

# Методы современной прикладной статистики

## 12. Анализ временных рядов.

Родионов Игорь Владимирович  
vecsell@gmail.com

Весна, 2018

# Постановка задачи

Временным рядом называется последовательность  $\{y_n\}_{n \geq 1}$ , зависящая от  $n$  (т.е. от времени).

Задача прогнозирования: найти такую функцию  $f$ , что

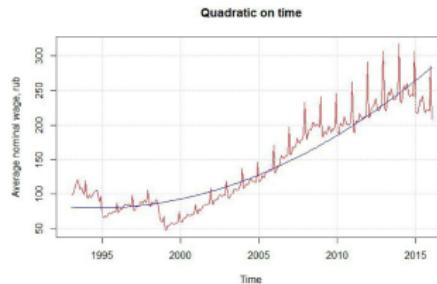
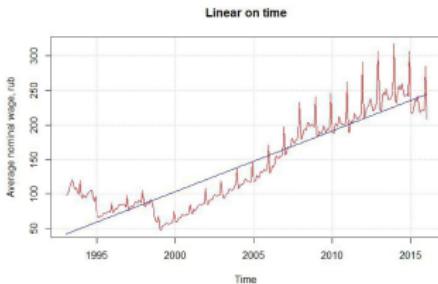
$$y_{n+d} \approx f(y_1, \dots, y_n, d) = \hat{y}_{n+d}, \quad d = 1, \dots, D.$$

Часто также стоит задача сделать прогноз не только по предыдущим значениям ряда, но и по другим имеющимся признакам.

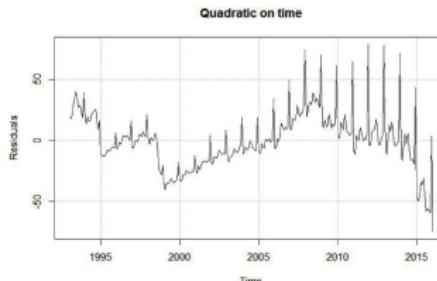
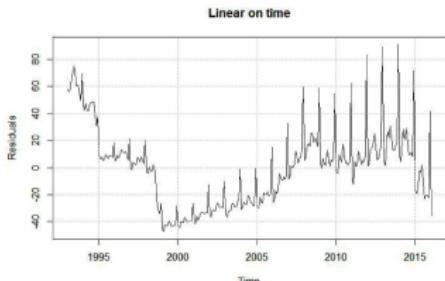
Простейшие методы типа добавить в модель в качестве признака время не всегда помогают, потому что наблюдения могут быть сильно зависимыми от своих соседей во временном ряду.

# Регрессия по времени

Сделаем регрессию по времени для наблюдений временного ряда



Видим, что остатки далеки от однородности



# Компоненты временного ряда

- Тренд – плавное долгосрочное изменение уровня ряда.
- Сезонность – циклические изменения уровня ряда с постоянным периодом.
- Циклы – изменения уровня ряда с переменным периодом (цикл жизни товара, экономические волны, периоды солнечной активности).
- Ошибка – непрогнозируемая случайная компонента ряда.

# STL-разложение

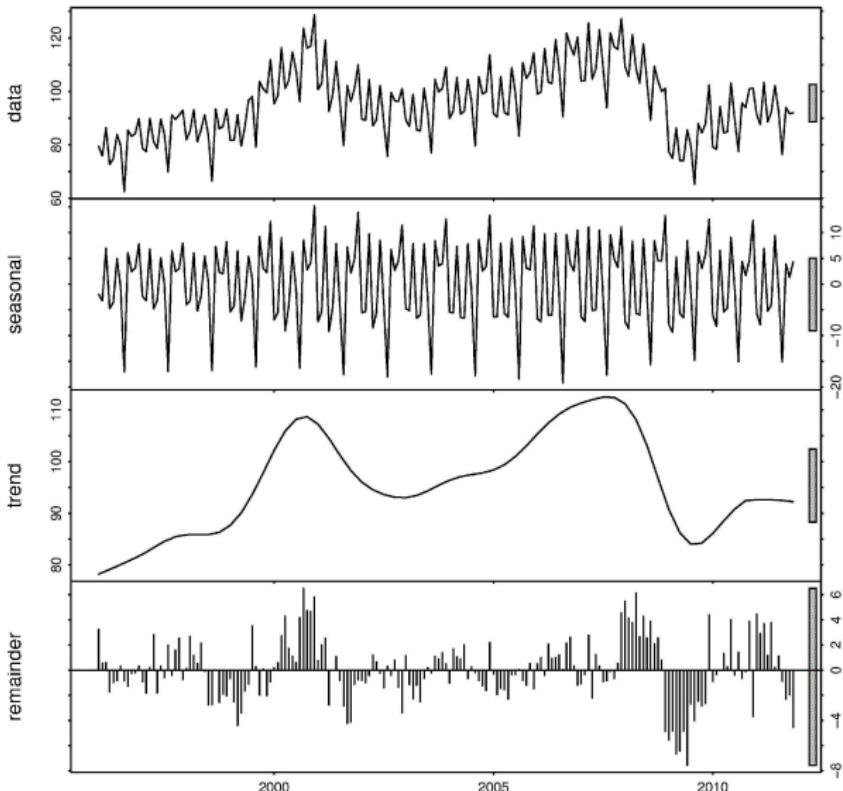
STL-разложение (STL-decomposition) – полезный визуальный метод анализа временных рядов.

Выделение тренда (+циклической составляющей) и сезонной компоненты основано на алгоритме loess (предшественнике LOWESS) робастной непараметрической регрессии.

После выделения тренда сезонность определяется с помощью сглаживающего среднего по нескольким циклам. Здесь используется входной параметр, задающий длину сезона, что является небольшим недостатком метода.

Ещё одним недостатком алгоритма является то, что он применим только для аддитивных моделей временных рядов.

# STL-разложение



# Критерии случайности

Если после визуального анализа всё же остаются сомнения в том, что перед нами временной ряд, то можно применить критерии случайности. Проверим гипотезу  $H_0 : \{X_i\}_{i=1}^n$  образуют выборку.

1) **Критерий инверсий.** Определим  $I = \sum_{i < j} I(X_i > X_j)$ . Тогда при  $n \geq 20$  и верной  $H_0$

$$I \approx N \left( \frac{n(n-1)}{4}, \frac{2n^3 + 3n^2 - 5n}{72} \right).$$

2) **Критерий Вальда-Волфовитца.** Пусть  $\hat{\mu}$  – выборочная медиана  $\{X_i\}_{i=1}^n$ . Определим  $Y_i = I(X_i \geq \hat{\mu})$  и  $W = \sum_i I(Y_i \neq Y_{i+1}) + 1$ . Пусть  $n_0$  и  $n_1$  – количество 0 и 1 в ряду  $\{Y_i\}_{i=1}^n$ . Тогда при верной  $H_0$  и  $n_0, n_1 > 20$

$$W \approx N \left( \frac{2n_0 n_1}{n_0 + n_1} + 1, \frac{2n_0 n_1 (2n_0 n_1 + n_0 + n_1)}{(n_0 + n_1)^2 (n_0 + n_1 - 1)} \right).$$

# Стационарность

Ряд  $\{X_n\}_{n \geq 1}$  – стационарный в узком смысле, если  $\forall s$   
 $\forall t_1, \dots, t_k$  выполнено

$$(X_{t_1}, \dots, X_{t_k}) \stackrel{d}{=} (X_{t_1+s}, \dots, X_{t_k+s}).$$

Ряд  $\{X_n\}_{n \geq 1}$  – стационарный в широком смысле, если  
 $E X_n = \text{const}$  и  $\text{cov}(X_n, X_m) = r(n - m)$ .

Как правило, пытаются сделать так, чтобы остатки временного ряда были нормальными, в этом случае первое и второе определения совпадают.

Нормальность проверяется с помощью простых критериев типа Харке-Бера, потому что остатки могут быть сильно зависимы.

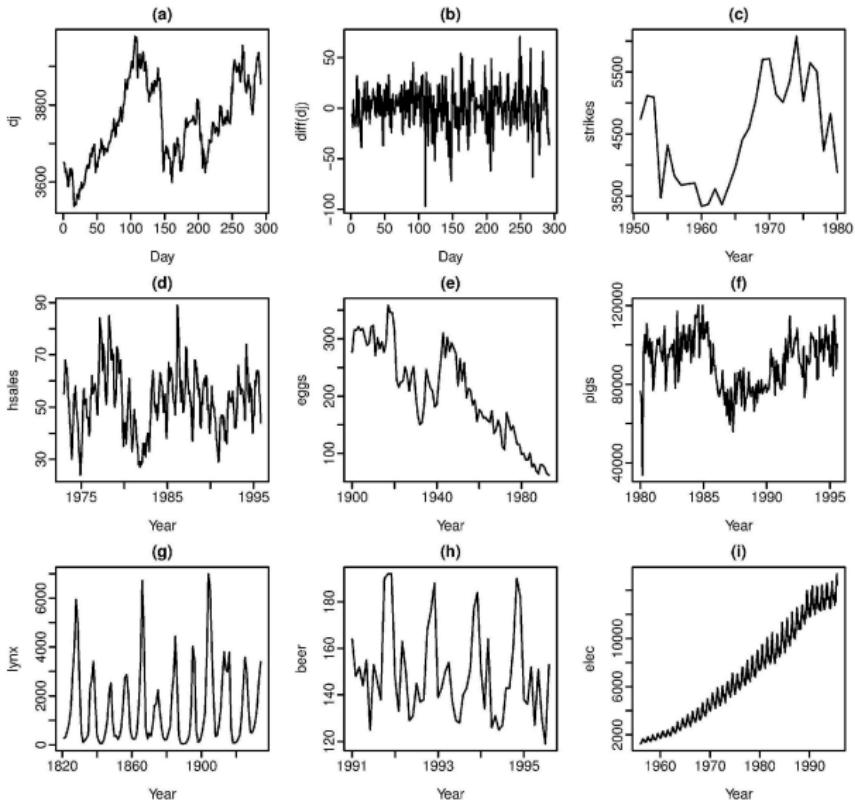
# Стационарность

Ряды с трендом или сезонностью не стационарны. Ряды с непериодическими циклами стационарны, поскольку нельзя предсказать заранее, где будут находиться максимумы и минимумы.

Стационарность так важна по той причине, что по стационарному ряду просто строить прогноз, так как мы полагаем, что его будущие статистические характеристики не будут отличаться от наблюдаемых текущих.

Какие из рядов на следующем рисунке являются стационарными?

# Примеры



# Как убрать нестационарность?

Путь первый (простой):

Пусть по графику ряда видно, что тренд похож на линейный, а период ряда явно просматривается и равен  $d$ . В таком случае строится ряд

$$MA_i = \frac{1}{d} \sum_{j=i-[d/2]}^{i+[d/2]} X_j.$$

Если разброс значений ряда не зависит от его значений, то стоит использовать аддитивную модель и вычесть из значений ряда  $MA_i$  там, где  $MA_i$  определен. Если зависит прямо пропорционально, то  $X_i$  делим на  $MA_i$ . После этой операции получаем значения  $Y_i$ .

# Как убрать нестационарность?

Далее, периодическая компонента вычисляется как

$$S_j = \frac{1}{[n/d] - 1} \sum_{i=j+dk, k \geq 0} Y_i,$$

$j = 0, \dots, d - 1$ , по тем значениям  $i$ , где  $Y_i$  определены.

После этого предсказания новых значений строятся следующим образом: строится регрессия  $MA_i$  по времени, получаем предсказания  $\widehat{MA}_i$ ,  $i > n$ . Если модель аддитивная, то к  $\widehat{MA}_i$  прибавляем  $S_{i \pmod d}$ , если мультипликативная, то домножаем.

Метод дает удовлетворительный результат, если остатки модели много меньше её основных компонент или если данных довольно мало.

# Как убрать нестационарность?

Путь второй (сложный):

- 1) Пусть период ряда равен  $d$ . Сезонным дифференцированием называется преобразование  $Y_t = X_t - X_{t-d}$ . Оно, как правило, помогает избавиться от сезонности в данных.
- 2) Дифференцированием ряда называется операция  $\Delta X_t = X_t - X_{t-1}$ . Если после этой операции ряд не стал стационарным, то нужно попробовать продифференцировать его ещё раз.
- 3) Можно попробовать убрать нестационарность с помощью регрессии значений ряда по времени, лагам и (если они есть) другим признакам.  $k$ -тым лагом значения  $X_t$  называется  $Y_t^{(k)} = X_t - X_{t-k}$ .

Сезонное дифференцирование стоит делать вначале – возможно, что после него ряд уже будет стационарным.



# Тест Дики-Фуллера

Рассмотрим модель AR(1) стационарного временного ряда:  $X_t = aX_{t-1} + \varepsilon_t$ , где  $\varepsilon_t$  – ошибка, не зависящая от значений временного ряда. В случае, если  $a = 1$ , ряд перестает быть стационарным (об этом чуть позже) – тогда говорят, что процесс имеет **единичный корень**.

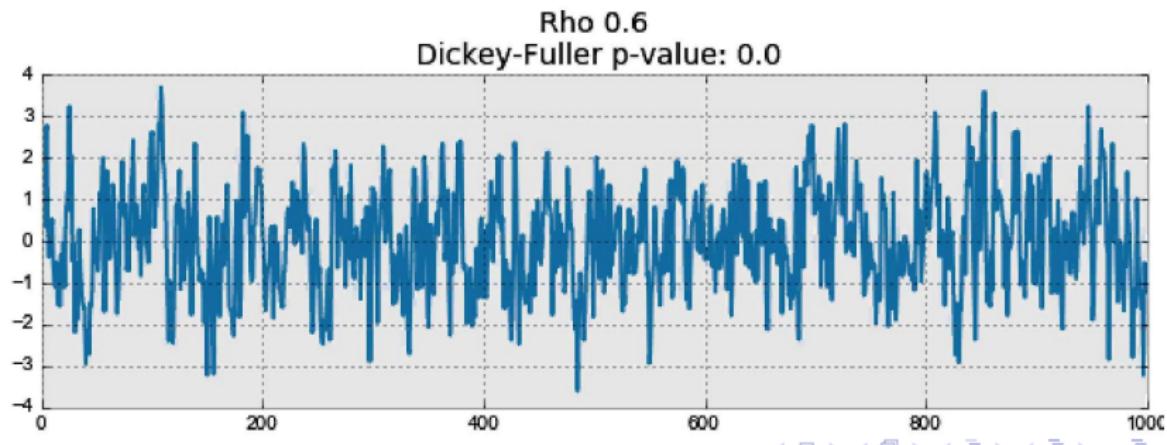
Рассмотрим теперь модель  $\Delta X_t = bX_{t-1} + \varepsilon_t$ . Если  $a = 1$ , то  $b = 0$ . Тест Дики-Фуллера (DF-test) проверяет гипотезу  $H_0 : b = 0$  для приведенной модели авторегрессии.

Статистика теста – обычная t-статистика для значимости коэффициентов линейной регрессии, однако распределение статистики при верности  $H_0$  будет другим.

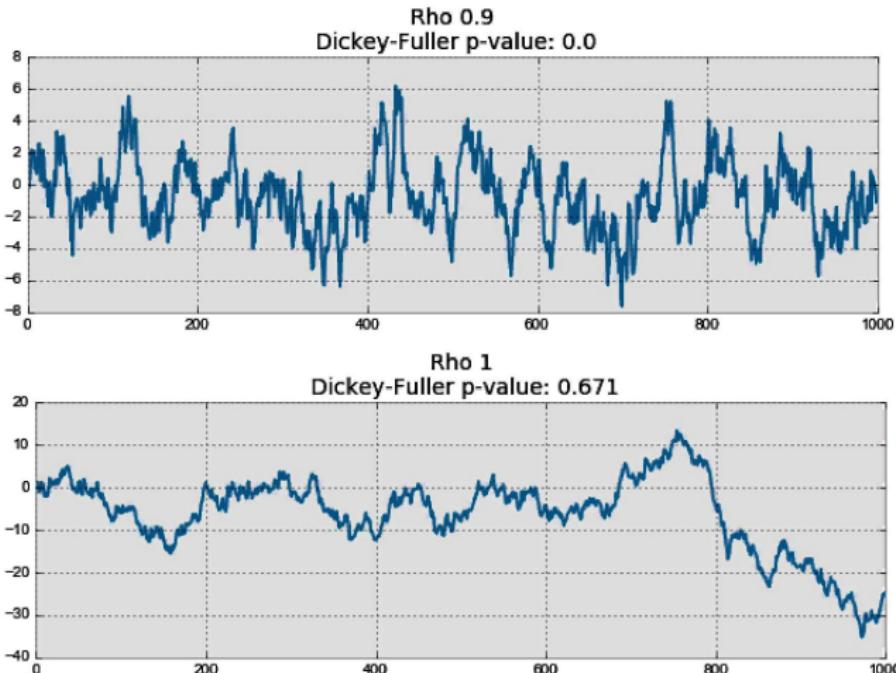
# Тест Дики-Фуллера

Объясним, почему наличие единичного корня у процесса так важно. Сгенерируем процесс по модели

$X_t = \rho X_{t-1} + e_t$ , где  $\{e_t\}_{t \geq 1}$  – независимые стандартные нормальные случайные величины. Случай, когда  $|\rho| > 1$  (“взрывной” процесс) легко распознаем визуально, поэтому не будем его рассматривать.



# Тест Дики-Фуллера



Видим, что при  $\rho = 1$  процесс не возвращается к своему среднему значению, а значит, не является стационарным.

## Другие тесты проверки стационарности

Если единичных корней у ряда несколько (т.е. чтобы достичь стационарности, нужно несколько раз продифференцировать ряд), то следует пользоваться расширенным тестом Дики-Фуллера (ADF).

Ещё одним тестом проверки стационарности является критерий KPSS (Kwiatkowski–Phillips–Schmidt–Shin test). Критерий проверяет стационарность аддитивной модели временного ряда  $X_t = c_t + \delta t + e_t$ ,  $c_t = c_{t-1} + u_t$ , где  $e_t$  – некий стационарный ряд,  $u_t \stackrel{d}{=} e_t$ , независим с  $e_t$ ,  $Eu_t = 0$ ,  $Du_t = \sigma^2$ .

# Модели авторегрессии

Перейдем теперь к вопросу, сколько раз нужно дифференцировать ряд, чтобы он стал похож на стационарный. Нам потребуется ввести несколько определений.

Процесс  $\varepsilon_t$  такой, что  $E\varepsilon_t = 0 \quad \forall t$ ,  $cov(X_t, X_s) = \sigma_\varepsilon^2 \delta(t - s)$ , где  $\delta(x)$  – символ Кронекера, называется **белым шумом**. В случае, если процесс является гауссовским, этот процесс имеет независимые значения.

# Модели авторегрессии

Процесс авторегрессии порядка  $p$  ( $AR(p)$ ) – это процесс, задаваемый как

$$X_t = a_1 X_{t-1} + \dots + a_p X_{t-p} + \varepsilon_t,$$

$a_p \neq 0$ ,  $\varepsilon_t$  – процесс белого шума,  $E\varepsilon_t = 0$ ,  $D\varepsilon_t = \sigma_\varepsilon^2$ , некоррелируемый с  $X_t$ .

Введем оператор запаздывания  $L$ :  $LX_t = X_{t-1}$ , тогда определение процесса  $AR(p)$  можно записать так:

$$a(L)X_t = \varepsilon_t,$$

где  $a(L) = 1 - (a_1 L + \dots + a_p L^p)$ .

Процесс авторегрессии  $AR(p)$  является стационарным (в широком смысле) тогда и только тогда, когда корни уравнения  $a(z) = 0$  лежат вне круга  $|z| \leq 1$  на комплексной плоскости.

# Модели авторегрессии

## Процесс

$$X_t = \varepsilon_t + b_1 \varepsilon_{t-1} + \dots + b_q \varepsilon_{t-q} = b(L) \varepsilon_t,$$

$b_q \neq 0$ , где  $\varepsilon_t$  – процесс белого шума, называется **процессом скользящего среднего порядка  $q$**  ( $MA(q)$ ). Такой процесс всегда является стационарным.

## Теорема.

- 1) Любой стационарный процесс  $AR(p)$ , заданный соотношением  $a(L)(X_t - \mu) = \varepsilon_t$ , можно представить в виде процесса  $MA(\infty)$ .
- 2) Обратно, если процесс  $X_t$  удовлетворяет  $MA$ -представлению  $X_t - \mu = b(L) \varepsilon_t$  и все корни уравнения  $b(z) = 0$  в комплексной плоскости по модулю больше 1, то  $X_t$  представим в виде процесса  $AR(\infty)$ .

# Модели авторегрессии

Будем называть  $X_t$  смешанным процессом авторегрессии ( $ARMA(p, q)$ ), если  $EX_t = 0$  и  $a(L)X_t = b(L)\varepsilon_t$ , т.е.

$$X_t = \sum_{j=1}^p a_j X_{t-j} + \sum_{k=0}^q b_k \varepsilon_{t-k},$$

где  $b_0 = 1$ ,  $a_p, b_q \neq 0$ .

**Теорема (разложение Вольда).**

Любой центрированный стационарный в широком смысле процесс  $X_t$  может быть представлен в виде

$$X_t = \sum_{j=1}^{+\infty} a_j X_{t-j} + \sum_{k=0}^{+\infty} b_k \varepsilon_{t-k},$$

где  $\varepsilon_t$  – процесс белого шума,  $cov(X_t, \varepsilon_{t-j}) = 0 \forall t, j$ .

Отсюда следует, что любой стационарный временной ряд можно приблизить моделью  $ARMA(p, q)$  сколь угодно точно.

Временной ряд описывается моделью  $ARIMA(p, q, d)$ , если ряд его дифференцирований  $\nabla^d X_t = (1 - L)^d X_t$  описывается моделью  $ARMA(p, q)$ , т.е.

$$a(L)\nabla^d X_t = b(L)\varepsilon_t.$$

Видим, что у данной модели  $d$  единичных корней, т.е. исходный процесс  $X_t$  не является стационарным. Если в модель добавить ещё и сезонное дифференцирование, то получится модель  $SARIMA$ .

Как оптимальным образом выбирать параметры  $p, q, d$  и длину периода?

# Автокорреляция

Пусть имеются наблюдения временного ряда  $\{x_t\}_{t=1}^T$ .

**Автокорреляционной функцией** (или  
автокорреляцией, ACF) этого ряда называется функция

$$r_\tau = \frac{\sum_{t=1}^{T-\tau} (x_t - \bar{x})(x_{t+\tau} - \bar{x})}{\sum_{t=1}^T (x_t - \bar{x})^2},$$

где  $\bar{x} = \frac{1}{T} \sum_{t=1}^T x_t$ . Легко заметить, что всегда  $r_0 = 1$ .

# Частичная автокорреляция

Частичной автокорреляционной функцией (PACF) ряда  $\{x_t\}_{t=1}^T$  называется функция

$$\varphi_h = \begin{cases} r_h = r(x_t, x_{t+h}), & h = 1 \\ r(x_{t+h} - x_{t+h}^{h-1}, x_t - x_t^{h-1}), & h > 1, \end{cases}$$

где  $r$  – оценка автокорреляционной функции с предыдущего слайда,  $x_t^{h-1}$  и  $x_{t+h}^{h-1}$  – линейная регрессия  $x_t$  и  $x_{t+h}$  соответственно по признакам  $x_{t+1}, \dots, x_{t+h-1}$ . По сути,  $PACF(h)$  – мера того, как сильно  $X_t$  влияет на  $X_{t+h}$ .

Анализ ACF и PACF – основной метод при определении параметров  $p$ ,  $d$  и  $q$  модели ARIMA.

# Определение параметров ARIMA

1) Если один из лагов PACF близок к 1, то стоит сделать дифференцирование ряда (если это первый лаг, то обычное дифференцирование, если это  $s$ -тый лаг при условии, что он значимее, чем первый, то сезонное).

Операцию следует повторять, пока не останется лагов, близких к единице. Для уверенности в том, что перед нами теперь стационарный ряд, стоит пользоваться тестами Дики-Фуллера и KPSS.

2) После приведения ряда к стационарному можно определить параметры  $p$  и  $q$  модели ARIMA. Параметр  $p$  равняется номеру последнего значимого лага функции PACF (после которого лаги экспоненциально убывают). Параметр  $q$  равняется номеру последнего значимого лага функции ACF.

# Определение качества модели

Пусть мы получили оценки членов временного ряда с помощью некоторой модели, и остатки в этой модели  $\{e_i\}_{i=1}^T$ . Определим  $\hat{\sigma}^2 = \frac{1}{T} \sum_{i=1}^T e_i^2$  – фактически, это оценка дисперсии ошибок. Пусть для описания ряда мы выбрали модель  $ARIMA(p, d, q)$ .

Критерий Акаике:

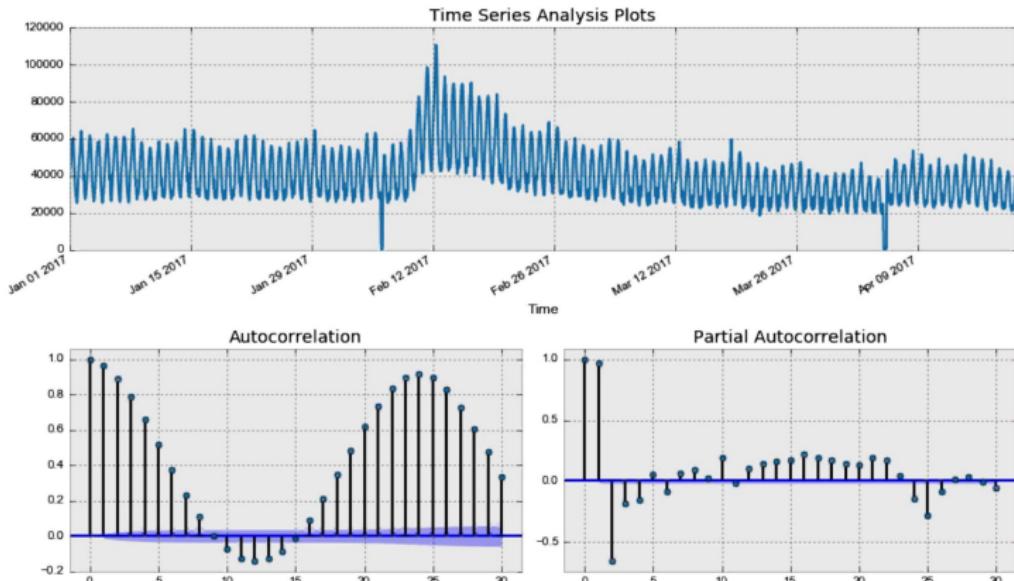
$$AIC_c = \ln \hat{\sigma}^2 + \frac{2(p+q+1)(p+q+2)}{T-q-p-2}.$$

Критерий Шварца:

$$BIC = \ln \hat{\sigma}^2 + \frac{(\ln T - 2)(p+q+1)}{T}.$$

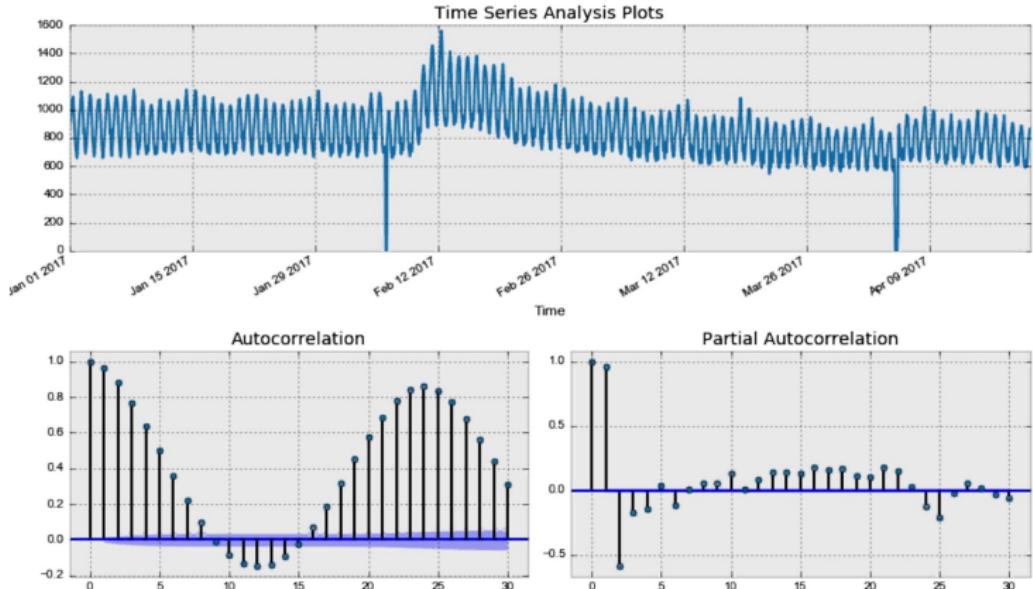
Как и ранее, чем меньше AIC (BIC), тем лучше считается модель.

# Пример



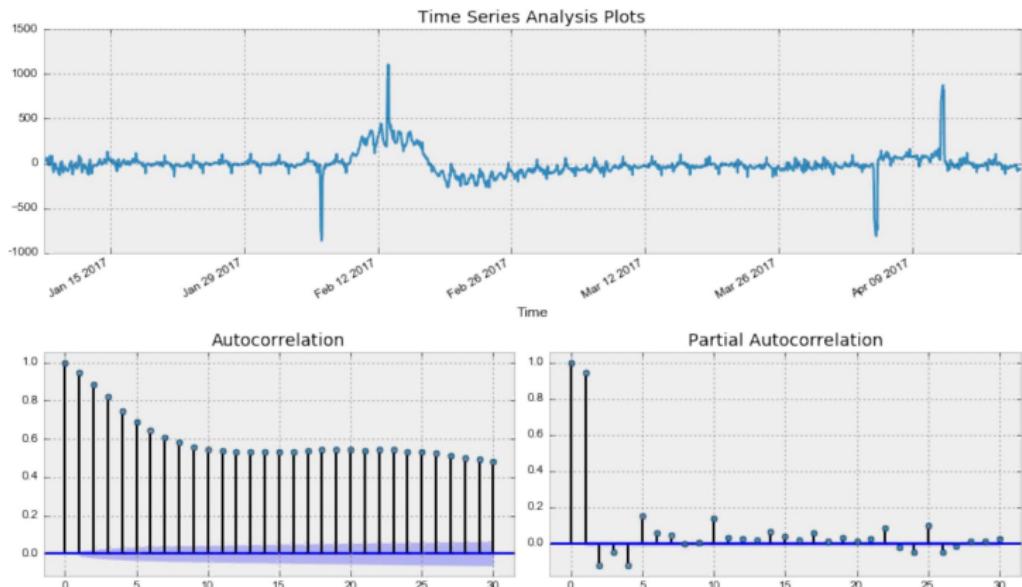
Видим почти единичный лаг PACF. Тест Дики-Фуллера дает  $p$ -значение 0.19, это значит, что есть единичный корень и ряд нестационарный, да и видно, что есть сезонность. Кроме того, есть гетероскедастичность - её можно попробовать исправить с помощью преобразования Бокса-Кокса.

# Пример



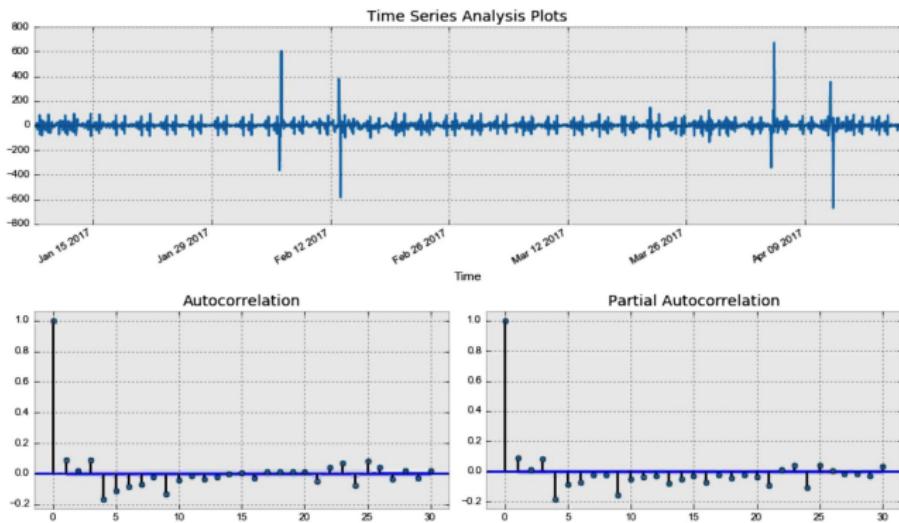
После применения преобразования Бокса-Кокса дисперсия стабилизировалась, но в остальном мало что изменилось. Р-значение теста Дики-Фуллера равно 0.07. Возьмем сезонное дифференцирование ряда.

# Пример

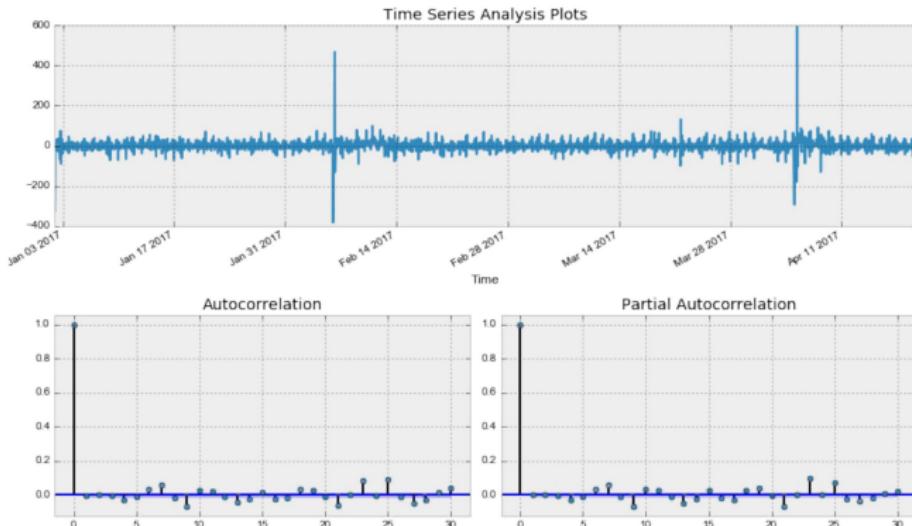


Теперь тест Дики-Фуллера отвергает гипотезу о нестационарности, но ACF выглядит плохо и в PACF остался лаг, близкий к 1. Сделаем обычное дифференцирование ряда.

# Пример

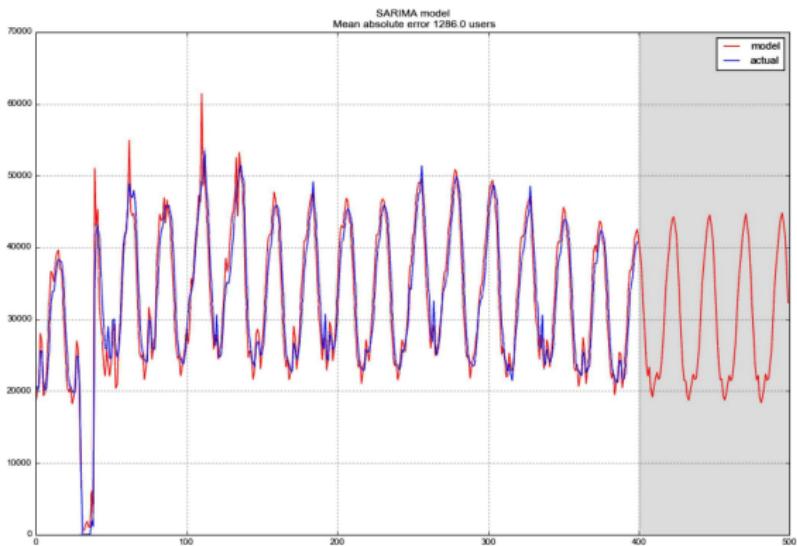


Наконец, мы получили стационарный ряд, причем стоит выбрать  $p \leq 4$  и  $q \leq 4$ , выбор можно осуществлять полным перебором. Выбор оптимальной модели можно осуществлять с помощью информационных критериев AIC и BIC или с помощью кросс-валидации.



Оптимальными (с точки зрения ВIC) параметрами модели оказались  $p = 4$  и  $q = 3$ , причем в модель нужно не забыть внести то, что мы делали сезонное и обычное дифференцирования. То, что первые  $L$  лагов автокорреляции получились незначимыми, можно проверить с помощью Q-критерия Льюнга-Бокса.

# Пример



Прогноз строится следующим образом: мы последовательно вычисляем  $\hat{x}_{T+1}, \hat{x}_{T+2}, \dots$ , новые ошибки  $\varepsilon_{T+1}, \dots$  полагаем равными 0, а вместо старых ошибок используем остатки модели. В итоге мы получили довольно точную модель, но суммарные затраты на приведение модели к адекватному виду могли такой точности и не стоить.

# Другие модели

Как уже упоминалось, затраты на настройку модели SARIMA часто не окупаются, и порой даже не удается привести модель к адекватному виду. Поэтому стоит рассмотреть другие возможности по предсказанию временных рядов.

Одним из таких путей является использование модели линейной регрессии. Из временного ряда вычленяется набор признаков-лагов, кроме того, из признака времени можно вычленить категориальные признаки дней недели, часов и тд. Модель приобретает больший смысл, если кроме самих значений временного ряда у нас есть и другие признаки. Хотя предположения Гаусса-Маркова, очевидно, не выполнены из-за автокоррелированности значений ряда, модель может давать неплохие результаты.

# Экспоненциальное сглаживание

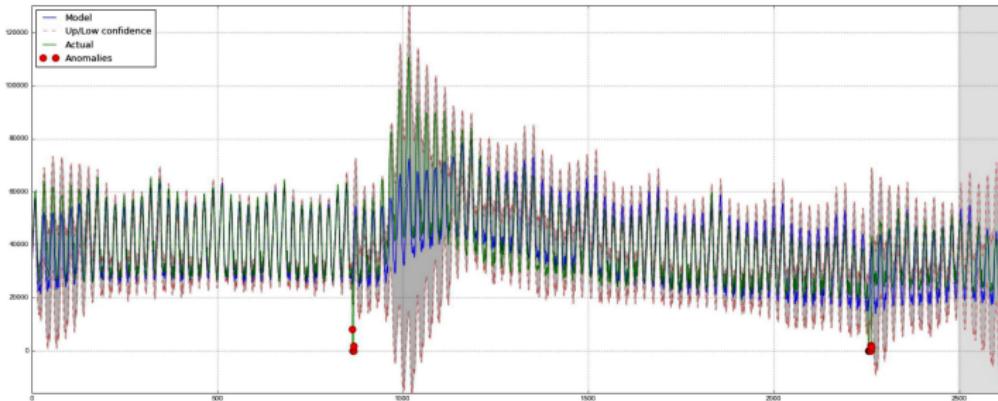
Другим выходом является использование моделей экспоненциального сглаживания. Различают аддитивные и мультипликативные модели, мы же рассмотрим аддитивную модель тройного экспоненциального сглаживания Хольта-Винтерса.

$$\begin{cases} l_t = \alpha(x_t - s_{t-L}) + (1 - \alpha)(l_{t-1} + b_{t-1}); \\ b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1}; \\ s_t = \gamma(x_t - l_t) + (1 - \gamma)s_{t-L}; \\ \hat{y}_{t+m} = l_t + mb_t + s_{t-L+1+((m-1)mod\ L)}. \end{cases}$$

Если занулить  $s_t$ , то получим модель двойного экспоненциального сглаживания, если занулить ещё и  $b_t$ , то получится модель простого экспоненциального сглаживания Брауна.

# Экспоненциальное сглаживание

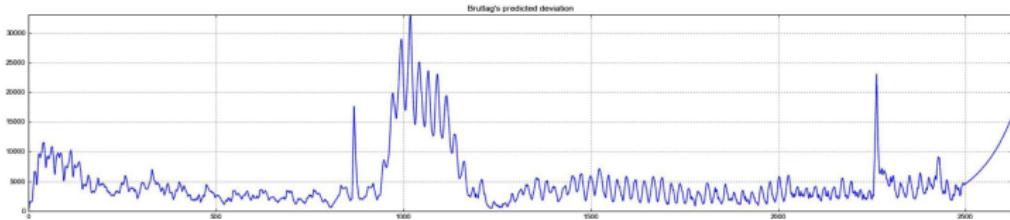
В модели Хольта-Винтерса  $l_t$  (level, intercept) отвечает за “уровень” ряда,  $b_t$  (trend, slope) – за тренд,  $s_t$  – за сезонность. Если сезонности нет, то лучше перейти к двойному экспоненциальному сглаживанию.



Судя по графику, модель неплохо описала исходный временной ряд, уловив недельную и дневную сезонность, и даже смогла поймать аномальные снижения, вышедшие за пределы доверительных интервалов.

# Экспоненциальное сглаживание

Если посмотреть на смоделированное отклонение, хорошо видно, что модель достаточно резко реагирует на значительные изменения в структуре ряда, но при этом быстро возвращает дисперсию к обычным значениям, “забывая” прошлое. Такая особенность позволяет неплохо и без значительных затрат на подготовку-обучение модели настроить систему по детектированию аномалий даже в достаточно шумных рядах.



# Finita!