Министерство науки и высшего образования Российской Федерации
**Федеральное государственное автономное образовательное учреждение высшего образования**
**«ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**
**(национальный исследовательский университет)**
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

# ОТЧЁТ ПО ЗАДАНИЮ №13

по дисциплине «Интеллектуальный анализ больших данных»

Тема: Коллективные аномалии

Выполнил
студент группы КЭ-120
Глизница Максим Николаевич
E-mail: letadllo@mail.ru

Челябинск 2021

## 1. Задание

Выполните поиск коллективных аномалий (выбросов) в двух различных наборах 2-х или 3-мерных данных с помощью двух любых приемов из следующего множества: метод вложенных циклов, метод решеток, кластеризация.

Выполните визуализацию полученных результатов в виде точечных графиков, использующих два цвета для отражения нормальных/аномальных точек.

Выполните поиск коллективных аномалий (выбросов) во временном ряде на основе понятия диссонанса.

Выполните визуализацию полученных результатов в виде точечных графиков, использующих два цвета для отражения нормальных/аномальных подпоследовательностей..

## 2. Краткие сведения о наборах данных

Использованные наборы данных:

Swiss banknote counterfeit detection (https://www.kaggle.com/chrizzles/swiss-banknote-conterfeit-detection). Содержит различные данные о банкнотах, такие как длина, ширина, отступы и т.д. Оригинальный набор содержит данные о 200 банкнотах, среди которых 100 – фальшивые, но для выявления аномалий было взято подмножество, содержащее 100 подлинных и 10 фальшивых банкнот.

Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009). Содержит информацию о произведённых партиях красного вина "Vinho Verde" (такую как pH и содержание хлоридов), а также оценку качества от 1 до 10 (реально в наборе данных присутствуют значения от 3 до 8). Как предлагается в пояснении к набору данных, высококачественным можно считать вино с оценкой 7 и выше. Таким образом, в наборе данных присутствуют 217 партий высокого качества и 1382 партий низкого качества.

US Unemployment Rate by County, 1990-2016 (https://www.kaggle.com/jayrav13/unemployment-by-county-us). Содержит данные об уровне безработицы в различных штатах США, ассоциированные с годом и месяцем.

## 3. Краткие сведения о средствах реализации

Для реализации методов была использована библиотека scikit-learn, включающая в себя множество алгоритмов для анализа данных, и библиотека

Matrix Profile, содержащая различные алгоритмы для выявления закономерностей и аномалий.

Репозиторий по дисциплине: https://github.com/Airpllane/DAAlgorithms.

Каталог для задания: 13. Outliers 2

## 4. Визуализация

На рис. 1 приведена визуализация первого набора данных.

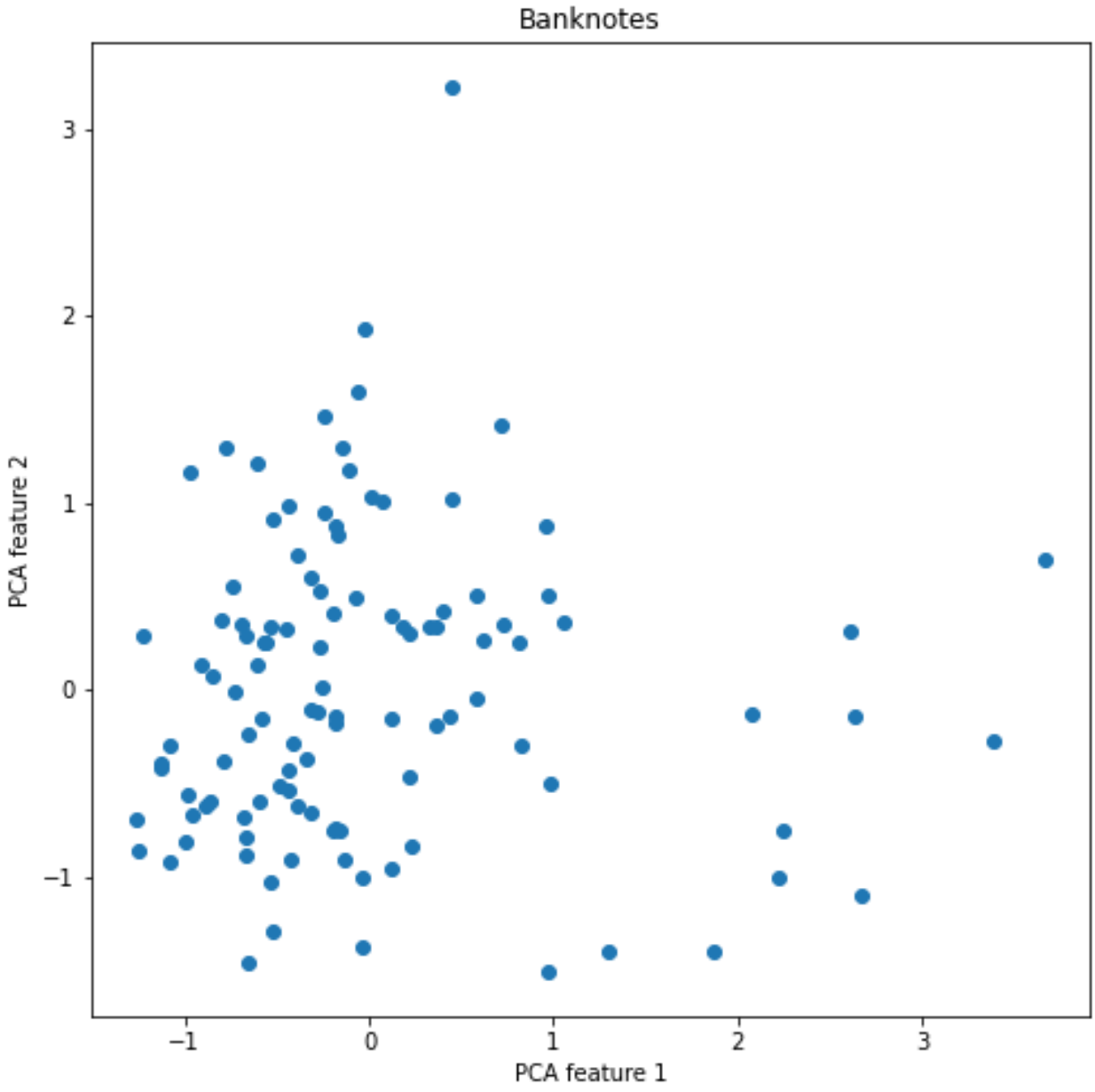


**Рис. 1.** Визуализация набора данных

Из рисунка можно увидеть, что многие экземпляры близки друг к другу, в то время как некоторые экземпляры удалены от этой группы.

Далее был выполнен поиск аномалий в этом наборе данных с помощью метода вложенных циклов и плотностной кластеризации. Результаты поиска приведены на рис. 2.

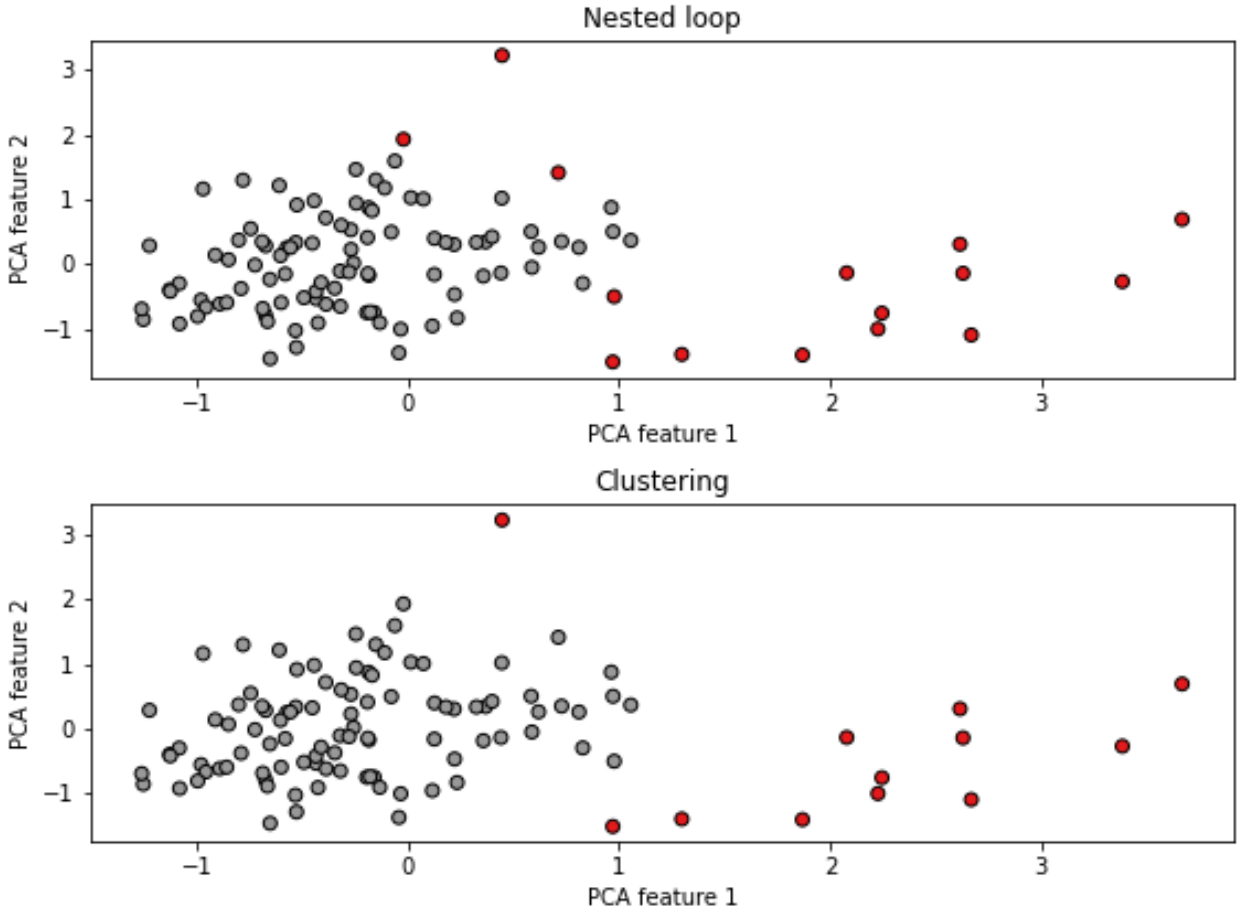


**Рис. 2.** Результаты поиска аномалий с помощью вложенных циклов и кластеризации

Из рисунка можно увидеть, что оба метода поиска смогли корректно выделить область, содержащую большую часть экземпляров, но также выделили некоторые сравнительно близкие к этой области экземпляры как аномалии.

Далее был загружен второй набор данных, визуализация которого приведена на рис. 3.

Видно, что экземпляры этого набора составляют плотную область, после чего становятся всё более разреженными по мере удаления от неё.

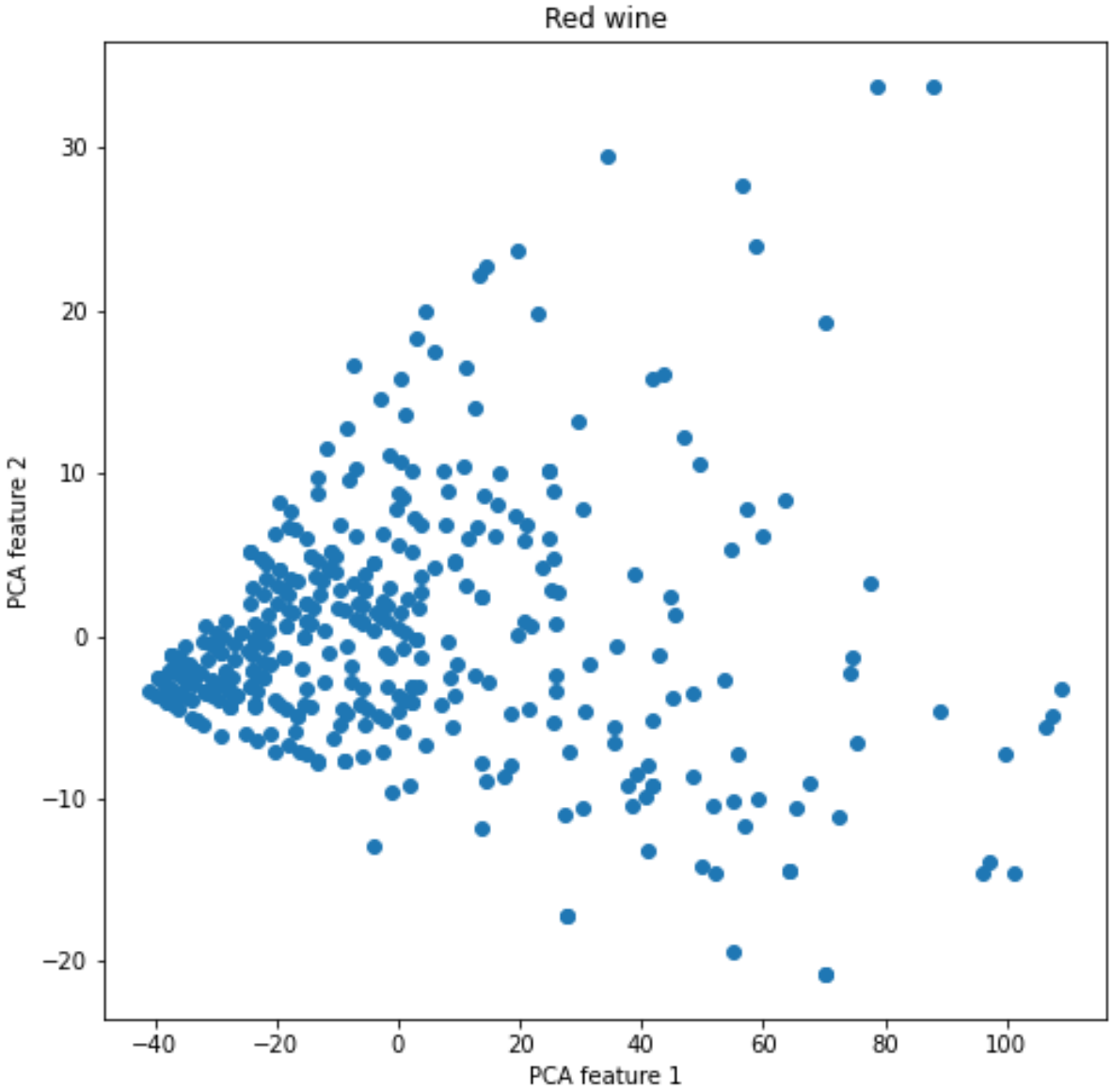


**Рис. 3.** Визуализация набора данных

Далее был выполнен поиск аномалий в этом наборе данных с помощью тех же алгоритмов. Результаты поиска приведены на рис. 4.

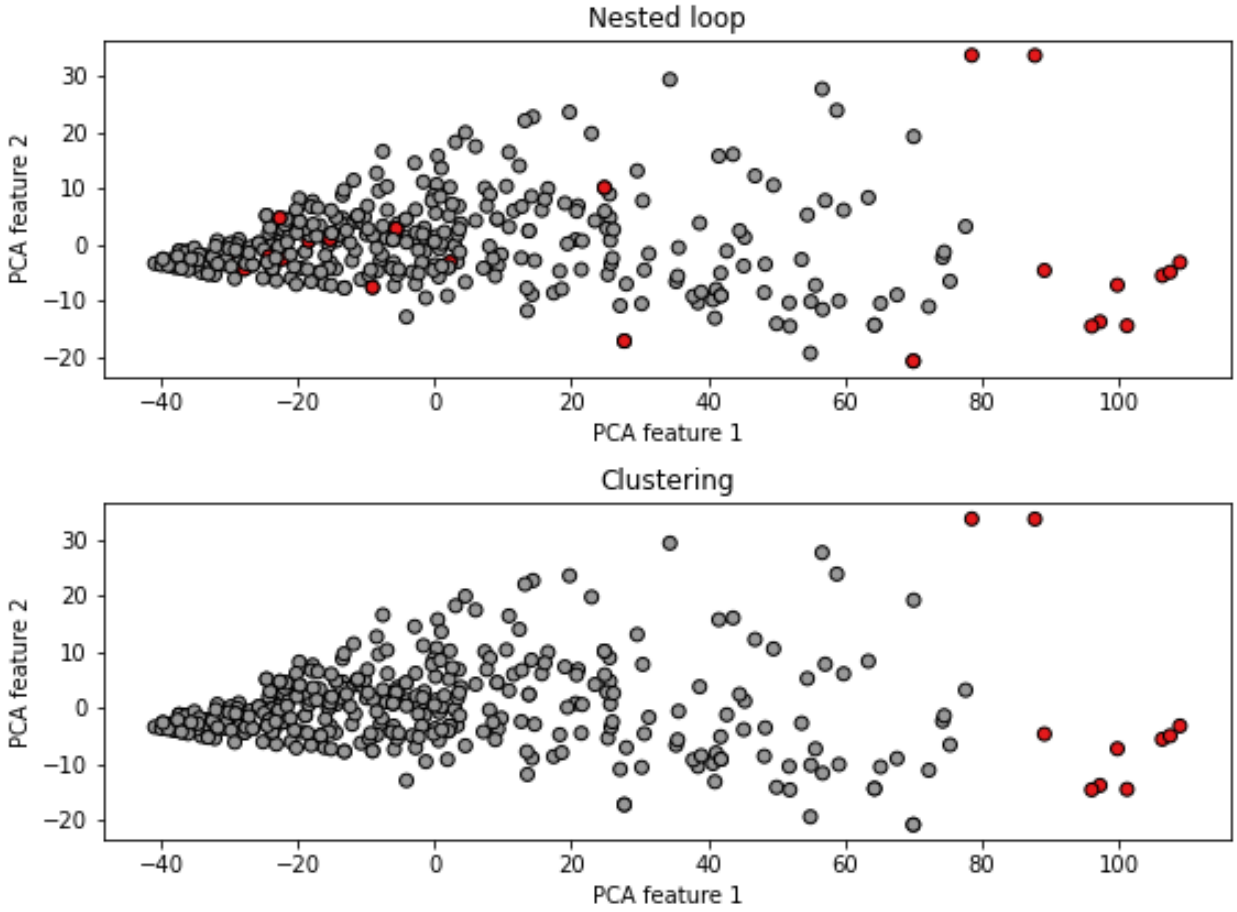


**Рис. 4.** Результаты поиска аномалий с помощью вложенных циклов и кластеризации

Из рисунка видно, что методы также способны корректно выделить «нормальную» зону в наборе данных. Конкретные пределы зоны зависят от выбранных параметров алгоритмов. Можно также увидеть, что метод вложенных циклов некорректно классифицирует некоторые экземпляры, находящиеся в «нормальной» зоне, как аномалии.

Таким образом, можно сделать вывод, что оба рассмотренных алгоритма можно использовать для поиска аномалий в двухмерных наборах данных, но при этом важно следить за значениями параметров алгоритмов. Также при использовании вложенных циклов есть риск классифицировать часть нормальных экземпляров как аномалии.

Далее был загружен третий набор данных, содержащий временной ряд. Визуализация набора приведена на рис. 5.

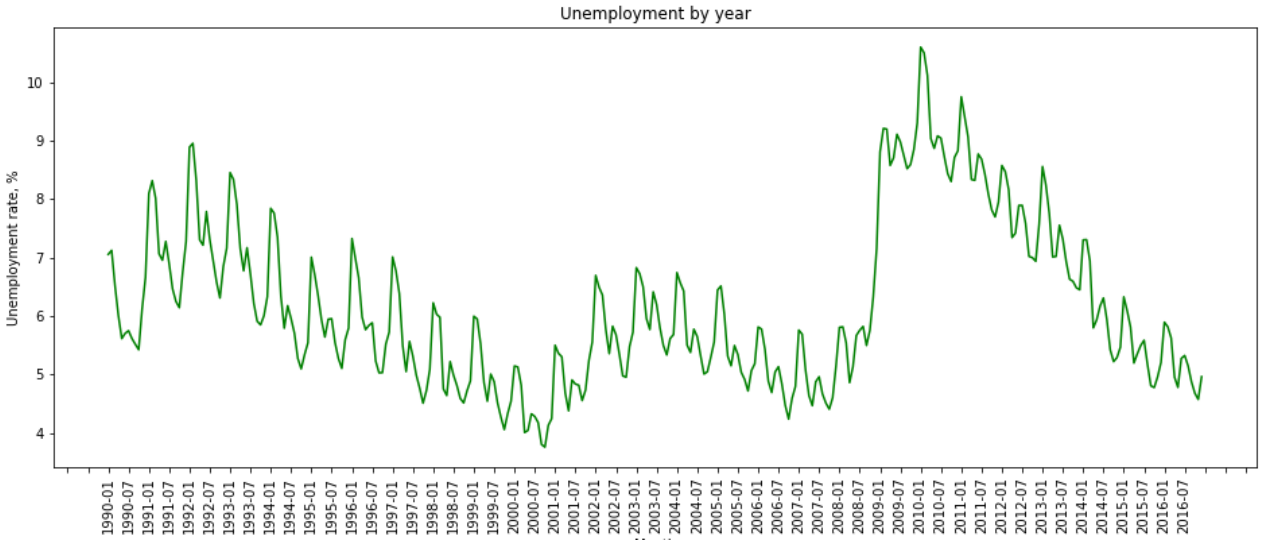


**Рис. 5.** Визуализация набора данных

В наборе можно увидеть скачки безработицы в 2001 и 2009 году, которые можно считать аномалиями во временном ряде.

Далее был выполнен поиск аномалий в этом наборе данных на основе понятия диссонанса. Количество отмечаемых аномалий было установлено как 2. Результаты поиска приведены на рис. 6 и рис. 7.

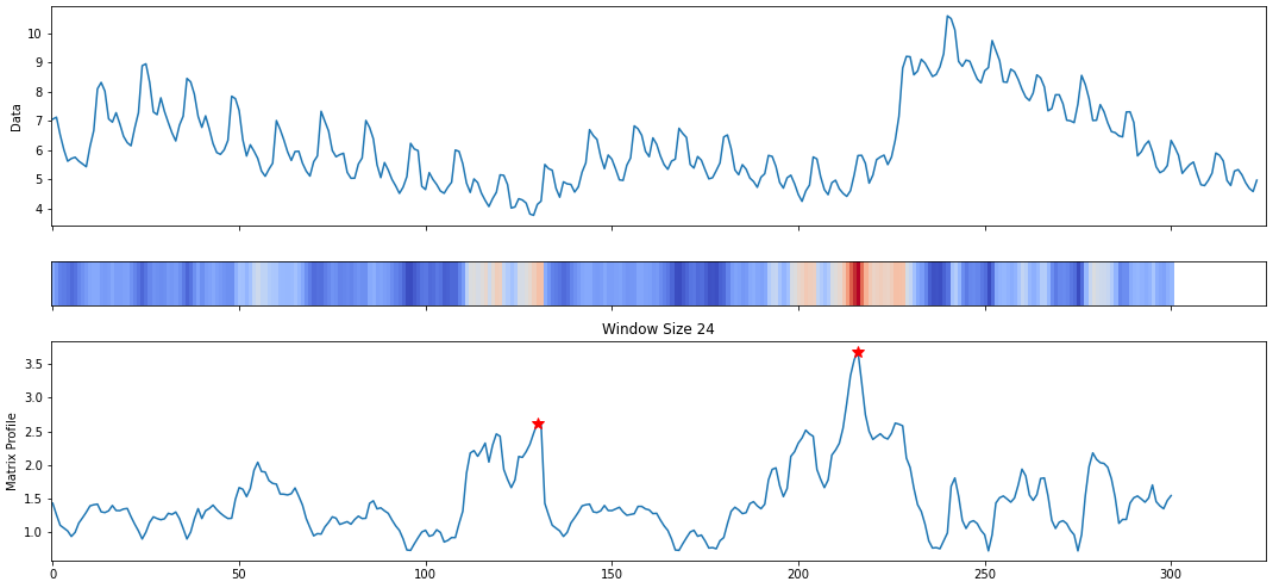


**Рис. 6.** График Matrix Profile

Как можно увидеть из этого графика, диссонансы были отмечены приблизительно в моменты скачков безработицы.

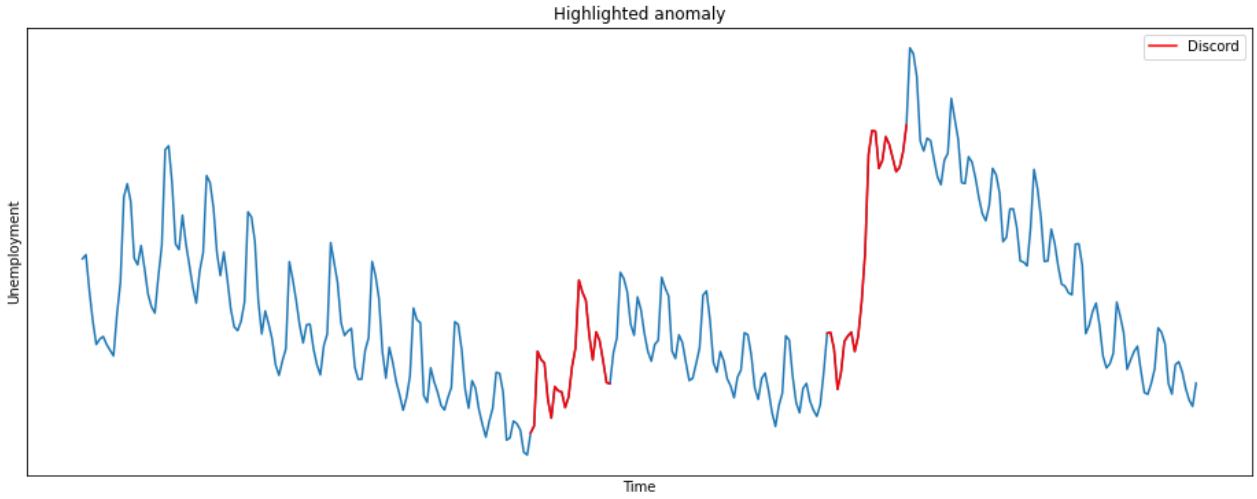


**Рис. 7.** Временные рамки аномалий

Как можно увидеть из этого рисунка, временные рамки скачков безработицы были отмечены корректно.

Таким образом, для поиска аномалий во временных рядах можно использовать графики Matrix Profile, которые позволяют найти диссонансные участки во временных рядах.