**ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ**

**НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**

**«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

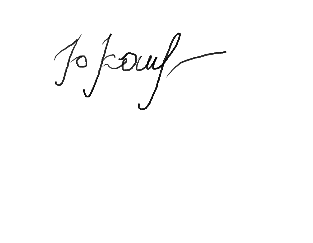
Факультет компьютерных наук

Образовательная программа «Прикладная математика и информатика»

УДК 004

**Отчет об исследовательском проекте**

на тему «Следы солнечной активности в температурных рядах данных»

**Выполнил**:

студент группы БПМИ188 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Кошелева Юлия Михайловна

Подпись И.О. Фамилия

\_\_\_\_\_\_\_\_\_\_23.06.2020\_\_\_\_\_\_\_\_

Дата

**Согласовано**:

руководитель проекта \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Имя, Отчество, Фамилия

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Должность / Место работы

Дата \_\_\_\_\_\_\_\_\_2020 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Оценка (по 10-тибалльной шкале) Подпись

**Москва 2020**

Аннотация

Цель исследования – найти схожие периодичности во временных рядах данных, характеризующих солнечную активность, и рядах, характеризующих уровень воды в океане. В работе рассматривается такой метод обработки временных рядов, как спектральный сингулярный анализ. В работе показаны результаты применения анализа на искусственных данных, помимо этого, показано, что с помощью анализа можно сказать по настоящим природным данным. В результате в рядах данных были выявлены несколько схожих гармоник.

Содержание

[Введение 4](#_Toc43834264)

[Часть 1. Ознакомление с методом 5](#_Toc43834265)

[Описание метода 5](#_Toc43834266)

[Исследование анализа сингулярного спектра 6](#_Toc43834267)

[Вывод 11](#_Toc43834268)

[Часть 2. Работа с временными рядами 12](#_Toc43834269)

[Данные солнечной активности 12](#_Toc43834270)

[Уровень моря 13](#_Toc43834271)

[Результаты сравнения 15](#_Toc43834272)

[Вывод 16](#_Toc43834273)

[Часть 3. Вариация длины окна 17](#_Toc43834274)

[Рассмотрение различных значений для длины окна 17](#_Toc43834275)

[Результат 20](#_Toc43834276)

[Заключение 22](#_Toc43834277)

[Список используемых источников 23](#_Toc43834278)

[Приложения 24](#_Toc43834279)

# Введение

В настоящее время существует множество способов и методов для анализа и прогнозирования временных рядов, с которыми люди сталкиваются постоянно. Например, современная проблема влияния солнечной активности на температуру на Земле является одной из важнейших. Мир озадачен климатическими изменениями, факторами, которые на это влияют, и прогнозированием дальнейших перемен. Для выявления зависимости между температурными изменениями и солнечной активностью используют так называемый анализ сингулярного спектра (Singular spectrum analysis).

Этот анализ базируется на преобразовании одномерного временного ряда в многомерный ряд с последующим применением к полученному многомерному временному ряду [метода главных компонент](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%B3%D0%BB%D0%B0%D0%B2%D0%BD%D1%8B%D1%85_%D0%BA%D0%BE%D0%BC%D0%BF%D0%BE%D0%BD%D0%B5%D0%BD%D1%82). Способ преобразования одномерного ряда представляет собой «свёртку». Для этой свёртки необходимо определить длину «окна». Чаще всего её выбор обусловлен признаками периодичности в исходных данных.

Анализ сингулярного спектра хорош тем, что подходит для решения множества задач, таких как нахождения общего тренда в ряде данных, подавление шума, выявление периодичностей, сглаживание. В связи с этим он широко применяется в климатологии, океанологии, геофизике, обработке изображений и т.д.

Выбрав тему, я поставила перед собой цель: исследовать ряды, характеризующие уровень воды в океане, и временные ряды, характеризующие солнечную активность, на зависимость посредством анализа сингулярного спектра. Чтобы приблизиться к этой цели мне нужно выполнить следующие задачи:

1. Ознакомиться с методом SSA-разложения

2. Рассмотреть искусственно полученную и применение к ней SSA

3. Применить анализ к данным солнечной активности и данным океанической активности

4. Найти схожие гармоники в разложениях

В первой части работы предметом моего исследования выступит влияние солнечной активности на земные процессы, в частности на уровень мирового океана, а объектом - SSA-разложение временных данных.

# Часть 1. Ознакомление с методом

### Описание метода

Метод обработки данных можно разделить на три последовательных этапа.

1. Построение траекторной матрицы.

Для построения траекторной матрицы нужно установить длину окна L. Далее нужно построить матрицу размера KхL, где каждый столбец под номером *i* будет соответствовать исходным данным с индексами от *i* до *i + L – 1*.

1. Сингулярное разложение

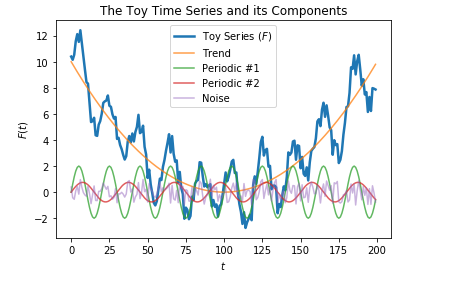
Далее к полученной траекторной матрице применяется метод сингулярного разложения. Среди полученных данных значения вектора *σ* будут сингулярными значениями компонент.

1. Получение компонент

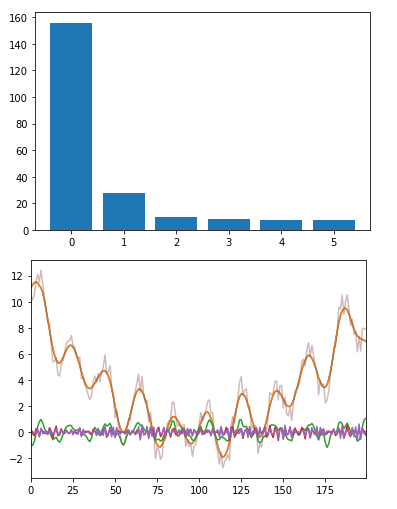
Далее к элементарным матрицам применяется усреднение по анти-диагоналям (таким образом получаем Ганкелевы матрицы). Получаем разложение исходного временного ряда данных на компоненты.

### Исследование анализа сингулярного спектра

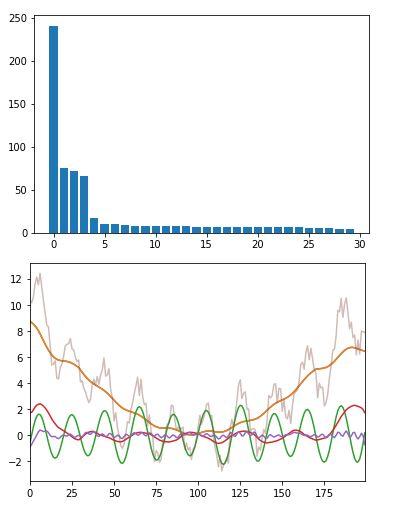
1. Изначально я построила функцию для дальнейшей работы. Эта функция складывалась из тренда(параболы), двух синусоид с разными коэффициентами и периодами и шума (значения этой компоненты генерировались рандомно). Функция стала суммой всех компонент.



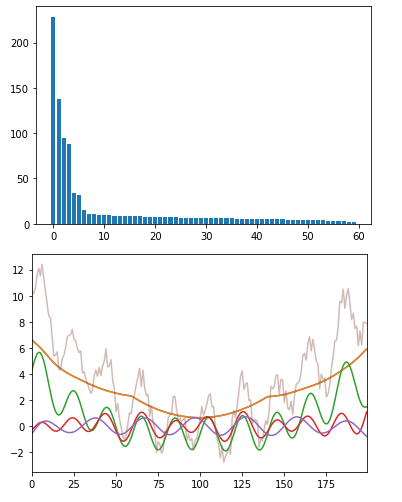
1. Далее к полученной функции несколько раз был применен SSA, но с переменной длиной окна. Соответственно, менялись и полученные компоненты. Параллельно для каждого разложения я строила график сингулярных значений. Сингулярное значение компоненты должно показывать «вклад» этой компоненты в исходную функцию.



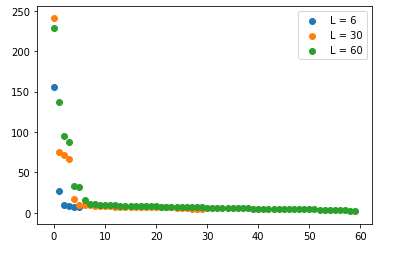
(разложение с длиной окна 6)



(разложение с длиной окна 20)



(разложение с длиной окна 60)



(График для сравнения сингулярных значений при смене длины окна)

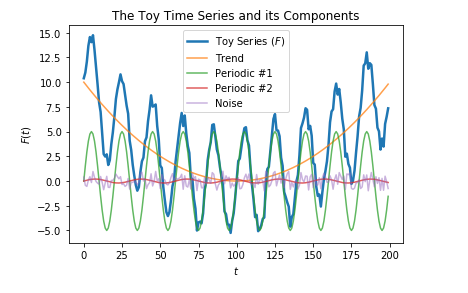
Вывод: SSA позволяет достаточно точно восстановить компоненты сигнала). Наилучший вариант разложения на компоненты достигается, когда характерные особенности функции попадают внутрь окна. В нашем примере это происходит, когда размер окна соизмерим с обоими периодами.

При длине окна равной 6 периодичности функции, которую создают синусоиды, не просматривается. В этом случае разложение имеет всего одну наиболее выделяющуюся компоненту, которая наиболее близка к самой функции, но не отображает ни тренд, ни периодичность, ни шум. При длине окна 60 на графике сингулярных значений появляется множество компонент, которые соизмеримы друг с другом, появляется множество синусоид, которые не несут в себе смысла и не относятся к изначальному разложению функции.

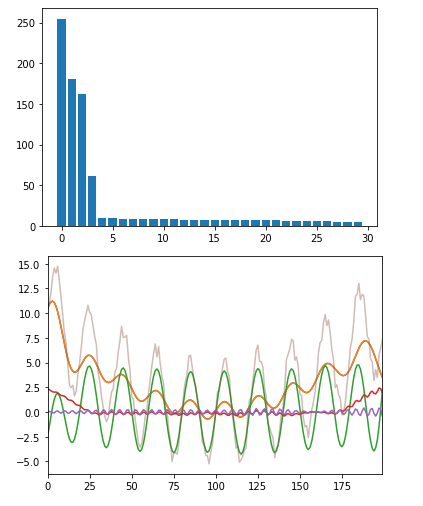
Наиболее точно восстановить исходные компоненты получается при длине окна 30. Также на графике сингулярных значений можно увидеть 4 наиболее выделяющихся компоненты. Это и есть тренд, синусоиды и шум.

Для дальнейшей работы я использовала L = 30, так как сочла эту длину окна оптимальной и наиболее точной для разложения на компоненты этой функции.

1. Далее я попробовала поменять коэффициенты синусоид для того, чтобы посмотреть, как изменятся сингулярные значения компонент при этом. Я увеличила коэффициент одной синусоиды с 2 до 5, и уменьшила у второй с 0.75 до 0.2.



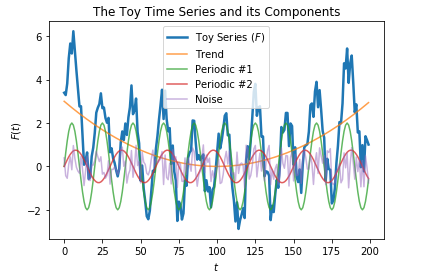
(график описанной функции и компонент, которые ее составляют)



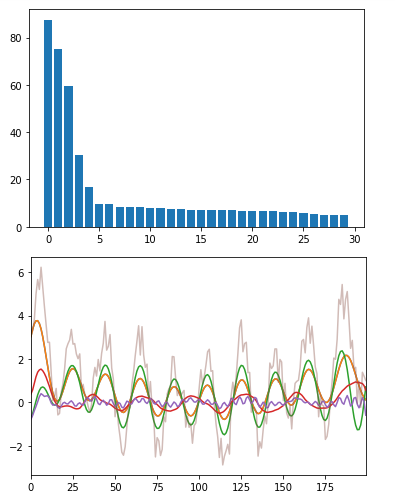
(график сингулярных значений и график компоненты после разложения)

Вывод: так как влияние одной синусоиды на функцию возросло, это отображается на графике сингулярных значений. Тренд на нем уже не выделяется так явно. Компонента тренда и компоненты синусоид имеют сравнимые сингулярные значения. Этот эффект прослеживается и на графике самой функции. На ней более заметны колебания, чем основной тренд.

1. Теперь я изменила коэффициент параболы, которая задает тренд, тем самым сделала ее более пологой.



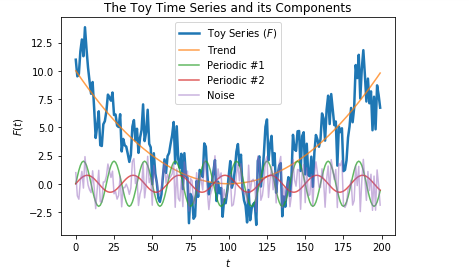
(График функции и составляющих ее компонент)

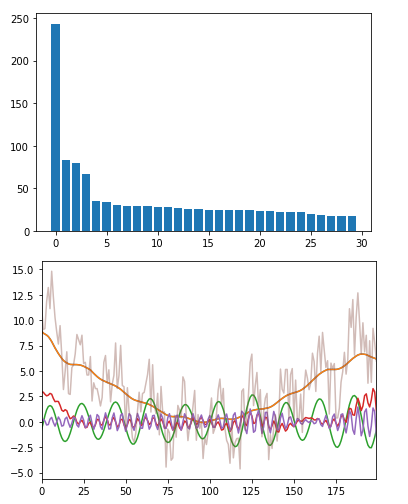


(график сингулярных значений и компонент после разложения)

Вывод: на графиках прослеживается эффект, схожий с пунктом 3. Парабола стала более пологой, соответственно на фоне горизонтальных синусоид она выделяется меньше. На графике сингулярных значений синусоиды наиболее приближены к тренду. Компоненты становятся соизмеримы друг с другом, что усложняет задачу анализу сингулярного спектра, так как выделение отдельных компонент из общей массы становится сложнее.

1. Увеличение коэффициента шума.





Вывод: шум заметно и значимо изменил вид функции, но на разложение это повлияло меньше, чем остальные преобразования. Тренд выделен четко и имеет существенное сингулярное значение, по сравнению с другими компонентами. В свою очередь шум приближен к синусоидам, но это также не мешает их выделению.

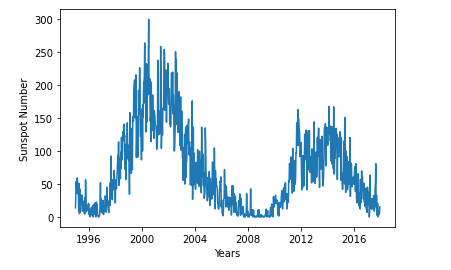
### Вывод

В ходе первых этапов работы (ознакомление с анализом сингулярного спектра) мной был изучен метод обработки временных рядов, активно применяемый на практике. Также путем построения искусственных данных, я выявила некоторые закономерности и зависимости между видом функции, подаваемой на вход SSA, и выходными данными анализа, а именно компонентами функции и их сингулярными значениями. В дальнейшем я смогу применить этот метод для поиска зависимостей между временными данными, характеризующими солнечную активность и характеризующими температуру на Земле.

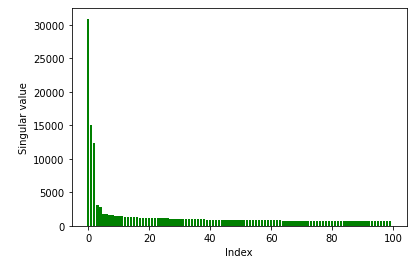
# Часть 2. Работа с временными рядами

Для анализа методом SSA и поиска закономерностей и периодичностей в данных я взяла промежуток времени с 1994 года по 2017 год.

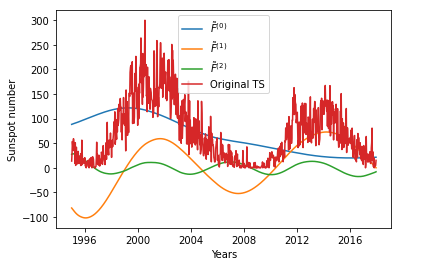
### Данные солнечной активности

Для начала рассмотрим данные, характеризующие солнечную активность. Как было сказано раннее, были использованы ежедневные данные. Так как источник данных уровня моря предоставлял измерения для промежутков длиной примерно в 10 дней, была выполнена предобработка ежедневных данных. В ходе предобработки был составлен новый массив со средними значениями за десятидневные промежутки времени. Таким образом данные солнечной активности представлены на рисунке 1.

(рис. 1)

К полученному ряду применяем метод SSA с окном, равным 11 годам:

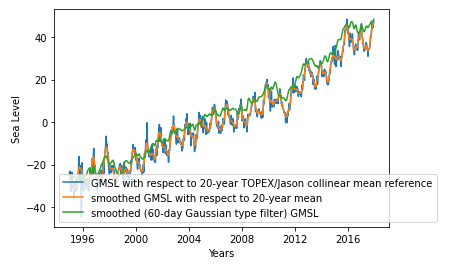
(рис. 2)

(рис. 3)

На графике сингулярных значений можно увидеть значимость первых трех компонент. Нулевая компонента является трендом для данного ряда. Вторая и третья имеют периодичности примерно 10–11 лет и 5-6 лет соответственно. Период в 11 лет соответствует современным представлениям о периодичности в рядах солнечной активности.

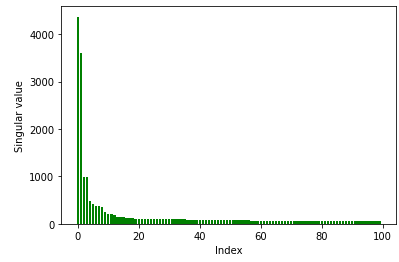
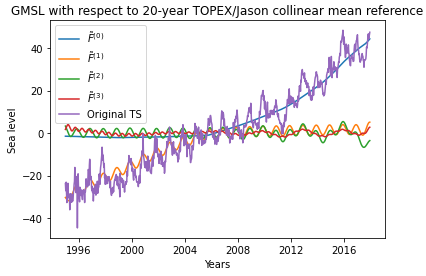
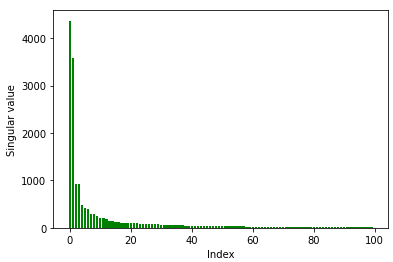
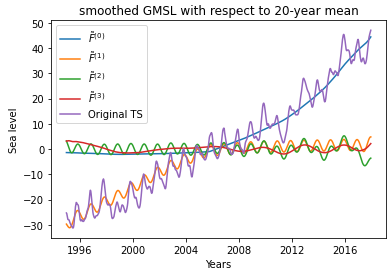
### Уровень моря

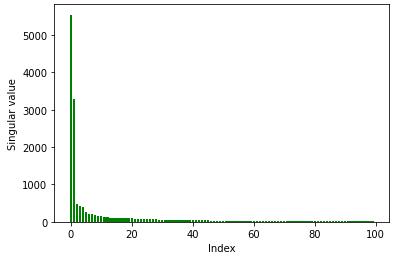
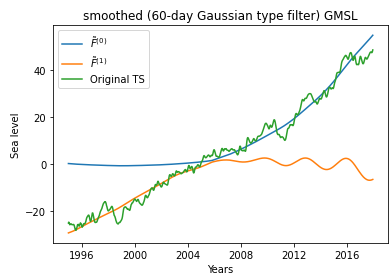
Теперь таким же образом рассмотрим временные ряды, характеризующие уровень моря. На одном графике отложим исходные данные, далее применим SSA ко всем данным с окном в 11 лет.

(рис. 4)

Показатели варьируются относительно друг друга, так как, как было описано выше, используются разные метода подсчета и усреднения данных.

При разложении первого типа данных на графике сингулярных значений стоит отметить, что первые четыре компоненты значительно больше остальных, исходя из этого в дальнейшем рассматриваются первые четыре компоненты разложения. Для второго типа подсчета уровня моря действия аналогичны. При рассмотрении третьего типа опять же по графику сингулярных значений можно заметить, что существенно отличны от шума исключительно первые две компоненты разложения.

(рис. 5)

(рис. 6)

(рис. 7)

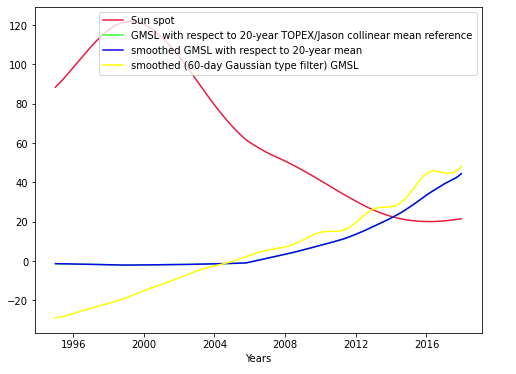
Глядя на графики компонент разложения, сложно увидеть в них определенную периодичность. Вторая и третья компоненты разложения говорят о периодичности примерно в один год. Более содержательную периодичность гармоник можно предположить во второй половине исследуемого временного промежутка. О такой периодичности свидетельствуют первая компонента разложения данных третьего типа и вторая компонента разложения данных первых двух типов. Но и эта гармоника имеет период не больше 3-4 лет. В дальнейшем к рядам был применен метод скользящего среднего для того, чтобы убрать вполне логичную и закономерную годовую периодичность.

При применении метода скользящего среднего, значимость компонент, показывающих периодичность сильно уменьшилась, но без этого метода эти компоненты показывают лишь периодичность длиной один год. Я могу предположить, что периодичность (имеется в виду не годовая) в случае рассмотрения уровня моря мала или даже незначительна. Возможно, она характеризуется большим периодом времени, чем взяты промежуток. Возможно, проблем в том, что ко всем типам данных так или иначе было применено сглаживание относительно двадцатилетнего среднего.

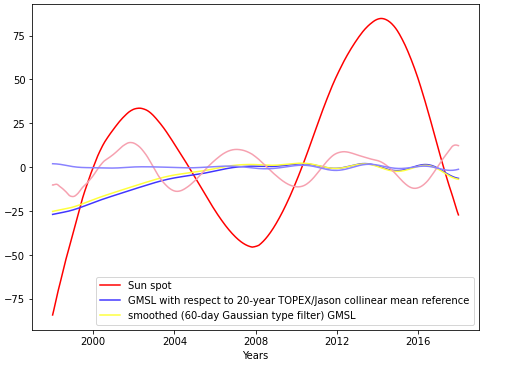
В процессе SSA возникла следующая проблема: данные уровня моря доступны и валидны только с 1993 года, соответственно для анализа доступен промежуток не более 26 лет, что сравнимо с длиной окна, которая равна 11 годам. Соответственно SSA работает некорректно. В дальнейшем, чтобы решить эту проблему, я попробую суммировать друг с другом нулевую и первую, третью и вторую компоненты.

### Результаты сравнения

Для начала сравним кривые трендов. У двух типов тренды совпали (это происходит, так как они представляют одни данные). Исходя из графика, я делаю вывод, что тренды не имеют между собой сходств. Общий тренд солнечной активности в заданный период стремится вниз в то время, как тренд уровня моря стремиться вниз, не имеет выраженных пиков, ни на одном участке не похож на тренд в данных солнечной активности.

(рис. 8)

Теперь перейдем к сравнению гармоник, входящих в данные. Собрав все гармонические компоненты на одном графике, делаю предположение, что схожих периодичностей в данных не прослеживается. В данных солнечной активности присутствует явная периодичность длиной 10-11 лет и менее явная периодичность приблизительно в два раза короче. В данных уровня моря периодичность прослеживается плохо, но даже если предположить, что она есть, ее можно выделить только во второй половине рассматриваемого временного отрезка и охарактеризовать периодом 3-4 года.

(рис. 9)

*(на данном графике разные цвета характеризуют разные данные, насыщенность оттенка характеризует порядок изображенной компоненты – чем ярче цвет, тем меньше номер компоненты в разложении)*

### Вывод

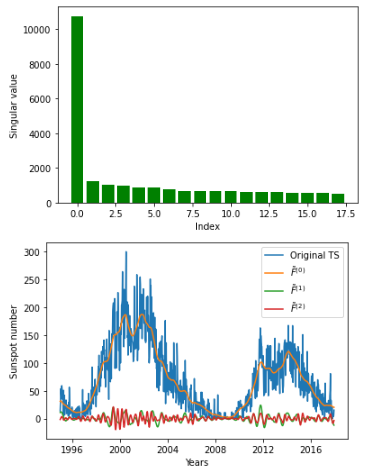
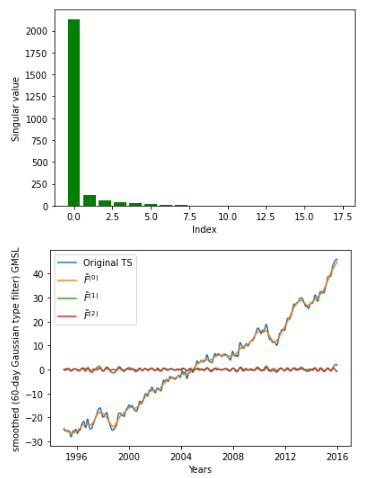
В процессе работы был проведен анализ двух временных рядов с помощью метода SSA и метода скользящего среднего. Были сопоставлены ряд, характеризующий солнечную активность, и ряд, характеризующий уровень моря на Земле за период времени с 1995 года по 2017. По итогам сравнения были сделаны выводы, что закономерности между рядами в описанных условиях прослеживаются плохо. И тренд, и периодические гармоники, входящие в функции, которые описывают представленные данные, имели расхожие характеристики, что говорит о малой связи между данными.

# Часть 3. Вариация длины окна

### Рассмотрение различных значений для длины окна

Так как длина окна в 11 лет не показала искомых результатов (что логично, так как временной промежуток предложенных данных мал, для такой длины окна), я попробовала исследовать другие значения, отличные от 11, для применения анализа сингулярного спектра.

Ранее уже была отмечена периодичность, соответствующая одному году, которая характерна для земных данных. Таким образом, далее будем обращать внимание преимущественно ряд, где годовая периодичность отсутствует (smoothed GMSL).

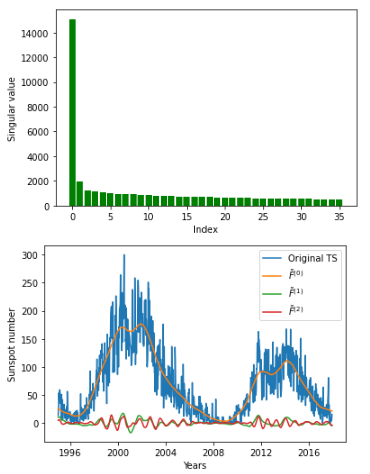
1. Длина окна – полгода.

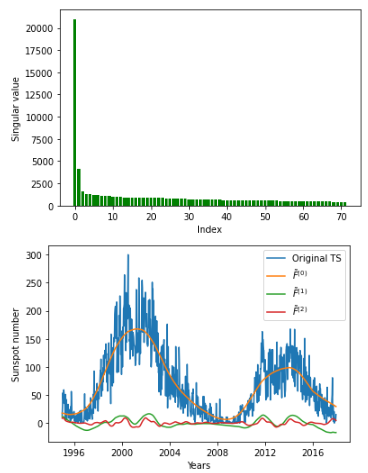
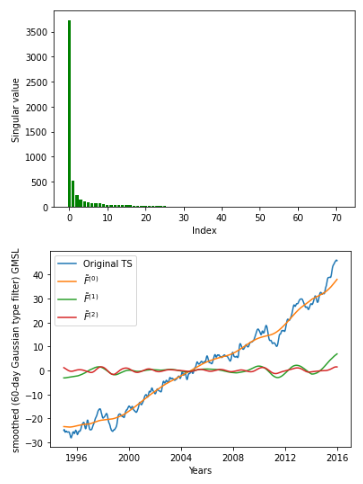
(рис. 3) (рис. 4)

На мой взгляд, период в полгода слишком короткий для рассматриваемых данных. На графике сингулярных значений первая компонента значительно больше остальных, другие компоненты выделить нельзя, поэтому я не могу ничего сказать о периодичности. Кроме того, на графиках компонент разложения первая компонента явно описывает тренд, остальные же ничего не говорят о данных.

1. Длина окна – 1 года

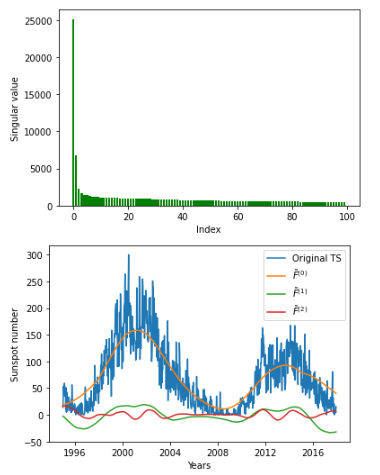
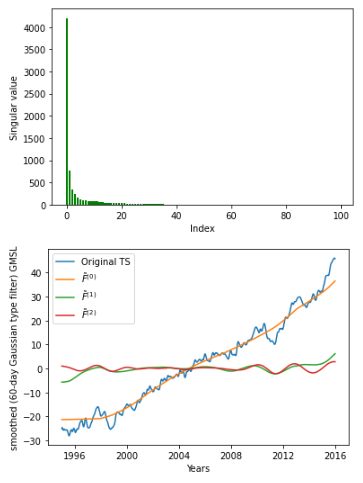
Я не рассматриваю годовой размер окна для рядов океанической активности, т.к. мы итак можем сказать, что такая периодичность присутствует в них. В рядах солнечной активности тоже можно проследить годовую периодичность в пиках активности (с 1999 года по 2004 год, с 2012 года по 2016 год). На графике сингулярных значений также видно отличие значения второй компоненты от последующих.

(рис. 5)

1. Длина окна – 2 года

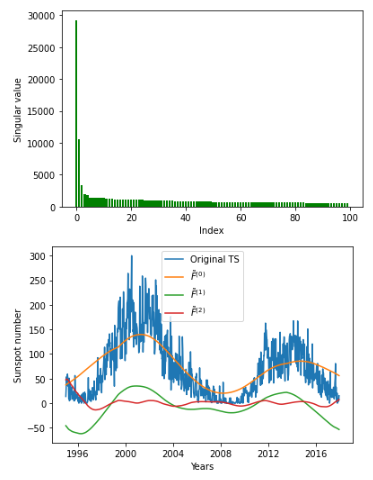
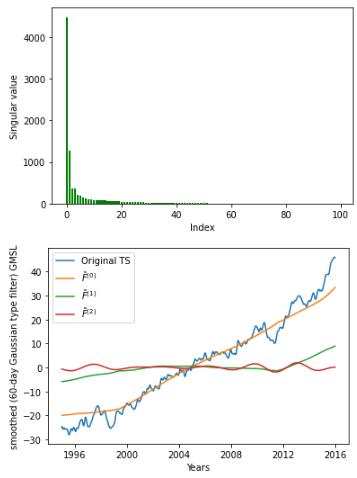
(рис. 6) (рис. 7)

При двухгодовой длине окна мы можем наблюдать значительный рост сингулярного значения второй компоненты по сравнению с предыдущими. Помимо этого, сопоставив два графика компонент (солнечной и океанической активности), можно сделать предположение о схожести второй и третьей компоненты разложения. Хотя в данных, характеризующих уровень воды в океане, нет явных периодов большей или меньше активности (значения просто возрастают), тем не менее мы можем наблюдать во второй и третьей компоненте следы двухгодовой периодичности. Эти колебания соответствуют примерно 2000 году и 2012. Аналогичные колебания прослеживаются во второй и третьей компоненте разложения ряда солнечной активности (соответственно около 2000 и 2012 годов).

1. Длина окна – 3 года

(рис. 8) (рис. 9)

При таком разложении в данных океанической активности не наблюдается чего-то нового. В данных, характеризующих солнечную активность появляется намёк на периодичность в 4.5 года (Если смотреть на вторую компоненту разложения). Поэтому дальше рассмотрим соответствующий размер окна.

1. Длина окна – 4.5 года

(рис. 10) (рис. 11)

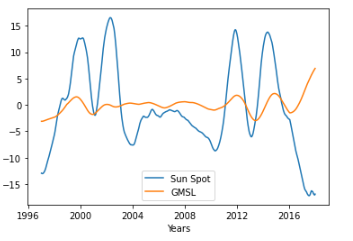
Могу предположить, что при таком разложении вторая компонента ряда солнечной активности близка к описанию известной одиннадцатилетней периодичности. В данных уровня океана подобного не наблюдается.

*Вывод:* вариация длины окна позволила нам выдвинуть несколько предположений. В первую очередь это предположение о двухлетней периодичности, которая прослеживается во всех рассматриваемых рядах, в периодах, соответствующих особо высокой активности Солнца. Помимо этого, мы заметили годовую периодичность в ряде солнечной активности и не нашли периодичности большей длины в рядах уровня океана, возможно, это связано с небольшим размером ряда данных.

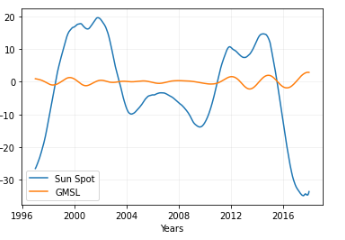
### Результат

В отдельности рассмотрим найденные периодичности для L = 2 года и L = 4 года.

1. L = 2 года



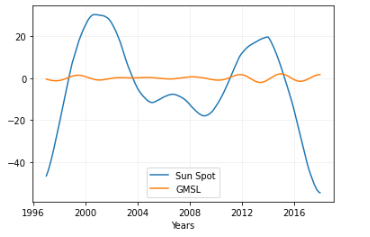
На данном графике представлены компоненты разложения, которые показывают гармонику, которая присутствует в рядах. Можно заметить, что хоть амплитуда колебаний и достаточно сильно различается, период колебаний схож и равен примерно двум - трём годам. Интересным я посчитала то, что этот период прослеживается не на всем исследуемом промежутке. В обоих случаях период прослеживается в промежутках от 1998 года до 2002 и от 2010 до 2014. Стоит заметить, что именно эти промежутки времени характерны наибольшей активности Солнца за весь рассматриваемый период. Таким образом, можно сказать, что на данном графике отражены следы солнечной активности в ряде океанической активности.

1. L = 3 года

При длине окна в 3 года результат меняется незначительно. Точно также прослеживается два периода, когда явно выражен период гармоник в 2-3 года.

1. L = 4 года

При такой длине окна период гармоники в ряде солнечной активности близится к четырем годам. Из графика ниже можно сделать вывод, что такая периодичность действительно присутствует в данных. Однако, подобного нельзя увидеть в ряде, характеризующем уровень воды в океане. Поэтому в данном случае мне сложно что-то сказать о взаимосвязи рядов.



# Заключение

В заключение стоит сказать, что в процессе выполнения работы мной были выполнены все поставленные задачи. Я ознакомилась с анализом сингулярного спектра, а также использовала его на практике для выявления схожих гармоник в природных данных. В отчете приведены выводы по каждой конкретной части работы. Стоит сказать, что анализ помог мне выявить гармоники с одинаковым периодом, но разной амплитудой в ряде данных солнечной активности и ряде, который характеризует уровень воды в океане. Я заметила, что эти ряды объединяет период в два – три года, который характерен временным промежуткам, когда Солнце проявляет наибольшую активность. Я сделала вывод, что такой результат может говорить о влиянии одного явление на другое в определенные периоды. Цель моего проекта выполнена.

# Список используемых источников

1. Интернет-источник

<https://www.kaggle.com/jdarcy/introducing-ssa-for-time-series-decomposition> - реализация метода, пример работы на искусственной функции

1. Интернет-источник <https://ru.wikipedia.org/wiki/SSA_(%D0%BC%D0%B5%D1%82%D0%BE%D0%B4)> – статья на википедии
2. Интернет-источник

<https://github.com/aj-cloete/pssa> - репозиторий с реализацией и SSA и примерами

1. Интернет-источник

<https://sealevel.nasa.gov/understanding-sea-level/key-indicators/global-mean-sea-level/> - данные уровня моря

1. Интернет-источник

<http://sidc.be/silso/eisninfo> - данные солнечной активности

1. Интернет-источник

<https://ru.wikipedia.org/wiki/%D0%A3%D1%80%D0%BE%D0%B2%D0%B5%D0%BD%D1%8C_%D0%BC%D0%BE%D1%80%D1%8F> – статья на википедии

# Приложения

1. Алгоритм проведения вычислений можно найти по ссылке: <https://github.com/Julykosheleva/SSA>