Практическая работа №6.
Сглаживание временных рядов

Постановка задачи

Имеется два файла, которые содержат разные временные ряды. В первом файле находится ряд с синусоидальным трендом. Во втором – с линейным.

Необходимо выделить трендовую составляющую из данных используя 4 метода:

1. Простое скользящее среднее (SMA).
2. Взвешенное скользящее среднее (WMA).
3. Экспоненциальное сглаживание (EMA).
4. Двойное экспоненциальное сглаживание (DEMA).

Каждый метод требует подбора некоторых параметров:

1. SMA – размер окна. WMA – размер окна. EMA – параметр сглаживания α. DEMA – параметр сглаживания вокруг тренда α и параметр сглаживания самого тренда γ.
2. Для весов в WMA использовать экспоненциальную весовую функцию с $\varepsilon = 0{,}3$.
3. Необходимо подобрать оптимальные значения соответствующих параметров, используя Q-статистику Льюнг-Бокса при lag = 5. Оптимальными параметрами будем считать те, что минимизируют приведенную статистику.
4. В качестве размеров окна $w = 2 * m + 1$ перебрать значения m = 3, 5, 7, 9. В качестве параметров сглаживания: α, γ = 0.1, 0.2, …, 0.9. Обратите внимание, что метод DEMA двухпараметрический, что требует выбрать оптимальную комбинацию сразу двух параметров α и γ.

5. После подбора оптимальных параметров провести тест Дарбина-Уотсона (m = 1, α = 0.95) на данных после исключения выделенного тренда для каждого метода и каждого ряда.
6. В отчете изобразить графики исходных данных, графики трендов при оптимальных параметрах у каждого метода для каждого ряда, расчетные формулы, а также результаты тестов Дарбина-Уотсона.

Ход работы

Рассматриваются следующие общие модели временных рядов:

1. Аддитивная:

$$x_t = T + S + C + E$$

2. Мультипликативная:

$$x_t = T * S * C * E$$

где T – тренд;

S – сезонность;

C – циклическая составляющая;

E – случайная составляющая.

Тренд

В большинстве задач анализа временных рядов под трендом понимается долгосрочная тенденция уровней ряда или, иначе говоря, рост или убывание среднего значения уровней на длинном промежутке времени.

Циклическая составляющая

Долгосрочные периодические колебания относительно тренда.

Сезонность

Краткосрочные периодические колебания уровней временного ряда относительно тренда и циклической составляющей.

Случайная составляющая

Некоторая непредсказуемая случайная добавка.

Метод простого скользящего среднего (Simple Moving Average, SMA)

Для расчета взвешенного скользящего среднего используется следующее выражение:

$$\widetilde{y_t} = \sum_{i=-m}^{m} \omega_i y_{t+i}$$

где y_t – исходный временной ряд;

$\widetilde{y_t}$ – сглаженный временной ряд;

ω_i – весовые коэффициенты;

m – количество членов ряда в сумме по одной стороне от центрального значения, размер окна сглаживания, который необходимо подобрать.

Выбор весовых коэффициентов ω_i лежит на плечах исследователя. Особый выбор этих коэффициентов порождает новые методы сглаживания.

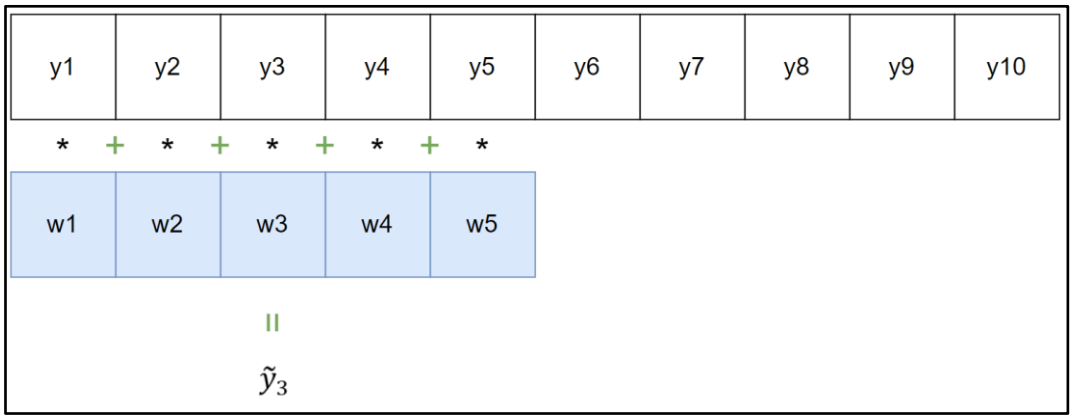
Простое скользящее среднее – это частный случай взвешенного с равными весовыми коэффициентами:

$$\omega_i = \frac{1}{2m+1}$$

Сумма весовых коэффициентов должна быть равна 1:

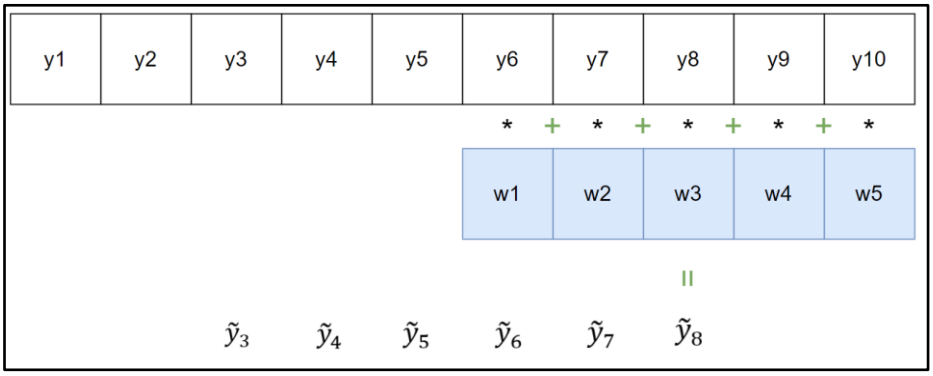
$$\sum \omega_i = 1$$

Существует понятная проблема с крайними членами ряда. Представим ситуацию скользящего среднего для ряда данных y_t с окном с $m = 2$. Изобразим на рисунке исходный ряд и окно сглаживания в виде массива, между которыми происходит операция свёртки:



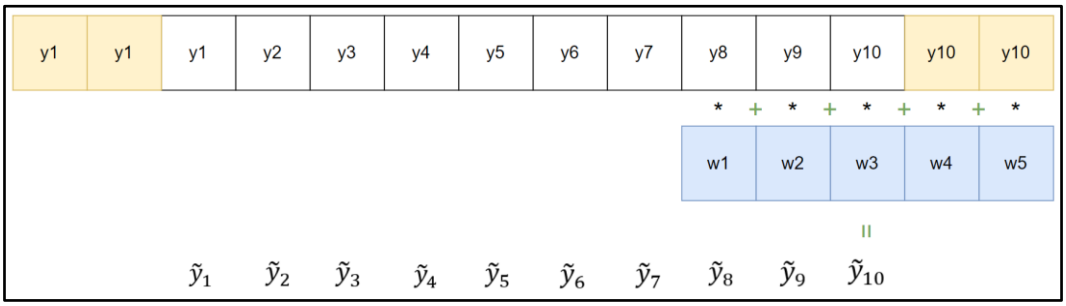
Как показано на рисунке выше, сглаживание при помощи операции свертки с весовым окном не дает в классическом виде возможность получить в сглаженном ряде данных такое же количество элементов массива. Для проведения операции такой свёртки, окну необходимы крайние элементы, а центральный отображается в сглаженный первый элемент. Другими словами, чем больше окно сглаживания у нас будет, тем большее количество членов ряду будет отниматься в результате сглаживания.

Полная картина такого сглаживания:



По изображению выше видно, что с каждой стороны ряда удаляются по $m = 2$ значений. Такую проблему в обработке изображений решают с помощью операции отступа (padding) на заданное количество единиц измерений с каждой стороны.

Мы можем добавить отступ с каждой стороны исходного ряда, чтобы нивелировать эффект сжатия ряда. Данный отступ необходимо делать первым значением в начале ряда и последним значением в конце ряда соответственно по m раз с каждой стороны, чтобы первые элементы сглаженного ряда не были занижены по уровню значений по сравнению с исходным рядом:



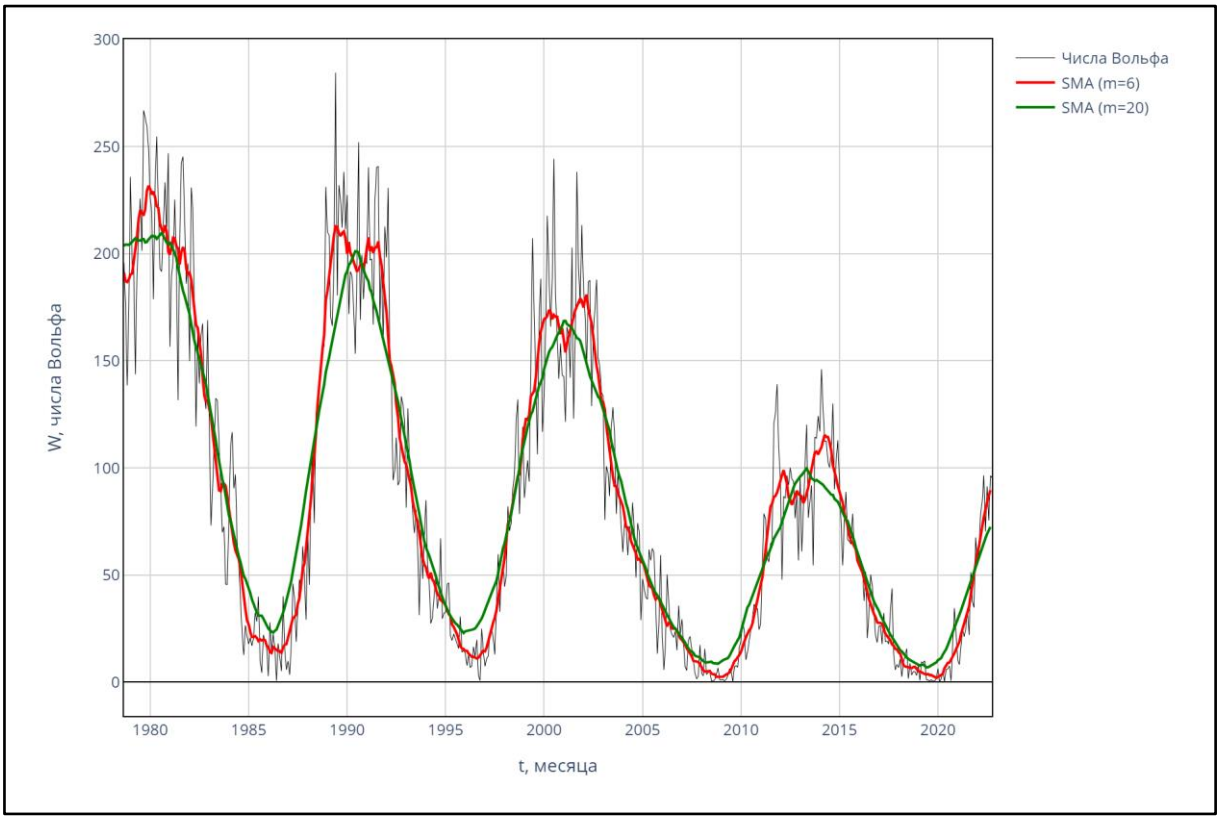
$$\tilde{y}_1 = y_1 \cdot (w_1 + w_2 + w_3) + y_2 \cdot w_4 + y_3 \cdot w_5$$

$$\tilde{y}_n = y_n \cdot (w_3 + w_4 + w_5) + y_{n-1} \cdot w_2 + y_{n-2} \cdot w_1$$

Стоит сказать, что при сравнимо больших значениях окна с размерами ряда, метод отступа будет порождать большие ошибки сглаживания, так как будет вносить нетипичные значения ряда. Так что злоупотреблять данной техникой не стоит.

Таким образом, получаем алгоритм простого скользящего среднего для сглаживания временного ряда с заданным размером окна m.

Продемонстрируем его работу на массиве данных среднемесячной динамики солнечной активности, которая определяется числами Вольфа. Построим график исходного временного ряда и графики его сглаживания по методу SMA с размером окна $m = 6$, $w = 2 \cdot m + 1 = 13$ и $m = 20$, $w = 2 \cdot m + 1 = 41$.



На графике видно, как сглаженная реализация все еще имеет в себе циклическую компоненту относительно тренда.

Познакомившись с простым методом скользящего среднего, далее переходим к взвешенному скользящему среднему.

Метод взвешенного скользящего среднего (Weighted Moving Average, WMA)

Метод взвешенного скользящего среднего работает идентично методу простого скользящего среднего, но необходимо определять саму весовую функцию метода сглаживания.

Весовая функция − метод определения значений весов исходя из определенного правила отображения номера элемента окна в его значение.

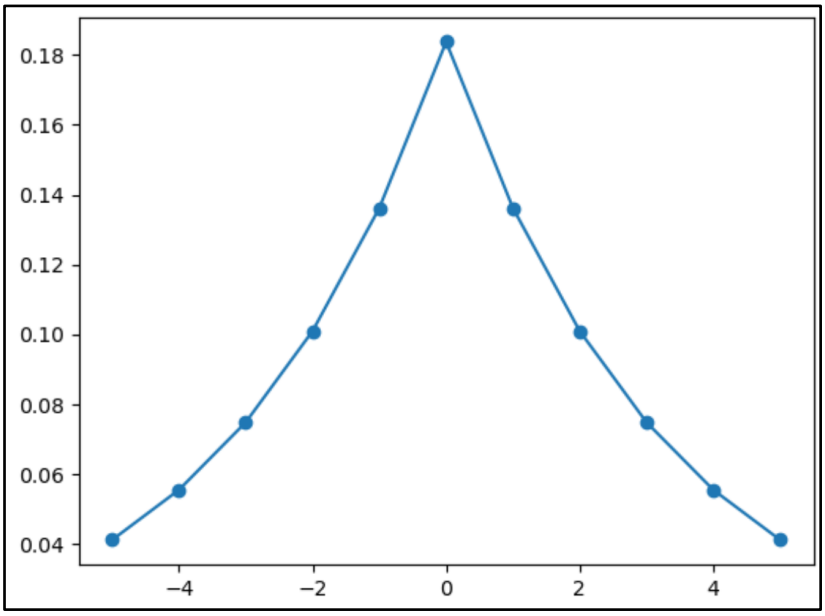
Весовых функций на практике встречается довольно много, и все их можно применять.

Из рисунка выше видно, что в методе простого скользящего среднего используется **П-окно**. В данной практике вам предлагается использовать экспоненциальную весовую функцию, также именуемую как **Пуассоновское сглаживание**:

$$\omega_i = \frac{e^{-\varepsilon \cdot |i|}}{\sum_{j=-m}^{m} e^{-\varepsilon \cdot |i|}}$$

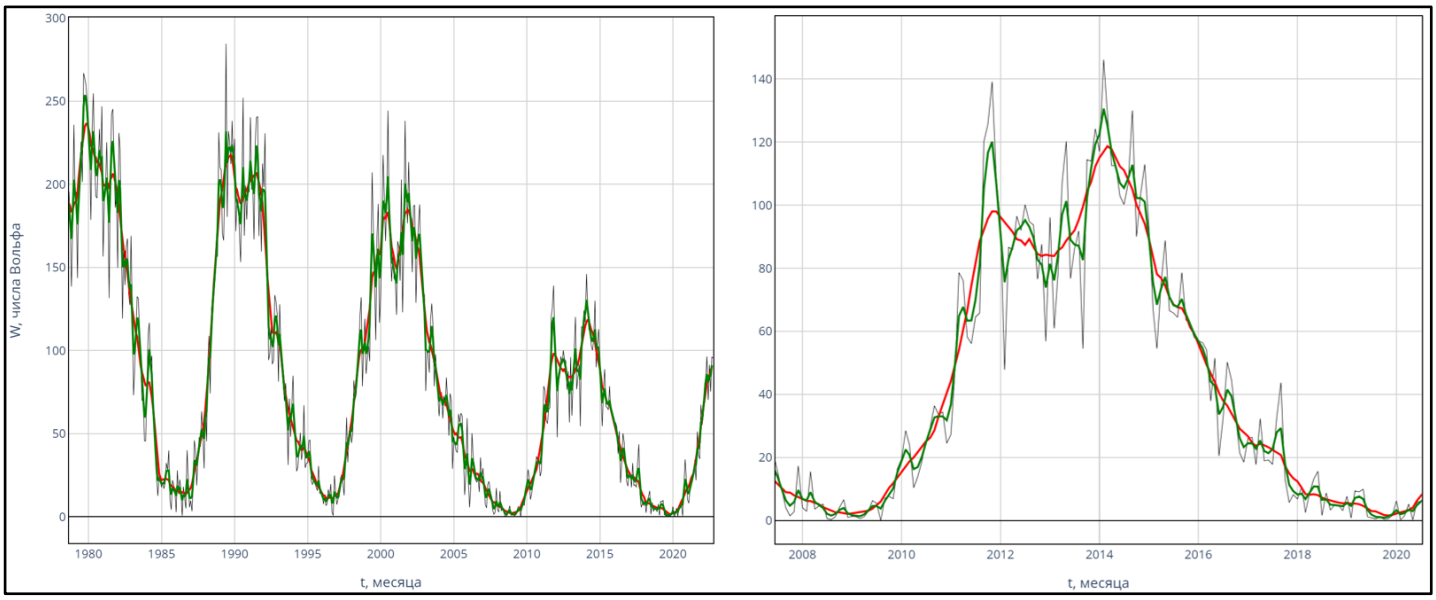
$$i = -m, (-m+1), \dots, m$$

$$\varepsilon = 0.3$$



Точками по оси ординат можно отследить значение веса в окне сглаживания алгоритма взвешенного скользящего среднего.

Далее пример работы данного алгоритма сглаживания:



Метод экспоненциального сглаживания (EMA). (Moving Average, EMA)

Известностью пользуется экспоненциальное сглаживание, в основе которого лежит расчет экспоненциальных средних. Целью такого сглаживания является передача большего веса последним уровням ряда, и меньшего веса более ранним.

Экспоненциальная средняя рассчитывается по рекуррентной формуле:

$$\tilde{y}_t = \alpha y_t + (1-\alpha)\tilde{y}_{t-1}$$

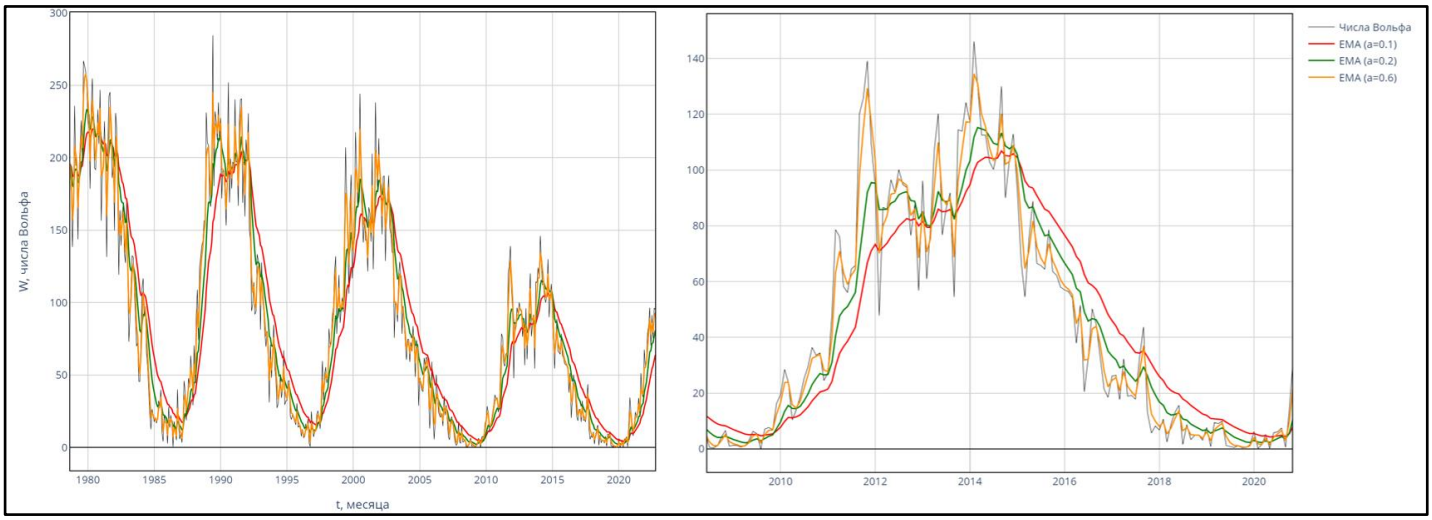
$$\tilde{y}_1 = y_1$$

где α − коэффициент сглаживания (сила сглаживания) принимает значения в диапазоне от 0 до 1.

При таком сглаживании усреднение ведется не по окну фиксированного размера, а по всему ряду от начала до текущего момента, при этом веса, с которыми учитываются давние измерения убывают экспоненциально

Коэффициент α влияет на степень восприятия истории. Чем ниже значение коэффициента, тем сильнее происходит сглаживание. Поскольку величины α и $(1-\alpha)$ взаимнообратные, следовательно смысл коэффициента α разнится с точностью до смены места их расстановки в зависимости.

Посмотрим работу алгоритма экспоненциального сглаживания на значениях ряда динамики чисел Вольфа для $\alpha = 0.1, 0.2$ и 0.6.



Метод двойного экспоненциального сглаживания (DEMA).

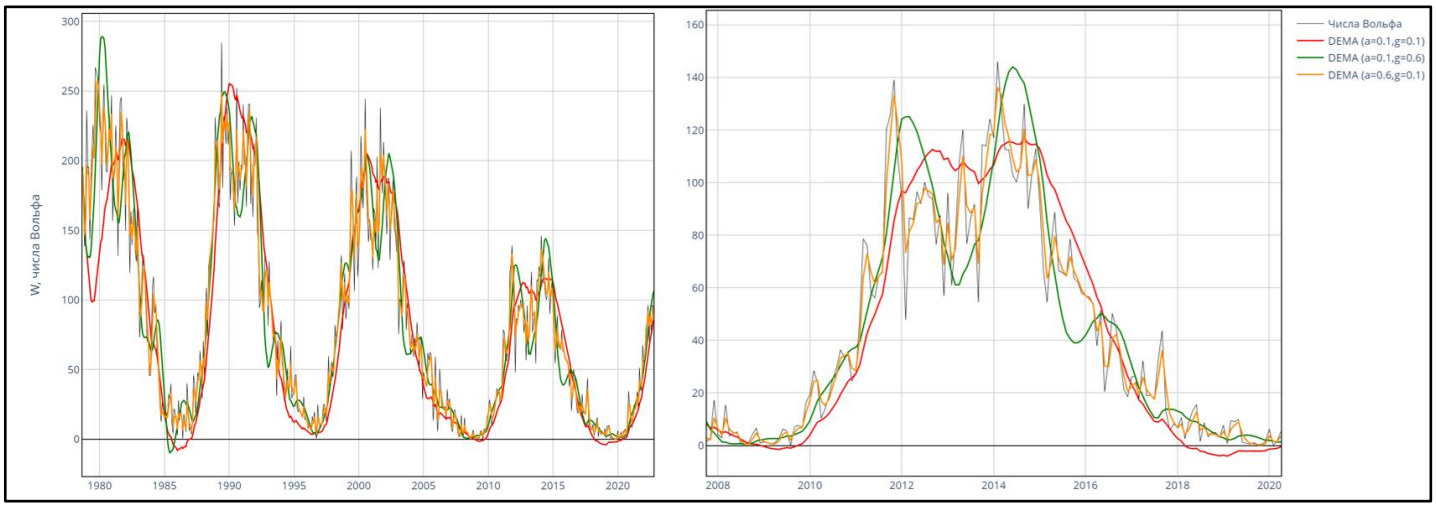
Двойное экспоненциальное сглаживание осуществляется по следующим формулам с коэффициентами α и γ, варьирующихся в пределах от 0 до 1:

$$\tilde{y}_t = \alpha \cdot y_t + (1-\alpha) \cdot (\tilde{y}_{t-1} + b_{t-1})$$

$$b_t = \gamma \cdot (\tilde{y}_t - \tilde{y}_{t-1}) + (1-\gamma) \cdot b_{t-1}$$

$$\tilde{y}_1 = y_1 \qquad b_1 = y_2 - y_1$$

Первый α отвечает за сглаживание ряда вокруг тренда, второй γ – за сглаживание самого тренда. Чем выше значения, тем больший вес будет отдаваться последним наблюдениям и тем менее сглаженным окажется модельный ряд. Комбинации параметров могут выдавать достаточно причудливые результаты.



Q-статистика Льюнг-Бокса

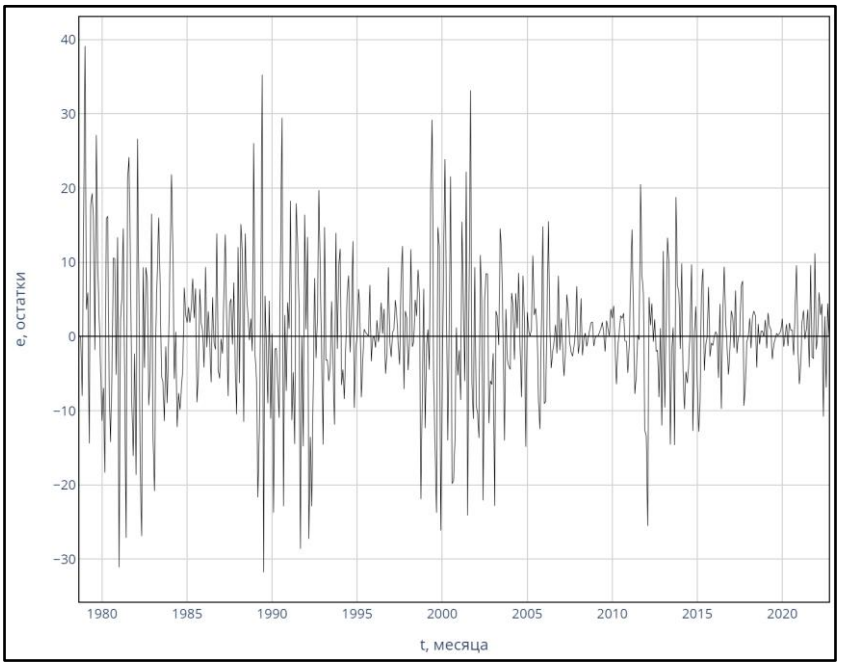
Временной ряд можно сглаживать по нескольким причинам, среди которых две очевидные:

1. Избавиться от случайных значений ряда вокруг тренда и сезонности, чтобы выделить их более значительно.
2. Выделить тренд, сгладив и случайные компоненты и сезонность вокруг тренда по известному количеству дней периода сезонной компоненты.

В каждом из выделенных случаев, проверка адекватности алгоритма сглаживания проводится по-разному. Например, можно оценить качество сглаживания, исследовав случайность остатков.

Для начала вычисляются остатки ряда данных:

$$e_t = x_t - \tilde{x}_t$$



Для второго случая, когда необходимо просто выделить тренд из данных достаточно проверить остатки на нулевое среднее. Если среднее остатков не находится около нуля, то тренд найден недостаточно хорошо.

Для первого случая проверка намного сложнее. Для остатков нам необходимо убедиться в отсутствии их автокоррелированности на заданное количество лагов назад. Если автокорреляционная функция остатков e_t резко убывает (и находится около нуля), то остатки не коррелируют друг с другом и, похоже, что они случайны.

В качестве статистики для проверки отсутствия автокорреляции остатков выступает Q-Статистика Льюнг-Бокса:

$$Q = n(n+2)\sum_{k=1}^{lag}\frac{r^2(k)}{n-k},$$

где r – выборочная оценка автокорреляционной функции:

$$r(k) = \frac{(n-k)\sum_{t=1}^{n-k} x_t x_{t+k} - \sum_{t=1}^{n-k} x_t \sum_{t=1}^{n-k} x_{t+k}}{\sqrt{(n-k)\sum_{t=1}^{n-k} x_t^2 - \left(\sum_{t=1}^{n-k} x_t\right)^2}\sqrt{(n-k)\sum_{t=1}^{n-k} x_{t+k}^2 - \left(\sum_{t=1}^{n-k} x_{t+k}\right)^2}}$$

Считается, что полученная величина имеет распределение χ^2 с lag степенями свободы.

H0: Автокорреляция остатков отсутствует.

H1: Автокорреляция остатков присутствует.

Если Q оказывается больше критического значения, то признается наличие автокорреляции до m-ого порядка в исследуемом ряду. Иначе считается, что автокорреляции нет и остатки признаются случайными.

Тест Дарбина-Уотсона

Еще один тест для проверки автокоррелированности остатков – тест Дарбина-Уотсона. Данный критерий применяется в ситуации, когда остатки связаны автокорреляционной зависимостью 1-го порядка.

Статистика вычисляется по следующей формуле:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Определяются критические значения d_L и d_U по специальным таблицам. Смотрим на столбец $m = 1$, это число факторных переменных.

n	m=1		m =2		m =3		m =4		m =5		m =6		m =7	
	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u
6	0,610	1,400												
7	0,7000	1,356	0,467	1,896										
8	0,763	1,332	0,359	1,777	0,368	2,287								
9	0,824	1,320	0,629	1,699	0,435	2,128	0,296	2,388						
10	0,879	1,320	0,697	1,641	0,525	2,016	0,356	2,414	0,243	2,822				
11	0,927	1,324	0,658	1,604	0,595	1,928	0,444	2,283	0,316	2,645	0,203	3,005		
12	0,971	1,331	0,812	1,576	0,658	1,864	0,512	2,177	0,379	2,506	0,268	2,832	0,171	3,149
13	1,010	1,340	0,861	1,562	0,715	1,816	0,574	2,094	0,445	2,390	0,328	2,692	0,230	2,985
14	1,045	1,330	0,905	1,551	0,767	1,779	0,632	2,030	0,505	2,296	0,389	2,572	0,286	2,848
15	1,077	1,361	0,946	1,543	0,814	1,750	0,685	1,977	0,562	2,220	0,447	2,472	0,343	2,727
16	1,106	1,371	0,982	1,539	0,857	1,728	0,734	1,935	0,615	2,157	0,502	2,388	0,398	2,624
17	1,133	1,381	1,015	1,536	0,897	1,710	0,779	1,900	0,664	2,104	0,554	2,318	0,451	1,537
18	1,158	1,391	1,046	1,535	0,933	1,696	0,820	1,872	0,710	2,060	0,603	2,257	0,502	2,461
19	1,180	1,401	1,074	1,536	0,967	1,685	0,859	1,848	0,752	2,023	0,649	2,206	0,549	2,396
20	1,201	1,411	1,100	1,537	0,998	1,676	0,894	1,828	0,792	1,991	0,692	2,162	0,595	2,339
21	1,221	1,420	1,125	1,538	1,026	1,669	0,927	1,812	0,829	1,964	0,732	2,124	0,637	2,290
22	1,239	1,429	1,147	1,541	1,053	1,664	0,958	1,797	0,863	1,940	0,769	2,090	0,677	2,246
23	1,257	1,437	1,168	1,543	1,078	1,660	0,986	1,785	0,895	1,920	0,804	2,061	0,715	2,208
24	1,273	1,446	1,188	1,546	1,101	1,656	1,013	1,775	0,925	1,902	0,837	2,035	0,751	2,174

n	m=1		m =2		m =3		m =4		m =5		m =6		m =7	
	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u	d_l	d_u
25	1,288	1,454	1,206	1,550	1,123	1,654	1,038	1,767	0,953	1,886	0,868	2,012	0,784	2,144
26	1,302	1,461	1,224	1,553	1,143	1,652	1,062	1,759	0,979	1,873	0,897	1,992	0,816	2,117
27	1,316	1,469	1,240	1,556	1,162	1,651	1,084	1,753	1,004	1,861	0,925	1,974	0,845	2,093
28	1,328	1,476	1,255	1,560	1,181	1,650	1,104	1,747	1,028	1,850	0,951	1,958	0,874	2,071
29	1,341	1,483	1,270	1,563	1,198	1,650	1,124	1,743	1,050	1,841	0,975	1,944	0,900	2,052
30	1,352	1,489	1,284	1,567	1,214	1,650	1,143	1,739	1,071	1,833	0,998	1,931	0,926	2,034
31	1,363	1,496	1,297	1,570	1,229	1,650	1,160	1,735	1,090	1,825	1,020	1,920	0,950	2,018
32	1,373	1,502	1,309	1,574	1,244	1,650	1,177	1,732	1,109	1,819	1,041	1,909	0,972	2,004
33	1,383	1,508	1,321	1,577	1,258	1,651	1,193	1,730	1,217	1,813	1,061	1,900	0,994	1,991
34	1,393	1,514	1,333	1,580	1,271	1,652	1,208	1,728	1,144	1,808	1,080	1,891	1,015	1,979
35	1,402	1,519	1,343	1,584	1,283	1,653	1,222	1,726	1,160	1,803	1,097	1,884	1,034	1,967
36	1,411	1,525	1,354	1,587	1,295	1,654	1,236	1,724	1,175	1,799	1,114	1,877	1,053	1,957
37	1,419	1,530	1,364	1,590	1,307	1,655	1,249	1,723	1,190	1,795	1,131	1,870	1,071	1,948
38	1,427	1,535	1,373	1,594	1,318	1,656	1,261	1,722	1,204	1,792	1,146	1,864	1,088	1,939
39	1,435	1,540	1,382	1,587	1,328	1,658	1,273	1,722	1,218	1,789	1,161	1,859	1,104	1,932
40	1,442	1,544	1,391	1,600	1,338	1,659	1,285	1,721	1,230	1,786	1,175	1,854	1,120	1,924
45	1,475	1,566	1,430	1,615	1,383	1,666	1,336	1,720	1,287	1,776	1,238	1,835	1,189	1,895
50	1,503	1,585	1,462	1,628	1,421	1,674	1,378	1,721	1,335	1,771	1,291	1,822	1,246	1,875
55	1,528	1,601	1,490	1,641	1,452	1,681	1,414	1,724	1,374	1,768	1,334	1,814	1,294	1,861
60	1,549	1,616	1,514	1,652	1,480	1,689	1,444	1,727	1,408	1,767	1,372	1,808	1,335	1,850
65	1,567	1,629	1,536	1,662	1,503	1,696	1,471	1,731	1,438	1,767	1,404	1,805	1,370	1,843
70	1,583	1,641	1,554	1,672	1,525	1,703	1,494	1,735	1,464	1,768	1,433	1,802	1,401	1,837

Статистика может принимать значения от 0 до 4:

$$0 \leq d \leq 4$$

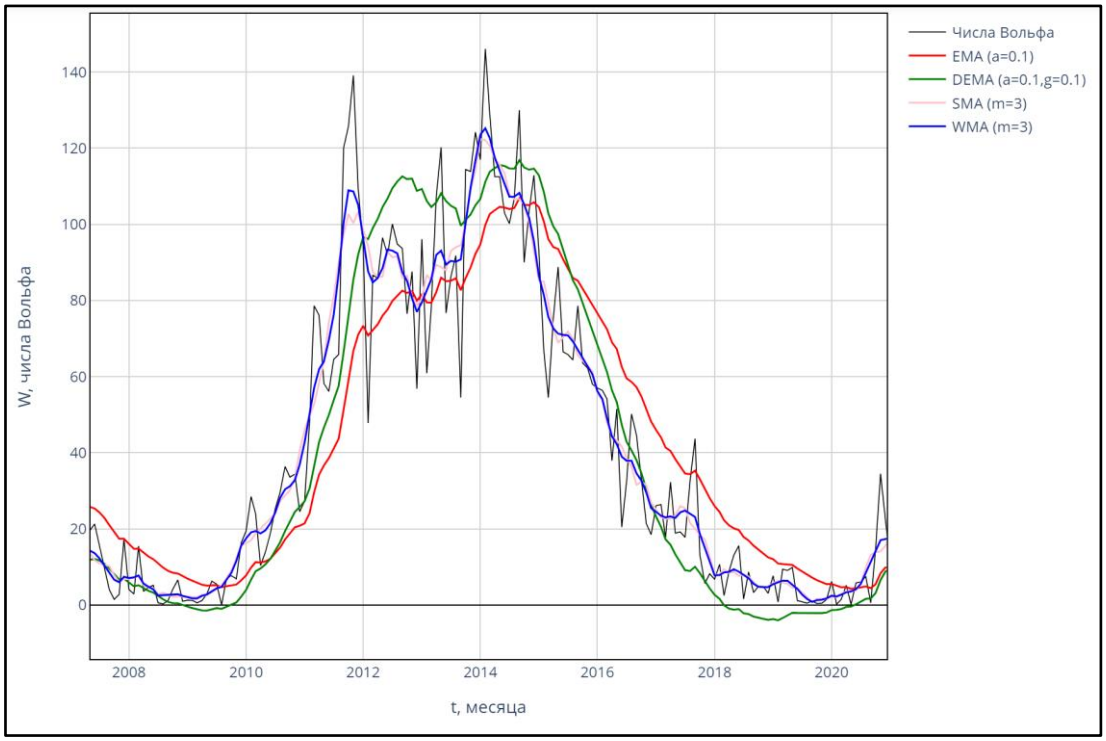
Для установления факта наличия автокорреляции удобно ввести вспомогательную величину $\tilde{d}$:

$$\tilde{d} = \begin{cases} d, \text{ если } 0 \leq d \leq 2 \\ 4 - d, \text{ если } 2 < d \leq 4 \end{cases}$$

Выделяются 3 области, куда может попасть значение статистики:

Значение $\tilde{d}$	**Вывод**
$0 \leq \tilde{d} < d_L$	Есть положительная автокорреляция
$d_L \leq \tilde{d} < d_U$	Неопределенность
$\tilde{d} \geq d_U$	Автокорреляция отсутствует

Критерий позволяет выявить только автокорреляцию 1-го порядка. Отклонение нулевой гипотезы не означает, что автокорреляции нет вообще, то есть возможно наличие автокорреляции более высоких порядков. Также данный критерий имеет зону неопределенности, когда нет оснований ни принимать, ни отвергать нулевую гипотезу.



Структура отчета

6 СГЛАЖИВАНИЕ ВРЕМЕННЫХ РЯДОВ

6.1 Постановка задачи

6.2 Ход выполнения работы

Графики и описание исходных данных.

6.2.1 SMA-сглаживание

В каждом подразделе сглаживание 2-х рядов, подбор параметров сглаживания с помощью q-статистики, изобразить на одном графике исходный ряд и сглаженный, тест Дарбина-Уотсона, расчетные формулы.

6.2.2 WMA-сглаживание

6.2.3 EMA-сглаживание

6.2.4 DEMA-сглаживание

6.3 Вывод