Поиск аномалий в полётных данных

К. О. Неклюдов

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

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

30 июня 2014 г.

Постановка задачи

Дано:

- X_{jt}^{i} – показания датчиков.
- i – индекс полёта.
- j – индекс датчика.
- t – момент времени.

Требуется:

- найти аномальные полёты;
- формализовать понятие аномальности;

при отсутствии информации об аномалиях от экспертов.

Постановка задачи

Предположения:

- Каждый полёт состоит из набора последовательных участков однородности – фаз.
- Каждая фаза может быть разбита на более мелкие участки однородности – сегменты.
- Аномалия – маловероятное событие:
  — внутри полёта;
  — на множестве полётов.

Обзор предыдущих работ

Во всех алгоритмах аномальность определяется расстоянием до ближайшего кластера.

- *G.Biswas, D.Mack, 2013* Иерархическая кластеризация. Метрика между полетами: Compression based dissimilarity. Данные: многомерные, размеченные.
- *S.Das,A.Srivastava, 2010* Одноклассовый SVM. Метрика между полетами: основана на LCS. Данные: многомерные, размеченные
- *V.Chandola, 2010* K ближайших соседей. Евклидова метрика. Данные: размеченные
- *Budalakoti, 2009* K-медоид кластеризация. Метрика между полетами: основана на LCS. Данные: одномерные дискретные неразмеченные. Использовалась экспертная оценка.

Этапы обработки полётных данных

Алгоритм состоит из последовательных этапов:

1. Выделение фаз полёта.
2. Проверка непрерывных датчиков на стационарность.
3. Дискретезация непрерывных датчиков.
4. Сегментация фаз.
5. Кластеризация сегментов.
6. Ранжирование полётов по аномальности.

Этап №1: Выделение фаз

Каждый полёт может быть разбит на участки однородности – фазы. Причём каждая фаза несёт **физический смысл**:

- стоянка
- буксировка от аэропорта
- руление до взлетной полосы
- руление до взлета
- взлет
- набор круизной высоты
- круиз
- снижение
- маневрирование
- приближение
- посадка
- руление до аэропорта

Описания фаз можно найти в документе:
`http://www.intlaviationstandards.org/Documents/PhaseofFlightDefinitions.pdf`

Этап №1: Выделение фаз

Для разбиения на фазы были выбраны датчики по физическому смыслу:

- Скорость самолёта
- Высота
- Угол тангажа
- Угол крена
- Угол атаки
- Расход топлива

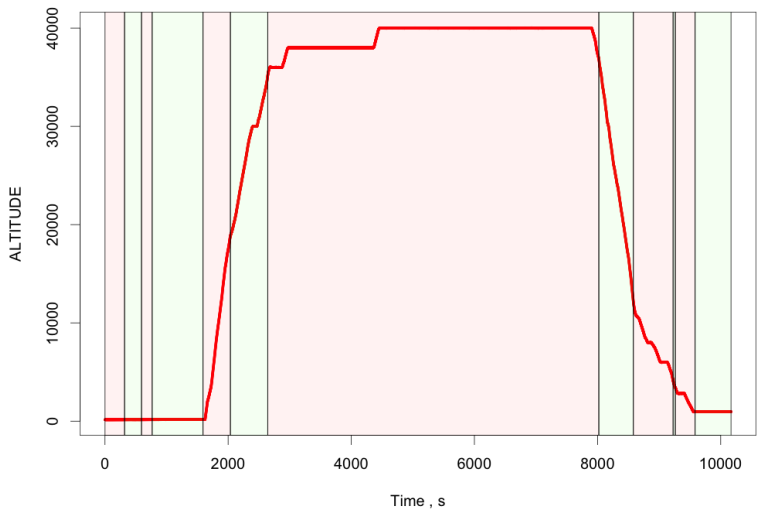
Этап №1: Выделение фаз

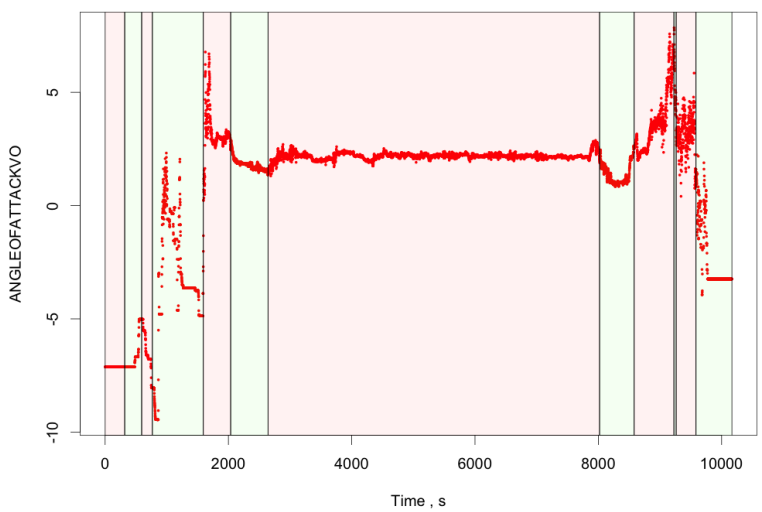
Вход: полёт $\left\{X_{jt}^{i}\right\}_{t=1}^{T_i}$, для заданного множества датчиков $j \in J_p$.
Выход: $\left\{t_1^i, t_2^i, \ldots, t_{12}^i\right\}$ — моменты времени, соответствующие началам фаз.
Метод:

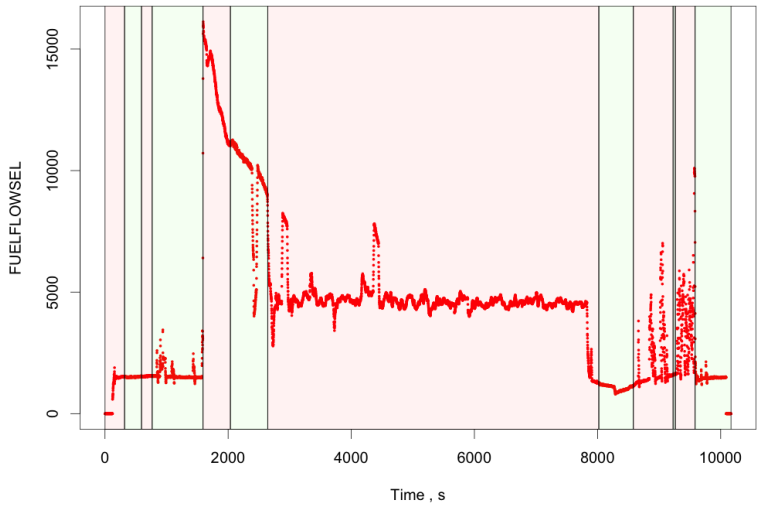
- **Иерархическая кластеризация** с ограничениями
- **Евклидова метрика** между объектами
- **Расстояние Уорда** между кластерами

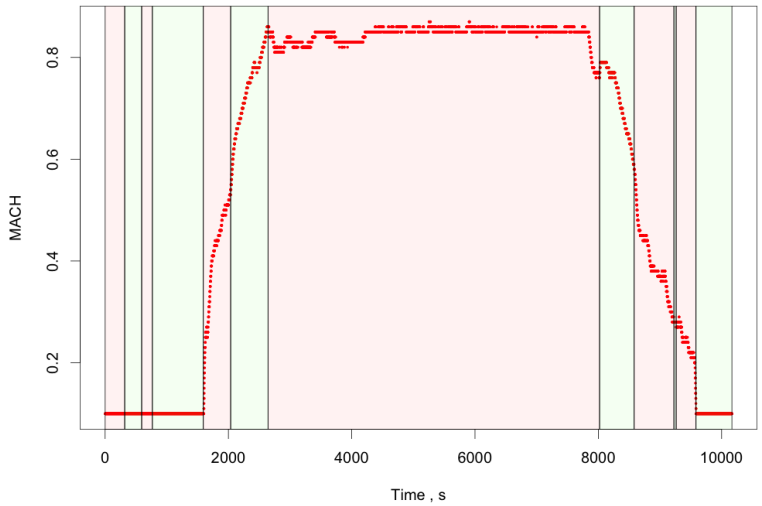
Ограничение: на каждой итерации могут объединяться только соседние по времени кластеры.

Показания датчиков, наложенные на разбиение по фазам:



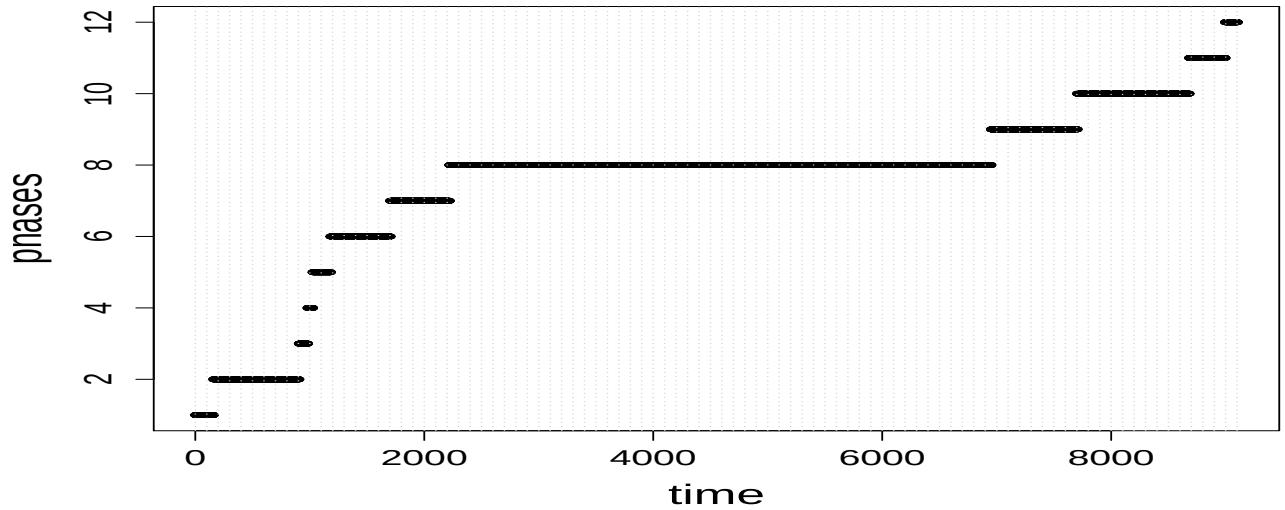


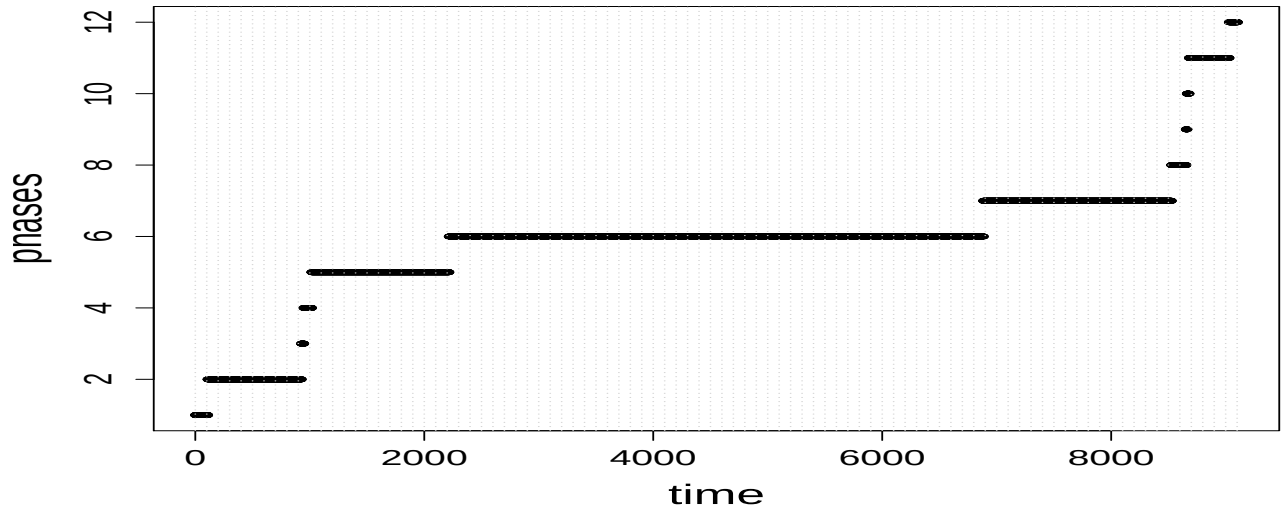




Вывод: разбиение на фазы соответствует однородным участкам

Сравнение выделенных фаз с заданной разметкой:





Вывод: разбиение на фазы согласуется с разметкой.

Этап №2: Стационарность

Вход: X_j – показания j-го датчика.

Выход: стационарные временные ряды разностей ΔX_j.

Метод:

- Проверка на стационарность: **KPSS тест**.
- Переход к попарным разностям в случае необходимости.

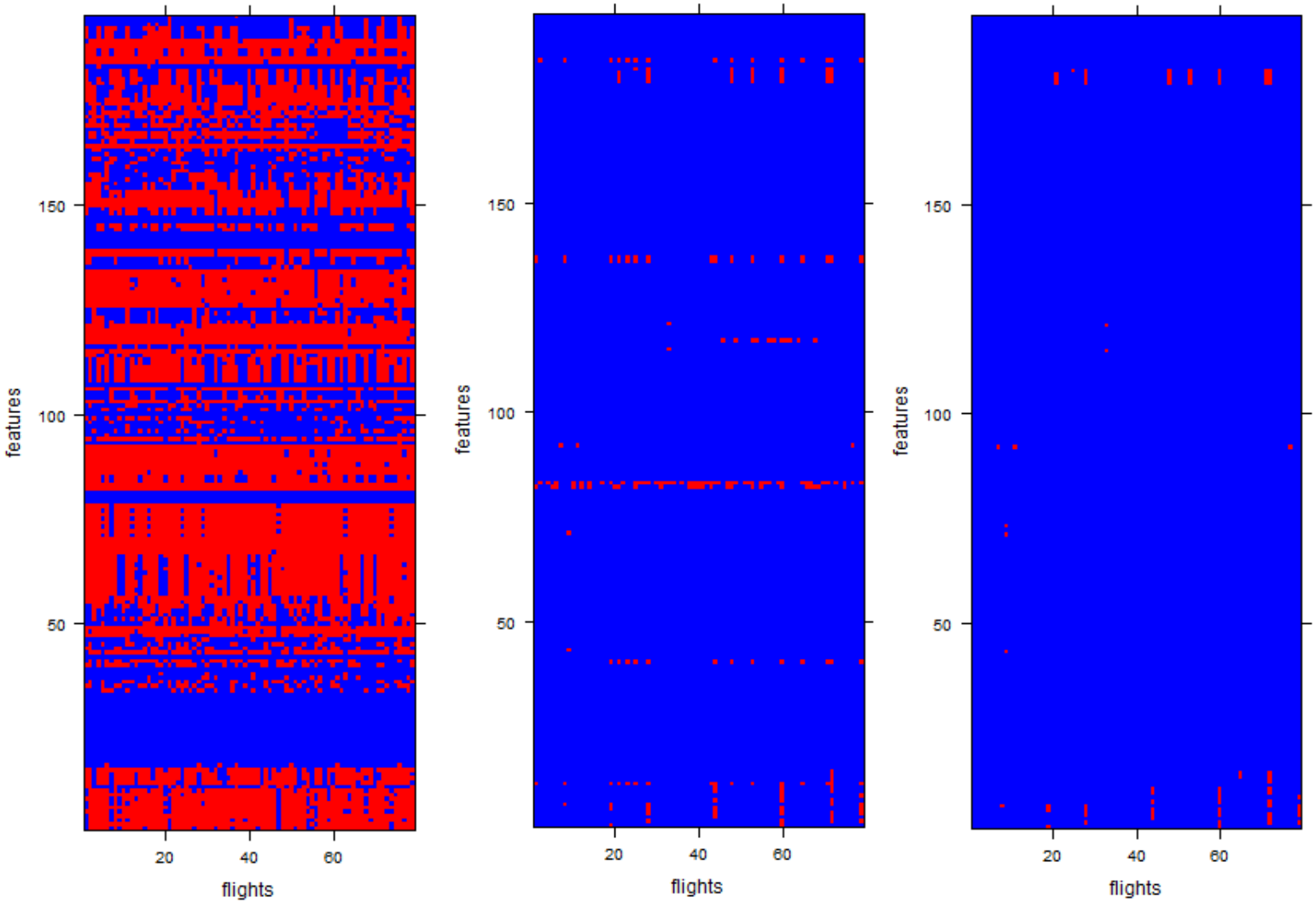
KPSS тест:

H_0 : временной ряд стационарен

H_1 : временной ряд имеет линейный тренд

Результаты KPSS теста.

Красные клетки соответствуют нестационарным рядам.



Вывод: получившиеся ряды стационарны.

Дискретизация

Вход: $\{X^i_{jt}\}$, где $j \in J_c$,
$t \in [t^i_k; t^i_{k+1}]$, где $[t^i_k; t^i_{k+1}]$ – k-ая фаза i-го полёта.
Выход: дискретизованные значения $\{X_{jt}\}$.

Сегментация

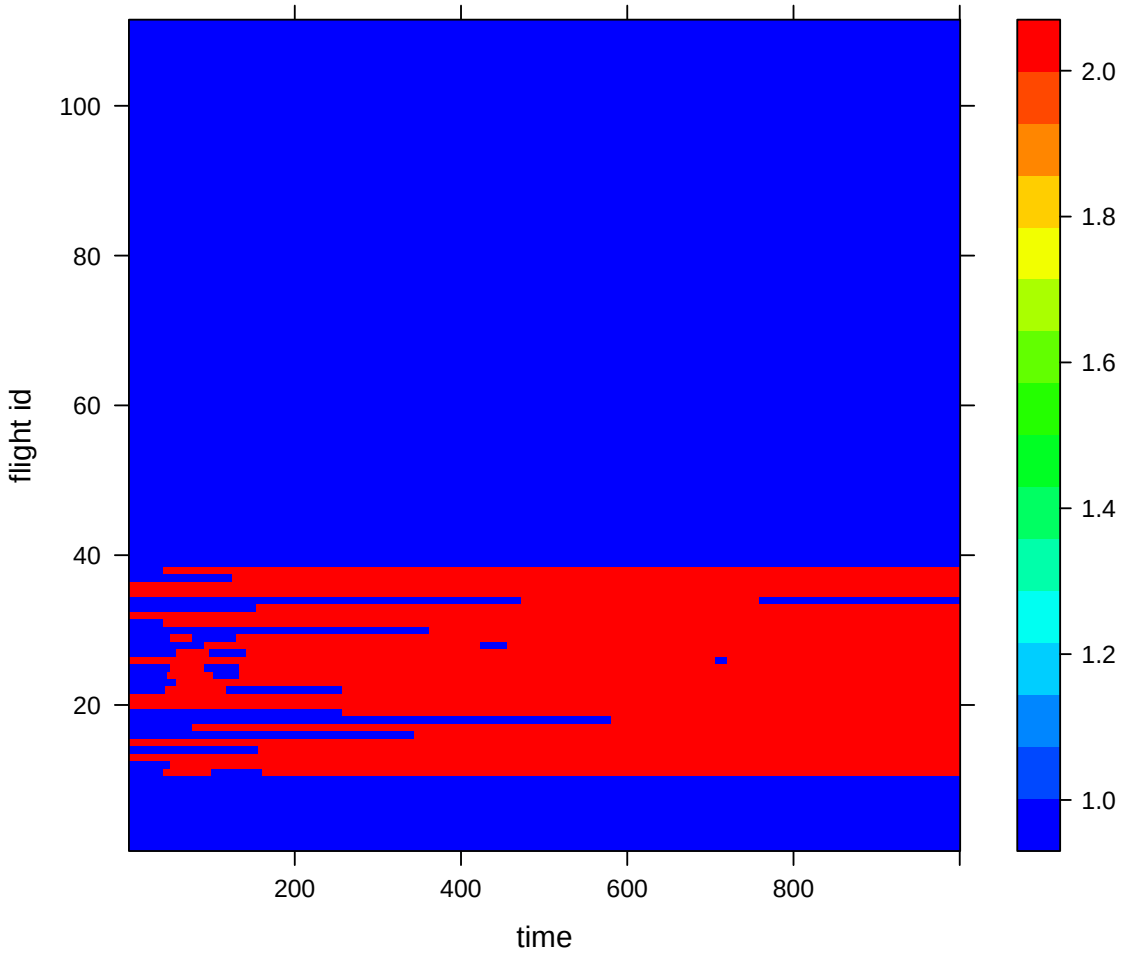
Вход: k-ая фаза полёта X^i.
Выход: $S_i = \{s_l\}_{l=1}^{L_i}$ – набор сегментов в фазе полёта X^i

Кластеризация сегментов

Вход: $\{s_l\}_{l=1}^{L}$ – множество всех сегментов. $L = \sum_{i=1}^{N} L_i$
Выход: метки кластеров для каждого сегмента s_l

Результаты кластеризации

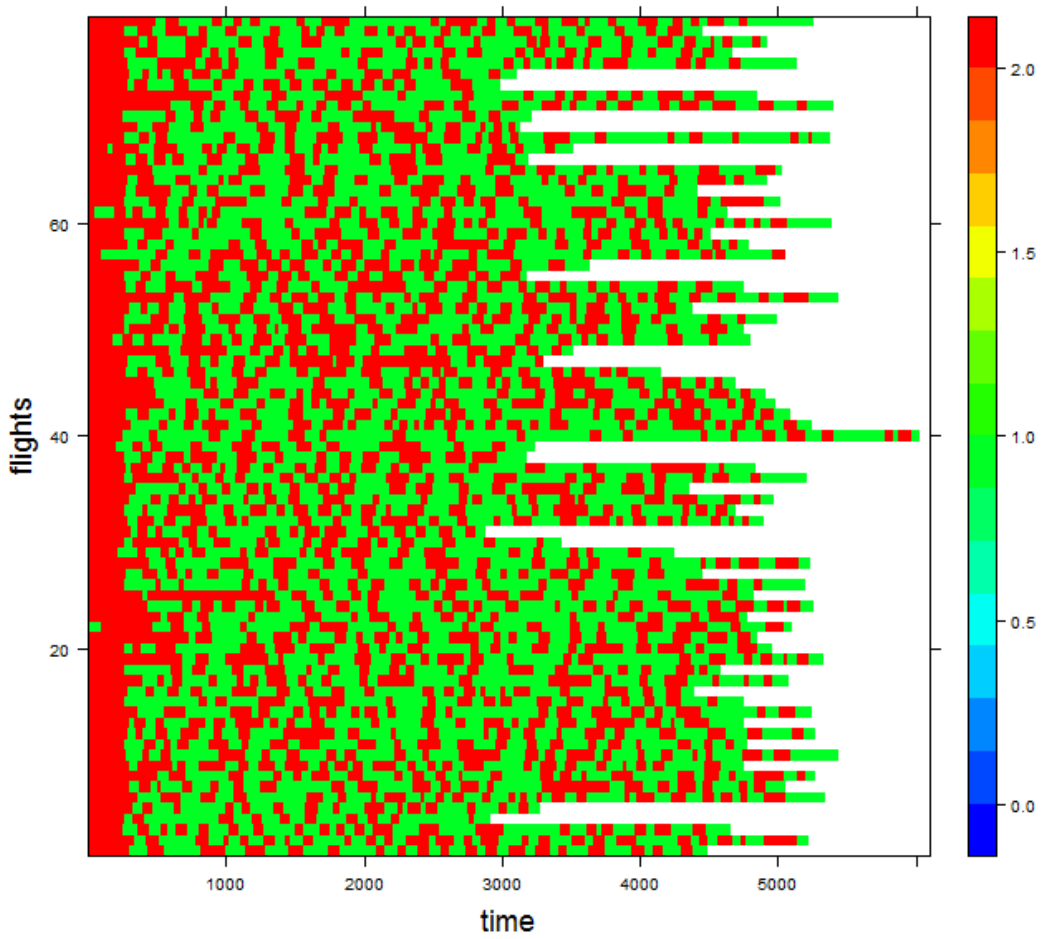
Кластеризация сегментов для двух кластеров:



Вывод: алгоритм различил типы самолётов.

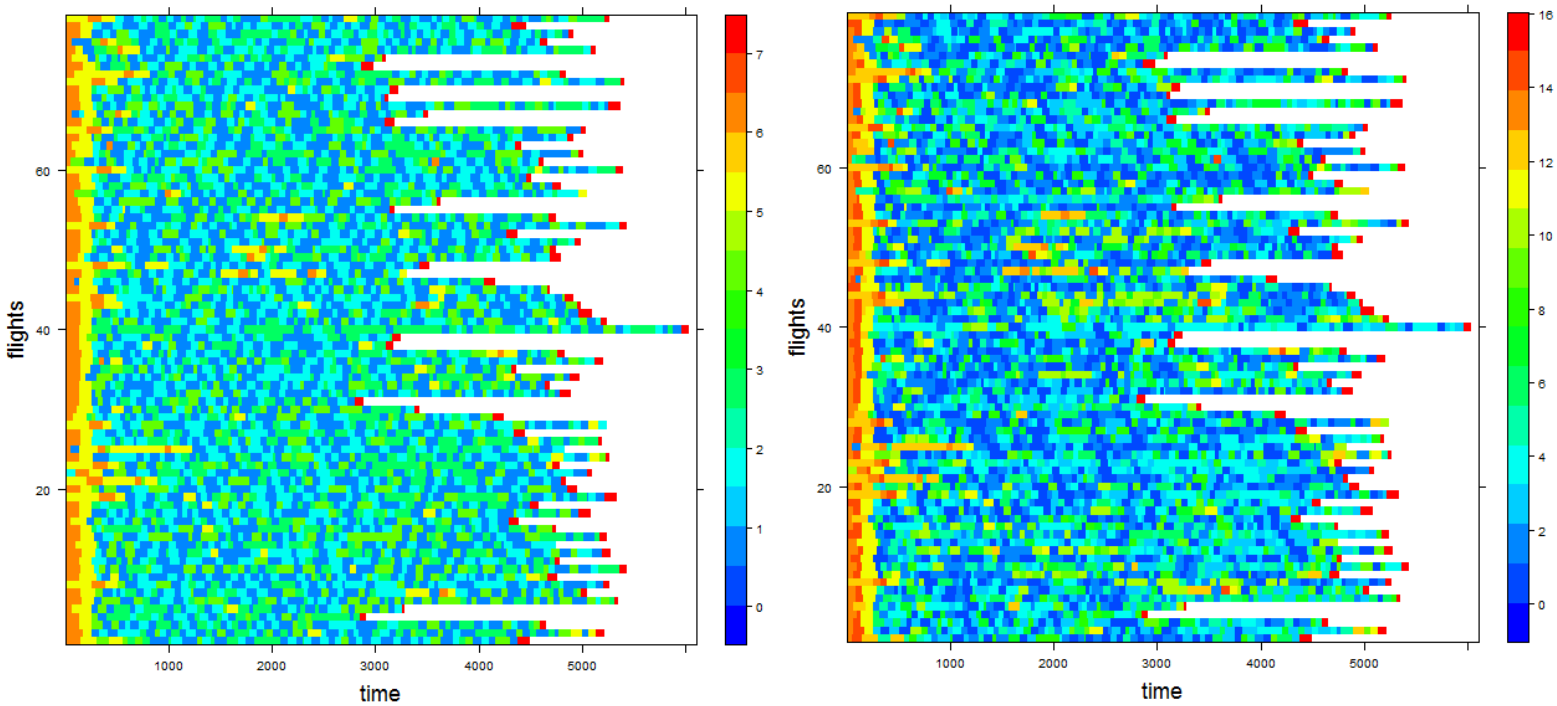
Результаты кластеризации

Для одного типа самолёта аналогичного разделения не наблюдается:



Результаты кластеризации

Графики для 7-ти и 15-ти кластеров.



Вывод: Круизные фазы большинства полетов схожи.

В то же время, несколько полетов заметно отличаются.

Этап №6: Ранжирование по аномальности

Вход: $\{\tilde{X}_i\}_{i=1}^{N}$ - набор фаз полётов, представленных в виде одномерных дискретных рядов.
Выход: список полётов, отранжированных по аномальности.
Метод:

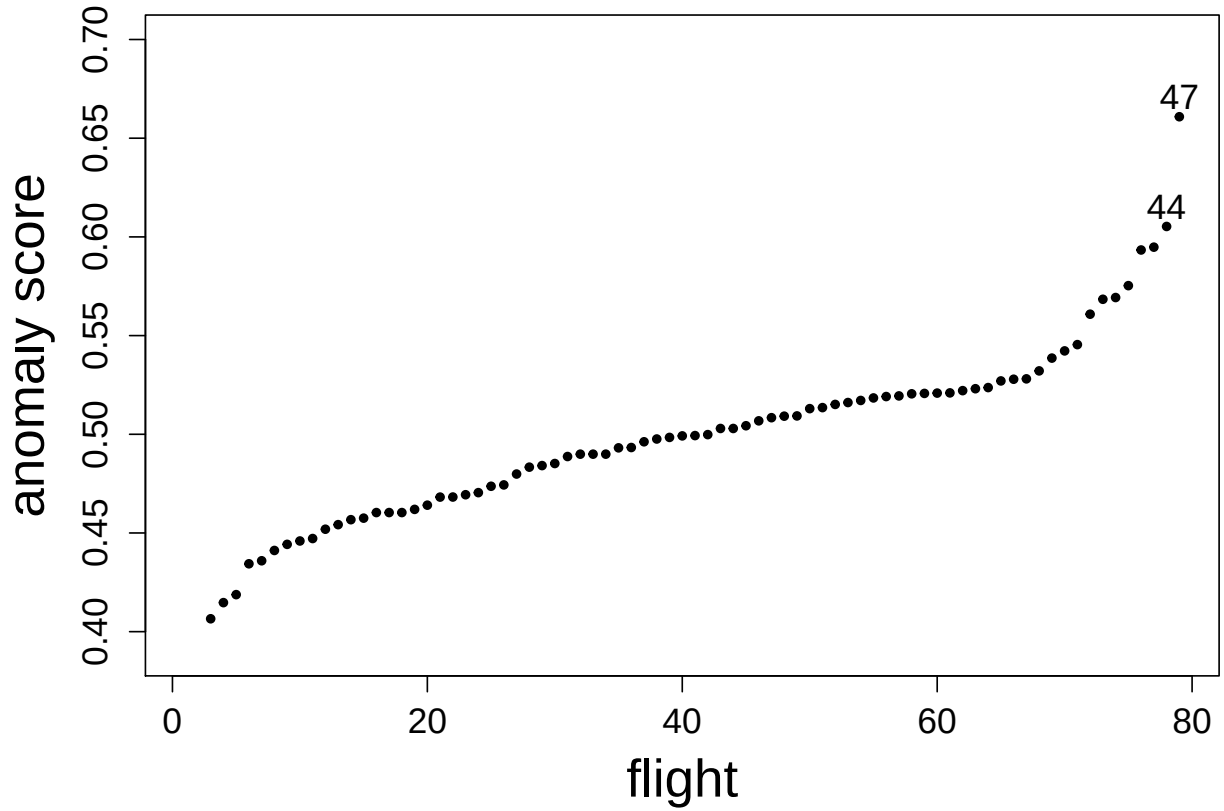
- Расстояние между полётами – **nLCS метрика**.
- Аномальность полёта – расстояние полёта до медианы кластера.

nLCS метрика:

$$nLCS(X, Y) = 1 - \frac{LCS(X, Y)}{\max(l_X, l_Y)}$$

Здесь $LCS(X, Y)$ – длина наибольшей общей подпоследовательности.

Полёты, отранжированные по аномальности.



Вывод: получены 5 полётов, которые можно интерпретировать как аномальные, но дальнейший анализ должны проводить эксперты. Наибольшее значение аномальности принимает полет под номером 47.

Устойчивость ранжирования

число кластеров =	2	3	4	5	6	7	8	9
топ 1	47	4	40	40	40	47	23	30
топ 2	19	40	4	4	4	23	47	23
топ 3	40	26	23	23	23	56	30	47
топ 4	68	5	26	26	5	65	56	56
топ 5	48	39	62	62	26	5	62	62

число кластеров =	10	11	12	13	14	15	16	17
топ 1	30	30	47	47	47	47	47	47
топ 2	23	23	62	23	44	44	44	44
топ 3	65	62	30	65	65	30	30	65
топ 4	62	65	23	13	30	65	65	43
топ 5	47	47	43	40	43	43	43	30

Заключение

Преимущества алгоритма:

1. Основной результат выдаётся в виде ранжированного списка, как в поисковых системах
2. Аномалии локализуются внутри фаз и сегментов
3. Не требуется экспертная разметка
4. Метод можно использовать как инструмент разведочного анализа данных