Практическая работа №6. Сглаживание временных рядов

Юрченков Иван Александрович, ассистент кафедры ПМ

2023-02-05

Постановка задачи

Для выполнения практического задания необходимо: 1. открыть папку, соответствующую своей группе; 2. открыть папку с вариантом, совпадающим с вашим номером в списке.

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

Необходимо выделить тренд используя 4 метода: 1. простое скользящее среднее (SMA); 2. взвешенное скользящее среднее (WMA) особо типа; 3. экспоненциальное сглаживание (EMA); 4. двойное экспоненциальное сглаживание (DEMA).

Каждый метод требует подбора некоторых параметров:

1. SMA и WMA - размер окна, EMA - параметр сглаживания A, DEMA - параметр сглаживания вокруг тренда A и параметр сглаживания самого тренда B.

Для весов в WMA использовать экспоненциальную весовую функцию:

$$\omega_i = \frac{e^{-\varepsilon \cdot |i|}}{\sum\limits_{j=-m}^{m} e^{-\varepsilon \cdot |j|}};\ i = -m, (-m+1), \dots, m;\ \varepsilon = 0.3.$$

2. Необходимо подобрать оптимальные значения соотвествующих параметров, используя Q-статистику Льюнг-Бокса при m = 5. Оптимальными параметрами будем считать те, что минимизируют приведенную статистику.
3. В качестве размеров окна $w = 2 * m + 1$ перебрать значения $m = 3, 5, 7, 9$; в качестве параметров сглаживания: $\alpha, \gamma = 0.1, 0.2, \dots, 0.9$. Обратите внимание, что метод DEMA двухпараметрический, что требует выбрать оптимальную комбинацию сразу двух параметров α, γ.
4. После подбора оптимальных параметров провести тест Дарбина-Уотсона $(m = 1, \alpha = 0.95)$ на данных после исключения выделенного тренда для каждого метода и каждого ряда.
5. 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.

Пример выполнения

Скользящее среднее

Метод простого скользящего среднего, с размером окна $w = 2 \cdot m + 1$, где m — количество членов ряда в сумме по одной стороне от центрального значения, является частным случаем метода взвешенного скользящего среднего с равными весовыми коэффициентами:

$$\widetilde{y_t} = \sum_{i=-m}^{m} \omega_i \cdot y_{t+i}, \quad \omega_i = \frac{1}{2 \cdot m + 1},$$

где y_t — исходные значения временного ряда в дискретных отсчётах t, ω_i — весовые коэффициенты окна сглаживания, $\widetilde{y_t}$ — сглаженный ряд данных y_t.

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:

Figure 1: Массив исходного ряда и окно свёртки на первом шаге

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.

Покажем также на рисунке полную картину такого сглаживания:

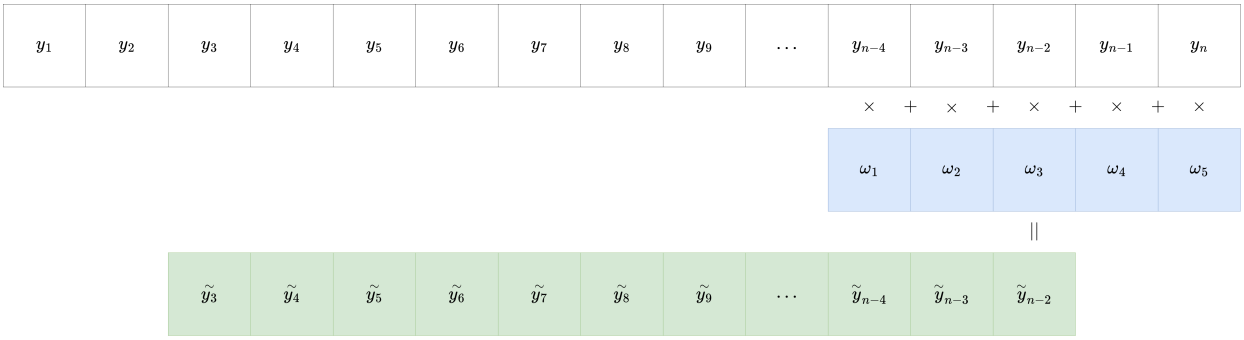


Figure 2: Массив исходного ряда и окно свёртки на последнем шаге

По изображению выше видно, что в нашем частном случае с обоих концов ряда удаляются по $m = 2$ значений с каждой стороны. Такую проблему в обработке изображений решают с помощью операции отступа (padding) на заданное количество единиц измерений с каждой стороны. С точки зрения обработки

изображений, к картинке с каждой стороны добавляют значение 0, для того, чтобы не вносить шум от краев изображения в модель обработки.

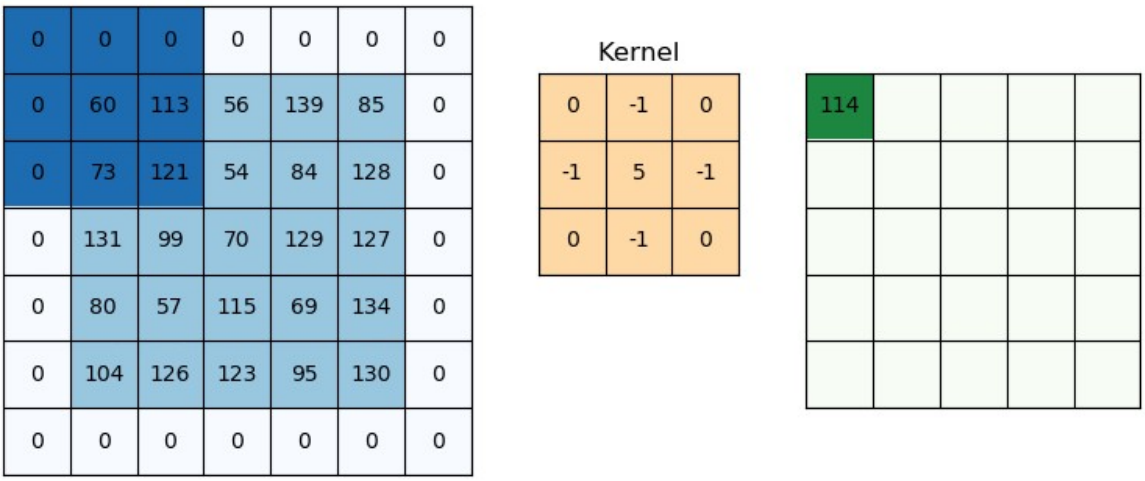


Figure 3: Отступ в обработке изображений для невелирования эффекта сжатия изображения в процессе свёртки

Так же, как и в задачах обработки изображений, мы имеем возможность добавить отступ с каждой стороны, чтобы невелировать эффект сжатия ряда. Однако, данный отступ необходимо делать первым значением в начале ряда и последним значением в конце ряда соответсвенно по m раз с каждой стороны, чтобы первые элементы сглаженного ряда не были занижены по уровню значений по сравнению с исходным рядом:

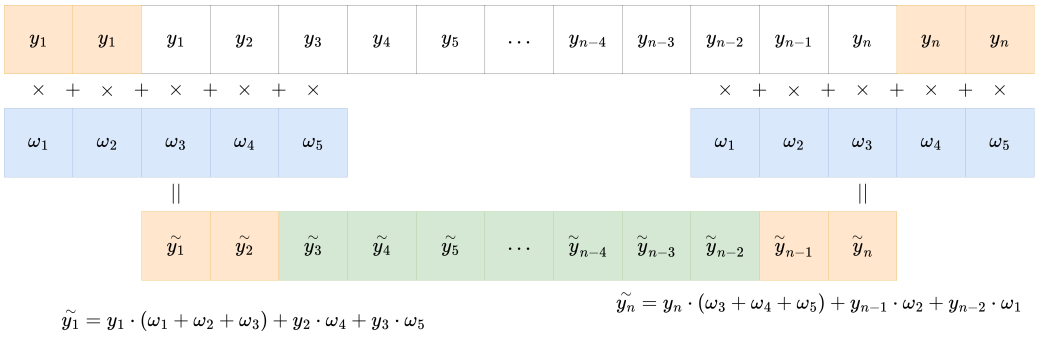


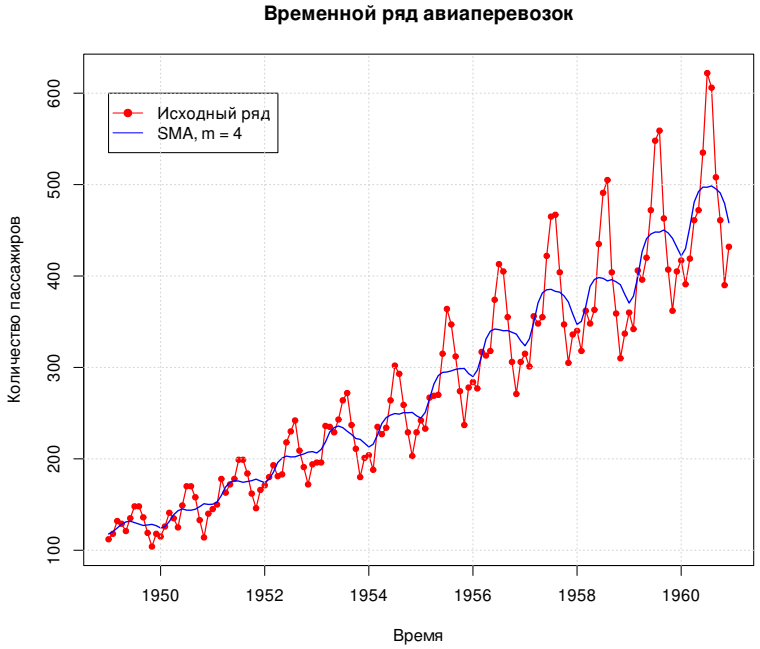
Figure 4: Свёртка одномерного временного ряда с отступом с каждой стороны

Стоит сказать, что при сравнимо с размерами ряда больших значениях окна, метод отступа будет

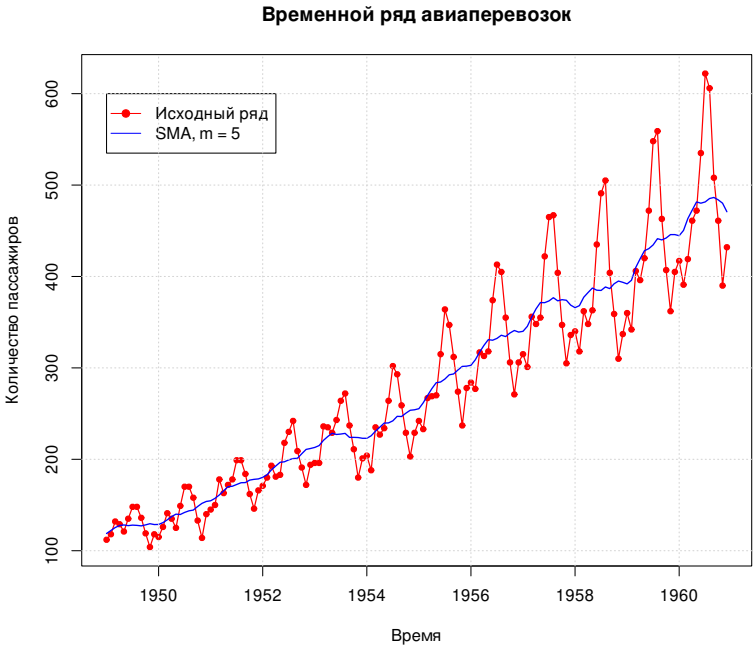
порождать большие ошибки сглаживания, так как будет вносить нетипичные значения ряда. Так что злоупотреблять данной техникой бывает опасно.

Таким образом, получаем алгоритм простого скользящего среднего для сглаживания временного ряда с заданным размером окна m. Продемонстрируем его работу на массиве данных авиаперевозок $AirPassengers$, а также на ряде данных уровня воды в озере Гурон.

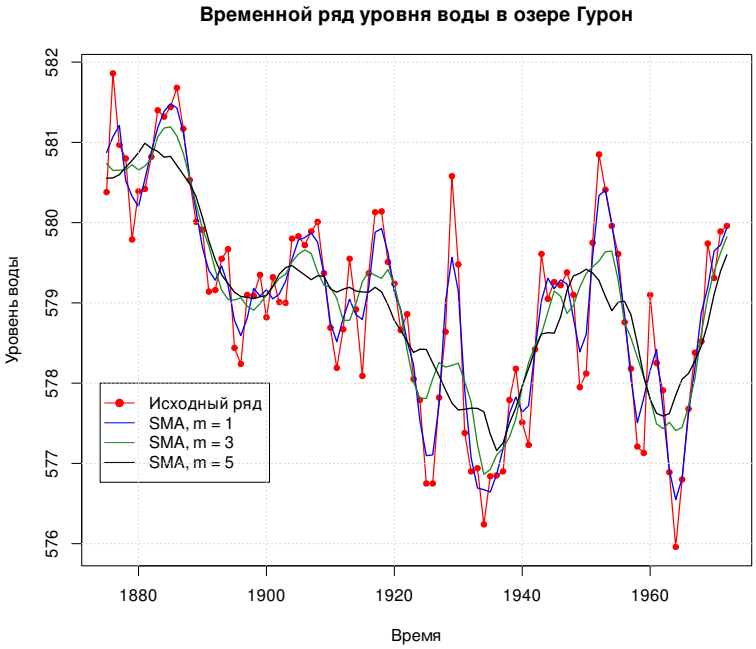
Построим для временного ряда данных $AirPassengers$ график и график его сглаживания по методу SMA с размером окна $m = 4, w = 2 \cdot m + 1 = 9$. Данный временной ряд не является стационарным, поскольку его значения разброса вокруг линейного тренда растут с ростом значения этого тренда (гетероскедастичность).



На графике видно, как сглаженная реализация все еще имеет в себе периодическую компоненту относительно тренда, и логично использовать близкие к $m = 5, 6$ значения окон для сглаживания, чтобы избавиться от влияния сезонности и рассмотреть именно сам тренд



Для значений временного ряда $LakeHuron$ уровня воды озера Гурон построим также график и реализацию сглаживания простым методом скользящего среднего.



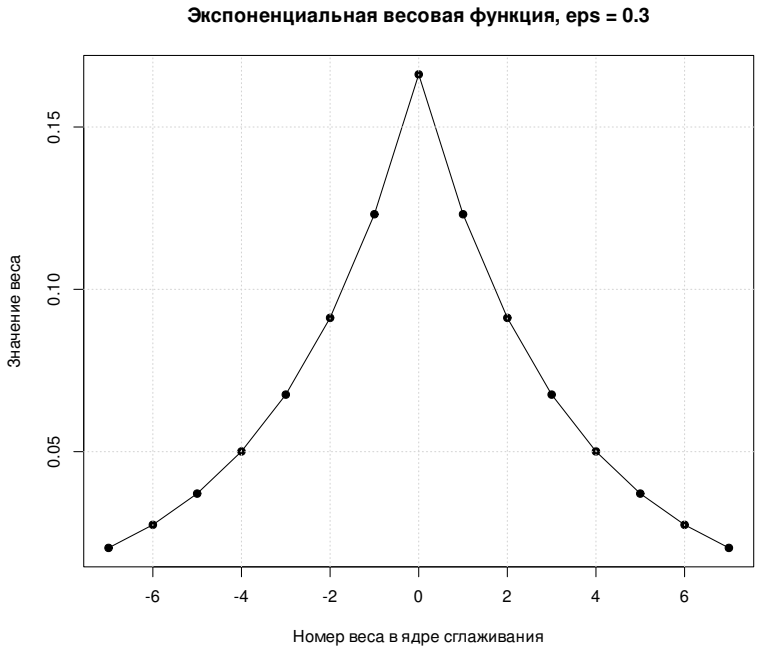
Познакомившись с простым методом скользящего среднего далее переходим к обобщенному взвешенному скользящему среднему.

Взвешенное скользящее среднее

Метод взвешенного скользящего среднего работает идентично методу простого скользящего среднего, за исключением необходимости 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:

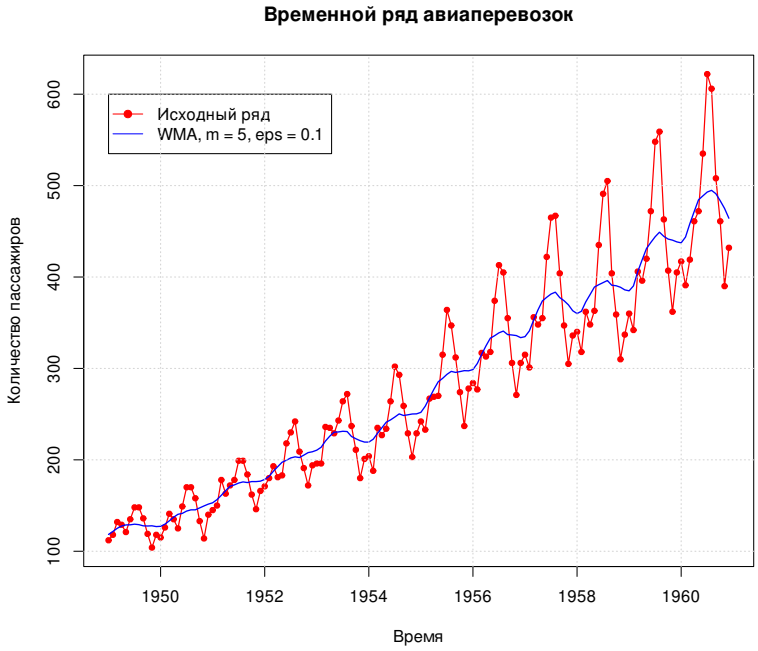
Из рисунка выше видно, что в методе простого скользящего среднего было использовано **П-окно**. В данной практике вам предлагается использовать экспоненциальную весовую функцию, также именуемую как **Пуассоновское сглаживание**:

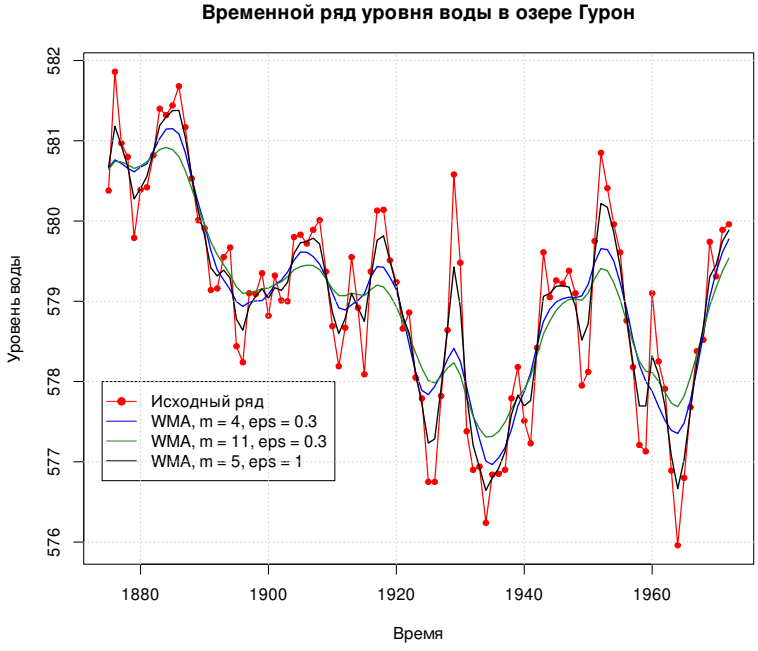


Черными точками по оси ординат можно отследить значение веса в окне сглаживания алгоритма

взвешенного скользящего среднего.

Приведем пример работы сглаживания по данным авиаперевозок и уровня воды в озере Гурон.





Экспоненциальное сглаживание

Экспоненциальное сглаживание - алгоритм, позволяющий рекуррентно оценивать новые значения сглаженного ряда по исходным данным, применяя следующие соотношения:

$$\tilde{y}_t = \begin{cases} y_t, & t = 1, \\ \tilde{y}_{t-1} + \alpha \cdot (y_t - \tilde{y}_{t-1}), & t > 1, \end{cases}$$

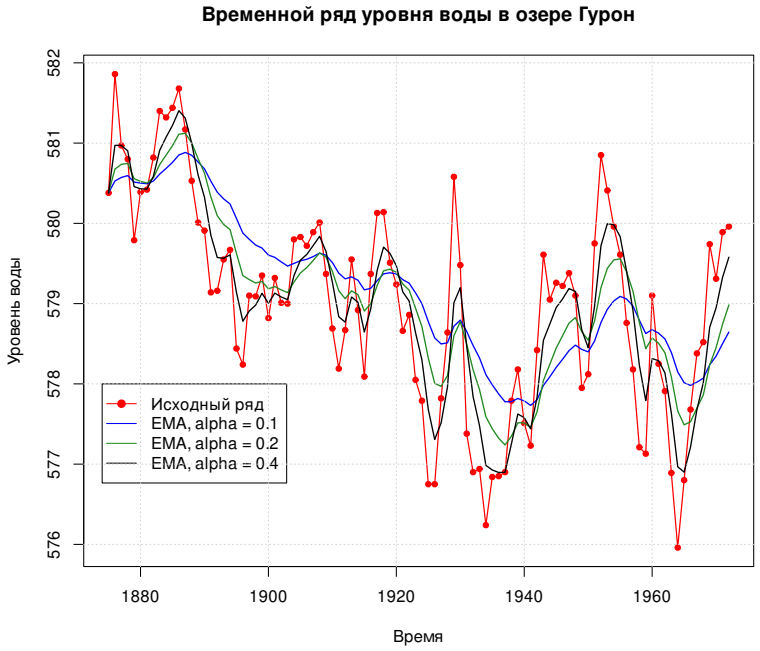
или

$$\widetilde{y}_t = \begin{cases} y_t, & t = 1, \\ \alpha \cdot y_t + (1 - \alpha) \cdot \widetilde{y}_{t-1}, & t > 1, \end{cases}$$

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

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

Покажем работу алгоритма сглаживания на значениях ряда уровня воды озера Гурон



Двойное экспоненциальное сглаживание

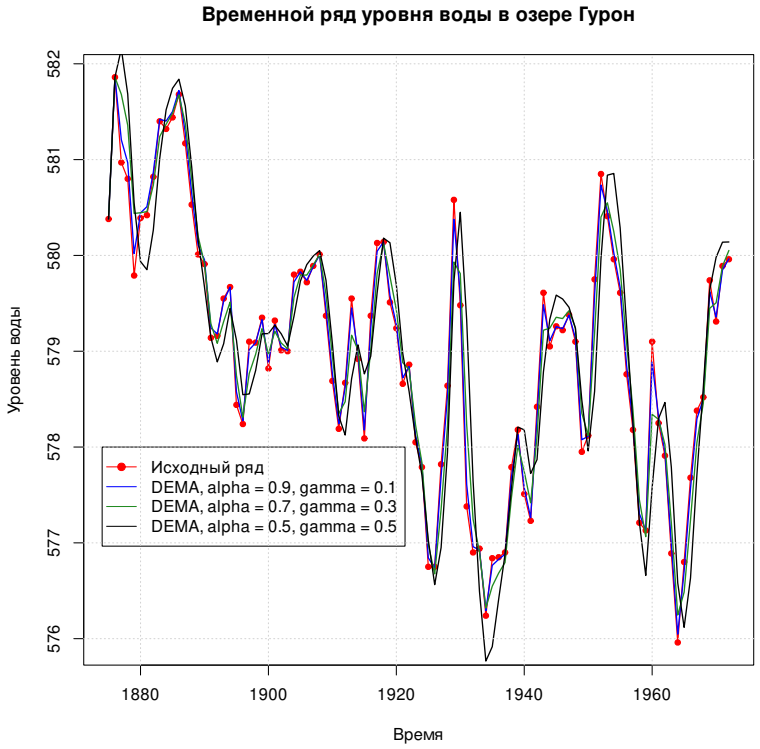
Двойное экспоненциальное сглаживание осуществляется по следующим формулам с коэффициентами α, γ, варьирующихся в пределах от 0 до 1 в действительной оси.

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

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

$$\widetilde{y}_1 = y_1, \;\; b_1 = y_2 - y_1$$

Продемонстрируем работу двойного экспоненциального сглаживания на примере временного ряда уровня воды в озере Гурон:



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, особенно если задавать их руками.

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 и выделить их более значительно.
2. Выделить тренд, сгладив и случайные компоненты и сезонность вокруг тренда по известному колчиеству дней периода сезонной компоненты.

В каждом из выделенных случаев, проверка адекватности алгоритма сглаживания проводится по-разному. Для начала выделяют остатки ряда данных

$$e_t = y_t - \widetilde{y_t}$$

, которые записываются отдельным временным рядом с теми же отсчётами. **Обратите внимание, что для методов сглаживания с окном (SMA, WMA) приходится либо рассматривать сглаженный ряд с дополненными исходными значениями ряда по краям, либо делать отступ перед сглаживанием**.

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. Проверки на выявления сезонности вместе с трендом не существует в общем употреблении.

Для первого случая проверка намного сложнее. Для остатков нам необходимо убедиться в отсутствии их авкоррелированности 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 отсутсвия автокорреляции остатков выступает Q-Статистика Льюнг-Бокса:

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

где $r(k)$ — выборочная оценка автокорреляционной функции в лаге k от текущего наблюдения:

$$r(k) = \frac{(n-k) \cdot \sum\limits_{t=1}^{n-k} x_t \cdot x_{t+k} - \sum\limits_{t=1}^{n-k} x_t \sum\limits_{t=1}^{n-k} x_{t+k}}{\sqrt{(n-k) \sum\limits_{t=1}^{n-k} x_t^2 - (\sum\limits_{t=1}^{n-k} x_t)^2} \cdot \sqrt{(n-k) \sum\limits_{t=1}^{n-k} x_{t+k}^2 - (\sum\limits_{t=1}^{n-k} x_{t+k})^2}}$$

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

В пакете R существует функция $Box.test(x, lag, type = "Ljung - Box")$, которая предоставляет возможность тестировать ряд на автокорреляцию. Необходимо посчитать остатки и провести тест.

```
Box.test(x = df$flights - df$flights_smoothed_DEMA, lag = 5, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  df$flights - df$flights_smoothed_DEMA
## X-squared = 190.95, df = 5, p-value < 2.2e-16
```

Темы вопросов на защиту практической работы

1. Понятие временного ряда (ВР). Стационарность и нестационарность ВР. Примеры ВР. Аддитивная и мультипликативная модели ВР
2. Методы сглаживания ВР: MA, WMA, SMA, EMA.
3. Лаговый оператор. Модели с распределенным лагом. Лаги Алмон.
4. Тест Дарбина-Уотсона.
5. Q-стастистика Льюнга-Бокса и применение для подбора параметров статистических моделей обработки данных.