Работа с временными рядами

Высшая Школа Цифровой Культуры
Университет ИТМО
dc@itmo.ru

Содержание

1 Временные ряды

1.1 Анализ временных рядов

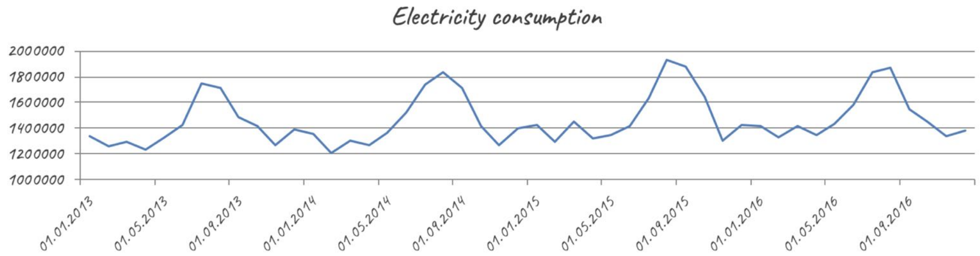
Временной ряд – последовательность наблюдений, упорядоченная по времени:

$$y_1, y_2, \dots, y_n$$

Где y_i – значения переменной в n равноотстоящих моментов времени:

$$t_1, t_2 \dots, t_n$$

Примерами временных рядов являются регулярно фиксируемые (каждый день, час, минуту и т.п.) данные о генерации электроэнергии, производстве продукции, продаже товаров, потреблению продукции, транспортных перевозках и т.п. Пример процесса, заданного временным рядом показан на рисунке.



Особенность анализа временных рядов заключается в том, что в отличие от других, независимых наблюдений, мы рассчитываем, что значения ряда в прошлом содержат сведения о его поведении в будущем.

В целом, задачи, с которыми сталкивается исследователь в процессе анализа временных рядов, можно разделить на два класса:

- анализ характерных **составляющих** временного ряда для понимания процессов, отображаемых рядом;
- **прогнозирование** поведения ряда в будущие периоды времени.

Рассмотрим последовательно эти задачи. Под составляющими временного ряда принято понимать следующее:

- **Тренд** – плавное, долгосрочное изменение уровня ряда.
- **Сезонность** – циклические изменения уровня ряда с постоянным периодом (например, ежемесячному потреблению электроэнергии соответствует период размером в 12 месяцев).

- **Цикл** – изменения уровня ряда с переменным периодом (например, экономические циклы, периоды солнечной активности 5-7 лет и т.п.).

- **Шум** – непрогнозируемая, случайная компонента ряда.

На рисунке приведено несколько примеров рядов с указанием визуально обнаруженных в них компонент.

Под прогнозированием ряда понимается построение (моделирование) такой функции f, которая на основе значений временного ряда $y_1, y_2, \ldots, y_t$ и дополнительного параметра h выдает прогнозное значение ряда для точек $t + h$:

$$f(y_1, y_2, \ldots, y_t, h) = \widehat{y}_{t+h},$$

где параметр h – это значение в интервале от 1 до H, H – **горизонт прогнозирования**. В зависимости от значения горизонта прогнозирования модели делят на **краткосрочные** и **долгосрочные**. Однако временная градация прогнозов является условной и чаще всего зависит от особенностей временного ряда.

В процессе решения упомянутых выше задач исследователь сталкивается с рядом проблем, которые следует учитывать в процессе анализа. Перечислим некоторые из них.

- Не всегда просто выявить скрытые закономерности в истории ряда (например, оценить продолжительность периодов, подобрать подходящую аналитическую функцию для тренда и т.п.).

- Закономерности (если они, действительно, есть) могут быть искажены шумами, присутствующими в данных. Такие искажения особенно характерны для данных, получаемых со всякого рода датчиков. Именно поэтому в анализе разработаны методы предварительной подготовки данных, ориентированные на удаление шумов с помощью специальных методов «сглаживания».

- Развитие динамики ряда в прошлом не гарантирует аналогичное поведение ряда в будущем, которое может значительно измениться под влиянием разного рода внешних факторов (например, динамика цен на нефть кардинально меняется при принятии решений об изменении квот на добычу нефти, смене правительств в нефтедобывающих странах и т.п.).

И, тем не менее, не смотря на указанные выше проблемы, строить модели временных рядов можно и нужно, так как даже с учетом упомянутых рисков они оказываются полезными для развития бизнеса в различных областях экономики. Хотя надо заметить, что в статистике существуют и более точные математические методы, позволяющие оценить, так называемый, **предсказательный интервал**, позволяющий определить диапазон, в котором предсказываемая величина окажется с вероятностью не меньше заданной.

Аддитивная модель

$$y_t = u_t + s_t + e_t$$

Мультипликативная модель

$$y_t = u_t \cdot s_t \cdot e_t$$

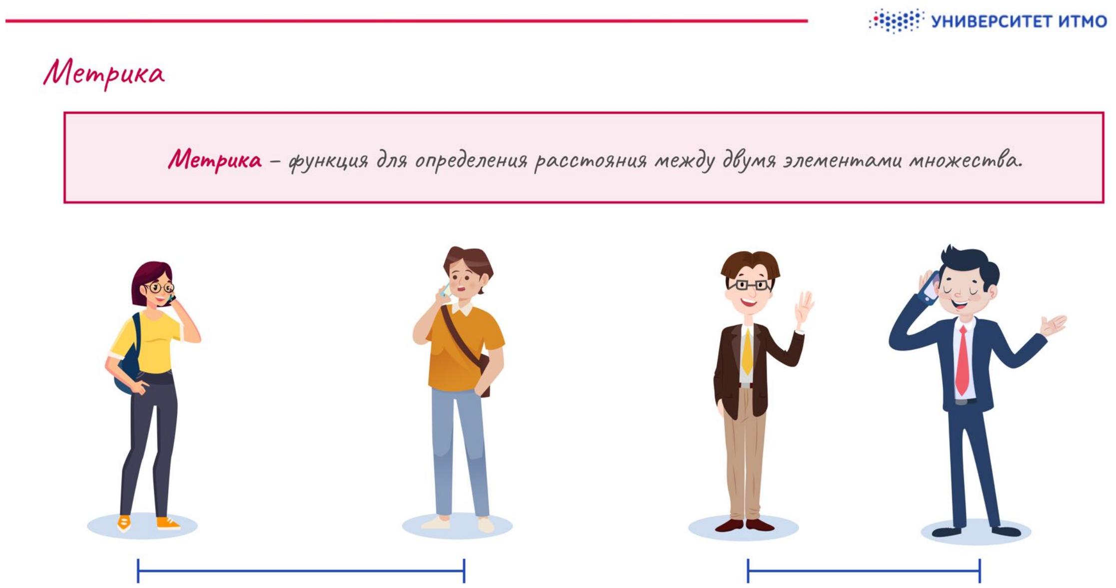
u_t – трендовая составляющая,

s_t – сезонная составляющая,

e_t – случайная составляющая.

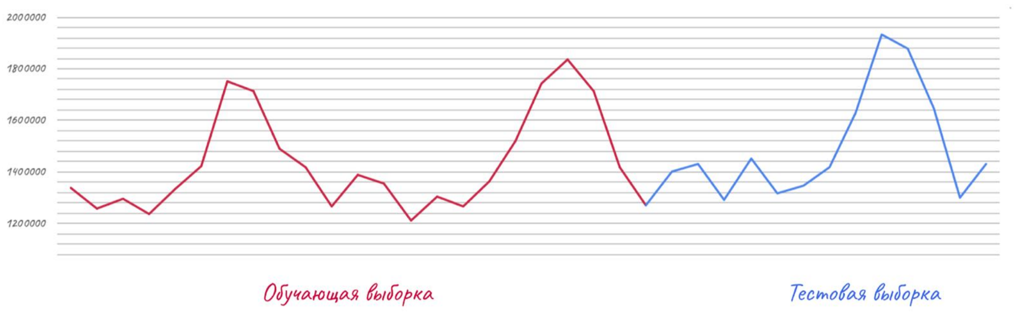
При построении моделей временного ряда принято различать два принципиально различающихся подхода: **аддитивный** и **мультипликативный**. Аддитивная модель предполагает прогнозирование ряда путем рекуррентного прибавления или вычитания некоторых приращений к известным значениям временного ряда. Мультипликативная модель предполагает прогнозирование ряда путем умножения известных членов ряда некоторые коэффициенты. Например, при построении аддитивной модели было определено, что среднемесячное увеличение спроса на некоторый товар составляет 100 единиц. Тогда прогнозное значение спроса в следующем месяце определяется как предыдущее значение ряда плюс 100 единиц. В мультипликативной модели увеличение спроса могло бы быть определено как повышение спроса на 10 процентов. Тогда прогнозное значение спроса в следующем месяце вычислялось бы как предыдущее, умноженное на 1.1. Сезонная закономерность, повторяющаяся в ряду с определенной периодичностью, также может быть определена аддитивным или мультипликативным образом.

В рамках данной лекции мы рассмотрим простейшие примеры моделирования временных рядов, однако прежде, чем мы перейдем к приемам построения моделей ряда, необходимо обсудить, как можно оценить качество построенной модели временного ряда и как можно сравнивать построенные модели между собой. Для такого рода сравнений принято использовать, так называемые, метрики качества. **Метрикой** называют функцию для определения расстояния между двумя элементами множества. Таких метрик существует достаточно много, в данном фрагменте лекции мы рассмотрим только некоторые из них.



1.2 Метрики качества

Но прежде, чем определяться с выбором метрики, необходимо понять к каким данным они будут применяться. Для того, чтобы оценить качество прогноза, нам нужны не только предсказанные значения, но и реальные данные временного ряда. Причем строить модель для проверки качества нужно на одной части данных, а проверять – на второй. Данные, на которых строится модель, принято называть обучающей выборкой, а данные, на которых модель проверяется, тестовой выборкой. Как поделить временной ряд на обучающую и тестовую выборки? В общем случае, это вопрос может решаться многими способами, однако, в случае временного ряда самое простое и адекватное решение – разделить ряд на три части, а затем первые две использовать для построения модели как обучающую выборку, а оставшуюся часть – для проверки качества построенной модели, т.е. как тестовую.



Теперь можно обсудить и сами метрики. Большинство метрик качества основано на понятии **ошибки прогноза** (будем в дальнейшем ее обозначать как e_t). Ошибка прогноза в момент времени t – это разность между предсказанным и реальным значением переменной в момент времени t, т.е.

$$e_t = \widehat{y}_t - y_t,$$

где $\widehat{y}_t$ – предсказанное значение, y_t – реальное значение переменной.

Все приведенные далее метрики основаны на предположении, что если модель обеспечивает небольшие суммарные ошибки в прошлом, то она же обеспечит небольшие суммарные ошибки в будущем.

Первая из метрик (MAE) с названием **средняя абсолютная ошибка**, получается как результат деления суммы абсолютных значений ошибок прогноза на количество точек тестовой выборки:

$$\mathsf{MAE} = \frac{1}{n}\sum_{t=1}^{n} |e_t|$$

Вторая (MSE) – **среднеквадратичная ошибка**, вычисляется как сумма квадратов ошибок прогноза, деленная на количество точек тестовой выборки:

$$\mathsf{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$$

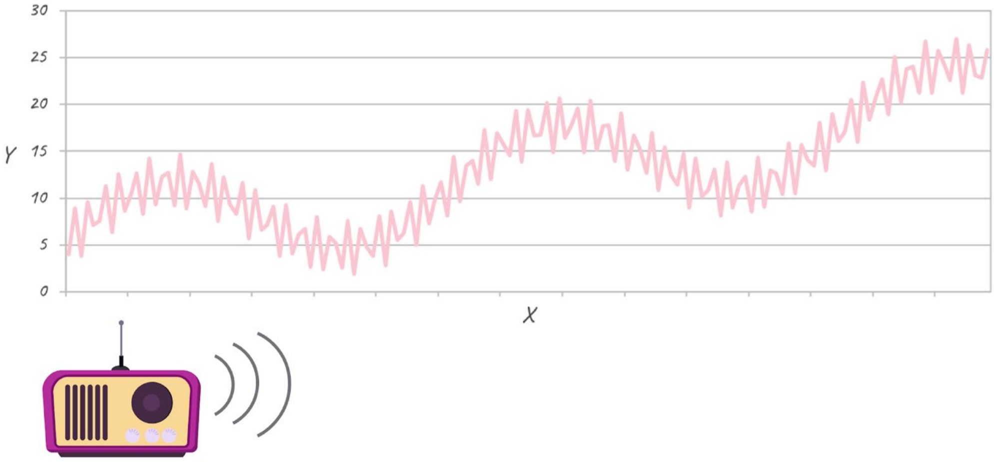
Третья (MAPE) – **средняя абсолютная процентная ошибка**, определяется как процентное соотношение суммы отношений ошибок прогноза и реальных значений временного ряда к количеству точек тестовой выборки. В приведенных формулах: e_t – ошибка прогноза, y_t – реальное значение переменной, n – количество точек тестовой выборки:

$$\mathsf{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{e_t}{y_t}\right| \cdot 100\%$$

Именно эти метрики мы и будем использовать в дальнейшем для оценки моделей временных рядов. При подборе подходящей модели будем ориентироваться на метрики с минимальными значениями.

1.3 Сглаживание временных рядов

Мы уже упоминали в предыдущем фрагменте лекции, что некоторые значения временных рядов активно сопровождаются шумами, т.е. случайными вариациями в той или иной форме. Как правило, это относится к рядам, которые формируются на основе показаний разного рода датчиков. Пример такого зашумленного ряда вы можете видеть на экране. Шумы мешают пониманию реальной структуры ряда и интерпретации процессов, образующих этот ряд, и поэтому существует множество методов, позволяющих избавиться от шумов. В этом фрагменте лекции мы рассмотрим несколько наиболее известных и применяемых методов, используемых на практике для удаления шумов.

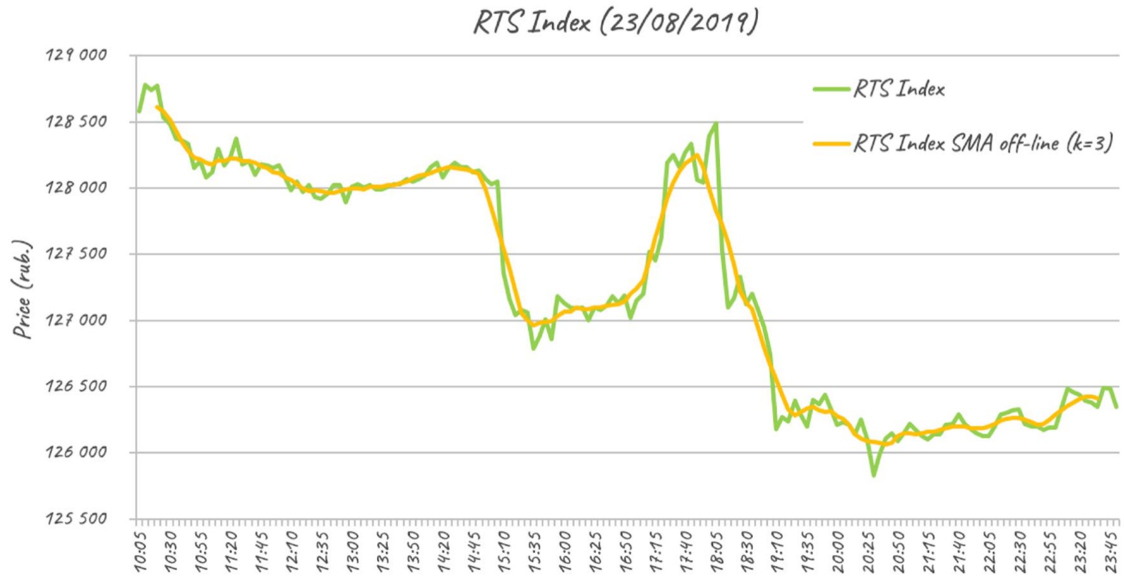


Первый из методов – **метод скользящего среднего**. Суть метода сводится к тому, что для каждого значения переменной ряда формируется окно из соседних значений ряда (в идеале k значений до сглаживаемого значения и k после), а затем на основании этих соседей и самого исходного значения ряда вычисляется среднее арифметическое). Формулу, по которой происходит сглаживание:

$$s_i = \frac{1}{2k+1} \sum_{j=-k}^{k} y_{i+j}$$

Здесь y_i – значение исходного ряда, s_i – значение сглаженного ряда, $2k+1$ – ширина окна. От ширины окна зависит степень сглаживания. При задании большой ширины окна сглаживание будет грубым и, возможно, будет потеряна полезная информация о динамике ряда. При задании небольшого окна в 5-7 точек могут остаться шумы. Универсальные значения для ширины скользящего окна задать невозможно – они сильно зависят от предметной области и от целей усреднения в каждом конкретном случае.

Сглаживание активно используется на этапе технического анализа биржевых котировок и встроено во все инструменты, предназначенные для биржевой аналитики. На экране приведен пример исходного и сглаженного графика котировок индекса RTS с применением метода скользящего среднего при значении k равном 3. Обратите внимание на то, что особенности расчета скользящего среднего не позволяют рассчитать сглаженное значение для k первых и k последних точек ряда. Приведенная выше формула расчета

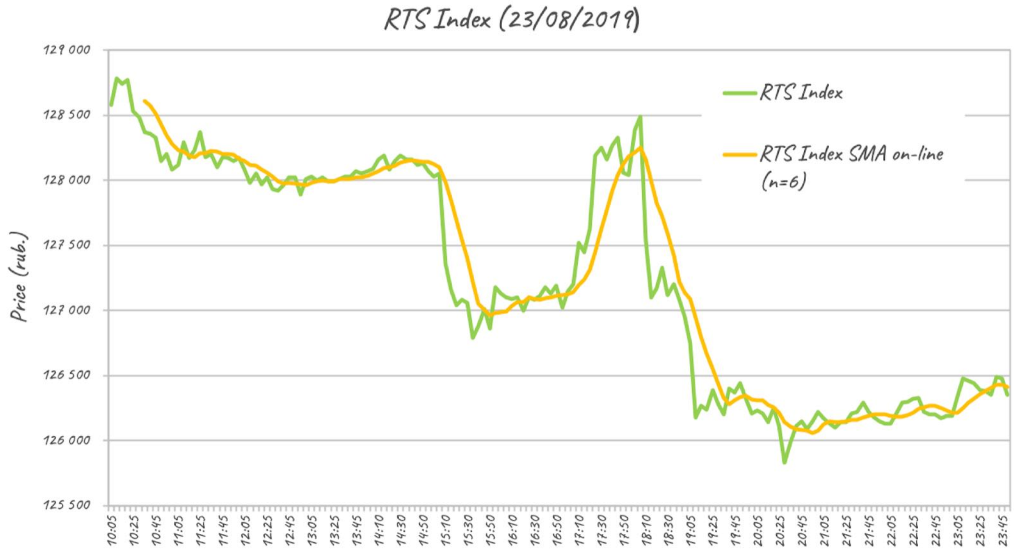


скользящего среднего подразумевает, что сглаживание происходит в режиме off-line (т.е. когда известны и предыдущие и следующие значения для каждой сглаживаемой точки ряда). Однако на практике иногда приходится сталкиваться с необходимостью сглаживания ряда в режиме on-line (когда известны только предыдущие значения и сама сглаживаемая точка ряда). Именно такое сглаживание применяется в биржевой деятельности при оперативном анализе котировок. Можно ли в режиме on-line применять метод скользящего среднего?

Можно. Только формула для расчета в этом случае будет указывать соседей, которые появились до сглаживаемого значения:

$$s_i = \frac{1}{n+1} \sum_{j=-n}^{0} y_{i+j},$$

где y_i – значение исходного ряда, s_i – значение сглаженного ряда, $n+1$ – ширина ряда. Для такого сглаживания характерно запаздывание в отображении сглаженного ряда, но, тем не менее, для некоторых задач этого метода сглаживания вполне достаточно. Для сглаживания в режиме on-line часто используется еще один популярный метод сглаживания. Это – **экспоненциальный метод сглаживания**. Он чрезвычайно прост в реализации, так как описывается рекуррентной формулой следующего вида.



Пусть $Y = \{y_1, \ldots, y_T\}$ – временной ряд. **Экспоненциальное** сглаживание ряда осуществляется по рекуррентной формуле:

$$s_t = \alpha y_t + (1-\alpha)s_{t-1},$$

где y_t – значение исходного ряда в точке t, s_{t-1} – значение сглаженного ряда в точке $t-1$, $\alpha \in (0,1)$ – коэффициент сглаживания. При этом начальное значение s_1 определяется как первая точка ряда:

$$s_1 = y_1.$$

Выбор коэффициента сглаживания α является решающим моментом при экспоненциальном сглаживании. Текущее сглаженное значение складывается из предыдущего и некоторой доли ошибки предыдущего сглаживания. Величина этой ошибки, которая используется для корректировки, определяется коэффициентом сглаживания α. Чем ближе значение α к 1, тем большая часть расхождения сглаживания и реального значения считается закономерной и используется для вычисления очередного значения. Чем ближе значение α к нулю, тем большая доля расхождения между сглаженным и реальным значением считается случайной и, соответственно, меньшая часть используется для вычисления очередного значения.

Формула для экспоненциального сглаживания может быть переписана в нереккурентном виде:

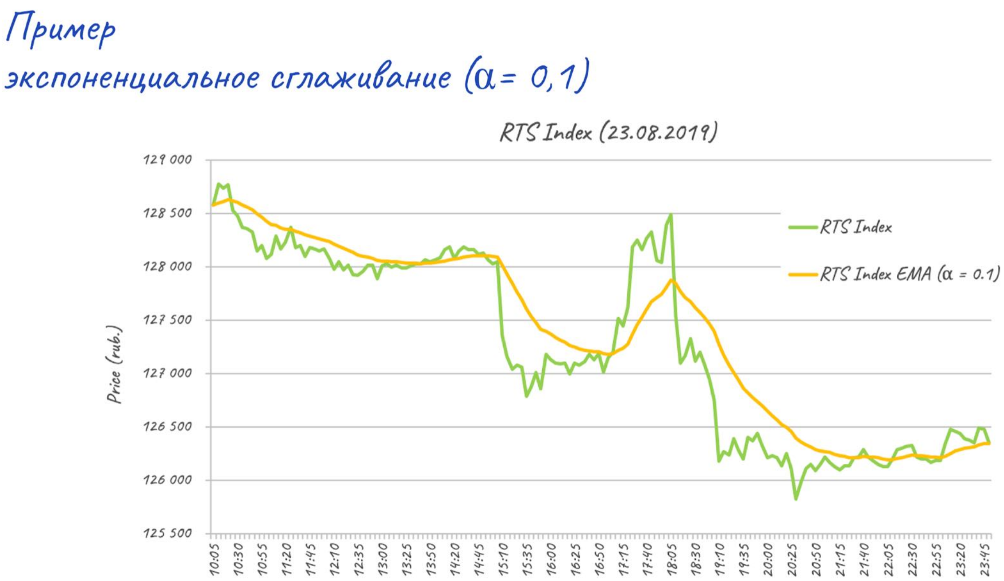
$$\mathbf{s}_t = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \alpha(1-\alpha)^3 y_{t-3} + \ldots$$

Из формулы на экране хорошо видно, что сглаженное значение представляет собой взвешенную сумму всех предыдущих значений ряда, причем коэффициенты уменьшаются по мере удаления значения ряда от текущего момента времени. Так, например, если $\alpha = 0.1$, то формула приобретает следующий вид:

$$s = 0.1y_t + 0.09y_{t-1} + 0.081y_{t-2} + 0.0729y_{t-3} + \ldots$$

Как правильно выбрать коэффициент сглаживания? Не существует четких формальных критериев выбора значения a. На практике (по крайней мере в биржевой деятельности) чаще всего используются значения α, лежащие в пределах от 0.1 до 0.3. Можно сказать, что значение коэффициента сглаживания отражает субъективное мнение исследователя относительно устойчивости изменения изучаемого показателя.

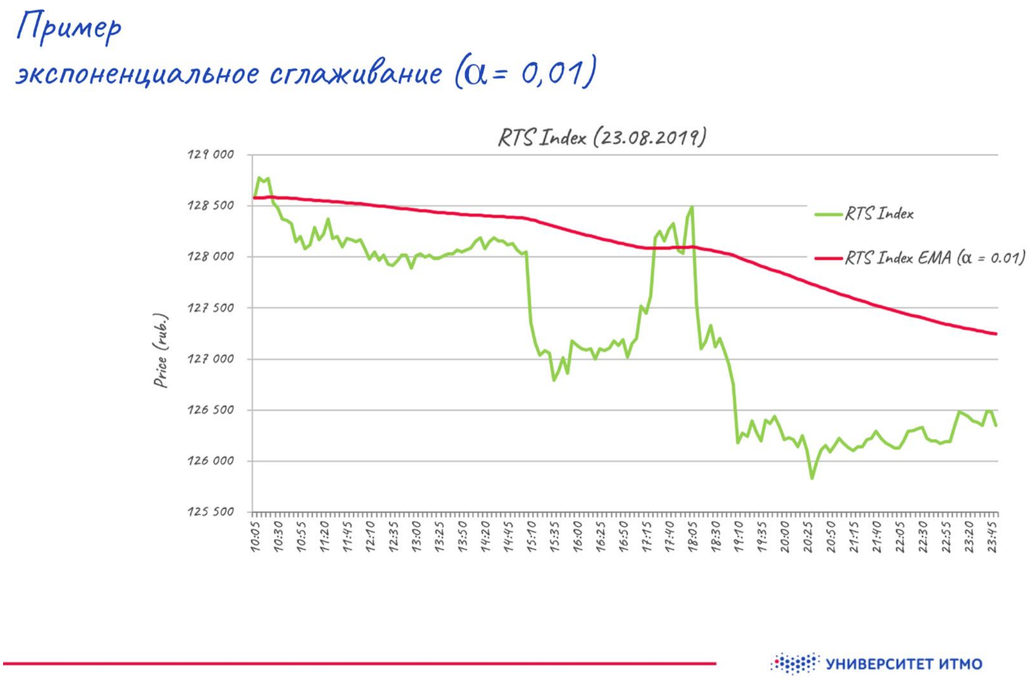
На рисунке приведены примеры экспоненциального сглаживания индекса RTS с различными значениями коэффициента сглаживания. На первом графике коэффициент $\alpha = 0.1$. На втором графике коэффициент α в 10 раз

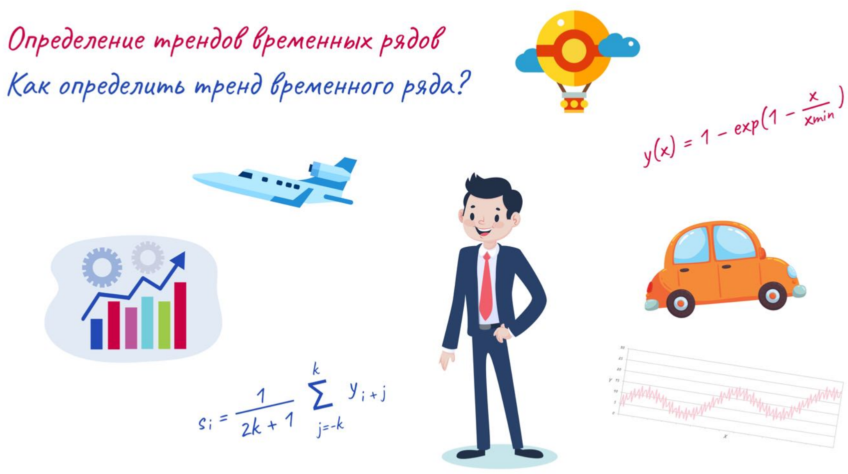


меньше – 0.01. Обратите внимание, как сильно повлиял выбор коэффициента сглаживания на результат.

1.4 Определение трендов временных рядов

В предыдущем фрагменте мы рассмотрели основные составляющие типичного временного ряда – тренд, сезонная составляющая и шумы. Мы уже обсудили, что в случае наличия шумов они могут быть удалены при предварительном анализе ряда с помощью специальных методов. И очистка от шума позволит лучше увидеть динамические тенденции ряда. Как выделить тренд и сезонную составляющую? Можно ли их определить аналитически (т.е. с помощью математической функции, зависящей от времени)? Что это дает? Если мы научимся описывать аналитически поведение временного ряда – мы сможем в дальнейшем прогнозировать поведение ряда. С точки зрения





реального бизнеса это значит, что мы можем планировать продажи автомобилей, количество пассажиров, потребителей различных услуг, посетителей ресторанов и т.п. А вот это – действительно интересно!

Итак, как определить тренд временного ряда? На практике для этого, как правило, используются следующие аналитические функции: линейная, полиномиальная, экспоненциальная, логарифмическая:

Разумеется, другие функции тоже возможны, но именно эти функции используются чаще других и встроены во многие существующие инструменты. Как определить, какая функция подходит в том или ином случае? Конечно, для этого существуют и формальные методы, но простейший способ – вспомнить, как выглядят графики соответствующих функций и по графику временного ряда подобрать подходящую по виду функцию. Приведем несколько примеров.

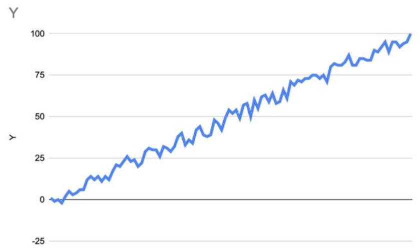
Однако для того, чтобы в дальнейшем моделировать поведение ряда, *Линейная* $f(x) = a + bx$

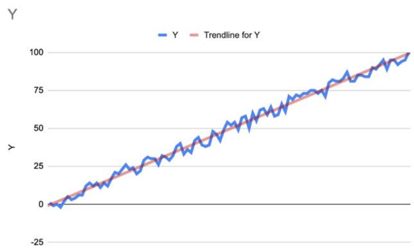
Полиномиальная $f(x) = a + b_1 x + b_2 x^2 + \ldots + b_k x^k$

Экспоненциальная $f(x) = ce^{a+bx}$

Логарифмическая $f(x) = a\log_b x + c$

Пример 1 (линейный тренд)





недостаточно выяснить тип функции, соответствующий линии тренда. Надо выяснить точные параметры функции. Как их узнать? Для каждого из упомянутого выше типа функций существуют подходящие математические методы, позволяющие определить эти параметры.

Так, например, для определения параметров линейного тренда можно воспользоваться, методом наименьших квадратов, который позволяет явно вычислить коэффициенты функции по формулам, приведенным на экране. Для определения параметров других трендов также существуют специальные математические методы. Но в рамках данной лекции мы не будем останавливаться на них подробно. Однако обратим внимание на то, что сами тренды и аналитические функции, которые им соответствуют, замечательно определяются простейшими инструментами.

$$y = ax + b,$$

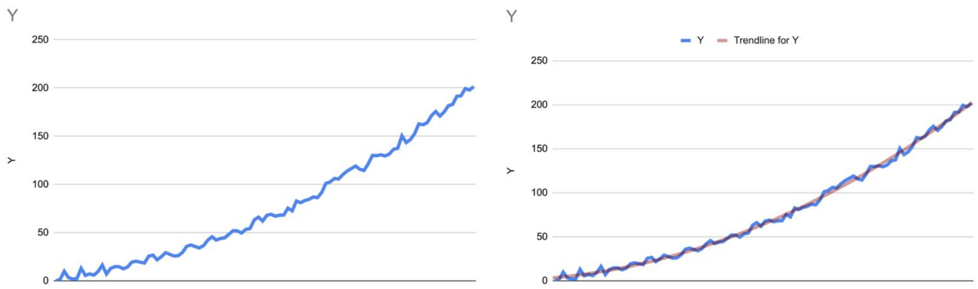
$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n},$$

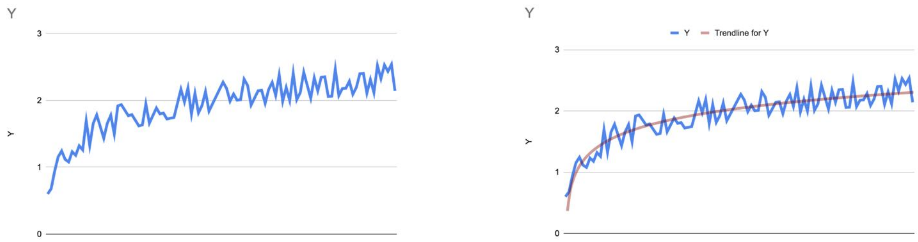
где n – количество измерений, y_i – элемент временного ряда, x_i – время.

Прежде, чем приступить к непосредственному построению трендов, обратим внимание еще на одно обстоятельство. Как подобрать подходящую линию тренда, если вариантов может быть несколько, и вы не уверены, какой из них лучше? Есть ли формальные критерии, позволяющие выбрать тип

Пример 2 (полиномиальный тренд)



Пример 3 (логарифмический тренд)

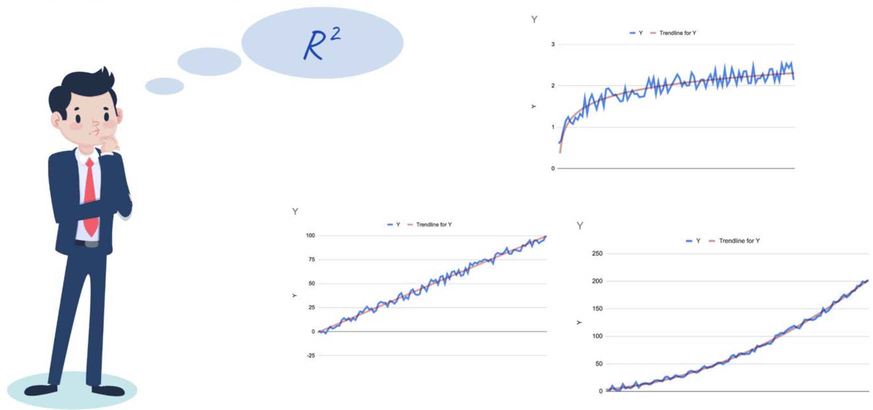


тренда? Оказывается, есть! Такой критерий называется – **коэффициент детерминации**. Обозначается – R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - f_i)^2}{\sum_{i=1}^{n} (y_i - y_{\text{avg}})^2},$$

где y_i – значения временного ряда в момент времени i, f_i – значение тренда в момент времени i, y_{avg} – среднее значение элементов временного ряда.

Как выбрать лучшую линию тренда?



Коэффициент детерминации может использоваться для оценки качества подобранного уравнения тренда. Он принимает значения от 0 до 1. Для приемлемых моделей тренда предполагается, что коэффициент детерминации должен быть хотя бы не меньше 0.5. Модели с коэффициентом детерминации выше 0.8 можно признать достаточно хорошими. Значение коэффициента детерминации $R^2 = 1$ означает функциональную зависимость между переменными (т.е. между исходным временным рядом и трендом). Точная формула для расчета коэффициента детерминации представлена на слайде.

Коэффициент детерминации реализован во многих инструментах, и мы будем его использовать для оценки качества построенных линий трендов.

1.5 Построение линии тренда

Кроме отображения уравнения тренда на диаграмме, в инструментах для обработки данных существуют функции, которые могут вычислять коэффициенты линейного тренда без прорисовки соответствующего графика. Это функции SLOPE(НАКЛОН) и INTERSECT(ОТРЕЗОК). Если уравнение для тренда задается как $y(x) = ax + b$, то функция SLOPE вычисляет коэффициент a, а INTERSECT – b. В некоторых обстоятельствах пользоваться этими функциями предпочтительнее, так как они выдают более точный результат, чем уравнение с округленными коэффициентами, которое мы видим на графике с трендом.

На рисунке приведен пример использования функции SLOPE в Google таблице: И функции INTERSECT в той же таблице. И продемонстрирован

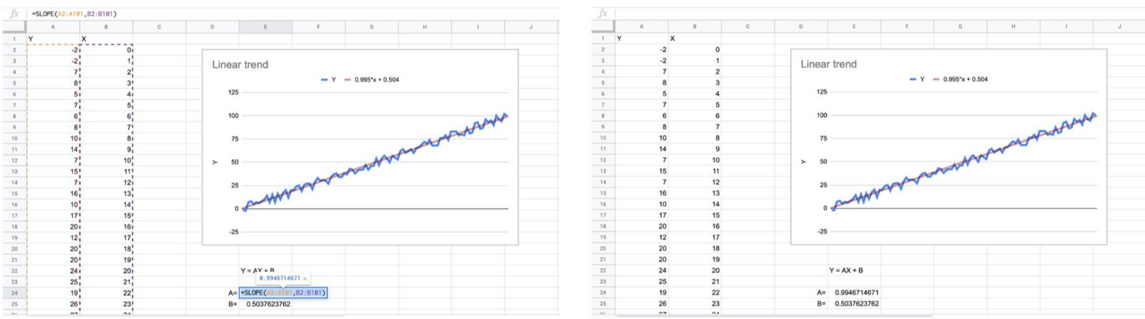


Рис. 1: Функция SLOPE

полученный результат. Сравните точность представления для коэффициента a в уравнении на графике и в результате применения функции. Итак, мы научились строить тренды и определять аналитические функции, которые им соответствуют. Как можно воспользоваться полученными знаниями в практических целях? Можно попробовать применить их для прогнозирования поведения несложных временных рядов, в которых, возможно, присутствуют шумы и тренды, но отсутствует сезонная составляющая. То есть мы можем смоделировать ряд на основе аналитической функции, соответствующей тренду. В нашем распоряжении есть подходящий ряд.

Это уже упоминавшийся ранее ряд с количеством автомобилей на 1000 человек в Центральном Федеральном Округе. Данные содержат сведения с

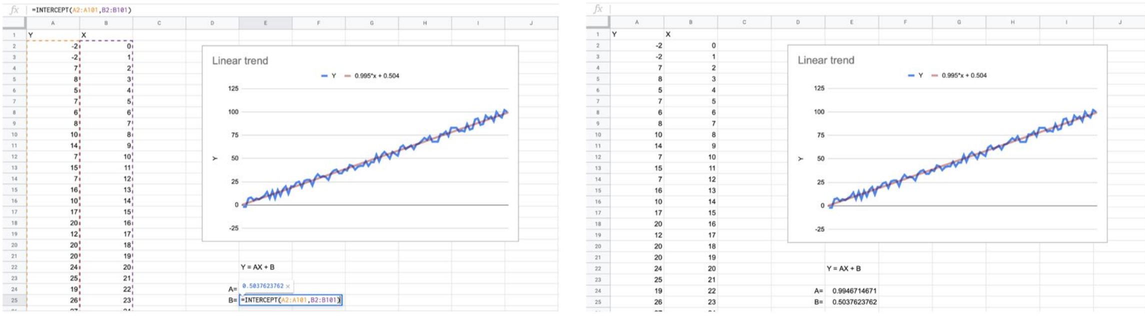
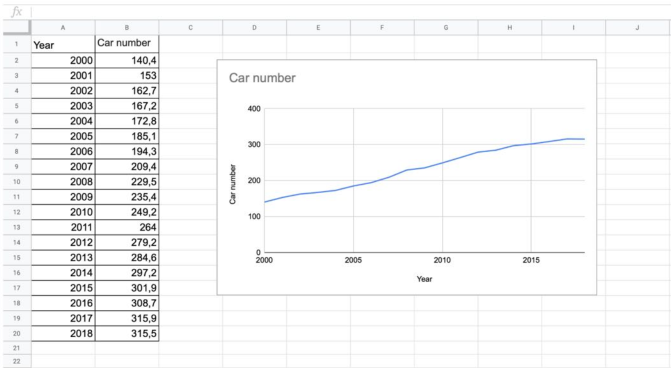


Рис. 2: Функция INTERSECT

2000 по 2018 годы. Построим линейный график, соответствующий этому ряду,

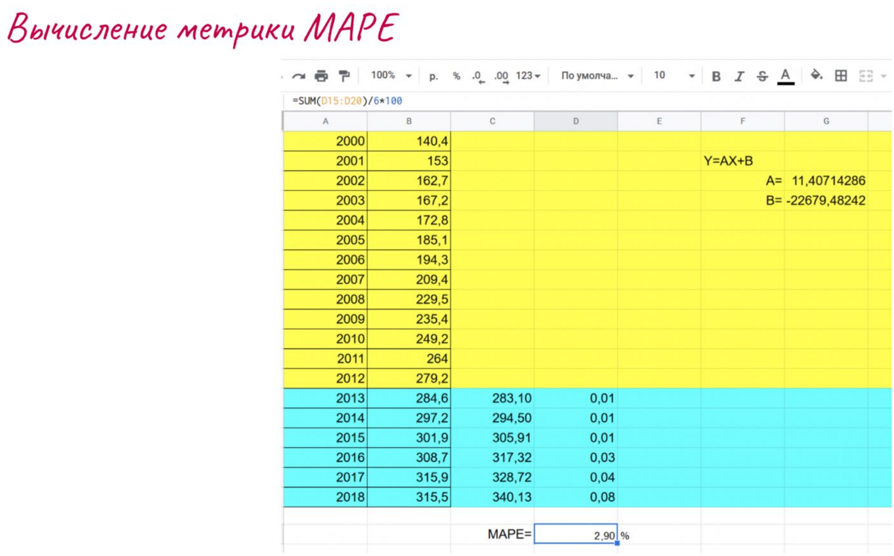
Количество автомобилей на 1000 человек в Центральном Федеральном Округе



и убедимся, что в нем нет сезонности. Значит, у нас есть основания моделировать ряд аналитической функцией вида $ax + b$.

Попробуем проверить, насколько эффективно моделируют точки ряда с помощью такой функции. Для этого разделим ряд на обучающую и тестовую выборки, определим параметры тренда на обучающей выборке, выполним прогноз на тестовой выборке, а потом оценим качество прогноза с помощью метрики MAPE.

Обучающую выборку выделим желтым цветом, тестовую – голубым. Определим параметры линейного тренда с помощью функций SLOPE и INTERSEPT. Вычислим прогнозные значения на тестовой выборке (т.е. с 2013 по 2018 гг.) по формуле $ax + b$. Результат вы можете видеть на экране в столбце Forecast. Отметим, что этот же результат в Google таблицах можно было бы получить и иным способом, т.к. существует встроенная функция FORECAST.LINEAR, которая выдает прогнозное значение на основе модели линейного тренда указанного временного ряда. Дальше можно оценить качество прогноза с помощью метрики MAPE. Для этого заведем в таблице специальный столбец с заголовком Error и вычислим в нем слагаемые числителя для метрики MAPE. После этого остается сложить полученные в



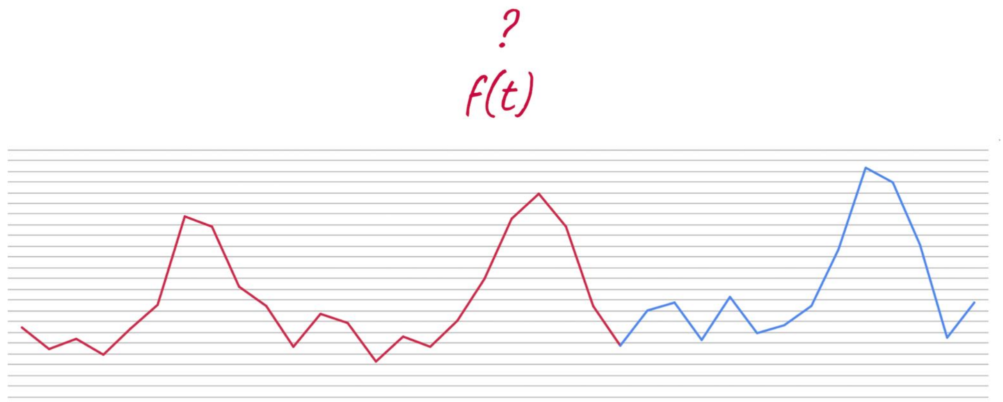
A	B	C	D	E	F	G
2000	140,4					
2001	153				Y=AX+B	
2002	162,7				A=	11,40714286
2003	167,2				B=	-22679,48242
2004	172,8					
2005	185,1					
2006	194,3					
2007	209,4					
2008	229,5					
2009	235,4					
2010	249,2					
2011	264					
2012	279,2					
2013	284,6	283,10	0,01			
2014	297,2	294,50	0,01			
2015	301,9	305,91	0,01			
2016	308,7	317,32	0,03			
2017	315,9	328,72	0,04			
2018	315,5	340,13	0,08			
		MAPE=	2,90	%		

столбце Error значения и разделить на количество элементов тестовой выборки. Результат 2.9% показывает среднее отклонение от реальных значений и выглядит достаточно убедительно. В следующем фрагменте мы рассмотрим, как моделируются ряды с сезонными составляющими.

1.6 Построение моделей для временных рядов с сезонными составляющими

Как определить сезонную составляющую временного ряда и как в дальнейшем ее использовать для моделирования ряда? Именно этот вопрос и является предметом обсуждения данного фрагмента. Многие временные ря-

ды экономического происхождения содержат сезонную составляющую в силу особенностей бизнес-процессов. Например, общая картина потребления электроэнергии периодически повторяется и зависит от месяца, посещаемость ресторанов также повторяется и зависит от дня недели, количество пассажиров в общественном транспорте зависит от времени суток и т.п. Эти процессы изначально позволяют определить продолжительность повторяющегося периода, который так важен для моделирования временных рядов с сезонной составляющей. В приведенных примерах длина периода равна 12 месяцам для потребления электроэнергии, 7 дням для ресторанов и 24 часам для общественного транспорта. А как следует поступать с рядами, с менее очевидными сезонными составляющими? Как определить длину повторяющегося периода и его сезонные компоненты? В общем случае для определения длины периода существуют достаточно сложные математические теории, однако мы упростим себе задачу и в рамках данного фрагмента будем предполагать, что длина периода нам точно известна, и, кроме того, исследуемый ряд уже избавлен от шумов. Итак, рассмотрим, как можно определять сезонные составляющие для периодического временного ряда и строить модели для его прогнозирования. Пусть исходные данные – временной ряд $f(t)$.



- Во-первых, рекомендуется построить аналитическое выражение для трендовой составляющей. Полученный ряд (т.е. тренд) для определенности назовем $d(t)$.

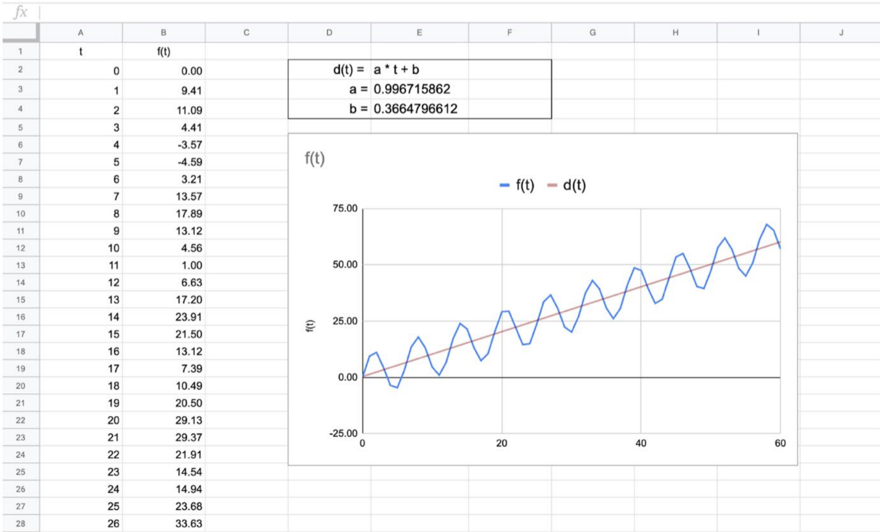
- Во-вторых, нужно выделить сезонную составляющую. Для этого потребуется вычесть трендовую составляющую из исходного временного ряда. То есть, построить ряд $r(t) = f(t) - d(t)$.

- И, наконец, можно строить модель для прогнозирования ряда на основе трендовой и сезонной составляющих.

Такая схема построения моделей является достаточно типичной для моделей с сезонной составляющей. Однако сами модели могут сильно отличаться в зависимости от особенностей сезонной составляющей. Продемонстрируем эти особенности на двух модельных примерах.

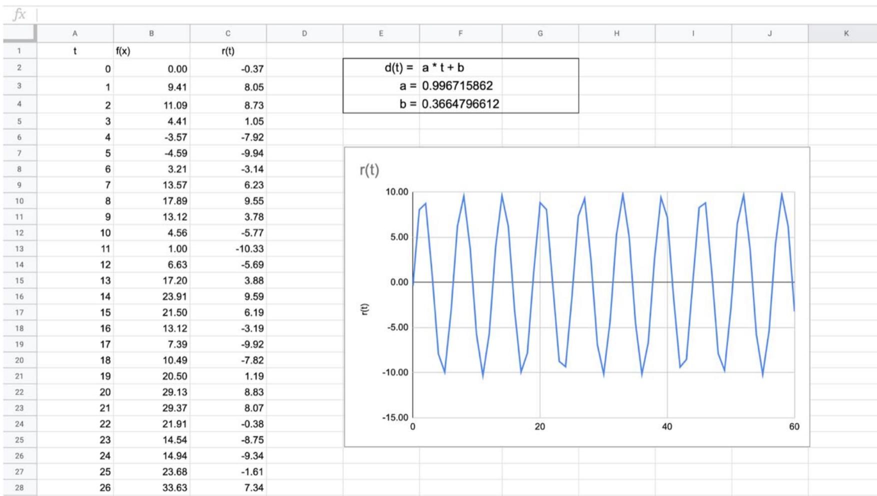
1.7 Ряд с сезонной составляющей постоянного размаха

На рисунке представлен исходный ряд с сезонной составляющей постоянного размаха. Кроме исходных данных построен график, соответствующий этим данным, линия тренда и параметры для линейного уравнения тренда. Периодичность такого ряда равна 6. Так как в нашем распоряжении есть па-



t	f(t)
0	0.00
1	9.41
2	11.09
3	4.41
4	-3.57
5	-4.59
6	3.21
7	13.57
8	17.89
9	13.12
10	4.56
11	1.00
12	6.63
13	17.20
14	23.91
15	21.50
16	13.12
17	7.39
18	10.49
19	20.50
20	29.13
21	29.37
22	21.91
23	14.54
24	14.94
25	23.68
26	33.63

раметры линейного тренда, нетрудно построить и сезонную составляющую ряда, которая получается простым вычитанием из исходного ряда трендовой составляющей. Результат в виде таблицы и графика вы можете видеть на рсиунке. Именно на этом графике отчетливо видно, почему мы назвали



t	f(x)	r(t)
0	0.00	-0.37
1	9.41	8.05
2	11.09	8.73
3	4.41	1.05
4	-3.57	-7.92
5	-4.59	-9.94
6	3.21	-3.14
7	13.57	6.23
8	17.89	9.55
9	13.12	3.78
10	4.56	-5.77
11	1.00	-10.33
12	6.63	-5.69
13	17.20	3.88
14	23.91	9.59
15	21.50	6.19
16	13.12	-3.19
17	7.39	-9.92
18	10.49	-7.82
19	20.50	1.19
20	29.13	8.83
21	29.37	8.07
22	21.91	-0.38
23	14.54	-8.75
24	14.94	-9.34
25	23.68	-1.61
26	33.63	7.34

этот ряд рядом с сезонной составляющей постоянного размаха. Локальные минимумы сезонной составляющей (точно так же, как и максимумы) лежат практически на одной линии параллельной оси времени.

Для ряда такого вида характерно, что сезонные компоненты для момента времени t такие же, как в момент времени $t - n$, где n – длина периода. А это дает возможность написать очень простую формулу для прогнозирования такого ряда на краткосрочный период (т.е. на 1 период вперед). Пусть последнее значение ряда определено в момент времени t. Тогда формула для краткосрочного прогнозирования в момент времени $t + k$ может состоять из двух слагаемых: значение трендовой составляющей в момент времени $t + k$ и сезонной составляющей в момент времени $t + k - n$.

Модель для краткосрочного прогнозирования ряда с сезонной компонентой постоянного размаха:

$$\mathsf{forecast}(t + k) = d(t + k) + r(t + k - n),$$

где t – момент времени, в который известна последняя точка исходного ряда, $t+k$ – момент, для которого осуществляется прогнозирование $k \leq n$, n – длина периода, $d(t + k)$ – тренд в точке $t + k$, $r(t + k - n)$ – сезонная составляющая в момент $t + k - n$.

Что должно измениться в модели для долгосрочного прогнозирования? Трудность состоит в том, что можно учитывать сезонную составляющую только на основе реально существующих значений ряда! То есть до момента времени t. Но и с этим можно справиться! Можно использовать в формуле последний известный период, даже если мы прогнозируем на несколько периодов вперед. Для этого в формуле, которую мы использовали для краткосрочного прогнозирования, достаточно в сезонной составляющей использовать не само значение k, которое указывает как далеко мы отошли от известных значений ряда, а его значение по модулю n, где n – длина периода.

Модель для долгосрочного прогнозирования временного ряда с сезонной компонентой постоянного размаха

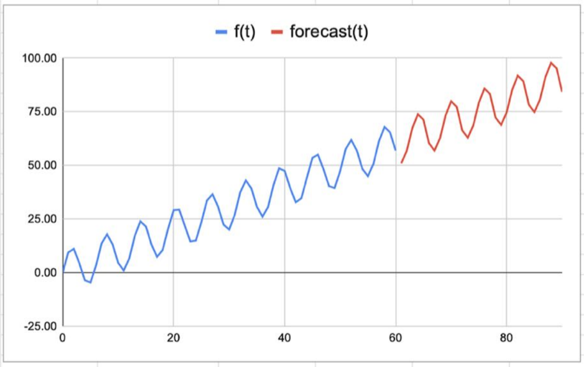
$$forecast(t + k) = d(t + k) + r(t + k \mod n - -n),$$

где t – момент времени, в который известна последняя точка исходного ряда, $t + k$ – момент, для которого осуществляется прогнозирование, n – длина периода, $d(t + k)$ – тренд в точке t+k, $r(t + k \mod n - n)$ – сезонная составляющая в момент $t + k - n$.

На рисунке можно видеть рассчитанный по этой формуле долгосрочный прогноз ряда (на 5 периодов вперед). Исходный ряд изображен синей линией, а прогнозные значения – красной. Результат выглядит достаточно убедительно. Однако, при этом надо понимать, что долгосрочный прогноз всегда хуже краткосрочного, так как чем ближе мы находимся к исходным данных, тем более основательны наши предположения о поведении ряда. Качество прогноза, разумеется, можно и нужно проверять с помощью метрик, которые мы уже обсуждали в данной лекции.

	t	f(t)	r(x)	forecast(t)
1	t	f(t)	r(x)	forecast(t)
2	0	0.00	-0.37	
3	1	9.41	8.05	
4	2	11.09	8.73	
5	3	4.41	1.05	
6	4	-3.57	-7.92	
7	5	-4.59	-9.94	
8	6	3.21	-3.14	
9	7	13.57	6.23	
10	8	17.89	9.55	
11	9	13.12	3.78	
12	10	4.56	-5.77	
13	11	1.00	-10.33	
14	12	6.63	-5.69	
15	13	17.20	3.88	
16	14	23.91	9.59	
17	15	21.50	6.19	
18	16	13.12	-3.19	
19	17	7.39	-9.92	
20	18	10.49	-7.82	
21	19	20.50	1.19	
22	20	29.13	8.83	
23	21	29.37	8.07	
24	22	21.91	-0.38	
25	23	14.54	-8.75	
26	24	14.94	-9.34	
27	25	23.68	-1.61	
28	26	33.63	7.34	
29	27	36.56	9.29	

d(t) =	a * t + b
a =	0.996715862
b =	0.3664796612
n =	6

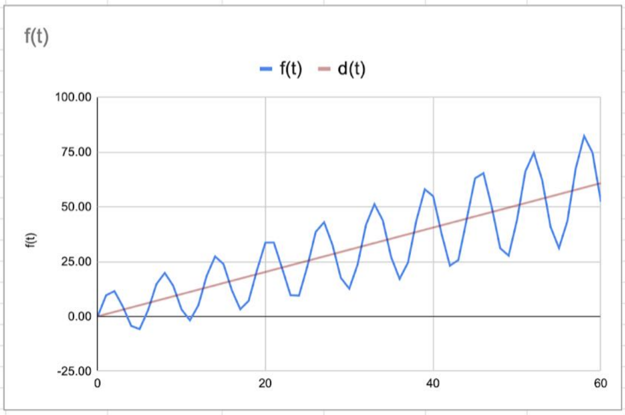


1.8 Ряд с сезонной компонентой растущего размаха

На рисунке представлен еще один временной ряд. Его отличительная особенность – наличие сезонной составляющей возрастающего размаха. Проведен первичный анализ данных. Построен график, соответствующий исходным данным, линия тренда и явно вычислены параметры для уравнения линейного тренда. Вычтем из исходного ряда трендовую составляющую. Ви-

	A	B
1		f(t)
2	0	0.00
3	1	9.63
4	2	11.55
5	3	4.52
6	4	-4.32
7	5	-5.79
8	6	2.79
9	7	14.72
10	8	19.87
11	9	14.05
12	10	3.20
13	11	-1.75
14	12	5.02
15	13	18.57
16	14	27.37
17	15	23.94
18	16	11.97
19	17	3.30
20	18	7.11
21	19	21.21
22	20	33.69
23	21	33.76
24	22	21.86
25	23	9.67
26	24	9.51
27	25	22.85
28	26	38.58
29	27	43.02

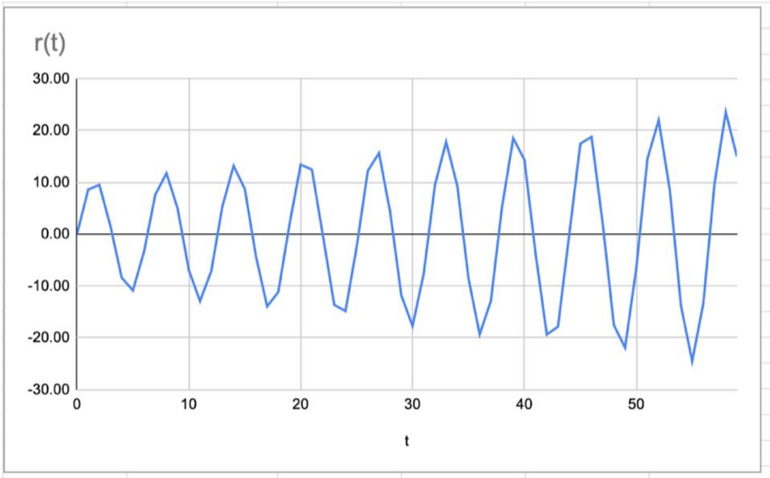
d(t) =	a * t + b
a =	1.01338677
b =	0.04186700655



зуализируем полученный результат. После визуализации растущий размах сезонных компонент ряда становится отчетливо виден.

На графике хорошо видно, что размах меняется со временем, причем достаточно равномерно. Модель краткосрочного прогнозирования, которая была использована для ряда с постоянным размахом, может быть использована и в этом случае, так как она все равно будет учитывать динамику изменения размаха, но с некоторым отставанием (будет запаздывать на один период).

Поэтому лучше все-таки формулу для краткосрочного прогнозирования второго ряда модифицировать и написать так, чтобы в ней учитывалось приращение сезонных компонент за 1 период. Более всего нас интересует поведение именно сезонных компонент, т.к. с трендовой составляющей мы можем справиться при помощи аналитического формулы. Как мы уже упоминали ранее, у нас есть два возможных подхода: аддитивный и мультипликативный. В аддитивном случае мы должны указать, на какое приращение изме-



няется компонента, а в мультипликативном – во сколько раз. Выберем второй, мультипликативный вариант. Остается совсем немного – выяснить, так называемый, коэффициент сезонности. Это та величина, на которую будут умножаться значения сезонных компонент последнего известного периода. Есть разные способы для вычисления такого коэффициента, но, пожалуй, простейший – сравнить размах последнего и предпоследнего периодов. Соотношение этих размахов и можно определить как коэффициент сезонности. Как вычислить размах в последнем и предпоследнем периодах? Он может быть определен как разница между максимальным и минимальным значениями точек периода. На экране приведен пример вычисления коэффициента сезонности в предположении, что длина периода – 6, а последняя известная точка сезонной составляющей содержится в ячейке C62.

После того, как коэффициент сезонности определен, нет никаких проблем в получении модели для краткосрочного прогнозирования. Формула

```
fx =(MAX(C57:C62)-MIN(C57:C62))/(MAX(C51:C56)-MIN(C51:C56))
```

	A	B	C	D	E	F	G
1	t	f(t)	r(t)	forecast(t)	d(t) =	a * t + b	
2	0	0.00	-0.04		a =	1.01338677	
3	1	9.63	8.57		b =	0.041867006	
4	2	11.55	9.48		n =	6	
5	3	4.52	1.44		t =	60	
6	4	-4.32	-8.42		i_season =	1.09	
7	5	-5.79	-10.90				

приведена на экране. Это практически тоже, что и для ряда с постоянным сезонным размахом, только сезонная компонента из последнего известного периода умножена на коэффициент сезонности.

Модель для краткосрочного прогнозирования ряда с сезонной компонентой возрастающего размаха t – момент времени, в который известна последняя точка исходного ряда, $t+k$ – момент, для которого осуществляется прогнозирование $k \leq n$, n – длина периода, $d(t+k)$ – тренд в точке $t+k$, $r(t+k-n)$ – сезонная составляющая в момент $t+k-n$, i_{season} – коэффициент сезонности.

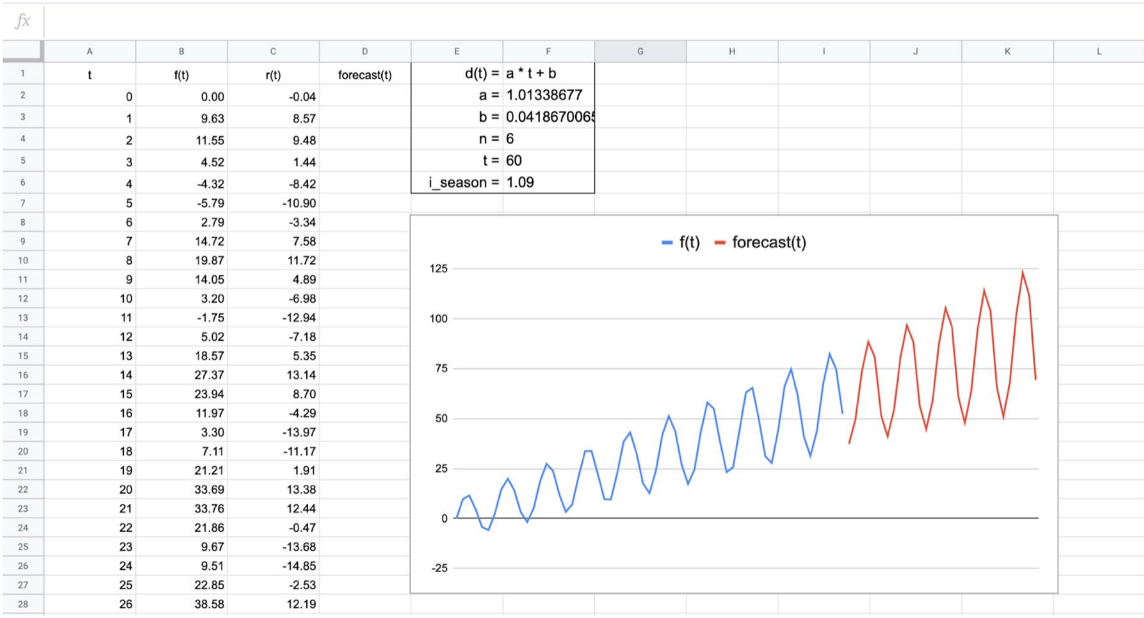
Что меняется в случае долгосрочного прогнозирования? Необходимо учитывать то приращение, которое увеличивает размах сезонных компонент, причем учитывать его необходимо на несколько периодов вперед. Если коэффициент сезонности хотя бы приблизительно можно считать одинаковым для всех периодов, то для построения модели можно использовать следующую формулу, которая приведена на экране. Обратите внимание на выражение k div $n+1$ (к деленное нацело на n плюс 1). Оно отражает относительный номер периода, для которого ведется прогноз. При этом нумерация таких периодов начинается после момента времени t.

Модель для долгосрочного прогнозирования с сезонной компонентой возрастающего размаха

$$forecast(t+k) = d(t+k) + r(t+k \mod n - n) \cdot {i_{season}}^{k \text{ div } n+1},$$

где t – момент времени, в который известна последняя точка исходного ряда, $t+k$ – момент, для которого осуществляется прогнозирование, n – длина периода, $d(t+k)$ – тренд в точке $t+k$ $r(t+k \mod n-n)$ – сезонная составляющая в момент $t+k-n$, i_{season} – коэффициент сезонности. На следующем рисунке можно видеть рассчитанный по этой формуле долгосрочный прогноз ряда с возрастающим размахом. Исходный ряд на графике изображен синей линией, а прогнозные значения – красной.

Разумеется, в случае если мы имеем дело с реальными рядами, нужно оценить качество прогноза и убедиться, что мы выбрали лучшую модель из возможных. Но мы не будем это делать в данном фрагменте, так как надеемся, что достаточно подробно говорили об оценке качества модели в предыдущем фрагменте лекции. Итак, мы рассмотрели несколько простей-



	A	B	C	D	E	F
1	t	f(t)	r(t)	forecast(t)	d(t) =	a * t + b
2	0	0.00	-0.04		a =	1.01338677
3	1	9.63	8.57		b =	0.0418670065
4	2	11.55	9.48		n =	6
5	3	4.52	1.44		t =	60
6	4	-4.32	-8.42		i_season =	1.09
7	5	-5.79	-10.90			
8	6	2.79	-3.34			
9	7	14.72	7.58			
10	8	19.87	11.72			
11	9	14.05	4.89			
12	10	3.20	-6.98			
13	11	-1.75	-12.94			
14	12	5.02	-7.18			
15	13	18.57	5.35			
16	14	27.37	13.14			
17	15	23.94	8.70			
18	16	11.97	-4.29			
19	17	3.30	-13.97			
20	18	7.11	-11.17			
21	19	21.21	1.91			
22	20	33.69	13.38			
23	21	33.76	12.44			
24	22	21.86	-0.47			
25	23	9.67	-13.68			
26	24	9.51	-14.85			
27	25	22.85	-2.53			
28	26	38.58	12.19			

ших способов моделирования значения рядов с сезонной компонентой. Разумеется, точность прогноза уменьшается по мере удаления горизонта прогноза от исторических данных. Тем не менее, даже эти модели могут оказаться полезными при прогнозировании поведения реальных экономических рядов.