Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет Компьютерных систем и сетей
Кафедра Информатики

Реферат
на тему:

Задача прогнозирования временных рядов. Основные характеристики временных рядов.

Студент М. С. Петрусевич
Проверил М. В. Стержанов

Минск 2019

Содержание

1 Задача прогнозирования временных рядов

Прогнозирование временных рядов является одним из важных факторов предсказания будущих значений, анализе трендов, циклов и сезонностей в определённых значениях. Для начала, следует рассмотреть само понятие временного ряда.

Временной ряд:

$$Y_1, Y_2 ... Y_t \in \mathbf{R} \tag{1.1}$$

, значения признака, измеренные через постоянные временные интервалы [1].

Ключевая особенность состоит в том, что измерения признака происходит во времени и между разными измерениями всегда проходит одинаковое количество времени. Т.к. если промежуток между отсчётами будет случайным, то в этом случае это будет являться случайным процессом и методы для обработки будут использоваться другие, нежели при работе с прогнозированием временных рядов.

В данном случае, мы рассматриваем прогнозирование вещественного скалярного ряда, т.е. измерения принадлежат множеству вещественных чисел (R).

Как простой пример временного ряда можно рассмотреть временной ряд заработных плат [2].

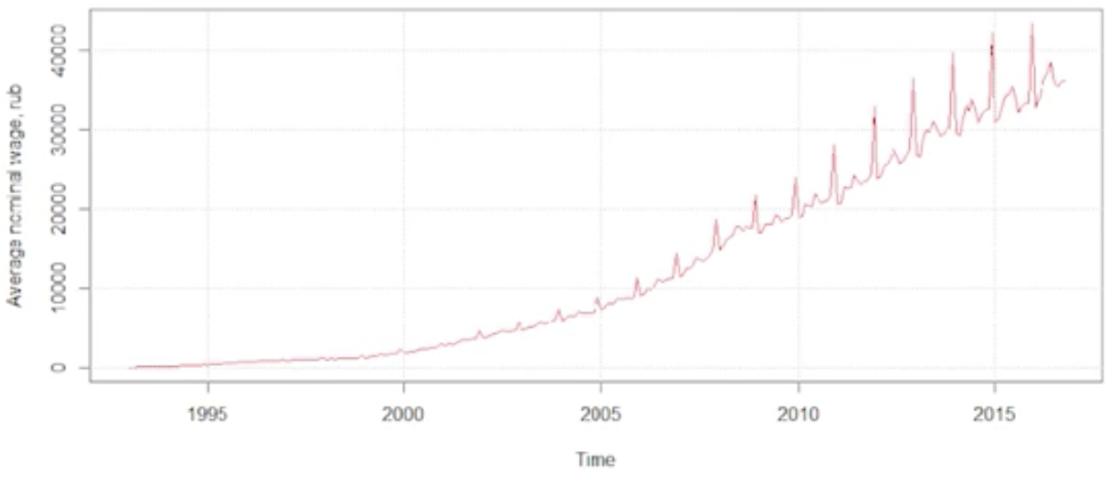

Рисунок 1.1 – Пример временного ряда

На данном рисунке видно, что, есть тренд к росту зарплат и пики, в декабре - месяце, где выдаются годовые премии. Длина ряда не является фиксированной величиной и может меняться, день, неделя, месяц, квартал, год. Сам по себе промежуток не имеет значения при прогнозировании.

Задачей прогнозирования временных рядов является поиск функции F, которая будет будет зависеть от всей известной информации к моменту прогнозирования, который обозначается T. На вход данная функция принимает все значения ряда, от Y_1 до Y_t. Также, функция принимает дополнительный параметр H, который показывает насколько вперёд необходимо прогнозировать ряд. Параметр h принимает значения от 1 до H ($h \in 1, 2, ...H$), где H называется горизонтом прогнозирования.

Помимо прогнозов, которые представляют собой число, при прогнозировании полезно добавлять интервал, который показывает вероятность, с которой будет выполнено предсказание. Такой интервал называется предсказательным.

Предсказательный интервал - интервал, в котором предсказываемая величина окажется с вероятностью не меньше заданной. В данном случае, не стоит его путать с доверительным интервалом, который является случайным интервалом, для фиксированного неслучайного параметра. Предсказательный интервал является очень полезным инструментом, т.к. он показывает заказчику прогноза, насколько можно быть уверенным в произведённом прогнозе. И в данном случае важно, данную степень неуверенности квантифицировать.

Как пример: в апреле 1997 в городе Гранд-Фокс, Северная Дакота, произошло наводнение. Город был защищён дамбой высотой 51 фут, согласно прогнозу, высота паводка должна была составить 49 футов, истинная же высота, оказалась 54. В результате этой ошибки было эвакуировано 75% населения города и нанесён ущерб на несколько миллиардов долларов. На исторических данных, точность прогнозов метеорологической службы составляла +- 9 футов.

Таким образом, выделим особенности задачи прогнозирования временных рядов:

– в классических задачах анализа данных предполагается независимость наблюдений;

– при прогнозировании временных рядов, прогноз строится на исторических данных.

В отличии от задач машинного обучения и статистики, где значения, как правило, являются простой выборкой, т.е. разные наблюдения, померенные на разных объектах, независимые одинаково распределённые. В то время как при прогнозировании временных рядов, данные устроены принципиально по другому - будущее зависит от прошлого, т.е. чем меньше прошлое похоже на шум, тем точнее можно будет сделать прогноз.

Лучше всего, в машинном обучении, решаются задачи с учителем, т.е.

есть определённом количество признаков и выходы, для данных признаков. В данном же случае, прогнозирования временных рядов, отсутсвуют x-ы, т.е. отсутствуют признаки, есть только y-ки [2].

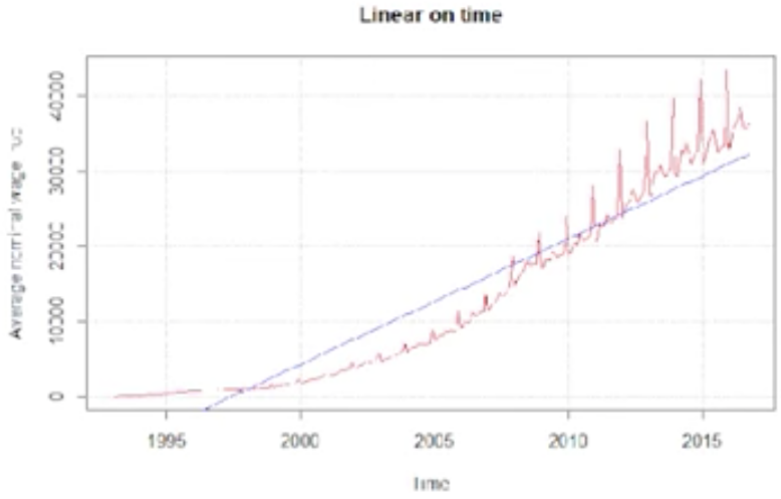

Рисунок 1.2 – Применение методов машинного обучения (линейное)

В данном рисунке видно, что предсказания (синяя линия) не будут достаточно точными, т.к. они явно не учитывают нешумовые всплески [2].

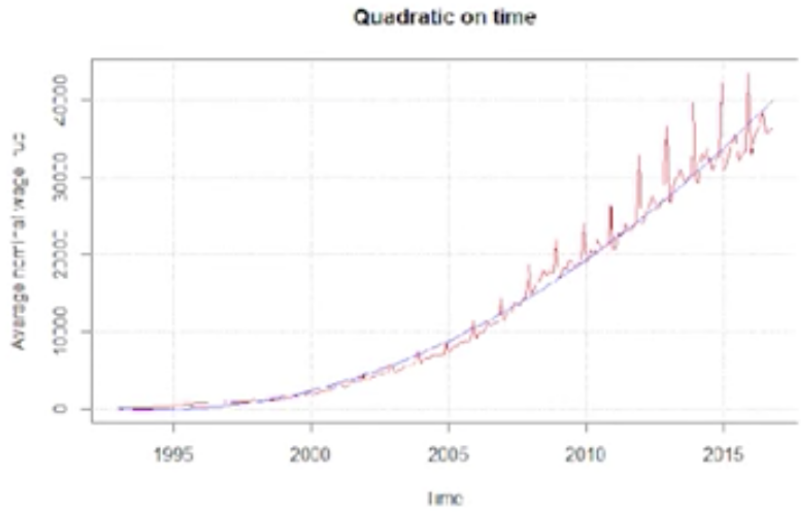

Рисунок 1.3 – Применение методов машинного обучения (квадратичное)

Также и применение квадратичной функции не приводит к точному предсказанию.

Ключевая особенность временного ряда заключается в том, что соседние значения не независимы. Квантифицировать это можно с помощью автокорреляции. Автокорреляция, это корреляция ряда с самим собой, сдвинутым на определённое количество отсчётов. То количество, на которое мы

сдвигаем отсчёт, называется *лагом* автокорреляции. Автокорреляция меняет свои значения от -1 до 1, 1 означает идеальную линейную зависимость с положительным знаком, -1 - линейная зависимость с отрицательным, 0 - отсутствие линейной зависимости.

Также необходимо рассмотреть компоненты временных рядов, т.е. то, из чего состоят ряды.

Тренд - плавное долгосрочное среднего изменения уровня ряда. Ряд может «колебаться» вокруг своего тренда.

Сезонность - циклические изменения уровня ряда с постоянным периодом. Например, если рассматриваются месячные ряды, то в них, скорее всего, будет годовая сезонность, т.е. то, что происходит в декабре этого года, будет похоже на то, что происходило в декабре предыдущего года.

Цикл - изменения уровня ряда с переменным периодом (например экономические циклы, периоды солнечной активности).

Ошибка - непрогнозируемая случайная компонента ряда.

Ошибку, также, можно описать как то, что нельзя описать любыми другими компонентами ряда. Рассмотрим несколько примеров.

В данном ряде можно рассмотреть тренд к падению количества контрактов сокровищницы США по дням [2].

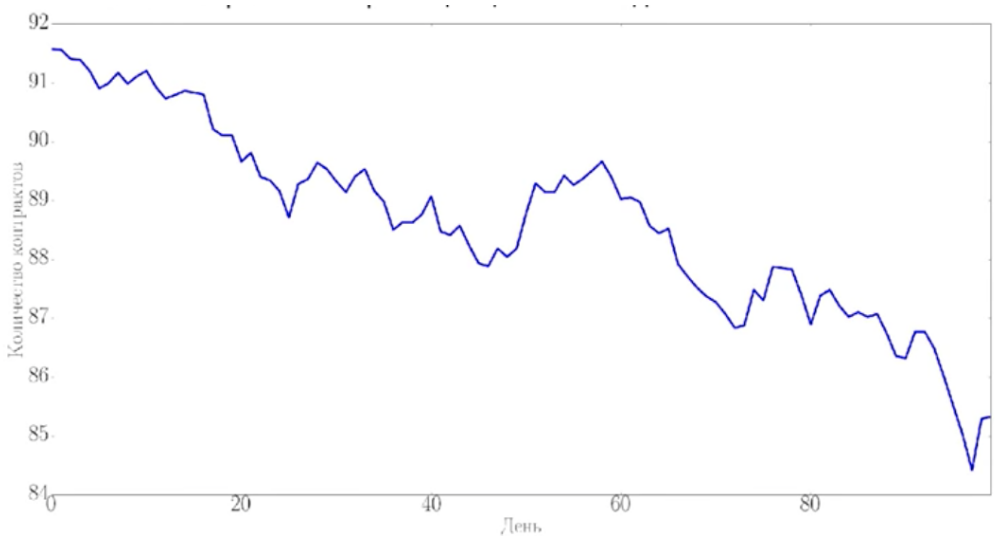

Рисунок 1.4 – Количества контрактов сокровищницы США

Это ряд, в котором можно отметить линейно понижающийся тренд. Можно сказать, что ряд совершает «колебания» вокруг своей линии тренда.

Далее можно рассмотреть ряд с объёмами производства электричества в Австралии [2].

В данном ряде есть ярковыраженный повышающийся тренд и кроме того, сильная годовая сезонность. Хорошо видно, что на графике происхо-

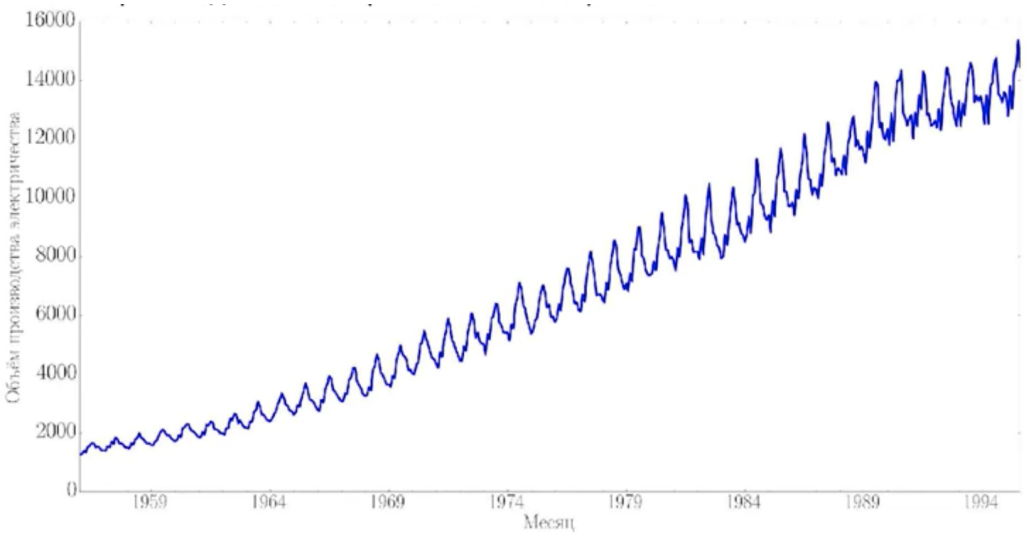

Рисунок 1.5 – Объём производства электричества в Австралии

дят колебания на середину лета - зиму в Австралии, с повышением потребления электричества.

На следующем графике представлен временной ряд продажи жилых домов [2].

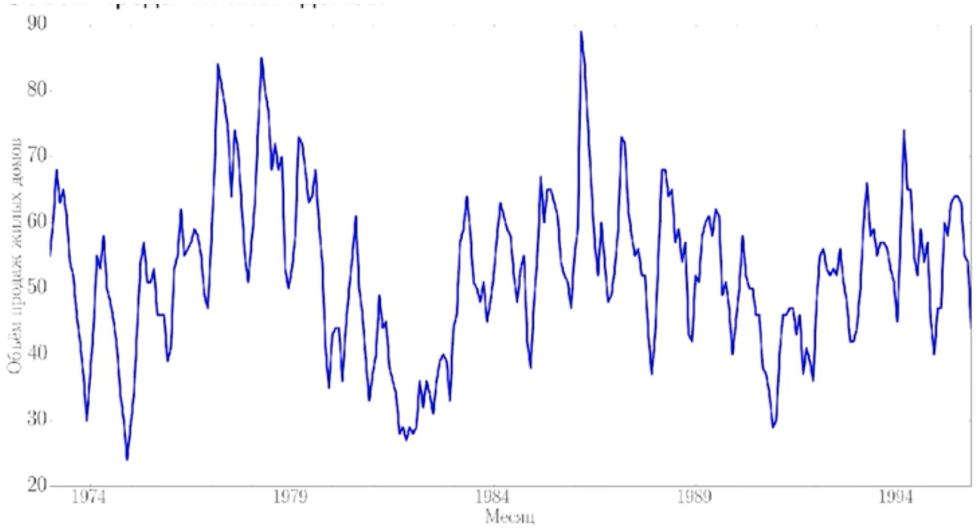

Рисунок 1.6 – Объём продажи жилых домов в США

На данном графике можно заметить годовую сезонность, длиной премерно равной году, и экономические циклы, которые можно отметить спадами и подъёмами объёмов продаж с нефиксированной временной длиной.

2 Автокорреляция во временных рядах

При обработке временных рядов необходимо учитывать наличие автокорреляции и авторегрессии, при которых значения последующего уровня ряда зависят от предыдущих значений [2].

Автокорреляция – явление взаимосвязи между рядами: первоначальным и этим же рядом сдвинутым относительно первоначального положения на h моментов времени.

Авторегрессия – регрессия, учитывающая влияние предыдущих уровней ряда на последующие ряды.

Количественно автокорреляцию можно измерить с помощью линейного коэффициента корреляции между уровнями исходного временного ряда и уровнями этого ряда, сдвинутыми на несколько шагов во времени.

Формула для расчета коэффициента автокорреляции имеет вид:

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \overline{y_1})(y_{t-1} - \overline{y_2})}{\sqrt{\sum_{t=2}^{n}(y_t - \overline{y_1})^2 \sum_{t=2}^{n}(y_{t-1} - \overline{y_2})^2}}, \tag{2.1}$$

где

$$\overline{y_1} = \frac{1}{n-1}\sum_{t=2}^{n} y_t, \overline{y_2} = \frac{1}{n-1}\sum_{t=2}^{n} y_{t-1} \tag{2.2}$$

Эту величину называют коэффициентом автокорреляции уровней ряда первого порядка, так как он измеряет зависимость между соседними уровнями ряда t и y_{t-1}. Аналогично можно определить коэффициенты автокорреляции второго и более высоких порядков. Так, коэффициент автокорреляции второго порядка характеризует тесноту связи между уровнями y_t и y_{t-2} и определяется по формуле:

$$r_2 = \frac{\sum_{t=3}^{n}(y_t - \overline{y_3})(y_{t-2} - \overline{y_4})}{\sqrt{\sum_{t=3}^{n}(y_t - \overline{y_3})^2 \sum_{t=3}^{n}(y_{t-2} - \overline{y_4})^2}}, \tag{2.3}$$

где

$$\overline{y_3} = \frac{1}{n-1}\sum_{t=3}^{n} y_t, \overline{y_4} = \frac{1}{n-1}\sum_{t=3}^{n} y_{t-2}. \tag{2.4}$$

Следует отметить, что $r_\tau \in [-1..1]$

Сдвиг между соседними уровнями или сдвинутыми на любое число периодов времени называютвременным лагом. С увеличением лага число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается. Считается целесообразным для обеспечения статистической

достоверности коэффициентов автокорреляции использовать правило – максимальный лаг должен быть не больше $\frac{n}{4}$.

Можно отметить следующие свойства автокорреляции [3]:

– коэффициент корреляции строится по аналогии с линейным коэффициентом корреляции и таким образом характеризует тесноту только линейной связи текущего и предыдущего уровней ряда. Поэтому по коэффициенту автокорреляции можно судить о наличии линейной (или близкой к линейной) тенденции. Для некоторых временных рядов, имеющих сильную нелинейную тенденцию (например, параболу второго порядка или экспоненту), коэффициент автокорреляции уровней исходного ряда может приближаться к нулю;

– по знаку коэффициента автокорреляции нельзя делать вывод о возрастающей или убывающей тенденции в уровнях ряда. Большинство временных рядов экономических данных содержат положительную автокорреляцию уровней, однако при этом могут иметь убывающую тенденцию.

Последовательность коэффициентов автокорреляции уровней первого, второго и т.д. порядков называют автокорреляционной функцией временного ряда. График зависимости ее значений от величины лага (порядка коэффициента автокорреляции) называется коррелограммой.

Анализ автокорреляционной функции и коррелограммы позволяет определить лаг, при котором автокорреляция наиболее высокая, а следовательно, и лаг, при котором связь между текущим и предыдущими уровнями ряда наиболее тесная, т.е. при помощи анализа автокорреляционной функции и коррелограммы можно выявить структуру ряда.

Если наиболее высоким оказался коэффициент автокорреляции первого порядка, исследуемый ряд содержит только тенденцию. Если наиболее высоким оказался коэффициент автокорреляции порядка τ , то ряд содержит циклические колебания с периодичностью в τ моментов времени. Если ни один из коэффициентов автокорреляции не является значимым, можно сделать одно из двух предположений относительно структуры этого ряда: либо ряд не содержит тенденции и циклических колебаний, либо ряд содержит сильную нелинейную тенденцию, для выявления которой нужно провести дополнительный анализ. Поэтому коэффициент автокорреляции уровней и автокорреляционную функцию целесообразно использовать для выявления во временном ряде наличия или отсутствия трендовой компоненты и циклической (сезонной) компоненты.

Для примера применения автокорреляции временного ряда можно рассмотреть следующие графики [2]:

На данном графике представлен временной ряд, показывающий объ-

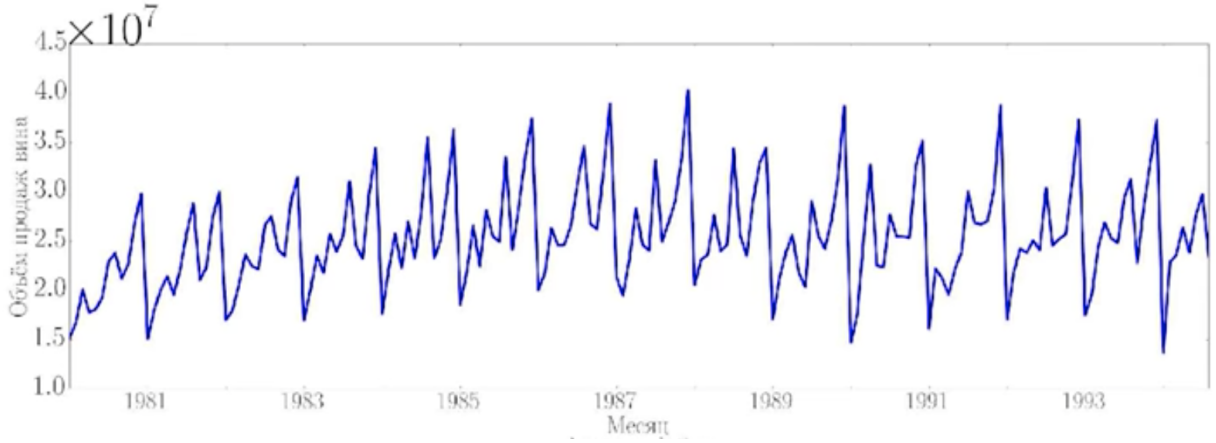

Рисунок 2.1 – Временной ряд представляющий продажи вина в Австралии

ёмы продаж вина в Австралии. Как видно, на нём присутствует ярковыраженная сезонность, выпадающая на декабрь каждого года, сопоставленный с католическим рождеством.

На следующем графике будет представлена автокорреляционная функция для этого временного ряда [2].

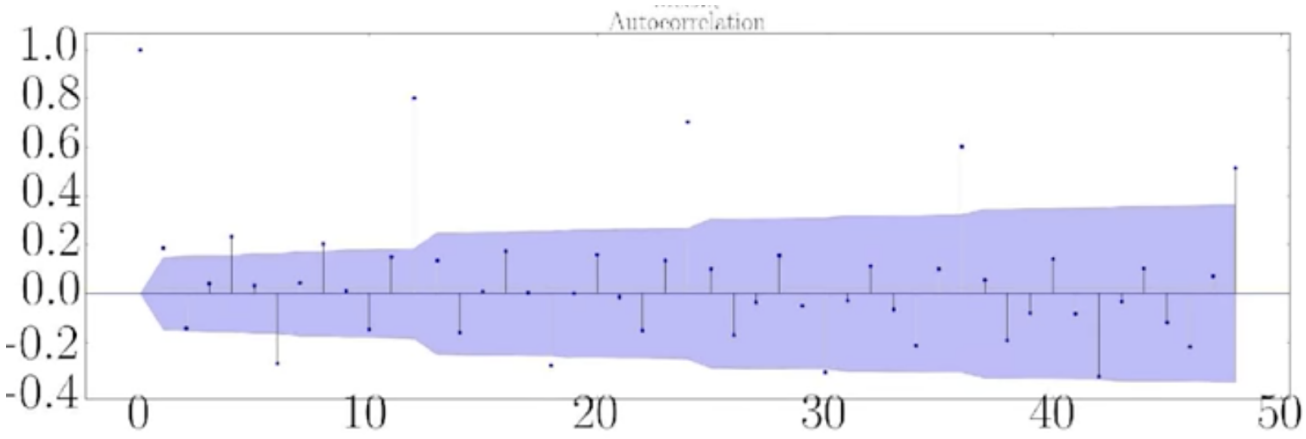

Рисунок 2.2 – График автокорреляционной функции

Автокорреляционная функция - это график автокорреляции при разных значениях лага.

На рисунке 2.2 каждый отрезок это значение автокорреляции. На графике видно, что есть пики на лагах кратных длине сезонного периода. Есть большая автокорреляция при значениях лага кратных 12 (12, 24, 36 и т.д.).

В ряде с контрактами сокровищницы США, рассмотренными ранее (1.4), если построить её график автокорреляционной функции можно увидеть следующую структуру [2]:

Можно отметить, что типичная структура автокорреляции у которой «сильный» тренд, т.е. существует большая автокорреляция при малых лагах

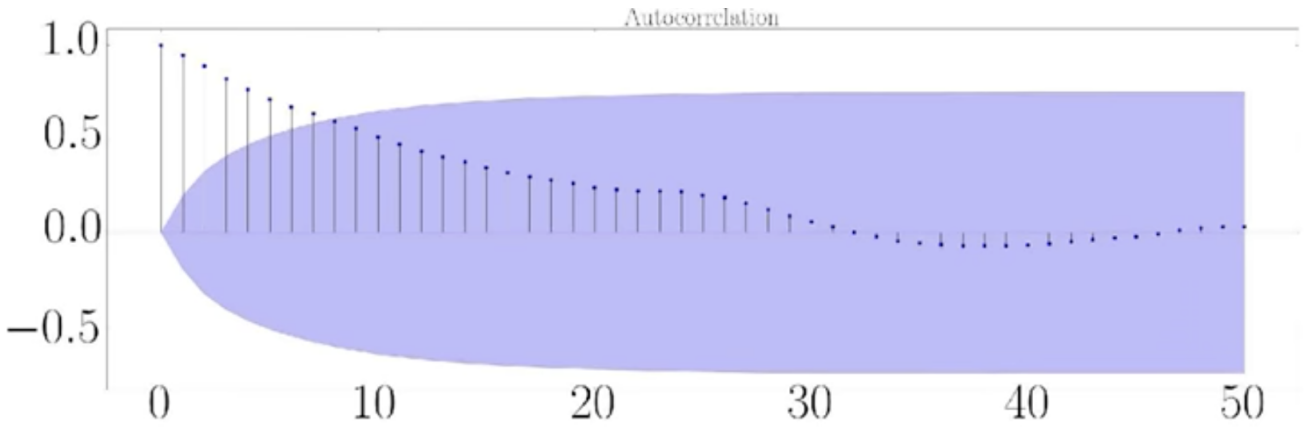

Рисунок 2.3 – График автокорреляционной функции временного ряда количества контрактов сокровищницы США

и она постепенно убывает и начинает синосуидально «колебаться» вокруг 0.

Стоит также рассмотреть график автокорреляции ранее описанного графика использования электричества в Австралии (1.5) [2].

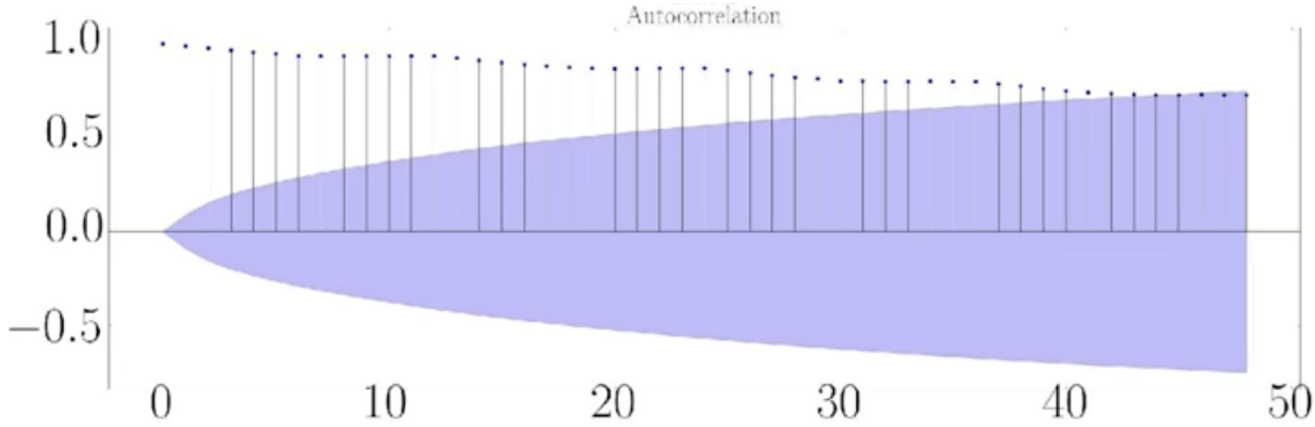

Рисунок 2.4 – График автокорреляционной функции временного ряда потребления электричества в Австралии

В этом ряде есть тренд и сезонность. Когда в ряде есть сезонность на лагах кратных сезонному периоду должны быть пики, однако, на данном графике их сложно разглядеть из за того, что в исходном ряде присутствует сильный тренд, который «забивает» общую картину.

Далее будет рассмотрен более сложный пример, на ранее рассмотренном ряде продажи жилых домов в США (1.6) [2].

На данном графике можно увидеть типичная автокорреляция для ряда, у которого присутствует сезонность и циклы. Циклы, в данном случае, приводят к тому, что постепенно происходит сдвиг местонахождения пи-

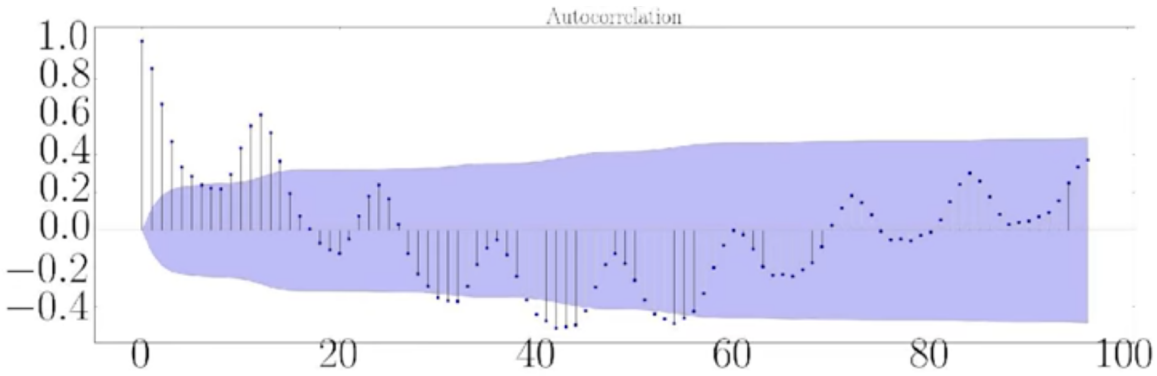

Рисунок 2.5 – График автокорреляционной функции временного ряда продажи домов в США

ка сезонного периода на некратное положение. В данном случае, первый пик находится на лаге равном 12, следующие не будет находится, в данном случае, не на 24. И так, постепенно происходит сдвиг пиков. Именно так циклы влияют на график автокорреляции - меняя положения пика сезонного периода.

И в заключение автокорреляционных графиков, стоит отметить график функции автокорреляции временного ряда с дневными значениями индекса Доу-Джонса [2].

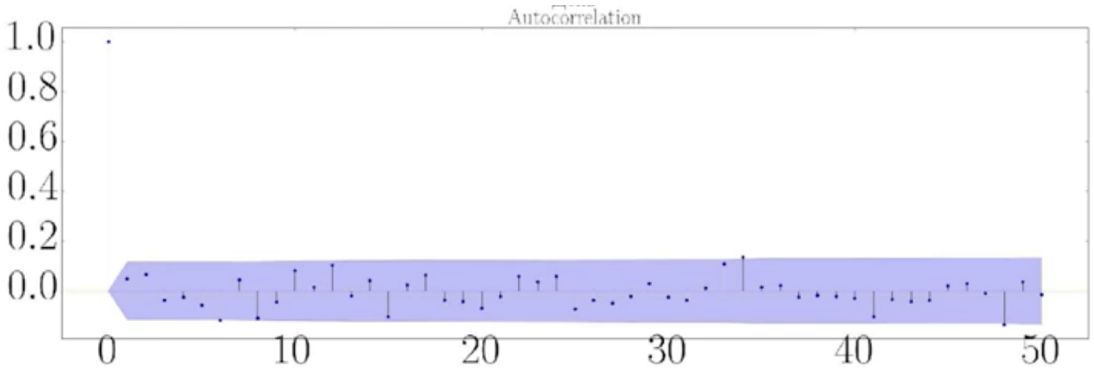

Рисунок 2.6 – График автокорреляционной функции временного ряда значений индекса Доу-Джонса

Данный график, показывает отсутствие какого либо наличия тренда, сезонов или циклов. В рассматриваемом примере, как и в любом примере, если даже сгенерировать для него шум, значения автокорреляционной функции будет в районе 0, что также можно отметить и на рисунке 2.6.

Значимость автокорреляции при каком-то фиксированном лаге можно проверить при помощи специальных критериев (например при помощи Q-критерия Льюинга-Бокса).

3 Стационарность временного ряда

Очень важным свойством временного ряда, помимо автокорреляции, является его стационарность.

Ряд $y_1, y_2 ... y_T$ стационарен, если любое распределение $y_t, ..., y_{t+s}$ независит от t, т.е. его свойства не зависят от времени. Т.е. если у нас есть окно производной длины s, то если мы его поставим в начале ряда, там будет совместное распределение y-ков, такое же, как и в конце этого ряда. Также, стоит упомянуть, что существуют и другие виды стационарности, например стационарность по среднему [4].

Про стационарность можно отметить, на основе тех компонентов рядов, которые мы выделили:

– ряды с трендами - нестационарны;

– ряды с сезонностью - нестационарны;

– наличие циклов в ряде не могут точно сказать, стационарен ряд или нет.

В общем случае, чтобы проверить - стационарен ли временной ряд, можно использовать статистические критерии. Существует множество данных критериев, далее будет рассмотрен один из них.

3.1 Критерий проверки стационарности - KPSS

Критерий проверки стационарности ряда называемый KPSS проверяет гипотезу о том, что ряд стационарен, против альтернативы, что он нестационарен [5].

Статистика для данного критерия выглядит следующим образом:

$$KPSS = \frac{1}{T^2} \sum_{i=1}^{T} (\sum_{t=1}^{i} y_t)^2 / \lambda^2, \tag{3.1}$$

при нулевой гипотезе, имеет табличное распределение (т.е. оно никак не выражается аналитически).

3.2 Критерий Дики-Фуллера

Следующим критерием, который будет рассморен является критерий Дики-Фуллера. Он устроен ровно наоборот. Критерий Дики-Фуллера проверяет гипотезу о том, что ряд нестационарен, против гипотезы, что он стационарен. Можно отметить, что таким образом устроено большинство статистических критериев работающих со стационарностью.

4 Вывод

В результате изучения такой сферы как предсказания временных рядов были рассмотрены основные характеристики временных рядов, которые используются при их прогнозировании. Можно сказать, что задача прогнозирования, как и любая другая задача, возникающая в процессе работы с данными — во многом творческая и уж точно исследовательская. Несмотря на обилие формальных метрик качества и способов оценки параметров, для каждого временного ряда часто приходится подбирать и пробовать что-то своё.

На данный момент классические методы предсказания могут быть заменены более современными методами машинного обучения. Однако это не означает, что описанные выше методы и подходы не подходят для предсказания временных рядов. Данные подходы позволяют обеспечивать достаточно точный результат прогнозирования временных рядов, учитывающий сезонности, циклы и тренды, максимально уменьшая влияние ошибки на предсказание.

Предсказания временных рядов позволяет делать достаточно точные прогнозы, которые способны удовлетворять требованиям бизнеса, такие как предсказания трендов в экономике, для прогноза сколько серверов понадобится online-сервису через год, каков будет спрос на каждый товар в гипермаркете, или для постановки целей и оценки работы команды.

Для прогноза временных рядов могут использоваться как ручные расчёты, так и заранее подготовленные библиотеки в программирование, например библиотека statsmodels в языке Python. Данный модуль предоставляет широкий набор средств и методов для проведения статистического анализа и эконометрики. В целом для небольших исследований пакет statsmodels может быть использован и получить, с достаточной долей точности.

Таким образом - использование предсказания временных рядов используя всё описанное выше достаточно легко может производиться современными средствами программирования и решать множество задач, где необходимо предсказать какие-либо события, где имеются заранее подготовленные замеры с фиксированной длиной промежутков замеров.

Список использованных источников

[1] Wikimedia. Временной ряд. https://ru.wikipedia.org/wiki/\T2A\CYRV\T2A\cyrr\T2A\cyre\T2A\cyrm\T2A\cyre\T2A\cyrn\T2A\cyrn\T2A\cyro\T2A\cyrishrt_\T2A\cyrr\T2A\cyrya\T2A\cyrd.

[2] MIPT. Data Mining In Action. — 2016. https://github.com/vkantor/MIPT_Data_Mining_In_Action_2016.

[3] R.J., Hyndman. Теория и практика в R. — 2018. https://www.otext.org/book/fpp.

[4] Website, Studopedia. Studopedia Website. — 2016. https://studopedia.org/3-10623.html.

[5] Анализ временных рядов. — 2011. http://statsoft.ru/home/textbook/modules/sttimser.html.