Прикладные задачи анализа данных

ОЦЕНКИ СРЕДНЕГО, ВЕРОЯТНОСТИ И ПЛОТНОСТИ. ВЕСОВЫЕ СХЕМЫ

Дьяконов А.Г.

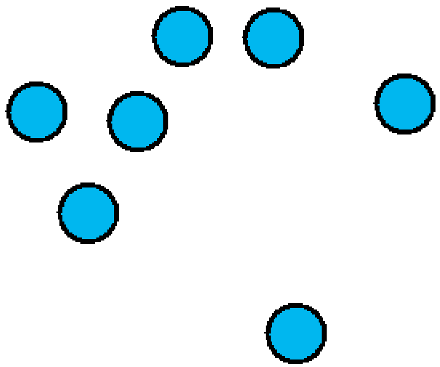
Московский государственный университет имени М.В. Ломоносова (Москва, Россия)

Что такое среднее?

Проблема выбросов

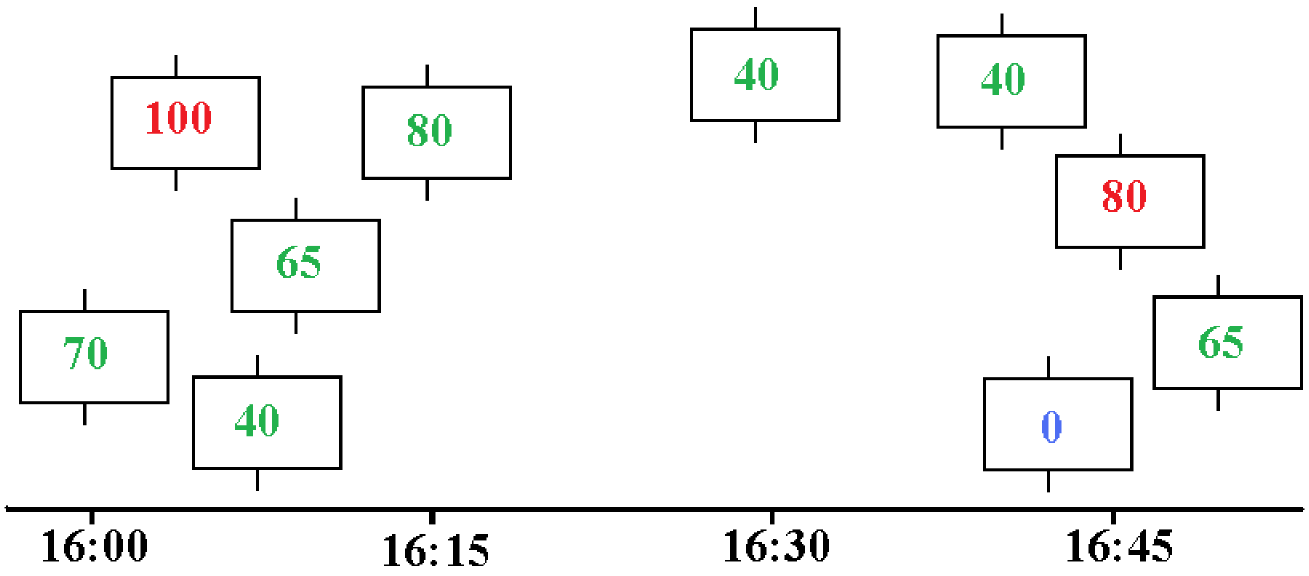
Проблема «виртуальных точек»

Проблема среднего в пространстве



Что здесь «медиана»?

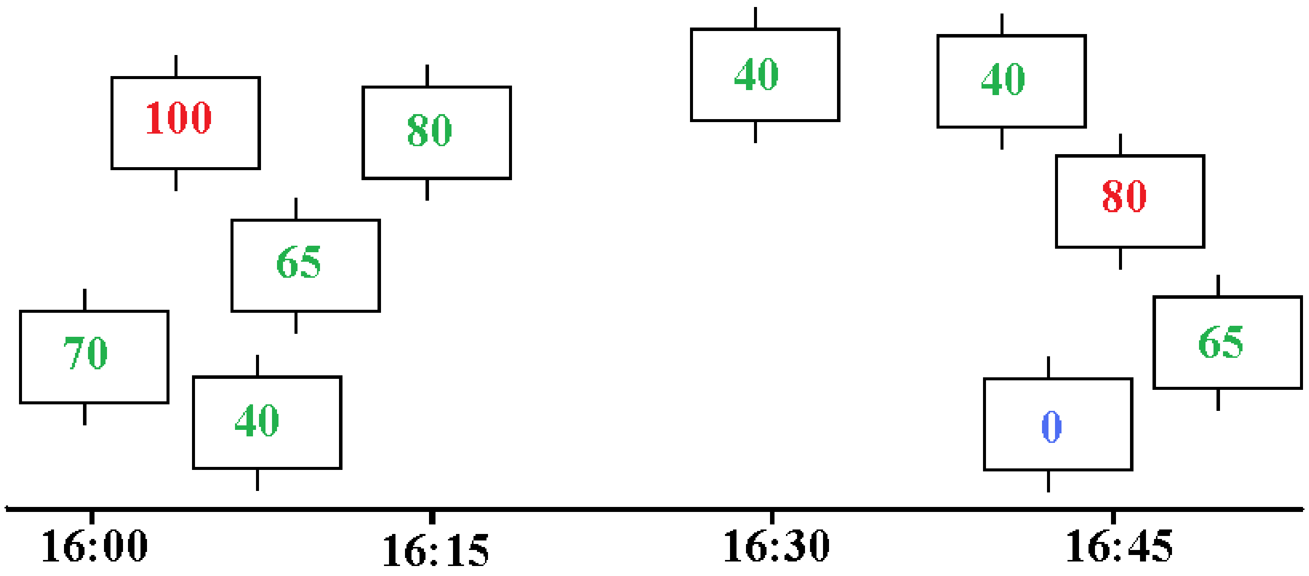
Пример: задача о пробках

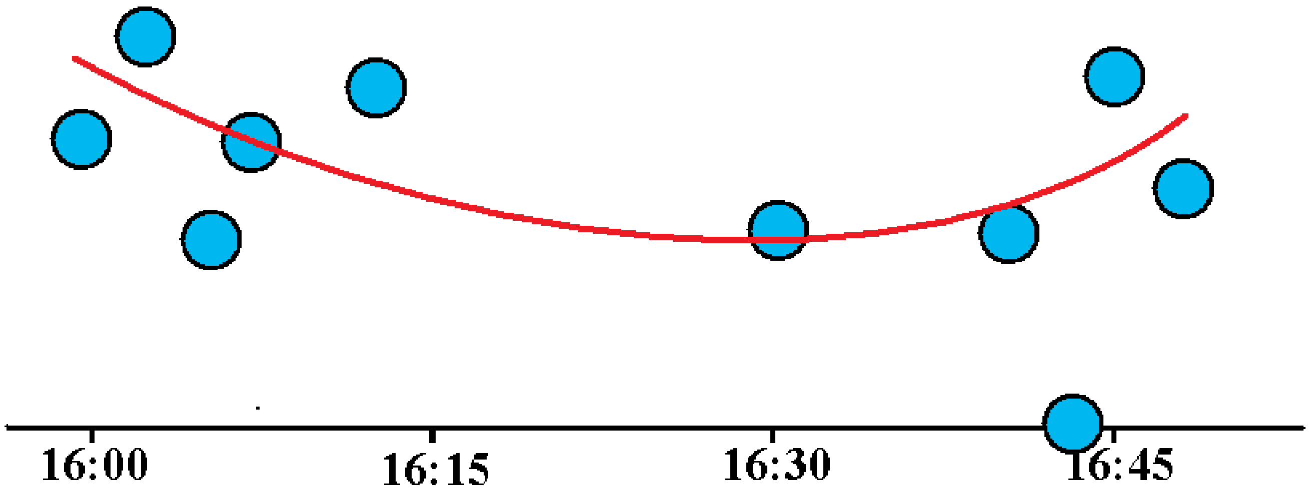


Нужно знать «среднюю» скорость на дороге
в каждый момент времени

т.е. + требование непрерывности

«Существенно двухмерное» усреднение





Стандартный способ

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^i$$

Алгоритм Шурыгина

1. Если $m \leq 2$, то пользуемся формулой (*). Выход.

2. Пусть $x^1 \leq \ldots \leq x^m$ (без ограничения общности).

3. Если $\dfrac{x^1 + x^m}{2} \leq x^2$, то удаляем из выборки x^1. Переходим к п.1 (с соответствующей перенумерацией объектов).

4. Если $\dfrac{x^1 + x^m}{2} \geq x^{m-1}$, то удаляем из выборки x^m. Переходим к п.1 (с соответствующей перенумерацией объектов).

5. Исключаем из выборки x^1, x^m, но добавляем в неё $\dfrac{x^1 + x^m}{2}$.

Практика: часто забываем о выбросах

Что минимизирует «среднее»

$$\sum_{i=1}^{m}(x^i-\mu)^2 \qquad \mu=\frac{1}{m}\sum_{i=1}^{m}x^i$$

$$\sum_{i=1}^{m}|x^i-\mu|$$

медиана

Для минимизации можно выбрать «что угодно»

$$\sum_{i=1}^{m}f(x^i,\mu)$$

μ **– оценка минимального контраста**

Оценка минимального контраста

Если после дифференцирования (здесь рассматриваем одномерный случай)

$$\sum_{i=1}^{m}\psi(x^i-\mu)=\sum_{i=1}^{m}(x^i-\mu)\xi(x^i-\mu)=0,$$

для некоторых функций ψ (оценочная функция) и ξ (весовая функция), то часто успешно применяется итеративный способ вычисления параметра μ по формуле

$$\mu=\frac{\sum_{i=1}^{m}x^i\xi(x^i-\mu)}{\sum_{i=1}^{m}\xi(x^i-\mu)}.$$

Принстонский эксперимент 1972 года

подбор различных функций

Мешалкин Л.Д. (1977) предлагал $\psi(y) = ye^{-\lambda y^2/2}$**, т.е.** $\xi(y) = e^{-\lambda y^2/2}$**.**

Система уравнений

для их поиска оценок среднего и матрицы ковариации для многомерного распределения:

$$\begin{cases} \sum_{i=1}^{m}(x^i - \mu)e^{-\lambda \cdot q_i/2} = 0, \\ \sum_{i=1}^{m}\left((x^i - \mu)(x^i - \mu)^{\mathrm{т}} - \frac{1}{1+\lambda}C\right) \cdot e^{-\lambda \cdot q_i/2} = 0, \end{cases}$$

$$q_i = (x^i - \mu)^{\mathrm{т}} C^{-1} (x^i - \mu)$$

Обобщение медианы на многомерный случай

$$\mu = \frac{\sum_{i=1}^{m} \frac{x^i}{\sqrt{q_i}}}{\sum_{i=1}^{m} \frac{1}{\sqrt{q_i}}}.$$

итерационный алгоритм

[см. Шурыгин]

Ещё пример функционала для минимизации

$$\mu = \frac{2}{q}\sum_{i=1}^{q}\frac{|\,y(x^i) - A(x^i)\,|}{y(x^i) + A(x^i)},$$

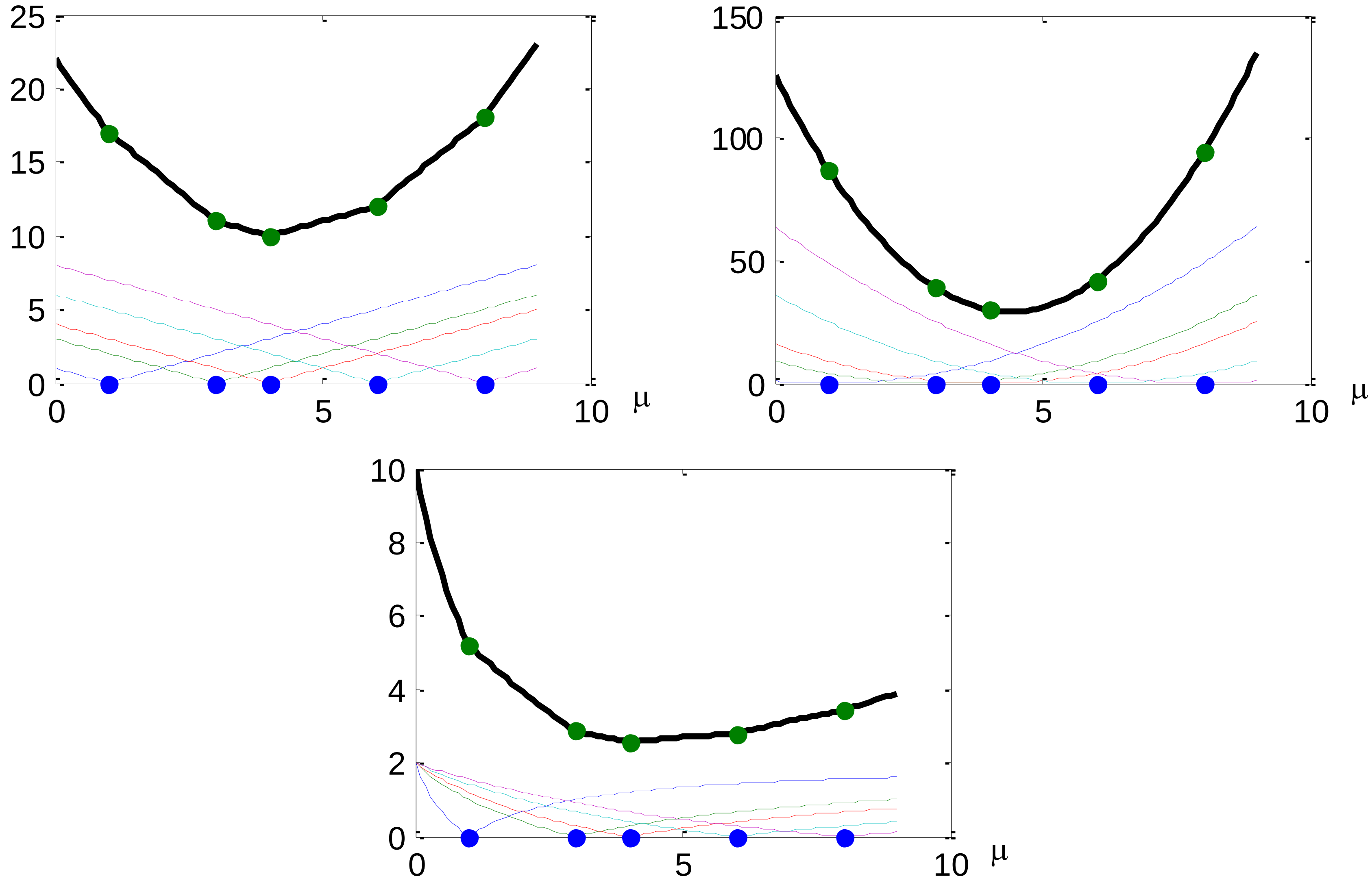
Symmetric mean absolute percentage error (SMAPE or sMAPE)

1 – 2	100 – 101	0 – 1
SMAPE = 67%	SMAPE = 1%	SMAPE = 200%

Начальники не знают, что такое проценты…

Применение SMAPE – прогноз временных рядов

Вопрос: что это за графики?



Практика: придумывать не функционал, а среднее

Среднее по А.Н.Колмогорову

$$\varphi^{-1}\left(\frac{\varphi(x_1)+\ldots+\varphi(x_n)}{n}\right)$$

среднее арифметическое $\varphi(x)=x$

среднее геометрическое $\varphi(x)=\log x$

среднее гармоническое $\varphi(x)=x^{-1}$

среднее квадратическое $\varphi(x)=x^2$

где медиана и мода?

что такое среднее по Коши?

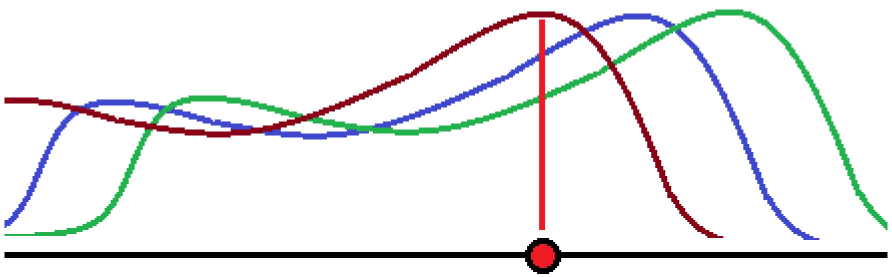
Оценивание вероятности

тоже, в некотором смысле, усреднение

Метод максимального правдоподобия

Есть выборка $x_1,\ldots,x_n$ **какое распределение** $\pi_\alpha(x)$ **?**

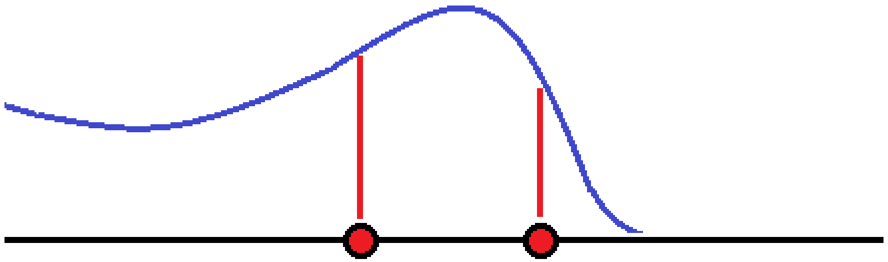
Пусть $m=1$**,** $\pi_\alpha(x)=\pi(x-\alpha)$ **какое распределение выбрать?**



$$\pi_\alpha(x_1)\to\max_\alpha$$

Метод максимального правдоподобия

Пусть $m = 2$



$$\pi_\alpha(x_1) \cdot \pi_\alpha(x_2) \to \max_\alpha$$

Общий случай:

$$\prod_{i=1}^{m} \pi_\alpha(x_i) \to \max_\alpha$$

Как максимизируют?

Случай биномиального распределения

$$\pi_p(x) = \begin{cases} p, & x = 1, \\ 1 - p, & x = 0. \end{cases}$$

$$\Pi = \prod_{i=1}^{n} \pi_p(x_i) = p^m (1-p)^{n-m} \sim$$

$$m \log p + (n-m) \log(1-p)$$

$$(\log \Pi)' = \frac{m}{p} - \frac{(n-m)}{1-p} = 0$$

$$p = \frac{m}{n}$$

Самый очевидный ответ для оценки вероятности!

Оценивание вероятности

тоже, в некотором смысле, усреднение

на практике есть априорная вероятность

$$\frac{m+\lambda\cdot p}{n+\lambda}$$

Вторая особенность практики

Не все эксперименты равнозначны!

$(1+3)/(1+2+1+1+3+2)$

Весовая схема

$$\frac{w_{i_1} + \ldots + w_{i_m}}{w_1 + \ldots + w_n}$$

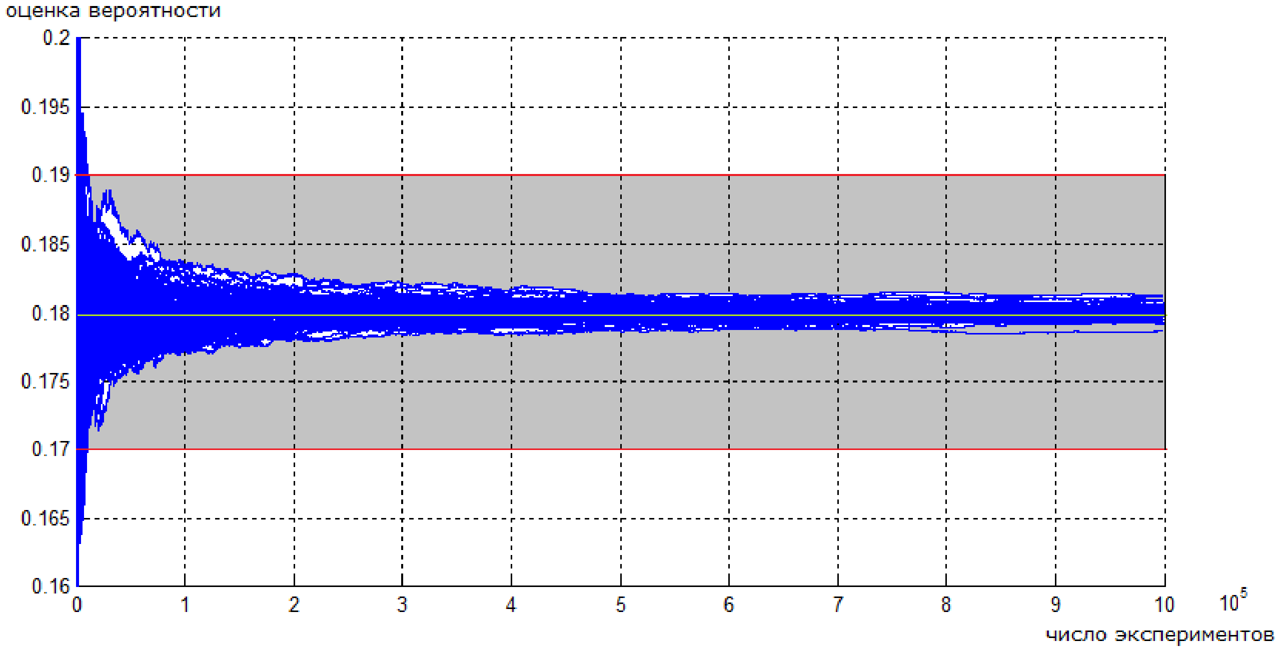
Веса (доверие) возникают даже, где нет эксперта

- есть временная ось
- есть «такие же условия»
- есть кластеры (и схожесть вообще)

Что ещё нужно знать про вероятности

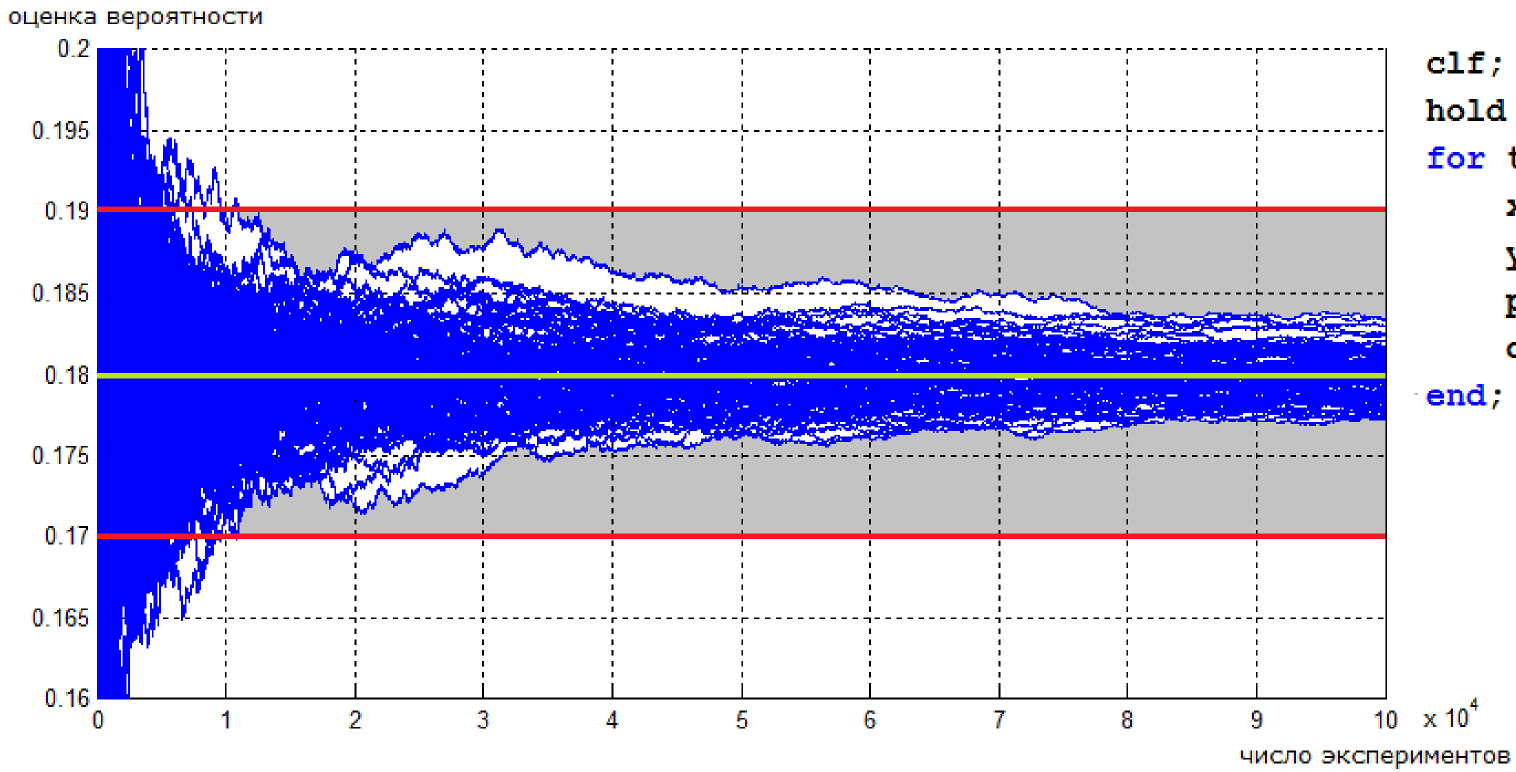
Объёмы выборок

Оцениваем вероятность в схеме Бернулли (неизвестная p=0.18)




```
clf;
hold on;
for t=1:100
    x = +(rand([1 1000000])<=0.18);
    y = cumsum(x)./(1:length(x));
    plot(y);
    disp(t)
end;
```

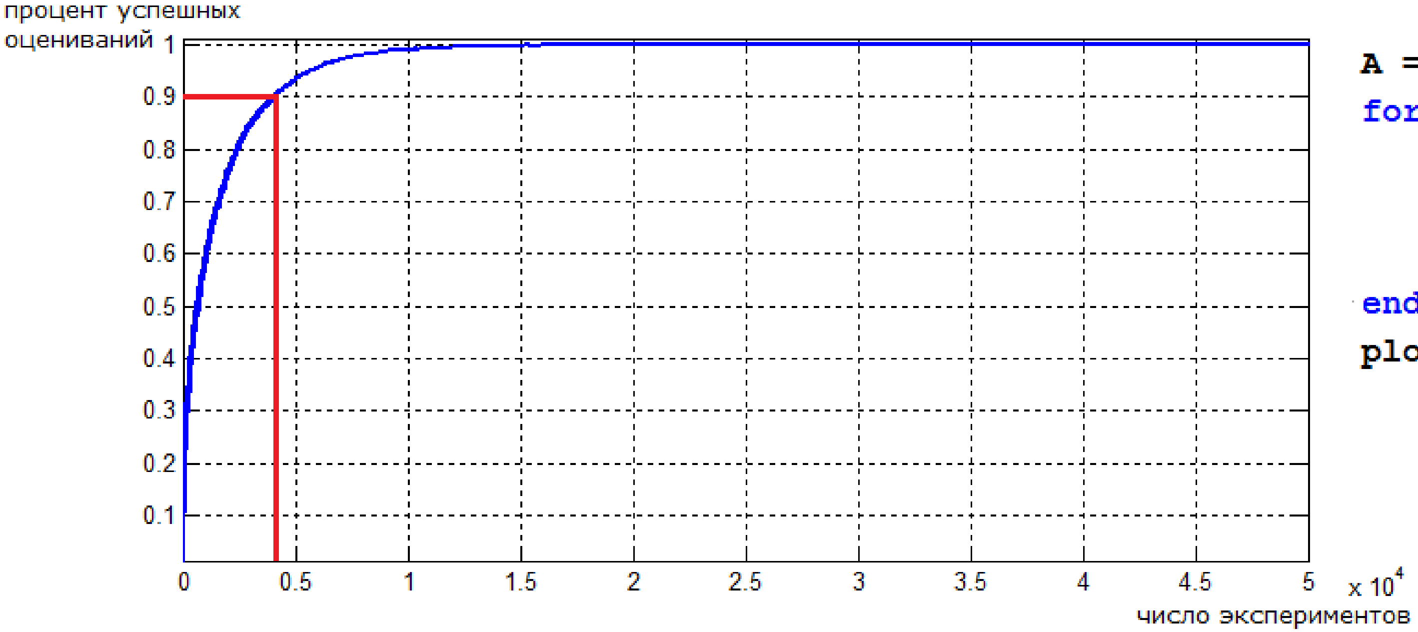
Объём выборки




```
clf;
hold on;
for t=1:100
    x = +(rand([1 1000000])<=0.18);
    y = cumsum(x)./(1:length(x));
    plot(y);
    disp(t)
end;
```

Выборки 10000 достаточно, но это чтобы оценить с точность ±0.01

Объём выборки

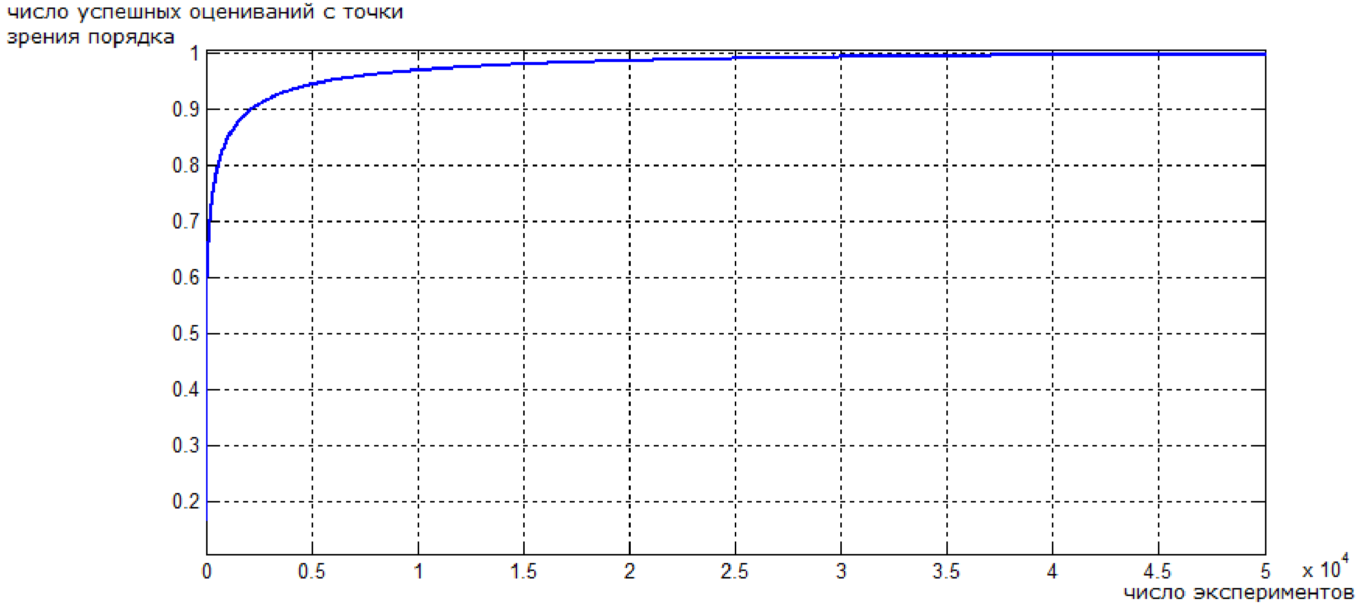



```
A = 0;
for t=1:10000
    x = +(rand([1 1000000])<=0.18);
    y = cumsum(x)./(1:length(x));
    A = A + (abs(y-0.18)<=0.01);
end;
plot(A(1:50000)/10000,'LineWidth', 2)
```

Классика статистики: есть точность,
а есть вероятность того, что мы оценили с этой точностью

Это её график

Объём выборки




```
A = zeros([1 1000000]);
P = linspace(0.16,0.22,12)';
for t=1:10000
    x = bsxfun(@le, rand([12 1000000]), P);
    x = cumsum(x')';
    y = bsxfun(@rdivide, x, 1:1000000);
    B = 0;
    for i=1:11
        for j=(i+1):12
            B = B + (y(i,:)<y(j,:));
        end;
    end;
    A = A+B/66;
end;
plot(A/10000,'LineWidth', 2)
```

Эксперименты в задаче со знаками зодиака

Задача

Прогнозирование визитов покупателей супермаркетов и сумм их покупок

http://www.kaggle.com/c/dunnhumbychallenge/

Международное соревнование «dunnhumby's Shopper Challenge»

Опишем лучший алгоритм из 287

#	Team Name	$10,000 • 279 teams	Score	Entries
1	D'yakonov Alexander (MSU, Moscow, Russia) *		18.83	68
2	NSchneider *		18.67	20
3	Ben Hamner *		18.57	19
4	William Cukierski		18.44	75

Дано: статистика визитов

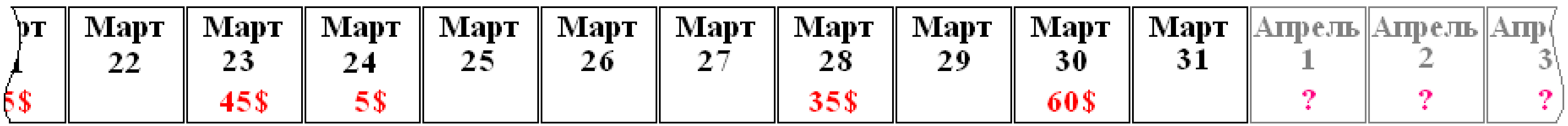
Предсказать: день **первого** визита + сумму покупки с точностью до 10 $

покупатель, дата визита, сумма
56, 2011-06-30, 35.01
56, 2011-06-08, 35.17
56, 2011-07-10, 24.12
56, 2011-07-12, 7.73
57, 2011-05-13, 29.38
57, 2011-05-19, 41.00
...

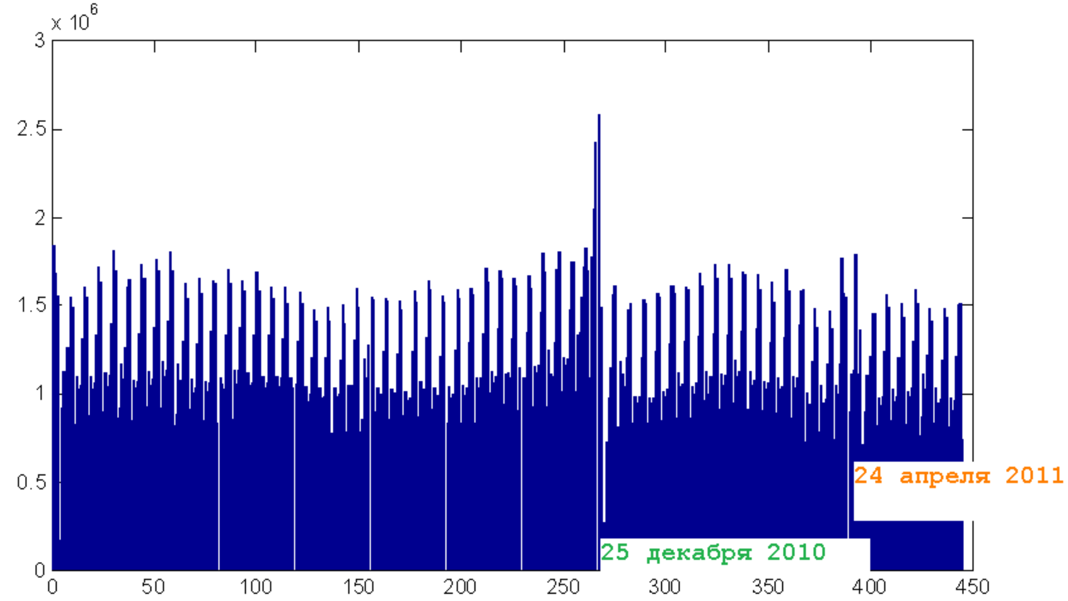
>100000 клиентов customers

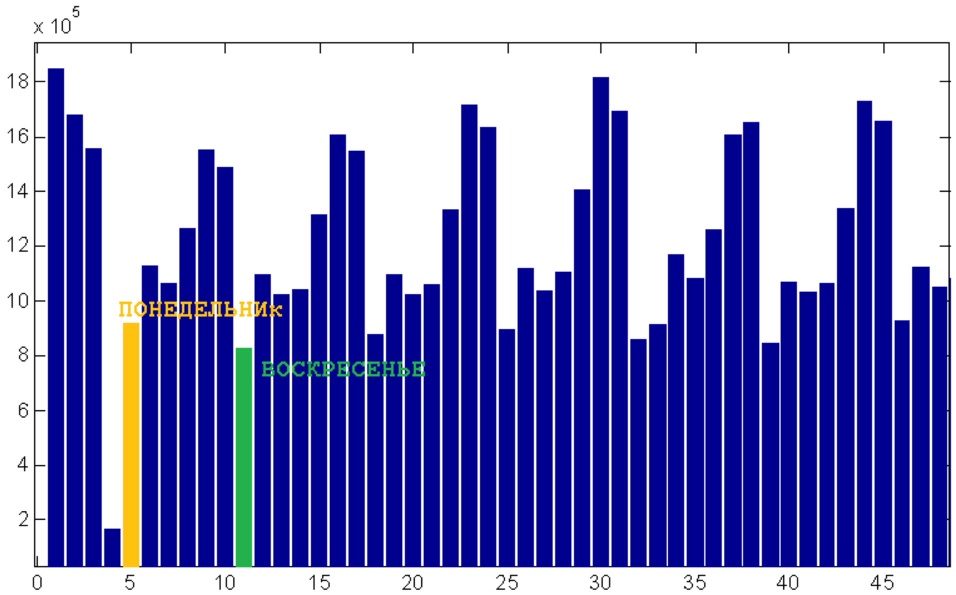
T = 1 год

Статистика визитов одного клиента:

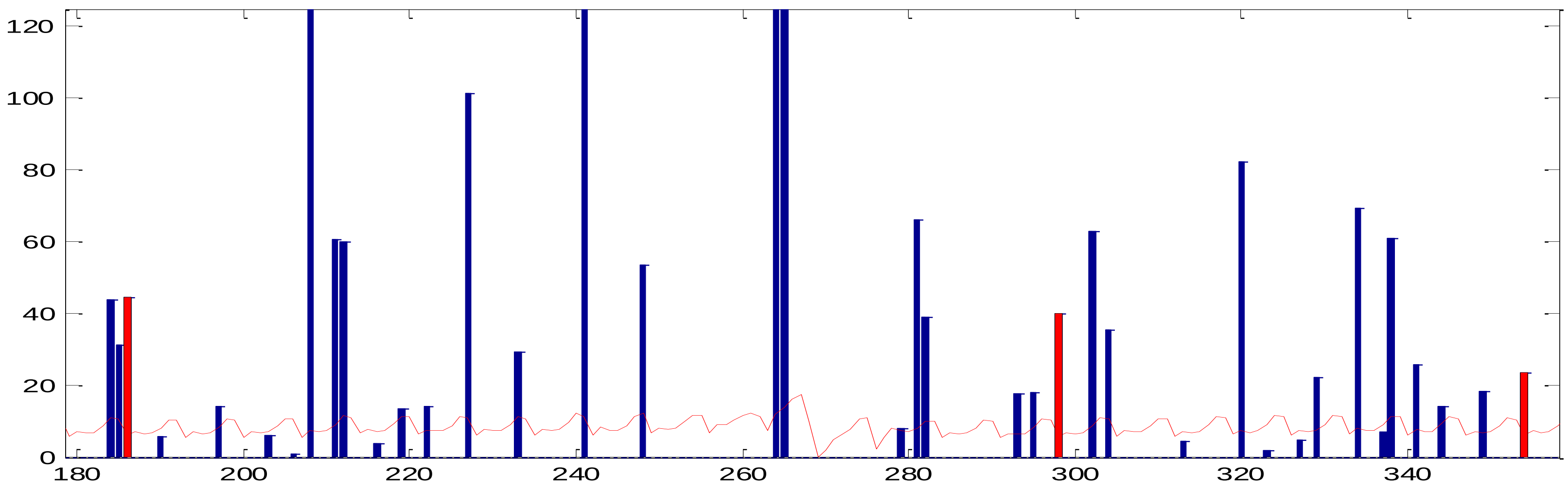


Суммы покупок всех клиентов





Покупки одного клиента

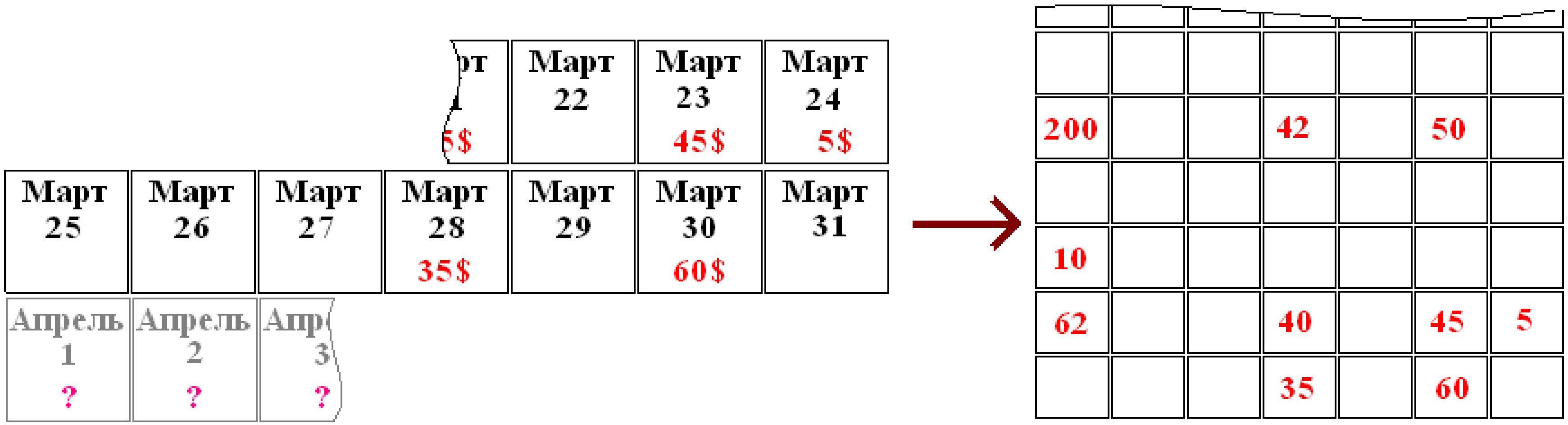


Предположение:

Все клиенты независимы

Будем анализировать каждого клиента отдельно

Разбиение на недели:



Матрица разбивки по неделям:

200			42		50	
10						
62			40		45	5
			35		60	

→

100						10
					18	
52			50			
200			42		50	
10						
62			40		45	5
			35		60	

Сработало устранение пустых недель...

Вероятностная модель поведения клиента

Матрица затрат: $S = \| s_{ij} \|_{d\times 7}$

Матрица визитов: $V = \| v_{ij} \|_{d\times 7}$, $v_{ij} = 1 \Leftrightarrow s_{ij} > 0$.

Вероятности визитов

оценки вероятностей... первых визитов

100						10
					18	
52			50			
200			42		50	
10						
62			40		45	5
			35		60	
5/N	0	0	4/N	0	4/N	2/N

вероятности визитов

5/N ((N-5)/N)·0=0

((N-5)/N)·1·0 = 0

((N-5)/N)·1·1·(4/N) ...

вероятности первых визитов

$$p_1 \qquad \tilde{p}_1 = p_1$$

$$p_2 \qquad \tilde{p}_2 = (1 - p_1) p_2$$

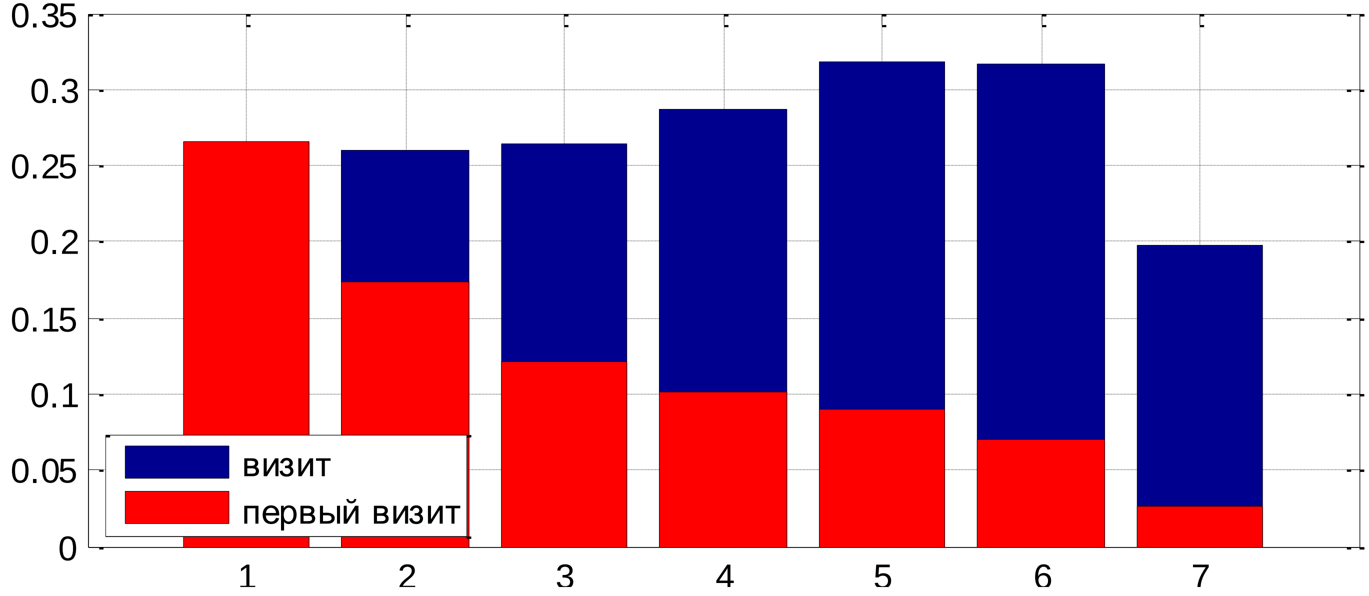
$$\ldots \qquad \ldots$$

$$p_7 \qquad \tilde{p}_7 = \prod_{i=1}^{6} (1 - p_i) p_7$$

Находим максимум вероятностей!

Предположение: Каждый клиент обязательно посетит магазин в течение следующей недели.

Процент визитов и первых визитов на неделе



«Более свежие» данные о клиенте важнее устаревших!

100						10	1
					18		2
52			50				3
200			42		50		4
10							5
62			40		45	5	6
			35		60		7

$$\frac{6+5+4+3+1}{7+6+\ldots+1}$$

вероятности визитов

$$\frac{7+6+4+3}{7+6+\ldots+1}$$

Весовые схемы!

Взвешенная схема оценки вероятности:

$$p_j = \sum_{i=1}^{d} w_i v_{ij},$$

$$w_1 \geq w_2 \geq \ldots \geq w_d \geq 0, \sum_{i=1}^{d} w_i = 1.$$

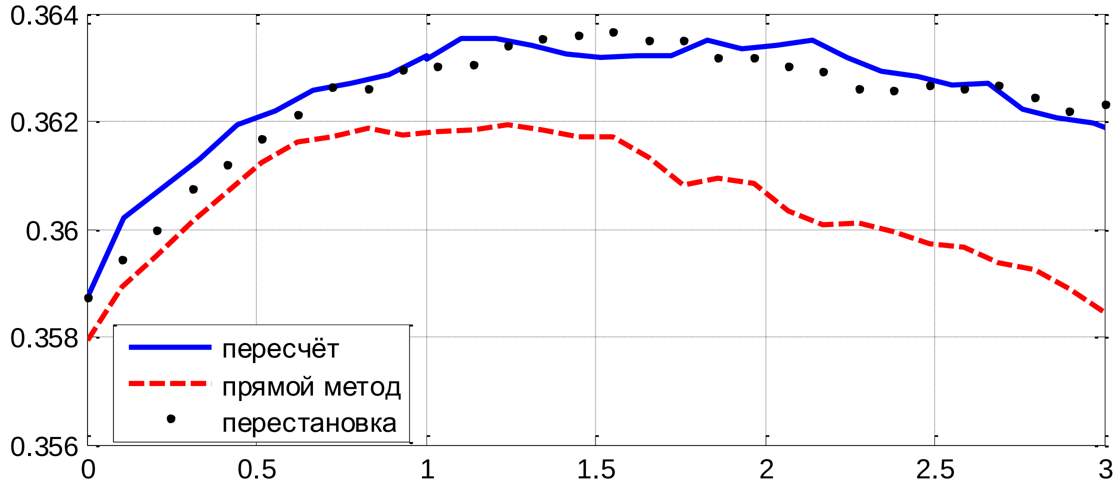
Способы

$$w_i^{\mathrm{N}} = \left(\frac{d-i+1}{d}\right)^{\delta}, \; i \in \{1,2,\ldots,d\},$$

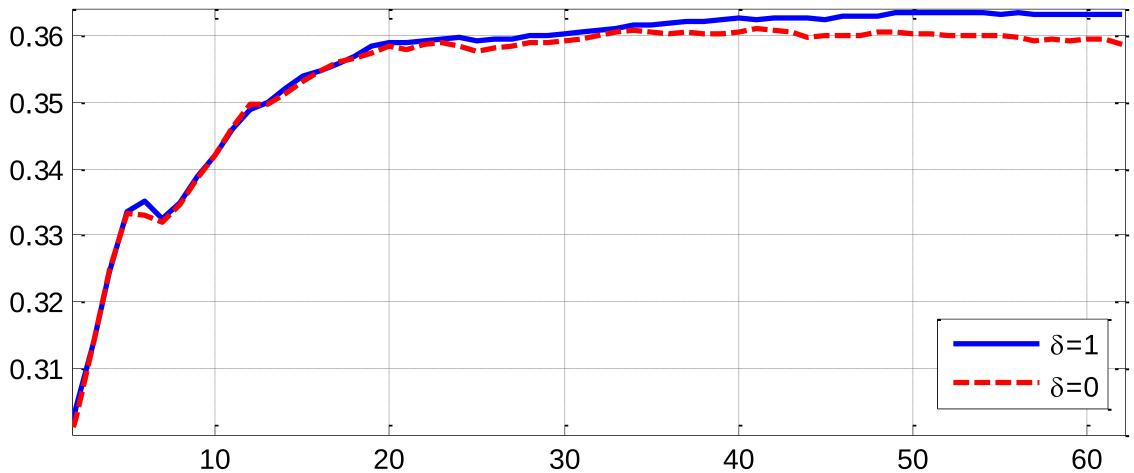
$$w_i = \frac{w_i^{\mathrm{N}}}{\sum_{i=1}^{d} w_i^{\mathrm{N}}}, \; i \in \{1,2,\ldots,d\}.$$ [просто нормировка]

Параметр $\delta \in [0,+\infty)$.

Веса – от равномерных к «агрессивным»

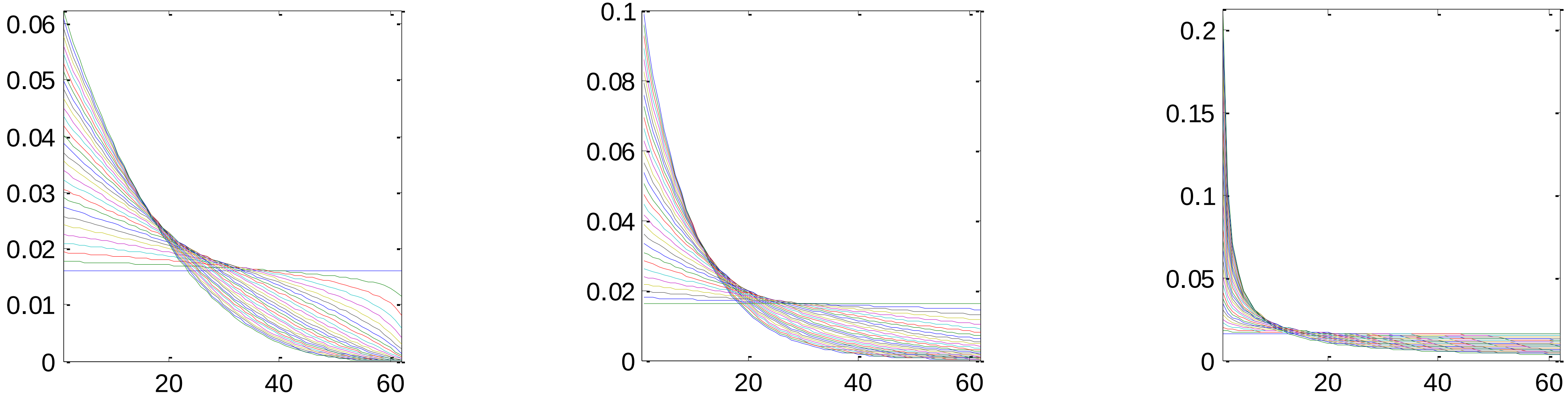


Зависимость качества прогноза от степени δ



Зависимость качества прогноза от числа учитываемых недель

Три разные весовые схемы



вес недели в зависимости от её номера

$$w_i^{\mathrm{N}} = \left(\frac{d-i+1}{d}\right)^{\delta}$$

$$w_i^{\mathrm{N}} = \lambda^i$$

$$w_i^{\mathrm{N}} = \frac{1}{i^{\gamma}},$$

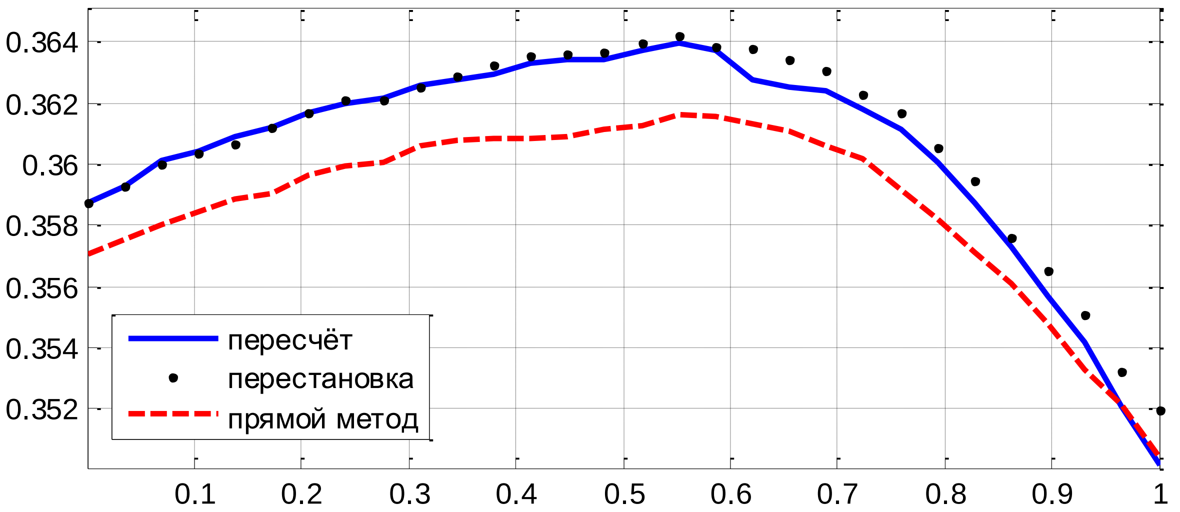
$$\delta \in [0,+\infty)$$

$$\lambda \in (0,1]$$

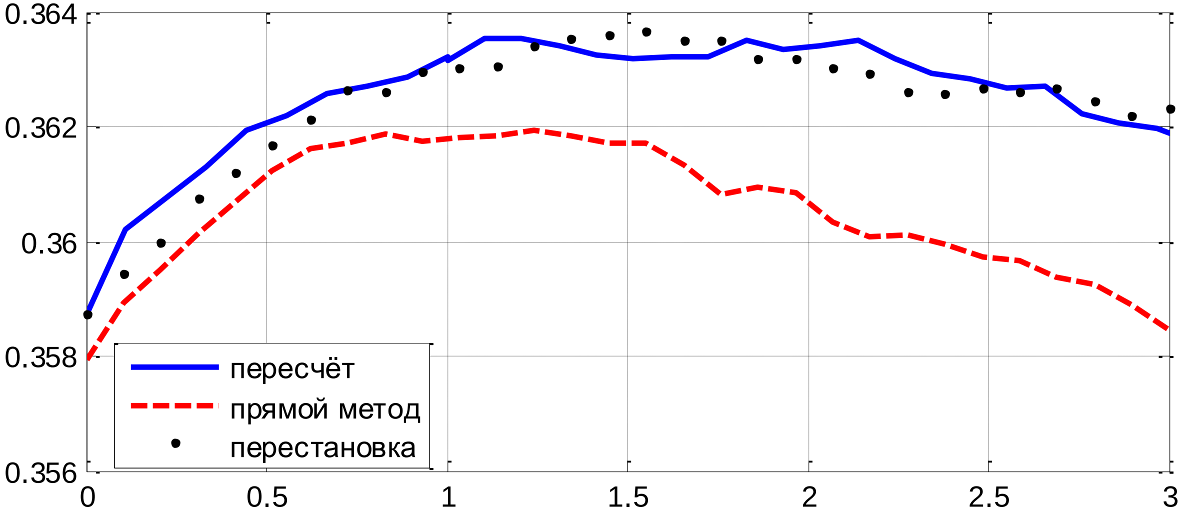
$$\gamma \in [0,+\infty)$$

Вопрос: какие ещё?

Принципиально всё одинаково...



Третья весовая схема



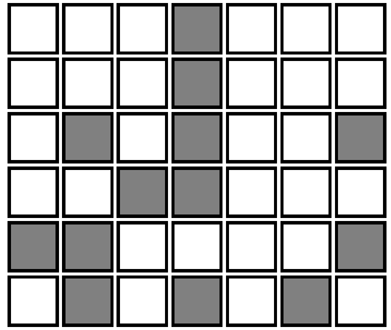
Первая весовая схема

Два способа оценки вероятности первого визита

Прямой метод

$$\tilde{p}_j^2 = \frac{1}{d} \left| \{ i \in \{1,2,\ldots,d\} : v_{i1} = \ldots = v_{i,j-1} = 0, v_{ij} = 1 \} \right|$$

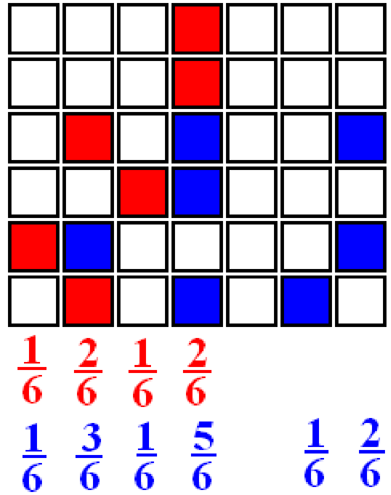
Более естественный, но хуже!



матрица первых визитов

$$V' = \| v'_{ij} \|_{d \times 7}$$

$$\tilde{p}_j^2 = \sum_{i=1}^{d} w_i v'_{ij}$$



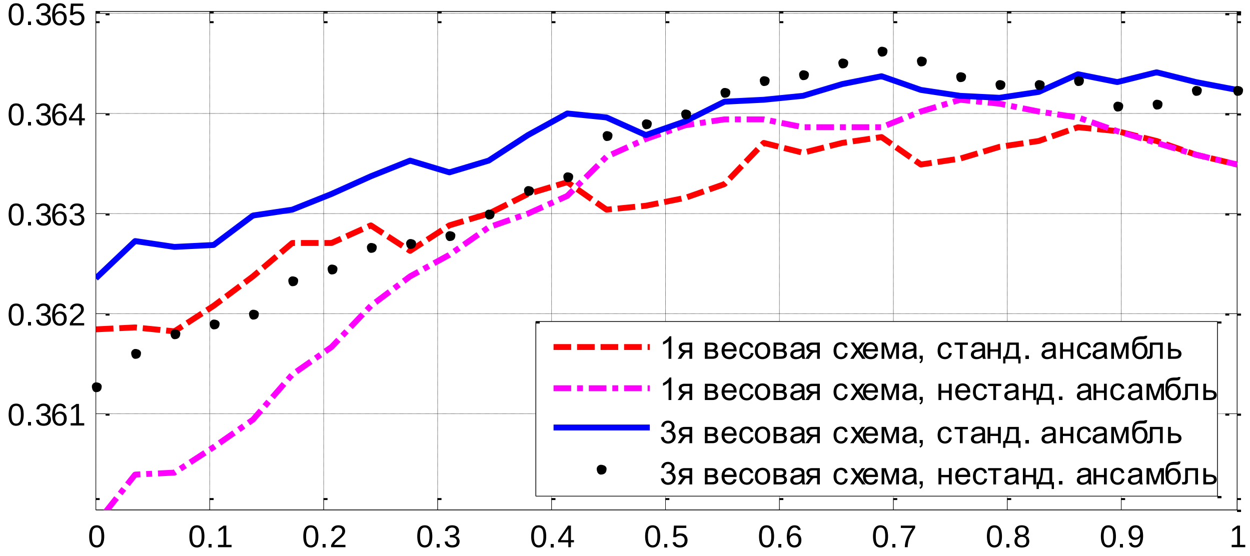
Ансамблирование

«Стандартный ансамбль» – взять выпуклую комбинацию:

$$\tilde{p}_j = \alpha \tilde{p}_j^1 + (1-\alpha)\tilde{p}_j^2, \quad \alpha \in [0,1].$$

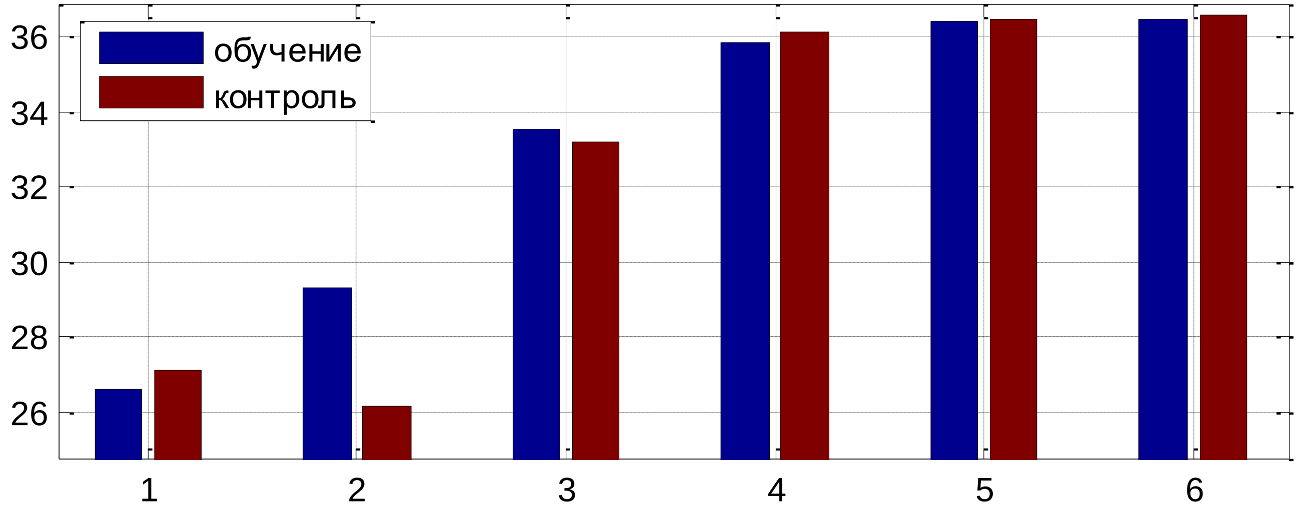
«Нестандартный ансамбль»

$$\alpha p_j + (1-\alpha)\tilde{p}_j^2 = \alpha \sum_{i=1}^{d} w_i v_{ij} + (1-\alpha)\sum_{i=1}^{d} w_i v'_{ij} = \sum_{i=1}^{d} w_i(\alpha v_{ij} + (1-\alpha)v'_{ij})$$



Качество ансамблирования от параметра $\alpha \in [0,1]$

Про переобучение



Качество на обучении и отложенном контроле для шести алгоритмов

1. Константный («клиент придёт на следующий день»),

2. Визит клиента как на прошлой неделе,

3. Вероятности (*) оценены по последним 5 неделям,

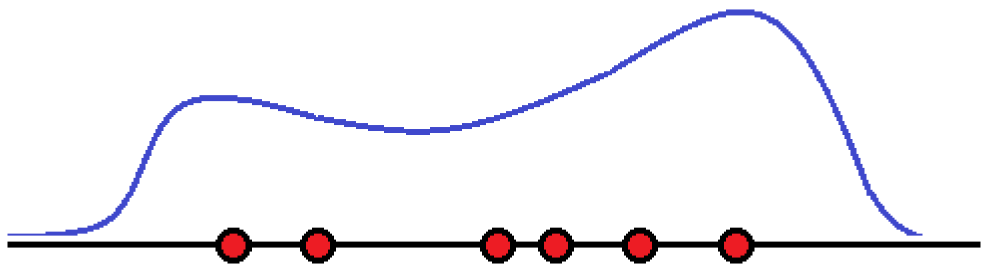
4. Вероятности оценены по всем неделям,

5. Оптимальные значения весов,

6. Оптимальное нестандартное ансамблирование.

Не усложнение, а сглаживание!

Восстановление плотности



Какие методы знаете?

Восстановление плотности

1. Параметрические

Плотность известна с точностью до параметров

2. Непараметрические

Вид плотности не известен

3. Восстановление смесей

Плотность = сумме плотностей

Непараметрические методы восстановления плотности

Метод окон Парзена:

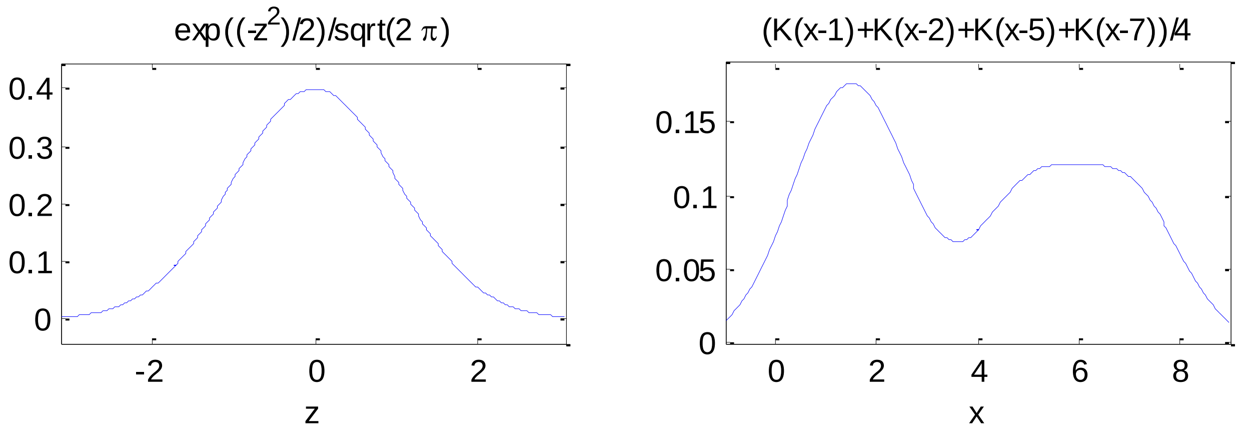
$$\frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{x-x^i}{h}\right),$$

Выборка $x^1,\ldots,x^m$
в пространстве R^d

где $K(x)$ **– функция окна.**

$$K((z_1,\ldots,z_d)) = \begin{cases} 1, & \forall j\in\{1,2,\ldots,d\} \mid z_i \mid \le 0.5 \\ 0, & \text{иначе.} \end{cases}$$

$$K(\tilde{z}) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\tilde{z}^{\mathrm{T}}\tilde{z}}{2}\right)$$

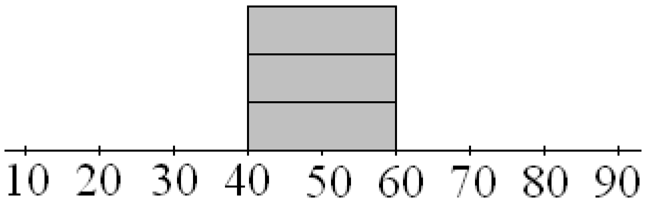


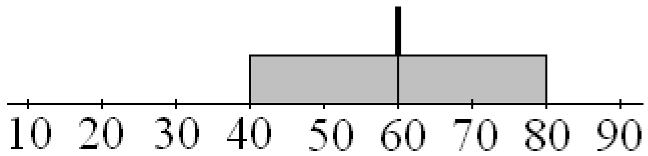
Предсказание суммы покупки

= непараметрическое восстановление плотности по Парзену

«Суммы ступенек» при покупках

50, 50, 50 **50, 70**



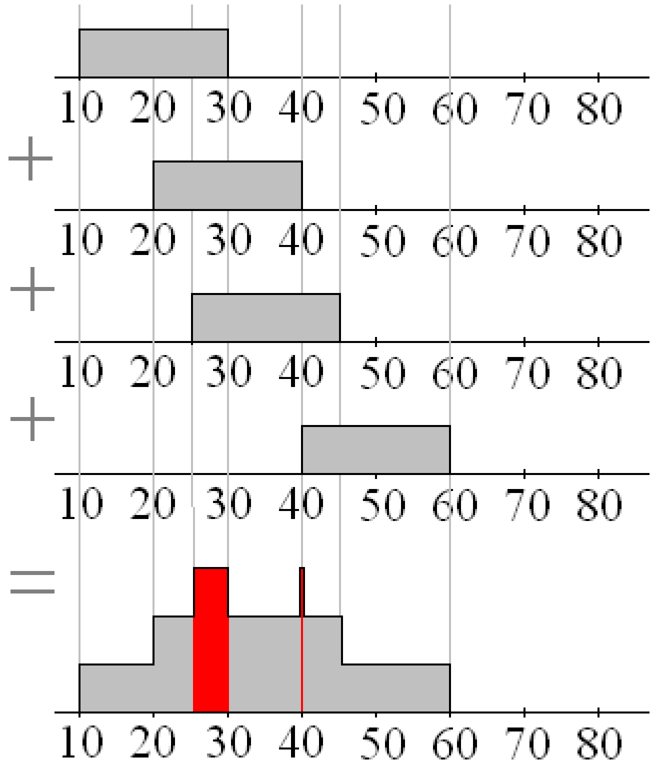


Наилучшая стратегия предсказания суммы при условии, что пользователь ведёт себя как раньше

т.е. это оценка среднего

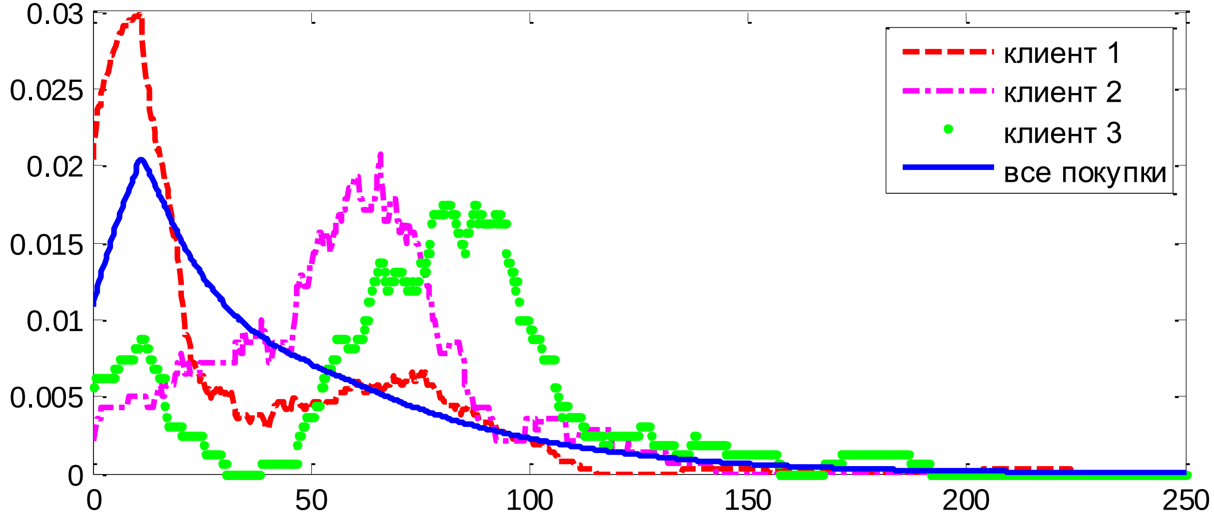
Прогноз с помощью моды

«Суммы ступенек» при покупках 20, 30, 35, 50 –

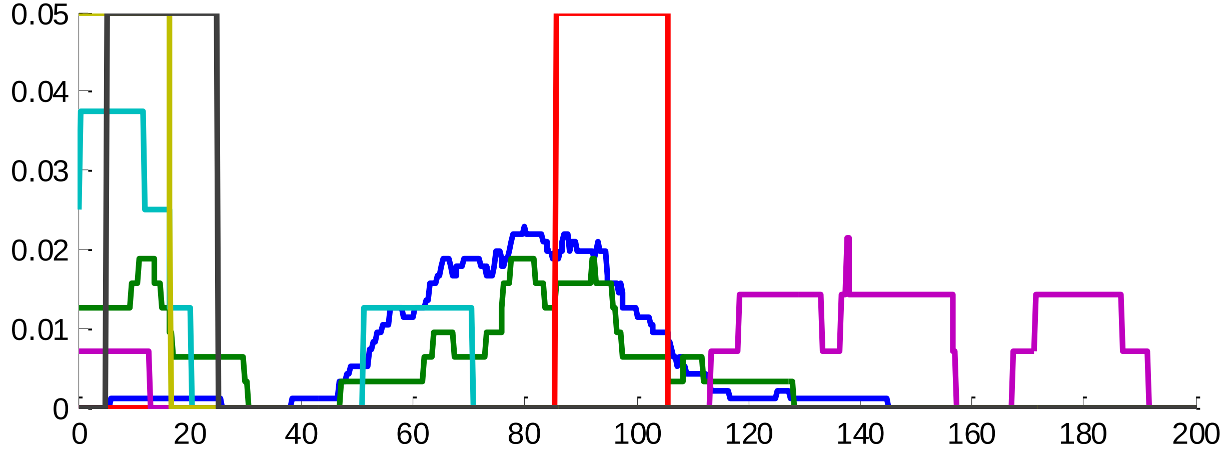


максимум достигается на отрезке [25, 30] и в точке 40.

Как выглядят плотности



Плотности распределения покупок



Плотности покупок одного пользователя в разные дни недели

И здесь сделаем весовую схему!

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} K(|s_i - x|)$$

$$2\int_{0}^{+\infty} K(x)\,dx = 1.$$

$$K(|s-x|) = \begin{cases} 1/2\varepsilon, & |s-x| \le \varepsilon, \\ 0, & |s-x| > \varepsilon. \end{cases}$$

Весовая схема:

$$f(x) = \sum_{i=1}^{m} w_i K(|s_i - x|)$$

Весовая схема

учёт времени, дня недели

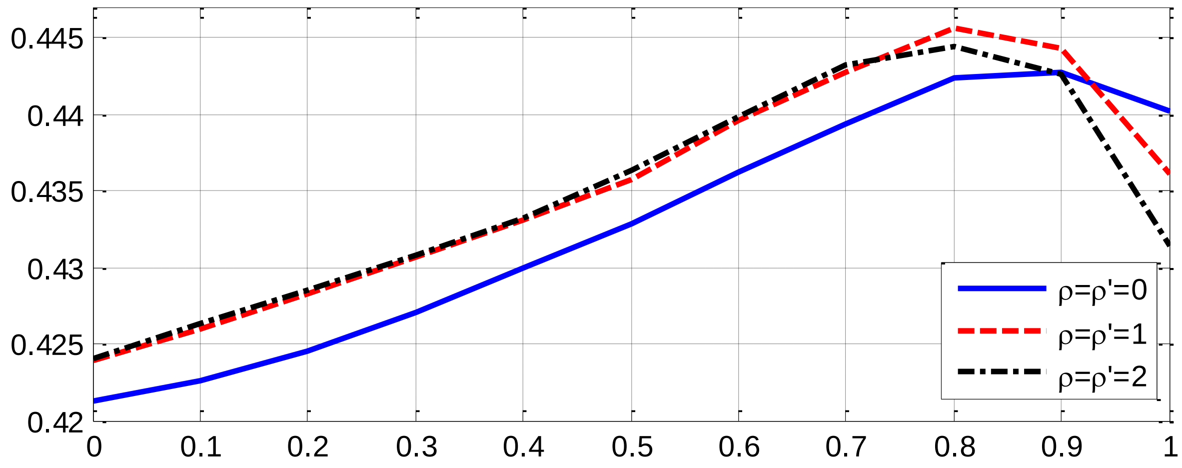
Пусть $s_1,\ldots,s_m$ **– все упорядоченные покупки пользователя,**
$s'_1,\ldots,s'_{m'}$ **– покупки, сделанные в этот день недели.**

Плотность будем восстанавливать для расширенного набора
$s'_1,\ldots,s'_{m'},s_1,\ldots,s_m$**.**

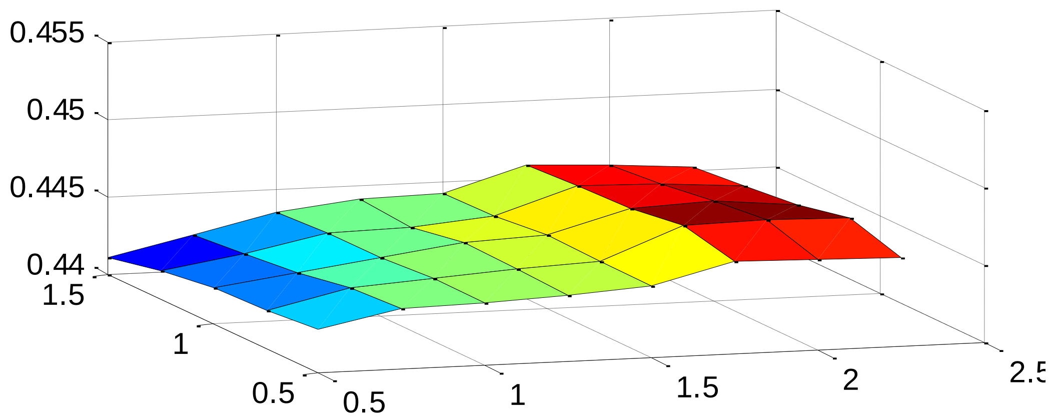
Веса:

$$s'_i \quad \leftrightarrow \quad \beta\frac{(m'-i+1)^{\rho'}}{\sum\limits_{j=1}^{m'} j^{\rho'}} \qquad\qquad s_i \quad \leftrightarrow \quad (1-\beta)\frac{(m-i+1)^{\rho}}{\sum\limits_{j=1}^{m} j^{\rho}}$$

Весовая схема



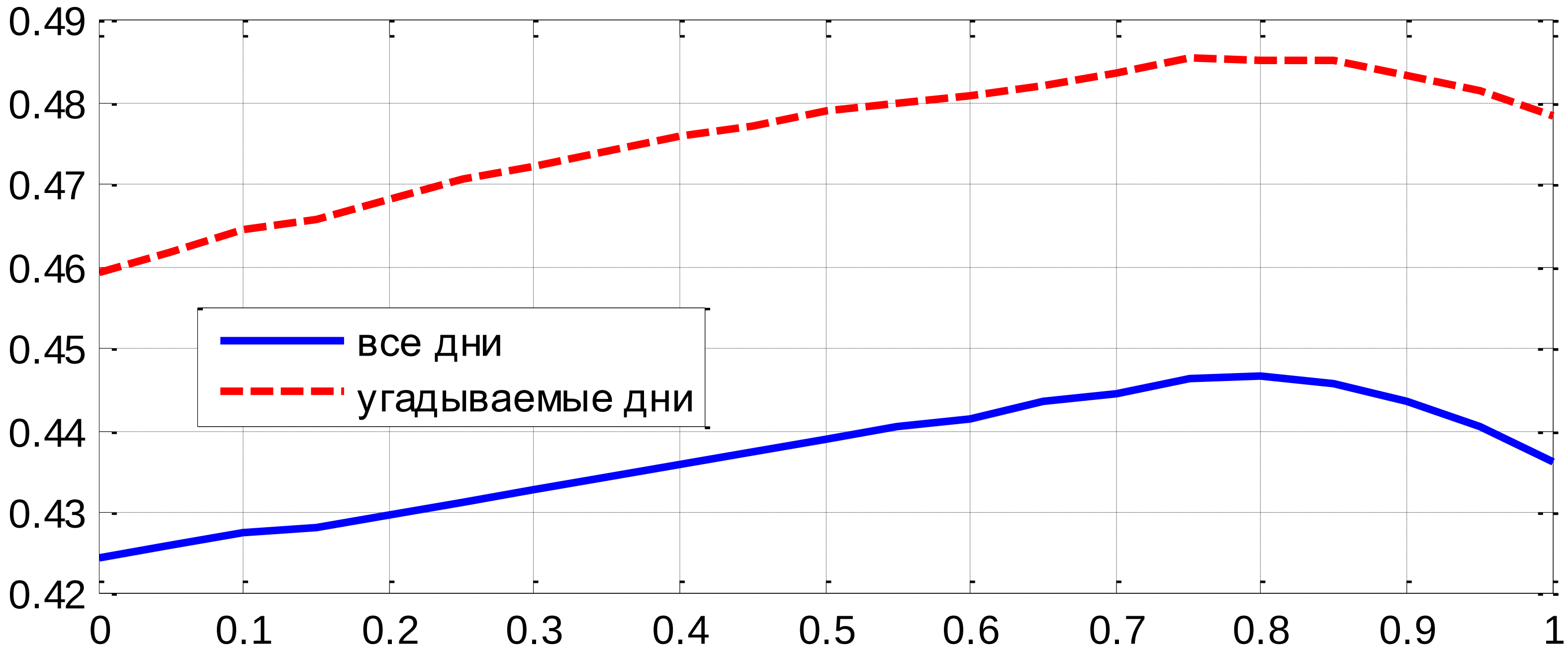
Качество прогноза суммы покупок от параметра β



Качество прогноза в зависимости от степеней при $\beta = 0.8$

Как настраивать, точнее где…

- на всей выборке
- на угадываемых днях (на остальных – бесполезно для функционала)



Качество прогноза суммы покупок от параметра β при $\rho = 0.7$, $\rho' = 1.6$.

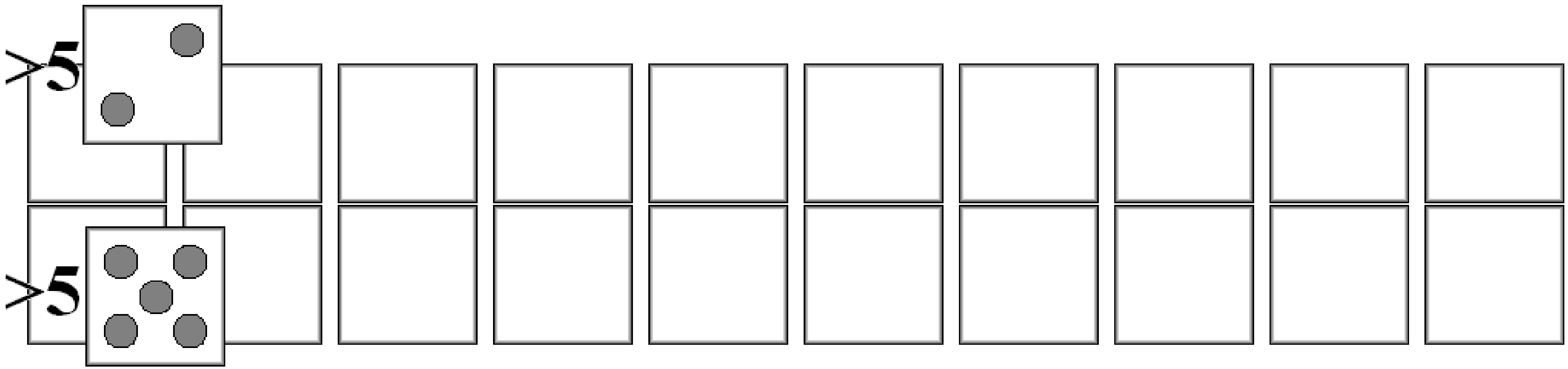
Улучшение алгоритма

Есть:

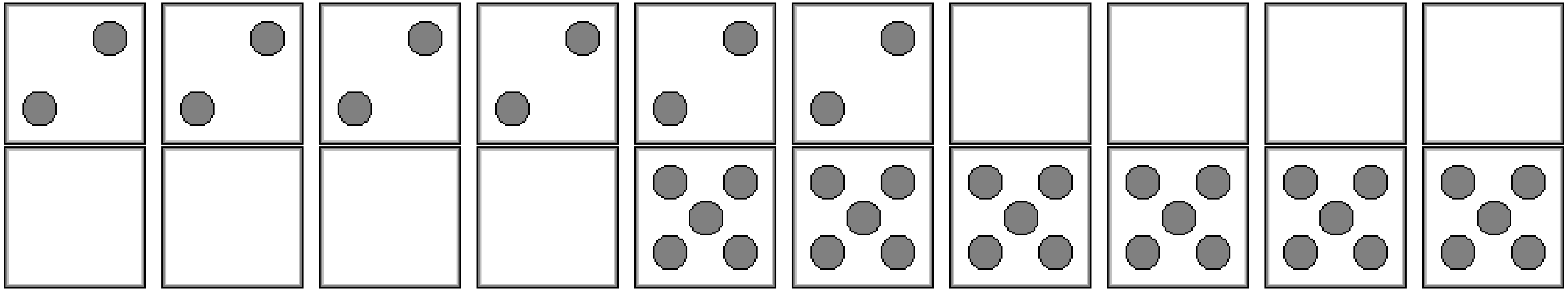
- **метод предсказания даты визита** (вероятностный пересчёт)
- **метод предсказания суммы покупки** (непараметрическое восстановление)

Можно ли так осуществить прогноз?

Все прогнозировали так…



Почему метод работает не очень хорошо...



«И» в условии не означает «И» в решении

Найти день И сумму.

Понедельник: 10\$, 50\$, 220\$, 100\$, 310\$, 5\$, 250\$, 75\$, 500\$

Вторник: 40\$, 42\$, 40\$

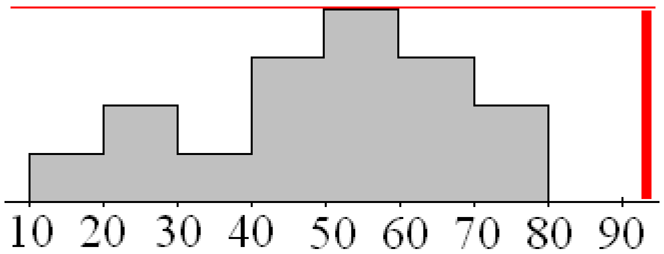
(вероятность угадать день) $*$ (вероятность угадать сумму)

$0.9 * 0.1 = 0.09$

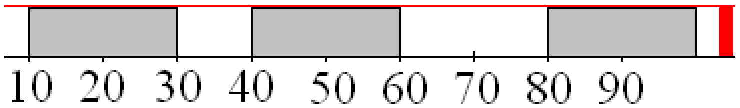
$0.1 * 1 = 0.1$ выгоднее ставить на вторник

Надо: вычислить вероятность угадывания дня и суммы

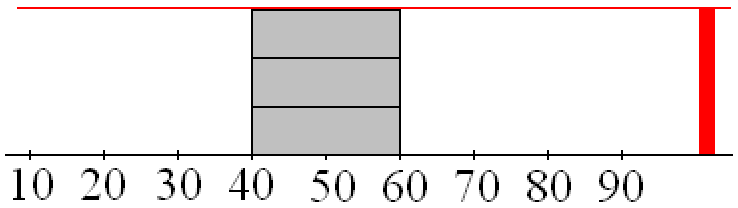
Как вычислить стабильность поведения клиента?



высота графика плотности



низкая стабильность



высокая стабильность

учёт стабильности = улучшение результата

Неполный учёт стабильности

$$\tilde{p}_j(q_j + h) \to \max_j$$

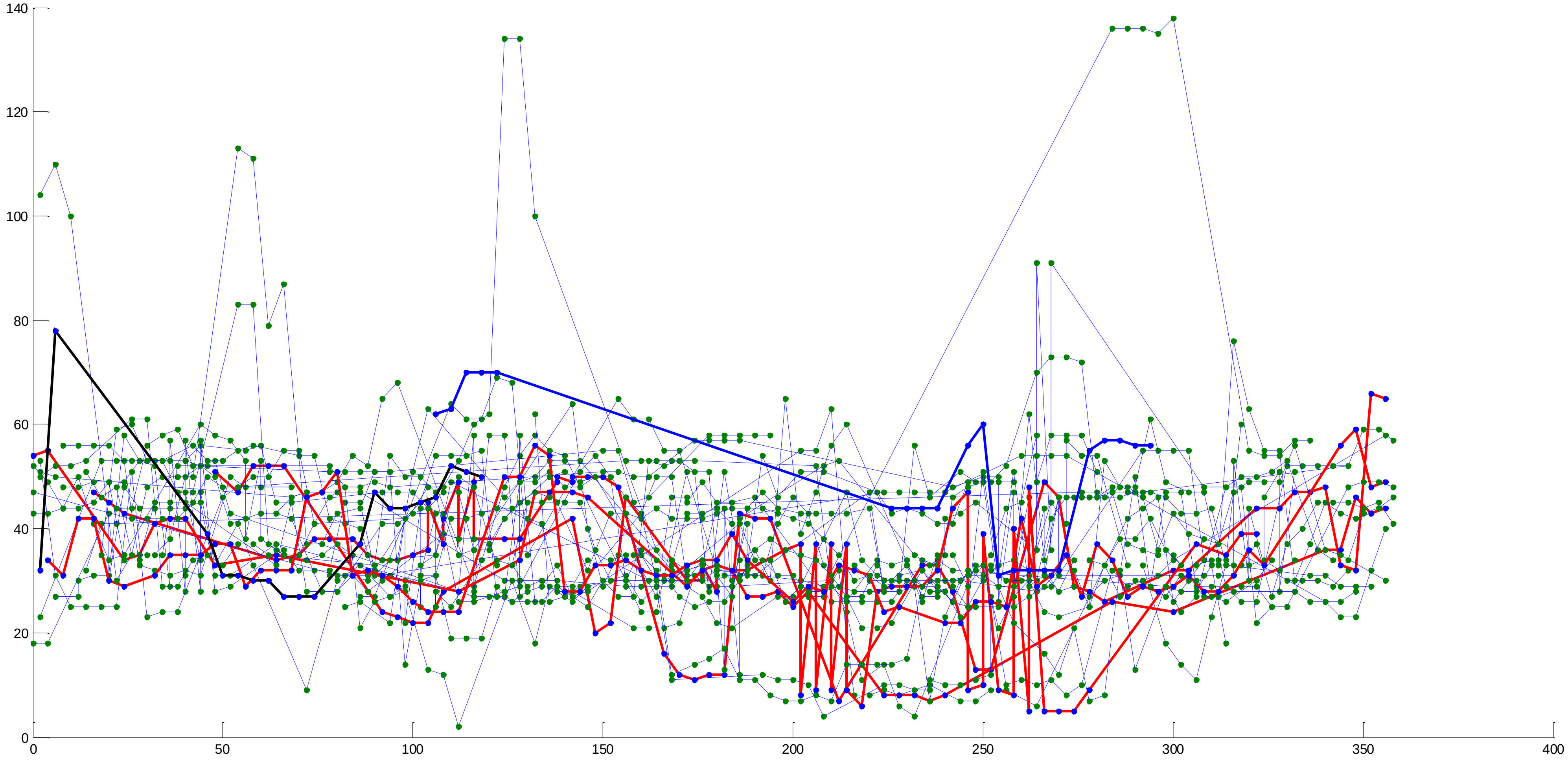
это и регуляризация

и ансамблирование $(\underbrace{\tilde{p}_j q_j}_{\max} + h\underbrace{\tilde{p}_j}_{\max})$

Качество предсказания поведения в зависимости от параметра h**.**

Вопрос: Как решать задачу о пробках?

Как выглядят данные:

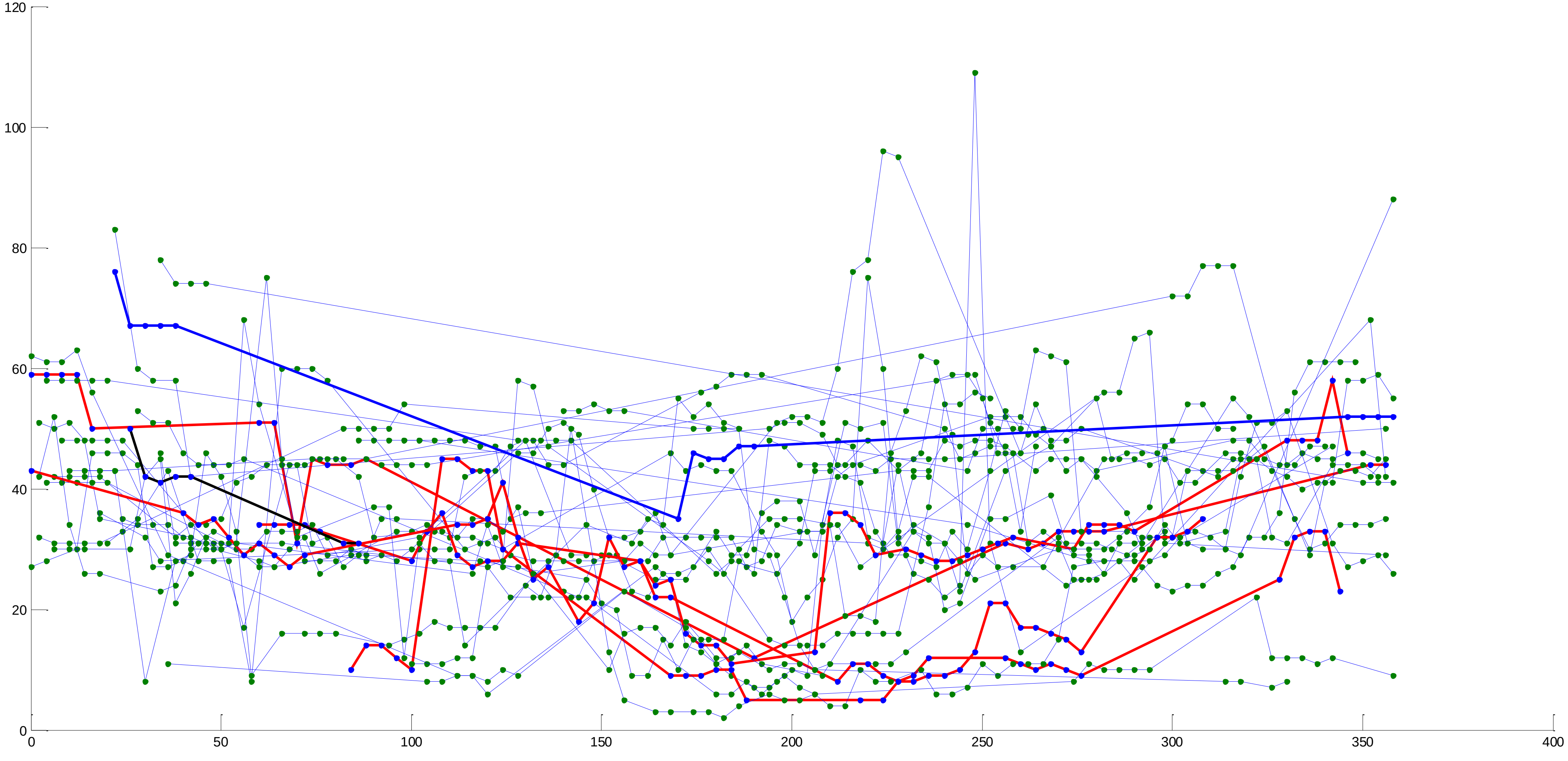


Чёрный – наш день,

Красный – этот день недели,

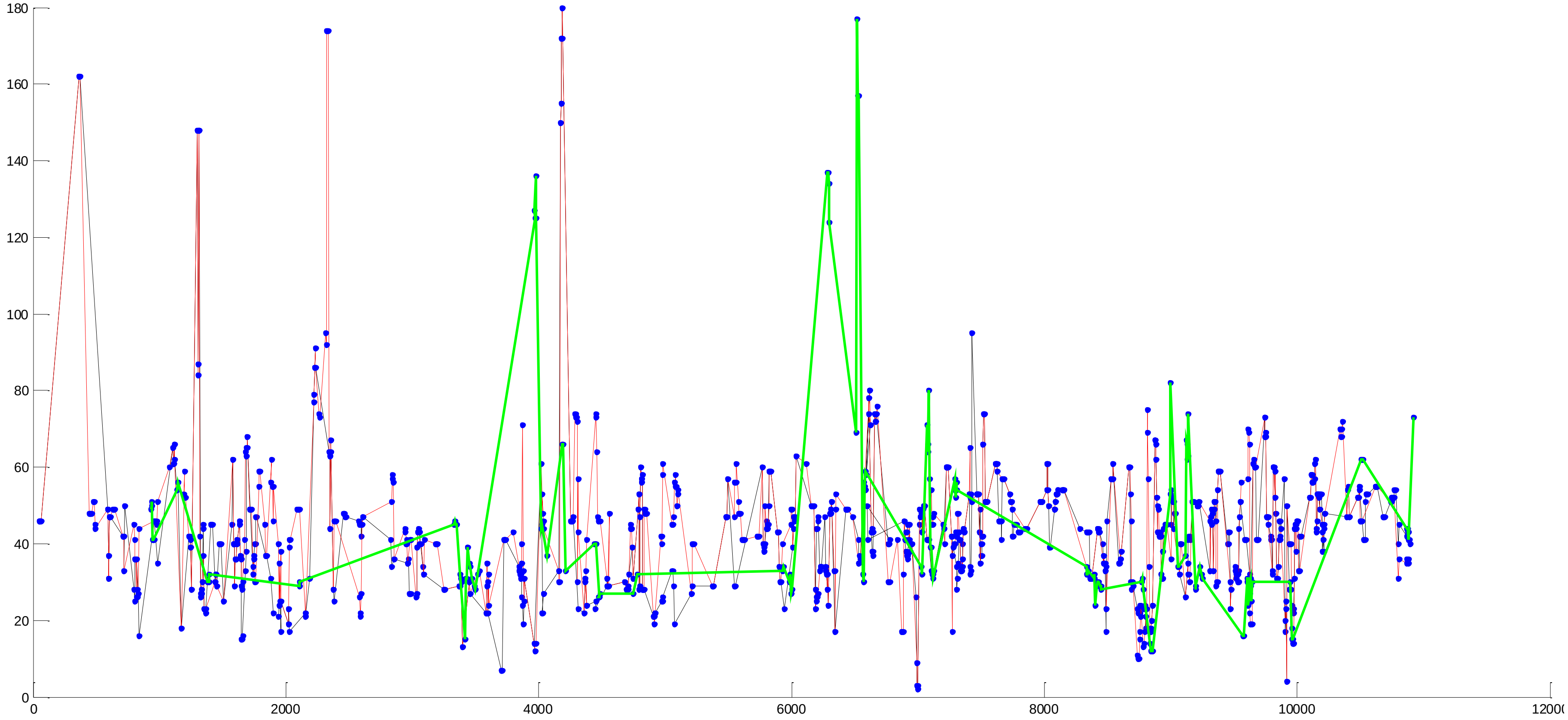
Синий – предыдущий день.

Другая дуга (граф ориентированный):



Замечаем странности:

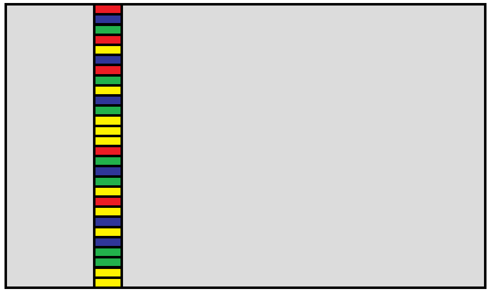
1. По некоторым дугам статистика совпадает
2. Или почти совпадает.
3. Скорость «теряется» при переходе на другую дугу.

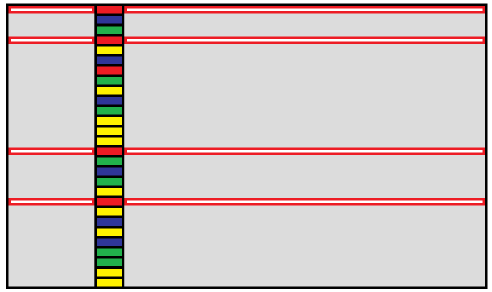


Разные дороги: чёрный, красный, зелёный.

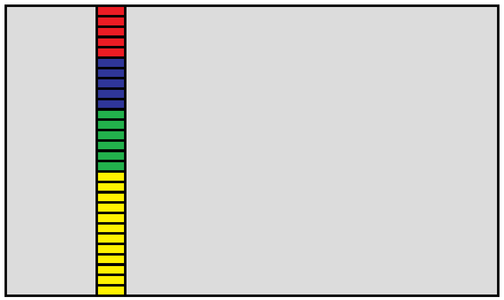
В процессе обработки данных открыл для себя приём:

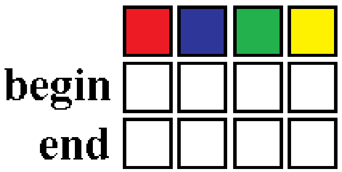
Выборка по факторам...



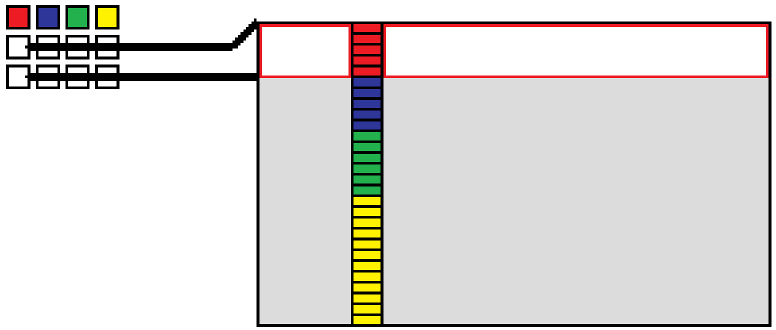


`M[M[:,i]==a,:]`



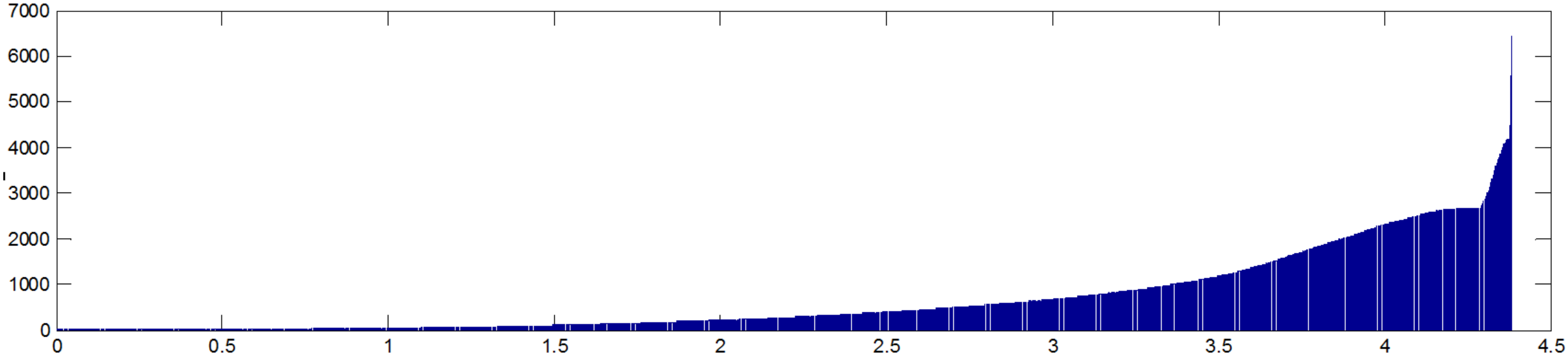


Сортировка



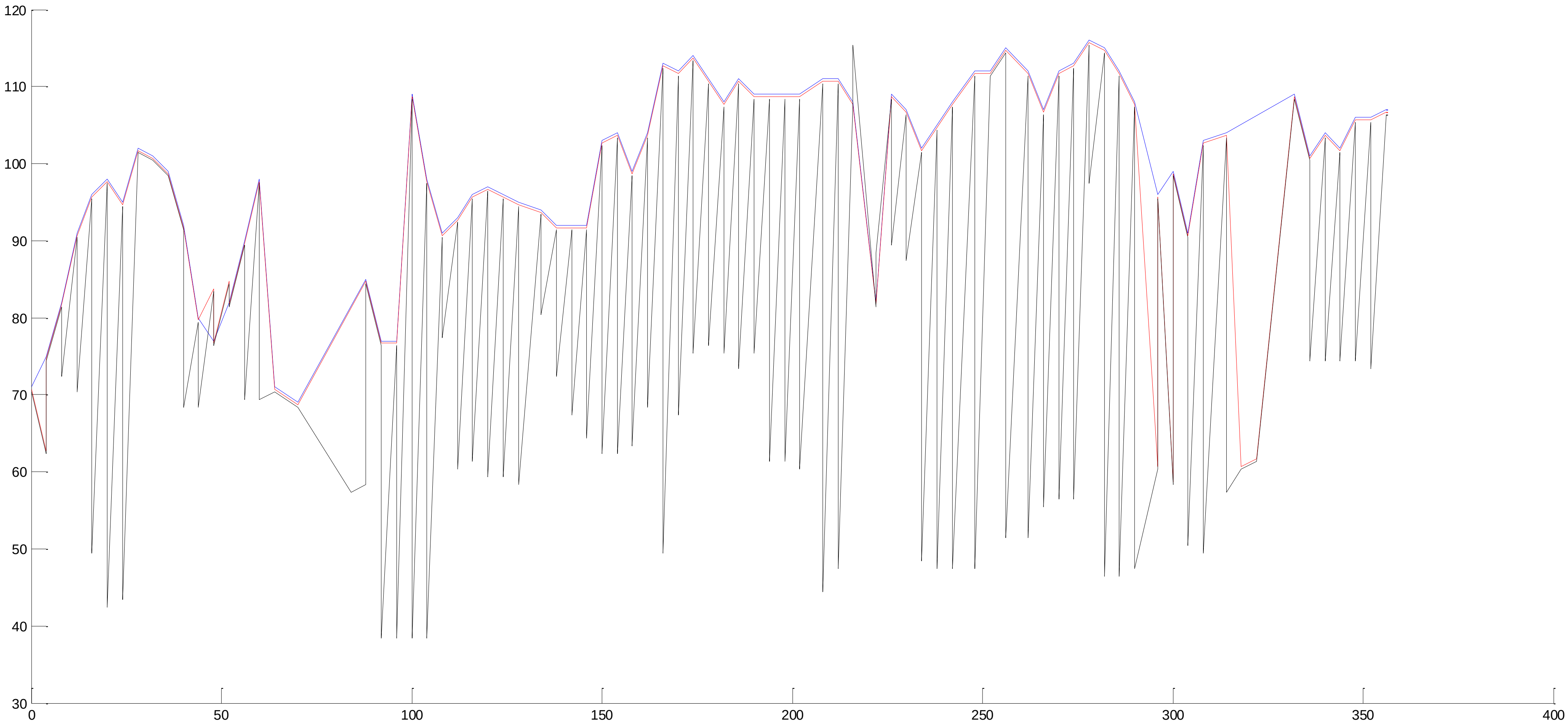
**Хранить
начала и концы разных
факторов.**

Распределение длин дорог



опять нет нормального распределения...

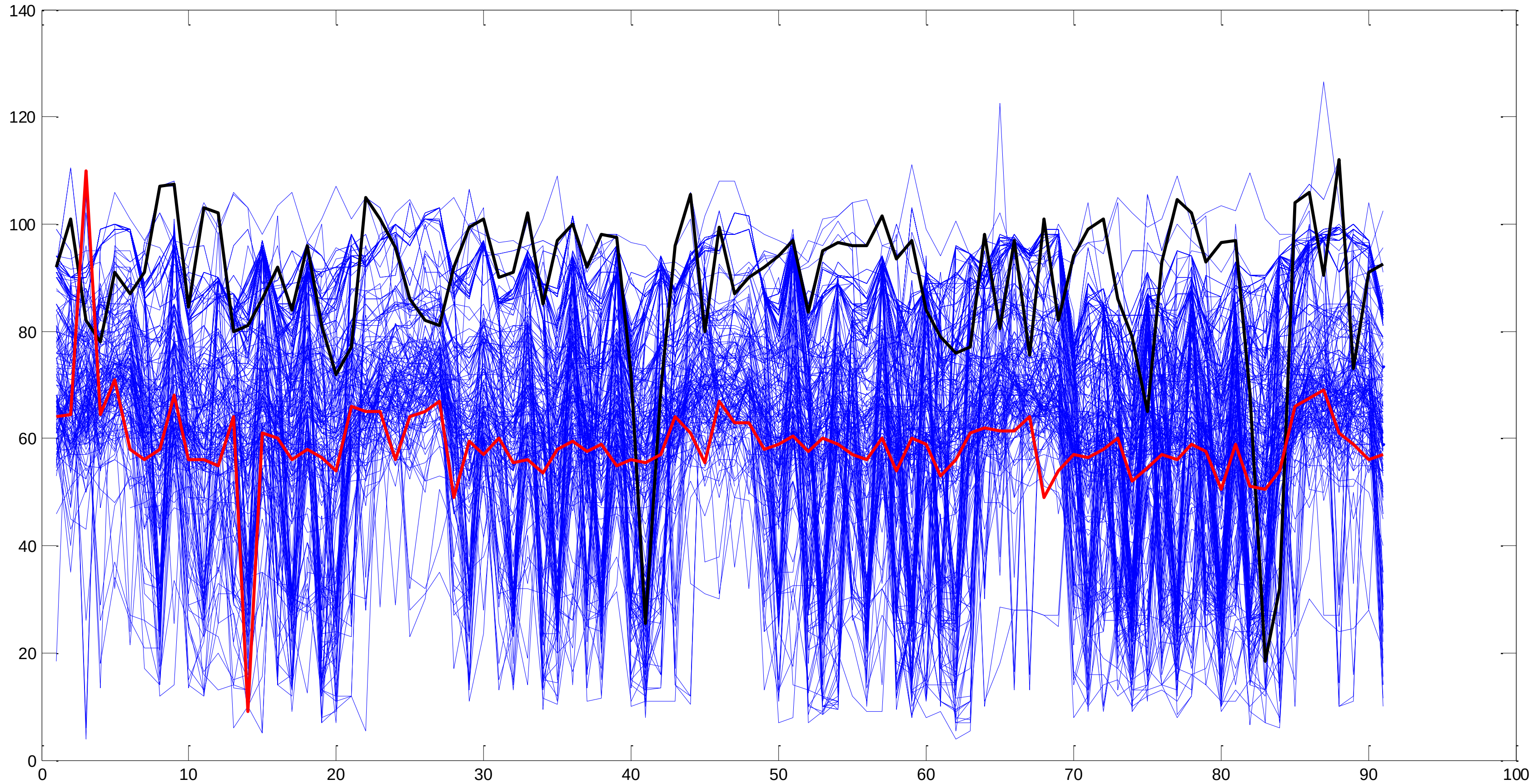
Данные с трёх дуг



Данные двух дуг совпадают,

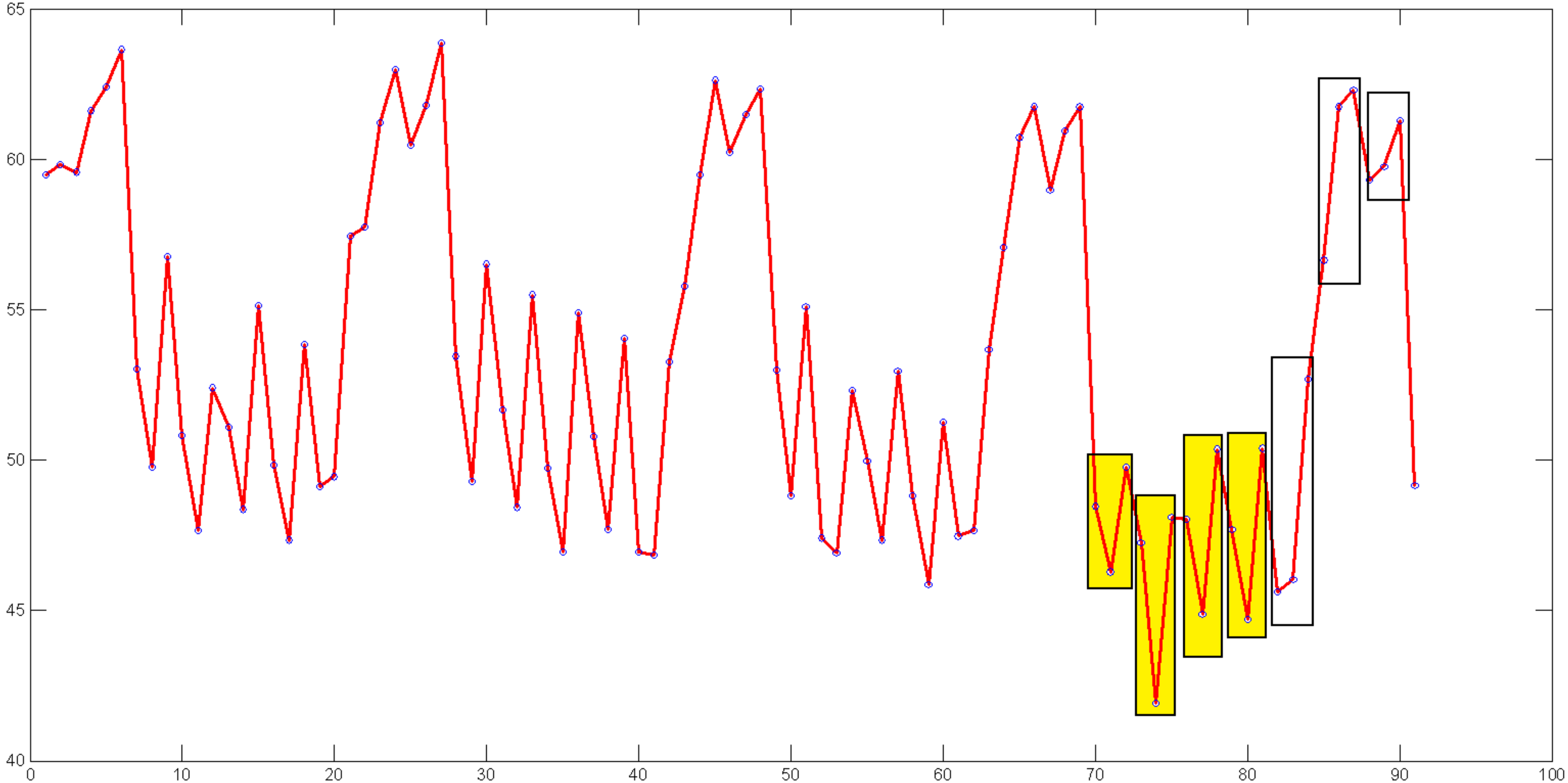
+ с половиной данных третьей дуги.

Медианные данные по всем дням



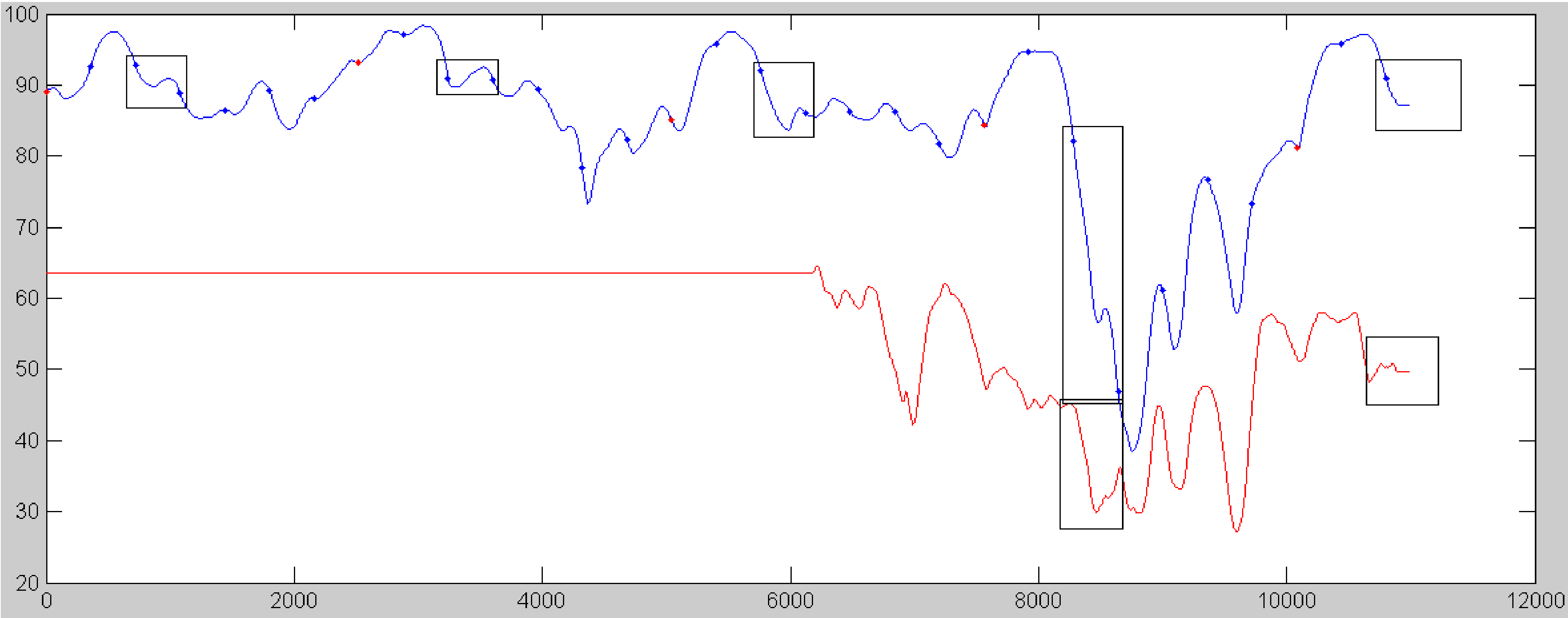
Что можно сказать?

Ответ: Идентифицировать дни недели.



и даже видна идея решения!

Иллюстрация сглаживания



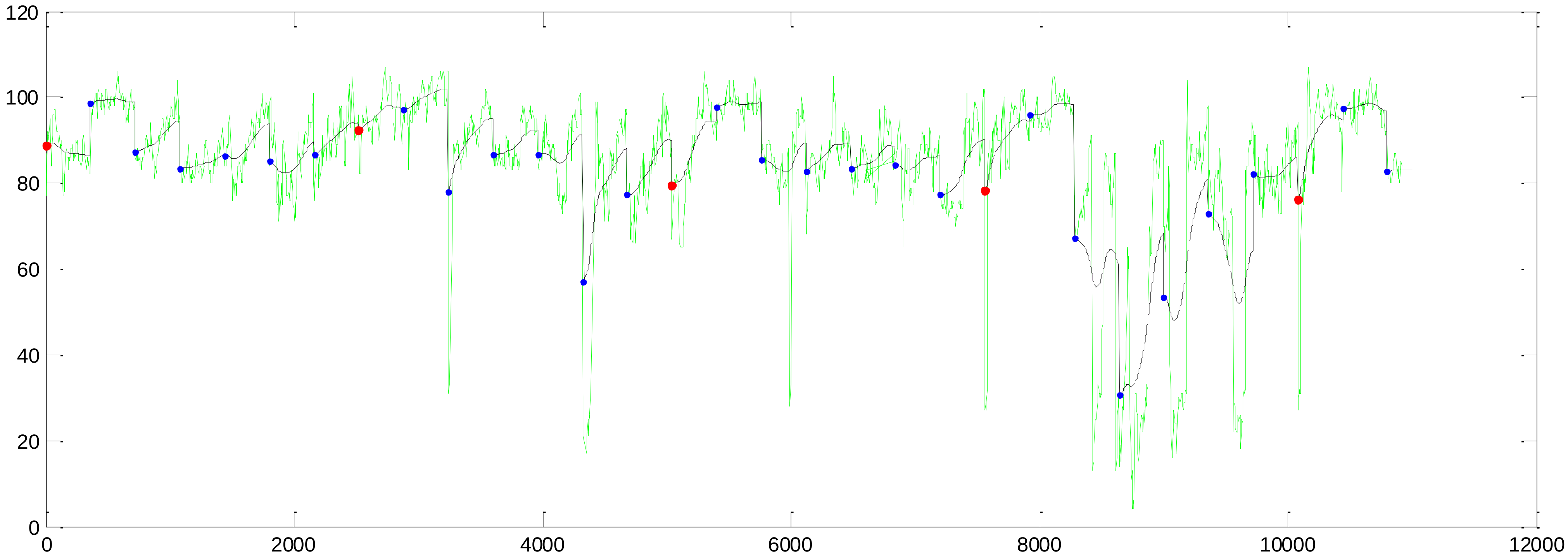
Данные по двум конкретным дорогам.

Выделены участки одного дня недели.

По красной нет достаточно статистики,

но она коррелирует с синей, по которой есть!

Пример сглаживания

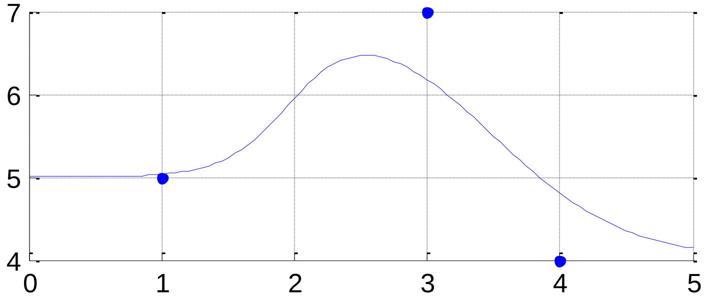


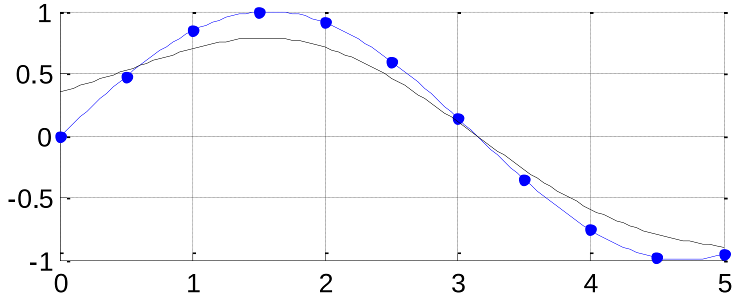
$$y(x)=\sum_{i=1}^{n} w_i y(x_i)$$

$$w_i = K(x,x_i) \approx^{\mathrm{N}} e^{-\rho(x,x_i)}$$

Формула Надарая-Ватсона – а ведь это тоже весовая схема!

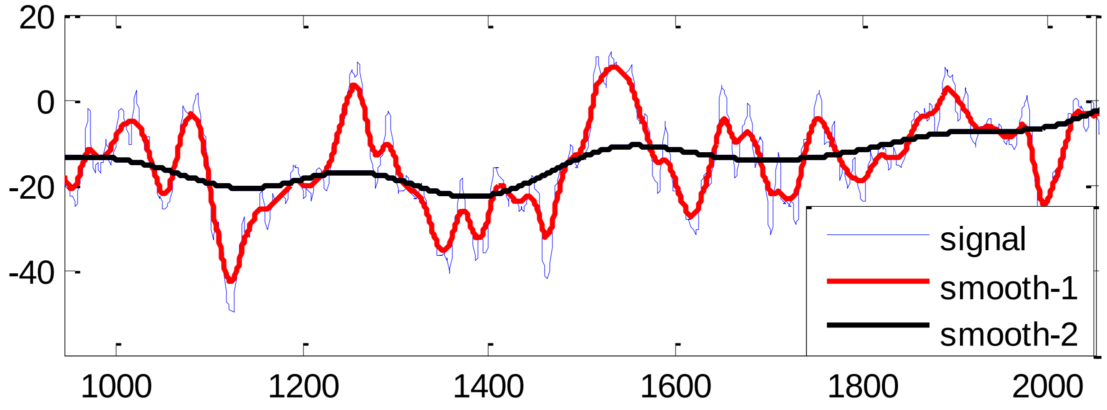
«Регрессия» по формуле Надарая-Ватсона



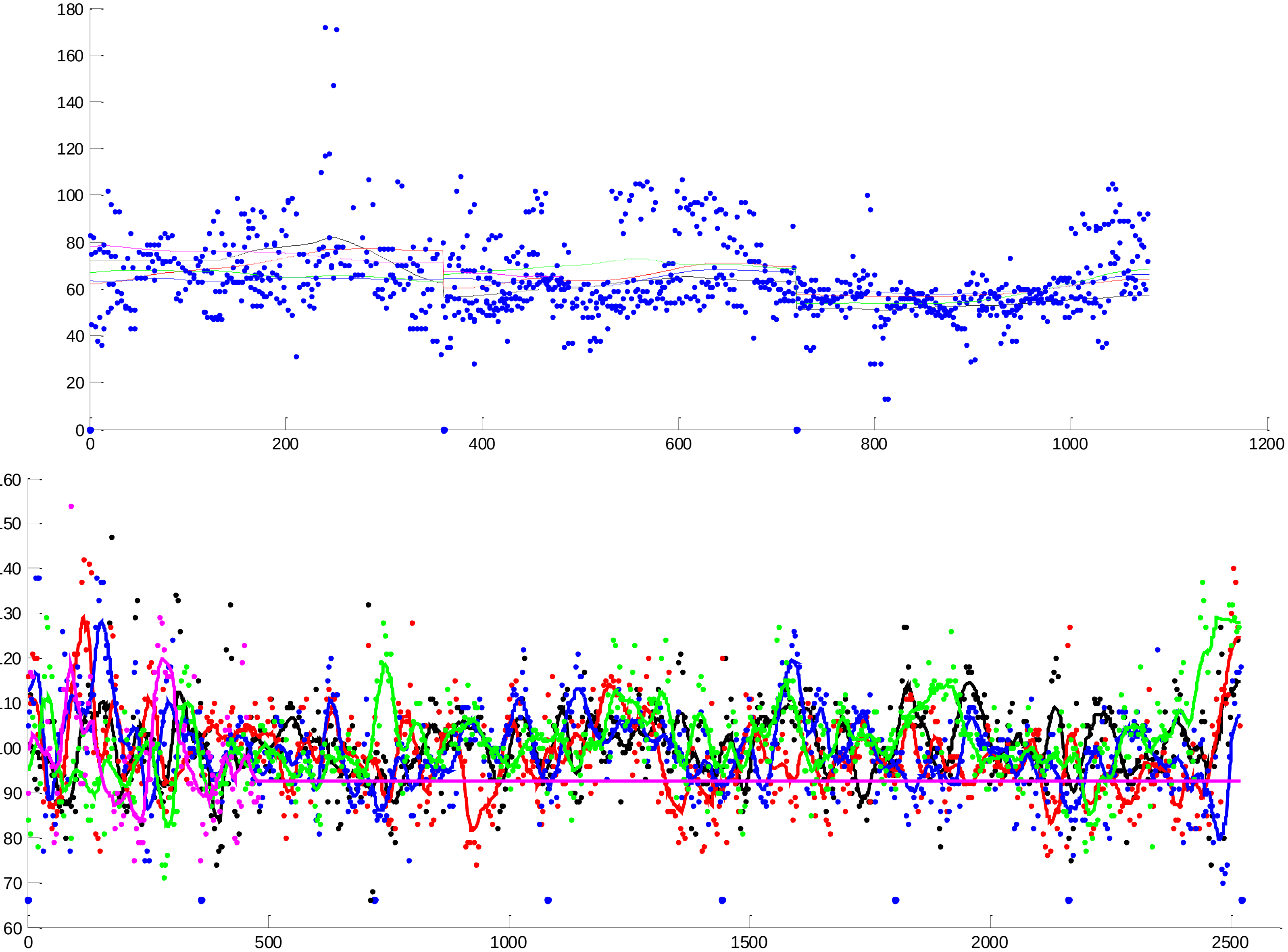


```
K = @(x) exp(-x.^2); % функция K
x = 0:0.05:5; % отрезок
f = sin(x); % истинные значения функции
X = x(1:10:end); % обучающая выборка - объекты
Y = sin(X); % - их метки
t = repmat(x',1,length(X)) - repmat(X,length(x),1);
t = arrayfun(K, t);
sumt = sum(t, 2);
t = sum(t.*repmat(Y,length(x),1), 2)./sumt; % значения
ф-лы Н-В
clf; hold on; grid on; % Графика
plot(x, f, 'b'); % как должно быть
scatter(X, Y, 20, 'filled'); % обучение
           plot(x, t, 'k'); % что получилось
```

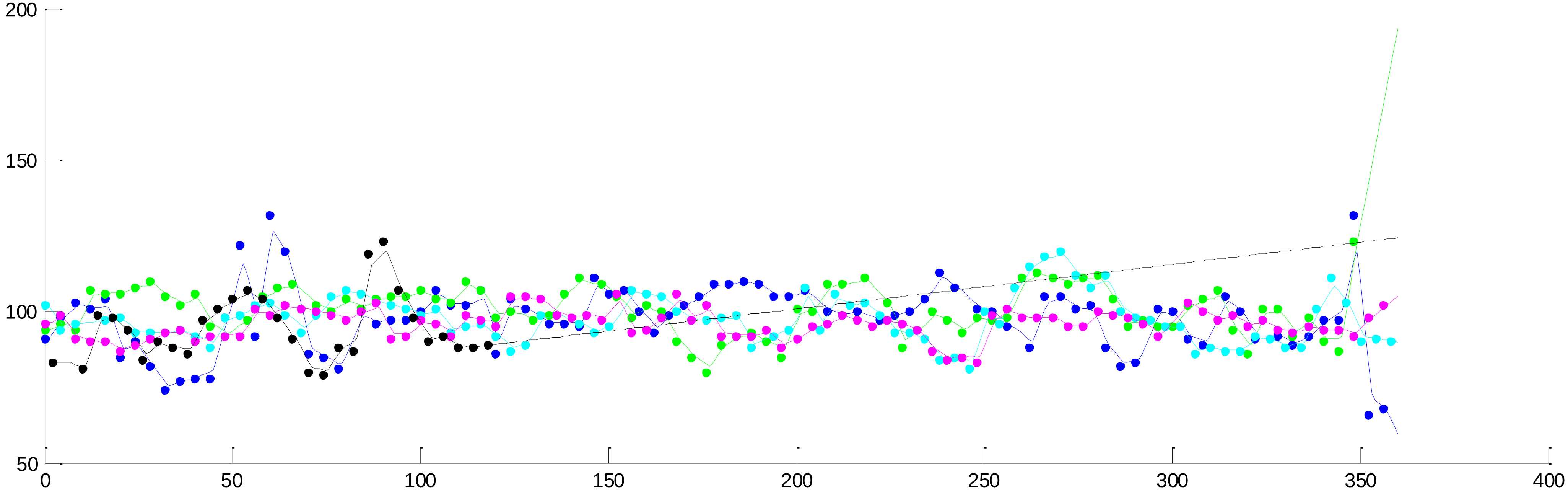
Сглаженная электрокортикограмма при различных *h*.



Зачем нужно сглаживать... скорость на одной дороге в разные дни



ЛИНЕЙНЫЙ Надарая-Ватсон достаточно опасный:



В обычном

- не проходит через точки

- почти всё считает выбросом

- не экстраполирует

- проблема подбора ширины окна (ядра)

Рецепт по усреднению:

Что усреднять:

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

Как – эксперименты!

Литература

- Шурыгин А.М. Математические методы прогнозирования // М., Горячая линия — Телеком, 2009, 180 с.

нужные фрагменты есть в http://www.machinelearning.ru/wiki/images/7/7e/Dj2010up.pdf

- Дьяконов А.Г. Прогноз поведения клиентов супермаркетов с помощью весовых схем оценок вероятностей и плотностей // Бизнес-информатика. 2014. № 1 (27). С. 68–77.

https://bijournal.hse.ru/data/2014/04/15/1320713004/8.pdf

- Оценка вероятности: когда к нам придёт клиент? // https://vimeo.com/119925869