

МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ
по научно-исследовательской работе
Тема: Разработка метода и алгоритмов детектирования аномалий в
природных данных

Студент гр. 8306

Щербаков И.Б.

Научный руководитель

Мандрикова Б.С.

Руководитель практики

Заславский М.М.

Санкт-Петербург

2023

ЗАДАНИЕ НА НАУЧНО-ИССЛЕДОВАТЕЛЬСКУЮ РАБОТУ

Студент Щербаков И.Б.

Группа 8306

Тема НИР: Разработка метода и алгоритмов детектирования аномалий в природных данных

Задание на НИР:

Улучшение метода считывания данных с нейтронных мониторов. Построение модели для идентификации аномалий в данных вариаций космических лучей.

Сроки выполнения НИР: 01.09.2023 – 20.12.2023

Дата сдачи отчета: 20.12.2023

Дата защиты отчета: 26.12.2023

Студент гр. 8306

Щербаков И.Б.

Научный руководитель

Мандрикова Б.С.

Руководитель практики

Заславский М.М.

АННОТАЦИЯ

В данной научно-исследовательской работе была изучена структура алгоритма ITD, основанного на преобразовании Фурье и вейвлет-преобразований). Реализованы модули, отвечающие за считывание данных с веб-сервиса, содержащего сведения с нейтронных мониторов. Реализовано отображение ключевых параметров для оценки качества определения аномалий: трёхмерный график, графики аппроксимаций и детализаций. Проведённая работа имеет практическое применение в области мониторинга космических лучей.

SUMMARY

In this research work the structure of ITD algorithm based on Fourier transform and wavelet transforms) was studied. Modules responsible for reading data from a web service containing information from neutron monitors were implemented. Display of key parameters for anomaly detection quality assessment was realized: three-dimensional graph, approximation and detail plots. This work has practical application in the field of cosmic ray monitoring.

СОДЕРЖАНИЕ

| | |
|---|--|
| ВВЕДЕНИЕ | |
| 1.АЛГОРИТМ INTRINSIC TIME-SCALE DECOMPOSITION | |
| 1.1. Введение в алгоритм | |
| 1.2. Сравнительный анализ алгоритмов Empirical Mode Decomposition и Intrinsic Time-Scale Decomposition | |
| 1.3. Сравнительный известных методов TFE-анализа..... | |
| 1.4. Результаты исследования метода ITD | |
| 2. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ | |
| 2.1. Реализация программного продукта..... | |
| Описание предполагаемого метода решения | |
| План работы на весенний семестр | |
| ЗАКЛЮЧЕНИЕ | |

ВВЕДЕНИЕ

Аномалии в природных данных представляют собой отклонения от ожидаемого нормального поведения и могут быть связаны с различными природными явлениями. В случае с космическими лучами, есть несколько основных причин, которые могут объяснить возникновение аномалий: солнечные вспышки, геомагнитные бури, солнечные циклы и солнечные ветры [1]. На основе проведенной научно-исследовательской работы в осеннем семестре, связанной со сравнением методов анализа нестационарных сигналов, было принято решение о использовании искусственных нейронных сетей для анализа данных с нейтронных мониторов.

При использовании веб-сервиса, хранящего данные нейтронных мониторов [10], возникла потребность в изучении инструмента парсинга данного веб-ресурса, для автоматизированного получения данных с различных станций в указанное для исследований время для того, чтобы проводить исследования по конкретной выборке значений.

Для более эффективного анализа временных рядов данных с нейтронных мониторов в рамках данного исследования был выбран и реализован метод Intrinsic Time-Scale Decomposition (ITD). ITD представляет собой инновационный подход к разложению временных рядов на внутренние временные масштабы, что позволяет выделять важные компоненты сигнала и отделять их от шумовых артефактов. Данный метод подходит для анализа нестационарных сигналов, так как он учитывает изменения в активности, которые могут происходить на различных временных шкалах.

1.АЛГОРИТМ INTRINSIC TIME-SCALE DECOMPOSITION

1.1. Введение в алгоритм

Метод Intrinsic Time-Scale Decomposition (ITD) эффективно преодолевает ограничения классических и более поздних подходов к частотно-временному анализу нестационарных сигналов. В отличие от методов, основанных на преобразовании Фурье, вейвлет-преобразовании или эмпирическом модальном разложении (EMD), ITD разлагает сигнал на высокочастотные компоненты с четко определенной мгновенной частотой и амплитудой, а также на монотонный тренд, аналогичный модам в EMD.

Особенностью алгоритма ITD является сохранение точной временной информации о критических точках сигнала и бегущих волнах с высоким временным разрешением, соответствующим временному масштабу экстремальных точек во входном сигнале. Применение ITD приводит к созданию нового класса мощных фильтров сигналов реального времени, способных извлекать и использовать информацию о мгновенной амплитуде, частоте и фазе, а также других морфологических особенностях сигнала.

Математически, алгоритм ITD представляет собой итеративный процесс декомпозиции исходного сигнала на высокочастотные («proper rotation» - H) и низкочастотные («baseline» - L) компоненты.

$$L_t^0 = X_t = L_t^D + \sum_{j=1}^D H_t^j; \quad L_t^j = L_t^{j+1} + H_t^{j+1}; \quad j = 0 \dots D$$

Рисунок 1. Формула алгоритма ITD

Пусть имеется сигнал X. Определим оператор \mathcal{L} , извлекающий из сигнала низкочастотную компоненту («baseline») таким образом, что остаток является высокочастотной компонентой («proper rotation»). Тогда сигнал X может быть записан следующим образом:

$$X_t = \mathcal{L}X_t + (1 - \mathcal{L})X_t = L_t + H_t$$

Рисунок 2. Формула записи сигнала X.

1.2. Сравнительный анализ алгоритмов Empirical Mode Decomposition и Intrinsic Time-Scale Decomposition

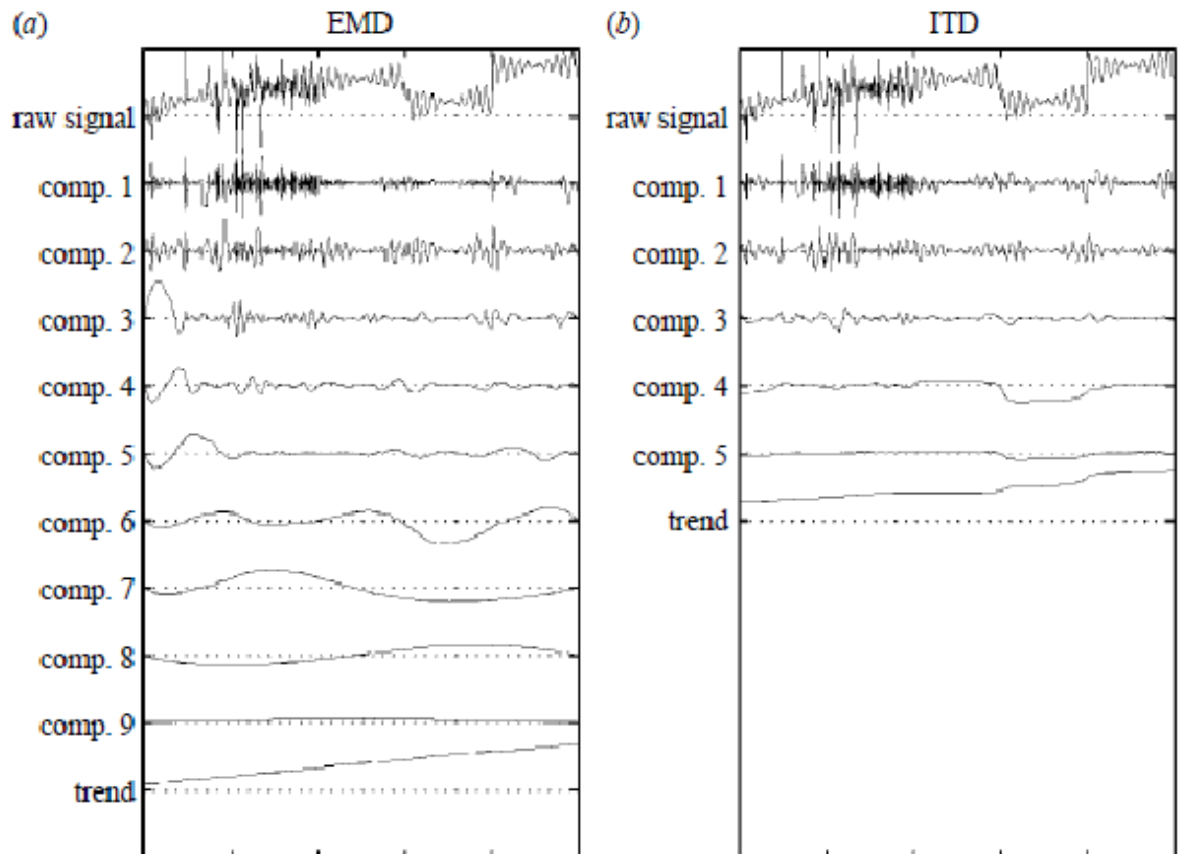


Рисунок 3. Работа алгоритмов EMD и ITD.

На рисунке 3 представлена работа алгоритмов ETD и ITD. Исходный сигнал обозначен как *raw signal*. Можно заметить, что алгоритм ITD позволяет разложить сигнал на меньшее число более стабильных составляющих. Краевые эффекты в EMD обусловлены интерполяцией сплайнами – что требует существенно большее время работы операций просеивания. Алгоритм ITD лишён этих особенностей и может работать в режиме реального времени.

1.3. Сравнительный известных методов TFE-анализа

| Преобразование | Фурье | Вейвлет | Гильберт |
|----------------------|---|---|---|
| Базис | Априорный | Априорный | Адаптивный |
| Частота | Свертка, глобальная, неопределенность | Свертка, региональная, неопределенность | Дифференциация, локальная, определенность |
| Представление | Энергия-частота | Энергия-время- частота | Энергия-время- частота |
| Нелинейность | Нет | Нет | Да |
| Нестационарность | Нет | Да | Да |
| Извлечение признаков | Нет | Дискретная форма: нет, непрерывная форма: да | Да |
| Теоретическая база | Нет | Полная теория | Да |

Алгоритмы EMD, ITD, а также преобразование Гильберта, широко применяются в следующих областях:

- Биомедицинские приложения;
- Нейробиология;
- Эпидемиология;
- Сейсмические исследования;
- Распознавание речи.

1.4. Результаты исследования метода ITD

Проведенный анализ существующих решений в области обработки нестационарных сигналов показывает необходимость модификации существующих и разработки новых методов адаптивной обработки сигналов с целью повышения их эффективности;

Показана целесообразность использования и дальнейшей модификации алгоритма Intrinsic Time-Scale Decomposition с целью использования в качестве основы в алгоритмах детектирования аномалий в природных данных.

2. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

2.1. Реализация программного продукта

Для извлечения данных с сайта нейтронных мониторов [10] была установлена и внедрена библиотека BeautifulSoup.

Перед использованием необходимо установить пакет BeautifulSoup и ознакомиться со способом запросов к веб-сервису, чтобы отобразить на странице требуемые данные.

HTTP ADDRESS

Another way to use NEST is to write the different parameters like stations, time, etc. in the http address:

Example 1:

```
http://nest.nmdb.eu/draw_graph.php?formchk=1&stations[]=KERG&stations[]=KIEL&output=ascii&tabchoice=ori&dtype=corr_for_efficiency&
date_choice=bydate
&start_year=2009&start_month=09&start_day=01&start_hour=00&start_min=00&end_year=2009
&end_month=09&end_day=05&end_hour=23&end_min=59&yunits=0
```

Example 2:

```
http://nest.nmdb.eu/draw_graph.php?formchk=1&stations[]=JUNG&output=plot&tabchoice=revori&dtype=corr_for_efficiency&odtype[]=uncorrected&
date_choice=
last&last_days=5&last_label=days_label&tresolution=10&envdata[]=measured_relative_humidity&yunits=0
```

Рисунок 3. Способы составления запросов к веб-сайту.

The list of main parameters:

| Parameter | Possible values |
|-------------|--|
| stations[] | APTY, AATB, ATHN, ARNM etc. |
| allstations | 1 (selects all the stations) |
| tabchoice | revori, ori, 1h |
| dtype | corr_for_efficiency, corr_for_pressure, uncorrected, pressure_mbar |
| tresolution | 0, 2, 5, 10, 30, 60, 120, 360, 720, 1440, 39276, 525969 (0 correspond to "best") |
| date_choice | last, bydate (, bygle, byforbush) |
| last_days | 1 to 9999 |
| last_label | days_label, hours_label or mins_label |
| force | 1 or 0 |
| yunits | 1 or 0 (% or counts/mbar for a single station query) |
| yscale | 1 or 0 |
| yscalemin | -9999 to 9999 or min |
| yscalemax | -9999 to 9999 or max |
| anomalous | 1 or 0 |
| shift | 0 to 999 |
| start_day | 1 to 31 |
| end_day | 1 to 31 |
| start_month | 1 to 12 |
| end_month | 1 to 12 |
| start_year | 1956 to curent year+1 |
| end_year | 1956 to current year +1 |
| start_hour | 0 to 23 |
| end_hour | 0 to 23 |
| start_min | 0 to 59 |
| end_min | 0 to 59 |
| output | plot, ascii, both |
| asciifrac | 1 or 0 |
| odtype[] | corr_for_efficiency, corr_for_pressure, uncorrected, pressure_mbar |
| ridtype | ssi, msi |
| envdata[] | measured_temperature_inside, measured_temperature_outside, measured_wind_speed_m_s, measured_relative_humidity |

Рисунок 4. Список всех параметров.

Для начала воспользуемся любой станцией, указав в параметре stations[], возьмём данные за два года, указав соответственно следующие атрибуты: start_day, end_day, start_month, end_month, start_year, end_year, start_hour, end_hour, start_min и end_min. Первоначально воспользуемся скорректированными от данными, указав показатель dtype как corr_for_efficiency.

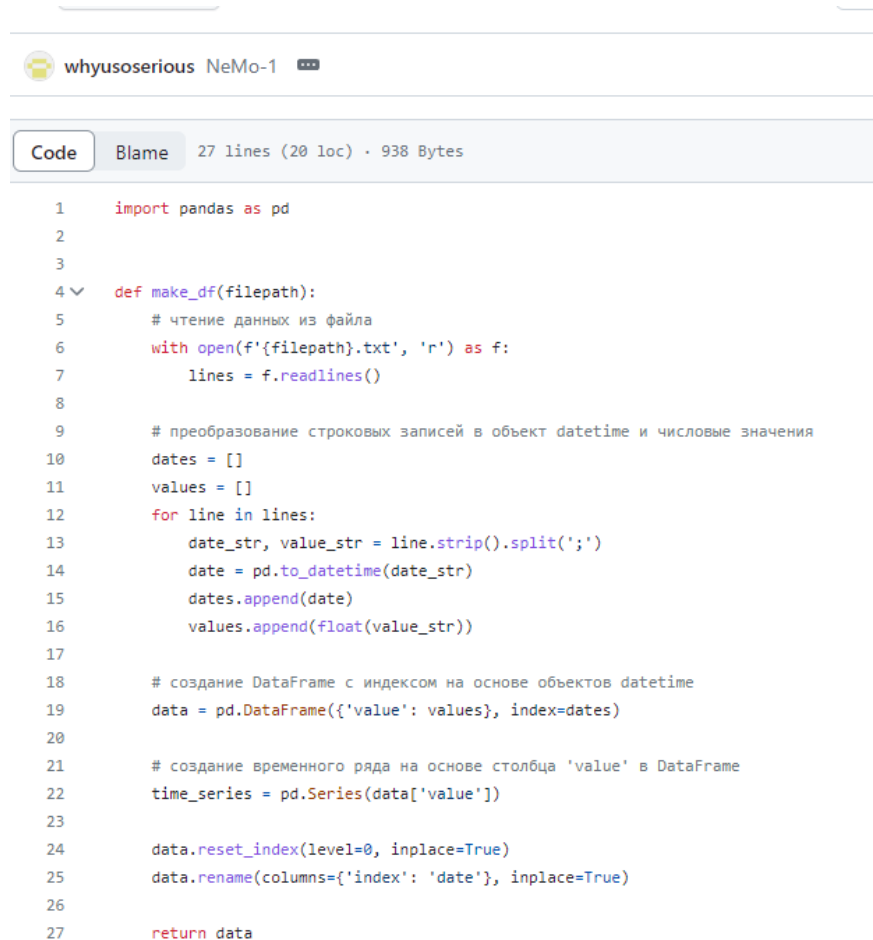
Реализованный код указан в приложении Б. Он импортирует требуемые модули, такие как: `urlopen` - для открытия URL-адреса и получения содержимого страницы. `Os` – для работы с файловой системой. `BeautifulSoup` – для непосредственного парсинга HTML-разметки. Указываем требуемые параметры для обращения. Для каждой указанной станции формируется необходимый URL-адрес. Извлекается текст из нужного блока HTML-разметки. Производится последующая обработка найденного текста: удаление лишних текстовых значений. Обработанный текст добавляется в директорию, которая содержит в себе название страницы, с созданием текстового файла. В случае отсутствия данных или другой ошибки выводится соответствующее сообщение. Всего значений получено – 34432.

The screenshot shows a GitHub repository for user `whyusoserious`, specifically the `nemo` repository. The file path is `nemo / 2021-2022 / Measured_data / ROME.txt`. The file is 942 KB and contains 34432 lines of code. The code is displayed in a table with line numbers from 1 to 23, showing a series of timestamp and value pairs.

| Line | Content |
|------|-----------------------------|
| 1 | 2021-01-01 00:00:00;127.655 |
| 2 | 2021-01-01 00:30:00;127.540 |
| 3 | 2021-01-01 01:00:00;128.091 |
| 4 | 2021-01-01 01:30:00;127.949 |
| 5 | 2021-01-01 02:00:00;127.673 |
| 6 | 2021-01-01 02:30:00;127.953 |
| 7 | 2021-01-01 03:00:00;128.846 |
| 8 | 2021-01-01 03:30:00;129.212 |
| 9 | 2021-01-01 04:00:00;129.104 |
| 10 | 2021-01-01 04:30:00;129.487 |
| 11 | 2021-01-01 05:00:00;129.499 |
| 12 | 2021-01-01 05:30:00;129.324 |
| 13 | 2021-01-01 06:00:00;129.659 |
| 14 | 2021-01-01 06:30:00;130.484 |
| 15 | 2021-01-01 07:00:00;130.050 |
| 16 | 2021-01-01 07:30:00;130.020 |
| 17 | 2021-01-01 08:00:00;130.815 |
| 18 | 2021-01-01 08:30:00;130.327 |
| 19 | 2021-01-01 09:00:00;130.359 |
| 20 | 2021-01-01 09:30:00;130.403 |
| 21 | 2021-01-01 10:00:00;129.536 |
| 22 | 2021-01-01 10:30:00;130.592 |
| 23 | 2021-01-01 11:00:00;130.312 |

Рисунок 5. Пример полученных данных.

Для дальнейшего анализа полученного временного ряда с помощью ITD необходимо подготовить соответствующий датасет на полученном текстовом файле.



The screenshot shows a code editor window with the title 'whyusoserious NeMo-1'. The editor has tabs for 'Code' and 'Blame', and a status bar indicating '27 lines (20 loc) · 938 Bytes'. The code is as follows:

```
1 import pandas as pd
2
3
4 def make_df(filepath):
5     # чтение данных из файла
6     with open(f'{filepath}.txt', 'r') as f:
7         lines = f.readlines()
8
9     # преобразование строковых записей в объект datetime и числовые значения
10    dates = []
11    values = []
12    for line in lines:
13        date_str, value_str = line.strip().split(';')
14        date = pd.to_datetime(date_str)
15        dates.append(date)
16        values.append(float(value_str))
17
18    # создание DataFrame с индексом на основе объектов datetime
19    data = pd.DataFrame({'value': values}, index=dates)
20
21    # создание временного ряда на основе столбца 'value' в DataFrame
22    time_series = pd.Series(data['value'])
23
24    data.reset_index(level=0, inplace=True)
25    data.rename(columns={'index': 'date'}, inplace=True)
26
27    return data
```

Рисунок 6. Код для подготовки данных к дальнейшему анализу.

Данные были разбиты по четырём временным отрезкам, для лучшей детализации данных. Используя Jupyter Notebook был написан код, который реализует ITD метод.

Полученные результаты были разбиты на несколько rotations, после применения оптимизации для устранения шумов, были получены высокочастотные и низкочастотные компоненты временных рядов, что позволило более детально рассмотреть изменения в сигналах.

С кодом можно ознакомиться в репозитории GitHub [9].

Прилагаются графики, иллюстрирующие результаты анализа данных с нейтронных мониторов с использованием ITD:

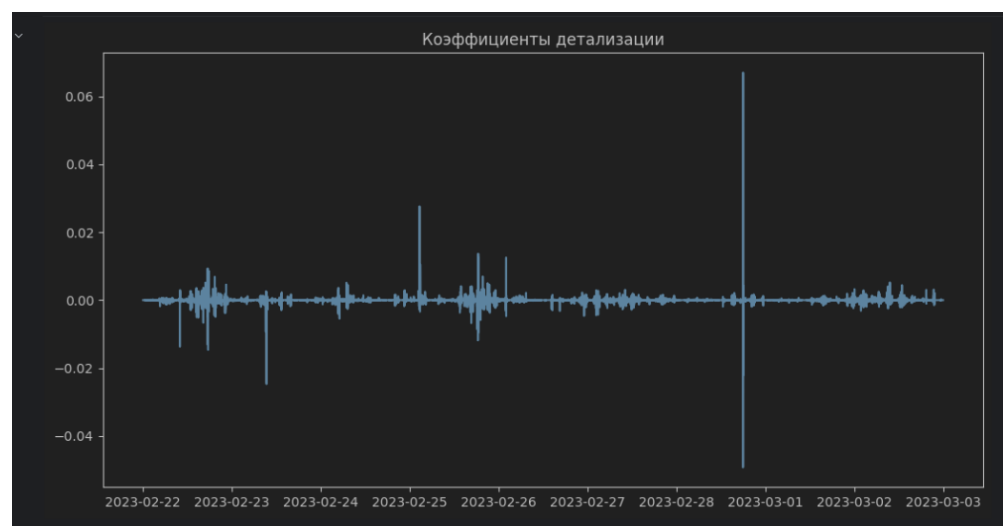
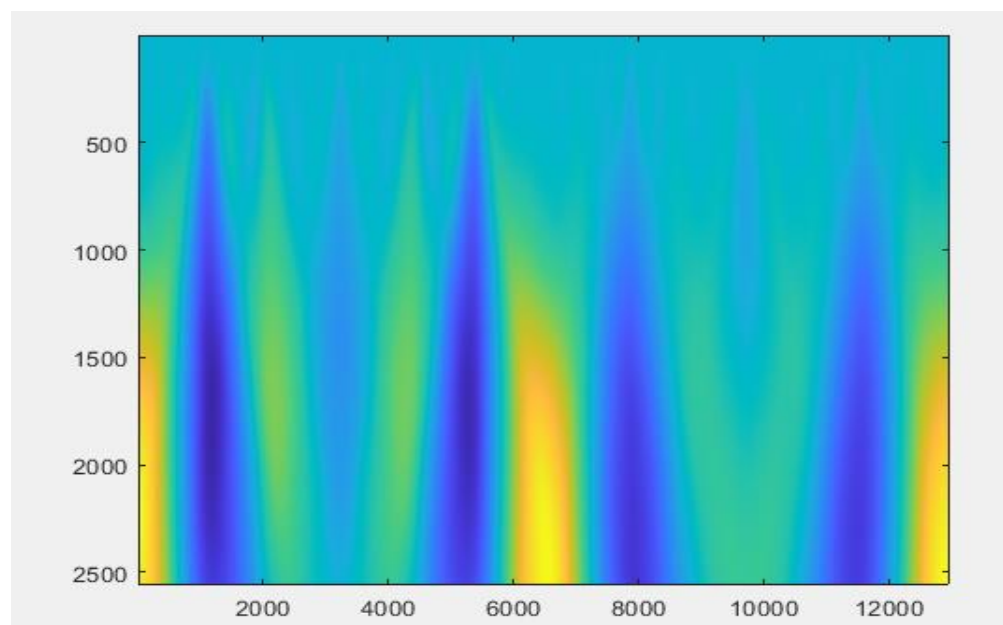
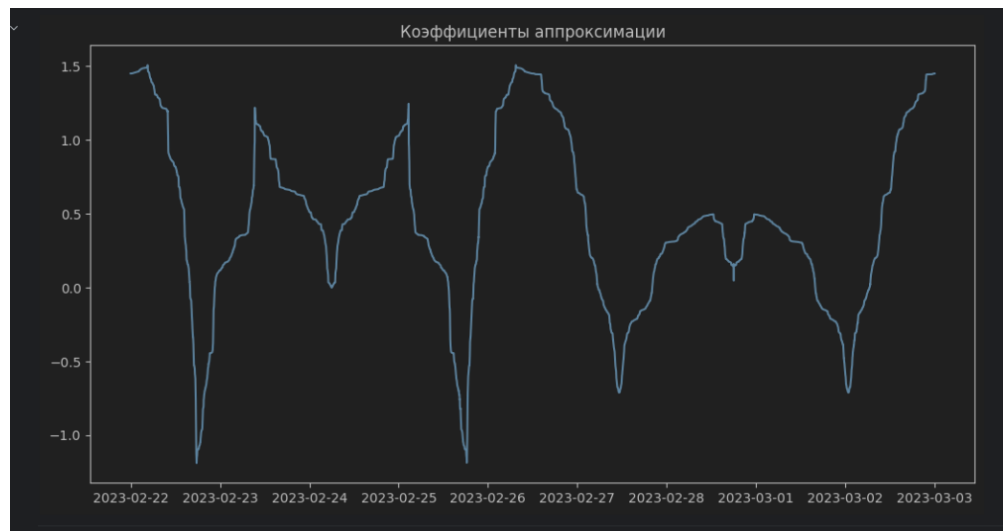


График 1. Даты (22.02.2023 – 03.03.2023)

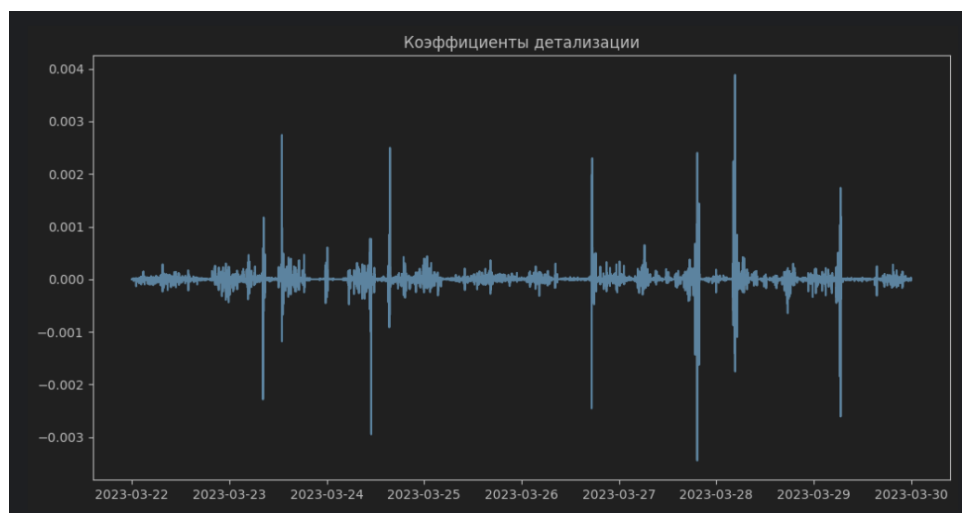
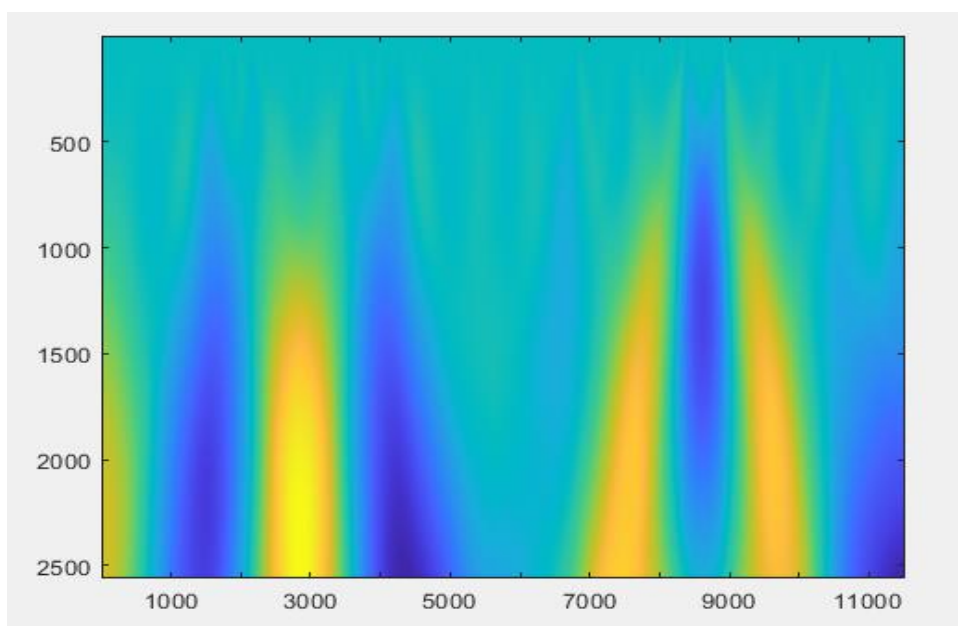
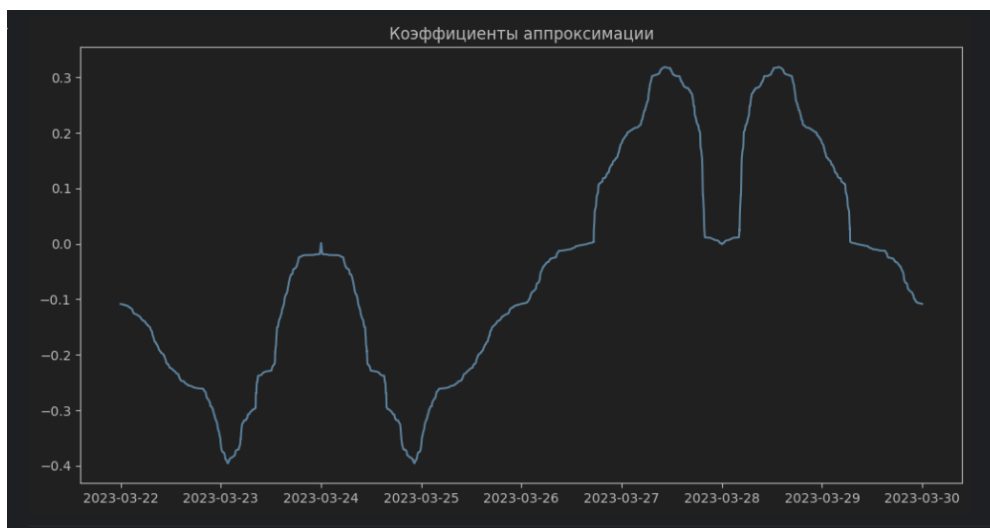


График 2 (22.03.2023 – 30.03.2023)

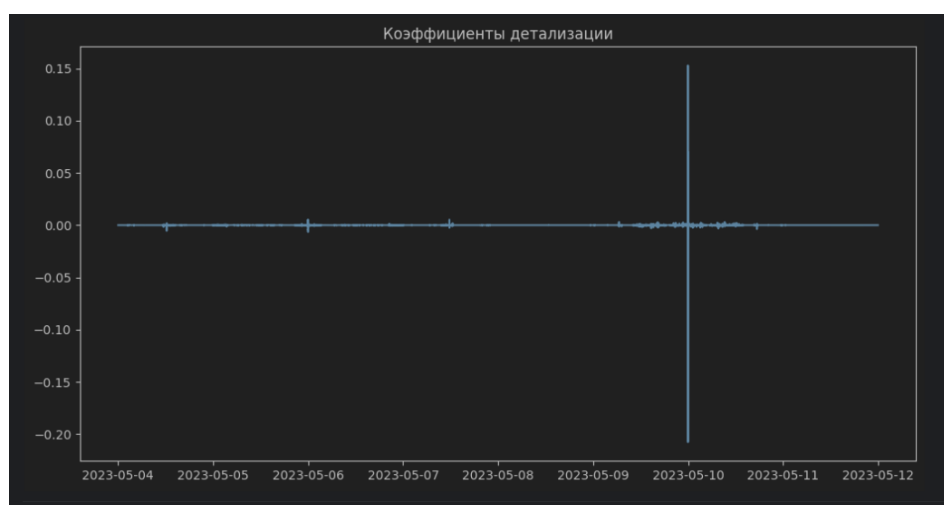
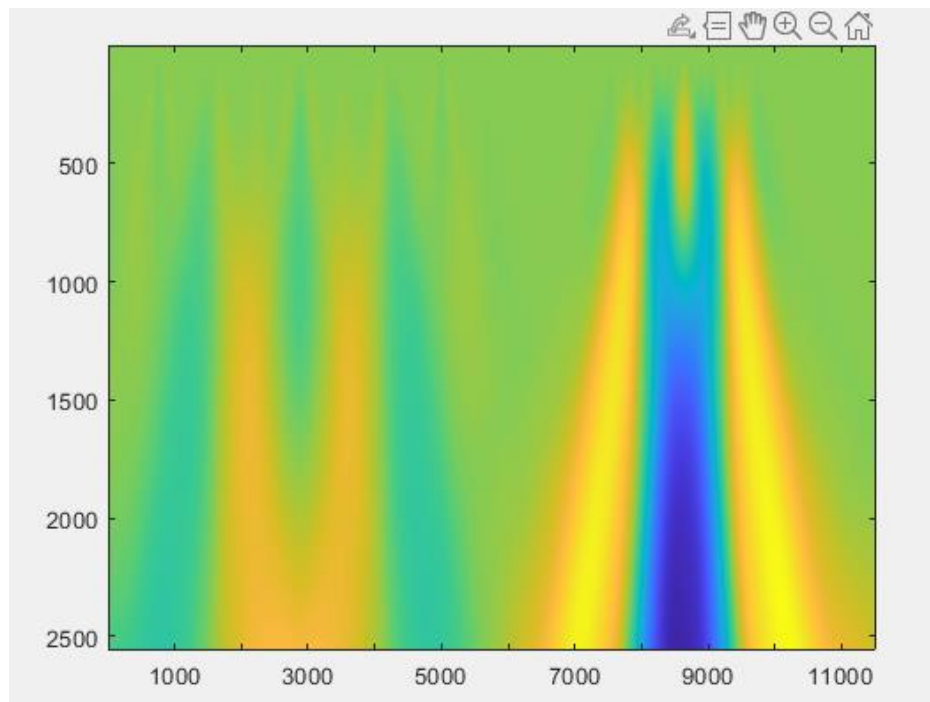
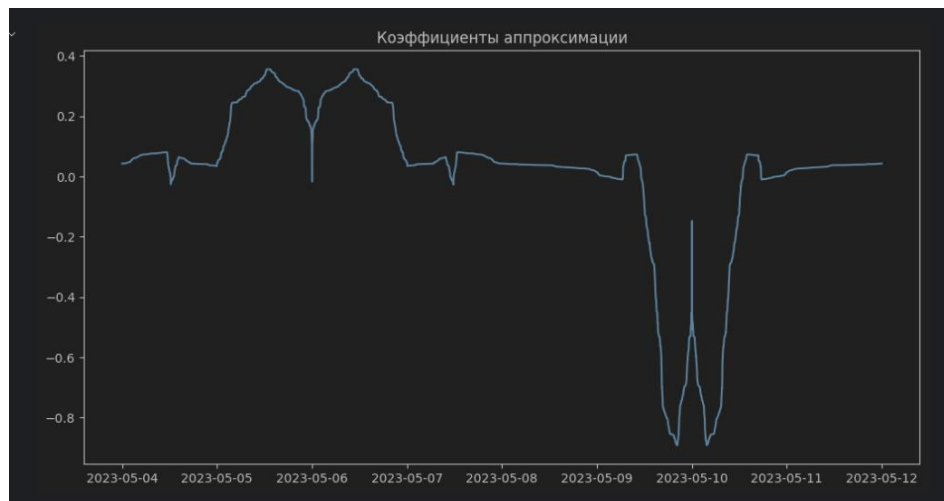


График 3 (04.05.2023 – 12.05.2023)

Описание предполагаемого метода решения

1. Архитектура: Использование ITD в качестве основного метода для декомпозиции временных рядов на высокочастотные и низкочастотные компоненты; Разработка модуля визуализации результатов для более наглядного анализа;
2. Технические детали: Программная реализация на языке программирования Python; Использование среды Jupyter Notebook для open-source: возможности вносить правки в нужные блоки кода и быстро получать экспериментальные данные;
3. Макет интерфейса: Возможность выбора определённой станции, даты и времени с целью дальнейшего анализа временных рядов; Графическое представление результатов с возможностью масштабирования и детального рассмотрения выделенных компонент.

План работы на весенний семестр

1. Провести сравнительный анализ с методами детектирования, которые используют схожий алгоритм ITD;
2. Улучшение алгоритма вычисления коэффициентов для идентификации аномальных значений временного ряда;
3. Написание научной статьи по теме магистерской диссертации.

ЗАКЛЮЧЕНИЕ

В ходе данной научно-исследовательской работы были разработаны ключевые методы для работы с данными нейтронных мониторов. Был проведен обзор алгоритмов декомпозиции исходного сигнала, на основе которого был выбран алгоритм ITD для дальнейшего детектирования аномальных значений временных рядов. Выбранный метод был успешно реализован. Код был загружен в удалённый репозиторий GitHub [9].

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Мирошниченко Л. И. Физика Солнца и солнечно-земных связей: учебное пособие // Л. И. Мирошниченко; под ред. М. И. Панасюка. — М.: Университетская книга, 2011, 345 с.
2. Блейхут, Р. Быстрые алгоритмы цифровой обработки сигналов / Р. Блейхут – Пер. с англ. – М.: Мир, 1989. – 448 с.
3. Франсуа Шолле Глубокое обучение на Python — М.: Питер, 2018, 400 с.
4. Voznesensky, A. Adaptive Signal Processing Algorithms Based on EMD and ITD / A. Voznesensky, D. Kaplun // IEEE Access – 2019. – vol. 7 – P. 171313-171321.
5. Ланнэ, А.А. Основы цифровой обработки сигналов: Учеб. пособие. / А. А. Ланнэ, Б.Д. Матюшкин, Д.А. Улахович – СПб.: ГУТ. 1998.
6. Пьюривал С. Основы разработки веб-приложений – М.: Питер, 2015, 272 с.
7. Николас Д.Х, Сюзанна С. Статистические методы в журнале// Nicholas J.H., Suzanne S. // IEEE Access – 2005. – P. 16267336
8. Сергиенко, А.Б. Цифровая обработка сигналов, 3-е изд.: Учебное пособие / А.Б. Сергиенко – СПб.: БХВ – Петербург, 2011 – 768 с.
9. Удалённый репозиторий // GitHub. URL: <https://github.com/whyusoserious/nemo> (дата обращения: 25.04.2023).
10. База данных нейтронных мониторов // NMDB. URL: <https://nmdb.eu/nest> (дата обращения: 20.04.2023).

ПРИЛОЖЕНИЕ А

ОТЗЫВ НАУЧНОГО РУКОВОДИТЕЛЯ

ОТЗЫВ

на научно-исследовательскую работу
на тему: «Разработка метода и алгоритмов детектирования аномалий в
природных данных»
студента Санкт-Петербургского государственного
электротехнического университета
«ЛЭТИ» им. В.И. Ульянова (Ленина)

Щербакова И.Б.

Щербаков И.Б. успешно выполнил поставленный план на осенний семестр. Щербаков И.Б. реализовал модули, определяющие считывание данных с веб-сервиса, содержащего сведения с нейтронных мониторов, изучил метод Внутренней временной декомпозиции (ITD), провел его сравнительный анализ с другими методами обработки и анализа нестационарных сигналов, программного реализовал метод в среде MatLab, провел эксперименты. Полученные оценки показали эффективность применения метода ITD к природным данным.

В ходе практики Щербаков И.Б. подтвердил высокий уровень знаний и навыков, приобретенных в процессе обучения. А также высокий уровень заинтересованности в работе и самостоятельности, внимательное отношение к замечаниям и рекомендациям руководителя.

Считаю, что научно-исследовательская работа Щербакова И.Б. полностью соответствует предъявляемым требованиям и выданному заданию, заслуживает оценки «отлично».

Руководитель: _____

Мандрикова

Мандрикова Б.С.