

Математические и технологические проблемы построения графика в параллельных осях

Студент: Тыцкий В.И.

Научный руководитель: Майсурадзе А.И.

МГУ имени М. В. Ломоносова, факультет ВМК, кафедра ММП

Оглавление

[Введение](#)

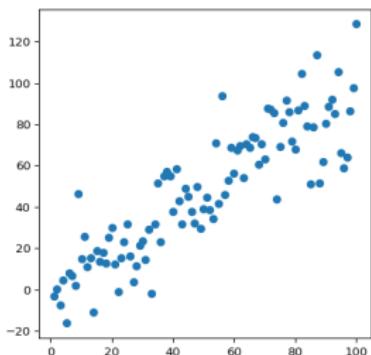
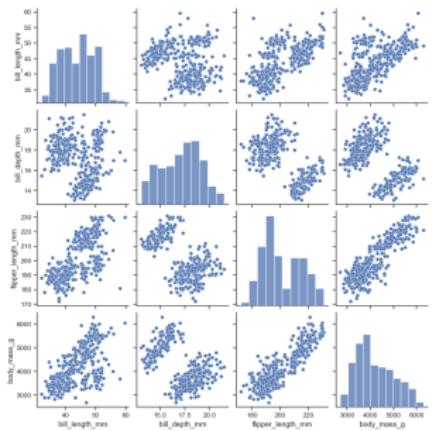
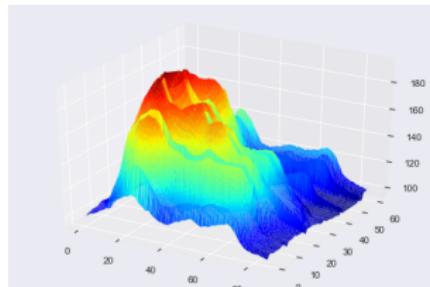
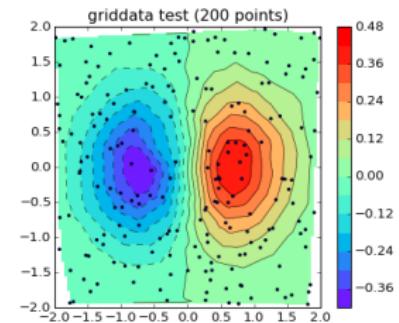
[Модификации](#)

[Проблемы построения](#)

[Методы выбора порядка](#)

[О библиотеке](#)

Примеры диаграмм



Историческая справка

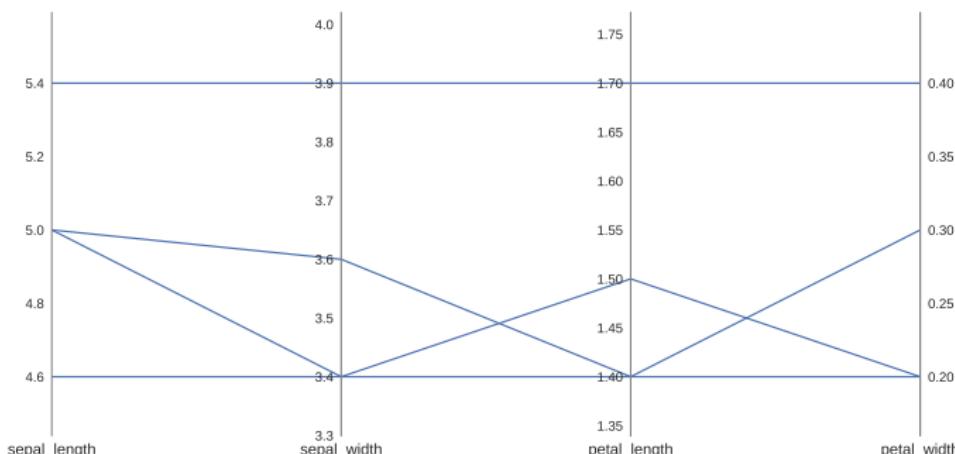
The value of data visualization is not seeing “zillions” of objects but rather recognizing relations among them.

Alfred Inselberg

- Параллельные координаты были известны еще в 19-ом веке
- В 1980-ых были популяризированы Альфредом Инсельбергом

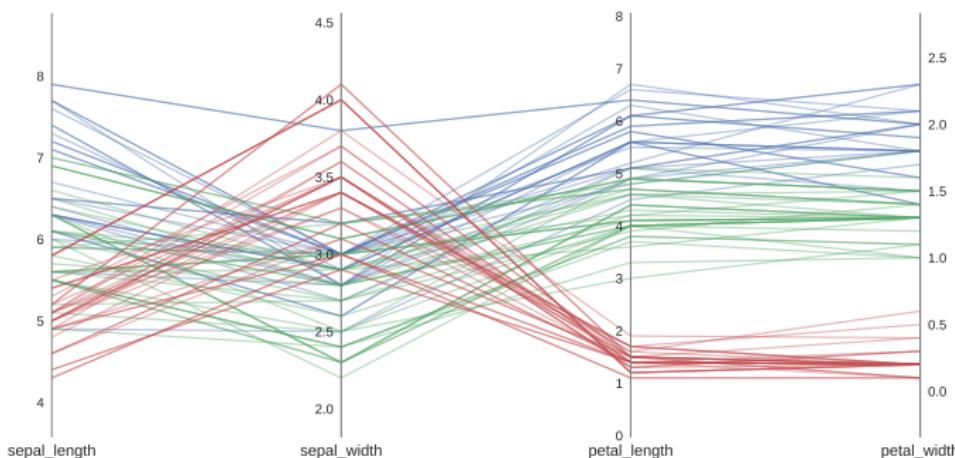
Классический график в параллельных осях

- Каждая линия соответствует объекту
- Каждая ось соответствует некоторой координате в пространстве
- Направление, порядок и масштаб осей может быть произвольным



Модификации: кластеры

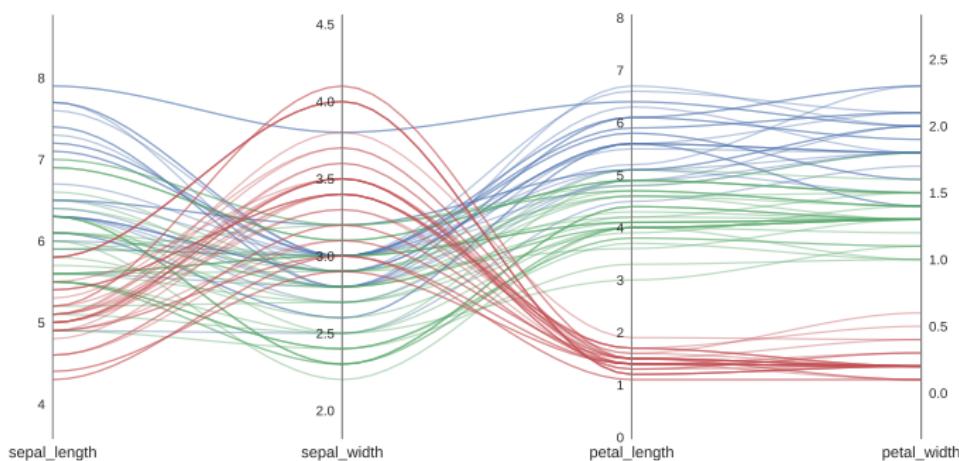
Кластер — класс родственных элементов статистической совокупности.



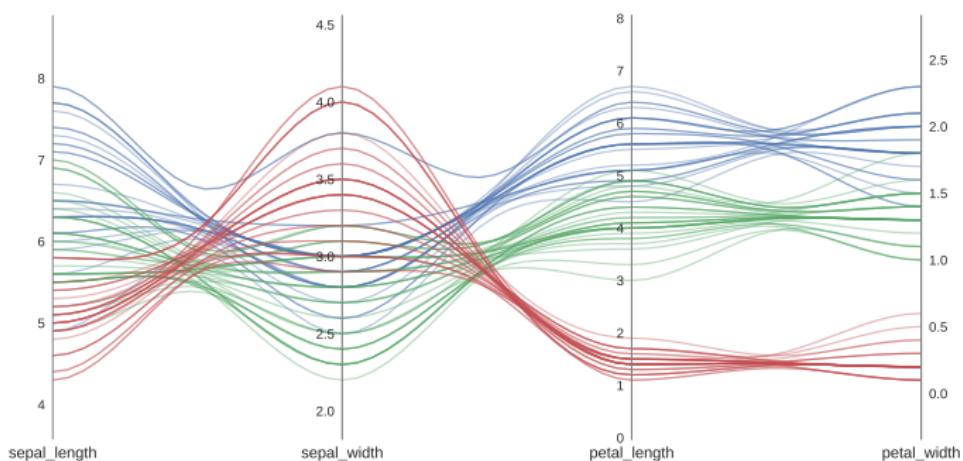
Чаще всего именно в таком виде используют график в параллельных осях.

Сглаживание линий

Человеку проще воспринимать гладкие линии, поэтому читаемость графика заметно возрастает.

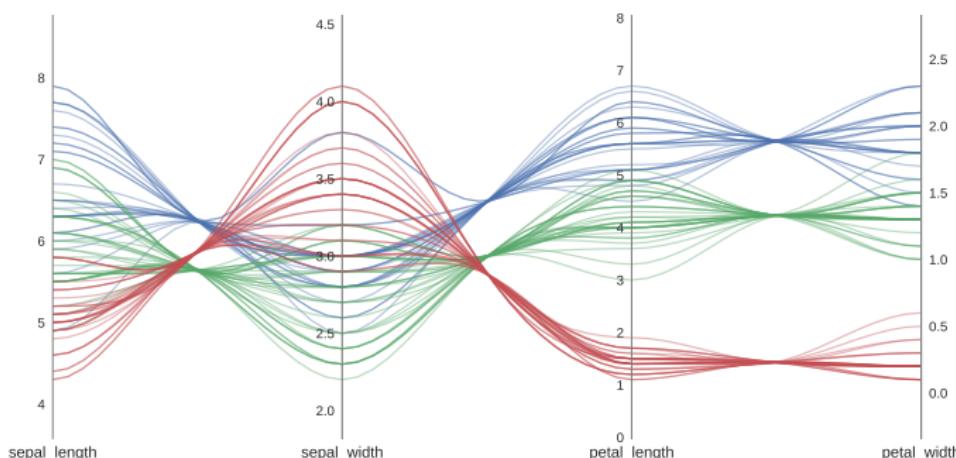


Связывание линий



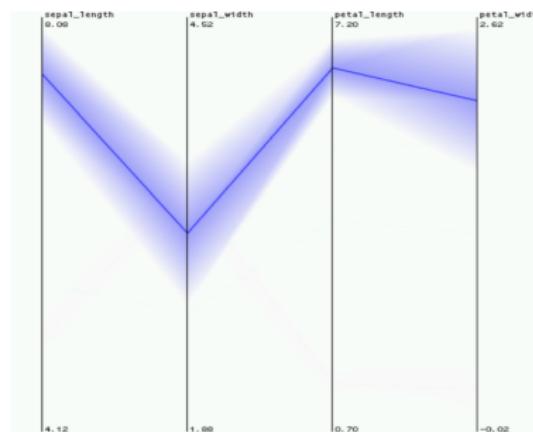
Связывание линий

Сильное связывание приводит к потере читаемости в рамках конкретного объекта, но увеличивает читаемость в рамках кластеров



Иерархические графики

Изображаем статистики распределений соответствующих кластеров (std, min, max, mean) вместо отрисовки каждого объекта.



Иерархические графики

Пусть $X = (x_1, \dots, x_n)$ – выборка, где $x_i \in \mathbb{R}^n$.

Назовем множество P m -разбиением множества X на m -подмножеств $\{P_1, \dots, P_m\}$ такое, что:

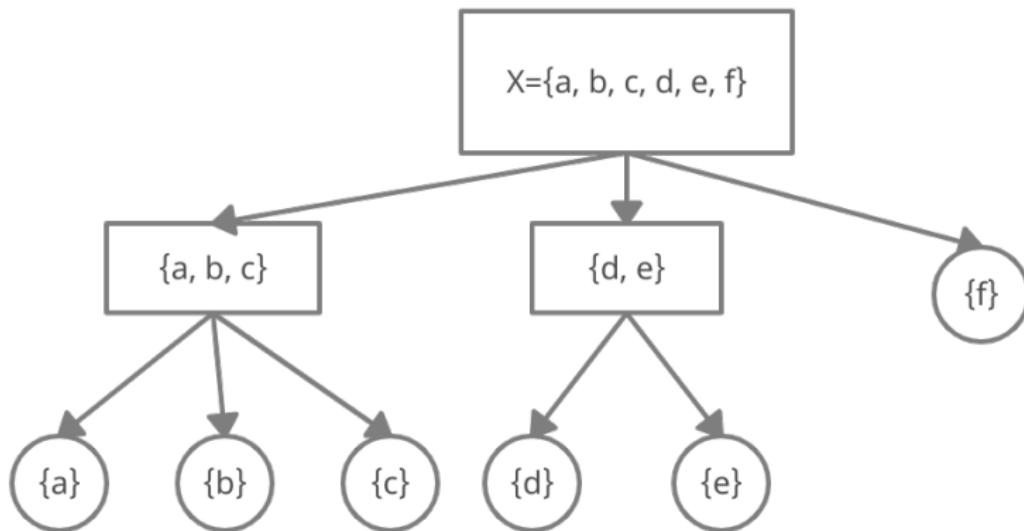
$$1. P_i \cap P_j = \emptyset, \quad \forall i, j = \overline{1, m}$$

$$2. \bigcup_{i=1}^m P_i = X$$

Организуем иерархическую структуру в виде дерева, где корню соответствует X , а каждая вершина сопоставлена элементу разбиения родительской вершины.

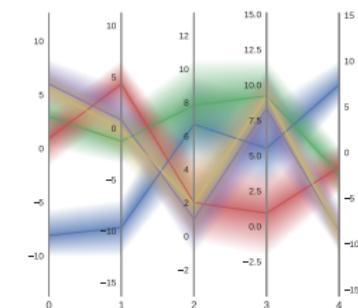
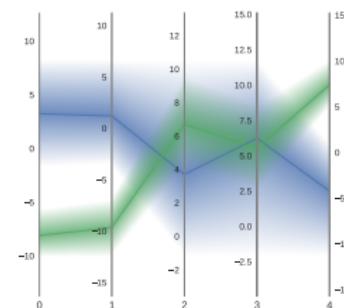
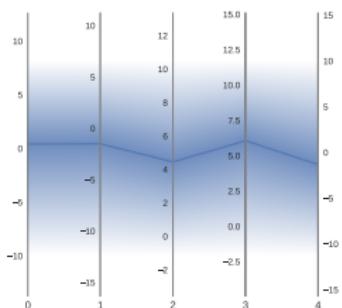
Синонимы: агломеративная кластеризация, иерархическая кластеризация

Пример иерархического разбиения

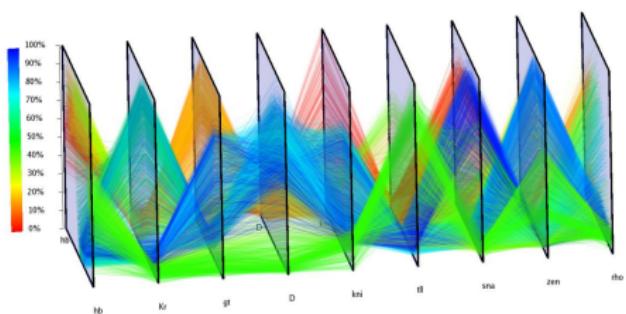


Иерархические графики

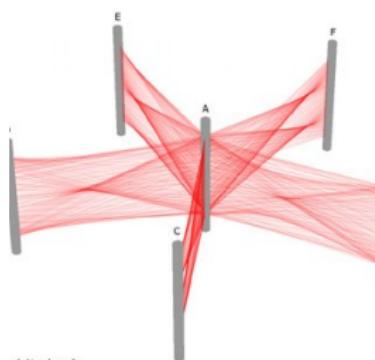
Регулируя глубину, мы добавляем/уменьшаем количество кластеров на графике



3D



3D parallel coordinates



3D multi-relational
parallel coordinates

Плюсы и минусы

Плюсы:

- Решаем проблему визуализации многомерных пространств
- Высокая вариативность
- Простая интерпретация
- Обнаружение аномалий (выбросов)

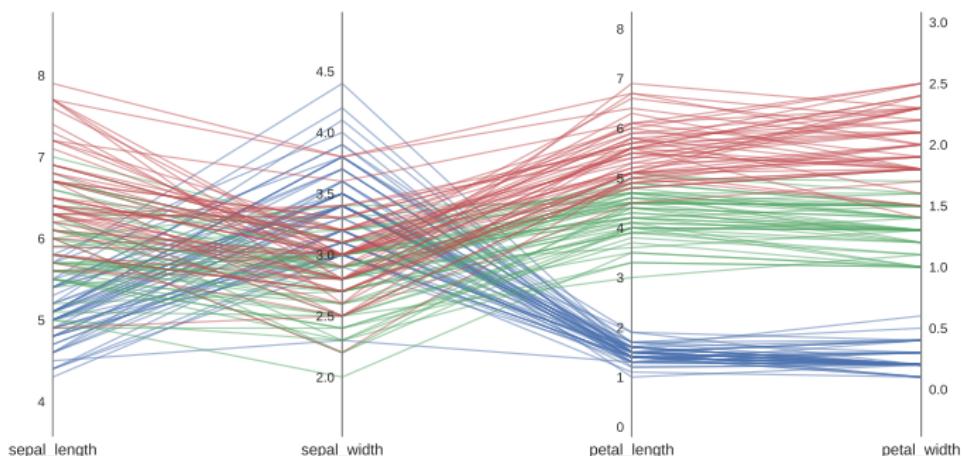
Минусы:

- Теряется читаемость на больших выборках
- Только вещественные признаки
- Много гиперпараметров
- Необходимость объяснять принцип построения графика

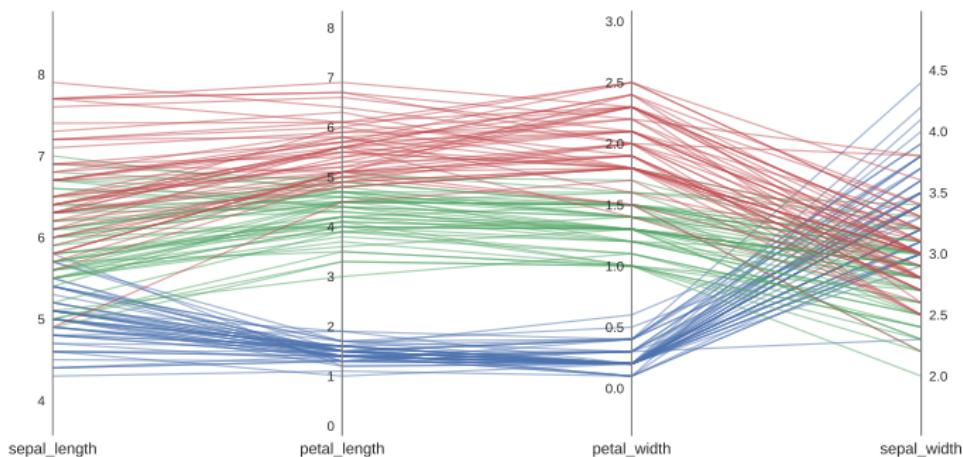
Естественные вопросы при построении

- В каком порядке расположить оси?
- В какую сторону направлять оси?
- Как много объектов отобразить?
- Какой масштаб выбрать для каждой оси?

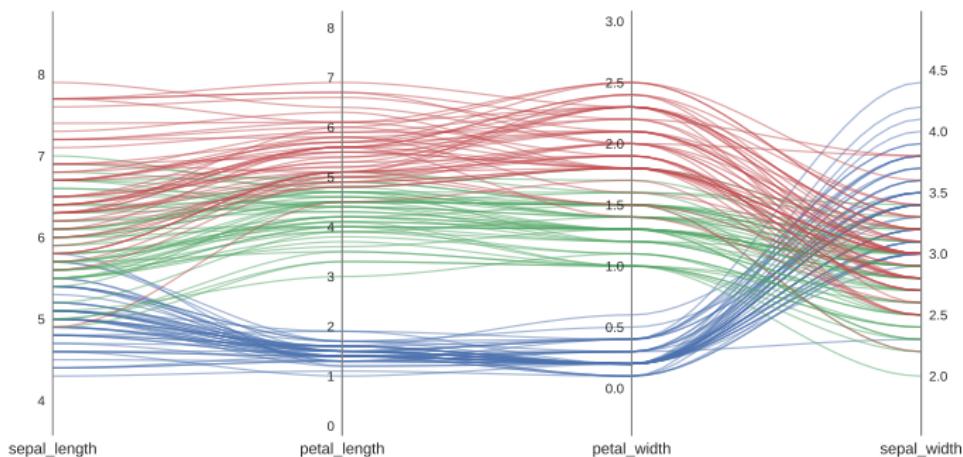
Выбор направлений и порядка осей



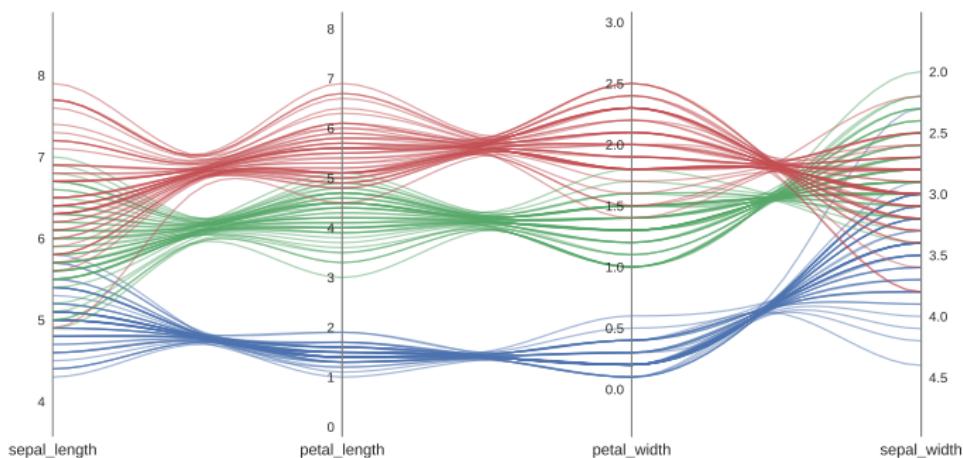
Выбор направлений и порядка осей



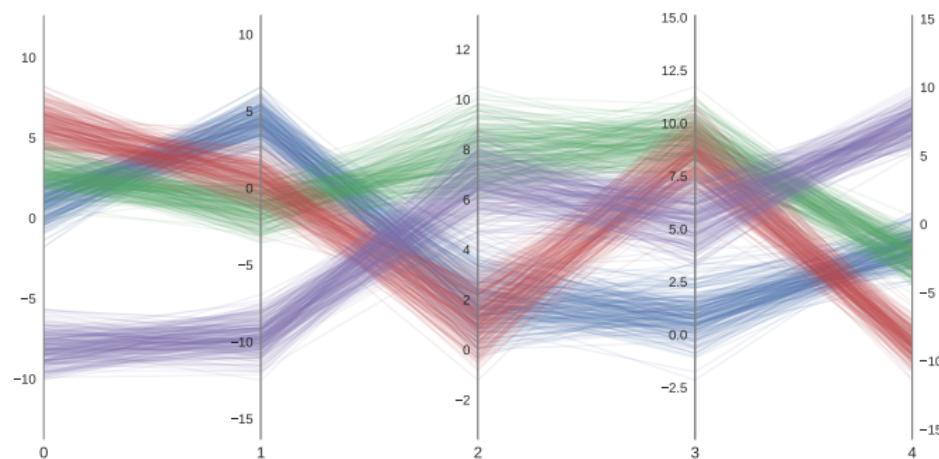
Выбор направлений и порядка осей



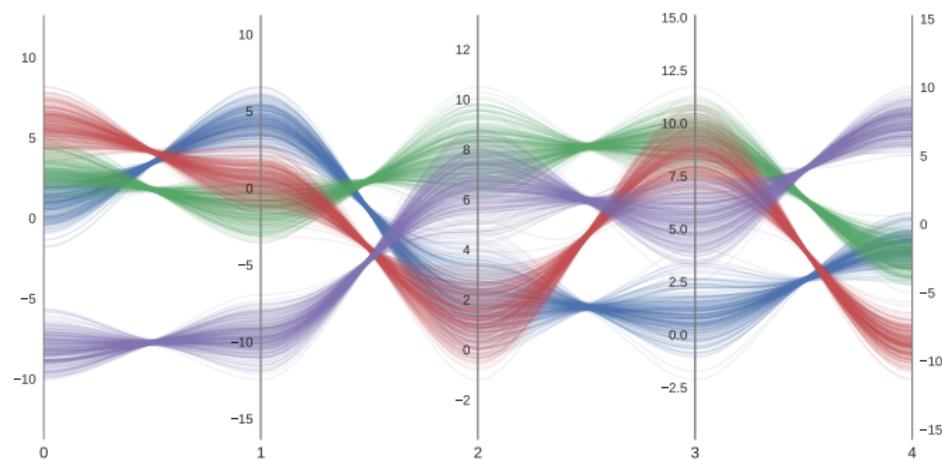
Выбор направлений и порядка осей



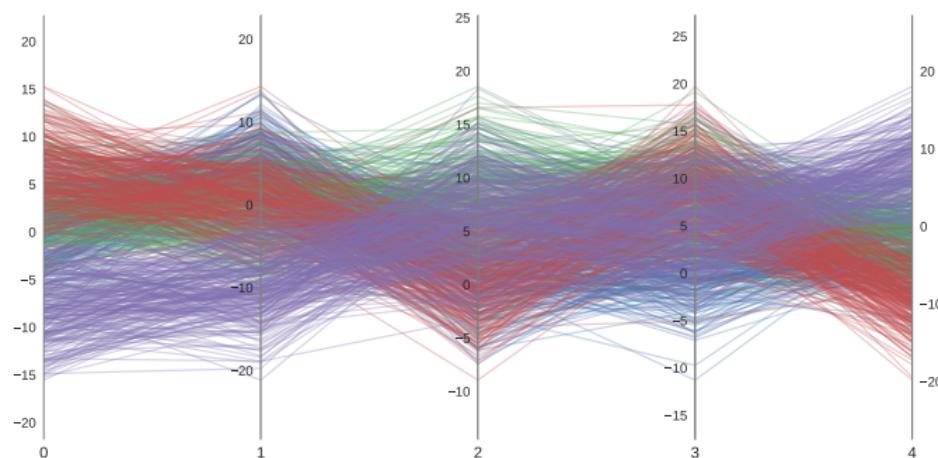
Влияние количества объектов на читаемость



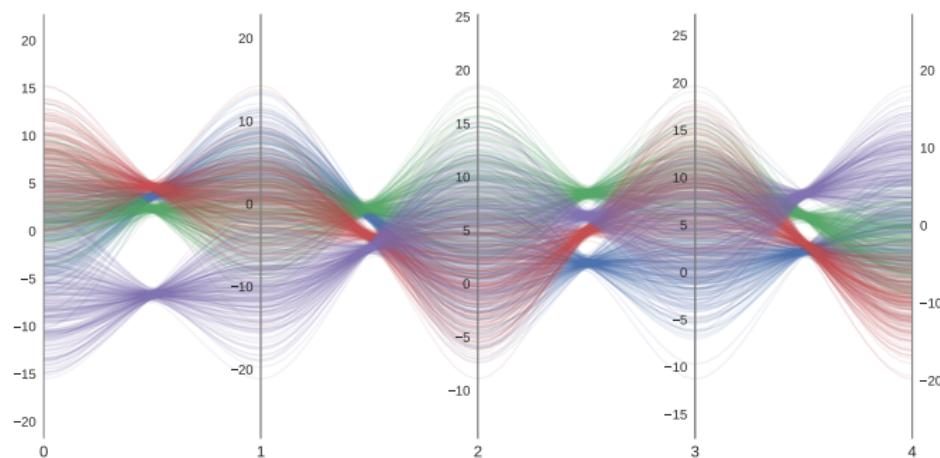
Влияние количества объектов на читаемость



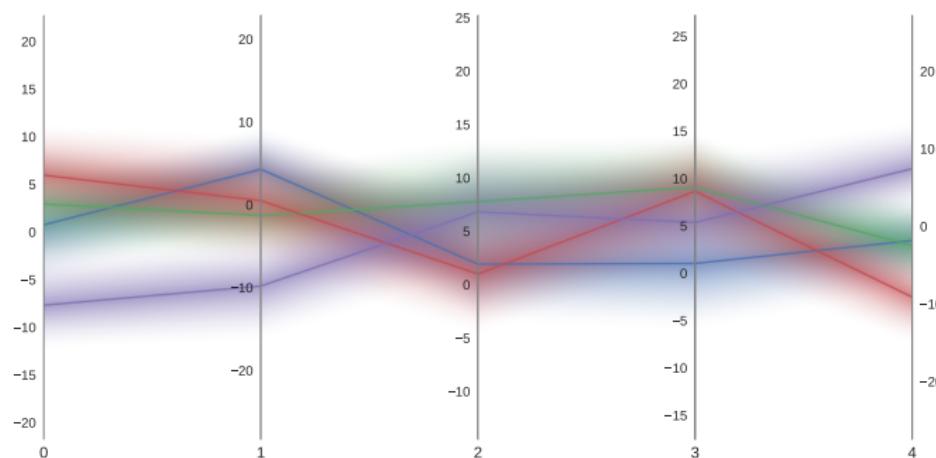
Влияние количества объектов на читаемость



Влияние количества объектов на читаемость



Влияние количества объектов на читаемость



Резюмируя

Методы решения проблем:

- Изменение степени прозрачности линий.
- Использование гладких линий.
- Связывание линий в рамках кластеров.
- Отображение лишь части объектов.
- Изменение порядка и направления осей.

"Правильный" порядок

Формализуем понятие "правильного" порядка осей

График хороший, если:

- Линии редко пересекаются между собой
- Есть монотонная зависимость между двумя соседними координатами
- Направление осей вверх
- Слишком шумные "зависимости" где-нибудь на последних осях

Корреляция

Пусть даны две выборки $X = (x_1, \dots, x_n)$, $Y = (y_1, \dots, y_n)$.

Корреляция Пирсона

$$\rho_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad |\rho_{XY}| \leq 1$$

Пусть R_i – ранг наблюдения x_i , S_i – ранг наблюдения y_i

Корреляция Спирмена

$$r_{XY} = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (S_i - \bar{S})^2}}, \quad |r_{XY}| \leq 1$$

$\rho_{XY} = 0$, $r_{XY} = 1$, где $X = Y^2$ и X симметрично распределена относительно нуля.

Минимизируемый функционал

Пусть $\pi = (\pi_1, \dots, \pi_n)$ перестановка множества $\{1, \dots, n\}$, где n размерность пространства.

Мы хотим максимизировать такой функционал:

$$\sum_{i=1}^{n-1} |r_{X^{\pi_i} X^{\pi_{i+1}}}| \rightarrow \min_{\pi}$$

где $r_{X^i X^j}$ – Корреляция Спирмена между i координатой и j координатой.

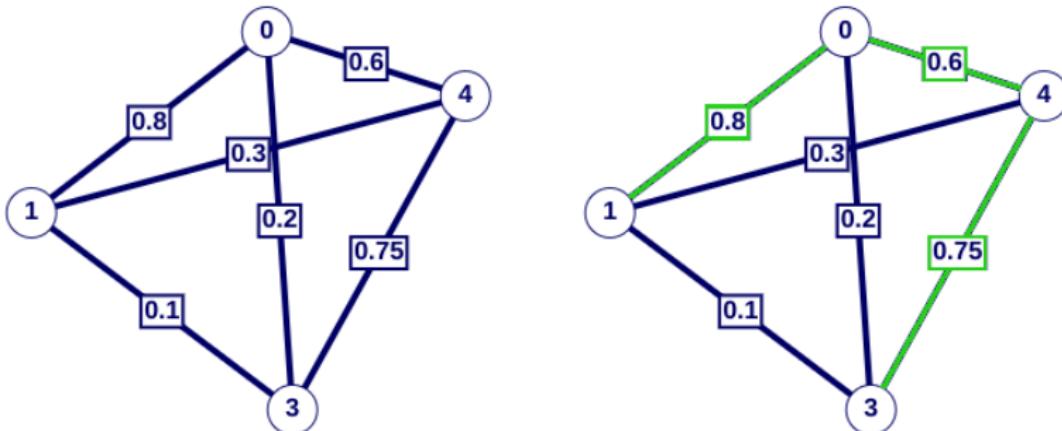
Задача о самом длинном пути

Это задача поиска простого пути максимальной длины в заданном графе. Является NP-трудной и не может быть решена за полиномиальное время для произвольных графов.

Пусть $X = (x_1, \dots, x_n)$ – выборка, где $x_i \in \mathbb{R}^n$.

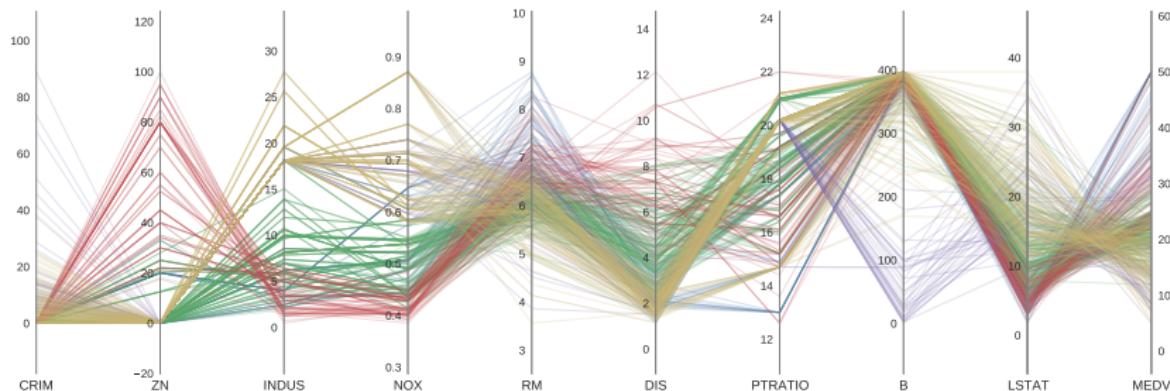
Построим связный граф $G(V, E)$, где каждая вершина $u^i \in V$ соответствует i -ой координате (i -ой оси на графике), а каждому ребру $\{u^i, u^j\} \in E$ сопоставим вес равный $|r_{X^i X^j}|$.

- Простейшим перебором задача решается за $O(n!)$
- Можно свести к задаче коммивояжера.
- С помощью методов динамического программирования можно улучшить асимптотику решения.



Пример использования

Классический вид

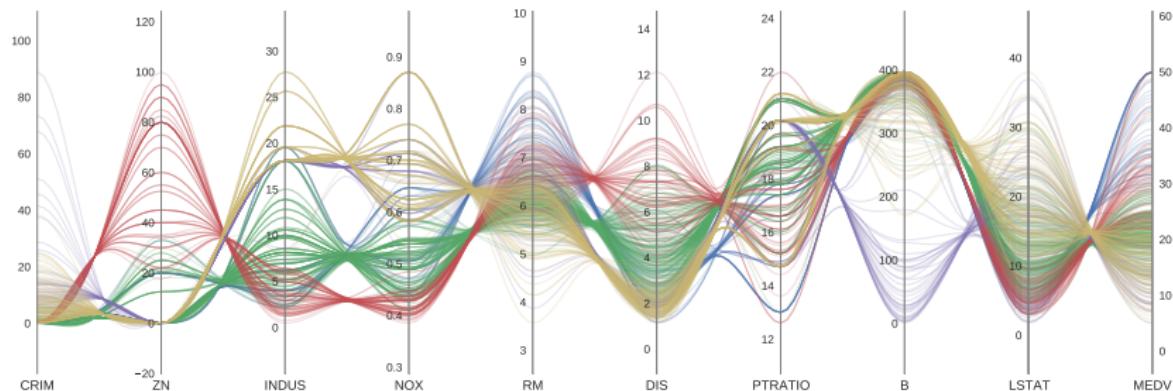


Датасет *boston housing*

Предварительно проведена кластеризация KMeans

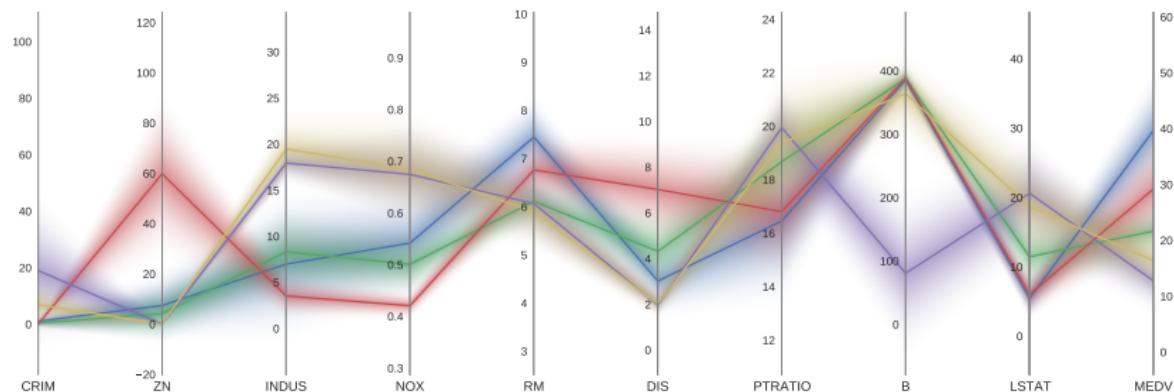
Пример использования

Со сплайнами и связыванием



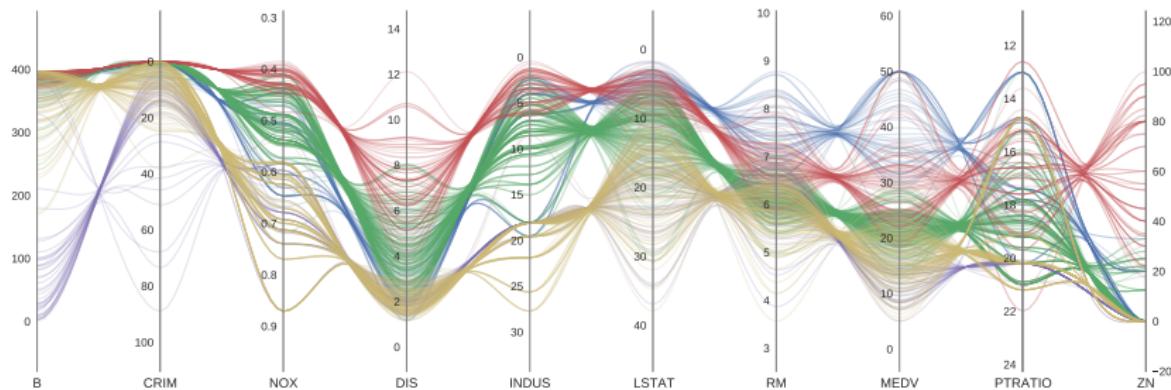
Пример использования

С агрегацией



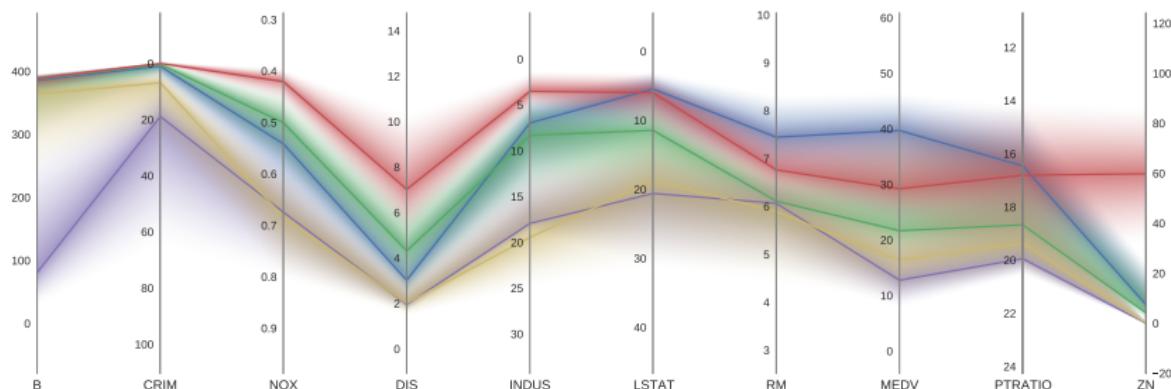
Пример использования

Сплайны + связывание + оптимальный порядок осей



Пример использования

Агрегация + оптимальный порядок осей



Обзор текущих средств

- На Python есть простейшая реализация лишь в библиотеке pandas!
- ELKI, GGobi, Mondrian, Orange и ROOT.
- Parcoords.js интерактивная библиотека на JavaScript.

Цели

- Дать возможность исследователям "безболезненно" использовать график в параллельных осях.
- Построение красивых и информативных графиков из "коробки".
- Реализация всевозможных видов данных графиков.

Технические подробности

- Статические графики.
- Библиотека пишется на языке Python на базе matplotlib.
- Простой высокоуровневый интерфейс. Как и в библиотеке seaborn методы могут принимать pandas.DataFrame, обычные питону массивы или списки – для всего единый интерфейс.

Возможности

Построение классических графиков в параллельных осях

- Возможность рисовать гладкие линии.
- Возможность "связывания" линий кластеров.
- Возможность "связывания" линий на основе близости.

Построение иерархических графиков

- Отрисовка полупрозрачного градиента.
- Работа с иерархическими кластерами.
- Изображение распределения с помощью градиента.

Дополнительные возможности

- выделение подмножества линий в диапазоне значений одной из осей.
- нахождение оптимального расположения осей.
- создание иерархических кластеров на основе входящей выборки.

Итоги

The screenshot shows a GitHub repository page for 'Tytskiy / hpcords'. The repository has 1 branch and 0 tags. It contains 29 commits from 3711392, with the latest commit being 'fix href' 6 days ago. The repository includes files for 'course', 'hpcords', 'tests', 'ignore', 'LICENSE', and 'README.md'. The 'README.md' file contains a section titled 'HPCoords' with a brief description: 'A library for visualizing a parallel coordinates and its variations and hierarchical parallel coordinates.' Below the README, there are links to 'Презентация' and 'Текст курсовой'. A note at the bottom says 'Простейший график с использованием библиотеки:'. On the right side of the page, there are sections for 'About', 'Readme', 'MIT License', 'Releases', 'Packages', and 'Languages'. The 'Languages' section shows TeX at 74.6% and Python at 25.4%.

- Реализовано API для визуализации всех основных графиков.
- Совсем скоро добавлю библиотеку в пакетные менеджеры.
- Работаю над реализацией дополнительных возможностей

Итоги

- Написана библиотека для визуализации графика в параллельных осях.
- Предложены новые способы улучшения читаемости графика.
- График в параллельных осях очень мощный инструмент визуализации.

Ссылки

<https://github.com/Tytskiy/hpcoords>

- [1] Inselberg A. *The plane with parallel coordinates.* The visual computer. 1985. – Т. 1. – №. 2. – С. 69-91.
- [2] Fua Y. H., Ward M. O., Rundensteiner E. A. *Hierarchical parallel coordinates for exploration of large datasets.* IEEE, 1999. – С. 43-508.
- [3] KYang, Jing; Peng, Wei; Ward, Matthew O.; Rundensteiner, Elke A. *Interactive Hierarchical Dimension Ordering Spacing and Filtering for Exploration of High Dimensional Datasets.* IEEE Symposium on Information Visualization, 2003. INFOVIS 2003: 3–4.