Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Фадеев Илья Владимирович

Выбор иерархических моделей в авторегрессионном прогнозировании

511656 - Математические и информационные технологии

Магистерская диссертация

Научный руководитель:
н.с. ВЦ РАН, к.ф.-м.н.
Стрижов Вадим Викторович

Москва

2013

Аннотация

В работе предложены методы анализа и прогнозирования периодических временных рядов с использованием полупараметрических моделей и инвариантных преобразований. Сегменты временного ряда кластеризуются на группы со схожей формой, для каждой группы оценивается форма и параметры полупараметрической модели. Для поиска моментов изменения формы сегментов с течением времени адаптируется алгоритм обнаружения разладок с помощью дискретной производной. Для прогнозирования используется двухуровневая иерархическая модель. В качестве примера использования предложенных методов рассматривается прогнозирование почасового потребления электроэнергии.

Содержание

Введение

В работе рассматриваются методы построения прогностических моделей, описывающих периодические временных ряды и включающие инвариантные преобразования. Временной ряд разбивается на сегменты, длина которых равна периоду. Сегменты рассматриваются как выборка в полупараметрической регрессионной модели.

Полупараметрические модели были предложены в работе [1] для анализа выборок, состоящих из регрессионных кривых. Примером такой выборки является набор зависимостей целевой переменной от времени: каждому объекту соответствует своя зависимость, заданная в некоторые моменты времени. В полупараметрической модели в пространстве кривых определяется параметрическое семейство преобразований. Предполагается, что существует кривая, называемая формой и общая для всех элементов выборки, такая, что каждый элемент выборки является результатом преобразования этой кривой с некоторыми параметрами. Таким образом, построение модели включает задачу непараметрической регрессии (вычисление формы) и параметрической регрессии (вычисление параметров преобразования).

Частным случаем полупараметрических моделей являются модели, инвариантные относительно формы. В таких моделях параметрическое семейство преобразований содержит четыре параметра: растяжение и сдвиг кривой вдоль оси времени и оси целевой переменной. В работах [4, 15, 3] предложен итеративный алгоритм вычисления формы и параметров такой модели: на каждой итерации при фиксированной форме параметры вычисляются методом наименьших квадратов, далее оценивается форма с помощью усреднения по формам всех кривых выборки. Модели, инвариантные относительно формы, исследуются также в работах [14, 8, 11]. Подклассом таких моделей являются модели сдвига вдоль оси времени [7, 5, 16].

Для оценки формы и параметров полупараметрических моделей используется минимизация функционала ошибки [10, 15, 8] или максимизация правдоподобия для заданной статистической модели [9, 14]. Кривые, составляющие выборку, являются аппроксимациями зависимости целевой переменной от времени с помощью ядерного сглаживания [15] или сплайнов [6, 8, 11], или рассматриваются как марковский случайный процесс [9]. Форма в полупараметрической модели оценивается с помощью метода Надарая-Ватсона [5], усреднением по кривым выборки [3] или рассматривается как матожидание марковского случайного процесса [9]. Для периодических

кривых форма оценивается рядом Фурье [7, 14].

В данной работе предлагаются методы оценки формы и параметров полупараметрических моделей для широкого класса семейств преобразований. Предложенные оценки не состоятельны, однако доказывается, что для достаточно больших выборок ошибка оценок не превосходит некоторой величины. Для более узкого класса преобразований схожие оценки использовались в [3].

Предполагается, что сегменты временного ряда могут быть разбиты на группы так, что каждой группе соответствует своя форма в полупараметрической модели. Для нахождения разбиения предлагается использовать кластеризацию или априорные предположения о последовательности сегментов временного ряда. При этом для принятия решения о разбиении группы сегментов на две подгруппы предлагается статистический критерий различимости форм для двух подгрупп.

В работе также решается задача поиска моментов времени, в которых происходит изменение формы сегментов временного ряда. Для этого адаптируется алгоритм обнаружения разладок с помощью дискретной производной и вычислением достижимого уровня значимости, предложенный в [13]. Алгоритм заключается в вычислении дискретной производной формы сегментов по времени, отборе потенциальных точек разладки среди локальных максимумом модуля дискретной производной, исключении из потенциальных точек разладки "ложных тревог"с помощью статистического критерия.

Работа построена следующим образом. В разделе 1 определяется полупараметрическая модель для сегментов временного ряда. Как и в работах [14, 3, 11], рассматривается задача однозначного определения формы и параметров модели, но для более широкого класса преобразований. В главе 2 предлагаются методы оценки формы полупараметрической модели, доказывается их устойчивость. Раздел 3 посвящена задаче разбиения сегментов временного ряда на группы с общей формой. В разделе 4 адаптируется алгоритм обнаружения разладок для поиска моментов изменения формы сегментов временного ряда. В главе 3.2 приведён критерий качества параметрического семейства преобразований, который используется при наличии разбиения сегментов на группы с общей формой, заданной экспертом. Этот критерий позволяет выбрать семейство преобразований, наилучшим образом отражающее экспертное представление о близости форм временных рядов. В разделе 5.2 предлагается иерар-

хическая двухуровневая модель для прогнозирования значений временного ряда. В разделе 5 приводится пример использования предложенных методов для анализа и прогнозирования почасового потребления электроэнергии.

1 Модель порождения данных

Рассматривается временной ряд с периодом n

$$y_t, \qquad t = 1, \ldots, T,$$

где $T = Nn$ — длина временного ряда, кратная периоду n, $N = \frac{T}{n}$ — число периодов.

Временной ряд y_t разбивается на сегменты, длина которых равна периоду n:

$$\mathbf{x}_i = (y_{(i-1)n+1}, y_{(i-1)n+2}, \ldots, y_{in}), \qquad i = 1, \ldots, N.$$

Пусть $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ — параметрическое семейство преобразований. Рассмотрим следующую модель порождения данных:

$$\mathbf{x}_i = \mathbf{f}(\mathbf{z}_0, \boldsymbol{\alpha}_i) + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\alpha}_i \in \mathbb{R}^m, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, E_n\sigma^2), \tag{1.1}$$

где $\mathbf{z}_0$ — некоторый временной ряд, $\boldsymbol{\alpha}_i$ — параметры преобразования, соответствующие каждому временному ряду, E_n — единичная матрица, σ^2 — дисперсия ошибки. Таким образом, временные ряды $\mathbf{x}_i$ являются результатом преобразования временного ряда $\mathbf{z}_0$ с аддитивным нормальным шумом.

Предположим также, что преобразования $\mathbf{f}$ разбивают пространство $\mathbb{R}^n$ на классы эквивалентности следующим образом: $\mathbf{x}_i \sim \mathbf{x}_j$ если и только если существует вектор $\boldsymbol{\alpha}$ такой, что $\mathbf{x}_j = \mathbf{f}(\mathbf{x}_i, \boldsymbol{\alpha})$.

Форма $\mathbf{z}_0$ и параметры $\boldsymbol{\alpha}_i$ модели (1.1) не заданы однозначно: любая форма из класса эквивалентности исходной $\tilde{\mathbf{z}}_0 \sim \mathbf{z}_0$ с соответствующим набором параметров $\tilde{\boldsymbol{\alpha}}_i$ определяет ту же модель. Действительно, из $\tilde{\mathbf{z}}_0 \sim \mathbf{z}_0 \sim \mathbf{f}(\mathbf{z}_0, \boldsymbol{\alpha}_i)$ следует, что существует вектор параметров $\tilde{\boldsymbol{\alpha}}_i$ такой, что $\mathbf{f}(\tilde{\mathbf{z}}_0, \tilde{\boldsymbol{\alpha}}_i) = \mathbf{f}(\mathbf{z}_0, \boldsymbol{\alpha}_i)$, и модель

$$\mathbf{x}_i = \mathbf{f}(\tilde{\mathbf{z}}_0, \tilde{\boldsymbol{\alpha}}_i) + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, E_n\sigma^2)$$

эквивалентна исходной модели (1.1) порождения данных.

Определим множество $\mathbb{Z} \subset \mathbb{R}^n$ так, чтобы оно содержало ровно по одному представителю от каждого класса эквивалентности преобразования $\mathbf{f}$. Тогда 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 преобразование $\mathbf{u} : \mathbb{R}^n \to \mathbb{R}^n$ и отображение $\mathbf{v} : \mathbb{R}^n \to \mathbb{R}^m$, такие, что

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

Введение ограничения $\mathbf{z}_0 \in \mathbb{Z}$ однозначно определяет $\mathbf{z}_0$ и $\boldsymbol{\alpha}_i$ модели (1.1):

$$\mathbf{x}_i = \mathbf{f}(\mathbf{z}_0, \boldsymbol{\alpha}_i) + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, E_n\sigma^2), \qquad \mathbf{z}_0 \in \mathbb{Z}. \tag{1.2}$$

Вектор $\mathbf{z}_0$ будем называть формой, соответствующей выборке $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, а вектор $\mathbf{z}_i = \mathbf{u}(\mathbf{x}_i)$ формой вектора $\mathbf{x}_i$. Заметим, что преобразование $\mathbf{u}$ инвариантно относительно $\mathbf{f}$, а форма $\mathbf{u}(\mathbf{x}_i)$ является инвариантом вектора $\mathbf{x}_i$.

1.1 Примеры семейств преобразований

Наиболее простые и хорошо интерпретируемые семейства инвариантных преобразований — растяжение-сдвиг вдоль оси значений временного ряда $\mathbf{f}(\mathbf{x}, \boldsymbol{\alpha}) = \alpha_1\mathbf{x} + \alpha_2$, растяжение-сдвиг вдоль оси времени, а также их суперпозиции.

Обобщением сдвига вдоль оси значений является прибавление полинома k-го порядка:

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

Множество $\mathbb{Z}$ можно задать преобразованиями

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

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

при этом $\mathbf{v}(\mathbf{x})$, $\mathbf{u}(\mathbf{x})$ вычисляются методом наименьших квадратов.

Обобщением растяжения являются преобразования вида

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

где g — семейство монотонных функций; например,

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

2 Оценка формы в полупараметрической модели

Предположим, что для любого набора векторов $\mathbf{z}_l \in \mathbb{Z}$ выполнено условие

$$\frac{1}{L}\sum_{l=1}^{L} \mathbf{z}_l \in \mathbb{Z}, \qquad \mathbf{z}_l \in \mathbb{Z}, \qquad l = 1, \ldots, L. \tag{2.1}$$

Тогда в качестве оценки формы $\mathbf{z}_0$ из модели (1.2) предлагается использовать среднее значение форм векторов $\mathbf{x}_i$

$$\mathbf{z}^* = \frac{1}{N}\sum_{i=1}^{N} \mathbf{u}(\mathbf{x}_i). \tag{2.2}$$

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

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

Тогда почти наверно существует минимальный размер выборки N_0 такой, что

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

где $\mathbf{z}^*$ — оценка (2.2), ε_0 — любое положительное число.

Доказательство. Заметим, что

$$|u_j(\mathbf{x}_i) - z_{0j}| \le \|\mathbf{u}(\mathbf{x}_i) - \mathbf{z}_0\| \le L\|\mathbf{x}_i - f(\mathbf{z}_0, \boldsymbol{\alpha}_i)\| = L\|\boldsymbol{\varepsilon}_i\|.$$

$$|\mathrm{E}[u_j(\mathbf{x}_i) - z_{0j}]| \le \mathrm{E}|u_j(\mathbf{x}_i) - z_{0j}| \le \mathrm{E}[L\|\boldsymbol{\varepsilon}_i\|] = L\mathrm{E}\|\boldsymbol{\varepsilon}_i\|. \tag{2.3}$$

$$\mathrm{V}[u_j(\mathbf{x}_i) - z_{0j}] = \mathrm{E}(u_j(\mathbf{x}_i) - z_{0j})^2 - (E[u_j(\mathbf{x}_i) - z_{0j}])^2 \le$$
$$\le \mathrm{E}(u_j(\mathbf{x}_i) - z_{0j})^2 \le \mathrm{E}[L^2\|\boldsymbol{\varepsilon}_i\|^2] = L^2\mathrm{E}\|\boldsymbol{\varepsilon}_i\|^2 = L^2\sigma^2 n. \tag{2.4}$$

Из модели порождения данных (1.2) следует, что

$$\frac{\boldsymbol{\varepsilon}_i}{\sigma} \sim \mathcal{N}(0, E_n\sigma^2),$$

— вектор с независимыми компонентами, имеющими стандартное нормальное распределение. Следовательно,

$$\left\| \frac{\boldsymbol{\varepsilon}_i}{\sigma} \right\|^2 \sim \chi^2(n).$$

Как известно, $E\chi^2(n) = n$, следовательно,

$$\mathrm{E}\|\boldsymbol{\varepsilon}_i\|^2 = \sigma^2 n,$$

$$\mathrm{E}\|\boldsymbol{\varepsilon}_i\| = \sqrt{\mathrm{E}\|\boldsymbol{\varepsilon}_i\|^2 - \mathrm{V}\|\boldsymbol{\varepsilon}_i\|} \leq \sqrt{\mathrm{E}\|\boldsymbol{\varepsilon}_i\|^2} = \sigma\sqrt{n}.$$

Из (2.3) и последнего неравенства следует, что

$$|\mathrm{E}[u_j(\mathbf{x}_i) - z_{0j}]| \leq L\sigma\sqrt{n}.$$

Поскольку матожидание среднего значения нескольких случайных величин не превосходит максимального значения матожидания, то

$$\left| \mathrm{E}\left[\frac{1}{N} \sum_{i=1}^{N} u_j(\mathbf{x}_i) - z_{0j} \right] \right| \leq L\sigma\sqrt{n}.$$

Воспользуемся усиленным законом больших чисел в формулировке Колмогорова, согласно которому

$$\frac{1}{N} \sum_{i=1}^{N} u_j(\mathbf{x}_i) - z_{0j} \xrightarrow{\text{almost sure}} \mathrm{E}\left[\frac{1}{N} \sum_{i=1}^{N} u_j(\mathbf{x}_i) - z_{0j} \right]$$

при выполнении условия

$$\sum_{i=1}^{\infty} \frac{1}{i^2} \mathrm{V}[u_j(\mathbf{x}_i) - z_{0j}] < \infty.$$

Последнее неравенство выполняется в силу (2.4) и того, что

$$\sum_{i=1}^{\infty} \frac{1}{i^2} \mathrm{V}[u_j(\mathbf{x}_i) - z_{0j}] \leq L^2\sigma^2 n \sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{L^2\sigma^2 n\pi^2}{6} < \infty.$$

Из сходимости почти наверное следует, что с вероятностью единица существует число N_0 такое, что для любого $N > N_0$

$$\frac{1}{N} \sum_{i=1}^{N} u_j(\mathbf{x}_i) - z_{0j} < L\sigma\sqrt{n} + \varepsilon_1,$$

где ε_1 — любое положительное число. В итоге получаем

$$\|\mathbf{z}^* - \mathbf{z}_0\| = \left\| \frac{1}{N} \sum_{i=1}^{N} u(\mathbf{x}_i) - \mathbf{z}_0 \right\| =$$

$$= \sqrt{\sum_{j=1}^{n} \left[\frac{1}{N} \sum_{i=1}^{N} u_j(\mathbf{x}_i) - z_{0j} \right]^2} \leq$$

$$\leq \sqrt{n}(L\sigma\sqrt{n} + \varepsilon_1) = L\sigma n + \varepsilon_0 \qquad \bullet$$

Не для любого множества $\mathbb{Z}$ выполнено условие (2.1). В общем случае в качестве оценки формы $\mathbf{z}_0$ предлагается использовать форму одного из векторов выборки:

$$\mathbf{z}^* = \underset{\mathbf{u}(\mathbf{x}_j)}{\operatorname{argmin}} \sum_{i=1}^{N} \|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)\|^2. \tag{2.5}$$

Теорема 2. Пусть преобразование $\mathbf{u}$ 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

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

где $\mathbf{z}^*$ — оценка (2.5), ε_0 — любое положительное число.

Доказательство. Из модели порождения данных (1.2) следует, что

$$\frac{\boldsymbol{\varepsilon}_i}{\sigma} \sim \mathcal{N}(0, E_n\sigma^2),$$

— вектор с независимыми компонентами, имеющими стандартное нормальное распределение. Следовательно,

$$\frac{\|\mathbf{x}_i - f(\mathbf{z}_0, \boldsymbol{\alpha}_i)\|^2}{\sigma^2} = \left\| \frac{\boldsymbol{\varepsilon}_i}{\sigma} \right\|^2 \sim \chi^2(n).$$

По условию теоремы,

$$\|\mathbf{u}(\mathbf{x}_i) - \mathbf{z}_0\|^2 \leq L^2 \|\mathbf{x}_i - f(\mathbf{z}_0, \boldsymbol{\alpha}_i)\|^2,$$

следовательно,

$$E\|\mathbf{u}(\mathbf{x}_i) - \mathbf{z}_0\|^2 \leq L^2\sigma^2 \mathrm{E}\chi^2(n) = L^2\sigma^2 n.$$

Согласно неравенству Маркова, для любого a

$$P\left(\|\mathbf{u}(\mathbf{x}_i) - \mathbf{z}_0\|^2 \geq a^2\right) \leq \frac{\mathrm{E}\|\mathbf{u}(\mathbf{x}_i) - \mathbf{z}_0\|^2}{a^2} \leq \frac{L^2\sigma^2 n}{a^2}.$$

Выберем $a > 0$ так, чтобы

$$P\left(\|\mathbf{u}(\mathbf{x}_i) - \mathbf{z}_0\|^2 \geq a^2\right) \leq \frac{L^2\sigma^2 n}{a^2} = \frac{1}{4}, \tag{2.6}$$

$$a^2 = 4L^2\sigma^2 n.$$

Определим разбиение индексов объектов выборки $I = \{1, \ldots, N\} = I_1 \cup I_2$ следующим образом:

$$I_1 = \{i : \|\mathbf{u}(\mathbf{x}_i) - \mathbf{z}_0\|^2 \leq a^2\},$$

$$I_2 = \{i : \|\mathbf{u}(\mathbf{x}_i) - \mathbf{z}_0\|^2 > a^2\}.$$

Из условия (2.6) следует, что существует число N_1 такое, что для любого размера выборки $N > N_1$

$$|I_2| < \frac{N}{3}. \tag{2.7}$$

В последующих рассуждениях предполагаем, что $N > N_1$ и, соответственно, выполнено условие (2.7).

Пусть $\mathbf{x}_\beta$ — некоторый объект выборки, $b = \|\mathbf{u}(\mathbf{x}_\beta) - \mathbf{z}_0\|$. Покажем, что при выполнении некоторого условия для b вектор $\mathbf{u}(\mathbf{x}_\beta)$ не может быть равен оценке $\mathbf{z}^*$ из (2.5). Для этого необходимо найти такой элемент выборки $\mathbf{x}_\alpha$, для которого

$$\sum_{i=1}^{N} \|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_\alpha)\|^2 < \sum_{i=1}^{N} \|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_\beta)\|^2. \tag{2.8}$$

Пусть α — произвольный индекс из I_1. Тогда для любого $i \in I_1$ из неравенства треугольника

$$\|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_\beta)\| \geq b - a,$$

$$\|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_\alpha)\| \leq 2a.$$

Учитывая данные неравенства и условие (2.7) получаем

$$\sum_{i\in I_1} \|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_\beta)\|^2 - \sum_{i\in I_1} \|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_\alpha)\|^2 \geq$$

$$\geq \sum_{i\in I_1} \left[(b-a)^2 - (2a)^2\right] > \frac{2N}{3}\left[(b-a)^2 - (2a)^2\right]. \tag{2.9}$$

Для сокращения записи введём следующие обозначения:

$$\rho_{ij} = \|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)\|,$$

$$\rho_{0i} = \|\mathbf{u}(\mathbf{x}_i) - \mathbf{z}_0\|.$$

В новых обозначениях

$$\sum_{i\in I_2} \|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_\beta)\|^2 - \sum_{i\in I_2} \|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_\alpha)\|^2 = \\ = \sum_{i\in I_2} \left(\rho_{\beta i}^2 - \rho_{\alpha i}^2\right) = \sum_{i\in I_2} (\rho_{\beta i} - \rho_{\alpha i})(\rho_{\beta i} + \rho_{\alpha i}). \quad (2.10)$$

Из неравенства треугольника

$$|\rho_{\beta i} - \rho_{\alpha i}| \le \rho_{\alpha\beta},$$

$$\rho_{\beta i} - \rho_{\alpha i} \ge -\rho_{\alpha\beta} \ge -(a+b). \quad (2.11)$$

С учётом условия (2.7)

$$\sum_{i\in I_2} (\rho_{\beta i} + \rho_{\alpha i}) \le \sum_{i\in I_2} (\rho_{0i} + b + \rho_{0i} + a) \le \frac{N}{3}(a+b) + 2\sum_{i\in I_2} \rho_{0i}. \quad (2.12)$$

Т. к. $\rho_{0i} > a$ для $i \in I_2$, то

$$\sum_{i\in I_2} \rho_{0i} \le \sum_{i\in I_2} \frac{\rho_{0i}^2}{a} = \frac{1}{a}\sum_{i\in I_2} \rho_{0i}^2 \le \frac{1}{a}\sum_{i=1}^{N} \rho_{0i}^2.$$

Далее из условия Липшица получаем

$$\sum_{i\in I_2} \rho_{0i} \le \frac{1}{a}\sum_{i=1}^{N} \rho_{0i}^2 \le \frac{1}{a}\sum_{i=1}^{N} \left(L^2\|\mathbf{x}_i - f(\mathbf{z}_0, \boldsymbol{\alpha}_i)\|^2\right) = \frac{L^2\sigma^2}{a}\sum_{i=1}^{N} \left\|\frac{\boldsymbol{\varepsilon}_i}{\sigma}\right\|^2.$$

Согласно усиленному закону больших чисел,

$$\frac{1}{N}\sum_{i=1}^{N} \left\|\frac{\boldsymbol{\varepsilon}_i}{\sigma}\right\|^2 \xrightarrow{\text{almost sure}} \mathrm{E}\left\|\frac{\boldsymbol{\varepsilon}_i}{\sigma}\right\|^2 = \mathrm{E}\chi^2(n) = n,$$

следовательно, для любого $\varepsilon_1 > 0$ существует такое N_2, что

$$\frac{1}{N}\sum_{i=1}^{N} \left\|\frac{\boldsymbol{\varepsilon}_i}{\sigma}\right\|^2 < n + \varepsilon_1, \qquad N > N_2.$$

В итоге получаем

$$\sum_{i\in I_2}\rho_{0i} < \frac{L^2\sigma^2(n+\varepsilon_0)N}{a}. \tag{2.13}$$

Подставим в (2.10) неравенства (2.11), (2.12) и (2.13):

$$\sum_{i\in I_2}\|\mathbf{u}(\mathbf{x}_i)-\mathbf{u}(\mathbf{x}_\beta)\|^2 - \sum_{i\in I_2}\|\mathbf{u}(\mathbf{x}_i)-\mathbf{u}(\mathbf{x}_\alpha)\|^2 \geq -(a+b)\left[\frac{N}{3}(a+b)+\varkappa\right], \tag{2.14}$$

$$\varkappa = \frac{L^2\sigma^2(n+\varepsilon_0)N}{a}.$$

Объединяя (2.9) и (2.14), получаем

$$\begin{aligned}\sum_{i=1}^{N}\|\mathbf{u}(\mathbf{x}_i)-\mathbf{u}(\mathbf{x}_\beta)\|^2 - \sum_{i=1}^{N}\|\mathbf{u}(\mathbf{x}_i)-\mathbf{u}(\mathbf{x}_\alpha)\|^2 = \\ = \sum_{i\in I_1}\|\mathbf{u}(\mathbf{x}_i)-\mathbf{u}(\mathbf{x}_\beta)\|^2 - \sum_{i\in I_1}\|\mathbf{u}(\mathbf{x}_i)-\mathbf{u}(\mathbf{x}_\alpha)\|^2 + \\ + \sum_{i\in I_2}\|\mathbf{u}(\mathbf{x}_i)-\mathbf{u}(\mathbf{x}_\beta)\|^2 - \sum_{i\in I_2}\|\mathbf{u}(\mathbf{x}_i)-\mathbf{u}(\mathbf{x}_\alpha)\|^2 \geq \\ \geq \frac{2N}{3}\left[(b-a)^2-(2a)^2\right] - (a+b)\left[\frac{N}{3}(a+b)+\varkappa\right].\end{aligned}$$

Необходимо определить, при каких ограничениях на b выполнено

$$\frac{2N}{3}\left[(b-a)^2-(2a)^2\right] - (a+b)\left[\frac{N}{3}(a+b)+\varkappa\right] > 0.$$

Обозначив

$$\tilde{\varkappa} = \frac{\varkappa}{a\,(N/3)}, \qquad v = \frac{b}{a},$$

перепишем последнее неравенство в виде

$$2\left[(v-1)^2-4\right] - (1+s)[(1+s)+\tilde{\varkappa}] > 0,$$

$$v^2 - (6-\tilde{\varkappa})v - 7 - \tilde{\varkappa} > 0,$$

$$v > 7 + \tilde{\varkappa},$$

$$b > a(7+\tilde{\varkappa}) = a\left(7+\frac{3(n+\varepsilon_1)}{4n}\right) = \frac{31a}{4}+\varepsilon_0 = \frac{31}{2}L\sigma\sqrt{n}+\varepsilon_0. \tag{2.15}$$

Таким образом, при выполнении условия (2.15), $N > \max(N_1, N_2)$, выполнено неравенство (2.8) и, следовательно,

$$\mathbf{u}(\mathbf{x}_\beta) \neq \underset{\mathbf{u}(\mathbf{x}_j)}{\operatorname{argmin}} \sum_{i=1}^{N}\|\mathbf{u}(\mathbf{x}_i)-\mathbf{u}(\mathbf{x}_j)\|^2. \qquad \bullet$$

Оценки (2.2) и (2.5) не являются состоятельными оценками формы $\mathbf{z}_0$, т. к. при увеличении размеров выборки ошибка оценок не стремится к нулю. Однако теоремы 1 и 2 утверждают, что для достаточной большой выборки ошибка $\|\mathbf{z}^* - \mathbf{z}_0\|$ не превосходит некоторого порога, размер которого пропорционален константе Липшица инвариантного преобразования $\mathbf{u}$.

3 Разбиение сегментов на группы со схожей формой

В модели (1.2) рассматривалась единая для всей выборки $\mathbf{x}_i$ форма $\mathbf{z}_0$. Теперь будем предполагать, что существует разбиение индексов $\mathcal{I} = \{1, \ldots, N\} = \mathcal{I}_1 \cup \ldots \cup \mathcal{I}_s$ такое, что

$$\mathbf{x}_i = \mathbf{f}(\mathbf{z}_{0k}, \boldsymbol{\alpha}_i) + \boldsymbol{\varepsilon}_i, \qquad i \in I_k, \qquad \mathbf{z}_{0k} \in Z, \qquad k = 1, \ldots, s,$$

т. е. каждому набору индексов $\mathcal{I}_k$ соответствует своя форма $\mathbf{z}_{0k}$.

Для нахождения разбиения $\mathcal{I}_1 \cup \ldots \cup \mathcal{I}_s$ предлагается использовать один из алгоритмов кластеризации, определив функцию расстояния между индексами $\rho_{ij} = \|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)\|$, $i, j \in I$ как расстояние между формами i-го и j-го временного ряда. Далее по каждой подвыборке $\mathbf{x}_i, i \in I_k$ оценивается форма $\mathbf{z}_{0k}$ по формуле (2.2) или (2.5). С увеличением количества кластеров s уменьшается количество временных рядов в каждом кластере и, следовательно, увеличивается ошибка при оценке формы $\mathbf{z}_{0k}$.

3.1 Критерий различимости форм для подгрупп

Для целей прогнозирования важна интерпретируемость разбиения множества индексов $\mathcal{I}$, т. к. она позволяет предсказать форму, соответствующую некоторому периоду в будущем. Например, разбиение $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$ временного ряда с суточной периодикой на будни $\mathcal{I}_1$ и выходные $\mathcal{I}_2$ позволяет при прогнозировании предсказать форму $\mathbf{z}_{01}$ для буднего дня и $\mathbf{z}_{02}$ для выходного дня. В качестве примеров таких разбиений можно рассмотреть следующие:

- фазы календарных и суточных периодик: время суток, будни/выходные, время года, праздники/рабочие дни;

- разбиение на последовательности подряд идущих индексов: $\mathcal{I} = \{1, \ldots, N_1\} \cup \{N_1 + 1, \ldots, N\}$.

Возникает вопрос о том, существует ли значимое различие между формами векторов с индексами из $\mathcal{I}_1$ и $\mathcal{I}_2$ для некоторого априорного разбиения $\mathcal{I}_1 \cup \mathcal{I}_2$. Если различие выявлено, то предлагается оценивать формы $\mathbf{z}_{01}$ и $\mathbf{z}_{02}$ для временных рядов с индексами из $\mathcal{I}_1$ и $\mathcal{I}_2$ соответственно вместо общей оценки формы для всех временных рядов.

Чтобы формализовать задачу, рассмотрим формы векторов $\mathbf{u}(\mathbf{x}_i), i \in \mathcal{I}_1$ и $\mathbf{u}(\mathbf{x}_i), i \in \mathcal{I}_2$ как выборки некоторых случайных векторов J_1 и J_2. Необходимо определить, существует ли статистически значимое различие между распределением J_1 и J_2. В качестве нулевой гипотезы рассмотрим предположение о равенстве матожидания межклассового расстояния $E_{12} = \mathrm{E}[\rho_{ij} | i \in \tilde{I}_1, j \in \tilde{I}_2]$ среднему значению матожиданий внутриклассовых расстояний $\frac{1}{2}(\mathrm{E}[\rho_{ij} | i, j \in \mathcal{I}_1] + \mathrm{E}[\rho_{ij} | i, j \in \mathcal{I}_2])$:

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

Оценки максимального правдоподобия матожиданий

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

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

Используя оценки максимального правдоподобия для дисперсий $V\rho_{ij}$, найдём дисперсию оценок матожиданий

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

$$\mathrm{V}\hat{E}_{11} \approx \frac{\sum_{i,j \in \mathcal{I}_1, i<j} (\rho_{ij} - \hat{E}_{11})^2}{|\mathcal{I}_1|^2 (|\mathcal{I}_1| - 1)^2/4},$$

$$\mathrm{V}\hat{E}_{22} \approx \frac{\sum_{i,j \in \mathcal{I}_2, i<j} (\rho_{ij} - \hat{E}_{22})^2}{|\mathcal{I}_2|^2 (|\mathcal{I}_2| - 1)^2/4}.$$

Среднеквадратичное отклонение $\hat{M}$

$$\widehat{se}_M = \sqrt{\mathrm{V}\hat{E}_{12} + \frac{1}{4}(\mathrm{V}\hat{E}_{11} + \mathrm{V}\hat{E}_{22})}.$$

Считая, что $\hat{M} = \hat{E}_{12} - (\hat{E}_{11} + \hat{E}_{22})/2 \sim \mathcal{N}(0, \widehat{se}_M^2)$, находим

$$\text{p–value} = 1 - \Phi\left(\frac{\hat{M}}{\widehat{se}_M}\right) \tag{3.1}$$

для нулевой гипотезы $M = 0$ против альтернативы $M > 0$, где Φ — функция стандартного нормального распределения. При уровне значимости 0,05 значение p–value $< 0,05$ свидетельствует о значимом различии в распределении форм векторов с индексами из $\mathcal{I}_1$ и $\mathcal{I}_2$. В этом случае предлагается оценивать формы $\mathbf{z}_{01}$ для $\mathbf{u}(\mathbf{x}_i), i \in \mathcal{I}_1$ и $\mathbf{z}_{02}$ для $\mathbf{u}(\mathbf{x}_i), i \in \mathcal{I}_2$ вместо общей оценки формы для всей выборки.

3.2 Критерий качества семейства преобразований

Выбор семейства преобразований $\mathbf{f}$ и множества $\mathbb{Z}$ должен отражать экспертные представления о близости форм временных рядов.

Предположим, что существует разбиение индексов

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

заданное экспертом, такое, что вектора $\mathbf{x}_i, i \in \mathcal{I}_j$ из одного класса имеют схожую форму, в то время как вектора из разных классов имеют разную форму. Необходимо выбрать преобразование $\mathbf{f}$ и множество $\mathbb{Z}$ так, чтобы расстояния между формами временных рядов $\|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)\|$ наиболее точно отражали представления эксперта, т. е. минимизировать расстояния между формами векторов из одного класса

$$\|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)\|, \qquad J(i) = J(j),$$

и максимизировать расстояния между формами векторов из разных классов

$$\|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)\|, \qquad J(i) \neq J(j),$$

где $J(i) = k \Longleftrightarrow i \in \mathcal{I}_k$.

Предлагается вычислить среднее внутриклассовое расстояние

$$F_1 = \frac{\sum_{i<j}[J(i) = J(j)]\|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)\|}{\sum_{i<j}[J(i) = J(j)]}$$

и среднее межклассовое расстояние

$$F_2 = \frac{\sum_{i<j}[J(i) = J(j)]\|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)\|}{\sum_{i<j}[J(i) = J(j)]}.$$

В качестве критерия качества семейства используется отношение:

$$S(\mathbf{f}) = \frac{F_2}{F_1}. \tag{3.2}$$

4 Поиск моментов изменения формы

Предположим, что существует набор целых чисел $0 = \tau_0 < \tau_1 < \ldots < \tau_K < \tau_{K+1} = N$ такой, что

$$\mathbf{x}_i = \mathbf{f}(\mathbf{z}_{0k}, \boldsymbol{\alpha}_i) + \boldsymbol{\varepsilon}_i, \qquad \tau_k < i \leq \tau_{k+1}, \qquad \mathbf{z}_{0k} \in \mathbb{Z}, \qquad k = 0, \ldots, K.$$

Числа $\tau_1, \ldots, \tau_K$ будем называть точками разладки; они разбивают последовательность временных рядов $\mathbf{x}_i, i \in \mathcal{I} = \{1, \ldots, N\}$ на подпоследовательности с индексами $\mathcal{I}_0 = \{1, \ldots, \tau_1\}, \ldots, \mathcal{I}_{K+1} = \{\tau_K + 1, \ldots, N\}$, и каждой подпоследовательности соответствует некоторая форма $\mathbf{z}_{0k}$.

Необходимо найти точки разладки $\tau_1, \ldots, \tau_K$ при условии, что их количество K неизвестно.

Для решения задачи предлагается адаптировать алгоритм поиска разладки с помощью дискретной производной и вычислением достижимого уровня значимости (Filtered Derivative with p–Values, FDp–V), предложенный в работе [13].

Определим дискретную производную в точке t как разницу между оценками форм $\mathbf{z}^*$, вычисленных по выборкам из двух скользящих окон ширины A слева и справа от точки t:

$$D(t, A) = \mathbf{z}^*(t, A) - \mathbf{z}^*(t - A, A), \tag{4.1}$$

где $\mathbf{z}^*(t, A)$ — оценка формы по выборке $\{\mathbf{x}_i : \tau_t < i \leq \tau_t + A\}$, вычисленная по формуле (2.2) или (2.5).

В соответствии с алгоритмом FDp–V локальные максимумы функции $\|D(t, A)\|$ по t рассматриваются как возможные точки разладки. Однако некоторые локальные максимумы могут не соответствовать точкам разладки и возникать вследствие неточности оценок $\mathbf{z}^*(t, A)$.

Алгоритм FDp–V состоит из двух шагов:

1. Поиск потенциальных точек разладки $\tilde{\tau}_k$ как локальных максимумов дискретной производной.

2. Отделение истинных точек разладки от "ложных тревог", применяя для каждой найденной точки $\tilde{\tau}_k$ статистический критерий.

Поиск потенциальных точек разладки. В качестве потенциальных точек разладки $\tilde{\tau}_k$ выбираются точки локальных максимумов функции $\|D(t, A)\|$ по t, в которых значение $\|D(t, A)\|$ превышает некоторый порог λ. Порог необходимо выбрать таким образом, чтобы вероятность появления в выборки "ложной тревоги"не превышала заданной величины p_1. В частности, при отсутствии истинных точек разладки необходимо выполнение условия

$$P(\max_t \|D(t, A)\| > \lambda) = p_1. \tag{4.2}$$

Чтобы оценить значение λ, удовлетворяющее условию (4.2), введём следующую статистическую модель. Рассмотрим $\mathbf{u}(\mathbf{x}_i)$ как независимые случайные вектора с соответствующими распределениями ξ_i. Нулевая гипотеза

$$H_0 : \qquad \xi_1 = \ldots = \xi_N$$

отвергается в пользу альтернативы

$$H_1 : \qquad \exists K, 0 = \tau_0 < \ldots < \tau_{K+1} = N :$$

$$\xi_1 = \ldots = \xi_{tau_1} \neq \xi_{\tau_1+1} = \ldots = \xi_{\tau_2} \neq \ldots \neq \xi_{\tau_K+1} = \ldots = \xi_N$$

при условии

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

Порог λ, задающий необходимую вероятность ошибки I рода

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

предлагается оценить с помощью бутстрепа. Для всех $i = 1, \ldots, M$, $M \sim 10^4 - 10^5$, выполним следующие шаги:

- Сгенерируем последовательность $\tilde{\mathbf{x}}_i$ с помощью случайной перестановки индексов исходной последовательности $\mathbf{x}_i$.

- Вычислим $S_i = \max_t \|D(t, A)\|$, где $D(t, A)$ подсчитана по последовательности $\tilde{\mathbf{x}}_i$.

В качестве оценки порога λ возьмём

$$\lambda = S_{(N(1-p_1))}, \tag{4.3}$$

где $S_{(1)}, \ldots, S_{(N)}$ — отсортированные по возрастанию значения S_i.

Таким образом, поиск потенциальных точек разладки включает следующие этапы:

1. Выбор из априорных соображений уровня значимости p_1 и размера окна A, оценка порога λ из (4.3).
2. Инициализация чисел $d_t = \|D(t, A)\|, t \in [A, N - A]$, счётчика $k = 0$.
3. До тех пор, пока $\max_t d_t > \lambda$, выполнять:
   - $k = k + 1$;
   - $\tilde{\tau}_k = \operatorname{argmax}_t d_t$;
   - $d_t = 0$ для всех $t \in (\tilde{\tau}_k - A; \tilde{\tau}_k + A)$.
4. Сортировка чисел $\tilde{\tau}_k$.

Исключение ложных тревог. Пусть $\tilde{\tau}_1, \ldots, \tilde{\tau}_{\tilde{K}}$ — отсортированная последовательность потенциальных точек разладки. Необходимо отфильтровать из последовательности ложные тревоги — точки, включенные в эту последовательность вследствие флукциаций функции $\|D(t, A)\|$. Поскольку точки $\tilde{\tau}_1, \ldots, \tilde{\tau}_{\tilde{K}}$ разбивают последовательность временных рядов $\mathbf{x}_i$ на подпоследовательности с индексами $\mathcal{I}_0 = \{1, \ldots, \tilde{\tau}_1\}, \ldots, \mathcal{I}_{\tilde{K}+1} = \{\tilde{\tau}_{\tilde{K}} + 1, \ldots, N\}$, то исключение точки $\tilde{\tau}_k$ из списка потенциальных точек разладки эквивалентно объединению подпоследовательностей временных рядов $\mathbf{x}_i, i \in \mathcal{I}_{k-1}$ и $\mathbf{x}_i, i \in \mathcal{I}_k$. В качестве критерия объединения используется критерий различимости форм на подвыборках (3.1).

Выполним следующие шаги:

1. Инициализируем $k = 1$, выбираем уровень значимости p_2.

2. До тех пор, пока $k \leq \tilde{K}$:

   - Вычисляем p–value по формуле (3.1), выбрав в качестве множества индексов $\mathcal{I}_1 = \{\tilde{\tau}_{k-1} + 1, \ldots, \tilde{\tau}_k\}$, $\mathcal{I}_2 = \{\tilde{\tau}_k + 1, \ldots, \tilde{\tau}_{k+1}\}$. (Считаем, что $\tilde{\tau}_0 = 0$, $\tilde{\tau}_{\tilde{K}+1} = N$.)
   - Если p–value $\geq p_2$, то исключаем $\tilde{\tau}_k$ из последовательности, уменьшая $\tilde{K}$ на единицу. В противном случае $k = k + 1$.

5 Вычислительный эксперимент

Рассмотрим применение методов, предложенных в работе, для анализа и прогнозирования потребления электроэнергии.

Пусть временной ряд y_t содержит данные о почасовом потреблении электроэнергии в Новосибирске. На рис. 1 показана зависимость потребления электроэнергии от времени в течение недели, с понедельника по воскресенье.

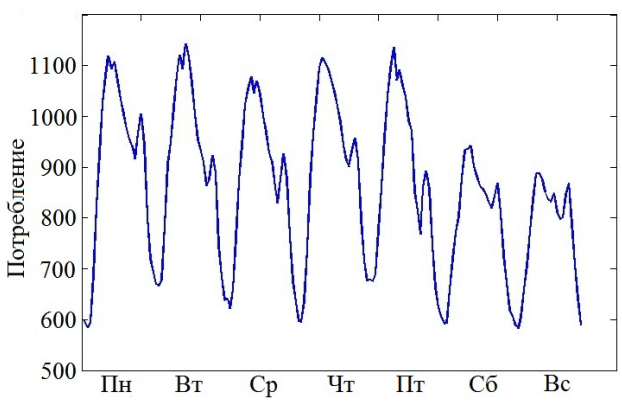

Рис. 1: Потребление электроэнергии в течение недели

Временной ряд y_t разбивается на сегменты $\mathbf{x}_i$ так, что компоненты $\mathbf{x}_i$ содержат данные о потреблении электроэнергии в течении каждого часа i-х суток в году. На рис. 2 показаны сегменты $\mathbf{x}_i$ для будних дней восьми последовательных недель.

Для построения полупараметрической модели выберем в качестве семейства преобразований сдвиг вдоль оси значений временного ряда:

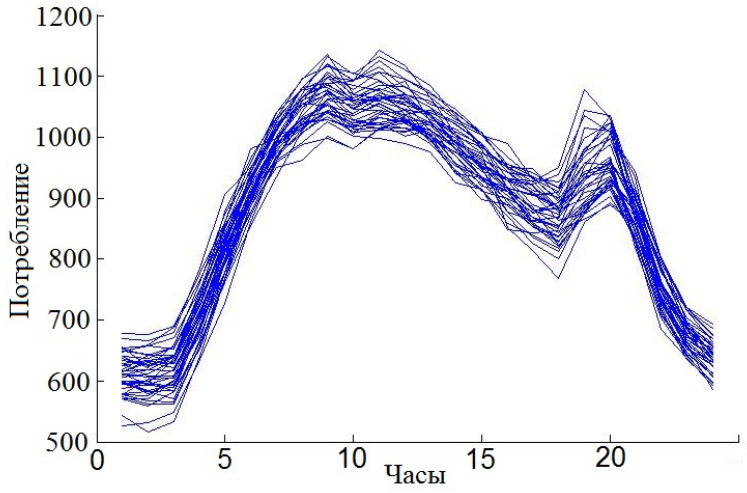

Рис. 2: Потребление электроэнергии в будние дни

$$\mathbf{f}(\mathbf{x}, \alpha) = \mathbf{x} + (\alpha, \ldots, \alpha)^{\mathsf{T}}.$$

Таким образом, согласно модели (1.2), все сегменты рассматриваются как один временной ряд, смещённый вдоль оси значений на некоторые величины α_i. Определим множество $\mathbb{Z}$ как

$$\mathbb{Z} = \{\mathbf{z} : \sum_{j=1}^{24} z_j = 0\}.$$

Тогда для сегмента $\mathbf{x}_i$ смещение

$$\alpha_i = v(\mathbf{x}_i) = \frac{1}{24} \sum_{j=1}^{24} x_j,$$

форма

$$\mathbf{u}(\mathbf{x}_i) = \mathbf{x}_i - (v(\mathbf{x}_i), \ldots, v(\mathbf{x}_i))^{\mathsf{T}}.$$

Очевидно, для множества $\mathbb{Z}$ выполнено условие (2.1), поэтому в качестве оценки формы $\mathbf{z}_0$ из модели (1.2) используем оценку (2.2). На рис. 3 изображены сегменты $\mathbf{x}_i$ для будних дней восьми последовательных недель (синие линии), и вычисленная по ним форма $\mathbf{z}_0$ (красная линия), смещённая вверх для компактности графика.

Смещения α_i в зависимости от номера буднего дня изображены на рис. 4 (выходные на графике пропущены).

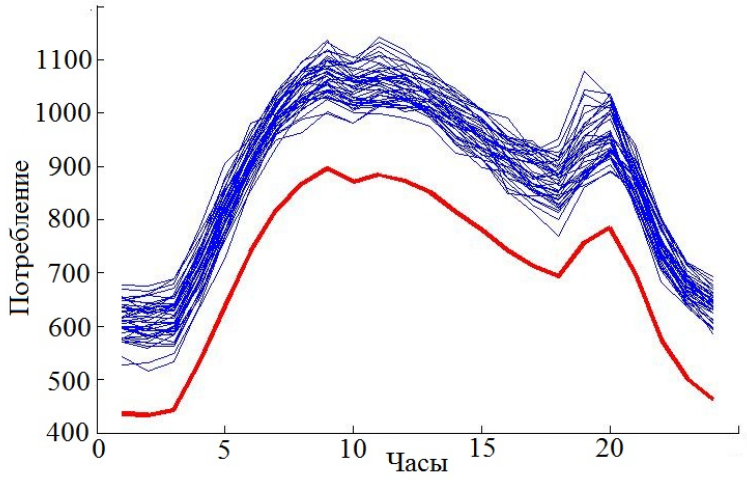

Рис. 3: Форма сегментов для будних дней

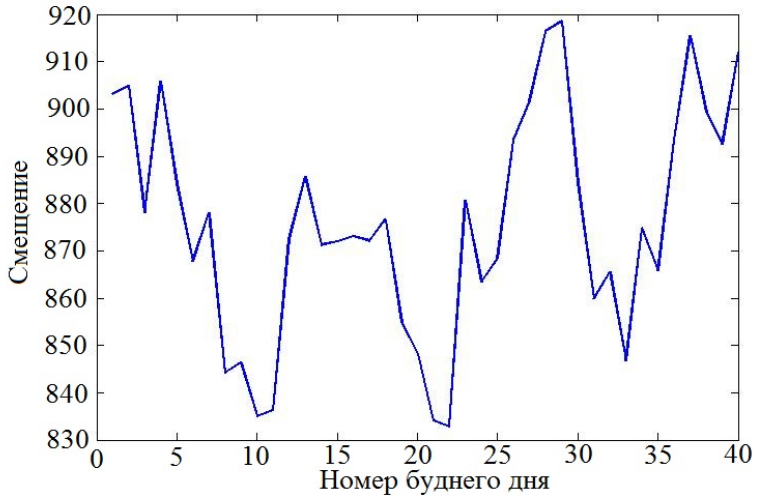

Рис. 4: Смещения сегментов для будних дней

5.1 Разбиение сегментов временного ряда по форме

Для выбора семейства преобразований $\mathbf{f}$ используем критерий качества (3.2). Необходимо задать априорное разбиение сегментов $\mathbf{x}_i$ на группы со схожей формой. На рис. 5 изображены графики сегментов $\mathbf{x}_i$ для будней (синие линии) и выходных (красные линии).

Предположив, что сегменты из двух групп имеют разные формы, определим априорное разбиение сегментов на будни и выходные. Далее приведены значения критерия $S(\mathbf{f})$ из (3.2) для некоторых семейств преобразований:

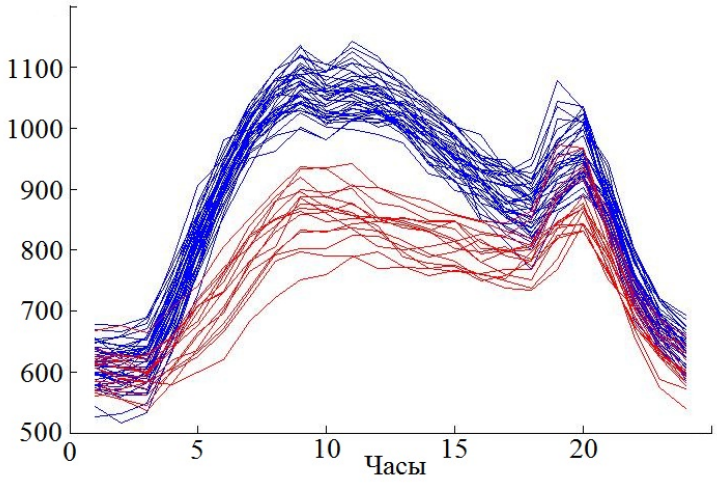

Рис. 5: Сегменты временного ряда: будни и выходные

1. $\mathbf{f}(\mathbf{x}, \alpha_0) = \mathbf{x} + (\alpha_0, \ldots, \alpha_0)^\intercal$: $S(\mathbf{f}) = 2.3851$;

2. $f_j(\mathbf{x}, \boldsymbol{\alpha}) = x_j + \alpha_0 + \alpha_1 j$, $j = 1, \ldots, n$: $S(\mathbf{f}) = 2.7449$;

3. $f_j(\mathbf{x}, \boldsymbol{\alpha}) = x_j + \alpha_0 + \alpha_1 j + \alpha_2 j^2$, $j = 1, \ldots, n$: $S(\mathbf{f}) = 1.7250$;

4. $f_j(\mathbf{x}, \boldsymbol{\alpha}) = x_j + \sum_{m=0}^{3} \alpha_m j^m$, $j = 1, \ldots, n$: $S(\mathbf{f}) = 1.3328$;

5. $f_j(\mathbf{x}, \boldsymbol{\alpha}) = x_j + \sum_{m=0}^{4} \alpha_m j^m$, $j = 1, \ldots, n$: $S(\mathbf{f}) = 1.3300$;

6. $f_j(\mathbf{x}, \boldsymbol{\alpha}) = x_j + \sum_{m=0}^{5} \alpha_m j^m$, $j = 1, \ldots, n$: $S(\mathbf{f}) = 1.1999$;

7. $\mathbf{f}(\mathbf{x}, \alpha_0) = \alpha_0 \mathbf{x}$: $S(\mathbf{f}) = 2.1014$;

8. $\mathbf{f}(\mathbf{x}, \boldsymbol{\alpha}) = \alpha_0 \mathbf{x} + \alpha_1$: $S(\mathbf{f}) = 1.9088$;

9. $f_j(\mathbf{x}, \boldsymbol{\alpha}) = \alpha_0 x_j + \alpha_1 + \alpha_2 j$, $j = 1, \ldots, n$: $S(\mathbf{f}) = 1.9790$.

Представленные значения $S(\mathbf{f})$ показывают, что если семейством преобразований является прибавление ко временному ряду полинома от времени, то качество разбиения на будни и выходные достигает максимума для линейной функции, и значительно убывает с увеличением числа параметров.

Чтобы визуально сравнить качество разбиения для различных преобразований, используем двумерное шкалирование. Определим расстояние между сегментами

$\rho(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)\|$ как расстояние между их формами. Сопоставляя каждому сегменту $\mathbf{x}_i$ точку $\mathbf{h}_i$ на плоскости, расположим точки $\mathbf{h}_i$ так, чтобы расстояния между ними $\rho(\mathbf{h}_i, \mathbf{h}_j)$ были близки к соответствующим расстояниям $\rho(\mathbf{x}_i, \mathbf{x}_j)$ между сегментами. Минимизируя $\sum_{i<j} |\rho(\mathbf{h}_i, \mathbf{h}_j) - \rho(\mathbf{h}_i, \mathbf{h}_j)|$ градиентными методами, находим требуемое расположение точек $\mathbf{h}_i$.

На рис. 6 изображено двумерное шкалирования сегментов $\mathbf{x}_i$ для семейства преобразований 2 из списка, приведённого выше. Синим точкам соответствуют будние дни, красным — выходные.

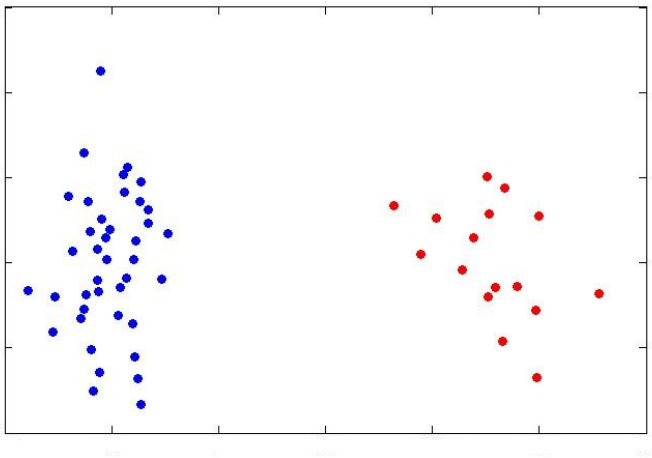

Рис. 6: Двумерное шкалирование для преобразований 2

На рис. 7 изображено двумерное шкалирование для семейства преобразований 9 с более низким качеством разбиения.

На рис. 8 изображено двумерное шкалирование для семейства преобразований 6, которое не обеспечивает разбиение на будни и выходные и, следовательно, не соответствует априорному представлению о близости форм сегментов.

Далее в эксперименте используется семейство преобразований 2.

Рассмотрим сегменты временного ряда $\mathbf{x}_i$, соответствующие 22-м последовательным неделям (154 сегмента). Определив расстояние между ними $\rho(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{u}(\mathbf{x}_i) - \mathbf{u}(\mathbf{x}_j)\|$, выполним иерархическую кластеризацию сегментов. На рис. 9 в каждой таблице представлен календарь, где каждой строке соответствует неделя, каждому столбцу — день недели, цвет ячейки соответствует кластеру соответствующего сегмента. В разных таблицах представлена различная степень разбиения на

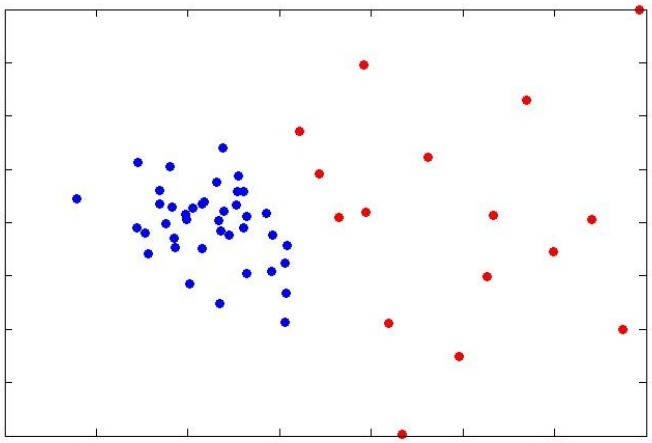

Рис. 7: Двумерное шкалирование для преобразований 9

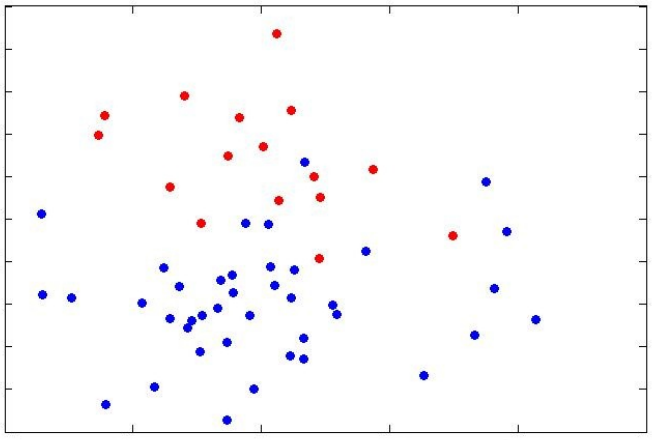

Рис. 8: Двумерное шкалирование для преобразований 6

кластеры; в схемах, расположенных напротив таблиц, показана иерархия кластеров.

Из таблиц видно, что форма временного ряда в дни государственных праздников (9, 10 мая и 13 июня) близка к форме в выходные, а 20 августа происходит изменение формы как для будней, так и для выходных.

Рассмотрим пример использования критерия различимости форм для двух групп сегментов временного ряда, предложенный в разделе 3.1. На рис. 10 представлены таблицы с календарями, аналогичными изображённым на рис. 9, где красным и синим цветами обозначены группы сегментов временного ряда и значение достижимого уровня значимости, подсчитанного по формуле (3.1) по соответствующим груп-

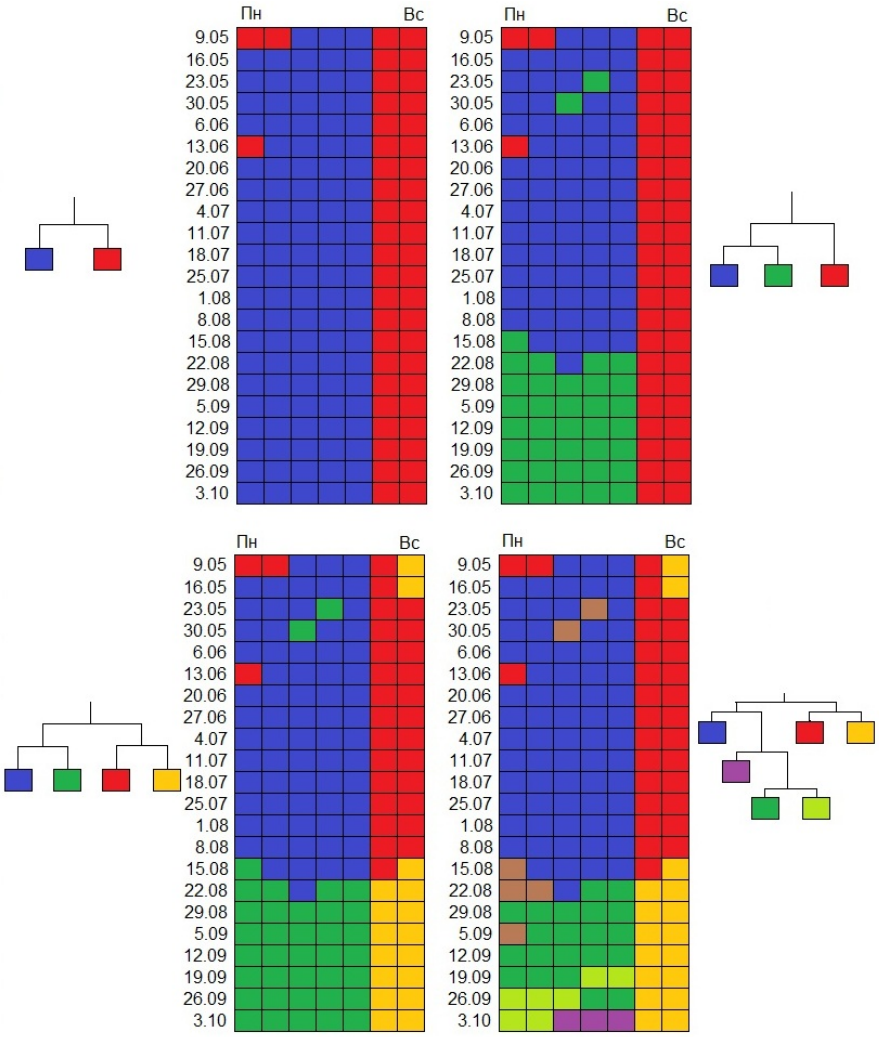

Рис. 9: Иерархическая кластеризация

пам. Значение достижимого уровня значимости, близкое к нулю, свидетельствует в значимом различии форм сегментов из разных групп. Значимое различие зафиксировано для будней и выходных, а также для понедельников и воскресений. Для других групп различие не выявлено.

На рис. 11 представлены достижимые уровни значимости для групп сегментов, соответствующих различным периодам времени.

Рассмотрим пример работы алгоритма поиска моментов изменения формы, предложенный в разделе 4. На первом этапе вычисляется дискретная производная $D(t, A)$ по определению (4.1) с окном $A = 10$, где оценка $\mathbf{z}^*(t, A)$ вычисляется по формуле (2.2). На рис. 12 показана норма дискретной производной $\|D(t, A)\|$ в зависимости от времени.

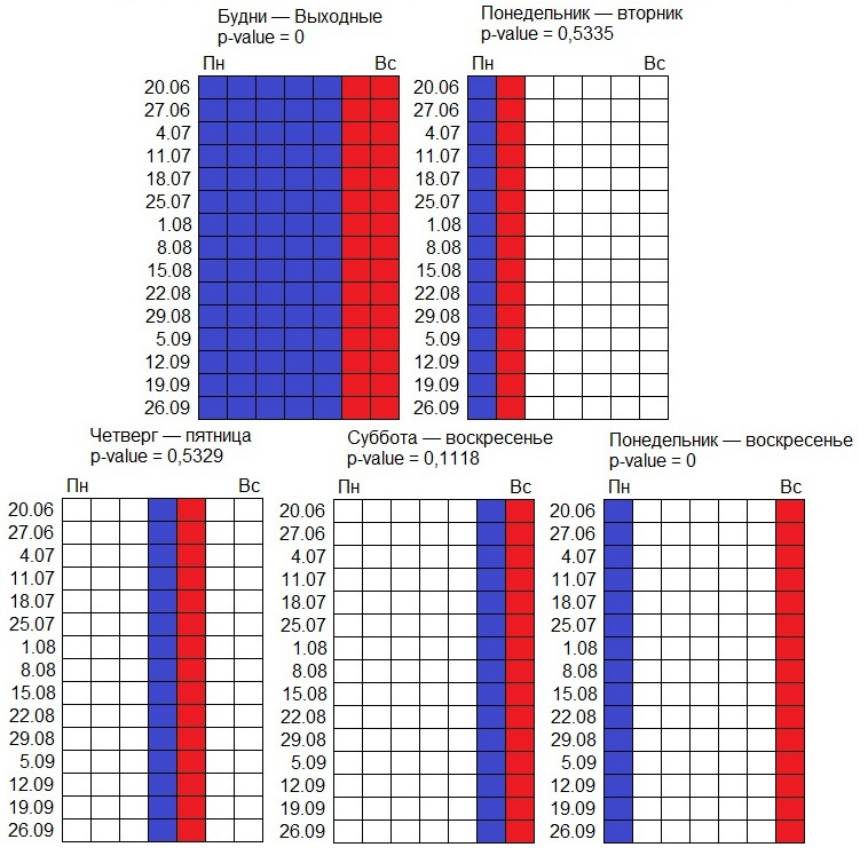

Рис. 10: Критерий различимости форм (1)

Далее с помощью бутстрепа вычисляется порог λ так, как описано в разделе 4. Локальные максимумы $\|D(t, A)\|$ по времени, лежащие выше порога, рассматриваются как потенциальные точки разладки. На рис. 13 отмечен порог и кружками показаны потенциальные точки разладки.

В соответствии с процедурой из раздела 4 из числа потенциальных точек разладки исключаются ложные тревоги. На рис. 14 ложные тревоги выделены красными кружками. Остальные точки рассматриваются как истинные точки разладки и разбивают временной ряд на отрезки, каждому из которых соответствует своя форма $\mathbf{z}_i$.

5.2 Иерархическая прогностическая модель

Предположим, что значения

$$y_t, \qquad t = 1, \ldots, T_0, \qquad T_0 = N_0 n$$

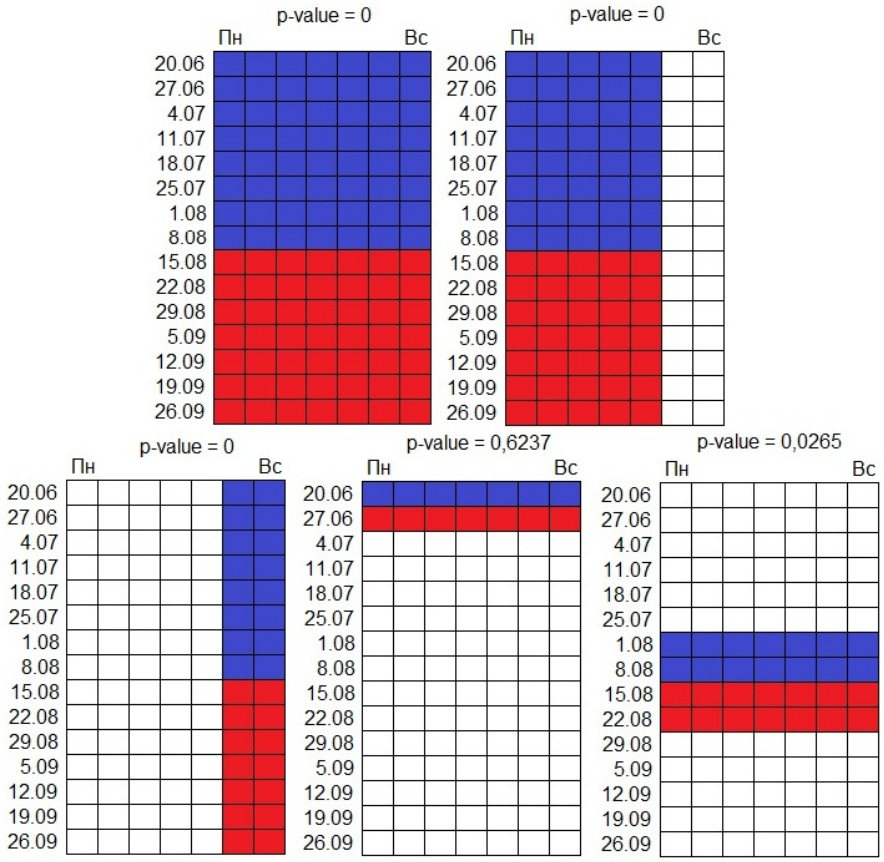

Рис. 11: Критерий различимости форм (2)

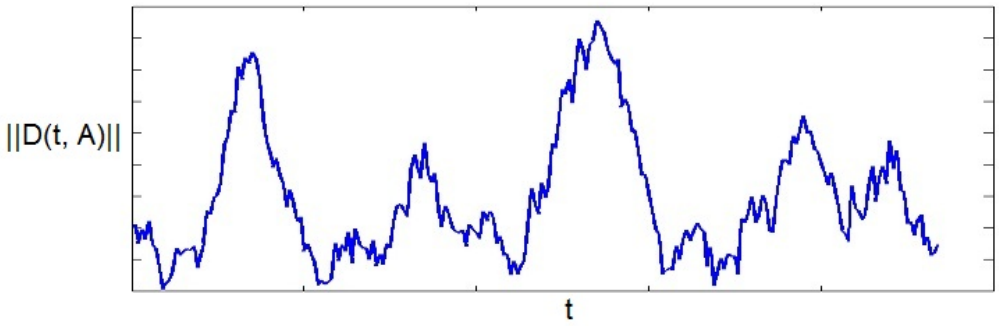

Рис. 12: Норма дискретной производной

временного ряда y_t известны, значения

$$y_t, \qquad t = T_0 + 1, \ldots, T = Nn$$

неизвестны и их необходимо спрогнозировать. Предлагается следующая двухуровневая прогностическая модель.

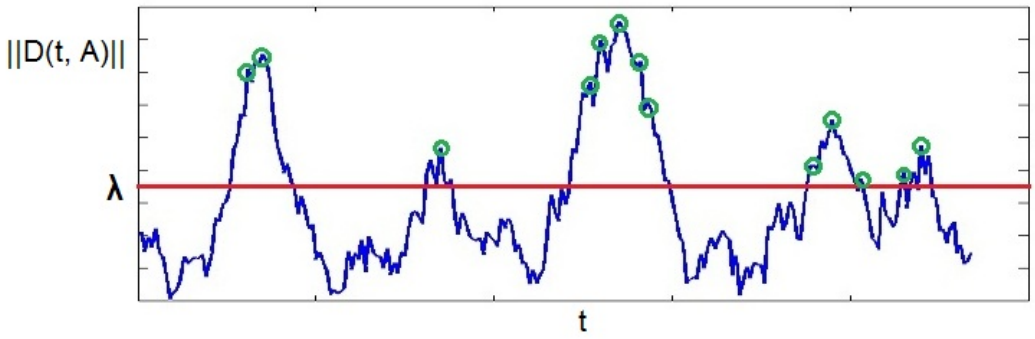

Рис. 13: Потенциальные точки разладки

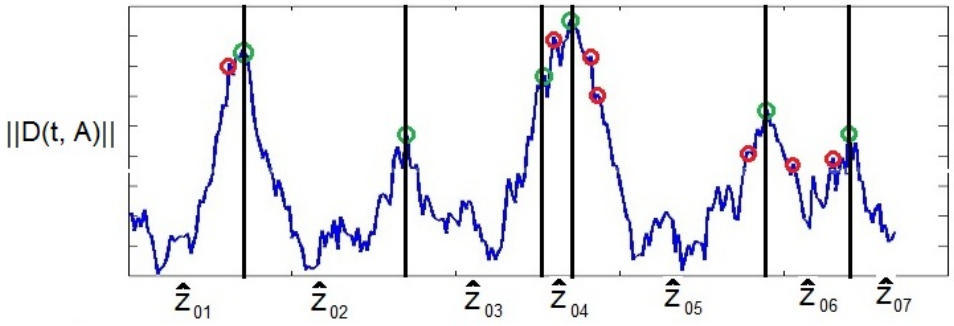

Рис. 14: Исключение ложных тревог

Временной ряд разбивается на сегменты, как в п. 1:

$$\mathbf{x}_i = (y_{(i-1)n+1}, y_{(i-1)n+2}, \ldots, y_{in}), \qquad i = 1, \ldots, N,$$

при этом значения $\mathbf{x}_i, i = 1, \ldots, N_0$ известны, значения $\mathbf{x}_i, i = N_0 + 1, \ldots, N$ неизвестны.

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

$$\mathbf{x}_i = \mathbf{f}(\mathbf{z}_{0k}, \boldsymbol{\alpha}_i) + \boldsymbol{\varepsilon}_i, \qquad i \in I_k, \qquad \mathbf{z}_{0k} \in Z, \qquad k = 1, \ldots, s.$$

Построение разбиения $\mathcal{I}_1 \cup \ldots \cup \mathcal{I}_s$ выполняется в два шага:

1. Построение априорного разбиения $\tilde{\mathcal{I}}_1 \cup \ldots \cup \tilde{\mathcal{I}}_{\tilde{s}}$ в соответствии с особенностями периодов исходного временного ряда (примеры таких разбиений представлены в п. 3). При этом множество индексов $\tilde{\mathcal{I}}_i$ разбивается на подмножества $\tilde{\mathcal{I}}_j$ и $\tilde{\mathcal{I}}_k$ только тогда, когда между формами сегментов с соответствующими индексами существует значимое различие. Достижимый уровень значимости вычисляется по формуле (3.1).

2. Каждая группа $\tilde{\mathcal{I}}_k$ разбивается на подгруппы с помощью алгоритма обнаружения разладок, предложенного в п. 4. Индексы, соответствующие сегментам $\mathbf{x}_i$ с неизвестными значениями временного ряда, относятся к последней по времени подгруппе.

Для каждой группы индексов $\mathcal{I}_k$ рассматривается многомерный временной ряд

$$\mathbf{v}(\mathbf{x}_i), i \in \mathcal{I}_k,$$

значения которого известны для $\mathbf{v}(\mathbf{x}_i), i \leq N_0$. Значения $\mathbf{v}(\mathbf{x}_i), N_0 < i \leq N$ оцениваются с помощью одного из алгоритмов прогнозирования многомерных временных рядов.

Таким образом, первый уровень иерархической модели образует полупараметрическая модель для сегментов временного ряда, связывающая целевые значения временного ряда с формой и параметрами полупараметрической модели. На втором уровне рассматриваются последовательности форм и параметров, соответствующие исходной последовательности сегментов.

Рассмотрим прогнозирование потребления электроэнергии на неделю. Сегменты временного ряда разбиваются на будни и выходные, далее для каждой из групп находятся моменты изменения формы с использованием метода, предложенного в разделе 4. Для каждой полученной группы сегментов выполняется прогнозирование значений параметров методом экспоненциального сглаживания. На рис. 15 показаны значения параметров для исторических данных в течение последних четырёх недель и прогноза: кружками отмечены параметры α_0, квадратами — параметры α_1. Красные линии отвечают форме, соответствующей выходным, синие — будням до момента последней разладки, зелёные — будням после разладки.

Заключение

В работе решена задача построения прогностических моделей, описывающих периодические временные ряды и включающие инвариантные преобразования. Сегменты временного ряда описывались полупараметрической моделью, для которой были предложены методы оценки формы. Доказано, что ошибка этих методов не превосходит величины, пропорциональной константе Липшица инвариантного преобра-

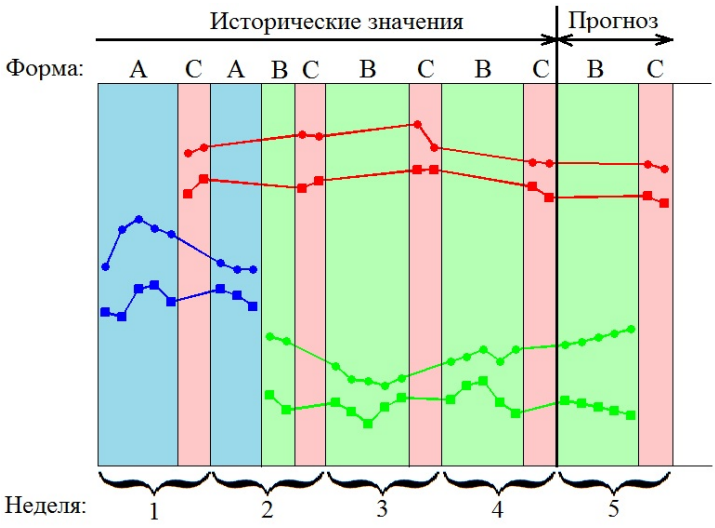

Рис. 15: Форма и параметры полупараметрической модели

зования. Было предложено использовать кластеризацию для разбиения сегментов временного ряда на группы со схожей формой. Предложен критерий качества семейства преобразований в полупараметрической модели. Для поиска моментов изменения формы сегментов модифицирован алгоритм обнаружения разладки с использованием дискретной производной и вычислением достижимого уровня значимости. Предложена иерархическая прогностическая модель, основанная на полупараметрической модели.

Вычислительный эксперимент, проведённый на данных о потреблении электроэнергии, показал, что формы суточных сегментов временного ряда в наибольшей степени различаются для будней и выходных, причём праздничные дни близки к выходным. Также форма заметно меняется с течением времени.

Таким образом, основные результаты работы:

- 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.

Список литературы

[1] Lawton W. Self modeling nonlinear regression. / *W. Lawton, E. Sylvestre, M. Maggio* // Technometrics. — 1972. — Vol. 14. — P. 513—532.

[2] Dalalyan A. S. Penalized Maximum Likelihood and Semiparametric Second–Order Efficiency / *A. S. Dalalyan, G. K. Golubev, A. B. Tsybakov* // The Annals of Statistics. — 2006. — Vol. 34. — P. 169–201.

[3] Kneip A. Model Estimation in Nonlinear Regression under Shape Invariance. / *A. Kneip, J. Engel* // The Annals of Statistics. — 1995. — Vol. 23. — P. 551–570.

[4] Kneip A. Convergence and Consistency Results for Self–Modeling Nonlinear Regression. / *A. Kneip, T. Gasser* // The Annals of Statistics. — 1988. — Vol. 16. — P. 82–112.

[5] Bercu B. A Robbins–Monro Procedure for Estimation in Semiparametric Regression Models. / *B. Bercu, Ph. Fraysse* // The Annals of Statistics. — 2012. — Vol. 40. — P. 666–693.

[6] Coull B. A. Self–Modeling Regression for Multivariate Curve Data. / *B. A. Coull, J. Staudenmayer* // Statistica Sinica. — 2005. — Vol. 14. — P. 695–711.

[7] Gamboa F. Semi–parametric Estimation of Shifts. / *F. Gamboa, J.–M. Loubes* // Electronic Journal of Statistics. — 2007. — Vol. 1 — P. 616–640.

[8] Hurtgen H. Semiparametric Shape–invariant Models for Periodic Data. / *H. Hurtgen, D. Gervini* // Journal of Applied Statistics. — 2009. — P. 1055–1065.

[9] Rice J. A. Estimating the Mean and Covariance Structure Nonparametrically when the Data are Curve. / *J. A. Rice, B. W. Silverman* // Journal of the Royal Statistical Society. — 1991. — Vol. 53. — P. 233–243.

[10] Wang K. Synchronizing Sample Curves Nonparametrically. / *K. Wang, T. Gasser* // The Annals of Statistics. — 1999. — Vol. 27. — P. 233–243.

[11] Lindstrom M. J. Self Modeling with Random Shift and Scale Parameters and a Free–knot Spline Shape Function. / *M. J. Lindstrom* // Statistic Medicine. — 1995. — P. 2009–2021.

[12] Lavielle M. Detection of Multiple Change–Points in Multivariate Time Series. / *M. Lavielle, G. Teyssiere* // Lithuanian Mathematical Journal. — 2006. — Vol. 46. — P. 287–306.

[13] Bertrand P. Off–line Detection of Multiple Change Points with the Filtered Derivative with p–Value Method. / *P. Bertrand, M. Fhima, A. Guillin* // Sequential Analysis. — 2010. — Vol. 26. — P. 439–460.

[14] Vimond M. Efficient Estimation for a Subclass of Shape Invariant Models. / *M. Vimond* // The Annals of Statistics. — 2010. — Vol. 38. — P. 1885–1912.

[15] Hardle W. Semiparametric Comparison of Regression Curves. / *W. Hardle, J. S. Marron* // The Annals of Statistics. — 1990. — Vol. 18. — P. 63–89.

[16] Wang Y. Shape Invariant Modelling of Circadian Rhythms with Random Effects and Smoothing Spline ANOVA Decompositions. / *Y. Wang, Ch. Ke, M. B. Brown* // Biometrics. — 2003. — P. 804—812.