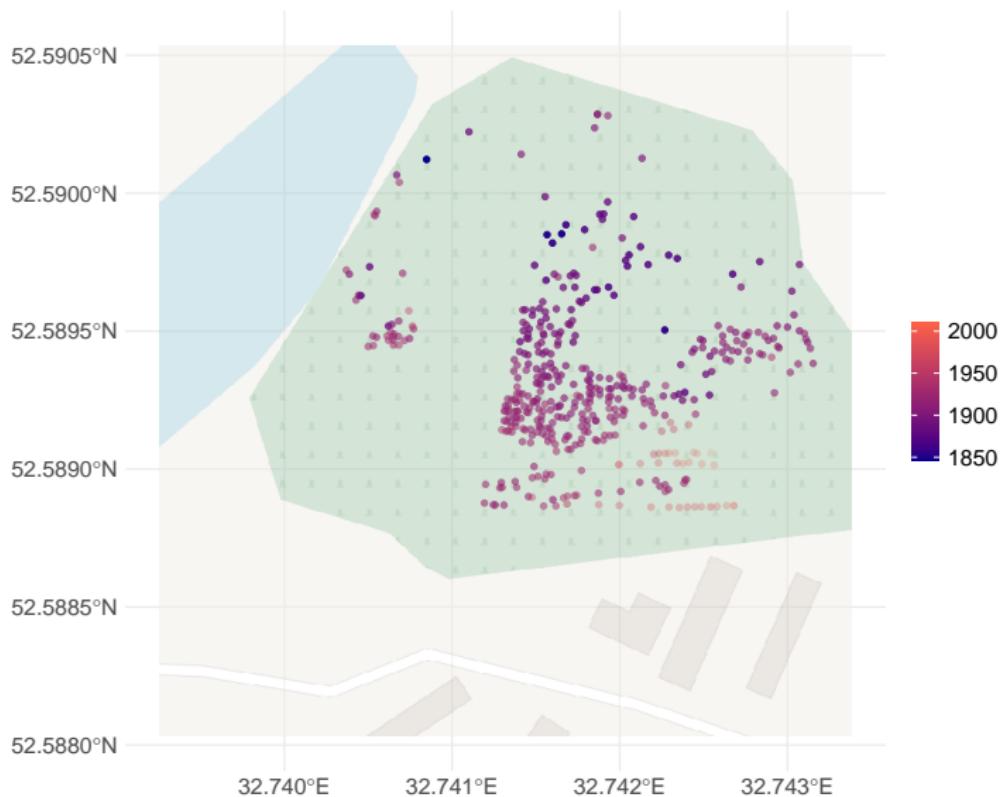
Кладбище Стародуб (данные полевого архива SFIRA)



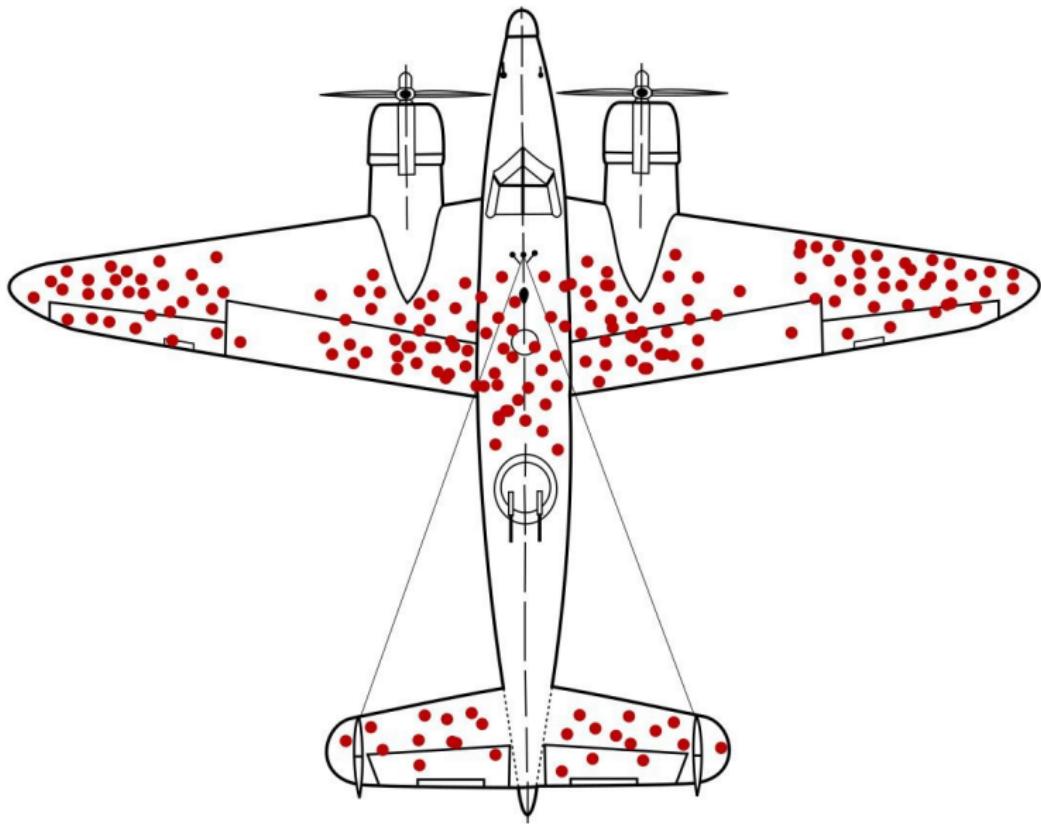
Кладбище Стародуб (данные полевого архива SFIRA)



Кладбище Стародуб (данные полевого архива SFIRA)



Ошибка выжившего: Абрахам Вальд



Картографическая проекция

Любое отображение некоторого небесного тела на плоскость называют картографической проекцией.

Если расстояния в ваших данных небольшие (особенно, если координаты близки к экватору), широту и долготу можно без страха использовать как оси в декартовой системе координат (она же — проекция Меркатора). Однако при работе с данными масштаба страны/континента/планеты такой подход будет накапливать ошибку из-за искажений одного из следующих типов:

- искажения длин;
- искажения углов;
- искажения площадей;
- искажения форм.

Картографическая проекция

Проекция Меркатора очень сильно искажает площади:



исходный

источник — Википедия



с сохранением площадей

Картографическая проекция

- [веб-приложение](#), помогающее выбрать подходящую проекцию
- [веб-приложение](#), которое показывает как изменяются объекты при преобразовании с сферы на одну из четырех проекций (Меркатора, цилиндрическую, Робинсона, Моллевайде)
- [Здесь](#) содержится список всех возможных проекций

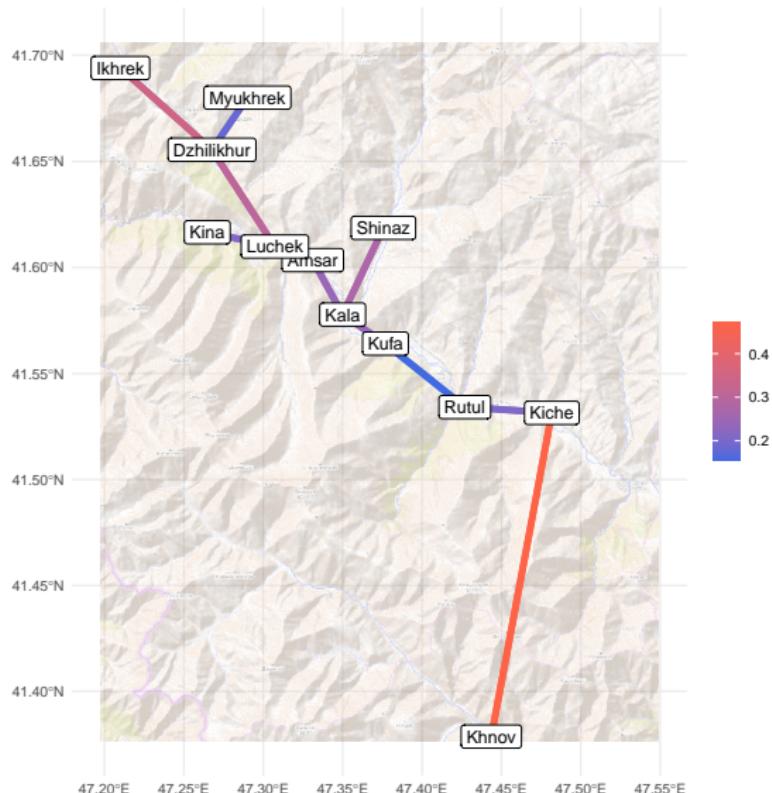
Моделирование пространственных отношений

Моделирование пространственных отношений позволяет отвечать на вопросы:

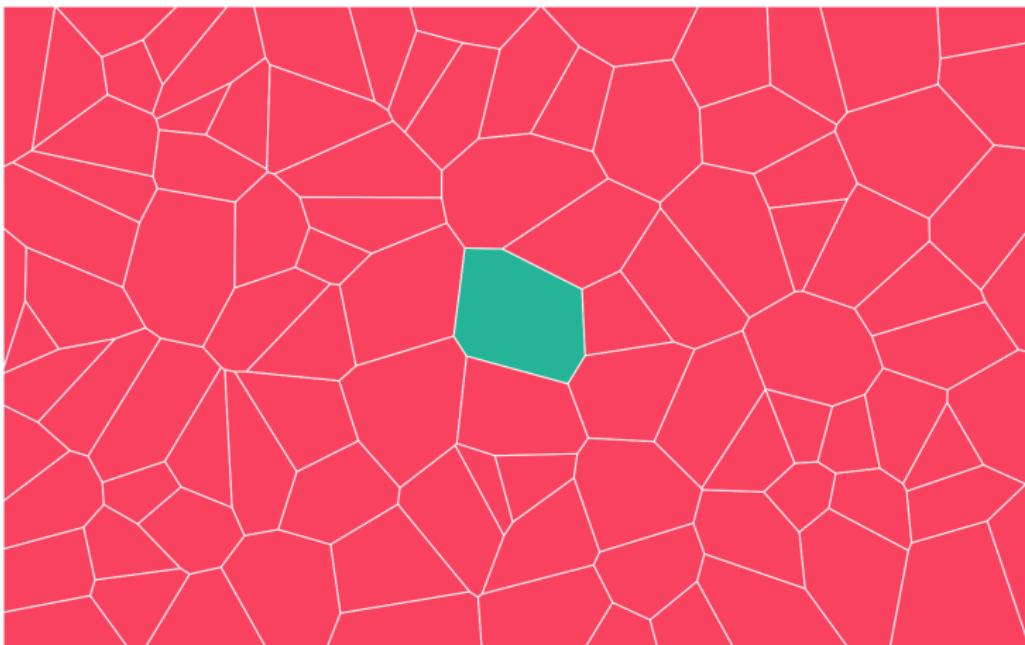
- Существует ли какая-то группировка значений исследуемой переменной в пространстве?
- Правда ли, что сходные значения имеют тенденцию находиться рядом?
- Можно ли выделить какие-то регионы концентрации каких-то из значений?

Однако для ответа на все эти вопросы мы прежде всего должны построить граф соседства.

Языковое сходство рутульских идиомов

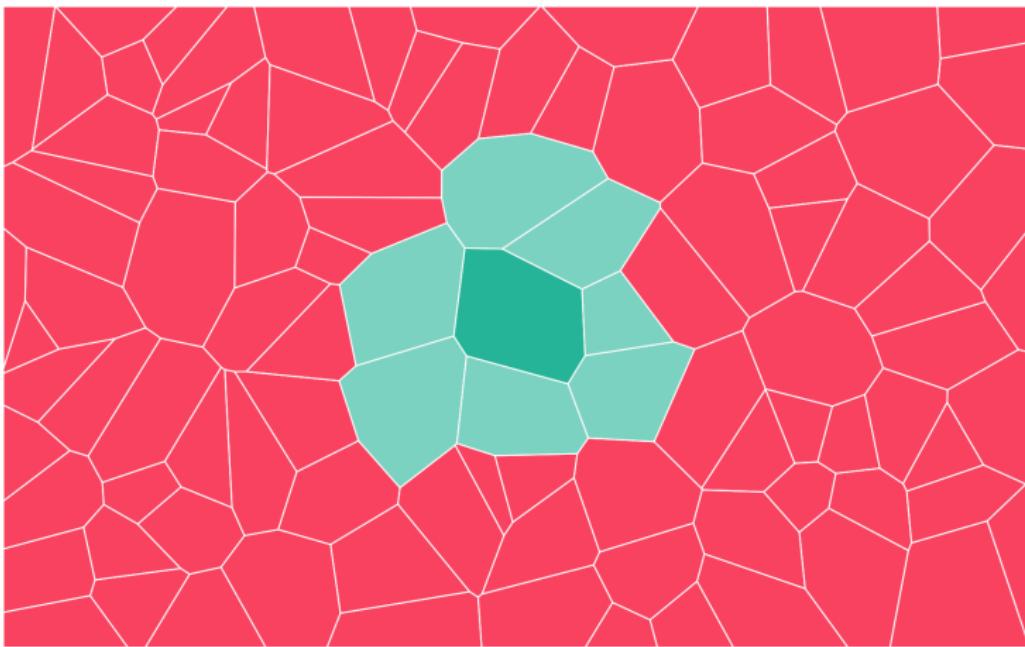


Как определить соседей?



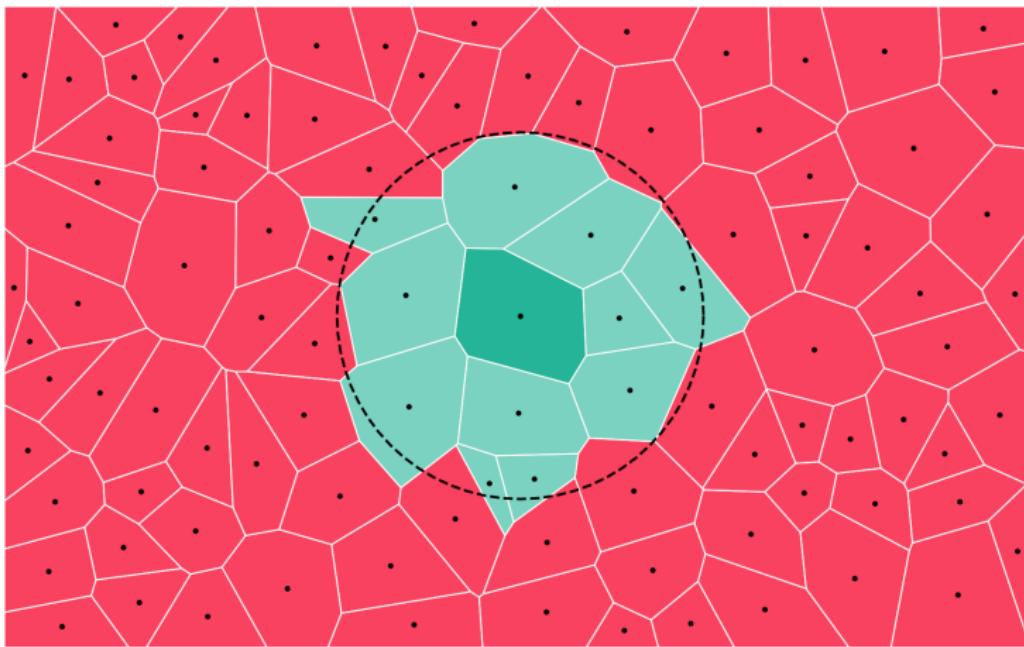
Из курса М. Фляйшманна “Spatial Data Science for Social Geography”

Как определить соседей?



Из курса М. Фляйшманна “Spatial Data Science for Social Geography”

Как определить соседей?



Из курса М. Фляйшманна “Spatial Data Science for Social Geography”

Пространственная автокорреляция

Степень в какой сходные значения находятся рядом.

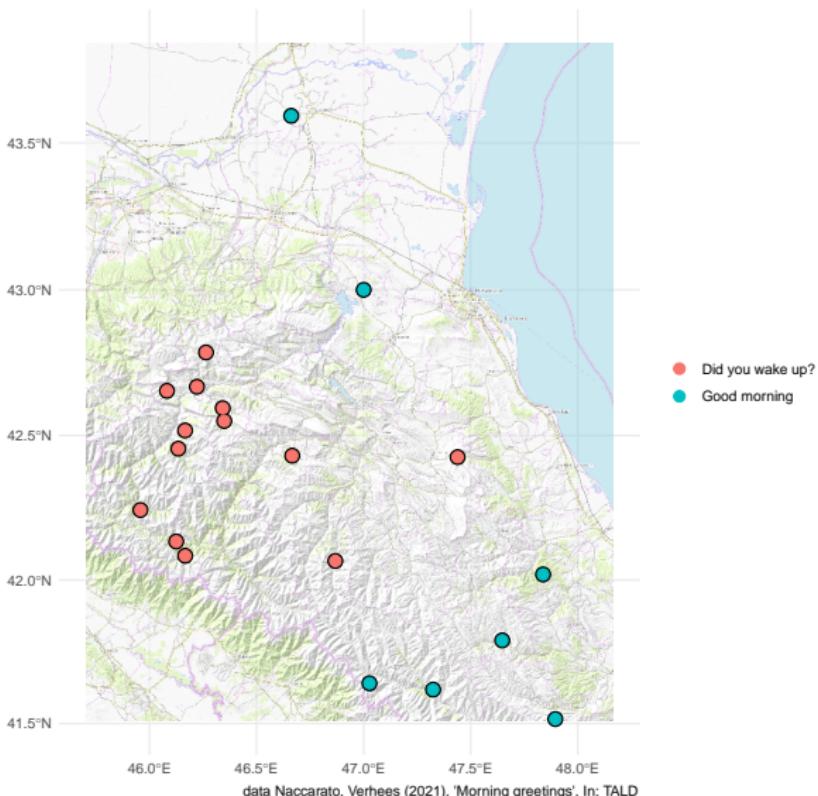
- положительная автокорреляция: похожие значения находятся рядом
- отрицательная автокорреляция: похожие значения находятся далеко друг от друга

Пространственная автокорреляция

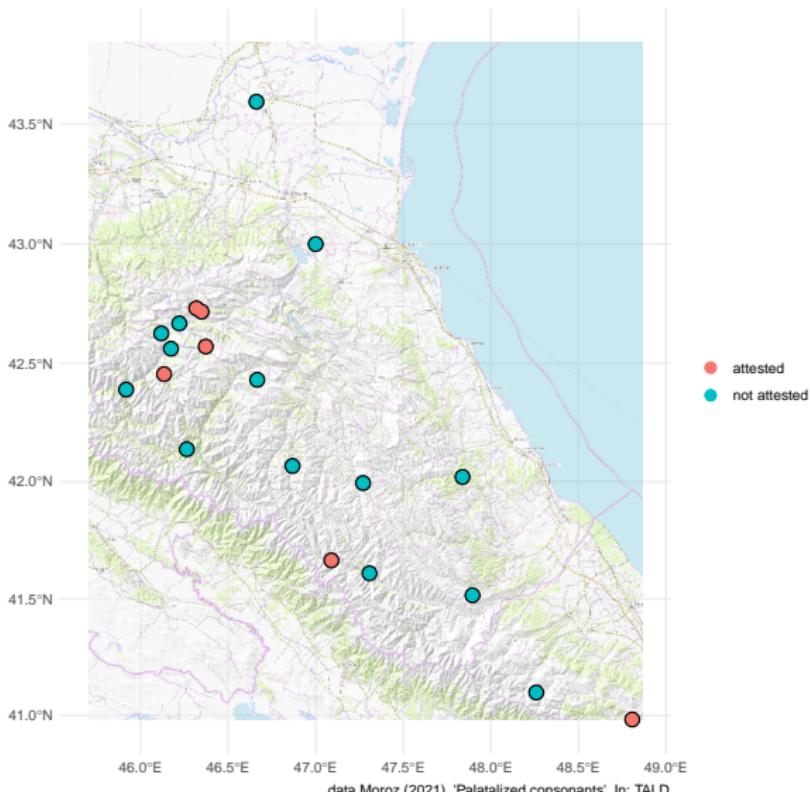
Степень в какой сходные значения находятся рядом.

- положительная автокорреляция: похожие значения находятся рядом
- отрицательная автокорреляция: похожие значения находятся далеко друг от друга
- глобальная: имеют ли значения тенденцию оказываться рядом с другими похожими/непохожими значениями;
- локальная: существует ли некоторый специфический фрагментом пространства, где наблюдается необычная концентрация похожими/непохожими значений.

Значение Moran I: 0.4763736



Значение Moran I: -0.1480726

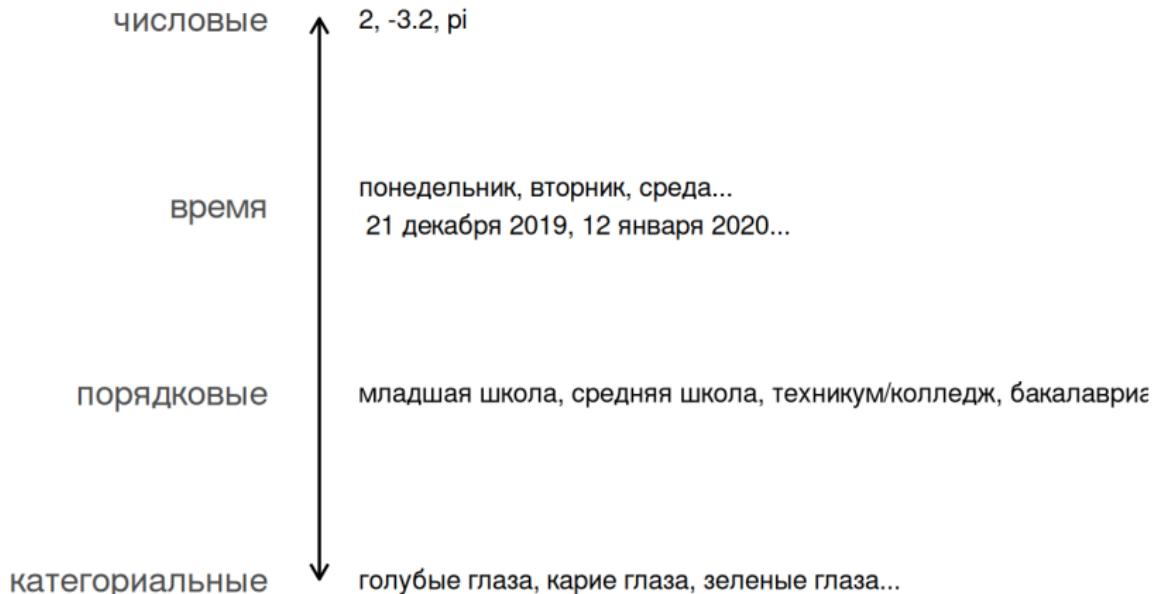


Мне хочется выразить благодарность

Евгению Николаевичу Матерову за его блог и телеграм-канал “Наука и данные” (<https://t.me/naukaidanne>), которые значительно упростили написание этой лекции, в частности за ссылку на курс Мартина Фляйшманна “Spatial Data Science for Social Geography”.

Временные данные

Переменные бывают разные



Переменные бывают разные

Кажется, что время — просто обычная числовая переменная, на которой определены все обычные операции сложения вычитания и т. п. Однако стоит держать в голове несколько фактов:

- Не каждый год содержит 365 дней. Существуют високосные годы.

Переменные бывают разные

Кажется, что время — просто обычная числовая переменная, на которой определены все обычные операции сложения вычитания и т. п. Однако стоит держать в голове несколько фактов:

- Не каждый год содержит 365 дней. Существуют високосные годы.
 - Не каждый день содержит 24 часа. Во многих странах используют переход на летнее и зимнее время.

Переменные бывают разные

Кажется, что время — просто обычная числовая переменная, на которой определены все обычные операции сложения вычитания и т. п. Однако стоит держать в голове несколько фактов:

- Не каждый год содержит 365 дней. Существуют високосные годы.
 - Не каждый день содержит 24 часа. Во многих странах используют переход на летнее и зимнее время.
 - Не в каждой минуте 60 секунд. Существует дополнительная секунда, которую добавляют чтобы компенсировать замедление во вращении земли (тогда после секунды 23:59:59 идет секунда 23:59:60).

Переменные бывают разные

- Григорианский календарь — не единственный календарь
 - тогда дней в году может быть не 365 (в исламском календаре 354–355 дней)
 - дней в неделе может быть не 7 (в исторических календарях, например, древнеегипетском)
 - месяцев в году может быть не 12

Високосная секунда

Данные расхождения с ожиданиями связаны с двумя возможными определениями суток:

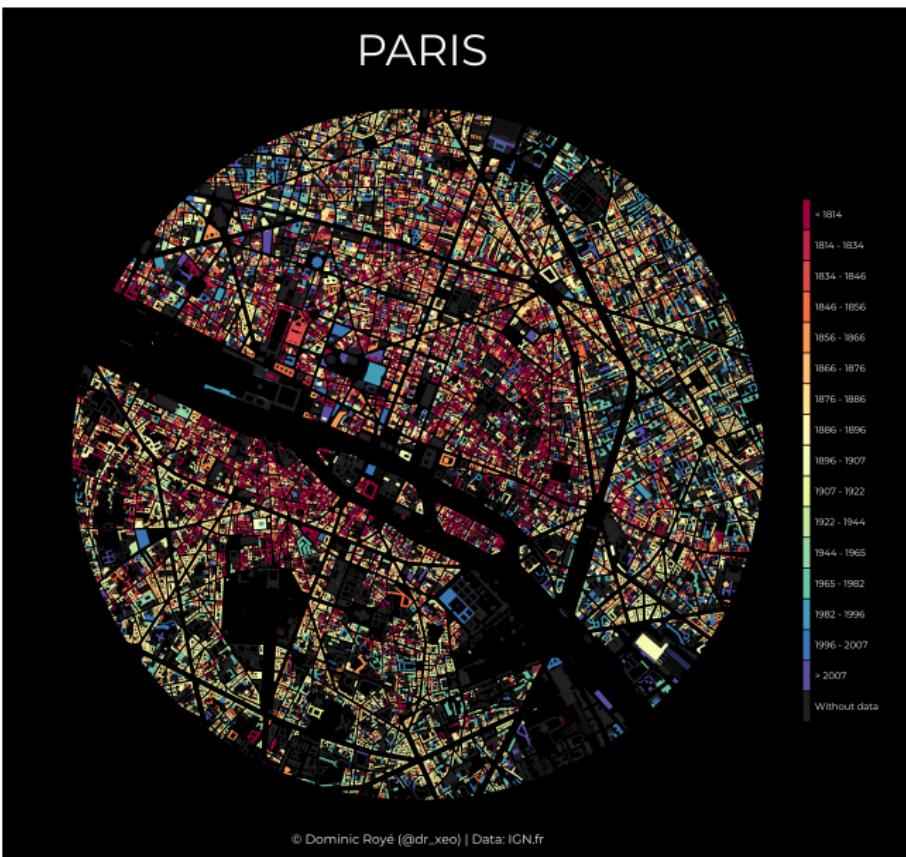
- период времени, за который Земля совершает оборот вокруг своей оси;
- период времени, равный 86 400 секундами (246060).

Так как секунду в какой-то момент определили без привязки к вращению Земли, ученым приходится периодически подкручивать время, добавляя високосные секунды.

Временные пояса



Время + география



Список календарей из Википедии

- Армелина
- Армянские: древнеармянский и христианский
- Ассирийский
- Ацтекский
- Бахаи
- Бенгальский
- Буддийский
- Вавилонский
- Византийский
- Восточнославянский
- Вьетнамский
- Гильбурда
- Григорианский
- Грузинский
- Дариский
- Древнегреческий
- Древнеегипетский
- Древнеперсидский
- Древнеславянский
- Еврейский
- Зороастрийский
- Индийские: древнеиндийский и единый
- Иники
- Иранский
- Ирландский
- Исламский
- Кельтский
- Киргизский
- Китайский
- Конта
- Коптский
- Малайский
- Майя
- Масонский
- Миньго
- Непальский
- Новоюлианский
- Пролетарийский: юлианский и григорианский
- Римский
- Румынский
- Рунический
- Симметричный
- Стабильный
- Тамильский
- Тайские: лунный и солнечный
- Тибетский
- Трёхсезонный
- Тувинский
- Туркменский
- Французский
- Хакасский
- Ханаанейский
- Хараппский
- Чучхе
- Шведский
- Шумерский
- Эфиопский
- Юлианский
- Яванский
- Японский

Проблемы разницы календарей

- необходима конвертация
- годы в разных системах могут начинаться в разное время
- информация о дне и месяце может быть опущена, и тогда приходиться конвертировать с неопределенностью

30.06.1938~28.07.1938 (данные полевого архива SFIRA)



... Хаим
... в месяц...
тамуз года
{5}698. Озерски.
АВРАМ [ХА]ИМОВ
Аврам Хаимов
О[ЗЕРС]КИЙ

30.06.1938~28.07.1938 (данные полевого архива SFIRA)



... Хаим
... в месяц...
тамуз года
{5}698. Озерски.
АВРАМ [ХА]ИМОВ
Аврам Хаимов
О[ЗЕРС]КИЙ

Если резчик не ошибся...

Моделирование со временем

Существует несколько типов моделей, которые пытаются предсказать нечто, как функция от времени, мы рассмотрим только следующие:

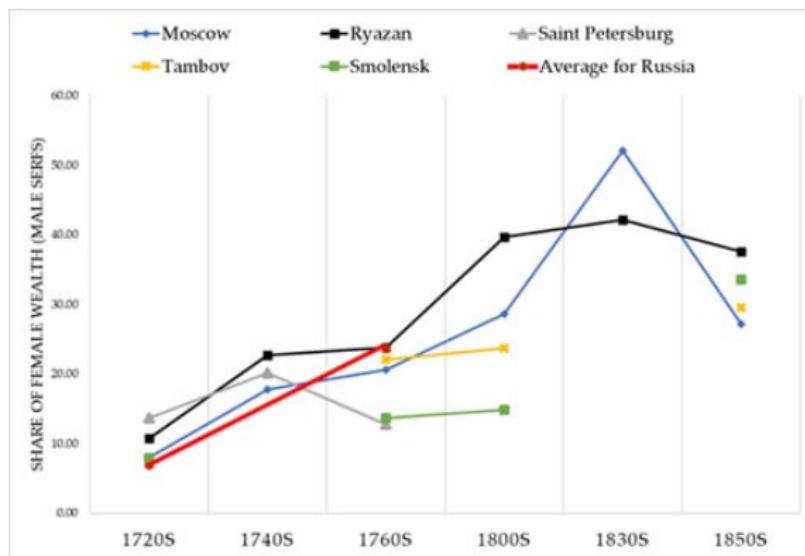
- временные ряды
- анализ выживаемости

Временные ряды

Любые упорядоченные наблюдения

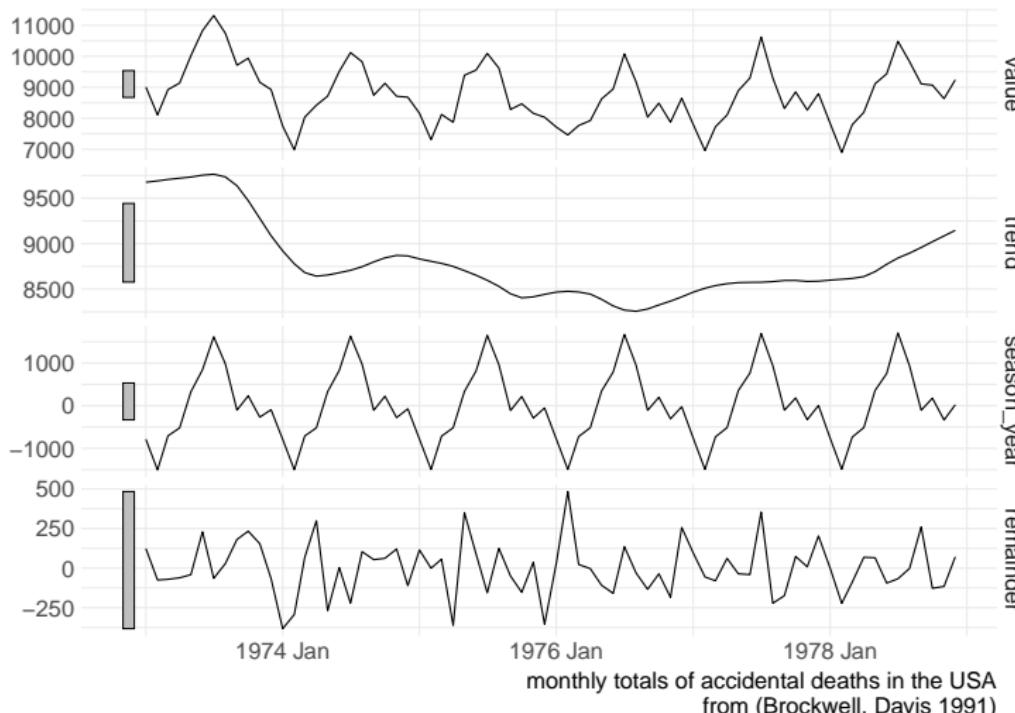
Временные ряды

Любые упорядоченные наблюдения



Женское землевладение в России из телеграм канала Елены Корчминой “Ревизская сказочница”

Временные ряды: разложение на тренд и сезонную составляющую



Временные ряды: анализ выживаемости

