Проблема понижения размерности в задаче поиска аномалий в многомерных временных рядах

Д. Д. Яшков

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

Научные руководители: К. В. Воронцов, В. А. Лобачёв

1 июля 2014 г.

Постановка задачи

Дано:

- X^i – объект, описывающийся многомерным временным рядом
- X^i_{jt} – показания временного ряда.
- i – индекс объекта, $i = \overline{1, N}$
- j – номер временного ряда, $j = \overline{1, J}$,
  $J = J_c \cup J_d$, где:
  J_c – непрерывные временные ряды;
  J_d – дискретные временные ряды.
- t – момент времени, $t = \overline{1, T_i}$

Требуется: Предложить преобразование исходных данных, уменьшающее размерность пространства (N, J, T), не теряя возможные аномалии.

Постановка задачи

Предположения:

- ряды описывают один и тот же процесс, происходящий с различными объектами, например: показания датчиков в течение различных полетов одного и того же самолета;
- время измерения и число временных рядов описывающих объекты велико;
- объекты можно разбить на участки однородности;
- аномалии – маловероятные события: — внутри объектов; — на множестве объектов.

Обзор предыдущих работ

Алгоритмы, работающие с многомерными временными рядами:

- *Santanu Das, 2010.* Алгоритм MKAD. Используется обобщение на многомерный случай метрики между последовательностями, основанной на *nLCS* – наибольшей общей подпоследовательности. Кластеризация объектов.
- *Srivastava2005.* Данные – многомерные бинарные временные ряды. Делается кластеризация каждого объекта по времени методом фон Мизеса-Фишера. Таким образом сводят каждый объект к одномерному дискретному ряду.
- *Tsay2000.* Данные – многомерные непрерывные. Настраивается модель VARIMA, моменты, когда у нее большие остатки – аномальные
- *Baragona2007.* Данные – многомерные непрерывные. Используется метод независимых компонент(ICA). аномальные моменты оказываются в первой компоненте.

Описание алгоритма

Алгоритм можно разбить на две части:

1. понижение размерности по времени:
   - дискретизация временных рядов;
   - сегментация объектов;
   - кластеризация сегментов.
2. понижение размерности по компонентам многомерного ряда.

Дискретизация временных рядов

Вход: X_j, где $j \in J_c$ – совокупность j-х временных рядов всех объектов;
n – размер алфавита в который будет преобразован ряд;
$\{p_k\}_{k=0}^{n}$ – вероятности для квантилей, $p_0 = 0$, $p_n = 1$.
Выход: преобразованный ряд X_j, $X_{jt} \in (a_1, \ldots, a_n)$.

- считаем эмпирическую функцию распределения F_j для X_j;
- считаем квантили q_k: $P(X_j < q_k) = F_j(q_k) = p_k$;
- в соответствии с полученными квантилями преобразуем исходные данные: если $X_j \in [q_{k-1}, q_k]$, то заменяем это значение на символ a_k.

Пример дикретизации временного ряда

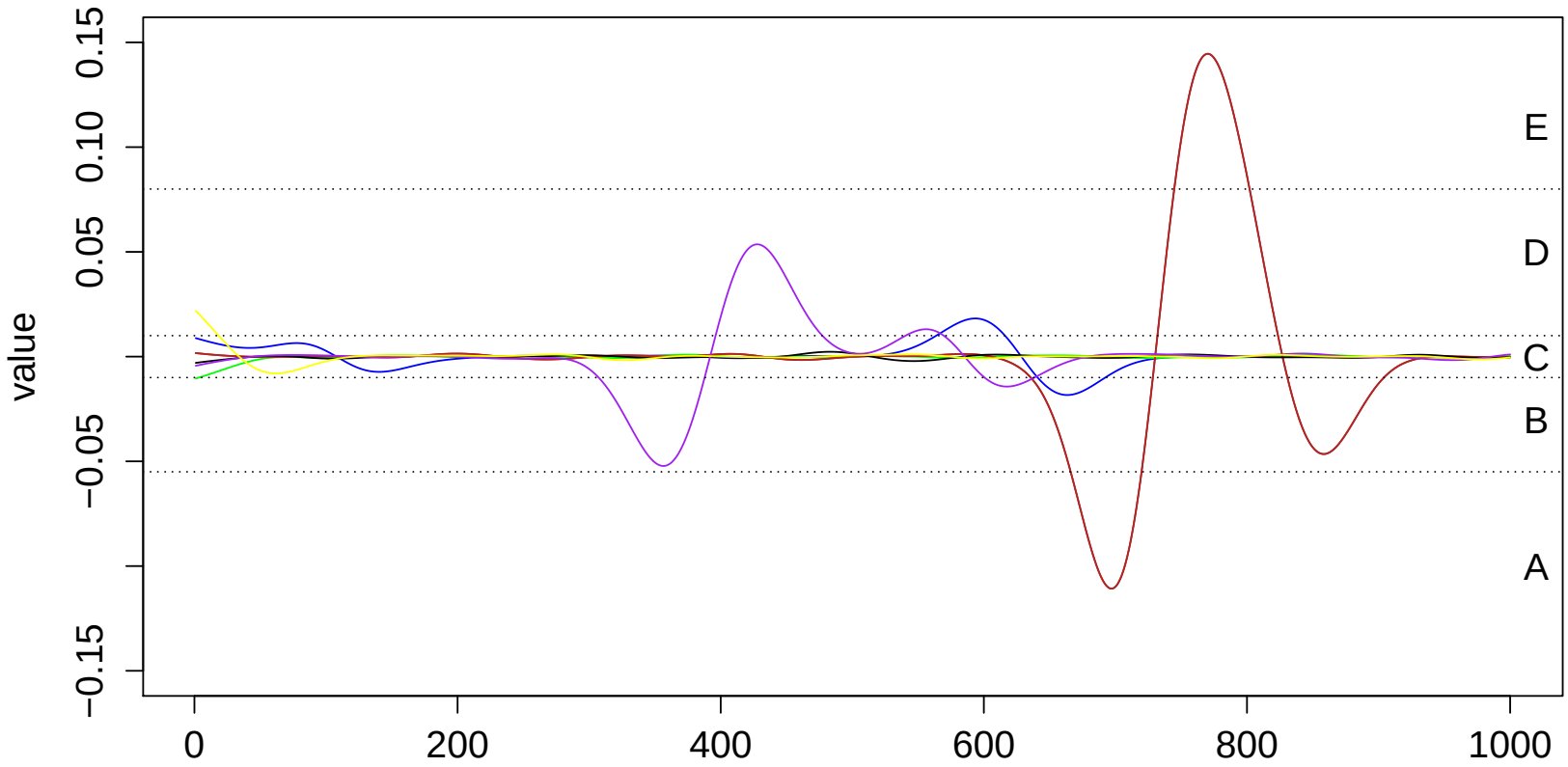


Рис.: Пример дискретизации временного ряда, разными цветами обозначен один и тот же временной ряд но для различных объектов. Пунктирные линии – квантили.

Сегментация объекта

Вход: $\{X_{jt}^i\}, j = \overline{1, J}, t = \overline{1, T_i}$ – объект;

w – ширина окна;

n –кличество итераций выбора локальных максимумов.

Выход: сегменты $\{S_m^i\}$, $m = 1, \ldots, M_i$.

Идея метода:

- проходим скользящим окном вдоль многомерного временного ряда;
- Для каждого момента времени считаем среднее расстояние до предыдущих векторов в окне;
- Моменты времени соответсвующие максимумам этого расстояния и будут границами сегментов.

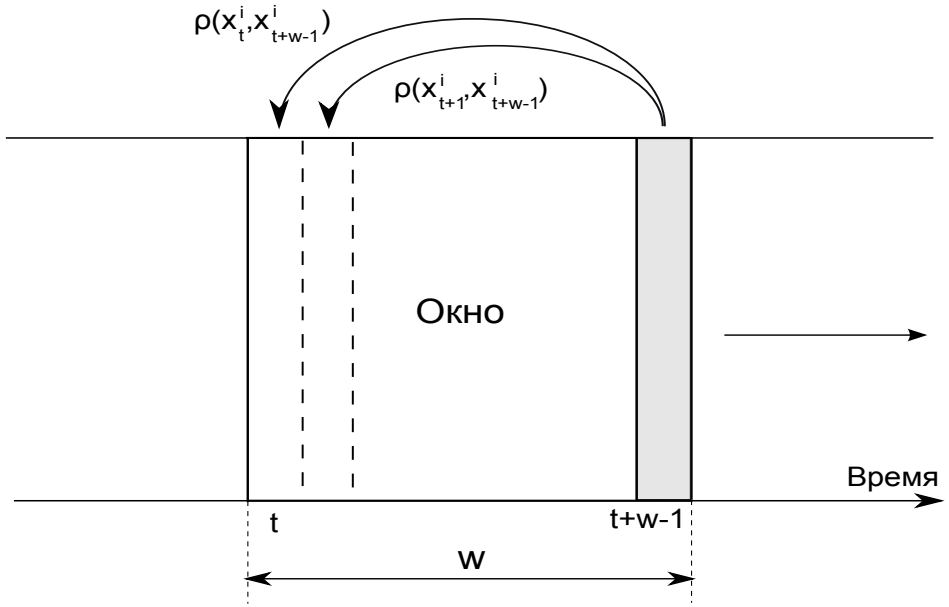


Рис.: Окно ширины w

Расстояние между символами изначально дискретных рядов ρ_d и дискретизованных ρ_c:

$$\rho_c(a_k, a_l) = \left\| \frac{p_k + p_{k-1}}{2} - \frac{p_l + p_{l-1}}{2} \right\|;$$

$$\rho_d(a_k, a_l) = |a_k \neq a_l|.$$

Расстояние между векторами значений многомерного дискретного временного ряда X^i в моменты времени t_1 и t_2:

$$\rho(X^i_{t_1}, X^i_{t_2}) = \sum_{j \in J_c} \rho_c(X^i_{jt_1}, X^i_{jt_2}) + 0.5 \sum_{j \in J_d} \rho_d(X^i_{jt_1}, X^i_{jt_2})$$

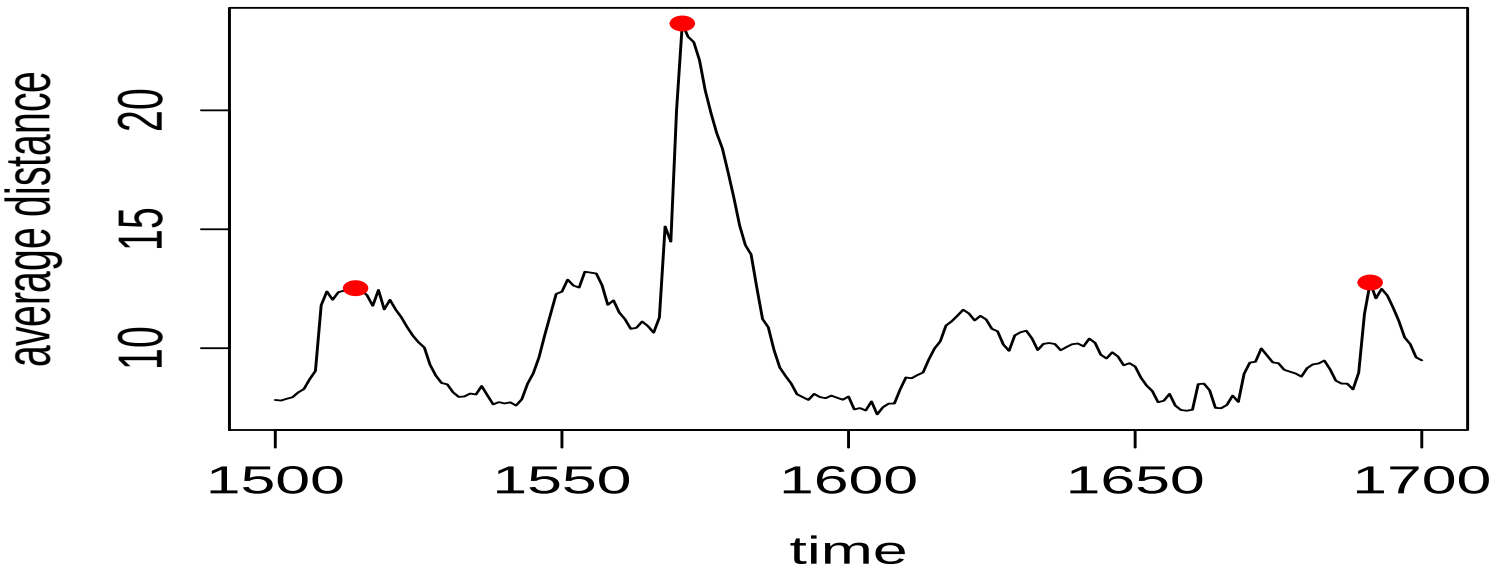


Рис.: Пример графика среднего расстояния, красные точки - главные локальные максимумы

Этап №5: Кластеризация сегментов

Задача: привести многомерные дискретные ряды к одномерным.

Вход: $\{S_m^i\}, m = \overline{1, M_i}, i = \overline{1, N}$ – множество всех сегментов.
Выход: Метки кластеров для каждого сегмента S_m^i
Метод:

- **Иерархическая кластеризация**
- **Метрика Хеллингера** между объектами
- **Расстояние Уорда** между кластерами

Метрика Хеллингера между сегментами:

$$\rho_H(S^1, S^2) = \frac{1}{J\sqrt{2}} \sum_{j=1}^{J} \sqrt{\sum_{k=1}^{n} \left(\sqrt{\nu_{kj}^1} - \sqrt{\nu_{kj}^2}\right)^2}$$

Здесь ν_{kj}^1 частота появления k-го символа алфавита в j-ом временном ряде сегмента S^1.

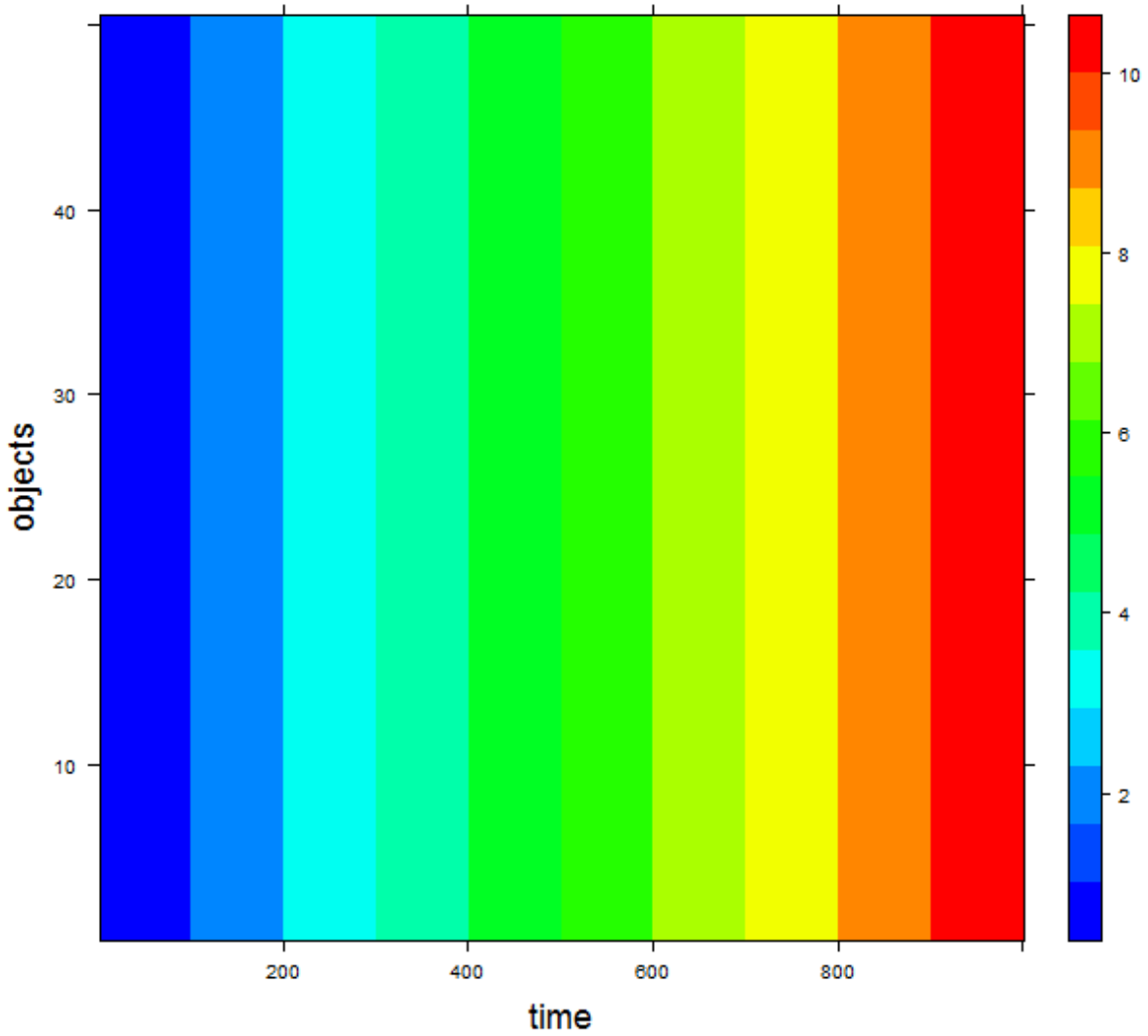


Рис.: “Идеальная” кластеризация сегментов

Вход: $X^i = \{S_m^i\}_{m=1}^{M_i}$, где $i = \overline{1, N}$ — объекты, представленные последовательностями сегментов.

Выход: метрика ρ

Функционал качества кластеризации:

$$Q(\rho) = \frac{2}{N^2} \sum_{i_1=1}^{N-1} \sum_{i_2=i_1+1}^{N} nLCS(X_\rho^{i_1}, X_\rho^{i_2}) \to \max,$$

$$nLCS\,(x, y) = \frac{|LCS(x, y)|}{\sqrt{l_x l_y}}.$$

$LCS\,(\cdot, \cdot)$ – наибольшаяя общая подпоследовательность;
X_ρ^i – объект, представленный **одномерным дискретным рядом** после кластеризации с метрикой ρ

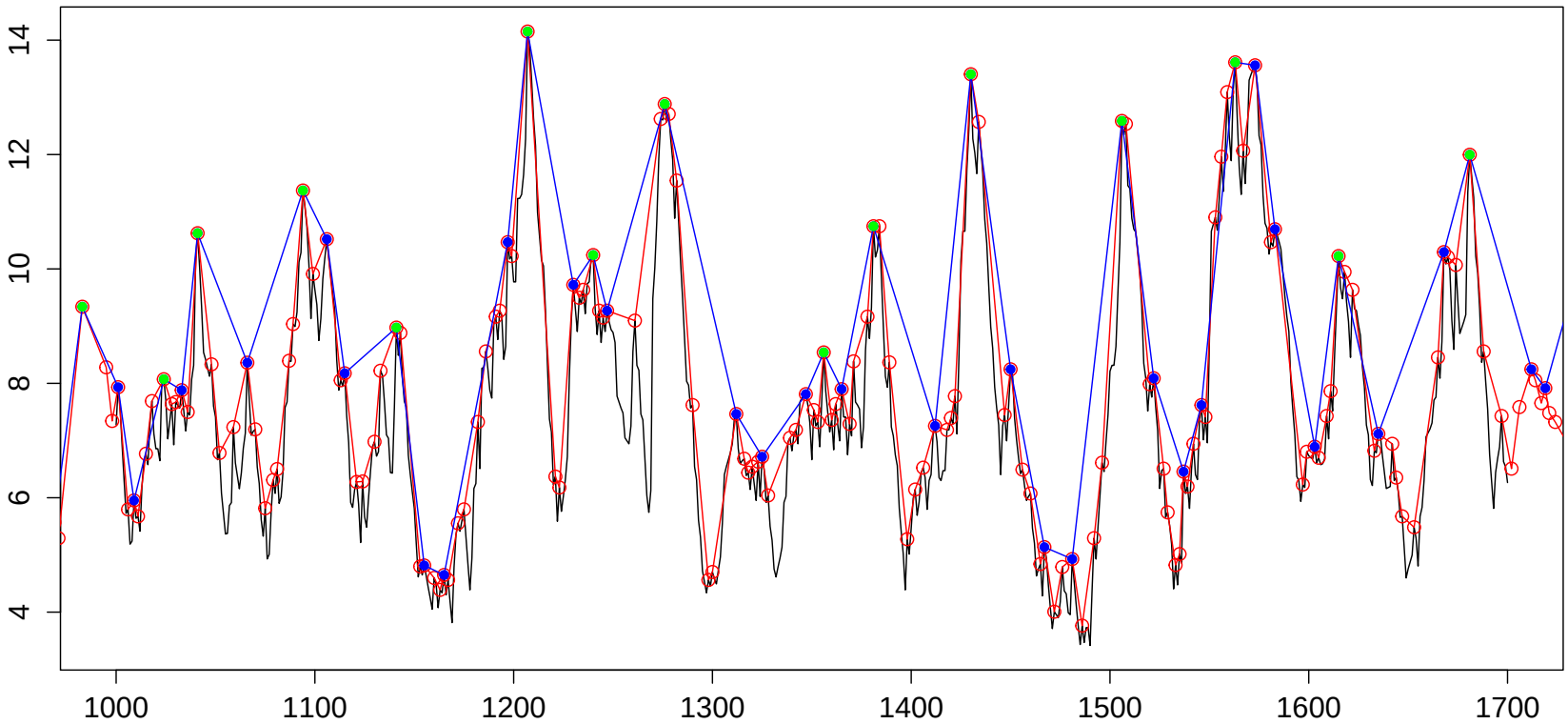
Данные:

- $N = 79$ круизные фазы полетов одного типа самолетов;
- $J = 304$ датчика, из них: $J_c = 195$ непрерывных; $J_d = 109$ дискретных;
- $T_i \in [3000, 6000]$ – продолжительность круизной фазы

Выбранные параметры:

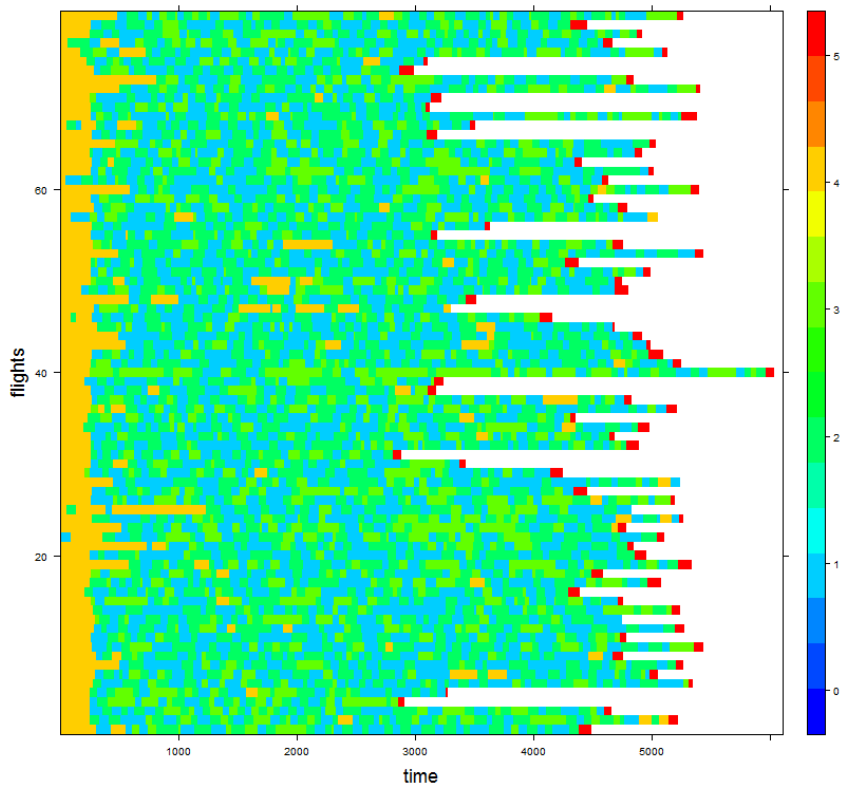
- пятибуквенный и одиннадцатибуквенный алфавиты;
- число итераций выбора локальных максимумов $n = 3$;
- число кластеров $C = \{5, 10, 20\}$;
- ширина окна при сегментации $w = 20$;

Рис.: График выделения локальных максимумов. Красные точки – первичные локальные максимумы, синие – вторичные, зеленые – итоговые.



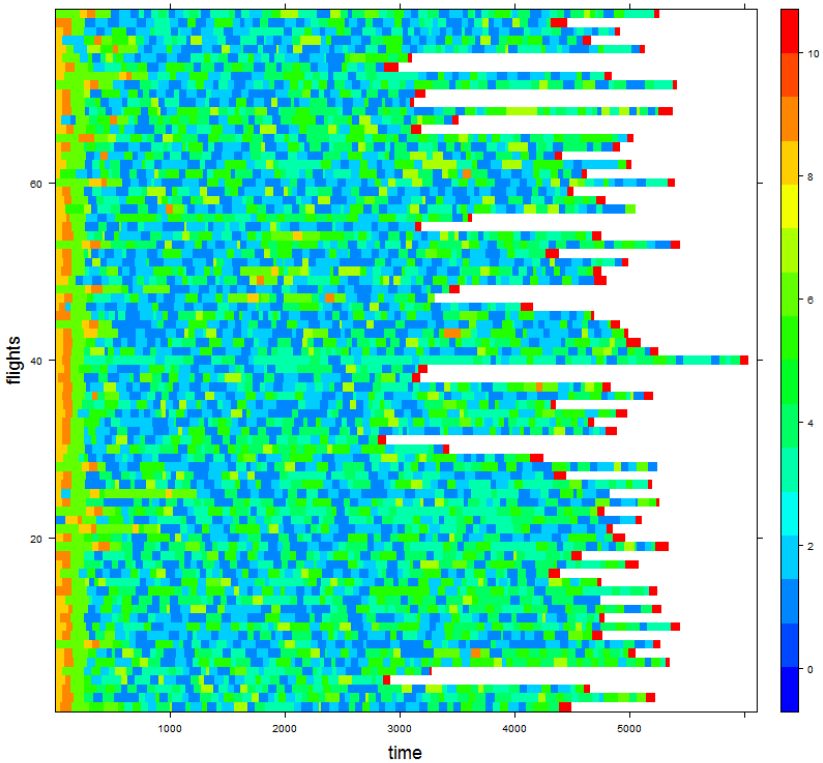
Вывод: После трёх итераций выделены практически все главные максимумы.

Рис.: Иллюстрация кластеризации сегментов для 5 кластеров. Цвет – номер кластера.

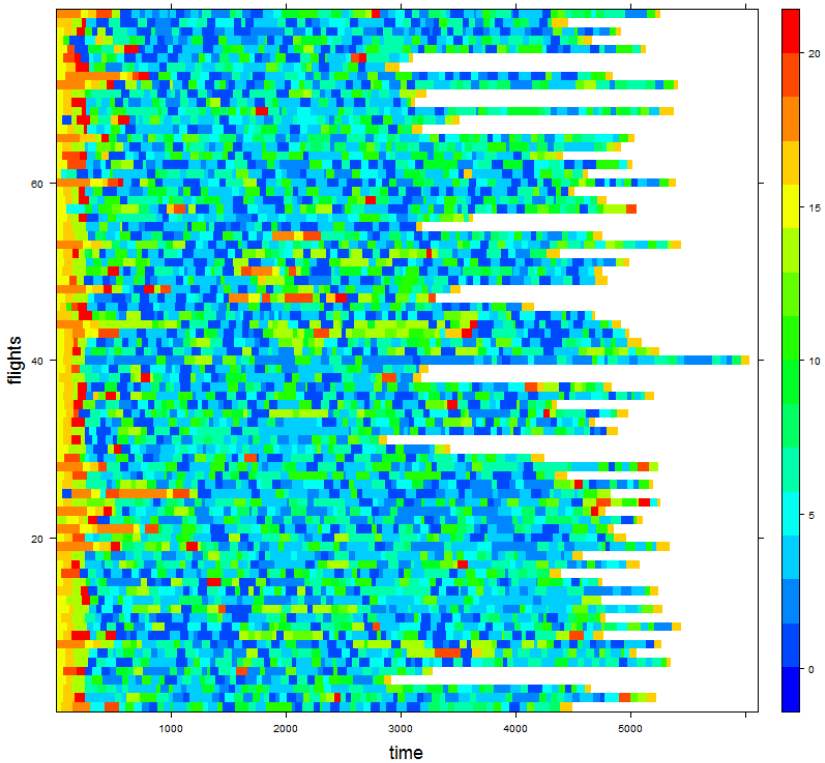


Вывод: Схожие по времени сегменты объединяются в один кластер.

Рис.: Иллюстрация кластеризации для 10 и 20 кластеров. Цвет – номер кластера.



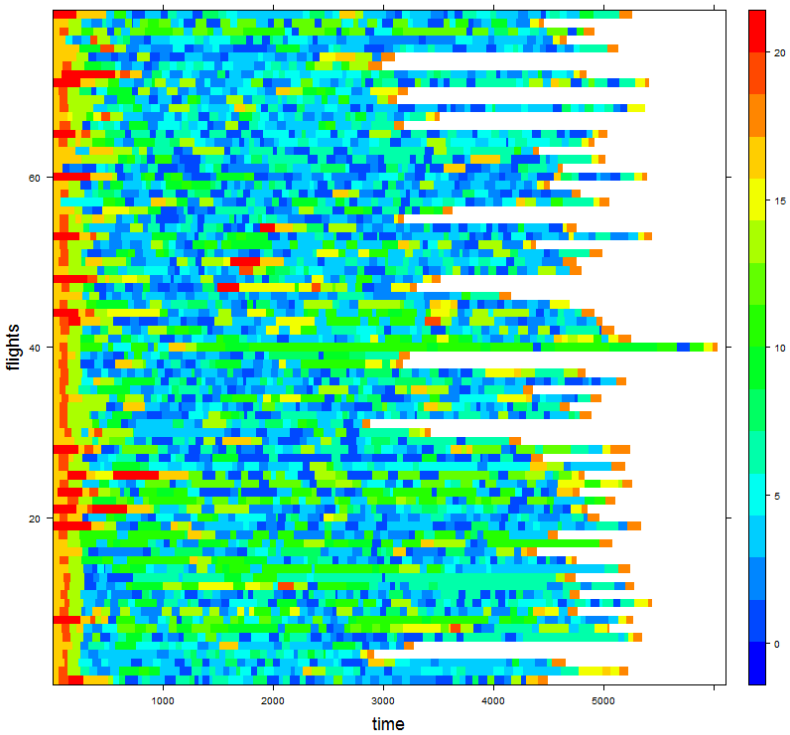
10 кластеров.



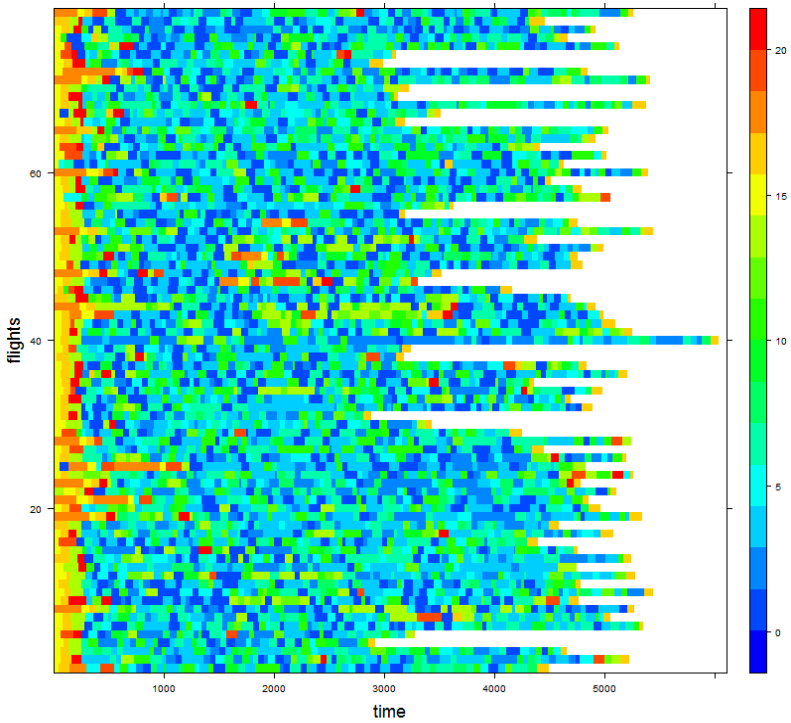
20 кластеров.

Вывод: при увеличении числа кластеров количество значительно отличающихся объектов увеличивается.

Рис.: Иллюстрация кластеризации для 20 кластеров при различном размере алфавита.



11-буквенный алфавит.



5-буквенный алфавит.

Вывод: увеличение размера алфавита не оказывает существенного влияния на итоговый результат.

Заключение

Преимущества алгоритма:

- аномалии локализуются внутри сегментов;
- уменьшается размерность исходной задачи по времени;
- возможность использовать весь комплекс алгоритмов для задачи поиска аномалий в одномерных дискретных рядах.