курс «Прикладные задачи анализа данных»

CASE: задача о пробках

Александр Дьяконов

1 сентября 2020 года

План лекции

Постановка задачи

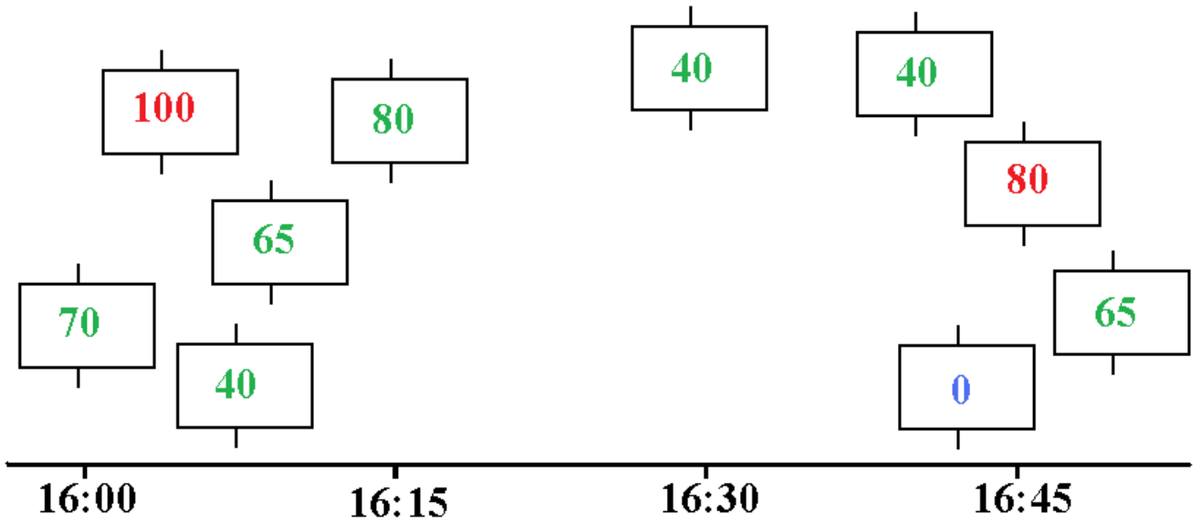
Двухмерное усреднение

Особенности данных

Специальное усреднение

в этой лекции будут плохие картинки

Задача о пробках

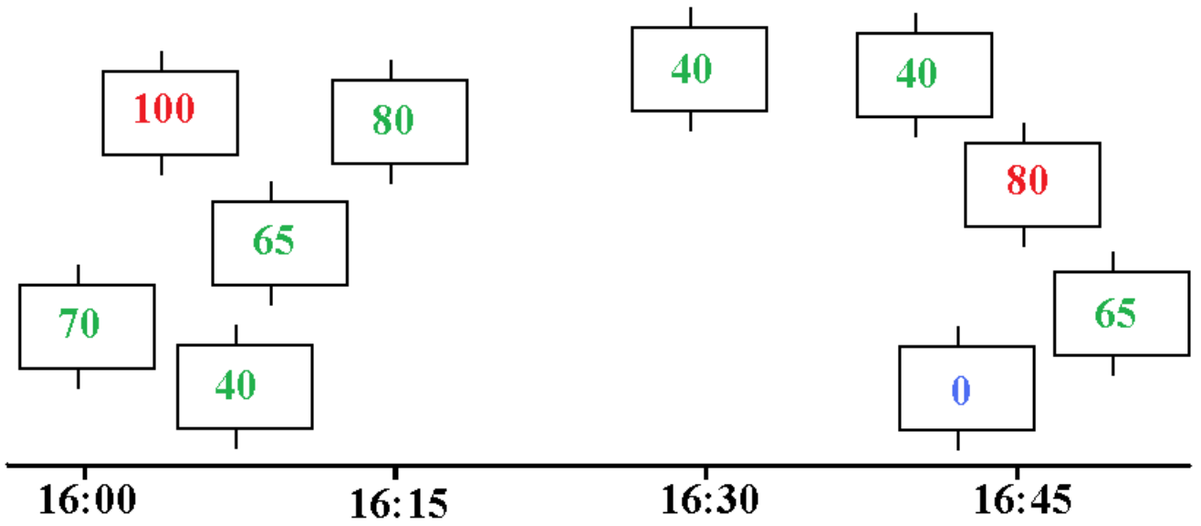


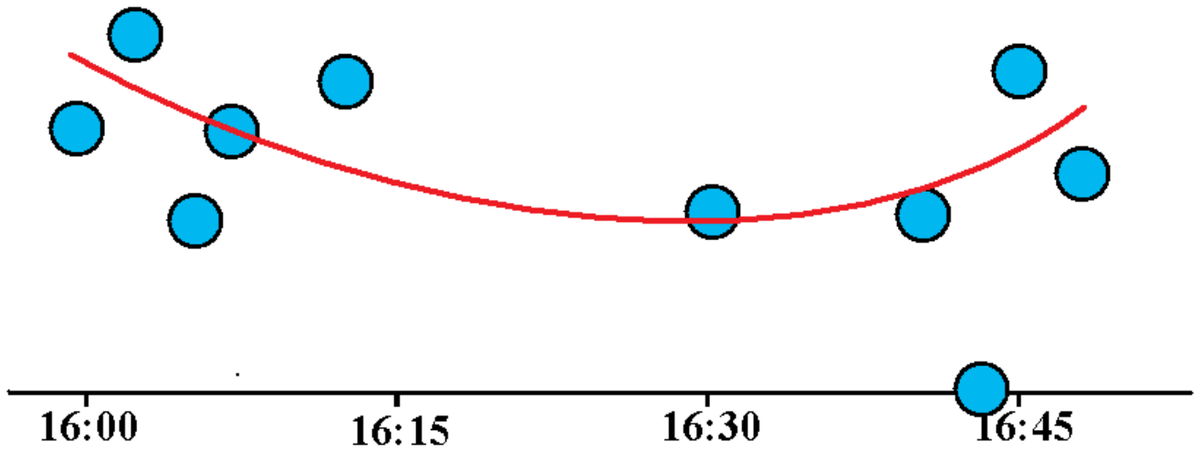
Нужно знать «среднюю» скорость на дороге

в каждый момент времени

т.е. + требование непрерывности

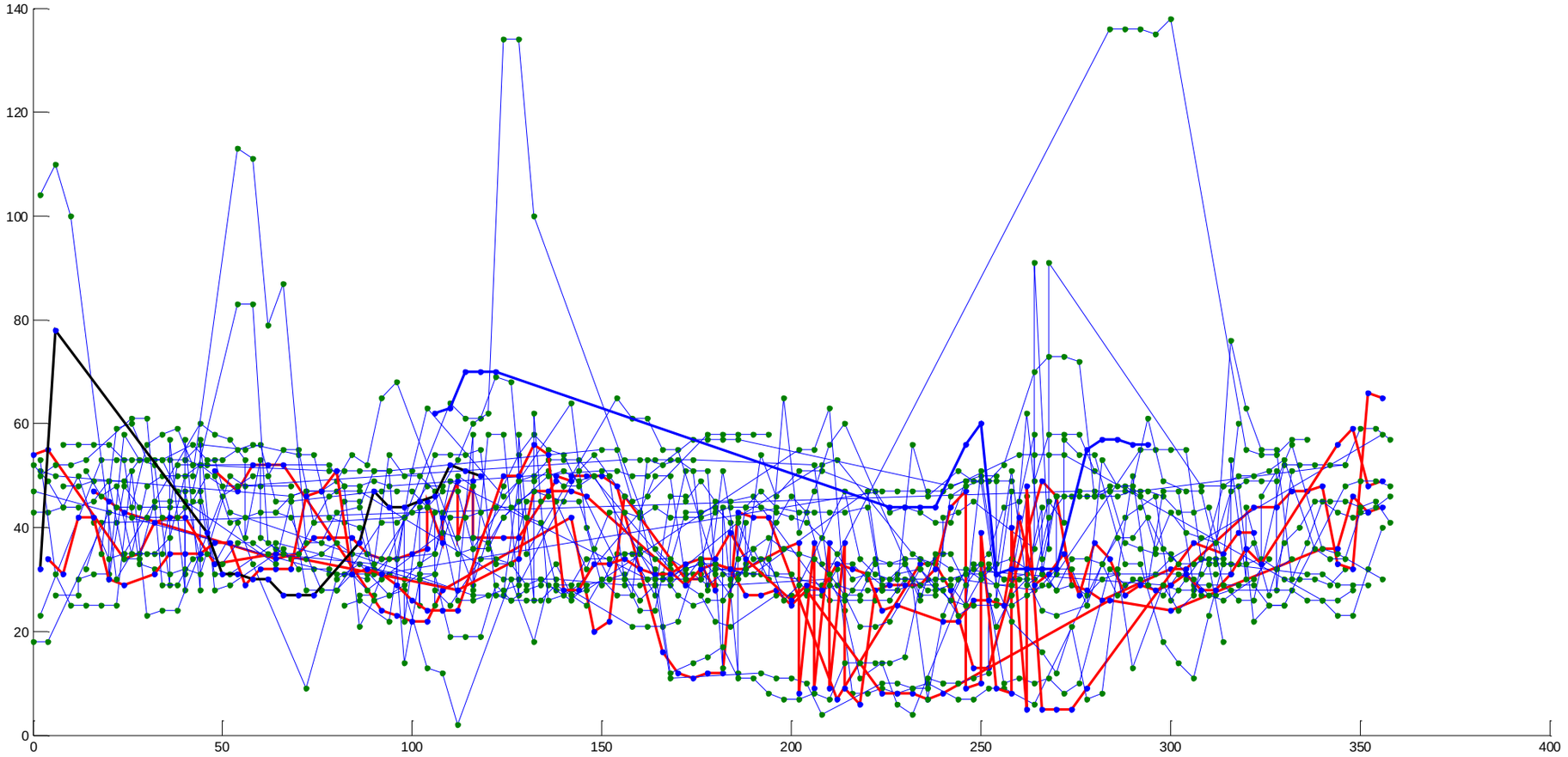
«Существенно двухмерное» усреднение





как бы Вы получали такой график?

Как выглядят данные



Чёрный – наш день,

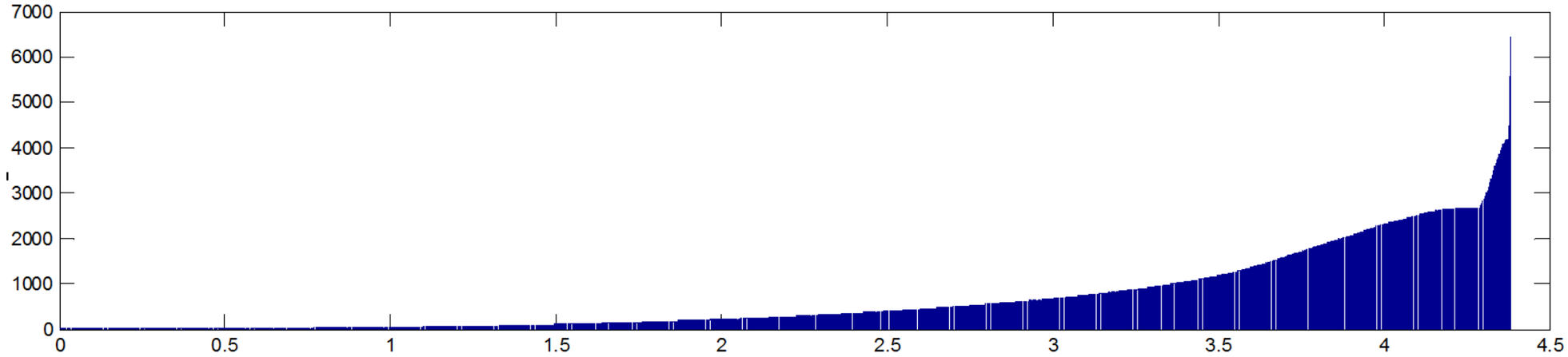
Красный – этот день недели,

Синий – предыдущий день.

Как выглядят данные

ориентированный граф дорог

Распределение длин дорог

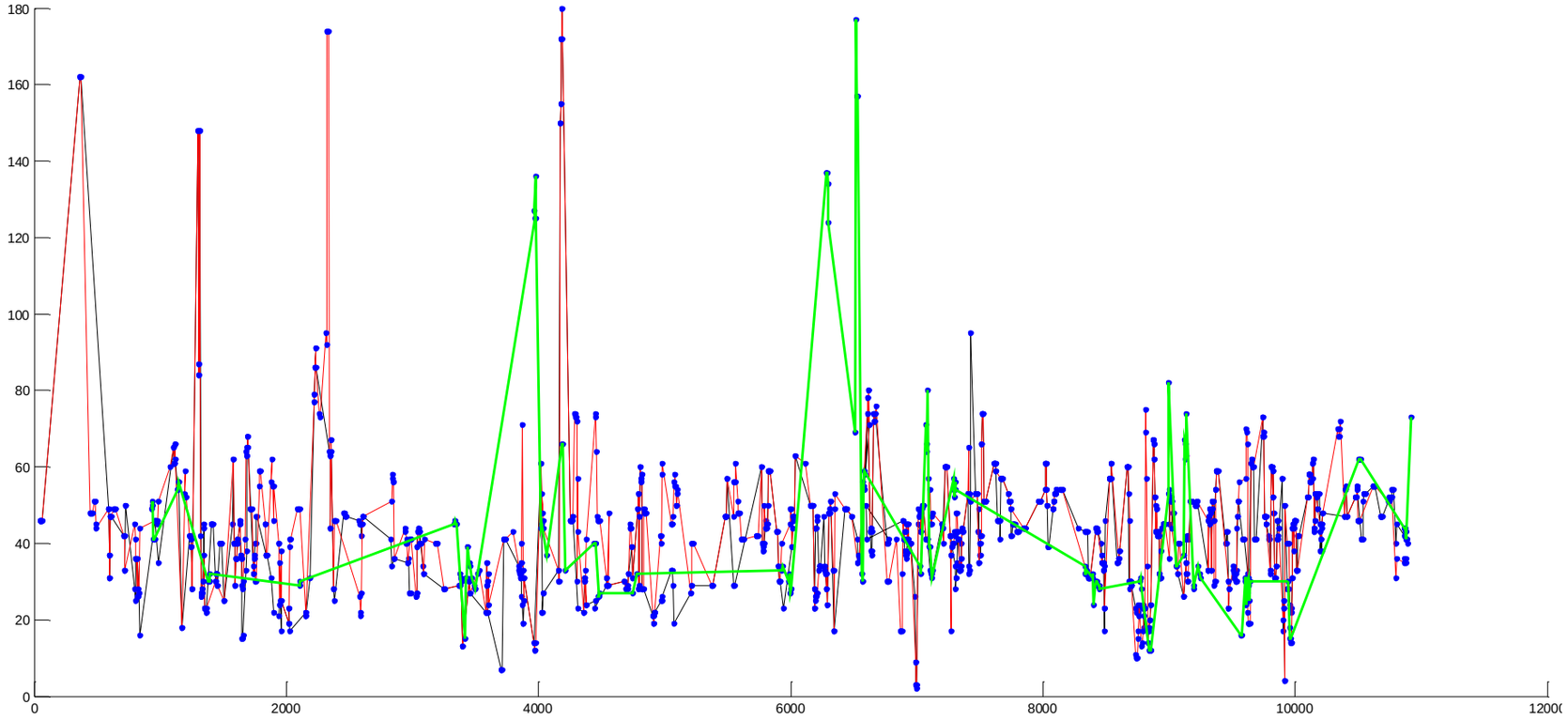


опять нет нормального распределения…

Замечаем странности

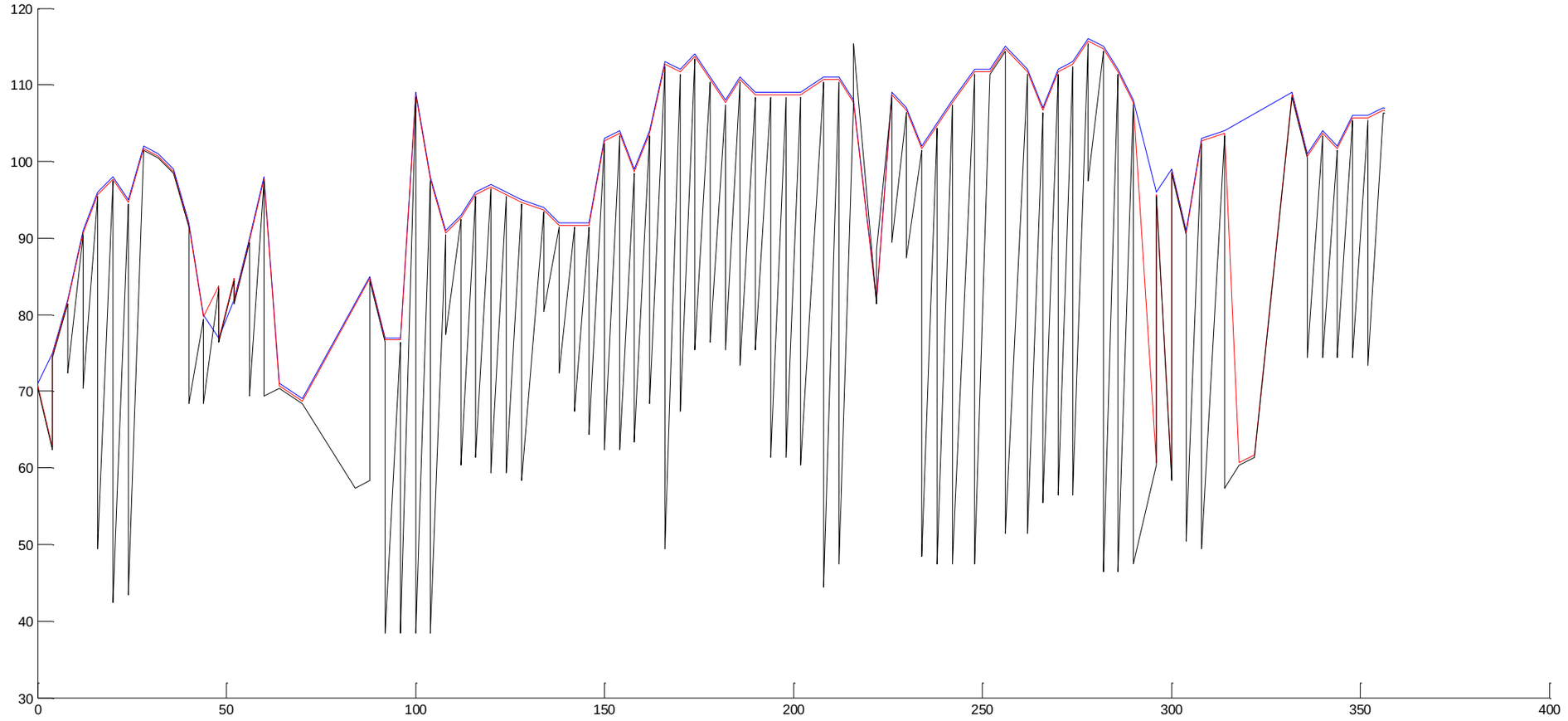
По некоторым дугам статистика совпадает или почти совпадает

Скорость «теряется» при переходе на другую дугу



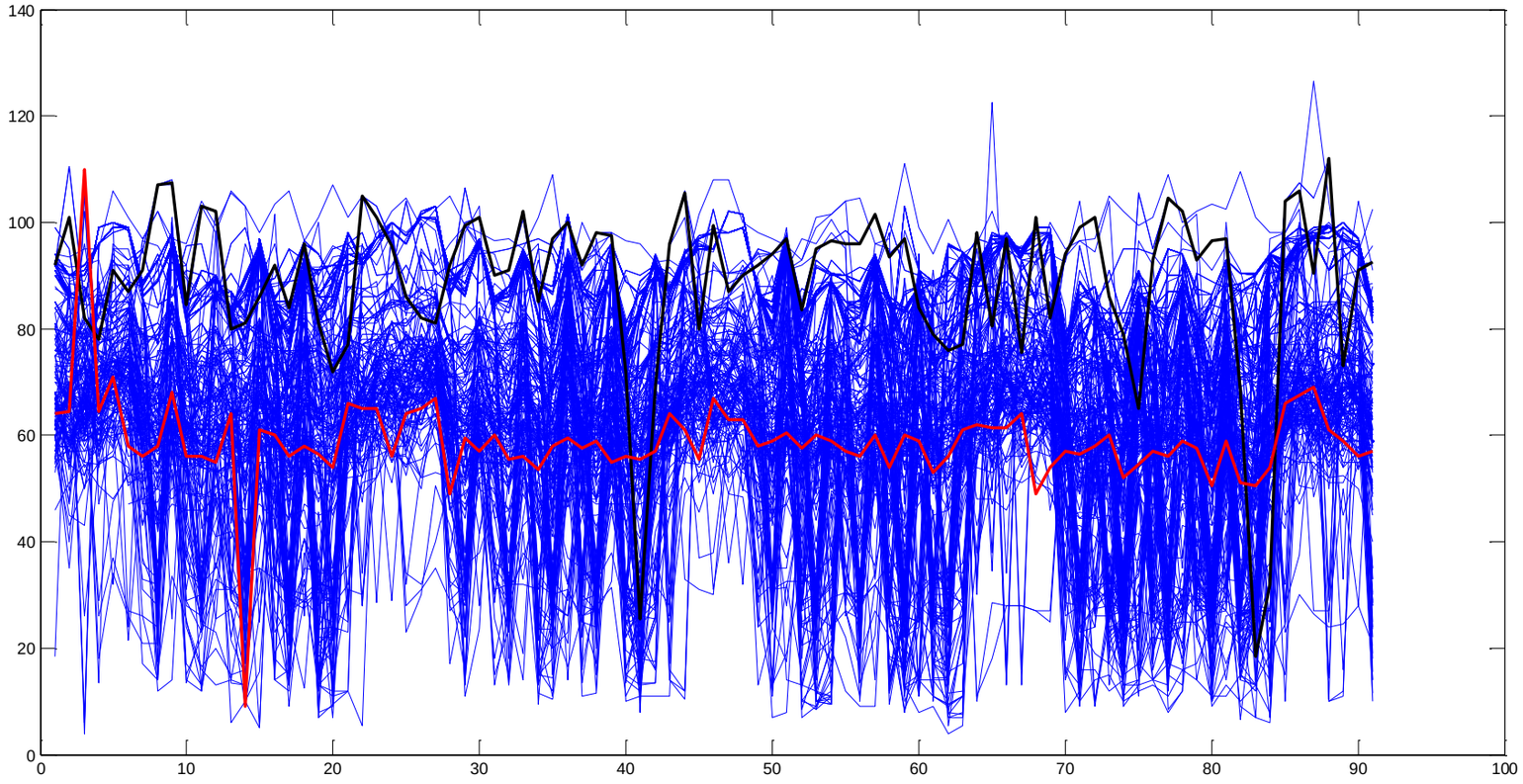
Разные дороги: чёрный, красный, зелёный.

Данные с трёх дуг



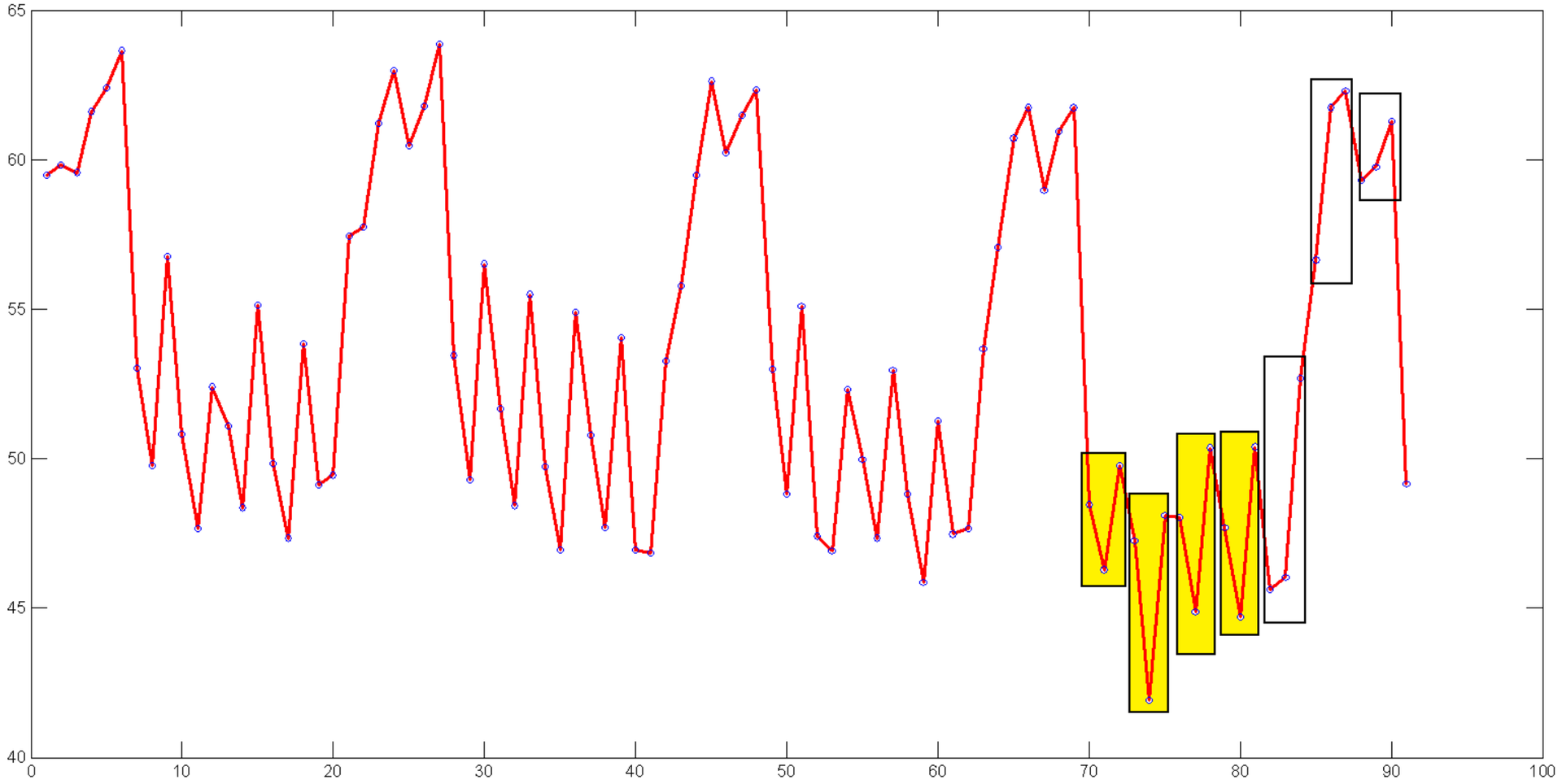
Данные двух дуг совпадают,
+ с половиной данных третьей дуги.

Медианные данные по всем дням



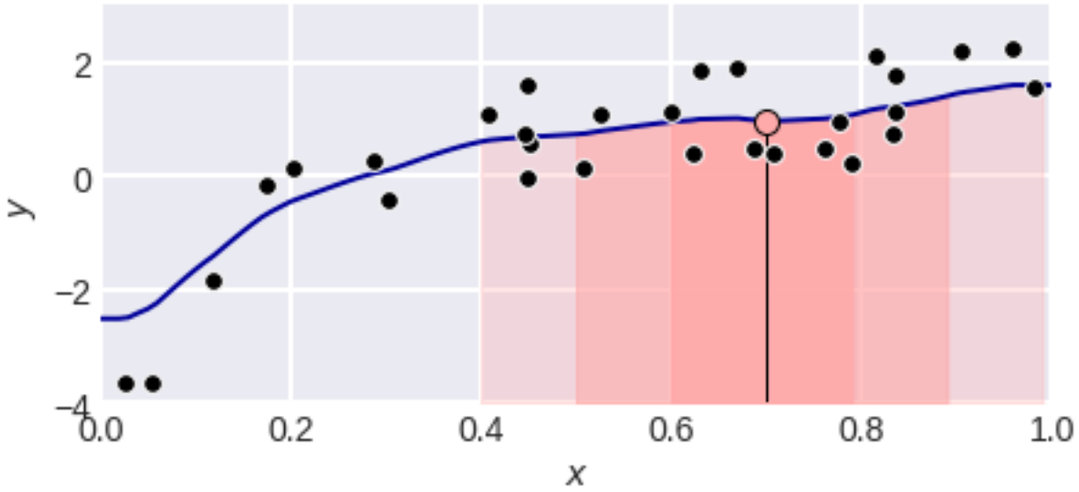
Что можно сказать?

Ответ: Идентифицировать дни недели.



и даже видна идея решения!

Регрессия Надарая-Ватсона (Nadaraya-Watson regression)



ответ – взвешенное усреднение целевых значений

$$a(x) = \frac{w_1(x)y_1 + \ldots + w_m(x)y_m}{w_1(x) + \ldots + w_m(x)}$$

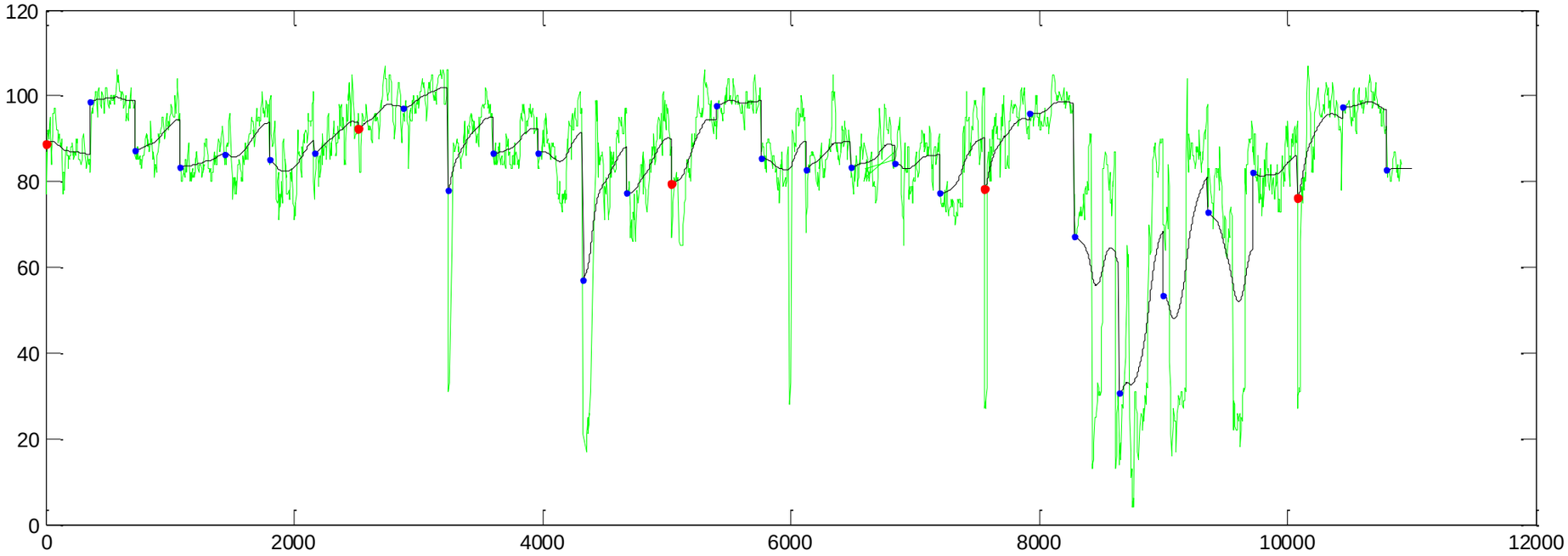
область применения, преимущества и недостатки?

Регрессия Надарая-Ватсона (Nadaraya-Watson regression)

$$a(x) = \frac{w_1(x)y_1 + \ldots + w_m(x)y_m}{w_1(x) + \ldots + w_m(x)}$$

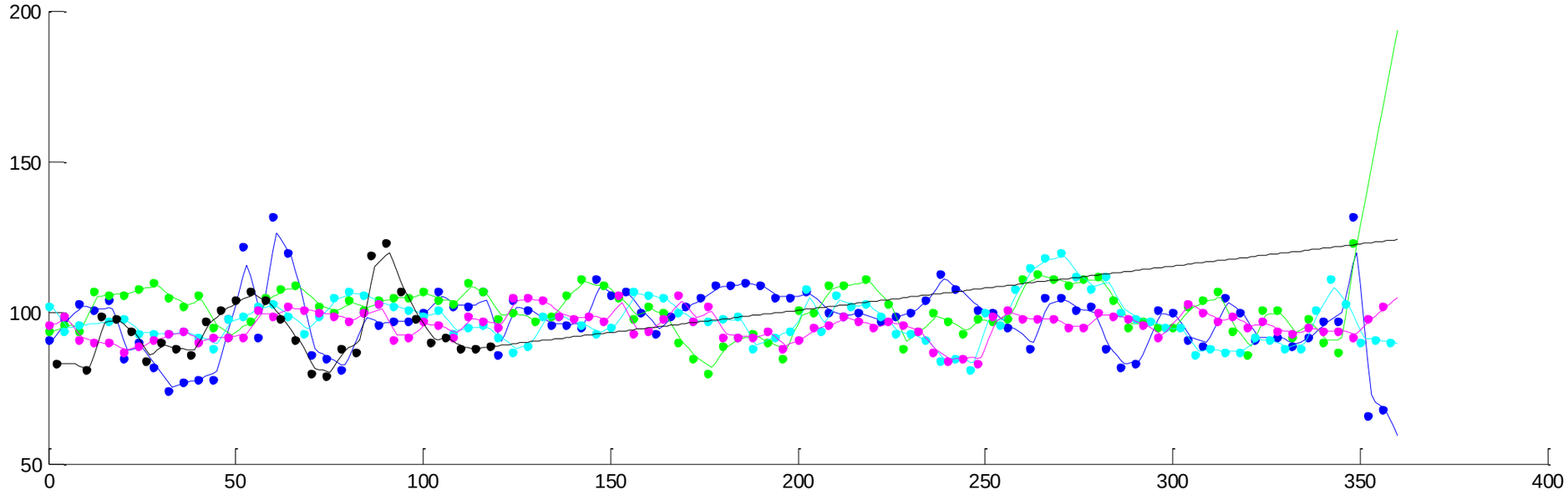
а ведь это тоже весовая схема!

Пример сглаживания



Линейная регрессия Надарая-Ватсона

достаточно опасная…



В обычном

- не проходит через точки

- почти всё считает выбросом

- не экстраполирует

- проблема подбора ширины окна (ядра)

Рецепт по усреднению

Что усреднять:

- **Данные этого дня**
- **Данные вчерашнего дня (тек. день - пн)**
- **Данные этого дня недели**

Как – эксперименты!

Итог

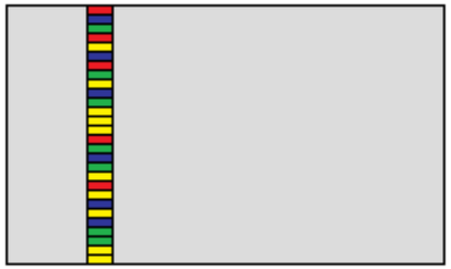
Усреднение нужно для приведения данных в порядок

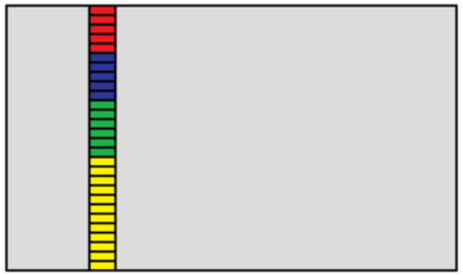
Есть специальные виды регрессии

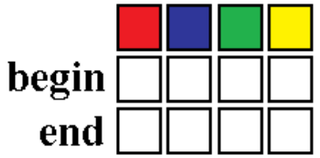
Данные не всегда соответствуют действительности

Полезные приёмы

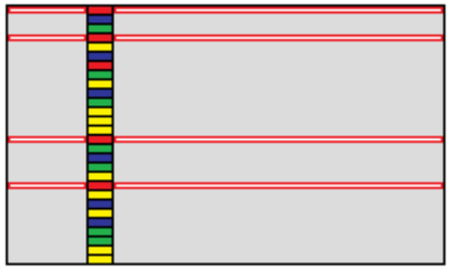
Переупорядочивание данных по факторам…



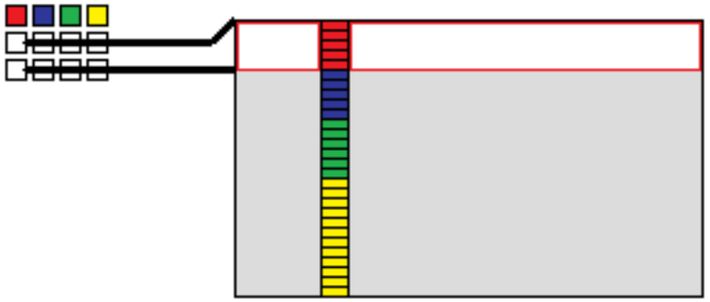




Сортировка



```
M[M[:, i]==a, :]
```



Хранить
начала и концы разных факторов.