# Выбор иерархических моделей в авторегрессионном прогнозировании

И. В. Фадеев

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель к.ф.-м.н., н.с. ВЦ РАН В. В. Стрижов

Москва,
2013 г.

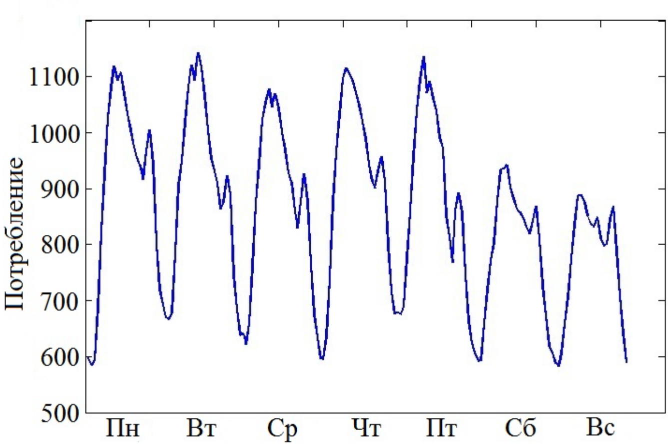
# Цели и методы исследования

Цель: разработать метод построения прогностических моделей, описывающих периодические временных ряды и включающие инвариантные преобразования.
Предмет исследования: пучки временных рядов.
Методы исследования: авторегрессионное прогнозирование, полупараметрическое и иерархическое моделирование.

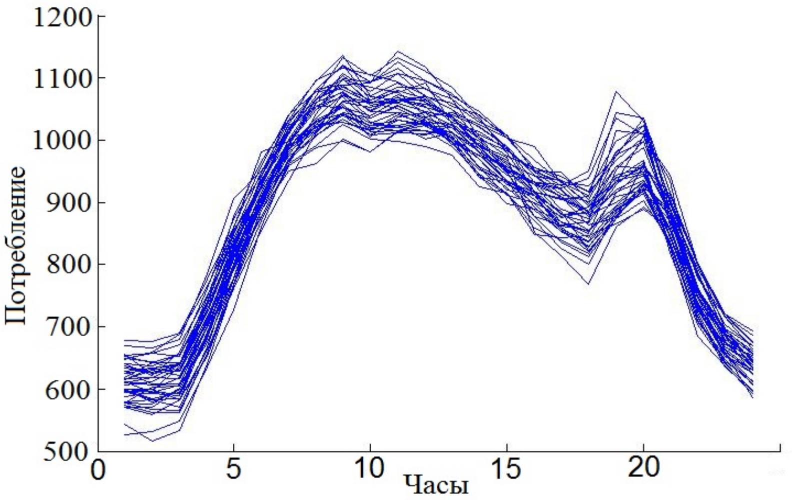
# Обзор литературы (1)

- Lawton, Sylvestre, Maggio (1972):
  Self modeling nonlinear regression —
  модель SEMOR, Shape-Invariant Model.
- Kneip, Engel (1995):
  Model estimation in nonlinear regression under shape invariance.
- Gamboa, Loubes (2007):
  Semi-parametric estimation of shifts.
- Hurtgen, Gervini (2008):
  Semiparametric shape-invariant models for periodic data.
- Vimond (2010):
  Efficient estimation for a subclass of shape invariant models.
- Bertrand, Fhima, Guillin (2010):
  Off-line detection of multiple change points with the Filtered Derivative with p-Value method.

# Потребление электроэнергии, неделя



# Потребление электроэнергии, будни



# Гипотеза порождения данных (1)

Рассматриваются $N$ временных рядов длины $n$:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in}), \qquad i = 1, \ldots, N.$$

Пусть

$$\mathbf{x}_i = \mathbf{f}(\mathbf{z}_0, \boldsymbol{\alpha}_i) + \varepsilon_i, \qquad \mathbf{z}_0 \in \mathbb{R}^n, \boldsymbol{\alpha}_i \in \mathbb{R}^m, \qquad \varepsilon_i \sim \mathcal{N}(0, E_n\sigma^2),$$

где $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ — параметрическое семейство преобразований, $\mathbf{z}_0$ — форма, соответствующая выборке $\mathbf{x}_i$.
Пусть преобразование $\mathbf{f}$ определяет в $\mathbb{R}^n$ отношение эквивалентности:

$$\mathbf{x}_i \sim \mathbf{x}_j \qquad \Longleftrightarrow \qquad \exists \boldsymbol{\alpha} : \mathbf{x}_j = \mathbf{f}(\mathbf{x}_i, \boldsymbol{\alpha}).$$

# Гипотеза порождения данных (2)

Определяется множество $\mathbb{Z} \subset \mathbb{R}^n$ так, что оно содержит ровно по одному представителю от каждого класса эквивалентности. Условие

$$\mathbf{z}_0 \in \mathbb{Z}$$

однозначно определяет форму $\mathbf{z}_0$ и параметры $\boldsymbol{\alpha}_i$.
Любой вектор $\mathbf{x}_i \in \mathbb{R}^n$ однозначно представим в виде

$$\mathbf{x}_i = \mathbf{f}(\mathbf{z}_i, \boldsymbol{\alpha}_i), \qquad \mathbf{z}_i \in \mathbb{Z},$$

что позволяет ввести преобразования $\mathbf{u} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{v} : \mathbb{R}^n \to \mathbb{R}^m$, такие, что

$$\mathbf{x}_i = \mathbf{f}(\mathbf{u}(\mathbf{x}), \mathbf{v}(\mathbf{x})), \qquad \mathbf{u}(\mathbf{x}) \in \mathbb{Z}.$$

$\mathbf{u}(\mathbf{x})$ — форма вектора $\mathbf{x}$ — инвариант относительно преобразования $\mathbf{f}$.

# Примеры семейств преобразований

- Прибавление полинома:

$$f_j(\mathbf{x}, \boldsymbol{\alpha}) = x_j + \sum_{m=0}^{k} \alpha_m j^m, \qquad j = 1, \ldots, n,$$

$$\mathbf{v}(\mathbf{x}) = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \sum_{j=1}^{n} \left( \sum_{m=0}^{k} \alpha_m j^m - x_j \right)^2,$$

$$u_j(\mathbf{x}) = x_j - \sum_{m=0}^{k} v_m(\mathbf{x}) j^m.$$

- Растяжение:

$$f_j(\mathbf{x}, \boldsymbol{\alpha}) = g(x_j, \boldsymbol{\alpha}), \qquad j = 1, \ldots, n,$$

где $g$ — семейство монотонных функций; например,

$$f_j(\mathbf{x}, \boldsymbol{\alpha}) = \alpha_0 x_j^{\alpha_1}, \qquad j = 1, \ldots, n.$$

# Оценка формы полупараметрической модели (1)

Пусть

$$\hat{\mathbf{z}}_0 = \frac{1}{N}\sum_{i=1}^{N} u(\mathbf{x}_i). \tag{1}$$

**Теорема 1.** Пусть преобразование $u$ удовлетворяет условию Липшица с константой $L$, т. е. для любых $\mathbf{x}_i$, $\mathbf{x}_j$

$$||\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)|| \leq L||\mathbf{x}_i - \mathbf{x}_j||.$$

Тогда почти наверно существует минимальный размер выборки $N_0$ такой, что

$$||\hat{\mathbf{z}}_0 - \mathbf{z}_0|| < L\sigma n + \varepsilon_0, \qquad \forall N : N > N_0,$$

где $\hat{\mathbf{z}}_0$ — оценка (1), $\varepsilon_0$ — любое положительное число.

# Оценка формы полупараметрической модели (2)

Пусть

$$\hat{\mathbf{z}}_0 = \operatorname*{argmin}_{\mathbf{u}(\mathbf{x}_j)} \sum_{i=1}^{N} ||\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)||^2. \qquad (2)$$

**Теорема 2.** Пусть преобразование **u** удовлетворяет условию Липшица с константой $L$, т. е. для любых $\mathbf{x}_i$, $\mathbf{x}_j$

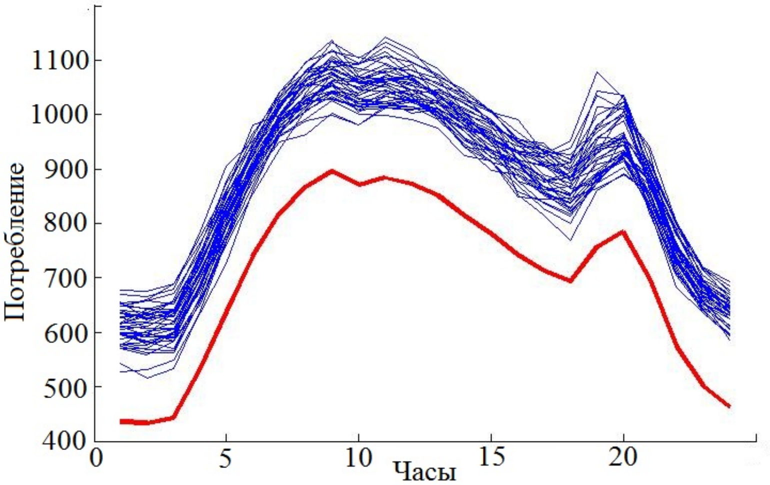
$$||\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)|| \leq L||\mathbf{x}_i - \mathbf{x}_j||.$$

Тогда почти наверно существует минимальный размер выборки $N_0$ такой, что

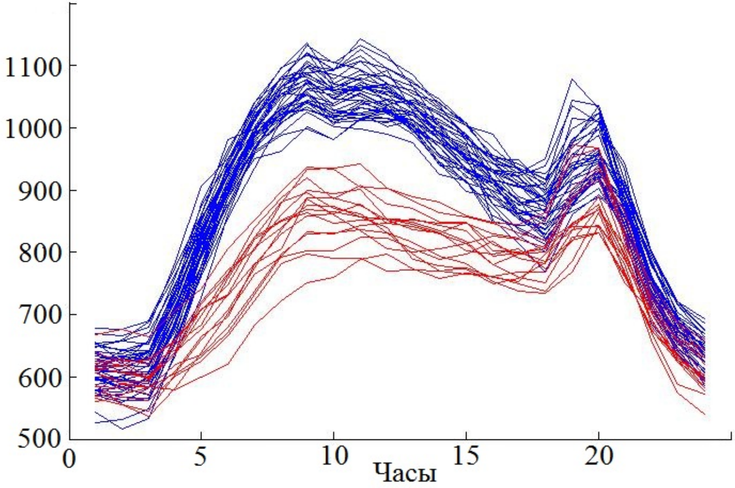
$$||\hat{\mathbf{z}}_0 - \mathbf{z}_0|| < \frac{31}{2} L\sigma\sqrt{n} + \varepsilon_0, \qquad \forall N : N > N_0,$$

где $\hat{\mathbf{z}}_0$ — оценка (2), $\varepsilon_0$ — любое положительное число.

# Форма сегментов в модели сдвига



# Потребление электроэнергии, будни и выходные



# Кластеризация выборки

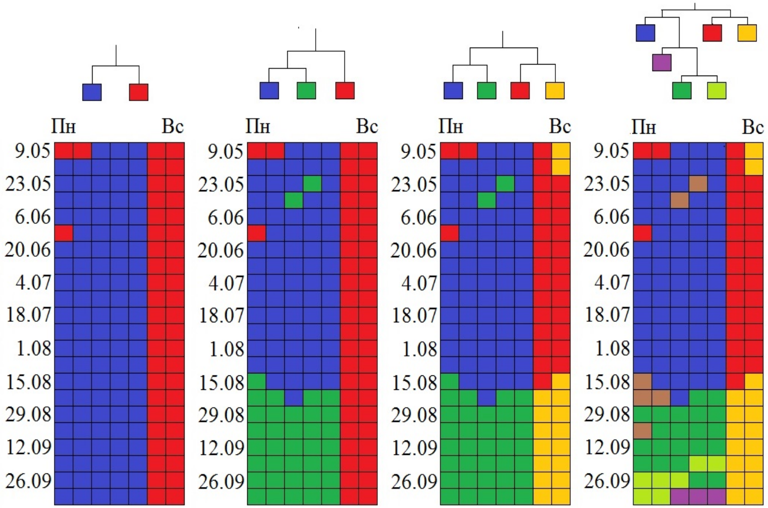
Пусть существует разбиение индексов
$\mathcal{I} = \{1, \ldots, N\} = \mathcal{I}_1 \cup \ldots \cup \mathcal{I}_s$ такое, что

$$\mathbf{x}_i = \mathbf{f}(\mathbf{z}_{0k}, \boldsymbol{\alpha}_i) + \varepsilon_i, \qquad i \in \mathcal{I}_k, \qquad \mathbf{z}_{0k} \in \mathbb{Z}, \qquad k = 1, \ldots, s.$$

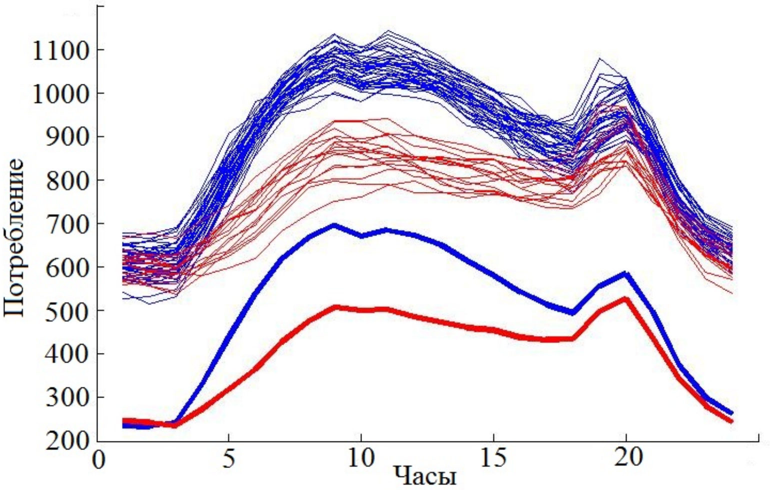
Для нахождения разбиения выполняем кластеризацию, определив функцию расстояния

$$\rho_{ij} = ||\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)||, \qquad i, j \in \mathcal{I}.$$

# Пример кластеризации



# Формы сегментов, будни и выходные



# Критерий различимости форм на подгруппах (1)

Пусть существует априорное разбиение индексов

$$\mathcal{I} = \tilde{\mathcal{I}}_1 \cup \tilde{\mathcal{I}}_2.$$

Нулевая гипотеза:

$$\frac{E[\rho_{ij}|i,j \in \tilde{\mathcal{I}}_1] + E[\rho_{ij}|i,j \in \tilde{I}_2]}{2} = E[\rho_{ij}|i \in \tilde{\mathcal{I}}_1, j \in \tilde{\mathcal{I}}_2],$$

$$D = E_{12} - \frac{E_{11} + E_{22}}{2} = 0.$$

Альтернатива:

$$D > 0.$$

Оценки матожиданий

$$\hat{E}_{12} = \frac{\sum_{i \in \tilde{\mathcal{I}}_1, j \in \tilde{\mathcal{I}}_2} \rho_{ij}}{|\mathcal{I}_1||\mathcal{I}_2|},$$

$$\hat{E}_{11} = \frac{\sum_{i,j \in \tilde{\mathcal{I}}_1, i<j} \rho_{ij}}{|\mathcal{I}_1|(|\mathcal{I}_1| - 1)/2}, \qquad \hat{E}_{22} = \frac{\sum_{i,j \in \tilde{\mathcal{I}}_2, i<j} \rho_{ij}}{|\mathcal{I}_2|(|\mathcal{I}_2| - 1)/2}.$$

# Критерий различимости форм на подгруппах (2)

Дисперсии оценок

$$\widehat{VE}_{12} = \frac{\hat{V}[\rho_{ij}|\tilde{\mathcal{I}}_1, j \in \tilde{\mathcal{I}}_2]}{|I_1||I_2|} = \frac{\sum_{i\in\tilde{\mathcal{I}}_1, j\in\tilde{\mathcal{I}}_2}(\rho_{ij} - \hat{E}_{12})^2}{|\mathcal{I}_1|^2|\mathcal{I}_2|^2},$$

$$\widehat{VE}_{11} = \frac{\sum_{i,j\in\tilde{\mathcal{I}}_1, i<j}(\rho_{ij} - \hat{E}_{11})^2}{|\mathcal{I}_1|^2(|\mathcal{I}_1| - 1)^2/4}, \widehat{VE}_{22} = \frac{\sum_{i,j\in\tilde{\mathcal{I}}_2, i<j}(\rho_{ij} - \hat{E}_{22})^2}{|\mathcal{I}_2|^2(|\mathcal{I}_2| - 1)^2/4},$$

$$\widehat{se}_D = \sqrt{\widehat{VE}_{12} + \frac{1}{4}(\widehat{VE}_{11} + \widehat{VE}_{22})}.$$

Используя центральную предельную теорему

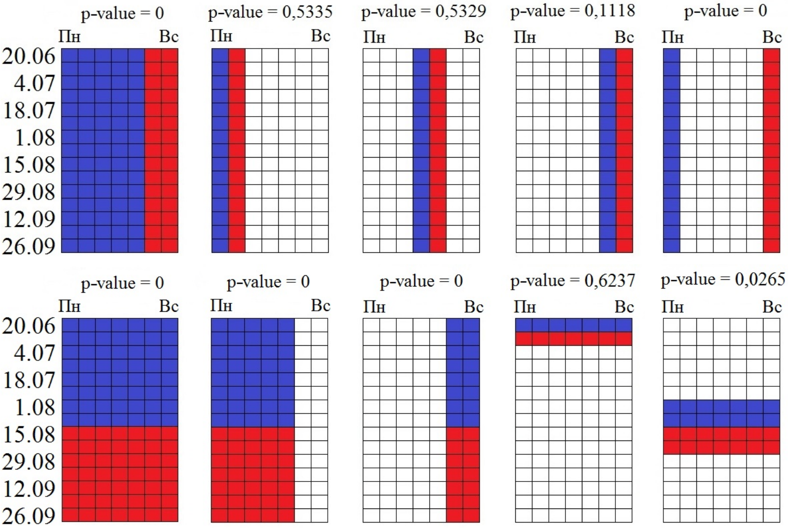
$$\hat{D} = \hat{E}_{12} - (\hat{E}_{11} + \hat{E}_{22})/2 \sim \mathcal{N}(0, \widehat{se}_D^2),$$

находим достижимый уровень значимости

$$\text{p-value} = 1 - \Phi\left(\frac{\hat{D}}{\widehat{se}_D}\right)$$

для нулевой гипотезы $D = 0$ против альтернативы $D > 0$.

# Критерий различимости форм на подгруппах (3)



# Критерий качества семейства преобразований

Пусть существует априорное разбиение индексов

$$\mathcal{I} = \{1, \ldots, N\} = \mathcal{I}_1 \cup \ldots \cup \mathcal{I}_s.$$

Необходимо выбрать преобразование f и множество Z так, чтобы минимизировать среднее внутриклассовое расстояние

$$F_1 = \frac{\sum_{i<j} [J(i) = J(j)] \rho_{ij}}{\sum_{i<j} [J(i) = J(j)]}$$

и максимизировать среднее межклассовое

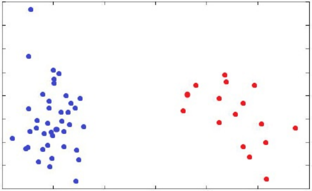
$$F_2 = \frac{\sum_{i<j} [J(i) = J(j)] \rho_{ij}}{\sum_{i<j} [J(i) = J(j)]},$$

где $J(i) = k \iff i \in \mathcal{I}_k$.
Предлагается критерий качества

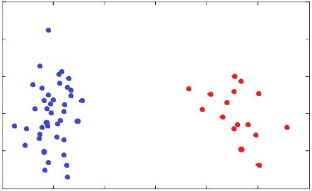
$$S(\mathbf{f}) = \frac{F_2}{F_1}.$$

# Двумерное шкалирование сегментов

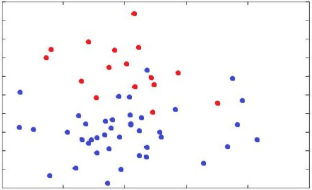
$f_i(\mathbf{x}, \boldsymbol{\alpha}) = x_j + \alpha_0,\ S(f) = 2,39$



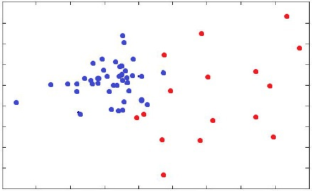
$f_j(\mathbf{x}, \boldsymbol{\alpha}) = x_j + \alpha_0 + \alpha_1 j,\ S(f) = 2,74$



$f_j(\mathbf{x}, \boldsymbol{\alpha}) = x_j + \sum_{m=0}^{5} \alpha_m j^m,\ S(f) = 1,2$



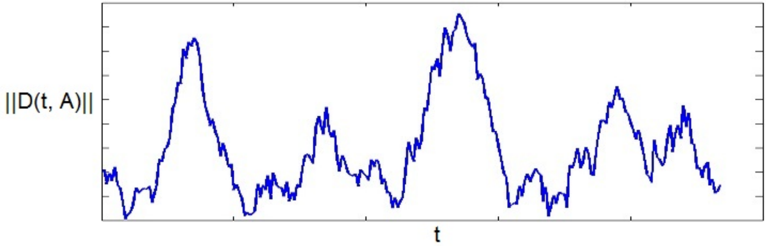
$f(\mathbf{x}, \boldsymbol{\alpha}) = \alpha_0 \mathbf{x},\ S(f) = 2,1$



Алгоритм поиска разладок с помощью дискретной производной и вычисления достижимых уровней значимости:
1) Вычисление дискретной производной

$$D(t, A) = \hat{\mathbf{z}}_0(t, A) - \hat{\mathbf{z}}_0(t - A, A),$$

где $\hat{\mathbf{z}}_0(t, A)$ — оценка формы по выборке $\{\mathbf{x}_i : t < i \leq t + A\}$.



2) Выбор потенциальных точек разладки $\tau_k$ как локальных максимум функции $||D(t,A)||$, лежащих выше порога $\lambda$:

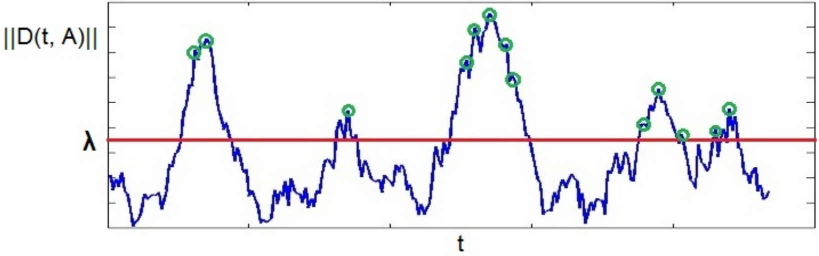
$$||D(\tau_k, A)|| > \lambda$$

.

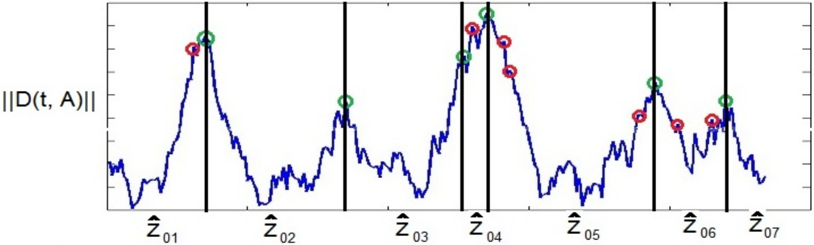
$\lambda$ выбирается так, чтобы при условии отсутствия разладки выполнялось

$$P(\max_t ||D(t,A)||) > \lambda) = p_1.$$

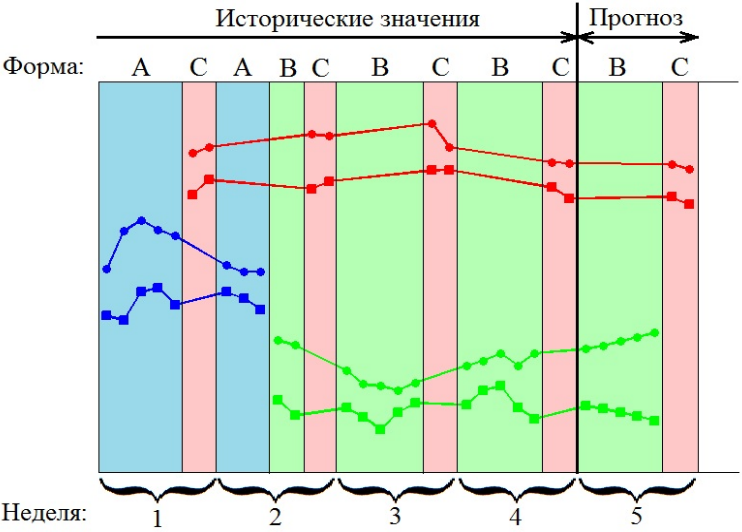
Значение $\lambda$ оценивается бутстрепом.



3) Исключение из точек $\tau_k$ ложных тревог с помощью критерия различимости форм на подвыборках.



# Прогнозирование



# Заключение

- Предложен метод оценки формы и параметров в полупараметрической регрессионной модели. Доказана устойчивость оценки формы.
- Рассмотрена задача кластеризации сегментов временных рядов схожей формы на примере потребления электроэнергии.
- Адаптирован алгоритм обнаружения разладки с помощью дискретной производной для работы с последовательностью временных рядов.
- Предложена двухуровневая иерархическая модель прогнозирования потребления электроэнергии.