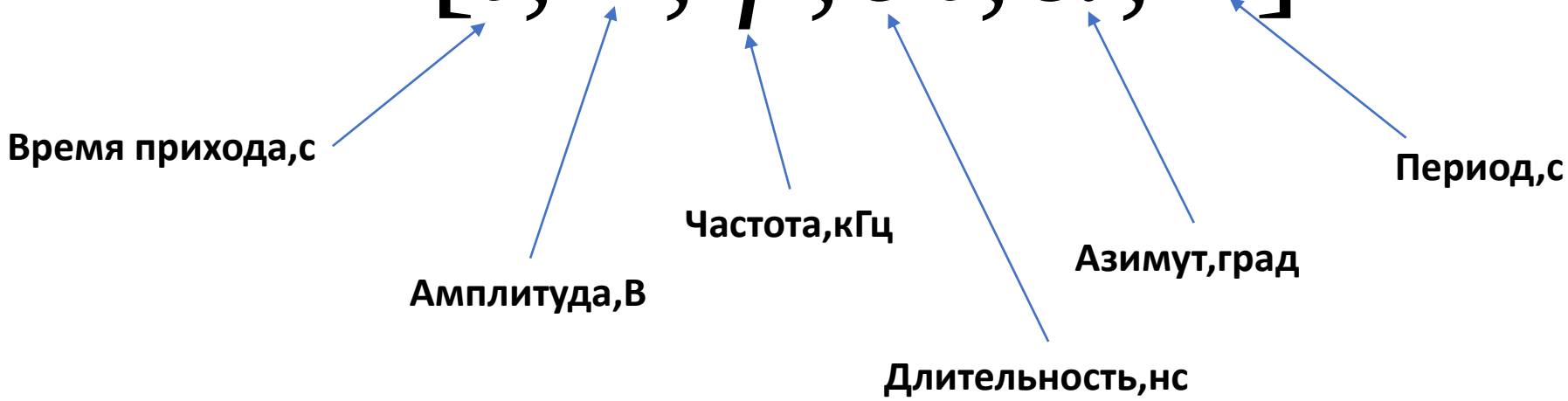
Постановка задачи

Имеется выборка принятых импульсов (отметок), характеризующаяся следующими параметрами:

$$[t, A, f, \delta t, \alpha, T]$$



Требуется определить кластеры сложных сигналов (паттернов), состоящих из n-ого числа отметок, где n > 1.

Методы решения задачи кластеризации

Метод	Основа алгоритма	Входные данные	Требует ли заранее знать количество кластеров?	Кластерные идентифицированные формы	Позволяет ли выделять выбросы?
Иерархическая кластеризация	Расстояние между объектами	Попарные расстояния между наблюдениями	Нет	Кластеры произвольной формы	Нет
k-средних	Расстояние между объектами и центроидами	Фактические наблюдения	Да	Сфероидальные кластеры с равной диагональной ковариацией	Нет
DBSCAN	**Плотность областей в данных**	**Фактические наблюдения или попарн.**	**Нет**	**Кластеры произвольной формы**	**Да**
Смешанные гауссовские модели	Смесь распределений Гаусса	Фактические наблюдения	Да	Сфероидальные кластеры с различными структурами ковариации	Да
Спектральная кластеризация	Граф связи между точками	Фактические наблюдения Или матрица подобия	Да, но можно оценить количество кластеров	Кластеры произвольной формы	Нет

Имитационная модель

Имитационная модель (ИМ) представляет собой набор последовательных произвольных импульсов – помеховую для алгоритма кластеризации среду, со следующим вектором состояния:

$$\vec{x} = [t,\ f\ , \delta t, T]^T$$

где частота и длительность импульса выбираются случайным образом из заданных списков значений:

- Для частоты [1,09; 1,5; 5,48; 9.8; 16] ГГц
- Для длительности [50; 100; 500; 20000; 65000] нс

Условимся, что размер выборки имитационной модели N будет равен 10000 импульсов.

Имитационная модель

Модель наблюдений описывается следующим образом:

$$\mathbf{Y} = \mathbf{X} + \boldsymbol{\varepsilon}$$

где $\boldsymbol{\varepsilon}$ - матрица шумов наблюдений размерностью (N*4).
$\mathbf{X}$ - матрица состояния размерностью (N*4). $\mathbf{Y}$ - матрица наблюдений размерностью (N*4).

PS. Где 1 и 4 компоненты характеризуются *N*(0,50*1e-9 с), 2 - *N*(0,1e3 Гц), а 3 компонента - *N*(0,10*1e-9 с).

Добавим в выборку импульсов ИМ два паттерна случайным образом

- Размер 1-ого паттерна возьмём равным **трем** импульсам;
- Размер 2-ого паттерна равным **семи** импульсам.

1 паттерн

	1	2	3	4
1	1.4000e-05	3.4000e-05	4.6000e-05	
2	9.8000e+09	9.7608e+09	9.8000e+09	
3	5.0000e-08	5.0000e-08	5.0000e-08	
4	1.4000e-05	2.0000e-05	1.2000e-05	

Отн. Время прихода, с

Частота, Гц

2 паттерн

	1	2	3	4	5	6	7	8
1	1.2000e-05	2.4000e-05	2.8000e-05	4.6000e-05	5.4000e-05	7.2000e-05	7.4000e-05	
2	9.8000e+09	9.8000e+09	9.8000e+09	9.7608e+09	9.8294e+09	9.8392e+09	9.8000e+09	
3	6.5000e-05	6.5000e-05	6.5000e-05	6.5000e-05	6.5000e-05	6.5000e-05	6.5000e-05	
4	1.2000e-05	1.2000e-05	4.0000e-06	1.8000e-05	8.0000e-06	1.8000e-05	2.0000e-06	

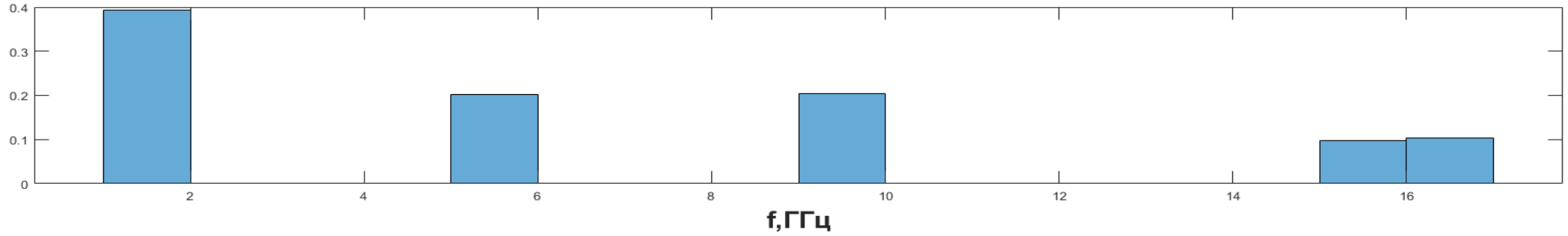
Длительность, с

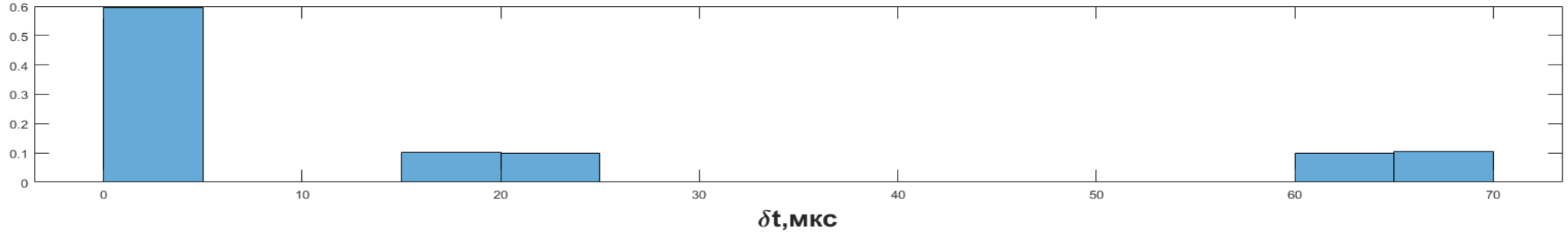
Период, с

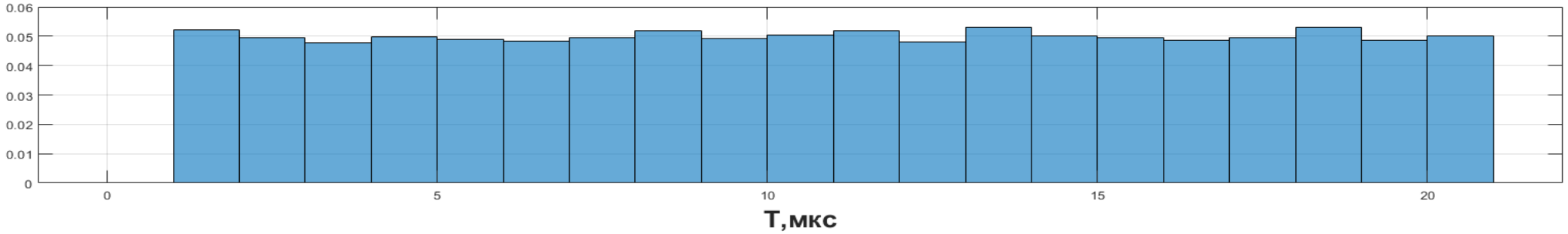
Гипотеза

Предполагается, что алгоритм кластеризации сможет выделить добавленные паттерны в отдельные кластеры, тем самым сформировав 2 кластера со схожими паттернами. Остальные импульсы он будет считать помехами и шумами (выбросами).

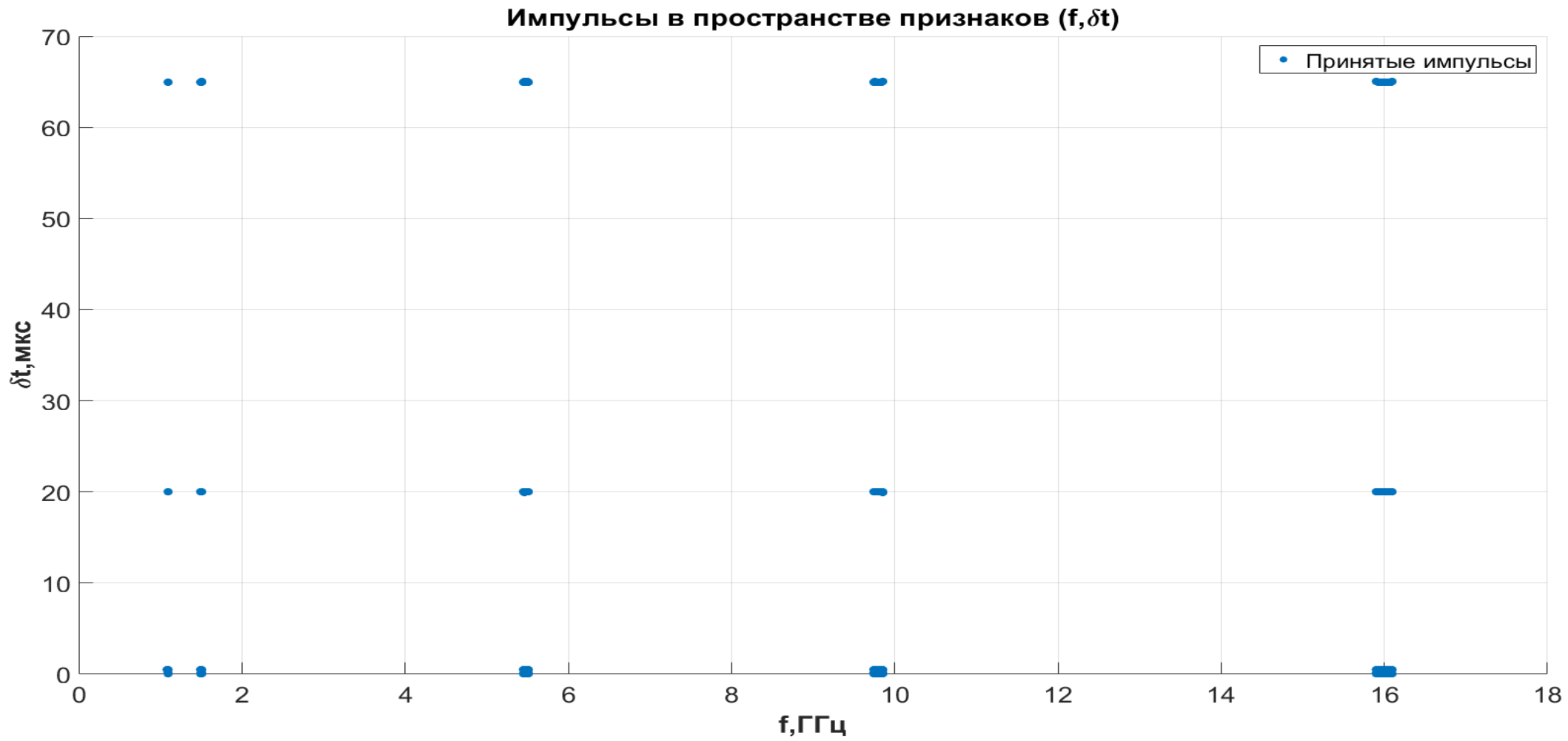
Вероятностные распределения параметров матрицы наблюдений (без времен прихода)



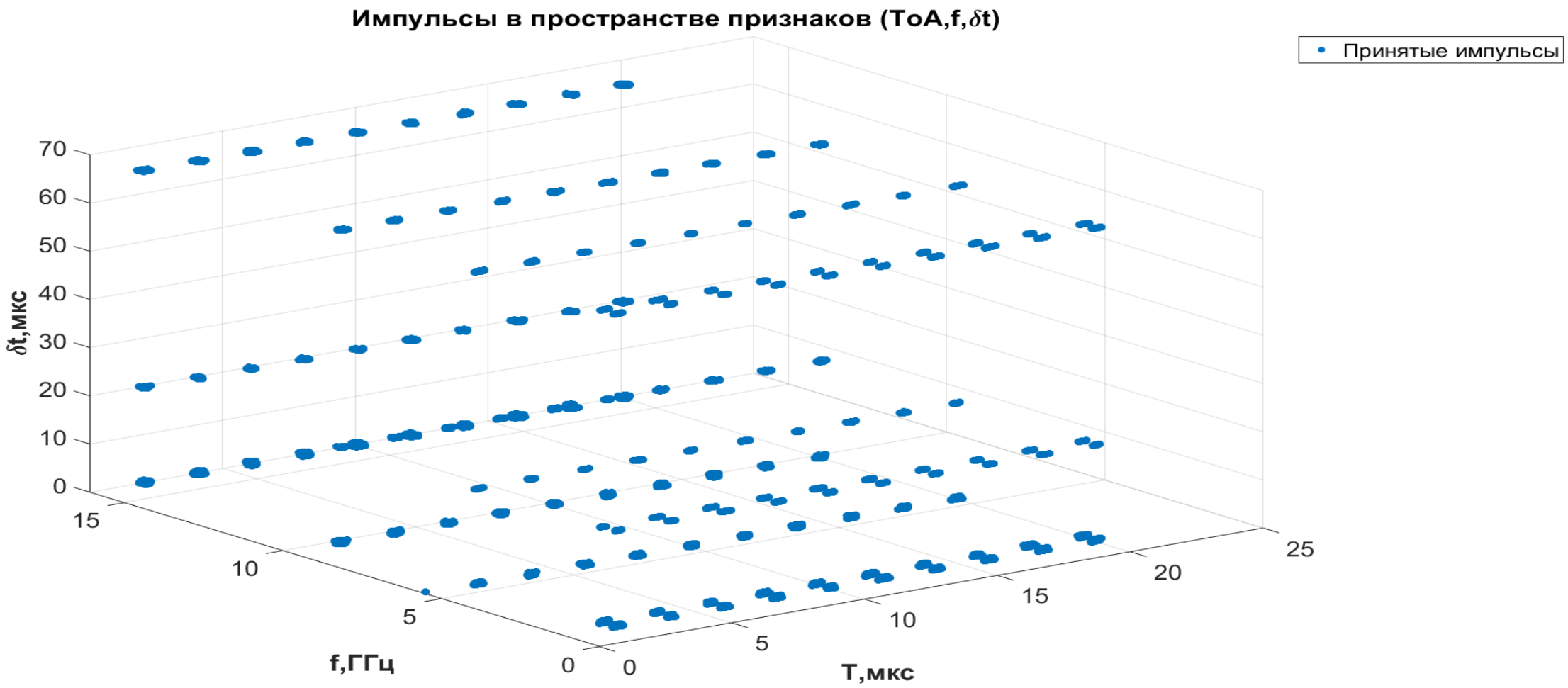




Двумерное представление импульсов в ИМ



Трехмерное представление импульсов в ИМ с осью значений периодов



Подготовка данных под кластеризацию

1) Первым делом нужно провести z-стандартизацию параметров матрицы наблюдений;

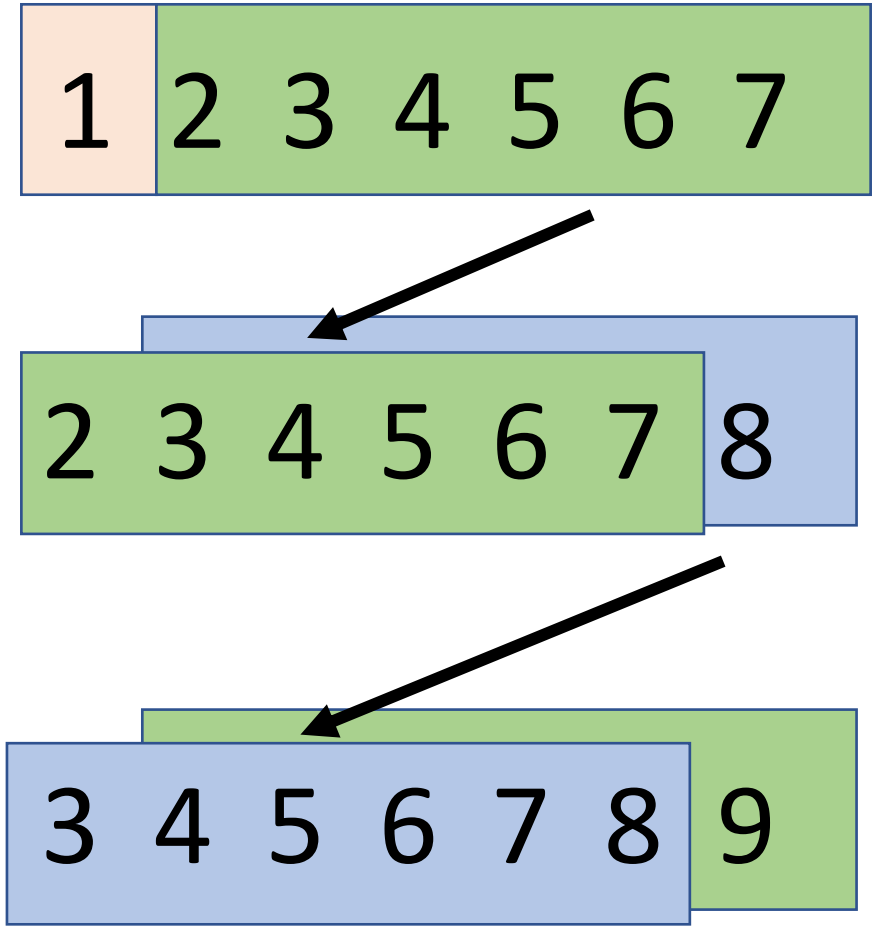
1.1) Параметр времени прихода использовать в кластеризации не будем, так он не является для нас информативным. Поэтому для кластеризации будем использовать следующие 3 параметра $[\ f\ ,\ \delta t\ ,\ T\ ]$

2) Чтобы алгоритм кластеризации выделял не просто отдельные импульсы в кластеры, а искал паттерны из импульсов требуется матрицу Y видоизменить, расширив ее до размерности (N*(3*размер_паттерна-1)). Назовем эту операцию «окном смещения»

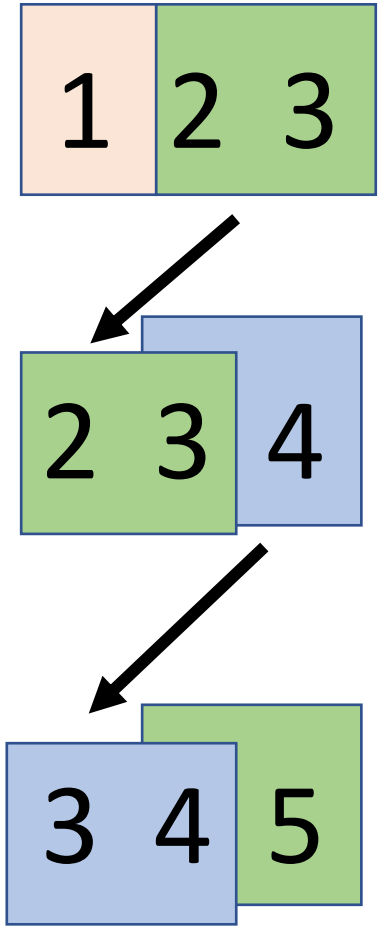
2.1) Операцию «окно смещения» необходимо проделать для двух случаев, так мы ищем два паттерна в выборке.

Окно смещения

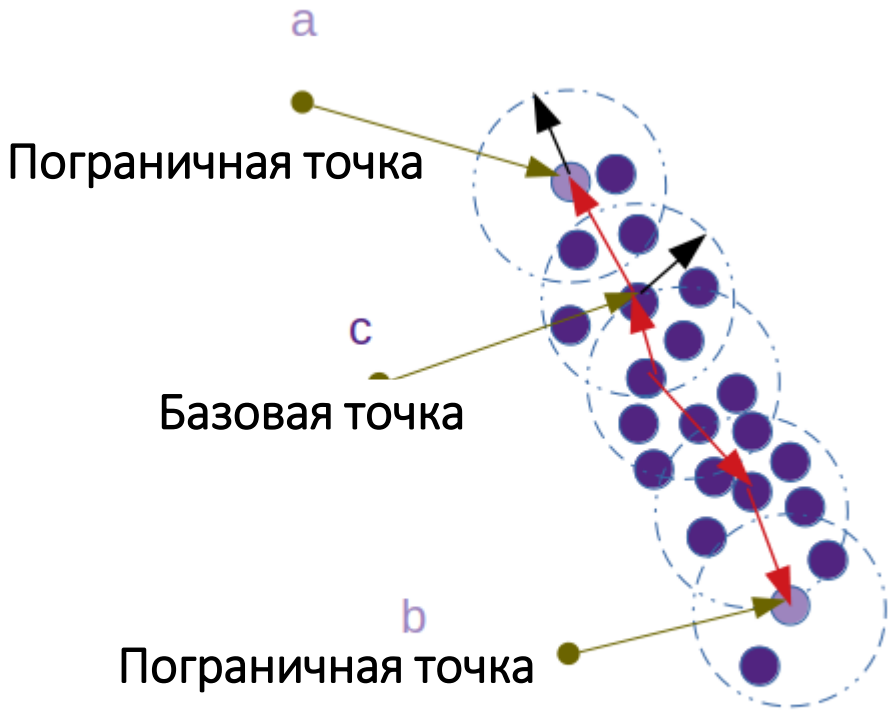
Для паттерна из 7 импульсов

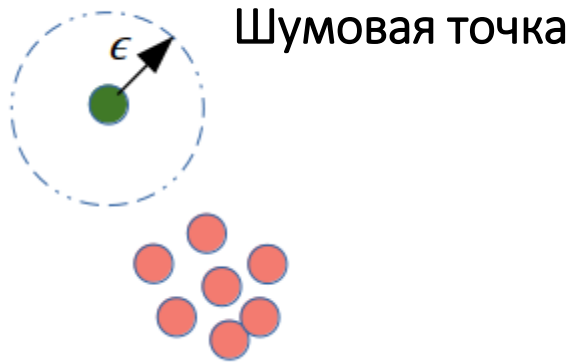


Для паттерна из 3 импульсов



Основная концепция алгоритма DBSCAN состоит в том, чтобы найти области высокой плотности, которые отделены друг от друга областями низкой плотности





Алгоритм DBSCAN идентифицирует три вида точек:

- Базовая точка — точка в кластере, которая имеет, по крайней мере, NN соседей в Epsilon окрестности.
- Пограничная точка — точка в кластере, которая имеет меньше, чем NN соседей в Epsilon окрестности.
- Шумовая точка — выброс, который не принадлежит никакому кластеру.

Подбор параметров алгоритма DBSCAN

Для работы алгоритма DBSCAN требуется задать следующие параметры:

1. Число соседей (NN)

2. Радиус поиска соседей (Epsilon)

NN на модели выбирается из учета минимального числа случайного повторения одного из двух паттернов в выборке (мы заранее знаем индексы паттернов в выборке и число их повторений в выборке).

Epsilon выбирается исходя из отсортированной по возрастанию матрицы попарных расстояний с Евклидовой метрикой.

Выбор NN

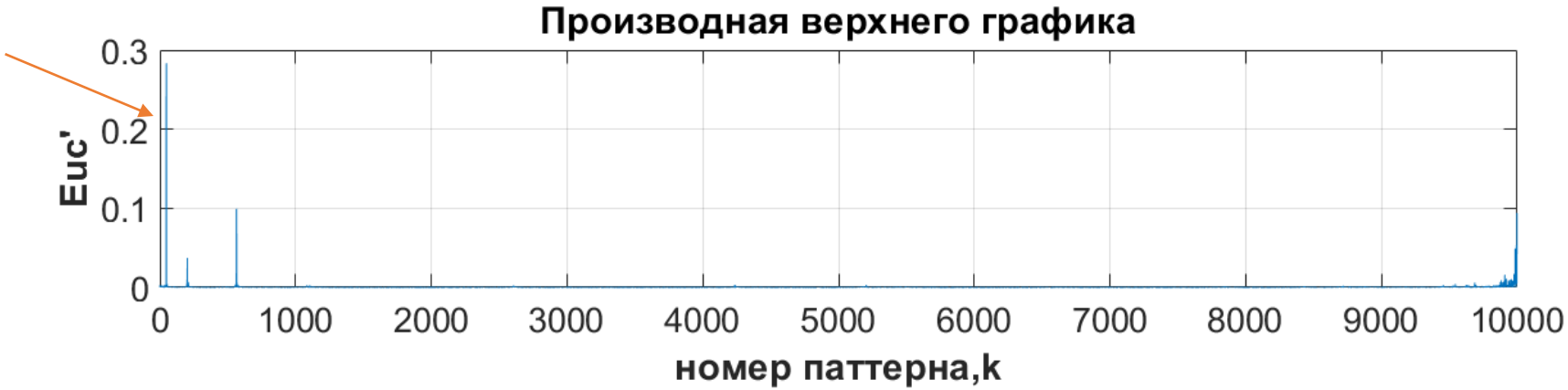
Предположим реализацию, где 1-ый паттерн длинной 3 импульса повторился 12 раз, а 2-ой паттерн длинной 7 импульсов 7 раз.

Тогда NN берем равным 7.

Выбор Epsilon

Epsilon выбирает как значение соответствующее максимальной производной в начале реализации

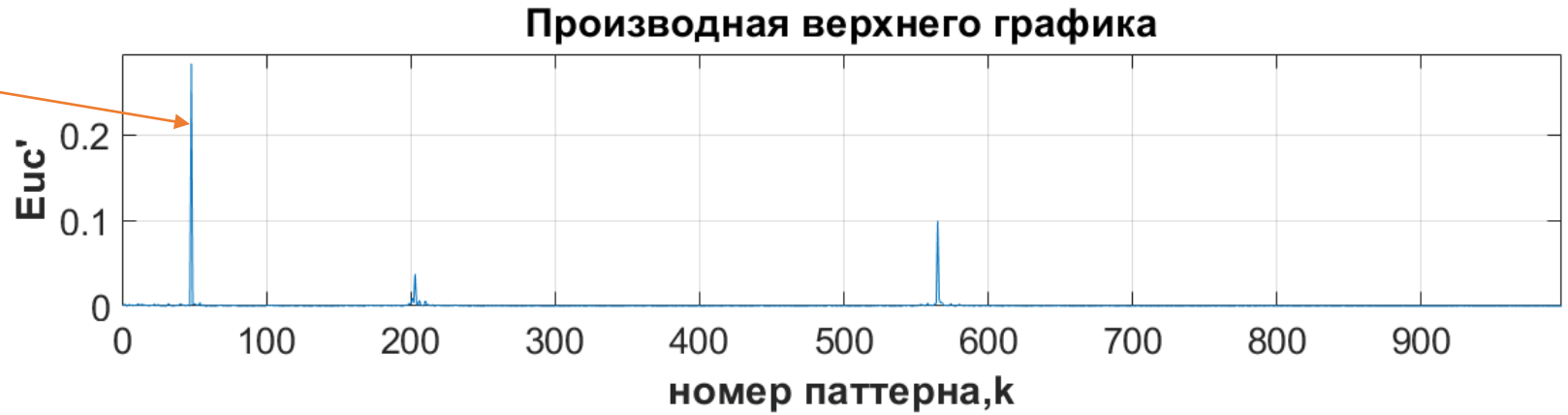
макс. производная



Выбор Epsilon

Epsilon

макс. производная



Оценки качества кластеризации

1) Точность правильного распознавания

$$accuracy = \frac{\hat{N}_{истинных\_паттернов}}{N_{истинных\_паттернов}}$$

2) Ложное распознавание

$$False\_alarm = \frac{\hat{N}_{ложных\_паттернов}}{N - \max(размер\_паттерна) + 1 - N_{истинных\_паттернов}}$$

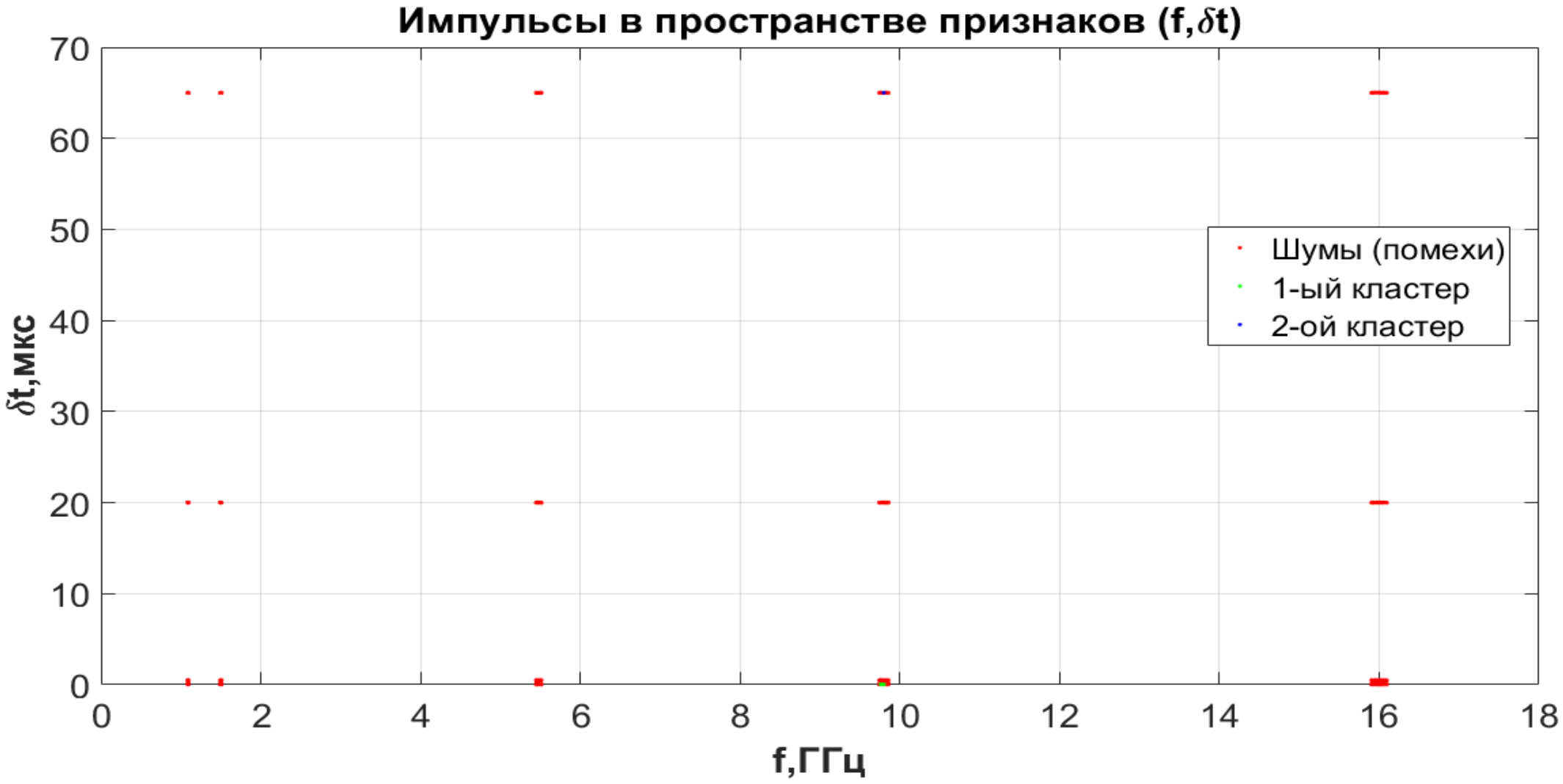
3) Число определенных кластеров $\hat{N}_{clusters}$

Результаты кластеризации на одной реализации

Алгоритм:
Число кластеров - 2
Позиции 1-ого кластера - 1388 1776 2739 2943 3656 4770 5761 7984 8327 8529 9537 9745
Позиции 2-ого кластера - 1806 2326 4045 4896 5576 6276 8874
Истина:
Число кластеров - 2
Позиции 1-ого кластера - 1388 1776 2739 2943 3656 4770 5761 7984 8327 8529 9537 9745
Позиции 2-ого кластера - 1806 2326 4045 4896 5576 6276 8874

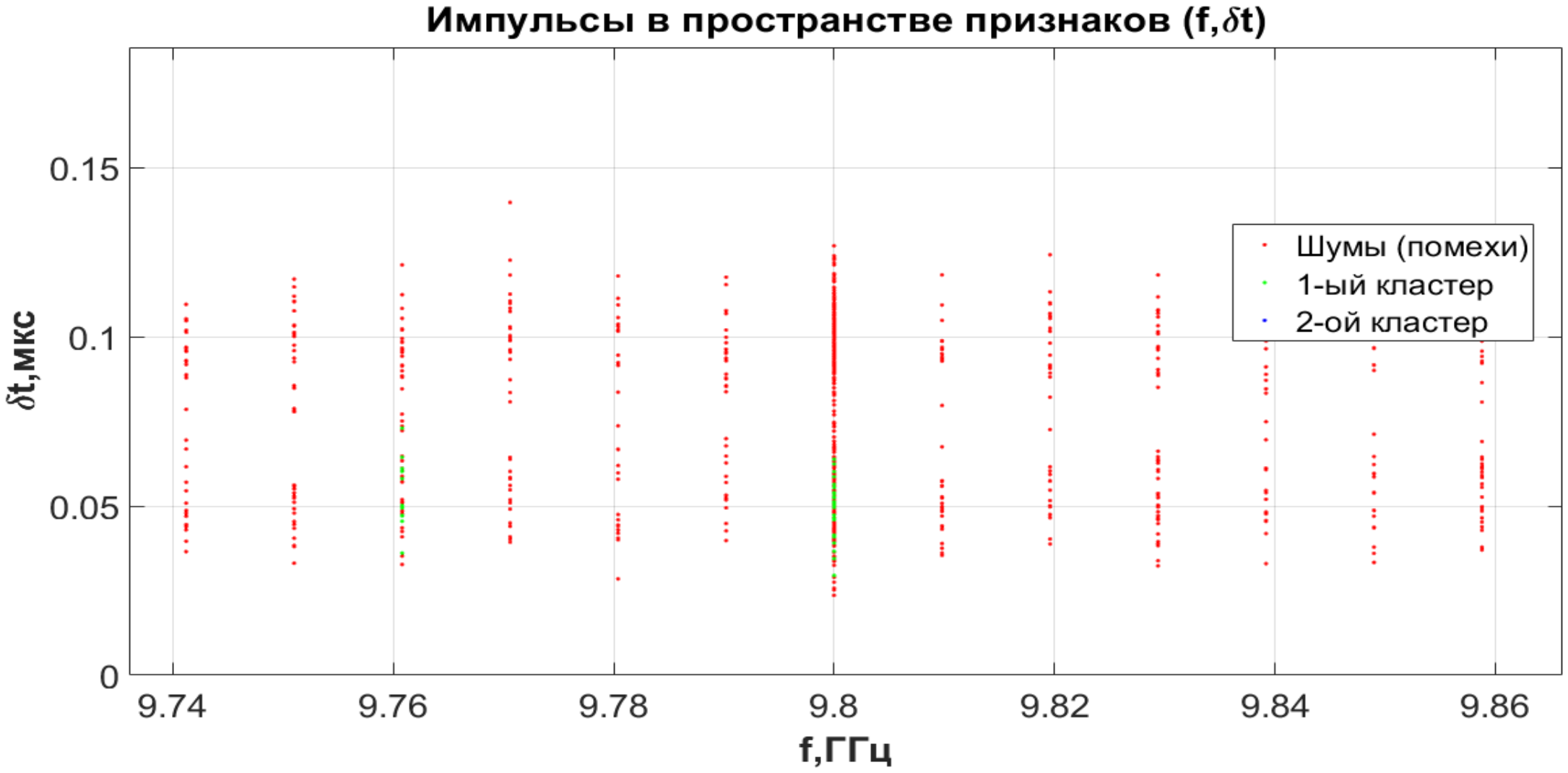
Реальных кластеров - 2 Выявили - 2
Точность распознавания паттернов алгоритмом на данной реализации составила - 100%
Процент определенных ложных паттернов - 0%

Визуализация кластеризации

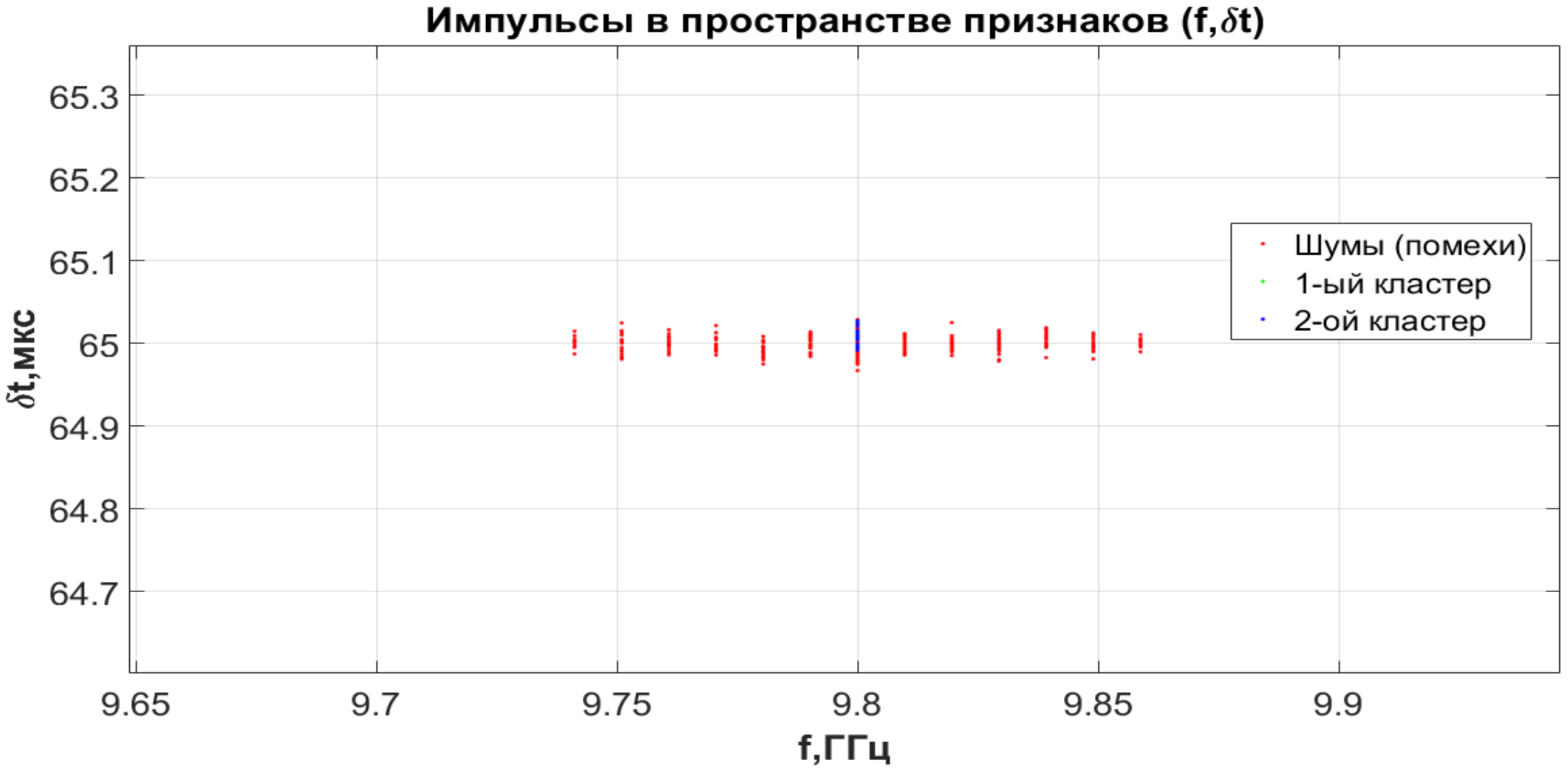


Приблизим 1-ый кластер

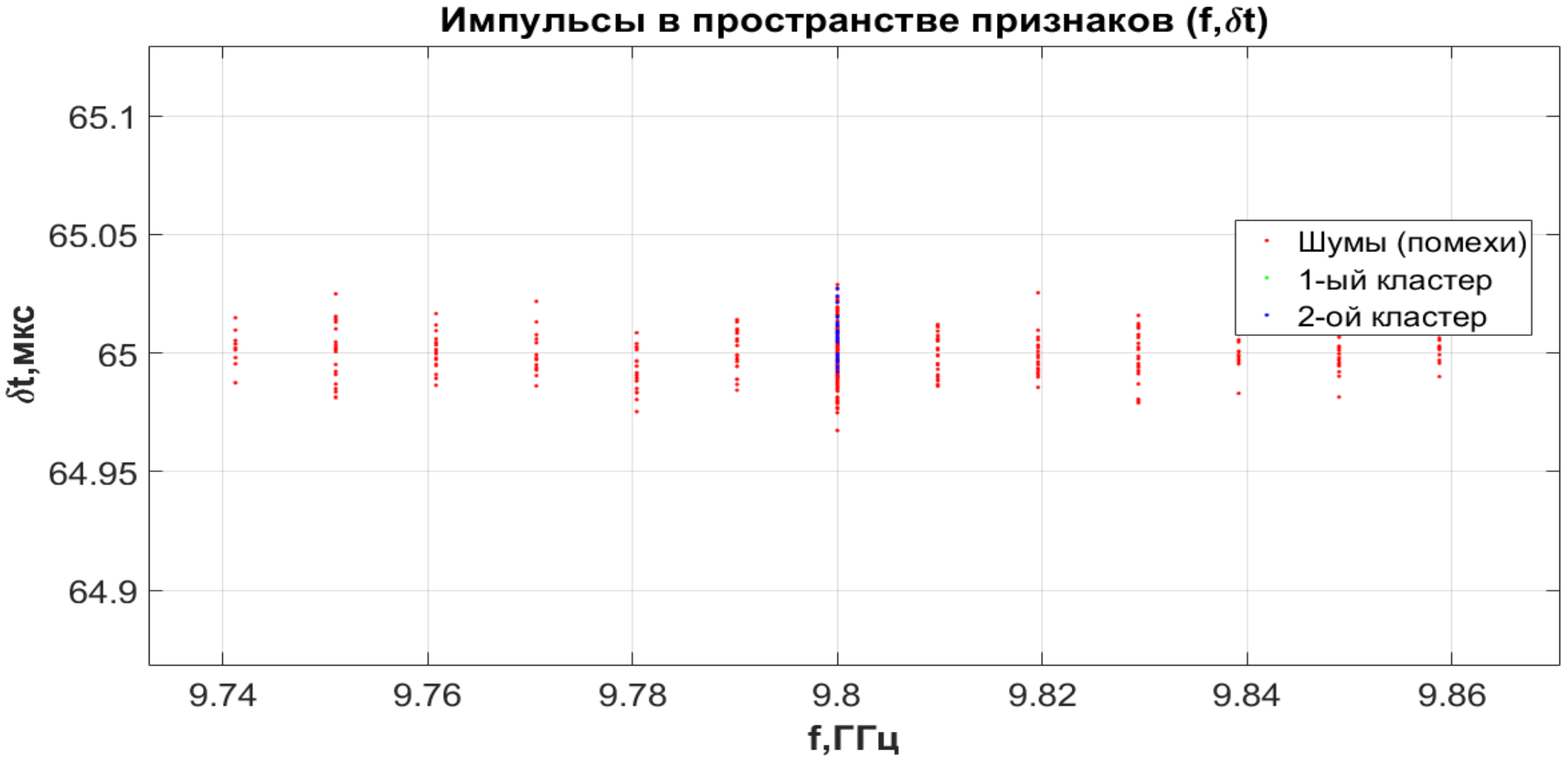
Еще сильнее


Импульсы в пространстве признаков (f,δt)
δt,мкс
0.15
0.1
0.05
0
9.74
9.76
9.78
9.8
9.82
9.84
9.86
f,ГГц
Шумы (помехи)
1-ый кластер
2-ой кластер

Приблизим 2-ой кластер



Еще сильнее приблизим



Результаты кластеризации на 100 реализациях

- Результат кластеризации на 100 случаях, средняя точность определения паттернов составила - 90.0185%
- Средний процент определенных ложных паттернов составил - 0.068572%
- Среднее число выявляемых кластеров составило - 2.4